Chapter 5
Moments and Laws of Large Numbers

The most important characteristic quantities of random variables are the median, expectation and variance. For large n, the expectation describes the typical approximate value of the arithmetic mean $(X_1 + \ldots + X_n)/n$ of i.i.d. random variables (law of large numbers). In Chapter 15, we will see how the variance determines the size of the typical deviations of the arithmetic mean from the expectation.

5.1 Moments

In the following, let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space.

Definition 5.1 Let X be a real-valued random variable.

(i) If $X \in \mathcal{L}^1(\mathbf{P})$, then X is called *integrable* and we call

$$\mathbf{E}[X] := \int X \, d\mathbf{P}$$

the *expectation* or *mean* of X. If $\mathbf{E}[X] = 0$, then X is called *centered*. More generally, we also write $\mathbf{E}[X] = \int X \, d\mathbf{P}$ if only X^- or X^+ is integrable.

(ii) If $n \in \mathbb{N}$ and $X \in \mathcal{L}^n(\mathbf{P})$, then the quantities

$$m_k := \mathbf{E}\big[X^k\big], \qquad M_k := \mathbf{E}\big[|X|^k\big] \quad \text{for any } k = 1, \ldots, n,$$

are called the kth *moments* and kth *absolute moments*, respectively, of X.

(iii) If $X \in \mathcal{L}^2(\mathbf{P})$, then X is called *square integrable* and

$$\mathbf{Var}[X] := \mathbf{E}\big[X^2\big] - \mathbf{E}[X]^2$$

is the *variance* of X. The number $\sigma := \sqrt{\mathbf{Var}[X]}$ is called the *standard deviation* of X. Formally, we sometimes write $\mathbf{Var}[X] = \infty$ if $\mathbf{E}[X^2] = \infty$.

A. Klenke, *Probability Theory*, Universitext, DOI 10.1007/978-1-4471-5361-0_5,

(iv) If $X, Y \in \mathcal{L}^2(\mathbf{P})$, then we define the *covariance* of X and Y by

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

X and Y are called *uncorrelated* if $\mathbf{Cov}[X, Y] = 0$ and *correlated* otherwise.

Remark 5.2

(i) The definition in (ii) is sensible since, by virtue of Theorem 4.19, $X \in \mathcal{L}^n(\mathbf{P})$ implies that $M_k < \infty$ for all $k = 1, \ldots, n$.
(ii) If $X, Y \in \mathcal{L}^2(\mathbf{P})$, then $XY \in \mathcal{L}^1(\mathbf{P})$ since $|XY| \le X^2 + Y^2$. Hence the definition in (iv) makes sense and we have

$$\mathbf{Cov}[X, Y] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

In particular, $\mathbf{Var}[X] = \mathbf{Cov}[X, X]$. ◊

We collect the most important rules of expectations in a theorem. All of these properties are direct consequences of the corresponding properties of the integral.

Theorem 5.3 (Rules for expectations) *Let $X, Y, X_n, Z_n, n \in \mathbb{N}$, be real integrable random variables on $(\Omega, \mathcal{A}, \mathbf{P})$.*

(i) *If $\mathbf{P}_X = \mathbf{P}_Y$, then $\mathbf{E}[X] = \mathbf{E}[Y]$.*
(ii) *(Linearity) Let $c \in \mathbb{R}$. Then $cX \in \mathcal{L}^1(\mathbf{P})$ and $X + Y \in \mathcal{L}^1(\mathbf{P})$ as well as*

$$\mathbf{E}[cX] = c\mathbf{E}[X] \quad \text{and} \quad \mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y].$$

(iii) *If $X \ge 0$ almost surely, then*

$$\mathbf{E}[X] = 0 \iff X = 0 \quad \text{almost surely}.$$

(iv) *(Monotonicity) If $X \le Y$ almost surely, then $\mathbf{E}[X] \le \mathbf{E}[Y]$ with equality if and only if $X = Y$ almost surely.*
(v) *(Triangle inequality) $|\mathbf{E}[X]| \le \mathbf{E}[|X|]$.*
(vi) *If $X_n \ge 0$ almost surely for all $n \in \mathbb{N}$, then $\mathbf{E}[\sum_{n=1}^{\infty} X_n] = \sum_{n=1}^{\infty} \mathbf{E}[X_n]$.*
(vii) *If $Z_n \uparrow Z$ for some Z, then $\mathbf{E}[Z] = \lim_{n\to\infty} \mathbf{E}[Z_n] \in (-\infty, \infty]$.*

Again probability theory comes into play when independence enters the stage; that is, when we exit the realm of linear integration theory.

Theorem 5.4 (Independent random variables are uncorrelated) *Let $X, Y \in \mathcal{L}^1(\mathbf{P})$ be independent. Then $(XY) \in \mathcal{L}^1(\mathbf{P})$ and $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$. In particular, independent random variables are uncorrelated.*

Proof Assume first that X and Y take only finitely many values. Then XY also takes only finitely many values and thus $XY \in \mathcal{L}^1(\mathbf{P})$. It follows that

$$\begin{aligned}
\mathbf{E}[XY] &= \sum_{z\in\mathbb{R}\setminus\{0\}} z\mathbf{P}[XY = z] \\
&= \sum_{z\in\mathbb{R}\setminus\{0\}} \sum_{x\in\mathbb{R}\setminus\{0\}} x\frac{z}{x}\mathbf{P}[X = x, Y = z/x] \\
&= \sum_{y\in\mathbb{R}\setminus\{0\}} \sum_{x\in\mathbb{R}\setminus\{0\}} xy\mathbf{P}[X = x]\mathbf{P}[Y = y] \\
&= \Big(\sum_{x\in\mathbb{R}} x\mathbf{P}[X = x]\Big)\Big(\sum_{y\in\mathbb{R}} y\mathbf{P}[Y = y]\Big) \\
&= \mathbf{E}[X]\mathbf{E}[Y].
\end{aligned}$$

For $N \in \mathbb{N}$, the random variables $X_N := (2^{-N}\lfloor 2^N|X|\rfloor) \wedge N$ and $Y_N := (2^{-N}\lfloor 2^N|Y|\rfloor) \wedge N$ take only finitely many values and are independent as well. Furthermore, $X_N \uparrow |X|$ and $Y_N \uparrow |Y|$. By the monotone convergence theorem (Theorem 4.20), we infer

$$\begin{aligned}
\mathbf{E}\big[|XY|\big] &= \lim_{N\to\infty} \mathbf{E}[X_N Y_N] = \lim_{N\to\infty} \mathbf{E}[X_N]\mathbf{E}[Y_N] \\
&= \Big(\lim_{N\to\infty} \mathbf{E}[X_N]\Big)\Big(\lim_{N\to\infty} \mathbf{E}[Y_N]\Big) = \mathbf{E}\big[|X|\big]\mathbf{E}\big[|Y|\big] < \infty.
\end{aligned}$$

Hence $XY \in \mathcal{L}^1(\mathbf{P})$. Furthermore, we have shown the claim in the case where X and Y are nonnegative. Hence (and since each of the families $\{X^+, Y^+\}$, $\{X^-, Y^+\}$, $\{X^+, Y^-\}$ and $\{X^-, Y^-\}$ is independent) we obtain

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

□

Theorem 5.5 (Wald's identity) *Let $T, X_1, X_2, \ldots$ be independent real random variables in $\mathcal{L}^1(\mathbf{P})$. Let $\mathbf{P}[T \in \mathbb{N}_0] = 1$ and assume that $X_1, X_2, \ldots$ are identically distributed. Define*

$$S_T := \sum_{i=1}^{T} X_i.$$

Then $S_T \in \mathcal{L}^1(\mathbf{P})$ and $\mathbf{E}[S_T] = \mathbf{E}[T]\mathbf{E}[X_1]$.

Proof Define $S_n = \sum_{i=1}^n X_i$ for $n \in \mathbb{N}_0$. Then $S_T = \sum_{n=1}^\infty S_n \mathbb{1}_{\{T=n\}}$. By Remark 2.15, the random variables S_n and $\mathbb{1}_{\{T=n\}}$ are independent for any $n \in \mathbb{N}$ and thus uncorrelated. This implies (using the triangle inequality; see Theorem 5.3(v))

$$\begin{aligned}\mathbf{E}\big[|S_T|\big] &= \sum_{n=1}^\infty \mathbf{E}\big[|S_n|\mathbb{1}_{\{T=n\}}\big] = \sum_{n=1}^\infty \mathbf{E}\big[|S_n|\big]\mathbf{E}[\mathbb{1}_{\{T=n\}}] \\ &\le \sum_{n=1}^\infty \mathbf{E}\big[|X_1|\big] n\mathbf{P}[T=n] = \mathbf{E}\big[|X_1|\big]\mathbf{E}[T].\end{aligned}$$

The same computation without absolute values yields the remaining part of the claim. □

We collect some basic properties of the variance.

Theorem 5.6 *Let $X \in \mathcal{L}^2(\mathbf{P})$. Then*:

(i) $\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] \ge 0$.
(ii) $\mathbf{Var}[X] = 0 \iff X = \mathbf{E}[X]$ *almost surely.*
(iii) *The map* $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto \mathbf{E}[(X - x)^2]$ *is minimal at* $x_0 = \mathbf{E}[X]$ *with* $f(\mathbf{E}[X]) = \mathbf{Var}[X]$.

Proof (i) This is a direct consequence of Remark 5.2(ii).

(ii) By Theorem 5.3(iii), we have $\mathbf{E}[(X - \mathbf{E}[X])^2] = 0 \iff (X - \mathbf{E}[X])^2 = 0$ a.s.

(iii) Clearly, $f(x) = \mathbf{E}[X^2] - 2x\mathbf{E}[X] + x^2 = \mathbf{Var}[X] + (x - \mathbf{E}[X])^2$. □

Theorem 5.7 *The map* $\mathbf{Cov} : \mathcal{L}^2(\mathbf{P}) \times \mathcal{L}^2(\mathbf{P}) \to \mathbb{R}$ *is a positive semidefinite symmetric bilinear form and* $\mathbf{Cov}[X, Y] = 0$ *if Y is almost surely constant. The detailed version of this concise statement is*: *Let* $X_1, \dots, X_m, Y_1, \dots, Y_n \in \mathcal{L}^2(\mathbf{P})$ *and* $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_n \in \mathbb{R}$ *as well as* $d, e \in \mathbb{R}$. *Then*

$$\mathbf{Cov}\left[d + \sum_{i=1}^m \alpha_i X_i, e + \sum_{j=1}^n \beta_j Y_j\right] = \sum_{i,j} \alpha_i \beta_j \, \mathbf{Cov}[X_i, Y_j]. \tag{5.1}$$

In particular, $\mathbf{Var}[\alpha X] = \alpha^2 \mathbf{Var}[X]$ *and the* Bienaymé formula *holds*,

$$\mathbf{Var}\left[\sum_{i=1}^m X_i\right] = \sum_{i=1}^m \mathbf{Var}[X_i] + \sum_{\substack{i,j=1 \\ i \ne j}}^m \mathbf{Cov}[X_i, X_j]. \tag{5.2}$$

For uncorrelated $X_1, \dots, X_m$, *we have* $\mathbf{Var}[\sum_{i=1}^m X_i] = \sum_{i=1}^m \mathbf{Var}[X_i]$.

