Probability theory

Peter Pfaffelhuber

Version: June 26, 2024

Prelude

These are the notes of a lecture which I gave at the University of Freiburg. After some elementary probability and measure theory, this course introduces some main concepts in (measure theoretic) probability theory. As a prerequisite, for measure theoretic contents, we refer to my manuscript *Measure theory for probabilists*. In particular, references to Chapters 1–5 are references to this manuscript.

The following books have guided me as references for the purpose of this manuscript.

- Durrett, Rick. Probability: Theory and Examples, Cambridge Series in Statistical and Probabilistic Mathematics, 2019
- Kallenberg, Olaf. Foundations of Modern Probability Theory. Springer, third edition, 2021
- Klenke, Achim. Probability theory. A comprehensive course. Springer, 2014

Throughout the manuscript, we will use a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ (recall from Definition 2.1). The integral with respect to $\mathbf{P}$ is denoted by $\mathbf{E}[\cdot] := \mathbf{P}[\cdot]$ (recall from Chapter 3). Further, we abbreviate $\mathcal{L}^p := \mathcal{L}^p(\mathbf{P})$ if this does not lead to confusion (recall these spaces from Chapter 4.

Our aim in the present course is to provide the most important probabilistic statements available. Fundamental to this is the concept of the random variable, which we will examine in Chapter 6 (see also Definition 3.3). We will often consider the case of E-valued random variables, where E is a Polish space (see Appendix A in the lecture notes on measure theory). The most influential theorems in probability theory are the strong law of large numbers (LLN, Theorem 8.21) and the central limit theorem (CLT, Theorem10.8). These two theorems are limit statements for random variables, and it is important to note that the type of convergence in both theorems is fundamentally different. While the strong LLN describes an almost sure convergence (refer to Remark 2.14), the CLT is a statement about convergence in distribution (i.e. about the weak convergence of the distributions of the random variables; see Chapter 9). Consequently, one of the tasks will be to understand the relationships between different types of convergence (see Chapter 7 and 9).

The present english version of this manuscript was written based on the German version with the help of DeepL.

Contents

6 Random variables

We usually use real-valued random variables $X : \Omega \to \mathbb{R}$ (i.e. Borel-measurable functions, i.e. random variables with values in $\mathbb{R}$, measurable with respect to the Borel σ-algebra in $\mathbb{R}$; recall from Definition 1.7)). We will now recall several concepts from measure theory about random variables and which we will need directly in the following. We will mainly deal with connecting the lecture to measure theory on one side, and *Basic Probability* on the other side.

6.1 Repetition

Recall that we assume throughout that a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is given. In Definition 3.5, we already gave several notions with a relationship to random variables, which we recall for completeness.

Remark 6.1 (Random variables and their distribution)**.** *Let $(\Omega', \mathcal{F}')$ be a measurable space.*

1. *Every $\mathcal{F}/\mathcal{F}'$-measurable function X is called (Ω'-valued)* random variable*. If $(\Omega', \mathcal{F}') = (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$, it is called* real-valued*. The σ-algebra $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{F}')\}$ is* the σ-algebra generated by X *(see Definition 3.3).*

2. *The probability measure $X_*\mathbf{P}$ on $\mathcal{F}'$ (i.e. the image measure of $\mathbf{P}$ under X; see Section 2.5) is called* distribution of X*. Furthermore, if Y is a random variable and $X_*\mathbf{P} = Y_*\mathbf{P}$ (i.e. $\mathbf{P}(X \in A') = \mathbf{P}(Y \in A')$ for all $A' \in \mathcal{F}'$), then X and Y are identically distributed and we write $X \stackrel{d}{=} Y$. However, this notation should be used with caution, as the equality $X \stackrel{d}{=} Y$ cannot be achieved by equivalence transformations to other statements. (For example $X \stackrel{d}{=} Y$ does in general not imply $X - Y \stackrel{d}{=} 0$).*

3. *For a family $(X_i)_{i\in I}$ of random variables, their* joint distribution *is given by $((X_i)_{i\in I})_*\mathbf{P}$. (This is the image measure under the mapping $(X_i)_{i\in I} : \omega \mapsto (X_i(\omega))_{i\in I}$).*

4. *We will use the following notation:* If X is a random variable with distribution $N(\mu, \sigma^2)$. *(This means means that $X : \Omega \mapsto \mathbb{R}$ is a measurable mapping and $X_*\mathbf{P} = \mu_{N(\mu,\sigma^2)}$; see example 2.22.) Then, we write $X \sim N(\mu, \sigma^2)$. Here, read '$\sim$' as* has distribution*.*

5. *Let λ be another measure on $\mathcal{F}$ and $f : \Omega \to \mathbb{R}$ with $f \geq 0$ almost everywhere and $\lambda[f] = 1$. Then, X has the density f with respect to μ if and only if $X_*\mathbf{P} = f \cdot \lambda$ (see Definition 4.13). Then, for $A \in \mathcal{F}$,*

$$\mathbf{P}(X \in A) = \mu[f, A].$$

In this case, for $g : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ that (see Lemma 4.14),

$$\mathbf{E}[g(X)] = (X_*\mathbf{P})[g] = (f \cdot \mu)[g] = \mu[fg],$$

if the right-hand side exists.

6. *Monotonicity and linearity of the integral means for random variables $X, Y \in \mathcal{L}^1$ and $a, b \in \mathbb{R}$:*

$$X \leq Y \text{ almost surely } \implies \mathbf{E}[X] \leq \mathbf{E}[Y],$$
$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y].$$

Furthermore, according to Proposition 3.21,

$$\mathbf{E}[X] < \infty \Longrightarrow \mathbf{P}(X < \infty) = 1.$$

Although we already have a σ-algebra $\mathcal{F}$, in further sections, especially in the introduction of the conditional expectation in Chapter 11, the σ-algebra generated by X will play a special role. Simply put, a real-valued random variable Y is $\sigma(X)$-measurable if and only if $Y = \varphi(X)$ for a Borel-measurable mapping φ. In other words, this means that the value of $Y(\omega)$ is known if you know $X(\omega)$, although you do not know what value ω has assumed. See also Exercise 3.38.

Lemma 6.2 (Measurability with respect to $\boldsymbol{\sigma(X)}$)**.** *Let $(\Omega', \mathcal{F}')$ be a measurable space and X a random variable with values in Ω', and $Z : \Omega \to \overline{\mathbb{R}}$. The, Z is $\sigma(X)$-measurable if and only if there is a $\mathcal{F}'/\mathcal{B}(\overline{\mathbb{R}})$-measurable mapping $\varphi : \Omega' \to \overline{\mathbb{R}}$ with $\varphi \circ X = Z$.*

Proof. '$\Longleftarrow$': clear
'$\Longrightarrow$': It suffices to consider the case $Z \geq 0$; otherwise, we write $Z = Z^+ - Z^-$. First, let $Z = 1_A$ for $A \in \sigma(X)$. Then there is an $A' \in \mathcal{F}'$ with $X^{-1}(A') = A$, i.e. $Z = 1_{X^{-1}(A')} = 1_{A'} \circ X$, i.e. $\varphi = 1_{A'}$ fulfills the statement. Due to linearity, the statement is also true for simple functions, i.e. finite linear combinations of indicator functions. In the general case, there are simple functions $Z_1, Z_2, \cdots \geq 0$ with $Z_n \uparrow Z$. In addition, there are $\mathcal{F}'$-measurable functions φ_n with $Z_n = \varphi_n \circ X$. Then $\varphi = \sup_n \varphi_n$ is again $\mathcal{F}'$-measurable and, since $Z \geq 0$, and

$$\varphi \circ X = (\sup_n \varphi_n) \circ X = \sup_n (\varphi_n \circ X) = \sup_n Z_n = Z.$$

□

We now briefly repeat the convergence theorems for integrals in the context of of random variables.

Proposition 6.3 (Integral convergence theorems)**.** *Let $X, X_1, X_2, \ldots$ be real-valued random variables.*

1. Lemma of Fatou, Theorem 3.27: *If $X_1, X_2, ... \geq 0$, then*

$$\liminf_{n\to\infty} \mathbf{E}[X_n] \geq \mathbf{E}[\liminf_{n\to\infty} X_n].$$

2. Monotone convergence, Theorem 3.26: *If $X_1, X_2, \cdots \in \mathcal{L}^1$ and $X_n \uparrow X$ almost surely, then*

$$\lim_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[X],$$

where both sides can take the value ∞.

3. Dominated convergence, Theorem 3.28: *Let $X_n \xrightarrow{n\to\infty} X$ almost surely, and Y another real-valued random variable with $|X_1|, |X_2|, \cdots \leq Y$ almost surely, and $\mathbf{E}[Y] < \infty$. Then,*

$$\mathbf{E}[X_n] \xrightarrow{n\to\infty} \mathbf{E}[X].$$

We now collect (and re-prove) already known inequalities. They often help to estimate probabilities or expected values. Most of the inequalities are already known from the lecture on *Basic Probability.*

Proposition 6.4 (Markov and Chebyshev inequality)**.** *1. Let X be a random variable with values in $\overline{\mathbb{R}}_+$ and $x \in \mathbb{R}_+$. Then the Markov inequality holds, i.e.,*

$$\mathbf{P}(X \geq x) \leq \frac{\mathbf{E}[X]}{x}.$$

2. If X is a real-valued random variable and $p, x \in \mathbb{R}_+$, then the Chebyshev inequality holds, i.e.

$$\mathbf{P}(|X| \geq x) \leq \frac{\mathbf{E}[|X|^p]}{x^p}.$$

Proof. 1. Since X is non-negative, we find $x \cdot 1_{X \geq x} \leq X$. Thus,

$$x \cdot \mathbf{P}(X \geq x) = \mathbf{E}[x \cdot 1_{X \geq x}] \leq \mathbf{E}[X],$$

and the inequality follows. The inequality in 2. follows from 1. by

$$\mathbf{P}(|X| \geq x) = \mathbf{P}(|X|^p \geq x^p) \leq \frac{\mathbf{E}[|X|^p]}{x^p}.$$

□

The next statement was already given in Proposition 4.2.

Proposition 6.5 (Minkowski and Hölder inequalities)**.** *Let X, Y be real-valued random variables.*

1. If $0 < p, q, r \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$. Then,

$$\mathbf{E}[|XY|^r]^{1/r} \leq \mathbf{E}[|X|^p]^{1/p} \cdot \mathbf{E}[|Y|^q]^{1/q} \qquad \textit{(Hölder inequality)} \tag{6.1}$$

Especially, if $p = q = 2$

$$\mathbf{E}[|XY|] \leq \mathbf{E}[|X|^2]^{1/2} \cdot \mathbf{E}[|Y|^2]^{1/2}. \qquad \textit{(Cauchy-Schwartz inequality)} \tag{6.2}$$

2. For $1 \leq p \leq \infty$,

$$\mathbf{E}[|X+Y|^p]^{1/p} \leq \mathbf{E}[|X|^p]^{1/p} + \mathbf{E}[|Y|^p]^{1/p}, \quad 1 \leq p \leq \infty \qquad \textit{(Minkowski inequality)} \tag{6.3}$$

Proposition 6.6 (Jensen's inequality)**.** *Let $I \subseteq \mathbb{R}$ be an open interval and $X \in \mathcal{L}^1$ with values in I and $\varphi : I \to \mathbb{R}$ convex.*[1] *Then,*

$$\mathbf{E}[\varphi(X)] \geq \varphi(\mathbf{E}[X]).$$

Proof. Since φ is convex, φ is continuous and

$$t \mapsto \frac{\varphi(tx + (1-t)y) - \varphi(y)}{t(x-y)}$$

[1] A mapping $\varphi : I \to \mathbb{R}$ is convex if $\varphi(tx + (1-t)y) \leq t\varphi(x) + (1-t)\varphi(y)$ for all $0 \leq t \leq 1$ and $x, y \in I$.

is monotonically decreasing for $y \leq x$. In particular, for $y \in I$ there exists

$$\lambda(y) := \lim_{x \downarrow y} \frac{\varphi(x) - \varphi(y)}{x - y} = \lim_{t \downarrow 0} \frac{\varphi(tx + (1-t)y) - \varphi(y)}{t(x-y)} \tag{6.4}$$

and

$$\frac{\varphi(x) - \varphi(y)}{x - y} \geq \lambda(y) \quad \Longrightarrow \quad \varphi(y) + \lambda(y)(x-y) \leq \varphi(x) \tag{6.5}$$

for all $x \in I$. (For $y > x$ one argues analogously as above).

Note, since I is an interval, we have $\mathbf{E}[X] \in I$. According to (6.5), for $x \in I$ with $y = \mathbf{E}[X]$

$$\varphi(x) \geq \varphi(\mathbf{E}[X]) + \lambda(\mathbf{E}[X])(x - \mathbf{E}[X])$$

and thus

$$\mathbf{E}[\varphi(X)] \geq \varphi(\mathbf{E}[X]) + \lambda(\mathbf{E}[X])\mathbf{E}[X - \mathbf{E}[X]] = \varphi\big(\mathbf{E}[X]\big). \qquad \square$$

Jensen's inequality can be used to show, for example, that $\mathcal{L}^q \subseteq \mathcal{L}^p$ for $p \leq q$. Alternatively, you can read this property from Proposition 4.3.

Lemma 6.7 ($\mathcal{L}^q$ and $\mathcal{L}^p$)**.** *Let $q > 0$ and $X \in \mathcal{L}^q$ be a reel-valued random variable. Then, for $p \leq q$*

$$\mathbf{E}[|X|^p] \leq \mathbf{E}[|X^q|]^{p/q}.$$

In particular, $\mathcal{L}^q \subseteq \mathcal{L}^p$.

Proof. The mapping $y \mapsto y^{p/q}$ is concave on $\mathbb{R}_+$, so with Jensen's inequality,

$$\mathbf{E}[|X|^p] = \mathbf{E}[(|X|^q)^{p/q}] \leq \mathbf{E}[|X|^q]^{p/q}. \qquad \square$$

6.2 Moments

From the lecture *Basic Probability*, we know terms such as expected value, variance and covariance. When we repeat them now, you will see that all calculation rules that you learned also apply in the measure theoretic sense. The only difference in the formulation is that $\mathbf{E}[\cdot]$ is the integral with respect to a probability measure.

Definition 6.8 (Moments)**.** *Let X, Y be real-valued random variables. Then, if it exists, $\mathbf{E}[X]$ is the* expected value *of X. Furthermore, if it exists, the* variance *of X is given by*

$$\mathbf{V}[X] := \mathbf{E}[(X - \mathbf{E}[X])^2].$$

If it exists, the covariance *of X and Y is given by*

$$\mathbf{COV}[X, Y] := \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])].$$

If $\mathbf{COV}[X, Y] = 0$, we say that X and Y are uncorrelated. Furthermore, $\mathbf{E}[X^p]$ for $p > 0$ is the p-th moment of X and $\mathbf{E}[(X - \mathbf{E}[X])^p]$ is the centered p-th moment of X.

We recall a few properties here.

Proposition 6.9 (Properties of the second moments)**.** *Let $X, Y \in \mathcal{L}^2$ be real-valued random variables. Then, $\mathbf{V}[X], \mathbf{V}[Y], |\mathbf{COV}[X,Y]| < \infty$ and*

$$\begin{aligned}\mathbf{V}[X] &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2,\\ \mathbf{COV}[X,Y] &= \mathbf{E}[XY] - \mathbf{E}[X]\cdot\mathbf{E}[Y].\end{aligned}$$

The Cauchy-Schwartz inequality holds, i.e.

$$\mathbf{COV}[X,Y]^2 \leq \mathbf{V}[X]\cdot\mathbf{V}[Y].$$

If $X_1, \ldots, X_n \in \mathcal{L}^2$, the identity of Bienamyé holds, i.e.

$$\mathbf{V}\Big[\sum_{k=1}^n X_k\Big] = \sum_{k=1}^n \mathbf{V}[X_k] + 2\sum_{1\leq k<l\leq n} \mathbf{COV}[X_k, X_l].$$

Proof. For the first statement, since $\mathbf{V}[X] = \mathbf{COV}[X,X]$ by definition, it is sufficient to show the second equation. This equation follows from the linearity of the expected value by means of

$$\begin{aligned}\mathbf{COV}[X,Y] &= \mathbf{E}[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])]\\ &= \mathbf{E}[XY] - \mathbf{E}[\mathbf{E}[X]Y] - \mathbf{E}[X\mathbf{E}[Y]] + \mathbf{E}[X]\mathbf{E}[Y]\\ &= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].\end{aligned}$$

The Cauchy-Schwartz inequality follows by applying Proposition 6.5 using the random variables $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$. In particular, $|\mathbf{COV}[X,Y]| < \infty$. For the equation of Bienamyé let wlog $\mathbf{E}[X_k] = 0$, $k = 1, \ldots, n$ (otherwise w use the random variables $X_k - \mathbf{E}[X_k]$). Then,

$$\begin{aligned}\mathbf{V}\Big[\sum_{k=1}^n X_k\Big] = \mathbf{E}\Big[\Big(\sum_{k=1}^n X_k\Big)^2\Big] &= \sum_{k=1}^n\sum_{l=1}^n \mathbf{E}[X_kX_l] = \sum_{k=1}^n \mathbf{E}[X_k^2] + 2\sum_{1\leq k<l\leq n}\mathbf{E}[X_kX_l]\\ &= \sum_{k=1}^n \mathbf{V}[X_k] + 2\sum_{1\leq k<l\leq n}\mathbf{COV}[X_kX_l].\end{aligned}$$

□

Proposition 6.10 (Moments of non-negative random variables)**.** *Let X be a random variable with values in $\mathbb{R}_+$. Then,*

$$\mathbf{E}[X^p] = p\int_0^\infty \mathbf{P}(X>t)t^{p-1}dt = p\int_0^\infty \mathbf{P}(X\geq t)t^{p-1}dt.$$

Proof. Note that both, $\mathbf{E}[.]$ and $\int\cdot dt$ are integrals. We use Fubini's theorem in order to be able to change the order of integration,

$$\mathbf{E}[X^p] = p\mathbf{E}\Big[\int_0^X t^{p-1}dt\Big] = p\int_0^\infty \mathbf{E}\Big[1_{X>t}t^{p-1}\Big]dt = p\int_0^\infty \mathbf{P}(X>t)t^{p-1}dt.$$

The proof of the second equation is analogous. □

6.3 Characteristic functions and Laplace transforms

We now introduce expected values of certain functions of random variables. This results in the characteristic function (of the distribution of real-valued random variables) and the Laplace transform (of the distribution of of non-negative random variables). (Both are covered in some courses on Basic Probability Theory, but not in all.) These functions are useful since they allow the calculation of moments (see Proposition 6.14). In addition, later in Proposition 9.25, we will show that these functions uniquely determine the underlying measure.

Definition 6.11 (Characteristic function and Laplace transform)**.**

1. *The* characteristic function *of an $\mathbb{R}^d$-valued random variable X is given by*

$$\psi_X := \psi_{X_*\mathbf{P}} := \begin{cases} \mathbb{R}^d & \to \mathbb{C}, \\ t & \mapsto \mathbf{E}[e^{itX}] := \mathbf{E}[\cos(tX)] + i\mathbf{E}[\sin(tX)], \end{cases}$$

   where $tx := \langle t, x\rangle$ is the scalar product in $\mathbb{R}^d$.

2. *The* Laplace transform *of X is given by*

$$\mathscr{L}_X := \mathscr{L}_{X_*\mathbf{P}} := \begin{cases} \mathbb{R}^d & \to \mathbb{R}, \\ t & \mapsto \mathbf{E}[e^{-tX}], \end{cases}$$

   provided the integral on the right-hand side exists. (The Laplace transform is used frequently for probability measures on $\mathbb{R}^d_+$.)

Proposition 6.12 (Properties of characteristic functions)**.** *Let X, Y be random variables with values in $\mathbb{R}^d$ and characteristic functions ψ_X, ψ_Y. Then,*

1. *$|\psi_X(t)| \le 1$ for each $t \in \mathbb{R}^d$ and $\psi_X(0) = 1$.*
2. *ψ_X is uniformly continuous.*
3. *$\psi_{aX+b}(t) = \psi_X(at)e^{ibt}$ for all $a \in \mathbb{R}, b \in \mathbb{R}^d$.*

Proof. 1. is clear. For uniform continuity, we have the bound

$$\begin{aligned} |e^{ihx} - 1| &= \sqrt{|\cos(hx) + i\sin(hx) - 1|^2} = \sqrt{(\cos(hx) - 1)^2 + \sin(hx)^2} \\ &= \sqrt{2(1 - \cos(hx))} = 2|\sin(hx/2)| \le |hx| \wedge 2. \end{aligned}$$

From this, 2. follows because of

$$\begin{aligned} \sup_{t\in\mathbb{R}^d} |\psi_X(t+h) - \psi_X(t)| &= \sup_{t\in\mathbb{R}^d} |\mathbf{E}[e^{i(t+h)X} - e^{itX}]| = \sup_{t\in\mathbb{R}^d} |\mathbf{E}[e^{itX}(e^{ihX} - 1)]| \\ &\le \mathbf{E}[|e^{ihX} - 1|] \le \mathbf{E}[|hX| \wedge 2] \xrightarrow{h\to 0} 0. \end{aligned}$$

For 3. we calculate using linearity

$$\mathbf{E}[e^{it(aX+b)}] = e^{itb}\mathbf{E}[e^{i(at)X}] = e^{itb}\psi_X(at).$$

□

Example 6.13 (Examples of characteristic functions functions). *1. The characteristic function of $X \sim B(n,p)$ is given by*

$$\psi_{B(n,p)}(t) = (1-p+pe^{it})^n.$$

Indeed: By definition,

$$\mathbf{E}[e^{itX}] = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} e^{itk} = (1-p+pe^{it})^n.$$

2. The characteristic function of $X \sim Poi(\gamma)$ is given by

$$\psi_{Poi(\gamma)} = e^{\gamma(e^{it}-1)}.$$

Indeed: We calculate

$$\psi_{Poi(\gamma)} = e^{-\gamma} \sum_{n=0}^{\infty} \frac{\gamma^n e^{itn}}{n!} = e^{\gamma(e^{it}-1)}.$$

3. The characteristic function of $X \sim N(\mu, \sigma^2)$ is given by

$$\psi_{N(\mu,\sigma^2)}(t) = e^{it\mu} e^{-\sigma^2 t^2/2}.$$

Indeed: We use that $X \sim \sigma Z + \mu$ for $Z \sim N(0,1)$. According to Proposition 6.12.2, it is sufficient to compute this assertion for $\mu = 0, \sigma^2 = 1$. For this case, using partial integration,

$$\frac{d}{dt}\psi_{N(0,1)}(t) = \frac{i}{\sqrt{2\pi}} \int x e^{-x^2/2} e^{itx} dx = -\frac{i}{\sqrt{2\pi}} \int e^{-x^2/2} ite^{itx} dx = -t\psi_{N(0,1)}(t).$$

This differential equation with $\psi_{N(0,1)}(0) = 1$ has the unique solution $\psi_{N(0,1)}(t) = e^{-t^2/2}$.

4. The Laplace transform of $X \sim \exp(\gamma)$ is given by

$$\mathscr{L}_{\exp(\gamma)}(t) = \frac{\gamma}{\gamma + t}.$$

Indeed: A straight-forward calculation reveals

$$\mathbf{E}[e^{-tX}] = \int_0^{\infty} \gamma e^{-\gamma x} e^{-tx} dx = \frac{\gamma}{\gamma + t}.$$

Characteristic functions and Laplace transforms are a simple tool to calculate the moments of random variables.

Proposition 6.14 (Characteristic function and moments). *Let X be a real-valued random variable.*

1. If $X \in \mathcal{L}^p$, then ψ_X is p-times continuously differentiable and for $k = 0, \ldots, p$,

$$\psi_X^{(k)}(t) = \mathbf{E}[(iX)^k e^{itX}].$$

In particular, $\psi_X^{(k)}(0) = i^k \mathbf{E}[X^k]$.

2. In particular, if $X \in \mathcal{L}^2$, then

$$\psi_X(t) = 1 + it\mathbf{E}[X] - \frac{t^2}{2}\mathbf{E}[X^2] + \varepsilon(t)t^2$$

as $\varepsilon(t) \xrightarrow{t\to 0} 0$.

Proof. 1. With $|X|^p$ also $|X|^p \vee 1$ is integrable. Thus, since $|X|^k \leq |X|^p \vee 1$, all $|X|^k$ can be dominated by an integrable random variable and the right-hand side exists. Since the statement is obviously true for $k = 0$, we assume assume that it is valid for $k < n$. Then

$$\Big|\frac{d^{k+1}}{dt^{k+1}}e^{itX}\Big| = \lim_{h\to 0}\Big|\frac{(iX)^k e^{i(t+h)X} - (iX)^k e^{itX}}{h}\Big| \leq \Big|X^k\frac{e^{ihX}-1}{h}\Big| \leq \big|X^{k+1}\big|.$$

Due to dominated convergence, the derivative and integral can be interchanged and it follows

$$\psi_X^{(k+1)}(t) = \mathbf{E}\Big[\frac{d}{dt}(iX)^k e^{itX}\Big] = \mathbf{E}[(iX)^{k+1}e^{itX}].$$

The continuity of the derivative also follows with dominated convergence.
2. For the estimation, we need the Taylor expansion of ψ_X with remainder term. We have

$$e^{itX} = 1 + itX - \frac{t^2X^2}{2}(\cos(\theta_1 tX) + i\sin(\theta_2 tX))$$

with random numbers θ_1, θ_2, so that $|\theta_1|, |\theta_2| \leq 1$. Therefore we get

$$\psi_X(t) = 1 + it\mathbf{E}[X] - \frac{t^2}{2}\mathbf{E}[X^2] + \varepsilon(t)t^2$$

with $2\varepsilon(t) = \mathbf{E}[X^2(1 - \cos(\theta_1 tX) + i\sin(\theta_2 tX))] \xrightarrow{t\to 0} 0$ from dominated convergence. □

Example 6.15 (Moments of the exponential and normal distribution)**.** *1. Let $X \sim \exp(\gamma)$. We know that $\mathscr{L}_{\exp(\gamma)}(t) = \gamma/(\gamma+t)$ from Example 6.13.4. From this and the last Proposition,*

$$\mathbf{E}[X^n] = (-1)^n\frac{d^n}{dt^n}\mathbf{E}[e^{-tX}]\big|_{t=0} = (-1)^n\frac{d^n}{dt^n}\frac{\gamma}{\gamma+t}\Big|_{t=0} = \frac{n!\gamma}{(\gamma+t)^{n+1}}\Big|_{t=0} = \frac{n!}{\gamma^n}.$$

2. For $X \sim N(\mu, \sigma^2)$, we already know $\psi_{N(\mu,\sigma^2)}(t) = e^{it\mu - \sigma^2t^2/2}$. For small t we develop this with

$$\psi_{N(\mu,\sigma^2)}(t) = 1 + it\mu - \sigma^2t^2/2 - \mu^2t^2/2 + \varepsilon(t)t^2$$

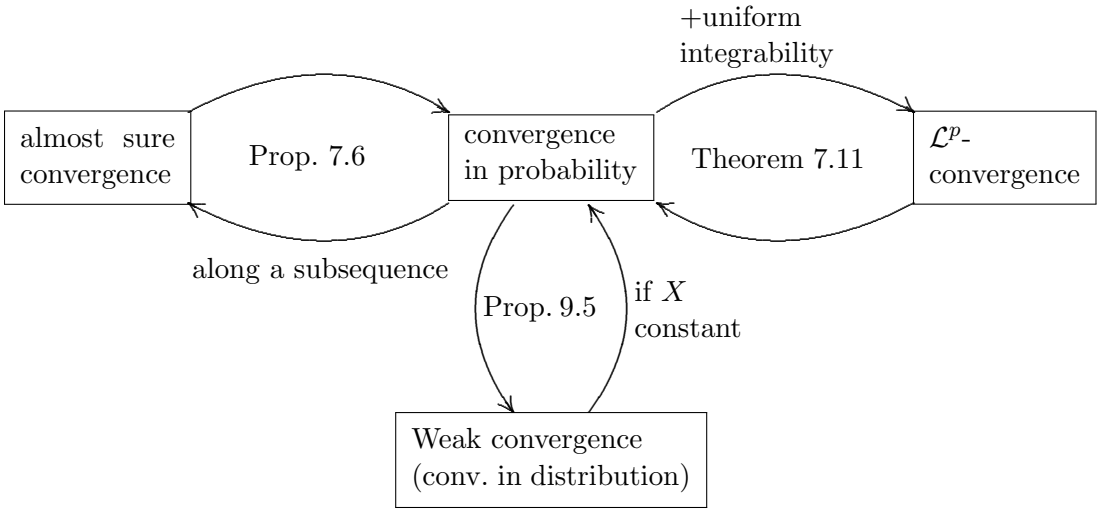
with $\varepsilon(t) \xrightarrow{t\to 0} 0$. From this one reads by means of Proposition 6.14.2 that

$$\mathbf{E}[X] = \mu, \qquad \mathbf{V}[X] = \mathbf{E}[X^2] - \mu^2 = \sigma^2.$$

7 Almost sure, stochastic and $\mathcal{L}^p$-convergence

It is already known from lectures on *Analysis* that there are different types of convergence, such as uniform and pointwise convergence. We will now discuss the most important types of convergence with respect to random variables. (For definitions, see below.)

In our course on measure theory, we have already seen almost sure convergence. In addition, we will discuss convergence in probability and the $\mathcal{L}^p$-convergence (see also Section 4). In Section 9, we will learn about convergence in distribution (which is the same as the weak convergence of the distributions of random variables). The following diagram summarizes all types of convergence:



7.1 Definition and examples

Let's start with some definitions.

Definition 7.1 (Almost sure convergence and convergence in probability)**.** *Let $X, X_1, X_2, \ldots$ be random variables with values in a metric space (E, r).*

1. *If*
$$\mathbf{P}(\lim_{n\to\infty} r(X_n, X) = 0) = 1,$$
we say that the sequence $X_1, X_2, \ldots$ converges almost surely to X and write $X_n \xrightarrow{n\to\infty}_{as} X$.

2. *If, for all $\varepsilon > 0$,*
$$\lim_{n\to\infty} \mathbf{P}(r(X_n, X) > \varepsilon) = 0,$$
we say that the sequence $X_1, X_2, \ldots$ converges to X in probability (or stochastically) and write $X_n \xrightarrow{n\to\infty}_{p} X$.

3. *If the random variables are real-valued and for some $p > 0$*
$$\lim_{n\to\infty} \mathbf{E}[|X_n - X|^p] = 0,$$
we say that the sequence $X_1, X_2, \ldots$ converges in $\mathcal{L}^p$ (or in the p-th mean) to X and also write $X_n \xrightarrow{n\to\infty}_{\mathcal{L}^p} X$.

Remark 7.2 (Properties of $\mathcal{L}^p$ convergence)**.** *From section 4 we already know a lot about the $\mathcal{L}^p$-convergence. For example, if $X, X_1, X_2, \dots$ is such that $X_n \xrightarrow{n\to\infty}_{\mathcal{L}^q} X$ and $p < q$, then $X_n \xrightarrow{n\to\infty}_{\mathcal{L}^p} X$ according to Proposition 4.7. In addition, the spaces $\mathcal{L}^p$ are complete according to Proposition 4.8. So, if for all $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that for all $m, n \geq n$*

$$\mathbf{E}[|X_n - X_m|^p] < \varepsilon,$$

then there is a random variable $X \in \mathcal{L}^p$ with $X_n \xrightarrow{n\to\infty}_{\mathcal{L}^p} X$.

