Stochastic Processes / Stochastik II
course notes
winter semester 2023/24

Markus Reiß
Humboldt-Universität zu Berlin

Preliminary version, 15. Februar 2024

Contents

1 Some important processes **1**
1.1 The Poisson process 1
1.2 Markov chains 6

2 General theory of stochastic processes **10**
2.1 Basic notions 10
2.2 Polish spaces and Kolmogorov's consistency theorem 12

3 The conditional expectation **19**
3.1 Orthogonal projections 19
3.2 Construction and properties 20

4 Martingale theory **28**
4.1 Martingales, sub- and supermartingales 28
4.2 Stopping times 32
4.3 Martingale inequalities and convergence 38
4.4 The Radon-Nikodym theorem and Kakutani's dichotomy 48

5 Ergodic theory and Markov chains **55**
5.1 Stationary and ergodic processes 55
5.2 Recurrence, transience and ergodicity of Markov chains 59
5.3 Ergodic theorems 68
5.4 Convergence of Markov chains and MCMC 72

6 Weak and functional convergence **82**
6.1 Fundamental properties 82
6.2 Tightness 86
6.3 Weak convergence on $C([0,T])$, $C(\mathbb{R}^+)$ 89

7 Invariance principle and the empirical process **94**

7.1 Invariance principle and Brownian motion 94

7.2 Empirical process and Brownian bridge 99

1 Some important processes

1.1 The Poisson process

1.1 Example. We count the number N_t of emissions of a radioactive substance during the time interval $[0,t]$ for $t \in [0,\infty)$. Since radioactivity is genuinely random, $N_t = N_t(\omega)$ is a random variable with values in $\mathbb{N}_0$. We write N_t, $N(t)$, $N_t(\omega)$, $N(t,\omega)$ synonymously.

1.2 Definition. Let $(S_k)_{k\geqslant 1}$ be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$ with $0 \leqslant S_1(\omega) \leqslant S_2(\omega) \leqslant \cdots$ for all $\omega \in \Omega$. Then $N = (N_t,\, t \geqslant 0)$ with

$$N_t := \sum_{k\geqslant 1} \mathbf{1}(S_k \leqslant t), \quad t \geqslant 0,$$

is called counting process (Zählprozess) with jump times (Sprungzeiten) (S_k).

1.3 Example. For the radioactive emissions physical modeling assumes that on small time intervals the probability for more than one emission is negligible and the probability for one emission is proportional to the length of the interval. Moreover, the number of emissions during disjoint time intervals is independent and its distribution only depends on the length of the time interval, not on the time points itself. Hence, we model N as a Poisson process in the following sense.

1.4 Definition. A counting process N is called Poisson process of intensity $\lambda > 0$ if

(i) $\mathbb{P}(N_{t+h} - N_t = 1) = \lambda h + o(h)$ for $h \downarrow 0$ and all $t \geqslant 0$;

(ii) $\mathbb{P}(N_{t+h} - N_t = 0) = 1 - \lambda h + o(h)$ for $h \downarrow 0$ and all $t \geqslant 0$;

(iii) (independent increments) $(N_{t_i} - N_{t_{i-1}})_{1\leqslant i\leqslant n}$ are independent for all $n \in \mathbb{N}$ and $0 = t_0 < t_1 < \cdots < t_n$;

(iv) (stationary increments) $N_t - N_s \overset{d}{=} N_{t-s}$ for all $t \geqslant s \geqslant 0$.

1.5 Remark. The 'small-o' $o(h)$ denotes a function f of h with $\lim_{h\downarrow 0} \frac{f(h)}{h} = 0$, in other words (i) is equivalent to $\lim_{h\downarrow 0} \frac{\mathbb{P}(N_{t+h}-N_t=1)}{h} = \lambda$. The notation $X \overset{d}{=} Y$ means that X and Y have the same law, i.e. $\mathbb{P}^X = \mathbb{P}^Y$.

1.6 Theorem. *For a counting process N with jump times (S_k) the following are equivalent:*

(a) N is a Poisson process;

(b) N satisfies conditions (iii),(iv) of a Poisson process and $N_t \sim \mathrm{Poiss}(\lambda t)$ holds for all $t \geqslant 0$ (setting $\mathrm{Poiss}(0) = \delta_0$);

(c) $T_1 := S_1$, $T_k := S_k - S_{k-1}$, $k \geqslant 2$, are i.i.d. $\mathrm{Exp}(\lambda)$-distributed random variables;

(d) *$N_t \sim \operatorname{Poiss}(\lambda t)$ holds for all $t \geqslant 0$ and the law of $(S_1, \dots, S_n)$ given $\{N_t = n\}$ has for all $n \in \mathbb{N}$ the density*

$$f(x_1, \dots, x_n) = \tfrac{n!}{t^n} \mathbf{1}(0 \leqslant x_1 \leqslant \cdots \leqslant x_n \leqslant t), \quad x_1, \dots, x_n \in \mathbb{R}. \tag{1.1}$$

(e) *N satisfies condition (iii) of a Poisson process, $\mathbb{E}[N_1] = \lambda$ and* (1.1) *is the density of $(S_1, \dots, S_n)$ given $\{N_t = n\}$ for all $n \in \mathbb{N}$, $t > 0$.*

1.7 Remark. Let $U_1, \dots U_n \sim U([0,t])$ i.i.d. and consider their order statistics $U_{(1)}, \dots, U_{(n)}$, i.e. $U_{(1)} = \min_i U_i$, $U_{(2)} = \min(\{U_1, \dots, U_n\} \setminus \{U_{(1)}\})$ etc. Then $(U_{(1)}, \dots, U_{(n)})$ has exactly density (1.1) ►Control.

The characterisations give rise to three simple methods to simulate a Poisson process: (i) The definition gives an approximation for small h (forgetting the $o(h)$-term) and we may recursively use $N_{(k+1)h} \approx N_{kh} + \varepsilon_k$ with independent $\operatorname{Bin}(1, \lambda h)$-distributed random variables ε_k. (ii) Part (c) just uses exponentially distributed inter-arrival times T_k and we may set $N_t = \sum_{k \geqslant 1} \mathbf{1}(T_1 + \cdots + T_k \leqslant t)$. (iii) By part (d) we simulate $N_T \sim \operatorname{Poiss}(\lambda T)$ at a specified right-end point $T > 0$ and then use N_T independent $U([0,T])$-distributed random variables as jump times in-between. Note that (c) ensures also the existence of a Poisson process because we can construct a probability space with independent $\operatorname{Exp}(\lambda)$-distributed random variables $(T_k)_{k \geqslant 1}$, giving rise to $(S_k)_{k \geqslant 1}$ and thus N.

Proof. We prove the equivalence by a circular argument.

(a)⇒(b) Put $p_n(t) = \mathbb{P}(N_t = n)$. By (i), (ii), (iii) we infer

$$p_0(t+h) = \mathbb{P}(N_t = 0, N_{t+h} - N_t = 0) = p_0(t)(1 - \lambda h + o(h)),$$

which implies

$$p_0'(t) = \lim_{h \downarrow 0} \frac{p_0(t+h) - p_0(t)}{h} = -\lambda p_0(t), \quad t \geqslant 0.$$

In view of $p_0(0) = 1$ (from (iv) with $t = s$) we obtain $p_0(t) = e^{-\lambda t}$.

Similarly, we have for $n \geqslant 1$:

$$\begin{aligned} p_n(t+h) &= \mathbb{P}(\{N_{t+h} = n\} \cap (\{N_t \leqslant n-2\} \cup \{N_t = n-1\} \cup \{N_t = n\})) \\ &= \mathbb{P}(N_t \leqslant n-2) o(h) + \mathbb{P}(N_t = n-1)(\lambda h + o(h)) \\ &\quad + \mathbb{P}(N_t = n)(1 - \lambda h + o(h)) \\ &= p_{n-1}(t) \lambda h + p_n(t)(1 - \lambda h) + o(h). \end{aligned}$$

This implies $p_n'(t) = -\lambda p_n(t) + \lambda p_{n-1}(t)$. Using $p_n(0) = 0$ we deduce inductively $p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$.

(b)⇒(c) Let $0 = b_0 \leqslant a_1 < b_1 \leqslant \dots \leqslant a_n < b_n$ and calculate

$$
\begin{aligned}
&\mathbb{P}\Big(\bigcap_{k=1}^{n}\{a_k < S_k \leqslant b_k\}\Big)\\
&= \mathbb{P}\Big(\bigcap_{k=1}^{n-1}\{N_{a_k} - N_{b_{k-1}} = 0, N_{b_k} - N_{a_k} = 1\} \cap \{N_{a_n} - N_{b_{n-1}} = 0, N_{b_n} - N_{a_n} \geqslant 1\}\Big)\\
&\overset{(iii),(iv)}{=} \Big(\prod_{k=1}^{n-1} \mathbb{P}(N_{a_k - b_{k-1}} = 0)\,\mathbb{P}(N_{b_k - a_k} = 1)\Big)\,\mathbb{P}(N_{a_n - b_{n-1}} = 0)\,\mathbb{P}(N_{b_n - a_n} \geqslant 1)\\
&= \Big(\prod_{k=1}^{n-1} \lambda(b_k - a_k) e^{-\lambda(b_k - a_k) - \lambda(a_k - b_{k-1})}\Big) e^{-\lambda(a_n - b_{n-1})}(1 - e^{-\lambda(b_n - a_n)})\\
&= \lambda^{n-1}\Big(\prod_{k=1}^{n-1}(b_k - a_k)\Big)(e^{-\lambda a_n} - e^{-\lambda b_n}).
\end{aligned}
$$

Taking derivatives with respect to $b_1, \dots, b_n$ we obtain the density of $(S_1, \dots, S_n)$:

$$f^{S_1,\dots,S_n}(b_1, \dots, b_n) = \lambda^n e^{-\lambda b_n}\mathbf{1}(0 \leqslant b_1 \leqslant b_2 \leqslant \dots \leqslant b_n),$$

noting the order $S_1 \leqslant \dots \leqslant S_n$ which implies the indicator function. Consequently, $(T_1, T_2, \dots, T_n) = (S_1, S_2 - S_1, \dots, S_n - S_{n-1})$ has density $\lambda^n e^{-\lambda(x_1 + \dots + x_n)}\mathbf{1}(x_1, \dots, x_n \geqslant 0)$ (density transformation ▶Control). The product density form implies that $T_1, \dots, T_n$ are independent (Stochastik I!) and that each T_i is Exp(λ)-distributed.

(c)⇒(d) We find $\mathbb{P}(N_t = 0) = \mathbb{P}(S_1 > t) = e^{-\lambda t}$ and

$$\mathbb{P}(N_t = n) = \mathbb{P}(N_t \geqslant n) - \mathbb{P}(N_t \geqslant n + 1) = \mathbb{P}(S_n \leqslant t) - \mathbb{P}(S_{n+1} \leqslant t).$$

Since $S_n = T_1 + \dots + T_n$ is $\Gamma(\lambda, n)$-distributed (Stochastik I !), we obtain

$$\mathbb{P}(N_t = n) = \int_0^t \Big(\frac{\lambda^n x^{n-1}}{(n-1)!} - \frac{\lambda^{n+1}x^n}{n!}\Big)e^{-\lambda x}dx = \frac{(\lambda x)^n}{n!}e^{-\lambda x}\Big|_{x=0}^{t} = \frac{(\lambda t)^n}{n!}e^{-\lambda t}$$

for $n \geqslant 1$ and we conclude $N_t \sim \text{Poiss}(\lambda t)$. By density transformation the joint density of $(S_1, \dots, S_{n+1})$ is for $s_{n+1} \geqslant s_n \geqslant \dots \geqslant s_1 \geqslant s_0 = 0$

$$f^{S_1,\dots,S_{n+1}}(s_1, \dots, s_{n+1}) = \prod_{k=1}^{n+1} \lambda e^{-\lambda(s_k - s_{k-1})} = \lambda^{n+1}e^{-\lambda s_{n+1}}.$$

Noting $\{N_t = n\} = \{S_n \leqslant t, S_{n+1} > t\}$ we consider $0 \leqslant a_1 < b_1 \leqslant \dots \leqslant a_n < b_n \leqslant t$ and obtain the conditional law of the jump times via

$$
\begin{aligned}
&\mathbb{P}(S_1 \in [a_1, b_1], \dots, S_n \in [a_n, b_n] \,|\, N_t = n)\\
&= \frac{\mathbb{P}(S_1 \in [a_1, b_1], \dots, S_n \in [a_n, b_n], S_{n+1} > t)}{\frac{(\lambda t)^n}{n!}e^{-\lambda t}}\\
&= \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \frac{n!}{t^n}\mathbf{1}(0 \leqslant s_1 \leqslant \dots \leqslant s_n \leqslant t)\, ds_n \cdots ds_1
\end{aligned}
$$

(use $\int_t^\infty \lambda^{n+1} e^{-\lambda s_{n+1}} ds_{n+1} = \lambda^n e^{-\lambda t}$), which identifies the integrand as the conditional density.

(d)⇒(e) $\mathbb{E}[N_1] = \lambda$ is direct from the assumption. For $0 = t_0 < t_1 < \cdots t_n = t$ and $k_1, \dots, k_n \in \mathbb{N}_0$ consider with $K := \sum_{l=1}^n k_l$

$$\begin{aligned}
&\mathbb{P}(\forall l = 1, \dots, n : \, N_{t_l} - N_{t_{l-1}} = k_l) \\
&= \mathbb{P}(N_t = K)\, \mathbb{P}(\forall l = 1, \dots, n : \, N_{t_l} - N_{t_{l-1}} = k_l \,|\, N_t = K) \\
&= \frac{(\lambda t)^K}{K!} e^{-\lambda t}\, \mathbb{P}(S_{k_1} \leqslant t_1 < S_{k_1+1}, \dots, S_K \leqslant t_n < S_{K+1} \,|\, N_t = K) \\
&= \frac{(\lambda t)^K}{K!} e^{-\lambda t} \frac{K!}{t^K} \prod_{l=1}^n \frac{(t_l - t_{l-1})^{k_l}}{k_l!} = \prod_{l=1}^n \frac{\big(\lambda(t_l - t_{l-1})\big)^{k_l}}{k_l!} e^{-\lambda(t_l - t_{l-1})} \\
&= \prod_{l=1}^n \mathbb{P}(N_{t_l} - N_{t_{l-1}} = k_l).
\end{aligned}$$

The last identity follows from the marginal probability

$$\mathbb{P}(N_{t_{l_0}} - N_{t_{l_0-1}} = k_{l_0}) = \frac{\big(\lambda(t_{l_0} - t_{l_0-1})\big)^{k_{l_0}}}{k_{l_0}!} e^{-\lambda(t_{l_0} - t_{l_0-1})}, \quad l_0 = 1, \dots, n,$$

obtained from the previous calculation by summing $\mathbb{P}(\forall l = 1, \dots, n : N_{t_l} - N_{t_{l-1}} = k_l)$ over $k_l \in \mathbb{N}_0$ for all $l \neq l_0$. Hence, $(N_{t_l} - N_{t_{l-1}})_{l \geqslant 1}$ are independent (the joint counting density is the product of the marginal counting densities).

(e)⇒(a) For $0 = t_0 < t_1 < \cdots t_n = t$ and $k_1, \dots, k_n \in \mathbb{N}_0$, $h > 0$, $m \geqslant k_1 + \cdots + k_n =: K$ note the shift invariance

$$\begin{aligned}
&\mathbb{P}(\forall l = 1, \dots, n : \, N_{t_l+h} - N_{t_{l-1}+h} = k_l \,|\, N_{t+h} = m) \\
&= \frac{m!}{(t+h)^m} \frac{h^{m-K}}{(m-K)!} \prod_{l=1}^m \frac{(t_l + h - (t_{l-1} + h))^{k_l}}{k_l!} \\
&= \mathbb{P}(\forall l = 1, \dots, n : \, N_{t_l} - N_{t_{l-1}} = k_l \,|\, N_{t+h} = m),
\end{aligned}$$

where we used that the first conditional probability is the probability that for m independent $U([0, t+h])$-distributed random variables $m - K$ end up in $[0, h]$ and k_l in $[t_{l-1} + h, t_l + h]$ for each l by the order statistics interpretation of Remark 1.7. We thus have

$$\mathbb{P}(\forall l : \, N_{t_l+h} - N_{t_{l-1}+h} = k_l, N_{t+h} = m) = \mathbb{P}(\forall l : \, N_{t_l} - N_{t_{l-1}} = k_l, N_{t+h} = m)$$

and summing up over all $m \geqslant k_1 + \cdots + k_n$ yields identity in law:

$$(N_{t_1+h} - N_{t_0+h}, \dots, N_{t_n+h} - N_{t_{n-1}+h}) \overset{d}{=} (N_{t_1} - N_{t_0}, \dots, N_{t_n} - N_{t_{n-1}}).$$

This gives (iv) (put $n = 1$ and observe $N_0 = 0$ a.s. due to the existence of a density for S_1) and for $0 < h < 1$

$$\mathbb{P}(N_h = 0) = \sum_{k=0}^\infty \mathbb{P}(N_1 = k)\, \mathbb{P}(N_1 - N_h = k \,|\, N_1 = k) = \sum_{k=0}^\infty \mathbb{P}(N_1 = k)(1-h)^k.$$

Because of $\sum_{k=0}^{\infty}\mathbb{P}(N_1=k)k=\mathbb{E}[N_1]=\lambda<\infty$ the function $p(h):=\mathbb{P}(N_h=0)$ is differentiable in $[0,1]$ with $p'(0)=-\lambda$. We conclude

$$\mathbb{P}(N_h=0)=\mathbb{P}(N_0=0)-\lambda h+o(h)=1-\lambda h+o(h).$$

By a similar argument, $\mathbb{P}(N_h=1)$ equals

$$\sum_{k=1}^{\infty}\mathbb{P}(N_1=k)\,\mathbb{P}(N_1-N_h=k-1\,|\,N_1=k)=\sum_{k=1}^{\infty}\mathbb{P}(N_1=k)k(1-h)^{k-1}h,$$

and this implies $\mathbb{P}(N_h=1)=\lambda h+o(h)$.

□

▷ **Control questions**

(a) Replace in the definition of a Poisson process conditions (i), (ii) by $\mathbb{P}(N_{t+h}-N_t=1)=\lambda h^2+o(h^2)$, $\mathbb{P}(N_{t+h}-N_t=0)=1-\lambda h^2+o(h^2)$ and show $\mathbb{P}(N_t=0)=1$ for such a process N.

This follows as in the proof for the implication (a)⇒(b): we have $p_0'(t)=\lim_{h\downarrow 0}\frac{\mathbb{P}(N_{t+h}-N_t=0)}{h}=0$ and with $p_0(0)=1$ this implies $p_0(t)=1$ for all $t\geqslant 0$. In fact, the condition $\mathbb{P}(N_{t+h}-N_t=0)=1-o(h)$ suffices for this. So, one could define a Poisson process of intensity $\lambda=0$ to be a process with $N_t=0$ all the time.

(b) Write down explicitly the density transformation step from $(S_1,\dots,S_n)$ to $(T_1,\dots,T_n)$ in proof part (b)⇒(c).

If we write $T=(T_1,\dots,T_n)^\top$, $S=(S_1,\dots,S_n)^\top$, then $T=g(S)$ holds with a linear function g whose inverse is $g^{-1}(t)=(t_1,t_1+t_2,\dots,t_1+\dots+t_n)^\top$. By the density transformation formula, we have for the densities $f^T(x)=f^S(g^{-1}(x))|\det(D_{g^{-1}}(x))|$ with the Jacobi matrix $(D_{g^{-1}}(x))_{i,j}=\mathbf{1}(j\leqslant i)$. Consequently, $\det(D_{g^{-1}}(X))=1$ holds and the result follows from $(g^{-1}(x))_n=x_1+\dots+x_n$ and $(g^{-1}(x))_i\leqslant(g^{-1}(x))_{i+1}\iff x_{i+1}\geqslant 0$.

(c) Derive the density of the order statistics $U_{(1)},\dots,U_{(n)}$ for independent $U_i\sim U([0,1])$ by showing for all $0\leqslant t_1<t_1+h_1\leqslant t_2<t_2+h_2\leqslant\dots\leqslant t_n<t_n+h_n\leqslant 1$:

$$\begin{aligned}&\mathbb{P}(U_{(1)}\in[t_1,t_1+h_1],\dots,U_{(n)}\in[t_n,t_n+h_n])\\&=\sum_{\pi\in S_n}\mathbb{P}(U_{\pi(1)}\in[t_1,t_1+h_1],\dots,U_{\pi(n)}\in[t_n,t_n+h_n])=n!\prod_{i=1}^{n}h_i\end{aligned}$$

with the set S_n of all permutations π of $\{1,\dots,n\}$.

Define the random permutation $\Pi\in S_n$ via $U_{\Pi(i)}=U_{(i)}$ for $i=1,\dots,n$. Π is almost surely uniquely defined ($\mathbb{P}(U_i=U_j)=0$ for $i\neq j$) and called *inverse rank*. Then $\{U_{(1)}\in[t_1,t_1+h_1],\dots,U_{(n)}\in[t_n,t_n+h_n]\}$ equals $\{U_{\Pi(1)}\in[t_1,t_1+h_1],\dots,U_{\Pi(n)}\in[t_n,t_n+h_n]\}$. We obtain

$$\begin{aligned}&\mathbb{P}(U_{(1)}\in[t_1,t_1+h_1],\dots,U_{(n)}\in[t_n,t_n+h_n])\\&=\sum_{\pi\in S_n}\mathbb{P}(U_{\pi(1)}\in[t_1,t_1+h_1],\dots,U_{\pi(n)}\in[t_n,t_n+h_n];\Pi=\pi).\end{aligned}$$

In the last probability we may omit the condition $\Pi = \pi$ because this follows from $U_{\pi(1)} \leqslant \cdots \leqslant U_{\pi(n)}$. Moreover, by independence this probability equals $\prod_{i=1}^n \mathbb{P}(U_{\pi(i)} \in [t_i, t_i + h_i]) = \prod_{i=1}^n h_i$, implying

$$\mathbb{P}(U_{(1)} \in [t_1, t_1 + h_1], \dots, U_{(n)} \in [t_n, t_n + h_n]) = n! \prod_{i=1}^n h_i.$$

This equals the density integral $\int_{t_1}^{t_1+h_1} \cdots \int_{t_n}^{t_n+h_n} f dx$ for $f(x) = n! \mathbf{1}(0 \leqslant x_1 \leqslant \cdots \leqslant x_n \leqslant 1)$. In view of the ordering $U_{(1)} \leqslant \cdots U_{(n)}$ this suffices to identify f as the density of the order statistics (compare class).

1.2 Markov chains

1.8 Example. In many models we can consider a process as *memoryless* in the sense that the future evolution only depends on the present value of the process and not on its history (the way it has developed before). The Poisson process is a typical example because at present time $t \geqslant 0$ its law at the future time $t + h$, $h > 0$, given the past and present values only depends on the present value, formally for any $0 < t_1 < \cdots < t_n < t$ with $t_0 =$, $i_0 = 0$, $t_{n+1} = t$:

$$\begin{aligned}
&\mathbb{P}(N_{t+h} = k \,|\, N_{t_1} = i_1, \dots, N_{t_n} = i_n, N_t = i_{n+1}) \\
&= \mathbb{P}(N_{t+h} - N_t = k - i_{n+1} \,|\, N_{t_1} = i_1, \dots, N_{t_{n+1}} = i_{n+1}) \\
&= \mathbb{P}(N_{t_{n+1}+h} - N_{t_{n+1}} = k - i_{n+1} \,|\, N_{t_1} - N_{t_0} = i_1 - i_0, \dots, N_{t_{n+1}} - N_{t_n} = i_{n+1} - i_n) \\
&= \mathbb{P}(N_{t_{n+1}+h} - N_{t_{n+1}} = k - i_{n+1}) \\
&= \mathbb{P}(N_{t_{n+1}+h} = k \,|\, N_{t_{n+1}} = i_{n+1}),
\end{aligned}$$

where we have only used (iii) and (iv) that a Poisson process has independent increments▶Control. The values $i_1, \dots, i_n, i, k$ are chosen so that all expressions are well defined. This property of memorylessness in law is called Markovianity and plays the same role as ordinary differential equations in analysis, compared to more general delay differential equations.

1.9 Definition. Let $T = \mathbb{N}_0$ (discrete time) or $T = [0, \infty)$ (continuous time) and S be a countable set (state space). Then random variables $(X_t)_{t \in T}$ with values in $(S, \mathcal{P}(S))$ form a Markov chain if for all $n \in \mathbb{N}$, $t_1 < t_2 < \cdots < t_{n+1}$, $i_1, \dots, i_{n+1} \in S$, satisfying $\mathbb{P}(X_{t_1} = i_1, \dots, X_{t_n} = i_n) > 0$, the Markov property is satisfied:

$$\mathbb{P}(X_{t_{n+1}} = i_{n+1} \,|\, X_{t_1} = i_1, \dots, X_{t_n} = i_n) = \mathbb{P}(X_{t_{n+1}} = i_{n+1} \,|\, X_{t_n} = i_n).$$

1.10 Remark. $\mathcal{P}(S) = \{A \,|\, A \subseteq S\}$ denotes the power set of S.

1.11 Definition. For a Markov chain X and $t_1 \leqslant t_2$, $i, j \in S$

$$p_{ij}(t_1, t_2) := \mathbb{P}(X_{t_2} = j \,|\, X_{t_1} = i) \text{ (or zero if not well-defined)}$$

defines the transition probability to reach state j at time t_2 from state i at time t_1. The transition matrix is given by

$$P(t_1, t_2) := (p_{ij}(t_1, t_2))_{i,j \in S}.$$

The transition matrix and the Markov chain are called time-homogeneous if $P(t_1, t_2)$ only depends on $t_2 - t_1$ (whenever well-defined). We then set $p_{ij}(t) := p_{ij}(t_1, t_1 + t)$, $P(t) := P(t_1, t_1 + t)$.

1.12 Example. As the discussion of the Poisson process shows, every process with independent increments is a Markov process (this is true in general, here we restrict to countable state space like $S = \mathbb{Z}$). Moreover, the Markov chain is time homogeneous. A discrete-time example is the random walk

$$S_n = \sum_{i=1}^{n} X_i \text{ with i.i.d. integer-valued random variables } X_i$$

for $n \geqslant 1$, $S_0 = 0$. Then the increments $S_n - S_{n-1}, \dots, S_1 - S_0$ equal $X_n, \dots, X_1$ and are by assumption independent and identically distributed, hence stationary ▶Control. A specific example is the simple (asymmetric) random walk with $\mathbb{P}(X_i = +1) = p$, $\mathbb{P}(X_i = -1) = q = 1 - p$, for which the first two transition matrices read

$$P(1) = \begin{pmatrix} \ddots & & & & \\ q & 0 & p & & \\ & \ddots & \ddots & \ddots & \\ & & q & 0 & p \\ & & & & \ddots \end{pmatrix}, \; P(2) = \begin{pmatrix} & \ddots & & & & & \\ q^2 & 0 & 2pq & 0 & p^2 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & q^2 & 0 & 2pq & 0 & p^2 \\ & & & & & & \ddots \end{pmatrix}$$

with all other entries zero. Using the Markov property (where?), we calculate

$$\begin{aligned} P(2)_{i,k} &= \mathbb{P}(S_{n+2} = k \mid S_n = i) \\ &= \sum_{j \in \mathbb{Z}} \mathbb{P}(S_{n+1} = j \mid S_n = i)\, \mathbb{P}(S_{n+2} = k \mid S_{n+1} = j) = \sum_{j \in \mathbb{Z}} P(1)_{i,j} P(1)_{j,k}. \end{aligned}$$

The important observation is that the *two-step transition matrix* $P(2)$ satisfies $P(2) = P(1)P(1)$ in terms of (infinite) matrix multiplication, which we now show in general.

1.13 Proposition. *The transition matrices of a Markov chain satisfy the Chapman-Kolmogorov equation*

$$\forall t_1 \leqslant t_2 \leqslant t_3 : P(t_1, t_3) = P(t_1, t_2) P(t_2, t_3) \text{ (matrix multiplication).}$$

In the time-homogeneous case this gives the semigroup property

$$\forall t, s \in T : P(t + s) = P(t) P(s),$$

in particular $P(n) = P(1)^n$ for $n \in \mathbb{N}$.

1.14 Remark. For an infinite state space S the matrix multiplication is well defined because all entries are in $[0, 1]$, the sum in each row is one (or smaller if some entries are not well defined) and thus $\sum_{j \in S} P(t_1, t_2)_{i,j} P(t_2, t_3)_{j,k}$ is absolutely summable.

Proof. By definition we obtain in case $\mathbb{P}(X_{t_1} = i) > 0$

$$\begin{aligned}
P(t_1, t_3)_{ij} &= \mathbb{P}(X_{t_3} = j \,|\, X_{t_1} = i) \\
&= \sum_{k \in S} \mathbb{P}(X_{t_3} = j, X_{t_2} = k \,|\, X_{t_1} = i) \\
&= \sum_{k \in S: \mathbb{P}(X_{t_1} = i, X_{t_2} = k) > 0} \mathbb{P}(X_{t_3} = j \,|\, X_{t_1} = i, X_{t_2} = k)\, \mathbb{P}(X_{t_2} = k \,|\, X_{t_1} = i) \\
&\overset{\text{Markov}}{=} \sum_{k \in S: \mathbb{P}(X_{t_1} = i, X_{t_2} = k) > 0} \mathbb{P}(X_{t_3} = j \,|\, X_{t_2} = k)\, \mathbb{P}(X_{t_2} = k \,|\, X_{t_1} = i) \\
&= \sum_{k \in S} P(t_2, t_3)_{kj} P(t_1, t_2)_{ik} \\
&= (P(t_1, t_2) P(t_2, t_3))_{ij}.
\end{aligned}$$

In case $\mathbb{P}(X_{t_1} = i) = 0$ we have $P(t_1, t_3)_{i,j} = P(t_1, t_2)_{i,k} = 0$ for all j, k and the formula holds as well.

For time-homogeneous Markov chains this reduces to $P(t_3 - t_1) = P(t_2 - t_1) P(t_3 - t_2)$ and substituting $t = t_2 - t_1$, $s = t_3 - t_2$ yields the assertion. □

1.15 Example (Ehrenfest model, 1907)**.** We consider a model for the physical diffusion of gas. N gas molecules are contained in two containers, which are connected by an (infinitesimally) small tube. In every time step, one molecule is picked at random and leaves its container via the tube to the other one. Let X_n denotes the number of molecules in container 1 at time n. Then at time $n+1$ the probability that a molecule of container 1 is picked equals X_n/N, which then leads to $X_{n+1} = X_n - 1$. This way we obtain a Markov chain (X_n) on $S = \{0, 1, \dots N\}$ with transition probabilities

$$\mathbb{P}(X_{n+1} = i - 1 \,|\, X_n = i) = i/N, \quad \mathbb{P}(X_{n+1} = i + 1 \,|\, X_n = i) = (N - i)/N$$

for $0 \leqslant i \leqslant N$. It is time-homogeneous with tri-diagonal transition matrix

$$P(1) = \begin{pmatrix} 0 & N/N & & & \\ 1/N & 0 & (N-1)/N & & \\ & \ddots & \ddots & \ddots & \\ & & (N-1)/N & 0 & 1/N \\ & & & N/N & 0 \end{pmatrix} \in \mathbb{R}^{(N+1) \times (N+1)}$$

and all other entries zero.

1.16 Definition. For a time-homogeneous Markov chain (X_t) a probability measure μ on $(S, \mathcal{P}(S))$ is called invariant initial distribution if the transition matrices satisfy

$$\sum_{i \in S} \mu(\{i\}) P(t)_{i,j} = \mu(\{j\}) \text{ for all } j \in S,\ t \in T.$$

1.17 Remark. If μ is the law of X_0 (initial distribution of (X_n)), then

$$\mathbb{P}(X_t = j) = \sum_{i \in S} \mathbb{P}(X_0 = i)\, \mathbb{P}(X_t = j \,|\, X_0 = i) = \sum_{i \in S} \mu(\{i\}) P(t)_{i,j}$$

holds (convention: $0 \bullet x := 0$ even if x is not well-defined or infinity) and for an invariant initial distribution μ the X_t have law μ for all $t \in T$, that is $\mathbb{P}^{X_t}$ is invariant in time t.

1.18 Lemma. *In discrete time, μ is already an invariant initial distribution of (X_n) if*

$$\sum_{i \in S} \mu(\{i\}) P(1)_{i,j} = \mu(\{j\}) \text{ for all } j \in S.$$

This condition is equivalent to the row vector $\vec{\mu} = (\mu(\{j\}))_{j \in S}$ being an eigenvector of $P(1)$ for the eigenvalue 1: $\vec{\mu} P(1) = \vec{\mu}$.

Proof. The last assertion is just row-vector matrix multiplication: $(\vec{\mu} P(1))_j = \sum_{i \in S} \vec{\mu}_i P(1)_{i,j}$. In this formulation the Chapman-Kolmogorov equation yields for all $n \geqslant 1$

$$\vec{\mu} P(n) = (\vec{\mu} P(1)) P(1)^{n-1} = \vec{\mu} P(n-1) = \cdots = \vec{\mu},$$

which is the vector-matrix formulation of the defining condition for an invariant initial distribution. □

1.19 Example (Ehrenfest model II)**.** In the Ehrenfest model the Binomial law $\mu = \mathrm{Bin}(N, 1/2)$ with $\mu(\{j\}) = \binom{N}{j} 2^{-N}$ is an invariant initial distribution because

$$\sum_{i=0}^{N} \mu(\{i\}) P(1)_{i,j} = \binom{N}{j-1} 2^{-N} \frac{N-(j-1)}{N} + \binom{N}{j+1} 2^{-N} \frac{j+1}{N} = \binom{N}{j} 2^{-N},$$

using $\binom{N-1}{j-1} + \binom{N-1}{j} = \binom{N}{j}$ and considering the border cases $j \in \{0, N\}$ separately. Note that the law of the molecule numbers X_n in container 1 remains constant in time, while the actual realisations of X_n change dynamically!

▷ **Control questions**

(a) Give precise arguments for each identity in the derivation of the Markov property for the Poisson process.

The first identity follows from the definition of conditional probabilities (the intersections of events are the same). For the second identity the conditioning events are identical. Independence of increments yields the third identity. The fourth identity follows as identities 1 to 3 by omitting the conditions involving N_{t_j}, $j = 1, \ldots, n$.

(b) Show that a discrete-time process $(S_n)_{n \geqslant 0}$ starting in $S_0 = 0$ has stationary and independent increments if and only if it can be written as $S_n = \sum_{i=1}^{n} X_i$ with i.i.d. random variables X_i. What about processes with independent increments only?

If $S_n = \sum_{i=1}^n X_i$ with i.i.d. random variables X_i holds, then $S_{t_i - t_{i-1}} = \sum_{j=t_{i-1}+1}^{t_i} X_j$ for integers $0 = t_0 < t_1 < \cdots < t_n$ are independent as measurable functions of independent (partitioned) X_j (proposition in Stochastik I). The increments are also stationary because $\sum_{j=t_{i-1}+1}^{t_i} X_j$ is distributed as $\sum_{j=+1}^{t_i - t_{i-1}} X_j$ for i.i.d. X_j. Conversely, define $X_i := S_i - S_{i-1}$ for a process (S_n) with independent, stationary increments and $S_0 = 0$. Then in particular $(S_1 - S_0, \ldots, S_n - S_{n-1}) = (X_1, \ldots, X_n)$ are independent and equal in law, hence i.i.d. By Stochastik I this means that the sequence (X_i) is i.i.d. because n is arbitrary.

(c) Let $(X_t)_{t \geqslant 0}$ be a time-continuous homogeneous Markov chain where the transition matrices are differentiable in time with *generator* $A := P'(0)$. Derive $P'(t) = AP(t) = P(t)A$ and $P(t) = \exp(At)$ (matrix exponential). What is A for the Poisson process?

By the Chapman-Kolmogorov equation we have $P(t+h) - P(t) = (P(h) - P(0))P(t) = P(t)(P(h) - P(0))$ so that taking derivatives $P'(t) = AP(t) = P(t)A$ follows. For the matrix exponential $\exp(At) = \sum_{m=0}^\infty \frac{t^m}{m!} A^m$ we also obtain $(\exp(At))' = A\exp(At) = \exp(At)A$. Since $P(0) = \exp(A0) = I$ (identity matrix) holds and the solution to the matrix differential equation is unique (proven as in the scalar case, at least for finite S), this shows $P(t) = \exp(At)$. For the Poisson process we have $P(t)_{i,j} = p_{j-i}(t)\mathbf{1}(j \geqslant i)$ with $p_n(t) = \mathbb{P}(N_t = n)$ as above. By properties (i), (ii) of a Poisson process we infer $p_0'(t) = -\lambda$, $p_1'(t) = \lambda$ and $p_n'(t) = 0$ for $n \geqslant 2$. The Poisson process generator is therefore given by $A_{i,i} = -\lambda$, $A_{i,i+1} = \lambda$ and all other entries $A_{i,j} = 0$, $i, j \in \mathbb{N}_0$.

2 General theory of stochastic processes

2.1 Basic notions

2.1 Definition. A family $X = (X_t,\, t \in T)$ of random variables on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is called stochastic process. We call X time-discrete if $T = \mathbb{N}_0$ and time-continuous if $T = \mathbb{R}_0^+ = [0, \infty)$. If all X_t take values in $(S, \mathscr{S})$, then $(S, \mathscr{S})$ is the state space (Zustandsraum) of X. For each fixed $\omega \in \Omega$ the mapping $t \mapsto X_t(\omega)$ is called sample path (Pfad), trajectory (Trajektorie) or Realisation (Realisierung) of X.

2.2 Lemma. *For a stochastic process $(X_t,\, t \in T)$ with state space $(S, \mathscr{S})$ the mapping $\bar{X} : \Omega \to S^T$ with $\bar{X}(\omega)(t) := X_t(\omega)$ is a $(S^T, \mathscr{S}^{\otimes T})$-valued random variable.*

2.3 Remark. Later on, we shall also consider smaller function spaces than S^T, e.g. $C(\mathbb{R}^+)$ instead of $\mathbb{R}^{\mathbb{R}^+}$. ▶Exercise

Proof. We have to show measurability. Since $\mathscr{S}^{\otimes T}$ is generated by the projections $\pi_t : S^T \to S$, $t \in T$, onto the t-th coordinate, $\bar{X}$ is measurable if all compositions $\pi_t \circ \bar{X} : \Omega \to S$ are measurable, but by definition $\pi_t \circ \bar{X} = X_t$, $t \in T$, are measurable as random variables. □

2.4 Definition. Given a stochastic process $(X_t,\, t \in T)$, the laws of the random vectors $(X_{t_1}, \dots, X_{t_n})$ with $n \geqslant 1$, $t_1, \dots, t_n \in T$ are called finite-dimensional distributions of X. We write $P^X_{t_1,\dots,t_n} := \mathbb{P}^{(X_{t_1},\dots,X_{t_n})}$.

2.5 Definition. Two processes $(X_t,\, t \in T)$, $(Y_t,\, t \in T)$ on $(\Omega, \mathscr{F}, \mathbb{P})$ are called

(a) indistinguishable (ununterscheidbar) if $\mathbb{P}(\forall\, t \in T:\; X_t = Y_t) = 1$;

(b) versions or modifications (Versionen, Modifikationen) of each other if we have $\forall\, t \in T:\; \mathbb{P}(X_t = Y_t) = 1$.

2.6 Remarks.

(a) Obviously, indistinguishable processes are versions of each other. The converse is in general false. For instance, defining counting processes via $N_t = \sum_{k=1}^{\infty} \mathbf{1}(S_k < t)$ yields left-continuous sample paths and the left- and right-continuous Poisson processes are versions of each other,▶Controlbut clearly distinguishable.

(b) If X is a version of Y, then X and Y share the same finite-dimensional distributions because countable intersections of sets of measure one have measure one and thus $\mathbb{P}(X_{t_1} = Y_{t_1}, \dots, X_{t_n} = Y_{t_n}) = 1$. Processes with the same finite-dimensional distributions need not even be defined on the same probability space and will in general not be versions of each other.

(c) If T is countable, then a version Y of X is also indistinguishable from X because countable intersections of 1-sets are 1-sets. Suppose $(X_t, t \geqslant 0)$ and $(Y_t, t \geqslant 0)$ are real-valued stochastic processes with right-continuous sample paths. Then they are indistinguishable already if they are versions of each other. ▶Exercise

2.7 Definition. A process $(X_t, t \geqslant 0)$ is called continuous if all sample paths are continuous. It is called stochastically continuous, if $t_n \to t$ always implies $X_{t_n} \xrightarrow{\mathbb{P}} X_t$ (convergence in probability).

2.8 Remark. Every continuous process is stochastically continuous since almost sure convergence implies stochastic convergence. On the other hand, the Poisson process is stochastically continuous, but obviously not continuous:

$$\forall \varepsilon \in (0,1):\; \lim_{t_n \to t} \mathbb{P}(|N_t - N_{t_n}| > \varepsilon) = \lim_{t_n \to t} (1 - e^{-\lambda |t - t_n|}) = 0.$$

Note that for stochastically continuous processes the finite-dimensional distributions vary continuously in time with respect to convergence in distribution.

2.9 Proposition. *Let $C([0,\infty))$ be equipped with the topology of uniform convergence on compacts using the metric $d(f,g) := \sum_{k \geqslant 1} 2^{-k} (\sup_{t \in [0,k]} |f(t) - g(t)| \wedge 1)$. Then:*

(a) $(C([0,\infty)), d)$ is a complete and separable metric space.

(b) The Borel σ-algebra is the smallest σ-algebra such that all coordinate projections $\pi_t : C([0,\infty)) \to \mathbb{R}$, $t \geqslant 0$, are measurable.

(c) *For any continuous stochastic process $(X_t,\, t \geqslant 0)$ on $(\Omega, \mathscr{F}, \mathbb{P})$ the mapping $\bar{X} : \Omega \to C([0,\infty))$ with $\bar{X}(\omega)_t := X_t(\omega)$ is Borel-measurable.*

(d) *The law of $\bar{X}$ is uniquely determined by the finite-dimensional distributions of X.*

Proof. ▶Exercise □

2.2 Polish spaces and Kolmogorov's consistency theorem

2.10 Definition. A metric space (S, d) is called Polish space if it is separable and complete. More generally, a separable topological space which is metrizable with a complete metric is called Polish. Canonically, it is equipped with its Borel σ-algebra $\mathfrak{B}_S$, generated by the open sets.

2.11 Example. In analysis it is shown that $\mathbb{R}^d$ with any norm, $(C([a,b];\mathbb{R}), \|\bullet\|_\infty)$, $\ell^p(\mathbb{N})$ and $L^p(\mathbb{R})$, $L^p([a,b])$ for $p \in [1,\infty)$ are separable Banach (complete normed) spaces and thus Polish. The rational numbers $\mathbb{Q}$ with Euclidean distance are not complete, the spaces ℓ^∞, $L^\infty([a,b])$ are examples of non-separable Banach spaces.

The Euclidean distance d on the set $(-1,1)$ is not complete (the Cauchy sequence $(1-1/n)_{n\geqslant 1}$ does not converge), but it generates a Polish topology: convergence is the same as under the metric(!) $\tilde{d}(x,y) = |\tan(x\pi/2) - \tan(y\pi/2)|$ and this metric is complete (it maps $(-1,1)$ homeomorphically to $\mathbb{R}$). Since Borel σ-algebras are generated by the open sets, they are the same under the metrics d and $\tilde{d}$. One can show (see e.g. Bauer, Measure and Integration Theory) that all closed and all open subsets of a Polish space are Polish.

2.12 Definition. For finitely or countably many metric spaces (S_k, d_k) the product space $\prod_k S_k$ is canonically equipped with the product metric $d((s_k),(t_k)) := \sum_k 2^{-k}(d_k(s_k,t_k) \wedge 1)$, which generates the product topology, in which a vector/sequence converges iff all coordinates converge.

2.13 Lemma. *Let (S_k, d_k) be separable (resp. complete) metric spaces, then the finite or countably infinite product space $\prod_k S_k$ with the product metric is again a separable (resp. complete) metric space.*

Proof. Separability: Choose for each k a countable dense set $D_k \subseteq S_k$. Then $\prod_k D_k$ is countable and dense in $\prod_k S_k$ because $d_k(s_k,t_k) < \varepsilon$ for all k implies $d(s,t) < \varepsilon$.

Completeness: For a Cauchy sequence $(s^{(n)})_{n\geqslant 1}$ in $(\prod_k S_k, d)$ each coordinate sequence $(s_k^{(n)})_{n\geqslant 1}$ is Cauchy in (S_k, d_k) by $d_k(s_k,t_k) \leqslant 2^k d(s,t)$. So by completeness of S_k, for each k there is $s_k \in S_k$ with $s_k^{(n)} \to s_k$. This implies $s^{(n)} \to s$ in $(\prod_k S_k, d)$ and thus completeness of $(\prod_k S_k, d)$. □

2.14 Lemma. *Let (S_k, d_k), $k \geqslant 1$, be separable metric spaces, then the Borel σ-algebra of the product satisfies $\mathfrak{B}_{\prod_{k=1}^K S_k} = \bigotimes_{k=1}^K \mathfrak{B}_{S_k}$, $\mathfrak{B}_{\prod_{k\geqslant 1} S_k} = \bigotimes_{k\geqslant 1} \mathfrak{B}_{S_k}$.*

Proof. Put $S = \prod_{k\geqslant 1} S_k$ and assume without loss of generality that infinitely many S_k are given (for finitely many $S_1, \dots, S_K$ the proof is even easier). Then $\bigotimes_{k\geqslant 1} \mathfrak{B}_{S_k}$ is the smallest σ-algebra such that the coordinate projections $\pi_i : S \to S_i$, $i \geqslant 1$, are measurable. Analogously, the product topology on S is the coarsest topology such that all π_i are continuous. Consequently, each π_i is in particular $\mathfrak{B}_S$-measurable, which implies $\mathfrak{B}_S \supseteq \bigotimes_{k\geqslant 1} \mathfrak{B}_{S_k}$.

S is separable by Lemma 2.13, and any open set $O \subseteq S$ is a countable union of open balls in S. Any such ball $B_r(s) = \{t \in S \mid \sum_k 2^{-k}(d_k(s_k, t_k) \wedge 1) < r\}$ can be represented as a countable union of cylinder sets (sets of the form $(\pi_1, \dots, \pi_K)^{-1}(B_K)$ with $B_K \in \bigotimes_{k=1}^K \mathfrak{B}_{S_k}$):

$$B_r(s) = \bigcup_{K\in\mathbb{N}} \Big\{ t \in S \,\Big|\, \sum_{k=1}^{K} 2^{-k}(d_k(s_k, t_k) \wedge 1) < r - 2^{-K} \Big\}.$$

By definition, cylinder sets lie in $\bigotimes_{k\geqslant 1} \mathfrak{B}_{S_k}$ (they even form a generator) and thus $O \in \bigotimes_{k\geqslant 1} \mathfrak{B}_{S_k}$ holds, proving $\mathfrak{B}_S \subseteq \bigotimes_{k\geqslant 1} \mathfrak{B}_{S_k}$. □

2.15 Remark. The $\supseteq$-relation holds for all topological spaces and products of any cardinality with the same proof. The $\subseteq$-property can already fail for the product of two non-separable spaces, see e.g. Elstrodt, Maß- und Integrationstheorie, § III.5.3.

2.16 Definition. A probability measure $\mathbb{P}$ on a metric space $(S, \mathfrak{B}_S)$ is called

(a) <u>tight</u> (straff) if $\forall \varepsilon > 0\, \exists K \subseteq S$ compact : $\mathbb{P}(K) \geqslant 1 - \varepsilon$,

(b) <u>regular</u> (regulär) if $\forall \varepsilon > 0$, $B \in \mathfrak{B}_S\, \exists K \subseteq B$ compact : $\mathbb{P}(B \setminus K) \leqslant \varepsilon$ and $\forall \varepsilon > 0$, $B \in \mathfrak{B}_S\, \exists O \supseteq B$ open : $\mathbb{P}(O \setminus B) \leqslant \varepsilon$.

2.17 Remark. Regularity and tightness have appeared already in Stochastik I for $S = \mathbb{R}$, either explicitly or implicitly during the construction of Lebesgue and Lebesgue-Stieltjes measures. We see that tight probability measures $\mathbb{P}$ can be approximated by their values inside compact sets via $\mathbb{P}(B) = \sup_{K \text{ compact}} \mathbb{P}(B \cap K)$, which will allow compactness arguments. Similarly, regular $\mathbb{P}$ are determined by their values on compact and open sets:

$$\mathbb{P}(B) = \sup_{K \subseteq B \text{ compact}} \mathbb{P}(K) \text{ and } \mathbb{P}(B) = \inf_{O \supseteq B \text{ open}} \mathbb{P}(O).$$

Verify this for Lebesgue measure on $[0, 1]$ and $B = \mathbb{Q} \cap [0, 1]$►Control.

2.18 Proposition. *Every probability measure on a Polish space is tight.*

Proof. Let $(s_n)_{n\geqslant 1}$ be a dense sequence in S and consider for any radius $\rho > 0$ the closed balls $\bar{B}_\rho(s_n)$ around s_n. Then $S = \bigcup_{n\geqslant 1} \bar{B}_\rho(s_n)$ and σ-continuity imply

$$\lim_{N\to\infty} \mathbb{P}\Big(\bigcup_{n=1}^{N} \bar{B}_\rho(s_n) \Big) = 1.$$

Now select for $\varepsilon > 0$ and every $\rho = 1/k$ an index N_k such that

$$\mathbb{P}\Big(\bigcup_{n=1}^{N_k} \bar{B}_{1/k}(s_n)\Big) \geqslant 1 - \varepsilon 2^{-k}.$$

Then $K := \bigcap_{k=1}^{\infty} \bigcup_{n=1}^{N_k} \bar{B}_{1/k}(s_n)$ is a closed subset, hence complete. Since for any $\delta > 0$ there is a finite cover of K by balls $\bar{B}_{1/k}(s_n)$ of diameter less than δ (K is *totally bounded*), any sequence in K has a subsequence which is Cauchy. By completeness, the Cauchy sequence converges and K is compact. By construction,

$$\mathbb{P}(S \setminus K) = \mathbb{P}\Big(\bigcup_{k=1}^{\infty} \bigcap_{n=1}^{N_k} \bar{B}_{1/k}(s_n)^C\Big) \leqslant \sum_{k=1}^{\infty} \varepsilon 2^{-k} = \varepsilon$$

holds. This shows tightness. □

▷ **Control questions**

(a) Show that the left- and right-continuous definitions of a Poisson process yield indeed versions of each other.

Let $(N_t, t \geqslant 0)$ denote the usual right-continuous Poisson process and $(\tilde{N}_t, t \geqslant 0)$ the left-continuous Poisson process. Then $\mathbb{P}(N_t \neq \tilde{N}_t)$ equals the probability $\mathbb{P}(\exists k : S_k = t)$ that there is a jump at time t. Since $S_k = T_1 + \cdots + T_k$ holds with independent $\operatorname{Exp}(\lambda)$-distributed T_i, the law of S_k is continuous (a Gamma-distribution) and $\mathbb{P}(S_k = t) = 0$ holds. The union of nullsets is a nullset so that $\mathbb{P}(\exists k : S_k = t) = 0$ follows.

(b) Why is a continuous process always stochastically continuous?

This follows directly from the fact that (almost) sure convergence implies convergence in probability.

(c) For Lebesgue measure λ find compact sets $K_n \subseteq \mathbb{Q} \cap [0,1]$ with $\lambda(K_n) \to \lambda(\mathbb{Q} \cap [0,1])$ and open sets $O_n \supseteq \mathbb{Q} \cap [0,1]$ with $\lambda(O_n) \to \lambda(\mathbb{Q} \cap [0,1])$.

Since every atom $\{x\}$, $x \in [0,1]$, is a λ-null set, so is every countable subset of $[0,1]$. Hence, $\lambda(\mathbb{Q}) = 0$ and $\lambda(K) = 0$ for any subset $K \subseteq \mathbb{Q}$, in particular any compact subset K. If we write $\mathbb{Q} \cap [0,1] = \{q_n \,|\, n \geqslant 1\}$ for an enumeration of the elements q_n of $\mathbb{Q} \cap [0,1]$, the sets $O_\varepsilon := \bigcup_{n \geqslant 1}(q_n - \varepsilon 2^{-n}, q_n + \varepsilon 2^{-n}) \cap [0,1]$ are open in $[0,1]$ (union of open sets). Their Lebesgue measure satisfies $\lambda(O_\varepsilon) \leqslant \sum_{n \geqslant 1} 2\varepsilon 2^{-n} = 2\varepsilon$. We thus have $O_{1/m} \supseteq \mathbb{Q} \cap [0,1]$ and $\lambda(O_{1/m}) \to 0 = \lambda(\mathbb{Q} \cap [0,1])$ as $m \to \infty$.

2.19 Theorem (Ulam, 1939). *Every probability measure on a Polish space (S,d) is regular.*

Proof. We consider the family of Borel sets

$$\mathcal{D} := \Big\{ B \in \mathfrak{B}_S \,\Big|\, \mathbb{P}(B) = \sup_{K \subseteq B \text{ compact}} \mathbb{P}(K) = \inf_{O \supseteq B \text{ open}} \mathbb{P}(O) \Big\}.$$

Consider any closed set $F \subseteq S$. By tightness, for any $\varepsilon > 0$ there is a compact set K_ε with $\mathbb{P}(K_\varepsilon) \geqslant 1 - \varepsilon$. Then $F \cap K_\varepsilon \subseteq F$ is compact with

$$\mathbb{P}(F \setminus (F \cap K_\varepsilon)) \leqslant \mathbb{P}(K_\varepsilon^C) \leqslant \varepsilon.$$

This shows $\mathbb{P}(F) = \sup_K \mathbb{P}(K)$ with $K \subseteq F$ compact. The open sets $O_n := \{s \in S \,|\, \inf_{x \in F} d(s,x) < 1/n\}$ satisfy $F = \bigcap_{n \geqslant 1} O_n$. By σ-continuity, we infer $\mathbb{P}(F) = \inf_{N \geqslant 1} \mathbb{P}(\bigcap_{n=1}^N O_n)$. Since finite intersections of open sets are open, we have shown the second regularity property and thus $F \in \mathcal{D}$.