Proof

$$\begin{aligned}
&\mathbf{Cov}\left[d+\sum_{i=1}^{m}\alpha_i X_i,\ e+\sum_{j=1}^{n}\beta_j Y_j\right]\\
&\quad=\mathbf{E}\left[\left(\sum_{i=1}^{m}\alpha_i\big(X_i-\mathbf{E}[X_i]\big)\right)\left(\sum_{j=1}^{n}\beta_j\big(Y_j-\mathbf{E}[Y_j]\big)\right)\right]\\
&\quad=\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_i\beta_j\mathbf{E}\big[\big(X_i-\mathbf{E}[X_i]\big)\big(Y_j-\mathbf{E}[Y_j]\big)\big]\\
&\quad=\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_i\beta_j\mathbf{Cov}[X_i,Y_j].
\end{aligned}$$

□

Theorem 5.8 (Cauchy–Schwarz inequality) *If $X, Y \in \mathcal{L}^2(\mathbf{P})$, then*

$$\big(\mathbf{Cov}[X,Y]\big)^2 \le \mathbf{Var}[X]\mathbf{Var}[Y].$$

Equality holds if and only if there are $a, b, c \in \mathbb{R}$ with $|a|+|b|+|c|>0$ and such that $aX+bY+c=0$ a.s.

Proof The Cauchy–Schwarz inequality holds for any positive semidefinite bilinear form and hence in particular for the covariance map. Using the notation of variance and covariance, a simple proof looks like this:

Case 1: $\mathbf{Var}[Y]=0$. Here the statement is trivial (choose $a=0$, $b=1$ and $c=-\mathbf{E}[Y]$).
Case 2: $\mathbf{Var}[Y]>0$. Let $\theta := -\frac{\mathbf{Cov}[X,Y]}{\mathbf{Var}[Y]}$. Then, by Theorem 5.6(i),

$$\begin{aligned}
0 &\le \mathbf{Var}[X+\theta Y]\mathbf{Var}[Y]\\
&= \big(\mathbf{Var}[X]+2\theta\mathbf{Cov}[X,Y]+\theta^2\mathbf{Var}[Y]\big)\mathbf{Var}[Y]\\
&= \mathbf{Var}[X]\mathbf{Var}[Y]-\mathbf{Cov}[X,Y]^2
\end{aligned}$$

with equality if and only if $X+\theta Y$ is a.s. constant. Now let $a=1$, $b=\theta$ and $c=-\mathbf{E}[X]-b\mathbf{E}[Y]$.

□

Example 5.9

(i) Let $p \in [0,1]$ and $X \sim \mathrm{Ber}_p$. Then

$$\mathbf{E}\big[X^2\big]=\mathbf{E}[X]=\mathbf{P}[X=1]=p$$

and thus $\mathbf{Var}[X]=p(1-p)$.

(ii) Let $n \in \mathbb{N}$ and $p \in [0,1]$. Let X be binomially distributed, $X \sim b_{n,p}$. Then

$$\begin{aligned}\mathbf{E}[X] &= \sum_{k=0}^{n} k\mathbf{P}[X=k] = \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k} \\ &= np \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1}(1-p)^{(n-1)-(k-1)} = np.\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbf{E}\big[X(X-1)\big] &= \sum_{k=0}^{n} k(k-1)\mathbf{P}[X=k] \\ &= \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^k (1-p)^{n-k} \\ &= np \cdot \sum_{k=1}^{n} (k-1)\binom{n-1}{k-1} p^{k-1}(1-p)^{(n-1)-(k-1)} \\ &= n(n-1)p^2 \cdot \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2}(1-p)^{(n-2)-(k-2)} \\ &= n(n-1)p^2.\end{aligned}$$

Hence $\mathbf{E}[X^2] = \mathbf{E}[X(X-1)] + \mathbf{E}[X] = n^2p^2 + np(1-p)$ and thus $\mathbf{Var}[X] = np(1-p)$.

The statement can be derived more simply than by direct computation if we make use of the fact that $b_{n,p} = b_{1,p}^{*n}$ (see Example 3.4(ii)). That is (see Theorem 2.31), $\mathbf{P}_X = \mathbf{P}_{Y_1+\ldots+Y_n}$, where $Y_1, \ldots, Y_n$ are independent and $Y_i \sim \mathrm{Ber}_p$ for any $i = 1, \ldots, n$. Hence

$$\begin{aligned}\mathbf{E}[X] &= n\mathbf{E}[Y_1] = np, \\ \mathbf{Var}[X] &= n\mathbf{Var}[Y_1] = np(1-p).\end{aligned} \tag{5.3}$$

(iii) Let $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, and let X be normally distributed, $X \sim \mathcal{N}_{\mu,\sigma^2}$. Then

$$\begin{aligned}\mathbf{E}[X] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x e^{-(x-\mu)^2/(2\sigma^2)}\, dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x+\mu) e^{-x^2/(2\sigma^2)}\, dx \\ &= \mu + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x e^{-x^2/(2\sigma^2)}\, dx = \mu.\end{aligned} \tag{5.4}$$

Similarly, we get $\mathbf{Var}[X] = \mathbf{E}[X^2] - \mu^2 = \ldots = \sigma^2$.

(iv) Let $\theta > 0$ and let X be exponentially distributed, $X \sim \exp_\theta$. Then

$$\mathbf{E}[X] = \theta \int_0^\infty x e^{-\theta x}\, dx = \frac{1}{\theta},$$

$$\mathbf{Var}[X] = -\theta^{-2} + \theta \int_0^\infty x^2 e^{-\theta x}\, dx = \theta^{-2}\left(-1 + \int_0^\infty x^2 e^{-x}\, dx\right) = \theta^{-2}.$$

◊

Theorem 5.10 (Blackwell–Girshick) *Let $T, X_1, X_2, \ldots$ be independent real random variables in $\mathcal{L}^2(\mathbf{P})$. Let $\mathbf{P}[T \in \mathbb{N}_0] = 1$ and let $X_1, X_2, \ldots$ be identically distributed. Define*

$$S_T := \sum_{i=1}^{T} X_i.$$

Then $S_T \in \mathcal{L}^2(\mathbf{P})$ and

$$\mathbf{Var}[S_T] = \mathbf{E}[X_1]^2 \mathbf{Var}[T] + \mathbf{E}[T]\mathbf{Var}[X_1].$$

Proof Define $S_n = \sum_{i=1}^n X_i$ for $n \in \mathbb{N}$. Then (as in the proof of Wald's identity) S_n and $\mathbb{1}_{\{T=n\}}$ are independent; hence S_n^2 and $\mathbb{1}_{\{T=n\}}$ are uncorrelated and thus

$$\begin{aligned}
\mathbf{E}\big[S_T^2\big] &= \sum_{n=0}^{\infty} \mathbf{E}\big[\mathbb{1}_{\{T=n\}} S_n^2\big] \\
&= \sum_{n=0}^{\infty} \mathbf{E}[\mathbb{1}_{\{T=n\}}]\mathbf{E}\big[S_n^2\big] \\
&= \sum_{n=0}^{\infty} \mathbf{P}[T=n]\big(\mathbf{Var}[S_n] + \mathbf{E}[S_n]^2\big) \\
&= \sum_{n=0}^{\infty} \mathbf{P}[T=n]\big(n\mathbf{Var}[X_1] + n^2\mathbf{E}[X_1]^2\big) \\
&= \mathbf{E}[T]\mathbf{Var}[X_1] + \mathbf{E}\big[T^2\big]\mathbf{E}[X_1]^2.
\end{aligned}$$

By Wald's identity (Theorem 5.5), we have $\mathbf{E}[S_T] = \mathbf{E}[T]\mathbf{E}[X_1]$; hence

$$\mathbf{Var}[S_T] = \mathbf{E}\big[S_T^2\big] - \mathbf{E}[S_T]^2 = \mathbf{E}[T]\mathbf{Var}[X_1] + \big(\mathbf{E}\big[T^2\big] - \mathbf{E}[T]^2\big)\mathbf{E}[X_1]^2,$$

as claimed. □

Exercise 5.1.1 Let X be an integrable real random variable whose distribution $\mathbf{P}_X$ has a density f (with respect to the Lebesgue measure λ). Show (using Theo-

rem 4.15) that

$$\mathbf{E}[X] = \int_{\mathbb{R}} x f(x) \lambda(dx).$$

Exercise 5.1.2 Let $X \sim \beta_{r,s}$ be a Beta-distributed random variable with parameters $r, s > 0$ (see Example 1.107(ii)). Show that

$$\mathbf{E}\big[X^n\big] = \prod_{k=0}^{n-1} \frac{r+k}{r+s+k} \quad \text{for any } n \in \mathbb{N}.$$

Exercise 5.1.3 Let $X_1, X_2, \ldots$ be i.i.d. nonnegative random variables. By virtue of the Borel–Cantelli lemma, show that

$$\limsup_{n\to\infty} \frac{1}{n} X_n = \begin{cases} 0 \text{ a.s.}, & \text{if } \mathbf{E}[X_1] < \infty, \\ \infty \text{ a.s.}, & \text{if } \mathbf{E}[X_1] = \infty. \end{cases}$$

Exercise 5.1.4 Let $X_1, X_2, \ldots$ be i.i.d. nonnegative random variables. By virtue of the Borel–Cantelli lemma, show that for any $c \in (0, 1)$

$$\sum_{n=1}^{\infty} e^{X_n} c^n \begin{cases} < \infty \text{ a.s.}, & \text{if } \mathbf{E}[X_1] < \infty, \\ = \infty \text{ a.s.}, & \text{if } \mathbf{E}[X_1] = \infty. \end{cases}$$

5.2 Weak Law of Large Numbers

Theorem 5.11 (Markov inequality, Chebyshev inequality) *Let X be a real random variable and let $f : [0, \infty) \to [0, \infty)$ be monotone increasing. Then for any $\varepsilon > 0$ with $f(\varepsilon) > 0$, the* Markov inequality *holds*,

$$\mathbf{P}\big[|X| \geq \varepsilon\big] \leq \frac{\mathbf{E}[f(|X|)]}{f(\varepsilon)}.$$

In the special case $f(x) = x^2$, we get $\mathbf{P}[|X| \geq \varepsilon] \leq \varepsilon^{-2}\mathbf{E}[X^2]$. In particular, if $X \in \mathcal{L}^2(\mathbf{P})$, the Chebyshev inequality *holds*:

$$\mathbf{P}\big[\big|X - \mathbf{E}[X]\big| \geq \varepsilon\big] \leq \varepsilon^{-2}\mathbf{Var}[X].$$

Proof We have

$$\begin{aligned} \mathbf{E}\big[f\big(|X|\big)\big] &\geq \mathbf{E}\big[f\big(|X|\big)\mathbb{1}_{\{f(|X|)\geq f(\varepsilon)\}}\big] \\ &\geq \mathbf{E}\big[f(\varepsilon)\mathbb{1}_{\{f(|X|)\geq f(\varepsilon)\}}\big] \\ &\geq f(\varepsilon)\mathbf{P}\big[|X| \geq \varepsilon\big]. \end{aligned}$$

□

Definition 5.12 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real random variables in $\mathcal{L}^1(\mathbf{P})$ and let $\widetilde{S}_n = \sum_{i=1}^n (X_i - \mathbf{E}[X_i])$.