Example 7.3 (Counterexamples)**.** *If we look at the diagram at the beginning of the chapter, we can see that convergence in probability follows from almost sure convergence, but not vice versa. Furthermore, convergence in probability follows from $\mathcal{L}^1$ convergence, but not even almost sure convergence implies $\mathcal{L}^1$ convergence. We first give two examples for these two cases.*

1. *Convergence in probability does not imply almost sure convergence: Let U be a $[0,1]$ uniformly distributed random variable. Further we set*

$$\begin{aligned} &A_1 = [0, \tfrac{1}{2}], && A_2 = [\tfrac{1}{2}, 1], \\ &A_3 = [0, \tfrac{1}{4}], && A_4 = [\tfrac{1}{4}, \tfrac{2}{4}], && A_5 = [\tfrac{2}{4}, \tfrac{3}{4}], && A_6 = [\tfrac{3}{4}, 1], \\ &\dots \end{aligned}$$

   and $X_n := 1_{U \in A_n}$. Then it is clear for $0 < \varepsilon < 1$

$$\lim_{n\to\infty} \mathbf{P}(|X_n| > \varepsilon) = \lim_{n\to\infty} \mathbf{P}(U \in A_n) = 0,$$

   i.e. $X_n \xrightarrow{n\to\infty}_p 0$, but for each $n \in \mathbb{N}$ there is an $m > n$ with $X_m = 1$. Therefore, $X_1, X_2, \dots$ does not converge almost surely to 0.

2. *Almost sure convergence does not imply $\mathcal{L}^1$ convergence: Let U again be a uniformly distributed random variable on $[0,1]$ random variable. Further, $B_n = [0, \frac{1}{n}]$ and $Y_n = n \cdot 1_{U \in B_n}$. Then $Y_n \xrightarrow{n\to\infty}_{fs} Y = \infty \cdot 1_{U=0}$, so $Y = 0$ is almost sure. On the other hand*

$$\mathbf{E}[Y_n] = n \cdot \mathbf{P}(U \in A_n) = 1,$$

   so $Y_1, Y_2, \dots$ does not converge to 0 in $\mathcal{L}^1$.

Lemma 7.4 (Limit in probability is (almost surely) unique)**.** *Let $X, Y, X_1, X_2, \dots$ be random variables with values in a metric space space (E, r) and $X_n \xrightarrow{n\to\infty}_p X$, as well as $X_n \xrightarrow{n\to\infty}_p Y$. Then $X = Y$ almost surely.*

Proof. For all $\varepsilon > 0$,

$$\begin{aligned} \mathbf{P}(r(X,Y) > 2\varepsilon) &\leq \mathbf{P}(r(X_n, X) > \varepsilon \text{ or } r(X_n, Y) > \varepsilon) \\ &\leq \mathbf{P}(r(X_n, X) > \varepsilon) + \mathbf{P}(r(X_n, Y) > \varepsilon) \xrightarrow{n\to\infty} 0. \end{aligned}$$

Therefore,

$$\mathbf{P}(X \neq Y) = \mathbf{P}\Big(\bigcup_{k=1}^{\infty} \Big\{ r(X,Y) > 1/k \Big\} \Big) \leq \sum_{k=1}^{\infty} \mathbf{P}(r(X,Y) > 1/k) = 0,$$

and the statement follows. □

7.2 Almost sure convergence and convergence in probability

We now show a result that relates almost sure convergence and convergence in probability.

Lemma 7.5 (Characterization of convergence in probability)**.** *Let $X, X_1, X_2, \dots$ be random variables with values in a metric space (E, r). Then,*

$$X_n \xrightarrow{n\to\infty}_p X \qquad \Longleftrightarrow \qquad \mathbf{E}[r(X_n, X) \wedge 1] \xrightarrow{n\to\infty} 0. \tag{7.1}$$

Proof. If $X_n \xrightarrow{n\to\infty}_p X$, then for all $\varepsilon > 0$,

$$\begin{aligned}\lim_{n\to\infty} \mathbf{E}[r(X_n, X) \wedge 1] &= \lim_{n\to\infty} \mathbf{E}[r(X_n, X) \wedge 1, r(X_n, X) \le \varepsilon] + \mathbf{E}[r(X_n, X) \wedge 1, r(X_n, X) > \varepsilon] \\ &\le \lim_{n\to\infty} \big(\varepsilon + \mathbf{P}(r(X_n, X) > \varepsilon)\big) = \varepsilon,\end{aligned}$$

which shows the right-hand side since $\varepsilon > 0$ was arbitrary. However, if the right-hand side applies, the Chebyshev inequality for $0 < \varepsilon \le 1$ implies that

$$\mathbf{P}(r(X_n, X) > \varepsilon) \le \frac{\mathbf{E}[r(X_n, X) \wedge 1]}{\varepsilon} \xrightarrow{n\to\infty} 0. \qquad \square$$

Proposition 7.6 (Convergence in probability and almost sure convergence)**.** *Let $X, X_1, X_2, \dots$ be random variables with values in a metric space (E, r). Then, the following are equivalent:*

1. $X_n \xrightarrow{n\to\infty}_p X$
2. *For each sequence $(n_k)_{k=1,2,\dots}$ there is a subsequence $(n_{k_\ell})_{\ell=1,2,\dots}$ with $X_{n_{k_\ell}} \xrightarrow{\ell\to\infty}_{as} X$.*

In particular,

$$X_n \xrightarrow{n\to\infty}_{fs} X \quad \longrightarrow \quad X_n \xrightarrow{n\to\infty}_p X.$$

Proof. 1. $\to$ 2.: Because of (7.1), for each sequence $(n_k)_{k=1,2,\dots}$, there is a subsequence $(n_{k_\ell})_{\ell=1,2,\dots}$ such that (using monotone convergence in the first equality)

$$\mathbf{E}\Big[\sum_{\ell=1}^{\infty} (r(X_{n_{k_\ell}}, X) \wedge 1)\Big] = \sum_{\ell=1}^{\infty} \mathbf{E}[r(X_{n_{k_\ell}}, X) \wedge 1] < \infty.$$

This means that

$$1 = \mathbf{P}\Big(\sum_{\ell=1}^{\infty} (r(X_{n_{k_\ell}}, X) \wedge 1) < \infty\Big) \le \mathbf{P}\big(\limsup_{\ell\to\infty} r(X_{n_{k_\ell}}, X) = 0\big) \le 1,$$

i.e. $X_{n_{k_\ell}} \xrightarrow{\ell\to\infty}_{fs} X$.
2. $\to$ 1.: Let's assume that 1. is not valid. (We must show that 2. cannot hold.) Because of (7.1), there is $\varepsilon > 0$ and a subsequence $(n_k)_{k=1,2,\dots}$, so that $\lim_{k\to\infty} \mathbf{E}[r(X_{n_k}, X) \wedge 1] > \varepsilon$. Assuming that there is a subsequence $(n_{k_\ell})_{\ell=1,2,\dots}$ such that $X_{n_{k_\ell}} \xrightarrow{\ell\to\infty}_{as} X$. Then also

$$\lim_{\ell\to\infty} \mathbf{E}[r(X_{n_{k_\ell}}, X) \wedge 1] = \mathbf{E}\big[\lim_{\ell\to\infty} r(X_{n_{k_\ell}}, X) \wedge 1\big] = 0$$

due to dominated convergence, i.e. a contradiction. So we have found a sequence $(n_k)_{k=1,2,\dots}$ for which there is no further subsequence $(n_{k_\ell})_{\ell=1,2,\dots}$ with $X_{n_{k_\ell}} \xrightarrow{\ell\to\infty}_{as} X$, and we have shown that 2. does not hold. $\square$

7.3 Convergence in probability and $\mathcal{L}^p$-convergence

In Example 7.3 we had already seen that almost sure convergence (as well as convergence in probability) does not imply $\mathcal{L}^1$-convergence. This is not surprising, since the theorem of dominated convergence states that a sequence $X_1, X_2, \ldots$, which converges almost surely to X and has an integrable dominating random variable converges in $\mathcal{L}^1$ to X. If the almost sure convergence implies the $\mathcal{L}^1$ convergence, one would not need to make the requirement of an integrable dominating random variable. In the following, we want to find conditions of the integrable dominating random variable in order to suffice for $\mathcal{L}^1$ convergence. See theorem 7.11 and Corollary 7.12. The concept of uniform integrability is central to this, see Definition 7.7.

Definition 7.7 (Uniform integrability)**.** *A family $(X_i)_{i\in I}$ is called* uniformly integrable *if*

$$\inf_K \sup_{i\in I} \mathbf{E}[|X_i|; |X_i| > K] = 0.$$

Example 7.8 (Uniform integrability)**.** *1. Let $Y \in \mathcal{L}^1$ and $(X_i)_{i\in I}$ with $\sup_i |X_i| \leq |Y|$. Then $(X_i)_{i\in I}$ is uniformly integrable since*

$$\sup_{i\in I} \mathbf{E}[|X_i|; |X_i| > K] \leq \mathbf{E}[|Y|; |Y| > K] \xrightarrow{K\to\infty} 0$$

by dominated convergence. In particular, every $Y \in \mathcal{L}^1$ is uniformly integrable.

2. *Every finite family $(X_i)_{i=1,\ldots,n}$ with $X_1, \ldots, X_n \in \mathcal{L}^1$ is uniformly integrable, because $\sup_{1\leq i\leq n} |X_i| \in \mathcal{L}^1$ and therefore, 1. applies with $Y = \sup_{1\leq i\leq n} |X_i|$.*

3. *Let us consider the example 7.3.2 Here, for $n > K$*

$$\mathbf{E}[|Y_n|; |Y_n| > K] = \mathbf{E}[Y_n] = 1.$$

In particular, $(Y_n)_{n=1,2,\ldots}$ is not uniformly integrable.

4. *Let $p > 1$. Then $(X_i)_{i\in I}$ with $X_i \in \mathcal{L}^p, i \in I$ is uniformly integrable if $\sup_{i\in I} ||X_i||_p < \infty$. This is because $K^{p-1}|X_i|1_{|X_i|>K} \leq |X_i|^p$, therefore*

$$\sup_{i\in I} \mathbf{E}[|X_i|; |X_i| > K] \leq \sup_{i\in I} \frac{\mathbf{E}[|X_i|^p]}{K^{p-1}} \xrightarrow{K\to\infty} 0.$$

.

Lemma 7.9 (Characterization of uniform integrability)**.** *Let $(X_i)_{i\in I}$ be a family of random variables. Then, the following are equivalent:*

1. *$(X_i)_{i\in I}$ is uniformly integrable.*

2. *It holds*

$$\sup_{i\in I} \mathbf{E}[|X_i|] < \infty \quad \textit{and} \quad \lim_{\varepsilon\to 0} \sup_{A:\mathbf{P}(A)<\varepsilon} \sup_{i\in I} \mathbf{E}[|X_i|; A] = 0.$$

.

3. *It holds*

$$\lim_{K\to\infty} \sup_{i\in I} \mathbf{E}[(|X_i| - K)^+] = 0.$$

4. *There exists a function* $f : \mathbb{R}_+ \to \mathbb{R}_+$ *such that* $\frac{f(x)}{x} \xrightarrow{x\to\infty} \infty$ *and* $\sup_{i\in I} \mathbf{E}[f(|X_i|)] < \infty$.

If one of the four statements is true, the function f *in 4. can be chosen to be monotonically increasing and convex.*

Proof. '1. $\to$ 2.': Let $\delta > 0$ be given and $K = K_\delta$ such that $\sup_{i\in I} \mathbf{E}[|X_i|; |X_i| > K] \le \delta$. Then, for $A \in \mathcal{F}$,

$$\mathbf{E}[|X_i|; A] = \mathbf{E}[|X_i|; A \cap \{|X_i| > K\}] + \mathbf{E}[|X_i|; A \cap \{|X_i| \le K\}] \le \delta + K \cdot \mathbf{P}(A).$$

In particular,

$$\sup_{i\in I} \mathbf{E}[|X_i|] = \sup_{i\in I} \mathbf{E}[|X_i|; \Omega] \le \delta + K < \infty$$

and

$$\sup_{A:\mathbf{P}(A)<\varepsilon} \sup_{i\in I} \mathbf{E}[|X_i|; A] \le \delta + K\varepsilon \xrightarrow{\varepsilon\to 0} \delta.$$

Since $\delta > 0$ was arbitrary,

$$\lim_{\varepsilon\to 0} \sup_{A:\mathbf{P}(A)<\varepsilon} \sup_{i\in I} \mathbf{E}[|X_i|; A] = 0.$$

'2. $\Rightarrow$ 3.': First, we note that $(|X_i| - K)^+ \le |X_i| 1_{|X_i|\ge K}$. Let $\varepsilon > 0$. Choose $K = K_\varepsilon$ large enough so that – using the Markov inequality –

$$\sup_{i\in I} \mathbf{P}(|X_i| > K) \le \sup_{i\in I} \frac{\mathbf{E}[|X_i|]}{K} < \varepsilon.$$

This means that 3. follows from

$$\begin{aligned}\lim_{K\to\infty} \sup_{i\in I} \mathbf{E}[(|X_i| - K)^+] &= \lim_{\varepsilon\to 0} \sup_{i\in I} \mathbf{E}[(|X_i| - K_\varepsilon)^+] \le \lim_{\varepsilon\to 0} \sup_{i\in I} \mathbf{E}[|X_i|; |X_i| > K_\varepsilon] \\ &\le \lim_{\varepsilon\to 0} \sup_{A:\mathbf{P}(A)<\varepsilon} \sup_{i\in I} \mathbf{E}[|X_i|; A] = 0.\end{aligned}$$

'3. $\Rightarrow$ 4.': There is a sequence $K_1, K_2, \dots$ with $K_n \uparrow \infty$ and $\sup_{i\in I} \mathbf{E}[(|X_i| - K_n)^+] \le 2^{-n}$. We set

$$f(x) := \sum_{n=1}^{\infty} (x - K_n)^+.$$

Then f is monotonically increasing and convex as a sum of convex functions. Furthermore, for $x \ge 2K_n$,

$$\frac{f(x)}{x} \ge \sum_{k=1}^{n} \left(1 - \frac{K_k}{x}\right) \ge \frac{n}{2},$$

thus $\frac{f(x)}{x} \xrightarrow{x\to\infty} \infty$. Because of monotone convergence,

$$\mathbf{E}[f(|X_i|)] = \sum_{n=1}^{\infty} \mathbf{E}[(|X_i| - K_n)^+] \le \sum_{n=1}^{\infty} 2^{-n} = 1.$$

'4. $\Rightarrow$ 1.': Set $a_K := \inf_{x\ge K} \frac{f(x)}{x}$, so that $a_K \xrightarrow{K\to\infty} \infty$. Thus,

$$\sup_{i\in I} \mathbf{E}[|X_i|; |X_i| \ge K] \le \frac{1}{a_K} \sup_{i\in I} \mathbf{E}[f(|X_i|); |X_i| \ge K] \le \frac{1}{a_K} \sup_{i\in I} \mathbf{E}[f(|X_i|)] \xrightarrow{K\to\infty} 0. \qquad \square$$

Example 7.10 (Differences and uniform integrability)**.** *For $X \in \mathcal{L}^1$, $(X_i)_{i\in I}$ is uniformly integrable iff $(X_i - X)_{i\in I}$ is uniformly integrable.*

To see this, let $(X_i)_{i\in I}$ be uniformly integrable. According to Example 7.8.2, X is uniformly integrable. Furthermore,

$$\sup_{i\in I} \mathbf{E}[|X_i - X|] \leq \mathbf{E}[|X|] + \sup_{i\in I} \mathbf{E}[|X_i|] < \infty$$

and

$$\sup_{A:\mathbf{P}(A)<\varepsilon} \sup_{i\in I} \mathbf{E}[|X_i - X|; A] \leq \sup_{A:\mathbf{P}(A)<\varepsilon} \sup_{i\in I} \mathbf{E}[|X_i|; A] + \sup_{A:\mathbf{P}(A)<\varepsilon} \mathbf{E}[|X|; A] \xrightarrow{\varepsilon\to 0} 0,$$

i.e. according to Lemma 7.9, $(X_i - X)_{i\in I}$ is uniformly integrable. The inverse follows analogously.

Theorem 7.11 (Convergence in probability and convergence in the p-th mean)**.** *Let $X_1, X_2, \ldots$ be a sequence in $\mathcal{L}^p$ with $1 \leq p < \infty$. The following statements are equivalent:*

1. *There is a measurable function $X \in \mathcal{L}^p$ with $X_n \xrightarrow{n\to\infty}_{\mathcal{L}^p} X$.*
2. *The family $(|X_i|^p)_{i=1,2,\ldots}$ is uniformly integrable and there is a measurable function X with $X_n \xrightarrow{n\to\infty}_{p} X$.*

If 1. or 2. applies, then the limits coincide almost surely.

Proof. 1.$\to$ 2.: First, due to Chebyshev's inequality for every $\varepsilon > 0$

$$\mathbf{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbf{E}[|X_n - X|^p]}{\varepsilon^p} = \frac{||X_n - X||_p^p}{\varepsilon^p} \xrightarrow{n\to\infty} 0,$$

i.e. convergence in probability applies. For the proof of uniform integrability, we use Lemma 7.9. Let $\varepsilon > 0$ and $N = N_\varepsilon$ such that $||X_n - X||_p < \varepsilon$ for $n \geq N$. Then for $A \in \mathcal{F}$, with Minkowski's inequality,

$$\begin{aligned}
\sup_{n\in\mathbb{N}}(\mathbf{E}[|X_n|^p; A])^{1/p} &= \sup_{n\in\mathbb{N}} ||X_n 1_A||_p \\
&\leq \sup_{n<N} ||X_n 1_A||_p + \sup_{n\geq N} ||(X_n - X)1_A||_p + ||X 1_A||_p \\
&\leq \sup_{n<N} (\mathbf{E}[|X_n|^p; A]) + \varepsilon + (\mathbf{E}[|X|^p; A]).
\end{aligned}$$

Using $A = \Omega$, we find $\sup_{n\in\mathbb{N}}(\mathbf{E}[|X_n|^p]) < \infty$. Moreover, since N is finite, we find

$$\lim_{\delta\to 0} \sup_{A:\mathbf{P}(A)<\delta} \sup_{n\in\mathbb{N}} \mathbf{E}[|X_n|^p; A] \leq \varepsilon^p.$$

Because $\varepsilon > 0$ was arbitrary, the assertion follows.
2.$\to$ 1.: Since $X_n \xrightarrow{n\to\infty}_{p} X$, according to proposition 7.6 there is a subsequence $n_1, n_2, \ldots$ with $X_{n_k} \xrightarrow{k\to\infty} X$ almost surely. With Fatou's Lemma,

$$\mathbf{E}[|X|^p] = \mathbf{E}[\liminf_{k\to\infty} |X_{n_k}|^p] \leq \sup_{n\in\mathbb{N}} \mathbf{E}[|X_n|^p] < \infty$$

because of Lemma 7.9. In particular, $X \in \mathcal{L}^p$. Just like in Example 7.10, $\{|X_n - X|^p : n \in \mathbb{N}\}$ is uniformly integrable. For every $\delta > 0$, due to convergence in probability,

$$\mathbf{P}(|X_n - X| > \delta) \xrightarrow{n\to\infty} 0.$$

From lemma 7.9 now follows with dominated convergence

$$\lim_{n\to\infty} \mathbf{E}[|X_n - X|^p] = \lim_{n\to\infty} \mathbf{E}[|X_n - X|^p; |X_n - X| > \delta] + \mathbf{E}[|X_n - X|^p; |X_n - X| \leq \delta] \leq \delta^p.$$

Since $\delta > 0$ was arbitrary, $X_n \xrightarrow{n\to\infty}_{\mathcal{L}^p} X$ follows. □

Corollary 7.12 ($\mathcal{L}^p$-convergence and uniform integrability)**.** *Let $1 \leq p < \infty$ and $X_1, X_2, \cdots \in \mathcal{L}^p$ and X be measurable with $X_n \xrightarrow{n\to\infty}_p X$. Then, the following are equivalent:*

1. $X_n \xrightarrow{n\to\infty}_{\mathcal{L}^p} X$,
2. $||X_n||_p \xrightarrow{n\to\infty} ||X||_p$,
3. *The family $(|X_n|^p)_{n=1,2,\dots}$ is uniformly integrable.*

Proof. The equivalence 1. $\Leftrightarrow$ 3. is clear from Theorem 7.11.
1. $\Rightarrow$ 2.: follows from Minkowski's inequality with

$$\big|||X_n||_p - ||X||_p\big| \leq ||X_n - X||_p \xrightarrow{n\to\infty} 0.$$

2. $\to$ 3.: For fixed K, we write

$$\mathbf{E}[|X_n|^p; |X_n| > K] \leq \mathbf{E}[|X_n|^p - (|X_n| \wedge (K - |X_n|)^+)^p] \xrightarrow{n\to\infty} \mathbf{E}[|X|^p - (|X| \wedge (K - |X|)^+)^p].$$

Convergence follows from $\mathbf{E}[|X_n|^p] \xrightarrow{n\to\infty} \mathbf{E}[|X|^p]$, and $(|X_n| \wedge (K - X_n|)^+)^p \xrightarrow{n\to\infty}_{\mathcal{L}^1} |X| \wedge (K - X)^+)^p$, since the convergence according to Proposition 7.6 is in probability, and $((|X_n| \wedge (K - |X_n|)^+)^p)_{n=1,2,\dots}$ is bounded, in particular uniformly integrable. Since $\mathbf{E}[|X|^p - (|X| \wedge (K - |X|)^+)^p] \xrightarrow{K\to\infty} 0$ after dominated convergence, $(|X_n|^p)_{n=1,2,\dots}$ is uniformly integrable. □

8 Independence and the strong law

With our knowledge on probability measures and σ-algebras we now shed light on the concept of independence. In particular, in this chapter we will prove the strong law of large numbers, see Theorem 8.21. On the way, we prove the Borel-Cantelli lemma (Theorem 8.8) and Kolmogorov's 0-1 law (Theorem 8.15).

8.1 Definition and simple properties

Already in the lecture *Basic Probability*, independent random variables were considered. The intuitive idea of independence is often correct, but should sometimes be treated with caution.

Definition 8.1 (Independence)**.** *1. A family of sets $(A_i)_{i\in I}$ with $A_i \in \mathcal{F}$ is called* independent *if*

$$\mathbf{P}\Big(\bigcap_{j\in J} A_j\Big) = \prod_{j\in J} \mathbf{P}(A_j) \tag{8.1}$$

for all $J \subseteq_f I$.[2]

[2]Recall that we write $J \subseteq_f I$ iff $J \subseteq I$ and J is finite.

2. *A family $(\mathcal{C}_i)_{i\in I}$ of set systems $\mathcal{C}_i \subseteq \mathcal{F}$ is called* independent *if* (8.1) *holds for all $J \subseteq_f I$ and $A_j \in \mathcal{C}_j, j \in J$.*

3. *A family of random variables $(X_i)_{i\in I}$ is called independent if $(\sigma(X_i))_{i\in I}$ is independent.*

We first deal with the question if there are probability spaces with an arbitrary number of independent random variables. Here we benefit from our knowledge of product measures.

Proposition 8.2 (Independence and product measures)**.** *A family $(X_i)_{i\in I}$ of random variables is independent iff for each $J \subseteq_f I$*

$$((X_i)_{i\in J})_*\mathbf{P} = \bigotimes_{i\in J}(X_i)_*\mathbf{P},$$

i.e. the joint distribution of each finite subfamily is the product distribution of the individual distributions.

Proof. By definition, the family $(X_i)_{i\in I}$ is independent if and only if for each $J \subseteq_f I$ and $A_i \in \mathcal{F}, i \in J$,

$$\mathbf{P}(X_i \in A_i, i \in J) = \prod_{i\in J}\mathbf{P}(X_i \in A_i).$$

The assertion now follows from the fact that $\mathbf{P}(X_i \in A_i) = (X_i)_*\mathbf{P}(A_i)$ (see Definition 2.23) and $\mathbf{P}(X_i \in A_i, i \in J) = ((X_i)_{i\in J})_*\mathbf{P}\big(\bigtimes_{i\in J} A_i\big)$ (see Corollary 5.14). □

Corollary 8.3 (Existence of uncountably many independent random variables)**.** *Let E be a Polish space and I an arbitrary index set. Let $(\Omega_i, \mathcal{F}_i, \mathbf{P}_i)$ be probability spaces and X_i an E-valued random variable, $i \in I$. Then there is a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a family $(Y_i)_{i\in I}$ E-valued, independent random variable with $Y_i \overset{d}{=} X_i$.*

Proof. It should be noted that $(((X_i)_{i\in J})_* \bigotimes_{i\in J}\mathbf{P}_i)_{J\subseteq_f I}$ is a projective family of probability measures on $(E, \mathcal{B}(E))$. Using Theorem 5.24 we find the projective limit $\mathbf{P}_I$. This is a probability measure on $(E^I, (\mathcal{B}(E))^I)$. Furthermore, with $\pi_i : E^I \to E$, the i-th projection, $(\pi_i)_*\mathbf{P}_I = (X_i)_*\mathbf{P}_i$, i.e. $\pi_i \overset{d}{=} X_i$. □

Lemma 8.4 (Functions of independent random variables)**.** *Let $(\Omega_i', \mathcal{F}_i'), (\Omega_i'', \mathcal{F}_i'')$, $i \in I$, measurable spaces. Let $(X_i)_{i\in I}$ be a family of independent random variables, $X_i : \Omega \to \Omega_i'$, and $\varphi_i : \Omega_i' \to \Omega_i''$ measurable, $i \in I$. Then the family $(\varphi_i(X_i))_{i\in I}$ is independent.*

Proof. According to Lemma 6.2, the random variable $\varphi_i(X_i)$ is measurable according to $\sigma(X_i)$, $i \in I$, i.e. $\sigma(\varphi_i(X_i)) \subseteq \sigma(X_i)$. Since $(\sigma(X_i))_{i\in I}$ is an independent family by assumption, the assertion follows from the definition of independence. □

Proposition 8.5 (Independent and Uncorrelated)**.** *Let $X, Y \in \mathcal{L}^1$ be independent, real-valued random variables. Then $XY \in \mathcal{L}^1$ and*

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

Proof. The assertion is clear if X and Y are indicator functions. Then, note that if the assertion applies to the pairs (X_i, Y_j), $i, j = 1, \dots, n$, then it also for $\sum_{i=1}^n X_i$ and $\sum_{j=1}^n Y_j$: Indeed, due to the linearity of the expected value,

$$\mathbf{E}\Big[\sum_{i=1}^n X_i \cdot \sum_{j=1}^n Y_j\Big] = \sum_{i=1}^n \sum_{j=1}^n \mathbf{E}[X_i Y_j] = \sum_{i=1}^n \sum_{j=1}^n \mathbf{E}[X_i]\mathbf{E}[Y_j] = \mathbf{E}\Big[\sum_{i=1}^n X_i\Big] \cdot \mathbf{E}\Big[\sum_{j=1}^n Y_j\Big].$$

So, since the assertion applies to indicator functions, it is also valid for simple functions, and thus with monotonic convergence also for non-negative measurable functions. The general case follows with the decomposition $X = X^+ - X^-$ and $Y = Y^+ - Y^-$. □

Example 8.6 (Uncorrelated, non-independent random variables)**.** *Let U be a random variable uniformly distributed on $[0, 1]$, $X = \cos(2\pi U)$ and $Y = \sin(2\pi U)$. Then $\mathbf{E}[X] = \mathbf{E}[Y] = 0$ and*

$$\mathbf{E}[XY] = \int_0^1 \cos(2\pi u)\sin(2\pi u)du = \tfrac{1}{2}\int_0^1 \sin(4\pi u)du = 0$$

and thus X, Y are are uncorrelated. However, $\{|X| < \varepsilon, |Y| < \varepsilon\} = \emptyset$ for $\varepsilon > 0$ is small enough and thus $\mathbf{P}(X^{-1}(-\varepsilon, \varepsilon), Y^{-1}(-\varepsilon, \varepsilon)) = 0 < \mathbf{P}(X^{-1}(-\varepsilon, \varepsilon)) \cdot \mathbf{P}(Y^{-1}(-\varepsilon, \varepsilon))$. This means that X and Y are not independent.

If there is a probability space and (countably) many events, you can ask yourself how many of these events will likely occur. The Borel-Cantelli lemma gives a sharp criterion for the occurrence of only finitely many events.

Definition 8.7 (Limsup of sets)**.** *For $A_1, A_2, \dots \in \mathcal{F}$,*

$$\limsup_{n\to\infty} A_n := \bigcap_{n\geq 1} \bigcup_{m\geq n} A_m$$

is the event infinitely many of the A_n occur*.*

Theorem 8.8 (Borel-Cantelli lemma)**.** *1. Let $A_1, A_2, ... \in \mathcal{F}$. Then,*

$$\sum_{n=1}^\infty \mathbf{P}(A_n) < \infty \Longrightarrow \mathbf{P}(\limsup_{n\to\infty} A_n) = 0.$$

2. If $A_1, A_2, \dots$ are independent,

$$\sum_{n=1}^\infty \mathbf{P}(A_n) = \infty \Longrightarrow \mathbf{P}(\limsup_{n\to\infty} A_n) = 1.$$

Proof. We start with 1. Because of the continuity of $\mathbf{P}$ from above (see Proposition 2.8),

$$\mathbf{P}(\limsup_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbf{P}\Big(\bigcup_{m\geq n} A_m\Big) \leq \lim_{n\to\infty} \sum_{m=n}^\infty \mathbf{P}(A_m) = 0$$

by assumption. For 2. we use that $\log(1-x) \leq -x$ for $x \in [0,1]$. From this and the continuity of $\mathbf{P}$ from below and the independence of $(A_n)_{n=1,2,\dots}$,

$$\begin{aligned}
\mathbf{P}((\limsup_{n\to\infty} A_n)^c) &= \mathbf{P}\Big(\bigcup_{n=1}^{\infty} \bigcap_{m\geq n} A_m^c\Big) \\
&= \lim_{n\to\infty} \mathbf{P}\Big(\bigcap_{m=n}^{\infty} A_m^c\Big) \\
&= \lim_{n\to\infty} \prod_{m=n}^{\infty} (1-\mathbf{P}(A_m)) \\
&= \lim_{n\to\infty} \exp\Big(\sum_{m=n}^{\infty} \log(1-\mathbf{P}(A_m))\Big) \\
&\leq \lim_{n\to\infty} \exp\Big(-\sum_{m=n}^{\infty} \mathbf{P}(A_m)\Big) \\
&= 0,
\end{aligned}$$

and the assertion follows. □

Example 8.9 (Infinite coin toss and geometric distributions)**.**

1. *We consider an infinite coin toss. This means that we have a probability space* $(\Omega, \mathcal{F}, \mathbf{P})$ *and independent random variables* $X_1, X_2, \dots$ *with values in* $\{$*heads, tails*$\}$*. The coin toss is fair, i.e.* $\mathbf{P}(X_n = \text{head}) = 1/2$*. We consider the events* $A_n = \{X_n = \text{head}\}$*. Since*

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \sum_{n=1}^{\infty} \tfrac{1}{2} = \infty$$

and the family $(A_n)_{n\in\mathbb{N}}$ *is independent, it follows from the Borel-Cantelli lemma that almost surely infinitely often* head *occurs.*

2. *We consider the same situation as in 1, but the events* $B_n := \{X_1 = \text{Kopf}\}$*. It is clear that the family* $(B_n)_{n\in\mathbb{N}}$ *is not independent. (For example* $\mathbf{P}(B_1 \cap B_2) = \mathbf{P}(B_1) = 1/2 \neq \frac{1}{4} = \mathbf{P}(B_1)\cdot\mathbf{P}(B_2)$*.) Just like in 1.* $\sum_{n=1}^{\infty} \mathbf{P}(B_n) = \infty$*. It is also clear that* $\mathbf{P}(\limsup_{n\to\infty} B_n) = \frac{1}{2}$*. It follows from this, that in the Borel-Cantelli lemma the condition of independence in 2. does not apply.*

3. *Let* $X_1, X_2, \dots$ *be geometrically distributed with the success parameter* p*. We consider the events* $A_n := \{X_n \geq n\}$ *and ask ourselves whether an infinite number of these events can occur. Since*

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \sum_{n=1}^{\infty} \mathbf{P}(X_n \geq n) = \sum_{n=1}^{\infty} (1-p)^{n-1} = \tfrac{1}{p} < \infty.$$

Therefore, almost surely only a finite number of the events $\{X_n \geq n\}$ *occur.*

8.2 Kolmogorov's 0-1 law

The Borel-Cantelli lemma is already a statement about when an event that depends on an infinite number of events is occur almost surely. We will now examine this situation further.