Furthermore▶Exercise, $\mathcal{D}$ is closed under set differences and countable unions. Since $S \in \mathcal{D}$ as a closed set, $\mathcal{D}$ is a σ-algebra containing the open sets, which implies $\mathcal{D} = \mathfrak{B}_S$, as asserted. □

2.20 Lemma. *Let $(X_t,\, t \in T)$ be a stochastic process with state space $(S, \mathscr{S})$ and denote by $\pi_{J \to I} : S^J \to S^I$ for $I \subseteq J \subseteq T$ the coordinate projection $\pi_{J \to I}((s_j)_{j \in J}) = (s_j)_{j \in I}$. Then the finite-dimensional distributions $(P_J^X)_{J \subseteq T \text{ finite}}$ satisfy the following consistency condition:*

$$\forall\, I \subseteq J \subseteq T \text{ with } I, J \text{ finite } \forall A \in \mathscr{S}^{\otimes I} :\ P_J^X(\pi_{J \to I}^{-1}(A)) = P_I^X(A). \tag{2.1}$$

Proof. We just write with $\bar{X}$ from Lemma 2.2:

$$\begin{aligned} P_I^X(A) &= \mathbb{P}((X_t)_{t \in I} \in A) = \mathbb{P}(\bar{X} \in \pi_{T \to I}^{-1}(A)) \\ &= \mathbb{P}(\bar{X} \in (\pi_{J \to I} \circ \pi_{T \to J})^{-1}(A)) = \mathbb{P}((X_t)_{t \in J} \in \pi_{J \to I}^{-1}(A)) \\ &= P_J^X(\pi_{J \to I}^{-1}(A)). \end{aligned}$$

□

2.21 Definition. Let $T \neq \varnothing$ be an index set and $(S, \mathscr{S})$ be a measurable space. Let for each finite subset $J \subseteq T$ a probability measure $\mathbb{P}_J$ on the product space $(S^J, \mathscr{S}^{\otimes J})$ be given. Then $(\mathbb{P}_J)_{J \subseteq T \text{ finite}}$ is called projective family if the following consistency condition is satisfied:

$$\forall I \subseteq J \subseteq T \text{ finite}, A \in \mathscr{S}^{\otimes I} :\ \mathbb{P}_I(A) = \mathbb{P}_J(\pi_{J \to I}^{-1}(A)).$$

2.22 Remark. The preceding lemma shows that the consistency condition is a necessary condition for the finite-dimensional distributions of a stochastic process. The main message of the next theorem is that it is also sufficient for the construction of a stochastic process on the product space of a Polish state space with prescribed finite-dimensional distributions.

2.23 Theorem (Kolmogorov's consistency/extension theorem; Daniell 1919, Kolmogorov 1933)**.** *Let $(S, \mathfrak{B}_S)$ be a Polish space, $T \neq \varnothing$ an index set and let $(\mathbb{P}_J)$ be a projective family of probability measures for S and T. Then there exists a unique probability measure $\mathbb{P}$ on the product space $(S^T, \mathfrak{B}_S^{\otimes T})$ satisfying*

$$\forall J \subseteq T \text{ finite}, B \in \mathfrak{B}_S^{\otimes J} :\ \mathbb{P}_J(B) = \mathbb{P}(\pi_{T \to J}^{-1}(B)).$$

Proof. Let $\mathfrak{A} := \bigcup_{J\subseteq T \text{ finite}} \pi_{T\to J}^{-1}(\mathfrak{B}_S^{\otimes J})$ be the algebra (check!) of cylinder sets on S^T, which generates $\mathfrak{B}_S^{\otimes T}$. Since $\mathfrak{A}$ is $\cap$-stable, $\mathbb{P}$ is uniquely determined by its values on the cylinder sets.

The existence of $\mathbb{P}$ follows from Caratheodory's extension theorem if $\mathbb{P}$ on $\mathfrak{A}$, as defined in the theorem, is a pre-measure (has all measure properties, but σ-additivity only if the countable union is again in $\mathfrak{A}$). The consistency of $(\mathbb{P}_J)$ ensures that $\mathbb{P}$ is well-defined on $\mathfrak{A}$ and additive: for disjoint $A, B \in \mathfrak{A}$ there are a finite $J \subseteq T$ and disjoint $A', B' \in \mathfrak{B}_S^{\otimes J}$ with $A = \pi_{T\to J}^{-1}(A')$, $B = \pi_{T\to J}^{-1}(B')$. Since $\mathbb{P}_J$ is a probability measure and standard set operations commute with taking preimages, we conclude

$$\mathbb{P}(A\cup B) = \mathbb{P}_J(A'\cup B') = \mathbb{P}_J(A') + \mathbb{P}_J(B') = \mathbb{P}(A) + \mathbb{P}(B).$$

Trivially, also $\mathbb{P}(S^T) = \mathbb{P}_J(S^J) = 1$ holds, using any finite $J \subseteq T$. It remains to show that $\mathbb{P}$ is σ-additive on $\mathfrak{A}$, which is (under finite additivity) equivalent to $\mathbb{P}(B_n) \to 0$ for any sequence $B_n \downarrow \varnothing$ of sets $B_n \in \mathfrak{A}$ (i.e. $B_{n+1} \subseteq B_n$ with $\bigcap_{n\geqslant 1} B_n = \varnothing$; σ-continuity at $\varnothing$).

We can write $B_n = \pi_{T\to J_n}^{-1}(A_n)$ for some finite $J_n \subseteq T$, $A_n \in \mathfrak{B}_S^{\otimes J_n}$. Without loss of generality we shall assume $J_n \subseteq J_{n+1}$ for all n. Now let $K_n \subseteq A_n$ be compact with $\mathbb{P}_{J_n}(A_n \setminus K_n) \leqslant \varepsilon 2^{-n}$ by Ulam's Theorem. Then $K_n' = \bigcap_{l=1}^{n-1} \pi_{J_n\to J_l}^{-1}(K_l) \cap K_n$ is compact in S^{J_n} as a closed subset of a compact set and $C_n = \pi_{T\to J_n}^{-1}(K_n') = \bigcap_{l=1}^{n} \pi_{T\to J_l}^{-1}(K_l) \subseteq B_n$ satisfies also $C_n \downarrow \varnothing$. Below we prove that there is already an $n_0 \in \mathbb{N}$ with $C_{n_0} = \varnothing$. From this and the decay of $\mathbb{P}(B_n)$ we conclude

$$\lim_{n\to\infty} \mathbb{P}(B_n) \leqslant \mathbb{P}(B_{n_0}) = \mathbb{P}(B_{n_0} \setminus C_{n_0}) \leqslant \sum_{l=1}^{n_0} \mathbb{P}_{J_l}(A_l \setminus K_l) \leqslant \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows $\mathbb{P}(B_n) \to 0$, as desired.

We prove the claim $\exists n_0 : C_{n_0} = \varnothing$ via reductio ad absurdum and a compactness argument, assuming that for all $n \in \mathbb{N}$ there is a $y_n \in C_n$. Since K_n' is compact in S^{J_n}, we can find a subsequence $(n_l^{(1)})$, such that $(\pi_{T\to J_1}(y_{n_l^{(1)}}))_{l\geqslant 1}$ converges in K_1', a further subsequence $(n_l^{(2)})$ such that $(\pi_{T\to J_2}(y_{n_l^{(2)}}))_{l\geqslant 1}$ converges in K_2' and so on. Along the diagonal sequence $(n_l^{(l)})_{l\geqslant 1}$ then $(\pi_{T\to J_m}(y_{n_l^{(l)}}))_{l\geqslant 1}$ converges in K_m' for all $m \geqslant 1$. Hence, $(\pi_{T\to\cup_{m\geqslant 1} J_m}(y_{n_l^{(l)}}))_{l\geqslant 1}$ converges in the product topology (metric) to some $z \in S^{\cup_{m\geqslant 1} J_m}$ (note: $\bigcup_{m\geqslant 1} J_m$ is countable). As $C_{n+1} \subseteq C_n$, $n \geqslant 1$, are nested, this implies $z \in \pi_{T\to\cup_{m\geqslant 1} J_m}(C_n)$ for all $n \geqslant 1$ and thus $z \in \pi_{T\to\cup_{m\geqslant 1} J_m}(\bigcap_{n\geqslant 1} C_n)$. This contradicts $\bigcap_{n\geqslant 1} C_n = \varnothing$ and the claim is proved. □

2.24 Corollary. *For any Polish space $(S, \mathfrak{B}_S)$ and any index set $T \neq \varnothing$ there exists to a prescribed projective family $(\mathbb{P}_J)$, $J \subseteq T$ finite, a stochastic process $(X_t, t \in T)$ whose finite-dimensional distributions are given by $(\mathbb{P}_J)$.*

Proof. By Kolmogorov's consistency theorem construct the probability measure $\mathbb{P}$ on $(S^T, \mathfrak{B}_S^{\otimes T})$ which satisfies $\mathbb{P}(\pi_{T\to\{t_1,\ldots,t_n\}}^{-1}(A)) = \mathbb{P}_{\{t_1,\ldots,t_n\}}(A)$ for all $n \in \mathbb{N}$,

$t_1, \dots, t_n \in T$, $A \in \mathfrak{B}_S^{\otimes n}$. Define X to be the coordinate process on $(S^T, \mathfrak{B}_S^{\otimes T}, \mathbb{P})$ via $X_t((s_\tau)_{\tau \in T}) := s_t$. Then X_t is measurable for every $t \in T$ and

$$\mathbb{P}((X_{t_1}, \dots, X_{t_n}) \in A) = \mathbb{P}(\pi^{-1}_{T \to \{t_1, \dots, t_n\}}(A)) = \mathbb{P}_{\{t_1, \dots, t_n\}}(A)$$

for all $A \in \mathfrak{B}_S^{\otimes n}$. □

2.25 Example (Markov chains)**.** Let $(S, \mathcal{P}(S))$ be a countable state space. Let an initial distribution $\mu^{(0)}$ on $\mathcal{P}(S)$ and transition probabilities $(p_{ij})_{i,j \in S}$ be given, that is a stochastic matrix $P = (p_{ij})_{i,j \in S}$ with non-negative entries and row sum equal to one. Then we want to construct a discrete time-homogeneous Markov chain $(X_n, n \geqslant 0)$ with $\mathbb{P}^{X_0} = \mu_0$ and $\mathbb{P}(X_n = j \,|\, X_{n-1} = i) = p_{ij}$ whenever $\mathbb{P}(X_{n-1} = i) > 0$. Note first that $(S, \mathcal{P}(S))$ becomes Polish if we consider the discrete metric $d(s,t) = \mathbf{1}(s \neq t)$ such that $\mathfrak{B}_S = \mathcal{P}(S)$►Control. Consider

$$\mu_n(A) = \sum_{i_0 \in S} \cdots \sum_{i_n \in S} \mathbf{1}_A(i_0, \dots, i_n) \mu^{(0)}_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}, \quad A \subseteq S^{n+1}.$$

Then, letting $T := \mathbb{N}_0$ and $\mu_{t_1, \dots, t_n}(B) := \mu_{t_n}(\pi^{-1}_{\{0,1,\dots,t_n\} \to \{t_1, \dots, t_n\}}(B))$ we see by induction that $(\mu_J)_{J \subseteq T}$ is a projective family iff $\mu_{n+1}(\pi^{-1}_{\{0,\dots,n+1\} \to \{0,\dots,n\}}(A)) = \mu_n(A)$ holds for all $n \geqslant 0$, $A \subseteq S^{n+1}$. The latter is easily verified ►Control and by the consistency theorem such a Markov chain always exists.

▷ **Control questions**

(a) Simplify, wherever possible, the proof of Kolmogorov's extension theorem in the case $T = \{0, 1 \dots, N\}$ finite (think of a Markov chain until time N).

This is trivial, just take $\mathbb{P} = \mathbb{P}_T$ which is prescribed because $T \subseteq T$ is finite. The defining property of $\mathbb{P}$ is given exactly by the consistency condition of the projective family $(\mathbb{P}_J)$. The construction of the finite-dimensional distributions is assumed by Kolmogorov's consistency theorem and must be done for each special case, compare the Markov chain and Gaussian process examples.

(b) Check that the discrete metric on a countable set S generates a Polish space with $\mathfrak{B}_S = \mathcal{P}(S)$.

For the discrete metric $d(s,t) = \mathbf{1}(s \neq t)$ the sets $\{t \in S \,|\, d(t,s) < 1/2\} = \{s\}$, $s \in S$, are open. This shows $\mathfrak{B}_S = \mathcal{P}(S)$ for countable S. Furthermore, every countable metric space is separable. Finally, a Cauchy sequence with respect to the discrete topology must be eventually constant (consider the Cauchy criterion with $\varepsilon = 1/2$) and thus converges. This proves that (S, d) is Polish.

(c) Prove in detail that $(\mu_J)_{J \subseteq T}$ is a projective family in the preceding example.

Let $J \subseteq J' \subseteq \mathbb{N}_0$. Then by definition, setting $|J|_\infty := \max_{t\in J} t$, $\mathbb{N}_J := \{0,1,\dots,|J|_\infty\}$, $\pi_{m\to n} := \pi_{\{0,\dots,m\}\to\{0,\dots,n\}}$

$$\begin{aligned}
\mu_J(B) &= \mu_{|J|_\infty}(\pi^{-1}_{\mathbb{N}_J\to J}(B))\\
&= \sum_{i_0\in S}\cdots\sum_{i_{|J|_\infty}\in S} \mathbf{1}_B\big(\pi_{\mathbb{N}_J\to J}(i_0,\dots,i_{|J|_\infty})\big)\mu^{(0)}_{i_0} p_{i_0,i_1}\cdots p_{i_{|J|_\infty-1},i_{|J|_\infty}}\\
&= \sum_{i_0\in S}\cdots\sum_{i_{|J'|_\infty}\in S} \mathbf{1}_B\big(\pi_{\mathbb{N}_{J'}\to J}(i_0,\dots,i_{|J'|_\infty})\big)\mu^{(0)}_{i_0} p_{i_0,i_1}\cdots p_{i_{|J'|_\infty-1},i_{|J'|_\infty}}\\
&= \sum_{i_0\in S}\cdots\sum_{i_{|J'|_\infty}\in S} \mathbf{1}_B\big(\pi_{J'\to J}(\pi_{\mathbb{N}_{J'}\to J'}(i_0,\dots,i_{|J'|_\infty})\big)\mu^{(0)}_{i_0} p_{i_0,i_1}\cdots p_{i_{|J|_\infty-1},i_{|J'|_\infty}}\\
&= \mu_{|J'|_\infty}(\pi^{-1}_{\mathbb{N}_{J'}\to J'}(\pi^{-1}_{J'\to J}(B)))\\
&= \mu_{J'}(\pi^{-1}_{J'\to J}(B)),
\end{aligned}$$

where the second identity follows from the fact that $\sum_{j\in S} p_{i,j} = 1$ holds for transition probabilities $p_{i,j}$. This gives the consistency condition.

2.26 Example (Gaussian processes)**.** Let $T \neq \varnothing$ be an index set, e.g. $\mathbb{N}_0$, $\mathbb{Z}$, $\mathbb{R}^+$ or $\mathbb{R}$, and $\mu : T \to \mathbb{R}$ be any function, $c : T^2 \to \mathbb{R}$ any symmetric, positive semi-definite function, that is $c(t,s) = c(s,t)$ and $\sum_{i,j=1}^n c(t_i,t_j)\alpha_i\alpha_j \geqslant 0$ for all $t,s \in T$, $n\in\mathbb{N}$, $t_1,\dots,t_n \in T$, $\alpha_1,\dots,\alpha_n\in\mathbb{R}$. Then there exists a process $(X_t, t\in T)$ whose finite-dimensional distributions are Gaussian:

$$P^X_{t_1,\dots,t_n} = N((\mu(t_1),\dots\mu(t_n))^\top, (c(t_i,t_j))_{i,j=1,\dots,n}).$$

Here, we understand $N(\mu,\Sigma)$ as the law of $Y = \mu + \Sigma^{1/2}Z$ for a standard-normal random vector Z or equivalently via the characteristic function $\varphi_{N(\mu,\Sigma)}(u) = \exp(i\langle u,\mu\rangle - \langle \Sigma u, u\rangle/2)$, $u\in\mathbb{R}^n$. The consistency condition follows from $AY \sim N(A\mu, A\Sigma A^\top)$ for any linear transformation (in particular, coordinate projection) $A : R^n \to \mathbb{R}^m$: for $m \leqslant n$, $B \in \mathfrak{B}_{\mathbb{R}^m}$, $\mu_n = (\mu(t_1),\dots\mu(t_n))^\top$, $C_n = (c(t_i,t_j))_{i,j=1,\dots,n}$, $Z_n \sim N(0,E_n)$ etc. we have

$$\begin{aligned}
&P^X_{t_1,\dots,t_n}(\pi^{-1}_{\{t_1,\dots,t_n\}\to\{t_1,\dots,t_m\}}(B)) = P(\mu_n + C_n^{1/2}Z_n \in \pi^{-1}_{\{t_1,\dots,t_n\}\to\{t_1,\dots,t_m\}}(B))\\
&= P(\pi_{\{t_1,\dots,t_n\}\to\{t_1,\dots,t_m\}}(\mu_n) + \pi_{\{t_1,\dots,t_n\}\to\{t_1,\dots,t_m\}}(C_n^{1/2}Z_n)\in B)\\
&= P(Y\in B) \text{ for } Y \sim N(\mu_m, \pi_{\{t_1,\dots,t_n\}\to\{t_1,\dots,t_m\}}C_n\pi^\top_{\{t_1,\dots,t_n\}\to\{t_1,\dots,t_m\}}).
\end{aligned}$$

The covariance matrix of Y just evaluates to C_m and so the last line equals $P^X_{t_1,\dots,t_m}(B)$. This Gaussian process thus exists by the consistency theorem. μ is called expectation function and c covariance function of X.

2.27 Corollary. *For any family $(\mathbb{P}_i)_{i\in I}$ of probability measures on $(S,\mathscr{S})$ there exists the product measure $\bigotimes_{i\in I}\mathbb{P}_i$ on $(S^I,\mathscr{S}^{\otimes I})$. In particular, a family $(X_i)_{i\in I}$ of independent random variables with prescribed laws $\mathbb{P}^{X_i}$ exists.*

Proof for $(S,\mathscr{S})$ Polish: for finite product measures the consistency condition holds because for all $B\in\mathfrak{B}_S^{\otimes J}$, $J\subseteq J'$:

$$\Big(\bigotimes_{j\in J'}\mathbb{P}_j\Big)(\pi^{-1}_{J'\to J}(B)) = \Big(\bigotimes_{j\in J}\mathbb{P}_j\Big)(B)\bullet\Big(\bigotimes_{j\in J'\setminus J}\mathbb{P}_j\Big)(S^{J'\setminus J}) = \Big(\bigotimes_{j\in J}\mathbb{P}_j\Big)(B).$$

Define $X_i : S^I \to S$ by $X_i((s_j)_{j\in I}) := s_i$. Then the assertion follows from the preceding corollary. For general measure spaces $(S, \mathscr{S})$ the proof is similar to that of Kolmogorov's consistency theorem, see e.g. Bauer, Probability Theory. □

2.28 Remark. Kolmogorov's consistency theorem (proved by Kolmogorov for $S = \mathbb{R}$ and countable T) does not hold for general measure spaces $(S, \mathcal{S})$, cf. the counterexample by Sparre Andersen, Jessen (1948), contrary to what the famous Joseph Doob had claimed before. The Ionescu-Tulcea Theorem, however, shows the existence of the probability measure on general measure spaces under a Markovian dependence structure, see e.g. Klenke (2008). Concerning the achievements of Doob in developing the theory of stochastic processes and also the coin tossing leading to the expression 'random variable' the historical survey `https://arxiv.org/pdf/0909.4213.pdf` by R. Getoor is worthwhile reading.

3 The conditional expectation

3.1 Remark. Conditional probabilities $\mathbb{P}(A \,|\, B)$ are only well-defined if $\mathbb{P}(B) > 0$. In particular, expressions like $\mathbb{P}(A \,|\, X = x)$ are not defined for continuously distributed random variables X. Nevertheless, we shall give sense to these conditional probabilities by first defining conditional expectations. We interprete them as best L^2-approximations, which requires to understand orthogonal projections in Hilbert spaces.

3.1 Orthogonal projections

3.2 Proposition. *Let L be a closed linear subspace of the Hilbert space H. Then for each $x \in H$ there is a unique $y_x \in L$ with $\|x - y_x\| = \operatorname{dist}(x, L) := \inf_{y\in L}\|x - y\|$.*

Proof. For $x \in L$ we have $y_x = x$. Otherwise, there is a sequence $(y_n) \subseteq L$ with $\|x - y_n\| \to \operatorname{dist}(x, L)$. Let us show that (y_n) is a Cauchy sequence. Note

$$\|y_n - y_m\|^2 = 2\|x - y_m\|^2 + 2\|x - y_n\|^2 - 4\|x - (y_m + y_n)/2\|^2.$$

Since $(y_m + y_n)/2 \in L$ and $\|x - (y_m + y_n)/2\| \leqslant (\|x - y_m\| + \|x - y_n\|)/2$, we see $\lim_{m,n\to\infty}\|x - (y_m + y_n)/2\| = \operatorname{dist}(x, L)$. From the identity above we thus conclude $\lim_{m,n\to\infty}\|y_n - y_m\|^2 = 0$.

By completeness of H and closedness of L, the Cauchy sequence (y_n) has a limit $y_x \in L$. By continuity of the norm, we deduce $\|x - y_x\| = \operatorname{dist}(x, L)$. For another element $z_x \in L$ with this property we obtain from the above identity

$$\|y_x - z_x\|^2 = 4\big(\operatorname{dist}(x, L)^2 - \|x - (y_x + z_x)/2\|^2\big).$$

Since by definition $\|x - (y_x + z_x)/2\| \geqslant \operatorname{dist}(x, L)$ holds, this shows $\|y_x - z_x\| = 0$ and uniqueness follows. □

3.3 Definition. For a closed linear subspace L of the Hilbert space H the orthogonal projection $P_L : H \to L$ onto L is defined by $P_L(x) = y_x$ with y_x from the previous proposition.

3.4 Lemma. *We have:*

(a) $P_L \circ P_L = P_L$ *(projection property);*

(b) $\forall\, x \in H :\ (x - P_L x) \in L^\perp = \{y \in H \,|\, \langle y, z\rangle = 0$ *for all* $z \in L\}$ *(orthogonality).*

Proof. By definition, $x \in L \Rightarrow P_L x = x$ and $P_L y \in L$ such that $P_L(P_L y) = P_L y$ for all $y \in H$ and (a) holds. For all $x \in H$, $y \in L$ we obtain

$$\|x - P_L x\|^2 \leqslant \|x - (P_L x + y)\|^2 = \|x - P_L x\|^2 + \|y\|^2 - 2\langle x - P_L x, y\rangle.$$

For all $\alpha \in \mathbb{R} \setminus \{0\}$ we therefore find

$$2\langle x - P_L x, \alpha y\rangle \leqslant \|\alpha y\|^2 \Rightarrow 2\,\mathrm{sgn}(\alpha)\langle x - P_L x, y\rangle \leqslant |\alpha|\|y\|^2.$$

Letting $\alpha \downarrow 0$ and $\alpha \uparrow 0$, this implies $\langle x - P_L x, y\rangle = 0$ and thus (b). □

3.5 Corollary. *We have:*

(a) Each $x \in H$ *can be decomposed uniquely as* $x = P_L x + (x - P_L x)$ *in the sum of an element of* L *and an element of* $L^\perp$*;*

(b) P_L *is selfadjoint:* $\langle P_L x, y\rangle = \langle x, P_L y\rangle$*;*

(c) P_L *is linear.*

Proof. For (a) it remains to prove uniqueness. Writing $x = y + z$ with $y \in L$, $z \in L^\perp$, we deduce $y - P_L x = (x - P_L x) - z \in L \cap L^\perp = \{0\}$ and thus $y = P_L x$, $z = x - P_L x$. Properties (b) and (c) follow by using the decomposition in (a). □

3.2 Construction and properties

3.6 Definition. For a random variables X on $(\Omega, \mathscr{F}, \mathbb{P})$ with values in $(S, \mathscr{S})$ we introduce the σ-algebra (!) $\sigma(X) := \{X^{-1}(A) \,|\, A \in \mathscr{S}\} \subseteq \mathscr{F}$, which is the smallest σ-algebra on Ω for which X is measurable. For a family of random variables X_i, $i \in I$, on $(\Omega, \mathscr{F}, \mathbb{P})$we denote by $\sigma(X_i, i \in I)$ the smallest σ-algebra on Ω for which all X_i are measurable. For a given probability space $(\Omega, \mathscr{F}, \mathbb{P})$ we set

$$\begin{aligned}
\mathcal{M} := \mathcal{M}(\Omega, \mathscr{F}) &:= \{X : \Omega \to \mathbb{R} \text{ measurable}\};\\
\mathcal{M}^+ := \mathcal{M}^+(\Omega, \mathscr{F}) &:= \{X : \Omega \to [0, \infty] \text{ measurable}\};\\
\mathcal{L}^p := \mathcal{L}^p(\Omega, \mathscr{F}, \mathbb{P}) &:= \{X \in \mathcal{M}(\Omega, \mathscr{F}) \,|\, \mathbb{E}[|X|^p] < \infty\};\\
L^p := L^p(\Omega, \mathscr{F}, \mathbb{P}) &:= \{[X] \,|\, X \in \mathcal{L}^p(\Omega, \mathscr{F}, \mathbb{P})\}\\
&\qquad \text{where } [X] := \{Y \in \mathcal{M}(\Omega, \mathscr{F}) \,|\, \mathbb{P}(X = Y) = 1\}.
\end{aligned}$$

3.7 Example. Suppose X and Y are real-valued random variables on the same discrete probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ with a positive counting density p. Then $A \mapsto \mathbb{P}(A \mid X = x)$ is a probability measure for all $x \in X(\Omega)$ (Stochastik I, note $\mathbb{P}(X = x) > 0$ because $p > 0$). Naturally, we can then define the conditional expectation

$$\varphi(x) := \mathbb{E}[Y \mid X = x] = \int Y \, d\,\mathbb{P}(\bullet \mid X = x) = \frac{\sum_{\omega \in \Omega} Y(\omega)\mathbf{1}(X(\omega) = x)p(\omega)}{\sum_{\omega\in\Omega} \mathbf{1}(X(\omega) = x)p(\omega)}.$$

We claim $\mathbb{E}[(Y - \varphi(X))^2] = \min_{\tilde{\varphi}} \mathbb{E}[(Y - \tilde{\varphi}(X))^2]$ for all functions $\tilde{\varphi} : S \to \mathbb{R}$. Indeed, we have

$$\mathbb{E}[(Y - \tilde{\varphi}(X))^2] = \sum_{x\in X(\Omega)} \Big(\sum_{\omega : X(\omega)=x} (Y(\omega) - \tilde{\varphi}(x))^2 p(\omega) \Big)$$

and the right-hand side is minimal in $\tilde{\varphi}$ whenever the inner sum is minimal for any given x. The quadratic functional $F(z) := \sum_{\omega:X(\omega)=x} (Y(\omega) - z)^2 p(\omega)$ is minimal when $\sum_{\omega:X(\omega)=x}(Y(\omega) - z)p(\omega) = 0$ holds, i.e for $z = \varphi(x)$.

As a simple specific example consider $S = X_1 + X_2$ the sum of two (independent and fair) dice with numbers X_1, X_2. What is $\mathbb{E}[S \mid X_1 = k]$ for $k = 1, \dots, 6$? Evaluating the above expression for φ in this case yields $\mathbb{E}[S \mid X_1 = k] = k + 3.5$. The calculation can be simplified by inserting $S = X_1 + X_2$ and splitting the numerator, which yields $\mathbb{E}[S \mid X_1 = k] = k + \mathbb{E}[X_2]$. Later, this will be a consequence of linearity: $\mathbb{E}[X_1 + X_2 \mid X_1] = X_1 + \mathbb{E}[X_2]$ for X_1, X_2 independent.

The discrete example shows that the conditional expectation is the best prediction of Y by a function $\tilde{\varphi}$ of X with respect to mean squared error $\mathbb{E}[(Y - \tilde{\varphi}(X))^2] = \|Y - \tilde{\varphi}(X)\|_{L^2}^2$. It is this property that allows to construct conditional expectations beyond the discrete setting. First, we characterise functions of the form $\omega \mapsto \tilde{\varphi}(X(\omega))$, which will allow us to condition more generally on σ-algebras instead of random variables X.

▷ **Control questions**

(a) Show that there exists a Gaussian process $(B_t, t \geqslant 0)$ with expectation function $\mathbb{E}[B_t] = 0$ (a *centred* process) and covariance function $\mathbb{E}[B_t B_s] = t \wedge s$, $t, s \geqslant 0$. B is called *Brownian motion.*
Hint: prove $\sum_{i,j=1}^n \alpha_i\alpha_j t_i \wedge t_j = \sum_{i=1}^n (\sum_{k=i}^n \alpha_k)^2 (t_i - t_{i-1})$ for $0 = t_0 < t_1 < \cdots < t_n$.

The formula follows inductively over n by looking at differences and using partial summation:

$$\sum_{i,j=1}^{n+1} \alpha_i\alpha_j t_i \wedge t_j - \sum_{i,j=1}^{n} \alpha_i\alpha_j t_i \wedge t_j = 2\sum_{i=1}^{n} \alpha_i\alpha_{n+1} t_i + \alpha_{n+1}^2 t_{n+1}$$
$$= \sum_{l=1}^{n+1} \Big(2 \sum_{i=l}^{n} \alpha_i \alpha_{n+1}(t_l - t_{l-1}) + \alpha_{n+1}^2 (t_l - t_{l-1}) \Big)$$
$$= \sum_{l=1}^{n+1} \Big(\Big(\sum_{i=l}^{n+1} \alpha_i \Big)^2 - \Big(\sum_{i=l}^{n} \alpha_i \Big)^2 \Big)(t_l - t_{l-1}).$$

The formula then shows that $(s,t) \mapsto s \wedge t$ is a positive-semidefinite (and symmetric) function. So, there exists a centred Gaussian process $(B_t, t \geqslant 0)$ with this covariance function. Later, we shall prove that we may even choose a continuous version of $(B_t, t \geqslant 0)$.

(b) What is an explicit expression for $\varphi(X)$ in the preceding example, noting that $\mathbb{E}[Y \mid X = X]$ is nonsense?

We write explicitly $\varphi(X(\omega))$ and obtain

$$\varphi(X(\omega)) = \mathbb{E}[Y \mid X = x]|_{x=X(\omega)} = \frac{\sum_{\omega' \in \Omega} Y(\omega')\mathbf{1}(X(\omega') = X(\omega))p(\omega')}{\sum_{\omega' \in \Omega} \mathbf{1}(X(\omega') = X(\omega))p(\omega')}.$$

(c) Show for general $Y \in \mathcal{L}^2$ that $\mathbb{E}[(Y-\mu)^2]$ is minimal at $\mu = \mathbb{E}[Y]$. Can you extend this to minimise $\mathbb{E}[(Y - \varphi(X))^2]$ in measurable φ when X and Y are independent?

The quadratic functional $\mathbb{E}[(Y-\mu)^2] = \mathbb{E}[Y^2] - 2\mu\,\mathbb{E}[Y] + \mu^2$ is minimal at $\mu = \mathbb{E}[Y]$ by basic calculus. Alternatively, use the bias-variance decomposition from Stochastik I. For independent X, Y we have by Fubini's Theorem $\mathbb{E}[(Y - \varphi(X))^2] = \int\int (y - \varphi(x))^2\, \mathbb{P}^Y(dy)\, \mathbb{P}^X(dx)$. For each x the inner integral is (by the same argument) minimal at $\varphi(x) = \mathbb{E}[Y]$ so that the constant function $\varphi(x) = \mathbb{E}[Y]$ minimises $\mathbb{E}[(Y - \varphi(X))^2]$. This will later be the identity $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$ for independent X, Y: there is no information in X to predict Y better than with its expectation.

3.8 Proposition (Factorisation Lemma)**.** *Let X be a $(S, \mathscr{S})$-valued and Y a real-valued random variable. Then Y is $\sigma(X)$-measurable if and only if there is a $(\mathscr{S}, \mathfrak{B}_{\mathbb{R}})$-measurable function $\varphi : S \to \mathbb{R}$ such that $Y = \varphi(X)$.*

Proof. Since X is $\sigma(X)$-measurable, so is any composition $Y = \varphi(X)$ with a measurable function φ. This gives one direction.

Conversely, let Y be $\sigma(X)$-measurable. We argue via measure-theoretic induction. For simple $Y = \sum_{k=1}^n a_k \mathbf{1}_{B_k}$ with $a_k \in \mathbb{R}$ we can assume without loss of generality (w.l.o.g.) that $B_k \cap B_l = \varnothing$ and $a_k \neq a_l$ for $k \neq l$. Then $B_k = Y^{-1}(\{a_k\})$ is in $\sigma(X)$ (since Y is $\sigma(X)$-measurable) and thus $B_k = X^{-1}(A_k)$ for some $A_k \in \mathscr{S}$. Therefore $\varphi := \sum_{k=1}^n a_k \mathbf{1}_{A_k}$ is $\mathscr{S}$-measurable and satisfies $\varphi(X) = Y$.

For $Y \in \mathcal{M}^+(\Omega, \sigma(X))$ we can find simple nonegative Y_n with $Y_n \uparrow Y$ and measurable $\varphi_n : S \to [0, \infty)$ with $Y_n = \varphi_n(X)$. Let us define inductively $\tilde{\varphi}_1 = \varphi_1$, $\tilde{\varphi}_{n+1}(x) = \max(\varphi_{n+1}(x), \tilde{\varphi}_n(x))$. Since $\varphi_{n+1}(X) = Y_{n+1} \geqslant Y_n = \varphi_n(X)$, we have $\varphi_{n+1}(x) \geqslant \varphi_n(x)$ for all $x \in X(\Omega)$ and thus $\tilde{\varphi}_n(X) = \varphi_n(X) = Y_n$. Moreover, $\tilde{\varphi}_n$ is measurable and $\tilde{\varphi}_{n+1}(x) \geqslant \tilde{\varphi}_n(x)$ holds for all x. Then $\varphi(x) := \lim_{n\to\infty} \tilde{\varphi}_n(x) \in [0, \infty]$ exists and is measurable. Hence, we have by definition $\varphi(X) = \lim_{n\to\infty} \tilde{\varphi}_n(X) = Y$.

For $Y \in \mathcal{M}(\Omega, \sigma(X))$ write $Y = Y^+ - Y^-$ with $Y^+, Y^- \in \mathcal{M}^+(\Omega, \sigma(X))$ and $Y^+ = \varphi^+(X)$, $Y^- = \varphi^-(X)$ with measurable φ^+, φ^-. Setting $\varphi(x) = (\varphi^+(x) - \varphi^-(x))\mathbf{1}(\varphi^+(x) < \infty, \varphi^-(x) < \infty)$, we check that φ is measurable and satisfies $\varphi(X) = Y$ (note that Y^+, Y^- do not take on the value $+\infty$). □

3.9 Lemma. *Let $\mathscr{G}$ be a sub-σ-algebra of $\mathscr{F}$. Then $L^2(\Omega, \mathscr{G}, \mathbb{P})$ is embedded as closed linear subspace in the Hilbert space $L^2(\Omega, \mathscr{F}, \mathbb{P})$.*

Proof. By definition, we have $\mathscr{L}^2(\Omega,\mathscr{G},\mathbb{P}) \subseteq \mathscr{L}^2(\Omega,\mathscr{F},\mathbb{P})$. For $f,g \in \mathscr{L}^2(\Omega,\mathscr{G},\mathbb{P})$ with $f = g$ $\mathbb{P}$-a.s. (i.e., $[f]_{\mathscr{G}} = [g]_{\mathscr{G}}$) we also have $f,g \in \mathscr{L}^2(\Omega,\mathscr{F},\mathbb{P})$ with $f = g$ $\mathbb{P}$-a.s. (i.e., $[f]_{\mathscr{F}} = [g]_{\mathscr{F}}$) and the equivalence classes are embedded in a well defined and isometric way. Since $L^2(\Omega,\mathscr{G},\mathbb{P})$ is a complete linear space (compare Stochastik I or Functional Analysis), its embedding is a complete linear subspace of $L^2(\Omega,\mathscr{F},\mathbb{P})$ and hence also closed. □

3.10 Remark. We now define the conditional expectation $\mathbb{E}[Y \,|\, X]$ as the best prediction of Y by a measurable function of X with respect to mean squared error. Due to the factorisation lemma this is equivalent to the best $\sigma(X)$-measurable prediction of Y. We are thus lead to consider more generally best $\mathscr{G}$-measurable predictions of Y for a sub-σ-algebra $\mathscr{G}$. This may be compared to the best prediction of Y by a linear function of X in regression analysis, compare Stochastik I.

3.11 Definition. Let X be a random variable on $(\Omega,\mathscr{F},\mathbb{P})$. Then for $Y \in L^2(\Omega,\mathscr{F},\mathbb{P})$ the conditional expectation (bedingte Erwartung) of Y given X is defined as the $L^2(\Omega,\mathscr{F},\mathbb{P})$-orthogonal projection of Y onto $L^2(\Omega,\sigma(X),\mathbb{P})$:

$$\mathbb{E}[Y \,|\, X] := P_{L^2(\Omega,\sigma(X),\mathbb{P})} Y.$$

If φ is a measurable function from the factorisation lemma with $\mathbb{E}[Y \,|\, X] = \varphi(X)$ a.s., we write $\mathbb{E}[Y \,|\, X = x] := \varphi(x)$ (conditional expected value, bedingter Erwartungswert).

More generally, for a sub-σ-algebra $\mathscr{G}$ the conditional expectation of $Y \in L^2(\Omega,\mathscr{F},\mathbb{P})$ given $\mathscr{G}$ is defined as

$$\mathbb{E}[Y \,|\, \mathscr{G}] = P_{L^2(\Omega,\mathscr{G},\mathbb{P})} Y.$$

3.12 Remark. The conditional expectations $\mathbb{E}[Y \,|\, X]$ or $\mathbb{E}[Y \,|\, \mathscr{G}]$ are only $\mathbb{P}$-almost surely defined (as elements in L^2). Similarly, $\mathbb{E}[Y \,|\, X = x]$ as a function of x is $\mathbb{P}^X$-almost surely uniquely defined.

3.13 Example. There are two extreme cases: For $\mathscr{G} = \mathscr{F}$ we obtain $\mathbb{E}[Y \,|\, \mathscr{F}] = Y$ because $P_{L^2(\Omega,\mathscr{F},P)}$ is the identity. For $\mathscr{G} = \{\varnothing, \Omega\}$ we find $\mathbb{E}[Y \,|\, \mathscr{G}] = \mathbb{E}[Y]$ because $L^2(\Omega,\mathscr{G},P)$ consists of P-a.s. constant random variables and $\mathbb{E}[Y] = \operatorname{argmin}_{c\in\mathbb{R}} \mathbb{E}[(Y-c)^2]$ (Stochastik I).

3.14 Lemma. *(Properties of the L^2-conditional expectation) For $Y \in L^2(\Omega,\mathscr{F},\mathbb{P})$ and a sub-σ-algebra $\mathscr{G} \subseteq \mathscr{F}$ the conditional expectation satisfies:*

(a) $\mathbb{E}[Y \,|\, \mathscr{G}] \in L^2(\Omega,\mathscr{F},\mathbb{P})$ *has a $\mathscr{G}$-measurable version: there is* $Z \in \mathscr{L}^2(\Omega,\mathscr{G},\mathbb{P})$ *with* $\mathbb{E}[Y \,|\, \mathscr{G}] = Z$ *$\mathbb{P}$-a.s.*

(b) $\mathbb{E}[Y \,|\, \mathscr{G}] = \operatorname{argmin}_{Z\in L^2(\Omega,\mathscr{G},\mathbb{P})} \mathbb{E}[(Y-Z)^2]$;

(c) $\forall\, \alpha \in \mathbb{R},\ Z \in L^2(\Omega,\mathscr{F},\mathbb{P}):\ \mathbb{E}[\alpha Y + Z \,|\, \mathscr{G}] = \alpha\,\mathbb{E}[Y \,|\, \mathscr{G}] + \mathbb{E}[Z \,|\, \mathscr{G}]$ *a.s.;*

(d) $\forall\, Z \in L^2(\Omega,\mathscr{G},\mathbb{P}):\ \mathbb{E}[\mathbb{E}[Y \,|\, \mathscr{G}]Z] = \mathbb{E}[YZ]$;

(e) $Y \geqslant 0$ *$\mathbb{P}$-a.s. implies* $\mathbb{E}[Y \,|\, \mathscr{G}] \geqslant 0$ *$\mathbb{P}$-a.s.*

Proof. Parts (a) and (b) follow immediately from the definition as an orthogonal projection onto the embedding of $L^2(\Omega,\mathscr{G},\mathbb{P})$ into $L^2(\Omega,\mathscr{F},\mathbb{P})$. The linearity of the projection implies (c). Its self-adjointness implies (d) via

$$\mathbb{E}[\mathbb{E}[Y\,|\,\mathscr{G}]Z] = \langle P_{L^2(\Omega,\mathscr{G},\mathbb{P})}Y, Z\rangle_{L^2} = \langle Y, P_{L^2(\Omega,\mathscr{G},\mathbb{P})}Z\rangle_{L^2} = \mathbb{E}[YZ]$$

for $Z\in L^2(\Omega,\mathscr{G},\mathbb{P})$. For (e) let $G=\{Z<0\}$ for a $\mathscr{G}$-measurable version Z of $\mathbb{E}[Y\,|\,\mathscr{G}]$. Then $G\in\mathscr{G}$ and we deduce from (d) and $Y\geqslant 0$

$$\mathbb{E}[Z\mathbf{1}_G] = \mathbb{E}[Y\mathbf{1}_G] \geqslant 0.$$

This shows $Z\geqslant 0$ $\mathbb{P}$-a.s. and thus $\mathbb{E}[Y\,|\,\mathscr{G}]\geqslant 0$ $\mathbb{P}$-a.s. □

3.15 Remark. The definition of the conditional expectation via orthogonal projections for L^2-random variables is for most cases sufficient. Note, however, that $L^2\subsetneq L^1$ and it would be natural to define conditional expectations for all L^1-random variables. To this end we first characterise the conditional expectation for $Y\in L^2$ in a way which is well defined even for $Y\in L^1$ and then extend it from L^2 to L^1 by approximation.

3.16 Lemma. *$\mathbb{E}[Y\,|\,\mathscr{G}]$ for $Y\in L^2(\Omega,\mathscr{F},\mathbb{P})$ is as an element of $L^2(\Omega,\mathscr{F},\mathbb{P})$ uniquely determined by the following properties:*

(a) $\mathbb{E}[Y\,|\,\mathscr{G}]$ has a $\mathscr{G}$-measurable version;

(b) $\forall\, G\in\mathscr{G}:\ \mathbb{E}[\mathbb{E}[Y\,|\,\mathscr{G}]\mathbf{1}_G] = \mathbb{E}[Y\mathbf{1}_G]$.

Proof. By Lemma 3.14(a,d) parts (a) and (b) hold for $\mathbb{E}[Y\,|\,\mathscr{G}]$, noting $Z=\mathbf{1}_G\in L^2(\Omega,\mathscr{G},\mathbb{P})$ for $G\in\mathscr{G}$.

Now suppose $Z\in L^2(\Omega,\mathscr{F},\mathbb{P})$ satisfies (a) and (b). Then using (b) also for $\mathbb{E}[Y\,|\,\mathscr{G}]$ gives

$$\mathbb{E}[(Z-\mathbb{E}[Y\,|\,\mathscr{G}])\mathbf{1}_G] = 0 \text{ for all } G\in\mathscr{G}.$$

Consider the events $G^> = \{Z>\mathbb{E}[Y\,|\,\mathscr{G}]\}$ and $G^< = \{Z<\mathbb{E}[Y\,|\,\mathscr{G}]\}$, which are in $\mathscr{G}$ if $\mathscr{G}$-measurable versions according to (a) are taken. This implies

$$\mathbb{E}[|Z-\mathbb{E}[Y\,|\,\mathscr{G}]|\mathbf{1}_{G^>\cup G^<}] = \mathbb{E}[(Z-\mathbb{E}[Y\,|\,\mathscr{G}])\mathbf{1}_{G^>}] - \mathbb{E}[(Z-\mathbb{E}[Y\,|\,\mathscr{G}])\mathbf{1}_{G^<}] = 0.$$

We deduce $|Z-\mathbb{E}[Y\,|\,\mathscr{G}]|=0$ $\mathbb{P}$-a.s., that is $Z=\mathbb{E}[Y\,|\,\mathscr{G}]$ $\mathbb{P}$-a.s. □

3.17 Theorem. *(general conditional expectation) Let $Y\in\mathcal{M}^+(\Omega,\mathscr{F})$ or $Y\in L^1(\Omega,\mathscr{F},\mathbb{P})$ and let $\mathscr{G}$ be a sub-σ-algebra of $\mathscr{F}$. Then there is a $\mathbb{P}$-a.s. unique element $\mathbb{E}[Y\,|\,\mathscr{G}]$ in $\mathcal{M}^+(\Omega,\mathscr{G})$ and $L^1(\Omega,\mathscr{G},\mathbb{P})$, respectively, such that*

$$\forall\, G\in\mathscr{G}:\ \mathbb{E}[\mathbb{E}[Y\,|\,\mathscr{G}]\mathbf{1}_G] = \mathbb{E}[Y\mathbf{1}_G].$$

Proof. First, consider $Y\in\mathcal{M}^+(\Omega,\mathscr{F})$. Then $Y_n:=\min(Y,n)\in L^2(\Omega,\mathscr{F},\mathbb{P})$ and $Y_n\uparrow Y$. For Y_n the conditional expectation is well defined and by monotonicity $\mathbb{E}[Y_{n+1}\,|\,\mathscr{G}]\geqslant\mathbb{E}[Y_n\,|\,\mathscr{G}]$, $n\geqslant 1$, holds $\mathbb{P}$-a.s. Then also the limit

$\mathbb{E}[Y \,|\, \mathscr{G}] := \lim_{n\to\infty} \mathbb{E}[Y_n \,|\, \mathscr{G}]$ exists $\mathbb{P}$-a.s. and has a $\mathscr{G}$-measurable version. Monotone convergence implies for $G \in \mathscr{G}$:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y \,|\, \mathscr{G}]\mathbf{1}_G] &= \mathbb{E}\Big[\lim_{n\to\infty} \mathbb{E}[Y_n \,|\, \mathscr{G}]\mathbf{1}_G\Big] = \lim_{n\to\infty} \mathbb{E}[\mathbb{E}[Y_n \,|\, \mathscr{G}]\mathbf{1}_G]\\ &= \lim_{n\to\infty} \mathbb{E}[Y_n \mathbf{1}_G] = \mathbb{E}\Big[\lim_{n\to\infty} Y_n \mathbf{1}_G\Big] = \mathbb{E}[Y\mathbf{1}_G].\end{aligned}$$

For $Y \in L^1(\Omega, \mathscr{F}, \mathbb{P})$ write $Y = Y^+ - Y^-$ with $Y^+, Y^- \in \mathcal{M}^+(\Omega, \mathscr{F})$ and set

$$\mathbb{E}[Y \,|\, \mathscr{G}] := \Big(\mathbb{E}[Y^+ \,|\, \mathscr{G}] - \mathbb{E}[Y^- \,|\, \mathscr{G}]\Big)\mathbf{1}\Big(\mathbb{E}[Y^+ \,|\, \mathscr{G}] < \infty, \mathbb{E}[Y^- \,|\, \mathscr{G}] < \infty\Big).$$

It is straightforward to check that the asserted properties are satisfied.

Concerning uniqueness assume that there are $Z, \tilde{Z} \in \mathcal{M}^+(\Omega, \mathscr{G})$ with $\mathbb{E}[Z\mathbf{1}_G] = \mathbb{E}[\tilde{Z}\mathbf{1}_G]$ for all $G \in \mathscr{G}$. Then for $G^> = \{Z > \tilde{Z}\}$ we deduce $\mathbb{E}[(Z - \tilde{Z})\mathbf{1}_{G^>}] = 0$, hence $\mathbb{P}(G^>) = 0$. For $G^< = \{Z < \tilde{Z}\}$ the same argument yields $\mathbb{P}(G^<) = 0$ and thus $\mathbb{P}(Z = \tilde{Z}) = 1$. The same lines also show uniqueness in $L^1(\Omega, \mathscr{G}, \mathbb{P})$. □

3.18 Definition. For $Y \in \mathcal{M}^+(\Omega, \mathscr{F})$ or $Y \in L^1(\Omega, \mathscr{F}, \mathbb{P})$ and a sub-σ-algebra $\mathscr{G}$ of $\mathscr{F}$ the (general) conditional expectation of Y given $\mathscr{G}$ is defined as $\mathbb{E}[Y \,|\, \mathscr{G}]$ from the preceding theorem. We put

$$\mathbb{E}[Y \,|\, (X_i)_{i\in I}] := \mathbb{E}[Y \,|\, \sigma(X_i,\, i \in I)]$$

for random variables X_i, $i \in I$.

▷ **Control questions**

(a) Verify the factorisation lemma for discrete random variables X, i.e. for countable $X(\Omega)$.

If X is discrete, then $\sigma(X) = \sigma(X^{-1}(\{s_k\}), k \geqslant 1)$ when $X(\Omega) = \{s_k \,|\, k \geqslant 1\}$ is an enumeration. If Y is $\sigma(X)$-measurable, then Y must be constant on each set $X^{-1}(\{s_k\})$ (because $Y^{-1}(\{y\}) \in \sigma(X^{-1}(\{s_k\}), k \geqslant 1)$ and all $X^{-1}(\{s_k\})$ are disjoint, $y \in \mathbb{R}$). Call this constant y_k and conclude $Y(\omega) = \sum_{k\geqslant 1} y_k \mathbf{1}(X(\omega) = s_k)$, which is a measurable function of X.

(b) Consider $L^2([0,1], \mathfrak{B}_{[0,1]}, \lambda)$ with the Lebesgue measure λ on $[0,1]$ and $\mathscr{G} = \sigma([(k-1)/n, k/n), k = 1, \dots, n)$. What are the elements of $L^2([0,1], \mathscr{G}, \lambda)$ and what is the difference with its embedding into $L^2([0,1], \mathfrak{B}_{[0,1]}, \lambda)$?

Since Lebesgue measure is positive for all events in $\mathscr{G}$ (union of intervals $[(k-1)/n, k/n))$ besides the empty set, the space $L^2([0,1], \mathscr{G}, \lambda)$ consists of all piecewise constant functions on the intervals $[(k-1)/n, k/n)$ (or very formally equivalence classes with just one element of that form). Its embedding into $L^2([0,1], \mathfrak{B}_{[0,1]}, \lambda)$ consists of all functions which are Lebesgue-almost everywhere equal to such a piecewise constant function (or more precisely, the equivalence classes of that form). For instance $f(x) = \mathbf{1}(x = 1/2)$ is not in $\mathcal{L}^2([0,1], \mathscr{G}, \lambda)$ or a member of an equivalence class of $L^2([0,1], \mathscr{G}, \lambda)$, but it lies in its embedding into $L^2([0,1], \mathfrak{B}_{[0,1]}, \lambda)$ (member of the equivalence class $[0]_{\mathfrak{B}_{[0,1]}}$).

(c) Let $Y \in L^2$ and $\mathscr{H} \subseteq \mathscr{G}$ be sub-σ-algebras. How does the *tower property* $\mathbb{E}[\mathbb{E}[Y \,|\, \mathscr{G}] \,|\, \mathscr{H}] = \mathbb{E}[Y \,|\, \mathscr{H}]$ follow from the projection property?

