Probability

Lecture Notes

Tomasz Tkocz*

These lecture notes were written for some parts of the undergraduate course 21-325 *Probability* that I taught at Carnegie Mellon University in Spring 2018 and 2019.

Special thanks to Kai Wen Wang who has enormously helped prepare these notes.

*Carnegie Mellon University; ttkocz@math.cmu.edu

Contents

1 Expectation

The goal of this section is to define expectation of random variables and establish its basic properties. We shall only consider real-valued random variables. Recall that a function $X : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **random variable** if for every $x \in \mathbb{R}$, the preimage $\{X \le x\} = \{\omega \in \Omega,\ X(\omega) \le x\} = X^{-1}((-\infty, x])$ is an event (belongs to the sigma-field $\mathcal{F}$).

A random variable X is called **simple** if its image $X(\Omega)$ is a finite set, that is

$$X = \sum_{k=1}^{n} x_k \mathbf{1}_{A_k},$$

for some distinct $x_1, \ldots, x_n \in \mathbb{R}$ (values) and events $A_1, \ldots, A_n$ which form a partition of Ω (we have, $A_k = \{X = x_k\}$).

The **expectation** of the simple random variable X, denoted $\mathbb{E}X$, is defined as

$$\mathbb{E}X = \sum_{k=1}^{n} x_k \mathbb{P}(A_k).$$

The expectation of a nonnegative random variable X is defined as

$$\mathbb{E}X = \sup\{\mathbb{E}Z,\ Z \text{ is simple and } Z \le X\}.$$

Note that $\mathbb{E}X \ge 0$ because we can always take $Z = 0$. We can have $\mathbb{E}X = +\infty$ (for instance, for a discrete random variable X with $\mathbb{P}(X = k) = \frac{1}{k(k-1)}$, $k = 2, 3, \ldots$). For an arbitrary random variable X, we write

$$X = X^+ - X^-,$$

where

$$X^+ = \max\{X, 0\} = X\mathbf{1}_{\{X \ge 0\}}$$

is the positive part of X and

$$X^- = -\min\{X, 0\} = -X\mathbf{1}_{\{X \le 0\}}$$

is the negative part of X. These are nonnegative random variables and the expectation of X is defined as

$$\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$$

provided that at least one of the quantities $\mathbb{E}X^+$, $\mathbb{E}X^-$ is finite (to avoid $\infty - \infty$). We say that X is **integrable** if $\mathbb{E}|X| < \infty$. Since $|X| = X^+ + X^-$, we have that X is integrable if and only if $\mathbb{E}X^+ < \infty$ and $\mathbb{E}X^- < \infty$.

One of the desired properties of expectation is linearity. It of course holds for simple random variables.

1.1 Theorem. *Let X and Y be simple random variables. Then $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$.*

Proof. Let $X = \sum_{k=1}^{m} x_k \mathbf{1}_{A_k}$ and $Y = \sum_{l=1}^{n} y_l \mathbf{1}_{B_l}$ for some reals x_k, y_l and events A_k and B_l are such that the A_k partition Ω and the B_l partition Ω. Then the events $A_k \cap B_l$, $k \leq m$, $l \leq n$ partition Ω and

$$X + Y = \sum_{k \leq m, l \leq n} (x_k + y_l) \mathbf{1}_{A_k \cap B_l}.$$

This is a simple random variable with

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_{k \leq m, l \leq n} (x_k + y_l) \mathbb{P}(A_k \cap B_l) \\
&= \sum_{k \leq m, l \leq n} x_k \mathbb{P}(A_k \cap B_l) + \sum_{k \leq m, l \leq n} y_l \mathbb{P}(A_k \cap B_l) \\
&= \sum_{k \leq m} x_k \sum_{l \leq n} \mathbb{P}(A_k \cap B_l) + \sum_{l \leq n} y_l \sum_{k \leq m} \mathbb{P}(A_k \cap B_l) \\
&= \sum_{k \leq m} x_k \mathbb{P}\left(A_k \cap \bigcup_{l \leq n} B_l\right) + \sum_{l \leq n} y_l \mathbb{P}\left(\bigcup_{k \leq m} A_k \cap B_l\right) \\
&= \sum_{k \leq m} x_k \mathbb{P}(A_k) + \sum_{l \leq n} y_l \mathbb{P}(B_l),
\end{aligned}$$

which is $\mathbb{E}X + \mathbb{E}Y$ and this finishes the proof. □

1.1 Nonnegative random variables

Our main goal is to prove linearity of expectation. We first establish a few basic properties of expectation for nonnegative random variables.

1.2 Theorem. *Let X and Y be nonnegative random variables. We have*

(a) if $X \leq Y$, then $\mathbb{E}X \leq \mathbb{E}Y$,

(b) for $a \geq 0$, $\mathbb{E}(a + X) = a + \mathbb{E}X$ and $E(aX) = a\mathbb{E}X$,

(c) if $\mathbb{E}X = 0$, then $X = 0$ a.s. (i.e. $\mathbb{P}(X = 0) = 1$)

(d) if A and B are events such that $A \subset B$, then $\mathbb{E}X\mathbf{1}_A \leq \mathbb{E}X\mathbf{1}_B$.

Proof. (a) Let $\varepsilon > 0$. By definition, there is a simple random variable Z such that $Z \leq X$ and $\mathbb{E}Z > \mathbb{E}X - \varepsilon$. Then also $Z \leq Y$, so by the definition of $\mathbb{E}Y$, we have $\mathbb{E}Z \leq \mathbb{E}Y$. Thus $\mathbb{E}X - \varepsilon < \mathbb{E}Y$. Sending ε to 0 finishes the argument.

(b) For a simple random variable Z, clearly $\mathbb{E}(a + Z) = a + \mathbb{E}Z$ and $\mathbb{E}(aZ) = a\mathbb{E}Z$. It remains to follow the proof of (a).

(c) For $n \geq 1$, we have $X \geq X\mathbf{1}_{\{X \geq 1/n\}} \geq \frac{1}{n}\mathbf{1}_{\{X \geq 1/n\}}$, so by (a) we get

$$0 = \mathbb{E}X \geq \mathbb{E}\frac{1}{n}\mathbf{1}_{\{X \geq 1/n\}} = \frac{1}{n}\mathbb{P}\left(X \geq 1/n\right),$$

thus $\mathbb{P}\left(X \geq 1/n\right) = 0$, so

$$\mathbb{P}\left(X > 0\right) = \mathbb{P}\left(\bigcap_{n \geq 1} \{X \geq 1/n\}\right) = \lim_{n \to \infty} \mathbb{P}\left(X \geq 1/n\right) = 0.$$

(d) follows immediately from (a). □

The following lemma gives a way to approximate nonnegative random variables with monotone sequences of simple ones.

1.3 Lemma. *If X is a nonnegative random variable, then there is a sequence (Z_n) of nonnegative simple random variables such that for every $\omega \in \Omega$, $Z_n(\omega) \leq Z_{n+1}(\omega)$ and $Z_n(\omega) \xrightarrow[n \to \infty]{} X(\omega)$.*

Proof. Define

$$Z_n = \sum_{k=1}^{n \cdot 2^n} \frac{k-1}{2^n} \mathbf{1}_{\{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\}} + n\mathbf{1}_{\{X \geq n\}}.$$

Fix $\omega \in \Omega$. Then $Z_n(\omega)$ is a nondecreasing sequence (check!). Since $n > X(\omega)$ for large enough n, we have for such n that $0 \leq X(\omega) - Z_n(\omega) \leq 2^{-n}$. □

The following is a very important and useful tool allowing to exchange the order of taking the limit and expectation for monotone sequences.

1.4 Theorem (Lebesgue's monotone convergence theorem)**.** *If X_n is a sequence of nonnegative random variables such that $X_n \leq X_{n+1}$ and $X_n \xrightarrow[n \to \infty]{} X$, then*

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

Proof. By 1.2 (a), $\mathbb{E}X_n \leq \mathbb{E}X_{n+1}$ and $\mathbb{E}X_n \leq \mathbb{E}X$, so $\lim_n \mathbb{E}X_n$ exists and is less than or equal to $\mathbb{E}X$. It remains to show that $\mathbb{E}X \leq \lim_n \mathbb{E}X_n$. Take a simple random variable Z such that $0 \leq Z \leq X$, with the largest value say K. Observe that for every $n \geq 1$ and $\varepsilon > 0$,

$$Z \leq (X_n + \varepsilon)\mathbf{1}_{\{Z < X_n + \varepsilon\}} + K\mathbf{1}_{\{Z \geq X_n + \varepsilon\}}. \tag{1.1}$$

Claim. For nonnegative random variables X, Y and an event A, we have

$$\mathbb{E}(X\mathbf{1}_A + Y\mathbf{1}_{A^c}) \leq \mathbb{E}X\mathbf{1}_A + \mathbb{E}Y\mathbf{1}_{A^c}.$$

Proof of the claim. Fix $\varepsilon > 0$. Take a simple random variable Z such that $Z \leq X\mathbf{1}_A + Y\mathbf{1}_{A^c}$ and $\mathbb{E}Z > \mathbb{E}(X\mathbf{1}_A + Y\mathbf{1}_{A^c}) - \varepsilon$. Note that

$$Z\mathbf{1}_A \leq X\mathbf{1}_A \qquad \text{and} \qquad Z\mathbf{1}_{A^c} \leq Y\mathbf{1}_{A^c}.$$

Thus by 1.2 (a),

$$\mathbb{E}Z\mathbf{1}_A \le \mathbb{E}X\mathbf{1}_A \qquad \text{and} \qquad \mathbb{E}Z\mathbf{1}_{A^c} \le \mathbb{E}Y\mathbf{1}_{A^c}.$$

Adding these two inequalities together and using that $\mathbb{E}Z\mathbf{1}_A + \mathbb{E}Z\mathbf{1}_{A^c} = \mathbb{E}Z$, which follows from linearity of expectation for simple random variables (Theorem 1.1), we get

$$\mathbb{E}(X\mathbf{1}_A + Y\mathbf{1}_{A^c}) - \varepsilon < \mathbb{E}Z \le \mathbb{E}X\mathbf{1}_A + \mathbb{E}Y\mathbf{1}_{A^c}.$$

Sending $\varepsilon \to 0$ finishes the argument. □

Applying the claim to (1.1), we obtain

$$\mathbb{E}Z \le \mathbb{E}X_n + \varepsilon + K\mathbb{P}\left(Z \ge X_n + \varepsilon\right).$$

The events $\{Z \ge X_n + \varepsilon\}$ form a decreasing family (because $X_n \le X_{n+1}$ and their intersection is $\{Z \ge X + \varepsilon\} = \varnothing$ (because $X_n \to X$ and $Z \le X$). Therefore taking $n \to \infty$ in the last inequality gives

$$\mathbb{E}Z \le \lim_n \mathbb{E}X_n + \varepsilon.$$

Taking the supremum over simple random variables $Z \le X$ gives

$$\mathbb{E}X \le \lim_n \mathbb{E}X_n + \varepsilon.$$

Letting $\varepsilon \to 0$, we finish the proof. □

As a corollary we obtain a result about the limit inferior of nonnegative random variables and its expectation.

1.5 Theorem (Fatou's lemma)**.** *If $X_1, X_2, \ldots$ are nonnegative random variables, then*

$$\mathbb{E}\liminf_{n\to\infty} X_n \le \liminf_{n\to\infty} \mathbb{E}X_n.$$

Proof. Let $Y_n = \inf_{k\ge n} X_k$. Then this is a nondecreasing sequence which converges to $\liminf_{n\to\infty} X_n$ and $Y_n \le X_n$. Note that

$$\liminf_{n\to\infty} \mathbb{E}X_n \ge \liminf_{n\to\infty} \mathbb{E}Y_n = \lim_{n\to\infty} \mathbb{E}Y_n,$$

where the last equality holds because the sequence $\mathbb{E}Y_n$, as nondecreasing, is convergent. By Lebesgue's monotone converge theorem,

$$\lim_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}\left(\lim_{n\to\infty} Y_n\right) = \mathbb{E}\liminf_{n\to\infty} X_n,$$

which in view of the previous inequality finishes the proof. □

We are ready to prove linearity of expectation for nonnegative random variables.

1.6 Theorem. *Let X and Y be nonnegative random variables. Then*

$$\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y.$$

Proof. By Lemma 1.3, there are nondecreasing sequences (X_n) and (Y_n) of nonnegative simple random variables such that $X_n \to X$ and $Y_n \to Y$. Then the sequence $(X_n + Y_n)$ is also monotone and $X_n + Y_n \to X + Y$. By Theorem 1.1,

$$\mathbb{E}(X_n + Y_n) = \mathbb{E}X_n + \mathbb{E}Y_n.$$

Letting $n \to \infty$, by the virtue of Lebesgue's monotone convergence theorem, we get in the limit $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$. □

1.2 General random variables

Key properties of expectation for general random variables are contained in our next theorem.

1.7 Theorem. *If X and Y are integrable random variables, then*

(a) $X + Y$ is integrable and $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$,

(b) $\mathbb{E}(aX) = a\mathbb{E}X$ for every $a \in \mathbb{R}$,

(c) if $X \le Y$, then $\mathbb{E}X \le \mathbb{E}Y$,

(d) $|\mathbb{E}X| \le \mathbb{E}|X|$.

Proof. (a) By the triangle inequality Theorem 1.2 (a) and Theorem 1.6,

$$\mathbb{E}|X+Y| \le \mathbb{E}(|X| + |Y|) = \mathbb{E}|X| + \mathbb{E}|Y|$$

and the right hand side is finite by the assumption, thus $X + Y$ is integrable.

To show the linearity, write $X + Y$ in two different ways

$$(X+Y)^+ - (X+Y)^- = X + Y = X^+ - X^- + Y^+ - Y^-,$$

rearrange

$$(X+Y)^+ + X^- + Y^- = (X+Y)^- + X^+ + Y^+,$$

to be able to use the linearity of expectation for nonnegative random variables (Theorem 1.6) and get

$$\mathbb{E}(X+Y)^+ + \mathbb{E}X^- + \mathbb{E}Y^- = \mathbb{E}(X+Y)^- + \mathbb{E}X^+ + \mathbb{E}Y^+,$$

which rearranged again gives $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$.

(b) We leave this as an exercise.

(c) Note that $X \le Y$ is equivalent to saying that $X^+ \le Y^+$ and $X^- \ge Y^-$ (if $X = X^+$, then $X \le Y$ implies that $Y = Y^+$, hence $X^+ \le Y^+$; similarly, if $Y = -Y^-$, then $X \le Y$ implies that $X = -X^-$, hence $X^- \ge Y^-$). It remains to use Theorem 1.2 (a).

(d) Since $-|X| \le X \le |X|$, by (c) we get $-\mathbb{E}|X| \le \mathbb{E}X \le \mathbb{E}|X|$, that is $|\mathbb{E}X| \le \mathbb{E}|X|$. $\square$

1.3 Discrete and continuous random variables

We show that the general definition of expectation we made agrees with the ad hoc definitions we made for discrete and continuous random variables in terms of their probability mass and probability density functions. We begin with discrete random variables.