Proposition 8.10 (Independence of generated $\boldsymbol{\sigma}$-algebras)**.** *Let $(\mathcal{C}_i)_{i\in I}$ be a family of independent, $\cap$-stable set systems. Then, $(\sigma(\mathcal{C}_i))_{i\in I}$ is also an independent family.*

Proof. Let $J = \{i_1, \dots, i_n\} \subseteq_f I$ and (wlog) $n > 1$. Then, (8.1) holds for any $A_{i_1}, \dots, A_{i_n}$ with $A_{i_k} \in \mathcal{C}_{i_k}, k = 1, \dots, n$. We keep $A_{i_2}, \dots, A_{i_n}$ fixed and define

$$\mathcal{D} := \{A_{i_1} \in \mathcal{F} : (8.1) \text{ holds}\}.$$

We will now show that $\mathcal{D}$ is a Dynkin system. Namely, if $A \subseteq B \in \mathcal{D}$, then $B \setminus A \in \mathcal{D}$, because

$$\begin{aligned}\mathbf{P}\Big((B \setminus A) \cap \bigcap_{k=2}^{n} A_{i_k}\Big) &= \mathbf{P}\Big(B \cap \bigcap_{k=2}^{n} A_{i_k}\Big) - \mathbf{P}\Big(A \cap \bigcap_{k=2}^{n} A_{i_k}\Big) \\ &= (\mathbf{P}(B) - \mathbf{P}(A)) \cdot \prod_{k=2}^{n} \mathbf{P}(A_{i_k}) \\ &= \mathbf{P}(B \setminus A) \cdot \prod_{k=2}^{n} \mathbf{P}(A_{i_k}).\end{aligned}$$

Furthermore, if $A_1, A_2, \dots \in \mathcal{D}$ with $A_1 \subseteq A_2 \subseteq A_2 \dots$, then due to the continuity of $\mathbf{P}$ from below,

$$\begin{aligned}\mathbf{P}\Big(\Big(\bigcup_{j=1}^{\infty} A_j\Big) \cap \bigcap_{k=2}^{n} A_{i_k}\Big) &= \sup_{j\in\mathbb{N}} \mathbf{P}\Big(A_j \cap \bigcap_{k=2}^{n} A_{i_k}\Big) \\ &= \sup_{j\in\mathbb{N}} \mathbf{P}(A_j) \cdot \prod_{k=2}^{n} \mathbf{P}(A_{i_k}) \\ &= \mathbf{P}\Big(\bigcup_{j=1}^{\infty} A_j\Big) \cdot \prod_{k=2}^{n} \mathbf{P}(A_{i_k}).\end{aligned}$$

Since $\mathcal{C}_{i_1}$ is $\cap$-stable and $\mathcal{C}_{i_1} \subseteq \mathcal{D}$, $\sigma(\mathcal{C}_{i_1}) \subseteq \mathcal{D}$ according to theorem 1.13. In particular, (8.1) applies for $A_{i_1} \in \sigma(\mathcal{C}_{i_1}), A_{i_2} \in \mathcal{C}_{i_2}, \dots, A_{i_n} \in \mathcal{C}_{i_n}$. Iterating the above procedure for $k = 2, \dots, n$, you get the statement. □

Corollary 8.11 (Independence of indicator functions)**.** *A family of sets $(A_i)_{i\in I}$ is independent if and only if the family of random variables $(1_{A_i})_{i\in I}$ is independent. In particular,*

$$\mathbf{P}\Big(\bigcap_{j\in J} B_j\Big) = \prod_{j\in J} \mathbf{P}(B_j)$$

for $J \subseteq_f I$, $B_j \in \{A_j, A_j^c\}, j \in J$.

Proof. For $i \in I$ let $\mathcal{C}_i = \{A_i\}$. Then $\sigma(1_{A_i}) = \{\emptyset, A_i, A_i^c, \Omega\} = \sigma(\mathcal{C}_i)$. Since $\mathcal{C}_i$ is trivially cut-stable, the statement follows from Proposition 8.10. □

Corollary 8.12 (Grouping). *Let $(\mathcal{F}_i)_{i\in I}$ be a family of independent σ-algebras. Further, let $\mathcal{I}$ be a partition of I, i.e. $\mathcal{I} = \{I_k, k \in K\}$ with $\biguplus_{k\in K} I_k = I$, so the I_k are disjoint and their union is I. Then, $(\sigma(\mathcal{F}_i : i \in I_k))_{k\in K}$ is also an independent system.*

Proof. The set system $\mathcal{C}_k := \big\{ \bigcap_{i\in J_k} A_i : J_k \subseteq_f I_k, A_i \in \mathcal{F}_i \big\}$ is $\cap$-stable and $\sigma(\mathcal{C}_k) = \sigma(\mathcal{F}_i : i \in I_k)$, $k \in K$. Since, according to the assumption, the family $(\mathcal{C}_k)_{k\in K}$ is independent, the assertion follows from Proposition 8.10. □

We now come to the main statement of this section, Kolmogorov's 0-1 law. For this we introduce a certain σ-algebra, the terminal σ-algebra.

Definition 8.13 (Terminal and trivial σ-algebras). *1. Let $\mathcal{F}_1, \mathcal{F}_2, ... \subseteq \mathcal{F}$ be a sequence of σ-algebras. Then*

$$\mathcal{T}(\mathcal{F}_1, \mathcal{F}_2, \dots) = \bigcap_{n\geq 1} \sigma\Big(\bigcup_{m>n} \mathcal{F}_m \Big)$$

the σ-algebra of terminal events *of $\mathcal{F}_1, \mathcal{F}_2, \dots$*

2. A σ-algebra $\widetilde{\mathcal{F}} \subseteq \mathcal{F}$ is called $\mathbf{P}$-trivial if $\mathbf{P}(A) \in \{0, 1\}$ for all $A \in \widetilde{\mathcal{F}}$.

Lemma 8.14 (Trivial σ-algebras). *1. A σ-algebra $\widetilde{\mathcal{F}}$ is $\mathbf{P}$-trivial if and only if $\widetilde{\mathcal{F}}$ is independent of itself.*

2. Let $\widetilde{\mathcal{F}}$ be a $\mathbf{P}$-trivial σ-algebra and X a $\widetilde{\mathcal{F}}$-measurable random variable with values in a separable metric space E. Then X is constant, almost surely.

Proof. 1. Let $\widetilde{\mathcal{F}}$ be $\mathbf{P}$-trivial and $A, B \in \widetilde{\mathcal{F}}$. Then $\mathbf{P}(A \cap B) = \mathbf{P}(A) \wedge \mathbf{P}(B) = \mathbf{P}(A) \cdot \mathbf{P}(B)$, therefore $\widetilde{\mathcal{F}}$ is independent of itself. If on the other hand, $\widetilde{\mathcal{F}}$ is independent of itself and $A \in \widetilde{\mathcal{F}}$, then $\mathbf{P}(A) = \mathbf{P}(A \cap A) = (\mathbf{P}(A))^2$, i.e. $\mathbf{P}(A) \in \{0, 1\}$.
2. For $n \in \mathbb{N}$, let $(B_{nj})_{j=1,2,\dots}$ be a countable covering of E with balls of radius $1/n$. Since $\widetilde{\mathcal{F}}$ is a $\mathbf{P}$-trivial σ-algebra then $\mathbf{P}(X \in B_{nj}) \in \{0, 1\}$ applies to all n, j. For $n \in \mathbb{N}$ let $J_n := \{j \in \mathbb{N} : \mathbf{P}(X \in B_{nj}) = 1\} \neq \emptyset$. Thus, due to the continuity from above, $\mathbf{P}\Big(X \in \bigcap_{n=1}^{\infty} \bigcap_{j\in J_n} B_{nj}\Big) = 1$. Since $\bigcap_{n=1}^{\infty} \bigcap_{j\in J_n} B_{nj}$ has at most one element, the assertion follows. □

Under independence, the σ-algebra of terminal events is particularly simple.

Theorem 8.15 (Kolmogorov's 0-1 law). *Let $\mathcal{F}_1, \mathcal{F}_2, \dots \subseteq \mathcal{F}$ be a sequence of independent σ-algebras. Then $\mathcal{T} := \mathcal{T}(\mathcal{F}_1, \mathcal{F}_2, \dots)$ $\mathbf{P}$-trivial.*

Proof. Let $\mathcal{T}_n := \sigma\Big(\bigcup_{m>n} \mathcal{F}_m \Big)$, $n = 1, 2, \dots$. According to Corollary 8.12, $(\mathcal{F}_1, \dots, \mathcal{F}_n, \mathcal{T}_n)$ are independent, $n = 1, 2, \dots$ This means that $(\mathcal{F}_1, \dots, \mathcal{F}_n, \mathcal{T})$ are also independent, $n = 1, 2, \dots$ and thus also $(\mathcal{T}, \mathcal{F}_1, \mathcal{F}_2, \dots)$. Again with Corollary 8.12, it follows that $(\mathcal{T}_0, \mathcal{T})$ are independent and, since $\mathcal{T} \subseteq \mathcal{T}_0$ it also follows that $\mathcal{T}$ is independent of itself. Therefore, the assertion follows from Lemma 8.14. □

8.3 Sums of independent random variables

Many important theorems in probability theory deal with independent random variables. In this lecture, these are in particular the Strong Law of Large Numbers (Theorem 8.21) and the Central Limit Theorem (Theorem 10.8). We present here important tools for analyzing sums of independent random variables. The first is the connection with the convolution of probability measures (see section 5.4).

Proposition 8.16 (Convolution is distribution of the independent sum)**.** *Let $X_1, \dots, X_n$ be independent, real-valued random variables. Then,*

$$(X_1 + \cdots + X_n)_* \mathbf{P} = (X_1)_* \mathbf{P} * \cdots * (X_n)_* \mathbf{P}.$$

Further, for the characteristic functions

$$\psi_{X_1+\cdots+X_n} = \psi_{X_1} \cdots \psi_{X_n}$$

and, if $X_1, \dots, X_n$ assume values in $\mathbb{R}_+$,

$$\mathscr{L}_{X_1+\cdots+X_n} = \mathscr{L}_{X_1} \cdots \mathscr{L}_{X_n}.$$

Proof. First of all, according to Proposition 8.2 $((X_1, \dots, X_n))_* \mathbf{P} = (X_1)_* \mathbf{P} \otimes \cdots \otimes (X_n)_* \mathbf{P}$. Thus, the first assertion already follows from Definition 5.17 of the convolution of measures. The further assertions follow from Proposition 8.5, since for example

$$\begin{aligned} \psi_{X_1+\cdots+X_n}(t) &= \mathbf{E}[e^{it(X_1+\cdots+X_n)}] = \mathbf{E}\big[e^{itX_1} \cdots e^{itX_n}\big] \\ &= \mathbf{E}[e^{itX_1}] \cdots \mathbf{E}[e^{itX_n}] = \psi_{X_1}(t) \cdots \psi_{X_n}(t). \end{aligned}$$

□

Kolmogorov's 0-1 law provides a very simple statement as to when sums of independent random variables are almost sure to converge.

Proposition 8.17 (Convergence of sums of independent random variables)**.** *Let $X_1, X_2, \dots$ be independent random variables and $S_n := X_1 + \cdots + X_n$.*

1. *Then,*
$$\mathbf{P}(\omega : S_n(\omega) \text{ converges for } n \to \infty) \in \{0, 1\}$$
2. *Further,*
$$\mathbf{P}(\omega : S_n(\omega)/n \text{ converges for } n \to \infty) \in \{0, 1\}.$$
If $\mathbf{P}(S_n/n \text{ converges}) = 1$, the limit value is almost surely constant.

Proof. Set $\mathcal{F}_i := \sigma(X_i)$, $i = 1, 2, \dots$ This means that the family $(\mathcal{F}_i)_{i=1,2,\dots}$ is independent. The set $\{\omega : S_n(\omega) \text{ converges for } n \to \infty\}$ is measurable with respect to $\mathcal{T}(\mathcal{F}_1, \mathcal{F}_2, \dots)$ and thus the first statement from Theorem 8.15 follows. In the same way it follows that $\mathbf{P}(S_n/n \text{ converges}) \in \{0, 1\}$. Let $S = \lim_{n\to\infty} S_n(n)/n$. Thus, for all $m = 1, 2, \dots$,

$$S = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \lim_{n\to\infty} \frac{X_m + \cdots + X_n}{n},$$

so S is measurable wrt $\sigma\Big(\bigcup_{k\geq m} \mathcal{F}_k\Big)$. This means that S is also $\mathcal{T}$-measurable and therefore almost surely constant according to Theorem 8.15 and Lemma 8.14. □

Proposition 8.18 (Maximum inequality of Kolmogorov)**.** *Let $X_1, X_2, \cdots \in \mathcal{L}^2$ be independent random variables. Then, for $K > 0$,*

$$\mathbf{P}\Big(\sup_{n\in\mathbb{N}}\Big|\sum_{k=1}^{n} X_k - \mathbf{E}[X_k]\Big| > K\Big) \leq \frac{\sum_{n=1}^{\infty}\mathbf{V}(X_n)}{K^2}.$$

Proof. Wlog, let $\mathbf{E}[X_k] = 0, k = 1, 2, \ldots$. We further set $S_n = X_1 + \cdots + X_n$ and $T := \inf\{n : |S_n| > K\}$. Then, $\mathbf{P}(\sup_n |S_n| > K) = \mathbf{P}(T < \infty)$. Because of Corollary 8.12, $S_k \cdot 1_{T=k}$ and $S_n - S_k$ are independent for $k \leq n$. Therefore

$$\begin{aligned}
\sum_{k=1}^{n}\mathbf{E}[X_k^2] = \mathbf{E}[S_n^2] &\geq \sum_{k=1}^{n}\mathbf{E}[S_n^2, T = k] \\
&= \sum_{k=1}^{n}\mathbf{E}[S_k^2 + (S_n - S_k + 2S_k)(S_n - S_k), T = k] \\
&\geq \sum_{k=1}^{n}\mathbf{E}[S_k^2, T = k] + 2\mathbf{E}[S_k(S_n - S_k), T = k] \\
&= \sum_{k=1}^{n}\mathbf{E}[S_k^2, T = k] \geq K^2\mathbf{P}(T \leq n)
\end{aligned}$$

Now follows the assertion with $n \to \infty$. □

Theorem 8.19 (Convergence criterion for series)**.** *Let $X_1, X_2, \cdots \in \mathcal{L}^2$ be independent random variables with $\sum_{n=1}^{\infty}\mathbf{V}[X_n] < \infty$. Then, $\sum_{k=1}^{n} X_k - \mathbf{E}[X_k]$ converges almost surely.*

Proof. Again, let $\mathbf{E}[X_k] = 0, k = 1, 2, \ldots$ and we write $S_n = X_1 + \cdots + X_n$. For $\varepsilon > 0$, according to Proposition 8.18,

$$\lim_{k\to\infty}\mathbf{P}(\sup_{n\geq k}|S_n - S_k| > \varepsilon) \leq \lim_{k\to\infty}\frac{\sum_{n=k+1}^{\infty}\mathbf{E}[X_n^2]}{\varepsilon^2} = 0.$$

Therefore, $\sup_{n\geq k}|S_n - S_k| \xrightarrow{k\to\infty}_p 0$. So, by Proposition 7.6, there is a subsequence $k_1, k_2, \ldots$ with $\sup_{n\geq k_i}|S_n - S_{k_i}| \xrightarrow{i\to\infty}_{fs} 0$. However, since $(\sup_{n\geq k}|S_n - S_k|)_{k=1,2,\ldots}$ is decreasing, $\sup_{n\geq k}|S_n - S_k| \xrightarrow{k\to\infty}_{fs} 0$ applies. This means, however, that $(S_n)_{n=1,2,\ldots}$ converges. □

8.4 The Strong Law of Large Numbers

In the lecture *Basic Probability*, we already proved the weak law of large numbers: if $X_1, X_2, \cdots \in \mathcal{L}^2$ are identically distributed and uncorrelated, then, for $\varepsilon > 0$

$$\mathbf{P}\Big(\frac{1}{n}\Big|\sum_{k=1}^{n}(X_k - \mathbf{E}[X_k])\Big| > \varepsilon\Big) \leq \frac{1}{\varepsilon^2}\mathbf{V}\Big[\frac{1}{n}\sum_{k=1}^{n}X_k\Big] = \frac{1}{\varepsilon^2 n^2}\sum_{k=1}^{n}\mathbf{V}[X_k] = \frac{\mathbf{V}[X_1]}{\varepsilon^2 n} \xrightarrow{n\to\infty} 0.$$

As we now know, this means in other terms,

$$\frac{1}{n}\sum_{k=1}^{n}X_k \xrightarrow{n\to\infty}_p \mathbf{E}[X_0].$$

We now want to improve this statement in two directions. On the one hand, we want to replace convergence in probability by almost sure convergence, and on the other hand only assume the existence of first moments (but not the existence of second moments). First, however, we define what exactly what we mean when we say that a sequence of random variables follows a law of large numbers.

Definition 8.20 (Law of large numbers)**.** *Let $X_1, X_2, \cdots \in \mathcal{L}^1$ be a sequence of real-valued random variables. We say that the sequence follows the weak law of large numbers if*

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - \mathbf{E}[X_k]) \xrightarrow{n\to\infty}_{p} 0.$$

The sequence satisfies the strong law of large numbers if

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - \mathbf{E}[X_k]) \xrightarrow{n\to\infty}_{fs} 0.$$

Theorem 8.21 (Strong law for independent random variables)**.** *A sequence $X_1, X_2, \cdots \in \mathcal{L}^1$ of independent and identically distributed random variables satisfies the strong law of large numbers, i.e.*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{n\to\infty}_{fs} \mathbf{E}[X_1].$$

Remark 8.22 (Weak law of large numbers)**.** *Since convergence in probability is implied by almost sure convergence (see Proposition 7.6), the sequence $X_1, X_2, \ldots$ from the theorem also satisfies the weak law of large numbers. Furthermore, the sequence $X_1^+, X_2^+, \ldots$ also satisfies the strong law and $\mathbf{E}[\frac{1}{n}(X_1^+ + \cdots X_n^+)] = \mathbf{E}[X_1^+]$. This means that the sequence $(\frac{1}{n}(X_1^+ + \cdots + X_n^+))_{n=1,2,\ldots}$ is uniformly according to Corollary 7.12. In the same way, the sequence of partial sums of the negative parts is uniformly integrable. It follows from Theorem 7.11 that $\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{n\to\infty}_{\mathcal{L}^1} \mathbf{E}[X_1]$.*

Remark 8.23 (Finite fourth and second moments)**.** *The difficulty in proving the strong law is that only may be used that $X_1 \in \mathcal{L}^1$. The proof is significantly easier if we use $X_1 \in \mathcal{L}^4$ or $X_1 \in \mathcal{L}^2$. We start with these two proofs and write $S_n := X_1 + \cdots + X_n$.*

1. *The case $X_1 \in \mathcal{L}^4$: Here you can get by without further aids: From the linearity of the expected value, it is clear that $\mathbf{E}[S_n/n] = \mathbf{E}[X_1]$. Wlog, let $\mathbf{E}[X_1] = 0$, otherwise you go to the random variables $X_1 - \mathbf{E}[X_1], X_2 - \mathbf{E}[X_2], \cdots \in \mathcal{L}^4$. First we calculate with the help of the independence of $(X_k)_{k=1,2,\ldots}$*

$$\mathbf{E}[S_n^4] = \sum_{k=1}^{n}\mathbf{E}[X_k^4] + 3\sum_{\substack{k,l=1\\k\neq l}}^{\infty}\mathbf{E}[X_k^2 X_l^2] \leq (n + 6n^2)\mathbf{E}[X_1^4]$$

because of the Cauchy-Schwartz inequality. From this,

$$\mathbf{E}\Big[\sum_{n=1}^{\infty}\Big(\frac{S_n}{n}\Big)^4\Big] \leq \sum_{n=1}^{\infty}\frac{n+6n^2}{n^4}\mathbf{E}[X_1^4] < \infty.$$

Therefore, $\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4 < \infty$ applies is almost sure, in particular $\frac{S_n}{n} \xrightarrow{n\to\infty}_{fs} 0$.

2. *The case $X_1 \in \mathcal{L}^2$: Here the convergence criterion for series, theorem 8.19 is of crucial help. We also need the following result:*

Lemma 8.24 (Kronecker Lemma)**.** *Let $x_1, x_2, \dots \in \mathbb{R}$, $y_1, y_2, \dots \in \mathbb{R}$ be monotone with $y_n \uparrow \infty$ and $\sum_{n=1}^{\infty} x_n/y_n < \infty$. Then, $\sum_{k=1}^{n} x_k/y_n \xrightarrow{n\to\infty} 0$.*

Proof. Let $z_0 = 0, z_n := \sum_{k=1}^{n} x_k/y_k$. Then $z_n \xrightarrow{n\to\infty} z_\infty < \infty$ and $x_k = y_k(z_k - z_{k-1})$. We write with $y_0 = 0$

$$\begin{aligned}\frac{\sum_{k=1}^{n} x_k}{y_n} &= \frac{1}{y_n}\sum_{k=1}^{n} y_k(z_k - z_{k-1}) = z_n + \frac{1}{y_n}\Big(\sum_{k=0}^{n-1} y_k z_k - \sum_{k=1}^{n} y_k z_{k-1}\Big)\\ &= z_n - \frac{1}{y_n}\Big(\sum_{k=1}^{n} y_k z_{k-1} - y_{k-1}z_{k-1}\Big)\\ &\xrightarrow{n\to\infty} z_\infty - z_\infty \cdot \lim_{n\to\infty}\frac{1}{y_n}\sum_{k=1}^{n} y_k - y_{k-1} = 0.\end{aligned}$$

□

Back to the proof of the strong law in the case $X_1 \in \mathcal{L}^2$. Wlog, let $\mathbf{E}[X_1] = 0$. Consider the sequence $X_1/1, X_2/2, \dots$ Because $\sum_{n=1}^{\infty} \mathbf{V}[X_n/n] = \mathbf{V}[X_1]\sum_{n=1}^{\infty} 1/n^2$ applies according to Theorem 8.19 that $\sum_{k=1}^{n} X_k/k$ almost surely converges. With Lemma 8.24 it follows that $S_n/n \xrightarrow{n\to\infty}_{fs} 0$.

Proof of theorem 8.21 if $X_1 \in \mathcal{L}^1$. It is sufficient to consider the case of non-negative random variables. In the general case, note that $X_1^+, X_2^+, \dots \in \mathcal{L}^1$ and $X_1^-, X_2^-, \dots \in \mathcal{L}^1$ fulfill the conditions of the theorem, and from $(X_1^+ + \dots + X_n^+)/n \xrightarrow{n\to\infty}_{fs} \mathbf{E}[X_1^+]$ and $(X_1^- + \dots + X_n^-)/n \xrightarrow{n\to\infty}_{fs} \mathbf{E}[X_1^-]$ the statement follows due to linearity of the expectation.

For $S_n = X_1 + \dots + X_n$ we will show that

$$\mathbf{E}[\limsup_{n\to\infty} S_n/n] \le \mathbf{E}[X_1]. \tag{8.2}$$

If this is true, then firstly

$$\begin{aligned}\mathbf{E}[\liminf_{n\to\infty} S_n/n] &\ge \mathbf{E}[\liminf_{n\to\infty}(X_1 \wedge k + \dots + X_n \wedge k)/n]\\ &= k - \mathbf{E}[\limsup_{n\to\infty}((k - X_1)^+ + \dots + (k - X_n)^+)/n]\\ &\ge \mathbf{E}[k - (k - X_1)^+] \xrightarrow{k\to\infty} \mathbf{E}[X_1].\end{aligned}$$

Secondly, then $\mathbf{E}[\limsup_{n\to\infty} S_n/n - \liminf_{n\to\infty} S_n/n] = 0$, i.e. $\limsup_{n\to\infty} S_n/n = \liminf_{n\to\infty} S_n/n = 0$ almost surely, since both $\liminf_{n\to\infty} S_n/n$ as well as $\limsup_{n\to\infty} S_n/n$ are terminal functions, and thus according to Theorem 8.15 and Lemma 8.14 are almost surely constant. Furthermore,

$$\liminf_{n\to\infty} S_n/n = \mathbf{E}[\liminf_{n\to\infty} S_n/n] \ge \mathbf{E}[X_1] \ge \mathbf{E}[\limsup_{n\to\infty} S_n/n] = \limsup_{n\to\infty} S_n/n,$$

from which the assertion follows.

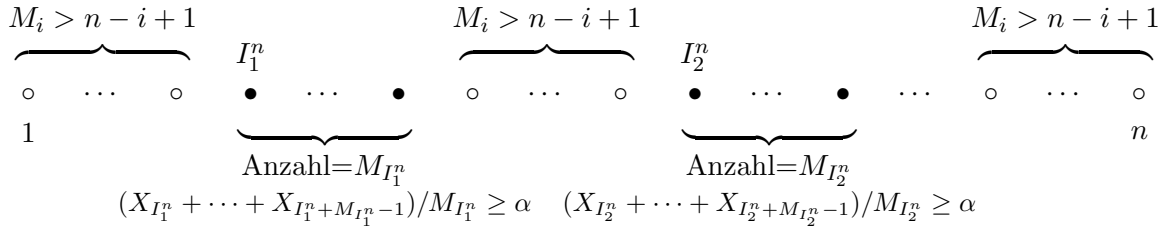


Figure 1: Illustration of M_i, I_j^n, introduced below (8.3). The size L_n is the number of contiguous areas of $\bullet$'s.

It therefore remains to show (8.2). Wlog let $\mathbf{E}[X_1] > 0$, otherwise $X_k = 0$ is almost sure, $k = 1, 2, \ldots$ and the statement is trivial. For this we will use

$$0 < \alpha < \mathbf{E}[\limsup_{n\to\infty} S_n/n] \qquad \Longrightarrow \qquad \alpha \leq \mathbf{E}[X_1] \tag{8.3}$$

can be proved. According to the assumption, for $i = 0, 1, 2, \ldots$

$$\alpha < \mathbf{E}[\limsup_{n\to\infty} S_n/n] = \limsup_{n\to\infty} S_n/n = \limsup_{n\to\infty}(X_{i+1} + \cdots X_{i+n})/n.$$

Thus,

$$M_i := \inf\{n \in \mathbb{N} : (X_i + \cdots + X_{i+n-1})/n \geq \alpha\}$$

is finite, almost surely, $i = 1, 2, \ldots$ The M_i's are identically distributed. We define recursively for $n = 1, 2, \ldots$ (see also Figure 1) $I_1^n = 0$ and for $j = 0, 1, 2, \ldots$ (with $M_0 := 0$)

$$I_{j+1}^n := \inf\{i \in \mathbb{N} : i \geq I_j^n + M_{I_j}^n, M_i \leq n - i + 1\}$$

with $\inf\emptyset = \infty$ and $L_n := \sup\{n \in \mathbb{N}_0 : I_j^n < \infty\}$. This means that for $1 \leq j \leq L_n$, $I_j^n + M_{I_j^n} \leq n$, i.e. $(X_{I_j^n} + \cdots + X_{I_j^n+M_{I_j^n}-1})/M_{I_j^n} \geq \alpha$. We now use this by means of

$$\begin{aligned}
\mathbf{E}[X_1] &= \mathbf{E}[(X_1 + \cdots + X_n)/n] \\
&\geq \tfrac{1}{n}\mathbf{E}\Big[\sum_{j=1}^{L_n} M_{I_j^n} \cdot (X_{I_j^n} + \cdots + X_{I_j^n+M_{I_j^n}-1})/M_{I_j^n}\Big] \\
&\geq \tfrac{\alpha}{n}\mathbf{E}\Big[\sum_{j=1}^{L_n} M_{I_j^n}\Big] = \alpha - \tfrac{\alpha}{n}\mathbf{E}\Big[n - \sum_{j=1}^{L_n} M_{I_j^n}\Big] \\
&\geq \alpha - \tfrac{\alpha}{n}\mathbf{E}\Big[\sum_{i=1}^{n} 1_{M_i > n-i+1}\Big] \\
&= \alpha\Big(1 - \tfrac{1}{n}\sum_{i=1}^{n} \mathbf{P}(M_i > i)\Big) \xrightarrow{n\to\infty} \alpha,
\end{aligned}$$

since $\big(\frac{1}{n}\sum_{i=1}^n \mathbf{P}(M_i > i)\big)_{n=1,2,\ldots}$ as Cesàro-Limes of $(\mathbf{P}(M_i > i))_{i=1,2,\ldots}$ because of the identity of the distributions of the M_i's converges to 0. Thus (8.3) is shown and the assertion is proven. □

We now give a simple application of the strong law. It often happens in statistics that a large number of independent, identically distributed, real-valued random variables must be studied. The Glivenko-Cantelli theorem (Theorem 8.26) states that the empirical distribution of the random variables almost surely converges to the underlying distribution.