For orthogonal projections onto closed subspace V, V' with $V \subseteq V'$ we have $P_V P_{V'} = P_V$ which follows e.g. via selfadjointness and $P_{V'}x = x$ for $x \in V \subseteq V'$:

$$\langle P_V P_{V'} v, w\rangle = \langle v, P_{V'} P_V w\rangle = \langle v, P_V w\rangle = \langle P_V v, w\rangle \text{ for all } v, w.$$

Since $L^2(\Omega, \mathscr{H}, \mathbb{P}) \subseteq L^2(\Omega, \mathscr{G}, \mathbb{P})$ holds (as subspaces of $L^2(\Omega, \mathscr{F}, \mathbb{P})$), we obtain

$$\mathbb{E}[\mathbb{E}[Y \,|\, \mathscr{G}] \,|\, \mathscr{H}] = P_{L^2(\Omega,\mathscr{H},\mathbb{P})} P_{L^2(\Omega,\mathscr{G},\mathbb{P})} Y = P_{L^2(\Omega,\mathscr{H},\mathbb{P})} Y = \mathbb{E}[Y \,|\, \mathscr{H}].$$

3.19 Proposition. *(Properties of the general conditional expectation) Let $Y \in L^1(\Omega, \mathscr{F}, \mathbb{P})$ and let $\mathscr{G}$ be a sub-σ-algebra of $\mathscr{F}$. Then:*

(a) $\mathbb{E}[\mathbb{E}[Y \,|\, \mathscr{G}]] = \mathbb{E}[Y]$*;*

(b) Y *$\mathscr{G}$-measurable* $\Rightarrow \mathbb{E}[Y \,|\, \mathscr{G}] = Y$ *a.s.;*

(c) $\alpha \in \mathbb{R}$, $Z \in L^1(\Omega, \mathscr{F}, \mathbb{P})$: $\mathbb{E}[\alpha Y + Z \,|\, \mathscr{G}] = \alpha\, \mathbb{E}[Y \,|\, \mathscr{G}] + \mathbb{E}[Z \,|\, \mathscr{G}]$ *a.s.;*

(d) $Y \geqslant 0$ *a.s.* $\Rightarrow \mathbb{E}[Y \,|\, \mathscr{G}] \geqslant 0$ *a.s.;*

(e) $Y_n \in \mathcal{M}^+(\Omega, \mathscr{F})$, $Y_n \uparrow Y$ *a.s.* $\Rightarrow$ $\mathbb{E}[Y_n \,|\, \mathscr{G}] \uparrow \mathbb{E}[Y \,|\, \mathscr{G}]$ *a.s. (monotone convergence);*

(f) $Y_n \in \mathcal{M}^+(\Omega, \mathscr{F}) \Rightarrow \mathbb{E}[\liminf_n Y_n \,|\, \mathscr{G}] \leqslant \liminf_n \mathbb{E}[Y_n \,|\, \mathscr{G}]$ *a.s. (Fatou's Lemma);*

(g) $Y_n \in \mathcal{M}(\Omega, \mathscr{F})$, $Y_n \to Y$, $|Y_n| \leqslant Z$ *with* $Z \in L^1(\Omega, \mathscr{F}, \mathbb{P})$: $\mathbb{E}[Y_n \,|\, \mathscr{G}] \to \mathbb{E}[Y \,|\, \mathscr{G}]$ *a.s. (dominated convergence);*

(h) $\mathscr{H} \subseteq \mathscr{G} \Rightarrow \mathbb{E}[\mathbb{E}[Y \,|\, \mathscr{G}] \,|\, \mathscr{H}] = \mathbb{E}[Y \,|\, \mathscr{H}]$ *a.s. (projection/tower property);*

(i) Z *$\mathscr{G}$-measurable,* $ZY \in L^1$: $\mathbb{E}[ZY \,|\, \mathscr{G}] = Z\, \mathbb{E}[Y \,|\, \mathscr{G}]$ *a.s.;*

(j) Y *independent of* $\mathscr{G}$: $\mathbb{E}[Y \,|\, \mathscr{G}] = \mathbb{E}[Y]$ *a.s.*

Proof.

(a) $\Omega \in \mathscr{G}$ implies $\mathbb{E}[\mathbb{E}[Y \,|\, \mathscr{G}]\mathbf{1}_\Omega] = \mathbb{E}[Y\mathbf{1}_\Omega]$ and thus the claim.

(b) Y satisfies the defining properties of $\mathbb{E}[Y \,|\, \mathscr{G}]$.

(c) The right-hand side satisfies the defining properties of $\mathbb{E}[\alpha Y + Z \,|\, \mathscr{G}]$, using linearity for the expectation.

(d) Let $G = \{\mathbb{E}[Y \,|\, \mathscr{G}] < 0\} \in \mathscr{G}$ for a $\mathscr{G}$-measurable version of the conditional expectation. $Y \geqslant 0$ a.s. therefore implies

$$\mathbb{E}[\mathbb{E}[Y \,|\, \mathscr{G}]\mathbf{1}_G] = \mathbb{E}[Y\mathbf{1}_G] \geqslant 0 \Rightarrow \mathbb{P}(G) = 0.$$

(e) Using (c), (d), we infer $\mathbb{E}[Y_n \,|\, \mathscr{G}] \uparrow U$ for some $U \in \mathcal{M}^+(\Omega, \mathscr{G})$. Monotone convergence for expectations shows

$$\forall G \in \mathscr{G}:\ \mathbb{E}[U\mathbf{1}_G] = \lim_{n\to\infty} \mathbb{E}[\mathbb{E}[Y_n \,|\, \mathscr{G}]\mathbf{1}_G] = \lim_{n\to\infty} \mathbb{E}[Y_n\mathbf{1}_G] = \mathbb{E}[Y\mathbf{1}_G].$$

We conclude $U = \mathbb{E}[Y \,|\, \mathscr{G}]$.

(f) ▶Exercise

(g) ▶Exercise

(h) The left-hand side satisfies the defining properties of $\mathbb{E}[Y \,|\, \mathscr{H}]$. For $Y \in L^2$ this is just the composition of orthogonal projections.▶Control

(i) For $Z = \mathbf{1}_{G'}$, $G' \in \mathscr{G}$, the right-hand side satisfies the defining properties of $\mathbb{E}[ZY \,|\, \mathscr{G}]$. Use measure-theoretic induction to extend the results to simple $\mathscr{G}$-measurable functions Z, to $Y, Z \in \mathscr{M}^+(\Omega, \mathscr{G})$ and finally to $Y, Z \in \mathcal{M}(\Omega, \mathscr{G})$ with $ZY \in L^1$.

(j) By independence of Y and $\mathbf{1}_G$ for $G \in \mathscr{G}$ and by Fubini's theorem we see that $\mathbb{E}[Y]$ satisfies the defining properties of $\mathbb{E}[Y \,|\, \mathscr{G}]$.

□

3.20 Example. Consider the compound Poisson process $X_t = \sum_{k=1}^{N_t} Y_k$ with a Poisson process N_t of intensity λ and i.i.d. random variables $Y_k \in L^1$, where (N_t) and (Y_k) are independent. What is $\mathbb{E}[X_t]$? This follows easily by conditioning on N_t:▶Control

$$\mathbb{E}[X_t] = \mathbb{E}[\mathbb{E}[X_t \,|\, N_t]] = \mathbb{E}\Big[\sum_{k=1}^{N_t} \mathbb{E}[Y_k \,|\, N_t]\Big] = \mathbb{E}\left[N_t\,\mathbb{E}[Y_1]\right] = \lambda t\,\mathbb{E}[Y_1].$$

3.21 Proposition (Jensen's Inequality). *If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex and Y, $\varphi(Y)$ are in L^1, then*

$$\varphi(\mathbb{E}[Y \,|\, \mathscr{G}]) \leqslant \mathbb{E}[\varphi(Y) \,|\, \mathscr{G}]$$

holds for any sub-σ-algebra $\mathscr{G}$ of $\mathscr{F}$.

Proof. We know for convex φ (compare Stochastik I)

$$\varphi(y) = \sup_{x\in\mathbb{R}} \big(\varphi(x) + \varphi'(x+)(y-x)\big), \quad y \in \mathbb{R}\,.$$

The monotonicity and linearity of conditional expectations, given by Proposition 3.19(c,d), imply for any $x \in \mathbb{R}$

$$\mathbb{E}[\varphi(Y) \,|\, \mathscr{G}] \geqslant \mathbb{E}\left[\varphi(x) + \varphi'(x+)(Y-x) \,\middle|\, \mathscr{G}\right] = \varphi(x) + \varphi'(x+)\big(\mathbb{E}[Y \,|\, \mathscr{G}] - x\big).$$

Taking the supremum over all $x \in \mathbb{R}$ on the right-hand side yields the assertion.

□

3.22 Definition. For a sub-σ-algebra $\mathscr{G} \subseteq \mathscr{F}$ define the conditional probability of $A \in \mathscr{F}$ given $\mathscr{G}$ as

$$\mathbb{P}(A \,|\, \mathscr{G}) = \mathbb{E}[\mathbf{1}_A \,|\, \mathscr{G}].$$

3.23 Remark. $\mathbb{P}(A \,|\, \mathscr{G})$ is only $\mathbb{P}$-a.s. defined. For fixed pairwise disjoint $A_n \in \mathscr{F}$, $n \geqslant 1$, we have the σ-additivity $\mathbb{P}(\bigcup_{n\geqslant 1} A_n \,|\, \mathscr{G}) = \sum_{n \geqslant 1} \mathbb{P}(A_n \,|\, \mathscr{G})$ $\mathbb{P}$-a.s., but there might be no version such that $A \mapsto \mathbb{P}(A \,|\, \mathscr{G})$ is indeed a probability measure on all of $\mathscr{F}$. For Polish spaces such a version, a so-called regular conditional probability or Markov kernel, always exists, see e.g. Klenke. In the case of densities, the conditional density gives a constructive way to define conditional probabilities, see the next example and the exercises.

3.24 Example. Let $f^{X,Y}$ be the joint density of two random variables X, Y with respect to some product measure $\mu \otimes \nu$ (like two-dimensional Lebesgue measure). We claim that for any event B

$$\mathbb{P}(Y \in B \,|\, X) = \frac{\int_B f^{X,Y}(X,y)\nu(dy)}{f^X(X)} \text{ with } f^X(x) = \int f^{X,Y}(x,y)\nu(dy) \tag{3.1}$$

holds $\mathbb{P}$-a.s., where the right-hand side is well defined $\mathbb{P}$-almost surely. Since densities are measurable the right-hand side is a measurable function in X, hence $\sigma(X)$-measurable. We have $\mathbb{P}(f^X(X) = 0) = \int \mathbf{1}(f^X(x) = 0) f^X(x)\mu(dx) = 0$ and the denominator is $\mathbb{P}$-almost surely strictly positive so that the right-hand side in formula (3.1) is $\mathbb{P}$-a.s. well defined.

Moreover, any $G \in \sigma(X)$ can be written as $G = X^{-1}(F)$ for an event F and we check

$$\begin{aligned}
\mathbb{E}\Big[\mathbf{1}_G \frac{\int_B f^{X,Y}(X,y)\nu(dy)}{f^X(X)}\Big] &= \int \mathbf{1}_F(x) \frac{\int \mathbf{1}_B(y) f^{X,Y}(x,y)\nu(dy)}{f^X(x)} f^X(x)\mu(dx) \\
&= \int\int \mathbf{1}_F(x)\mathbf{1}_B(y) f^{X,Y}(x,y)\,\nu(dy)\mu(dx) \\
&= \mathbb{E}[\mathbf{1}_F(X)\mathbf{1}_B(Y)] = \mathbb{E}[\mathbf{1}_G \mathbf{1}_B(Y)]
\end{aligned}$$

and the right-hand side in (3.1) is indeed a version of the conditional probability $\mathbb{P}(Y \in B \,|\, X) = \mathbb{E}[\mathbf{1}_B(Y) \,|\, X]$.

4 Martingale theory

4.1 Martingales, sub- and supermartingales

4.1 Definition. A sequence $(\mathscr{F}_n)_{n \geqslant 0}$ of sub-σ-algebras of $\mathscr{F}$ is called filtration if $\mathscr{F}_n \subseteq \mathscr{F}_{n+1}$, $n \geqslant 0$, holds. $(\Omega, \mathscr{F}, \mathbb{P}, (\mathscr{F}_n))$ is called filtered probability space.

4.2 Definition. A sequence $(M_n)_{n \geqslant 0}$ of random variables on a filtered probability space $(\Omega, \mathscr{F}, \mathbb{P}, (\mathscr{F}_n))$ forms a martingale (submartingale, supermartingale) if:

(a) $M_n \in L^1$, $n \geqslant 0$;

(b) M_n is $\mathscr{F}_n$-measurable, $n \geqslant 0$ (adapted);

(c) $\mathbb{E}[M_{n+1} \,|\, \mathscr{F}_n] = M_n$ (resp. $\mathbb{E}[M_{n+1} \,|\, \mathscr{F}_n] \geqslant M_n$ for submartingale, resp. $\mathbb{E}[M_{n+1} \,|\, \mathscr{F}_n] \leqslant M_n$ for supermartingale).

If $\mathscr{F}_n = \sigma(M_0, \dots, M_n)$ holds, then $(\mathscr{F}_n)$ is the natural filtration of M, notation $(\mathscr{F}_n^M)$.

4.3 Remark.

(a) For martingales (M_n) the expectation is constant: $\mathbb{E}[M_n] = \mathbb{E}[\mathbb{E}[M_n \,|\, \mathscr{F}_{n-1}]] = \mathbb{E}[M_{n-1}] = \cdots = \mathbb{E}[M_0]$. Similarly, for submartingales the expectation is increasing and for supermartingales the expectation is decreasing.

(b) Suppose (M_n) is even an L^2-martingale (i.e., $M_n \in L^2$ for all $n \geqslant 0$). Then the martingale differences $\Delta_n M := M_{n+1} - M_n$, $\Delta_m M := M_{m+1} - M_m$ for $m < n$ are uncorrelated:

$$\begin{aligned}\mathbb{E}[\Delta_m M \Delta_n M] &= \mathbb{E}\Big[\mathbb{E}[\Delta_m M \Delta_n M \,|\, \mathscr{F}_{m+1}]\Big] = \mathbb{E}\Big[\Delta_m M \,\mathbb{E}[\Delta_n M \,|\, \mathscr{F}_{m+1}]\Big] \\ &= \mathbb{E}\Big[\Delta_m M \,\mathbb{E}\big[\mathbb{E}[\Delta_n M \,|\, \mathscr{F}_n] \,\big|\, \mathscr{F}_{m+1}\big]\Big] = 0\end{aligned}$$

since $\Delta_m M$ is $\mathscr{F}_{m+1}$-measurable, $\mathscr{F}_{m+1} \subseteq \mathscr{F}_n$ and $\mathbb{E}[\Delta_n M \,|\, \mathscr{F}_n] = 0$ by the martingale property.

It turns out that the martingale property is exactly the right generalisation of sums of independent random variables to still provide a rich theory, e.g. a law of large numbers. Many more complicated processes are analysed by a decomposition or transformation using martingales.

4.4 Example.

(a) Let $(X_k)_{k \geqslant 1}$ be independent random variables with $X_k \in L^1$, $\mathbb{E}[X_k] = 0$, $k \geqslant 1$. Put $S_0 = 0$, $S_n = \sum_{k=1}^n X_k$. For $n \geqslant 1$ we have $\mathscr{F}_n^S = \sigma(X_1, \dots, X_n)$ and $(S_n)_{n \geqslant 0}$ is a martingale with respect to its natural filtration. Similarly, $(S_n)_{n \geqslant 0}$ is a submartingale if $\mathbb{E}[X_k] \geqslant 0$ for all $k \geqslant 1$ and a supermartingale if $\mathbb{E}[X_k] \leqslant 0$ for all $k \geqslant 1$.

(b) Let $(X_k)_{k \geqslant 1}$ be independent random variables with $X_k \in L^1$, $\mathbb{E}[X_k] = 1$, $k \geqslant 1$. Put $P_0 = 1$, $P_n = \prod_{k=1}^n X_k$. For $n \geqslant 1$ we have $\mathscr{F}_n^P \subseteq \sigma(X_1, \dots, X_n)$ and $(P_n)_{n \geqslant 0}$ is a martingale with respect to its natural filtration because $\mathbb{E}[X_{n+1} \,|\, \mathscr{F}_n^P] = \mathbb{E}[X_{n+1}]$ by independence and thus

$$\mathbb{E}[P_{n+1} \,|\, \mathscr{F}_n^P] = \mathbb{E}\Big[\prod_{k=1}^{n+1} X_k \,\Big|\, \mathscr{F}_n^P\Big] = \mathbb{E}[X_{n+1} \,|\, \mathscr{F}_n^P] \prod_{k=1}^{n} X_k = P_n.$$

In Stochastik I we have seen that for $\mathbb{P}(X_k = 3/2) = \mathbb{P}(X_k = 1/2) = 1/2$ we have $P_n \to 0$ $\mathbb{P}$-a.s. although $\mathbb{E}[P_n] = 1$ for all $n \geqslant 0$.

(c) Let $X \in L^1$ and $(\mathscr{F}_n)$ be any filtration, then $M_n := \mathbb{E}[X \,|\, \mathscr{F}_n]$ defines a martingale with respect to $(\mathscr{F}_n)$.▶Control

4.5 Definition. A martingale (M_n) with respect to a filtration $(\mathscr{F}_n)$ is called closable (abschließbar), if there exists an $X \in L^1$ with $M_n = \mathbb{E}[X \,|\, \mathscr{F}_n]$, $n \geqslant 0$.

▷ **Control questions**

(a) Give all formal details for the calculation of $\mathbb{E}[X_t]$ in Example 3.20.

We write $X_t = \sum_{k=1}^{\infty} Y_k \mathbf{1}(k \leqslant N_t)$ and observe that the series converges a.s. Hence, we obtain

$$\mathbb{E}[X_t \,|\, N_t] = \lim_{K\to\infty} \mathbb{E}\Big[\sum_{k=1}^{K} Y_k \mathbf{1}(k \leqslant N_t) \,\Big|\, N_t\Big] = \lim_{K\to\infty} \sum_{k=1}^{K} \mathbb{E}[Y_k] \mathbf{1}(k \leqslant N_t)$$
$$= N_t \,\mathbb{E}[Y_1]$$

by dominated convergence for conditional expectations. For this note that $\sum_{k=1}^{K} Y_k \mathbf{1}(k \leqslant N_t)$ is dominated by $\sum_{k=1}^{\infty} |Y_k| \mathbf{1}(k \leqslant N_t)$ and monotone convergence in the above equation line for $|Y_k|$ instead of Y_k shows that this has finite expectation $\lambda t \,\mathbb{E}[|Y_1|]$.

(b) What is called a *martingale* (*Martingal*) in real life?

This is part of the horse-gear to control the head movements of the horse.

(c) How does the martingale property of $M_n = \mathbb{E}[X \,|\, \mathscr{F}_n]$ for $X \in L^1$ and a filtration $(\mathscr{F}_n)$ follow from the tower property of conditional expectations?

We have $\mathbb{E}[M_{n+1} \,|\, \mathscr{F}_n] = \mathbb{E}[\mathbb{E}[X \,|\, \mathscr{F}_{n+1}] \,|\, \mathscr{F}_n] = \mathbb{E}[X \,|\, \mathscr{F}_n] = M_n$ by the tower property and (M_n) is a martingale.

4.6 Definition. A process $(X_n)_{n\geqslant 1}$ is predictable (vorhersehbar) (with respect to a filtration $(\mathscr{F}_n)$) if each X_n is $\mathscr{F}_{n-1}$-measurable. For a predictable process (X_n) and a martingale (or more general: adapted process) (M_n) the martingale transform (or discrete stochastic integral) $((X\bullet M)_n)_{n\geqslant 0}$ is defined by

$$(X\bullet M)_0 := 0, \quad (X\bullet M)_n := \sum_{k=1}^{n} X_k (M_k - M_{k-1}).$$

4.7 Remark. Any predictable process is adapted.

4.8 Proposition. *For a predictable process (X_n) in L^p and an L^q-martingale (M_n) (i.e., $X_n \in L^p$, $M_n \in L^q$ for all n) with $\frac{1}{p}+\frac{1}{q} = 1$ the process $((X\bullet M)_n)_{n\geqslant 0}$ is again a martingale. If M is a submartingale instead of a martingale and in addition $X_n \geqslant 0$ a.s., $n \geqslant 1$, then the process $(X\bullet M)$ is again a submartingale.*

Proof. By Hölder's inequality we have

$$\mathbb{E}[|(X\bullet M)_n|] \leqslant \sum_{k=1}^{n} \mathbb{E}[|X_k||M_k - M_{k-1}|] \leqslant \sum_{k=1}^{n} \|X_k\|_{L^p} (\|M_k\|_{L^q} + \|M_{k-1}\|_{L^q}) < \infty$$

and thus $(X\bullet M)_n \in L^1$. By definition and adaptedness of X, M it follows that $(X\bullet M)$ is adapted. By the predictability of X and the martingale property of M we conclude

$$\mathbb{E}[(X\bullet M)_{n+1} - (X\bullet M)_n \,|\, \mathscr{F}_n] = \mathbb{E}[X_{n+1}(M_{n+1} - M_n) \,|\, \mathscr{F}_n]$$
$$= X_{n+1} \,\mathbb{E}[M_{n+1} - M_n \,|\, \mathscr{F}_n] = 0,$$

hence $\mathbb{E}[(X\bullet M)_{n+1}\,|\,\mathscr{F}_n]=(X\bullet M)_n$.

For a submartingale M and $X_{n+1}\geqslant 0$ we note $X_{n+1}\,\mathbb{E}[M_{n+1}-M_n\,|\,\mathscr{F}_n]\geqslant 0$ and we conclude by the same argument that $(X\bullet M)_n$ forms a submartingale. □

4.9 Remark. Interpreting the martingale as a fair game in the sense that M_n-M_{n-1} is the payoff in round n for a stake of 1 Euro, the result can be interpreted as saying that under a predictable *investment* strategy of X_n Euros in round n a fair game remains fair (the expected gain in round n given all information until round $n-1$ is zero). Obviously, this need not be the case if X_n is not predictable, knowing future outcomes of M.

4.10 Lemma. *If (M_n) is a martingale and $\varphi:\mathbb{R}\to\mathbb{R}$ convex with $\varphi(M_n)\in L^1$, $n\geqslant 0$, then $\varphi(M_n)$ is a submartingale. The same is true for a submartingale (M_n) and an increasing convex function $\varphi:\mathbb{R}\to\mathbb{R}$.*

In particular, (M_n^2) is a submartingale for an L^2-martingale (M_n).

Proof. By Jensen's inequality $\mathbb{E}[\varphi(M_{n+1})\,|\,\mathscr{F}_n]\geqslant\varphi(\mathbb{E}[M_{n+1}\,|\,\mathscr{F}_n])=\varphi(M_n)$ holds $\mathbb{P}$-a.s. for a martingale (M_n) and a convex function φ. If (M_n) is a submartingale, then $\mathbb{E}[M_{n+1}\,|\,\mathscr{F}_n]\geqslant M_n$ holds $\mathbb{P}$-a.s. and the last equality becomes an $\geqslant$-inequality for increasing φ. The last assertion follows from the convexity of $\varphi(x)=x^2$. □

4.11 Theorem (Doob decomposition)**.** *Given a submartingale (X_n), there exists a martingale (M_n) and a predictable increasing (i.e., $A_{n+1}\geqslant A_n$ a.s.) process (A_n) such that*

$$X_n=X_0+M_n+A_n,\quad n\geqslant 1;\qquad M_0=A_0=0.$$

This decomposition is a.s. unique and $A_n=\sum_{k=1}^n\mathbb{E}[X_k-X_{k-1}\,|\,\mathscr{F}_{k-1}]$.

Proof. The process $A_n:=\sum_{k=1}^n\mathbb{E}[X_k-X_{k-1}\,|\,\mathscr{F}_{k-1}]$, $n\geqslant 1$, is by definition predictable, in L^1 and increasing (for this use that X is a submartingale). Now define $M_n:=X_n-X_0-A_n$, $n\geqslant 1$, $M_0:=0$. Then M_n is in L^1 and adapted since X_n, A_n are so. Moreover,

$$\mathbb{E}[M_{n+1}-M_n\,|\,\mathscr{F}_n]=\mathbb{E}[(X_{n+1}-X_n)-(A_{n+1}-A_n)\,|\,\mathscr{F}_n]=0$$

holds and A,M give a Doob decomposition of X.

To prove uniqueness, suppose $X_n=X_0+M_n'+A_n'$ is another Doob decomposition of X. Then $M_n-M_n'=A_n'-A_n$, $n\geqslant 1$, holds as well as $M_0-M_0'=A_0'-A_0=0$. This shows that $(M_n-M_n')_{n\geqslant 0}$ is a predictable martingale starting in zero. Yet, a predictable martingale is easily shown to be a.s. constant▶Control and thus $M_n-M_n'=0$ a.s. This shows $M_n=M_n'$, $A_n=A_n'$ a.s. □

4.12 Definition. The predictable process (A_n) in the Doob decomposition of (X_n) is called compensator of (X_n). For an L^2-martingale (M_n) the compensator of the submartingale (M_n^2) is called quadratic variation of (M_n), denoted by $\langle M\rangle_n$.

4.13 Lemma. *We have* $\langle M\rangle_n = \sum_{k=1}^n \mathbb{E}[(M_k - M_{k-1})^2 \,|\, \mathscr{F}_{k-1}]$, $n \geqslant 1$.

Proof. Using the definition and martingale property of M we conclude

$$\begin{aligned}\langle M\rangle_{n+1} - \langle M\rangle_n &= \mathbb{E}[M_{n+1}^2 - M_n^2 \,|\, \mathscr{F}_n]\\ &= \mathbb{E}[(M_{n+1} - M_n)^2 + 2M_n(M_{n+1} - M_n) \,|\, \mathscr{F}_n]\\ &= \mathbb{E}[(M_{n+1} - M_n)^2 \,|\, \mathscr{F}_n].\end{aligned}$$

With $\langle M\rangle_0 = 0$ the claim follows. □

4.14 Example.

(a) Let $(X_k)_{k\geqslant 1}$ be independent random variables with $X_k \in L^2$, $\mathbb{E}[X_k] = 0$, $k \geqslant 1$. Put $S_0 = 0$, $S_n = \sum_{k=1}^n X_k$. Then $(S_n)_{n\geqslant 0}$ is an L^2-martingale with respect to its natural filtration and with quadratic variation

$$\langle S\rangle_n = \sum_{k=1}^n \mathbb{E}[X_k^2 \,|\, \mathscr{F}_{k-1}^S] = \sum_{k=1}^n \mathbb{E}[X_k^2] = \mathrm{Var}(S_n).$$

In particular, the quadratic variation of S is deterministic because S has independent increments.

(b) For an L^{2q}-martingale M and a predictable process X in L^{2p} with $\frac{1}{p} + \frac{1}{q} = 1$ the martingale transform $X{\bullet}M$ can be checked to be an L^2-martingale. Its quadratic variation is

$$\begin{aligned}\langle X{\bullet}M\rangle_n &= \sum_{k=1}^n \mathbb{E}[X_k^2(M_k - M_{k-1})^2 \,|\, \mathscr{F}_{k-1}^X]\\ &= \sum_{k=1}^n X_k^2(\langle M\rangle_k - \langle M\rangle_{k-1}) = (X^2{\bullet}\langle M\rangle)_n.\end{aligned}$$

So, the quadratic variation of $X{\bullet}M$ is again represented as a discrete stochastic integral with 'integrand' X^2 and 'integrator' $\langle M\rangle$.

4.2 Stopping times

4.15 Definition. A map $\tau : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ is called stopping time (Stoppzeit) with respect to a filtration $(\mathscr{F}_n)$ if $\{\tau = n\} \in \mathscr{F}_n$ holds for all $n \geqslant 0$.

4.16 Remark. A stopping time is a random time index which *cannot look into the future*, as will become clear from the following properties and examples.

4.17 Lemma.

(a) A stopping time is an $([0,\infty], \mathfrak{B}_{[0,\infty]})$*-valued random variable.*

(b) A map $\tau : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ *is a stopping time if and only if* $\{\tau \leqslant n\} \in \mathscr{F}_n$ *for all* $n \geqslant 0$.

(c) Every deterministic time $\tau(\omega) = n_0$ *is a stopping time.*

(d) *For stopping times σ and τ also $\sigma \wedge \tau$, $\sigma \vee \tau$ and $\sigma + \tau$ are stopping times.*

4.18 Remark. The open sets in $[0,\infty]$ are the open sets in $[0,\infty)$ and their unions with sets of the form $(x,\infty]$ for $x \in [0,\infty)$. This defines the natural convergence in $[0,\infty]$, in particular the notion of $a_n \to \infty$. The Borel-σ-algebra $\mathfrak{B}_{[0,\infty]}$ is then generated by these open sets and contains all Borel sets B of $[0,\infty)$ as well as all $B \cup \{\infty\}$.

Proof.

(a) For $m \in \mathbb{N}_0$ we have $\tau^{-1}(\{m\}) = \{\tau = m\} \in \mathscr{F}_m \subseteq \mathscr{F}$ and thus also $\tau^{-1}(\{\infty\}) = \tau^{-1}(\mathbb{N}_0)^{\complement} \in \mathscr{F}$. For any $B \in \mathfrak{B}_{[0,\infty]}$ we conclude $\tau^{-1}(B) = \bigcup_{m \in B \cap (\mathbb{N}_0 \cup \{\infty\})} \tau^{-1}(\{m\}) \in \mathscr{F}$.

(b) If $\{\tau = k\} \in \mathscr{F}_k$ for all $k \geqslant 0$, then $\{\tau \leqslant n\} = \bigcup_{k=0}^n \{\tau = k\} \in \mathscr{F}_n$ for all $n \geqslant 0$. Conversely, if $\{\tau \leqslant k\} \in \mathscr{F}_k$ for all $k \geqslant 0$, then $\{\tau = n\} = \{\tau \leqslant n\} \setminus \{\tau \leqslant n-1\} \in \mathscr{F}_n$ for all $n \geqslant 1$ and $\{\tau \leqslant 0\} = \{\tau = 0\}$ trivially.

(c) We have $\{\tau = n\} = \varnothing \in \mathscr{F}_n$ for all $n \neq n_0$ and also $\{\tau = n_0\} = \Omega \in \mathscr{F}_{n_0}$.

(d) Use part (b) to see $\{\sigma \wedge \tau \leqslant n\} = \{\sigma \leqslant n\} \cup \{\tau \leqslant n\} \in \mathscr{F}_n$ and $\{\sigma \vee \tau \leqslant n\} = \{\sigma \leqslant n\} \cap \{\tau \leqslant n\} \in \mathscr{F}_n$. Moreover, $\{\sigma + \tau = n\} = \bigcup_{k=0}^n \{\sigma = k\} \cap \{\tau = n-k\} \in \mathscr{F}_n$ holds.

□

4.19 Example. Let $(X_n)_{n\geqslant 0}$ be an $(\mathscr{F}_n)$-adapted $(S,\mathscr{S})$-valued process and $B \in \mathscr{S}$. Then the entrance time into B

$$\tau_B := \inf\{n \geqslant 0 \,|\, X_n \in B\} \text{ with } \inf \varnothing := \infty$$

is an $(\mathscr{F}_n)$-stopping time: $\{\tau_B \leqslant n\} = \bigcup_{k=0}^n \{X_k \in B\} \in \mathscr{F}_n$. For $k \in \mathbb{N}$ also $\tau_{B,k} = \inf\{n \geqslant k \,|\, X_{n-k} \in B\} = \tau_B + k$ is a stopping time, but usually $\tau_{B,-k} = \inf\{n \geqslant 0 \,|\, X_{n+k} \in B\}$ is *not* a stopping time.

▷ **Control questions**

(a) Show that a predictable martingale is a.s. constant.

By predacibility $\mathbb{E}[M_{n+1} \,|\, \mathscr{F}_n] = M_{n+1}$ so that the martingale property implies $M_n = M_{n+1}$ a.s.

(b) Why can the quadratic variation be written as $\langle M \rangle_n = \sum_{k=1}^n \operatorname{Var}(M_k \,|\, \mathscr{F}_{k-1})$ in terms of conditional variances (compare exercises)? What is $\operatorname{Var}(M_k \,|\, \mathscr{F}_{k-1})$ for martingales which are sums of independent random variables?

We have $\mathbb{E}[M_k \,|\, \mathscr{F}_{k-1}] = M_{k-1}$ so that $\mathbb{E}[(M_k - M_{k-1})^2 \,|\, \mathscr{F}_{k-1}] = \operatorname{Var}(M_k \,|\, \mathscr{F}_{k-1})$ holds. If $M_k - M_{k-1}$ is independent of $\mathscr{F}_{k-1}$, then this formula gives $\operatorname{Var}(M_k \,|\, \mathscr{F}_{k-1}) = \operatorname{Var}(M_k - M_{k-1})$, compare Example 4.14(a). This is a case where the conditional variance is much smaller than the variance!

(c) What are examples of stopping times in an infinitely long coin tossing experiment where $\mathscr{F}_n$ is generated by the first n coin tosses?

The first time, when head appears, is a stopping time or the first time three times heads has appeared in a 'run', but the time when a run of three heads starts is not a stopping time.

4.20 Theorem (Optional Stopping)**.** *Let (M_n) be a (sub/super-)martingale and τ a stopping time. Then the stopped process $(M_n^\tau) := (M_{\tau\wedge n})$ is again a (sub/super-)martingale.*

Proof. Note first $\mathbb{E}[|M_{\tau\wedge n}|] \leqslant \mathbb{E}[\sum_{k=0}^n |M_k|] < \infty$. Put $C_n := \mathbf{1}(\tau \geqslant n)$, $n \geqslant 1$. Then (C_n) is a bounded predictable process, noting $\{\tau \geqslant n\} = \{\tau \leqslant n-1\}^\complement \in \mathscr{F}_{n-1}$, and satisfies

$$(C\bullet M)_n = \sum_{k=1}^n C_k(M_k - M_{k-1}) = \sum_{k=1}^{\tau\wedge n}(M_k - M_{k-1}) = M_{\tau\wedge n} - M_0.$$

If M is a martingale, then Proposition 4.8 shows that $(C\bullet M)$ is a martingale and thus also M^τ. Since $C_n \geqslant 0$, we can deduce in the submartingale case that $C\bullet M$ is also a submartingale and so is M^τ. For a supermartingale M consider the submartingale $-M$ and conclude. □

4.21 Example (doubling strategy)**.** Consider a fair game where you can win or lose 1 Euro for a stake of 1 Euro in each round. We model this by i.i.d. random variables $(\varepsilon_k)_{k\geqslant 1}$ with $\mathbb{P}(\varepsilon = +1) = \mathbb{P}(\varepsilon = -1) = 1/2$ ((ε_k) is called a *Rademacher sequence*) and the martingale $S_0 := 0$, $S_n = \sum_{k=1}^n \varepsilon_k$, $n \geqslant 1$, with respect to its natural filtration $(\mathscr{F}_n^S)$. In each round we double the stake and invest $C_n = 2^{n-1}$ Euro (which is deterministic, hence predictable). We obtain the martingale $M_n = (C\bullet S)_n$ so that $\mathbb{E}[M_n] = \mathbb{E}[M_0] = 0$. Now we stop playing after the first win, hence consider $\tau = \inf\{k \geqslant 1 \,|\, \varepsilon_k = +1\}$. Before τ we shall have lost $\sum_{k=1}^{\tau-1} 2^{k-1} = 2^{\tau-1} - 1$ Euro, but in round τ we win $2^{\tau-1}$ Euro so that $M_\tau = 1$ holds almost surely (τ is a.s. finite, i.e. $\mathbb{P}(\tau < \infty) = 1$, since it is geometrically distributed). These kind of examples where in a fair game you can still win (and the other lose) with probability one, have attracted a lot of attention and have become known as *Sankt Petersburg paradoxon* (Daniel Bernoulli, 1738). This cannot happen if the number of games is finite or equivalently the stopping time τ is a.s. bounded, i.e. $\mathbb{P}(\tau \leqslant n) = 1$ for some $n \in \mathbb{N}$, because then $\mathbb{E}[M_\tau] = \mathbb{E}[M_{\tau\wedge n}] = \mathbb{E}[M_{\tau\wedge 0}] = \mathbb{E}[M_0] = 0$ by optional stopping. In the sequel, we shall understand M_τ in more detail. Please be aware of the difference between finite and bounded stopping times.

4.22 Definition. For a stopping time τ the σ-algebra of τ-history (τ-Vergangenheit) is defined by $\mathscr{F}_\tau := \{A \in \mathscr{F} \,|\, \forall n \geqslant 0 : \; A \cap \{\tau \leqslant n\} \in \mathscr{F}_n\}$.

4.23 Remark. The definition of $\mathscr{F}_\tau$ is quite implicit, but it encodes events that occur until τ stops►EXERCISE. We just need some key properties.

4.24 Lemma. *$\mathscr{F}_\tau$ is a σ-Algebra and τ is $\mathscr{F}_\tau$-measurable.*

Proof. The axioms of a σ-algebra are checked directly. For all $n \geqslant 0$ the event $\{\tau = m\} \cap \{\tau \leqslant n\}$ is either empty (if $n < m$) or equal to $\{\tau = m\} \in \mathscr{F}_m \subseteq \mathscr{F}_n$ (if $n \geqslant m$). This shows $\{\tau = m\} \in \mathscr{F}_\tau$ and as in Lemma 4.17(a) for $\mathscr{F}$ we conclude that τ is $\mathscr{F}_\tau$-measurable. □

4.25 Lemma. *For stopping times σ and τ with $\sigma \leqslant \tau$ we have $\mathscr{F}_\sigma \subseteq \mathscr{F}_\tau$.*

Proof. For $A \in \mathscr{F}_\sigma$ we have $A \cap \{\sigma \leqslant n\} \in \mathscr{F}_n$ for all $n \geqslant 0$. Since $\{\tau \leqslant n\} \in \mathscr{F}_n$ holds by the stopping time property, we conclude from $\sigma \leqslant \tau$ that

$$A \cap \{\tau \leqslant n\} = \big(A \cap \{\sigma \leqslant n\}\big) \cap \{\tau \leqslant n\} \in \mathscr{F}_n.$$

Hence, $A \in \mathscr{F}_\tau$. □

4.26 Lemma. *For an adapted $(S, \mathscr{S})$-valued process (X_n) and a finite stopping time τ the random variable (!) X_τ is $\mathscr{F}_\tau$-measurable.*

Proof. For any $B \in \mathscr{S}$ we have to check $\{\omega \in \Omega \,|\, X_{\tau(\omega)}(\omega) \in B\} \in \mathscr{F}_\tau$, which is equivalent to

$$\forall n \geqslant 0: \ \{X_\tau \in B\} \cap \{\tau \leqslant n\} \in \mathscr{F}_n.$$

Writing $\{X_\tau \in B\} \cap \{\tau \leqslant n\} = \bigcup_{k=0}^n (\{X_k \in B\} \cap \{\tau = k\})$ we see that $\{X_k \in B\}, \{\tau = k\}$ lie in $\mathscr{F}_k \subseteq \mathscr{F}_n$ for $k \leqslant n$ and therefore $\{X_\tau \in B\} \in \mathscr{F}_\tau$. □

4.27 Remark. Because of $\mathscr{F}_\tau \subseteq \mathscr{F}$ wee see that X_τ is a random variable which is not clear a priori because ω enters at two different places in $\omega \mapsto X_{\tau(\omega)}(\omega)$.

4.28 Theorem (Optional Sampling)**.** *Let (M_n) be a martingale (submartingale) and σ, τ bounded stopping times with $\sigma \leqslant \tau$ (i.e., $\exists R \in \mathbb{N}\, \forall \omega : \sigma(\omega) \leqslant \tau(\omega) \leqslant R$). Then $\mathbb{E}[M_\tau \,|\, \mathscr{F}_\sigma] = M_\sigma$ (resp. $\mathbb{E}[M_\tau \,|\, \mathscr{F}_\sigma] \geqslant M_\sigma$) holds.*

Proof. (martingale case) Because of $\tau \leqslant R$ we see from above $M_\tau = M_{\tau \wedge R} \in L^1$. We have to show $\mathbb{E}[M_\sigma \mathbf{1}_A] = \mathbb{E}[M_\tau \mathbf{1}_A]$ for all $A \in \mathscr{F}_\sigma$. Putting $\rho := \sigma \mathbf{1}_A + \tau \mathbf{1}_{A^\complement}$, we have

$$\{\rho = n\} = (A \cap \{\sigma = n\}) \cup (A^\complement \cap \{\tau = n\}) \in \mathscr{F}_n,$$

using $A^\complement \in \mathscr{F}_\sigma \subseteq \mathscr{F}_\tau$. Hence, ρ is a bounded stopping time as well and

$$\mathbb{E}[M_\rho] = \mathbb{E}[M_{\rho \wedge R}] = \mathbb{E}[M_{\rho \wedge 0}] = \mathbb{E}[M_0] = \mathbb{E}[M_{\tau \wedge R}] = \mathbb{E}[M_\tau]$$

by optional stopping implies $\mathbb{E}[M_\rho - M_\tau] = 0$. By definition of ρ, we have $M_\rho - M_\tau = (M_\sigma - M_\tau)\mathbf{1}_A$ and $\mathbb{E}[M_\sigma \mathbf{1}_A] = \mathbb{E}[M_\tau \mathbf{1}_A]$ follows. □

4.29 Proposition. *Let (M_n) be a martingale and τ a finite stopping time. Then $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$ holds under one of the following conditions:*

(a) τ is bounded;

(b) $(M_{\tau \wedge n})_{n \geqslant 0}$ is dominated ($|M_{\tau \wedge n}| \leqslant Y$ for all n and some $Y \in L^1$);

(c) $\mathbb{E}[\tau] < \infty$ and $(\mathbb{E}[|M_{n+1} - M_n| \,|\, \mathscr{F}_n])_{n \geqslant 0}$ is uniformly bounded.

Proof.

(a) There is some $R \in \mathbb{N}$ with $\tau \leqslant R$. Optional stopping therefore yields $\mathbb{E}[M_\tau] = \mathbb{E}[M_{\tau\wedge R}] = \mathbb{E}[M_0]$.

(b) We have $M_{\tau\wedge n} \to M_\tau$ as $n \to \infty$ and dominated convergence together with (a) implies $\mathbb{E}[M_\tau] = \lim_{n\to\infty} \mathbb{E}[M_{\tau\wedge n}] = \lim_{n\to\infty} \mathbb{E}[M_0] = \mathbb{E}[M_0]$.

(c) By a telescoping sum expansion we obtain

$$|M_{\tau\wedge n} - M_0| \leqslant \sum_{k=1}^{\tau\wedge n} |M_k - M_{k-1}| \leqslant \sum_{k=1}^{\infty} |M_k - M_{k-1}| \mathbf{1}(k \leqslant \tau) =: Z.$$

By assumption there is $R \in \mathbb{N}$ with $\mathbb{E}[|M_k - M_{k-1}|\mathbf{1}(k \leqslant \tau) \,|\, \mathscr{F}_{k-1}] \leqslant R\mathbf{1}(k \leqslant \tau)$ for all k, noting $\{k \leqslant \tau\} \in \mathscr{F}_{k-1}$. We therefore deduce

$$\mathbb{E}[Z] = \mathbb{E}\Big[\sum_{k=1}^{\infty} \mathbb{E}\left[|M_k - M_{k-1}|\mathbf{1}(k \leqslant \tau) \,\middle|\, \mathscr{F}_{k-1}\right]\Big] \leqslant \mathbb{E}\Big[\sum_{k=1}^{\infty} R\mathbf{1}(k \leqslant \tau)\Big].$$

The right-hand side equals $R\,\mathbb{E}[\tau]$, which is finite by assumption. With $Y = Z + |M_0| \in L^1$ we thus have $|M_{\tau\wedge n}| \leqslant Y$ for all n and applying part (b) yields the assertion.

□

4.30 Example. If we consider a uniformly bounded predictable 'betting strategy' (C_n) in the martingale $M_n = (C\bullet S)_n$ of Example 4.21 instead of the unbounded $C_n = 2^{n-1}$, then

$$\mathbb{E}[|M_{n+1} - M_n| \,|\, \mathscr{F}_n^S] = C_{n+1}\,\mathbb{E}[|\varepsilon_{n+1}| \,|\, \mathscr{F}_n^S] \leqslant C_{n+1}$$

holds (assuming $C_n \geqslant 0$) and part (c) of the proposition implies $\mathbb{E}[M_\tau] = 0$ for any stopping time τ with finite expectation. The doubling strategy above therefore also profits from the potentially infinite capital the gambler can spend.

4.31 Corollary (Wald's Identity). *Let $(X_k)_{k\geqslant 1}$ be $(\mathscr{F}_k)$-adapted random variables such that $\sup_k \mathbb{E}[|X_k|] < \infty$, $\mathbb{E}[X_k] = \mu \in \mathbb{R}$ and X_k is independent of $\mathscr{F}_{k-1}$, $k \geqslant 1$. Then for $S_n := \sum_{k=1}^n X_k$, $S_0 = 0$ and every $(\mathscr{F}_k)$-stopping time τ with $\mathbb{E}[\tau] < \infty$ we have $\mathbb{E}[S_\tau] = \mathbb{E}[\tau]\mu$.*

4.32 Remark. If additionally all X_k are in L^2 and $\mathbb{E}[X_k] = 0$, $\mathrm{Var}(X_k) = \sigma^2$, $k \geqslant 1$, holds, then also the second Wald identity is valid: $\mathrm{Var}(S_\tau) = \mathbb{E}[\tau]\sigma^2$; see Bauer, Satz 17.7.

Proof. $M_n = S_n - n\mu$, $n \geqslant 0$, forms an $(\mathscr{F}_n)$-martingale due to $\mathbb{E}[M_{n+1} - M_n \,|\, \mathscr{F}_n] = \mathbb{E}[X_{n+1} - \mu] = 0$. Moreover,

$$\mathbb{E}[|M_{n+1} - M_n| \,|\, \mathscr{F}_n] = \mathbb{E}[|X_{n+1} - \mu|] \leqslant \sup_{k\geqslant 1} \mathbb{E}[|X_k|] + |\mu|$$

holds such that by Proposition 4.29(c) $\mathbb{E}[M_\tau] = 0$, hence $\mathbb{E}[S_\tau] = \mathbb{E}[\tau]\mu$. □

4.33 Example.

(a) Let (X_k) be i.i.d., $X_k \in L^1$, and τ be an $(\mathscr{F}_n^X)$-stopping time. Then Wald's identity applies and gives $\mathbb{E}[S_\tau] = \mathbb{E}[\tau]\,\mathbb{E}[X_1]$. In the Rademacher case $\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = 1/2$ one can show that $\tau = \inf\{n \geqslant 0 \,|\, S_n = 1\}$ is an almost surely finite stopping time ►EXERCISE. Then $S_\tau = 1$ a.s., but $\mathbb{E}[X_1] = 0$ such that we conclude $\mathbb{E}[\tau] = \infty$.

(b) Let (X_k) be i.i.d., $X_k \in L^1$, and τ be a random time independent of (X_k). Then for the filtration $\mathscr{F}_n = \sigma(\tau, X_1, \dots, X_n)$ τ is an $(\mathscr{F}_n)$-stopping time, Wald's identity applies and gives $\mathbb{E}[S_\tau] = \mathbb{E}[\tau]\,\mathbb{E}[X_1]$ (compare Example 3.20 with the compound Poisson process). This is an example of a filtration which is not natural for X.

4.34 Example (random walk)**.** Let $(X_k)_{k \geqslant 1}$ be independent with $\mathbb{P}(X_k = 1) = p$, $\mathbb{P}(X_k = -1) = q = 1 - p$ for $p \in (0,1)$ and define $S_0 = 0$, $S_n = \sum_{k=1}^n X_k$, $n \geqslant 1$. Then $(S_n)_{n \geqslant 0}$ defines a simple random walk. Let $a < 0 < b$ with $a, b \in \mathbb{Z}$ be given and consider the $(\mathscr{F}_n^X)$-stopping time

$$\tau = \inf\{n \geqslant 0 \,|\, S_n \in \{a, b\}\}.$$

Let $R = |a| + b$ and observe that $(S_{mR} - S_{(m-1)R})_{m \geqslant 1}$ are i.i.d. and $\mathbb{P}(S_{mR} - S_{(m-1)R} = R) > 0$. Hence, the time $\sigma = \inf\{m \geqslant 1 \,|\, S_{mR} - S_{(m-1)R} = R\}$ is geometrically distributed and has finite expectation. From $\tau \leqslant R\sigma$ we conclude $\mathbb{E}[\tau] < \infty$.

Consider the asymmetric case $p \neq 1/2$ first. Then $((q/p)^{S_n})_{n \geqslant 0}$ forms an $(\mathscr{F}_n^X)$-martingale due to

$$\mathbb{E}[(q/p)^{S_{n+1}} \,|\, \mathscr{F}_n^X] = (q/p)^{S_n}\, \mathbb{E}[(q/p)^{X_{n+1}}] = (q/p)^{S_n}(p(q/p) + q(p/q)) = (q/p)^{S_n}$$

(this is a typical example of a transformed process forming a martingale!). Because of $|S_{\tau \wedge n}| \leqslant |a| \vee b$ the $S_{\tau \wedge n}$, $n \geqslant 1$, are dominated and by Proposition 4.29(b) we infer $\mathbb{E}[(q/p)^{S_\tau}] = \mathbb{E}[(q/p)^{S_0}] = 1$. This gives the two equations

$$\mathbb{P}(S_\tau = a)(q/p)^a + \mathbb{P}(S_\tau = b)(q/p)^b = 1, \quad \mathbb{P}(S_\tau = a) + \mathbb{P}(S_\tau = b) = 1.$$

We can thus calculate the probability whether S_n first hits a or b:

$$\mathbb{P}(S_\tau = a) = \frac{(q/p)^b - 1}{(q/p)^b - (q/p)^a}, \quad \mathbb{P}(S_\tau = b) = \frac{1 - (q/p)^a}{(q/p)^b - (q/p)^a}.$$

Using $\mathbb{E}[\tau] < \infty$, Wald's identity gives $\mathbb{E}[S_\tau] = (p - q)\,\mathbb{E}[\tau]$ and we can solve for $\mathbb{E}[\tau]$:

$$\mathbb{E}[\tau] = \frac{a((q/p)^b - 1) + b(1 - (q/p)^a)}{(p - q)((q/p)^b - (q/p)^a)}.$$

As a special case note that for $p > q$ and $a \downarrow -\infty$ we find $\mathbb{E}[\tau] \uparrow \frac{b}{p-q}$. Using a monotone convergence argument, this is the expectation for the one-sided stopping time $\inf\{n \geqslant 0 \,|\, X_n = b\}$.►CONTROL

For the symmetric simple random walk with $p = q = 1/2$ Wald's identity yields directly $\mathbb{E}[S_\tau] = 0$ and thus $\mathbb{P}(S_\tau = a) = \frac{b}{|a|+b}$, $\mathbb{P}(S_\tau = b) = \frac{|a|}{|a|+b}$. The second Wald identity then gives $\operatorname{Var}(S_\tau) = \mathbb{E}[\tau]$ and thus $\mathbb{E}[\tau] = |a|b$. Note that here $\mathbb{E}[\tau] \uparrow \infty$ as $a \downarrow -\infty$.

▷ **Control questions**

(a) Consider the infinite Bernoulli experiment $(\Omega, \mathscr{F}, \mathbb{P})$ with $\Omega = \{0,1\}^{\mathbb{N}}$, $\mathscr{F} = (\mathcal{P}(\{0,1\}))^{\otimes \mathbb{N}}$ and the natural filtration $(\mathscr{F}_n^X)$ of $X_k(\omega) = \omega_k$, $k \geqslant 1$ (kth coin toss). Find elementary events $\{\omega\}$ for $\omega \in \Omega$ which do and which do not lie in $\mathscr{F}_\tau$ for $\tau = \inf\{n \geqslant 1 \,|\, X_n = 1\}$ (first time of success).

We have $\{\omega\} \cap \{\tau = n\} = \{\omega\}$ if $\omega_1 = \cdots = \omega_{n-1} = 0$, $\omega_n = 1$, otherwise the intersection is empty. Since no elementary event $\{\omega\}$ lies in $\mathscr{F}_n$ (events in $\mathscr{F}_n$ are determined by the first n coordinates only), we conclude that $\{\omega\} \in \mathscr{F}_\tau$ holds if and only if $\omega = \{0, \dots, \}$ is the null sequence.

(b) How can the investment in the doubling strategy example be modified to obtain even infinite gains on average, i.e. $\mathbb{E}[M_\tau] = \infty$?

Since τ is geometrically distributed (with parameter $1/2$), We have $M_\tau = C_\tau - \sum_{k=1}^{\tau-1} C_k$ and $\mathbb{E}[M_\tau] = \sum_{n \geqslant 1} 2^{-n}(C_n - \sum_{k=1}^{n-1} C_k)$. For $C_n = 3^n$ this yields $\mathbb{E}[M_\tau] \geqslant \sum_{n \geqslant 1} 2^{-n}(3^n - 3^n/2) = \infty$ and we win arbitrary much on average.

(c) How is the precise argument to obtain $\mathbb{E}[\inf\{n \geqslant 0 \,|\, X_n = b\}] = \frac{b}{p-q}$ in the preceding example?

(d) Writing $\tau_{a,b} = \inf\{n \geqslant 0 \,|\, S_n \in \{a,b\}\}$, we have $\tau_{a,b} \uparrow \tau_b = \inf\{n \geqslant 0 \,|\, S_n = b\} \in [1, \infty]$ as $a \to -\infty$. By monotone convergence therefore $\mathbb{E}[\tau_{a,b}] \uparrow \mathbb{E}[\tau_b]$ follows. For $p > q$ the terms $(q/p)^a$ dominate the explicit formula for $a \to -\infty$ which gives $\mathbb{E}[\tau_{a,b}] \uparrow b/(p-q)$, hence $\mathbb{E}[\tau_b] = b/(p-q)$.

4.3 Martingale inequalities and convergence

4.35 Remark. One of the key problems in stochastrics is to control the maximum over random variables. For martingales there are very tight bounds.

4.36 Proposition (Maximal inequality)**.** *Any submartingale (M_n) satisfies*

$$\forall \alpha > 0,\, n \geqslant 0 : \, \mathbb{P}\Big(\max_{0 \leqslant k \leqslant n} M_k \geqslant \alpha\Big) \leqslant \frac{1}{\alpha}\, \mathbb{E}\Big[M_n \mathbf{1}\Big(\max_{0 \leqslant k \leqslant n} M_k \geqslant \alpha\Big)\Big].$$

In particular, any martingale (M_n) satisfies

$$\forall \alpha > 0,\, n \geqslant 0 : \, \mathbb{P}\Big(\max_{0 \leqslant k \leqslant n} |M_k| \geqslant \alpha\Big) \leqslant \tfrac{1}{\alpha}\, \mathbb{E}[|M_n|].$$

Proof. Put $\tau := \inf\{k \geqslant 0 \,|\, M_k \geqslant \alpha\}$. Then τ is a stopping time with $\{\max_{0 \leqslant k \leqslant n} M_k \geqslant \alpha\} = \{\tau \leqslant n\}$. We obtain:

$$\begin{aligned}\mathbb{P}\Big(\max_{0 \leqslant k \leqslant n} M_k \geqslant \alpha\Big) &= \mathbb{E}[\mathbf{1}(\tau \leqslant n)] \leqslant \mathbb{E}\Big[\frac{M_{\tau \wedge n}}{\alpha} \mathbf{1}(\tau \leqslant n)\Big] \qquad (4.1)\\ &\leqslant \frac{1}{\alpha}\, \mathbb{E}[M_n \mathbf{1}(\tau \leqslant n)] = \frac{1}{\alpha}\, \mathbb{E}\Big[M_n \mathbf{1}\Big(\max_{0 \leqslant k \leqslant n} M_k \geqslant \alpha\Big)\Big],\end{aligned}$$

where the last bound follows by optional sampling▶CONTROL of the submartingale $(M_n)_{n \geqslant 0}$ at times $\tau \wedge n$ and n. This gives the inequality for submartingales.