1.8 Theorem. *Let X be a discrete random variable taking values $\ldots < x_{-2} < x_{-1} < x_0 \le 0 < x_1 < x_2 < \ldots$. Then $\mathbb{E}X$ exists if and only if at least one of the series $\sum_{k=-\infty}^{0} x_k \mathbb{P}(X = x_k)$, $\sum_{k=1}^{\infty} x_k \mathbb{P}(X = x_k)$ converges and then*

$$\mathbb{E}X = \sum_{k=-\infty}^{\infty} x_k \mathbb{P}(X = x_k).$$

Proof. It suffices to show that for a nonnegative discrete random variable X taking values $0 \le x_1 < x_2 < \ldots$, we have $\mathbb{E}X = \mu$ with $\mu = \sum_{k=1}^{\infty} x_k \mathbb{P}(X = x_k) \in [0, \infty]$. Let Z be a simple nonnegative random variable such that $Z \le X$. First we show that then $\mathbb{E}Z \le \mu$, which by taking the supremum over Z gives $\mathbb{E}X \le \mu$. Suppose $Z = \sum_{k=1}^{n} z_k \mathbf{1}_{A_k}$ for some events A_k which partition Ω and nonnegative numbers z_k. Denote $B_k = \{X = x_k\}$. Observe that the condition $Z \le X$ for $\omega \in A_k$ gives

$$z_k \le \min_{\omega \in A_k} X(\omega) = \min_{j : A_k \cap B_j \neq \varnothing} x_j.$$

Since the B_j partition Ω, we have $\mathbb{P}(A_k) = \sum_{j : A_k \cap B_j \neq \varnothing} \mathbb{P}(A_k \cap B_j)$. Consequently,

$$\begin{aligned}
\mathbb{E}Z = \sum_{k=1}^{n} z_k \mathbb{P}(A_k) &= \sum_{k=1}^{n} z_k \sum_{j : A_k \cap B_j \neq \varnothing} \mathbb{P}(A_k \cap B_j) \\
&\le \sum_{k=1}^{n} \min_{j : A_k \cap B_j \neq \varnothing} x_j \sum_{j : A_k \cap B_j \neq \varnothing} \mathbb{P}(A_k \cap B_j) \\
&\le \sum_{k=1}^{n} \sum_{j : A_k \cap B_j \neq \varnothing} x_j \mathbb{P}(A_k \cap B_j) \\
&\le \sum_{k=1}^{n} \sum_{j \ge 1} x_j \mathbb{P}(A_k \cap B_j) \\
&= \sum_{j \ge 1} x_j \mathbb{P}(B_j) = \mu
\end{aligned}$$

(in the second last equality we exchanged the order of summation and used that the A_k partition Ω). This shows $\mathbb{E}X \le \mu$.

To obtain the opposite inequality, for every $\varepsilon > 0$, we have to find a simple random variable Z such that $\mathbb{E}Z > \mu - \varepsilon$. Since $\mu = \lim_{n\to\infty} \sum_{k=1}^{n} x_k \mathbb{P}(X = x_k)$, it suffices to take $Z = \sum_{k=1}^{n} x_k \mathbf{1}_{\{X=x_k\}}$ for n large enough. □

1.9 Remark. Under the assumptions of the previous theorem, for every function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}g(X) = \sum g(x_k)\mathbb{P}(X = x_k).$$

Indeed, it suffices to apply the theorem to the random variable $g(X)$ (which is also discrete).

Now we settle the continuous case.

1.10 Theorem. *Let X be a continuous random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a Borel function. Then*

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x.$$

In particular,

$$\mathbb{E}X = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x.$$

Proof. It is enough to consider nonnegative functions g. When $g = \mathbf{1}_A$ for a Borel subset A of $\mathbb{R}$, we have $\mathbb{E}g = \mathbb{E}\mathbf{1}_A = \mathbb{P}(X \in A) = \int_A f(x)\mathrm{d}x = \int \mathbf{1}_A(x)f(x)\mathrm{d}x = \int g(x)f(x)\mathrm{d}x$. By linearity, the identity also holds for g being any linear combination of indicator functions, that is for g being a simple function. By Lemma 1.3, any Borel function g is a pointwise limit of a nondecreasing sequence of simple functions g_n. Then the random variable $g(X)$ is the limit of $g_n(X)$ (which is also a monotone sequence). Thus we can apply Lebesgue's monotone convergence theorem to both sides, which finishes the argument (to argue about $\int g_n f \to \int gf$, we in fact use a version of Lebesgue's theorem for Lebesgue's integrals, which by their construction, has the same proof as Theorem 1.4). □

1.4 Lebesgue's dominated convergence theorem

We finish this chapter with one more limit theorem, quite useful in various applications; we also show one of them.

1.11 Theorem (Lebesgue's dominated convergence theorem)**.** *If (X_n) is a sequence of random variables and X is a random variable such that for every $\omega \in \Omega$, we have $X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega)$ and there is an integrable random variable Y such that $|X_n| \le Y$, then*

$$\mathbb{E}|X_n - X| \xrightarrow[n\to\infty]{} 0.$$

In particular,

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

Proof. Since $|X_n| \le Y$, taking $n \to \infty$ yields $|X| \le Y$. In particular, X is integrable as well. By the triangle inequality,

$$|X_n - X| \le 2Y$$

and Fatou's lemma (Theorem 1.5) gives

$$\begin{aligned}\mathbb{E}(2Y) = \mathbb{E}\liminf(2Y - |X_n - X|) &\le \liminf \mathbb{E}(2Y - |X_n - X|) \\ &= 2\mathbb{E}Y - \limsup \mathbb{E}|X_n - X|.\end{aligned}$$

As a result, $\limsup \mathbb{E}|X_n - X| \le 0$, so

$$\mathbb{E}|X_n - X| \xrightarrow[n\to\infty]{} 0.$$

In particular, since by Theorem 1.7 (d),

$$|\mathbb{E}(X_n - X)| \le \mathbb{E}|X_n - X|,$$

we get that the left hand side goes to 0, that is $\mathbb{E}X_n \to \mathbb{E}X$. □

As application, we show a necessary condition for the expectation of a nonegative random variable to be finite, in terms of the rate of decay of its tail function. To motive this condition, recall the following formula for the expectation: for a nonnegative random variable X, we have

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t. \tag{1.2}$$

This can be justified as follows,

$$\mathbb{E}X = \mathbb{E}\int_0^X \mathrm{d}t = \mathbb{E}\int_0^\infty \mathbf{1}_{t<X}\,\mathrm{d}t = \int_0^\infty \mathbb{E}\mathbf{1}_{t<X}\,\mathrm{d}t = \int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t,$$

where the second last equality follows from Fubini's theorem (in this case, essentially linearity of expectation; it would be clear, if the integral $\int_0^\infty$ was a finite sum).

1.12 Remark. In particular, formula (1.2) justifies a desired fact that the expectation is determined by distribution, that is if X and Y are random variables with the same distribution (the same cumulative distribution functions), then $\mathbb{E}X = \mathbb{E}Y$.

Since the integral $\int_1^\infty \frac{\mathrm{d}t}{t}$ is $+\infty$, in view of the above formula we can suspect that if $\mathbb{E}X < \infty$, then $\mathbb{P}(X > t)$ goes to 0 "faster" than $\frac{1}{t}$. This is indeed true, and to show it, we use Lebesgue's dominated convergence theorem.

1.13 Theorem. *If X is a nonnegative random variable such that $\mathbb{E}X < \infty$, then*

$$t\mathbb{P}(X > t) \xrightarrow[t\to\infty]{} 0.$$

Proof. Fix a sequence t_n going to ∞ as $n \to \infty$. Let $X_n = t_n \mathbf{1}_{\{X > t_n\}}$. We have

$$t_n \mathbb{P}\left(X_n > t_n\right) = t_n \mathbb{E}\mathbf{1}_{\{X > t_n\}} = \mathbb{E}X_n.$$

We want to show that $\mathbb{E}X_n \xrightarrow[n\to\infty]{} 0$. Since $X_n \xrightarrow[n\to\infty]{} 0$ at each ω and $X_n \leq X\mathbf{1}_{\{X > t_n\}} \leq X$, which means that the sequence (X_n) is pointwise upperbounded by the integrable random variable X, by Lebesgue's dominated converge theorem, $\lim \mathbb{E}X_n = \mathbb{E}(\lim_n X_n) = 0$. □

2 Inequalities

2.1 Basic probabilistic inequalities

One of the simplest and very useful probabilistic inequalities is a tail bound by expectation: the so-called Chebyshev's inequality.

2.1 Theorem (Chebyshev's inequality)**.** *If X is a nonnegative random variable, then for every $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{1}{t}\mathbb{E}X.$$

Proof. Since $X \geq X\mathbf{1}_{\{X \geq t\}} \geq t\mathbf{1}_{\{X \geq t\}}$, taking the expectation yields

$$\mathbb{E}X \geq \mathbb{E}t\mathbf{1}_{\{X \geq t\}} = t\mathbb{P}(X \geq t).$$

□

There are several variants, easily deduced from Chebyshev's inequality by monotonicity of certain functions. For a nonnegative random variable X and $t > 0$, using the power function x^p, $p > 0$, we get

$$\mathbb{P}(X \geq t) = \mathbb{P}(X^p \geq t^p) \leq \frac{1}{t^p}\mathbb{E}X^p. \tag{2.1}$$

For a real-valued random variable X, every $t \in \mathbb{R}$ and $\lambda > 0$, using the exponential function $e^{\lambda x}$, we have

$$\mathbb{P}(X \geq t) = \mathbb{P}(\lambda X \geq \lambda t) \leq \frac{1}{e^{\lambda t}}\mathbb{E}e^{\lambda X}. \tag{2.2}$$

For a real-valued random variable X, every $t \in \mathbb{R}$, using the square function x^2 and variance, we have

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{1}{t^2}\mathbb{E}|X - \mathbb{E}X|^2 = \frac{1}{t^2}\operatorname{Var}(X). \tag{2.3}$$

Our next inequality, the so-called Hölder's inequality, is a very effective inequality used to factor out the expectation of a product

2.2 Theorem (Hölder's inequality)**.** *Let $p, q \geq 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$ (when $p = 1$, then $q = \infty$). For random variables X and Y, we have*

$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}.$$

In particular, when $p = q = 2$, this gives the Cauchy-Schwarz inequality

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}|X|^2}\sqrt{\mathbb{E}|Y|^2}.$$

Proof. The key ingredient is an elementary inequality for numbers.

Claim. For $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$ and $x, y \geq 0$, we have

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}.$$

Proof. By the concavity of the log function, we have

$$\log\left(\frac{x^p}{p}+\frac{y^q}{q}\right)\geq\frac{1}{p}\log x^p+\frac{1}{q}\log y^q=\log xy.$$

□

Setting $x=\frac{|X|^p}{(\mathbb{E}|X|^p)^{1/p}}$, $y=\frac{|Y|^q}{(\mathbb{E}|Y|^q)^{1/q}}$, taking the expectation and simplifying yields the desired inequality. □

Recall that for $p\neq 0$, the pth moment of a random variable X, denoted $\|X\|_p$, is defined as

$$\|X\|_p=\left(\mathbb{E}|X|^p\right)^{1/p}.$$

Hölder's inequality gives the following helpful variational formula for $p\geq 1$.

$$\|X\|_p=\sup\{\mathbb{E}XY,\ Y \text{ is a random variable with } \mathbb{E}|Y|^q\leq 1\},\tag{2.4}$$

where $\frac{1}{p}+\frac{1}{q}=1$. To see that the supremum does not exceed the pth moment, simply apply Theorem 2.2. To see the opposite inequality, consider $Y=\operatorname{sgn}(X)|X|^{p-1}\|X\|_p^{-p/q}$. Then $\mathbb{E}XY=\|X\|_p$, so in fact we can write "max "instead of "sup "in (2.4). Using this linearisation, we can effortlessly establish the triangle inequality for the pth moment, the so-called Minkowski's inequality.

2.3 Theorem (Minkowski's ineuality)**.** *Let $p\geq 1$. Let X and Y be random variables. Then*

$$\|X+Y\|_p\leq\|X\|_p+\|Y\|_p.$$

Proof. Invoking (2.4),

$$\|X+Y\|_p=\sup\{\mathbb{E}(X+Y)Z,\ \mathbb{E}|Z|^q\leq 1\}.$$

By linearity, $\mathbb{E}(X+Y)Z=\mathbb{E}XZ+\mathbb{E}YZ$. Using that $\sup\{f+g\}\leq\sup f+\sup g$ and applying again (2.4) finishes the proof. □

2.4 Remark. For every $0<p<1$ Minkowski's inequality fails (for instance, take X and Y to be i.i.d. $\mathrm{Ber}(\alpha)$). Let us derive its analogue. Observe that for $0<p<1$ and every real numbers x, y, we have

$$|x+y|^p\leq|x|^p+|y|^p.\tag{2.5}$$

If $x+y=0$, the inequality is trivial. Otherwise, note that $|t|^p\geq|t|$ for $|t|\leq 1$, so using this and the triangle inequality yields

$$\left(\frac{|x|}{|x+y|}\right)^p+\left(\frac{|y|}{|x+y|}\right)^p\geq\frac{|x|}{|x+y|}+\frac{|y|}{|x+y|}=\frac{|x|+|y|}{|x+y|}\geq\frac{|x+y|}{|x+y|}=1.$$

Given two random variables, applying (2.5) for $x = X(\omega)$, $y = Y(\omega)$ and taking the expectation gives

$$\mathbb{E}|X+Y|^p \le \mathbb{E}|X|^p + \mathbb{E}|Y|^p, \qquad p \in (0,1]. \tag{2.6}$$

In other words,

$$\|X+Y\|_p^p \le \|X\|_p^p + \|Y\|_p^p, \qquad p \in (0,1]. \tag{2.7}$$

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $p > 0$ define the L_p **space** of random variables having finite pth moment,

$$L_p = L_p(\Omega, \mathcal{F}, \mathbb{P}) = \{X : \Omega \to \mathbb{R}, X \text{ is a random variable such that } \mathbb{E}|X|^p < \infty\}.$$

Minkowski's inequality and Remark 2.4 assert that L_p, $p > 0$ is a linear space. Moreover, for $p \ge 1$, the pth moment $\|\cdot\|_p$ is a norm on L_p meaning that

1) $\|X\|_p = 0$ if and only if $X = 0$ a.s.

2) $\|\lambda X\|_p = |\lambda|\|X\|_p$, $X \in L_p$, $\lambda \in \mathbb{R}$ (it is homogeneous)

3) $\|X+Y\|_p \le \|X\|_p + \|Y\|_p$ (it satisfies the triangle inequality)

(property 1) follows from Theorem 1.2 (c)). Consequently, $d(X,Y) = \|X-Y\|_p$ is a metric on L_p and is used to measure how close random variables are.