Definition 8.25 (Empirical distribution)**.** *Let $X_1, X_2, \dots$ be random variables. For $n = 1, 2, \dots$ the distribution is called (random) probability distribution*

$$\widehat{\mu}_n := \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k}$$

the empirical distribution *of $X_1, \dots, X_n$. If the random variables are real-valued, then in addition*

$$\widehat{F}_n(x) := \frac{1}{n} \sum_{k=1}^{n} 1_{X_k \le x},$$

the empirical distribution function *of $X_1, \dots, X_n$.*

Theorem 8.26 (Glivenko-Cantelli Theorem)**.** *Let $X_1, X_2, \dots$ be independent, real-valued random variables with identical distribution with distribution function F. Then,*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} widehas F_n(x) - F(x)| \xrightarrow{n\to\infty}_{fs} 0.$$

Proof. For $x \in \mathbb{R}$ and $n = 1, 2, \dots$ let $Y_n(x) := 1_{X_n \le x}$ and $Z_n(x) := 1_{X_n < x}$. According to Theorem 8.21, for each $x \in \mathbb{R}$

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} Y_k(x) \xrightarrow{n\to\infty}_{fs} \mathbf{E}[Y_1(x)] = \mathbf{P}(X_1 \le x) = F(x),$$
$$\widehat{F}_n(x-) = \frac{1}{n} \sum_{k=1}^{n} Z_k(x) \xrightarrow{n\to\infty}_{fs} \mathbf{E}[Z_1(x)] = \mathbf{P}(X_1 < x) = F(x-).$$

We must show that these limits hold uniformly for all $x \in \mathbb{R}$. For $N = 1, 2, \dots$ and $j = 0, \dots, N$ we set

$$x_j^N := \inf\{x \in \mathbb{R} : F(x) \ge j/N\}$$

and

$$R_n^N := \max_{j=1,\dots,N-1} \big(|\widehat{F}_n(x_j^N) - F(x_j^N)| + |\widehat{F}_n(x_j^N-) - F(x_j^N-)| \big).$$

For $N = 1, 2, \dots$, therefore, $R_n^N \xrightarrow{n\to\infty}_{fs} 0$. Furthermore, for $x \in (x_{j-1}^N, x_j^N)$

$$\widehat{F}_n(x) \le \widehat{F}_n(x_j^N) \le \widehat{F}_n(x_j^N) + R_n^N \le F(x) + R_n^N + \tfrac{1}{N},$$
$$\widehat{F}_n(x) \ge \widehat{F}_n(x_{j-1}^N) \ge F(x_{j-1}^N) - R_n^N \ge F(x) - R_n^N - \tfrac{1}{N},$$

thus, for each $N = 1, 2, \dots$

$$\sup_{x\in\mathbb{R}} widehat F_n(x) - F(x)| \le \tfrac{1}{N} + R_n^N \xrightarrow{n\to\infty}_{fs} \tfrac{1}{N}.$$

Since the left-hand side does not depend on N, the assertion follows with $N \to \infty$. □

9 Weak convergence

For measurable spaces, we have often used the Borel σ-algebra, i.e. the σ-algebra that is generated by a topology. In this section we will often assume that the topological space is Polish, i.e. separable and metrizable by a complete metric; recall from Definition A.1 in the manuscript on measure theory. To save us some work, we will assume throughout that (E, r) is a metric space and sometimes we will assume that it is complete and separable.

For a measurable mapping $f : E \to \mathbb{R}$ and a measure μ on $\mathcal{B}(E)$ (the Borel's σ-algebra of E) we will use throughout this and the next chapter the notation

$$\mu[f] := \int f d\mu.$$

9.1 Definition and simple properties

So far, we have dealt with different types of convergence of random variables. The convergence in distribution of random variables is the same as the weak convergence of the distributions of random variables. For the motivation behind the following definitions, let us recall a fact: in a metric space (E, r) we have $x_n \xrightarrow{n\to\infty} x$ if and only if $f(x_n) \xrightarrow{n\to\infty} f(x)$ for all continuous functions on E (i.e. $f \in \mathcal{C}(E, \mathbb{R})$).

Definition 9.1 (Weak convergence and convergence in distribution)**.**

1. *We denote by $\mathcal{P}(E)$ the set of probability measures on $\mathcal{B}(E)$ and with $\mathcal{P}_{\leq 1}(E)$ the set of finite measures μ on $\mathcal{B}(E)$ with $\mu(E) \leq 1$. Further, $\mathcal{C}_b(E)$ is the set of the real-valued, bounded, continuous functions on E and $\mathcal{C}_c(E) \subseteq \mathcal{C}_b(E)$ is the set of the real-valued, bounded continuous functions on E with compact support.*

2. *A sequence $\mathbf{P}_1, \mathbf{P}_2, \cdots \in \mathcal{P}(E)$ converges* weakly *to $\mathbf{P} \in \mathcal{P}(E)$, if*

$$\mathbf{P}_n[f] \xrightarrow{n\to\infty} \mathbf{P}[f] \tag{9.1}$$

   for all $f \in \mathcal{C}_b(E)$. We then write

$$\mathbf{P}_n \xRightarrow{n\to\infty} \mathbf{P}.$$

3. *Let $\mu_1, \mu_2, \cdots \in \mathcal{P}_{\leq 1}$ and μ be a measure on E. If (9.1) only applies to all $f \in \mathcal{C}_c(E)$, we say that μ_n converges* vagely *to μ. We then write*

$$\mu_n \xRightarrow{n\to\infty}_v \mu.$$

4. *Let $X, X_1, X_2, \ldots$ be random variables on probability spaces $(\Omega, \mathcal{A}, \mathbf{P})$, $(\Omega_1, \mathcal{A}_1, \mathbf{P}_1)$, $(\Omega_2, \mathcal{A}_2, \mathbf{P}_2), \ldots$ with values in E. Then, $X_1, X_2, \ldots$ converges* in distribution *to X if $(X_n)_*\mathbf{P}_n \xRightarrow{n\to\infty} X_*\mathbf{P}$. We then write*

$$X_n \xRightarrow{n\to\infty} X.$$

Remark 9.2. *1. Note that for random variables $X, X_1, X_2, \ldots$ with values in E, we have $X_n \xRightarrow{n\to\infty} X$ if*

$$\mathbf{P}[f(X_n)] \xrightarrow{n\to\infty} \mathbf{P}[f(X)]$$

for all $f \in \mathcal{C}_b(E)$. Many of the following results can therefore be formulated in two ways: either by means of probability distributions, or by means of random variables. The connection here is always that the statement about the probability distributions is also a statement about the distributions of the random variables.

2. *The weak limit of probability measures must again be a probability measure, since $1 \in \mathcal{C}_b(E)$. The vague limit of probability measures does not necessarily have to be a probability measure, since $1 \notin \mathcal{C}_c(E)$ if E is not compact; see also Example 9.3.1. After all, the vague limit is in $\mathcal{P}_{\leq 1}(E)$, as Lemma 9.12 shows.*

3. *We already know the almost sury convergence, the convergence in probability, and the convergence in $\mathcal{L}^p$ of random variables $X_1, X_2, \dots$ to X. The difference to convergence in distribution is that the latter does not require that the random variables are defined on the same probability space.*

4. *By Definition 9.1, the topology of weak convergence on $\mathcal{P}(E)$ is the weakest (i.e. the smallest) topology for which $\mathbf{P} \mapsto \mathbf{P}[f]$ for all $f \in \mathcal{C}_b(E)$ is continuous.*

Example 9.3. 1. *Let $x, x_1, x_2, \dots \in \mathbb{R}$ with $x_n \xrightarrow{n\to\infty} x$ and $\mathbf{P} = \delta_x, \mathbf{P}_1 = \delta_{x_1}, \mathbf{P}_2 = \delta_{x_2}, \dots$ Then, $\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \mathbf{P}$, since*

$$\mathbf{P}_n[f] = f(x_n) \xrightarrow{n\to\infty} f(x) = \mathbf{P}[f]$$

for all $f \in \mathcal{C}_b(\mathbb{R})$.

If the sequence $x_1, x_2, \dots$ diverges, for example $x_n = n$, then $\mathbf{P}_n \xrightarrow{n\to\infty}_v 0$ (this is the 0-measure on $\mathcal{B}(\mathbb{R})$), since

$$\mathbf{P}_n[f] = f(x_n) \xrightarrow{n\to\infty} 0 = 0[f]$$

for all $f \in \mathcal{C}_c(\mathbb{R})$. However, weak convergence does not hold, since $\mathbf{P}_n[1] = 1 \neq 0 = 0[1]$.

2. *Let $X, X_1, X_2, \dots$ be identically distributed. Then $X_n \overset{n\to\infty}{\Longrightarrow} X$, but in general the convergence is neither almost sure, nor in probability nor in $\mathcal{L}^p$ for any $p > 0$.*

3. *As we will see, the Central Limit Theorem (Theorem 10.8), is a result about convergence in distribution. In its simplest form, the theorem of deMoivre-Laplace (see also Remark 9.8 and Example 9.34), it states: let $p \in (0,1)$, $X_n \sim B(n,p), n = 1, 2, \dots$ and $X \sim N(0,1)$. Then,*

$$\frac{X_n - np}{\sqrt{np(1-p)}} \overset{n\to\infty}{\Longrightarrow} X.$$

4. *Similarly, the Poisson approximation of the binomial distribution is a statement about convergence in distribution (see the course in* Basic Probability *and Theorem 10.5): let $X_n \sim B(n, p_n), n = 1, 2, \dots$ with $n \cdot p_n \xrightarrow{n\to\infty} \lambda$ and $X \sim Poi(\lambda)$. Then,*

$$X_n \xrightarrow{n\to\infty} X.$$

Lemma 9.4 (Uniqueness of the weak limit)**.** *Let $\mathbf{P}, \mathbf{Q}, \mathbf{P}_1, \mathbf{P}_2, \dots \in \mathcal{P}(E)$ with $\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \mathbf{P}$ and $\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \mathbf{Q}$. Then $\mathbf{P} = \mathbf{Q}$.*

Proof. According to Proposition 2.11 it suffices to show that $\mathbf{P}(A) = \mathbf{Q}(A)$ for all closed $A \subseteq E$. (The set of all closed sets is a $\cap$-stable generator of $\mathcal{B}(E)$.) So let $A \subseteq E$ be closed. We set

$$r(x, A) := \inf_{y \in A} r(x, y)$$

and

$$f_m(x) \mapsto (1 - m \cdot r(x, A))^+.$$

for $m = 1, 2, \ldots$. Then $f_m \xrightarrow{m \to \infty} 1_A$, since A is is closed. Then, using cominated convergence,

$$\mathbf{P}(A) = \lim_{m \to \infty} \mathbf{P}[f_m] = \lim_{m \to \infty} \lim_{n \to \infty} \mathbf{P}_n[f_m] = \lim_{m \to \infty} \lim_{m \to \infty} \mathbf{Q}[f_m] = \mathbf{Q}(A)$$

and the assertion follows. □

Recall the initial figure of Chapter 7. A sequence of random variables can converge almost surely, in probability, in $\mathcal{L}^p$ or in distribution. Convergence in distribution is the weakest of these terms in the following sense.

Proposition 9.5 (Convergence in probability and in distribution)**.** *Let $X, X_1, X_2, \ldots$ be random variables with values in E. If $X_n \xrightarrow{n \to \infty}_p X$, then $X_n \xRightarrow{n \to \infty} X$. If X is constant, the inversion also applies.*

Proof. Let $X_n \xrightarrow{n \to \infty}_p X$. Suppose that there is an $f \in \mathcal{C}_b(E)$ such that $\lim_{n \to \infty} \mathbf{P}[f(X_n)] \neq \mathbf{P}[f(X)]$. Then there is a subsequence $(n_k)_{k=1,2,\ldots}$ and a $\varepsilon > 0$ with

$$\lim_{k \to \infty} |\mathbf{P}[f(X_{n_k})] - \mathbf{P}[f(X)]| > \varepsilon. \tag{9.2}$$

Because of $X_n \xrightarrow{n \to \infty}_p X$ and Proposition 7.6 there is a subsequence $(n_{k_\ell})_{\ell=1,2,\ldots}$ such that $X_{n_{k_\ell}} \xrightarrow{\ell \to \infty} X$ almost surely. By dominated convergence, this would imply

$$\lim_{\ell \to \infty} \mathbf{P}[f(X_{n_{k_\ell}})] = \mathbf{P}[f(X)]$$

in contradiction to (9.2).

For the inverse, let $X = s \in E$. Note that $x \mapsto r(x, s) \wedge 1$ is a bounded, continuous function and therefore

$$\mathbf{P}[r(X_n, s) \wedge 1] \xrightarrow{n \to \infty} \mathbf{P}[r(X, s) \wedge 1] = 0.$$

Thus, $X_n \xrightarrow{n \to \infty}_p X$ holds because of (7.1). □

Theorem 9.6 (Portmanteau theorem)**.** *Let $X, X_1, X_2, \ldots$ be random variables with values in E. The following conditions are equivalent:*

(i) $X_n \xRightarrow{n \to \infty} X$

(ii) $\mathbf{P}[f(X_n)] \xrightarrow{n \to \infty} \mathbf{P}[f(X)]$ *for all bounded, Lipschitz-continuous functions f.*

(iii) $\liminf_{n \to \infty} \mathbf{P}(X_n \in G) \geq \mathbf{P}(X \in G)$ *for all open $G \subseteq E$.*

(iv) $\limsup_{n \to \infty} \mathbf{P}(X_n \in F) \leq \mathbf{P}(X \in F)$ *for all completed $F \subseteq E$.*

(v) $\lim_{n\to\infty} \mathbf{P}(X_n \in B) = \mathbf{P}(X \in B)$ *for all* $B \in \mathcal{B}(E)$ *with*[3] $\mathbf{P}(X \in \partial B) = 0$.

Proof. $(i) \to (ii)$: clear.
$(ii) \Rightarrow (iv)$: Let $F \subseteq E$ be closed and $f_1, f_2, \ldots$ Lipschitz-continuous such that $f_k \downarrow 1_F$. (For example, one chooses $\varepsilon_k \downarrow 0$ and $f_k(x) = (1 - \frac{1}{\varepsilon_k} r(x,F))^+$, where $r(x,F) := \inf_{y\in F} r(x,y)$.) This means that

$$\limsup_{n\to\infty} \mathbf{P}(X_n \in F) \le \inf_{k=1,2,\ldots} \limsup_{n\to\infty} \mathbf{P}[f_k(X_n)] = \inf_{k=1,2,\ldots} \mathbf{P}[f_k(X)] = \mathbf{P}(X \in F).$$

$(iii) \iff (iv)$: That is clear. For $(iii) \Rightarrow (iv)$, set $F := E \setminus G$ and for $(iv) \Rightarrow (iii)$, set $G := E \setminus F$.
$(iii) \Rightarrow (i)$: Let $f \ge 0$ be continuous. By Proposition 6.10 and Fatou's lemma,

$$\begin{aligned}\mathbf{P}[f(X)] &= \int_0^\infty \mathbf{P}(f(X) > t)dt \le \int_0^\infty \liminf_{n\to\infty} \mathbf{P}(f(X_n) > t)dt \\ &\le \liminf_{n\to\infty} \int_0^\infty \mathbf{P}(f(X_n) > t)dt = \liminf_{n\to\infty} \mathbf{P}[f(X_n)].\end{aligned}$$

For $-c < f < c$, since $-f + c \ge is 0$,

$$\begin{aligned}\limsup_{n\to\infty} \mathbf{P}[f(X_n)] &= c - \liminf_{n\to\infty} \mathbf{P}[-f(X_n) + c] \le c - \mathbf{P}[-f(X) + c] = \mathbf{P}[f(X)] \\ &\le \liminf_{n\to\infty} \mathbf{P}[f(X_n)],\end{aligned}$$

thus $\mathbf{P}[f(X_n)] \xrightarrow{n\to\infty} \mathbf{P}[f(X)]$.
$(iii), (iv) \to (v)$ For $B \in \mathcal{B}(E)$,

$$\mathbf{P}(X \in B^\circ) \le \liminf_{n\to\infty} \mathbf{P}(X_n \in B^\circ) \le \limsup_{n\to\infty} \mathbf{P}(X_n \in \overline{B}) \le \mathbf{P}(X \in \overline{B}).$$

Given $\mathbf{P}(X \in \partial B) = \mathbf{P}(X \in \overline{B}) - \mathbf{P}(X \in B^\circ) = 0$, therefore $\mathbf{P}(X_n \in B) \xrightarrow{n\to\infty} \mathbf{P}(X \in B)$.
$(v) \to (iv)$: Assume (v) is true and $F \subseteq E$ is closed. We write $F^\varepsilon := \{x \in E : r(x,F) \le \varepsilon\}$ for $\varepsilon > 0$. The sets $\partial F^\varepsilon \subseteq \{x : r(x,F) = \varepsilon\}$ are disjoint, so

$$\mathbf{P}(X \in \partial F^\varepsilon) = 0 \tag{9.3}$$

for Lebesgue-almost every ε. Let $\varepsilon_1, \varepsilon_2, \ldots$ denote a sequence with $\varepsilon_k \downarrow 0$ such that (9.3) holds for all $\varepsilon_1, \varepsilon_2, \ldots$. This means that

$$\limsup_{n\to\infty} \mathbf{P}(X_n \in F) \le \inf_{k=1,2,\ldots} \limsup_{n\to\infty} \mathbf{P}(X_n \in F^{\varepsilon_k}) = \inf_{k=1,2,\ldots} \mathbf{P}(X \in F^{\varepsilon_k}) = \mathbf{P}(X \in F).$$

□

Corollary 9.7 (Convergence of distribution functions)**.** *Let* $\mathbf{P}, \mathbf{P}_1, \mathbf{P}_2, \cdots \in \mathcal{P}(\mathbb{R})$ *with distribution functions* $F, F_1, F_2, \ldots$ *Then* $\mathbf{P}_n \xrightarrow{n\to\infty} \mathbf{P}$ *exactly if* $F_n(x) \xrightarrow{n\to\infty} F(x)$ *for all continuity points* x *of* F.

[3] For the closure $\overline{B}$ and the interior B° denote here $\partial B := \overline{B} \setminus B^\circ$ the edge of B.

Proof. '⇒': If x is a continuity point of F, then $\mathbf{P}(\partial(-\infty;x]) = \mathbf{P}(\{x\}) = 0$. This means that – according to Theorem 9.6 (direction $(i) \Rightarrow (v)$) – that

$$F_n(x) = \mathbf{P}_n((-\infty;x]) \xrightarrow{n\to\infty} \mathbf{P}((-\infty;x]) = F(x).$$

'⇐': According to Theorem 9.6 (direction $(ii) \Rightarrow (i)$), it suffices to show that $\mathbf{P}_n[f] \xrightarrow{n\to\infty} \mathbf{P}[f]$ for all bounded, Lipschitz functions f. Wlog, we assume that $|f| \leq 1$ and f has Lipschitz constant 1. For $\varepsilon > 0$ choose $N \in \mathbb{N}$ and continuity points $y_0 < \cdots < y_N$ of F, so that $F(y_0) < \varepsilon$, $F(y_N) > 1 - \varepsilon$ and $y_i - y_{i-1} < \varepsilon$ for $i = 1, \ldots, N$. Then $F_n(y_i) \xrightarrow{n\to\infty} F(y_i)$ and

$$f \leq 1_{(-\infty,y_0]} + 1_{(y_N,\infty)} + \sum_{i=1}^{N-1} (f(y_i) + \varepsilon) 1_{(y_i,y_{i+1}]},$$

as well as

$$\begin{aligned}\limsup_{n\to\infty} \mathbf{P}_n[f] &\leq \limsup_{n\to\infty} F_n(y_0) + 1 - F_n(y_N) + \sum_{i=1}^{N} (f(y_i) + \varepsilon)(F_n(y_i) - F_n(y_{i-1}) \\ &\leq 3\varepsilon + \sum_{i=1}^{N} f(y_i)(F(y_i) - F(y_{i-1})) \leq 4\varepsilon + \mathbf{P}[f].\end{aligned}$$

With $\varepsilon \to 0$ and by replacing f with $1 - f$, we find $\mathbf{P}_n[f] \xrightarrow{n\to\infty} \mathbf{P}[f]$. □

Remark 9.8 (The Theorem of deMoivre-Laplace)**.** *In Example 9.3 we claimed that deMoivre-Laplace's Theorem makes a statement about weak convergence. The Theorem states that for $B(n,p)$-distributed random variables X_n, $n = 1, 2, ...$,*

$$\mathbf{P}\Big(\frac{X_n - np}{\sqrt{np(1-p)}} \leq x\Big) \xrightarrow{n\to\infty} \Phi(x),$$

where Φ is the distribution function of the standard normal distribution. As Corollary 9.7 shows, this means exactly the convergence in distribution to a standard normal distribution.

Corollary 9.9 (Slutzky's Theorem)**.** *Let $X, X_1, X_2, ..., Y_1, Y_2, ...$ be random variables with values in E. If $X_n \xrightarrow{n\to\infty} X$ and $r(X_n, Y_n) \xrightarrow{n\to\infty}_p 0$, then $Y_n \xrightarrow{n\to\infty} X$.*

Proof. Let $f : E \to \mathbb{R}$ be bounded and Lipschitz-continuous with Lipschitz constant L. Then,

$$|f(x) - f(y)| \leq L \cdot r(x,y) \wedge (2||f||_\infty)$$

for all $x, y \in E$. From this,

$$\limsup_{n\to\infty} \mathbf{E}[f(X_n) - f(Y_n)] \leq \limsup_{n\to\infty} \mathbf{E}[L \cdot r(X_n, Y_n) \wedge (2||f||_\infty)] = 0$$

according to Lemma 7.5. Thus,

$$\limsup_{n\to\infty} \big|\mathbf{E}[f(Y_n)] - \mathbf{E}[f(X)]\big| \leq \limsup_{n\to\infty} \big|\mathbf{E}[f(Y_n)] - \mathbf{E}[f(X_n)]\big| + \big|\mathbf{E}[f(X_n)] - \mathbf{E}[f(X)]\big| = 0,$$

and the claimed convergence follows with Theorem 9.6. □

Theorem 9.10 (Continuous mapping theorem)**.** *Let E be separable, (E', r') another metric space and $\varphi : E \to E'$ measurable and $U_\varphi \subseteq E$ the set of discontinuity points of φ.*

1. *If $\mathbf{P}, \mathbf{P}_1, \mathbf{P}_2, \dots \in \mathcal{P}(E)$ and $\mathbf{P}(U_\varphi) = 0$ and $\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \mathbf{P}$, then $\varphi_* \mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \varphi_* \mathbf{P}$.*
2. *If $X, X_1, X_2, \dots$ are random variables with values in E and $\mathbf{P}(X \in U_\varphi) = 0$ and $X_n \overset{n\to\infty}{\Longrightarrow} X$, then also $\varphi(X_n) \overset{n\to\infty}{\Longrightarrow} \varphi(X)$.*

Proof. First, we note that 2. is an application of 1. if one sets $\mathbf{P}_n = (X_n)_* \mathbf{P}$. The set U_φ is Borel-measurable, since

$$U_\varphi^{\delta,\varepsilon} = \{x \in E : \exists y, z \in B_\delta(x), r'(\varphi(y), \varphi(z)) > \varepsilon\}$$

is Borel-measurable (here the separability of E is included) and

$$U_\varphi = \bigcup_{n=1}^{\infty} \bigcap_{k=1}^{\infty} U_\varphi^{1/k, 1/n}.$$

Let $G \subseteq E'$ be open and $x \in \varphi^{-1}(G) \cap U_\varphi^c$. Since φ is continuous in x, there is a $\delta > 0$ with $\varphi(y) \in G$ (i.e. $y \in \varphi^{-1}(G)$) for all y with $r(x, y) < \delta$. Therefore, $\varphi^{-1}(G) \cap U_\varphi^c \subseteq (\varphi^{-1}(G))^\circ$. This follows with Theorem 9.6 (direction $(i) \Rightarrow (iii)$)

$$\begin{aligned}\varphi_* \mathbf{P}(G) &= \mathbf{P}(\varphi^{-1}(G)) = \mathbf{P}(\varphi^{-1}(G) \cap U_\varphi^c) \leq \mathbf{P}((\varphi^{-1}(G))^\circ) \\ &\leq \liminf_{n\to\infty} \mathbf{P}_n((\varphi^{-1}(G))^\circ) \leq \liminf_{n\to\infty} \mathbf{P}_n(\varphi^{-1}(G)) = \liminf_{n\to\infty} \varphi_* \mathbf{P}_n(G).\end{aligned}$$

Again due to theorem 9.6 (direction $(iii) \Rightarrow (i)$), this implies $\varphi_* \mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \varphi_* \mathbf{P}$. □

Apart from the vague convergence, convergence in distribution is the weakest form of convergence. However, there is a connection with almost sure convergence, as the following theorem shows.

Theorem 9.11 (Weak and almost sure convergence, Skorohod)**.** *Let $X, X_1, X_2, \dots$ be random variables with values in a complete and separable space (E, r). Then, $X_n \overset{n\to\infty}{\Longrightarrow} X$ holds if and only if there is a probability space on which random variables $Y, Y_1, Y_2, \dots$ are defined with $Y_n \xrightarrow{n\to\infty}_{fs} Y$ and $Y \overset{d}{=} X, Y_1 \overset{d}{=} X_1, Y_2 \overset{d}{=} X_2, \dots$*

Proof. '⇐': This is clear, since almost sure convergence implies weak convergence (see Proposition 9.5).
'⇒': We extend the probability space on which X is defined, and we set $Y = X$. Let $E = \{1, \dots, m\}$ be finite, U be uniformly distributed on $[0, 1]$ and independent of Y, and $W_1, W_2, \dots$ independent with

$$\mathbf{P}(W_n = k) = \frac{\mathbf{P}(X_n = k) - \mathbf{P}(X = k) \wedge \mathbf{P}(X_n = k)}{1 - \sum_{l=1}^m \mathbf{P}(X = l) \wedge \mathbf{P}(X_n = l)}.$$

We set $Y_n = k$ if either

$$X = k \text{ and } U \leq \frac{\mathbf{P}(X_n = k)}{\mathbf{P}(X = k)}$$

or

$$X = l \text{ and } U > \frac{\mathbf{P}(X_n = l)}{\mathbf{P}(X = l)} \text{ and } W_n = k.$$

Then

$$\begin{aligned}\mathbf{P}(Y_n = k) &= \mathbf{P}(X = k) \cdot \frac{\mathbf{P}(X_n = k)}{\mathbf{P}(X = k)} \wedge 1 \\ &+ \sum_{l=1}^{m} \mathbf{P}(X = l) \cdot \Big(1 - \frac{\mathbf{P}(X_n = l)}{\mathbf{P}(X = l)}\Big)^{+} \frac{\mathbf{P}(X_n = k) - \mathbf{P}(X = k) \wedge \mathbf{P}(X_n = k)}{1 - \sum_{l'=1}^{m} \mathbf{P}(X = l') \wedge \mathbf{P}(X_n = l')} \\ &= \mathbf{P}(X_n = k) \wedge \mathbf{P}(X = k) \\ &\quad + \sum_{l=1}^{m} (\mathbf{P}(X = l) - \mathbf{P}(X_n = l) \wedge \mathbf{P}(X = l)) \frac{\mathbf{P}(X_n = k) - \mathbf{P}(X = k) \wedge \mathbf{P}(X_n = k)}{1 - \sum_{l'=1}^{m} \mathbf{P}(X = l') \wedge \mathbf{P}(X_n = l')} \\ &= \mathbf{P}(X_n = k).\end{aligned}$$

Thus $Y_n \stackrel{d}{=} X_n$. Since according to the condition $\mathbf{P}(X_n = k) \xrightarrow{n\to\infty} \mathbf{P}(X = k)$, the almost sure convergence follows.

For general E, let $p = 1, 2, ...$ and choose a partition of E in sets $B_1, B_2, \dots$ in E with $\mathbf{P}(Y \in \partial B_k) = 0$ and diameter at most 2^{-p}. Choose m large enough, so that $\mathbf{P}(Y \notin B_0) < 2^{-p}$ with $B_0 := E \setminus \bigcup_{k \le m} B_k$. For $k = 1, 2, \dots$, define random variables $\widetilde{Z}, \widetilde{Z}_1, \widetilde{Z}_2, \dots$ such that $\widetilde{Z} = k$ exactly when $Y \in B_k$ and $\widetilde{Z}_n = k$ if $Y_n \in B_k$. Then $\widetilde{Z}_n \xRightarrow{n\to\infty} \widetilde{Z}$. Since $\widetilde{Z}, \widetilde{Z}_1, \dots$ only takes values in a finite set, we can use random variables $Z, Z_1, Z_2, \dots$ with $Z_n \xrightarrow{n\to\infty}_{fs} Z$. Furthermore, let $W_{n,k}$ be random variables with distribution $\mathbf{P}[X_n \in .|X_n \in B_k]$ and $\widetilde{Y}_{n,p} = \sum_k W_{n,k} 1_{Z_n = k}$, so that $\widetilde{Y}_{n,p} \stackrel{d}{=} X_n$ for all n. It is now clear

$$\Big\{ r(\widetilde{Y}_{n,p}, Y) > 2^{-p} \Big\} \subseteq \{Z_n \neq Z\} \cup \{Y \in B_0\}.$$

Since $Z_n \xrightarrow{n\to\infty}_{fs} Z$ and $\mathbf{P}\{Y \in B_0\} < 2^{-p}$, for each p there are numbers $n_1 < n_2 < \dots$ with

$$\mathbf{P}\Big(\bigcup_{n \ge n_p} \big\{ r(\widetilde{Y}_{n,p}, Y) > 2^{-p} \big\} \Big) < 2^{-p}$$

for all p. With the Borel-Cantelli lemma we get

$$\sup_{n \ge n_p} r(\widetilde{Y}_{n,p}, Y) \le 2^{-p}$$

for almost all p. We therefore define $Y_n := \widetilde{Y}_{n,p}$ for $n_p \le n < n_{p+1}$ and note that $X_n \stackrel{d}{=} Y_n \xrightarrow{n\to\infty}_{fs} Y$. □

9.2 Prohorov' Theorem

In this section, we first examine the concept of vague convergence. We will restrict ourselves to the space $E = \mathbb{R}$. (Most of the statements shown here are still valid in locally compact spaces). It is already clear that weak convergence of distributions implies vague convergence (since all continous functions with compact support are bounded), and that the weak convergence is equivalent to the convergence of the distribution functions (Corollary 9.7). The main result here is the theorem of Helly (Theorem 9.13), which states that every sequence of probability measures has a vaguely convergent subsequence.

We then examine the question when a sequence of probability measures also has weakly convergent subsequence. This leads us to the notion of tightness of probability measures and Prohorov's theorem (Theorem 9.19).
As we have already seen in Remark 9.2.1, it can be that the vague limit measure of probability measures is not a probability measure. However, the following result shows that the limit measure has total mass at most 1.

Lemma 9.12 (Mass loss at vague convergence)**.** *Let $\mathbf{P}_1, \mathbf{P}_2, \cdots \in \mathcal{P}(\mathbb{R})$ and μ a measure on $\mathcal{B}(\mathbb{R})$ with $\mathbf{P}_n[f] \xrightarrow{n\to\infty} \mu[f], f \in \mathcal{C}_c(\mathbb{R})$, then $\mu \in \mathcal{P}_{\leq 1}(\mathbb{R})$ applies.*

Proof. Let $f_1, f_2, \cdots \in \mathcal{C}_c(\mathbb{R})$ with $f_k \uparrow 1$. Then with monotonic convergence

$$\mu(\mathbb{R}) = \sup_{k\in\mathbb{N}} \mu[f_k] = \sup_{k\in\mathbb{N}} \limsup_{n\to\infty} \mathbf{P}_n[f_k] \leq 1.$$

. □

Theorem 9.13 (Helly's theorem)**.** *Let $\mathbf{P}_1, \mathbf{P}_2, \cdots \in \mathcal{P}(\mathbb{R})$. Then there is a subsequence $(n_k)_{k=1,2,\dots}$ and a $\mu \in \mathcal{P}_{\leq 1}(\mathbb{R})$ with $\mathbf{P}_{n_k} \xrightarrow{k\to\infty}_v \mu$.*

Proof. Let $F_1, F_2, \dots$ be the distribution functions of $\mathbf{P}_1, \mathbf{P}_2, \dots$ Further, let $(x_1, x_2, \dots)$ be a count of $\mathbb{Q}$. Since $[0,1]$ is compact, for each sequence there is $(F_n(x_i))_{n=1,2,\dots}$ a convergent subsequence. By means of a diagonal argument, there is a sequence $(n_k)_{k=1,2,\dots}$ such that $(F_{n_k}(x_i))_{k=1,2,\dots}$ for all i against a limit $G(x_i)$ converges to $\mathbb{Q}$. We define

$$F(x) := \inf\{G(r) : r \in \mathbb{Q}, r > x\}.$$

Since all F_n and therefore G have non-negative increments, the same applies to F. From the definition of F and the monotonicity of G, it also follows that F is right-continuous. According to Proposition 2.19, there is a measure μ on $\mathbb{R}$ with $\mu((x,y]) = F(y) - F(x)$ for all $x, y \in \mathbb{R}, x \leq y$. It remains to show that $\mathbf{P}_n[f] \xrightarrow{n\to\infty} \mu[f]$ for all $f \in \mathcal{C}_c(\mathbb{R})$. Wlog we can assume that $f \geq is 0$.