For martingales (M_n) apply the maximal inequality to the submartingale $(|M|_n)$ ($\varphi(x) = |x|$) is convex) and bound the indicator bei one. □

4.37 Remark. Markov's inequality (Stochastik I) just gives $\mathbb{P}(|M_n| \geqslant \alpha) \leqslant \frac{1}{\alpha}\,\mathbb{E}[|M_n|]$ so that the maximal inequality bounds an a priori much larger event by the same value.

4.38 Theorem (Doob's L^p-inequality)**.** *Any L^p-martingale or positive L^p-submartingale (M_n) (i.e. $M_n \in L^p$ for all n) with $p > 1$ satisfies*

$$\Big\| \max_{0\leqslant k\leqslant n} |M_k| \Big\|_{L^p} \leqslant \frac{p}{p-1} \|M_n\|_{L^p}.$$

Proof. Write $M_n^* = \max_{0\leqslant k\leqslant n}|M_k|$ and note that $(|M_n|)$ is a submartingale under our assumptions. Then by the maximal inequality for $(|M_n|)$ of Proposition 4.36 and by Tonelli-Fubini's theorem we obtain

$$\begin{aligned} \tfrac{1}{p}\,\mathbb{E}[(M_n^*)^p] &= \mathbb{E}\Big[\int_0^{M_n^*} x^{p-1} dx\Big] = \int_0^\infty x^{p-1}\,\mathbb{P}(M_n^* \geqslant x)\,dx \\ &\leqslant \int_0^\infty x^{p-2}\,\mathbb{E}[|M_n|\mathbf{1}(M_n^* \geqslant x)]\,dx \\ &= \mathbb{E}\Big[|M_n| \int_0^{M_n^*} x^{p-2}\,dx\Big] = \tfrac{1}{p-1}\,\mathbb{E}[|M_n|(M_n^*)^{p-1}]. \end{aligned}$$

By Hölder inequality for $p^{-1} + q^{-1} = 1 \iff q = p/(p-1)$ we thus have

$$\mathbb{E}[(M_n^*)^p] \leqslant \tfrac{p}{p-1}\,\mathbb{E}[|M_n|(M_n^*)^{p-1}] \leqslant \tfrac{p}{p-1}\,\mathbb{E}[|M_n|^p]^{1/p}\,\mathbb{E}[(M_n^*)^p]^{(p-1)/p}.$$

Dividing by $\mathbb{E}[(M_n^*)^p]^{(p-1)/p}$, the assertion follows. □

4.39 Example. An analogue of Doob's L^p inequality cannot hold in the case $p = 1$. On $([0,1], \mathfrak{B}_{[0,1]}, \lambda)$ (with Lebesgue measure λ) consider $M_n = 2^n\mathbf{1}_{[0,2^{-n})}$. Then (M_n) is a non-negative martingale with respect to its natural filtration $(\mathcal{F}_n^M)$▶Control. We have $\|M_n\|_{L^1} = 1$ for all n, but (sketch the function!) $\|\max_{0\leqslant k\leqslant n} M_k\|_{L^1} = 2^n 2^{-n} + \sum_{k=0}^{n-1} 2^k 2^{-k-1} = 1 + n/2 \to \infty$ as $n \to \infty$.

4.40 Definition. The number of upcrossings (aufsteigende Überquerungen) on an interval $[a,b]$ by a process (X_k) until time n is defined by $U_n^{[a,b]} := \sup\{k \geqslant 1 \,|\, \tau_k \leqslant n\}$, where inductively $\tau_0 := 0$, $\sigma_{k+1} := \inf\{\ell \geqslant \tau_k \,|\, X_\ell \leqslant a\}$, $\tau_{k+1} := \inf\{\ell \geqslant \sigma_{k+1} \,|\, X_\ell \geqslant b\}$.

4.41 Remark. Convince yourself that σ_1 is the first time the process (X_k) has been below a, τ_1 is the first time after σ_1 it is above b etc. So, $U_n^{[a,b]}$ really counts the number of times the process has moved from below a to above b. We shall see that (sub)martingales only allow for a small number of upcrossings $U_n^{[a,b]}$ and that this will yield an almost sure convergence result.

4.42 Proposition (Upcrossing Inequality)**.** *The upcrossings of a submartingale (X_n) satisfy $\mathbb{E}[U_n^{[a,b]}] \leqslant \frac{1}{b-a}\,\mathbb{E}[(X_n - a)_+]$.*

Proof. We shall assume in the sequel $a = 0$ and $X_n \geqslant 0$. This can be done without loss of generality because otherwise it suffices to consider $Y_n = (X_n -$

$a)_+$, $n \geqslant 0$, which forms by Jensen's inequality also a submartingale with $Y_n \geqslant 0$ and whose upcrossings on $[0, b-a]$ equals $U_n^{[a,b]}$ (upcrossings of X on $[a,b]$).

By definition of the upcrossing stopping times a telescoping sum yields

$$\mathbb{E}[X_n] = \mathbb{E}[X_{\sigma_1 \wedge n}] + \sum_{k=1}^{n} \mathbb{E}[X_{\tau_k \wedge n} - X_{\sigma_k \wedge n}] + \sum_{k=1}^{n} \mathbb{E}[X_{\sigma_{k+1} \wedge n} - X_{\tau_k \wedge n}].$$

Since (X_n) is a non-negative submartingale by assumption, optional sampling shows that all summands in this decomposition are non-negative. From

$$\sum_{k=1}^{n} (X_{\tau_k \wedge n} - X_{\sigma_k \wedge n}) = \sum_{k=1}^{U_n^{[0,b]}} (X_{\tau_k} - X_{\sigma_k}) \geqslant b U_n^{[0,b]}$$

we thus infer $\mathbb{E}[X_n] \geqslant b\, \mathbb{E}[U_n^{[0,b]}]$, the upcrossing inequality for $a = 0$. □

4.43 Theorem (First martingale convergence theorem)**.** *Let (M_n) be a (sub-/super-)martingale with $\sup_n \mathbb{E}[|M_n|] < \infty$ ((M_n) is L^1-bounded). Then $M_\infty := \lim_{n\to\infty} M_n$ exists a.s. and M_∞ is in L^1.*

4.44 Remark. If (M_n) is a submartingale, it is L^1-bounded already if $\sup_n \mathbb{E}[(M_n)_+]$ is finite because

$$\mathbb{E}[(M_n)_-] = \mathbb{E}[(M_n)_+] - \mathbb{E}[M_n] \leqslant \mathbb{E}[(M_n)_+] - \mathbb{E}[M_0]$$

holds. Let us also emphasize that (M_n) need not converge in L^1 to M_∞.

Proof. Let (M_n) be a submartingale. By monotonicity $U_n^{[a,b]}$ converges to some $U^{[a,b]} \in \mathbb{N}_0 \cup \{\infty\}$ as $n \to \infty$ and by monotone convergence and the upcrossing inequality

$$\mathbb{E}[U^{[a,b]}] \leqslant \frac{1}{b-a} \lim_{n\to\infty} \mathbb{E}[(M_n - a)_+] \leqslant \frac{1}{b-a} \Big(\sup_n \mathbb{E}[|M_n|] + |a| \Big) < \infty.$$

This shows $\mathbb{P}(U^{[a,b]} = \infty) = 0$. For $a < b$ and the event

$$\Lambda_{a,b} = \Big\{ \limsup_{n\to\infty} M_n \geqslant b,\, \liminf_{n\to\infty} M_n \leqslant a \Big\}$$

this implies $\mathbb{P}(\Lambda_{a,b}) = 0$ and thus

$$\mathbb{P}\Big(\limsup_{n\to\infty} M_n > \liminf_{n\to\infty} M_n \Big) = \mathbb{P}\Big(\bigcup_{a<b, a,b\in\mathbb{Q}} \Lambda_{a,b} \Big) = 0.$$

Hence, (M_n) converges $\mathbb{P}$-almost surely to some M_∞ with values in $\mathbb{R} \cup \{\pm\infty\}$. Fatou's Lemma gives

$$\mathbb{E}[|M_\infty|] = \mathbb{E}\Big[\liminf_{n\to\infty} |M_n| \Big] \leqslant \liminf_{n\to\infty} \mathbb{E}[|M_n|] < \infty.$$

This shows that M_∞ is a.s. finite and in L^1.

For supermartingales (M_n) apply the result to the submartingales $(-M_n)$. □

4.45 Corollary. *Each non-negative (super)martingale (M_n) converges $\mathbb{P}$-a.s. to some M_∞ with $\mathbb{E}[M_\infty] \leqslant \lim_{n\to\infty} \mathbb{E}[M_n] = \inf_n \mathbb{E}[M_n]$.*

Proof. The decay $\mathbb{E}[M_{n+1}] \leqslant \mathbb{E}[M_n]$ and $M_n \geqslant 0$ show $\sup_n \mathbb{E}[|M_n|] = \mathbb{E}[M_0] < \infty$. By the first martingale convergence theorem we obtain $M_n \to M_\infty$ $\mathbb{P}$-a.s. By Fatou's Lemma

$$\mathbb{E}[M_\infty] \leqslant \liminf_{n\to\infty} \mathbb{E}[M_n] = \lim_{n\to\infty} \mathbb{E}[M_n] = \inf_n \mathbb{E}[M_n]$$

follows. □

4.46 Example (A fair game where you lose in the long run)**.** Consider the multiplicative martingale $P_0 := 1$, $P_n = \prod_{i=1}^n X_i$, $n \geqslant 1$, with independent random variables (X_i), satisfying $\mathbb{P}(X_i = 3/2) = \mathbb{P}(X_i = 1/2) = 1/2$. The corollary ensures that (P_n) converges a.s. The strong law of large numbers applied to $\frac{1}{n}\log(P_n)$ shows more precisely $P_n \to P_\infty = 0$ $\mathbb{P}$-a.s. (Stochastik I). Yet, $\mathbb{E}[P_n] = 1$ holds for all $n \geqslant 0$ and thus $P_n \to P_\infty$ does not hold in L^1.

4.47 Definition. A family $(X_i)_{i\in I}$ of random variables is uniformly integrable (gleichgradig integrierbar) if

$$\lim_{R\to\infty} \sup_{i\in I} \mathbb{E}[|X_i| \mathbf{1}_{\{|X_i|>R\}}] = 0.$$

4.48 Lemma.

(a) $(X_i)_{i\in I}$ is uniformly integrable if and only if $(X_i)_{i\in I}$ is L^1-bounded and $\forall\, \varepsilon > 0\; \exists\, \delta > 0 :\; \mathbb{P}(A) < \delta \Rightarrow \sup_{i\in I} \mathbb{E}[|X_i| \mathbf{1}_A] < \varepsilon$.

(b) If $(X_i)_{i\in I}$ is L^p-bounded ($\sup_{i\in I} \mathbb{E}[|X_i|^p] < \infty$) for some $p > 1$, then $(X_i)_{i\in I}$ is uniformly integrable.

(c) If $|X_i| \leqslant Y$ holds for all $i \in I$ and some $Y \in L^1$ ($(X_i)_{i\in I}$ is dominated), then $(X_i)_{i\in I}$ is uniformly integrable.

Proof. For (a) see ▶Exercise. For (b) note

$$\mathbb{E}[|X_i| \mathbf{1}_{\{|X_i|>R\}}] \leqslant \mathbb{E}[|X_i| (|X_i|/R)^{p-1}] = R^{-(p-1)} \|X_i\|_{L^p}^p$$

where the right-hand side tends to zero for $R \to \infty$, uniformly over i. Part (c) follows directly from $|X_i| \mathbf{1}_{\{|X_i|>R\}} \leqslant Y \mathbf{1}_{\{Y>R\}}$ and dominated convergence. □

▷ **Control questions**

(a) How is the optional sampling theorem used precisely in the proof of the maximal inequality?

We have to show $\mathbb{E}[M_{\tau\wedge n} \mathbf{1}(\tau \leqslant n)] \leqslant \mathbb{E}[M_n \mathbf{1}(\tau \leqslant n)]$ for a submartingale (M_n). Adding $\mathbb{E}[M_n \mathbf{1}(\tau > n)]$ on both sides, this is equivalent to $\mathbb{E}[M_{\tau\wedge n}] \leqslant \mathbb{E}[M_n]$. Since $\tau \wedge n$ and n are bounded stopping times, optional sampling gives $M_{\tau\wedge n} \leqslant \mathbb{E}[M_n \,|\, \mathscr{F}_{\tau\wedge n}]$ and it remains to take expectations of both sides.

(b) Show formally that (M_n) in Example 4.39 is a martingale.

We claim $\mathbb{E}[M_{n+1}\,|\,\mathscr{F}_n^M] = M_n$. $\mathscr{F}_n^M$ is generated by the intervals $[0,2^{-n}), [2^{-n}, 2^{-n+1}), \ldots, [1/2, 1]$ so that $\mathbb{E}[M_{n+1}\,|\,\mathscr{F}_n^M] = c\mathbf{1}_{[0,2^{-n})}$ must hold a.s. for some $c \in \mathbb{R}$ (M_{n+1} vanishes on the complement). Taking expectations, we see $1 = \mathbb{E}[M_{n+1}] = c2^{-n}$ so that $c = 2^n$ and the claim holds.

(c) Find an example of uniformly integrable $(X_i)_{i\geqslant 1}$ where $Y := \sup_i |X_i|$ is not in L^1, thus showing that condition (c) in the lemma is only sufficient, not necessary.
Hint: Example 4.39 with a smaller factor.

Consider $M_n = (n+1)^{-1}2^n\mathbf{1}_{[0,2^{-n})}$, $n \geqslant 0$, for $q \in (1,2)$ in Example 4.39. Then $\mathbb{E}[|M_n|] = \mathbb{E}[M_n] = (n+1)^{-1} \to 0$ and $(M_n)_{n\geqslant 1}$ is uniformly integrable. On the other hand, we have $\sup_{n\geqslant 1}|M_n| = \sum_{n=1}^{\infty}\sum_{k=0}^{n-1}(k+1)^{-1}2^k\mathbf{1}_{[2^{-n},2^{-n+1})}$ a.e. and because of $\sum_{k=0}^{n-1}(k+1)^{-1}2^k \geqslant n^{-1}(2^n-1)$ we obtain $\mathbb{E}[\sup_{n\geqslant 1}|M_n|] \geqslant \sum_{n=1}^{\infty} n^{-1}(1-2^{-n}) = \infty$.

4.49 Remark. From Stochastik I and measure theory we know the dominated convergence theorem which ensures convergence of expected values or integrals. The domination condition is quite strong, e.g. often it does not apply for proving that sums S_n of independent random variables with $S_n \to S_\infty$ a.s. (or in probability) satisfy $\mathbb{E}[S_n] \to \mathbb{E}[S_\infty]$ (when it is true). In view of the first martingale convergence theorem we know that L^1-boundedness yields the almost sure-convergence and it turns out that uniform integrability is a necessary and sufficient condition to guarantee also L^1-convergence, hence convergence of expected values. In functional analysis, the Dunford-Pettis Theorem asserts more generally that a family $(X_i)_{i\in I}$ is weakly relatively compact in $L^1(\mathbb{P})$ if and only if it is uniformly integrable.

4.50 Theorem (Vitali, 1907). *Let $(X_n)_{n\geqslant 0}$ be random variables in $L^1(\mathbb{P})$ with $X_n \xrightarrow{\mathbb{P}} X$ (in probability). Then the following statements are equivalent:*

(a) $(X_n)_{n\geqslant 0}$ *is uniformly integrable;*

(b) $X_n \to X$ *in* $L^1(\mathbb{P})$*;*

(c) $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|] < \infty$.

Proof. To show (a)⇒(b), we can assume w.l.o.g. that $X_n \to X$ $\mathbb{P}$-a.s. by the classical subsubsequence argument: if $X_n \xrightarrow{\mathbb{P}} X$, but (X_n) did not converge to X in L^1, then there would be a subsequence (n_k) and $\varepsilon > 0$ such that $\|X_{n_k} - X\|_{L^1} \geqslant \varepsilon$ for all k and by Stochastik I a subsubsequence (n_{k_l}) such that $X_{n_{k_l}} \to X$ $\mathbb{P}$-a.s., for which, however, we now prove $X_{n_{k_l}} \to X$ in L^1, a contradiction to $\|X_{n_{k_l}} - X\|_{L^1} \geqslant \varepsilon$.

Since (X_n) is L^1-bounded, Fatou's Lemma shows $\mathbb{E}[|X|] \leqslant \liminf_{n\to\infty}\mathbb{E}[|X_n|] < \infty$. For $\varepsilon > 0$ we can choose by uniform integrability some $R > 0$ with

$$\sup_n \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] + \mathbb{E}[|X|\mathbf{1}(|X| > R)] < \frac{\varepsilon}{2}.$$

Put $\varphi_R(x) = (-R) \vee (x \wedge R)$ (clipping by $-R$ and R). By dominated convergence (use $\|\varphi_R\|_\infty \leqslant R < \infty$ and $|\varphi_R(X_n) - \varphi_R(X)| \to 0$ a.s.) there is $n_0 \in \mathbb{N}$ with $\mathbb{E}[|\varphi_R(X_n) - \varphi_R(X)|] < \varepsilon/2$ for all $n \geqslant n_0$. Consequently,

$$\mathbb{E}[|X_n - X|] \leqslant \mathbb{E}[|\varphi_R(X_n) - \varphi_R(X)|] + \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] + \mathbb{E}[|X|\mathbf{1}(|X| > R)] < \varepsilon$$

holds for all $n \geqslant n_0$. Since $\varepsilon > 0$ was arbitrary, (b) follows.

The implication (b)⇒(c) follows immediately by the continuity of the L^1-norm via $||X_n| - |X|| \leqslant |X_n - X|$.

For (c)⇒(a) put $\psi_R(x) = |x|$ for $|x| \leqslant R - 1$, $\psi_R(x) = 0$ for $|x| \geqslant R$ and interpolate linearly on $[-R, -R+1]$ and $[R-1, R]$. Then ψ_R is continuous and satisfies $\psi_R(x) \leqslant |x|\mathbf{1}(|x| \leqslant R)$ such that for any n, R

$$\mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] \leqslant \mathbb{E}[|X_n| - \psi_R(X_n)].$$

Since ψ_R is bounded and continuous and $X_n \to X$ in distribution (Stochastik I), we have $\mathbb{E}[\psi_R(X_n)] \to \mathbb{E}[\psi_R(X)]$. Letting first $n \to \infty$ and then $R \to \infty$ we thus have

$$\lim_{R\to\infty} \limsup_{n\to\infty} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] \leqslant \lim_{R\to\infty} \mathbb{E}[|X| - \psi_R(X)] = 0.$$

Using monotonicity, we obtain uniform integrability:

$$\begin{aligned}
&\lim_{R\to\infty} \sup_{n\geqslant 1} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] \\
&\leqslant \inf_{R>0} \inf_{n_0\geqslant 1} \Big(\sum_{n=1}^{n_0} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] + \sup_{n>n_0} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] \Big) \\
&= \inf_{n_0\geqslant 1} \lim_{R\to\infty} \Big(\sum_{n=1}^{n_0} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] + \sup_{n>n_0} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] \Big) \\
&= \inf_{n_0\geqslant 1} \inf_{R>0} \sup_{n>n_0} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] \\
&= \inf_{R>0} \inf_{n_0\geqslant 1} \sup_{n>n_0} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] \\
&= \inf_{R>0} \limsup_{n\to\infty} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] = 0,
\end{aligned}$$

where $\lim_{R\to\infty} \mathbb{E}[|X_n|\mathbf{1}(|X_n| > R)] = 0$ was used (dominated convergence because of $X_n \in L^1$). □

4.51 Theorem (Second martingale convergence theorem)**.**

(a) If (M_n) is a uniformly integrable martingale, then (M_n) converges a.s. and in L^1 to some $M_\infty \in L^1$. (M_n) is closable with $M_n = \mathbb{E}[M_\infty \,|\, \mathscr{F}_n]$.

(b) If (M_n) is a closable martingale, with $M_n = \mathbb{E}[M \,|\, \mathscr{F}_n]$ say, then (M_n) is uniformly integrable and (a) holds with $M_\infty = \mathbb{E}[M \,|\, \mathcal{F}_\infty]$ where $\mathscr{F}_\infty = \sigma(\mathscr{F}_n, n \geqslant 1)$.

Proof.

(a) The first martingale convergence theorem in view of L^1-boundedness (Lemma 4.48(a)) and Vitali's theorem ensure almost sure and then L^1-convergence. Furthermore, for any $m \geqslant n \geqslant 1$ and $A \in \mathscr{F}_n$ we have $\mathbb{E}[M_n \mathbf{1}_A] = \mathbb{E}[M_m \mathbf{1}_A]$. L^1-convergence implies

$$|\mathbb{E}[M_m \mathbf{1}_A - M_\infty \mathbf{1}_A]| \leqslant \|M_m - M_\infty\|_{L^1} \to 0 \text{ as } m \to \infty,$$

hence $\mathbb{E}[M_\infty \mathbf{1}_A] = \mathbb{E}[M_n \mathbf{1}_A]$. This shows $M_n = \mathbb{E}[M_\infty \,|\, \mathscr{F}_n]$. In fact, we have shown the general fact that L^1-convergence implies convergence of the conditional expectations.

(b) Assume first $M \geqslant 0$ with $\mathbb{E}[M] > 0$. Then $M_n = \mathbb{E}[M \,|\, \mathscr{F}_n] \geqslant 0$ holds a.s and $M_\infty = \lim_{n\to\infty} M_n$ exists a.s. with $\mathbb{E}[M_\infty] \leqslant \lim_{n\to\infty} \mathbb{E}[M_n] = \mathbb{E}[M]$ (Corollary 4.45). On the other hand, dominated convergence and $\mathbb{E}[M \,|\, \mathscr{F}_n] \wedge R \geqslant \mathbb{E}[M \wedge R \,|\, \mathscr{F}_n]$ (by monotonicity or by Jensen's inequality with $x \mapsto x \wedge R$ concave) give for any $R > 0$

$$\mathbb{E}[M_\infty \wedge R] = \lim_{n\to\infty} \mathbb{E}[M_n \wedge R] \geqslant \mathbb{E}[M \wedge R].$$

This implies $\mathbb{E}[M_\infty] \geqslant \sup_{R>0} \mathbb{E}[M \wedge R] = \mathbb{E}[M]$, whence $\mathbb{E}[M_\infty] = \mathbb{E}[M] = \lim_{n\to\infty} \mathbb{E}[M_n]$. By Vitali's Theorem 4.50(c), we infer that (M_n) is uniformly integrable and from part (a) that $M_n = \mathbb{E}[M_\infty \,|\, \mathscr{F}_n]$.

The limit M_∞ is $\mathscr{F}_\infty$-measurable and satisfies for any $A \in \mathscr{F}_n$, $n \geqslant 1$:

$$\mathbb{E}[M_\infty \mathbf{1}_A] = \mathbb{E}[M_n \mathbf{1}_A] = \mathbb{E}[\mathbb{E}[M \,|\, \mathscr{F}_n] \mathbf{1}_A] = \mathbb{E}[M \mathbf{1}_A].$$

Hence, the probability measures(!) $\mathbb{Q}_1(A) := \mathbb{E}[M_\infty \mathbf{1}_A]/\mathbb{E}[M]$ and $\mathbb{Q}_2(A) := \mathbb{E}[M\mathbf{1}_A]/\mathbb{E}[M]$, $A \in \mathscr{F}_\infty$, coincide on $\bigcup_{m\geqslant 1} \mathscr{F}_m$. As the latter is an $\cap$-stable generator of $\mathscr{F}_\infty$, $\mathbb{Q}_1$ and $\mathbb{Q}_2$ agree everywhere. By definition, this means $M_\infty = \mathbb{E}[M \,|\, \mathscr{F}_\infty]$ a.s. This gives the result for $M \geqslant 0$ (if $M \geqslant 0$ and $\mathbb{E}[M] \leqslant 0$, then $M_n = M = 0$ a.s. and it is trivial). For general M consider $M_n^+ := \mathbb{E}[M^+ \,|\, \mathscr{F}_n]$, $M_n^- := \mathbb{E}[M^- \,|\, \mathscr{F}_n]$ separately.

□

4.52 Remark. The second martingale convergence theorem neatly characterizes the uniformly integrable martingales as closable martingales. The following result is very important in concrete situations. Note again that it is not valid for $p = 1$.▶Control

4.53 Corollary. *Let $p > 1$. Every L^p-bounded martingale (M_n) (i.e. $\sup_n \mathbb{E}[|M_n|^p] < \infty$) converges for $n \to \infty$ a.s. and in L^p, hence also in L^1.*

Proof. ▶Exercise □

4.54 Example. Recall from Analysis I the harmonic series $\sum_{k=1}^\infty \frac{1}{k} = \infty$ and the alternating harmonic series $\sum_{k=1}^\infty (-1)^{k+1} \frac{1}{k} = \log 2$. What about random signs in front of $\frac{1}{k}$? The random harmonic sum $S_n = \sum_{k=1}^n \frac{\varepsilon_k}{k}$ with a Rademacher sequence (ε_k) (i.i.d., $\mathbb{P}(\varepsilon_k = 1) = \mathbb{P}(\varepsilon_k = -1) = 1/2$) forms an L^2-bounded martingale due to $\mathbb{E}[S_n^2] = \sum_{k=1}^n k^{-2} \leqslant \sum_{k=1}^\infty k^{-2} = \frac{\pi^2}{6}$. Hence

$S_n \to S_\infty$ holds a.s. and in L^2. More generally, for $(a_k) \in \ell^2$ deterministic we have that $S_n = \sum_{k=1}^n \varepsilon_k a_k$ is an L^2-bounded martingale converging in L^2 and $\mathbb{P}$-a.s., e.g. $a_k = k^{-\alpha}$ is eligible for any $\alpha > 1/2$.

In the case of random (A_k) in $S_n = \sum_{k=1}^n \varepsilon_k A_k$ we could choose $A_k = \varepsilon_k / k$ and $S_n = \sum_{k=1}^n k^{-1}$ diverges. If (A_k) is predictable, however, and we set $M_n = \sum_{k=1}^n \varepsilon_k$, then $S_n = \sum_{k=1}^n \varepsilon_k A_k = (A \bullet M)_n$ is an L^2-martingale with quadratic variation $\langle S \rangle_n = (A^2 \bullet \langle M \rangle)_n = \sum_{k=1}^n A_k^2$. Using $\mathbb{E}[S_n^2] = \mathbb{E}[\langle S \rangle_n] = \sum_{k=1}^n \mathbb{E}[A_k^2]$, we conclude that (S_n) is an L^2-bounded martingale if and only if $\sum_{k=1}^\infty \mathbb{E}[A_k^2] < \infty$. In the following we obtain a more precise convergence result for L^2-martingales.

4.55 Definition. For $A, B \in \mathscr{F}$ we write $A \subseteq B$ $\mathbb{P}$-a.s. if $\mathbb{P}(A \setminus B) = 0$.

4.56 Proposition. *Let (M_n) be an L^2-martingale. Then:*

$$\Big\{ \lim_{n\to\infty} \langle M \rangle_n < \infty \Big\} \subseteq \Big\{ \lim_{n\to\infty} M_n \textit{ exists} \Big\} \quad \mathbb{P}\textit{-a.s.}$$

4.57 Remark. If the increments of (M_n) are uniformly bounded, then also the relation '$\supseteq$' holds $\mathbb{P}$-a.s., see Williams 12.13.

4.58 Example. The martingale $S_n = (A \bullet M)_n$ with predictable (A_n) from the previous example converges for $\mathbb{P}$-almost all ω such that $\sum_{k=1}^\infty A_k(\omega)^2$ is finite.

Proof. W.l.o.g. assume $M_0 = 0$ (otherwise consider the martingale $(M_n - M_0)_{n \geqslant 0}$). Since $\langle M \rangle_n$ is predictable, $\tau_k := \inf\{n \geqslant 0 \,|\, \langle M \rangle_{n+1} > k\}$ are stopping times for each $k \in \mathbb{N}$. For later note $\langle M \rangle_{\tau_k} \leqslant k$ in view of $\langle M \rangle_0 = 0$. The stopped quadratic variation $\langle M \rangle_n^{\tau_k} := \langle M \rangle_{\tau_k \wedge n}$ is predictable: for $n \geqslant 1$, $B \in \mathfrak{B}_\mathbb{R}$

$$\{\langle M \rangle_{\tau_k \wedge n} \in B\} = \bigcup_{l=0}^{n-1} \{\tau_k = l,\, \langle M \rangle_l \in B\} \cup \{\tau_k \geqslant n,\, \langle M \rangle_n \in B\} \in \mathscr{F}_{n-1}$$

holds. By definition and optional stopping $(M^2 - \langle M \rangle)_{\tau_k \wedge n}$ is a martingale such that

$$\begin{aligned} \mathbb{E}[(M_{n+1}^{\tau_k})^2 - \langle M \rangle_{n+1}^{\tau_k} \,|\, \mathscr{F}_n] &= \mathbb{E}[(M^2 - \langle M \rangle)_{\tau_k \wedge (n+1)} \,|\, \mathscr{F}_n] \\ &= (M^2 - \langle M \rangle)_{\tau_k \wedge n} = (M_n^{\tau_k})^2 - \langle M \rangle_n^{\tau_k}, \end{aligned}$$

Hence, $\langle M \rangle^{\tau_k}$ is the quadratic variation of M^{τ_k}: $\langle M^{\tau_k} \rangle = \langle M \rangle^{\tau_k}$ (the quadratic variation of the stopped martingale is the stopped quadratic variation).

Now, $\mathbb{E}[(M_n^{\tau_k})^2] = \mathbb{E}[\langle M \rangle_n^{\tau_k}] \leqslant k$ holds for all $n \geqslant 0$ and $(M_n^{\tau_k})_n$ is an L^2-bounded martingale. This shows that $\lim_{n\to\infty} M_n^{\tau_k}$ exists a.s. for all k. The identity $\{\exists k \geqslant 1 : \tau_k = \infty\} = \{\lim_{n\to\infty} \langle M \rangle_n < \infty\}$ implies the assertion. □

▷ Control questions

(a) Let $(f_n)_{n \geqslant 1}$ be probability densities on $[0,1]$ with $f_n(x) \to f(x)$ for Lebesgue-almost all $x \in [0,1]$. Why does Vitali's Theorem show $\|f_n - f\|_{L^1} \to 0$ if and only if f is a probability density itself? [Remark: in that case the laws converge in total-variation distance!]

Probability densities are nonnegative and integrate to one. Interpreting Lebesgue measure on $[0,1]$ as a uniform probability, Vitali's Theorem shows $\|f_n - f\|_{L^1} \to 0$ if and only if $\int_0^1 f(x)dx = 1$. Since f is as a pointwise limit also nonnegative and measurable, the last property is equivalent to f being a probability density. This result is known as *Scheffé's Theorem*.

(b) What is an example for $M_\infty \neq M$ in the second martingale convergence theorem?

Let $\mathscr{F}_n = \{\varnothing, \Omega\}$ be trivial for all n. Then $M_n = \mathbb{E}[M \,|\, \mathscr{F}_n] = \mathbb{E}[M]$ a.s. is constant and $M_\infty = \mathbb{E}[M]$ as well. For a non-trivial random variable M this implies $M_\infty \neq M$ a.s.

(c) Why does Corollary 4.53 not hold for $p = 1$?

The multiplicative martingale $P_n = \prod_{k=1}^n X_k$ with $\mathbb{P}(X_k = 3/2) = \mathbb{P}(X_k = 1/2) = 1/2$ from Example 4.4(b) satisfies $\mathbb{E}[P_n] = 1$, but $P_\infty = 0$ a.s. so that (P_n) is L^1-bounded (by one), but not uniformly integrable due to $\mathbb{E}[P_\infty] \neq \lim_{n\to\infty} \mathbb{E}[P_n]$.

4.59 Lemma (Kronecker's Lemma)**.** *Let (a_n), (c_n) be sequences of real numbers with $a_n \uparrow \infty$ and such that $\lim_{N\to\infty} \sum_{n=1}^N \frac{c_n}{a_n}$ exists. Then $\lim_{N\to\infty} \frac{1}{a_N} \sum_{n=1}^N c_n = 0$ holds.*

Proof. Set $d_N := \sum_{n=1}^N \frac{c_n}{a_n}$ and $b_n := a_n - a_{n-1} \geqslant 0$, $n \geqslant 1$, where $a_0 := 0$, and let N be so large that $a_N > 0$. Then partial summation gives for any $1 \leqslant k < N$

$$\begin{aligned}\frac{1}{a_N}\Big|\sum_{n=1}^N c_n\Big| &= \frac{1}{a_N}\Big|\sum_{n=1}^N a_n(d_n - d_{n-1})\Big| = \frac{1}{a_N}\Big|-\sum_{n=1}^N (a_n - a_{n-1})d_{n-1} + a_N d_N\Big| \\ &= \Big|\sum_{n=1}^N \frac{b_n}{a_N}(d_N - d_{n-1})\Big| \leqslant \Big|\sum_{n=1}^k \frac{b_n}{a_N}(d_N - d_{n-1})\Big| + \max_{k\leqslant n\leqslant N} |d_N - d_n|,\end{aligned}$$

using $\sum_{n=k+1}^N \frac{b_n}{a_N} \in [0,1]$. Letting first $N \to \infty$, then $k \to \infty$ the right-hand side tends to zero. □

4.60 Example. Let $(X_i)_{i\geqslant 1}$ be i.i.d. random variables in L^2 with $\mathbb{E}[X_i] = 0$. Then $M_n := \sum_{i=1}^n \frac{X_i}{i}$, $M_0 := 0$, forms a martingale with $\mathbb{E}[M_n^2] = \sum_{i=1}^n i^{-2}\,\mathbb{E}[X_i^2] \leqslant \frac{\pi^2}{6}\,\mathbb{E}[X_1^2]$. Consequently, (M_n) is an L^2-bounded martingale and converges by Corollary 4.53 almost surely. Kronecker's Lemma with $a_n = n$ and $c_n = X_n(\omega)$ thus shows that also $\frac{1}{n}\sum_{i=1}^n X_i \to 0$ a.s. holds, which is the strong law of large numbers. Next, we obtain a more general and more precise result.

4.61 Lemma. *Let $(M_n)_{n\geqslant 0}$ and $(A_n)_{n\geqslant 0}$ be processes, (A_n) non-negative and increasing with $A_n \uparrow A_\infty \in \mathbb{R}\cup\{\infty\}$. Then*

$$\{A_\infty = \infty\} \cap \Big\{ \lim_{n\to\infty} ((1+A)^{-1}{\bullet}M)_n \textit{ exists in } \mathbb{R} \Big\} \subseteq \Big\{ \lim_{n\to\infty} \frac{M_n}{A_n} = 0 \Big\}.$$

Proof. Put $a_n = 1 + A_n(\omega)$, $c_n = (M_n - M_{n-1})(\omega)$. Then by definition for all ω in the left-hand side event $a_n \uparrow \infty$ holds and $\sum_{n=1}^N c_n/a_n$ converges as $N \to \infty$. Kronecker's Lemma implies $\frac{1}{a_N}\sum_{n=1}^N c_n \to 0$. This means $\frac{M_N(\omega)-M_0(\omega)}{1+A_N(\omega)} \to 0$ and thus $(M_N/A_N)(\omega) \to 0$, as asserted. □

4.62 Corollary (Strong law of large numbers for L^2-martingales)**.** *An L^2-martingale (M_n) satisfies for any increasing function $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ with $\int_0^\infty (1+\rho(t))^{-2} dt < \infty$*

$$\Big\{ \lim_{n\to\infty} \langle M \rangle_n = \infty \Big\} \subseteq \Big\{ \lim_{n\to\infty} \frac{M_n}{\rho(\langle M \rangle_n)} = 0 \Big\} \quad \mathbb{P}\text{-}a.s.$$

4.63 Remark. We may choose $\rho(t) = t^\alpha$ for any $\alpha > 1/2$ or even $\rho(t) = \sqrt{t(\log(1+t))^\beta}$ for any $\beta > 1$.

Proof. Consider $X_n = ((1+\rho(\langle M \rangle)^{-1} \bullet M)_n$. Then (X_n) is an L^2-martingale with

$$\langle X \rangle_n = ((1+\rho(\langle M \rangle))^{-2} \bullet \langle M \rangle)_n = \sum_{k=1}^n \frac{\langle M \rangle_k - \langle M \rangle_{k-1}}{(1+\rho(\langle M \rangle_k))^2} \leqslant \sum_{k=1}^n \int_{\langle M \rangle_{k-1}}^{\langle M \rangle_k} (1+\rho(t))^{-2} dt.$$

This shows $\lim_{n\to\infty} \langle X \rangle_n \leqslant \int_0^\infty (1+\rho(t))^{-2} dt < \infty$ $\mathbb{P}$-a.s. Proposition 4.56 implies that (X_n) converges $\mathbb{P}$-a.s. and Lemma 4.61 with $A_n = \rho(\langle M \rangle_n)$ yields the result. □

4.64 Example. Let $(X_i)_{i\geqslant 1}$ be independent L^2-random variables with $\mathbb{E}[X_i] = 0$, $\operatorname{Var}(X_i) = \sigma_i^2$. Then $S_n = \sum_{i=1}^n X_i$ is an L^2-martingale with $\langle S \rangle_n = \operatorname{Var}(S_n) = \sum_{i=1}^n \sigma_i^2$. If $\sum_{i=1}^\infty \sigma_i^2 = \infty$, then we conclude

$$\forall \beta > 1 : \; \frac{S_n - \mathbb{E}[S_n]}{\operatorname{Var}(S_n)^{1/2}(\log(1+\operatorname{Var}(S_n)))^{\beta/2}} \to 0 \quad \mathbb{P}\text{-a.s.}$$

If $\operatorname{Var}(X_i) = \sigma^2 > 0$ independent of i, then $n^{-1/2}\log(n)^{-\beta/2}(S_n - \mathbb{E}[S_n]) \to 0$ a.s., that is $S_n - \mathbb{E}[S_n]$ converges 'almost' with rate $n^{-1/2}$ to zero. The *law of the iterated logarithm* shows in the i.i.d. case that the critical scaling between a.s. convergence and divergence is $\sqrt{n \log(\log n)}$.

4.65 Remark. The scaling with $\rho(t) = t^{1/2}$ gives rise to a martingale central limit theorem. If (M_n) is an L^2-martingale with $M_0 = 0$, $\langle M \rangle_n \to \infty$, then

$$\frac{M_n}{\langle M \rangle_n^{1/2}} \xrightarrow{d} N(0,1)$$

holds, provided $\langle M \rangle_n / \mathbb{E}[M_n^2] \to 1$ and the conditional Lindeberg condition is satisfied:

$$\forall \varepsilon > 0 : \; \frac{1}{\mathbb{E}[M_n^2]} \sum_{k=1}^n \mathbb{E}\Big[(M_k - M_{k-1})^2 \mathbf{1}\big((M_k - M_{k-1})^2 \geqslant \varepsilon^2\, \mathbb{E}[M_n^2]\big) \,\Big|\, \mathscr{F}_{k-1} \Big] \xrightarrow{\mathbb{P}} 0,$$

cf. A. Shiryaev, Probability, Springer, Thm. VII.8.4.

4.66 Definition. A process $(M_{-n})_{n\geqslant 0}$ is called backward martingale (Rückwärtsmartingal) with respect to $(\mathscr{F}_{-n})_{n\geqslant 0}$ with σ-algebras $\mathscr{F}_{-n-1}\subseteq\mathscr{F}_{-n}$ if $M_{-n}\in L^1$, M_{-n} is $\mathscr{F}_{-n}$-measurable and $\mathbb{E}[M_{-n}\,|\,\mathscr{F}_{-n-1}]=M_{-n-1}$ hold for all $n\geqslant 0$.

4.67 Remark. While a martingale (M_n) provides finer information for $n\to\infty$ because the filtration increases, a backward martingale (M_{-n}) provides coarser information as $n\to\infty$ and in case $\bigcap_{n\geqslant 0}\mathscr{F}_{-n}=\{\varnothing,\Omega\}$ a potential limit $M_{-\infty}$ must be necessarily constant (deterministic).

4.68 Theorem. *Every backward martingale $(M_{-n})_{n\geqslant 0}$ converges for $n\to\infty$ a.s. and in L^1.*

Proof. Denote by $U_{-n}^{[a,b]}$ the upcrossings on $[a,b]$ of $(M_{-k},k=0,\dots,n)$. The upcrossing inequality gives $\mathbb{E}[U_{-n}^{[a,b]}]\leqslant\mathbb{E}[(M_0-a)_+]/(b-a)$ because it relies on the martingale property for finitely many time indices only. With $U_{-n}^{[a,b]}\uparrow U^{[a,b]}$ as $n\uparrow\infty$ monotone convergence shows $\mathbb{E}[U^{[a,b]}]\leqslant\mathbb{E}[(M_0-a)_+]/(b-a)<\infty$ for any $a<b$. Now, $U^{[a,b]}$ counts the upcrossings on $[a,b]$ of $(M_{-n},n\geqslant 0)$. As in the first martingale convergence theorem this implies $\mathbb{P}$-a.s. convergence $M_{-n}\to M_{-\infty}$ for some random variable $M_{-\infty}\in L^1$. Now, $M_{-n}=\mathbb{E}[M_0\,|\,\mathscr{F}_{-n}]$ holds and the same argument as for the second martingale convergence theorem shows that $(M_{-n},n\geqslant 0)$ is uniformly integrable▶Control. Vitali's Theorem thus implies $M_{-n}\to M_{-\infty}$ in L^1. □

4.69 Corollary. *(Kolmogorov's strong law of large numbers) For i.i.d. random variables $(X_i)_{i\geqslant 1}$ in L^1 we have*

$$\frac{1}{n}\sum_{i=1}^n X_i\xrightarrow{\mathbb{P}\text{-}a.s.\ and\ L^1}\mathbb{E}[X_1].$$

Proof. Put $S_0=0$, $S_n=\sum_{i=1}^n X_i$, $n\geqslant 1$ and $\mathscr{F}_{-n}=\sigma(S_k,k\geqslant n)$, $n\geqslant 0$. Then $\mathscr{F}_{-n-1}\subseteq\mathscr{F}_{-n}$ holds for $n\geqslant 0$ and S_n is $\mathscr{F}_{-n}$-measurable. From

$$S_n=\mathbb{E}[S_n\,|\,\mathscr{F}_{-n}]=\sum_{i=1}^n\mathbb{E}[X_i\,|\,\mathscr{F}_{-n}]$$

and the fact that $(X_i)_{i=1,\dots,n}$ has the same law as $(X_{\pi(i)})_{i=1,\dots,n}$ for any permutation π of $\{1,\dots,n\}$ ((X_i) are *exchangeable*), while S_k for $k\geqslant n$ is invariant under each π, we conclude that $\mathbb{E}[X_i\,|\,\mathscr{F}_{-n}]$ does not depend on $i\in\{1,\dots,n\}$ and thus equals S_n/n. By definition $M_{-n}:=\mathbb{E}[X_1\,|\,\mathscr{F}_{-n}]$ forms a backward martingale. We conclude that $S_n/n=M_{-n}\to M_{-\infty}$ converges $\mathbb{P}$-a.s. and in L^1. By Kolmogorov's 0-1 law (Stochastik I, ▶Exercise) the limit $M_{-\infty}$ must be $\mathbb{P}$-a.s. constant. From $\mathbb{E}[X_1]=\mathbb{E}[M_{-n}]\to\mathbb{E}[M_{-\infty}]$ we infer $M_{-\infty}=\mathbb{E}[X_1]$ $\mathbb{P}$-a.s. □

4.4 The Radon-Nikodym theorem and Kakutani's dichotomy

4.70 Definition. Let μ and ν be measures on the measurable space $(\Omega,\mathscr{F})$. Then μ is absolutely continuous (absolutstetig) with respect to ν, notation

$\mu \ll \nu$, if $\forall A \in \mathscr{F} : \nu(A) = 0 \Rightarrow \mu(A) = 0$. μ and ν are equivalent (äquivalent), notation $\mu \sim \nu$, if $\mu \ll \nu$ and $\nu \ll \mu$. If there is an $A \in \mathscr{F}$ with $\nu(A) = 0$ and $\mu(A^C) = 0$, then μ and ν are singular (singulär), notation $\mu \perp \nu$.

4.71 Example.

(a) A probability measure $\mathbb{P}$ on $\mathfrak{B}_{\mathbb{R}^d}$ with Lebesgue density f is absolutely continuous with respect to Lebesgue measure λ: $\lambda(A) = 0 \Rightarrow \mathbb{P}(A) = \int_A f(x)\lambda(dx) = 0$.

More generally, any measure μ with a ν-density f, i.e. $\mu(A) = \int_A f d\nu$, satisfies $\mu \ll \nu$. If $f > 0$ ν-almost everywhere holds, then $\mu(A) = 0$ implies $\int f\mathbf{1}_A d\nu = 0$ and thus $\nu(A) = \nu(\{f\mathbf{1}_A > 0\}) = 0$. This shows $\nu \ll \mu$, hence equivalence $\nu \sim \mu$.

(b) The measures λ and δ_0 on $\mathfrak{B}_{\mathbb{R}}$ are singular: $\lambda(\{0\}) = 0$, $\delta_0(\{0\}^C) = 0$.

▷ **Control questions**

(a) Why do we need $\langle M\rangle_n \to \infty$ in the strong law for L^2-martingales? Provide a simple example where $\sup_n \langle M\rangle_n < \infty$ and $M_n/\langle M\rangle_n$ does not converge to zero.

For an L^1-random variable X we have that $M_n = X$ is a (constant) martingale with respect to $\mathscr{F}_n = \mathscr{F}$. Its quadratic variation is $\langle M\rangle_n = X^2$. Hence, $M_n/\langle M\rangle_n = X^{-1}$ does not convergence to zero (it even needs not be well-defined)

(b) Why do we have for backward martingales (M_{-n}) that $M_{-n} \to \mathbb{E}[M_{-1} \,|\, \mathscr{F}_{-\infty}]$ a.s. and in L^1 with $\mathscr{F}_{-\infty} = \bigcap_{n\geqslant 1} \mathscr{F}_{-n}$?

By the martingale property $M_{-n} = \mathbb{E}[M_{-1} \,|\, \mathscr{F}_{-n}]$ and by backward martingale convergence $M_{-n} \to M_{-\infty}$ a.s. and in L^1 for some $M_{-\infty} \in L^1$ hold. By the filtration property M_{-n} is $\mathscr{F}_{-m}$-measurable for all $n \geqslant m$ so that $M_{-\infty}$ is also $\mathscr{F}_{-m}$-measurable for all m. This shows that $M_{-\infty}$ is $\mathscr{F}_{-\infty}$-measurable. For $A \in \mathscr{F}_{-\infty}$ we have by L^1-convergence

$$\mathbb{E}[M_{-\infty}\mathbf{1}_A] = \lim_{n\to\infty} \mathbb{E}[M_{-n}\mathbf{1}_A] = \lim_{n\to\infty} \mathbb{E}[\mathbb{E}[M_{-1}\mathbf{1}_A \,|\, \mathscr{F}_{-n}]] = \mathbb{E}[M_{-1}\mathbf{1}_A],$$

whence $M_{-\infty} = \mathbb{E}[M_{-1} \,|\, \mathscr{F}_{-\infty}]$ holds.

4.72 Lemma. *A finite measure μ is absolutely continuous with respect to a measure ν if and only if*

$$\forall \varepsilon > 0 \,\exists \delta > 0 \,\forall A \in \mathscr{F} : \; \nu(A) < \delta \Rightarrow \mu(A) < \varepsilon.$$

Proof. To prove '$\Leftarrow$' suppose $\nu(A) = 0$. Then $\nu(A) < \delta$ holds for all $\delta > 0$ and thus $\mu(A) < \varepsilon$ for all $\varepsilon > 0$. This implies $\mu(A) = 0$, hence $\mu \ll \nu$.

The implication '$\Rightarrow$' is shown by contradiction. Assume there are $A_n \in \mathscr{F}$ and $\varepsilon > 0$ with $\nu(A_n) \leqslant 2^{-n}$ and $\mu(A_n) \geqslant \varepsilon$. Consider the event $A_\infty = \bigcap_{m\geqslant 1} \bigcup_{n\geqslant m} A_n$ that infinitely many events A_n occur. Then for all $n_0 \in \mathbb{N}$

$$\nu(A_\infty) = \int \limsup_{n\to\infty} \mathbf{1}_{A_n} d\nu \leqslant \int \sum_{n>n_0} \mathbf{1}_{A_n} d\nu = \sum_{n>n_0} \nu(A_n) \leqslant \sum_{n>n_0} 2^{-n} = 2^{-n_0}$$

holds, which for $n_0 \to \infty$ yields $\nu(A_\infty) = 0$ (for $\nu = \mathbb{P}$ this is Borel-Cantelli). On the other hand, Fatou's Lemma gives

$$\mu(A_\infty) = \int \limsup_{n\to\infty} (1 - \mathbf{1}_{A_n^C})\, d\mu \geqslant \mu(\Omega) - \liminf_{n\to\infty} \int \mathbf{1}_{A_n^C}\, d\mu = \limsup_{n\to\infty} \mu(A_n) \geqslant \varepsilon.$$

This contradicts $\mu \ll \nu$. □

4.73 Remark.

(a) It is necessary to ask for finite μ as the following counterexample shows: take ν the Lebesgue measure on $((0,1], \mathfrak{B}_{(0,1]})$ and $\mu(dx) = x^{-1}dx$. Then $\mu \ll \nu$ holds, while $\nu((0,\delta)) = \delta$ and $\mu((0,\delta)) = \infty$ hold for all $\delta \in (0,1)$.

(b) If F is the distribution function of a probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$, absolutely continuous with respect to Lebesgue measure, then the lemma says that there is for any $\varepsilon > 0$ a $\delta > 0$ such that

$$\forall n \in \mathbb{N}; a_1 \leqslant b_1 \leqslant \cdots \leqslant a_n \leqslant b_n : \sum_{i=1}^n (b_i - a_i) < \delta \Rightarrow \sum_{i=1}^n (F(b_i) - F(a_i)) < \varepsilon.$$

In real analysis one says that the function F is *absolutely continuous* and thus *weakly differentiable* with derivative f (the density or the Radon-Nikodym derivative of $\mathbb{P}$, see below). Note that for Cantor measure $\mathbb{P}$ the distribution function F is continuous, but not absolutely continuous in that sense.

(c) If a finite measure μ has a density f with respect to a probability measure $\mathbb{P}$, then the preceding lemma shows $\forall \varepsilon > 0\ \exists \delta > 0 : \mathbb{P}(A) < \delta \Rightarrow \mathbb{E}_\mathbb{P}[\mathbf{1}_A f] < \varepsilon$. In view of Lemma 4.48(a), the density f, considered as a random variable under $\mathbb{P}$, is uniformly integrable (which is clear for one L^1-random variable anyway▶Control). The remarkable fact of the following Radon-Nikodym Theorem is that absolute continuity suffices to ensure the existence of a density. Its martingale proof is constructive in the sense that we build the density iteratively for finitely generated σ-algebras and then take the limit.

4.74 Theorem (Radon-Nikodym Theorem (Vitali 1905, Lebesgue 1910, Radon 1913, Nikodym 1930, von Neumann 1940))**.** *Let ν be a σ-finite measure and μ a measure on $\mathscr{F}$ with $\mu \ll \nu$, then there is an $f \in \mathcal{M}^+(\Omega, \mathscr{F})$ such that*

$$\mu(A) = \int_A f\, d\nu \text{ for all } A \in \mathscr{F}.$$

4.75 Definition. The function f in the Radon-Nikodym theorem is called Radon-Nikodym derivative, density or likelihood function of μ with respect to ν, notation $f = \frac{d\mu}{d\nu}$.

Proof. We give the proof in the case μ finite, $\nu = \mathbb{P}$ and $\mathscr{F} = \sigma(F_n, n \geqslant 1)$ for some $F_n \subseteq \Omega$ ($\mathscr{F}$ *separable*, e.g. a Borel σ-algebra of a Polish space ▶Control); see Williams for the general case.