Another general and helpful inequality is about convex functions. Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is convex if $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$ for every $\lambda \in [0,1]$ and $x, y \in \mathbb{R}$. By induction, this can be extended to

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \le \sum_{i=1}^n \lambda_i f(x_i)$$

for every $\lambda_1, \dots, \lambda_n \ge$ such that $\sum_{i=1}^n \lambda_i = 1$ and every $x_1, \dots, x_n \in \mathbb{R}$. The weights λ_i can of course be interpreted in probabilistic terms: if X is a random variable taking the value x_i with probability λ_i, then $\sum \lambda_i x_i = \mathbb{E}X$, whereas $\sum \lambda_i f(x_i) = \mathbb{E}f(X)$, so we have

$$f(\mathbb{E}X) \le \mathbb{E}f(X).$$

This generalises to arbitrary random variables and is called Jensen's inequality.

2.5 Theorem (Jensen's inequality)**.** *If $f : \mathbb{R} \to \mathbb{R}$ is a convex function and X is a random variable such that both $\mathbb{E}X$ and $\mathbb{E}f(X)$ exist, then*

$$f(\mathbb{E}X) \le \mathbb{E}f(X).$$

Proof. Suppose f is differentiable. Then by convexity, a tangent line at x_0 is below the graph, so

$$f(x) \ge f(x_0) + f'(x_0)(x - x_0)$$

(which holds for every x_0 and x). We set $x = X$, $x_0 = \mathbb{E}X$ and take the expectation of both sides to get

$$\mathbb{E}f(X) \geq \mathbb{E}\left[f(\mathbb{E}X) + f'(\mathbb{E}X)(X - \mathbb{E}X)\right] = f(\mathbb{E}X) + f'(\mathbb{E}X)\mathbb{E}(X - \mathbb{E}X) = f(\mathbb{E}X).$$

If f is not differentiable, this argument requires more work, but there is another concise argument using the fact that a convex function is a pointwise supremum of a family of linear functions – we skip the details) □

2.6 Example. Let $0 < p < q$. Take $r = \frac{q}{p}$ and $f(x) = |x|^r$ which is convex. Thus for a random variable X which is in L_q, using Jensen's inequality, we have

$$\mathbb{E}|X|^q = \mathbb{E}f(|X|^p) \geq f(\mathbb{E}|X|^p) = (\mathbb{E}|X|^p)^{q/p},$$

equivalently,

$$\|X\|_q \geq \|X\|_p.$$

In other words, the function $p \mapsto \|X\|_p$ of moments of the random variable X is nondecreasing.

We finish this section with a concentration bound for weighted sums of independent random signs. It is a nice application of exponential Chebyshev's inequality (2.2) and illustrates how it takes advantage of independence.

2.7 Theorem (Bernstein's inequality)**.** *Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ be independent random signs and let $a_1, a_2, \ldots, a_n$ be real numbers. Then for every $tt \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i\varepsilon_i\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right\}.$$

Proof. Let $S = \sum a_i\varepsilon_i$. Note that S is a symmetric random variable, that is S has the same distribution as $-S$. Thus

$$\mathbb{P}\left(|S| \geq t\right) = \mathbb{P}\left(\{S \geq t\} \cup \{-S > t\}\right) = \mathbb{P}\left(S \geq t\right) + \mathbb{P}\left(-S \geq t\right) = 2\mathbb{P}\left(S \geq t\right).$$

Exponential Chebyshev's inequality (2.2) for every $\lambda > 0$ yields

$$\mathbb{P}\left(S \geq t\right) = \mathbb{P}\left(e^{\lambda S} \geq e^{\lambda t}\right) \leq e^{-\lambda t}\mathbb{E}e^{\lambda S}.$$

The last expectation can be computed thanks to independence,

$$\mathbb{E}e^{\lambda S} = \mathbb{E}\prod_{i=1}^n e^{\lambda a_i\varepsilon_i} = \prod_{i=1}^n \mathbb{E}e^{\lambda a_i\varepsilon_i} = \prod_{i=1}^n \frac{e^{\lambda a_i} + e^{-\lambda a_i}}{2}.$$

Using an elementary inequality $\frac{e^x + e^{-x}}{2} \leq e^{x^2/2}$, $x \in \mathbb{R}$, we get

$$\mathbb{E}e^{\lambda S} \leq \prod_{i=1}^n e^{\lambda^2 a_i^2/2} = e^{\lambda^2\sigma^2/2},$$

where $\sigma^2 = \sum a_i^2$. Putting these together yields

$$\mathbb{P}(S \geq t) \leq e^{-\lambda t + \lambda^2 \sigma^2/2}.$$

This holds for every $\lambda > 0$. Choose λ such that the right hand side is as small as possible, that is $\lambda = t/\sigma^2$, to get the assertion. □

2.2 Application in analysis: Weierstrass' theorem

Fix $p \in [0,1]$. Let $\delta_1, \ldots, \delta_n$ be i.i.d. Ber(p) random variables. Let $S_n = \delta_1 + \ldots + \delta_n$. Using similar arguments as in Theorem 2.7, it can be shown that $\mathbb{P}(|S_n - \mathbb{E}S_n| > nt) \leq 2e^{-nt^2/4}$, that is the probability that $\frac{S_n}{n}$ deviates from its expectation $\mathbb{E}\frac{S_n}{n} = p$ by more than a fixed t is exponentially small in n. This means that for large n, $\frac{S_n}{n}$ is approximately p (we say $\frac{S_n}{n}$ concentrates around its expectation p). Guided by this observation, we can give a constructive proof of Weierstrass' theorem about uniform approximation of continuous functions with polynomials.

2.8 Theorem (Weierstrass)**.** *Let $f : [0,1] \to \mathbb{R}$ be a continuous function. For every $\varepsilon > 0$, there is a polynomial Q such that $|f - Q| < \varepsilon$ on $[0,1]$.*

Proof. Fix $p \in [0,1]$ and let $S_n \sim \mathrm{Bin}(p,n)$ be as above. Define

$$Q(p) = \mathbb{E}f\left(\frac{S_n}{n}\right) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) p^k (1-p)^{n-k}$$

As explained, $\frac{S_n}{n}$ concentrates around p for large n, thus it is reasonable to hope that $Q(p)$ will be approximately $\mathbb{E}f(p) = f(p)$. Note that Q as a function of p is a polynomial of degree n.

Now we show that for every $\varepsilon > 0$, there is n such that $|Q(p) - f(p)| < \varepsilon$ for every $p \in [0,1]$. We have,

$$\begin{aligned}
|Q(p) - f(p)| = \left|\mathbb{E}f\left(\frac{S_n}{n}\right) - f(p)\right| &\leq \mathbb{E}\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \\
&= \mathbb{E}\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \mathbf{1}_{\{|\frac{S_n}{n} - p| \geq n^{-1/4}\}} \\
&\quad + \mathbb{E}\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \mathbf{1}_{\{|\frac{S_n}{n} - p| < n^{-1/4}\}}.
\end{aligned}$$

The function f, as continuous on $[0,1]$, is bounded on $[0,1]$, say $|f| \leq M$. Using this and Chebyshev's inequality (2.3), we get a good estimate for the first term,

$$\begin{aligned}
\mathbb{E}\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \mathbf{1}_{\{|\frac{S_n}{n} - p| \geq n^{-1/4}\}} &\leq 2M\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq n^{-1/4}\right) \\
&= 2M\mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}\frac{S_n}{n}\right| \geq n^{-1/4}\right) \\
&\leq 2M\frac{1}{n^{-1/2}} \operatorname{Var}\left(\frac{S_n}{n}\right) \\
&= 2Mn^{1/2}\frac{np(1-p)}{n^2} \leq \frac{M}{2n^{1/2}}.
\end{aligned}$$

(in the last inequality we used $p(1-p) \le \frac{1}{4}$). This is less than $\varepsilon/2$ for n large enough.

To bound the second term, we use that f is uniformly continuous on $[0,1]$, that is, there is δ such that $|f(x) - f(y)| < \varepsilon/2$ for all $x, y \in [0,1]$ such that $|x-y| < \delta$. For n large, such that $n^{-1/4} < \delta$, we thus get

$$\mathbb{E}\left| f\left(\frac{S_n}{n}\right) - f(p)\right| \mathbf{1}_{\{|\frac{S_n}{n}-p|<n^{-1/4}\}} < \mathbb{E}\frac{\varepsilon}{2}\mathbf{1}_{\{|\frac{S_n}{n}-p|<n^{-1/4}\}} \le \frac{\varepsilon}{2}.$$

Combined with the previous bound, we get $|Q - f| < \varepsilon/2 + \varepsilon/2 = \varepsilon$. □

2.3 Application in combinatorics: 1st and 2nd moment method

Simple probabilistic inequalities are of use in combinatorics. We mention two such applications in situations when we want to show that a nonnegative integer-valued random variable (e.g. counting the number of some combinatorial objects) is zero with high probability (1st moment method) and complementary, is positive with high probability (2nd moment method).

Consider the following problem: we put m balls independently uniformly at random into n boxes. How large does m have to be (with respect to n) so that with high probability (w.h.p.) there are no empty boxes (that is, with probability going to 1 as n goes to ∞)?

1st moment method

Claim. If X is a nonnegative integer-valued random variable, then

$$\mathbb{P}(X > 0) \le \mathbb{E}X.$$

This follows immediately from Chebyshev's inequality (Theorem 2.1) because $\{X > 0\} = \{X \ge 1\}$.

Let X be the number of empty boxes after putting m balls independently uniformly at random into n boxes,

$$X = \sum_{i=1}^{n} X_i,$$

where X_i is the indicator random variable of the event "ith box being empty". Clearly,

$$\mathbb{E}X = \sum_{i=1}^{n} \mathbb{E}X_i = n\mathbb{E}X_1 = n\mathbb{P}(X_1 = 1) = n\left(1 - \frac{1}{n}\right)^m.$$

If $m \ge (1+\varepsilon)n\log n$, we get from $1 - x \le e^{-x}$,

$$\mathbb{E}X \le ne^{-m/n} \le n^{-\varepsilon}.$$

We have obtained that for every $\varepsilon > 0$, if $m \ge (1+\varepsilon)n\log n$, then

$$\mathbb{P}(\text{there are empty boxes}) = \mathbb{P}(X > 0) \le \mathbb{E}X \le n^{-\varepsilon},$$

equivalently

$$\mathbb{P}(\text{no empty boxes}) \ge 1 - n^{-\varepsilon}.$$

2nd moment method

Claim. If X is a nonnegative integer-valued random variable, then

$$\mathbb{P}(X > 0) \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}X^2}.$$

This follows from the Cauchy-Schwarz inequality (Theorem 2.2) because

$$\mathbb{E}X = \mathbb{E}X\mathbf{1}_{\{X>0\}} \leq \sqrt{\mathbb{E}X^2}\sqrt{\mathbb{E}\mathbf{1}^2_{\{X>0\}}} = \sqrt{\mathbb{E}X^2}\sqrt{\mathbb{P}(X>0)}.$$

Let us apply this again to X being the number of empty boxes. We have

$$\mathbb{E}X^2 = \mathbb{E}\sum_{i,j\leq n} X_iX_j = n\mathbb{E}X_1 + n(n-1)\mathbb{E}X_1X_2 = n\left(1-\frac{1}{n}\right)^m + n(n-1)\mathbb{E}X_1X_2$$

and

$$\mathbb{E}X_1X_2 = \mathbb{P}(X_1 = X_2 = 1) = \left(1-\frac{2}{n}\right)^m.$$

Thus

$$\begin{aligned}
\frac{(\mathbb{E}X)^2}{\mathbb{E}X^2} = \frac{n^2\left(1-\frac{1}{n}\right)^{2m}}{n\left(1-\frac{1}{n}\right)^m + n(n-1)\left(1-\frac{2}{n}\right)^m} &\geq \frac{n^2\left(1-\frac{2}{n}\right)^m}{n\left(1-\frac{1}{n}\right)^m + n^2\left(1-\frac{2}{n}\right)^m} \\
&= \frac{1}{1+\frac{1}{n}\left(\frac{1-\frac{1}{n}}{1-\frac{2}{n}}\right)^m} \\
&= \frac{1}{1+\frac{1}{n}\left(1+\frac{1}{n-2}\right)^m} \\
&> 1-\frac{1}{n}\left(1+\frac{1}{n-2}\right)^m \\
&\geq 1-\frac{1}{n}e^{\frac{m}{n-2}}.
\end{aligned}$$

If $m \leq (1-\varepsilon)n\log n$,

$$\mathbb{P}(X>0) \geq 1-\frac{1}{n}e^{(1-\varepsilon)\frac{n}{n-2}\log n} = 1-\frac{1}{n}e^{(1-\varepsilon)\log n}e^{(1-\varepsilon)\frac{2\log n}{n-2}} = 1-n^{-\varepsilon}e^{(1-\varepsilon)\frac{2\log n}{n-2}},$$

so for large n we get

$$\mathbb{P}(\text{there are empty boxes}) = \mathbb{P}(X>0) \geq 1-2n^{-\varepsilon}.$$

Summarising, we have obtained a rather tight answer: $m = n\log n$ is the number of balls at which the probability of having empty boxes transitions from being very large to being very small as n goes to ∞.

2.4 Existence by averaging

Sometimes to prove existence of a combinatorial or geometric object, it suffices to average, that is take expectation. Specifically, if $X : \Omega \to \mathbb{R}$ is a random variable such that

$\mathbb{E}X > a$ for some $a \in \mathbb{R}$, then there exists $\omega \in \Omega$ such that $X(\omega) > a$ (otherwise, $X \leq a$, so $\mathbb{E}X \leq a$). Moreover, when X is discrete, if $\mathbb{E}X \geq a$, then $X(\omega) \geq a$ for some $\omega \in \Omega$ (otherwise $\mathbb{E}X = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x) < \sum_{x \in X(\Omega)} a\mathbb{P}(X = x) = a$). To illustrate this, we consider the following example.

2.9 Example. Let $v_1, \ldots, v_m$ be unit vectors in $\mathbb{R}^n$. We show that there are signs $\varepsilon_1, \ldots, \varepsilon_m \in \{-1, 1\}$ such that

$$|\varepsilon_1 v_1 + \ldots + \varepsilon_m v_m| \geq \sqrt{m}.$$

Here and throughout, $|x| = \sqrt{x_1^2 + \ldots + x_n^2}$ is the Euclidean norm of a vector x in $\mathbb{R}^n$.