(i) We say that $(X_n)_{n\in\mathbb{N}}$ fulfills the *weak law of large numbers* if

$$\lim_{n\to\infty} \mathbf{P}\left[\left|\frac{1}{n}\widetilde{S}_n\right| > \varepsilon\right] = 0 \quad \text{for any } \varepsilon > 0.$$

(ii) We say that $(X_n)_{n\in\mathbb{N}}$ fulfills the *strong law of large numbers* if

$$\mathbf{P}\left[\limsup_{n\to\infty}\left|\frac{1}{n}\widetilde{S}_n\right| = 0\right] = 1.$$

Remark 5.13 The strong law of large numbers implies the weak law. Indeed, if $A_n^\varepsilon := \{|\frac{1}{n}\widetilde{S}_n| > \varepsilon\}$ and $A = \{\limsup_{n\to\infty} |\frac{1}{n}\widetilde{S}_n| > 0\}$, then clearly

$$A = \bigcup_{m\in\mathbb{N}} \limsup_{n\to\infty} A_n^{1/m};$$

hence $\mathbf{P}[\limsup_{n\to\infty} A_n^\varepsilon] = 0$ for $\varepsilon > 0$. By Fatou's lemma (Theorem 4.21), we obtain

$$\begin{aligned}\limsup_{n\to\infty} \mathbf{P}\big[A_n^\varepsilon\big] &= 1 - \liminf_{n\to\infty} \mathbf{E}[\mathbb{1}_{(A_n^\varepsilon)^c}] \\ &\le 1 - \mathbf{E}\Big[\liminf_{n\to\infty} \mathbb{1}_{(A_n^\varepsilon)^c}\Big] = \mathbf{E}\Big[\limsup_{n\to\infty} \mathbb{1}_{A_n^\varepsilon}\Big] = 0.\end{aligned}$$

◊

Theorem 5.14 *Let $X_1, X_2, \ldots$ be uncorrelated random variables in $\mathcal{L}^2(\mathbf{P})$ with $V := \sup_{n\in\mathbb{N}} \mathbf{Var}[X_n] < \infty$. Then $(X_n)_{n\in\mathbb{N}}$ fulfills the weak law of large numbers. More precisely, for any $\varepsilon > 0$, we have*

$$\mathbf{P}\left[\left|\frac{1}{n}\widetilde{S}_n\right| \ge \varepsilon\right] \le \frac{V}{\varepsilon^2 n} \quad \textit{for all } n \in \mathbb{N}. \tag{5.5}$$

Proof Without loss of generality, assume $\mathbf{E}[X_i] = 0$ for all $i \in \mathbb{N}$ and thus $\widetilde{S}_n = X_1 + \ldots + X_n$. By Bienaymé's formula (Theorem 5.7), we obtain

$$\mathbf{Var}\left[\frac{1}{n}\widetilde{S}_n\right] = n^{-2} \sum_{i=1}^n \mathbf{Var}[X_i] \le \frac{V}{n}.$$

By Chebyshev's inequality (Theorem 5.11), for any $\varepsilon > 0$,

$$\mathbf{P}\big[|\widetilde{S}_n/n| \ge \varepsilon\big] \le \frac{V}{\varepsilon^2 n} \overset{n\to\infty}{\longrightarrow} 0.$$

□

Example 5.15 (Weierstraß's approximation theorem) Let $f : [0,1] \to \mathbb{R}$ be a continuous map. By Weierstraß's approximation theorem, there exist polynomials f_n of

degree at most n such that

$$\|f_n - f\|_\infty \overset{n\to\infty}{\longrightarrow} 0,$$

where $\|f\|_\infty := \sup\{|f(x)| : x \in [0,1]\}$ denotes the supremum norm of $f \in C([0,1])$ (the space of continuous functions $[0,1] \to \mathbb{R}$).

We present a probabilistic proof of this theorem. For $n \in \mathbb{N}$, define the polynomial f_n by

$$f_n(x) := \sum_{k=0}^{n} f(k/n)\binom{n}{k} x^k (1-x)^{n-k} \quad \text{for } x \in [0,1].$$

f_n is called the *Bernstein polynomial* of order n.

Fix $\varepsilon > 0$. As f is continuous on the compact interval $[0,1]$, f is uniformly continuous. Hence there exists a $\delta > 0$ such that

$$\big|f(x) - f(y)\big| < \varepsilon \quad \text{for all } x, y \in [0,1] \text{ with } |x-y| < \delta.$$

Now fix $p \in [0,1]$ and let $X_1, X_2, \ldots$ be independent random variables with $X_i \sim \mathrm{Ber}_p$, $i \in \mathbb{N}$. Then $S_n := X_1 + \ldots + X_n \sim b_{n,p}$ and thus

$$\mathbf{E}\big[f(S_n/n)\big] = \sum_{k=0}^{n} f(k/n)\mathbf{P}[S_n = k] = f_n(p).$$

We get

$$\big|f(S_n/n) - f(p)\big| \le \varepsilon + 2\|f\|_\infty \mathbb{1}_{\{|(S_n/n)-p| \ge \delta\}}$$

and thus (by Theorem 5.14 with $V = p(1-p) \le \frac{1}{4}$)

$$\begin{aligned}\big|f_n(p) - f(p)\big| &\le \mathbf{E}\big[\big|f(S_n/n) - f(p)\big|\big] \\ &\le \varepsilon + 2\|f\|_\infty \mathbf{P}\left[\left|\frac{S_n}{n} - p\right| \ge \delta\right] \\ &\le \varepsilon + \frac{\|f\|_\infty}{2\delta^2 n}\end{aligned}$$

for any $p \in [0,1]$. Hence $\|f_n - f\|_\infty \overset{n\to\infty}{\longrightarrow} 0$. ◊

Exercise 5.2.1 (Bernstein–Chernov bound) Let $n \in \mathbb{N}$ and $p_1, \ldots, p_n \in [0,1]$. Let $X_1, \ldots, X_n$ be independent random variables with $X_i = \mathrm{Ber}_{p_i}$ for any $i = 1, \ldots, n$. Define $S_n = X_1 + \ldots + X_n$ and $m := \mathbf{E}[S_n]$. Show that, for any $\delta > 0$, the following two estimates hold:

$$\mathbf{P}\big[S_n \ge (1+\delta)m\big] \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^m$$

and

$$\mathbf{P}\big[S_n \le (1-\delta)m\big] \le \exp\left(-\frac{\delta^2 m}{2}\right).$$

Hint: For S_n, use Markov's inequality with $f(x) = e^{\lambda x}$ for some $\lambda > 0$ and then find the λ that optimizes the bound.

5.3 Strong Law of Large Numbers

We show Etemadi's version [47] of the strong law of large numbers for identically distributed, pairwise independent random variables. There is a zoo of strong laws of large numbers, each of which varies in the exact assumptions it makes on the underlying sequence of random variables. For example, the assumption that the random variables be identically distributed can be waived if other assumptions are introduced such as bounded variances. We do not strive for completeness but show only a few of the statements.

In order to illustrate the method of the proof of Etemadi's theorem, we first present (and prove) a strong law of large numbers under stronger assumptions.

Theorem 5.16 *Let $X_1, X_2, \ldots \in \mathcal{L}^2(\mathbf{P})$ be pairwise independent (that is, X_i and X_j are independent for all $i, j \in \mathbb{N}$ with $i \neq j$) and identically distributed. Then $(X_n)_{n\in\mathbb{N}}$ fulfills the strong law of large numbers.*

Proof The random variables $(X_n^+)_{n\in\mathbb{N}}$ and $(X_n^-)_{n\in\mathbb{N}}$ again form pairwise independent families of square integrable random variables (compare Remark 2.15(ii)). Hence, it is enough to consider $(X_n^+)_{n\in\mathbb{N}}$. Thus we henceforth assume $X_n \ge 0$ almost surely for all $n \in \mathbb{N}$.

Let $S_n = X_1 + \ldots + X_n$ for $n \in \mathbb{N}$. Fix $\varepsilon > 0$. For any $n \in \mathbb{N}$, define $k_n = \lfloor (1+\varepsilon)^n \rfloor \ge \frac{1}{2}(1+\varepsilon)^n$. Then, by Chebyshev's inequality (Theorem 5.11),

$$\begin{aligned}
&\sum_{n=1}^{\infty} \mathbf{P}\left[\left|\frac{S_{k_n}}{k_n} - \mathbf{E}[X_1]\right| \ge (1+\varepsilon)^{-n/4}\right] \\
&\quad\le \sum_{n=1}^{\infty} (1+\varepsilon)^{n/2} \mathbf{Var}\big[k_n^{-1} S_{k_n}\big] \\
&\quad= \sum_{n=1}^{\infty} (1+\varepsilon)^{n/2} k_n^{-1} \mathbf{Var}[X_1] \\
&\quad\le 2\mathbf{Var}[X_1] \sum_{n=1}^{\infty} (1+\varepsilon)^{-n/2} < \infty. \qquad (5.6)
\end{aligned}$$

Thus, by the Borel–Cantelli lemma, for $\mathbf{P}$-a.a. ω, there is an $n_0 = n_0(\omega)$ such that

$$\left| \frac{S_{k_n}}{k_n} - \mathbf{E}[X_1] \right| < (1+\varepsilon)^{-n/4} \quad \text{for all } n \geq n_0,$$

whence

$$\limsup_{n\to\infty} \left| k_n^{-1} S_{k_n} - \mathbf{E}[X_1] \right| = 0 \quad \text{almost surely.}$$

Note that $k_{n+1} \leq (1+2\varepsilon)k_n$ for sufficiently large $n \in \mathbb{N}$. For $l \in \{k_n, \ldots, k_{n+1}\}$, we get

$$\frac{1}{1+2\varepsilon} k_n^{-1} S_{k_n} \leq k_{n+1}^{-1} S_{k_n} \leq l^{-1} S_l \leq k_n^{-1} S_{k_{n+1}} \leq (1+2\varepsilon) k_{n+1}^{-1} S_{k_{n+1}}.$$

Now $1 - (1+2\varepsilon)^{-1} \leq 2\varepsilon$ implies

$$\begin{aligned} \limsup_{l\to\infty} \left| l^{-1} S_l - \mathbf{E}[X_1] \right| &\leq \limsup_{n\to\infty} \left| k_n^{-1} S_{k_n} - \mathbf{E}[X_1] \right| + 2\varepsilon \limsup_{n\to\infty} k_n^{-1} S_{k_n} \\ &\leq 2\varepsilon \mathbf{E}[X_1] \quad \text{almost surely.} \end{aligned}$$

Hence the strong law of large numbers is in force. □

The similarity of the variance estimates in the weak law of large numbers and in (5.6) suggests that in the preceding theorem the condition that the random variables $X_1, X_2, \ldots$ be identically distributed could be replaced by the condition that the variances be bounded (see Exercise 5.3.1).