Put $\mathscr{F}_n = \sigma(F_1, \dots, F_n)$. Then $\mathscr{F}_n$ consists of finitely many events only (intersections of the F_i and finite unions). In particular, there are finitely many 'atoms' $A_1^{(n)}, \dots, A_{r_n}^{(n)} \in \mathscr{F}_n$ with $\bigcup_{i=1}^{r_n} A_i^{(n)} = \Omega$, $A_i^{(n)} \cap A_j^{(n)} = \varnothing$ for $i \neq j$ and $\mathscr{F}_n = \sigma(A_1^{(n)}, \dots, A_{r_n}^{(n)})$. Then $\frac{\mu(A_i^{(n)})}{\mathbb{P}(A_i^{(n)})}$ should be the $\mathbb{P}$-density of μ on each $A_i^{(n)}$, restricted to $\mathscr{F}_n$. More precisely, we set

$$M_n := \sum_{i=1}^{r_n} \Big(\frac{\mu(A_i^{(n)})}{\mathbb{P}(A_i^{(n)})} \mathbf{1}(\mathbb{P}(A_i^{(n)}) > 0) \Big) \mathbf{1}_{A_i^{(n)}}.$$

Then $M_n \in \mathcal{M}^+(\Omega, \mathscr{F}_n)$ and for $F \in \mathscr{F}_n$ we have $\int_F M_n d\,\mathbb{P} = \sum_i \mu(A_i^{(n)} \cap F) = \mu(F)$. This shows indeed $M_n = \frac{d\mu|_{\mathscr{F}_n}}{d\,\mathbb{P}|_{\mathscr{F}_n}}$ and (M_n) is an $(\mathscr{F}_n)$-martingale because $\mathbb{E}[M_n \mathbf{1}_F] = \mu(F) = \mathbb{E}[M_{n+1}\mathbf{1}_F]$ for $F \in \mathscr{F}_n$ means $\mathbb{E}[M_{n+1} \,|\, \mathscr{F}_n] = M_n$.

From Lemma 4.72 with $\nu = \mathbb{P}$, $A = \{M_n > R\}$ for any $\varepsilon > 0$ there is a $\delta > 0$ such that $\mathbb{P}(M_n > R) < \delta$ implies $\mu(M_n > R) < \varepsilon$. By Markov's inequality this holds in particular for $R > \mu(\Omega)/\delta$: $\mathbb{P}(M_n > R) \leqslant \frac{\mathbb{E}[M_n]}{R} < \delta$. For this R then $\sup_n \mathbb{E}[M_n \mathbf{1}(M_n > R)] = \sup_n \mu(M_n > R) < \varepsilon$ holds. We conclude that (M_n) is uniformly integrable and hence forms a closable martingale with $M_n = \mathbb{E}[M_\infty \,|\, \mathscr{F}_n]$ for $M_\infty = \lim_{n\to\infty} M_n$.

Define the finite measure $\rho(A) = \int_A M_\infty d\,\mathbb{P}$ on $\mathscr{F}$. For $A \in \mathscr{F}_n$

$$\rho(A) = \mathbb{E}[\mathbf{1}_A \, \mathbb{E}[M_\infty \,|\, \mathscr{F}_n]] = \mathbb{E}[\mathbf{1}_A M_n] = \mu(A)$$

follows. Consequently ρ and μ coincide on the $\cap$-stable generator $\bigcup_{n\geqslant 1} \mathscr{F}_n$ of $\mathscr{F}$ and have the same mass $\rho(\Omega) = \mu(\Omega)$. The uniqueness theorem for finite measures gives $\rho = \mu$ on $\mathscr{F}$ and the theorem follows with $f = M_\infty$. □

4.76 Corollary (Lebesgue decomposition). *For σ-finite measures μ, ν on $(\Omega, \mathscr{F})$ we can decompose $\mu = \mu_1 + \mu_2$ with σ-finite measures $\mu_1 \ll \nu$ and $\mu_2 \perp \nu$.*

4.77 Remark. It is easy to see that this decomposition is unique unless $\nu = 0$.

Proof. Put $\rho = \mu + \nu$. Since $\nu \ll \rho$ and ρ is σ-finite, there is a Radon-Nikodym derivative $f = \frac{d\nu}{d\rho}$. Set $\mu_1(A) := \mu(A \cap \{f > 0\})$, $\mu_2(A) := \mu(A \cap \{f = 0\})$, $A \in \mathscr{F}$. Then $\mu = \mu_1 + \mu_2$ holds and

$$\nu(\{f = 0\}) = \int_{\{f=0\}} f \, d\rho = 0, \quad \mu_2(\{f = 0\}^C) = \mu(\varnothing) = 0 \Rightarrow \nu \perp \mu_2.$$

For $A \in \mathscr{F}$ with $\nu(A) = 0$, on the other hand, we have

$$\int_A f \, d\rho = 0 \Rightarrow \int_{A \cap \{f > 0\}} f \, d\mu = 0 \Rightarrow \mu(A \cap \{f > 0\}) = 0,$$

which implies $\mu_1 \ll \nu$. □

4.78 Theorem (Kakutani 1948). *Let $(X_k)_{k\geqslant 1}$ be independent random variables with $X_k \geqslant 0$ and $\mathbb{E}[X_k] = 1$. Then $M_n := \prod_{k=1}^n X_k$, $M_0 = 1$ is a non-negative martingale converging a.s. to some M_∞. The following statements are equivalent:*

(a) $\mathbb{E}[M_\infty]=1$*;*

(b) $M_n\to M_\infty$ *in* L^1*;*

(c) (M_n) *is uniformly integrable;*

(d) $\prod_{k=1}^\infty \mathbb{E}[X_k^{1/2}]>0$*;*

(e) $\sum_{k=1}^\infty (1-\mathbb{E}[X_k^{1/2}])<\infty$.

If one (then all) statement fails to hold, then $M_\infty=0$ *holds a.s. (Kakutani's dichotomy).*

4.79 Remark. Here, we use constantly the martingale property for products from Example 4.4(b) and the Corollary 4.45 to the first martingale convergence theorem. Note also the concrete application to the case $\mathbb{P}(X_k=3/2)=\mathbb{P}(X_k=1/2)=1/2$ from Example 4.4(b).

Proof. First note $a_k:=\mathbb{E}[X_k^{1/2}]\leqslant \mathbb{E}[X_k]^{1/2}=1$ by Jensen's inequality and $a_k=\mathbb{E}[X_k^{1/2}]>0$ from $\mathbb{P}(X_k>0)>0$ due to $\mathbb{E}[X_k]=1$. In particular, the product $\prod_{k=1}^\infty a_k$ converges always to a limit in $[0,1]$ due to $0<a_k\leqslant 1$. The equivalence (a) $\iff$ (b) $\iff$ (c) is due to Vitali's Theorem. The equivalence (d) $\iff$ (e) is shown in analysis (consider $\log(\prod_k \mathbb{E}[X_k^{1/2}])$▶CONTROL).

(a)⇒(d): Define $a_k:=\mathbb{E}[X_k^{1/2}]$ and $N_n:=\prod_{k=1}^n a_k^{-1}X_k^{1/2}$, $n\geqslant 1$, $N_0=1$. Then (N_n) is a non-negative martingale with $M_n^{1/2}/\prod_{k=1}^n a_k=N_n\to N_\infty$ $\mathbb{P}$-a.s. for some $N_\infty\in L^1$. Since the nonnegative martingale (M_n) satisfies $M_n\to M_\infty$ $\mathbb{P}$-a.s. with $\mathbb{E}[M_\infty]=1$ by (a), we have

$$M_\infty=N_\infty^2\prod_{k=1}^\infty a_k^2\ \mathbb{P}\text{-a.s.}\Rightarrow \mathbb{E}\Big[N_\infty^2\Big(\prod_{k=1}^\infty a_k\Big)^2\Big]=1\Rightarrow \prod_{k=1}^\infty a_k>0.$$

(d)⇒(a): The martingale (N_n) from above satisfies under (d) $\mathbb{E}[N_n^2]=\mathbb{E}[M_n]/\prod_{k=1}^n a_k^2\leqslant \prod_{k=1}^\infty a_k^{-2}<\infty$. As an L^2-bounded martingale it converges in L^2 and $\mathbb{E}[N_n^2]\to\mathbb{E}[N_\infty^2]=\mathbb{E}[M_\infty]\prod_{k=1}^\infty a_k^{-2}$. This implies $\mathbb{E}[M_n]\to\mathbb{E}[M_\infty]$.

If (a)-(e) do not hold, then $\prod_{k=1}^\infty a_k=0$ and the argument in (d)⇒(a) shows $M_\infty=N_\infty^2\prod_{k=1}^\infty a_k^2=0$ $\mathbb{P}$-a.s. □

▷ **Control questions**

(a) Why is a finite family $(X_i)_{1\leqslant i\leqslant n}$ of L^1-random variables X_i always uniformly integrable?

Dominated convergence shows for $X_i\in L^1$ that $\lim_{R\to\infty}\mathbb{E}[|X_i|\mathbf{1}(|X_i|>R)]=0$. We conclude by

$$\lim_{R\to\infty}\max_{1\leqslant i\leqslant n}\mathbb{E}[|X_i|\mathbf{1}(|X_i|>R)]\leqslant \lim_{R\to\infty}\sum_{i=1}^n\mathbb{E}[|X_i|\mathbf{1}(|X_i|>R)]=\sum_{i=1}^n 0=0.$$

(b) Why is $\mathfrak{B}_{\mathbb{R}}$ and more generally the Borel-σ-algebra of a Polish space always separable?

Let D be a countable dense set of the Polish space, then the set of open balls $\mathscr{E} := \{B_r(d)\, d \in D, r \in \mathbb{Q}^+\}$ is countable and clearly inside the Borel σ-algebra. For any open set O and any $x \in O$ there is an $r > 0$ with $B_{2r}(x) \subseteq O$. This radius r can be chosen rational in $\mathbb{Q}^+$. Moreover, by denseness there is a $d \in D$ with $x \in B_r(d)$ and thus $B_r(d) \subseteq B_{2r}(x) \subseteq O$. This shows $O = \bigcup_{n\geqslant 1} B_{r_n}(d_n)$ for suitable $r_n \in \mathbb{Q}^+$ and $d_n \in D$. Hence, all open sets are in $\sigma(\mathscr{E})$ and $\mathscr{E}$ generates the Borel-σ algebra.

(c) Show $\prod_{k=1}^\infty a_k > 0 \iff \sum_{k=1}^\infty (1-a_k) < \infty$ for real numbers $a_k \in (0,1]$ and thus establish equivalence between (d) and (e) in Kakutani's Theorem.

Taking logarithms we have $\prod_{k=1}^\infty a_k > 0 \iff \sum_{k=1}^\infty \log(a_k) > -\infty$. By concavity of the log-function we have $\log(a_k) \leqslant a_k - 1$. Therefore, $\prod_{k=1}^\infty a_k > 0$ implies $\sum_{k=1}^\infty (1-a_k) < \infty$. Conversely, if $\sum_{k=1}^\infty (1-a_k) < \infty$, then $\lim_{k\to\infty}(1-a_k) = 0$ and $\lim_{k\to\infty} \frac{-\log(a_k)}{1-a_k} = \lim_{x\to 0} \frac{-\log(1-x)}{x} = 1$ (use L'Hôpital's rule). This implies that also $-\sum_{k=1}^\infty \log(a_k)$ converges to a finite value ('limit comparison test'), hence $\prod_{k=1}^\infty a_k > 0$.

4.80 Definition. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures with densities $p = \frac{d\mathbb{P}}{d\mu}$, $q = \frac{d\mathbb{Q}}{d\mu}$ for some *dominating* measure μ (e.g. $\mu = \mathbb{P} + \mathbb{Q}$). Then their Hellinger distance is defined as

$$H(\mathbb{P},\mathbb{Q}) := \Big(\int (\sqrt{p} - \sqrt{q})^2 d\mu\Big)^{1/2} = \|\sqrt{p} - \sqrt{q}\|_{L^2(\mu)}.$$

4.81 Remark. One can show that this definition does not depend on the dominating measure μ. H defines a metric on the set of all probability measures on $(\Omega, \mathscr{F})$.

4.82 Lemma. *The formula $H^2(\mathbb{P},\mathbb{Q}) = 2(1 - \int \sqrt{pq}\, d\mu)$ holds and if $\mathbb{Q} \ll \mathbb{P}$ then $H^2(\mathbb{P},\mathbb{Q}) = 2(1 - \mathbb{E}_{\mathbb{P}}[(\frac{d\mathbb{Q}}{d\mathbb{P}})^{1/2}])$.*

Proof. The first identity follows from the binomial formula:

$$H^2(\mathbb{P},\mathbb{Q}) = \int \Big(p - 2\sqrt{pq} + q\Big)\, d\mu = 1 - 2\int \sqrt{pq}\, d\mu + 1.$$

For $\mathbb{Q} \ll \mathbb{P}$ consider $\mu = \mathbb{P}$ and substitute $q = \frac{d\mathbb{Q}}{d\mathbb{P}}$, $p = 1$. □

4.83 Theorem. *Let $(\mathbb{P}_n)_{n\geqslant 1}$, $(\mathbb{Q}_n)_{n\geqslant 1}$ be two sequences of probability measures on $(\Omega, \mathscr{F})$ with $\mathbb{Q}_n \ll \mathbb{P}_n$, $n \geqslant 1$. Then for the product measures on $(\Omega^{\mathbb{N}}, \mathscr{F}^{\otimes \mathbb{N}})$ we have*

$$\bigotimes_{n=1}^\infty \mathbb{Q}_n \ll \bigotimes_{n=1}^\infty \mathbb{P}_n \iff \sum_{n=1}^\infty H^2(\mathbb{P}_n, \mathbb{Q}_n) < \infty.$$

Otherwise, we have singularity $\bigotimes_{n=1}^\infty \mathbb{Q}_n \perp \bigotimes_{n=1}^\infty \mathbb{P}_n$ (Kakutani's dichotomy).

4.84 Example. In particular, for $\mathbb{P}_n = \mathbb{P}$ and $\mathbb{Q}_n = \mathbb{Q}$ with $\mathbb{P} \neq \mathbb{Q}$ the infinite product measures $\mathbb{P}^{\otimes \mathbb{N}}$ and $\mathbb{Q}^{\otimes \mathbb{N}}$ are singular. Note $H(\mathbb{P},\mathbb{Q}) > 0$ for $\mathbb{P} \neq \mathbb{Q}$.

Proof. Introduce the coordinate projections $X_n : \Omega^{\mathbb{N}} \to \Omega$, $X_n((\omega_m)_m) := \omega_n$ and recall that under $\mathbb{P} = \bigotimes_{n=1}^{\infty} \mathbb{P}_n$ the X_n are independent random variables with laws $\mathbb{P}_n$. Introduce the product densities

$$\Lambda_n(\omega) = \frac{d(\mathbb{Q}_1 \otimes \cdots \otimes \mathbb{Q}_n)}{d(\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n)}(\omega_1, \dots, \omega_n) = \prod_{k=1}^{n} \frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}(X_k(\omega)), \quad \omega \in \Omega^{\mathbb{N}}.$$

Then $\Lambda_n = \prod_{k=1}^{n} \frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}(X_k)$ forms a non-negative martingale with respect to $\mathscr{F}_n = \sigma(X_1, \dots, X_n)$ under $\mathbb{P}$ as in Kakutani's Theorem 4.78. Lemma 4.82 yields

$$\mathbb{E}_{\mathbb{P}}\Big[\Big(\frac{d\,\mathbb{Q}_n}{d\,\mathbb{P}_n}(X_n)\Big)^{1/2}\Big] = \mathbb{E}_{\mathbb{P}_n}\Big[\Big(\frac{d\,\mathbb{Q}_n}{d\,\mathbb{P}_n}\Big)^{1/2}\Big] = 1 - \tfrac{1}{2}H^2(\mathbb{P}_n, \mathbb{Q}_n),$$

The condition in Theorem 4.78(e) is satisfied if and only if $\sum_{n=1}^{\infty} H^2(\mathbb{P}_n, \mathbb{Q}_n) < \infty$. In that case (Λ_n) is a closable martingale with $\Lambda_n = \mathbb{E}[\Lambda_\infty \,|\, \mathscr{F}_n]$. Set

$$\mathbb{Q}(A) := \int_A \Lambda_\infty \, d\,\mathbb{P}, \quad A \in \mathscr{F}^{\otimes \mathbb{N}}.$$

Then $\mathbb{Q}$ is a probability measure (note $\mathbb{E}_{\mathbb{P}}[\Lambda_\infty] = 1$ by Kakutani) and for $A \in \mathscr{F}_n$, i.e. $A = (X_1, \dots, X_n)^{-1}(B)$ for some $B \in \mathscr{F}^{\otimes n}$, we have

$$\begin{aligned}\mathbb{Q}(A) &= \mathbb{E}_{\mathbb{P}}[\mathbb{E}_{\mathbb{P}}[\Lambda_\infty | \mathscr{F}_n]\,\mathbf{1}_A] = \mathbb{E}_{\mathbb{P}}[\Lambda_n \mathbf{1}_B(X_1, \dots, X_n)] \\ &= \mathbb{E}_{\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n}\Big[\frac{d(\mathbb{Q}_1 \otimes \cdots \otimes \mathbb{Q}_n)}{d(\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n)}\mathbf{1}_B\Big] = (\mathbb{Q}_1 \otimes \cdots \otimes \mathbb{Q}_n)(B).\end{aligned}$$

Consequently, $\mathbb{Q}$ satisfies the definition of the product measure $\bigotimes_{n=1}^{\infty} \mathbb{Q}_n$ (Stochastik 1) and by uniqueness ($\bigcup_n \mathscr{F}_n$ is an $\cap$-stable generator of $\mathscr{F}^{\otimes \mathbb{N}}$) $\mathbb{Q} = \bigotimes_{n=1}^{\infty} \mathbb{Q}_n$ follows. With Λ_∞ we have exhibited its density with respect to $\mathbb{P}$ such that $\mathbb{Q} \ll \mathbb{P}$ holds.

In the case $\sum_{n=1}^{\infty} H^2(\mathbb{P}_n, \mathbb{Q}_n) = \infty$ Kakutani's Theorem 4.78 gives $\Lambda_\infty = 0$ $\mathbb{P}$-a.s. Consider

$$\Lambda_n' := \prod_{k=1}^{n} \Big(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}(X_k)\Big)^{-1} \mathbf{1}\Big(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}(X_k) > 0\Big).$$

Then (Λ_n') forms a non-negative supermartingale with respect to $\mathbb{Q} = \bigotimes_{n=1}^{\infty} \mathbb{Q}_n$ and $(\mathscr{F}_n)$ due to the independence of (X_k) and

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\Big[\Big(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}(X_k)\Big)^{-1}\mathbf{1}\Big(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}(X_k) > 0\Big)\Big] &= \mathbb{E}_{\mathbb{P}_k}\Big[\Big(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}\Big)^{-1}\mathbf{1}\Big(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k} > 0\Big)\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}\Big] \\ &= \mathbb{P}_k\Big(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k} > 0\Big) \in [0,1].\end{aligned}$$

By the first martingale convergence Theorem 4.43, $\Lambda_n' \to \Lambda_\infty'$ holds $\mathbb{Q}$-a.s. with some $\Lambda_\infty' \in L^1(\mathbb{Q})$. Because of

$$\mathbb{Q}\Big(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}(X_k) = 0\Big) = \mathbb{Q}_k\Big(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k} = 0\Big) = \int_{\{\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k} = 0\}} \frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k}\, d\,\mathbb{P}_k = 0,$$

we have $\Lambda_n'\Lambda_n = 1$ $\mathbb{Q}$-a.s. This implies $\mathbb{Q}(\Lambda_n \to 0) = \mathbb{Q}(\Lambda_n' \to \infty) = 0$. Together with $\Lambda_\infty = 0$ $\mathbb{P}$-a.s., i.e. $\mathbb{P}(\Lambda_n \to 0) = 1$, this shows $\mathbb{P} \perp \mathbb{Q}$. □

4.85 Example. Consider i.i.d. random variables $(X_k)_{k\geqslant 1}$ and $(Y_k)_{k\geqslant 1}$ with laws $\mathbb{P}^X := \mathbb{P}^{X_k}$, $\mathbb{P}^Y := \mathbb{P}^{Y_k}$, which are different, but equivalent. Then $\sum_{k=1}^{\infty} H^2(\mathbb{P}^{X_k}, \mathbb{P}^{Y_k}) = \infty$ holds and the proof shows that we have by symmetry $\Lambda_n := \frac{d\,\mathbb{P}^{(Y_1,\dots,Y_n)}}{d\,\mathbb{P}^{(X_1,\dots,X_n)}} \to 0$ $\mathbb{P}^{(X_k)_k}$-a.s. and $\Lambda_n \to \infty$ $\mathbb{P}^{(Y_k)_k}$-a.s. The *likelihood process* (Λ_n) for n observations is therefore a natural criterion to discriminate between $\mathbb{P}^{(X_k)_k}$ and $\mathbb{P}^{(Y_k)_k}$ (it gives even rise to optimal tests, compare the Neyman-Pearson Lemma in statistics).

▷ **Control questions**

(a) Assume even $\mathbb{P}^n \sim \mathbb{Q}^n$ in Corollary 4.83. Show that Λ_n' in the proof is the density of $\bigotimes_{k=1}^n \mathbb{P}_k$ with respect to $\bigotimes_{k=1}^n \mathbb{Q}_k$. Can you interpret Λ_n' without this assumption as a density of some measure? *Hint:* Lebesgue decomposition.

Since Λ_n is the density of $\bigotimes_{k=1}^n \mathbb{Q}_k$ with respect to $\bigotimes_{k=1}^n \mathbb{P}_k$ ►Exercise shows that $\Lambda_n' = \Lambda_n^{-1}$ is $\bigotimes_{k=1}^n \mathbb{Q}_k$-a.s. the density of $\bigotimes_{k=1}^n \mathbb{P}_k$ with respect to $\bigotimes_{k=1}^n \mathbb{Q}_k$. More generally, the Lebesgue decomposition shows that $\mathbb{P}_k = \mathbb{P}_{k,1} + \mathbb{P}_{k,2}$ with $\mathbb{P}_{k,1} \ll \mathbb{Q}_k$, $\mathbb{P}_{k,2} \perp \mathbb{Q}_k$ and $\mathbb{P}_{k,1}(A) = \mathbb{Q}_k(A \cap \{d\,\mathbb{Q}_k\,/d\,\mathbb{P}_k > 0\})$ (compare the proof and note $\mathbb{Q}_k \ll \mathbb{P}_k$). Hence, $\bigotimes_{k=1}^n \mathbb{P}_{k,1}$ has the density $\prod_{k=1}^n (\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k} \mathbf{1}(\frac{d\,\mathbb{Q}_k}{d\,\mathbb{P}_k} > 0))^{-1} = \Lambda_n'$ with respect to $\bigotimes_{k=1}^n \mathbb{Q}_k$. We can interpret Λ_∞' as the density of the infinite product measure $\bigotimes_{k=1}^\infty \mathbb{P}_{k,1}$ with respect to $\bigotimes_{k=1}^\infty \mathbb{Q}_k$.

5 Ergodic theory and Markov chains

5.1 Example (Ehrenfest model III)**.** Recall the Ehrenfest model $(X_n, n \geqslant 0)$ on the state space $S = \{0, 1, \dots, N\}$ from Examples 1.15 and 1.19. The Binomial distribution $\mu = \mathrm{Bin}(N, 1/2)$ is an invariant initial distribution for this Markov chain, but the X_n fluctuate randomly in n. The so called *ergodic hypothesis* in physics is that the relative frequencies (or the time average) $\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{1}(X_t = j)$, $T \in \mathbb{N}$, nevertheless converges for $T \to \infty$ almost surely to the invariant probability (or space average) $\mu(\{j\})$ for all states $j \in S$. Note that starting with the initial distribution μ we clearly have $\mathbb{E}[\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{1}(X_t = j)] = \mu(\{j\})$, but the X_t are neither independent nor martingale differences. We thus need to establish a type of strong law for large numbers under a different concept called ergodicity.

What is more, under additional assumptions, satisfied by the Ehrenfest model, the T-step transition probabilities $p_{ij}(T) = \mathbb{P}(X_T = j \mid X_0 = i)$ converge to $\mu(\{j\})$ as $T \to \infty$ regardless of the initial condition $X_0 = i$. These Markov chains *forget* their initial condition asymptotically and their laws approximate the invariant distribution for large T, which is exploited by Markov chain Monte Carlo (MCMC) methods.

5.1 Stationary and ergodic processes

5.2 Definition. A stochastic process $(X_t,\, t \in T)$ with $T \in \{\mathbb{N}_0, \mathbb{Z}, \mathbb{R}^+, \mathbb{R}\}$ is stationary (stationär) if $(X_{t_1}, \dots, X_{t_n}) \overset{d}{=} (X_{t_1+s}, \dots, X_{t_n+s})$ (equality of the laws) holds for all $n \geqslant 1$, $t_1, \cdots, t_n \in T$ and $s \in T$.

5.3 Example.

(a) A time-homogeneous Markov chain is stationary if and only if its initial distribution is invariant. ▶Exercise

(b) Let $X_t = A\sin(\omega t+U)$, $t \geqslant 0$, with $A, \omega \in \mathbb{R}$ and $U \sim U([0, 2\pi])$ (periodic signal with random phase). Then $X_t = \operatorname{Im}(Z_t)$ holds with $Z_t = Ae^{i(\omega t+U)}$ and
$$(Z_{t_1+s}, \dots, Z_{t_n+s}) = Ae^{i(\omega s+U)}(e^{i\omega t_1}, \dots, e^{i\omega t_n}).$$
Since $e^{i(\omega s+U)} \overset{d}{=} e^{iU}$ is uniformly distributed on the unit sphere S^1, we conclude $(Z_{t_1+s}, \dots, Z_{t_n+s}) \overset{d}{=} (Z_{t_1}, \dots, Z_{t_n})$ and (Z_t) is stationary. This implies that also (X_t) is stationary.

(c) Let $(X_t, t \in T)$ be a Gaussian process with expectation function $\mu(t)$ and covariance function $c(t, s)$. If (X_t) is stationary, then $X_t \overset{d}{=} X_s$ implies $\mu(t) = \mu(s)$ and μ must be constant. Furthermore, $(X_t, X_{t+u}) \overset{d}{=} (X_s, X_{s+u})$ implies $c(t, t + u) = c(s, s + u)$ and the covariance function satisfies $c(t, s) = c(0, |t - s|)$. It is easy to see that these properties of μ and c conversely imply that (X_t) is stationary.

More generally, any process (X_t) with $X_t \in L^2$, constant expectation function $\mu(t)$ and covariance function $c(t, s) = c(0, |t-s|)$, only depending on the distance of time points, is called <u>weakly stationary</u>. It need not be stationary in our (strict) sense ▶Control.

5.4 Definition. A measurable map $T : \Omega \to \Omega$ on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is called <u>measure-preserving</u> (maßerhaltend) if $\mathbb{P}(T^{-1}(A)) = \mathbb{P}(A)$ holds for all $A \in \mathscr{F}$.

5.5 Remark. The next lemma shows that measure-preserving maps generate stationary processes and stationary processes induce a measure-preserving left shift on path space.

5.6 Lemma. *Let $(S, \mathcal{S})$ be a measurable space.*

(a) Every S-valued stationary process $X = (X_n, n \geqslant 0)$ induces a measure-preserving transformation ϑ on $(S^{\mathbb{N}_0}, \mathcal{S}^{\otimes \mathbb{N}_0}, \mathbb{P}^X)$ via
$$\vartheta((x_0, x_1, x_2, \dots)) = (x_1, x_2, \dots) \text{ (\underline{left shift})}.$$

(b) For an S-valued random variable Y and a measure-preserving map T on $(\Omega, \mathscr{F}, \mathbb{P})$ the process $X_n(\omega) := Y(T^n(\omega))$, $n \geqslant 0$, ($T^0 := \mathrm{Id}$) is stationary.

Proof.

(a) Let $\pi_{\{0,\dots,n\}}((\omega_k)_{k\geqslant 0}) = (\omega_0, \dots, \omega_n)$ denote the projection onto the first $(n + 1)$ coordinates. Consider a cylinder set $A = \pi^{-1}_{\{0,\dots,n\}}(B_n)$ with $B_n \in \mathcal{S}^{\otimes(n+1)}$. Then
$$\mathbb{P}^X(A) = \mathbb{P}((X_0, \dots, X_n) \in B_n) \overset{!}{=} \mathbb{P}((X_1, \dots, X_{n+1}) \in B_n) = \mathbb{P}(\vartheta{\circ}X \in A)$$

follows from $(X_0,\dots,X_n) \overset{d}{=} (X_1,\dots,X_{n+1})$. Since the cylinder sets form an $\cap$-stable generator of $\mathcal{S}^{\otimes \mathbb{N}_0}$ the probability measures $\mathbb{P}^X$ and $\mathbb{P}^{\vartheta\circ X}$ coincide on $\mathcal{S}^{\otimes \mathbb{N}_0}$ and ϑ is measure-preserving, writing $\mathbb{P}^X(A) = \mathbb{P}^{\vartheta\circ X}(A) = \mathbb{P}(\vartheta\circ X \in A) = \mathbb{P}(X \in \vartheta^{-1}(A)) = \mathbb{P}^X(\vartheta^{-1}(A))$.

(b) We obtain for $A \in \mathcal{S}^{\otimes n}$ and $0 \leqslant t_1 < \dots < t_n$, using measure preservation $\mathbb{P}(T^{-m}(\bullet)) = \mathbb{P}(\bullet)$:

$$\begin{aligned}
\mathbb{P}((X_{t_1+m},\dots,X_{t_n+m}) \in A) &= \mathbb{P}((Y\circ T^{t_1+m},\dots,Y\circ T^{t_n+m}) \in A)\\
&= \mathbb{P}(T^{-m}((Y\circ T^{t_1},\dots,Y\circ T^{t_n})^{-1}(A)))\\
&\overset{!}{=} \mathbb{P}((Y\circ T^{t_1},\dots,Y\circ T^{t_n})^{-1}(A))\\
&= \mathbb{P}((X_{t_1},\dots,X_{t_n}) \in A).
\end{aligned}$$

□

5.7 Definition. An event A is (almost) invariant with respect to a measure-preserving map T on $(\Omega,\mathscr{F},\mathbb{P})$ if $T^{-1}(A) = A$ $\mathbb{P}$-a.s., that is $\mathbb{P}(T^{-1}(A)\Delta A) = 0$ holds with the symmetric difference $A_1\Delta A_2 := (A_1\setminus A_2)\cup(A_2\setminus A_1)$.

The σ-algebra▶Control of all (almost) invariant events is denoted by $\mathscr{I}_T$. T is ergodic if $\mathscr{I}_T$ is trivial, i.e. $\mathscr{I}_T = \{A \in \mathscr{F} \mid \mathbb{P}(A) \in \{0,1\}\}$ holds. A stationary process whose left shift ϑ in Lemma 5.6(a) is ergodic will be called ergodic process.

5.8 Remark. Null and one sets are always invariant events. An ergodic transformation leaves no other events fixed.

5.9 Example. Consider $\Omega = \{1,2,3,4,5\}$, $\mathscr{F} = \mathcal{P}(\Omega)$ and the permutation $T \in S_5$ with cycles $T = (1,2,3)(4,5)$ (i.e., $1 \overset{T}{\mapsto} 2 \overset{T}{\mapsto} 3 \overset{T}{\mapsto} 1$ and $4 \overset{T}{\mapsto} 5 \overset{T}{\mapsto} 4$). Then T preserves the measure $\mathbb{P}$ whenever $\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\})$ and $\mathbb{P}(\{4\}) = \mathbb{P}(\{5\})$. If all probabilities are non-zero, then $\mathscr{I}_T = \{\varnothing,\Omega,\{1,2,3\},\{4,5\}\}$ holds. For $X_n(\omega) := \mathbf{1}_{\{1,4\}}(T^n(\omega))$ we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} X_i(\omega) = \begin{cases} 1/3, & \text{if } \omega \in \{1,2,3\},\\ 1/2, & \text{if } \omega \in \{4,5\}.\end{cases}$$

Note that the limit can be written as $\mathbb{E}[X_0 \mid \mathscr{I}_T]$. This is a first concrete example of an ergodic theorem.

5.10 Lemma. *Let $\mathscr{I}_T$ be the invariant σ-algebra with respect to some measure-preserving transformation T on $(\Omega,\mathscr{F},\mathbb{P})$. Then:*

(a) A (real-valued) random variable Y is $\mathscr{I}_T$-measurable if and only if it is $\mathbb{P}$-a.s. invariant, i.e. $\mathbb{P}(Y\circ T = Y) = 1$. In particular, T is ergodic if and only if each $\mathbb{P}$-a.s. invariant and bounded random variable is $\mathbb{P}$-a.s. constant.

(b) For each invariant event $A \in \mathscr{I}_T$ there exists a strictly invariant event B (i.e. with $T^{-1}(B) = B$ exactly) such that $\mathbb{P}(A\Delta B) = 0$. In particular, T is ergodic if $\mathbb{P}(B) \in \{0,1\}$ holds for any strictly invariant set B.

Proof. ▶Exercise □

5.11 Example.

(a) Suppose $(X_n)_{n\geqslant 0}$ are i.i.d. $(S,\mathscr{S})$-valued random variables. Then they form a stationary process X. Moreover, $\vartheta^{-1}(A)=A$ for the left shift ϑ on $(S^{\mathbb{N}_0},\mathcal{S}^{\otimes\mathbb{N}_0},\mathbb{P}^X)$ implies $A\in\sigma(\pi_k,k\geqslant n)$ for all $n\in\mathbb{N}$ with the projection π_k on the k-th coordinate because for $\vartheta_n=\vartheta\circ\cdots\circ\vartheta$ (n-fold composition, n-fold left shift) we have $\omega\in\vartheta_n^{-1}(A)\iff(\pi_n(\omega),\pi_{n+1}(\omega),\ldots)\in A$ and thus $A=\vartheta^{-1}(A)=\cdots=\vartheta_n^{-1}(A)\in\sigma(\pi_k,\,k\geqslant n)$. This means that A lies in the terminal (asymptotic) σ-algebra of (π_n). Since (π_n) under $\mathbb{P}^X$ are distributed as (X_n) under $\mathbb{P}$, they are independent and Kolmogorov's 0-1 law implies $\mathbb{P}(A)\in\{0,1\}$. By Lemma 5.10(b) ϑ is ergodic.

(b) Let $\Omega=[0,1)$, $\mathscr{F}=\mathfrak{B}_\Omega$, $\mathbb{P}$ Lebesgue measure on $[0,1)$ and $T(\omega)=(\omega+r)$ mod 1 for some fixed $r\in\mathbb{R}$ (model for rotation by an angle $2\pi r$; note x mod $1:=x-\lfloor x\rfloor$ is the non-integer part of x). Translation invariance of Lebesgue measure ensures that T is measure-preserving. There are two cases:

(i) $r=p/q\in\mathbb{Q}$ rational with $p,q\in\mathbb{N}$: Consider $A=\bigcup_{k=0}^{q-1}[k/q,(k+1/2)/q)$. Then $A=(A+p/q)$ mod 1 and A is T-invariant, but $\mathbb{P}(A)=1/2$ holds. T is not ergodic (it is *periodic*).

(ii) $r\in\mathbb{R}\setminus\mathbb{Q}$ irrational: Let $f:[0,1)\to\mathbb{R}$ be bounded, measurable and $\mathbb{P}$-a.s. invariant, i.e. $f\circ T=f$ Lebesgue-almost everywhere. Since $f\in L^2([0,1))$ holds, there is a Fourier series expansion $f(x)=\sum_{k\in\mathbb{Z}}c_ke^{2\pi ikx}$ in $L^2([0,1))$ with coefficients $(c_k)\in\ell^2$. This gives

$$f\circ T(x)=\sum_{k\in\mathbb{Z}}c_ke^{2\pi ik(x+r)}=\sum_{k\in\mathbb{Z}}(c_ke^{2\pi ikr})e^{2\pi ikx}.$$

The invariance $f=f\circ T$ and the uniqueness of the Fourier coefficients (c_k) implies $c_k=c_ke^{2\pi ikr}$ for all $k\in\mathbb{Z}$. Due to $e^{2\pi ikr}\neq 1$ for $k\in\mathbb{Z}\setminus\{0\}$ (r is irrational!), we infer $c_k=0$ for $k\neq 0$ and f must be $\mathbb{P}$-a.s. constant. By Lemma 5.10(a) T is ergodic. This is Weyl's Equidistribution Theorem (1909), which has many connections to number theory, harmonic analysis and pseudo-random number generation.

▷ **Control questions**

(a) Find a weakly stationary process which is not (strictly) stationary in our sense. *Hint:* Use random variables with the same expectation and variance, but different laws.

Take $X_{2m}\sim\mathrm{Exp}(1)$ and $X_{2m+1}\sim N(1,1)$, $m\geqslant 0$, all independent. Then $\mathbb{E}[X_n]=1$, $\mathrm{Var}(X_n)=1$ and $\mathrm{Cov}(X_n,X_{n+k})=0$ for all $n,k\geqslant 0$ and $(X_n)_{n\geqslant 0}$ is a weakly stationary process. It is not (strictly) stationary because X_n and X_{n+1} have different laws.

(b) Check that $\mathscr{I}_T$ forms a σ-algebra.

Obviously, $\Omega \in \mathscr{I}_T$ and for $A \in \mathscr{I}_T$ we have $T^{-1}(A^C) = (T^{-1}(A))^C = A^C$ $\mathbb{P}$-a.s. and thus $A^C \in \mathscr{I}_T$. Finally, for $A_n \in \mathscr{I}_T$ the rules for preimages give $T^{-1}(\bigcup_n A_n) = \bigcup_n T^{-1}(A_n) = \bigcup_n A_n$ $\mathbb{P}$-a.s. and $\mathscr{I}_T$ is indeed a σ-algebra.

(c) For which measures $\mathbb{P}$ is T in Example 5.9 ergodic? Conclude that then $\frac{1}{n}\sum_{i=0}^{n-1} X_i \to \mathbb{E}[X_0]$ $\mathbb{P}$-a.s.
Hint: Lemma 5.10(b).

The strictly invariant events are $\varnothing, \Omega, \{1,2,3\}, \{4,5\}$ and they have 0-1 probability if and only if $\mathbb{P}(\{1,2,3\}) \in \{0,1\}$. Hence, there are two probability measure $\mathbb{P}$ rendering T ergodic: $\mathbb{P}_1(\{1\}) = \mathbb{P}_1(\{2\}) = \mathbb{P}_1(\{3\}) = 1/3$ and $\mathbb{P}_2(\{4\}) = \mathbb{P}_2(\{5\}) = 1/2$, all other probabilities being zero. Under $\mathbb{P}_1$ we start in $\{1,2,3\}$ almost surely and $\frac{1}{n}\sum_{i=0}^{n-1} X_i \to 1/3 = \mathbb{E}_1[X_0]$ holds $\mathbb{P}_1$-a.s. In the same manner $\mathbb{P}_2(\omega \in \{4,5\}) = 1$ and $\frac{1}{n}\sum_{i=0}^{n-1} X_i \to 1/2 = \mathbb{E}_2[X_0]$ holds $\mathbb{P}_2$-a.s.

5.2 Recurrence, transience and ergodicity of Markov chains

In this section $(X_n,\, n \geqslant 0)$ always denotes a time-homogeneous Markov chain with discrete state space $(S, \mathcal{P}(S))$, realized as coordinate process on $\Omega = S^{\mathbb{N}_0}$ with product σ-algebra $\mathscr{F} = \mathcal{P}(S)^{\otimes \mathbb{N}_0}$, filtration $\mathscr{F}_n = \sigma(X_0, \dots, X_n)$ and measure $\mathbb{P}_\mu$, where μ denotes the initial distribution of X_0. We write short $\mathbb{P}_x := \mathbb{P}_{\delta_x}$ for a Markov chain starting in $x \in S$. In particular, we have $\mathbb{P}_i(X_n = j) = p_{ij}(n)$, the n-step transition probability from i to j.

5.12 Proposition (Generalised Markov property)**.** *For a non-negative random variable $Y : \Omega \to \mathbb{R}$ and the n-fold left shift $\vartheta_n = \vartheta^n$ on Ω we have*

$$\mathbb{E}_\mu[Y \circ \vartheta_n \,|\, \mathscr{F}_n] = h(X_n)\ \mathbb{P}_\mu\text{-a.s. for } h(x) := \mathbb{E}_x[Y].$$

5.13 Remark. This abstract formulation generalises the Markov property from probabilities to expectations and allows in particular that Y may depend on all future values of X_n, not only on finitely many. The result extends to $Y \in \bigcap_{x\in S} L^1(\mathbb{P}_x)$ (then also $Y \in L^1(\mathbb{P}_\mu)$). From now on we shall write $h(X_n) = \mathbb{E}_{X_n}[Y]$.

Proof. First let $Y = \mathbf{1}(X_0 = i_0, \dots, X_m = i_m)$ for some $i_0, \dots, i_m \in S$. Then for $A \in \mathscr{F}_n$, i.e. $A = \{(X_0, \dots, X_n) \in B\}$ for some $B \subseteq S^{n+1}$:

$$\begin{aligned}
&\mathbb{E}_\mu[(Y \circ \vartheta_n)\mathbf{1}_A] = \mathbb{E}_\mu[\mathbf{1}(X_n = i_0, \dots, X_{m+n} = i_m)\mathbf{1}((X_0, \dots, X_n) \in B)] \\
&= \sum_{(b_0,\dots,b_n)\in B} \mathbb{P}_\mu(X_0 = b_0, \dots, X_n = b_n)\times \\
&\qquad\qquad \mathbb{P}_\mu(X_n = i_0, \dots, X_{m+n} = i_m \,|\, X_0 = b_0, \dots, X_n = b_n) \\
&= \sum_{(b_0,\dots,b_n)\in B} \mathbb{P}_\mu(X_0 = b_0, \dots, X_n = b_n)\, \mathbb{P}_\mu(X_n = i_0, \dots, X_{m+n} = i_m \,|\, X_n = b_n) \\
&= \sum_{(b_0,\dots,b_n)\in B} \mathbb{P}_\mu(X_0 = b_0, \dots, X_n = b_n)\, \mathbb{P}_{b_n}(X_0 = i_0, \dots, X_m = i_m) \\
&= \sum_{(b_0,\dots,b_n)\in B} \mathbb{E}_\mu[\mathbf{1}(X_0 = b_0, \dots, X_n = b_n)\, \mathbb{E}_{b_n}[Y]] \\
&= \sum_{(b_0,\dots,b_n)\in B} \mathbb{E}_\mu[\mathbf{1}(X_0 = b_0, \dots, X_n = b_n)\, \mathbb{E}_{X_n}[Y]] = \mathbb{E}_\mu[\mathbb{E}_{X_n}[Y]\mathbf{1}_A],
\end{aligned}$$

where we used the Markov property in line 3 and the time homogeneity in line 4. Taking sums over all elements of $D_m \subseteq S^{m+1}$, the identity extends to $Y = \mathbf{1}_{C_m}$ with $C_m = (X_0, \dots, X_m)^{-1}(D_m) \in \mathscr{F}_m$:

$$\mathbb{E}_\mu[(\mathbf{1}_{C_m} \circ \vartheta_n)\mathbf{1}_A] = \mathbb{E}_\mu[\mathbb{E}_{X_n}[\mathbf{1}((X_0, \dots, X_m) \in D_m)]\mathbf{1}_A] = \mathbb{E}_\mu[\mathbb{E}_{X_n}[\mathbf{1}_{C_m}]\mathbf{1}_A].$$

Therefore the finite measures $C \mapsto \mathbb{E}_\mu[(\mathbf{1}_C \circ \vartheta_n)\mathbf{1}_A]$ and $C \mapsto \mathbb{E}_\mu[\mathbb{E}_{X_n}[\mathbf{1}_C]\mathbf{1}_A]$, $C \in \mathscr{F}$, coincide on the $\cap$-stable generator $\bigcup_{m\geqslant 0} \mathscr{F}_m$, containing the full set Ω. By the uniqueness theorem for measures, they are equal. This establishes $\mathbb{E}_\mu[Y \circ \vartheta_n \,|\, \mathscr{F}_n] = \mathbb{E}_{X_n}[Y]$ for all indicators $Y = \mathbf{1}_C$, $C \in \mathscr{F}$. It remains to apply measure-theoretic induction (take linear combinations, monotone limits) to obtain the result for any non-negative measurable Y. □

5.14 Remark. What will become very useful in the sequel is that the preceding result also extends to stopping times τ, in its easiest form we shall obtain $\mathbb{E}_\mu[f(X_{\tau+1}) \,|\, \mathscr{F}_\tau] = \mathbb{E}_{X_\tau}[f(X_1)]$, letting $Y_n = f(X_1)$ below, which intuitively means that after τ the Markov chain evolves by the standard transitions starting from the initial value X_τ.

5.15 Proposition (Strong Markov property)**.** *For a stopping time τ, the σ-algebra $\mathscr{F}_\tau$ of τ-history and a non-negative random variable $Y : \Omega \to \mathbb{R}$ we have*

$$\mathbb{E}_\mu[Y \circ \vartheta_\tau \,|\, \mathscr{F}_\tau] = \mathbb{E}_{X_\tau}[Y]\ \mathbb{P}_\mu\text{-a.s. on } \{\tau < \infty\},$$

i.e. $\mathbb{E}_\mu[(Y \circ \vartheta_\tau)\mathbf{1}(\tau < \infty) \,|\, \mathscr{F}_\tau] = \mathbb{E}_{X_\tau}[Y]\mathbf{1}(\tau < \infty)\ \mathbb{P}_\mu$*-a.s.*

Proof. Let $A \in \mathscr{F}_\tau$. Then splitting into the events $\{\tau = n\}$ we obtain due to

$A\cap\{\tau=n\}\in\mathscr{F}_n$ by definition of $\mathscr{F}_\tau$:

$$\begin{aligned}\mathbb{E}_\mu[(Y\circ\vartheta_\tau)\mathbf{1}(A\cap\{\tau<\infty\})]&=\sum_{n=0}^{\infty}\mathbb{E}_\mu[(Y\circ\vartheta_n)\mathbf{1}(A\cap\{\tau=n\})]\\&=\sum_{n=0}^{\infty}\mathbb{E}_\mu[\mathbb{E}[Y\circ\vartheta_n\,|\,\mathscr{F}_n]\mathbf{1}(A\cap\{\tau=n\})]\\&=\sum_{n=0}^{\infty}\mathbb{E}_\mu[\mathbb{E}_{X_n}[Y]\mathbf{1}(A\cap\{\tau=n\})]\\&=\mathbb{E}_\mu\big[\mathbb{E}_{X_\tau}[Y]\mathbf{1}(A\cap\{\tau<\infty\})\big],\end{aligned}$$

using the generalised Markov property, which yields the assertion. □

5.16 Definition. For $y\in S$ let $\tau_y^0:=0$ and for $k\geqslant 1$

$$\tau_y^k:=\inf\{n>\tau_y^{k-1}\,|\,X_n=y\}\in\mathbb{N}\cup\{+\infty\}$$

the time of the kth return to y, setting $\tau_y:=\tau_y^1$. For $x,y\in S$ define

$$\rho_{xy}:=\mathbb{P}_x(\tau_y<\infty)$$

and call the state y recurrent if $\rho_{yy}=1$ and transient if $\rho_{yy}<1$.

5.17 Remark. Note that $\tau_y=\inf\{n\geqslant 1\,|\,X_n=y\}$ is the first time *after* time zero to return to y. A Markov chain starting in a recurrent state y almost surely returns to it. The next result below, which is essentially based on the strong Markov property, shows that the Markov chain then even returns infinitely often to y and that states x with a positive probability of being visited after y will then also be recurrent (*recurrence is infectious*). Later we shall see that stationary Markov chains only visit recurrent states.

5.18 Example. Consider a Markov chain on $S=\{1,2,3\}$ with one-step transition matrix

$$P(1)=\begin{pmatrix}* & * & 0\\ 0 & * & *\\ 0 & * & *\end{pmatrix}\text{ where $*$ denotes a non-zero entry.}$$

Then the state 1 is transient because from states 2 and 3 the chain does not return to 1: $\mathbb{P}_1(\tau_1=\infty)\geqslant\mathbb{P}_1(X_1=2)>0$. The states 2 and 3 are recurrent, e.g $\mathbb{P}_2(\tau_2=\infty)=\mathbb{P}_2(\forall n\geqslant 1: X_n=3)=\lim_{N\to\infty}p_{23}(1)\prod_{n=2}^{N}p_{33}(1)=0$ because $p_{33}(1)=1-p_{32}(1)<1$. We note further $\rho_{12}=\rho_{13}=\rho_{23}=1$, $\rho_{21}=\rho_{31}=0$▶Control.

5.19 Theorem. *Consider states $x,y\in S$ and let $N(y):=\sum_{n=1}^{\infty}\mathbf{1}(X_n=y)$ be the number of visits to y. Then:*

(a) $\mathbb{P}_x(\tau_y^k<\infty)=\rho_{xy}\rho_{yy}^{k-1}$, in particular $\mathbb{P}_y(X_n=y$ infinitely often$)=1$ holds for recurrent y.

(b) *If y is transient, then $\mathbb{E}_x[N(y)] = \frac{\rho_{xy}}{1-\rho_{yy}} < \infty$ holds for all x. If y is recurrent, then $\mathbb{E}_y[N(y)] = \infty$ holds.*

(c) *If y is recurrent and $\rho_{yx} > 0$, then x is recurrent and $\rho_{xy} = \rho_{yx} = 1$.*

5.20 Remark. While the formal proof is relatively abstract, the results are very intuitive: (a) says that for visiting y k times after start in x we have first to go to y once and then $k-1$ times from y to y; from that it follows that the probability of visiting y at least k times is $\rho_{xy}\rho_{yy}^{k-1}$ and summing over k yields the expected value in (b); for (c) we remark that for recurrent y and $\rho_{yx} > 0$ we must return from x to y a.s. ($\rho_{xy} = 1$ to avoid going from y to x without ever returning to y) and then we shall also return from x to x via y a.s.

▷ **Control questions**

(a) How does the original Markov property follow from the generalised Markov property?

Put $Y = \mathbf{1}(X_1 = j)$ for some $j \in S$. Then:

$\mathbb{E}_\mu[\mathbf{1}(X_{n+1} = j) \,|\, X_0, \dots, X_n] = \mathbb{E}_\mu[Y \circ \vartheta_n \,|\, \mathscr{F}_n] = \mathbb{E}_{X_n}[Y] = \mathbb{P}_{X_n}(X_1 = j)$.

Evaluating the conditional expectation on $(X_0 = i_0, \dots, X_n = i_n)$ yields

$$\begin{aligned}\mathbb{P}(X_{n+1} = j \,|\, X_0 = i_0, \dots, X_n = i_n) &= \mathbb{P}(X_1 = j \,|\, X_0 = i_n)\\ &= \mathbb{P}(X_{n+1} = j \,|\, X_n = i_n),\end{aligned}$$

where we omit the subscript μ, only looking at values i_0 with $\mathbb{P}_\mu(X_0 = i_0) > 0$ so that everything is well defined.

(b) For the state space $S = \{1, 2, 3\}$ construct examples of Markov chains with k recurrent and $3-k$ transient states for $k = 0, 1, 2, 3$ whenever possible.

Example 5.18(a) can be used for $k = 2$. Example 5.18(c) proves that there is always a recurrent state so that $k = 0$ cannot happen. For $k = 3$ take a Markov chain where all states are connected, i.e. $\rho_{xy} > 0$ for all $x, y \in S$ (simplest example $p_{xy} = 1/3$ for all x, y). For $k = 1$ let $p_{13} = p_{2,3} = p_{3,3} = 1$ and all other transition probabilities zero. Then the states 1 and 2 are transient, while 3 is recurrent.

(c) Check the values of ρ_{xy} in Example 5.18(a).

Abstractly, this follows from Theorem 5.19. Concretely, we can argue that $1-\rho_{12} = 1-\rho_{13} = \mathbb{P}_1(\forall n : X_n = 1) = \lim_{N\to\infty} p_{11}^N = 0$ since $p_{11} = 1-p_{12} < 1$. Similarly, $1 - \rho_{23} = \mathbb{P}_2(\forall n : X_n = 2) = 0$ holds. On the other hand, $\rho_{21} = \mathbb{P}_2(\exists n : X_n = 1) = 0$ holds and also $\rho_{31} = \mathbb{P}_3(\exists n : X_n = 1) = 0$ since there is no path from state 2 or 3 to 1. A more formal argument would rely on the Chapman-Kolmogorov equations which give ($(1, 0, 0)^\top$ is an eigenvector)

$$P(n) = P(1)^n = \begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \end{pmatrix}, \quad n \geqslant 2,$$

thus implying for the n-step transition probabilities $p_{21}(n) = p_{31}(n) = 0$, $n \geqslant 1$, and hence $\mathbb{P}_2(\exists n : X_n = 1) = \mathbb{P}_3(\exists n : X_n = 1) = 0$.

Proof.

(a) Since the result for $k = 1$ is just the definition, we suppose $k \geqslant 2$. Let $Y = \mathbf{1}(\exists n \geqslant 1 : X_n = y) = \mathbf{1}(\tau_y < \infty)$. Then

$$Y \circ \vartheta_{\tau_y^{k-1}} = \mathbf{1}(\exists n \geqslant 1 : X_{n+\tau_y^{k-1}} = y) = \mathbf{1}(\tau_y^k < \infty)$$

holds. The strong Markov property implies on $\{\tau_y^{k-1} < \infty\}$

$$\mathbb{E}_x[Y \circ \vartheta_{\tau_y^{k-1}} \,|\, \mathscr{F}_{\tau_y^{k-1}}] = \mathbb{E}_{X_{\tau_y^{k-1}}}[Y] = \mathbb{E}_y[Y] = \mathbb{P}_y(\tau_y < \infty) = \rho_{yy}.$$

Taking expectations and noting $\mathbf{1}(\tau_y^k < \infty) = \mathbf{1}(\tau_y^k < \infty)\mathbf{1}(\tau_y^{k-1} < \infty)$ yield

$$\begin{aligned}\mathbb{P}_x(\tau_y^k < \infty) &= \mathbb{E}_x\Big[\mathbb{E}_x[\mathbf{1}(\tau_y^k < \infty) \,|\, \mathscr{F}_{\tau_y^{k-1}}]\mathbf{1}(\tau_y^{k-1} < \infty)\Big] \\ &= \mathbb{E}_x\Big[\mathbb{E}_x[Y \circ \vartheta_{\tau_y^{k-1}} \,|\, \mathscr{F}_{\tau_y^{k-1}}]\mathbf{1}(\tau_y^{k-1} < \infty)\Big] \\ &= \mathbb{E}_x[\rho_{yy}\mathbf{1}(\tau_y^{k-1} < \infty)] = \rho_{yy}\,\mathbb{P}_x(\tau_y^{k-1} < \infty).\end{aligned}$$

The claimed identity now follows by induction using $\mathbb{P}_x(\tau_y^1 < \infty) = \rho_{xy}$. If y is recurrent, then $\mathbb{P}_y(\tau_y^k < \infty) = \rho_{yy}^k = 1$ for all $k \geqslant 1$. By intersection $\mathbb{P}_y(X_n = y \text{ infinitely often}) = \mathbb{P}_y(\forall k \geqslant 1 : \tau_y^k < \infty) = 1$ follows.