Let $\varepsilon_1, \ldots, \varepsilon_m$ be i.i.d. random signs. Consider

$$X = |\varepsilon_1 v_1 + \ldots + \varepsilon_m v_m|^2.$$

Using the linearity of the standard scalar product $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$, $x, y \in \mathbb{R}^n$ (note $|x|^2 = \langle x, x \rangle$), we have

$$\begin{aligned}
\mathbb{E}X = \mathbb{E}\left\langle \sum_{i=1}^m \varepsilon_i v_i, \sum_{i=1}^m \varepsilon_i v_i \right\rangle &= \mathbb{E}\left(\sum_{i,j \leq n} \varepsilon_i \varepsilon_j \langle v_i, v_j \rangle \right) \\
&= \mathbb{E}\left(\sum_{i=1}^m \varepsilon_i^2 \langle v_i, v_i \rangle + \sum_{i \neq j} \varepsilon_i \varepsilon_j \langle v_i, v_j \rangle \right) \\
&= \sum_{i=1}^m (\mathbb{E}\varepsilon_i^2) |v_i|^2 + \sum_{i \neq j} (\mathbb{E}\varepsilon_i \varepsilon_j) \langle v_i, v_j \rangle.
\end{aligned}$$

Since $|v_i|^2 = 1$, $\mathbb{E}\varepsilon_i^2 = 1$ and by independence, $\mathbb{E}\varepsilon_i \varepsilon_j = \mathbb{E}\varepsilon_i \mathbb{E}\varepsilon_j = 0$ for $i \neq j$, we obtain

$$\mathbb{E}X = m.$$

Therefore, we can conclude that exist ω such that $X(\omega) \geq m$, equivalently, the choice of signs $\varepsilon_1(\omega), \ldots, \varepsilon_m(\omega)$ such that

$$|\varepsilon_1(\omega) v_1 + \ldots + \varepsilon_m(\omega) v_m| \geq \sqrt{m}.$$

3 Notions of convergence for random variables

3.1 Definitions and relationships

A sequence of random variables (X_n) converges to a random variable X

a) **almost surely** if $\mathbb{P}\left(\{\omega \in \Omega,\ \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1$, denoted $X_n \xrightarrow[n\to\infty]{a.s.} X$

b) **in probability** if for every $\varepsilon > 0$, $\mathbb{P}\left(|X_n - X| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0$, denoted $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$

c) **in** L_p, $p > 0$, if $\mathbb{E}|X_n - X|^p \xrightarrow[n\to\infty]{} 0$, denoted $X_n \xrightarrow[n\to\infty]{L_p} X$.

For instance, let $\Omega = \{1, 2\}$ and $\mathbb{P}(1) = \mathbb{P}(2) = \frac{1}{2}$, $X_n(1) = -1/n$, $X_n(2) = 1/n$. We have

a) $X_n \xrightarrow[n\to\infty]{a.s.} 0$ because $X_n(\omega) \to 0$ for every $\omega \in \Omega$,

b) $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$ because $\mathbb{P}\left(|X_n| > \varepsilon\right) = \mathbb{P}\left(\frac{1}{n} > \varepsilon\right) \to 0$,

c) $X_n \xrightarrow[n\to\infty]{L_p} 0$ because $\mathbb{E}|X_n|^p = 2\frac{1}{2}\frac{1}{n^p} \to 0$.

We have two results, saying that the convergence in probability is the weakest among the three.

3.1 Theorem. *If a sequence of random variables (X_n) converges to X a.s. then it also converges in probability, but in general* not *conversely.*

Proof. By the definition of the limit of a sequence,

$$\{\lim_n X_n = X\} = \bigcap_{l\geq 1} \bigcup_{N\geq 1} \bigcap_{n\geq N} \left\{|X_n - X| < \frac{1}{l}\right\}.$$

For any events A_l, $\mathbb{P}\left(\bigcap_{l\geq 1} A_l\right) = 1$ if and only if $\mathbb{P}(A_l) = 1$ for all $l \geq 1$. Therefore, $X_n \xrightarrow[n\to\infty]{a.s.} 0$ is equivalent to: for every $l \geq 1$,

$$\mathbb{P}\left(\bigcup_{N\geq 1} \bigcap_{n\geq N} \left\{|X_n - X| < \frac{1}{l}\right\}\right) = 1.$$

By monotonicity with respect to N,

$$\mathbb{P}\left(\bigcup_{N\geq 1} \bigcap_{n\geq N} \left\{|X_n - X| < \frac{1}{l}\right\}\right) = \lim_{N\to\infty} \mathbb{P}\left(\bigcap_{n\geq N} \left\{|X_n - X| < \frac{1}{l}\right\}\right).$$

Finally, observe that by the inclusion $\bigcap_{n\geq N}\left\{|X_n - X| < \frac{1}{l}\right\} \subset \left\{|X_N - X| < \frac{1}{l}\right\}$, we have

$$1 = \lim_{N\to\infty} \mathbb{P}\left(\bigcap_{n\geq N} \left\{|X_n - X| < \frac{1}{l}\right\}\right) \leq \lim_{N\to\infty} \mathbb{P}\left(\left\{|X_N - X| < \frac{1}{l}\right\}\right),$$

so passing to the complements, for every $l \geq 1$,

$$0 \leq \lim_{N\to\infty} \mathbb{P}\left(\left\{|X_N - X| < \frac{1}{l}\right\}\right) \leq 0.$$

Therefore, for every $\varepsilon > 0$, $\lim_{N\to\infty} \mathbb{P}(\{|X_N - X| \geq \varepsilon\}) = 0$, that is $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. The following example of a sequence convergent in probability but not a.s. finishes the proof.

3.2 Example. Let $\Omega = [0,1]$ and $\mathbb{P}(\cdot)$ be the uniform probability measure. Let $X_1 = 1$, $X_2 = \mathbf{1}_{[0,1/2]}$, $X_3 = \mathbf{1}_{[1/2,1]}$, $X_4 = \mathbf{1}_{[0,1/4]}$, $X_5 = \mathbf{1}_{[1/4,1/2]}$, $X_6 = \mathbf{1}_{[1/2,3/4]}$, $X_7 = \mathbf{1}_{[3/4,1]}$, etc., $X_{2^n}, X_{2^n+1}, \ldots, X_{2^{n+1}-1}$ are indicators of a wandering interval of length 2^{-n} shifting to right by 2^{-n} every increment of the index. We have

a) $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$ because for every $\varepsilon > 0$, $\mathbb{P}(|X_n| > \varepsilon) \leq 2^{-k}$ when $2^k \leq n < 2^{k+1}$, which goes to 0 as n goes to ∞.

b) $X_n \overset{a.s}{\nrightarrow} 0$ because for every $\omega \in (0,1)$, the sequence $(X_n(\omega))$ contains infinitely many 0 and 1, so it is not convergent; moreover, if $X_n \xrightarrow[n\to\infty]{a.s} X$ for some random variable X other than 0, then by Theorem 3.1, $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ and from the uniqueness of limits in probability (homework!), $X = 0$ a.s., contradiction.

c) $X_n \xrightarrow[n\to\infty]{L_p} 0$ because $\mathbb{E}|X_n|^p = 2^{-kp}$ when $2^k \leq n < 2^{k+1}$, which goes to 0 as n goes to ∞.

□

3.3 Theorem. *If a sequence of random variables (X_n) converges to X in L_p for some $p > 0$, then it also converges in probability, but in general* not *conversely.*

Proof. By Chebyshev's inequality (2.1),

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{1}{\varepsilon^p}\mathbb{E}|X_n - X|^p \xrightarrow[n\to\infty]{} 0,$$

so $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$. The following example of a sequence convergent in probability but not in L_p finishes the proof. □

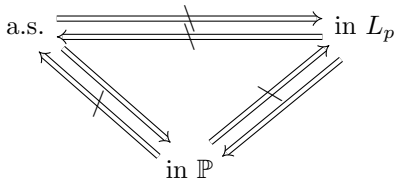
3.4 Example. Let $\Omega = [0,1]$ and $\mathbb{P}(\cdot)$ be the uniform probability measure. Let $X_n = n^{1/p}\mathbf{1}_{[0,1/n]}$. We have

a) $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$ because for every $\varepsilon > 0$, $\mathbb{P}(|X_n| > \varepsilon) \leq \frac{1}{n}$ which goes to 0 as n goes to ∞

b) $X_n \overset{L_p}{\nrightarrow} 0$ because $\mathbb{E}|X_n|^p = n\frac{1}{n} = 1$; moreover, if $X_n \xrightarrow[n\to\infty]{L_p} X$ for some random variable X other than 0, then by Theorem 3.3, $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ and from the uniqueness of limits in probability (homework!), $X = 0$ a.s., contradiction.

c) $X_n \xrightarrow[n\to\infty]{a.s.} 0$ because for every $\omega > 0$, the sequence $X_n(\omega)$ becomes eventually constant 0.

Theorems (3.1), (3.3) and Examples 3.2, 3.4 can be summarised in the following diagram.



3.2 Properties

We record a few basic algebraic properties of the three notions of convergence (homework!).

1) If X_n converges to X a.s./in probability/in L_p and Y_n converges to Y a.s./in probability/in L_p, then $X_n + Y_n$ converges to $X + Y$ a.s./in probability/in L_p.

2) If X_n converges to X a.s./in probability and Y_n converges to Y a.s./in probability, then $X_n \cdot Y_n$ converges to $X \cdot Y$ a.s./in probability.

3) If $0 < p < q$ and X_n converges to X in L_q, then X_n converges to X in L_p.

Immediately, 1) and 2) for the almost sure convergence follow from those statements for sequences of numbers since the intersection of two events of probability 1 is of probability 1.

Property 1) for L_p convergence follows from Minkowski's inequality (Theorem 2.3) and Property 3) follows from the monotonicity of moments (Example 2.6).

Establishing 1) and 2) directly from definition is cumbersome. Instead, we first prove a convenient equivalent condition for convergence in probability in terms of almost sure convergence. En route to this result, we need the Borel-Cantelli lemma, allowing to deduce when only finitely many events occur with probability one.

3.5 Lemma. *If $A_1, A_2, \ldots$ are events such that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then*

$$\mathbb{P}(\text{infinitely many } A_n \text{ occur}) = 0.$$

Proof. By the union bound,

$$\begin{aligned}\mathbb{P}(\text{infinitely many } A_n \text{ occur}) = \mathbb{P}\left(\bigcap_{N\geq 1}\bigcup_{n>N} A_n\right) &= \lim_{N\to\infty} \mathbb{P}\left(\bigcup_{n>N} A_n\right) \\ &\leq \lim_{N\to\infty} \sum_{n>N} \mathbb{P}(A_n),\end{aligned}$$

and the right hand side is zero by the assumption. □

3.6 Theorem (Riesz)**.** *If a sequence (X_n) of random variables converges to a random variable X in probability, then there is a subsequence $(X_{n_k})_k$ which converges to X almost surely.*

Proof. Since for every ε, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$, then we can find an index n_1 such that $\mathbb{P}\left(|X_{n_1} - X| > 2^{-1}\right) < 2^{-1}$. By the same logic, we can find an index $n_2 > n_1$ such that $\mathbb{P}\left(|X_{n_2} - X| > 2^{-2}\right) < 2^{-2}$, etc. We get a subsequence $(X_{n_k})_k$ such that $\mathbb{P}\left(|X_{n_k} - X| > 2^{-k}\right) < 2^{-k}$ for every k. Since the series $\sum_{k=1}^{\infty} \mathbb{P}\left(|X_{n_k} - X| > 2^{-k}\right)$ converges, by the Borel-Cantelli lemma (Lemma 3.5), with probability 1 only finitely many events $A_k = \{|X_{n_k} - X| > 2^{-k}\}$ occur. When this happens, $X_{n_k} \to X$, so $X_{n_k} \xrightarrow[k\to\infty]{} X$. □

3.7 Theorem. *A sequence (X_n) of random variables converges to a random variable X in probability if and only if every subsequence $(X_{n_k})_k$ contains a further subsequence $(X_{n_{k_l}})_l$ which converges to X almost surely.*

Proof. ($\Rightarrow$) It follows directly from Theorem 3.6.

($\Leftarrow$) If (X_n) does not converge to X in probability, then there is $\varepsilon > 0$ such that $\mathbb{P}(|X_n - X| > \varepsilon) \nrightarrow 0$. Consequently, there is $\varepsilon' > 0$ and a subsequence (X_{n_k}) for which $\mathbb{P}(|X_{n_k} - X| > \varepsilon) > \varepsilon'$. By the assumption, there is a subsequence $(X_{n_{k_l}})_l$ convergent to X almost surely, in particular, in probability, so $\mathbb{P}\left(|X_{n_{k_l}} - X| > \varepsilon\right) \to 0$. This contradiction finishes the proof. □

Going back to the algebraic properties 1) and 2) for convergence in probability, we can easily justify them using that they hold for convergence almost surely. For 1), say $S_n = X_n + Y_n$ does not converge in probability to $S = X + Y$. Then as in the proof of Theorem 3.7, $\mathbb{P}(|S_{n_k} - S| > \varepsilon) > \varepsilon'$ for some $\varepsilon, \varepsilon' > 0$ and a subsequence (n_k). Using Theorem 3.7, there is a further subsequence (n_{k_l}) such that $(X_{n_{k_l}})_l$ converges to X a.s. and a further subsequence (for simplicity, denote it the same) such that $(Y_{n_{k_l}})_l$ converges to Y a.s.. Then $S_{n_{k_l}} \xrightarrow{a.s.} S$, which contradicts $\mathbb{P}(|S_{n_k} - S| > \varepsilon) > \varepsilon'$.

4 Laws of large numbers

Suppose we roll a die n times and the outcomes are $X_1, X_2, \ldots, X_n$. We expect that the average $\frac{X_1+\ldots+X_n}{n}$ should be approximately 3.5 (the expectation) as n becomes large. Laws of large numbers establish that rigorously, in a fairly general situation.

Formally, we say that a sequence of random variables $X_1, X_2, \ldots$ satisfies the **weak law of large numbers** if $\frac{X_1+\ldots+X_n}{n} - \mathbb{E}\frac{X_1+\ldots+X_n}{n}$ converges to 0 in probability and the sequence satisfies the **strong law of large numbers** if the convergence is almost sure. In particular, for a sequence of identically distributed random variables, we ask whether $\frac{X_1+\ldots+X_n}{n} \xrightarrow[n\to\infty]{} \mathbb{E}X_1$. Consider two examples when no reasonable law of large numbers holds and the opposite.

4.1 Example. Let $X_1, X_2, \ldots$ be i.i.d. standard Cauchy random variables. Then it can be checked that $\bar{S}_n = \frac{X_1+\ldots+X_n}{n}$ has the same distribution as X_1, so $\bar{S}_n$ is a "well spread out" random variable which in no reasonable sense should be close to its expectation (which in fact does not exists!), or any other constant.

4.2 Example. Let $\varepsilon_1, \varepsilon_2, \ldots$ be i.i.d. symmetric random signs, that is $\mathbb{P}(\varepsilon_i = \pm 1) = \frac{1}{2}$. Let $\bar{S}_n = \frac{\varepsilon_1+\ldots+\varepsilon_n}{n}$. By Bernstein's inequality (Theorem 2.7), $\mathbb{P}\left(|\bar{S}_n| > t\right) \le 2e^{-nt^2/2}$, so the series $\sum_{n=1}^{\infty} \mathbb{P}\left(|\bar{S}_n| > t\right)$ converges, so $\bar{S}_n \xrightarrow[n\to\infty]{a.s.} 0 = \mathbb{E}\varepsilon_1$ (check!). In other words, the sequence (ε_n) satisfies the strong law of large numbers.

4.1 Weak law of large numbers

Using the second moment, we can easily get the weak law of large numbers for uncorrelated random variables with uniformly bounded variance.