We can weaken the condition in Theorem 5.16 in a different direction by requiring integrability only instead of square integrability of the random variables.

Theorem 5.17 (Etemadi's strong law of large numbers (1981)) *Let* $X_1, X_2, \ldots \in \mathcal{L}^1(\mathbf{P})$ *be pairwise independent and identically distributed. Then* $(X_n)_{n\in\mathbb{N}}$ *fulfills the strong law of large numbers.*

We follow the proof in [39, Section 2.4]. Define $\mu = \mathbf{E}[X_1]$ and $S_n = X_1 + \ldots + X_n$. We start with some preparatory lemmas. (For the "a.s." notation see Definition 1.68.)

Lemma 5.18 *For* $n \in \mathbb{N}$, *define* $Y_n := X_n \mathbb{1}_{\{|X_n| \leq n\}}$ *and* $T_n = Y_1 + \ldots + Y_n$. *The sequence* $(X_n)_{n\in\mathbb{N}}$ *fulfills the strong law of large numbers if* $T_n/n \overset{n\to\infty}{\longrightarrow} \mu$ *a.s.*

Proof By Theorem 4.26, we have $\sum_{n=1}^{\infty} \mathbf{P}[|X_n| > n] \leq \mathbf{E}[|X_1|] < \infty$. Thus, by the Borel–Cantelli lemma,

$$\mathbf{P}[X_n \neq Y_n \text{ for infinitely many } n] = 0.$$

Hence there is an $n_0 = n_0(\omega)$ with $X_n = Y_n$ for all $n \geq n_0$, whence for $n \geq n_0$

$$\frac{T_n - S_n}{n} = \frac{T_{n_0} - S_{n_0}}{n} \overset{n\to\infty}{\longrightarrow} 0. \qquad \square$$

Lemma 5.19 $2x \sum_{n>x} n^{-2} \leq 4$ *for all* $x \geq 0$.

Proof For $m \in \mathbb{N}$, by comparison with the corresponding integral, we get

$$\sum_{n=m}^{\infty} n^{-2} \leq m^{-2} + \int_m^{\infty} t^{-2}\, dt = m^{-2} + m^{-1} \leq \frac{2}{m}. \qquad \square$$

Lemma 5.20 $\sum_{n=1}^{\infty} \frac{\mathbf{E}[Y_n^2]}{n^2} \leq 4\mathbf{E}[|X_1|]$.

Proof By Theorem 4.26,

$$\mathbf{E}\big[Y_n^2\big] = \int_0^{\infty} \mathbf{P}\big[Y_n^2 > t\big]\, dt.$$

Substituting $x = \sqrt{t}$, we obtain

$$\mathbf{E}\big[Y_n^2\big] = \int_0^{\infty} 2x\mathbf{P}\big[|Y_n| > x\big]\, dx \leq \int_0^{n} 2x\mathbf{P}\big[|X_1| > x\big]\, dx.$$

By Lemma 5.19, for $m \to \infty$,

$$f_m(x) = \left(\sum_{n=1}^{m} n^{-2}\mathbb{1}_{\{x<n\}}\right) 2x\mathbf{P}\big[|X_1| > x\big] \uparrow f(x) \leq 4\mathbf{P}\big[|X_1| > x\big].$$

Hence, by the monotone limit theorem, we can interchange the summation and the integral and obtain

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\mathbf{E}[Y_n^2]}{n^2} &\leq \sum_{n=1}^{\infty} n^{-2} \int_0^{\infty} \mathbb{1}_{\{x<n\}} 2x\mathbf{P}\big[|X_1| > x\big]\, dx \\
&= \int_0^{\infty} \left(\sum_{n=1}^{\infty} n^{-2}\mathbb{1}_{\{x<n\}}\right) 2x\mathbf{P}\big[|X_1| > x\big]\, dx \\
&\leq 4\int_0^{\infty} \mathbf{P}\big[|X_1| > x\big]\, dx = 4\mathbf{E}\big[|X_1|\big].
\end{aligned} \qquad \square$$

Proof of Theorem 5.17 As in the proof of Theorem 5.16, it is enough to consider the case $X_n \geq 0$. Fix $\varepsilon > 0$ and let $\alpha = 1 + \varepsilon$. For $n \in \mathbb{N}$, define $k_n = \lfloor \alpha^n \rfloor$. Note that $k_n \geq \alpha^n/2$. Hence, for all $m \in \mathbb{N}$ (with $n_0 = \lceil \log m / \log \alpha \rceil$),

$$\sum_{n:k_n \geq m} k_n^{-2} \leq 4 \sum_{n=n_0}^{\infty} \alpha^{-2n} = 4\alpha^{-2n_0}\big(1 - \alpha^{-2}\big)^{-1} \leq 4\big(1 - \alpha^{-2}\big)^{-1} m^{-2}. \tag{5.7}$$

The aim is to employ Lemma 5.20 to refine the estimate (5.6) for $(Y_n)_{n\in\mathbb{N}}$ and $(T_n)_{n\in\mathbb{N}}$. For $\delta > 0$, Chebyshev's inequality yields (together with (5.7))

$$\begin{aligned}
&\sum_{n=1}^{\infty} \mathbf{P}\big[\big|T_{k_n} - \mathbf{E}[T_{k_n}]\big| > \delta k_n\big] \\
&\quad\le \delta^{-2} \sum_{n=1}^{\infty} \frac{\mathbf{Var}[T_{k_n}]}{k_n^2} \\
&\quad= \delta^{-2} \sum_{n=1}^{\infty} k_n^{-2} \sum_{m=1}^{k_n} \mathbf{Var}[Y_m] = \delta^{-2} \sum_{m=1}^{\infty} \mathbf{Var}[Y_m] \sum_{n:k_n\ge m} k_n^{-2} \\
&\quad\le 4\big(1-\alpha^{-2}\big)^{-1} \delta^{-2} \sum_{m=1}^{\infty} m^{-2} \mathbf{E}\big[Y_m^2\big] < \infty \quad \text{by Lemma 5.20.}
\end{aligned}$$

(In the third step, we could change the order of summation since all summands are nonnegative.) Letting $\delta \downarrow 0$, we infer by the Borel–Cantelli lemma

$$\lim_{n\to\infty} \frac{T_{k_n} - \mathbf{E}[T_{k_n}]}{k_n} = 0 \quad \text{almost surely.} \tag{5.8}$$

By the monotone convergence theorem (Theorem 4.20), we have

$$\mathbf{E}[Y_n] = \mathbf{E}[X_1 \mathbb{1}_{\{X_1\le n\}}] \overset{n\to\infty}{\longrightarrow} \mathbf{E}[X_1].$$

Hence $\mathbf{E}[T_{k_n}]/k_n \overset{n\to\infty}{\longrightarrow} \mathbf{E}[X_1]$. By (5.8), we also have $T_{k_n}/k_n \overset{n\to\infty}{\longrightarrow} \mathbf{E}[X_1]$ a.s. As in the proof of Theorem 5.16, we also get (since $Y_n \ge 0$)

$$\lim_{l\to\infty} \frac{T_l}{l} = \mathbf{E}[X_1] \quad \text{almost surely.}$$

By Lemma 5.18, this implies the claim of Theorem 5.17. □

Example 5.21 (Monte Carlo integration) Let $f : [0,1] \to \mathbb{R}$ be a function and assume we want to determine the value of its integral $I := \int_0^1 f(x)\,dx$ numerically. Assume that the computer generates numbers $X_1, X_2, \ldots$ that can be considered as independent random numbers, uniformly distributed on $[0,1]$. For $n \in \mathbb{N}$, define the estimated value

$$\widehat{I}_n := \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Assuming $f \in \mathcal{L}^1([0,1])$, the strong law of large numbers yields $\widehat{I}_n \overset{n\to\infty}{\longrightarrow} I$ a.s.

Note that the last theorem made no statement on the *speed of convergence*. That is, we do not have control on the quantity $\mathbf{P}[|\widehat{I}_n - I| > \varepsilon]$. In order to get more

precise estimates for the integral, we need additional information; for example, the value $V_1 := \int f^2(x)\,dx - I^2$ if $f \in \mathcal{L}^2([0,1])$. (For bounded f, V_1 can easily be bounded.) Indeed, in this case, $\mathbf{Var}[\widehat{I}_n] = V_1/n$; hence, by Chebyshev's inequality,

$$\mathbf{P}\big[|\widehat{I}_n - I| > \varepsilon n^{-1/2}\big] \le V_1/\varepsilon^2.$$

Hence the error is at most of order $n^{-1/2}$. The central limit theorem will show that the error is indeed exactly of this order.

If f is smooth in some sense, then the usual numerical procedures yield better orders of convergence. Hence *Monte Carlo simulation* should be applied only if all other methods fail. This is the case in particular if $[0,1]$ is replaced by $G \subset \mathbb{R}^d$ for very large d. ◊

Definition 5.22 (Empirical distribution function) Let $X_1, X_2, \dots$ be real random variables. The map $F_n : \mathbb{R} \to [0,1]$, $x \mapsto \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,x]}(X_i)$ is called the *empirical distribution function* of $X_1, \dots, X_n$.

Theorem 5.23 (Glivenko–Cantelli) *Let $X_1, X_2, \dots$ be i.i.d. real random variables with distribution function F, and let F_n, $n \in \mathbb{N}$, be the empirical distribution functions. Then*

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} \big|F_n(x) - F(x)\big| = 0 \quad \textit{almost surely.}$$

Proof Fix $x \in \mathbb{R}$ and let $Y_n(x) = \mathbb{1}_{(-\infty,x]}(X_n)$ and $Z_n(x) = \mathbb{1}_{(-\infty,x)}(X_n)$ for $n \in \mathbb{N}$. Additionally, define the left-sided limits $F(x-) = \lim_{y\uparrow x} F(y)$ and similarly for F_n. Then each of the families $(Y_n(x))_{n\in\mathbb{N}}$ and $(Z_n(x))_{n\in\mathbb{N}}$ is independent. Furthermore, $\mathbf{E}[Y_n(x)] = \mathbf{P}[X_n \le x] = F(x)$ and $\mathbf{E}[Z_n(x)] = \mathbf{P}[X_n < x] = F(x-)$. By the strong law of large numbers, we thus have

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n Y_i(x) \overset{n\to\infty}{\longrightarrow} F(x) \quad \text{almost surely}$$

and

$$F_n(x-) = \frac{1}{n}\sum_{i=1}^n Z_i(x) \overset{n\to\infty}{\longrightarrow} F(x-) \quad \text{almost surely.}$$

Formally, define $F(-\infty) = 0$ and $F(\infty) = 1$. Fix some $N \in \mathbb{N}$ and define

$$x_j := \inf\big\{x \in \overline{\mathbb{R}} : F(x) \ge j/N\big\}, \quad j = 0, \dots, N,$$

and

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

As shown above, $R_n \overset{n\to\infty}{\longrightarrow} 0$ almost surely. For $x \in (x_{j-1}, x_j)$, we have (by definition of x_j)

$$F_n(x) \le F_n(x_j-) \le F(x_j-) + R_n \le F(x) + R_n + \frac{1}{N}$$

and

$$F_n(x) \ge F_n(x_{j-1}) \ge F(x_{j-1}) - R_n \ge F(x) - R_n - \frac{1}{N}.$$

Hence

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} \big|F_n(x) - F(x)\big| \le \frac{1}{N} + \limsup_{n\to\infty} R_n = \frac{1}{N}.$$

Letting $N \to \infty$, the claim follows. □

Example 5.24 (Shannon's theorem) Consider a source of information that sends a sequence of independent random symbols $X_1, X_2, \dots$ drawn from a finite alphabet E (that is, from an arbitrary finite set E). Let p_e be the probability of the symbol $e \in E$. Formally, the $X_1, X_2, \dots$ are i.i.d. E-valued random variables with $\mathbf{P}[X_i = e] = p_e$ for $e \in E$.