(b) We note the identity $\{N(y) \geqslant k\} = \{\tau_y^k < \infty\}$ and conclude for transient y by (a)

$$\mathbb{E}_x[N(y)] = \sum_{k=1}^{\infty} \mathbb{P}_x(N(y) \geqslant k) = \sum_{k=1}^{\infty} \mathbb{P}_x(\tau_y^k < \infty) = \sum_{k=1}^{\infty} \rho_{xy}\rho_{yy}^{k-1} = \frac{\rho_{xy}}{1-\rho_{yy}}.$$

If y is recurrent, then by (a) $N(y) = \infty$ holds $\mathbb{P}_y$-a.s. so that $\mathbb{E}_y[N(y)] = \infty$.

(c) Consider the non-trivial case $x \neq y$ and set $\tau = \tau_x \wedge \tau_y$. We use $\tau_y = \infty \iff \forall n \geqslant 0 : X_{n+\tau} \neq y$ on $\{\tau < \infty\}$, condition on $\mathscr{F}_\tau$ and obtain by the strong Markov property

$$\begin{aligned}0 = \mathbb{P}_y(\tau_y = \infty) \geqslant \mathbb{P}_y(\tau_y = \infty, \tau < \infty)& \\ &= \mathbb{E}_y[\mathbb{E}_y[\mathbf{1}(\forall\, n \geqslant 0 : X_n \circ \vartheta_\tau \neq y) \,|\, \mathscr{F}_\tau]\mathbf{1}(\tau < \infty)] \\ &= \mathbb{E}_y[\mathbb{E}_{X_\tau}[\mathbf{1}(\forall\, n \geqslant 0 : X_n \neq y)]\mathbf{1}(\tau < \infty)] \\ &= \mathbb{E}_y[\mathbb{E}_x[\mathbf{1}(\forall\, n \geqslant 0 : X_n \neq y)]\mathbf{1}(\tau_x < \infty)] \\ &= \mathbb{P}_x(\tau_y = \infty)\,\mathbb{P}_y(\tau_x < \infty) = (1-\rho_{xy})\rho_{yx},\end{aligned}$$

where we used that $\tau = \tau_y$ implies

$$\mathbb{E}_{X_\tau}[\mathbf{1}(\forall\, n \geqslant 0 : X_n \neq y)] = \mathbb{E}_y[\mathbf{1}(\forall\, n \geqslant 0 : X_n \neq y)] = 0$$

such that only the case $\tau = \tau_x$ contributes to the expectation. Hence, $\rho_{yx} > 0$ implies $\rho_{xy} = 1$.

Starting in y we can split the chain in excursions between τ_y^{k-1} and τ_y^k, $k \geqslant 1$, and obtain by the strong Markov property and $\mathbb{P}_y(\tau_y^{k-1} < \infty) = 1$

$$\begin{aligned}
\mathbb{E}_y[N(x)] &= \sum_{k\geqslant 1} \mathbb{E}_y\Big[\textstyle\sum_{n=\tau_y^{k-1}+1}^{\tau_y^k} \mathbf{1}(X_n = x)\Big] \\
&= \sum_{k\geqslant 1} \mathbb{E}_y\Big[\Big(\textstyle\sum_{n=1}^{\tau_y} \mathbf{1}(X_n = x)\Big) \circ \vartheta_{\tau_y^{k-1}}\Big] \\
&= \sum_{k\geqslant 1} \mathbb{E}_y\Big[\textstyle\sum_{n=1}^{\tau_y} \mathbf{1}(X_n = x)\Big] \in \{0, \infty\}.
\end{aligned}$$

By assumption we have $\mathbb{P}_y(N(x) \geqslant 1) = \rho_{yx} > 0$ so that $E_y[N(x)] = \infty$ holds. Part (b) implies that x cannot be transient. Hence, x is recurrent and reversing the roles of x and y then also yields $\rho_{yx} = 1$.

□

5.21 Example.

(a) Let $S_0 = x$, $S_n = S_{n-1} + X_n$ for $n \geqslant 1$ with independent $(X_n)_{n\geqslant 1}$ and $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = -1) = 1 - p$ be a simple random walk starting in $x \in \mathbb{Z}$. If $p > 1/2$, then by the strong law of large numbers $\frac{1}{n}(S_n - x) \to \mathbb{E}[X_1] > 0$ a.s., hence $S_n \to +\infty$ a.s. This implies that S_n visits x only finitely many times and by Theorem 5.19(a) below this means that x is transient. A symmetric argument shows that also for $p < 1/2$ all states x are transient. For the symmetric random walk with $p = 1/2$, however, we shall see that all states are recurrent. We have $\mathbb{P}_x(\tau_y < \infty) = 1$ for all $y > x$ ►Exercise(put $b = y - x$ in the problem) and symmetrically for all $y < x$. This shows in particular $\mathbb{P}_x(\exists n_+, n_- \geqslant 1 : S_{n_+} = x + 1, S_{n_-} = x - 1) = 1$, but due to $S_n - S_{n-1} \in \{-1, +1\}$ this means that almost surely there is some $n \in \mathbb{N}$ between n_+ and n_- with $S_n = x$ such that x is recurrent.

(b) In the Ehrenfest model we have $\rho_{xy} > 0$ for all $x, y \in S = \{0, \dots, N\}$ (there is a monotone path from x to y in $|x - y|$ steps). If all states $y \in S$ were transient, then Theorem 5.19(b) below would imply $\sum_{y\in S} \mathbb{E}_x[N(y)] < \infty$. By definition, however, $\sum_{y\in S} N(y) = \sum_{n\geqslant 1} 1 = \infty$ holds, a contradiction. Hence, there exists a recurrent state y. By Theorem 5.19(c) we conclude that all states are recurrent.

Remark that in (b) we have derived the general result that on a finite state space there always exists at least one recurrent state, which by Example (a) is not always true on infinite state spaces.

5.22 Proposition. *If there exists an invariant initial distribution π, then all transient states y satisfy $\pi(\{y\}) = 0$.*

Proof. By Theorem 5.19(b) we have for transient y

$$\mathbb{E}_\pi[N(y)] = \sum_{x\in S} \pi(\{x\})\, \mathbb{E}_x[N(y)] = \sum_{x\in S} \pi(\{x\}) \frac{\rho_{xy}}{1 - \rho_{yy}} \leqslant \frac{1}{1 - \rho_{yy}} < \infty$$

because $\rho_{xy} \leqslant 1$ and π is a probability measure. By definition of $N(y)$ and invariance of π, however,

$$\mathbb{E}_\pi[N(y)] = \sum_{n=1}^{\infty} \mathbb{E}_\pi[\mathbf{1}(X_n = y)] = \sum_{n=1}^{\infty} \pi(\{y\}) \in \{0, \infty\}$$

and $\mathbb{E}_\pi[N(y)]$ is finite only if $\pi(\{y\}) = 0$. □

5.23 Remark. A stationary Markov chain thus almost surely only visits recurrent states. Transient states are only visited finitely often by any Markov chain so that their probability must decrease during the evolution of the chain. Note that recurrence is a property of the transition probabilities, not the initial distribution. For the existence of an invariant initial distribution recurrence of all states does clearly not suffice as the symmetric random walk on $\mathbb{Z}$ demonstrates▶Control.

We shall now investigate the σ-algebra $\mathscr{I}_\vartheta$ of invariant events for the left-shift ϑ of a stationary Markov chain. By the preceding result, we can restrict to recurrent Markov chains in the following sense.

5.24 Definition. We call a Markov chain recurrent if all its states are recurrent. For two recurrent states $x, y \in S$ we write $x \sim y$ if $\rho_{xy} > 0$ (by Theorem 5.19(c) equivalent to $\rho_{xy} = \rho_{yx} = 1$) and then say that x and y are connected (or communicate). If all states are pairwise connected, then the Markov chain is called irreducible.

5.25 Remark. By definition of recurrence $x \sim x$ holds and Theorem 5.19(c) shows symmetry $x \sim y \iff y \sim x$. By the strong Markov property▶Control we also have transitivity $x \sim y$, $y \sim z \Rightarrow x \sim z$ such that $\sim$ defines an equivalence relation with equivalence classes (or *connected components*) $[x] := \{y \in S \,|\, y \sim x\}$. By definition, we have $X_0 \in [x] \Rightarrow X_n \in [x]$ a.s. for all $n \in \mathbb{N}$, a recurrent Markov chain remains in one connected component all the time. If there are several equivalence classes, we can thus reduce the complete dynamics to one class depending on the initial state. This explains the notion of irreducibility.

5.26 Theorem. *Let $(X_n, n \geqslant 0)$ be a recurrent Markov chain with invariant initial distribution π and the σ-algebra $\mathscr{I}_\vartheta$ of invariant events for the left shift ϑ on $S^{\mathbb{N}_0}$. For $A \in \mathscr{I}_\vartheta$ define $B := \{x \in S \,|\, P_x(A) = 1\}$. Then $A = \{X_0 \in B\} = \{[X_0] \subseteq B\}$ holds $\mathbb{P}_\pi$-a.s.*

5.27 Remark. In more detail, the identities $A = \{(X_n)_{n\geqslant 0} \,|\, X_0 \in B\} = \{(X_n)_{n\geqslant 0} \,|\, [X_0] \subseteq B\}$ hold $\mathbb{P}_\pi$-a.s. In particular, the theorem shows that up to $\mathbb{P}_\pi$-null sets the invariant events A are contained in $\mathscr{F}_0$: modulo null sets an invariant event just describes the connected components in which the chain starts.

Proof. Suppose $A = \vartheta^{-1}(A)$, that is A is strictly invariant, which suffices for the proof by Lemma 5.10(b). Then $\mathbf{1}_A = \mathbf{1}_A \circ \vartheta_n$ holds for the n-fold left shift ϑ_n. The Markov property yields

$$\mathbb{E}_\pi[\mathbf{1}_A \,|\, \mathscr{F}_n] = \mathbb{E}_\pi[\mathbf{1}_A \circ \vartheta_n \,|\, \mathscr{F}_n] = \mathbb{E}_{X_n}[\mathbf{1}_A] = \mathbb{P}_{X_n}(A), \quad n \geqslant 0.$$

The left-hand side defines a bounded martingale. By the second martingale convergence theorem and $\mathscr{F} = \sigma(\bigcup_{n\in\mathbb{N}} \mathscr{F}_n)$ it converges $\mathbb{P}_\pi$-a.s. for $n \to \infty$ to $\mathbf{1}_A$ (*Lévy's 0-1 law*). This shows

$$\lim_{n\to\infty} \mathbb{P}_{X_n}(A) = \mathbf{1}_A((X_n)_{n\geqslant 0}) \in \{0,1\} \ \mathbb{P}_\pi\text{-a.s.}$$

By recurrence we have $\mathbb{P}_x(X_n = y \text{ infinitely often}) = 1$ for all $y \in [x]$. The convergence can therefore only take place if $\pi(\{x\}) > 0$ and $y \in [x]$ imply $\mathbb{P}_y(A) = \mathbb{P}_x(A)$. Consequently, $x \mapsto \mathbb{P}_x(A)$ is π-a.s. constant on all equivalence classes and there equal to zero or one. With $B = \{x \in S \mid \mathbb{P}_x(A) = 1\}$ the limit of the convergence yields $\mathbf{1}_A((X_n)_{n\geqslant 0}) = \mathbf{1}_B(X_0) = \mathbf{1}([X_0] \subseteq B)$ $\mathbb{P}_\pi$-a.s. □

5.28 Corollary. *A recurrent Markov chain with invariant initial distribution π is ergodic if and only if π-almost all states are connected ($\pi([x]) = 1$ for some $x \in S$).*

In particular, a recurrent and irreducible stationary Markov chain is ergodic.

Proof. If π-almost all states are connected, then $\mathbb{P}_\pi(C \subseteq [X_0]) = 1$ holds for the deterministic set $C = \{x \in S \,|\, \pi(\{x\}) > 0\}$. By Theorem 5.26 any invariant set A satisfies $\mathbb{P}_\pi(A = \{[X_0] \subseteq B\}) = 1$ with deterministic $B \subseteq S$. This implies $\mathbb{P}_\pi(A) = \mathbf{1}(C \subseteq B) \in \{0,1\}$ and the Markov chain is ergodic.

If there is a connected component $[x]$ with $\pi([x]) \in (0,1)$, consider the event $A := \{[X_0] = [x]\}$. Then $A = \{[X_1] = [x]\}$ $\mathbb{P}_\pi$-a.s. by connectivity and A is ϑ-invariant with $\mathbb{P}_\pi(A) = \pi([x]) \in (0,1)$. Hence, the Markov chain is not ergodic. □

▷ **Control questions**

(a) Assume that π is an invariant initial distribution for the symmetric random walk on $\mathbb{Z}$. Why does that mean $\pi(\{x\}) = \pi(\{y\})$ for all $x, y \in \mathbb{Z}$, implying that π cannot exist?

If π is invariant, then $\pi(\{x\}) = \pi(\{x-1\})p_{x-1,x} + \pi(\{x+1\})p_{x+1,x} = \frac{1}{2}(\pi(\{x-1\}) + \pi(\{x+1\})$ holds for all $x \in \mathbb{Z}$. This shows $\pi(\{x+1\}) - \pi(\{x\}) = \pi(\{x\}) - \pi(\{x-1\})$ and $x \mapsto \pi(\{x+1\}) - \pi(\{x\})$ is constant. If this constant were non-zero, then π would become negative at some $x \in \mathbb{Z}$. If it was zero, then $x \mapsto \pi(\{x\})$ would be constant, but this cannot give a probability measure on $\mathbb{Z}$ (total mass is zero or infinity).

(b) Show $\rho_{xy} = \rho_{yz} = 1 \Rightarrow \rho_{xz} = 1$ and thus transitivity of $\sim$.

We have $\tau_z \leqslant \tau_y + \tau_z \circ \vartheta_{\tau_y}$ (the right-hand side is the time to reach z via y). By assumption $\mathbb{P}_x(\tau_y < \infty) = \rho_{xy} = 1$ and by the strong Markov property (using $\mathbb{P}_x(\tau_y < \infty) = 1$ and $X_{\tau_y} = y$ $\mathbb{P}_x$-a.s.)

$$\begin{aligned}\mathbb{P}_x(\tau_z \circ \vartheta_{\tau_y} < \infty) &= \mathbb{E}_x[\mathbb{E}_x[\mathbf{1}(\tau_z \circ \vartheta_{\tau_y} < \infty) \,|\, \mathscr{F}_{\tau_y}]] = \mathbb{E}_x[\mathbb{P}_{X_{\tau_y}}(\tau_z < \infty)] \\ &= \mathbb{P}_y(\tau_z < \infty) = \rho_{yz} = 1.\end{aligned}$$

We conclude $\rho_{xz} = \mathbb{P}_x(\tau_z < \infty) = 1$.

(c) Why is the set $\mathscr{I}_\vartheta$ of invariant events huge for the left shift ϑ and does still have the simple structure of the preceding theorem?
Note: $\{\limsup_{n\to\infty} X_n \in B\}$ is strictly ϑ-invariant for any Borel set B.

The assertion that $\{\limsup_{n\to\infty} X_n \in B\}$ is strictly ϑ-invariant for any Borel set B follows as in the first step of the proof for Birkhoff's ergodic theorem. Hence, $\mathscr{I}_\vartheta$ has at least the cardinality of the Borel-σ-algebra, which is uncountable. Most invariant sets, however, will have $\mathbb{P}_\pi$-probability zero because they cannot be reached by a Markov chain. An example would be $B = \{x\}$ above for some $x \in S$. Then by recurrence $\{\limsup_{n\to\infty} X_n = x\} = \{X_0 \in [x]\}$ $\mathbb{P}_\pi$-a.s. if $y \leqslant x$ holds for all $y \in [x]$. If there is a $y > x$ with $y \sim x$, then $\mathbb{P}_\pi(\limsup_{n\to\infty} X_n = x) = 0$. The Markov structure and recurrence are sufficient to describe the invariant events already by the connected component $[X_0]$ of the initial value up to null sets.

5.29 Example. Consider a Markov chain on $S = \{1,2,3\}$ with one-step transition matrix

$$P(1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$

Then the Markov chain is recurrent with connected components $\{1\}$ and $\{2,3\}$. An initial distribution π is invariant if and only if $\pi(\{2\}) = \pi(\{3\})$▶Control. Then an invariant $A \in \mathscr{I}_\vartheta$ can a.s. be written as $\{[X_0] \subseteq B\}$. So, the invariant events are $\{\varnothing, \Omega, \{X_0 = 1\}, \{X_0 \in \{2,3\}\}\}$ up to null sets. The Markov chain is ergodic if $\pi(\{1\}) = 1$ or if $\pi(\{2,3\}) = 1$. Otherwise, the chain is not ergodic. Note, however, that all invariant distributions π are obtained as convex combinations of the two ergodic initial distributions: $\pi = \alpha\delta_1 + (1-\alpha)\frac{\delta_2+\delta_3}{2}$ for some $\alpha \in [0,1]$, which holds in wider generality, see below.

5.30 Remark. So far, we have established that an ergodic Markov chain is recurrent and irreducible modulo null sets. The example of the symmetric random walk shows that these conditions are not sufficient for ergodicity. There is a complete characterisation available, see e.g. Klenke.

A recurrent state $x \in S$ is called <u>positive-recurrent</u> if $\mathbb{E}_x[\tau_x] < \infty$. For an irreducible Markov chain one can show equivalence between:

(a) all states are positive-recurrent;

(b) there is a positive-recurrent state;

(c) there is an invariant initial distribution;

(d) there is exactly one invariant initial distribution;

(e) the Markov chain is ergodic under the invariant initial distribution.

In that case the invariant initial distribution satisfies $\pi(\{x\}) = \frac{1}{\mathbb{E}_x[\tau_x]}$ for all $x \in S$. For a finite state space S all five properties are always satisfied.

5.3 Ergodic theorems

5.31 Remark. We come to the main results for measure-preserving transformations and thus stationary processes. The following lemma is not very intuitive, but provides the key ingredient for the proof of the ergodic theorem.

5.32 Lemma (Maximal ergodic lemma)**.** *Let $Y \in L^1$ and T be measure-preserving on $(\Omega, \mathscr{F}, \mathbb{P})$. Denoting $S_n := \sum_{k=0}^{n-1} Y \circ T^k$, $S_0 := 0$ and $M_n := \max\{S_0, \dots, S_n\}$, we have $\mathbb{E}[Y\mathbf{1}_{\{M_n>0\}}] \geqslant 0$.*

Proof. We have

$$Y + M_n \circ T = \max_{0\leqslant k\leqslant n}(Y + S_k \circ T) = \max(S_1, \dots, S_{n+1}).$$

Because of $S_0 = 0$ we obtain further

$$(Y + M_n \circ T)\mathbf{1}(M_n > 0) = \max(S_0, S_1, \dots, S_{n+1})\mathbf{1}(M_n > 0) \geqslant M_n\mathbf{1}(M_n > 0).$$

Since T is measure-preserving, this yields

$$\mathbb{E}[(Y + M_n \circ T)\mathbf{1}(M_n > 0)] \geqslant \mathbb{E}[M_n\mathbf{1}(M_n > 0)] = \mathbb{E}[M_n] = \mathbb{E}[M_n \circ T].$$

From this $\mathbb{E}[Y\mathbf{1}(M_n > 0)] \geqslant 0$ follows, using $(M_n \circ T)\mathbf{1}(M_n \leqslant 0) \geqslant 0$. □

5.33 Theorem (Birkhoff's ergodic theorem, 1931)**.** *Let $X \in L^1$ and T be measure-preserving on $(\Omega, \mathscr{F}, \mathbb{P})$. Then:*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} X \circ T^k = \mathbb{E}[X \mid \mathscr{I}_T] \qquad \mathbb{P}\text{-a.s. and in } L^1.$$

If T is even ergodic, then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} X \circ T^k = \mathbb{E}[X] \qquad \mathbb{P}\text{-a.s. and in } L^1.$$

Proof. We set $A_n := \frac{1}{n}\sum_{k=0}^{n-1} X \circ T^k$, $\overline{A} := \limsup_{n\to\infty} A_n$, $\underline{A} := \liminf_{n\to\infty} A_n$ and split the proof in several steps.

$\overline{A}, \underline{A}$ are (strictly) T-invariant: We have $\frac{n+1}{n}A_{n+1} = A_n \circ T + \frac{1}{n}X$. From $\frac{1}{n}X \to 0$ we deduce

$$\overline{A} = \limsup_{n\to\infty} A_{n+1} = \limsup_{n\to\infty} \tfrac{n+1}{n}A_{n+1} = \limsup_{n\to\infty} A_n \circ T = \overline{A} \circ T.$$

Analogously, $\underline{A} = \underline{A} \circ T$ follows.

$\overline{A} = \underline{A}$ $\mathbb{P}$-a.s.: Apply the maximal ergodic lemma to $Y = (X - b)\mathbf{1}(\underline{A} < a, \overline{A} > b)$ for some $a < b$. In the notation of the lemma we have

$$\begin{aligned}\lim_{n\to\infty} \mathbf{1}(M_n > 0) &= \mathbf{1}\Big(\sup_{n\geqslant 1} S_n > 0\Big) = \mathbf{1}\Big(\sup_{n\geqslant 1} \tfrac{1}{n}S_n > 0\Big)\\ &= \mathbf{1}\Big(\sup_{n\geqslant 1}(A_n - b)\mathbf{1}(\underline{A} < a, \overline{A} > b) > 0\Big) = \mathbf{1}\Big(\underline{A} < a, \overline{A} > b\Big).\end{aligned}$$

By the maximal ergodic lemma and dominated convergence we deduce

$$0 \leqslant \mathbb{E}\left[(X-b)\mathbf{1}(\underline{A}<a,\overline{A}>b)\mathbf{1}(M_n>0)\right] \to \mathbb{E}\left[(X-b)\mathbf{1}(\underline{A}<a,\overline{A}>b)\right].$$

Consequently, the limit is non-negative and

$$\mathbb{E}[X\mathbf{1}(\underline{A}<a,\overline{A}>b)] \geqslant b\,\mathbb{P}(\underline{A}<a,\overline{A}>b).$$

The analogous argument for $Y=(a-X)\mathbf{1}(\underline{A}<a,\overline{A}>b)$ yields

$$\mathbb{E}[X\mathbf{1}(\underline{A}<a,\overline{A}>b)] \leqslant a\,\mathbb{P}(\underline{A}<a,\overline{A}>b),$$

implying $\mathbb{P}(\underline{A}<a,\overline{A}>b)=0$. Taking the union of all these null events with $a,b\in\mathbb{Q}$ and $a<b$, we conclude $\mathbb{P}(\underline{A}<\overline{A})=0$ and thus $\underline{A}=\overline{A}$ $\mathbb{P}$-a.s.

(A_n) **is uniformly integrable:** Since T is measure-preserving, $X\circ T^k$ has the same distribution as X such that $(X\circ T^k)_{k\geqslant 0}$ is uniformly integrable. By Lemma 4.48(a) there is a $\delta>0$ for given $\varepsilon>0$ such that $\mathbb{P}(B)<\delta$ implies $\mathbb{E}[|X\circ T^k|\mathbf{1}_B]<\varepsilon$ for all $k\geqslant 0$. Then by triangle inequality for events B with $\mathbb{P}(B)<\delta$

$$\sup_n \mathbb{E}[|A_n|\mathbf{1}_B] \leqslant \sup_n \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}[|X\circ T^k|\mathbf{1}_B] < \varepsilon$$

follows. Noting $\mathbb{E}[|A_n|]\leqslant\mathbb{E}[|X|]$, we conclude that (A_n) is uniformly integrable.

Convergence to $\mathbb{E}[X\,|\,\mathscr{I}_T]$**:** Since $A_n\to\overline{A}$ $\mathbb{P}$-a.s. and (A_n) is uniformly integrable, the convergence also holds in L^1. Moreover, L^1-convergence implies L^1-convergence of the conditional expectations:

$$\mathbb{E}\left[|\mathbb{E}[A_n\,|\,\mathscr{I}_T]-\mathbb{E}[\overline{A}\,|\,\mathscr{I}_T]|\right] \leqslant \mathbb{E}[\mathbb{E}[|A_n-\overline{A}|\,|\,\mathscr{I}_T]] = \|A_n-\overline{A}\|_{L^1}\to 0.$$

We have $\mathbb{E}[X\circ T^k\,|\,\mathscr{I}_T]=\mathbb{E}[X\,|\,\mathscr{I}_T]$, checking for $B\in\mathscr{I}_T$

$$\mathbb{E}[(X\circ T^k)\mathbf{1}_B]=\mathbb{E}[(X\circ T^k)(\mathbf{1}_B\circ T^k)]=\mathbb{E}[X\mathbf{1}_B].$$

We conclude $\mathbb{E}[X\,|\,\mathscr{I}_T]=\mathbb{E}[A_n\,|\,\mathscr{I}_T]\to\mathbb{E}[\overline{A}\,|\,\mathscr{I}_T]=\overline{A}$ in L^1 an thus $\overline{A}=\mathbb{E}[X\,|\,\mathscr{I}_T]$ $\mathbb{P}$-a.s.

T **ergodic implies** $\mathbb{E}[X\,|\,\mathscr{I}_T]=\mathbb{E}[X]$ $\mathbb{P}$**-a.s.:** For ergodic T events $B\in\mathscr{I}_T$ satisfy $\mathbb{P}(B)\in\{0,1\}$ and thus $\mathbb{E}[\mathbb{E}[X]\mathbf{1}_B]=\mathbb{E}[X\mathbf{1}_B]$.

□

5.34 Example. The Ehrenfest model under the invariant initial distribution $\pi=\mathrm{Bin}(N,1/2)$ is an irreducible and recurrent stationary Markov chain. By Corollary 5.28 the chain is ergodic and Birkhoff's ergodic theorem yields

$$\frac{1}{n}\sum_{k=0}^{n-1}f(X_k)\to\mathbb{E}_\pi[f(X_0)]=\sum_{i=0}^{N}\binom{N}{i}2^{-N}f(i)$$

for any function $f : \{0, \dots, N\} \to \mathbb{R}$. We have shown that the physicists' ergodic hypothesis applies in this case.

Let $f : \mathbb{R} \to \mathbb{R}$ be a 1-periodic measurable function with $\int_0^1 |f(x)| dx < \infty$. Then for all $r \in \mathbb{R}$ and Lebesgue-almost all $x \in \mathbb{R}$ the average $\frac{1}{n} \sum_{k=0}^{n-1} f(x+kr)$ converges. If $r = p/q$ with $p, q \in \mathbb{N}$ is rational, then the limit is $\frac{1}{q} \sum_{l=0}^{q-1} f(x + lp/q)$. If r is irrational, then the limit is $\int_0^1 f(y)dy$. This follows from Example 5.11(b) and Birkhoff's ergodic theorem by noting $f(x+kr \mod 1) = f(x+kr)$ for 1-periodic f and by evaluating $\mathbb{E}[f(X_0) \,|\, \mathscr{I}_T]$▶Control.

▷ **Control questions**

(a) How does the derivation of the invariant initial distributions in Example 5.29 work precisely?

$\vec{\pi} P(1) = \vec{\pi}$ is equivalent to the equations $\pi(\{1\}) = \pi(\{1\})$, $\frac{1}{2}(\pi(\{2\}) + \pi(\{3\}) = \pi(\{2\}) = \pi(\{3\})$. By Lemma 1.18 this gives all invariant initial distributions.

(b) Extend Lemma 5.10(a) to unbounded random variables Y.

We know from the lemma that T measure-preserving gives $\mathbb{P}(Y = Y \circ T) = 1$ for any $\mathscr{I}_T$-random variable Y. If Y is not $\mathbb{P}$-a.s. constant, then there is a Borel set B with $\mathbb{P}(Y \in B) \in (0,1)$. Now, $A := \{Y \in B\}$, $T^{-1}(A) = \{Y \circ T \in B\}$ and $Y = Y \circ T$ $\mathbb{P}$-a.s imply $A = T^{-1}(A)$ $\mathbb{P}$-a.s. and A is invariant. Hence T is not ergodic. The contraposition then asserts that for ergodic T, Y must be $\mathbb{P}$-a.s. constant. The proof in this direction never uses that Y is bounded.

(c) Evaluate $\mathbb{E}[f(X_0) \,|\, \mathscr{I}_T]$ for the shift by r in Example 5.34.

A Borel set $B \subseteq [0,1)$ is strictly invariant if $B = \{x - r \mod 1 \,|\, x \in B\}$. Similarly to Lemma 5.10(a), a function f is then $\mathscr{I}_T$-measurable if and only if $f(x) = f((x+r) \mod 1)$ for Lebesgue-almost all $x \in [0,1)$. For $r = p/q \in \mathbb{Q}$ and 1-periodic f we see that $g(x) := \frac{1}{q} \sum_{l=0}^{q-1} f(x + lr)$ is $\mathscr{I}_T$-measurable. Moreover, for any bounded $\mathscr{I}_T$-measurable function h (extended to be 1-periodic on $\mathbb{R}$) we have

$$\mathbb{E}[g(X_0)h(X_0)] = \frac{1}{q} \sum_{l=0}^{q-1} \int_0^1 f(x+lr)h(x)\,dx = \frac{1}{q} \sum_{l=0}^{q-1} \int_0^1 f(x+lr)h(x+lr)dx$$
$$= \int_0^1 f(x)h(x)\,dx = \mathbb{E}[f(X_0)h(X_0)].$$

This shows $g(X_0) = \mathbb{E}[f(X_0) \,|\, \mathscr{I}_T]$. Of course, the ergodic theorem follows in this case directly from the periodicity of the arguments (and even surely).

5.35 Theorem (L^p-ergodic theorem, L^2-version by von Neumann 1932)**.** *For $X \in L^p$, $p \geqslant 1$, and measure-preserving T on $(\Omega, \mathscr{F}, \mathbb{P})$ we have*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} X \circ T^k = \mathbb{E}[X \,|\, \mathscr{I}_T] \qquad \mathbb{P}\text{-a.s. and in } L^p.$$

Proof. ▶Exercise □

5.36 Corollary. *Let $(X_k,\, k \geqslant 0)$ be an ergodic process in L^1 (i.e. $X_k \in L^1$ and the left shift on $(\mathbb{R}^{\mathbb{N}_0}, \mathfrak{B}_{\mathbb{R}}^{\otimes \mathbb{N}_0}, \mathbb{P}^X)$ is ergodic). Then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} X_k = \mathbb{E}[X_1] \qquad \mathbb{P}\text{-a.s. and in } L^1.$$

In particular, Kolmogorov's strong law of large number for (X_n) in L^1 follows.

Proof. Combine Example 5.11(a) and the Ergodic Theorems 5.33, 5.35. □

5.37 Corollary. *A measure-preserving transformation T on $(\Omega, \mathscr{F}, \mathbb{P})$ is ergodic if and only if*

$$\forall A, B \in \mathscr{F}: \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{P}(A \cap T^{-k}(B)) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Proof. ▶Exercise □

5.38 Remark. We change perspective, fix a transformation T and consider the set of all probability measures that render T measure-preserving. This set of invariant probabilities has a very nice geometric structure, which we saw already in Example 5.29: it is convex and at its extremal points T is ergodic.

5.39 Definition. Let $T : \Omega \to \Omega$ be measurable on $(\Omega, \mathscr{F})$. Each probability measure μ on $\mathscr{F}$ with $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \mathscr{F}$ is called invariant with respect to T. If T is even ergodic on $(\Omega, \mathscr{F}, \mu)$, then also μ is called ergodic. The set of all T-invariant probability measures is denoted by $\mathscr{M}_T$.

5.40 Lemma. *$\mathscr{M}_T$ is convex (maybe empty).*

Proof. For $\mu_1, \mu_2 \in \mathscr{M}_T$, $\alpha \in (0,1)$ consider $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$. Then μ is again a probability measure and satisfies for $A \in \mathscr{F}$

$$\mu(T^{-1}(A)) = \alpha\mu_1(T^{-1}(A)) + (1-\alpha)\mu_2(T^{-1}(A)) = \alpha\mu_1(A) + (1-\alpha)\mu_2(A) = \mu(A).$$

Hence, $\mu \in \mathscr{M}_T$ and $\mathscr{M}_T$ is convex. □

5.41 Proposition. *If μ and ν are distinct ergodic measures, then they are singular: $\mu \perp \nu$.*

Proof. Choose $A \in \mathscr{F}$ with $\mu(A) \neq \nu(A)$. The ergodic theorem implies

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_A \circ T^k \to \begin{cases} \mu(A), & \mu\text{-a.s.,} \\ \nu(A), & \nu\text{-a.s.} \end{cases}$$

Hence, for $\Omega_\mu = \{\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_A \circ T^k = \mu(A)\}$ we have $\mu(\Omega_\mu) = 1$, $\nu(\Omega_\mu) = 0$. □

5.42 Definition. A point x in a convex set C is called extremal if x is not a strict convex combination of other points in C: $x, y, z \in C$ and $x = \alpha y + (1-\alpha) z$ for $\alpha \in (0,1)$ implies $y = z = x$.

5.43 Example. The corners of a triangle and more generally of any convex polygon are exactly the extremal points. The sphere is the set of extremal points of the ball.

5.44 Theorem. *The ergodic measures form exactly the extremal points of the convex set $\mathscr{M}_T$.*

Proof. Suppose first that μ is not ergodic. Then there is some strictly invariant $A \in \mathscr{F}$ with $T^{-1}(A) = A$ and $\mu(A) \in (0,1)$ by Lemma 5.10(b). Introduce the probability measures $\mu_1 = \mu(\bullet \,|\, A)$, $\mu_2 = \mu(\bullet \,|\, A^C)$. Then $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$ holds with $\alpha = \mu(A) \in (0,1)$. Moreover,

$$\mu_1(T^{-1}(B)) = \frac{\mu(T^{-1}(B) \cap A)}{\mu(A)} = \frac{\mu(T^{-1}(B \cap A))}{\mu(A)} = \frac{\mu(B \cap A)}{\mu(A)} = \mu_1(B)$$

for $B \in \mathscr{F}$ shows that μ_1 is T-invariant. Similarly, μ_2 is T-invariant. Therefore μ is a strict convex combination of $\mu_1, \mu_2 \in \mathscr{M}_T$ and thus not extremal.

Next, let us show that $\mu, \nu \in \mathscr{M}_T$, $\nu \ll \mu$ and μ ergodic implies $\mu = \nu$. Indeed, the ergodic theorem yields μ-a.s. and because of $\nu \ll \mu$ also ν-a.s.

$$\forall A \in \mathscr{F} : \ \frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_A \circ T^k \to \mu(A).$$

Dominated convergence yields $\mu = \nu$:

$$\forall A \in \mathscr{F} : \ \nu(A) = \int \Big(\frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_A \circ T^k\Big)\, d\nu \to \int \mu(A)\, d\nu = \mu(A).$$

Now, if μ is ergodic and $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$ holds with $\mu_1, \mu_2 \in \mathscr{M}_T$, $\alpha \in (0,1)$, then we have $\mu_1, \mu_2 \ll \mu$, thus implying $\mu_1 = \mu_2 = \mu$. This shows that μ is an extremal point of $\mathscr{M}_T$. □

5.45 Corollary. *If T possesses exactly one invariant probability measure μ, then μ is ergodic.*

Proof. μ is an extremal point of $\mathscr{M}_T = \{\mu\}$. □

5.46 Remark. The same geometric structure is present for the set of invariant distributions of a recurrent Markov chain. It is convex, the ergodic invariant distributions have positive probability exactly on one connected component and each invariant distribution is a (possibly infinite) convex combination of the ergodic invariant distributions ▶Exercise. In particular, an irreducible and recurrent Markov chain has at most one invariant distribution, which is then ergodic.

5.4 Convergence of Markov chains and MCMC

In continuation of Section 5.2 we consider time-homogeneous Markov chains on the canonical path space $\Omega = S^{\mathbb{N}_0}$. In addition, we focus on finite state spaces

$S = \{1, \dots, M\}$. Then $P(n) \in \mathbb{R}^{M \times M}$ is the matrix of n-step transition probabilities $p_{ij}(n)$, satisfying the Chapman-Kolmogorov equation $P(n) = P(1)^n$. For an invariant initial distribution π we have $\vec{\pi} P(1) = \vec{\pi}$ for the row vector $\vec{\pi} = (\pi(\{1\}), \dots \pi(\{M\}))$ by Lemma 1.18 and $\vec{\pi}$ is a left eigenvector of $P(1)$ with eigenvalue one.

5.47 Theorem. *Let $x \in S$ be a recurrent state and S be finite. Then*

$$\pi_x(\{y\}) := \frac{\mathbb{E}_x[\sum_{n=0}^{\tau_x - 1} \mathbf{1}(X_n = y)]}{\mathbb{E}_x[\tau_x]}, \quad y \in S,$$

is an invariant initial distribution.

5.48 Remark. In the proof we show $\mathbb{E}_x[\tau_x] < \infty$ for finite state space S. In general, this is not true for any recurrent state and states with this property are called *positive recurrent.* $\pi_x(\{y\})$ is the expected number of visits to y on an excursion of the Markov chain from x relative to the expected length of the excursion. If there is only one invariant measure π, then $\pi = \pi_x$ holds for all recurrent $x \in S$ and in particular for a recurrent Markov chain the identity $\pi(\{y\}) = \frac{1}{\mathbb{E}_y[\tau_y]}$, $y \in S$, follows.

Proof. Consider $y \in S$, $y \neq x$, with $\rho_{xy} > 0$. Then both, $\mathbb{P}_x(\tau_x > \tau_y) > 0$ and $\mathbb{P}_y(\tau_x < \tau_y) > 0$ hold because otherwise starting in x the chain would a.s. never visit y. By the strong Markov property we have

$$\begin{aligned}\mathbb{E}_y\Big[\sum_{n=0}^{\tau_x - 1} \mathbf{1}(X_n = y)\Big] &= 1 + \mathbb{E}_y\Big[\mathbf{1}(\tau_x > \tau_y) \sum_{n=\tau_y}^{\tau_x - 1} \mathbf{1}(X_n = y)\Big] \\ &= 1 + \mathbb{P}_y(\tau_x > \tau_y)\, \mathbb{E}_y\Big[\sum_{n=0}^{\tau_x - 1} \mathbf{1}(X_n = y)\Big]\end{aligned}$$

and consequently

$$\mathbb{E}_y\Big[\sum_{n=0}^{\tau_x - 1} \mathbf{1}(X_n = y)\Big] = \frac{1}{\mathbb{P}_y(\tau_x < \tau_y)} < \infty.$$

The same argument for start in $x \neq y$ yields

$$E(y) := \mathbb{E}_x\Big[\sum_{n=0}^{\tau_x - 1} \mathbf{1}(X_n = y)\Big] = \mathbb{P}_x(\tau_x > \tau_y)\, \mathbb{E}_y\Big[\sum_{n=0}^{\tau_x - 1} \mathbf{1}(X_n = y)\Big] = \frac{\mathbb{P}_x(\tau_x > \tau_y)}{\mathbb{P}_y(\tau_x < \tau_y)}.$$

For $y \in S$ with $\rho_{xy} = 0$ the chain never visits y and $E(y) = 0$, while $E(x) = 1$ by definition. This yields $\mathbb{E}_x[\tau_x] = \sum_{y \in S} E(y) < \infty$ and π_x is well-defined.

To show invariance, we deduce for $z \neq x$

$$\begin{aligned}E(z) &= \mathbb{E}_x\Big[\sum_{n=1}^{\tau_x} \sum_{y \in S} \mathbf{1}(X_{n-1} = y, X_n = z)\Big] \\ &= \sum_{y \in S} p_{yz}(1)\, \mathbb{E}_x\Big[\sum_{n=1}^{\tau_x} \mathbf{1}(X_{n-1} = y)\Big] = \sum_{y \in S} E(y) p_{yz}(1),\end{aligned}$$

while for $z = x$

$$E(x) = 1 = \mathbb{E}_x \Big[\sum_{n=1}^{\tau_x} \mathbf{1}(X_n = x)\Big] = \mathbb{E}_x \Big[\sum_{n=1}^{\tau_x} \sum_{y \in S} \mathbf{1}(X_{n-1} = y, X_n = x)\Big]$$

$$= \sum_{y \in S} p_{yx}(1)\, \mathbb{E}_x \Big[\sum_{n=1}^{\tau_x} \mathbf{1}(X_{n-1} = y)\Big] = \sum_{y \in S} E(y) p_{yx}(1).$$

In the calculations we can take out the transition probabilities $p_{yz}(1)$, $p_{yx}(1)$ by writing $\mathbf{1}(n \leqslant \tau_x) = \prod_{k=1}^{n-1} \mathbf{1}(X_k \neq x)$. Since the denominator in the definition of π_x does not depend on y, this implies the invariance of π_x. □

5.49 Lemma. *An irreducible Markov chain on a finite state space is recurrent and has exactly one invariant initial distribution π, which is ergodic.*

5.50 Remark. Here we need not assume recurrence for the definition of irreducibility in the sense that we only require $\rho_{xy} > 0$ for all $x, y \in S$.

Proof. By the argument in Example 5.18(c) there is at least one recurrent state. By irreducibility, all states are connected so that Theorem 5.19(c) shows that all states are recurrent. Theorem 5.47 gives the existence of an invariant measure. By Corollary 5.28 an irreducible recurrent Markov chain starting in an invariant initial distribution is ergodic. This invariant distribution is unique by Corollary 5.28, compare Remark 5.46. □

▷ **Control questions**

(a) Consider the translation T by $r = 1/2$ in Example 5.11(b). What are invariant measures besides the Lebesgue measure? Can you determine $\mathscr{M}_T$ and the ergodic measures?

Let $\mathbb{P}_1$ be a probability measure on the Borel sets of $[0, 1/2)$ and define $\mathbb{P}_2(B) := \mathbb{P}_1(B - 1/2)$ for the Borel sets of $[1/2, 1)$. Then $\mathbb{P}(A) = (\mathbb{P}_1(A \cap [0, 1/2)) + \mathbb{P}_2(A \cap [1/2, 1)))/2$ is a probability measure on $[0, 1]$, which is invariant for $T(x) = (x + 1/2) \mod 1$ (T translates $[0, 1/2)$ to $[1/2, 1)$ and vice versa). It is easily seen that all invariant measures $\mathbb{P}$ have this form using for $B \subseteq [1/2, 1)$ that $\mathbb{P}(B) = \mathbb{P}(T^{-1}(B)) = \mathbb{P}(B - 1/2)$. The extremal points are the point measures $\mathbb{P} = (\delta_x + \delta_{x+1/2})/2$ for $x \in [0, 1/2)$, which are ergodic: if B is strictly invariant under T, then $x \in B \iff (x + 1/2) \in B$ and thus $\mathbb{P}(B) \in \{0, 1\}$.

(b) Is the set $\mathscr{M}_T$ closed in the finite case $|\Omega| < \infty$, when representing invariant measures μ on Ω as vectors in $\mathbb{R}^{|\Omega|}$?

Yes: if $\mu_n \in \mathscr{M}_T$ satisfy $\mu_n(\{\omega\}) \to \mu(\{\omega\})$ for all $\omega \in \Omega$, then

$$\mu(T^{-1}(A)) = \lim_{n \to \infty} \mu_n(T^{-1}(A)) = \lim_{n \to \infty} \mu_n(A) = \mu(A)$$

follows for all events A. In fact, the argument shows that more generally $\mathscr{M}_T$ is sequentially closed for the topology of pointwise convergence for probability measures. This is stronger than weak convergence which requires a metric on Ω (and then provides a rich theory for continuous T!).

(c) Is the following true: If $vP = \lambda v$ holds for $P \in \mathbb{R}^{M\times M}$, a left (row) eigenvector $v \in \mathbb{R}^m$ and an eigenvalue λ, then there is a right (column) eigenvector $w \in \mathbb{R}^M$ with $Pw = \lambda w$ and vice versa?

Yes, because $v(P-I) = 0$ with the identity matrix I and $v \neq 0$ means that $P - I$ cannot have full rank ($v^\top$ is not in its image) and thus by linear algebra it must have a non-trival kernel of dimension 1 or larger, which forms an eigenspace for the eigenvalue 1 of P. The converse is also true. More abstractly, $v^\top$ is an eigenvector of the adjoint $P^\top$ and P and $P^\top$ share the same eigenvalues (e.g., consider the characteristic equation $\det(\lambda I - P) = \det(\lambda I - P^\top) = 0$).

5.51 Definition. Let π be the invariant distribution of an irreducible Markov chain (X_n) on $S = \{1,\dots,M\}$. Then with $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$

$$\|f\|_\pi := \Big(\sum_{x=1}^M \pi(x)|f(x)|^2\Big)^{1/2} = \mathbb{E}_\pi[|f(X_0)|^2]^{1/2}, \quad f : S \to \mathbb{K}$$

defines the $L^2(\pi)$-norm of f and $(Pf)(x) := \sum_{y=1}^M p_{xy}(1) f(y) = \mathbb{E}_x[f(X_1)]$ is the (Markov) transition operator.

5.52 Remark. Since the chain is irreducible, $\rho_{xy} > 0$ for an $x \in S$ with $\pi(\{x\}) > 0$ implies $\mathbb{P}_x(X_n = y) > 0$ for some n and thus by invariance $\pi(\{y\}) = \mathbb{P}_\pi(X_n = y) > 0$ for all y. Hence, $L^2(\pi) = \mathscr{L}^2(\pi) = \{f : S \to \mathbb{R}\}$ holds for finite S (there are no non-empty π-null sets and all functions are measurable). Any function $f : S \to \mathbb{K}$ is characterised by the vector $\vec{f} = (f(1),\dots,f(M))^\top \in \mathbb{K}^M$ and we have $P(1)\vec{f} = \overrightarrow{Pf}$, that is Pf is obtained by matrix-vector multiplication. Yet, the stochastic interpretation $Pf(x) = \mathbb{E}_x[f(X_1)]$ is easier to formulate for functions. We shall profit from the linear algebra approach to Markov chains and it is convenient throughout to interpret column vectors as functions (and row vectors as probability measures) on $S = \{1,\dots,M\}$.

5.53 Example. The transition operator P for the Ehrenfest model on $S = \{0,\dots,N\}$ satisfies $(Pf)(x) = \frac{x}{N} f(x-1) + \frac{N-x}{N} f(x+1)$ for $x \in S$, setting $f(-1) := f(N+1) := 0$ (or arbitrary). This shows that Pf at x is just an average of f at its neighbours $x-1$, $x+1$. For functions f with $Pf = f$ we conclude that $f(x)$ is an average of $f(x-1)$ and $f(x+1)$ for all x, hence f must be constant on S (e.g., consider $\max_x f(x)$). Consequently, 1 is an eigenvalue of P and the constant functions are the only associated eigenfunctions.

5.54 Lemma (Properties of the Markov transition operator P)**.**

(a) *$\|Pf\|_\pi \leqslant \|f\|_\pi$ holds for all $f \in L^2(\pi)$ and all eigenvalues λ of P satisfy $|\lambda| \leqslant 1$.*

(b) *If $\lambda \in \mathbb{C}$ is an eigenvalue of P with $|\lambda| = 1$, then there is a smallest number $d \in \mathbb{N}$ with $\lambda^d = 1$ (λ is a dth-unit root) and $p_{xx}(n) = 0$ holds for all $x \in S$ and all n that are not multiples of d.*

(c) *$\lambda = 1$ is always an eigenvalue of P with multiplicity one. Its eigenspace consists of the constant functions $f = c\mathbf{1}$ with $c \in \mathbb{R}$ (or $\mathbb{C}$).*

Proof.

(a) Use the definitions and the Cauchy-Schwarz inequality to bound

$$\|Pf\|_\pi^2 = \mathbb{E}_\pi[|\mathbb{E}_{X_0}[f(X_1)]|^2] \leqslant \mathbb{E}_\pi[\mathbb{E}_{X_0}[|f(X_1)|^2]] = \mathbb{E}_\pi[|f(X_1)|^2] = \|f\|_\pi^2.$$

From $Pf = \lambda f$ for some non-zero function f we infer $\|Pf\|_\pi = |\lambda| \|f\|_\pi$ and thus $|\lambda| \leqslant 1$ by the inequality.

(b) Let λ be an eigenvalue of P with eigenfunction f, i.e. $Pf = \lambda f$, $f \neq 0$ and $|\lambda| = 1$. Then $\|Pf\|_\pi = \|f\|_\pi$ implies equality in the Cauchy-Schwarz inequality. This holds only if $f(X_1)$ is $\mathbb{P}_x$-a.s. constant for all $x \in S$▶Control. Fix some state x with $f(x) \neq 0$. Then $Pf(x) = \mathbb{E}_x[f(X_1)] = \lambda f(x)$ implies $f(y) = \lambda f(x)$ for all y with $p_{xy}(1) > 0$. From $P^n f(x) = \mathbb{E}_x[f(X_n)] = \lambda^n f(x)$ we equally deduce $f(y) = \lambda^n f(x)$ for all states y with $p_{xy}(n) > 0$. Since x is irreducible, for each $y \in S$ there is an n_y with $p_{xy}(n_y) > 0$ and thus $f(y) = \lambda^{n_y} f(x) \neq 0$. The case $y = x$ yields $\lambda^{n_x} = 1$. For $d := \min\{n \in \mathbb{N} \mid \lambda^n = 1\} \leqslant n_x$ and $i = 1, \dots, d-1$, $k \in \mathbb{N}_0$ we have $f(x) \neq \lambda^i f(x) = \lambda^{kd+i} f(x) \Rightarrow p_{xx}(kd+i) = 0$.

(c) For constant functions $f = c\mathbf{1}$ we clearly have $(Pf)(x) = \mathbb{E}_x[f(X_1)] = c = f(x)$, hence $Pf = f$, and 1 is an eigenvalue with eigenfunction f. The argument in (b) shows conversely that for an eigenfunction f with eigenvalue 1, we have $f(y) = 1^n f(x) = f(x)$ if $p_{xy}(n) > 0$. By irreducibility, there is such an n for all $x, y \in S$ so that f must be constant.

□

5.55 Example.

(a) Consider the *rotation* with transitions $p_{x,x+1} = 1$, $x = 1, \dots, M-1$, $p_{M,1} = 1$, where you go deterministically $1 \mapsto 2 \mapsto \cdots \mapsto S \mapsto 1$. This is an irreducible Markov chain with uniform invariant distribution $\pi(\{x\}) = \frac{1}{M}$. The transition operator is $Pf(x) = f((x \mod M) + 1)$. For an eigenvalue λ with eigenfunction f we derive from $P^M f = f$ that $\lambda^M = 1$. We claim that $\lambda_j = e^{2\pi i j/M}$ for $j = 0, \dots, M-1$ are eigenvalues of P. Indeed, for $f_j(x) = \lambda_j^x$, $x = 1, \dots, M$, we obtain $Pf_j(x) = \lambda_j^{(x \mod M)+1} = \lambda_j^{x+1} = \lambda_j f_j(x)$. Obviously, $p_{xx}(n) > 0$ is equivalent to $n = kM$ for $k \in \mathbb{N}$ and the chain is *M-periodic*.

(b) In the Ehrenfest model let λ be an eigenvalue of P with $|\lambda| = 1$ and eigenfunction f. Then $\lambda f(x) = \frac{x}{N} f(x-1) + \frac{N-x}{N} f(x+1)$ holds. Taking absolute values shows $|f(x)| \leqslant \frac{x}{N}|f(x-1)| + \frac{N-x}{N}|f(x+1)|$. In extension of Example 5.53 this can only happen if equality holds (consider again $\max_x |f(x)|$) and $|f|$ is constant on S. Moreover, the equality $|\alpha_1 z_1 + \alpha_2 z_2| = \alpha_1 |z_1| + \alpha_2 |z_2|$ for $\alpha_1, \alpha_2 > 0$, $z_1, z_2 \in \mathbb{C}$ only holds if $z_1 = |z_1| e^{i\varphi}$, $z_2 = |z_2| e^{i\varphi}$ with the same angle $\varphi \in [0, 2\pi)$. This shows that $f(x)$ is constant for all odd x and for all even x, respectively. By a short reflection, we conclude that the eigenvalue must be $\lambda = 1$ with constant eigenfunction f or $\lambda = -1$ with eigenfunction $f(x) = c(-1)^x$ for some $c \in \mathbb{C}$. Obviously, we have $p_{xx}(n) > 0$ if and only if n is even so that the chain is 2-periodic..

(c) For a Markov chain where each transition $p_{xy}(1)$ for $x \neq y$ has positive probability and $|S| > 3$, we have $p_{xx}(n) > 0$ for all $x \in S$ and $n \geqslant 2$ via a path $x \mapsto x_1 \mapsto \cdots \mapsto x_{n-1} \mapsto x$ with $x_i \neq x_{i+1}$ and $x \notin \{x_1, x_{n-1}\}$. The same holds for any Markov chain when the probability $p_{xx}(1) > 0$ to stay in a state is positive for all $x \in S$, compare Google PageRank. For these Markov chains the transition operator has eigenvalues λ with $|\lambda| < 1$ except for the trivial eigenvalue $\lambda = 1$. They are *aperiodic* according to the definition below.