4.3 Theorem (The L_2 law of large numbers)**.** *Let $X_1, X_2, \ldots$ be random variables such that $\mathbb{E}|X_i|^2 < \infty$ for every i. If*

$$\frac{1}{n^2}\operatorname{Var}(X_1 + \ldots + X_n) \xrightarrow[n\to\infty]{} 0,$$

then denoting $S_n = X_1 + \ldots + X_n$,

$$\frac{S_n}{n} - \mathbb{E}\frac{S_n}{n} \xrightarrow[n\to\infty]{L_2} 0.$$

In particular, this holds when the X_i are uncorrelated with bounded variance, that is $\operatorname{Var}(X_i) \le M$ for every i for some M.

Proof. We have

$$\mathbb{E}\left|\frac{S_n}{n} - \mathbb{E}\frac{S_n}{n}\right|^2 = \frac{1}{n^2}\mathbb{E}|S_n - \mathbb{E}S_n|^2 = \frac{1}{n^2}\operatorname{Var}(X_1 + \ldots + X_n) \xrightarrow[n\to\infty]{} 0.$$

Since

$$\operatorname{Var}(X_1 + \ldots + X_n) = \sum_{i=1}^{n}\operatorname{Var}(X_i) + 2\sum_{1\le i<j\le n}\operatorname{Cov}(X_i, X_j),$$

when the X_i are uncorrelated with bounded variance, we have

$$\frac{1}{n^2}\operatorname{Var}(X_1+\ldots+X_n) \le \frac{Mn}{n^2} = \frac{M}{n}$$

which goes to 0 as $n \to \infty$. □

Since convergence in L_2 implies convergence in probability, the above is in fact stronger then a weak law of large numbers.

4.4 Example. Let X be a random vector in $\mathbb{R}^n$ uniformly distributed on the cube $[-1,1]^n$, that is $X = (X_1, \ldots, X_n)$ with the X_i being i.i.d. uniform on $[-1,1]$. The assumptions of the above L_2 law of large numbers are satisfied for $X_1^2, X_2^2, \ldots$, so in particular

$$\frac{X_1^2+\ldots+X_n^2}{n} - \mathbb{E}X_1^2 \xrightarrow[n\to\infty]{\mathbb{P}} 0$$

Note that $\mathbb{E}X_1^2 = \frac{1}{3}$. By definition, this convergence in probability means that for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{X_1^2+\ldots+X_n^2}{n} - \frac{1}{3}\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0,$$

or equivalently,

$$\mathbb{P}\left(\sqrt{n(1/3-\varepsilon)} < \sqrt{X_1^2+\ldots+X_n^2} < \sqrt{n(1/3+\varepsilon)}\right) \xrightarrow[n\to\infty]{} 1.$$

In words, a random point in a high dimensional cube is typically near the boundary of the Euclidean ball centered at 0 of radius $\sqrt{n/3}$.

For completeness, we state without proof the weak law of large numbers for i.i.d. sequences under optimal assumptions on integrability.

4.5 Theorem (The weak law of large numbers)**.** *If $X_1, X_2, \ldots$ are i.i.d. random variables such that $t\mathbb{P}(|X_1| > t) \xrightarrow[t\to\infty]{} 0$, then*

$$\frac{X_1+\ldots+X_n}{n} - \mu_n \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

where $\mu_n = \mathbb{E}X_1 \mathbf{1}_{\{|X_1|\le n\}}$.

4.2 Strong law of large numbers

The main goal is to prove the following strong law of large numbers for i.i.d. sequences with optimal assumptions on integrability, which is due to Kolmogorov.

4.6 Theorem (Kolmogorov's strong law of large numbers)**.** *If $X_1, X_2, \ldots$ are i.i.d. random variables such that $\mathbb{E}|X_1| < \infty$, then*

$$\frac{X_1+\ldots+X_n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}X_1.$$

To prove this, we first need to develop a few tools. The first one is a useful fact about deducing convergence of averages of sequences of numbers from convergence of series.

4.7 Lemma (Kronecker)**.** *Let (a_n) be a sequence of real numbers. If the series $\sum_{n=1}^{\infty} \frac{a_n}{n}$ converges, then*

$$\frac{a_1 + \ldots + a_n}{n} \xrightarrow[n\to\infty]{} 0.$$

Proof. Let $s_n = \sum_{k=1}^{n} \frac{a_k}{k}$. Then $s_1 = a_1$, $s_n - s_{n-1} = \frac{a_n}{n}$, $n \geq 2$, so

$$\begin{aligned}\frac{a_1 + \ldots + a_n}{n} &= \frac{s_1 + 2(s_2 - s_1) + 3(s_3 - s_2) + \ldots + n(s_n - s_{n-1})}{n} \\ &= \frac{ns_n - s_1 - s_2 - \ldots - s_{n-1}}{n}.\end{aligned}$$

Fix $\varepsilon > 0$. Since (s_n) is a convergent sequence, it is bounded, say $|s_n| \leq M$ for every n, and by the Cauchy criterion, there is N such that for $n, m > N$, we have $|s_n - s_m| < \varepsilon$. Consequently, for $n > N$,

$$\begin{aligned}&\left|\frac{ns_n - s_1 - s_2 - \ldots - s_{n-1}}{n}\right| \\ &= \left|\frac{(N+1)s_n - s_1 - \ldots - s_N}{n} + \frac{s_n - s_{N+1}}{n} + \ldots + \frac{s_n - s_{n-1}}{n}\right| \\ &\leq \frac{(2N+1)M}{n} + \frac{(n-N-1)\varepsilon}{n}\end{aligned}$$

which is less than, say 2ε for n large enough. □

The second tool we need is a classical maximal inequality (tail bound) for partial sums of independent random random variables, due to Kolmogorov.

4.8 Theorem (Kolmogorov's maximal inequality)**.** *If $X_1, X_2, \ldots, X_n$ are independent random variables such that $\mathbb{E}|X_i|^2 < \infty$ and $\mathbb{E}X_i = 0$ for every i, the for $t > 0$, we have*

$$\mathbb{P}\left(\max_{1\leq k\leq n} |X_1 + \ldots + X_k| \geq t\right) \leq \frac{1}{t^2}\mathbb{E}(X_1 + \ldots + X_n)^2.$$

Proof. Denote $S_0 = 0$ and $S_k = X_1 + \ldots + X_k$, $1 \leq k \leq n$. Fix $t > 0$ and consider the events

$$A_k = \{|S_j| < t \text{ for all } j < k \text{ and } |S_k| \geq k\}$$

(A_k means that $j = k$ is the first index for a partial sum S_j to be at least t). These are disjoint events and

$$\bigcup_{j=1}^{n} A_j = \{\max_{1\leq j\leq n} |S_j| \geq t\}.$$

Moreover, note that the event A_k depends only on $X_1, \ldots, X_k$ (hence is independent of $X_{k+1}, \ldots, X_n$). We have

$$\mathbb{E}S_n^2 \geq \mathbb{E}S_n^2 \mathbf{1}_{\bigcup A_k} = \sum_{k=1}^{n} \mathbb{E}S_n^2 \mathbf{1}_{A_k},$$

where the equality holds because the A_k are disjoint. Writing $S_n = S_k + (S_n - S_k)$ and squaring yields

$$\mathbb{E}S_n^2 \geq \sum_{k=1}^{n} \left(\mathbb{E}S_k^2 \mathbf{1}_{A_k} + 2\mathbb{E}S_k(S_n - S_k)\mathbf{1}_{A_k} + \mathbb{E}(S_n - S_k)^2 \mathbf{1}_{A_k} \right)$$

Note that by independence, $\mathbb{E}S_k(S_n - S_k)\mathbf{1}_{A_k} = \mathbb{E}(S_k \mathbf{1}_{A_k})(S_n - S_k) = \mathbb{E}S_k \mathbf{1}_{A_k} \cdot \mathbb{E}(S_n - S_k) = \mathbb{E}S_k \mathbf{1}_{A_k} \cdot 0 = 0$. Bounding the last term trivially by 0 thus gives

$$\mathbb{E}S_n^2 \geq \sum_{k=1}^{n} \mathbb{E}S_k^2 \mathbf{1}_{A_k}.$$

On the event A_k, $|S_k| \geq t$, so $\mathbb{E}S_k^2 \mathbf{1}_{A_k} \geq \mathbb{E}t^2 \mathbf{1}_{A_k} = t^2 \mathbb{P}(A_k)$ and we finally obtain

$$\mathbb{E}S_n^2 \geq \sum_{k=1}^{n} t^2 \mathbb{P}(A_k) = t^2 \mathbb{P}\left(\bigcup A_k \right) = t^2 \mathbb{P}\left(\max_{1 \leq j \leq n} |S_j| \geq t \right),$$

which is the desired inequality. □

To use Kronecker's lemma, we establish a convenient criterion for almost sure convergence of series of independent random variables.

4.9 Lemma. *Let $X_1, X_2, \ldots$ be independent random variables such that for every i, $\mathbb{E}X_i = 0$ and $\mathbb{E}|X_i|^2 < \infty$. If $\sum_{n=1}^{\infty} \mathrm{Var}(X_i)$ converges, then $\sum_{n=1}^{\infty} X_n$ converges a.s.*

Proof. We want to show that $\mathbb{P}(\sum_{n=1}^{\infty} X_n \text{ diverges}) = 0$. By the Cauchy criterion, $\sum_{n=1}^{\infty} X_n$ diverges if and only if there is $\varepsilon > 0$ such that for every $N \geq 1$ there are $n > m > N$ with $|X_m + \ldots + X_n| \geq \varepsilon$. Thus

$$\begin{aligned} \mathbb{P}\left(\sum_{n=1}^{\infty} X_n \text{ diverges} \right) &= \mathbb{P}\left(\bigcup_{l \geq 1} \bigcap_{N \geq 1} \left\{ \sup_{n > m > N} |X_m + \ldots + X_n| \geq \frac{1}{l} \right\} \right) \\ &\leq \sum_{l=1}^{N} \mathbb{P}\left(\bigcap_{N \geq 1} \left\{ \sup_{n > m > N} |X_m + \ldots + X_n| \geq \frac{1}{l} \right\} \right). \end{aligned}$$

It suffices that every term in the last sum is zero. Observe that by simple monotonicity of the events involved,

$$\begin{aligned} &\mathbb{P}\left(\bigcap_{N \geq 1} \left\{ \sup_{n > m > N} |X_m + \ldots + X_n| \geq \frac{1}{l} \right\} \right) \\ &\quad = \lim_{N \to \infty} \mathbb{P}\left(\sup_{n > m > N} |X_m + \ldots + X_n| \geq \frac{1}{l} \right) \\ &\quad = \lim_{N \to \infty} \mathbb{P}\left(\bigcup_{n > N} \left\{ \max_{n > m > N} |X_m + \ldots + X_n| \geq \frac{1}{l} \right\} \right) \\ &\quad = \lim_{N \to \infty} \lim_{n \to \infty} \mathbb{P}\left(\max_{n > m > N} |X_m + \ldots + X_n| \geq \frac{1}{l} \right). \end{aligned}$$

By Kolmogorov's maximal inequality (Theorem 4.8),

$$\mathbb{P}\left(\max_{n>m>N}|X_m+\ldots+X_n|\geq\frac{1}{l}\right)\leq\frac{1}{(1/l)^2}\operatorname{Var}(X_{N+1}+\ldots+X_n)=l^2\sum_{k=N+1}^{n}\operatorname{Var}(X_k),$$

which after taking consecutively the limits $n\to\infty$ and then $N\to\infty$ gives an upper bound by $\lim_{N\to\infty}\sum_{k=N+1}^{\infty}\operatorname{Var}(X_k)=0$ because the series $\sum\operatorname{Var}(X_k)$ converges by the assumption. This finishes the proof. □

Now we are ready to prove Kolmogorov's strong law of large numbers (Theorem 4.6). To give a good picture of our strategy, we first show how to quickly prove it under a stronger assumption of the finite second moment.

4.10 Remark. Let, as in Theorem 4.6, $X_1, X_2, \ldots$ be i.i.d. random variables such that $\mathbb{E}|X_1|^2<\infty$ (which clearly implies that $\mathbb{E}|X_1|<\infty$). We would like to show that $\frac{X_1+\ldots+X_n}{n}\xrightarrow[n\to\infty]{a.s.}\mathbb{E}X_1$. This is equivalent to

$$\frac{(X_1-\mathbb{E}X_1)+\ldots+(X_n-\mathbb{E}X_n)}{n}\xrightarrow[n\to\infty]{a.s.}0.$$

By Kronecker's lemma (Lemma 4.7), it is enough to show that

$$\sum_{n=1}^{\infty}\frac{X_n-\mathbb{E}X_n}{n}\quad\text{converges a.s.}$$

By Lemma 4.9, it is enough to show that

$$\sum_{n=1}^{\infty}\operatorname{Var}\left(\frac{X_n-\mathbb{E}X_n}{n}\right)<\infty.$$

This is clear because $\operatorname{Var}\left(\frac{X_n-\mathbb{E}X_n}{n}\right)=\frac{1}{n^2}\operatorname{Var}(X_n)=\frac{1}{n^2}\operatorname{Var}(X_1)$.

Proof of Theorem 4.6. Consider the truncated random variables

$$Y_n=X_n\mathbf{1}_{\{|X_n|\leq n\}}.$$

We have

$$\frac{X_1+\ldots+X_n}{n}-\mathbb{E}X_1=R_n+S_n+T_n,$$

where

$$\begin{aligned}R_n&=\frac{(X_1-Y_1)+\ldots+(X_n-Y_n)}{n},\\S_n&=\frac{(Y_1-\mathbb{E}Y_1)+\ldots+(Y_n-\mathbb{E}Y_n)}{n},\\T_n&=\frac{\mathbb{E}Y_1+\ldots+\mathbb{E}Y_n}{n}-\mathbb{E}X_1.\end{aligned}$$

We shall show that each of these sequences converges to 0 a.s. and this will finish the proof.

Sequence (T_n). Here we use a simple fact that for a sequence of numbers (a_n), if $\lim_n a_n = a$, then $\lim_n \frac{a_1+\ldots+a_n}{n} = a$. Let $a_n = \mathbb{E}Y_n$. Since the X_i are identically distributed,

$$a_n = \mathbb{E}X_n\mathbf{1}_{\{|X_n|\le n\}} = \mathbb{E}X_1\mathbf{1}_{\{|X_1|\le n\}}$$

By Lebesgue's dominated convergence theorem ($|X_1\mathbf{1}\{|X_1|\le n\}|$ is dominated by $|X_1|$, which is integrable),

$$\lim_n a_n = \lim_n \mathbb{E}X_1\mathbf{1}_{\{|X_1|\le n\}} = \mathbb{E}(X_1 \lim_n \mathbf{1}_{\{|X_1|\le n\}}) = \mathbb{E}X_1.$$

This shows that $T_n \to 0$.