For any $\omega \in \Omega$ and $n \in \mathbb{N}$, let

$$\pi_n(\omega) := \prod_{i=1}^{n} p_{X_i(\omega)}$$

be the probability that the observed sequence $X_1(\omega), \dots, X_n(\omega)$ occurs. Define $Y_n(\omega) := -\log(p_{X_n(\omega)})$. Then $(Y_n)_{n\in\mathbb{N}}$ is i.i.d. and $\mathbf{E}[Y_n] = H(p)$, where

$$H(p) := -\sum_{e\in E} p_e \log(p_e)$$

is the *entropy* of the distribution $p = (p_e)_{e\in E}$ (compare Definition 5.25). By the strong law of large numbers, we infer Shannon's theorem:

$$-\frac{1}{n}\log \pi_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \overset{n\to\infty}{\longrightarrow} H(p) \quad \text{almost surely.}$$

♢

*Entropy and Source Coding Theorem**

We briefly discuss the importance of π_n and the entropy. How can we quantify the *information* inherent in a message $X_1(\omega), \dots, X_n(\omega)$? This information can be measured by the length of the shortest sequence of zeros and ones by which the message can be encoded. Of course, you do not want to invent a new code for

every message but rather use one code that allows for the shortest average coding of the messages for the particular information source. To this end, associate with each symbol $e \in E$ a sequence of zeros and ones that when concatenated yield the message. The length $l(e)$ of the sequence that codes for e may depend on e. Hence, for efficiency, those symbols that appear more often get a shorter code than the more rare symbols. The Morse alphabet is constructed similarly (the letters "e" and "t", which are the most frequent letters in English, have the shortest codes ("dot" and "dash"), and the rare letter "q" has the code "dash-dash-dot-dash"). However, the Morse code also consists of gaps of different lengths that signal ends of letters and words. As we want to use only zeros and ones (and no gap-like symbols), we have to arrange the code in such a way that no code is the beginning of the code of a different symbol. For example, we could not encode one symbol with 0110 and a different one with 011011. A code that fulfills this condition is called a *binary prefix code*. Denote by $c(e) \in \{0,1\}^{l(e)}$ the code of e, where $l(e)$ is its length. We can represent the codes of all letters in a tree.

Let us construct a code $C = (c(e), e \in E)$ that is efficient in the sense that it minimizes the expected length of the code (of a random symbol)

$$L_p(C) := \sum_{e \in E} p_e l(e).$$

We first define a specific code and then show that it is almost optimal. As a first step, we enumerate $E = \{e_1, \ldots, e_N\}$ such that $p_{e_1} \geq p_{e_2} \geq \ldots \geq p_{e_N}$. Define $\ell(e) \in \mathbb{N}$ for any $e \in E$ by

$$2^{-\ell(e)} \leq p_e < 2^{-\ell(e)+1}.$$

Let $\tilde{p}_e = 2^{-\ell(e)}$ for any $e \in E$ and let $\tilde{q}_k = \sum_{l<k} \tilde{p}_{e_l}$ for $k = 1, \ldots, N$.

By construction, $\ell(e_l) \leq \ell(e_k)$ for all $l \leq k$; hence the binary representation of $\tilde{q}_k$ has at most $\ell(e_k)$ digits:

$$\tilde{q}_k = \sum_{i=1}^{\ell(e_k)} c_i(e_k) 2^{-i}.$$

Here the numbers $c_1(e_k), \ldots, c_{\ell(e_k)}(e_k) \in \{0,1\}$ are uniquely determined.

Clearly, $\tilde{q}_l \geq \tilde{q}_k + 2^{-\ell(e_k)}$ for any $l > k$; hence

$$\big(c_1(e_k), \ldots, c_{\ell(e_k)}(e_k)\big) \neq \big(c_1(e_l), \ldots, c_{\ell(e_k)}(e_l)\big) \quad \text{for all } l > k.$$

Thus $C = (c(e), e \in E)$ is a prefix code.

For any $b > 0$ and $x > 0$, denote by $\log_b(x) := \frac{\log(x)}{\log(b)}$ the logarithm of x to base b. By construction, $-\log_2(p_e) \leq l(e) \leq 1 - \log_2(p_e)$. Hence the expected length is

$$-\sum_{e \in E} p_e \log_2(p_e) \leq L_p(C) \leq 1 - \sum_{e \in E} p_e \log_2(p_e).$$

The length of this code for the first n symbols of our random information source is thus approximately $-\sum_{k=1}^{n} \log_2(p_{X_k(\omega)}) = -\log_2 \pi_n(\omega)$. Here we have the connection to Shannon's theorem. That theorem thus makes a statement about the length of a binary prefix code needed to transmit a long message.

Now, is the code constructed above optimal, or are there codes with smaller mean length? The answer is given by the source coding theorem for which we prepare with a definition and a lemma.

Definition 5.25 (Entropy) Let $p = (p_e)_{e\in E}$ be a probability distribution on the countable set E. For $b > 0$, define

$$H_b(p) := -\sum_{e\in E} p_e \log_b(p_e)$$

with the convention $0\log_b(0) := 0$. We call $H(p) := H_e(p)$ ($e = 2.71\ldots$ Euler's number) the *entropy* and $H_2(p)$ the *binary entropy* of p.

Note that, for infinite E, the entropy need not be finite.

Lemma 5.26 (Entropy inequality) *Let b and p be as above. Further, let q be a sub-probability distribution; that is, $q_e \geq 0$ for all $e \in E$ and $\sum_{e\in E} q_e \leq 1$. Then*

$$H_b(p) \leq -\sum_{e\in E} p_e \log_b(q_e) \tag{5.9}$$

with equality if and only if $H_b(p) = \infty$ or $q = p$.

Proof Without loss of generality, we can do the computation with $b = e$; that is, with the natural logarithm. Note that $\log(1 + x) \leq x$ for $x > -1$ with equality if and only if $x = 0$. If in (5.9) the left-hand side is finite, then we can subtract the right-hand side from the left-hand side and obtain

$$\begin{aligned} H(p) + \sum_{e\in E} p_e \log(q_e) &= \sum_{e:p_e>0} p_e \log(q_e/p_e) \\ &= \sum_{e:p_e>0} p_e \log\left(1 + \frac{q_e - p_e}{p_e}\right) \\ &\leq \sum_{e:p_e>0} p_e \frac{q_e - p_e}{p_e} = \sum_{e\in E} (q_e - p_e) \leq 0. \end{aligned}$$

If $q \neq p$, then there is an $e \in E$ with $p_e > 0$ and $q_e \neq p_e$. If this is the case, then strict inequality holds if $H(p) < \infty$. □

Theorem 5.27 (Source coding theorem) *Let $p = (p_e)_{e\in E}$ be a probability distribution on the finite alphabet E. For any binary prefix code $C = (c(e), e \in E)$, we have $L_p(C) \geq H_2(p)$. Furthermore, there is a binary prefix code C with $L_p(C) \leq H_2(p) + 1$.*

Proof The second part of the theorem was shown in the above construction. Now assume that a prefix code is given. Let $L = \max_{e\in E} l(e)$. For $e \in E$, let

$$C_L(e) = \big\{c \in \{0,1\}^L : c_k = c_k(e) \text{ for } k \leq l(e)\big\}$$

the set of all dyadic sequences of length L that start like $c(e)$. Since we have a prefix code, the sets $C_L(e)$, $e \in E$, are pairwise disjoint and $\bigcup_{e\in E} C_L(e) \subset \{0,1\}^L$. Hence, if we define $q_e := 2^{-l(e)}$, then (note that $\#C_L(e) = 2^{L-l(e)}$)

$$\sum_{e\in E} q_e = 2^{-L} \sum_{e\in E} \#C_L(e) \leq 1.$$

By Lemma 5.26, we have $L_p(C) = \sum_{e\in E} p_e l(e) = -\sum_{e\in E} p_e \log_2(q_e) \geq H_2(p)$. □

Exercise 5.3.1 Show the following improvement of Theorem 5.16: If $X_1, X_2, \ldots \in \mathcal{L}^2(\mathbf{P})$ are pairwise independent with bounded variances, then $(X_n)_{n\in\mathbb{N}}$ fulfills the strong law of large numbers.

Exercise 5.3.2 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent identically distributed random variables with $\frac{1}{n}(X_1 + \ldots + X_n) \overset{n\to\infty}{\longrightarrow} Y$ almost surely for some random variable Y. Show that $X_1 \in \mathcal{L}^1(\mathbf{P})$ and $Y = \mathbf{E}[X_1]$ almost surely.

Hint: First show that

$$\mathbf{P}\big[|X_n| > n \text{ for infinitely many } n\big] = 0 \quad \Longleftrightarrow \quad X_1 \in \mathcal{L}^1(\mathbf{P}).$$

Exercise 5.3.3 Let E be a finite set and let p be a probability vector on E. Show that the entropy $H(p)$ is minimal (in fact, zero) if $p = \delta_e$ for some $e \in E$. It is maximal (in fact, $\log(\#E)$) if p is the uniform distribution on E.

Exercise 5.3.4 (Subadditivity of Entropy) For $i = 1, 2$, let E^i be a finite set and p^i a probability vector on E^i. Let p be a probability vector on $E^1 \times E^2$ with marginals p^1 and p^2. That is,

$$\sum_{e^2\in E^2} p_{(e^1,e^2)} = p^1_{e^1} \quad \text{and} \quad \sum_{f^1\in E^1} p_{(f^1,f^2)} = p^2_{f^2} \quad \text{for all } e^1 \in E^1,\ f^2 \in E^2.$$

Show that $H(p) \leq H(p^1) + H(p^2)$.