5.56 Definition. An irreducible Markov chain is called aperiodic if the greatest common divisor (*größter gemeinsamer Teiler*) of $\{n \in \mathbb{N} \mid p_{xx}(n) > 0\}$ equals 1 for some state x.

5.57 Remark. By Lemma 5.54(b) $\lambda = 1$ is the only eigenvalue of the transition operator P with $|\lambda| = 1$ for an aperiodic Markov chain. For aperiodic Markov chains the n-step transition matrix $P(n) = P^n$ then has the trivial eigenvalue 1 and all other eigenvalues have the form λ^n for $|\lambda| < 1$. This makes the convergence $\lim_{n\to\infty}(P^n f)(x) = 0$ for f with $\mathbb{E}_\pi[f] = 0$ plausible, at least for diagonalisable P. Next, a version of this is proved rigorously.

5.58 Theorem (Convergence for aperiodic chains)**.** *For an irreducible and aperiodic Markov chain on a finite state space S*

$$\lim_{n\to\infty} p_{xy}(n) = \pi(\{y\}), \quad x, y \in S,$$

holds with the invariant initial distribution π. The convergence is exponentially fast in n.

Proof. By Remark 5.57 the eigenvalues λ of the transition operator P of an irreducible and aperiodic Markov chain satisfy $|\lambda| < 1$ or $\lambda = 1$ (with eigenfunction $\mathbf{1}$). Observe that $\mathbb{E}_\pi[f(X_0)] = 0$ implies $\mathbb{E}_\pi[f(X_1)] = 0$ by invariance, in other words $\langle f, \mathbf{1}\rangle_\pi = 0 \Rightarrow \langle Pf, \mathbf{1}\rangle_\pi = 0$. This means that $L_0^2(\pi) := \{f : S \to \mathbb{R} \mid \langle f, \mathbf{1}\rangle_\pi = 0\}$ is invariant under P: $P(L_0^2(\pi)) \subseteq L_0^2(\pi)$. Then all eigenvalues λ of the restriction $P|_{L_0^2(\pi)}$ satisfy $|\lambda| < 1$. The spectral radius result from linear algebra gives for $f \in L_0^2(\pi)$

$$\limsup_{n\to\infty} \|P^n f\|_\pi^{1/n} \leqslant \max\{|\lambda| \mid \lambda \text{ is eigenvalue of } P|_{L_0^2(\pi)}\} < 1$$

(if the eigenvectors diagonalise P, then this is clear, otherwise consider the Jordan form; observe $\|f\|_\pi^{1/n} \to 1$ for $f \neq \mathbf{0}$). For any $g : S \to \mathbb{K}$ we obtain $f := g - \langle g, \mathbf{1}\rangle_\pi \mathbf{1} \in L_0^2(\pi)$ and thus

$$\|P^n g - \mathbb{E}_\pi[g(X_0)]\|_\pi = \|P^n(g - \langle g, \mathbf{1}\rangle_\pi \mathbf{1})\|_\pi \to 0 \text{ exponentially fast.}$$

For $g = \mathbf{1}_{\{y\}}$ we obtain in particular $\sum_{x\in S} \pi(x)(p_{xy}(n) - \pi(\{y\}))^2 \to 0$, whence $p_{xy}(n) \to \pi(\{y\})$ follows for all $x, y \in S$. □

▷ **Control questions**

(a) Why does $|\mathbb{E}_x[f(X_1)]|^2 = \mathbb{E}_x[|f(X_1)|^2]$ imply that $f(X_1) = \mathbb{E}_x[f(X_1)]$ is constant $\mathbb{P}_x$-a.s.?

Just use the variance (here real case): $\mathrm{Var}_x(f(X_1)) = \mathbb{E}_x[f(X_1)^2] - \mathbb{E}_x[f(X_1)]^2 = 0$ implies $P_x(f(X_1) = \mathbb{E}_x[f(X_1)]) = 1$, see Stochastik I.

(b) Consider an invariant, but not ergodic distribution π for the non-irreducible Markov chain of Example 5.29. Show that the eigenspace to the eigenvalue 1 of the transition operator P is two-dimensional. Can you generalise?

Any function f with $f(2) = f(3)$ satisfies $Pf(x) = f(x)$, $x = 1, 2, 3$, and is thus an eigenfunction of P to the eigenvalue 1. More generally, any function which is constant on each connected component of a recurrent Markov chain remains invariant. So, $\lambda = 1$ has multiplicity 1 as an eigenvalue of P if and only if a recurrent Markov chain is irreducible if and only if there is only one invariant distribution.

(c) If $(X_n, n \geqslant 0)$ is the 2-periodic Markov chain of the Ehrenfest model, is the embedded Markov chain $(X_{2n}, n \geqslant 0)$ aperiodic?
Yes, that is trivially the case because $\mathbb{P}_x(X_2 = x) > 0$ for all x (go to the next state and then back). So, the transition probabilities $p_{xy}(2n)$ along the even steps converge to the invariant distribution on $\{y \in S \,|\, |y - x| \text{ is even}\}$, which is also a Binomial distribution. In fact, the Ehrenfest model 'forgets' the initial state asymptotically except for its property of being even or odd.

5.59 Remark. In linear algebra symmetric matrices have additional structure. This can be exploited for the transition matrix / operator of a Markov chain. Interestingly, it turns out that this leads to the property that the law of the Markov chain is the same when time is reversed, that is $(X_0, \dots, X_n) \overset{d}{=} (X_n, \dots, X_0)$.

5.60 Definition. An initial distribution μ of a Markov chain with one-step transition probabilities $p_{x,y}(1)$ is called reversible if

$$\forall x, y \in S: \; \mu(\{x\}) p_{xy}(1) = \mu(\{y\}) p_{yx}(1).$$

Then the Markov chain is said to be (time) reversible or in detailed balance.

5.61 Proposition. *A reversible initial distribution π is invariant and the Markov chain (X_n) satisfies*

$$\mathbb{P}_\pi(X_0 = x_0, \dots, X_n = x_n) = \mathbb{P}_\pi(X_0 = x_n, \dots, X_n = x_0)$$

for all $x_0, \dots, x_n \in S$. The associated transition operator P is self-adjoint:

$$\forall f, g : S \to \mathbb{R}: \; \langle Pf, g \rangle_\pi = \langle f, Pg \rangle_\pi.$$

Proof. ▶Exercise □

5.62 Remark. A reversible transition matrix $P(1)$ itself is usually not symmetric, but its (basis) transformation $\mathrm{diag}(\vec{\pi})^{1/2} P(1) \,\mathrm{diag}(\vec{\pi})^{-1/2}$ with entries

$\pi(\{x\})^{1/2}p_{xy}(1)\pi(\{y\})^{-1/2} = \pi(\{y\})^{1/2}p_{yx}(1)\pi(\{x\})^{-1/2}$ is symmetric, provided $\pi(\{x\}) > 0$ for all x. A better viewpoint is that the linear map P (defined via $P(1)$) can be represented by a symmetric matrix in any $L^2(\pi)$-orthonormal basis.

From linear algebra we know that then P can be diagonalised with real eigenvalues and an $L^2(\pi)$-orthonormal basis of eigenvectors. Lemma 5.54 shows for reversible irreducible Markov chains that the eigenvalues λ of P satisfy $\lambda \in [-1,1]$ and that $\lambda = -1$ can only occur if $p_{xx}(2n-1) = 0$ holds for all $x \in S$, $n \in \mathbb{N}$ (the chain is 2-periodic). Otherwise the Markov chain can be shown to be aperiodic.

5.63 Example. Let π be the uniform distribution on S. Then π is reversible whenever the transition matrix $P(1)$ is symmetric: $p_{xy}(1) = p_{yx}(1)$ for all $x,y \in S$. Consider a reflected random walk on $S = \{1,\dots,M\}$ with $p_{xy}(1) > 0$ only if $|x-y| \leqslant 1$ (you can only go left, go right or stay). If the chain satisfies $p_{x,x+1}(1) = p_{x,x-1}(1)$ in the interior ($x = 2,\dots,M-1$), then by symmetry all these probabilities coincide and are equal to some $q \in [0,1/2]$. Symmetry implies then also at the boundary $p_{12}(1) = p_{M,M-1}(1) = q$ and we must thus have $p_{xx}(1) = 1-2q$ for $x = 2,\dots,M-1$ and $p_{xx}(1) = 1-q$ for $x \in \{1,M\}$. This gives a reversible Markov chain where the uniform distribution is invariant.

5.64 Lemma. *Let (X_n) be a reversible irreducible Markov chain with invariant distribution π. Consider the undirected graph $G = (S,E)$ with vertices $x \in S$ and edge set $E = \{\{x,y\}\,|\,x,y \in S, \mathbb{P}_\pi(X_0 = x, X_1 = y) > 0\}$. Assign the probability $\tilde p_{\{x,y\}} := \mathbb{P}_\pi(X_0 = x, X_1 = y)$ to each edge $\{x,y\} \in E$. Then $(\tilde p_e)_{e\in E}$ defines uniquely the law of the Markov chain via*

$$\pi(\{x\}) = \sum_{y\in S:\{x,y\}\in E} \tilde p_{\{x,y\}}, \quad p_{xy}(1) = \frac{\tilde p_{\{x,y\}}}{\pi(\{x\})}, \quad x,y \in S. \tag{5.1}$$

Proof. Note that $\pi(\{x\}) > 0$ holds for all x because (X_n) is irreducible, compare Remark 5.52. By definition of reversibility, we have $\tilde p_{\{x,y\}} = \pi(\{x\})p_{xy}(1)$ and the formulas follow from $\sum_{y\in S} p_{xy}(1) = 1$. □

5.65 Example. For the reversible random walk of Example 5.63 with $q \in (0,1/2)$ we have $E = \{\{x,y\}\,|\,x,y \in S,\, |x-y| \leqslant 1\}$ and $\tilde p_{\{x,x+1\}} = M^{-1}q$ for $x = 1,\dots,S-1$, $\tilde p_{\{0\}} = \tilde p_{\{M\}} = M^{-1}(1-q)$ as well as $\tilde p_{\{x\}} = M^{-1}(1-2q)$ for $x = 2,\dots,S-1$.

5.66 Lemma. *Let $(\tilde p_e)_{e\in E}$ be any (counting density of a) probability distribution on the edge set E of an undirected graph $G = (S,E)$ with $\sum_{y:\{x,y\}\in E} \tilde p_{\{x,y\}} > 0$ for all $x \in S$. Then equations* (5.1) *define a reversible (not necessarily irreducible) Markov chain on S.*

Proof. By assumption, $\pi(\{x\}) > 0$ follows for all $x \in S$ and thus $\pi(\{x\})p_{xy}(1) = \tilde p_{\{x,y\}} = \pi(\{y\})p_{yx}(1)$. Hence, π is reversible. □

5.67 Remark. The basic idea of Markov chain Monte Carlo (MCMC) methods is to construct an irreducible aperiodic Markov chain $(X_n, n \geqslant 0)$ for a given invariant distribution π and to simulate π-distributed random variables approximately by $X(n)$ for an arbitrary starting value $x \in S$ and large n because by the convergence in Theorem 5.58 we know $X(n) \xrightarrow{d} \pi$ as $n \to \infty$. In typical applications like Bayesian posterior computation or energy functionals in statistical physics the distribution π is only known up to a norming constant (whose numerical computation would require a summation over a huge state space S).

The prominent Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, Teller 1953) starts off with an irreducible 'proposal' Markov chain with transition matrix $Q(1) = (q_{xy}(1))_{x,y \in S}$ and obtains a π-reversible Markov chain by changing the transition probabilities from $q_{xy}(1)$ to $p_{xy}(1)$ by an 'accept-reject step' which involves only the ratio $\pi(\{y\})/\pi(\{x\})$ and is thus independent of a norming constant.

The idea is quite intuitive: Assume $q_{xy}(1) > 0 \iff q_{yx}(1) > 0$ for all $x, y \in S$. Consider the undirected graph $G = (S, E)$ with $E = \{\{x, y\} \mid q_{xy}(1) > 0\}$. Each edge $\{x, y\} \in E$ with $x \neq y$ is assigned the probability

$$\tilde{p}_{\{x,y\}} := \min\big(\pi(\{x\})q_{xy}(1), \pi(\{y\})q_{yx}(1)\big) \in [0, 1],$$

which enforces the symmetry for π-reversible transitions, when we assign all remaining transition probability from x to the loop probability $p_{xx}(1)$.

More algorithmically, when the proposal Markov chains proposes to go from x to $y \neq x$ we go to y ('accept') or stay in x ('reject') because by Formula (5.1) $p_{xy}(1) = \min(q_{xy}(1), \frac{\pi(\{y\})}{\pi(\{x\})}q_{yx}(1)) \leqslant q_{xy}(1)$ for $x \neq y$ and we accept a transition to y with probability $p_{xy}(1)/q_{xy}(1) \in [0, 1]$ (by an independent random experiment). This is formalised in the next statement, where we say that a transition matrix is aperiodic if the generated Markov chain is aperiodic.

5.68 Theorem (Metropolis Markov chain)**.** *Consider a distribution π on S with $\pi(\{x\}) > 0$ for all $x \in S$ and an irreducible Markov chain on S with transition probabilities $q_{xy}(1)$, $x, y \in S$, satisfying $q_{xy}(1) > 0$ if and only if $q_{yx}(1) > 0$ for $x, y \in S$. Then the Markov chain with transition probabilities*

$$p_{xy}(1) := \begin{cases} \min\big(q_{xy}(1), \frac{\pi(\{y\})}{\pi(\{x\})}q_{yx}(1)\big), & \text{if } x \neq y, \\ 1 - \sum_{z \neq x} p_{xz}(1), & \text{if } x = y \end{cases}$$

is reversible with respect to π and irreducible. If the transition matrix $Q(1)$ is aperiodic or if π is not reversible with respect to $Q(1)$, then the transition matrix $P(1)$ is aperiodic.

Proof. ▶Exercise □

5.69 Example. Suppose X is a $\mathrm{Bin}(n, p)$-distributed random variable where the success probability p is itself drawn at random from a discrete set $S \subseteq (0, 1)$ according to a distribution μ on S. Then the joint law of (X, p) is given by

$$\mathbb{P}(X = k, p = \rho_i) = \mu(\{\rho_i\})\binom{n}{k}\rho_i^k(1 - \rho_i)^{n-k}, \quad k = 0, \ldots, n;\ \rho_i \in S.$$

In Bayesian statistics we observe X and prescribe the *prior distribution* μ for the unknown parameter p. We base our inference (estimation, testing, credible intervals) on the *posterior distribution* of p given the observation of X. By the Bayes formula ('$\propto$' means proportional to)

$$\pi(\{\rho_i\}) := \mathbb{P}(p=\rho_i \,|\, X=k) = \frac{\mathbb{P}(X=k, p=\rho_i)}{\sum_j \mathbb{P}(X=k,p=\rho_j)} \propto \mu(\{\rho_i\})\rho_i^k(1-\rho_i)^{n-k}$$

holds with a constant independent of i, where k is the observed value of X. Let us suppose now that $\rho_i = (i-1/2)/M$, $i=1,\dots,M$ and $\mu(\{\rho_i\}) = 1/M$ is a discretisation of the uniform distribution $U([0,1])$. Then we obtain π as the invariant distribution of a Metropolis chain with transition probabilities ($i \neq j$)

$$p_{\rho_i\rho_j}(1) = \min\Big(q_{\rho_i\rho_j}(1), \tfrac{\rho_j^k(1-\rho_j)^{n-k}}{\rho_i^k(1-\rho_i)^{n-k}} q_{\rho_j\rho_i}(1)\Big).$$

If the proposal Markov chain is the symmetric random walk on S, interpreted as torus with $\rho_0=\rho_M$, $\rho_{M+1}=\rho_1$, we have $q_{\rho_i\rho_{i\pm1}}(1)=1/2$ and the transition probabilities $p_{\rho_i\rho_i\pm1}(1)$ simplify to

$$p_{\rho_i\rho_i\pm1}(1) = \frac{1}{2}\min\Big(1, \Big(1\pm\frac{1}{i-1/2}\Big)^k\Big(1\pm\frac{-1}{M-i+1/2}\Big)^{n-k}\Big),\ i=2,\dots,M-1,$$

and similarly at the boundary $i\in\{1,M\}$. This yields an aperiodic irreducible chain which is easy to simulate ►Exercise.

▷ **Control questions**

(a) Simplify the proof of Theorem 5.58 for reversible irreducible Markov chains by expressing P in an orthonormal basis of eigenfunctions (=eigenvectors).

Since P is self-adjoint on $L^2(\pi)$ there is an $L^2(\pi)$-orthonormal basis $(f_1,\dots,f_M)$ of eigenfunctions with $Pf_i=\lambda_i f_i$. We choose $\lambda_1=1$, $f_1=\mathbf{1}$ (an eigenpair we know already) and observe $|\lambda_i|<1$ for $i\geqslant 2$ by aperiodicity. In the basis $(f_1,\dots,f_M)$ we have the diagonal matrix representation $P=\operatorname{diag}(\lambda_1,\dots,\lambda_M)$ and $P^n=\operatorname{diag}(\lambda_1^n,\dots,\lambda_M^n)$. This shows $P^n\to\operatorname{diag}(1,0,\dots,0)$ (entrywise or in any matrix norm). We obtain for $f=\sum_{i=1}^M\langle f,f_i\rangle_\pi f_i\in L^2(\pi)$ that $P^nf\to\langle f,f_1\rangle_\pi=\mathbb{E}_\pi[f]$. For $f=\mathbf{1}_{\{y\}}$ this yields $P_nf(x)=\mathbb{E}_x[\mathbf{1}_{\{y\}}(X_n)]=p_{xy}(n)\to\mathbb{E}_\pi[f(X_0)]=\pi(\{y\})$, noting $\pi(\{x\})>0$ for all x.

(b) Let π be the invariant initial distribution of an irreducible Markov chain $(X_n, n\geqslant 0)$ with transition probabilities $p_{xy}(1)$. Show that the time-reversed process $\tilde{X}_n = X_{N-n}$, $n=0,\dots,N$, is a Markov chain with initial distribution π and transition probabilities $\tilde{p}_{xy}(1)=\frac{\pi(\{y\})}{\pi(\{x\})}p_{yx}(1)$.

We have for $x_0,\dots,x_N\in S$

$$\begin{aligned}
&\mathbb{P}_\pi(\tilde{X}_0=x_0,\dots,\tilde{X}_N=x_N) = \mathbb{P}_\pi(X_0=x_N,\dots,X_N=x_0)\\
&= \pi(\{x_N\})\prod_{i=1}^N p_{x_i,x_{i-1}}(1) = \pi(\{x_N\})\prod_{i=1}^N\Big(\frac{\pi(\{x_{i-1}\})}{\pi(\{x_i\})}\tilde{p}_{x_{i-1},x_i}(1)\Big)\\
&= \pi(\{x_0\})\prod_{i=1}^N \tilde{p}_{x_{i-1},x_i}(1),
\end{aligned}$$

which yields the claim.

(c) Give an example of a 2-periodic, reversible, irreducible Markov chain, i.e $p_{xx}(2n-1)=0$ for all $x\in S$, $n\in\mathbb{N}$. Is $\lambda=-1$ an eigenvalue of P and if so, what is a corresponding eigenfunction?

Consider the symmetric random walk on $S=\{1,\dots,M\}$, M even and S interpreted as a torus, with $p_{x,x\pm1}(1)=1/2$ (identifying $1-1=S, S+1=1$). This is obviously a 2-periodic, reversible and irreducible Markov chain with transition operator

$$Pf(x)=\tfrac{1}{2}(f(x-1)+f(x+1)).$$

We have $Pf=-f$ iff $f(x)=-\frac{1}{2}(f(x-1)+f(x+1))$ for all $x\in S$. This implies $f(x)=\frac{1}{4}(f(x-2)+f(x)+f(x+2))$, hence $f(x)=\frac{1}{2}(f(x-2)+f(x+2))$. From this we conclude that f is constant on all even x and on all odd x (e.g. consider the maximum n these x). Hence, the functions $f(x)=c(-1)^x$, $c\in\mathbb{R}$, form the one-dimensional eigenspace of P to the eigenvalue $\lambda=-1$.

5.70 Remark. At the end of this chapter it is worthwhile to see the usage of MCMC methods in practice, e.g. `https://www.youtube.com/watch?v=h1NOS_wxgGg`.

6 Weak and functional convergence

6.1 Fundamental properties

Throughout $(S,\mathfrak{B}_S)$ denotes a metric space with Borel σ-algebra. The space of all bounded continuous and real-valued functions on S is denoted by $C_b(S)$.

6.1 Definition. Probability measures $\mathbb{P}_n$ converge weakly (schwach) to a probability measure $\mathbb{P}$ on $(S,\mathfrak{B}_S)$ if

$$\forall f\in C_b(S):\ \lim_{n\to\infty}\int_S f\,d\,\mathbb{P}_n=\int_S f\,d\,\mathbb{P}$$

holds, notation $\mathbb{P}_n\xrightarrow{w}\mathbb{P}$. $(S,\mathfrak{B}_S)$-valued random variables X_n converge in distribution (or in law, in Verteilung) to some random variable X if $\mathbb{P}^{X_n}\xrightarrow{w}\mathbb{P}^X$ holds, i.e.

$$\forall f\in C_b(S):\ \lim_{n\to\infty}\mathbb{E}[f(X_n)]=\mathbb{E}[f(X)].$$

Notation $X_n\xrightarrow{d}X$ or $X_n\xrightarrow{d}\mathbb{P}^X$.

6.2 Example. For $x_n\to x$ in S the point measures δ_{x_n} converge weakly to δ_x. Note that for $x_n\neq x$, $n\geqslant 1$, we have $0=\delta_{x_n}(\{x\})\not\to\delta_x(\{x\})=1$. In general, we cannot expect that $\mathbb{P}_n\xrightarrow{w}\mathbb{P}$ implies $\mathbb{P}_n(A)\to\mathbb{P}(A)$ for an event A.

6.3 Lemma (Continuous mapping). *If $g:S\to T$ is continuous, T another metric space, then: $X_n\xrightarrow{d}X\Rightarrow g(X_n)\xrightarrow{d}g(X)$.*

Proof. For $f\in C_b(T)$ we have $f\circ g\in C_b(S)$. Hence, $X_n\xrightarrow{d}X$ implies $\mathbb{E}[f(g(X_n))]\to\mathbb{E}[f(g(X))]$ and therefore $g(X_n)\xrightarrow{d}g(X)$. □

6.4 Example. If real-valued random variables X_n satisfy $X_n \xrightarrow{d} N(0,1)$, then $aX_n + b \xrightarrow{d} N(b, a^2)$ follows from $aZ + b \sim N(b, a^2)$ for $Z \sim N(0,1)$. Furthermore, $X_n^2 \xrightarrow{d} \chi_1^2$ (χ^2-distribution with one degree of freedom) follows from $Z^2 \sim \chi_1^2$ for $Z \sim N(0,1)$.

6.5 Definition. Let $(X_n), X$ be random variables with values in a Polish space $(S, \mathfrak{B}_S)$. Then (X_n) <u>converges in probability</u> or <u>stochastically</u> to X, notation $X_n \xrightarrow{\mathbb{P}} X$, if

$$\forall\, \varepsilon > 0: \lim_{n\to\infty} \mathbb{P}(d(X_n, X) > \varepsilon) = 0.$$

6.6 Remark. We need that S is Polish (separable suffices) to justify that $d(X_n, X)$ is a real-valued random variable ▶Exercise. Note that $X_n \xrightarrow{\mathbb{P}} X$ is therefore equivalent to the stochastic convergence $d(X_n, X) \xrightarrow{\mathbb{P}} 0$ for real-valued random variables. This allows to transfer many results for stochastic convergence from $\mathbb{R}$ to general Polish S. In particular, stochastic convergence is metrisable and $X_n \xrightarrow{\mathbb{P}} X$ implies $X_n \xrightarrow{d} X$ ▶Exercise.

6.7 Theorem (Portmanteau Theorem, Alexandrov 1940)**.** *For probability measures $(\mathbb{P}_n)_{n\in\mathbb{N}}$, $\mathbb{P}$ on $(S, \mathfrak{B}_S)$ the following are equivalent:*

(a) $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$;

(b) $\int f\, d\mathbb{P}_n \to \int f\, d\mathbb{P}$ *holds for all bounded Lipschitz-continuous functions* $f: S \to \mathbb{R}$;

(c) $\forall\, U \subseteq S$ *open*: $\liminf_{n\to\infty} \mathbb{P}_n(U) \geqslant \mathbb{P}(U)$;

(d) $\forall\, F \subseteq S$ *closed*: $\limsup_{n\to\infty} \mathbb{P}_n(F) \leqslant \mathbb{P}(F)$;

(e) $\forall\, A \in \mathfrak{B}_S$ *with* $\mathbb{P}(\partial A) = 0$: $\lim_{n\to\infty} \mathbb{P}_n(A) = \mathbb{P}(A)$.

6.8 Remark. Recall that f is Lipschitz-continuous if there is a constant $L \geqslant 0$ with $|f(x) - f(y)| \leqslant L d(x, y)$ for all $x, y \in S$. The topological boundary $\partial A = \bar{A} \setminus A^\circ$ is the difference between the closure $\bar{A}$ and the interior A° of A. Intuitively, under weak convergence $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ probability mass of $\mathbb{P}_n$ can move continuously and thus for open sets U it might get lost at the boundary: $\mathbb{P}(U) \leqslant \liminf_{n\to\infty} \mathbb{P}_n(U)$, while $\mathbb{P}(\bar{U}) \geqslant \limsup_{n\to\infty} \mathbb{P}_n(U)$ (put $F = \bar{U} \supseteq U$ in (d)).

For probability measures on $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$ the Portmanteau Theorem shows that $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ implies $F_n(x) := \mathbb{P}_n((-\infty, x]) \to \mathbb{P}((-\infty, x]) = F(x)$ for all $x \in \mathbb{R}$ with $\mathbb{P}(\partial(-\infty, x]) = \mathbb{P}(\{x\}) = 0$. These x are exactly the continuity points of the distribution function F of $\mathbb{P}$ and we find back the result from Stochastik I on the pointwise convergence of the distribution functions at continuity points.

Observe that on $(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d})$ the convergence of the characteristic functions suffices already to ensure weak convergence (Stochastik I), that is convergence for the specific test functions $f(x) = \cos(\langle u, x\rangle)$ and $f(x) = \sin(\langle u, x\rangle)$, $u \in \mathbb{R}^d$, which form a small subclass of bounded Lipschitz-continuous functions.

Proof. (a)$\Rightarrow$(b): This follow directly from the definition because Lipschitz-continuous functions are continuous.

(b)$\Rightarrow$(c): Let $U \subseteq S$ be open, $F = S \setminus U$. Then $x \mapsto \operatorname{dist}_F(x) = \inf_{y\in F} d(x,y)$ is Lipschitz-continuous: for arbitrary $x, z \in S$ and $y_n \in F$ with $d(x, y_n) \to \operatorname{dist}_F(x)$ deduce by triangle inequality

$$\operatorname{dist}_F(z) \leqslant \inf_{n\geqslant 1} d(z, y_n) \leqslant \inf_{n\geqslant 1}(d(z,x) + d(x, y_n)) = d(z,x) + \operatorname{dist}_F(x),$$

so that by symmetry $|\operatorname{dist}_F(z) - \operatorname{dist}_F(x)| \leqslant d(x,z)$ follows. Therefore $f_m(x) := (m \operatorname{dist}_F(x)) \wedge 1$, $m \in \mathbb{N}$, are bounded Lipschitz-continuous functions which satisfy $f_m \uparrow \mathbf{1}_U$ as $m \uparrow \infty$. From (b) we deduce for any $m \in \mathbb{N}$

$$\liminf_{n\to\infty} \mathbb{P}_n(U) \geqslant \liminf_{n\to\infty} \int f_m \, d\,\mathbb{P}_n = \int f_m \, d\,\mathbb{P}\,.$$

Monotone convergence gives $\lim_{m\to\infty} \int f_m d\,\mathbb{P} = \int \mathbf{1}_U d\,\mathbb{P} = \mathbb{P}(U)$ and thus (c).

(c) $\Longleftrightarrow$ (d) follows directly by taking complements.

(c,d)$\Rightarrow$(e): For all $A \in \mathfrak{B}_S$ we have by (c) and (d)

$$\begin{aligned}\mathbb{P}(A^\circ) &\leqslant \liminf_{n\to\infty} \mathbb{P}_n(A^\circ) \leqslant \liminf_{n\to\infty} \mathbb{P}_n(A)\\ &\leqslant \limsup_{n\to\infty} \mathbb{P}_n(A) \leqslant \limsup_{n\to\infty} \mathbb{P}_n(\bar{A}) \leqslant \mathbb{P}(\bar{A}).\end{aligned}$$

If $\mathbb{P}(\partial A) = \mathbb{P}(\bar{A}) - \mathbb{P}(A^\circ) = 0$ holds, then we have equality everywhere and (e) follows.

(e)$\Rightarrow$ (a): Let $f \in C_b(S)$. Since the preimages $(f^{-1}(\{y\}))_{y\in\mathbb{R}}$ are pairwise disjoint, there are at most countably many y with $\mathbb{P}(f^{-1}(\{y\})) > 0$▶Control. Without loss of generality assume $\mathbb{P}(f^{-1}(\{0\})) = 0$ and consider

$$B_{k,\varepsilon} = f^{-1}([k\varepsilon, (k+1)\varepsilon)), \quad k \in \mathbb{Z},\, \varepsilon > 0.$$

By continuity of f, we have $\partial B_{k,\varepsilon} \subseteq f^{-1}(\{k\varepsilon\}) \cup f^{-1}(\{(k+1)\varepsilon\})$ and there are only countably many pairs (k, ε) with $\mathbb{P}(\partial B_{k,\varepsilon}) > 0$. Consequently, we can choose a sequence $\varepsilon_m \downarrow 0$ with $\mathbb{P}(\partial B_{k,\varepsilon_m}) = 0$ for all m, k. Using

$$\sum_k k\varepsilon_m \mathbf{1}_{B_{k,\varepsilon_m}} \leqslant f \leqslant \sum_k (k+1)\varepsilon_m \mathbf{1}_{B_{k,\varepsilon_m}},$$

where k runs through the finite set $\{k \in \mathbb{Z} \,|\, |k| \leqslant \|f\|_\infty/\varepsilon_m + 2\}$, we use $\mathbb{P}_n(B_{k,\varepsilon_m}) \to \mathbb{P}(B_{k,\varepsilon_m})$ by (e) and obtain

$$\begin{aligned}\int f \, d\,\mathbb{P} - \varepsilon_m &\leqslant \sum_k k\varepsilon_m \, \mathbb{P}(B_{k,\varepsilon_m}) = \lim_{n\to\infty} \sum_k k\varepsilon_m \, \mathbb{P}_n(B_{k,\varepsilon_m})\\ &\leqslant \liminf_{n\to\infty} \int f \, d\,\mathbb{P}_n \leqslant \limsup_{n\to\infty} \int f \, d\,\mathbb{P}_n \leqslant \lim_{n\to\infty} \sum_k (k+1)\varepsilon_m \, \mathbb{P}_n(B_{k,\varepsilon_m})\\ &= \sum_k (k+1)\varepsilon_m \, \mathbb{P}(B_{k,\varepsilon_m}) \leqslant \int f \, d\,\mathbb{P} + \varepsilon_m.\end{aligned}$$

With $\varepsilon_m \downarrow 0$ we obtain $\lim_{n\to\infty} \int f \, d\,\mathbb{P}_n = \int f \, d\,\mathbb{P}$. □

6.9 Remark.

(a) The proof yields a simple argument why the limit for weak convergence is unique. We have constructed for open sets U functions $f_m \in C_b(S)$ with $f_m \uparrow \mathbf{1}_U$. Consequently, if $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ and $\mathbb{P}_n \xrightarrow{w} \mathbb{Q}$, then $\lim_{n\to\infty} \int f_m d\,\mathbb{P}_n = \int f_m d\,\mathbb{P} = \int f_m d\,\mathbb{Q}$ holds. By monotone convergence in m this shows $\int \mathbf{1}_U d\,\mathbb{P} = \int \mathbf{1}_U d\,\mathbb{Q}$ so that $\mathbb{P}$ and $\mathbb{Q}$ agree on all open sets. Open sets form an $\cap$-stable generator of the Borel σ-algebra and the uniqueness theorem for probability measures yields $\mathbb{P} = \mathbb{Q}$.

(b) Convergence in distribution is not compatible with addition or multiplication: $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$ does not(!) imply $X_n + Y_n \xrightarrow{d} X + Y$ or $X_n Y_n \xrightarrow{d} XY$ because the joint distribution of (X, Y) cannot be inferred. A simple counterexample is given by $X_n = X$, $Y_n = -X$ for $X \sim N(0,1)$ satisfying $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} X$, while $X_n + Y_n = 0$ does not converge in law to $X + X \sim N(0,4)$ and $X_n Y_n = -X^2$ does not converge in law to X^2. Slutsky's Lemma, which often refers just to the corollary below, yields an important sufficient condition to conclude the convergence of sums and products.

6.10 Lemma (Slutsky, 1925)**.** *Let (S,d) be Polish. We have for $(S, \mathfrak{B}_s)$-valued random variables (X_n), (Y_n), X*

$$X_n \xrightarrow{d} X,\ d(X_n, Y_n) \xrightarrow{\mathbb{P}} 0 \Rightarrow Y_n \xrightarrow{d} X.$$

Proof. Let $f \in C_b(S)$ be Lipschitz-continuous with Lipschitz constant L. Then for any $\varepsilon > 0$

$$\begin{aligned}\limsup_{n\to\infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(Y_n)]| &\leqslant \limsup_{n\to\infty} \mathbb{E}[|f(X_n) - f(Y_n)|] \\ &\leqslant \limsup_{n\to\infty} \mathbb{E}[|f(X_n) - f(Y_n)| \mathbf{1}(d(X_n, Y_n) > \varepsilon)] + L\varepsilon \\ &\leqslant 2\|f\|_\infty \limsup_{n\to\infty} \mathbb{P}(d(X_n, Y_n) > \varepsilon) + L\varepsilon = L\varepsilon.\end{aligned}$$

With $\varepsilon \downarrow 0$ we conclude $\lim_{n\to\infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(Y_n)]| = 0$. This yields

$$\lim_{n\to\infty} \mathbb{E}[f(Y_n)] = \lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

By the Portmanteau Theorem 6.7, we obtain $Y_n \xrightarrow{d} X$. □

6.11 Corollary (Slutsky's Lemma)**.** *Consider $(S, \mathfrak{B}_s)$-valued random variables (X_n), (Y_n), X and $a \in S$ for (S,d) Polish. Then*

$$X_n \xrightarrow{d} X,\ Y_n \xrightarrow{\mathbb{P}} a \Rightarrow (X_n, Y_n) \xrightarrow{d} (X, a)$$

holds. In particular, for $S = \mathbb{R}$ we have $X_n Y_n \xrightarrow{d} aX$ and $X_n + Y_n \xrightarrow{d} X + a$.

Proof. Note that the space S^2, equipped with the product metric $d_2((x_1,x_2),(y_1,y_2)) = d(x_1,y_1) + d(x_2,y_2)$, is again Polish. On one hand, we have $d_2((X_n,Y_n),(X_n,a)) = d(Y_n,a) \xrightarrow{\mathbb{P}} 0$ due to $Y_n \xrightarrow{\mathbb{P}} a$. On the other hand, $(X_n,a) \xrightarrow{d} (X,a)$ follows from

$$f \in C_b(S^2) \Rightarrow f(\bullet,a) \in C_b(S) \Rightarrow \mathbb{E}[f(X_n,a)] \to \mathbb{E}[f(X,a)].$$

Applying Slutsky's Lemma to the S^2-valued random variables (X_n,Y_n) and (X_n,a), we conclude $(X_n,Y_n) \xrightarrow{d} (X,a)$.

Noting that $(x,y) \mapsto x+y$, $(x,y) \mapsto xy$ are both continuous from $\mathbb{R}^2$ to $\mathbb{R}$, the continuous mapping theorem shows that $(X_n,Y_n) \xrightarrow{d} (X,a)$ implies $X_n + Y_n \xrightarrow{d} X + a$, $X_nY_n \xrightarrow{d} Xa$. □

6.2 Tightness

6.12 Definition. A family $(\mathbb{P}_i)_{i\in I}$ of probability measures on $(S,\mathfrak{B}_S)$ is called (weakly) relatively compact if each sequence $(\mathbb{P}_{i_k})_{k\geqslant 1}$ has a weakly convergent subsequence. This means that there is a probability measure $\mathbb{P}$ (not necessarily in $(\mathbb{P}_i)_{i\in I}$) and a subsequence (i_{k_l}) such that $\mathbb{P}_{i_{k_l}} \xrightarrow{w} \mathbb{P}$ as $l \to \infty$.

The family $(\mathbb{P}_i)_{i\in I}$ is (uniformly) tight (straff) if for any $\varepsilon > 0$ there is a compact set $K_\varepsilon \subseteq S$ such that $\mathbb{P}_i(K_\varepsilon) \geqslant 1 - \varepsilon$ for all $i \in I$.

6.13 Remarks. One can show that the set $M(S)$ of all probability measures on a Polish space $(S,\mathfrak{B}_S)$ under weak convergence is metrisable, see ►Exercise. In particular, sequential compactness and compactness are identical and 'relatively compact' just means that the closure is compact. Compare with the Heine-Borel Theorem which says that bounded subsets are relatively compact in $\mathbb{R}^d$.

Ulam's Theorem (already Proposition 2.18) shows that on a Polish space one probability measure is always tight and thus also any finite family $(\mathbb{P}_1,\ldots,\mathbb{P}_n)$.

In Stochastik I we have proved Helly's Theorem that probability measures on $\mathfrak{B}_\mathbb{R}$ are weakly relatively compact if and only if they are tight. We prove this for general Polish spaces, which requires more (topological) work.

▷ **Control questions**

(a) Find examples of probability measures and sets U, F where strict inequality holds in the Portmanteau Theorem 6.7(c,d).

For $x_n \to x$ and $x_n \neq x$ we have seen above $0 = \delta_{x_n}(F) < \delta_x(F) = 1$ for the closed set $F = \{x\}$. Obviously, the complement $U = \mathbb{R}\backslash\{x\}$ is an example for the reverse inequality.

(b) Let $(B_y)_{y\in\mathbb{R}}$ be pairwise disjoint events. Prove that $\mathbb{P}(B_y) > 0$ can only hold for countably many y.
Hint: show first that $\mathbb{P}(B_y) > 1/n$, $n \in \mathbb{N}$, holds only for finitely many y.

Let $Y_n = \{y \in \mathbb{R} \mid \mathbb{P}(B_y) \geqslant 1/n\}$. Since the B_y are pairwise disjoint, we have $1 = P(\Omega) \geqslant \sum_{y\in Y_n} \mathbb{P}(B_y) \geqslant |Y_n| n^{-1}$ and Y_n has at most n elements (formally, you should reduce uncountable Y_n to a countable subset Y_n first). This implies that $\bigcup_{n\geqslant 1} Y_n = \{y \in \mathbb{R} \mid \mathbb{P}(B_y) > 0\}$ is at most countably infinite.

(c) When do we have $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{\mathbb{P}} a \Rightarrow X_n/Y_n \xrightarrow{d} X/a$?

This holds whenever $a \neq 0$. This follows from Slutsky's Lemma by setting $Z_n = Y_n^{-1}\mathbf{1}(Y_n \neq 0)$ and checking $Z_n \xrightarrow{\mathbb{P}} a^{-1}$.

6.14 Theorem. *Any weakly relatively compact family of probability measures on a Polish space is tight.*

Proof. Similar to Ulam's Theorem ►Exercise. □

6.15 Theorem (Prohorov, 1956)**.** *Any tight family $(\mathbb{P}_i)_{i\in I}$ of probability measures on a separable metric space is weakly relatively compact.*

Proof. Note first that it suffices to prove that there is a weakly converging subsequence in $(\mathbb{P}_i)_{i\in I}$ because for any sequence $(\mathbb{P}_{i_n})_{n\geqslant 1}$ in $(\mathbb{P}_i)_{i\in I}$ we can then consider $I' = \{i_n \,|\, n \geqslant 1\}$.

By tightness we may choose nested compact sets $K_m \subseteq K_{m+1}$ with $\mathbb{P}_i(K_m) > 1 - m^{-1}$ for $m \in \mathbb{N}$, $i \in I$. In particular $\lim_{m\to\infty} \mathbb{P}_i(K_m) = 1$ holds. Let D be a countable dense subset of S and consider an open set G. Then for each $x \in G$ there is a $d \in D$ and $r \in \mathbb{Q}^+$ with $x \in B_r(d) \subseteq \overline{B_r(d)} \subseteq G$, where $B_r(d)$ is the open ball with diameter r around d. Introduce the countable family

$$\mathscr{H} := \Big\{ \bigcup_{j=1,\dots,J} \overline{B_{r_j}(d_j)} \cap K_m \,\Big|\, J \geqslant 1,\, d_j \in D,\, r_j \in \mathbb{Q}^+,\, m \geqslant 1 \Big\} \cup \{\varnothing\}.$$

Then $\mathbb{P}_i(G) = \lim_{m\to\infty} \mathbb{P}_i(G \cap K_m) = \sup_{H\in\mathscr{H}, H\subseteq G} \mathbb{P}_i(H)$, $i \in I$, follows by σ-continuity of $\mathbb{P}_i$.

By compactness of $[0,1]$, for $H \in \mathscr{H}$ and any sequence (i_n) in I there is a subsequence (i_n^H) of (i_n) such that $\mathbb{P}_{i_n^H}(H)$ converges for $n \to \infty$. By a diagonal sequence argument there is thus a subsequence $(\mathbb{P}_n)_{n\geqslant 1}$ of $(\mathbb{P}_i)_{i\in I}$ with $\lim_{n\to\infty} \mathbb{P}_n(H) = \alpha(H)$ for all $H \in \mathscr{H}$ and some $\alpha : \mathscr{H} \to [0,1]$. Below we construct a probability measure $\mathbb{P}$ on $\mathfrak{B}_S$ with $\mathbb{P}(G) = \sup_{H\in\mathscr{H}, H\subseteq G} \alpha(H)$ for open sets G. From this we conclude

$$\begin{aligned}\mathbb{P}(G) &= \sup_{H\in\mathscr{H}, H\subseteq G} \alpha(H) = \sup_{H\in\mathscr{H}, H\subseteq G} \lim_{n\to\infty} \mathbb{P}_n(H) \\ &\leqslant \liminf_{n\to\infty} \sup_{H\in\mathscr{H}, H\subseteq G} \mathbb{P}_n(H) = \liminf_{n\to\infty} \mathbb{P}_n(G).\end{aligned}$$

By the Portmanteau Theorem 6.7 we conclude weak convergence $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ and $(\mathbb{P}_i)_{i\in I}$ is weakly relatively compact.

The following construction of $\mathbb{P}$ relies on Caratheodory's Theorem and is optional for this course. First observe that $\alpha : \mathscr{H} \to [0,1]$ satisfies

$H_1 \subseteq H_2 \Rightarrow \alpha(H_1) \subseteq \alpha(H_2)$ (monotone),

$H_1 \cap H_2 = \varnothing \Rightarrow \alpha(H_1 \cup H_2) = \alpha(H_1) + \alpha(H_2)$ (finitely additive),

$H_1, H_2 \in \mathscr{H}$ arbitrary $\Rightarrow \alpha(H_1 \cup H_2) \leqslant \alpha(H_1) + \alpha(H_2)$ (sub-additive).

For open sets G define $\beta(G) := \sup_{H\in\mathscr{H}, H\subseteq G} \alpha(H)$. Then β is monotone with $\beta(\varnothing) = 0$. For arbitrary $M \subseteq S$ set $\gamma(M) := \inf_{M\subseteq G, G \text{ open}} \beta(G)$ so that $\gamma(G) = \beta(G)$ for open sets G and γ is monotone. We shall prove that γ is an outer measure and that all closed sets F are γ-measurable. This shows by Caratheodory's Theorem that $\mathbb{P} := \gamma|_{\mathfrak{B}_S}$ is a measure. It is even a probability measure because by compactness $K_m \in \mathscr{H}$ (K_m can be covered by finitely many balls $B_r(d)$) and thus

$$1 \geqslant \gamma(S) = \beta(S) \geqslant \sup_{m\geqslant 1} \alpha(K_m) \geqslant \sup_{m\geqslant 1}(1 - m^{-1}) = 1.$$

For γ to be an outer measure it remains to prove $\gamma(\bigcup_{n\geqslant 1} A_n) \leqslant \sum_{n\geqslant 1}\gamma(A_n)$ for arbitrary $A_n \subseteq S$.

First, we show that for closed F, open G and $H \in \mathscr{H}$ with $F \subseteq G\cap H$ there is $H_0 \in \mathscr{H}$ with $F \subseteq H_0 \subseteq G$. For every $x \in F$ choose $d_x \in D$, $r_x \in \mathbb{Q}^+$ with $x \in B_{r_x}(d_x) \subseteq \overline{B_{r_x}(d_x)} \subseteq G$. Since $H \subseteq K_m$ for some $m \in \mathbb{N}$, $F \subseteq H \subseteq K_m$ is compact as a closed subset of a compact set and there is a finite set $\mathcal{X}$ with $F \subseteq \bigcup_{x\in\mathcal{X}} B_{r_x}(d_x) \cap K_m$. Then $H_0 := \bigcup_{x\in\mathcal{X}} \overline{B_{r_x}(d_x)} \cap K_m \in \mathscr{H}$ satisfies $F \subseteq H_0 \subseteq G$.

Second, we show $\beta(G_1 \cup G_2) \leqslant \beta(G_1) + \beta(G_2)$ for open sets G_1, G_2. For $H \in \mathscr{H}$ with $H \subseteq G_1 \cup G_2$ the sets

$$F_1 := \{x \in H \mid \operatorname{dist}_{G_1^C}(x) \geqslant \operatorname{dist}_{G_2^C}(x)\} \subseteq G_1,$$
$$F_2 := \{x \in H \mid \operatorname{dist}_{G_2^C}(x) \geqslant \operatorname{dist}_{G_1^C}(x)\} \subseteq G_2$$

are closed ($F_1 = H \cap (\operatorname{dist}_{G_1^C} - \operatorname{dist}_{G_2^C})^{-1}([0,\infty))$ and the preimage of a closed set under a continuous functions is closed, F_2 analogously). By the first step $F_1 \subseteq H_1 \subseteq G_1$ and $F_2 \subseteq H_2 \subseteq G_2$ hold for some $H_1, H_2 \in \mathscr{H}$. We deduce

$$\alpha(H) = \alpha(F_1 \cup F_2) \leqslant \alpha(H_1) + \alpha(H_2) \leqslant \beta(G_1) + \beta(G_2).$$

This shows $\beta(G_1 \cup G_2) = \sup_{H\in\mathscr{H}, H\subseteq G_1\cup G_2} \alpha(H) \leqslant \beta(G_1) + \beta(G_2)$.

In a third step we prove $\beta(\bigcup_{n\geqslant 1} G_n) \leqslant \sum_{n\geqslant 1}\beta(G_n)$ for open sets G_n. This follows from $H \subseteq \bigcup_{n\geqslant 1} G_n$ and $H \in \mathscr{H}$ by $H \subseteq \bigcup_{n=1}^N G_n$ for some $N \in \mathbb{N}$ due to compactness of H and

$$\alpha(H) \leqslant \beta\Big(\bigcup_{n=1}^N G_n\Big) \leqslant \sum_{n=1}^N \beta(G_n) \leqslant \sum_{n\geqslant 1}\beta(G_n).$$

Finally, we prove $\gamma(\bigcup_{n\geqslant 1} A_n) \leqslant \sum_{n\geqslant 1}\gamma(A_n)$ for any $A_n \subseteq S$. Choose open sets $G_n \supseteq A_n$ with $\beta(G_n) < \gamma(A_n) + \varepsilon 2^{-n}$. Then by step 3

$$\gamma\Big(\bigcup_{n\geqslant 1} A_n\Big) \leqslant \beta\Big(\bigcup_{n\geqslant 1} G_n\Big) \leqslant \sum_{n\geqslant 1}\beta(G_n) < \sum_{n\geqslant 1}\gamma(A_n) + \varepsilon.$$

It remains to let $\varepsilon \to 0$ and γ is confirmed to be an outer measure.

At last, we must check that any closed set F is γ-measurable, that is

$$\forall\, M \subseteq S: \; \gamma(M) \geqslant \gamma(M \cap F) + \gamma(M \cap F^C).$$

Let G be open and $\varepsilon > 0$. Choose $H_0, H_1 \in \mathscr{H}$ with $H_1 \subseteq F^C \cap G$, $\alpha(H_1) > \beta(F^C \cap G) - \varepsilon$ and $H_0 \subseteq H_1^C \cap G$, $\alpha(H_0) > \beta(H_1^C \cap G) - \varepsilon$. Then $H_0 \cap H_1 = \varnothing$ and $H_0 \cup H_1 \subseteq G$ so that

$$\begin{aligned}\beta(G) &\geqslant \alpha(H_0 \cup H_1) = \alpha(H_0) + \alpha(H_1) > \beta(H_1^C \cap G) + \beta(F^C \cap G) - 2\varepsilon \\ &\geqslant \gamma(F \cap G) + \gamma(F^C \cap G) - 2\varepsilon.\end{aligned}$$

Letting $\varepsilon \to 0$ we see $\beta(G) \geqslant \gamma(F \cap G) + \gamma(F^C \cap G)$. This yields

$$\gamma(M) = \inf_{G \supseteq M \text{ open}} \beta(G) \geqslant \inf_{G \supseteq M \text{ open}} (\gamma(F \cap G) + \gamma(F^C \cap G)) \geqslant \gamma(F \cap M) + \gamma(F^C \cap M).$$

Therefore all closed sets F are γ-measurable and by Caratheodory's Theorem also the generated σ-algebra of Borel sets. □

6.16 Corollary (Prohorov). *On a Polish space a family of probability measures is weakly relatively compact if and only if it is tight.*

Proof. Combine Theorems 6.14 and 6.15. □

6.3 Weak convergence on $C([0,T])$, $C(\mathbb{R}^+)$

In the sequel C stands for $C([0,T]) = \{f : [0,T] \to \mathbb{R} \mid f \text{ continuous}\}$ or $C(\mathbb{R}^+) = \{f : [0,\infty) \to \mathbb{R} \mid f \text{ continuous}\}$, equipped with the supremum norm and the uniform convergence on compact sets ($f_n \to f$ if $\sup_{t \in [0,T]} |f_n(t) - f(t)| \to 0$ for all $T > 0$), respectively. Then C forms a Polish space. By an exercise the Borel σ-algebra $\mathfrak{B}_C$ is generated by the coordinate projections $\pi_t : C \to \mathbb{R}$, $\pi_t(f) = f(t)$ for all $t \in [0,T]$ and $t \geqslant 0$, respectively. Hence, a probability measure $\mathbb{P}$ on C is uniquely determined by its finite-dimensional distributions $(\mathbb{P}^{\pi_{t_1,\dots,t_m}})_{m \in \mathbb{N}, 0 \leqslant t_1 < \dots < t_m}$ with $\pi_{t_1,\dots,t_m}(f) = (f(t_1), \dots, f(t_m))$.

6.17 Theorem. *A sequence $(\mathbb{P}_n)$ of probability measures on $\mathfrak{B}_C$ converges weakly to $\mathbb{P}$ if and only if all finite-dimensional distributions $\mathbb{P}_n^{\pi_{t_1,\dots,t_m}}$ converge weakly to $\mathbb{P}^{\pi_{t_1,\dots,t_m}}$ and $(\mathbb{P}_n)$ is tight.*

Proof. Since C is Polish, $(\mathbb{P}_n)$ is weakly relatively compact if and only if it is tight by Prohorov's Theorem. Therefore, $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ implies that $(\mathbb{P}_n)$ is tight. Since all $\pi_{t_1,\dots,t_m} : S \to \mathbb{R}^m$ are continuous, the continuous mapping theorem, applied to probability measures, gives $\mathbb{P}_n^{\pi_{t_1},\dots,\pi_{t_m}} \xrightarrow{w} \mathbb{P}^{\pi_{t_1,\dots,t_m}}$.

Conversely, suppose $(\mathbb{P}_n)$ is tight and consider the set $\mathcal{Q}$ of all probability measures $\mathbb{Q}$ such that $\mathbb{P}_{n_k} \xrightarrow{w} \mathbb{Q}$ for some subsequence (n_k). By tightness, $\mathcal{Q}$ is not empty. By continuous mapping, also the finite-dimensional distributions of $\mathbb{P}_{n_k}$ converge to those of $\mathbb{Q}$. By assumption, they converge to the finite-dimensional distributions of $\mathbb{P}$. By the above uniqueness result, we infer $\mathcal{Q} = \{\mathbb{P}\}$ and $\mathbb{P}$ is the only accumulation point of $(\mathbb{P}_n)$. This proves already $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ by a general result for compact metric spaces▶Control, but we provide a concrete proof by contradiction.