It is $F_n(x) \xrightarrow{n\to\infty} F(x)$ at all continuity points x of F by construction. There is a countable set $D \subseteq \mathbb{R}$ such that F is continuous on D^c is continuous. This means that $\mathbf{P}_n(U) \xrightarrow{n\to\infty} \mu(U)$ for all finite unions U of intervals with vertices in D^c. Now let $B \subseteq \mathbb{R}$ be open and bounded. Let $U_1, U_2, \dots$ and $V_1, V_2, \dots$ be sequences of finite unions of open intervals with vertices in C such that such that $U_k \uparrow B$, $V_k \downarrow \overline{B}$. Then,

$$\begin{aligned}\mu(B) &= \lim_{k\to\infty} \mu(U_k) = \lim_{k\to\infty} \liminf_{n\to\infty} \mathbf{P}_n(U_k) \leq \liminf_{n\to\infty} \mathbf{P}_n(B) \\ &\leq \limsup_{n\to\infty} \mathbf{P}_n(B) \leq \lim_{k\to\infty} \limsup_{n\to\infty} \mathbf{P}_n(V_k) = \lim_{k\to\infty} \mu(V_k) = \mu(\overline{B}).\end{aligned}$$

Since $\mu(f = t) > 0$ for at most countably many t, and since $\mathbf{P}_n(f > t) \leq 1_{t\geq ||f||}$, it follows with dominated convergence

$$\begin{aligned}\mu[f] &= \int_0^\infty \mu(f > t)dt \leq \int_0^\infty \liminf_{n\to\infty} \mathbf{P}_n(f > t)dt = \liminf_{n\to\infty} \int_0^\infty \mathbf{P}_n(f > t)dt = \liminf_{n\to\infty} \mathbf{E}_n[f] \\ &\leq \limsup_{n\to\infty} \mathbf{P}_n[f] = \limsup_{n\to\infty} \int_0^\infty \mathbf{P}_n(f > t)dt = \int_0^\infty \limsup_{n\to\infty} \mathbf{P}_n(f > t)dt \leq \int_0^\infty \mu(f \geq t) \\ &= \mu[f].\end{aligned}$$

□

We now return to the case of a general metric space (E, r). To show the existence of accumulation points in the sense of weak convergence, it must be ensured that limit measures are again limit measures are again probability measures. In particular no mass is lost at the boundary crossing as in the case of vague convergence (see Lemma 9.12). Here, the concept of *tightness* is central.

Definition 9.14 (Tightness)**.** *Let $\mathcal{K}$ be the system of all compact sets in E. A family $(\mathbf{P}_i)_{i\in I}$ in $\mathcal{P}(E)$ is* tight*, if*

$$\sup_{K\in\mathcal{K}} \inf_{i\in I} \mathbf{P}_i(K) = 1.$$

A family $(X_i)_{i\in I}$ of E-valued random variables is tight *if $((X_i)_*\mathbf{P})_{i\in I}$ is tight, i.e.*

$$\sup_{K\in\mathcal{K}} \inf_{i\in I} \mathbf{P}(X_i \in K) = 1.$$

Remark 9.15 (Equivalent formulations)**.** *1. The definition of the tightness of a family $(\mathbf{P}_i)_{i\in I}$ in $\mathcal{P}(E)$ is equivalent to the following condition: for all $\varepsilon > 0$ there exists $K \subseteq E$ compact with $\inf_{i\in I} \mathbf{P}_i(K) \geq 1 - \varepsilon$.*

2. If $E = \mathbb{R}^d$, a family $(\mathbf{P}_i)_{i\in I}$ is tight if and only if

$$\sup_{r>0} \inf_{i\in I} \mathbf{P}_i(B_r(0)) = 1,$$

where $B_r(0)$ is the sphere around 0 with radius r.

3. In Lemma 2.9 we have shown that $\mathbf{P} \in \mathcal{P}(E)$ is tight if (E, r) is complete and is separable. It also follows that every finite family of probability measures on the Borel's σ-algebra of a Polish space is tight.

4. Further, a countable family $(\mathbf{P}_i)_{i=1,2,\dots}$ is of probability measures on a Polish space (E, r) is tight if and only if

$$\sup_{K\in\mathcal{K}} \liminf_{i=1,2,\dots} \mathbf{P}_i(K) = 1.$$

Proof. '$\Rightarrow$': This is clear, since $\liminf_{i=1,2,\dots} \mathbf{P}_i(K) \geq \inf_{i=1,2,\dots} \mathbf{P}_i(K) = 1$.

'$\Leftarrow$': Let $\varepsilon > 0$ and K such that $\liminf_{i=1,2,\dots} \mathbf{P}_i(K) \geq 1 - \varepsilon/2$. Choose N such that $\inf_{i=N+1,N+2,\dots} \mathbf{P}_i(K) \geq 1 - \varepsilon$ and $K_1, \dots, K_N$ compact such that $\mathbf{P}_i(K_i) \geq 1 - \varepsilon$ for $i = 1, \dots, N$. Since $\widetilde{K} = K \cup K_1 \cup \cdots \cup K_N$ is compact and $\inf_{i=1,2,\dots} \mathbf{P}_i(\widetilde{K}) \geq 1 - \varepsilon$ the tightness of $(\mathbf{P}_i)_{i=1,2,\dots}$ follows. □

Example 9.16 (Tight sets of probability measures)**.** *1. If E is compact, every family of probability measures on $\mathcal{B}(E)$ is tight.*

2. A family $(X_i)_{i\in I}$ of real-valued random variables with

$$\sup_{i\in I} \mathbf{P}[|X_i|] < \infty,$$

is tight. This is because

$$\inf_{r>0} \sup_{i\in I} \mathbf{P}(|X_i| \geq r) \leq \inf_{r>0} \sup_{i\in I} \frac{\mathbf{P}[|X_i|]}{r} = 0.$$

.

3. The family $(\delta_n)_{n=1,2,\dots}$, where δ_n is the Dirac measure on n, is not tight.

Lemma 9.17 (Vague convergence and tightness)**.** *Let $\mathbf{P}_1, \mathbf{P}_2, \dots \in \mathcal{P}(\mathbb{R})$ and $\mu \in \mathcal{P}_{\leq 1}(\mathbb{R})$ with*

$$\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow}_v \mu.$$

Then

$$\mu(\mathbb{R}) = 1 \qquad \Longleftrightarrow \qquad (\mathbf{P}_n)_{n=1,2,\dots} \text{ is tight}$$

. In this case, $\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \mu$.

Proof. For $r > 0$ choose a $g_r \in \mathcal{C}_c(\mathbb{R})$, $1_{B_r(0)} \leq g_r \leq 1_{B_{r+1}(0)}$. Then $(\mathbf{P}_n)_{n=1,2,\dots}$ is tight if and only if

$$\sup_{r>0} \liminf_{n\to\infty} \mathbf{P}_n[g_r] = 1.$$

'⇒': Since μ is continuous from below, we find

$$1 = \sup_{r>0} \mu(B_r(0)) \leq \sup_{r>0} \mu[g_r] = \sup_{r>0} \liminf_{n\to\infty} \mathbf{P}_n[g_r] \leq 1.$$

'⇐': Let $(\mathbf{P}_n)_{n=1,2,\dots}$ be be tight. Then, from Lemma 9.12,

$$1 \geq \mu(\mathbb{R}) = \sup_{r>0} \mu(B_r(0)) = \sup_{r>0} \mu[g_r] = \sup_{r>0} \liminf_{n\to\infty} \mathbf{P}_n[g_r] = 1.$$

It remains to show the weak convergence. Assuming that $(\mathbf{P}_n)_{n=1,2,\dots}$ is tight and $f \in \mathcal{C}_b(\mathbb{R})$. Then,

$$\begin{aligned}\limsup_{n\to\infty} |\mathbf{P}_n[f] - \mu[f]| &\leq \inf_{r>0} \limsup_{n\to\infty} \big(|\mathbf{P}_n[f - fg_r]| + |\mathbf{P}_n[fg_r]| - \mu[fg_r]| + |\mu[f - fg_r]|\big) \\ &\leq ||f|| \inf_{r>0} \limsup_{n\to\infty} \mathbf{P}_n(B_r(0)^c) + \inf_{r>0} \mu[B_r(0)^c] = 0,\end{aligned}$$

and $\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \mu$ follows. □

Corollary 9.18 (Weak convergence and tightness)**.** *Let $\mathbf{P}, \mathbf{P}_1, \mathbf{P}_2, \dots \in \mathcal{P}(\mathbb{R})$. If $\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \mathbf{P}$, then $(\mathbf{P}_n)_{n\in\mathbb{N}}$ is tight.*

Proof. Since weak convergence of $\mathbf{P}_1, \mathbf{P}_2, \dots$ to $\mathbf{P}$ implies vague convergence, for $\mathbf{P}, \mathbf{P}_1, \mathbf{P}_2, \dots$ the conditions of Lemma 9.17 and $\mathbf{P}(\mathbb{R}) = 1$ are satisfied. Therefore, $(\mathbf{P}_n)_{n\in\mathbb{N}}$ is tight. □

To determine the weak convergence of probability measures Theorem 9.6 is helpful. We now turn to the question whether a sequence of probability measures can have an accumulation point. This means that there is a subsequence that converges weakly to a probability measure.

Theorem 9.19 (Prohorov's theorem)**.** *Let (E, r) be complete and separable and $(\mathbf{P}_i)_{i\in I}$ a family in $\mathcal{P}(E)$. The following are equivalent:*

1. *The family $(\mathbf{P}_i)_{i\in I}$ is relatively compact with respect to the topology of weak convergence, i.e. every sequence in $(\mathbf{P}_i)_{i\in I}$ has a weakly convergent subsequence.*
2. *For every $\varepsilon > 0$ there is an $N \in \mathbb{N}$ and $x_1, \dots, x_N \in E$, so that*

$$\inf_{i\in I} \mathbf{P}_i\Big(\bigcup_{k=1}^{N} B_\varepsilon(x_k)\Big) \geq 1 - \varepsilon.$$

3. The family $(\mathbf{P}_i)_{i\in I}$ is tight.

Proof. Let $x_1, x_2, \dots$ be a dense subsequence in E (which exists since (E,r) is separable).
1. $\Rightarrow$ 2.: Suppose 2. is not true. Then there is $\varepsilon > 0$ and for each $N = 1, 2, \dots$ a $\mathbf{P}_{i_N}$ with $\mathbf{P}_{i_N}\Big(\bigcup_{k=1}^N B_\varepsilon(x_k)\Big) \le 1-\varepsilon$. By relative compactness, there would then be some subsequence $(\mathbf{P}_{i_M})_{M=1,2,\dots}$ which is weakly convergent to some $\mathbf{P} \in \mathcal{P}(E)$. Thus, because of Theorem 9.6 $((i) \Rightarrow (iii))$ we find that

$$1 = \mathbf{P}(E) = \sup_{N\in\mathbb{N}} \mathbf{P}\Big(\bigcup_{k=1}^N B_\varepsilon(x_i)\Big) \le \sup_{N\in\mathbb{N}} \liminf_{M\to\infty} \mathbf{P}_{i_M}\Big(\bigcup_{k=1}^N B_\varepsilon(x_i)\Big) \le 1-\varepsilon,$$

thus a contradiction.
2. $\Rightarrow$ 3.: Let $\varepsilon > 0$. For $j = 1, 2, \dots$ we choose $x_{j1}, \dots, x_{jN_j}$ such that

$$\inf_{i\in I} \mathbf{P}_i\Big(\bigcup_{k=1}^{N_j} B_{\varepsilon 2^{-j}}(x_{jk})\Big) > 1 - \varepsilon 2^{-j}.$$

We further set

$$K := \bigcap_{j=1}^\infty \bigcup_{k=1}^{N_j} B_{\varepsilon 2^{-j}}(x_{jk}).$$

Then $K \subseteq E$ is totally bounded by construction, according to Proposition A.9 therefore relatively compact, so $\overline{K}$ is compact. Furthermore

$$\sup_{i\in I} \mathbf{P}_i(\overline{K}^c) \le \sup_{i\in I} \sum_{j=1}^\infty \mathbf{P}_i\Big(\bigcap_{k=1}^{N_j} (B_{\varepsilon 2^{-j}}(x_{jk}))^c\Big) \le \varepsilon.$$

Thus the family $(\mathbf{P}_i)_{i\in I}$ is tight.
3. $\Rightarrow$ 1.: Let $\mathbf{P}_1, \mathbf{P}_2, \dots$ be a sequence in the family of the family $(\mathbf{P}_i)_{i\in I}$. The aim is to find a convergent subsequence. For this purpose, we choose compact sets $K_1 \subseteq K_2 \subseteq \dots \subseteq E$ with $\inf_{n=1,2,\dots} \mathbf{P}_n(K_j) \ge 1 - 1/j$. Further, we choose the system of compact sets

$$\mathcal{K} := \Big\{ \bigcup_{k=1}^N K_{j_k} \cap \overline{B_{\varepsilon_k}(x_k)} : N, j_k \in \mathbb{N}, \varepsilon_k \in \mathbb{Q}^+ \Big\}.$$

Since $\mathcal{K}$ is countable, we can use a diagonal argument in order to create a subsequence $\mathbf{P}_{n_1}, \mathbf{P}_{n_2}, \dots$ from $\mathbf{P}_1, \mathbf{P}_2, \dots$ so that $\mathbf{P}_{n_k}(A)$ converges for all $A \in \mathcal{K}$. Define the set function μ on $\mathcal{K}$ by

$$\mu(A) = \lim_{k\to\infty} \mathbf{P}_{n_k}(A), \qquad A \in \mathcal{K}.$$

Our goal is to construct a probability measure $\mathbf{P}$, such that, for all open sets B,

$$\mathbf{P}(B) = \sup_{\mathcal{K}\ni A\subseteq B} \mu(A). \tag{9.4}$$

Indeed, if we find such a $\mathbf{P}$, we can write for B open

$$\mathbf{P}(B) = \sup_{\mathcal{K}\ni A\subseteq B} \lim_{k\to\infty} \mathbf{P}_{n_k}(A) \le \liminf_{k\to\infty} \mathbf{P}_{n_k}(B),$$

and $\mathbf{P}_{n_k} \xrightarrow{k\to\infty} \mathbf{P}$ follows by Theorem 9.6. In order to find $\mathbf{P}$, we are going to construct an outer measure γ, and show that the open sets are γ-measurable. Then, $\mathbf{P}$ can be defined via γ on the σ-algebra of all measurbale sets; see Lemma 6.2.

We first extend μ to all open sets (giving rise to β below), and directly construct γ by setting

$$\gamma(C) := \inf_{B\supseteq C \text{ open}} \beta(B), \qquad \beta(B) =: \sup_{\mathcal{K}\ni K\subseteq B} \mu(K).$$

So, β is defined on all open sets, and, by construction, β is monotone, additive, sub-additive, and $\gamma = \beta$ on all open sets.
We claim that

$$\gamma \text{ is an outer measure and all closed sets are } \gamma\text{-measurable.} \tag{9.5}$$

(Recall that C is measurable with respect to the outer measure γ, if $\gamma(S) \geq \gamma(S\cap C)+\gamma(S\cap C^c)$ for all $S \subseteq E$; see Definition 2.1.6 and sub-additivity of γ). Then, we write for B open $\mathbf{P}(B) = \gamma(B) = \beta(B) = \sup_{K\ni A\subseteq B} \mu(A)$, i.e. (9.4) follows.
In order to show (9.5), we proceed in steps:
Step 1: *If $F \subseteq B \cap K$ is closed, with B open and $K \in \mathcal{K}$, then there is $K' \in \mathcal{K}$ with $F \subseteq K' \subseteq B$.*
For each $x \in F$, choose $\varepsilon(x) \in \mathbb{Q}$ such that $B_{\varepsilon(x)}(x) \subseteq B$. Since $(B_{\varepsilon(x)}(x))_{x\in F}$ is an open cover of $F \cap K$, which is compact, there must be a finite subcover, i.e. some $F = F \cap K \subseteq \bigcup_{n=1}^N \overline{B_{\varepsilon(x_n)}(x_n)} \cap K \subseteq B$. We can now read off the required K'.
Step 2: *β is σ-sub-additive on the open sets.*
For finite sub-additivity, let B_1, B_2 be open, and $\mathcal{K} \ni K \subseteq B_1 \cup B_2$. Define

$$F_1 := \{x \in K : r(x, B_1^c) \geq r(x, B_2^c\}, \quad F_2 := \{x \in K : r(x, B_2^c) \geq r(x, B_1^c)\}.$$

Note that $F_1 \subseteq B_1$: Indeed, if $x \in F_1 \subseteq K \subseteq B_1 \cup B_2$ and $x \in B_2 \setminus B_1$, then $0 = r(x, B_1^c) < r(x, B_2^c)$ since B_2^c is closed, which is a contradiction. Analogously, $F_2 \subseteq B_2$.
So, for $i = 1, 2$, we find $F_i \subseteq B_i \cap K$, and we find $K_i \in \mathcal{K}$ with $F_i \subseteq K_i \subseteq B_i$ with Step 1. So, note that $F_1 \cup F_2 = K$, and we can write

$$\mu(K) \leq \mu(K_1 \cup K_2) \leq \mu(K_1) + \mu(K_2) \leq \beta(B_1) + \beta(B_2).$$

Finite sub-additivity follows by taking the supremum over $\mathcal{K} \ni K \subseteq B_1 \cup B_2$ on the left hand side. For σ-sub-additivity, take $\mathcal{K} \ni K \subseteq \bigcup_{n=1}^\infty B_n$. Since K is compact, choose n_0 such that $K \subseteq \bigcup_{n=1}^{n_0} B_n$ and write

$$\mu(K) \leq \beta\Big(\bigcup_{n=1}^{n_0} B_n\Big) \leq \sum_{n=1}^{n_0} \beta(B_n) \leq \sum_{n=1}^{\infty} \beta(B_n).$$

Then, σ-sub-additivity by taking the supremum over $\mathcal{K} \ni K \subseteq \bigcup_{n=1}^\infty B_n$ on the left hand side.
Step 3: *γ is an outer measure.*
Since $\gamma(\emptyset) = 0$ and γ is monotone by construction, it remains to show σ-sub-additivity. If $\S_1, S_2, ... \subseteq E$, let $\varepsilon > 0$ and choose $B_1 \subseteq S_1, B_2 \subseteq S_2, ...$ open with $\beta(B_n) < \gamma(S_n) + \varepsilon/2^n$. Then, using Step 2,

$$\gamma\Big(\bigcup_{n=1}^{\infty} S_n\Big) \leq \beta\Big(\bigcup_{n=1}^{\infty} S_n\Big) \leq \sum_{n=1}^{\infty} \beta(B_n) \leq \varepsilon + \sum_{n=1}^{\infty} \gamma(S_n).$$

The assertion follows by letting $\varepsilon \downarrow 0$.
Step 4: Closed sets are γ-measurable.
It suffices to show

$$\beta(B) \geq \gamma(F \cap B) + \gamma(F^c \cap B)$$

for F closed and B open. Once this is shown, consider an arbitrary S and $B \supseteq S$ open. Then, $\beta(B) \geq \gamma(F \cap B) + \gamma(F^c \cap B) \geq \gamma(F \cap S) + \gamma(F^c \cap S)$ by monotonicity og γ. From here, the assertion follows by taking $\inf_{B \subseteq S \text{ open}}$ on the left hand side.
So, let F be closed and B be open and $\varepsilon > 0$. Choose $K_1, K_2 \in \mathcal{K}$ with $K_1 \subseteq F^c \cap B$ and $K_2 \subseteq K_1^c \cap B$ (in particular, K_1, K_2 are disjoint) with $\mu(K_1) > \beta(F^c \cap B) - \varepsilon$ and $\mu(K_2) > \beta(K_1^c \cap B) - \varepsilon$. Then, since $\beta(K_1^c \cap B) \geq \gamma(F \cap B)$

$$\beta(B) \geq \mu(K_1 \cup K_2) = \mu(K_1) + \mu(K_2) > \gamma(F^c \cap B) + \gamma(K_1^c \cap B) - 2\varepsilon.$$

By letting $\varepsilon \to 0$, this concludes the proof, i.e. (iii)⇒(i) is shown. □

9.3 Separating classes of functions

Now we will introduce separating classes of functions. In particular, this will shed some light on the usefulness of characteristic functions and Laplace transforms of distributions (see Definition 6.11). These are based on two specific classes of functions that are separating.

Definition 9.20 (Classes of functions separating points and separating function classes)**.**

1. *A function class $\mathcal{M} \subseteq \mathcal{C}(E)$ is said to separate points in E if for all $x, y \in E$ with $x \neq y$ there exists an $f \in \mathcal{M}$ with $f(x) \neq f(y)$.*

2. *A class of functions $\mathcal{M} \subseteq \mathcal{C}(E)$ is called separating in $\mathcal{P}(E)$ if from $\mathbf{P}, \mathbf{Q} \in \mathcal{P}(E)$ and*

$$\mathbf{P}[f] = \mathbf{Q}[f] \text{ for all } f \in \mathcal{M}$$

implies that $\mathbf{P} = \mathbf{Q}$.

Example 9.21. *1. The class of functions $\mathcal{M} := \mathcal{C}_b(E)$ is both, separating points and separating. Namely, if $x \neq y$, then $z \mapsto r(x, z) \wedge 1$ is a bounded, continuous function that separates x and y. Furthermore, if $\mathbf{P}, \mathbf{Q} \in \mathcal{P}(E)$ and $\mathbf{P} \neq \mathbf{Q}$, then there is an open ball A with $\mathbf{P}(A) \neq \mathbf{Q}(A)$. Let $f_1, f_2, \ldots$ be a sequence in $\mathcal{C}_b(E)$ with $f_n \uparrow 1_A$. If $\mathbf{P}[f_n] = \mathbf{Q}[f_n]$ for all $n = 1, 2, \ldots$, then it would also*

$$\mathbf{P}(A) = \lim_{n \to \infty} \mathbf{P}[f_n] = \lim_{n \to \infty} \mathbf{Q}[f_n] = \mathbf{Q}(A)$$

in contradiction to the assumption.

2. The class of functions $\{x \mapsto cx : c \in \mathbb{R}\}$ of all linear functions separates points, but is not separating.

The next result requires the Stone-Weierstrass theorem, which we repeat first.

Definition 9.22 (Algebra)**.** *A set system $\mathcal{M} \subseteq \mathcal{C}(E)$ is called an algebra, if $1 \in \mathcal{M}$, and if $\alpha, \beta \in \mathbb{R}$ and it contains f, g it also contains $\alpha f + \beta g$, as well as fg.*

Theorem 9.23 (Stone-Weierstrass)**.** *Let (E, r) be compact and $\mathcal{M} \subseteq C_b(E)$ an algebra separating points. Then, $\mathcal{M}$ is dense in $\mathcal{C}_b(E)$ with respect to the supremum norm.*

Proof. See some lecture on *Analysis*. □

Theorem 9.24 (Algebras separating points and separating algebras)**.**
Let (E, r) be complete and separable. If $\mathcal{M} \subseteq \mathcal{C}_b(E)$ separates points and is such that $f, g \in \mathcal{M}$ implies $fg \in \mathcal{M}$. Then $\mathcal{M}$ is is separating.

Proof. Let $\mathbf{P}, \mathbf{Q} \in \mathcal{P}(E)$. Without restriction, $1 \in \mathcal{M}$, since $\mathbf{P}[1] = \mathbf{Q}[1]$ always holds. Thus $\mathcal{M}$ is wlog an algebra. Let $\varepsilon > 0$ and K be compact such that $\mathbf{P}(K) > 1-\varepsilon$, $\mathbf{Q}(K) > 1-\varepsilon$. For $g \in \mathcal{C}_b(E)$, according to the Stone-Weierstrass Theorem 9.23 there is a sequence $(g_n)_{n=1,2,\dots}$ in $\mathcal{M}$ with

$$\sup_{x \in K} |g_n(x) - g(x)| \xrightarrow{n \to \infty} 0. \tag{9.6}$$

Now,

$$\begin{aligned}
\left|\mathbf{P}[ge^{-\varepsilon g^2}] - \mathbf{Q}[ge^{-\varepsilon g^2}]\right| \leq{}& \left|\mathbf{P}[ge^{-\varepsilon g^2}] - \mathbf{P}[ge^{-\varepsilon g^2}; K]\right| \\
&+ \left|\mathbf{P}[ge^{-\varepsilon g^2}; K] - \mathbf{P}[g_n e^{-\varepsilon g_n^2}; K]\right| \\
&+ \left|\mathbf{P}[g_n e^{-\varepsilon g_n^2}; K] - \mathbf{P}[g_n e^{-\varepsilon g_n^2}]\right| \\
&+ \left|\mathbf{P}[g_n e^{-\varepsilon g_n^2}] - \mathbf{Q}[g_n e^{-\varepsilon g_n^2}]\right| \\
&+ \left|\mathbf{Q}[g_n e^{-\varepsilon g_n^2}] - \mathbf{Q}[g_n e^{-\varepsilon g_n^2}; K]\right| \\
&+ \left|\mathbf{Q}[g_n e^{-\varepsilon g_n^2}] - \mathbf{Q}[ge^{-\varepsilon g^2}; K]\right| \\
&+ \left|\mathbf{Q}[ge^{-\varepsilon g^2}; K] - \mathbf{Q}[ge^{-\varepsilon g^2}]\right|
\end{aligned}$$

We restrict the first term by

$$\left|\mathbf{P}[ge^{-\varepsilon g^2}] - \mathbf{P}[ge^{-\varepsilon g^2}; K]\right| \leq \frac{C}{\sqrt{\varepsilon}} \mathbf{P}(K^c) \leq C\sqrt{\varepsilon}$$

with $C = \sup_{x \geq 0} x e^{-x^2}$; analogous to the third, fifth and last terms. The second and penultimate terms converge to 0 for $n \to \infty$ due to (9.6). Since $\mathcal{M}$ is an algebra, $g_n e^{-\varepsilon g_n^2}$ can be approximated by functions in $\mathcal{M}$, which means that the fourth term for $n \to \infty$ converges to 0. This means that

$$\left|\mathbf{P}[g] - \mathbf{Q}[g]\right| = \lim_{\varepsilon \to 0} \left|\mathbf{P}[ge^{-\varepsilon g^2}] - \mathbf{Q}[ge^{-\varepsilon g^2}]\right| \leq 4C \lim_{\varepsilon \to 0} \sqrt{\varepsilon} = 0.$$

Since g was arbitrary and $\mathcal{C}_b(E)$ is separating, $\mathbf{P} = \mathbf{Q}$ follows. □

We now come back to the characteristic function and the Laplace transform. As already mentioned, the usefulness of the characteristic function and the Laplace transforms is due to the fact that they are distribution-determining.

Proposition 9.25 (Characteristic function distribution-determining)**.**
A probability measure $\mathbf{P} \in \mathcal{P}(\mathbb{R}^d)$ ($\mathbf{P} \in \mathcal{P}(\mathbb{R}^d_+)$) is uniquely characterized by the characteristic function $\psi_{\mathbf{P}}$ (the Laplace transform $\mathscr{L}_{\mathbf{P}}$).

Proof. We show the statement only for characteristic functions that are proven for Laplace transforms is proven analogously. We establish that the set $\mathcal{M} := \{x \mapsto e^{itx}; t \in \mathbb{R}^d\}$ in $\mathbb{R}^d$ separates points. Since $\mathcal{M} \subseteq \mathcal{C}_b(\mathbb{R}^d)$ and is closed under product formation, it is also separating according to theorem 9.24. This finishes the proof. □

Corollary 9.26 (Independence and characteristic function)**.** *1. A family $(X_j)_{j\in I}$ of real-valued random variables is is independent if and only if for all $J \subseteq_f I$*

$$\mathbf{E}\Big[\prod_{j\in J} e^{it_j X_j}\Big] = \prod_{j\in J} \mathbf{E}[e^{it_j X_j}] \tag{9.7}$$

for all $(t_j)_{j\in J} \in \mathbb{R}^J$ is valid.

2. A family $(X_j)_{j\in I}$ of random variables with values in $\mathbb{R}_+$ is independent if and only if for all $J \subseteq_f I$

$$\mathbf{E}\Big[\prod_{j\in I} e^{-t_j X_j}\Big] = \prod_{j\in J} \mathbf{E}[e^{-t_j X_j}]$$

for all $(t_j)_{j\in J} \in \mathbb{R}^J$ applies.

Proof. We only show the first statement, the second follows analogously. If $(X_j)_{j\in I}$ is independent, then according to Lemma 8.4, the random variables $(e^{it_j X_j})_{j\in I}$ for all $(t_j)_{j\in J} \in \mathbb{R}^J$ are independent. Thus, (9.7) follows from Proposition 8.5. Conversely, the following applies. On the one hand, the left-hand side of (9.7) represents the characteristic function of the distribution $((X_j)_{j\in J})_*\mathbf{P}$. On the other hand, the right side of (9.7) is the characteristic function of $\bigotimes_{j\in J}(X_j)_*\mathbf{P}$. Since the characteristic function according to Proposition 9.25 is the joint distribution of $(X_j)_{j\in J}$ is uniquely determined, $((X_j)_{j\in J})_*\mathbf{P} = \bigotimes_{j\in J}(X_j)_*\mathbf{P}$. The independence of $(X_j)_{j\in I}$ thus follows from Proposition 8.2. □

9.4 Lévy's theorem

We now want to analyze the relationship between weak convergence and the convergence of the characteristic functions of the underlying distributions. Let $\mathbf{P}, \mathbf{P}_1, \mathbf{P}_2, \cdots \in \mathcal{P}(\mathbb{R}^d)$. How to get from Proposition 9.27, the weak convergence follows $\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \mathbf{P}$ follows from the pointwise convergence of the characteristic functions, $\psi_{\mathbf{P}_n}(t) \overset{n\to\infty}{\longrightarrow} \psi_{\mathbf{P}}(t)$, $t \in \mathbb{R}^d$, given $(\mathbf{P}_n)_{n\in\mathbb{N}}$ is tight. The decisive factor is that the tightness of the family $(\mathbf{P}_n)_{n\in\mathbb{N}}$ can also be read from the characteristic functions as we will show in Proposition 9.32. This leads to the statement of Lévy's continuity theorem (Theorem 9.33), which states when the pointwise limit of characteristic functions is again a characteristic function of a probability measure.