Exercise 5.3.5 Let $b \in \{2, 3, 4, \ldots\}$. A b-adic prefix code is defined in a similar way as a binary prefix code; however, instead of 0 and 1, now all numbers $0, 1, \ldots, b-1$

are admissible. Show that the statement of the source coding theorem holds for b-adic prefix codes with $H_2(p)$ replaced by $H_b(p)$.

Exercise 5.3.6 We want to check the efficiency of the Morse alphabet. To this end we need a table of the Morse code as well as the frequencies of the letters in a typical text. The following frequencies for letters in German texts are taken from [11, p. 10]. The frequencies for other languages can be found easily, e.g., at Wikipedia.

Letter	Morse code	Frequency	Letter	Morse code	Frequency
A	.-	0.0651	N	-.	0.0978
B	-...	0.0189	O	---	0.0251
C	-.-.	0.0306	P	.--.	0.0079
D	-..	0.0508	Q	--.-	0.0002
E	.	0.1740	R	.-.	0.07
F	..-.	0.0166	S	...	0.0727
G	--.	0.0301	T	-	0.0615
H	....	0.0476	U	..-	0.0435
I	..	0.0755	V	...-	0.0067
J	.---	0.0027	W	.--	0.0189
K	-.-	0.0121	X	-..-	0.0003
L	.-..	0.0344	Y	-.--	0.0004
M	--	0.0253	Z	--..	0.0113

Here '.' denotes a short signal while '-' denotes a long signal. Each letter is finished by a pause sign. Thus the Morse code can be interpreted as a ternary prefix code.

Determine the average code length of a letter and compare it with the entropy H_3 in order to check the efficiency of the Morse code.

5.4 Speed of Convergence in the Strong LLN

In the weak law of large numbers, we had a statement on the speed of convergence (Theorem 5.14). In the strong law of large numbers, however, we did not. As we required only first moments, in general, we cannot expect to get any useful statements. However, if we assume the existence of higher moments, we get reasonable estimates on the rate of convergence.

The core of the weak law of large numbers is Chebyshev's inequality. Here we present a stronger inequality that claims the same bound but now for the maximum over all partial sums until a fixed time.

Theorem 5.28 (Kolmogorov's inequality) *Let $n \in \mathbb{N}$ and let $X_1, X_2, \ldots, X_n$ be independent random variables with $\mathbf{E}[X_i] = 0$ and $\mathbf{Var}[X_i] < \infty$ for $i = 1, \ldots, n$. Further, let $S_k = X_1 + \ldots + X_k$ for $k = 1, \ldots, n$. Then, for any $t > 0$,*

$$\mathbf{P}\big[\max\{S_k : k = 1, \ldots, n\} \geq t\big] \leq \frac{\mathbf{Var}[S_n]}{t^2 + \mathbf{Var}[S_n]}. \tag{5.10}$$

Furthermore, Kolmogorov's inequality holds:

$$\mathbf{P}\big[\max\big\{|S_k| : k = 1, \ldots, n\big\} \geq t\big] \leq t^{-2}\mathbf{Var}[S_n]. \tag{5.11}$$

In Theorem 11.2 we will see Doob's inequality, which is a generalization of Kolmogorov's inequality.

Proof We decompose the probability space according to the first time τ at which the partial sums exceed the value t. Hence, let

$$\tau := \min\big\{k \in \{1, \ldots, n\} : S_k \geq t\big\}$$

and $A_k = \{\tau = k\}$ for $k = 1, \ldots, n$. Further, let

$$A = \biguplus_{k=1}^{n} A_k = \big\{\max\{S_k : k = 1, \ldots, n\} \geq t\big\}.$$

Let $c \geq 0$. The random variable $(S_k + c)\mathbb{1}_{A_k}$ is $\sigma(X_1, \ldots, X_k)$-measurable and $S_n - S_k$ is $\sigma(X_{k+1}, \ldots, X_n)$-measurable. By Theorem 2.26, the two random variables are independent, and

$$\mathbf{E}\big[(S_k + c)\mathbb{1}_{A_k}(S_n - S_k)\big] = \mathbf{E}\big[(S_k + c)\mathbb{1}_{A_k}\big]\mathbf{E}[S_n - S_k] = 0.$$

Clearly, the events $A_1, \ldots, A_n$ are pairwise disjoint; hence $\sum_{k=1}^{n} \mathbb{1}_{A_k} = \mathbb{1}_A \leq 1$. We thus obtain

$$\begin{aligned}
\mathbf{Var}[S_n] + c^2 &= \mathbf{E}\big[(S_n + c)^2\big] \\
&\geq \mathbf{E}\left[\sum_{k=1}^{n} (S_n + c)^2 \mathbb{1}_{A_k}\right] = \sum_{k=1}^{n} \mathbf{E}\big[(S_n + c)^2 \mathbb{1}_{A_k}\big] \\
&= \sum_{k=1}^{n} \mathbf{E}\big[\big((S_k + c)^2 + 2(S_k + c)(S_n - S_k) + (S_n - S_k)^2\big)\mathbb{1}_{A_k}\big] \\
&= \sum_{k=1}^{n} \mathbf{E}\big[(S_k + c)^2 \mathbb{1}_{A_k}\big] + \sum_{k=1}^{n} \mathbf{E}\big[(S_n - S_k)^2 \mathbb{1}_{A_k}\big] \\
&\geq \sum_{k=1}^{n} \mathbf{E}\big[(S_k + c)^2 \mathbb{1}_{A_k}\big].
\end{aligned} \tag{5.12}$$

Since $c \geq 0$, we have $(S_k + c)^2 \mathbb{1}_{A_k} \geq (t + c)^2 \mathbb{1}_{A_k}$. Hence we can continue (5.12) to get

$$\mathbf{Var}[S_n] + c^2 \geq \sum_{k=1}^{n} \mathbf{E}\big[(t + c)^2 \mathbb{1}_{A_k}\big] = (t + c)^2 \mathbf{P}[A].$$

For $c = \mathbf{Var}[S_n]/t \geq 0$, we obtain

$$\mathbf{P}[A] \leq \frac{\mathbf{Var}[S_n] + c^2}{(t + c)^2} = \frac{c(t + c)}{(t + c)^2} = \frac{tc}{t^2 + tc} = \frac{\mathbf{Var}[S_n]}{t^2 + \mathbf{Var}[S_n]}.$$

This shows (5.10). In order to show (5.11), choose

$$\bar{\tau} := \min\big\{k \in \{1, \ldots, n\} : |S_k| \geq t\big\}.$$

Let $\bar{A}_k = \{\bar{\tau} = k\}$ and $\bar{A} = \{\bar{\tau} \leq n\}$. We cannot now continue (5.12) as above with $c > 0$. However, if we choose $c = 0$, then $S_k^2 \mathbb{1}_{\bar{A}_k} \geq t^2 \mathbb{1}_{\bar{A}_k}$. The same calculation as in (5.12) does then yield $\mathbf{P}[\bar{A}] \leq t^{-2} \mathbf{Var}[S_n]$. □

From Kolmogorov's inequality, we derive the following sharpening of the strong law of large numbers.

Theorem 5.29 *Let $X_1, X_2, \ldots$ be independent random variables with $\mathbf{E}[X_n] = 0$ for any $n \in \mathbb{N}$ and $V := \sup\{\mathbf{Var}[X_n] : n \in \mathbb{N}\} < \infty$. Then, for any $\varepsilon > 0$,*

$$\limsup_{n \to \infty} \frac{|S_n|}{n^{1/2} (\log(n))^{(1/2)+\varepsilon}} = 0 \quad \textit{almost surely.}$$

Proof Let $k_n = 2^n$ and $l(n) = n^{1/2} (\log(n))^{(1/2)+\varepsilon}$ for $n \in \mathbb{N}$. Then we have $l(k_{n+1})/l(k_n) \overset{n \to \infty}{\longrightarrow} \sqrt{2}$. Hence, for $n \in \mathbb{N}$ sufficiently large and $k \in \mathbb{N}$ with $k_{n-1} \leq k \leq k_n$, we have $|S_k|/l(k) \leq 2|S_k|/l(k_n)$. Hence, it is enough to show for every $\delta > 0$ that

$$\limsup_{n \to \infty} l(k_n)^{-1} \max\big\{|S_k| : k \leq k_n\big\} \leq \delta \quad \text{almost surely.} \tag{5.13}$$

For $\delta > 0$ and $n \in \mathbb{N}$, define $A_n^\delta := \{\max\{|S_k| : k \leq k_n\} > \delta l(k_n)\}$. Kolmogorov's inequality yields

$$\sum_{n=1}^{\infty} \mathbf{P}\big[A_n^\delta\big] \leq \sum_{n=1}^{\infty} \delta^{-2} \big(l(k_n)\big)^{-2} V k_n = \frac{V}{\delta^2 (\log 2)^{1+2\varepsilon}} \sum_{n=1}^{\infty} n^{-1-2\varepsilon} < \infty.$$

The Borel–Cantelli lemma then gives $\mathbf{P}[\limsup_{n \to \infty} A_n^\delta] = 0$ and hence (5.13). □

In Chapter 22, we will see that for independent identically distributed, square integrable, centered random variables $X_1, X_2, \ldots$, the following strengthening holds,

$$\limsup_{n\to\infty} \frac{|S_n|}{\sqrt{2n\mathbf{Var}[X_1]\log(\log(n))}} = 1 \quad \text{almost surely.}$$

Hence, in this case, the speed of convergence is known precisely. If the $X_1, X_2, \ldots$ are not independent but only pairwise independent, then the rate of convergence deteriorates, although not drastically. Here we cite without proof a theorem that was found independently by Rademacher (1922) [141] and Menshov (1923) [113].

Theorem 5.30 (Rademacher–Menshov) *Let $X_1, X_2, \ldots$ be uncorrelated, square integrable, centered random variables and let $(a_n)_{n\in\mathbb{N}}$ be an increasing sequence of nonnegative numbers such that*

$$\sum_{n=1}^{\infty} (\log n)^2 a_n^{-2} \mathbf{Var}[X_n] < \infty. \tag{5.14}$$

Then

$$\limsup_{n\to\infty} \left| a_n^{-1} \sum_{k=1}^{n} X_k \right| = 0 \quad \textit{almost surely.}$$

Proof See, for example, [128]. □

Remark 5.31 Condition (5.14) is sharp in the sense that for any increasing sequence $(a_n)_{n\in\mathbb{N}}$ with $\sum_{n=1}^{\infty} a_n^{-2}(\log n)^2 = \infty$, there exists a sequence of pairwise independent, square integrable, centered random variables $X_1, X_2, \ldots$ with $\mathbf{Var}[X_n] = 1$ for all $n \in \mathbb{N}$ such that

$$\limsup_{n\to\infty} \left| a_n^{-1} \sum_{k=1}^{n} X_k \right| = \infty \quad \text{almost surely.}$$

See [22]. There an example of [163] (see also [164, 165]) for orthogonal series is developed further. See also [117]. ◊

For random variables with infinite variance, the statements about the rate of convergence naturally get weaker. For example (see [8]), see the following theorem.