Sequence (R_n). Let $A_n = \{X_n \neq Y_n\} = \{|X_n| > n\}$. Since

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1| > n) \le \mathbb{E}|X_1|,$$

by the Borel-Cantelli lemma (Lemma 3.5), with probability 1 only finitely many A_n occur. If that is the case, then the sequences $(X_n(\omega))$ and $(Y_n(\omega)$ are eventually the same, which implies that

$$R_n(\omega) = \frac{(X_1(\omega) - Y_1(\omega)) + \ldots + (X_n(\omega) - Y_n(\omega))}{n} \xrightarrow[n\to\infty]{} 0.$$

This shows that $R_n \xrightarrow[n\to\infty]{a.s.} 0$.

Sequence (S_n). We proceed as in Remark 4.10: by Kronecker's lemma and Lemma 4.9 it suffices to show that

$$\sum_{n=1}^{\infty} \mathrm{Var}\left(\frac{Y_n - \mathbb{E}Y_n}{n}\right) < \infty.$$

Note that

$$\begin{aligned}
\mathrm{Var}(Y_n) \le \mathbb{E}Y_n^2 = \mathbb{E}X_n^2\mathbf{1}_{\{|X_n|\le n\}} &= \mathbb{E}X_1^2\mathbf{1}_{\{|X_1|\le n\}} \\
&= \sum_{k=1}^{n} \mathbb{E}X_1^2\mathbf{1}_{\{k-1<|X_1|\le k\}} \\
&\le \sum_{k=1}^{n} k\mathbb{E}|X_1|\mathbf{1}_{\{k-1<|X_1|\le k\}}.
\end{aligned}$$

Changing the order of summation and using that $\sum_{n=k} \frac{1}{n^2} \le \frac{2}{k}$, we obtain

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathrm{Var}\left(\frac{Y_n - \mathbb{E}Y_n}{n}\right) &= \sum_{n=1}^{\infty} \frac{1}{n^2}\mathrm{Var}(Y_n) \\
&\le \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{k=1}^{n} k\mathbb{E}|X_1|\mathbf{1}_{\{k-1<|X_1|\le k\}} \\
&= \sum_{k=1}^{\infty} \left(\sum_{n=k}^{\infty} \frac{1}{n^2}\right) k\mathbb{E}|X_1|\mathbf{1}_{\{k-1<|X_1|\le k\}} \\
&\le 2\sum_{k=1}^{\infty} \mathbb{E}|X_1|\mathbf{1}_{\{k-1<|X_1|\le k\}} = 2\mathbb{E}|X_1|
\end{aligned}$$

which is finite. □

We finish this chapter with an application of the strong law of large numbers to computation of certain integrals.

4.11 Example. Let $I_n = \int_0^1 \dots \int_0^1 \frac{x_1^3+\dots+x_n^3}{x_1+\dots+x_n} \mathrm{d}x_1 \dots \mathrm{d}x_n$. We shall find $\lim_n I_n$.

Let $X_1, \dots, X_n$ be i.i.d. random variables uniform on $[0,1]$. The density of the vector $X = (X_1, \dots, X_n)$ is $f(x) = \prod_{i=1}^n \mathbf{1}_{[0,1]}(x_i)$, so

$$I_n = \mathbb{E}\frac{X_1^3 + \dots + X_n^3}{X_1 + \dots + X_n} = \mathbb{E}\frac{\frac{X_1^3+\dots+X_n^3}{n}}{\frac{X_1+\dots+X_n}{n}}.$$

By the strong law of large numbers (Theorem 4.6),

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}X_1 = \frac{1}{2}$$

and similarly

$$\frac{X_1^3 + \dots + X_n^3}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}X_1^3 = \frac{1}{4}.$$

Thus

$$\frac{X_1^3 + \dots + X_n^3}{X_1 + \dots + X_n} = \frac{\frac{X_1^3+\dots+X_n^3}{n}}{\frac{X_1+\dots+X_n}{n}} \xrightarrow[n\to\infty]{a.s.} \frac{1/4}{1/2} = \frac{1}{2}.$$

Moreover, we have a simple bound $\left|\frac{X_1^3+\dots+X_n^3}{X_1+\dots+X_n}\right| \le 1$ (because the $0 \le X_i \le 1$), so by Lebesgue's dominated convergence theorem,

$$\lim_{n\to\infty} I_n = \lim_{n\to\infty} \mathbb{E}\frac{X_1^3 + \dots + X_n^3}{X_1 + \dots + X_n} = \mathbb{E} \lim_{n\to\infty} \frac{X_1^3 + \dots + X_n^3}{X_1 + \dots + X_n} = \mathbb{E}\frac{1}{2} = \frac{1}{2}.$$

5 Central limit theorem

Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}|X_1|^2 < \infty$. By the strong law of large numbers,

$$Y_n = \frac{X_1 + \ldots + X_n}{n} - \mathbb{E}X_1$$

converges to 0 a.s. By our (too generous to get this convergence) assumption, we can compute

$$\mathrm{Var}(Y_n) = \frac{\mathrm{Var}(X_1 + \ldots + X_n)}{n^2} = \frac{n\,\mathrm{Var}(X_1)}{n^2} = \frac{\mathrm{Var}(X_1)}{n},$$

so Y_n concentrates around its expectation, which is 0 and in a sense it is not surprising that Y_n goes to 0. What happens if we zoom in, that is rescale appropriately so that the variance of Y_n is fixed (when fluctuations of Y_n have a fixed size, as opposed to decaying like $1/n$ as earlier)? Consider

$$Z_n = \frac{Y_n}{\sqrt{\mathrm{Var}(Y_n)}} = \sqrt{n}\frac{1}{\sqrt{\mathrm{Var}(X_1)}}\left(\frac{X_1 + \ldots + X_n}{n} - \mathbb{E}X_1\right)$$

which has variance 1 for all n. What "limit distribution" does Z_n have as $n \to \infty$ (if any)? This is addressed by the central limit theorem which says that the limiting distribution exists and is Gaussian! (If it exists and is universal, that is the same for all i.i.d. sequences, then it has to be Gaussian because when the X_i are standard Gaussian, Z_n is also standard Gaussian.) To make things rigorous, first we need to develop a notion of convergence in distribution, the most important type of convergence in probability theory.

We say that a sequence of random variables (X_n) converges to a random variable X **in distribution**, denoted $X_n \xrightarrow[n\to\infty]{d} X$, if

$$F_{X_n}(t) \xrightarrow[n\to\infty]{} F_X(t) \qquad \text{for every point of continuity of } F_X.$$

Here as usual, $F_Y(t) = \mathbb{P}(Y \le t)$ is the distribution function of a random variable Y. Note that this notion of convergence only depends on the distribution functions of random variables involved and not on their particular realisations as functions on a probability space (in fact, they can be defined on different probability spaces). This is in strong contrast to, for instance almost sure convergence – see also Theorem 5.4. Particularly, if $X_n \xrightarrow[n\to\infty]{d} X$ and say X' is another random variable with the same distribution as X (i.e. $F_{X'} = F_X$), then we can also write $X_n \xrightarrow[n\to\infty]{d} X'$.

5.1 Theorem (Central limit theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}|X_1|^2 < \infty$. Then the sequence (Z_n) of normalised sums*

$$Z_n = \frac{X_1 + \ldots + X_n - n\mathbb{E}X_1}{\sqrt{n\,\mathrm{Var}(X_1)}}$$

converges in distribution to a standard Gaussian random variable.

To give a complete proof of this fundamental limit theorem, we shall first study properties of convergence in distribution and then build up a tool, borrowing ideas from Fourier analysis, of so-called characteristic functions of random variables which captures the convergence in distribution and easily allows to take advantage of independence.

5.1 Convergence in distribution

We begin with two simple examples.

5.2 Example. Let ε be a symmetric random sign. Consider the sequence $(X_n)_{n=1}^{\infty} = (\varepsilon, -\varepsilon, \varepsilon, -\varepsilon, \ldots)$. Since $-\varepsilon$ has the same distribution as ε, we have $F_{X_n} = F_\varepsilon$ for every n, so $X_n \xrightarrow{d} \varepsilon$. On the other hand, the sequence (X_n) does not converge in probability, for suppose $X_n \xrightarrow{\mathbb{P}} X$ for some random variable X. Then for n, m large enough $\mathbb{P}(|X_n - X_m| > 1) \leq \mathbb{P}(|X_n - X| > 1/2) + \mathbb{P}(|X - X_m| > 1/2) \leq 1/4$. Taking n and m of different parity, we get $\mathbb{P}(|X_n - X_m| > 1) = \mathbb{P}(|2\varepsilon| > 1) = 1$, a contradiction. It turns out that convergence in probability implies convergence in distribution, but we will be able to show this a bit later.

5.3 Example. Let X be a random variable and consider the sequence $X_n = X + \frac{1}{n}$. For any reasonable definition of "convergence in distribution" we should have $X_n \to X$. Note that for a fixed $t \in \mathbb{R}$, we have

$$\lim F_{X_n}(t) = \lim \mathbb{P}(X_n \leq t) = \lim \mathbb{P}\left(X \leq t - \frac{1}{n}\right) = F(t-),$$

which is $F(t)$ if and only if t is a continuity point of F. This explains why in the definition we make this exclusion.

We have a curious relationship between convergence in distribution and almost sure convergence.

5.4 Theorem. *If a sequence of random variables (X_n) converges in distribution to a random variable X, then there are random variables Y_n and Y such that Y_n has the same distribution as X_n, Y has the same distribution as X and $Y_n \to Y$ a.s.*

Proof. Let $F_n = F_{X_n}$ be the distribution function of X_n and let $F = F_X$ be the distribution function of X. Let $\Omega = (0,1)$, $\mathcal{F}$ be the Borel subsets of $(0,1)$ and $\mathbb{P}(\cdot)$ be uniform. For every $x \in (0,1)$ define the "inverse" distribution functions

$$Y_n(x) = \sup\{y \in \mathbb{R},\ F_n(y) < x\}$$

and similarly

$$Y(x) = \sup\{y \in \mathbb{R},\ F(y) < x\}.$$

By the construction, $F_{Y_n} = F_n$ and $F_Y = F$. Note that Y_n and Y are nondecreasing right-continuous functions whose only discontinuities are jumps which happen at at most

countably many points. If we let Ω_0 to be the set of points where Y is continuous, then $\mathbb{P}(\Omega_0) = 1$. Fix $x \in \Omega_0$. We claim that $Y_n(x) \to Y(x)$, which then gives $Y_n \to Y$ a.s. We have

1. $\liminf Y_n(x) \geq Y(x)$, for suppose $y < Y(x)$ is a continuity point of F; then $F(y) < x$ (since $x \in \Omega_0$), so for large n, $F_n(y) < x$ and by the definition of the supremum, $y \leq Y_n(x)$. Taking lim inf, we get $\liminf Y_n(x) \geq y$ for every $y < Y(x)$, so $\liminf Y_n(x) \geq Y(x)$.

2. $Y(x) \geq \limsup Y_n(x)$, for suppose $y > Y(x)$ is a continuity point of F; then $F(y) > x$, so for large n, $F_n(y) > x$ which gives $y \geq Y_n(x)$. Taking lim sup finishes the argument.

□

We show an equivalent definition of convergence in distribution in terms of bounded continuous test functions.

5.5 Theorem. *A sequence (X_n) of random variables converges in distribution to a random variable X if and only if for every bounded continuous function $g : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$.*

Proof. ($\Rightarrow$) Let Y_n and Y be as in Theorem 5.4, $Y_n \xrightarrow{a.s.} Y$. Since g is continuous, we also have $g(Y_n) \xrightarrow{a.s.} g(Y)$, so by Lebesgue's dominated convergence theorem (g is bounded),

$$\mathbb{E}g(X_n) = \mathbb{E}g(Y_n) \to \mathbb{E}g(Y) = \mathbb{E}g(X)$$

(recall Remark 1.12).

($\Leftarrow$) For parameters $t \in \mathbb{R}$ and $\varepsilon > 0$ define the continuous bounded functions

$$g_{t,\varepsilon}(x) = \begin{cases} 1, & x \leq t, \\ 1 - \frac{x-t}{\varepsilon}, & t < x \leq t + \varepsilon, \\ 0, & x > t + \varepsilon. \end{cases}$$

The idea is that these functions are continuous approximations of indicator functions. We have, $\mathbf{1}_{\{x \leq t\}} \leq g_{t,\varepsilon}(x) \leq \mathbf{1}_{\{x \leq t+\varepsilon\}}$. Consequently,

$$\begin{aligned} \limsup \mathbb{P}(X_n \leq t) = \limsup \mathbb{E}\mathbf{1}_{\{X_n \leq t\}} &\leq \limsup \mathbb{E}g_{t,\varepsilon}(X_n) \\ &= \mathbb{E}g_{t,\varepsilon}(X) \leq \mathbb{E}\mathbf{1}_{\{X \leq t+\varepsilon\}} = \mathbb{P}(X \leq t + \varepsilon). \end{aligned}$$

Letting $\varepsilon \to 0$ gives

$$\limsup F_{X_n}(t) \leq F_X(t).$$

On the other hand, since

$$\begin{aligned}\liminf \mathbb{P}\left(X_n \le t\right) = \liminf \mathbb{E}\mathbf{1}_{\{X_n \le t\}} &\ge \liminf \mathbb{E}g_{t-\varepsilon,\varepsilon}(X_n) \\ &= \mathbb{E}g_{t-\varepsilon,\varepsilon}(X) \ge \mathbb{E}\mathbf{1}_{\{X \le t-\varepsilon\}} = \mathbb{P}\left(X \le t-\varepsilon\right)\end{aligned}$$

after taking $\varepsilon \to 0$, we get

$$\liminf F_{X_n}(t) \ge F_X(t-).$$

If t is a point of continuity of F_X, $F_X(t-) = F_X(t)$ and we obtain $\lim F_{X_n}(t) = F_X(t)$, which means $X_n \xrightarrow{d} X$. □

Being able to extract convergent subsequences often helps. Since distribution functions are bounded, this is always possible, as stated in the next theorem.

5.6 Theorem (Helly's selection theorem)**.** *If $(F_n)_n$ is a sequence of distribution functions, then there is a subsequence $(F_{n_k})_k$ and a right-continuous nondecreasing function $F : \mathbb{R} \to [0,1]$ such that $F_{n_k}(t) \xrightarrow[k\to\infty]{} F(t)$ for every point t of continuity of F.*

5.7 Remark. In general, F may not be a distribution function – it may happen that $F(\infty) < 1$ or $F(-\infty) > 0$.

Proof. To construct the desired subsequence we use a standard diagonal argument. Let $q_1, q_2, \ldots$ be a sequence of all rationals. Since the sequence $F_n(q_1)$ is bounded, it has a convergent subsequence, say $F_{n_k^{(1)}}(q_1)$ converges to $G(q_1)$. Then we look at the sequence $F_{n_k^{(1)}}(q_2)$ which is bounded, so it has a convergent subsequence, say $F_{n_k^{(2)}}(q_2)$ converges to $G(q_2)$, etc. We obtain subsequences $(n_k^{(l)})$ such that $(n_k^{(l+1)})$ is a subsequence of $(n_k^{(l)})$ and $F_{n_k^{(l)}}(q_l)$ converges to $G(q_l)$. Choose the diagonal subsequence $n_k = n_k^{(k)}$. Then $F_{n_k^{(k)}}(q_l)$ converges to $G(q_l)$ for every l. The function $G : \mathbb{Q} \to [0,1]$ obtained as the limit is nondecreasing. We extend it to the nondecreasing function $F : \mathbb{R} \to [0,1]$ by

$$F(x) = \inf\{G(q),\ q \in \mathbb{Q}, q > x\}, \qquad x \notin \mathbb{Q}.$$

The function F, as monotone, satisfies $F(x-) \le F(x) \le F(x+)$ for every x. At the points x, where F is not right-continuous, we modify it and set $F(x) = F(x+)$ (there are at most countably many such points).