Proposition 9.27 (Separating class of functions and weak convergence)**.** *Let (E, r) be complete and separable and $\mathbf{P}, \mathbf{P}_1, \mathbf{P}_2, \cdots \in \mathcal{P}(E)$. Then the following are equivalent:*

1. $\mathbf{P}_n \overset{n\to\infty}{\Longrightarrow} \mathbf{P}$.

2. $(\mathbf{P}_n)_{n=1,2,\ldots}$ is tight and there is a separating family $\mathcal{M} \subseteq \mathcal{C}_b(E)$ with

$$\mathbf{P}_n[f] \overset{n\to\infty}{\longrightarrow} \mathbf{P}[f] \text{ for all } f \in \mathcal{M}.$$

Proof. 1. $\Rightarrow$ 2. According to Corollary 9.18, we have that $(\mathbf{P}_n)_{n=1,2,\dots}$ is tight. The second part of 2. holds because of the definition of weak convergence.
2. $\Rightarrow$ 1. Suppose $(\mathbf{P}_n)_{n=1,2,\dots}$ is tight and $\mathbf{P}_1, \mathbf{P}_2, \dots$ does not converge weakly to $\mathbf{P}$. Then there is $\varepsilon > 0$, some $f \in \mathcal{C}_b(E)$ and a subsequence $(n_k)_{k=1,2,\dots}$ such that

$$\mathbf{P}_{n_k}[f] - \mathbf{P}[f]| > \varepsilon \text{ for all } k. \tag{9.8}$$

According to theorem 9.19 there is a subsequence $(n_{k_\ell})_{\ell=1,2,\dots}$ and a $\mathbf{Q} \in \mathcal{P}(E)$, such that $\mathbf{P}_{n_{k_\ell}} \overset{\ell\to\infty}{\Longrightarrow} \mathbf{Q}$. Because of (9.8),

$$|\mathbf{P}[f] - \mathbf{Q}[f]| \geq |\liminf_{\ell\to\infty} \big(\mathbf{P}[f] - \mathbf{P}_{n_{k_\ell}}[f]|) + \liminf_{\ell\to\infty}(\mathbf{P}_{n_{k_\ell}}[f] - \mathbf{Q}[f])| > \varepsilon,$$

in particular $\mathbf{P} \neq \mathbf{Q}$. On the other hand, for all $g \in \mathcal{M}$ we have

$$\mathbf{P}[g] = \lim_{\ell\to\infty} \mathbf{P}_{n_{k_\ell}}[g] = \mathbf{Q}[g].$$

Since $\mathcal{M}$ is separating, this is a contradiction and 1. is shown. □

Let $\mathbf{P} \in \mathcal{P}(\mathbb{R})$ and $\psi_{\mathbf{P}}$ be its characteristic function. We first show an estimate, which is important to relate tightness and $\psi_{\mathbf{P}}$.

Lemma 9.28 (Tightness and the characteristic function function)**.** *Let $\mathbf{P} \in \mathcal{P}(\mathbb{R})$. Then for all $r > 0$*

$$\mathbf{P}((-\infty; -r] \cup [r; \infty)) \leq \frac{r}{2} \int_{-2/r}^{2/r} (1 - \psi_{\mathbf{P}}(t))dt, \tag{9.9}$$

Proof. It is $\sin(x)/x \leq 1$ for $x \leq 2$ and $\sin x \leq x/2$ for $x \geq 2$. Let X be a random variable with distribution $\mathbf{P}$. Therefore, for every $c > 0$ according to Fubini,

$$\begin{aligned}
\int_{-c}^{c} (1 - \psi_{\mathbf{P}}(t))dt &= \mathbf{P}\Big[\int_{-c}^{c} (1 - e^{itX})dt\Big] = \mathbf{P}\Big[2c - \frac{1}{iX} e^{itX}\Big|_{t=-c}^{c}\Big] \\
&= 2c\mathbf{P}\Big[1 - \frac{\sin(cX)}{cX}\Big] \\
&\geq 2c\mathbf{P}\Big[1 - \frac{\sin(cX)}{cX}; |cX| \geq 2\Big] \\
&\geq c \cdot \mathbf{P}(|cX| \geq 2) = c\mathbf{P}((-\infty; -\tfrac{2}{c}] \cup [\tfrac{2}{c}; \infty)),
\end{aligned}$$

and the assertion follows with $c = 2/r$. □

Definition 9.29 (Uniform continuity)**.** *We repeat a definition from calculus. A set $\mathcal{M} \subseteq \mathcal{C}(\mathbb{R}^d)$ is called uniformly continuous in $x \in \mathbb{R}^d$ if*

$$\sup_{f\in\mathcal{M}} |f(y) - f(x)| \xrightarrow{y\to x} 0.$$

Remark 9.30 (Equivalent condition for sequences)**.** *If $\mathcal{M} = \{f_1, f_2, \dots\}$, then the condition*

$$\limsup_{n\to\infty} |f_n(y) - f_n(x)| \xrightarrow{y\to x} 0$$

equivalent.

Lemma 9.31 (Uniform integrability and convergence)**.** *Let $f_1, f_2, \cdots \in \mathcal{C}(\mathbb{R}^d)$, so that $f_n \xrightarrow{n\to\infty} f$ pointwise for a function $f : \mathbb{R}^d \to \mathbb{R}$. Then f is continuous in 0 iff $(f_n)_{n=1,2,\dots}$ is uniformly continuous in 0.*

Proof. If $(f_n)_{n=1,2,\dots}$ is uniformly continuous in 0, then

$$|f(t) - f(0)| = |\lim_{n\to\infty}(f_n(t) - f_n(0))| \le \limsup_{n\to\infty} |f_n(t) - f_n(0)| \xrightarrow{t\to 0} 0.$$

Conversely, if f is continuous in 0, then

$$\limsup_{n\to\infty} |f_n(t)-f_n(0)| \le \limsup_{n\to\infty} |f_n(t)-f(t)|+|f(t)-f(0)|+|f(0)-f_n(0) = |f(t)-f(0)| \xrightarrow{t\to 0} 0.$$

□

Proposition 9.32 (Tightness and uniformity continuity)**.** *Let $(\mathbf{P}_i)_{i\in I}$ be a family in $\mathcal{P}(\mathbb{R}^d)$. If $(\psi_{\mathbf{P}_i})_{i\in I}$ is uniformly continuous in 0, then $(\mathbf{P}_i)_{i\in I}$ is tight.*

Proof. It suffices to show that $((\pi_k)_*\mathbf{P}_i)_{i\in I}$ is tight for all projections $\pi_1, \dots, \pi_d$. Apparently, $\psi_{(\pi_k)_*\mathbf{P}_i}(t) = \psi_{\mathbf{P}_i}(te_k)$, if e_k is the k-th unit vector. It is therefore sufficient to prove the assertion in the case $d = 1$. Since $\psi_{\mathbf{P}_i}(0) = 1$ for all $i \in I$, we conclude from uniform continuity that

$$\sup_{i\in I} |1 - \psi_{\mathbf{P}_i}(t)| \xrightarrow{t\to 0} 0,$$

thus, see Remark 9.15,

$$\begin{aligned}
\sup_{r>0} \inf_{i\in I} \mathbf{P}_i([-r; r]) &\ge 1 - \inf_{r>0} \sup_{i\in I} \frac{r}{2} \int_{-2/r}^{2/r} (1 - \psi_{\mathbf{P}_i}(t))dt \\
&\ge 1 - \inf_{r>0} \frac{r}{2} \int_{-2/r}^{2/r} \sup_{i\in I} |1 - \psi_{\mathbf{P}_i}(t)|dt \\
&\ge 1 - 2 \inf_{r>0} \sup_{t\in[0;2/r]} \sup_{i\in I} |1 - \psi_{\mathbf{P}_i}(t)| = 1.
\end{aligned}$$

This shows the assertion. □

Theorem 9.33 (Lévy's continuity theorem)**.** *Let $\mathbf{P}_1, \mathbf{P}_2, \cdots \in \mathcal{P}(\mathbb{R}^d)$ and $\psi : \mathbb{R}^d \to \mathbb{C}$, so that $\psi_{\mathbf{P}_n}(t) \xrightarrow{n\to\infty} \psi(t)$ for all $t \in \mathbb{R}^d$. If ψ is continuous in 0, then $\mathbf{P}_n \xRightarrow{n\to\infty} \mathbf{P}$ for a $\mathbf{P} \in \mathcal{P}(\mathbb{R}^d)$ with $\psi_{\mathbf{P}} = \psi$.*

Proof. Since $\psi_{\mathbf{P}_n}$ converges pointwise to a function ψ which is continuous in 0, it follows from Lemma 9.31 that $(\psi_{\mathbf{P}_n})_{n=1,2,\dots}$ is uniformly continuous in 0. With Proposition 9.32 it follows that $(\mathbf{P}_n)_{n=1,2,\dots}$ is tight. Let $(n_k)_{k=1,2,\dots}$ be a subsequence and $\mathbf{P} \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathbf{P}_{n_k} \xRightarrow{k\to\infty} \mathbf{P}$. Since $x \mapsto e^{itx}$ is a continuous, bounded function, it follows that $\psi_{\mathbf{P}_{n_k}}(t) \xrightarrow{k\to\infty} \psi_{\mathbf{P}}(t)$ for all $t \in \mathbb{R}^d$. On the other hand, since $\psi_{\mathbf{P}_n}(t) \xrightarrow{n\to\infty} \psi(t)$, and $\psi_{\mathbf{P}} = \psi$ follows. This identifies ψ as a characteristic function of $\mathbf{P}$ and since this uniquely determines $\mathbf{P}$, we find $\mathbf{P}_n \xRightarrow{n\to\infty} \mathbf{P}$. □

Example 9.34 (Theorem of deMoivre-Laplace)**.** *Let $S_n \sim B(n,p)$. The Theorem of deMoivre-Laplace states that*

$$S_n^* := \frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{n\to\infty} N(0,1). \tag{9.10}$$

We now want to show this again with the help of characteristic functions, i.e. $\psi_{S_n^*} \xrightarrow{n\to\infty} \psi_{N(0,1)}$ pointwise. To do this, we use Proposition 6.12.3 and write with $q := 1-p$ and $C_1, C_2, \dots \in \mathbb{C}$ with $\limsup_{n\to\infty} |C_n| < \infty$

$$\begin{aligned}
\psi_{S_n^*}(t) &= \exp\Big(-it\sqrt{\frac{np}{q}}\Big) \cdot \psi_{B(n,p)}\Big(\frac{t}{\sqrt{npq}}\Big) \\
&= \exp\Big(-it\sqrt{\frac{np}{q}}\Big)\Big(q + p\exp\Big(\frac{it}{\sqrt{npq}}\Big)\Big)^n \\
&= \Big(q\exp\Big(-it\sqrt{\frac{p}{nq}}\Big) + p\exp\Big(it\sqrt{\frac{q}{np}}\Big)\Big)^n \\
&= \Big(1 - qit\sqrt{\frac{p}{nq}} - q\frac{t^2}{2}\frac{p}{nq} + pit\sqrt{\frac{q}{np}} - p\frac{t^2}{2}\frac{q}{np} + \frac{C_n}{n^{3/2}}\Big)^n \\
&= \Big(1 - \frac{t^2}{2}\frac{1}{n} + \frac{C_n}{n^{3/2}}\Big)^n \xrightarrow{n\to\infty} e^{-\frac{t^2}{2}} = \psi_{N(0,1)}(t).
\end{aligned}$$

The result now follows from Theorem 9.33.

Lévy's continuity theorem can also be formulated with Laplace transforms. We state the theorem without proof:

Theorem 9.35 (Lévy's continuity theorem for Laplace transforms)**.** *Let $\mathbf{P}_1, \mathbf{P}_2, \dots \in \mathcal{P}(\mathbb{R}_+^d)$ and $\mathscr{L} : \mathbb{R}^d \to [0,1]$, so that $\mathscr{L}_{\mathbf{P}_n}(t) \xrightarrow{n\to\infty} \mathscr{L}(t)$ for all $t \in \mathbb{R}^d$. If $\mathscr{L}$ is continuous in 0, then $\mathbf{P}_n \xrightarrow{n\to\infty} \mathbf{P}$ for a $\mathbf{P} \in \mathcal{P}(\mathbb{R}^d)$ with $\mathscr{L}_{\mathbf{P}} = \mathscr{L}$.*

Example 9.36 (Convergence of the geometric to the exponential distribution)**.** *Let $X_n \sim \mu_{geo(p_n)}$ be distributed and $n \cdot p_n \xrightarrow{n\to\infty} \lambda$. Then*

$$\begin{aligned}
\mathscr{L}_{X_n/n}(t) = \mathbf{P}[e^{-tX_n/n}] &= \sum_{k=1}^{\infty} (1-p_n)^{k-1} p_n e^{-tk/n} \\
&= p_n e^{-t/n} \frac{1}{1-(1-p_n)e^{-t/n}} \\
&= \frac{\lambda}{n(1-(1-\lambda/n)(1-t/n))} + o(1/n) \\
&\xrightarrow{n\to\infty} \frac{\lambda}{\lambda+t}.
\end{aligned}$$

Therefore, $\frac{X_n}{n} \xrightarrow{n\to\infty} Y$, where $Y \sim \mu_{exp(\lambda)}$, since

$$\mathscr{L}_{exp(\lambda)}(t) = \int_0^\infty \lambda e^{-\lambda a} e^{-ta} da = \frac{\lambda}{\lambda+t}.$$

10 Weak limit laws

We will now apply our knowledge of weak convergence and characteristic functions in special situations. In Section 10.1 we are concerned with statements about when the sum of random variables converges against a Poisson distributed random variable. In section 10.2 we will apply the central Lindeberg-Feller's central limit theorem, which provides a characterization for the weak convergence against a normal distribution. Section 10.3 finally deals with extensions for the case of multidimensional random variables.

10.1 Poisson convergence

We already know the statement that $B(n, p_n)$ for $n \cdot p_n \xrightarrow{n\to\infty} \lambda$ converges weakly against $\mathrm{Poi}(\lambda)$ for large n; see Example 10.1. In this section we generalize this statement; see Theorem 10.5.

Example 10.1 (Poisson approximation of the binomial distribution)**.** *Let $p_1, p_2, \cdots \in [0,1]$ be such that $n \cdot p_n \xrightarrow{n\to\infty} \lambda$. Then we already know from* Basic probability *that*

$$B(n, p_n)(\{k\}) \xrightarrow{n\to\infty} Poi(\lambda)(\{k\}).$$

In other words, this is a statement about weak convergence:

$$B(n, p_n) \xRightarrow{n\to\infty} Poi(\lambda). \tag{10.1}$$

Lévy's theorem provides another way to prove this result. We recall the characteristic functions of the binomial and Poisson distribution from Example 6.13. We write directly

$$\begin{aligned}\psi_{B(n,p_n)}(t) &= \Big(1 - p_n\big(1 - e^{it}\big)\Big)^n \\ &= \Big(1 - \frac{n \cdot p_n}{n}\big(1 - e^{it}\big)\Big)^n \\ &\xrightarrow{n\to\infty} \exp\big(-\lambda(1 - e^{it})\big) = \psi_{Poi(\lambda)}(t).\end{aligned}$$

In particular, the characteristic functions of the binomial distributions converge pointwise to a function that is continuous in 0, namely the characteristic function of the Poisson distribution. With Theorem 9.33 this imples (10.1)*.*

In the following, we will see that the weak convergence to a Poisson distribution is even more general. For this we will use use generating functions.

Remark 10.2 (Generating function)**.** *Consider a random variable X with values in $\mathbb{Z}_+$ and define the* generating function

$$z \mapsto \varphi_X(z) := \mathbf{P}[z^X] = \sum_{k=0}^{\infty} z^k \mathbf{P}[X = k].$$

We note that for $z \in [0,1]$ this is related to the Laplace transform of X because (with $z = e^{-t}$)

$$\mathscr{L}_X(t) = \mathbf{P}[e^{-tX}] = \mathbf{P}[z^X] = \varphi_X(z).$$

In particular, the following two properties of Laplace transforms carry ofer to generating functions.

1. Generating functions determine the distribution, see Proposition 9.25: *The distribution of X is uniquely determined by $z \mapsto \varphi_X(z)$ for $z \in [0,1]$.*

2. Weak convergence equivalent to the convergence of the generating functions, see Theorem 9.33: *Let $X_1, X_2, \dots$ be a sequence of random variables with values in $\mathbb{Z}_+$ such that $\varphi_{X_n}(z) \xrightarrow{n\to\infty} \varphi(z)$ for $z \in [0,1]$ for a function φ that is continuous from below in 1. Then $X_n \xRightarrow{n\to\infty} X$ for a random variable X with generating function φ.*

Sometimes generating functions are practical tools. By their definition, they are power series with radius of convergence $r \geq 1$. It is known that inside the radius of convergence, the derivative and sum are interchanged. So if $r > 1$, for example, we write

$$\varphi'_X(1) = \sum_{k=0}^{\infty} k z^{k-1} \mathbf{P}(X = k)\Big|_{z=1} = \sum_{k=0}^{\infty} k \mathbf{P}(X = k) = \mathbf{P}[X].$$

Analogous calculations for higher derivatives are also possible.

Definition 10.3 (Asymptotic negligibility)**.** *A triagonal family of random variables $(X_{nj})_{n=1,2,\dots,n,j=1,\dots,m_n}$ with $m_1, m_2, \dots \in \mathbb{N}$ is* asymptotically negligible*, if the random variables $X_{n1}, \dots, X_{n,m_n}$ are independent for each $n = 1, 2, \dots$, and*

$$\sup_{j=1,\dots,m_n} \mathbf{P}(|X_{nj}| > \varepsilon) \xrightarrow{n\to\infty} 0 \tag{10.2}$$

for all $\varepsilon > 0$. If $X_{ij} \geq 0$ for all i, j, then $m_n = \infty$ is also permitted.

Remark 10.4 (Equivalent formulation)**.** 1. *For a triagonal family of random variables $(X_{nj})_{n=1,2,\dots,n,j=1,\dots,m_n}$, (10.2) holds iff*

$$\sup_{j=1,\dots,m_n} \mathbf{E}[|X_{nj}| \wedge 1] \xrightarrow{n\to\infty} 0.$$

2. *Let $(X_{nj})_{n=1,2,\dots,n,j=1,\dots,m_n}$ be a triangular of $\mathbb{Z}_+$-valued random variables. Then (10.2) holds iff*

$$\inf_{z\in[0,1]} \inf_{j=1,\dots,m_n} \varphi_{X_{nj}}(z) = \inf_{j=1,\dots,m_n} \varphi_{X_{nj}}(0) = \inf_{j=1,\dots,m_n} \mathbf{P}(|X_{nj}| = 0) \xrightarrow{n\to\infty} 1. \tag{10.3}$$

Theorem 10.5 (Poisson convergence)**.** *Let $(X_{nj})_{n=1,2,\dots,n,j=1,\dots,m_n}$ be a family of asymptotically negligible random variables with values in $\mathbb{Z}_+$ and $X \sim Poi(\lambda)$. Then,*

$$\sum_{j=1}^{m_n} X_{nj} \xRightarrow{n\to\infty} X$$

iff

1. $\displaystyle\sum_{j=1}^{m_n} \mathbf{P}(X_{nj} > 1) \xrightarrow{n\to\infty} 0$

2. $\displaystyle\sum_{j=1}^{m_n} \mathbf{P}(X_{nj} = 1) \xrightarrow{n\to\infty} \lambda.$

We prepare the proof with a lemma.

Lemma 10.6. *Let $(\lambda_{nj})_{n=1,2,\dots,j=1,\dots,m_n}$ be a triangular family of asymptotically negligible, non-negative constants and $\lambda \in [0;\infty]$. Then,*

$$\prod_{j=1}^{m_n}(1-\lambda_{nj}) \xrightarrow{n\to\infty} e^{-\lambda} \qquad \Longleftrightarrow \qquad \sum_{j=1}^{m_n}\lambda_{nj} \xrightarrow{n\to\infty} \lambda$$

Proof. First note that $\log(1-x) = -x + \varepsilon(x)$ for $x > 0$ with $\varepsilon(x)/x \xrightarrow{x\to 0} 0$. Since $\sup_{j=1,\dots,m_n}\lambda_{nj} < 1$ for large n, the left hand side is equivalent to

$$-\lambda = \lim_{n\to\infty}\sum_{j=1}^{m_n}\log(1-\lambda_{nj}) = -\lim_{n\to\infty}\sum_{j=1}^{m_n}\lambda_{nj}\big(1-\tfrac{\varepsilon(\lambda_{nj})}{\lambda_{nj}}\big) = -\lim_{n\to\infty}\sum_{j=1}^{m_n}\lambda_{nj},$$

as

$$\sup_{j=1,\dots,m_n}\tfrac{\varepsilon(\lambda_{nj})}{\lambda_{nj}} \xrightarrow{n\to\infty} 0.$$

From this, the right hand side is immediate. □

Proof of Theorem 10.5. We denote by $\varphi_{n,j}$ the generating function of $X_{n,j}$. According to Remark 10.2.2, the weak convergence in the theorem is equivalent to pointwise convergence of $\prod_{j=1}^{m_n}\varphi_{nj}(z) \xrightarrow{n\to\infty} e^{-\lambda(1-z)}$, since

$$\varphi_X(z) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}z^k = e^{-\lambda(1-z)}.$$

By Lemma 10.6 this is true iff

$$A_n(z) := \sum_{j=1}^{m_n}(1-\varphi_{nj}(z)) \xrightarrow{n\to\infty} \lambda(1-z), \tag{10.4}$$

since the family $(1-\varphi_{nj}(z))_{n=1,2,\dots,j=1,\dots,m_n}$ for each $z \in [0,1]$ after (10.3) is asymptotically negligible. We decompose $A_n(z) = A_n^1(z) + A_n^2(z)$ with

$$A_n^1(z) = \sum_{k=1}^{\infty}(1-z)\sum_{j=1}^{m_n}\mathbf{P}(X_{nj}=k) = (1-z)\sum_{j=1}^{m_n}\mathbf{P}(X_{nj}>0),$$
$$A_n^2(z) = \sum_{k=2}^{\infty}(z-z^k)\sum_{j=1}^{m_n}\mathbf{P}(X_{nj}=k).$$

First, $z(1-z) \le z - z^k \le z$ for all $k = 2,3,\dots$ This means that

$$z(1-z)\sum_{j=1}^{m_n}\mathbf{P}(X_{nj}>1) \le A_n^2(z) \le z\sum_{j=1}^{m_n}\mathbf{P}(X_{nj}>1). \tag{10.5}$$

Let us now turn to the proof of the assertion.

'⇒': Let (10.4) hold. For $z = 0$ this means, since $\varphi_{nj}(0) = \mathbf{P}(X_{nj} = 0)$, that

$$\sum_{j=1}^{m_n} \mathbf{P}(X_{nj} > 0) = \sum_{j=1}^{m_n} (1 - \varphi_{nj}(0)) \xrightarrow{n\to\infty} \lambda.$$

Therefore, $A_n^1(z) \xrightarrow{n\to\infty} \lambda(1-z)$ for $z \in [0,1]$. But then $A_n^2(z) \xrightarrow{n\to\infty} 0$ must apply to $z \in [0,1]$. Because of (10.5) this means that 1. is valid. The statement 2. follows from this by subtraction.
'⇐': So 1. and 2. apply. It is clear that $A_n^2(z) \xrightarrow{n\to\infty} 0$ by (10.5). Then $A_n^1(z) \xrightarrow{n\to\infty} (1-z)\lambda$ by 2., i.e. (10.4) is shown. □

Example 10.7 (Convergence of geometric distributions against Poisson)**.** *Let X_{nj}, $j = 1, \ldots, n, n = 1, 2, \ldots$ be geometrically distributed with parameter p_n (i.e. $\mathbf{P}(X_{nj} = k) = (1-p_n)^{k-1}p_n$, see Example 2.2.4) and $Y_{nj} = X_{nj} - 1$. (Thus, Y_{nj} is the number of* failures *before the first* success*). We set $Y_n := \sum_{j=1}^n Y_{nj}$, which is as distributed as the number of failures before the nth success. If $Y \sim Poi(\lambda)$ and $(1-p_n)\cdot n \xrightarrow{n\to\infty} \lambda$, then $Y_n \xrightarrow{n\to\infty} Y$. Since*

$$\sum_{j=1}^{n} \mathbf{P}(Y_{nj} = 1) = n(1-p_n)p_n \xrightarrow{n\to\infty} \lambda,$$
$$\sum_{j=1}^{n} \mathbf{P}(Y_{nj} > 1) = n(1-p_n)^2 \xrightarrow{n\to\infty} 0,$$

Theorem 10.5 gives the result.

10.2 The Central Limit Theorem

The central limit theorem, Theorem 10.8, generalizes the Theorem of deMoivre Laplace. The generalization consists of the fact that any sums of independent (not necessarily identically distributed) random variables converge weakly to a normally distributed random variable if they satisfy the *Lindeberg condition* (see 2. in Theorem 10.8).

Theorem 10.8 (Central limit theorem of Lindeberg-Feller)**.** *Let $(X_{nj})_{n=1,2,\ldots,j=1,\ldots,m_n}$ be a family of random variables such that for $n = 1, 2, \ldots$ the random variables $X_{n1}, \ldots, X_{nm_n}$ are independent. Assume that*

$$\sum_{j=1}^{m_n} \mathbf{E}[X_{nj}] \xrightarrow{n\to\infty} \mu, \qquad \sum_{j=1}^{m_n} \mathbf{V}[X_{nj}] \xrightarrow{n\to\infty} \sigma^2$$

and $X \sim N(\mu, \sigma^2)$. Then the following statements are equivalent:

1. $\sum_{j=1}^{m_n} X_{nj} \xrightarrow{\mathcal{D}} X$ *and* $\sup_{j=1,\ldots,m_n} \mathbf{V}[X_{nj}] \xrightarrow{n\to\infty} 0$,

2. $\sum_{j=1}^{m_n} \mathbf{E}[(X_{nj} - \mathbf{E}[X_{nj}])^2; |X_{nj} - \mathbf{E}[X_{nj}]| > \varepsilon] \xrightarrow{n\to\infty} 0$ *for all $\varepsilon > 0$.*

Before we prove the central limit theorem, we refer to the special case of identically distributed random variables, which was already discussed in the lecture *Basic Probability.*

Corollary 10.9 (Central limit theorem for identically distributed random variables)**.** *Let $X_1, X_2, \dots$ be independent and identically distributed with $\mathbf{E}[X_1] = \mu, \mathbf{V}[X_1] = \sigma^2 > 0$. Let $S_n := \sum_{k=1}^n X_k$ and $X \sim N(0,1)$. Then,*

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{n\to\infty} X.$$

Proof. Let $m_n = n$ and $X_{nj} = \frac{X_j - \mu}{\sqrt{n\sigma^2}}$. Then the family $(X_{nj})_{n=1,2,\dots,j=1,\dots,n}$ fulfills the conditions of Theorem 10.8 with $\mu = 0, \sigma^2 = 1$. Furthermore

$$\sum_{j=1}^n \mathbf{E}[X_{nj}^2; |X_{nj}| > \varepsilon] = \frac{1}{\sigma^2}\mathbf{E}[(X_1 - \mu)^2; |X_1 - \mu| > \varepsilon\sqrt{n\sigma^2}] \xrightarrow{n\to\infty} 0$$

due to dominated convergence. □

The Lindeberg condition is often not easy to verify. The stronger Lyapunoff condition is often simpler.

Remark 10.10 (Lyapunoff condition)**.** *The family $(X_{nj})_{n=1,2,\dots,j=1,\dots,m_n}$ from Theorem 10.8 satisfies the Lyapunoff condition if for some $\delta > 0$*

$$\sum_{j=1}^{m_n} \mathbf{E}\big[|X_{nj} - \mathbf{E}[X_{nj}]|^{2+\delta}\big] \xrightarrow{n\to\infty} 0.$$

Under the conditions of Theorem 10.8, the Lyapunoff condition implies the Lindeberg condition. To see this, let wlog $\mathbf{E}[X_{nj}] = 0$. For all $\varepsilon > 0$,

$$x^2 1_{|x|>\varepsilon} \le \frac{|x|^{2+\delta}}{\varepsilon^\delta} 1_{|x|>\varepsilon} \le \frac{|x|^{2+\delta}}{\varepsilon^\delta}.$$

If the Lyapunoff condition applies, the Lindeberg condition follows from

$$\sum_{j=1}^{m_n} \mathbf{E}[X_{nj}^2; |X_{nj}| > \varepsilon] \le \frac{1}{\varepsilon^\delta} \sum_{j=1}^{m_n} \mathbf{E}[|X_{nj}|^{2+\delta}] \xrightarrow{n\to\infty} 0.$$

The proof of Theorem 10.8 is based on the clever use of the characteristic functions of the random variable random variable X_{nj} and Taylor approximations. We prepare the proof of the theorem with two lemmas.

Lemma 10.11 (An estimate)**.** *For complex numbers $z_1, \dots, z_n, z_1', \dots, z_n'$ with $|z_i| \le 1, |z_i'| \le 1$ for $i = 1, \dots, n$,*

$$\Big| \prod_{k=1}^n z_k - \prod_{k=1}^n z_k' \Big| \le \sum_{k=1}^n |z_k - z_k'|. \tag{10.6}$$

Proof. For $n = 1$ the equation is obviously correct. Moreover, if (10.6) is valid for an n, then

$$\Big|\prod_{k=1}^{n+1} z_k - \prod_{k=1}^{n+1} z'_k\Big| \le \Big|z_{n+1}\Big(\prod_{k=1}^{n} z_k - \prod_{k=1}^{n} z'_k\Big)\Big| + \Big|(z_{n+1} - z'_{n+1})\prod_{k=1}^{n} z'_k\Big|$$
$$\le \sum_{k=1}^{n} |z_k - z'_k| + |z_{n+1} - z'_{n+1}|.$$

From this the assertion follows. □

Lemma 10.12 (Taylor approximation of the exponential function)**.** *Let $t \in \mathbb{C}$ and $n \in \mathbb{Z}_+$. Then,*

$$\Big|e^{it} - \sum_{k=0}^{n} \frac{(it)^k}{k!}\Big| \le \frac{2|t|^n}{n!} \wedge \frac{|t|^{n+1}}{(n+1)!}. \tag{10.7}$$

Proof. Denote by $h_n(t)$ the difference on the left-hand side. For $n = 0$, (10.7) follows from

$$|h_0(t)| = \Big|\int_0^t e^{is} ds\Big| \le \int_0^t |e^{is}| ds = |t|$$

and

$$|h_0(t)| \le |e^{it}| + 1 = 2.$$

In general, the following applies to $t \in \mathbb{R}, n \in \mathbb{N}$

$$\Big|\int_0^t h_n(s) ds\Big| = \Big| - i(e^{it} - 1) + i\sum_{k=0}^{n} \frac{(it)^{k+1}}{(k+1)!}\Big| = \Big|ie^{it} - i\sum_{k=0}^{n+1} \frac{(it)^k}{k!}\Big| = |h_{n+1}(t)|,$$

and (10.7) follows by induction. □

Remark 10.13 (notation)**.** *In the following proof, we will use for functions a and b the notation $a \lesssim b$ iff there is a constant C with $a \le Cb$.*

Proof of theorem 10.8. Wlog let $\mathbf{E}[X_{nj}] = \mu = 0$ and $\sigma^2 = 1$; otherwise we replace X_{nj} by $\frac{X_{nj} - \mathbf{E}[X_{nj}]}{\sqrt{\sigma^2}}$. Let $\sigma_{nj}^2 := \mathbf{V}[X_{nj}]$ and $\sigma_n^2 := \sum_{j=1}^{m_n} \sigma_{nj}^2 \xrightarrow{n\to\infty} 1$. Denote by ψ_{nj} the characteristic function of X_{nj}.
2. $\Rightarrow$ 1. Since for every $\varepsilon > 0$

$$\sup_{j=1,\dots,m_n} \sigma_{nj}^2 \le \varepsilon^2 + \sup_{j=1,\dots,m_n} \mathbf{E}[X_{nj}^2; |X_{nj}| > \varepsilon] \le \varepsilon^2 + \sum_{j=1}^{m_n} \mathbf{E}[X_{nj}^2; |X_{nj}| > \varepsilon] \xrightarrow{n\to\infty} \varepsilon^2, \tag{10.8}$$

the second part of 1. is already shown.