Theorem 5.32 (Baum and Katz (1965)) *Let $\gamma > 1$ and let $X_1, X_2, \ldots$ be i.i.d. Define $S_n = X_1 + \ldots + X_n$ for $n \in \mathbb{N}$. Then*

$$\sum_{n=1}^{\infty} n^{\gamma-2} \mathbf{P}\big[|S_n|/n > \varepsilon\big] < \infty \quad \textit{for any } \varepsilon > 0$$
$$\iff \quad \mathbf{E}\big[|X_1|^{\gamma}\big] < \infty \quad \textit{and} \quad \mathbf{E}[X_1] = 0.$$

Exercise 5.4.1 Let $X_1, \ldots, X_n$ be independent real random variables and let $S_k = X_1 + \ldots + X_k$ for $k = 1, \ldots, n$. Show that for $t > 0$ *Etemadi's inequality* holds:

$$\mathbf{P}\Big[\max_{k=1,\ldots,n} |S_k| \geq t\Big] \leq 3 \max_{k=1,\ldots,n} \mathbf{P}\big[|S_k| \geq t/3\big].$$

5.5 The Poisson Process

We develop a model for the number of clicks of a Geiger counter in the (time) interval $I = (a, b]$. The number of clicks should obey the following rules. It should

- be random and independent for disjoint intervals,
- be homogeneous in time in the sense that the number of clicks in $I = (a, b]$ has the same distribution as the number of clicks in $c + I = (a + c, b + c]$,
- have finite expectation, and
- have no double points: At any point of time, the counter makes at most one click.

We formalize these requirements by introducing the following notation:

$$\mathcal{I} := \big\{(a, b] : a, b \in [0, \infty), a \leq b\big\},$$

$$\ell\big((a, b]\big) := b - a \qquad \big(\text{the length of the interval } I = (a, b]\big).$$

For $I \in \mathcal{I}$, let N_I be the number of clicks after time a but no later than b. In particular, we define $N_t := N_{(0,t]}$ as the total number of clicks until time t. The above requirements translate to: $(N_I, I \in \mathcal{I})$ being a family of random variables with values in $\mathbb{N}_0$ and with the following properties:

(P1) $N_{I \cup J} = N_I + N_J$ if $I \cap J = \emptyset$ and $I \cup J \in \mathcal{I}$.
(P2) The distribution of N_I depends only on the length of I: $\mathbf{P}_{N_I} = \mathbf{P}_{N_J}$ for all $I, J \in \mathcal{I}$ with $\ell(I) = \ell(J)$.
(P3) If $\mathcal{J} \subset \mathcal{I}$ with $I \cap J = \emptyset$ for all $I, J \in \mathcal{J}$ with $I \neq J$, then $(N_J, J \in \mathcal{J})$ is an independent family.
(P4) For any $I \in \mathcal{I}$, we have $\mathbf{E}[N_I] < \infty$.
(P5) $\limsup_{\varepsilon \downarrow 0} \varepsilon^{-1} \mathbf{P}[N_\varepsilon \geq 2] = 0$.

The meaning of (P5) is explained by the following calculation. Define

$$\lambda := \limsup_{\varepsilon \downarrow 0} \varepsilon^{-1} \mathbf{P}[N_\varepsilon \geq 2].$$

For any $n \in \mathbb{N}$ and $\varepsilon > 0$, we have

$$\mathbf{P}[N_{2^{-n}} \geq 2] \geq \big\lfloor 2^{-n}/\varepsilon \big\rfloor \mathbf{P}[N_\varepsilon \geq 2] - \big\lfloor 2^{-n}/\varepsilon \big\rfloor^2 \mathbf{P}[N_\varepsilon \geq 2]^2.$$

Hence

$$2^n \mathbf{P}[N_{2^{-n}} \geq 2] \geq \lambda - 2^{-n} \lambda^2 \overset{n \to \infty}{\longrightarrow} \lambda.$$

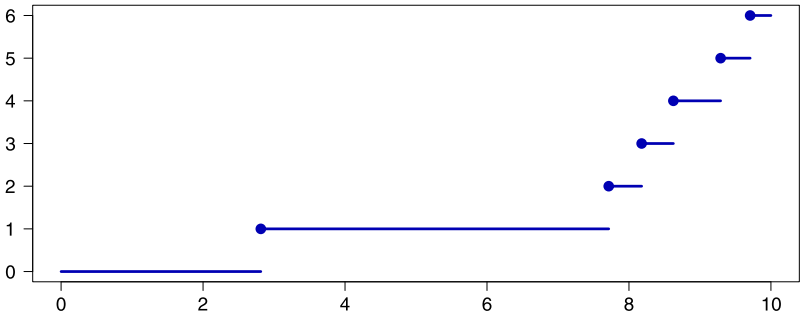

Fig. 5.1 Simulation of a Poisson process with rate $\alpha = 0.5$

Then (because $(1 - a_k/k)^k \overset{k\to\infty}{\longrightarrow} e^{-a}$ if $a_k \overset{k\to\infty}{\longrightarrow} a$)

$$\begin{aligned}
&\mathbf{P}\big[\text{there is a double click in } (0,1]\big] \\
&= \lim_{n\to\infty} \mathbf{P}\left[\bigcup_{k=0}^{2^n-1} \{N_{(k2^{-n},(k+1)2^{-n}]} \geq 2\}\right] \\
&= 1 - \lim_{n\to\infty} \mathbf{P}\left[\bigcap_{k=0}^{2^n-1} \{N_{(k2^{-n},(k+1)2^{-n}]} \leq 1\}\right] \\
&= 1 - \lim_{n\to\infty} \prod_{k=0}^{2^n-1} \mathbf{P}[N_{(k2^{-n},(k+1)2^{-n}]} \leq 1] \\
&= 1 - \lim_{n\to\infty} \big(1 - \mathbf{P}[N_{2^{-n}} \geq 2]\big)^{2^n} \\
&= 1 - e^{-\lambda}.
\end{aligned}$$

Hence we have to postulate $\lambda = 0$. This, however, is exactly (P5).

The following theorem shows that properties (P1)–(P5) characterize the random variables $(N_I, I \in \mathcal{I})$ uniquely and that they form a Poisson process.

Definition 5.33 (Poisson process) A family $(N_t, t \geq 0)$ of $\mathbb{N}_0$-valued random variables is called a *Poisson process* with intensity $\alpha \geq 0$ if $N_0 = 0$ and if:

(i) For any $n \in \mathbb{N}$ and any choice of $n+1$ numbers $0 = t_0 < t_1 < \ldots < t_n$, the family $(N_{t_i} - N_{t_{i-1}}, i = 1, \ldots, n)$ is independent.
(ii) For $t > s \geq 0$, the difference $N_t - N_s$ is Poisson-distributed with parameter $\alpha(t-s)$; that is,

$$\mathbf{P}[N_t - N_s = k] = e^{-\alpha(t-s)} \frac{(\alpha(t-s))^k}{k!} \quad \text{for all } k \in \mathbb{N}_0.$$

See Fig. 5.1 for a computer simulation of a Poisson process.

The *existence* of the Poisson process has not yet been shown. We come back to this point in Theorem 5.36.

Theorem 5.34 *If* $(N_I, I \in \mathcal{I})$ *has properties* (P1)–(P5), *then* $(N_{(0,t]}, t \geq 0)$ *is a Poisson process with intensity* $\alpha := \mathbf{E}[N_{(0,1]}]$. *If, on the other hand,* $(N_t, t \geq 0)$ *is a Poisson process, then* $(N_t - N_s, (s,t] \in \mathcal{I})$ *has properties* (P1)–(P5).

Proof First assume that $(N_t, t \geq 0)$ is a Poisson process with intensity $\alpha \geq 0$. Then, for $I = (a, b]$, clearly $\mathbf{P}_{N_I} = \mathrm{Poi}_{\alpha(b-a)} = \mathrm{Poi}_{\alpha\ell(I)}$. Hence (P2) holds. By (i), we have (P3). Clearly, $\mathbf{E}[N_I] = \alpha\ell(I) < \infty$; thus we have (P4). Finally, $\mathbf{P}[N_\varepsilon \geq 2] = 1 - e^{-\alpha\varepsilon} - \alpha\varepsilon e^{-\alpha\varepsilon} = f(0) - f(\alpha\varepsilon)$, where $f(x) := e^{-x} + xe^{-x}$. The derivative is $f'(x) = -xe^{-x}$, whence

$$\lim_{\varepsilon \downarrow 0} \varepsilon^{-1} \mathbf{P}[N_\varepsilon \geq 2] = -\alpha f'(0) = 0.$$

This implies (P5).

Now assume that $(N_I, I \in \mathcal{I})$ fulfills (P1)–(P5). Define $\alpha(t) := \mathbf{E}[N_t]$. Then (owing to (P2))

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

As $t \mapsto \alpha(t)$ is monotone increasing, this implies linearity: $\alpha(t) = t\alpha(1)$ for any $t \geq 0$. Letting $\alpha := \alpha(1)$, we obtain $\mathbf{E}[N_I] = \alpha\ell(I)$. It remains to show that $\mathbf{P}_{N_t} = \mathrm{Poi}_{\alpha t}$. In order to apply the Poisson approximation theorem (Theorem 3.7), for fixed $n \in \mathbb{N}$, we decompose the interval $(0, t]$ into 2^n disjoint intervals of equal length,

$$I^n(k) := \big((k-1)2^{-n}t, k2^{-n}t\big], \quad k = 1, \ldots, 2^n.$$

Now define $X^n(k) := N_{I^n(k)}$ and

$$\overline{X}^n(k) := \begin{cases} 1, & \text{if } X^n(k) \geq 1, \\ 0, & \text{else.} \end{cases}$$

By properties (P2) and (P3), the random variables $(X^n(k), k = 1, \ldots, 2^n)$ are independent and identically distributed. Hence also $(\overline{X}^n(k), k = 1, \ldots, 2^n)$ are i.i.d., namely $\overline{X}^n(k) \sim \mathrm{Ber}_{p_n}$, where $p_n = \mathbf{P}[N_{2^{-n}t} \geq 1]$.