If $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ were not true, then there would be a subsequence (n_k) and some $\varepsilon > 0, f \in C_b(S)$ such that $|\int f \, d\mathbb{P}_{n_k} - \int f \, d\mathbb{P}| \geqslant \varepsilon$ for all k. From (n_k), however, we could extract by tightness a subsubsequence (n_{k_l}) with $\mathbb{P}_{n_{k_l}} \xrightarrow{w} \mathbb{P}$ and thus $|\int f \, d\mathbb{P}_{n_{k_l}} - \int f \, d\mathbb{P}| \to 0$. This contradiction proves $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$. □

6.18 Remark. The previous theorem is the workhorse for proving convergence in law for continuous processes $(X_n(t), t \geqslant 0)$ to $(X(t), t \geqslant 0)$. One shows finite-dimensional convergence $(X_n(t_1), \dots, X_n(t_m)) \xrightarrow{d} (X(t_1), \dots, X(t_m))$ for any $0 \leqslant t_1 < \dots < t_m$, which is often easy, and then establishes tightness of the laws $\mathbb{P}^{X_n}$ in C, which is usually more involved and for which we shall study criteria in the following. Note that convergence in law in C allows asymptotic results for many path-dependent functionals, for example $\max_{0 \leqslant t \leqslant T} X_n(t) \xrightarrow{d} \max_{0 \leqslant t \leqslant T} X(t)$ follows from $X_n \xrightarrow{d} X$ in C and the continuity of $f \mapsto \max_{0 \leqslant t \leqslant T} f(t)$, $f \in C$, via the continuous mapping theorem.

▷ **Control questions**

(a) Do $f_n : \mathbb{R}^+ \to \mathbb{R}$ with $f_n(x) = e^{x/n}$ converge pointwise, uniformly on compact sets and/or uniformly on $\mathbb{R}^+$?

We have $f_n \to f$ uniformly on compacts for $f(x) = 1$: $\sup_{x \in [0,R]} |f_n(x) - f(x)| = e^{R/n} - 1 \to 0$ for all fixed $R > 0$. Yet, $\|f_n - f\|_\infty = \infty$ on the positive axis and there cannot be uniform convergence.

(b) Find an example of continuous processes $(X_n(t), t \in [0,1])$ whose finite-dimensional distributions converge to those of $(X(t), t \in [0,1])$, but whose laws in C do not converge.
Hint: Look at the deterministic situation.

The standard example from analysis is $X_n(t) = t^n$ which converges pointwise to $X(t) = \mathbf{1}_{\{1\}}(t)$, $t \in [0,1]$.

(c) Show that a sequence (x_n) in a compact metric space with exactly one accumulation point x converges to x.

For $\varepsilon_m \downarrow 0$ consider a finite cover $\bigcup_{n=1}^{N_m} B_{\varepsilon_m}(x_{n,m})$ of the compact space K with open balls of radius ε_m. Suppose w.l.o.g. $x \in B_{\varepsilon_m}(x_{1,m})$ for all m. Then $K \setminus B_{\varepsilon_m}(x_{1,m})$ can only contain finitely many sequence elements x_n (otherwise there would be another accumulation point away from x). This shows that for each m there is some $n_m \in \mathbb{N}$ with $x_n \in B_{2\varepsilon_m}(x)$ for all $n \geqslant n_m$. This proves $x_n \to x$. One could, of course, also apply a subsubsequence argument as above in the proof.

6.19 Definition. For $f \in C([0,T])$ and $\delta > 0$ the modulus of continuity (Stetigkeitsmodul) is defined as

$$\omega_\delta(f) := \max\{|f(s) - f(t)| \mid s, t \in [0,T], \ |s-t| \leqslant \delta\}.$$

6.20 Theorem (Arzelà-Ascoli)**.** *A subset $A \subseteq C([0,T])$ is relatively compact if and only if*

(a) $\sup_{f \in A} |f(0)| < \infty$ *and*

(b) $\lim_{\delta \to 0} \sup_{f \in A} \omega_\delta(f) = 0$ *(equi-continuity, gleichgradige Stetigkeit).*

Relative compactness in $C(\mathbb{R}^+)$ holds if (a) holds and (b) is satisfied for all $T > 0$, i.e. $\lim_{\delta \to 0} \sup_{f \in A} \max\{|f(s) - f(t)| \mid s, t \in [0,T], \ |s-t| \leqslant \delta\} = 0$ *holds for all $T > 0$.*

Proof. See e.g. H. Heuser, *Lehrbuch der Analysis I*, Teubner or T. Bühler, D. Salamon, *Functional Analysis*, AMS Graduate Studies in Mathematics. □

6.21 Example. Suppose $A \subseteq C([0,T])$ consists of functions f with $\|f\|_\infty \leqslant R$ and $|f(s)-f(t)| \leqslant L|s-t|$ for all $0 \leqslant s,t \leqslant T$ and some constants $R, L > 0$, that is A is uniformly bounded and has a uniform Lipschitz constant. Then A is relatively compact because $\sup_{f\in A}|f(0)| \leqslant R$ and $\sup_{f\in A}\omega_\delta(f) \leqslant L\delta \to 0$ for $\delta \to 0$. Note that by the mean value theorem differentiable functions f have Lipschitz constant $L = \sup_{t\in[0,T]}|f'(t)|$ so that in particular any family of uniformly bounded functions with uniformly bounded derivatives is relatively compact in $C([0,T])$.

6.22 Corollary. *A sequence $(\mathbb{P}_n)_{n\geqslant 1}$ of probability measures on $\mathfrak{B}_{C([0,T])}$ is tight if and only if*

- *(a)* $\lim_{R\to\infty}\limsup_{n\to\infty}\mathbb{P}_n(\{|f(0)|>R\}) = 0$ *and*
- *(b)* $\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}_n(\{\omega_\delta(f)\geqslant\varepsilon\}) = 0$ *for all* $\varepsilon > 0$.

6.23 Remark. Events $\{f \in C \,|\, \text{some property of } f\} \subseteq C$ are, as usually in stochastics, abbreviated by $\{\text{some property of } f\}$. We keep, however, the braces within probabilities.

Proof. If $(\mathbb{P}_n)$ is tight, then for $\eta > 0$ there is a compact set K_η with $\mathbb{P}_n(K_\eta) \geqslant 1-\eta$, $n \geqslant 1$. By the Arzelà-Ascoli Theorem, for any $\varepsilon > 0$ we have

$$K_\eta \subseteq \{|f(0)| < R_\eta, \omega_{\delta_{\eta,\varepsilon}}(f) < \varepsilon\}$$

for sufficiently large $R_\eta > 0$ and small $\delta_{\eta,\varepsilon} > 0$. With $\eta \downarrow 0$ we deduce

$$\lim_{R\to\infty}\sup_n \mathbb{P}_n(\{|f(0)| \geqslant R\}) = 0, \quad \lim_{\delta\to 0}\sup_{n\to\infty}\mathbb{P}_n(\{\omega_\delta(f)\geqslant\varepsilon\}) = 0,$$

which implies (a), (b) due to $\limsup \leqslant \sup$.

Conversely, given (a), (b) and $\eta > 0$ choose $R > 0$ with

$$\limsup_{n\to\infty}\mathbb{P}_n(\{|f(0)| > R\}) \leqslant \eta/2$$

and for each $k \in \mathbb{N}$ some $\delta_k > 0$ with

$$\limsup_{n\to\infty}\mathbb{P}_n(\{\omega_{\delta_k}(f) > 1/k\}) \leqslant \eta 2^{-k-1}.$$

By the other direction in the Arzelà-Ascoli Theorem, the set $K_\eta = \{|f(0)| \leqslant R\} \cap \bigcap_{k\geqslant 1}\{\omega_{\delta_k}(f) \leqslant 1/k\}$ is relatively compact and satisfies $\limsup_{n\to\infty}\mathbb{P}_n(K_\eta^C) \leqslant \eta/2 + \sum_{k\geqslant 1}\eta 2^{-k-1} = \eta$. This shows for the compact set $\overline{K}_\eta$: $\liminf_{n\to\infty}\mathbb{P}_n(\overline{K}_\eta) \geqslant 1-\eta$. Pick $n_0 \in \mathbb{N}$ such that for $n \geqslant n_0$ we have $\mathbb{P}_n(\overline{K}_\eta) \geqslant 1-2\eta$ and then some compact set C_η with $\mathbb{P}_n(C_\eta) \geqslant 1-2\eta$ for the finitely many indices $n = 1, \ldots, n_0-1$. Then $\inf_{n\geqslant 1}\mathbb{P}_n(\overline{K}_\eta \cup C_\eta) \geqslant 1-2\eta$ holds with the compact set $\overline{K}_\eta \cup C_\eta$. This shows tightness. □

6.24 Lemma. *A sequence $(\mathbb{P}_n)_{n\geqslant 1}$ of probability measures on $\mathfrak{B}_{C([0,T])}$ is already tight if*

(a) $\lim_{R\to\infty}\limsup_n \mathbb{P}_n(\{|f(0)|>R\})=0$ *and*

(b') *for all* $\varepsilon>0$

$$\lim_{r\to\infty}\limsup_{n\to\infty}\sup_{t\in[0,T-2^{-r}]} 2^r\,\mathbb{P}_n\Big(\Big\{\max_{s\in[t,t+2^{-r}]}|f(s)-f(t)|\geqslant\varepsilon\Big\}\Big)=0.$$

6.25 Remark. Tightness on $\mathfrak{B}_{C(\mathbb{R}^+)}$ follows if conditions (a), (b') are satisfied for all $T>0$. The main advantage of (b') compared to (b) is that the maximum over t is pulled out of the probability at the cost of the additional factor 2^r.

Proof. To prove (b')$\Rightarrow$ (b) in Corollary 6.22, let $\varepsilon,\eta>0$ and choose $r,n_0\in\mathbb{N}$ with

$$\forall\, n\geqslant n_0,\; t\in[0,T-2^{-r}]:\; 2^r\,\mathbb{P}_n\Big(\Big\{\max_{s\in[t,t+2^{-r}]}|f(s)-f(t)|\geqslant\varepsilon/2\Big\}\Big)\leqslant\frac{\eta}{2T}.$$

Suppose $\omega_{2^{-r-1}}(f)\geqslant\varepsilon$ for some $f\in C([0,T])$. Then there are $t<s\leqslant t+2^{-r-1}$ with $|f(t)-f(s)|\geqslant\varepsilon$. For $k\in\{0,\dots,\lceil 2^{r+1}T\rceil-2\}$ with $k2^{-r-1}\leqslant t<s\leqslant(k+2)2^{-r-1}$ this implies $|f(t)-f(k2^{-r-1})|\geqslant\varepsilon/2$ or $|f(s)-f(k2^{-r-1})|\geqslant\varepsilon/2$. Consequently, we obtain for $n\geqslant n_0$

$$\begin{aligned}\mathbb{P}_n(\{\omega_{2^{-r-1}}(f)\geqslant\varepsilon\})&\leqslant\sum_{k=0}^{\lceil 2^{r+1}T\rceil-2}\mathbb{P}_n\Big(\Big\{\max_{s\in[k2^{-r-1},(k+2)2^{-r-1}]}|f(s)-f(k2^{-r-1})|\geqslant\varepsilon/2\Big\}\Big)\\&\leqslant 2^{r+1}T2^{-r}\frac{\eta}{2T}=\eta.\end{aligned}$$

For $\eta\downarrow 0$, noting the monotonicity of $\delta\mapsto\mathbb{P}_n(\{\omega_\delta(f)\geqslant\varepsilon\})$, this gives condition (b). □

6.26 Theorem (Kolmogorov, Centsov 1956)**.** *Let* $(X_n(t),\,0\leqslant t\leqslant T)$*,* $n\geqslant 1$*, be continuous processes. Then their laws* $\mathbb{P}^{X_n}$ *are tight on* $C([0,T])$ *if*

(a) $\lim_{R\to\infty}\limsup_n \mathbb{P}(\{|X_n(0)|>R\})=0$ *and*

(b") $\exists\alpha,\,\beta>0,\,K>0\,\forall\,n\geqslant 1,\,s,t\in[0,T]:\mathbb{E}[|X_n(s)-X_n(t)|^\alpha]\leqslant K|s-t|^{1+\beta}$.

6.27 Remark. Condition (b") is usually much easier to check than (b) or (b'). In particular, for even integers α we may neglect the absolute value. Below, we shall apply it for $\alpha=4$ and $\beta=1$.

Observe that for Lipschitz continuous processes X_n, i.e. $|X_n(s)-X_n(t)|\leqslant L_n|s-t|$ with real random variables L_n for all s,t, the Kolmogorov-Centsov Theorem yields tightness as soon as $\sup_{n\geqslant1}\mathbb{E}[L_n^\alpha]<\infty$ for some $\alpha>1$ (choose $\beta=\alpha-1$ in (b")) and (a) hold.

Proof. Putting $\mathbb{P}_n=\mathbb{P}^{X_n}$, condition (a) in Lemma 6.24 is verified directly. For notational simplicity let us consider in condition (b') the probability $\mathbb{P}_n(\{\max_{s\in[t,t+2^{-r}]}|f(s)-f(t)|\geqslant\varepsilon\})$ for $t=0$ only. All arguments will remain valid for general t if s is replaced by $s-t$ in the sequel.

We apply the famous *chaining argument*. Consider $D = \{k2^{-m} \mid m \in \mathbb{N},\, k \in \mathbb{N}_0\}$ (dyadic numbers) and $\gamma = (\beta - \delta)/\alpha > 0$ for some $\delta \in (0, \beta)$. By Markov's inequality and the assumption, we obtain for $c > 0$

$$\begin{aligned}&\mathbb{P}(|X_n(k2^{-j}) - X_n((k-1)2^{-j})| \geqslant c2^{-\gamma j})\\ &\leqslant \mathbb{E}\left[|X_n(k2^{-j}) - X_n((k-1)2^{-j})|^{\alpha}\right] c^{-\alpha}2^{\alpha\gamma j} \leqslant c^{-\alpha}K2^{-(1+\delta)j}.\end{aligned}$$

For $s \in [0, 2^{-r}) \cap D$ we write $s = \sum_{l=r+1}^{m} b_l 2^{-l}$ for some $m \in \mathbb{N}$ and $b_l \in \{0, 1\}$. Introducing $s_j = \sum_{l=r+1}^{j} b_l 2^{-l}$ for $j \leqslant m$, we note $|s_{j+1} - s_j| \leqslant 2^{-(j+1)}$ and $s_m = s$. With $s_r := 0$ we have the telescoping sum

$$X_n(s) - X_n(0) = \sum_{j=r+1}^{m} (X_n(s_j) - X_n(s_{j-1})).$$

If $|X_n(s_j) - X_n(s_{j-1})| < c2^{-\gamma j}$ with $c = 2 - 2^{1-\gamma}$ holds for all j, then $|X_n(s) - X_n(0)| < c\sum_{j>r} 2^{-\gamma j} = 2^{-\gamma r}$ follows by evaluating the geometric series. This shows

$$\bigcup_{s \in [0,2^{-r}) \cap D} \left\{|X_n(s) - X_n(0)| \geqslant 2^{-\gamma r}\right\} \subseteq \bigcup_{\substack{j \geqslant r+1 \\ 1 \leqslant k \leqslant 2^{j-r}}} \left\{|X_n(k2^{-j}) - X_n((k-1)2^{-j})| \geqslant c2^{-\gamma j}\right\}.$$

Using that D is dense and X_n is continuous, we conclude

$$\begin{aligned}&\mathbb{P}\Big(\max_{s \in [0,2^{-r}]} |X_n(s) - X_n(0)| > 2^{-\gamma r}\Big) \leqslant \mathbb{P}\Big(\bigcup_{s \in [0,2^{-r}) \cap D} \left\{|X_n(s) - X_n(0)| \geqslant 2^{-\gamma r}\right\}\Big)\\ &\leqslant \sum_{j \geqslant r+1} \sum_{k=1}^{2^{j-r}} \mathbb{P}\Big(|X_n(k2^{-j}) - X_n((k-1)2^{-j})| \geqslant c2^{-\gamma j}\Big)\\ &\leqslant \sum_{j \geqslant r+1} \sum_{k=1}^{2^{j-r}} c^{-\alpha}K2^{-(1+\delta)j} = c^{-\alpha}K2^{-r} \sum_{j \geqslant r+1} 2^{-\delta j} = c^{-\alpha}K2^{-r}\frac{2^{-\delta(r+1)}}{1 - 2^{-\delta}}.\end{aligned}$$

Thus, uniformly over $n \in \mathbb{N}$ and $t \in [0, T - 2^{-r}]$ we obtain

$$2^r\, \mathbb{P}\Big(\max_{s \in [t, t+2^{-r}]} |X_n(s) - X_n(t)| > 2^{-\gamma r}\Big) \leqslant c^{-\alpha}K(2^{\delta} - 1)^{-1}2^{-\delta r} \to 0$$

as $r \to \infty$. By monotonicity, this convergence continues to hold if '$> 2^{-\gamma r}$' inside the probability is replaced by '$\geqslant \varepsilon$' for some fixed $\varepsilon > 0$, which yields condition (b') to be proved. □

6.28 Remark. By the Arzelà-Ascoli Theorem for $C(\mathbb{R}^+)$, the Kolmogorov-Centsov Theorem extends to the case $C(\mathbb{R}^+)$, when the conditions hold for all $T > 0$, that is the moment condition is satisfied for all $s, t \geqslant 0$.

▷ **Control questions**

(a) Suppose $(X_n(t), t \geqslant 0)$ are uniformly γ-Hölder continuous in the sense that there are deterministic $\alpha \in (0,1)$, $L > 0$ such that a.s. $|X_n(t) - X_n(s)| \leqslant L|t-s|^\gamma$ for all $s,t \geqslant 0$, $n \in \mathbb{N}$. Are the laws of $(X_n(t), t \geqslant 0)$ tight in $C(\mathbb{R}^+)$?

Yes, if we assume tightness of $(X_n(0))_{n\geqslant 1}$, that is condition (a) of Corollary 6.22. This follows from $\sup_n \omega_\delta(X_n) \leqslant L\delta^\gamma$ a.s. so that condition (b) of Corollary 6.22 is satisfied.

(b) The Kolmogorov-Centsov Theorem does not hold in the case $\beta = 0$. Try to find yourself or in the literature a counterexample.

The Poisson process $N(t)$ of intensity λ satisfies $\mathbb{E}[|N(t) - N(s)|] = \lambda|t-s|$. This remains to hold for linear interpolations. We can choose linear interpolations $\tilde{N}_n(t)$ of $N(t)$ in the neighbourhood of the jumps of $N(t)$ so that $\lim_{n\to\infty} \tilde{N}_n(t) = N(t)$ follows. Hence, the laws of $\tilde{N}$ do not converge in $C(\mathbb{R}^+)$ to a law of a continuous process. Tightness fails.

(c) Deduce from the proof of the Kolmogorov-Centsov Theorem that the processes X_n must be a.s. γ-Hölder continuous with $\gamma \in (0, \beta/\alpha)$.

By a union bound the last display in the proof (line -4) implies $\mathbb{P}(\max_{s,t:t\leqslant s\leqslant t+\delta_r} |X_n(s) - X_n(t)| > 2\delta_r^\gamma) \to 0$ for $\delta_r \to 0$. From this one can deduce by a contradiction argument that X_n must have γ-Hölder continuous paths.

7 Invariance principle and the empirical process

7.1 Invariance principle and Brownian motion

7.1 Definition. A process $(B_t,\, t \geqslant 0)$ is called Brownian motion (Brownsche Bewegung) if

(a) $B_0 = 0$ and $B_t \sim N(0,t)$, $t > 0$, holds;

(b) the increments are stationary and independent: for $0 \leqslant t_0 < t_1 < \cdots < t_m$ we have

$$(B_{t_1} - B_{t_0}, \ldots, B_{t_m} - B_{t_{m-1}}) \sim N(0, \operatorname{diag}(t_1 - t_0, \ldots, t_m - t_{m-1})).$$

(c) B has continuous sample paths.

7.2 Remark. The existence of Brownian motion is non-trivial. Without the continuity assumption this follows from the construction of Gaussian processes by Kolmogorov's consistency theorem. Yet, one can show that the set of continuous paths is not even measurable with respect to the product σ-algebra.

We had seen that discrete-time processes with stationary and independent increments are exactly given by random walks $S_n = \sum_{k=1}^n X_k$ with i.i.d. random variables X_k. If we assume that the X_k are in L^2 and standardised ($\mathbb{E}[X_k] = 0$, $\operatorname{Var}(X_k) = 1$), we have $\mathbb{E}[S_n] = 0$ and $\operatorname{Var}(aS_{bn}) = a^2bn$ for $a, b \geqslant 0$ with $bn \in \mathbb{N}_0$. This means that we can rescale the random walk by a shrinking factor $a = n^{-1/2}$ in space to obtain $\operatorname{Var}(n^{-1/2}S_{bn}) = b$ for all $n \in \mathbb{N}$. Rescaling also time $Y_n(t) := n^{-1/2} S_{nt}$, the central limit theorem then even shows $Y_n(t) \xrightarrow{d} N(0,t)$ for rational numbers $t \geqslant 0$ and along a sequence $n \to \infty$ with $nt \in \mathbb{N}_0$. Obviously, the Y_n have still independent and stationary increments (for all t where defined). In short, by rescaling space and time of a random walk, we zoom away from the original and obtain processes $Y_n(t)$ which have asymptotically the law of Brownian motion. We make this precise by using interpolations of Y_n in the sequel.

Below, we shall prove existence of Brownian motion as a by-product of the result that laws of the interpolated Y_n converge in $C([0,1])$ or $C(\mathbb{R}^+)$ towards a limit law under which the coordinate projections form a Brownian motion.

7.3 Lemma. *Suppose $(X_k)_{k \geqslant 1}$ are i.i.d., $X_k \in L^2$, $\mathbb{E}[X_k] = 0$, $\operatorname{Var}(X_k) = 1$. Consider $S_n := \sum_{k=1}^n X_k$, $S_0 = 0$ and the rescaled, linearly interpolated random walk*

$$Y_n(t) := \frac{1}{\sqrt{n}} \Big(S_{\lfloor nt \rfloor} + (nt - \lfloor nt \rfloor) X_{\lfloor nt \rfloor + 1} \Big), \quad t \geqslant 0.$$

Then the finite-dimensional distributions of Y_n converge to those of a Brownian motion.

Proof. Because of $Y_n(0) = B_0 = 0$ we just consider increments along $0 = t_0 < t_1 < \ldots < t_m$. We write

$$Y_n(t_j) = \sum_{i=1}^{j} Z_i^{(n)} + \frac{nt_j - \lfloor nt_j \rfloor}{\sqrt{n}} X_{\lfloor nt_j \rfloor + 1} \text{ with } Z_i^{(n)} = \frac{1}{\sqrt{n}} \sum_{k=\lfloor nt_{i-1} \rfloor + 1}^{\lfloor nt_i \rfloor} X_k.$$

The central limit theorem yields for $n \to \infty$

$$\frac{\sqrt{n}}{\sqrt{\lfloor nt_i \rfloor - \lfloor nt_{i-1} \rfloor}} Z_i^{(n)} \xrightarrow{d} N(0,1).$$

Since $\frac{\sqrt{n}}{\sqrt{\lfloor nt_i \rfloor - \lfloor nt_{i-1} \rfloor}} \to \frac{1}{\sqrt{t_i - t_{i-1}}}$ holds, Slutsky's Lemma and scalings of the normal distribution imply $Z_i^{(n)} \xrightarrow{d} \bar{Z}_i \sim N(0, t_i - t_{i-1})$. Now observe that $Z_1^{(n)}, \ldots, Z_m^{(n)}$ are independent for each n. A consequence of weak convergence is that then $(Z_1^{(n)}, \ldots, Z_m^{(n)}) \xrightarrow{d} (\bar{Z}_1, \ldots, \bar{Z}_m)$ holds with independent $\bar{Z}_1, \ldots, \bar{Z}_m$. By continuous mapping, we conclude

$$\begin{aligned}(Z_1^{(n)}, Z_1^{(n)} + Z_2^{(n)}, \ldots, Z_1^{(n)} + \cdots + Z_m^{(n)}) &\xrightarrow{d} (\bar{Z}_1, \bar{Z}_1 + \bar{Z}_2, \ldots, \bar{Z}_1 + \cdots + \bar{Z}_m) \\ &\overset{d}{=} (B_{t_1}, B_{t_2}, \ldots, B_{t_m}).\end{aligned}$$

Let us remark that this step follows more easily by the multivariate central limit theorem.

Finally, observe $\mathbb{E}[(\frac{nt_j - \lfloor nt_j \rfloor}{\sqrt{n}} X_{\lfloor nt_j \rfloor + 1})^2] \leqslant \frac{1}{n} \mathbb{E}[X_1^2] \to 0$ such that another application of Slutsky's Lemma gives $(Y_n(t_1), Y_n(t_2), \ldots, Y_n(t_m)) \xrightarrow{d} (B_{t_1}, B_{t_2}, \ldots, B_{t_m})$. □

7.4 Theorem (Invariance principle, functional CLT, Donsker 1951)**.** *Suppose $(X_k)_{k\geqslant 1}$ are i.i.d., $X_k \in L^2$, $\mathbb{E}[X_k]=0$, $\operatorname{Var}(X_k)=1$ and set $S_n := \sum_{k=1}^n X_k$, $S_0=0$. The rescaled, linearly interpolated random walk*

$$Y_n(t) := \frac{1}{\sqrt{n}}\Big(S_{\lfloor nt\rfloor} + (nt-\lfloor nt\rfloor)X_{\lfloor nt\rfloor+1}\Big), \quad t\geqslant 0,$$

satisfies $Y_n \xrightarrow{d} B$ with a Brownian motion $(B_t,\, t\geqslant 0)$ and convergence in distribution on $(C(\mathbb{R}^+), \mathfrak{B}_{C(\mathbb{R}^+)})$. In particular, Brownian motion exists.

Proof. Here we give the proof under the additional assumption $X_k \in L^4$, which permits an application of the Kolmogorov-Centsov criterion. Due to $Y_n(0)=B_0=0$ and Lemma 7.3 it suffices to check tightness via

$$\exists K>0\,\forall\, s,t\geqslant 0:\ \mathbb{E}[(Y_n(t)-Y_n(s))^4]\leqslant K(t-s)^2.$$

From tightness it follows in particular that the limit law (so-called Wiener measure) exists on $(C(\mathbb{R}^+), \mathfrak{B}_{C(\mathbb{R}^+)})$. Under this law $B_t(\omega)=\omega(t)$ for $\omega\in\Omega=C(\mathbb{R}^+)$ is clearly continuous in t and has the correct finite-dimensional distributions such that $(B_t, t\geqslant 0)$ forms a Brownian motion. Let us write for $t>s$

$$Y_n(t)-Y_n(s)=\frac{1}{\sqrt{n}}(S_{\lfloor nt\rfloor}-S_{\lfloor ns\rfloor})+\frac{nt-\lfloor nt\rfloor}{\sqrt{n}}X_{\lfloor nt\rfloor+1}-\frac{ns-\lfloor ns\rfloor}{\sqrt{n}}X_{\lfloor ns\rfloor+1}.$$

We shall use $(A+B)^4\leqslant 2^3(A^4+B^4)$ several times, but in general just write C_i, $i=1,\dots$ for some numerical constants.

In the case $t-s\geqslant \frac{1}{n}$ we have $\mathbb{E}[(\frac{nt-\lfloor nt\rfloor}{\sqrt{n}}X_{\lfloor nt\rfloor+1})^4]\leqslant n^{-2}\,\mathbb{E}[X_1^4]\leqslant C_1(t-s)^2$ and similarly for the term in s instead of t. By the independence of (X_k) and $\mathbb{E}[X_k]=0$ we have for $L>l$

$$\begin{aligned}\mathbb{E}[(S_L-S_l)^4]&=\sum_{k=l+1}^{L}\mathbb{E}[X_k^4]+2\sum_{l+1\leqslant k_1<k_2\leqslant L}\mathbb{E}[X_{k_1}^2]\,\mathbb{E}[X_{k_2}^2]\\&=(L-l)\,\mathbb{E}[X_1^4]+(L-l)(L-l+1)\,\mathbb{E}[X_1^2]^2\leqslant C_2(L-l)^2.\end{aligned}$$

For $n(t-s)\geqslant 1$ this shows

$$\mathbb{E}[(\tfrac{1}{\sqrt{n}}(S_{\lfloor nt\rfloor}-S_{\lfloor ns\rfloor}))^4]\leqslant C_2n^{-2}(\lfloor nt\rfloor-\lfloor ns\rfloor)^2\leqslant C_3(t-s)^2.$$

We conclude $\mathbb{E}[(Y_n(t)-Y_n(s))^4]\leqslant C_4(t-s)^2$ provided $t-s\geqslant\frac{1}{n}$.

If $t-s<\frac{1}{n}$ and $\lfloor nt\rfloor=\lfloor ns\rfloor$ holds, then $Y_n(t)-Y_n(s)=\sqrt{n}(t-s)X_{\lfloor nt\rfloor+1}$. This gives $\mathbb{E}[(Y_n(t)-Y_n(s))^4]\leqslant C_5n^2(t-s)^4\leqslant C_5(t-s)^2$.

If $t-s<\frac{1}{n}$ and $\lfloor nt\rfloor=\lfloor ns\rfloor+1$ holds, then we have $Y_n(t)-Y_n(s)=\frac{nt-\lfloor nt\rfloor}{\sqrt{n}}X_{\lfloor nt\rfloor+1}+\frac{\lfloor nt\rfloor-ns}{\sqrt{n}}X_{\lfloor nt\rfloor}$. This implies

$$\begin{aligned}\mathbb{E}[(Y_n(t)-Y_n(s))^4]&\leqslant C_6\Big(\frac{(nt-\lfloor nt\rfloor)^4}{n^2}\,\mathbb{E}[X_1^4]+\frac{(\lfloor nt\rfloor-ns)^4}{n^2}\,\mathbb{E}[X_1^4]\Big)\\&\leqslant C_7n^{-2}(nt-ns)^4\leqslant C_7(t-s)^2.\end{aligned}$$

With $K=C_4\vee C_5\vee C_7$ the Kolmogorov-Centsov criterion is satisfied. □

▷ **Control questions**

(a) Show that $X(t) = a^{-1/2}B_{at}$, $t \geqslant 0$, for a Brownian motion B and $a > 0$ is again a Brownian motion (scale invariance).

Obviously, X has again stationary and independent increments and $X(0) = 0$ holds. Since $X(t) \sim N(0, a^{-1}at) = N(0,t)$ and X is again continuous, it is a Brownian motion.

(b) Show that $B_T - B_{T-t}$, $0 \leqslant t \leqslant T$, is also again a Brownian motion on the interval $[0,T]$ (time reversal).

$t \mapsto B_T - B_{T-t}$ is continuous and zero for $t = 0$. Its increments are $B_T - B_{T-t_1}, B_{T-t_1} - B_{T-t_2}, \dots, B_{T-t_{m-1}} - B_{T-t_m}$ for $0 \leqslant t_1 < \cdots < t_m \leqslant T$ and hence again independent and stationary. Finally, $B_T - T_{T-t} \overset{d}{=} B_t \sim N(0,t)$ holds and we obtain again a Brownian motion.

(c) Show that $(B_1(t) + B_2(t))/\sqrt{2}$, $t \geqslant 0$, for two independent Brownian motions B_1 and B_2 is again a Brownian motion. Can you check this also via random walks and Donsker's Theorem?

$(B_1(t) + B_2(t))/\sqrt{2}$ starts in zero, has stationary and independent increments and is continuous. Because of $\mathbb{E}[(B_1(t)+B_2(t))^2/2] = t$ we also have $(B_1(t) + B_2(t))/\sqrt{2} \sim N(0,t)$ and thus again a Brownian motion.
Via Donsker's Invariance Principle one could argue that two independent interpolated random walks $Y_{1,n}(t)$, $Y_{2,n}(t)$ converge (by independence) jointly to the law of two independent Brownian motions $(B_1(t), B_2(t))_{t\geqslant 0}$. On the other hand, $(Y_{1,n}(t) + Y_{2,n}(t))/\sqrt{2}$ is also an interpolated random walk converging to the law of some Brownian motion $B(t)$. By continuous mapping we must have $(B_1(t) + B_2(t))/\sqrt{2} \overset{d}{=} B(t)$ in $C(\mathbb{R}^+)$.

7.5 Proposition (Reflection principle for simple random walk)**.** *Consider the simple symmetric random walk $S_n = \sum_{k=1}^n X_k$ with independent random variables $(X_k)_{k\geqslant 1}$ and $\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = 1/2$. Then its running maximum $M_n = \max(S_1, \dots, S_n)$, $n \geqslant 1$, satisfies*

$$\mathbb{P}(M_n \geqslant a) = 2\,\mathbb{P}(S_n > a) + \mathbb{P}(S_n = a), \quad a \in \mathbb{N}\,.$$

7.6 Remark. It is remarkable that the law of the maximum M_n can be found so simply and that the maximum is not so much larger. Note that by symmetry we obtain $\mathbb{P}(M_n \geqslant a) = \mathbb{P}(|S_n| > a) + \mathbb{P}(S_n = a)$ as well.

The intuition is that after the first time τ_a where $(S_m)_{m\geqslant 0}$ reaches the level a the random walk $S_m = a + (S_m - a)$ between τ_a and n and its reflection $a - (S_m - a)$ have the same law. From this we derive $\mathbb{P}(S_n > a, \tau_a < n) = \mathbb{P}(S_n < a, \tau_a < n) = \frac{1}{2}(\mathbb{P}(\tau_a < n) - \mathbb{P}(S_n = a, \tau_a < n))$. Adding $\pm\,\mathbb{P}(\tau_a = n)$ on the right-hand side, $\mathbb{P}(S_n > a) = \mathbb{P}(S_n > a, \tau_a < n) = \frac{1}{2}(\mathbb{P}(\tau_a \leqslant n) - \mathbb{P}(S_n = a))$ follows. The proof is just more detailed.

Proof. Consider the stopping time $\tau_a := \inf\{k \geqslant 1 \,|\, S_k = a\} \wedge (n+1)$ so that $\{M_n \geqslant a\} = \{\tau_a \leqslant n\}$ holds. Moreover, $S_n - S_i$ is independent of $(X_1, \dots, X_i)$ and hence of the event $\{\tau_a = i\}$ for $1 \leqslant i < n$. By symmetry we have $\mathbb{P}(S_k >$

$0) = \mathbb{P}(S_k < 0) = \frac{1}{2}(1 - \mathbb{P}(S_k = 0))$ for all $k \in \mathbb{N}$. With $S_0 := 0$ we obtain the result:

$$\begin{aligned}
\mathbb{P}(S_n > a) = \mathbb{P}(S_n > a, \tau_a \leqslant n) &= \sum_{i=1}^n \mathbb{P}(S_n > a, \tau_a = i) \\
&= \sum_{i=1}^n \mathbb{P}(S_n - S_i > 0, \tau_a = i) = \sum_{i=1}^n \mathbb{P}(S_n - S_i > 0)\,\mathbb{P}(\tau_a = i) \\
&= \sum_{i=1}^n \mathbb{P}(S_{n-i} > 0)\,\mathbb{P}(\tau_a = i) = \sum_{i=1}^n \tfrac{1}{2}\,\mathbb{P}(S_{n-i} \neq 0)\,\mathbb{P}(\tau_a = i) \\
&= \frac{1}{2}\sum_{i=1}^n \Big(1 - \mathbb{P}(S_{n-i} = 0)\Big)\,\mathbb{P}(\tau_a = i) \\
&= \frac{1}{2}\Big(\,\mathbb{P}(\tau_a \leqslant n) - \sum_{i=1}^n \mathbb{P}(S_n = a, \tau_a = i)\Big) \\
&= \tfrac{1}{2}\,\mathbb{P}(M_n \geqslant a) - \tfrac{1}{2}\,\mathbb{P}(S_n = a, \tau_a \leqslant n) = \tfrac{1}{2}\,\mathbb{P}(M_n \geqslant a) - \tfrac{1}{2}\,\mathbb{P}(S_n = a).
\end{aligned}$$

More abstractly, we can view (S_n) as a Markov chain and use the strong Markov property for an alternative proof.

Let $F(t, (s_m)_{m\geqslant 0}) := \mathbf{1}(t \leqslant n)(\mathbf{1}(s_{n-t} > a) + \frac{1}{2}\mathbf{1}(s_{n-t} = a))$ and note

$$F(\tau_a, (S_{\tau_a+m})_{m\geqslant 0}) = \mathbf{1}(S_n > a) + \tfrac{1}{2}\mathbf{1}(S_n = a)$$

due to $S_n \geqslant a \Rightarrow \tau_a \leqslant n$. Moreover, $\mathbb{E}_a[F(t, (S_m)_{m\geqslant 0})] = 1/2$ holds for $t \leqslant n$ by the above symmetry argument. With $\mathbb{E}_x$ denoting the expectation when starting in x, we use that τ_a is $\mathscr{F}_{\tau_a}$-measurable and the strong Markov property to arrive at

$$\begin{aligned}
\mathbb{P}(S_n > a) + \tfrac{1}{2}\,\mathbb{P}(S_n = a) &= \mathbb{E}_0[\mathbb{E}[F(\tau_a, (S_{\tau_a+m})_{m\geqslant 0}) \,|\, \mathscr{F}_{\tau_a}]] \\
&= \mathbb{E}_0[\mathbb{E}[F(t, (S_m)_{m\geqslant 0} \circ \vartheta_{\tau_a}) \,|\, \mathscr{F}_{\tau_a}]|_{t=\tau_a}] \\
&= \mathbb{E}_0[\mathbb{E}_a[F(t, (S_m)_{m\geqslant 0})]|_{t=\tau_a}] \\
&= \mathbb{E}_0[\tfrac{1}{2}\mathbf{1}(\tau_a \leqslant n)] = \tfrac{1}{2}\,\mathbb{P}(M_n \geqslant a).
\end{aligned}$$

This is again the result. □

7.7 Remark. The preceding result extends to all symmetric random walks. Can we generalise this result to other random walks, at least approximately? The invariance principle will allow to obtain the same formula asymptotically for $\mathbb{P}(M_n \geqslant \sqrt{n}a)$. Moreover, it yields the law of the running maximum of Brownian motion.

7.8 Proposition (Reflection principle). *Let $(X_k)_{k\geqslant 1}$ be a sequence of i.i.d. random variables in L^2 with $\mathbb{E}[X_k] = 0$, $\mathbb{E}[X_k^2] = 1$. Set $S_n := \sum_{k=1}^n X_k$, $\tilde{M}_n := \frac{1}{\sqrt{n}} \max_{1\leqslant i\leqslant n} S_i$. Then $\tilde{M}_n \xrightarrow{d} |B_1|$ holds with $B_1 \sim N(0,1)$. For the Brownian motion B we have* $\max_{0\leqslant t\leqslant 1} B_t \overset{d}{=} |B_1|$.

Proof. With the linear interpolations Y_n from Theorem 7.4 $\tilde{M}_n = \max_{0\leqslant t\leqslant 1} Y_n(t)$ holds (the maximum is attained at some $t = i/n$ and not at the interpolated values). Since $f \mapsto \max_{0\leqslant t\leqslant 1} f(t)$ is continuous on $C([0,1])$ (or $C(\mathbb{R}^+)$), the continuous mapping device yields

$$\tilde{M}_n \xrightarrow{d} \max_{0\leqslant t\leqslant 1} B(t)$$

for a Brownian motion B. The latter is independent of the random walk specification (that is the invariance principle) and we may in particular consider the simple symmetric random walk.

Proposition 7.5 and Remark 7.6 state for $\sqrt{n}x \in \mathbb{N}$

$$\mathbb{P}(\tilde{M}_n \geqslant x) = \mathbb{P}(|Y_n(1)| > x) + \mathbb{P}(Y_n(1) = x).$$

This extends to all real $x > 0$ because $\tilde{M}_n$ and $Y_n(1)$ only take values in $n^{-1/2}\,\mathbb{Z} = \{n^{-1/2}k \,|\, k \in \mathbb{Z}\}$ and the probabilities do not change for $x \in ((k-1)n^{-1/2}, kn^{-1/2}]$. Again by continuous mapping and Donsker's theorem, $|Y_n(1)| \xrightarrow{d} |B(1)|$ follows. This implies (recall $B(1) \sim N(0,1)$)

$$\mathbb{P}(|Y_n(1)| > x) \to \mathbb{P}(|B(1)| > x), \quad \mathbb{P}(Y_n(1) = x) \to \mathbb{P}(B(1) = x) = 0.$$

So, $\mathbb{P}(\tilde{M}_n \geqslant x) \to \mathbb{P}(|B(1)| \geqslant x)$ follows for all $x > 0$. We conclude $\tilde{M}_n \xrightarrow{d} |B(1)|$ and therefore $|B(1)| \overset{d}{=} \max_{0\leqslant t\leqslant 1} B(t)$. □

7.2 Empirical process and Brownian bridge

7.9 Definition. For i.i.d. real-valued random variables $X_1, \dots, X_n$ with distribution function F the (random) function

$$F_n(x) := \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}(X_k \leqslant x), \quad x \in \mathbb{R},$$

is called empirical distribution function (empirische Verteilungsfunktion). The associated empirical process (empirischer Prozess) is given by

$$G_n(x) := \sqrt{n}(F_n(x) - F(x)), \quad x \in \mathbb{R}\,.$$

7.10 Lemma. *For $x_1, \dots, x_m \in \mathbb{R}$ and $n \to \infty$ we have*

$$(G_n(x_1), \dots, G_n(x_m)) \xrightarrow{d} N(0,\Sigma) \text{ with } \Sigma = \big(F(x_i \wedge x_j) - F(x_i)F(x_j)\big)_{i,j=1,\dots,m}.$$

Proof. We have $(G_n(x_1), \dots, G_n(x_m)) = \frac{1}{\sqrt{n}}\sum_{k=1}^n Y_k$ with i.i.d. random vectors $Y_k = (\mathbf{1}(X_k \leqslant x_i) - F(x_i))_{i=1,\dots,m}$. We obtain $\mathbb{E}[Y_k] = 0$ and

$$\begin{aligned}\mathbb{E}[(Y_k)_i(Y_k)_j] &= \mathbb{E}[\mathbf{1}(X_k \leqslant x_i)\mathbf{1}(X_k \leqslant x_j)] - F(x_i)F(x_j)\\ &= F(x_i \wedge x_j) - F(x_i)F(x_j).\end{aligned}$$

Consequently, Y_k has covariance matrix Σ and the standard multivariate central limit theorem (compare Stochastik I) yields the claim. □

7.11 Definition. The Brownian bridge $(B_t^0, t \in [0,1])$ is the centred and continuous Gaussian process with $\operatorname{Cov}(B_s^0, B_t^0) = s \wedge t - st$ for $s, t \in [0,1]$.

7.12 Remark. The lemma shows that the finite-dimensional distributions of the empirical process $(G_n(x), 0 \leqslant x \leqslant 1)$ in the case $X_k \sim U([0,1])$ i.i.d. and $F(x) = x$ converge to those of the Brownian bridge B^0. The empirical process, however, is not continuous and we shall use a continuous interpolation $\tilde{G}_n$ to prove even functional convergence in $C([0,1])$.

A continuous Brownian bridge process can be constructed as $B_t^0 = B_t - tB_1$, $t \in [0,1]$, with a Brownian motion B. To see this, it suffices to check the covariances using $\operatorname{Cov}(B_s, B_t) = s \wedge t$. One can show that a Brownian bridge has the law of a Brownian motion on $[0,1]$ conditional on the event $\{B_1 = 0\}$.

7.13 Theorem (Donsker Theorem for empirical processes)**.** *For independent $U([0,1])$-distributed random variables $X_1, \dots, X_n$ consider the linear interpolation $\tilde{F}_n : [0,1] \to [0,1]$ of the empirical distribution function F_n satisfying $\tilde{F}_n(X_k) = F_n(X_k)$, $k = 1, \dots, n$, $\tilde{F}_n(0) = 0$, $\tilde{F}_n(1) = 1$.*

For the interpolated empirical process $\tilde{G}_n = \sqrt{n}(\tilde{F}_n - F)$ we have convergence to a Brownian bridge B^0 in distribution on $C([0,1])$: $\tilde{G}_n \xrightarrow{d} B^0$.

7.14 Corollary (Kolmogorov-Smirnov)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables with continuous distribution function F. Then their empirical distribution function F_n satisfies with a Brownian bridge B^0*

$$\sup_{x \in \mathbb{R}} \sqrt{n} |F_n(x) - F(x)| \xrightarrow{d} \max_{0 \leqslant t \leqslant 1} |B_t^0|.$$

Proof. Since F is continuous, $U_k = F(X_k)$, $k = 1, \dots, n$, are independent $U([0,1])$-distributed random variables and thus a.s.

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}(U_k \leqslant F(x)) = F_n^U(F(x)), \text{where } F_n^U(t) := \frac{1}{n} \sum_{k=1}^n \mathbf{1}(U_k \leqslant t).$$

Even more, the maximum of the modulus of the empirical process is the same for (X_k) and (U_k):

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = \sup_{x \in \mathbb{R}} |F_n^U(F(x)) - F(x)| = \sup_{t \in [0,1]} |F_n^U(t) - t|.$$

Using $\|\tilde{G}_n^U - G_n^U\|_\infty \leqslant n^{-1/2}$ for the interpolated and original empirical process of the (U_k), we thus arrive at

$$\sup_{x \in \mathbb{R}} \sqrt{n} |F_n(x) - F(x)| = \sup_{t \in [0,1]} |G_n^U(t)| = \sup_{t \in [0,1]} |\tilde{G}_n^U(t)| + O(n^{-1/2}).$$

By Theorem 7.13, continuous mapping and Slutsky's Lemma we conclude

$$\sup_{x \in \mathbb{R}} \sqrt{n} |F_n(x) - F(x)| \xrightarrow{d} \sup_{t \in [0,1]} |B_t^0|.$$

Since B^0 is continuous, the supremum is actually a maximum, as asserted. □

7.15 Remark. This is the basis of the Kolmogorov-Smirnov goodness-of-fit test where we test the null hypothesis that the cumulative distribution function is given by F, using the test

$$\varphi_n := \mathbf{1}\Big(\sup_{x\in\mathbb{R}} \sqrt{n}|F_n(x) - F(x)| > \kappa_\alpha\Big)$$

for some critical value $\kappa_\alpha > 0$ depending on the desired level α of the test. Usually, κ_α is chosen as a quantile of the law of the limit $\max_{0\leqslant t\leqslant 1}|B_t^0|$. The maximum of a Brownian bridge can be shown to be described by the *Kolmogorov distribution* and κ_α is calculated from

$$\alpha = \mathbb{P}\Big(\max_{0\leqslant t\leqslant 1}|B_t^0| > \kappa_\alpha\Big) = 1 - \sum_{j\in\mathbb{Z}}(-1)^j e^{-2j^2\kappa_\alpha^2}.$$

Proof of Donsker's Theorem for empirical processes. Lemma 7.10 together with Slutsky's Lemma due to $\|\tilde{G}_n - G_n\|_\infty = n^{-1/2}$ (the jumps of G_n have size $n^{-1/2}$) show that the finite-dimensional distributions of $\tilde{G}_n$ converge to those of a Brownian bridge. We shall establish tightness via the Kolmogorov-Centsov criterion by showing for some constant $K > 0$

$$\mathbb{E}[(\tilde{G}_n(t) - \tilde{G}_n(s))^4] \leqslant K(t-s)^2, \quad s,t \in [0,1]. \tag{7.1}$$

By Theorem 6.17 this yields the convergence in law in $C([0,1])$. In the sequel $C_1, C_2, \dots$ denote some numerical constants.

First case: $t - s \geqslant n^{-1}$. Then $\|\tilde{G}_n - G_n\|_\infty = n^{-1/2}$ implies

$$\begin{aligned}\mathbb{E}[(\tilde{G}_n(t) - \tilde{G}_n(s))^4] &\leqslant C_1\Big(\mathbb{E}[(G_n(s) - G_n(t))^4] + n^{-2}\Big)\\ &\leqslant C_1\Big(\mathbb{E}[(G_n(s) - G_n(t))^4] + (t-s)^2\Big).\end{aligned}$$

We can write $G_n(t) - G_n(s) = n^{-1/2}\sum_{k=1}^n (Z_k - \mathbb{E}[Z_k])$ with independent $Z_k = \mathbf{1}(s < X_k \leqslant t) \sim \mathrm{Bin}(1, t-s)$. By expanding the fourth power of the sum and noting that the expectation of $Z_k - \mathbb{E}[Z_k]$ vanishes, we obtain

$$\begin{aligned}\mathbb{E}[(G_n(t) - G_n(s))^4] &= \frac{n}{n^2}\,\mathbb{E}[(Z_1 - \mathbb{E}[Z_1])^4] + \frac{n^2-n}{n^2}\,\mathbb{E}[(Z_1 - \mathbb{E}[Z_1])^2]^2\\ &\leqslant n^{-1}(t-s) + (t-s)^2 \leqslant 2(t-s)^2.\end{aligned}$$

This yields (7.1) in this first case.

Second case: $0 \leqslant t - s \leqslant n^{-1}$. For $h = (t-s)^{1/2}$ introduce the events

$$A_k = \{F_n(t + h/2) - F_n(s - h/2) = k/n\}, \quad k \in \mathbb{N}_0\,.$$

On A_0 there is no X_k in $(s - h/2, t + h/2]$ and $\tilde{F}_n$ is linear there with maximal slope $n^{-1}(t - s + h)^{-1}$ so that

$$\tilde{F}_n(t) - \tilde{F}_n(s) \leqslant n^{-1}(t-s+h)^{-1}(t-s) \text{ on } A_0.$$

On A_1 there is exactly one sample point X_k in $(s-h/2, t+h/2]$ and $\tilde{F}_n(t) - \tilde{F}_n(s)$ is maximal when this X_k is in the center $(s+t)/2$ of the interval. This shows

$$\tilde{F}_n(t) - \tilde{F}_n(s) \leqslant 2n^{-1}(t-s+h)^{-1}(t-s) \text{ on } A_1.$$

Generally we obtain

$$\tilde{F}_n(t) - \tilde{F}_n(s) \leqslant \frac{(k+1)(t-s)}{n(t-s+h)} \text{ on } A_k.$$

With $Z := \#\{X_k \in (s-h/2, t+h/2] \,|\, k = 1, \dots, n\} \sim \mathrm{Bin}(n, t-s+h)$ we thus find

$$\mathbb{E}[(\tilde{F}_n(t) - \tilde{F}_n(s))^4] \leqslant \mathbb{E}[(Z+1)^4] \frac{(t-s)^4}{n^4(t-s+h)^4}.$$

Noting $\mathbb{E}[Z(Z-1)\cdots(Z-m+1)] = (n(n-1)\cdots(n-m+1))(t-s+h)^m$, $m \in \mathbb{N}$, we have

$$\mathbb{E}[(Z+1)^4] \leqslant C_2(\mathbb{E}[Z^4]+1) \leqslant C_3 \sum_{m=0}^{4} n^m (t-s+h)^m \leqslant C_4\big(n^4(t-s+h)^4 + 1\big).$$

Inserting also $h = (t-s)^{1/2}$, we arrive at

$$\mathbb{E}[(\tilde{F}_n(t) - \tilde{F}_n(s))^4] \leqslant C_5\Big((t-s)^4 + \frac{(t-s)^2}{n^4}\Big).$$

Due to $F(t) - F(s) = t - s$ and $n^2 \leqslant (t-s)^{-2}$ this yields for the empirical process

$$\mathbb{E}[(\tilde{G}_n(t) - \tilde{G}_n(s))^4] \leqslant C_6\Big(n^2(t-s)^4 + n^{-2}(t-s)^2\Big) \leqslant 2C_6(t-s)^2$$

and establishes (7.1) also in the second case. □

Markus Reiß

Course *Stochastic Processes*
Winter 2023/24
Humboldt-Universität zu Berlin

List of exam-related questions

(a) Formulate and give main steps in the proof: Chapman-Kolmogorov equation, Ulam's Theorem, Kolmogorov's consistency theorem, Factorisation Lemma, existence and properties of conditional expectations, Doob decomposition and quadratic variation, optional stopping and optional sampling theorems, Wald identity, martingale inequalities, 1st martingale convergence theorem, Vitali's Theorem, 2nd martingale convergence theorem, strong law for L^2-martingales, backward martingale convergence theorem, Radon-Nikodym theorem, Lebesgue decomposition, Kakutani's Theorem, strong Markov property, recurrence and transience of Markov chains, condition for ergodicity of Markov chains, Birkhoff's ergodic theorem, properties of Markov transition operator, convergence theorem for aperiodic Markov chains, Continuous Mapping Theorem, Portmanteau Theorem, Slutsky Lemma, Prokhorov Theorem, Kolmogorov-Centsov criterion for weak convergence in $C([0,T])$, Donsker Theorems.

(b) Which ways exist to construct a Poisson process? What is the Markov property? Why can a stochastic process be considered as an $(S^T, \mathscr{S}^{\otimes T})$-valued random variable? Why is the law of a continuous process determined by its finite dimensional distributions? What are different notions of equality for stochastic processes? How is the conditional expectation in L^2 constructed? What is the meaning of $\mathbb{E}[Y \mid X = x]$? What is $(X{\bullet}M)_n$ for X predictable, M martingale and what are its properties? When does $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$ hold for M martingale, τ stopping time? What are sufficient conditions for uniform integrability? Does uniform integrability imply tightness of the laws? How many ways do you know to prove the classical strong law of large numbers? What are equivalent characterisations for T being ergodic? What do irreducible, recurrent, transient, aperiodic, reversible, invariant mean for a Markov chain, its states or its initial distribution? How do we obtain a Metropolis Markov chain? What is the relationship between tightness and compactness? What is the implication of the Portmanteau Lemma for distribution functions of real-valued random variables? What is the relationship between $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ for laws on $C([0,T])$ and convergence of the finite-dimensional distributions? How can we prove existence of Brownian motion?

(c) Give examples and counter-examples for: Polish spaces, (sub-/super-) martingales, predictable processes, stopping times, uniformly integrable random variables, absolutely continuous and singular measures, stationary and ergodic processes, irreducible and aperiodic Markov chains, recurrent, transient states, invariant initial distributions, weak convergence, tight laws.

(d) Review all major examples of the lecture like Poisson process, Ehrenfest model, simple random walk, different martingales, Gaussian processes, etc.

(e) For each result in point (a) find examples and possibly counter-examples where assertions do not hold. Where do the conditions in the theorems enter in the proof?

(f) Consider again in detail the exercise problems!