Let $(Z_{nj})_{n=1,2,\dots,j=1,\dots,m_n}$ be independent random variables with $Z_{nj} \sim N(0, \sigma_{nj}^2)$. This means that $Z_n = \sum_{j=1}^{m_n} Z_{nj} \sim N(0, \sigma_n^2)$. In particular, the following applies thus $Z_n \xrightarrow{n\to\infty} X$, which can be derived directly from the form of the characteristic functions of the normal distribution, Example 6.13.3 can be read off. Let $\widetilde{\psi}_{nj}$ be the characteristic function of Z_{nj}. Then it suffices to show, see Theorem 9.33, that

$$\prod_{j=1}^{n_j} \psi_{nj}(t) - \prod_{j=1}^{m_n} \widetilde{\psi}_{nj}(t) \xrightarrow{n\to\infty} 0 \tag{10.9}$$

for all t. Using Lemma 10.11 and Lemma 10.12 we write

$$\begin{aligned}\Big|\prod_{j=1}^{m_n}\psi_{nj}(t)-\prod_{j=1}^{m_n}\widetilde{\psi}_{nj}(t)\Big| &\le \sum_{j=1}^{m_n}|\psi_{nj}(t)-\widetilde{\psi}_{nj}(t)| \\ &\le \sum_{j=1}^{m_n}|\psi_{nj}(t)-1+\tfrac{1}{2}t^2\sigma_{nj}^2|+\sum_{j=1}^{m_n}|\widetilde{\psi}_{nj}(t)-1+\tfrac{1}{2}t^2\sigma_{nj}^2| \\ &\lesssim 2\sum_{j=1}^{m_n}\mathbf{E}[X_{nj}^2(1\wedge|X_{nj}|)+\sum_{j=1}^{m_n}|e^{-\frac{1}{2}\sigma_{nj}^2t^2}-1+\tfrac{1}{2}t^2\sigma_{nj}^2|.\end{aligned}$$

Furthermore,

$$\sum_{j=1}^{m_n}\mathbf{E}[X_{nj}^2(1\wedge|X_{nj}|)]\le\varepsilon\sum_{j=1}^{m_n}\sigma_{nj}^2+\sum_{j=1}^{m_n}\mathbf{E}[X_{nj}^2;|X_{nj}|>\varepsilon]\xrightarrow{n\to\infty}\varepsilon$$

and

$$\sum_{j=1}^{m_n}|e^{-\frac{1}{2}\sigma_{nj}^2t^2}-1+\tfrac{1}{2}t^2\sigma_{nj}^2|\lesssim\sum_{j=1}^{m_n}\sigma_{nj}^4\le\sigma_n^2\sup_{j=1,\dots,m_n}\sigma_{nj}^2\xrightarrow{n\to\infty}0$$

because of (10.8). This means (10.9) is already proven.
1. $\Rightarrow$ 2. According to the second part of 1. for each $\varepsilon>0$ with the Chebyshev inequality

$$\sup_{j=1,\dots,m_n}\mathbf{P}[|X_{nj}|>\varepsilon]\le\sup_{j=1,\dots,m_n}\frac{\sigma_{nj}^2}{\varepsilon^2}\xrightarrow{n\to\infty}0. \tag{10.10}$$

With Lemma 10.12,

$$\sup_{j=1,\dots,m_n}|\psi_{nj}(t)-1|\le\sup_{j=1,\dots,m_n}\mathbf{E}[2\wedge|t\cdot X_{nj}|]\le 2\sup_{j=1,\dots,m_n}\mathbf{P}[|X_{nj}|>\varepsilon]+\varepsilon|t|\xrightarrow{n\to\infty}\varepsilon|t|.$$

In particular, $\sum_{j=1}^{m_n}\log\psi_{nj}(t)$ is defined for every t if n is large enough. From 1.

$$\sum_{j=1}^{m_n}\log\psi_{nj}(t)\xrightarrow{n\to\infty}-\frac{t^2}{2}. \tag{10.11}$$

Furthermore, because $\psi_{nj}'(0)=i\mathbf{E}[X_{nj}]=0,\psi_{nj}''(0)=-\mathbf{V}[X_{nj}]=-\sigma_{nj}^2$ with the help of a Taylor expansion of ψ_{nj} around 0

$$|\psi_{nj}(t)-1|\lesssim\sigma_{nj}^2|t|^2$$

and

$$\begin{aligned}\Big|\sum_{j=1}^{m_n}\log\psi_{nj}(t)-\sum_{j=1}^{m_n}(\psi_{nj}(t)-1)\Big| &\lesssim \sum_{j=1}^{m_n}|\psi_{nj}(t)-1|^2 \\ &\lesssim\sum_{j=1}^{m_n}(\sigma_{nj}^2)^2|t|^4\lesssim|t|^4\sup_{j=1,\dots,m_n}\sigma_{nj}^2\xrightarrow{n\to\infty}0.\end{aligned} \tag{10.12}$$

Since the convergence of an imaginary series follows from the convergence of its real and imaginary parts, we deduce from (10.11) and (10.12) because $\mathrm{Re}(\psi_{nj}(t)) = \mathbf{E}[\cos(tX_{nj})]$

$$\sum_{j=1}^{m_n} \mathbf{E}[\cos(tX_{nj}) - 1] \xrightarrow{n\to\infty} -\frac{t^2}{2}$$

For $\varepsilon > 0$ is now because of $0 \le 1 - \cos(\theta) \le \frac{\theta^2}{2}$

$$\begin{aligned}
0 \le \limsup_{n\to\infty} \sum_{j=1}^{m_n} \mathbf{E}[X_{nj}^2; |X_{nj}| > \varepsilon] &= \limsup_{n\to\infty}\ 1 - \sum_{j=1}^{m_n} \mathbf{E}[X_{nj}^2; |X_{nj}| \le \varepsilon] \\
&\le \limsup_{n\to\infty}\ 1 - \frac{2}{t^2} \sum_{j=1}^{m_n} \mathbf{E}[1 - \cos(tX_{nj}); |X_{nj}| \le \varepsilon] \\
&= \limsup_{n\to\infty} \frac{2}{t^2} \sum_{j=1}^{m_n} \mathbf{E}[1 - \cos(tX_{nj}); |X_{nj}| > \varepsilon] \qquad (10.13) \\
&\le \limsup_{n\to\infty} \frac{2}{t^2} \sum_{j=1}^{m_n} \mathbf{P}[|X_{nj}| > \varepsilon] \\
&\le \frac{2}{\varepsilon^2 t^2} \limsup_{n\to\infty} \sum_{j=1}^{m_n} \sigma_{nj} = \frac{2}{\varepsilon^2 t^2}.
\end{aligned}$$

Since $t, \varepsilon > 0$ were arbitrary, 2. is shown, if in the the last inequality chain $t \to \infty$ is considered. □

10.3 Multidimensional limit laws

So far, we have only considered weak limit theorems (Theorems 10.5 and 10.8) for the case of $\mathbb{R}$–valued random variables. We now generalize this to $\mathbb{R}^d$–valued random variables. In particular, we give a variant of the multidimensional central limit theorem.

Definition 10.14 (Multidimensional normal distribution)**.** *Let $\mu \in \mathbb{R}^d$ and $C \in \mathbb{R}^{d\times d}$ be a strictly positive definite symmetric matrix.*[4,5]*. The* d-dimensional normal distribution *with expected value μ and covariance matrix C is the probability measure $N_{\mu,C}$ on $\mathbb{R}^d$ with density*

$$f_{\mu,C}(x) = \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp\Big(-\tfrac{1}{2}(x-\mu)C^{-1}(x-\mu)^\top \Big).$$

Proposition 10.15 (Properties of the multidimensional normal distribution)**.** *Let $\mu \in \mathbb{R}^d$, $C = AA^\top \in \mathbb{R}^{d\times d}$ a strictly positive definite symmetric matrix and I the d-dimensional unit matrix. The following are equivalent:*

1. *$X \sim N_{\mu,C}$;*
2. *$tX^\top \sim N_{t\mu^\top, tCt^\top}$ for each $t \in \mathbb{R}^d$;*

[4]We denote row vectors by x and column vectors by $x^\top$.

[5]Strictly positive definite means $xCx^\top > 0$ for all $x \in \mathbb{R}^d$. From linear algebra it is known that for a strictly positive definite matrix C there is always an invertible matrix A with $C = AA^\top$

3. $\psi_X(t) = e^{it\mu^\top} e^{-\frac{1}{2}tCt^\top}$ *for each* $t \in \mathbb{R}^d$.

In each of these cases

4. $X \stackrel{d}{=} AY + \mu$ *for* $Y \sim N_{0,I}$,

5. $\mathbf{E}[X_i] = \mu_i$ *for* $i = 1, \ldots, d$,

6. $\mathbf{COV}[X_i, X_j] = C_{ij}$ *for* $i, j = 1, \ldots, d$.

Proof. First, let $X \sim N_{\mu,C}$. We first show 4.-6. The property 4. is an application of the transformation theorem. For $B \in \mathcal{B}(\mathbb{R}^d)$ and $T : y \mapsto Ay^\top + \mu^\top$,

$$\begin{aligned} N_{0,I}(T^{-1}(B)) &= \frac{1}{\sqrt{(2\pi)^d}} \int_{T^{-1}(B)} e^{-\frac{1}{2}yy^\top} dy \\ &\stackrel{y=A^{-1}(x-\mu)}{=} \frac{1}{\sqrt{(2\pi)^d}} \frac{1}{\det A} \int_B \exp\big(-\tfrac{1}{2}(x-\mu)(A^\top)^{-1}A^{-1}(x-\mu)^\top\big) dx \\ &= \frac{1}{\sqrt{(2\pi)^d \det C}} \int_B \exp\big(-\tfrac{1}{2}(x-\mu)C^{-1}(x-\mu)^\top\big) dx \\ &= N_{\mu,C}(B). \end{aligned}$$

5. follows from 4. with

$$\mathbf{E}[X_i] = \mathbf{E}[\pi_i(AY + \mu)] = \pi_i\mu = \mu_i,$$

where π_i is the projection onto the i-th coordinate.
6. also follows from 4. with

$$\begin{aligned} \mathbf{COV}[X_i, X_j] &= \mathbf{E}[(\pi_i AY^\top)(\pi_j AY^\top)] = \mathbf{E}[(A_{i\cdot}Y^\top)(A_{j\cdot}Y^\top)] = \mathbf{E}[A_{i\cdot}Y^\top Y A_{j\cdot}^\top] \\ &= A_{i\cdot}A_{j\cdot}^\top = (AA^\top)_{ij} = C_{ij}. \end{aligned}$$

We now come to the equivalence of 1.-3.: '1. $\Rightarrow$ 2.': Since $X \stackrel{d}{=} AY^\top + \mu^\top$ as in 4. $tX^\top = tAY^\top + t\mu^\top$ as a linear combination of (one-dimensional) normal distributions is normally distributed again. The expected value is obviously $t\mu^\top$ and the variance

$$\mathbf{V}[tX^\top] = \mathbf{E}[(tAY^\top)^2] = \mathbf{E}[tAY^\top Y A^\top t^\top] = tAA^\top t^\top = tCt^\top.$$

'2. $\Rightarrow$ 3.': Since $tX^\top \sim N_{t\mu^\top, tCt^\top}$, the statement follows from example 6.13.3.
'3. $\Rightarrow$ 1.': This follows from Proposition 9.25. □

Remark 10.16 (Special cases)**.** *1. If C in Definition 10.14 is positive, but not strictly positive definite (i.e. there is $x \in \mathbb{R}^d$ with $x \neq 0$ and $xCx = 0$), one cannot determine $N_{\mu,C}$ by specifying the density as in the definition above. In this case $N_{\mu,C}$ is defined by specifying the characteristic function, i.e. function, i.e. $N_{\mu,C}$ is the uniquely determined distribution on $\mathbb{R}^d$ with $\psi_{N_{\mu,C}}(t) = e^{it\mu}e^{-\frac{1}{2}tCt^\top}$.*

2. If $Y \sim N_{0,I}$ and A is an orthogonal matrix, then also $X := AY \sim N_{0,I}$. This follows from Proposition 10.15, if you write $I = AA^\top$ and use 4. is used.

Proposition 10.17 (Cramér-Wold Device)**.** *If $X, X_1, X_2, \dots$ are random variables with values in $\mathbb{R}^d$. Then $X_n \xRightarrow{n\to\infty} X$ applies if and only if $tX_n \xRightarrow{n\to\infty} tX$ for all $t \in \mathbb{R}^d$ (where $(t,x) \mapsto tx$ is the scalar product in $\mathbb{R}^d$).*

Proof. '$\Rightarrow$': Let $t \in \mathbb{R}^d$ and $f \in \mathcal{C}_b(\mathbb{R})$. Then $f(t\cdot) \in \mathcal{C}_b(\mathbb{R}^d)$. This means that $\mathbf{E}[f(tX_n)] \xrightarrow{n\to\infty} \mathbf{E}[f(tX)]$, i.e. $tX_n \xrightarrow{n\to\infty} tX$.
'$\Leftarrow$': Let π_i be the projection onto the ith coordinate. Since $(\pi_i X_n)_{n=1,2,\dots}$ according to Corollary 9.18 is tight for all i, you can see that $(X_n)_{n=1,2,\dots}$ is tight. Since $\{x \mapsto e^{itx} : t \in \mathbb{R}^d\}$ is a separating class of functions, the assertion follows from $\mathbf{E}[e^{itX_n}] \xrightarrow{n\to\infty} \mathbf{E}[e^{itX}]$ for all $t \in \mathbb{R}^d$ and Proposition 9.27. □

Theorem 10.18 (Multidimensional central limit theorem)**.** *Let $X_1, X_2, \dots$ be independent, identical distributed random variables with values in $\mathbb{R}^d$ with $\mathbf{E}[X_n] = \mu \in \mathbb{R}^d$ and $\mathbf{COV}[X_{n,i}, X_{n,j}] = C_{ij}$ for $i,j = 1,\dots,d$ and $S_n = \sum_{i=1}^n X_i$. If $X \sim N_{0,C}$, then*

$$\frac{S_n - n\mu}{\sqrt{n}} \xRightarrow{n\to\infty} X.$$

Proof. We apply the one-dimensional central limit theorem, Corollary 10.9, to the independent, identically distributed random variables $tX_1, tX_2, \dots$. This provides

$$t\frac{S_n - n\mu}{\sqrt{n}} \xRightarrow{n\to\infty} tX.$$

Since t was arbitrary, the statement follows from Proposition 10.17. □

11 The conditional expectation

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. We write $\mathcal{L}^1 := \mathcal{L}^1(\mathbf{P})$ for the set of all real random variables whose expected value exists. In this chapter we again use the notation $\mathbb{E}[\cdot]$ for the integral with respect to the probability measure $\mathbf{P}$, as well as $\mathcal{L}^p := \mathcal{L}^p(\mathbf{P})$.

11.1 Motivation

Define as in *Elementary Probability* for $A, G \in \mathcal{A}$ and $\mathbf{P}(G) > 0$

$$\mathbf{P}(A|G) := \frac{\mathbf{P}(A \cap G)}{\mathbf{P}(G)}$$

and analogously the *conditional expectation*

$$\mathbf{E}[X|G] := \frac{\mathbf{E}[X; G]}{\mathbf{P}(G)}.$$

Then $\mathbf{P}(A|G) = \mathbf{E}[1_A|G]$. This relationship means that conditional expectations can be used to calculate conditional probabilities. In particular, the notion of conditional expectation is more general than the notion of conditional probability.

In this chapter, we will use the conditional expectation $\mathbf{E}[X|\mathcal{G}]$ for a random variable X and a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. Here, $\mathbf{E}[X|\mathcal{G}]$ is a $\mathcal{G}$-measurable random variable. As a simple

example, $\{G_1, G_2, \dots\} \subseteq \mathcal{F}$ is a partition of Ω with $\mathbf{P}(G_i) > 0$ for $i = 1, 2, \dots$ and $\mathcal{G}$ the generated σ algebra. Then we set for $X \in \mathcal{L}^1$

$$\mathbf{E}[X|\mathcal{G}](\omega) := \sum_{i=1}^{\infty} \mathbf{E}[X|G_i] \cdot 1_{G_i}(\omega). \tag{11.1}$$

The following therefore applies: for $\omega \in G_i$, the random variable $\mathbf{E}[X|\mathcal{G}]$ is given by $\mathbf{E}[X|\mathcal{G}](\omega) = \mathbf{E}[X|G_i] = \mathbf{E}[X; G_i]/\mathbf{P}(G_i)$. In particular it is constant on G_i, $i = 1, 2, \dots$ In other words, $\mathbf{E}[X|\mathcal{G}]$ is measurable with respect to $\mathcal{G}$. The following also applies to $J \subseteq \mathbb{N}$ and $A = \bigcup_{j\in J} G_j \in \mathcal{G}$

$$\begin{aligned} \mathbf{E}[\mathbf{E}[X|\mathcal{G}]; A] &= \mathbf{E}\Big[\sum_{i=1}^{\infty} \mathbf{E}[X|G_i] 1_{G_i} 1_A\Big] \\ &= \sum_{j\in J} \mathbf{E}\big[\mathbf{E}[X|G_j] 1_{G_j}\big] \\ &= \sum_{j\in J} \mathbf{E}[X|G_j] \cdot \mathbf{P}(G_j) \\ &= \mathbf{E}[X; A]. \end{aligned} \tag{11.2}$$

In particular, with $J = \mathbb{N}$ therefore $\mathbf{E}\big[\mathbf{E}[X|\mathcal{F}]\big] = \mathbf{E}[X]$. The definition of the conditional expectation (11.1) can be generalized with the help of the property (11.2) to any σ-algebras $\mathcal{G} \subseteq \mathcal{F}$.

Example 11.1 (Binomial distribution with random success probability)**.** *Let X be uniformly distributed on $[0, 1]$, i.e. the distribution of X has density $1_{[0;1]}$. Given $X = x$ let $Y_1, \dots, Y_n$ be a sequence of Bernoulli distributed random variables with probability of success x. Therefore, $Y = Y_1 + \cdots + Y_n$ is binomially distributed with n and x, i.e. Y counts the number of successes in n independent experiments with probability of success x. Intuitively, it is clear what*

$$\mathbf{P}(Y = k|X) = \binom{n}{k} X^k (1 - X)^{n-k}$$

should mean. However, this has not yet been defined, since $\mathbf{P}(X = x) = 0$. However, it is worth noting that the right side is a $\sigma(X)$-measurable random variable (since it is a function of X; see Lemma 6.2).

11.2 Definition and properties

We now formally define the conditional expectation $\mathbf{E}[X|\mathcal{G}]$ for $\mathcal{G} \subseteq \mathcal{F}$. As mentioned above, this is a $\mathcal{G}$-measurable random variable whose expectations are as in (11.2) match those of X.

Theorem 11.2 (Existence and properties of the conditional expectation)**.** *Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Then there is an almost surely unique linear operator $\mathbf{E}[.|\mathcal{G}] : \mathcal{L}^1 \to \mathcal{L}^1$ such that $\mathbf{E}[X|\mathcal{G}]$ for all $X \in \mathcal{L}^1$ a $\mathcal{G}$-measurable random variable with*

1. *$\mathbf{E}\big[\mathbf{E}[X|\mathcal{G}]; A\big] = \mathbf{E}[X; A]$ for all $A \in \mathcal{G}$.*

Further,

2. *$\mathbf{E}[X|\mathcal{G}] \geq 0$ if $X \geq 0$.*

3. $\mathbf{E}\big[|\mathbf{E}[X|\mathcal{G}]|\big] \leq \mathbf{E}[|X|]$.

4. *If* $0 \leq X_n \uparrow X$ *for* $n \to \infty$, *then also* $\mathbf{E}[X_n|\mathcal{G}] \uparrow \mathbf{E}[X|\mathcal{G}]$ *in* $\mathcal{L}^1$ *if all expectations exist.*

5. *If* X *is a* $\mathcal{G}$*-measurable function, then* $\mathbf{E}[XY|\mathcal{G}] = X\mathbf{E}[Y|\mathcal{G}]$ *if all expectations exist.*

6. $\mathbf{E}\big[X\mathbf{E}[Y|\mathcal{G}]\big] = \mathbf{E}\big[\mathbf{E}[X|\mathcal{G}]Y\big] = \mathbf{E}\big[\mathbf{E}[X|\mathcal{G}]\mathbf{E}[Y|\mathcal{G}]\big]$ *if all expectations exist.*

7. *If* $\mathcal{H} \subseteq \mathcal{G}$, *then* $\mathbf{E}\big[\mathbf{E}[X|\mathcal{G}]|\mathcal{H}\big] = \mathbf{E}[X|\mathcal{H}]$.

8. *If* X *is independent of* $\mathcal{G}$, *then* $\mathbf{E}[X|\mathcal{G}] = \mathbf{E}[X]$.

Proof. 1. in the case $X \in \mathcal{L}^2$: Let M be the closed linear subspace of $\mathcal{L}^2$, which consists of all functions which, except for a zero set, correspond to a $\mathcal{G}$-measurable function. According to Proposition 4.10 there are almost surely unique functions $Y \in M, Z \perp M$ with $X = Y + Z$. We define $\mathbf{E}[X|\mathcal{G}] := Y$. This means that $X - \mathbf{E}[X|\mathcal{G}] \perp M$, i.e. $\mathbf{E}[X - \mathbf{E}[X|\mathcal{G}]; A] = 0$ for $A \in \mathcal{G}$, from which 1. for $X \in \mathcal{L}^2$ follows.
3. in the case $X \in \mathcal{L}^2$: Choose $A := \{\mathbf{E}[X|\mathcal{G}] \geq 0\}$. According to 1.,

$$\mathbf{E}[|\mathbf{E}[X|\mathcal{G}]|] = \mathbf{E}[\mathbf{E}[X|\mathcal{G}]; A] - \mathbf{E}[\mathbf{E}[X|\mathcal{G}]; A^c] = \mathbf{E}[X; A] - \mathbf{E}[X; A^c] \leq \mathbf{E}[|X|].$$

1. in the case $X \in \mathcal{L}^1$: If $X \in \mathcal{L}^1 \supset \mathcal{L}^2$, then choose $X_1, X_2, \cdots \in \mathcal{L}^2$ with $||X_n - X||_1 \xrightarrow{n\to\infty} 0$ (such that $|X_n| := |X| \wedge n$), and define $\mathbf{E}[X|\mathcal{G}] := \lim_{n\to\infty} \mathbf{E}[X_n|\mathcal{G}]$. This limit value exists in $\mathcal{L}^1$, since because of 3.

$$\mathbf{E}[|\mathbf{E}[X_n|\mathcal{G}] - \mathbf{E}[X_m|\mathcal{G}]|] = \mathbf{E}[|\mathbf{E}[X_n - X_m|\mathcal{G}]|] \leq \mathbf{E}[|X_n - X_m|] \xrightarrow{n,m\to\infty} 0$$

the sequence $(\mathbf{E}[X_n|\mathcal{G}])_{n=1,2,\ldots}$ is a Cauchy sequence and $\mathcal{L}^1$ is complete. Furthermore, this means that $||\mathbf{E}[X_n|\mathcal{G}] - \mathbf{E}[X|\mathcal{G}]||_1 \xrightarrow{n\to\infty} 0$. Furthermore, for $A \in \mathcal{G}$

$$\begin{aligned}|\mathbf{E}[X - \mathbf{E}[X|\mathcal{G}]; A]| &\leq \mathbf{E}[|X1_A - X_n1_A|] \\ &\quad + |\mathbf{E}[X_n - \mathbf{E}[X_n|\mathcal{G}]; A]]| \\ &\quad + \mathbf{E}[|\mathbf{E}[X_n|\mathcal{G}]1_A - \mathbf{E}[X|\mathcal{G}]1_A|] \\ &\quad \xrightarrow{n\to\infty} 0\end{aligned}$$

due to dominated convergence and 1. follows in the case $X \in \mathcal{L}^1$.
3. in the case $X \in \mathcal{L}^1$. Here, too, you can see through an approximation argument if $X_1, X_2, \cdots \in \mathcal{L}^2$ with $X_n \xrightarrow{n\to\infty}_{\mathcal{L}^1} X$,

$$\mathbf{E}[|\mathbf{E}[X|\mathcal{G}]|] = \lim_{n\to\infty} \mathbf{E}[|\mathbf{E}[X_n|\mathcal{G}]|] \leq \lim_{n\to\infty} \mathbf{E}[|X_n|] = \mathbf{E}[|X|],$$

since, due to the inverse triangle inequality, approximately,

$$\mathbf{E}[||\mathbf{E}[X_n|\mathcal{G}]| - |\mathbf{E}[X|\mathcal{G}]||] \leq \mathbf{E}[|\mathbf{E}[X|\mathcal{G}] - \mathbf{E}[X_n|\mathcal{G}]|] \xrightarrow{n\to\infty} 0.$$

.
2. set $A = \{\mathbf{E}[X|\mathcal{G}] \leq 0\}$ and thus

$$0 \geq \mathbf{E}[\mathbf{E}[X|\mathcal{G}]; A] = \mathbf{E}[X; A] \geq 0,$$

thus because of $\mathbf{E}[X|\mathcal{G}]1_A \leq 0$ also $\mathbf{E}[X|\mathcal{G}]1_A = 0$ is almost sure.

4. Due to monotone convergence, $||X_n - X||_1 \xrightarrow{n\to\infty} 0$, i.e. with 3.,

$$\mathbf{E}[|\mathbf{E}[X_n|\mathcal{G}] - \mathbf{E}[X|\mathcal{G}]|] = \mathbf{E}[|\mathbf{E}[X_n - X|\mathcal{G}]|] \leq \mathbf{E}[|X_n - X|] \xrightarrow{n\to\infty} 0.$$

6. in the case $X, Y \in \mathcal{L}^2$. According to the definition of the conditional expectation, $\mathbf{E}[X|\mathcal{G}], \mathbf{E}[Y|\mathcal{G}] \in M$ if M is the linear subspace of $\mathcal{L}^2$ which contains functions which, apart from a zero set with a $\mathcal{G}$-measurable function. Furthermore $X - \mathbf{E}[X|\mathcal{G}] \perp M$. Thus

$$\mathbf{E}[(X - \mathbf{E}[X|\mathcal{G}])\mathbf{E}[Y|\mathcal{G}]] = 0.$$

6. in the case $X, Y \in \mathcal{L}^1$. Choose $X_1, Y_1, X_2, Y_2, \cdots \in \mathcal{L}^2$ with $X_n \uparrow X, Y_n \uparrow Y$. Because of 4. and dominated convergence, if all expectations exist,

$$\mathbf{E}[(X - \mathbf{E}[X|\mathcal{G}])\mathbf{E}[Y|\mathcal{G}]] = \lim_{n\to\infty} \mathbf{E}[(X_n - \mathbf{E}[X_n|\mathcal{G}])\mathbf{E}[Y_n|\mathcal{G}]] = 0.$$

5. because of 1. is $\mathbf{E}[X|\mathcal{G}]1_A = X1_A$ for $A \in \mathcal{G}$, almost surely. This means that

$$\mathbf{E}[XY; A] = \mathbf{E}[X\mathbf{E}[Y|\mathcal{G}]; A]$$

after 6. from this follows after 1. already $\mathbf{E}[XY|\mathcal{G}] = X\mathbf{E}[Y|\mathcal{G}]$.

Since $\mathcal{H} \subseteq \mathcal{G}$, for $A \in \mathcal{H}$,

$$\mathbf{E}[\mathbf{E}[X|\mathcal{G}]; A] = \mathbf{E}[X; A] = \mathbf{E}[\mathbf{E}[X|\mathcal{H}]; A]$$

after 1. From here follows but $\mathbf{E}[\mathbf{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbf{E}[X|\mathcal{H}]$.

8. Certainly, $\mathbf{E}[X]$ is measurable with respect to $\mathcal{G}$. For $A \in \mathcal{G}$,

$$\mathbf{E}[\mathbf{E}[X|\mathcal{G}]; A] = \mathbf{E}[X; A] = \mathbf{E}[X]\mathbf{E}[1_A] = \mathbf{E}\big[\mathbf{E}[X]; A\big]$$

and thus $\mathbf{E}[X|\mathcal{G}] = \mathbf{E}[X]$. □

Remark 11.3 (Interpretation and alternative proof)**.** *1. Let $X \in \mathcal{L}^2$. As the proof of 1. in Theorem 11.2 shows, $X - \mathbf{E}[X|\mathcal{G}]$ is perpendicular to the linear subspace of all $\mathcal{G}$-measurable functions. In particular $\mathbf{E}[X|\mathcal{G}]$ is the $\mathcal{G}$-measurable random variable that (in terms of the $\mathcal{L}^2$ norm) is closest to the random variable X comes closest. Therefore, we can say that $\mathbf{E}[X|\mathcal{G}]$ is the best estimate of X if information from the σ algebra $\mathcal{G}$ is available.*

2. The almost surely unambiguous existence of the conditional expectation with the property 1. in Theorem 11.2 can be proved differently than above with the help of the theorem of Radon-Nikodým (Corollary 4.17):

Let $X \geq 0$ first. Set $\widetilde{\mathbf{P}} := \mathbf{P}|_{\mathcal{G}}$, the restriction of $\mathbf{P}$ to $\mathcal{G}$, and $\mu(.) := \widetilde{\mathbf{E}}[X; .]$ a finite measure. Then obviously $\mu \ll \widetilde{\mathbf{P}}$ applies. The theorem of Radon-Nikodým ensures that μ is a density with respect to $\widetilde{\mathbf{P}}$, i.e. there is a $\mathcal{G}$-measurable random variable Z with

$$\mathbf{E}[X; A] = \widetilde{\mathbf{E}}[X; A] = \mu(A) = \widetilde{\mathbf{E}}[Z; A] = \mathbf{E}[Z; A]$$

for all $A \in \mathcal{G}$. Thus Z fulfills the properties of 1. from theorem 11.2. The general case (i.e. X can also take can also assume negative values) then follows with the decomposition $X = X^+ - X^-$.