It remains to check that F_{n_k} converges to F at its points of continuity. Let x be such a point and let q, r be rationals such that $q < x < r$. Then

$$\begin{aligned}F(q) = G(q) = \liminf_k F_{n_k}(q) &\le \liminf_k F_{n_k}(x) \\ &\le \limsup_k F_{n_k}(x) \le \limsup_k F_{n_k}(r) = G(r) = F(r).\end{aligned}$$

Letting $q, r \to x$, we get $F(q), F(r) \to F(x)$, so $\liminf_k F_{n_k}(x) = \limsup_k F_{n_k}(x) = F(x)$. □

To capture when the limiting function is a distribution function of a random variable, we need the notion of tightness. A sequence (X_n) of random variables is **tight** if for every $\varepsilon > 0$, there is $M > 0$ such that $\mathbb{P}(|X_n| \leq M) > 1 - \varepsilon$ for every n.

5.8 Remark. If there is $\delta > 0$ such that $C = \sup_n \mathbb{E}|X_n|^\delta < \infty$, then the sequence (X_n) is tight. Indeed, by Chebyshev's inequality,

$$\mathbb{P}(|X_n| > M) \leq M^{-\delta}\mathbb{E}|X_n|^\delta \leq \frac{C}{M^\delta}$$

which is less than ε for M large enough.

The main result of this section is the following compactness type result. It gives a necessary and sufficient condition for existence of convergent subsequences in distribution in terms of tightness.

5.9 Theorem. *A sequence of random variables (X_n) is tight if and only every subsequence $(X_{n_k})_k$ has a subsequent $(X_{n_{k_l}})_l$ which converges in distribution to some random variable.*

Proof. Let F_n be the distribution function of X_n.

($\Rightarrow$) By Helly's theorem applied to $(F_{n_k})_k$, there is a subsequence $(F_{n_{k_l}})_l$ which converges to a right-continuous nondecreasing function $F : \mathbb{R} \to [0, 1]$ pointwise at the points of continuity of F. It remains to check that F is a distribution function, that is $F(-\infty) = 0$ and $F(+\infty) = 1$. By tightness, there is $M > 0$ such that $F_n(M) - F_n(-M) > 1 - \varepsilon$, for every n and we can further arrange that $-M$ and M are points of continuity of F. Taking $n = n_{k_l}$ and letting $l \to \infty$, we get $F(M) - F(-M) \geq 1 - \varepsilon$. Since ε is arbitrary and F is monotone, this yields $F(-\infty) = 0$ and $F(+\infty) = 1$.

($\Leftarrow$) If (X_n) is not tight, there is $\varepsilon > 0$ and an increasing sequence of indices n_k such that $\mathbb{P}(|X_{n_k}| \leq k) \leq 1 - \varepsilon$ for every k. By the assumption, $X_{n_{k_l}} \xrightarrow[l\to\infty]{d} X$. Let $x < 0 < y$ be points of continuity of F_X. Then

$$F_X(y) - F_X(x) = \lim_l (F_{n_{k_l}}(y) - F_{n_{k_l}}(x)) \leq \limsup_l (F_{n_{k_l}}(k_l) - F_{n_{k_l}}(-k_l)) \leq 1 - \varepsilon.$$

Taking $x \to -\infty$ and $y \to \infty$ gives $1 \leq 1 - \varepsilon$, a contradiction. □

5.2 Characteristic functions

The **characteristic function** of a random variable X is the function $\phi_X : \mathbb{R} \to \mathbb{C}$ defined as

$$\phi_X(t) = \mathbb{E}e^{itX}, \qquad t \in \mathbb{R}.$$

(For complex valued random variables, say $Z = X + iY$, we of course define $\mathbb{E}Z = \mathbb{E}X + i\mathbb{E}Y$.) Since $e^{ix} = \cos x + i \sin x$, $x \in \mathbb{R}$ is a complex number of modulus 1, e^{itX} is a bounded random variable hence its expectation exists, so ϕ_X is well-defined on $\mathbb{R}$.

For starters, two examples. For a symmetric random sign ε,

$$\phi_\varepsilon(t) = \mathbb{E}e^{it\varepsilon} = \frac{e^{it} + e^{-it}}{2} = \cos t.$$

For an exponential random variable X with parameter λ,

$$\phi_X(t) = \mathbb{E}e^{itX} = \int_{-\infty}^{\infty} e^{itx} f_X(x)\mathrm{d}x = \int_0^{\infty} \lambda e^{(it-\lambda)x}\mathrm{d}x = \lambda \frac{e^{-\lambda x}e^{itx}}{it-\lambda}\bigg|_0^{\infty} = \frac{\lambda}{\lambda - it}$$

(when taking the limit $x \to \infty$, we use that e^{itx} is bounded).

Basic properties of characteristic functions

We gather several basic properties in the following theorem.

5.10 Theorem. *Let X be a random variable with characteristic function ϕ_X. Then*

(i) $|\phi_X(t)| \le 1$, $t \in \mathbb{R}$,

(ii) $\phi_X(0) = 1$,

(iii) ϕ_X *is uniformly continuous,*

(iv) if $\mathbb{E}|X|^n < \infty$ *for some positive integer* n*, then the* n*th derivative* $\phi_X^{(n)}$ *exists, equals* $\phi_X^{(n)}(t) = i^n \mathbb{E}X^n e^{itX}$ *and is uniformly continuous.*

Proof. (i) and (ii) are clear because $|\phi_X(t)| = |\mathbb{E}e^{itX}| \le \mathbb{E}|e^{itX}| = 1$ and $\phi_X(0) = \mathbb{E}e^{i\cdot 0\cdot X} = 1$.

(iii) For every $t, h \in \mathbb{R}$,

$$|\phi_X(t+h) - \phi_X(t)| = |\mathbb{E}e^{itX}(e^{ihX} - 1)| \le \mathbb{E}|e^{ihX} - 1| \xrightarrow[h\to 0]{} 0$$

where the limit is justified by Lebesgue's dominated convergence theorem ($|e^{ihX} - 1| \to 0$ pointwise and the sequence is bounded by 2). This implies the continuity of ϕ_X at t. The continuity is uniform because the bound does not depend on t.

(iv) Fix n such that $\mathbb{E}|X|^n < \infty$. First, we inductively show that for $0 \le k \le n$,

$$\phi_X^{(k)}(t) = \mathbb{E}(iX)^k e^{itX}.$$

This is clear for $k = 0$ and for $k < n$, inductively, we have

$$\begin{aligned}\phi_X^{(k+1)}(t) = \lim_{h\to 0} \frac{\phi_X^{(k)}(t+h) - \phi_X^{(k)}(t)}{h} &= \lim_{h\to 0} \mathbb{E}\left[(iX)^k e^{itX} \frac{e^{ihX} - 1}{h}\right] \\ &= \mathbb{E}\left[(iX)^k e^{itX} \lim_{h\to 0} \frac{e^{ihX} - 1}{h}\right].\end{aligned}$$

The last equality is justified by Lebesgue's dominated convergence theorem because

$$\left|(iX)^k e^{itX} \frac{e^{ihX} - 1}{h}\right| \le |X|^k |X| = |X|^{k+1}$$

and by the assumption $\mathbb{E}|X|^{k+1} < \infty$; we also used that for $t \in \mathbb{R}$, $|e^{it} - 1| \leq |t|$ which can be justified as follows

$$|e^{it} - 1| = \left| \frac{1}{i} \int_0^t e^{ix} \mathrm{d}x \right| \leq \int_0^t |e^{ix}| \mathrm{d}x = t$$

when $t \geq 0$ and similarly for $t < 0$. Finally, $\lim_{h \to 0} \frac{e^{ihX} - 1}{h} = iX$ which finishes the inductive argument. Having the formula, uniform continuity follows as in (iii). □

5.11 Example. Let X be a standard Gaussian random variable. We have,

$$\phi_X(t) = \int_{\mathbb{R}} e^{itx} e^{-x^2/2} \frac{\mathrm{d}x}{\sqrt{2\pi}} = e^{-t^2/2} \int_{\mathbb{R}} e^{-(x-it)^2/2} \frac{\mathrm{d}x}{\sqrt{2\pi}} = e^{-t^2/2},$$

where the last step would need proper justification (e.g., integrating along an appropriate contour and using $\int_{\mathbb{R}} e^{-x^2/2} \frac{\mathrm{d}x}{\sqrt{2\pi}}$). Instead, we use Theorem 5.10 (iv),

$$\phi'_X(t) = i\mathbb{E}Xe^{itX} = -\mathbb{E}X \sin(tX) + i\mathbb{E}X \cos(tX).$$

Since X is symmetric and cos is even, $\mathbb{E}X \cos(tX) = 0$ and integrating by parts,

$$\begin{aligned} \phi'_X(t) = -\mathbb{E}X \sin(tX) = -\int x \sin(tx) e^{-x^2/2} \frac{\mathrm{d}x}{\sqrt{2\pi}} &= \int \sin(tx)(e^{-x^2/2})' \frac{\mathrm{d}x}{\sqrt{2\pi}} \\ &= -t \int \cos(tx) e^{-x^2/2} \frac{\mathrm{d}x}{\sqrt{2\pi}} \end{aligned}$$

which is $-t\mathbb{E}\cos(tX) = -t\mathbb{E}e^{itX} = -t\phi_X(t)$ (by the symmetry of X, again, $\mathbb{E}\sin(tX) = 0$), so $\phi'_X(t) = -t\phi_X(t)$. That is, $\phi'_X(t) = -t\phi_X(t)$, equivalently, $(e^{t^2/2}\phi_X(t))' = 0$ which finally gives $e^{t^2/2}\phi_X(t) = \phi_X(0) = 1$.

If $Y \sim N(\mu, \sigma^2)$, then $Y = \mu + \sigma X$ and we thus get

$$\phi_Y(t) = \mathbb{E}e^{it(\mu + \sigma X)} = e^{it\mu}\mathbb{E}e^{i(t\sigma)X} = e^{it\mu - \sigma^2 t^2/2}. \tag{5.1}$$

Note a simple but very powerful observation involving independence.

5.12 Theorem. *If X and Y are independent random variables, then*

$$\phi_{X+Y} = \phi_X \cdot \phi_Y.$$

Proof. Clearly, $\mathbb{E}e^{it(X+Y)} = \mathbb{E}e^{itX}e^{itY} = \mathbb{E}e^{itX}\mathbb{E}e^{itY}$. □

Two crucial properties of characteristic functions are: 1) they determine the distribution 2) they capture convergence in distribution. Specifically, we have the following two theorems.

5.13 Theorem. *Random variables X and Y have the same distribution (that is, $F_X = F_Y$) if and only if they have the same characteristic functions $\phi_X = \phi_Y$.*

5.14 Theorem (Lévy's continuity theorem)**.** *Let (X_n) be a sequence of random variables such that for every $t \in \mathbb{R}$, $\phi_{X_n}(t) \xrightarrow[n \to \infty]{} \phi(t)$ for some function $\phi : \mathbb{R} \to \mathbb{C}$ which is continuous at $t = 0$. Then there is a random variable X such that $\phi = \phi_X$ and $X_n \xrightarrow{d} X$.*

5.15 Remark. The converse also holds: if $X_n \xrightarrow{d} X$, then $\phi_{X_n}(t) \to \phi_X(t)$ for every $t \in \mathbb{R}$. Indeed, by Theorem 5.5, $\mathbb{E}\sin(tX_n) \to \mathbb{E}\sin(tX)$ and the same for the cos function, so $\phi_{X_n}(t) = \mathbb{E}\cos(tX_n) + i\mathbb{E}\sin(tX_n) \to \phi_X(t)$.

5.16 Example. In Levy's theorem the continuity assumption is necessary. Let G be a standard Gaussian random variable and consider the sequence $X_n = nG$. We have $\phi_{X_n}(t) = \phi_{nG}(t) = \phi_G(nt) = e^{-n^2t^2/2}$, so

$$\phi_{X_n}(t) \to \begin{cases} 0, & t \neq 0, \\ 1, & t = 1. \end{cases}$$

The limiting function is discontinuous at 0. The sequence X_n does not converge in distribution because $F_{X_n}(t) = \mathbb{P}(G \leq t/n) \to \mathbb{P}(G \leq 0) = 1/2$, but the limit is not a distribution function (an alternative argument: by Remark 5.15, if $X_n \xrightarrow{d} X$, then ϕ_{X_n} would converge to a characteristic function which is continuous).

We prove these results in the next two subsections.

Inversion formulae

En route to proving Theorem 5.13, we establish an inversion formula, quite standard in Fourier analysis. We first need a lemma.

5.17 Lemma. *For two independent random variables X and Y and every $t \in \mathbb{R}$, we have*

$$\mathbb{E}e^{-itY}\phi_X(Y) = \mathbb{E}\phi_Y(X - t).$$

Proof. Changing the order of taking expectation, we have

$$\mathbb{E}_Y e^{-itY}\phi_X(Y) = \mathbb{E}_Y e^{-itY}\mathbb{E}_X e^{iYX} = \mathbb{E}_{X,Y} e^{iY(X-t)} = \mathbb{E}_X\mathbb{E}_Y e^{iY(X-t)} = \mathbb{E}_X\phi_Y(X - t).$$

□

5.18 Theorem (Inversion formula)**.** *For a random variable X, at every point x of continuity of its distribution function F_X, we have*

$$F_X(x) = \lim_{a\to\infty} \int_{-\infty}^{x} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ist}\phi_X(s)e^{-\frac{s^2}{2a^2}}\,\mathrm{d}s \right) \mathrm{d}t.$$

Proof. Let G be a standard Gaussian random variable, independent of X. For $a > 0$, consider $X_a = X + a^{-1}G$. Since X_a converges pointwise to X as $a \to \infty$, by Lebesgue's dominated convergence theorem $\mathbb{E}g(X_a) \to \mathbb{E}g(X)$ for every bounded continuous function g, thus $X_a \xrightarrow{d} 0$ as $a \to \infty$ (Theorem 5.5). Consequently, for every continuity point x of F_X, we have

$$F_X(x) = \lim_{a\to\infty} F_{X_a}(x).$$

Let us find the distribution function of X_a. We have,

$$F_{X_a}(x) = \mathbb{P}\left(X + a^{-1}G \leq x\right) = \mathbb{E}_{X,G}\mathbf{1}_{\{X+a^{-1}G\leq x\}} = \mathbb{E}_X\mathbb{E}_G\mathbf{1}_{\{X+a^{-1}G\leq x\}}$$
$$= \mathbb{E}_X\mathbb{P}\left(X + a^{-1}G \leq x\right).$$

For any $y \in \mathbb{R}$, the density of $y + a^{-1}G$ at t is $\frac{a}{\sqrt{2\pi}}e^{-a^2(t-y)^2/2}$, thus

$$F_{X_a}(x) = \mathbb{E}_X \int_{-\infty}^{x} \frac{a}{\sqrt{2\pi}} e^{-a^2(t-X)^2/2}\mathrm{d}t = \int_{-\infty}^{x} \mathbb{E}_X \frac{a}{\sqrt{2\pi}} e^{-a^2(t-X)^2/2}\mathrm{d}t.$$

Note that $e^{-a^2s^2/2}$ is the characteristic function of aG at s (Example 5.11), so by Lemma 5.17,

$$\mathbb{E}_X \frac{a}{\sqrt{2\pi}} e^{-a^2(t-X)^2/2} = \frac{a}{\sqrt{2\pi}}\mathbb{E}_X\phi_{aG}(X-t) = \frac{a}{\sqrt{2\pi}}\mathbb{E}e^{-itaG}\phi_X(aG).$$

Writing this explicity using the density of aG yields

$$\frac{a}{\sqrt{2\pi}}\mathbb{E}e^{-itaG}\phi_X(aG) = \frac{a}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}a}\int_{-\infty}^{\infty} e^{-its}\phi_X(s)e^{-\frac{s^2}{2a^2}}\mathrm{d}s$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ist}\phi_X(s)e^{-\frac{s^2}{2a^2}}\mathrm{d}s.$$

Plugging this back,

$$F_{X_a}(x) = \int_{-\infty}^{x}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ist}\phi_X(s)e^{-\frac{s^2}{2a^2}}\mathrm{d}s\right)\mathrm{d}t,$$

which combined with $F_X(x) = \lim_{a\to\infty} F_{X_a}(x)$ remarked earlier finishes the proof. □

Now we can prove that characteristic functions determine distribution.