Finally, let $N_t^n := \sum_{k=1}^{2^n} \overline{X}^n(k)$. Then $N_t^n \sim b_{2^n, p_n}$. Clearly, $N_t^{n+1} - N_t^n \geq 0$. Now, by (P5),

$$\mathbf{P}\big[N_t \neq N_t^n\big] \leq \sum_{k=1}^{2^n} \mathbf{P}\big[X^n(k) \geq 2\big] = 2^n\, \mathbf{P}[N_{2^{-n}t} \geq 2] \overset{n\to\infty}{\longrightarrow} 0. \tag{5.15}$$

Hence $\mathbf{P}[N_t = \lim_{n\to\infty} N_t^n] = 1$. By the monotone convergence theorem, we get

$$\alpha t = \mathbf{E}[N_t] = \lim_{n\to\infty} \mathbf{E}\big[N_t^n\big] = \lim_{n\to\infty} p_n 2^n.$$

Using the Poisson approximation theorem (Theorem 3.7), we infer that, for any $l \in \mathbb{N}_0$,

$$\mathbf{P}[N_t = l] = \lim_{n\to\infty} \mathbf{P}\big[N_t^n = l\big] = \mathrm{Poi}_{\alpha t}\big(\{l\}\big).$$

Hence $\mathbf{P}_{N_t} = \mathrm{Poi}_{\alpha t}$. □

At this point, we still have to show that there are Poisson processes at all. We present a general two-step construction principle that will be used in a similar form later in Chapter 24 in a more general setting. In the first step, we determine the (random) number of jumps in $(0, 1]$. In the second step, we distribute these jumps uniformly and independently on $(0, 1]$. Strictly speaking, this gives the Poisson process only on the time interval $(0, 1]$, but it is clear how to move on: We perform the same procedure independently for each of the intervals $(1, 2]$, $(2, 3]$ and so on and then collect the jumps (see also Exercise 5.5.1).

Let $\alpha > 0$ and let L be a Poi_α random variable. Further, let $X_1, X_2, \ldots$ be independent random variables, that are uniformly distributed on $(0, 1]$, i.e., $X_k \sim \mathcal{U}_{(0,1]}$ for each k. We assume that $\{L, X_1, X_2, \ldots\}$ is an independent family of random variables. We now define $N = (N_t)_{t\in[0,1]}$ by

$$N_t := \sum_{l=1}^{L} \mathbb{1}_{(0,t]}(X_l) \quad \text{for } t \in [0, 1]. \tag{5.16}$$

Theorem 5.35 *The family N of random variables defined in* (5.16) *is a Poisson process with intensity α (and time set $[0, 1]$).*

Proof We have to show that the increments of N in finitely many pairwise disjoint intervals are independent and Poisson distributed. Hence let $m \in \mathbb{N}$ and $0 = t_0 < t_1 < \ldots < t_m = 1$. We use the abbreviations $p_i := t_i - t_{i-1}$ and $\lambda_i = \alpha \cdot (t_i - t_{i-1})$ and show that

$$(N_{t_i} - N_{t_{i-1}})_{i=1,\ldots,m} \quad \text{is independent} \tag{5.17}$$

and

$$N_{t_i} - N_{t_{i-1}} \sim \mathrm{Poi}_{\lambda_i} \quad \text{for all } i = 1, \ldots, m. \tag{5.18}$$

This is equivalent to showing that for each choice of $k_1, \ldots, k_m \in \mathbb{N}_0$, we have

$$\mathbf{P}[N_{t_i} - N_{t_{i-1}} = k_i \quad \text{for any } i = 1, \ldots, m] = \prod_{i=1}^{m} \left(e^{-\lambda_i} \frac{\lambda_i^{k_i}}{k_i!} \right). \tag{5.19}$$

Write

$$M_{n,i} := \#\{l \le n : t_{i-1} < X_l \le t_i\} = \sum_{l=1}^{n} \mathbb{1}_{(t_{i-1},t_i]}(X_l).$$

By Exercise 2.2.3, the vector $(M_{n,1}, \dots, M_{n,m})$ is multinomially distributed with parameters n and $p = (p_1, \dots, p_m)$. That is, if we assume $n := k_1 + \dots + k_m$, then

$$\mathbf{P}[M_{n,1} = k_1, \dots, M_{n,m} = k_m] = \frac{n!}{k_1! \dots k_m!} p_1^{k_1} \dots p_m^{k_m}.$$

In order to show (5.19), note that the event in (5.19) implies $L = n$ and that L and $(M_{n,1}, \dots, M_{n,m})$ are independent. Hence we have

$$\begin{aligned}
&\mathbf{P}[N_{t_i} - N_{t_{i-1}} = k_i \text{ for } i = 1, \dots, m] \\
&\quad = \mathbf{P}\big[\{N_{t_i} - N_{t_{i-1}} = k_i \text{ for } i = 1, \dots, m\} \cap \{L = n\}\big] \\
&\quad = \mathbf{P}\big[\{M_{n,1} = k_1, \dots, M_{n,m} = k_m\} \cap \{L = n\}\big] \\
&\quad = \mathbf{P}[M_{n,1} = k_1, \dots, M_{n,m} = k_m] \cdot \mathbf{P}[L = n] \\
&\quad = \frac{n!}{k_1! \dots k_m!} p_1^{k_1} \dots p_m^{k_m} e^{-\alpha} \frac{\alpha^n}{n!} \\
&\quad = \prod_{i=1}^{m} \left(e^{-\lambda_i} \frac{\lambda_i^{k_i}}{k_i!} \right).
\end{aligned}$$

□

We close this section by presenting a further, rather elementary and instructive construction of the Poisson process based on specifying the waiting times between the clicks of the Geiger counter, or, more formally, between the points of discontinuity of the map $t \mapsto N_t(\omega)$. At time s, what is the probability that we have to wait another t time units (or longer) for the next click? Since we modeled the clicks as a Poisson process with intensity α, this probability can easily be computed:

$$\mathbf{P}[N_{(s,s+t]} = 0] = e^{-\alpha t}.$$

Hence the waiting time for the next click is exponentially distributed with parameter α. Furthermore, the waiting times should be independent. We now take the waiting times as the starting point and, based on them, construct the Poisson process.

Let $W_1, W_2, \dots$ be an independent family of exponentially distributed random variables with parameter $\alpha > 0$; hence $\mathbf{P}[W_n > x] = e^{-\alpha x}$. We define

$$T_n := \sum_{k=1}^{n} W_k$$

and interpret W_n as the waiting time between the $(n-1)$th click and the nth click. T_n is the time of the nth click. Appealing to this intuition we define the number of clicks until time t by

$$N_t := \#\{n \in \mathbb{N}_0 : T_n \le t\}.$$

Hence

$$\{N_t = k\} = \{T_k \le t < T_{k+1}\}.$$

In particular, N_t is a random variable; that is, measurable.

Theorem 5.36 *The family* $(N_t, t \ge 0)$ *is a Poisson process with intensity* α.

Proof (We follow the proof in [59, Theorem 3.34].) We must show that for any $n \in \mathbb{N}$ and any sequence $0 = t_0 < t_1 < \ldots < t_n$, we have that $(N_{t_i} - N_{t_{i-1}}, i = 1, \ldots, n)$ is independent and $N_{t_i} - N_{t_{i-1}} \sim \mathrm{Poi}_{\alpha(t_i - t_{i-1})}$. We are well aware that it is not enough to show this for the case $n = 2$ only. However, the notational complications become overwhelming for $n \ge 3$, and the idea for general $n \in \mathbb{N}$ becomes clear in the case $n = 2$. Hence we restrict ourselves to the case $n = 2$.

Hence we show for $0 < s < t$ and $l, k \in \mathbb{N}_0$ that

$$\mathbf{P}[N_s = k, N_t - N_s = l] = \left(e^{-\alpha s}\frac{(\alpha s)^k}{k!}\right)\left(e^{-\alpha(t-s)}\frac{(\alpha(t-s))^l}{l!}\right). \tag{5.20}$$

This implies that N_s and $(N_t - N_s)$ are independent. Furthermore, by summing over $k \in \mathbb{N}_0$, this yields $N_t - N_s \sim \mathrm{Poi}_{\alpha(t-s)}$.

By Corollary 2.22, the distribution $\mathbf{P}_{(W_1,\ldots,W_{k+l+1})}$ has the density

$$x \mapsto \alpha^{k+l+1} e^{-\alpha S_{k+l+1}(x)},$$

where $S_n(x) := x_1 + \ldots + x_n$. It is sufficient to consider $l \ge 1$ since we get the $l = 0$ term from the fact that the probability measure has total mass one. Hence, let $l \ge 1$. We compute

$$\begin{aligned}
&\mathbf{P}[N_s = k, N_t - N_s = l] \\
&\quad = \mathbf{P}[T_k \le s < T_{k+1}, T_{k+l} \le t < T_{k+l+1}] \\
&\quad = \int_0^\infty \ldots \int_0^\infty dx_1 \ldots dx_{k+l+1} \\
&\qquad \times \alpha^{k+l+1} e^{-\alpha S_{k+l+1}(x)} \mathbb{1}_{\{S_k(x) \le s < S_{k+1}(x)\}} \mathbb{1}_{\{S_{k+l}(x) \le t < S_{k+l+1}(x)\}}.
\end{aligned}$$

Starting with x_{k+l+1}, we integrate successively. In the first step, substitute $z = S_{k+l+1}(x)$ to obtain

$$\int_0^\infty dx_{k+l+1}\, \alpha e^{-\alpha S_{k+l+1}(x)} \mathbb{1}_{\{S_{k+l+1}(x) > t\}} = \int_t^\infty dz\, \alpha e^{-\alpha z} = e^{-\alpha t}.$$

Now keep $x_1, \ldots, x_k$ fixed and substitute for the remaining variables by letting $y_1 = S_{k+1}(x) - s$, $y_2 = x_{k+2}, \ldots, y_l = x_{k+l}$ to obtain

$$\int_0^\infty \ldots \int_0^\infty dx_{k+1} \ldots dx_{k+l} \, \mathbb{1}_{\{s < S_{k+1}(x) \le S_{k+l} \le t\}}$$
$$= \int_0^\infty \ldots \int_0^\infty dy_1 \ldots dy_l \, \mathbb{1}_{\{y_1 + \ldots + y_l \le t-s\}} = \frac{(t-s)^l}{l!}.$$

(The last identity can be obtained, for example, by induction on l.) Now integrate the remaining variables $x_1, \ldots, x_k$ to get

$$\int_0^\infty \ldots \int_0^\infty dx_1 \ldots dx_k \, \mathbb{1}_{\{S_k(x) \le s\}} = \frac{s^k}{k!}.$$

In total, we have

$$\mathbf{P}[N_s = k, N_t - N_s = l] = e^{-\alpha t} \alpha^{k+l} \frac{s^k}{k!} \frac{(t-s)^l}{l!};$$

hence (5.20) holds. □

Exercise 5.5.1 Let L_n, X_k^n, $k, n \in \mathbb{N}$ be independent random variables with $L_n \sim \mathrm{Poi}_\alpha$ and $X_k^n \sim \mathcal{U}_{(n-1,n]}$ (the uniform distribution on $(n-1, n]$) for all $k, n \in \mathbb{N}$. Define

$$N_t := \#\big\{(k, n) \in \mathbb{N}^2 : k \le L_n \text{ and } X_k^n \le t\big\}.$$

Show that $(N_t)_{t \ge 0}$ is a Poisson process with intensity α.

Exercise 5.5.2 Let $T > 0$ and let $X_1, X_2, \ldots$ be i.i.d. random variables that are uniformly distributed on $[0, 1]$. Let

$$N := \max\{n \in \mathbb{N}_0 : X_1 + \ldots + X_n \le T\}$$

and compute $\mathbf{E}[N]$.