To prove the (almost sure) uniqueness of the conditional expectation, let Z' be another $\mathcal{G}$-measurable random variable with random variable with $\mathbf{E}[Z'; A] = \mathbf{E}[X; A]$ for all $A \in \mathcal{G}$. Then $B := \{Z' - \mathbf{E}[X|\mathcal{G}] > 0\} \in \mathcal{G}$ and $\mathbf{E}[\mathbf{E}[X|\mathcal{G}] - Z'; B] = \mathbf{E}[X - X; B] = 0$ and likewise $\mathbf{E}[\mathbf{E}[X|\mathcal{G}] - Z'; B^c] = 0$. This therefore means $Z' = \mathbf{E}[X|\mathcal{G}]$, almost surely.

Proposition 11.4 (Jensen's inequality for conditional expectations)**.** *Let I be an open interval, $\mathcal{G} \subseteq \mathcal{A}$ and $X \in \mathcal{L}^1$ with values in I and $\varphi : I \to \mathbb{R}$ is convex. Then,*

$$\mathbf{E}[\varphi(X)|\mathcal{G}] \geq \varphi(\mathbf{E}[X|\mathcal{G}]).$$

Proof. The proof is analogous to that of Jensen's inequality in the unconditional case, Proposition 6.6: Since I is open, $\mathbf{E}[X|\mathcal{G}] \in I$, almost surely. We recall the definition of λ in (6.4). Further, as in (6.5) for $x \in I$,

$$\varphi(x) \geq \varphi(\mathbf{E}[X|\mathcal{G}]) + \lambda(\mathbf{E}[X|\mathcal{G}])(x - \mathbf{E}[X|\mathcal{G}])$$

and thus

$$\begin{aligned}\mathbf{E}[\varphi(X)|\mathcal{G}] &\geq \mathbf{E}[\varphi(\mathbf{E}[X|\mathcal{G}])|\mathcal{G}] + \mathbf{E}[\lambda(\mathbf{E}[X|\mathcal{G}]) \cdot (X - \mathbf{E}[X|\mathcal{G}])|\mathcal{G}] \\ &= \varphi\big(\mathbf{E}[X|\mathcal{G}]\big).\end{aligned}$$

□

Lemma 11.5 (Uniform integrability and conditional expectation)**.** *Let $X \in \mathcal{L}^1$. Then the family $(\mathbf{E}[X|\mathcal{G}])_{\mathcal{G}\subseteq\mathcal{A}}$ is uniformly integrable.*

Proof. Since $\{X\}$ is uniformly integrable, according to Lemma 7.9 there is a monotonically increasing convex function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\frac{\varphi(x)}{x} \xrightarrow{x\to\infty} \infty$ and $\mathbf{E}[\varphi(|X|)] < \infty$. With Theorem 11.2.3, we obtain

$$\sup_{\mathcal{F}\subseteq\mathcal{A}} \mathbf{E}[\varphi(|\mathbf{E}[X|\mathcal{F}]|)] \leq \mathbf{E}[\varphi(|X|)] < \infty.$$

This means that $\{\mathbf{E}[X|\mathcal{F}] : \mathcal{F} \subseteq \mathcal{A}\ \sigma\text{-algebra}\}$ is uniformly integrable, again according to Lemma 7.9. □

Theorem 11.6 (Dominated and monotone convergence for conditional expectations)**.** *Let $\mathcal{G} \subseteq \mathcal{F}$ and $X_1, X_2, \cdots \in \mathcal{L}^1$. Assume one of the following:*

1. *Let $X \in \mathcal{L}^1$ such that $X_n \uparrow X$, almost surely.*
2. *If $Y \in \mathcal{L}^1$ such that $|X_n| \leq |Y|$ for all n, and $X_n \xrightarrow{n\to\infty} X$ almost surely.*

Then

$$\mathbf{E}[X_n|\mathcal{G}] \xrightarrow{n\to\infty} \mathbf{E}[X|\mathcal{G}]$$

almost surely and in $\mathcal{L}^1$.

Proof. For the $\mathcal{L}^1$-convergence one has in both cases with Theorem 11.2.3

$$\begin{aligned}\mathbf{E}\big[\big|\mathbf{E}[X_n|\mathcal{G}] - \mathbf{E}[X|\mathcal{G}]\big|\big] &= \mathbf{E}\big[\big|\mathbf{E}[X_n - X|\mathcal{G}]\big|\big] \\ &\leq \mathbf{E}[|X_n - X|] \xrightarrow{n\to\infty} 0.\end{aligned}$$

We divide the almost sure convergence into the two cases: in case 1. it is clear from Theorem 11.2.2 that $\mathbf{E}[X_n|\mathcal{G}]$ grows monotonically. Furthermore, for $A \in \mathcal{F}$ with the theorem of monotone convergence

$$\mathbf{E}\big[\sup_n \mathbf{E}[X_n|\mathcal{G}]; A\big] = \sup_n \mathbf{E}\big[\mathbf{E}[X_n|\mathcal{G}]; A\big] = \sup_n \mathbf{E}[X_n; A] = \mathbf{E}[\sup_n X_n; A] = \mathbf{E}[X; A].$$

However, this shows that $\sup_n \mathbf{E}[X_n|\mathcal{G}] = \mathbf{E}[X|\mathcal{G}]$, almost surely. In case 2. we set

$$Y_n := \sup_{k\geq n} X_k \downarrow \limsup_n X_n = X \text{ almost surely,}$$
$$Z_n := \inf_{k\geq n} X_k \uparrow \liminf_n X_n = X \text{ almost surely.}$$

Thus $-Y \leq Z_n \leq X_n \leq Y_n \leq Y$, i.e. in particular $Y_1, Z_1, Y_2, Z_2, \cdots \in \mathcal{L}^1$, so according to 1.,

$$\mathbf{E}[X|\mathcal{G}] = \lim_{n\to\infty} \mathbf{E}[Z_n|\mathcal{G}] \leq \lim_{n\to\infty} \mathbf{E}[X_n|\mathcal{G}] \leq \lim_{n\to\infty} \mathbf{E}[Y_n|\mathcal{G}] = \mathbf{E}[X|\mathcal{G}],$$

almost surely. In particular, $\mathbf{E}[X_n|\mathcal{G}] \xrightarrow{n\to\infty} \mathbf{E}[X|\mathcal{G}]$, almost surely. □

11.3 The case $\mathcal{G} = \sigma(X)$

In the case $\mathcal{G} = \sigma(X)$, $\mathbf{E}[Y|X] := \mathbf{E}[Y|\sigma(X)]$ is the expectation of Y, given that the random variable X is fixed. This is a function of X, as Proposition 11.7 shows.

Proposition 11.7 (Conditioning on a random variable)**.** *Let $(\Omega', \mathcal{F}')$ be a measurable space, X a random variable with values in Ω' and $Y \in \mathcal{L}^1$. Then there exists a $\mathcal{F}'/\mathcal{B}(\mathbb{R})$-measurable mapping $\varphi : \Omega' \to \mathbb{R}$ with $\mathbf{E}[Y|X] = \varphi(X)$.*

Proof. Clear according to Lemma 6.2. □

Example 11.8 (Random success probability)**.** *Let us consider the question posed in Example 11.1 regarding the existence of the conditional probability $\mathbf{P}(Y = k|X)$, where X is uniform on $[0,1]$ and X is independently binomially distributed with n and X. We now show (the intuitive equation)*

$$\mathbf{P}(Y = k|X) = \binom{n}{k} X^k (1-X)^{n-k}. \tag{11.3}$$

Let $A = \{X \in I\}$ for $I \in \mathcal{B}([0,1])$, i.e. A is a $\sigma(X)$-measurable quantity. Then,

$$\mathbf{E}[1_{Y=k}; A] = \mathbf{P}(Y = k, X \in I) = \int_I \binom{n}{k} x^k (1-x)^{n-k} dx = \mathbf{E}\Big[\binom{n}{k} X^k (1-X)^{n-k}; A\Big]$$

However, this means that (11.3) *is true.*

Example 11.9 (Sums of independent identically distributed random variables)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables, $\mu = \mathbf{E}[X_1]$ and $S_n := X_1 + \cdots + X_n$. Then*

$$\mathbf{E}[S_n|X_1] = \mathbf{E}[X_1|X_1] + \mathbf{E}[X_2 + \cdots + X_n|X_1] = X_1 + (n-1)\mu,$$
$$\mathbf{E}[X_1|S_n] = \tfrac{1}{n}\sum_{i=1}^n \mathbf{E}[X_i|S_n] = \tfrac{1}{n}\mathbf{E}[S_n|S_n] = \tfrac{1}{n}S_n.$$

In the second calculation, for example, for $X = S_n$ and $Y = X_1$ the function φ from Proposition 11.7 is given by $\varphi(x) = \frac{1}{n}x$.

Example 11.10 (Buffon's needle problem)**.** *On a plane, vertical lines are at a horizontal distance of 1. Needles, also of length 1, are thrown onto the plane; see Figure 2. Let us consider a needle. We set*

$$Z := \begin{cases} 1, & \textit{if the needle intersects a straight line} \\ 0, & \textit{otherwise} \end{cases}.$$

The center of the needle X is away from the left straight line and the (extension of the)

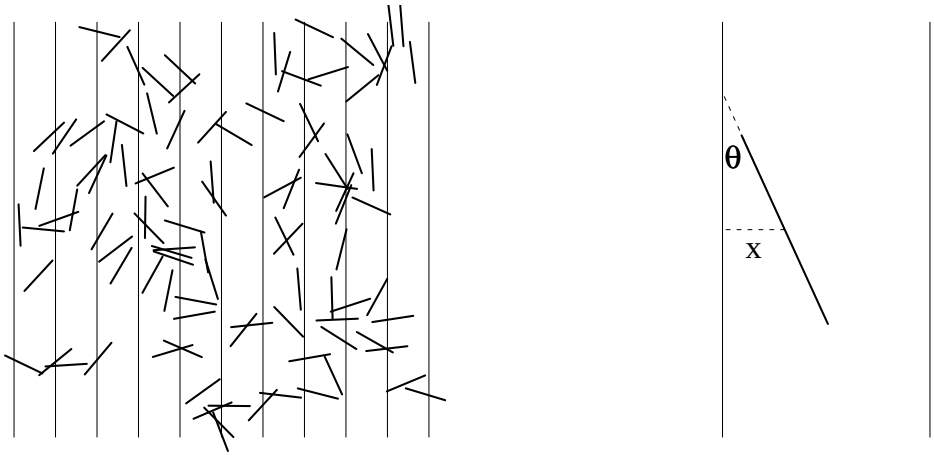


Figure 2: Sketch of Buffon's needle problem

needle makes an angle Θ with the straight line. This means that X is uniform on $[0;1]$, Θ is uniformly independent on $[0;\frac{\pi}{2}]$ and

$$\mathbf{P}(Z=1|\Theta) = \mathbf{P}(X \le \tfrac{1}{2}\sin(\Theta) \textit{ or } X \ge 1 - \tfrac{1}{2}\sin(\Theta)Theta) = \sin(\Theta).$$

This means that

$$\mathbf{P}(Z=1) = \mathbf{E}[\mathbf{P}(Z=1|\Theta)] = \mathbf{E}[\sin(\Theta)] = \frac{2}{\pi}\int_0^{\pi/2}(\sin(\theta)d\theta = \frac{2}{\pi}.$$

This can be interpreted as follows: if you want to determine by simulation (i.e. by a Monte Carlo method) to find the numerical value of π you can simulate Buffon's needles. Since each individual needle has the probability $\frac{2}{\pi}$ of hitting a vertical line, is approximately

$$\pi \approx \frac{2}{\textit{proportion of needles that hit a vertical line}}$$

according to the law of large numbers.

Example 11.11 (Search in lists)**.** *Consider n names of people who come from r different cities. Each person comes (independently of any other) with probability p_j from city j, $j = 1,\dots,r$. The names (together with other personal data) are entered in r different (unordered) lists. If you now want a (random, according to the probabilities $p_1,\dots,p_r$) person in the list,*

first determine the list, you first determine the city j from which the person comes from and then search the list j for the person's name. Until you realize that the name does not appear in the list you have to compare the person to be found with names on the list. The question now is: How many times on average do you have to compare the name of the person to be found with names on the list without success until you finally know that the person is not on the list?

We first define a few random variables:
J *: number of the city from which the person to be searched comes*
L *: number of unsuccessful comparisons until the name of the person to be found is found*
Z_j *: number of people from city* j
and $Z = (Z_1, \dots, Z_r)$*. In order to determine* $\mathbf{E}[L]$*, we first determine*

$$\mathbf{P}(L = a|J, Z) = 1_{Z_J = a}$$

and thus

$$\mathbf{P}(L = a|Z) = \sum_{j=1}^{r} p_j 1_{Z_j = a}.$$

From this we conclude

$$\mathbf{E}[L|Z] = \sum_{a=1}^{\infty} \sum_{j=1}^{r} a \cdot p_j \cdot 1_{Z_j = a} = \sum_{j=1}^{r} p_j Z_J$$

and therefore

$$\mathbf{E}[L] = \mathbf{E}[\mathbf{E}[L|Z]] = \sum_{j=1}^{r} p_j \mathbf{E}[Z_j] = n \cdot \sum_{j=1}^{r} p_j^2.$$

Example 11.12 (Mixture of Poisson distributions)**.** *Let* $\lambda > 0$ *and* $\lambda \sim \exp(\lambda)$ *and for a given* λ *let* $X \sim Poi(\lambda)$*. We now show that* $X + 1 \sim geo(1/(1+\lambda))$*.*
Because: According of Proposition 9.25, the distribution is determined by the characteristic function. First of all, the characteristic function of $Y \sim geo(p)$

$$t \mapsto \mathbf{E}[e^{itY}] = \sum_{k=0}^{\infty} (1-p)^{k-1} p e^{itk} = p e^{it} \sum_{k=0}^{\infty} \big((1-p)e^{it}\big)^k = \frac{p e^{it}}{1-(1-p)e^{it}} = \frac{\frac{p}{1-p} e^{it}}{\frac{1}{1-p} - e^{it}}.$$

We calculate with Example 6.13.2 for $t \in \mathbb{R}$

$$\mathbf{E}[e^{it(X+1)}] = e^{it} \mathbf{E}\big[\mathbf{E}[e^{itX}|\Lambda]\big] = e^{it} \mathbf{E}\big[e^{-(1-e^{it})\Lambda}\big] = \frac{\lambda e^{it}}{1+\lambda-e^{it}},$$

so that the assertion with $\lambda = p/(1-p)$ *or* $p = 1/(1+\lambda)$ *follows.*

11.4 Conditional independence

In Section 8, we have already learned about the independence of σ algebras (or of random variables). Conditional expectations and independence are closely related, as the next lemma shows.

Lemma 11.13 (Conditional probability and independence)**.** *The* σ*-algebras* $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ *are independent if and only if* $\mathbf{P}(G|\mathcal{H}) = \mathbf{P}(G)$ *for all* $G \in \mathcal{G}$*.*

Proof. '⇒': Let $\mathcal{G}$ and $\mathcal{H}$ be independent. Then, for $G \in \mathcal{G}, H \in \mathcal{H}$,

$$\mathbf{E}[\mathbf{P}(G), H] = \mathbf{P}(G \cap H) = \mathbf{E}[\mathbf{P}(G|\mathcal{H}), H].$$

This means that $\mathbf{P}(G|\mathcal{H}) = \mathbf{P}(G)$ according to the definition of the conditional expectation.

'⇐': So if $\mathbf{P}(G|\mathcal{H}) = \mathbf{P}(G)$, it follows for $H \in \mathcal{H}$

$$\mathbf{P}(G \cap H) = \mathbf{E}[1_G, H] = \mathbf{E}[\mathbf{P}(G|\mathcal{H}), H] = \mathbf{E}[\mathbf{P}(G), H] = \mathbf{P}(G) \cdot \mathbf{P}(H).$$

□

The concept of independence is often also required in a conditional form. Let's start with an important example.

Example 11.14 (Markov chains)**.** *Let E be a countable set. A Markov chain $\mathcal{X} = (X_t)_{t=0,1,2,\dots}$ is a family of E-valued random variables such that for all $A \subseteq E$*

$$\mathbf{P}(X_{t+1} \in A|X_0, \dots, X_t) = \mathbf{P}(X_{t+1} \in A|X_t). \tag{11.4}$$

This means: if you want to know the distribution of X_{t+1}, and the information of the random variable X_t is already available, the information about the random variables $X_0, \dots, X_{t-1}$ does not provide any additional information. One also says:

Given X_t, X_{t+1} is independent of $X_0, \dots, X_{t-1}$.

Or in terms of σ-algebras:

Given $\sigma(X_t)$, $\sigma(X_{t+1})$ is independent of $\sigma(X_0, \dots, X_{t-1})$.

One can also say in this case: given the present (that is the state at time t, X_t) the future (i.e. X_{t+1}) is independent of the past (these are the states $X_0, \dots, X_{t-1}$).

A simple example of a Markov chain is the one-dimensional random walk: let $Y_1, Y_2, \dots$ be independent and identically distributed such that $\mathbf{P}(Y_1 = 1) = p$ and $\mathbf{P}(Y_1 = -1) = q$ for a $p \in [0, 1]$. Further, let $X_0 = 0$ and $X_t = Y_1 + \cdots + Y_t$. Then $(X_t)_{t\geq 0}$ is a Markov chain, because

$$\mathbf{P}(X_{t+1} = k|X_0, \dots, X_t) = \begin{cases} p, & k = X_t + 1, \\ q, & k = X_t - 1. \end{cases}$$

In particular, the right-hand side defines an X_t-measurable random variable and is therefore equal to $\mathbf{P}(X_{t+1} = k|X_t)$.

Definition 11.15 (Conditional independence)**.** *Let $\mathcal{G} \subseteq \mathcal{F}$. A family $(\mathcal{C}_i)_{i\in I}$ of set systems with $\mathcal{C}_i \subseteq \mathcal{F}$ is called* independently given $\mathcal{G}$ *if*

$$\mathbf{P}\Big(\bigcap_{j\in J} A_j|\mathcal{G}\Big) = \prod_{j\in J} \mathbf{P}(A_j|\mathcal{G}) \tag{11.5}$$

applies to all $J \subseteq_f I$ and $A_j \in \mathcal{C}_j, j \in J$.

Similarly, conditional independence is de fined for random variables. Let Y be a random variable. A family $(X_i)_{i\in I}$ of random variables is independent given $\mathcal{G}$ (or Y) if $(\sigma(X_i))_{i\in I}$ is independent given $\mathcal{G}$ (resp. $\sigma(Y)$).

Example 11.16 (Simple cases)**.** *Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra and $(\mathcal{C}_i)_{i\in I}$ a family of set systems.*

1. *If $\mathcal{G} = \mathcal{F}$, then $(\mathcal{C}_i)_{i\in I}$ is always independent given $\mathcal{G}$.*
2. *If $\mathcal{G} = \{\emptyset, \Omega\}$, then $(\mathcal{C}_i)_{i\in I}$ is independent given $\mathcal{G}$ if and only if $(\mathcal{C}_i)_{i\in I}$ are independent.*

Example 11.17 (Binomial distribution with random success probability)**.** *We look again at the coin toss with random success probability from Example 11.1 and 11.8. Here X was uniformly distributed on $[0,1]$ and, given X, $Y_1, \dots, Y_n$ are Bernoulli distributed. Now it should hold that $(Y_1, \dots, Y_n)$ are independent given X. Just like in Example 11.8, we calculate for $A = \{X \in I\}$ and for some $I \in \mathcal{B}([0,1])$ and $y_1, \dots, y_n \in \{0,1\}$ and $k := y_1 + \cdots + y_n$*

$$\begin{aligned}\mathbf{E}[1_{Y_1=y_1,\dots,Y_n=y_n}, A] &= \mathbf{P}(Y_1 = y_1, \dots, Y_n = y_n, X \in I) \\ &= \int_I x^{y_1+\cdots+y_n}(1-x)^{n-y_1-\cdots-y_n}dx = \mathbf{E}\big[X^k(1-X)^{n-k}, A\big],\end{aligned}$$

so

$$\mathbf{P}(Y_1 = y_1, \dots, Y_n = y_n|X) = X^k(1-X)^{n-k}.$$

Analogously, one shows for $i = 1, \dots, n$

$$\mathbf{P}(Y_i = y_1|X) = X^{y_i}(1-X)^{1-y_i}.$$

From this follows

$$\mathbf{P}(Y_1 = y_1, \dots, Y_n = y_n|X) = \prod_{i=1}^n \mathbf{P}(Y_i = y_i|X),$$

so $(Y_1, \dots, Y_n)$ are independent given X.

Lemma 11.13 also exists in the following version, in which the independence is replaced by conditional independence.

Proposition 11.18 (Conditional probability and conditional independence)**.** *Let $\mathcal{K} \subseteq \mathcal{F}$ be a σ-algebra. The σ-algebras $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ are independent given $\mathcal{K}$ if and only if $\mathbf{P}(G|\sigma(\mathcal{H}, \mathcal{K})) = \mathbf{P}(G|\mathcal{K})$ for all $G \in \mathcal{G}$.*

Proof. '$\Rightarrow$': If $\mathcal{G}$ and $\mathcal{H}$ are independent given $\mathcal{K}$, then for $G \in \mathcal{G}, H \in \mathcal{H}, K \in \mathcal{K}$

$$\mathbf{E}[\mathbf{P}(G|\mathcal{K}), H \cap K] = \mathbf{E}[\mathbf{P}(G|\mathcal{K})\mathbf{P}(H|\mathcal{K}), K] = \mathbf{E}[\mathbf{P}(G \cap H|\mathcal{K}), K] = \mathbf{P}(G \cap H \cap K).$$

Now we can show that the set system

$$\mathcal{D} := \{A \in \sigma(\mathcal{H}, \mathcal{K}) : \mathbf{E}[\mathbf{P}(G|\mathcal{K}), A] = \mathbf{P}(G \cap A)\}$$

is a $\cap$-stable Dynkin system with $\mathcal{D} \supseteq \mathcal{H}, \mathcal{K}$. Now it follows from Theorem 1.13 that $\mathcal{D} = \sigma(\mathcal{H}, \mathcal{K})$, from which $\mathbf{P}(G|\sigma(\mathcal{H}, \mathcal{K})) = \mathcal{P}(G|\mathcal{K})$ follows.
'$\Leftarrow$': So if $\mathbf{P}(G|\sigma(\mathcal{H}, \mathcal{K})) = \mathbf{P}(G|\mathcal{K})$, it follows for $H \in \mathcal{H}$

$$\mathbf{P}(G \cap H|\mathcal{K}) = \mathbf{E}[\mathbf{P}(G|\sigma(\mathcal{H}, \mathcal{K})), H|\mathcal{K}] = \mathbf{E}[\mathbf{P}(G|\mathcal{K}), H|\mathcal{K}] = \mathbf{P}(G|\mathcal{K}) \cdot \mathbf{P}(H|\mathcal{K}).$$

□

Example 11.19 (Markov chains)**.** *Let's look again at the Markov chain $(X_t)_{t=0,1,2,\dots}$ from Example 11.14. For fixed t we set $\mathcal{G} = \sigma(X_{t+1}), \mathcal{H} = \sigma(X_0, \dots, X_{t-1}), \mathcal{K} = \sigma(X_t)$. The Markov property (11.4) now says for $G \in \mathcal{G}, H \in \mathcal{H}, K \in \mathcal{K}$ that $\mathbf{P}(G|\sigma(\mathcal{H}, \mathcal{K})) = \mathbf{P}(G|\mathcal{K})$. According to Proposition 11.18 this means that X_{t+1} and $(X_0, \dots, X_{t-1})$ are independent given X_t.*

11.5 Regular version of the conditional distribution

We have seen in Section 11.1 how the conditional probability $\mathbf{P}(A|\mathcal{G}) := \mathbf{E}[1_A|\mathcal{G}]$ for a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ is defined. However, this does *not* mean that we have a probability measure $A \mapsto \mathbf{P}(A|\mathcal{G})$; see the next remark. In most cases, however, one can define such a (random, $\mathcal{G}$-measurable) measure, the (or better: a) regular version of the conditional distribution.

Remark 11.20 (Conditional probabilities and conditional distributions)**.** *Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra and $A_1, A_2, \dots \in \mathcal{F}$ with $A_i \cap A_j = \emptyset$. Then, for $B \in \mathcal{G}$*

$$\begin{aligned}\mathbf{E}\Big[\mathbf{P}\Big(\bigcup_{n=1}^{\infty} A_n|\mathcal{G}\Big); B\Big] &= \mathbf{E}\big[\mathbf{E}[1_{\bigcup_{n=1}^{\infty} A_n}|\mathcal{G}]; B\big] = \mathbf{E}[1_{\bigcup_{n=1}^{\infty} A_n}; B] \\ &= \mathbf{E}\Big[\sum_{n=1}^{\infty} 1_{A_n}; B\Big] = \sum_{n=1}^{\infty} \mathbf{E}[1_{A_n}; B] \\ &= \sum_{n=1}^{\infty} \mathbf{E}\big[\mathbf{P}(A_n|\mathcal{G}); B\big] = \mathbf{E}\big[\sum_{n=1}^{\infty} \mathbf{P}(A_n|\mathcal{G}); B\big]\end{aligned}$$

and therefore

$$\mathbf{P}\Big(\bigcup_{n=1}^{\infty} A_n|\mathcal{G}\Big) = \sum_{n=1}^{\infty} \mathbf{P}(A_n|\mathcal{G}) \tag{11.6}$$

$\mathbf{P}$-almost surely. This means that there is a zero set (depending on $A_1, A_2, \dots$) so that (11.6) *applies to all ω outside this zero set. However, since there are uncountably many sequences $A_1, A_2, \dots \in \mathcal{F}$, there does not have to be a zero set N, so that* (11.6) *holds for every choice of $A_1, A_2, \dots \in \mathcal{F}$ outside of N. However, if there is such an N, we will say that a* regular version of the conditional distribution *of $\mathbf{P}$ given $\mathcal{G}$ exists. We will give conditions for this in Theorem 11.23.*

We recall the concept of the stochastic kernel; see Definition 5.9.

Definition 11.21 (Regular version of the conditional distribution)**.** *Let $(\Omega', \mathcal{F}')$ be a measurable space, Y an Ω'-valued measurable random variable and $\mathcal{G} \subseteq \mathcal{F}$. A stochastic kernel $\kappa_{Y,\mathcal{G}}$ from $(\Omega, \mathcal{G})$ to $(\Omega', \mathcal{F}')$ is called* regular version of the conditional distribution *of Y, given $\mathcal{G}$, if*

$$\kappa_{Y,\mathcal{G}}(\omega, B) = \mathbf{P}(Y \in B|\mathcal{G})(\omega)$$

for $\mathbf{P}$-almost all ω and every $B \in \mathcal{F}'$.

Remark 11.22 (Distribution conditional on a random variable)**.** *1. For the stochastic kernel from Definition 11.21 it is sufficient to use property (ii) from Definition 5.9 only for a $\cap$-stable generator $\mathcal{C}$ of $\mathcal{F}$. This is because*

$$\mathcal{D} := \{A' \in \mathcal{F}' : \omega \mapsto \kappa(\omega, A') \text{ is } \mathcal{A}\text{-measurable}\}$$

is alwaysa Dynkin system. Thus, according to Theorem 1.13, $\mathcal{D} = \sigma(\mathcal{C})$.

2. Let $\mathcal{G} = \sigma(X)$ for a random variable X in Definition 11.21.2. Then, if $\kappa_{Y,\sigma(X)}$ is a regular version of the conditional expectation of Y given $\sigma(X)$, then $\omega \mapsto \kappa_{Y,\sigma(X)}(\omega, A')$

$\sigma(X)$-is measurable for all $A' \in \mathcal{A}'$. This means that, according to Proposition 11.7, there is a $\sigma(X)/\mathcal{B}([0;1])$-measurable map $\varphi_{A'} : \Omega \to [0;1]$ with $\varphi_{A'} \circ X = \kappa_{Y,\sigma(X)}(., A')$. We then set

$$\kappa_{Y,X}(x, A') := \varphi_{A'}(x)$$

and say $\kappa_{Y,X}$ is the regular version of the conditional distribution of Y given X.

Theorem 11.23 (Existence of the regular version of the conditional distribution)**.** *Let (E, r) be a complete and separable metric space space equipped with Borel's σ-algebra, $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra and Y a (according to $\mathcal{F}$ measurable) random variable with values in E. Then there exists a regular version of the conditional distribution of Y given $\mathcal{G}$.*

Before we can prove the theorem, we need a property (Proposition 11.25) over complete, separable metric spaces.

Definition 11.24 (Borel space)**.** *1. Two metric spaces $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ are called* isomorphic *if there is a bijective, according to $\mathcal{F}/\mathcal{F}'$-measurable mapping $\varphi : \Omega \to \Omega'$ exists such that φ^{-1} is $\mathcal{F}'/\mathcal{F}$-measurable.*

2. A measurable space $(\Omega, \mathcal{F})$ is called Borel space *if there is a Borel set $A \in \mathcal{B}(\mathbb{R})$ exists such that $(\Omega, \mathcal{F})$ and $(A, \mathcal{B}(A))$ are isomorphic.*

Proposition 11.25 (Polish and Borel spaces)**.** *Every complete and separable metric space (E, r), equipped with the Borel's σ-algebra, is a Borel space.*

Proof. See, for example, Dudley, Real analysis and probability, Theorem 13.1.1. □

Proof of theorem 11.23. We prove the theorem under the weaker condition that E, equipped with the Borel σ-algebra, is a Borel space. Wlog, we can therefore assume that $E \in \mathcal{B}(\mathbb{R})$ is. The strategy of our proof consists of finding a distribution function of the conditional distribution by first fixing it for rational values before extending it to all real numbers.

For $r \in \mathbb{Q}$, let F_r be a version of $\mathbf{P}(Y \leq r|\mathcal{G})$ (i.e. $F_r = \mathbf{P}(Y \leq r|\mathcal{G})$ almost surely. Let $A \in \mathcal{F}$ be such that for $\omega \in A$ the mapping $r \mapsto F_r(\omega)$ is non-increasing with limits 1 and 0 at $\pm\infty$. Since A is given by countably many conditions, all of which are almost certainly fulfilled, $\mathbf{P}(A) = 1$. Now define for $x \in \mathbb{R}$

$$F_x(\omega) := 1_A(\omega) \cdot \inf_{r>x} F_r(\omega) + 1_{A^c}(\omega) \cdot 1_{x \geq 0}.$$

Thus, $x \mapsto F_x(\omega)$ is a distribution function for all ω. Define

$$\kappa(\omega, .) := \text{measure defined by } x \mapsto F_x(\omega).$$

For $r \in \mathbb{Q}$ and $B = (-\infty; r]$,

$$\omega \mapsto \kappa(\omega, B) = 1_A(\omega) \cdot \mathbf{P}(Y \leq r|\mathcal{G})(\omega) + 1_{A^c}(\omega) \cdot 1_{r \geq 0} \tag{11.7}$$

is $\mathcal{F}$-measurable. Since $\{(-\infty; r] : r \in \mathbb{Q}\}$ is a $\cap$-stable generator of $\mathcal{B}(\mathbb{R})$, according to Remark 11.22 the mapping $\omega \mapsto \kappa(\omega, B)$ is measurable for all $B \in \mathcal{F}$. Therefore, κ is a stochastic kernel.

It remains to show that κ is a regular version of the conditional distribution. Since (11.7) is based on a $\cap$-stable generator of $\mathcal{E}$, for $\omega \in A$

$$\kappa(\omega, B) = \mathbf{P}(Y \in B|\mathcal{G})(\omega).$$

In other words, κ is a regular version of the conditional distribution. □