Proof of Theorem 5.13. By Theorem 5.18, $F_X(x) = F_Y(x)$ for every $x \in \mathbb{R} \setminus B$, where B is the union of the discontinuity points of F_X and the discontinuity points of F_Y. For $x \in B$, take $x_n > x$ such that $x_n \in \mathbb{R} \setminus B$ and $x_n \to x$ (it is possible since B is at most countable). Then $F_X(x_n) = F_Y(x_n)$ and by right-continuity, $F_X(x) = F_Y(x)$. □

The inversion formula from Theorem 5.18 gives us several other interesting corollaries. Since the characteristic function determines distribution, it should be possible to reconstruct densities from characteristic functions.

5.19 Theorem. *If X is a random variable such that $\int_{\mathbb{R}} |\phi_X| < \infty$, then X has density f given by*

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-isx}\phi_X(s)\mathrm{d}s$$

which is bounded and uniformly continuous.

5.20 Remark. If X is a continuous random variable with density f, then clearly

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{its} f(s)\mathrm{d}s$$

The two formulae have the same form!

Proof. For two continuity points $x < y$ of F_X, we have from Theorem 5.18,

$$F_X(y) - F_X(x) = \lim_{a\to\infty} \int_x^y \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ist} \phi_X(s) e^{-\frac{s^2}{2a^2}} \mathrm{d}s \right) \mathrm{d}t.$$

Since $|e^{-ist}\phi_X(s)e^{-\frac{s^2}{2a^2}}| \le |\phi_X(s)|$, that is the integrand is dominated by $|\phi_X|$ which is integrable on $[x, y] \times \mathbb{R}$, by Lebesgue's dominated convergence theorem,

$$F_X(y) - F_X(x) = \int_x^y \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ist} \phi_X(s) \mathrm{d}s \right) \mathrm{d}t$$

which gives that X has density given by the promised formula. The rest follows as for characteristic functions (recall the proof of Theorem 5.10 (iii)). □

5.21 Corollary. *If X is a continuous random variable with density f_X and characteristic function ϕ_X which is nonnegative, then $\int_{\mathbb{R}} \phi_X < \infty$ if and only if f is bounded.*

Proof. If $\int_{\mathbb{R}} \phi_X < \infty$, then by Theorem 5.19, f is bounded. Conversely, let as in the proof of Theorem 5.19, G be a standard Gaussian random variable independent of X. Then the density of $X + a^{-1}G$ at x equals

$$\int_{\mathbb{R}} f_X(x - y) f_{a^{-1}G}(y) \mathrm{d}y.$$

On the other hand, it equals $\frac{\mathrm{d}}{\mathrm{d}x} F_{X_a}(x)$ and from the last identity in the proof of Theorem 5.19, this becomes

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isx} \phi_X(s) e^{-\frac{s^2}{2a^2}} \mathrm{d}s.$$

For $x = 0$ we thus get

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(s) e^{-\frac{s^2}{2a^2}} \mathrm{d}s = \int_{\mathbb{R}} f_X(-y) f_{a^{-1}G}(y) \mathrm{d}y.$$

If f_X is bounded by, say M, we obtain that the right hand side is bounded by M, so

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(s) e^{-\frac{s^2}{2a^2}} \mathrm{d}s \le M.$$

As $a \to \infty$, by Lebesgue's monotone convergence theorem, the left hand side converges to $\frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X$, which proves that $\int_{\mathbb{R}} \phi_X \le 2\pi M$. □

5.22 Example. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables uniform on $[-1, 1]$. Then $X_1 + \ldots + X_n$ for $n \ge 2$ has density

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cos(tx) \left(\frac{\sin t}{t} \right)^n \mathrm{d}t.$$

Indeed, note that $\phi_{X_i}(t) = \frac{\sin t}{t}$, so $\phi_{X_1+\ldots+X_n}(t) = \left(\frac{\sin t}{t}\right)^n$ which is integrable for $n \ge 2$ and the formula follows from Theorem 5.19.

We finish with two Fourier analytic identities.

5.23 Theorem (Parseval's identities)**.** *If X and Y are continuous random variables with densities f_X and f_Y, then*

(i) $\int_{\mathbb{R}} |\phi_X|^2 < \infty$ *if and only if* $\int_{\mathbb{R}} f_X^2 < \infty$ *and then*

$$\int_{\mathbb{R}} f_X^2 = \frac{1}{2\pi} \int_{\mathbb{R}} |\phi_X|^2,$$

(ii) if $\int_{\mathbb{R}} f_X^2 < \infty$ *and* $\int_{\mathbb{R}} f_Y^2 < \infty$*, then*

$$\int_{\mathbb{R}} f_X f_Y = \frac{1}{2\pi} \int_{\mathbb{R}} \phi_X \overline{\phi_Y}.$$

Proof. (i) Let X' be an independent copy of X. Consider $\tilde{X} = X - X'$. We have,

$$\phi_{\tilde{X}}(t) = \phi_X(t)\phi_{-X'}(t) = \phi_X(t)\phi_{-X}(t) = \phi_X(t)\overline{\phi_X(t)} = |\phi_X(t)|^2.$$

On the other hand, $\tilde{X}$ is continuous with density given by convolution,

$$f_{\tilde{X}}(y) = (f_X \star f_{-X})(y) = \int_{\mathbb{R}} f_X(x) f_{-X}(y-x)\mathrm{d}x.$$

It can be seen from here that if $\int f_X^2 < \infty$, then by the Cauchy-Schwarz inequality, $f_{\tilde{X}}$ is bounded. Then by Corollary 5.21, $\phi_{\tilde{X}} = |\phi_X|^2$ is integrable. Conversely, if $|\phi_X|^2$ is integrable, then from Theorem 5.19 applied to $\tilde{X}$, we get

$$f_{\tilde{X}}(0) = \frac{1}{2\pi} \int_{\mathbb{R}} \phi_{\tilde{X}} = \frac{1}{2\pi} \int_{\mathbb{R}} |\phi_X|^2.$$

Since

$$f_{\tilde{X}}(0) = (f_X \star f_{-X})(0) = \int_{\mathbb{R}} f_X(x) f_{-X}(0-x)\mathrm{d}x = \int_{\mathbb{R}} f_X(x) f_X(x)\mathrm{d}x = \int_{\mathbb{R}} f_X^2,$$

we get that $\int f_X^2 = \frac{1}{2\pi} \int |\phi_X|^2$. In particular, f_X^2 is integrable.

(ii) Apply (i) to the density $\frac{f_X + f_Y}{2}$. □

Characteristic functions and convergence in distribution

Our goal is to prove Theorem 5.14. We start with a lemma that will help us get tightness.

5.24 Lemma. *For a random variable X and $\delta > 0$,*

$$\mathbb{P}\left(|X| > \frac{2}{\delta}\right) \leq \frac{1}{\delta} \int_{-\delta}^{\delta} [1 - \phi_X(t)]\mathrm{d}t.$$

Proof. Note that

$$\int_{-\delta}^{\delta} [1 - \phi_X(t)]\mathrm{d}t = \int_{-\delta}^{\delta} [1 - \mathbb{E}e^{itX}]\mathrm{d}t = 2\delta - \mathbb{E}\int_{-\delta}^{\delta} e^{itX}\mathrm{d}t = 2\delta - \mathbb{E}\frac{e^{i\delta X} - e^{-i\delta X}}{iX}$$
$$= 2\delta - 2\mathbb{E}\frac{\sin(\delta X)}{X}$$

(this incidentally shows that the a priori complex number $\int_{-\delta}^{\delta}[1-\phi_X(t)]\mathrm{d}t$ is real). Thus

$$\frac{1}{\delta}\int_{-\delta}^{\delta}[1-\phi_X(t)]\mathrm{d}t = 2\mathbb{E}\left[1-\frac{\sin(\delta X)}{\delta X}\right].$$

Using $|\sin x| \le |x|$, we have $1-\frac{\sin x}{x} \ge 0$, so

$$\begin{aligned}\frac{1}{\delta}\int_{-\delta}^{\delta}[1-\phi_X(t)]\mathrm{d}t &\ge 2\mathbb{E}\left[\left(1-\frac{\sin(\delta X)}{\delta X}\right)\mathbf{1}_{\{|\delta X|>2\}}\right]\\ &= 2\mathbb{E}\left[\left(1-\frac{\sin(\delta|X|)}{\delta|X|}\right)\mathbf{1}_{\{|\delta X|>2\}}\right],\end{aligned}$$

where is the last equality we used that $\frac{\sin x}{x}$ is even. Crudely, $-\sin(\delta|X|) \ge -1$, hence

$$\begin{aligned}\frac{1}{\delta}\int_{-\delta}^{\delta}[1-\phi_X(t)]\mathrm{d}t &\ge 2\mathbb{E}\left[\left(1-\frac{1}{\delta|X|}\right)\mathbf{1}_{\{|\delta X|>2\}}\right] \ge 2\mathbb{E}\left[\frac{1}{2}\mathbf{1}_{\{|\delta X|>2\}}\right]\\ &= \mathbb{P}\left(|\delta X|>2\right).\end{aligned}$$

□

Proof of Theorem 5.14. Since $|\phi_{X_n}(t)| \le 1$ for every t, we have the same for the limit, $|\phi(t)| \le 1$ for every t.

Step 1 (tightness). Since ϕ is continuous at 0 and $\phi(0) = \lim_n \phi_{X_n}(0) = 1$, for every $\varepsilon > 0$, there is $\delta > 0$ such that $|1-\phi(t)| < \varepsilon$ for $|t| < \delta$, so

$$\frac{1}{\delta}\int_{-\delta}^{\delta}|1-\phi(t)|\mathrm{d}t \le 2\varepsilon.$$

By Lebesgue's dominated convergence theorem,

$$\frac{1}{\delta}\int_{-\delta}^{\delta}|1-\phi_{X_n}(t)|\mathrm{d}t \xrightarrow[n\to\infty]{} \frac{1}{\delta}\int_{-\delta}^{\delta}|1-\phi(t)|\mathrm{d}t,$$

so for large n,

$$\frac{1}{\delta}\int_{-\delta}^{\delta}|1-\phi_{X_n}(t)|\mathrm{d}t < 3\varepsilon.$$

By Lemma 5.24, we obtain

$$\mathbb{P}\left(|X_n| > \frac{2}{\delta}\right) < \varepsilon.$$

This shows that the sequence (X_n) is tight. By Theorem 5.9, there is a subsequence (X_{n_k}) which converges in distribution to a random variable, say X. This is our candidate for the limit of (X_n).

Step 2 ($\phi = \phi_X$). Since $X_{n_k} \xrightarrow{d} X$, we get $\phi_{X_{n_k}} \to \phi_X$ at every point, but also $\phi_{X_{n_k}} \to \phi$ at every point, so $\phi = \phi_X$, which proves that ϕ is a characteristic function.

Step 3 ($X_n \xrightarrow{d} X$). If this is not the case, then, by Theorem 5.5, there is a bounded continuous function g such that $\mathbb{E}g(X_n) \nrightarrow \mathbb{E}g(X)$. Therefore, there is $\varepsilon > 0$ and a

sequence m_k such that $|\mathbb{E}g(X_{m_k}) - \mathbb{E}g(X)| > \varepsilon$. Since (X_n) is tight, using Theorem 5.9 again, there is a convergent subsequence $X_{m_{k_l}}$ to some random variable, say X'. As in Step 2, $\phi_{X'} = \phi = \phi_X$, so X' has the same distribution as X (Theorem 5.13) and $|\mathbb{E}g(X_{m_{k_l}}) - \mathbb{E}g(X')| = |\mathbb{E}g(X_{m_{k_l}}) - \mathbb{E}g(X)| > \varepsilon$ contradicts that $X_{m_{k_l}} \xrightarrow{d} X'$. □

5.3 Proof of the central limit theorem

We shall need several elementary lemmas about complex numbers.

5.25 Lemma. *If $z_1, \ldots, z_n$ and $w_1, \ldots, w_n$ are complex numbers all with modulus at most θ, then*

$$\left| \prod_{j=1}^{n} z_j - \prod_{j=1}^{n} w_j \right| \leq \theta^{n-1} \sum_{j=1}^{n} |z_j - w_j|.$$

Proof. We proceed by induction on n. For $n = 1$, we have equality. For $n > 1$, we have

$$\begin{aligned}
\left| \prod_{j=1}^{n} z_j - \prod_{j=1}^{n} w_j \right| &= \left| z_1 \prod_{j=2}^{n} z_j - w_1 \prod_{j=2}^{n} w_j \right| \\
&\leq \left| z_1 \prod_{j=2}^{n} z_j - z_1 \prod_{j=2}^{n} w_j \right| + \left| z_1 \prod_{j=2}^{n} w_j - w_1 \prod_{j=2}^{n} w_j \right| \\
&= |z_1| \left| \prod_{j=2}^{n} z_j - \prod_{j=2}^{n} w_j \right| + \left| \prod_{j=2}^{n} w_j \right| |z_1 - w_1| \\
&\leq \theta \left| \prod_{j=2}^{n} z_j - \prod_{j=2}^{n} w_j \right| + \theta^{n-1} |z_1 - w_1|
\end{aligned}$$

and the inductive assumption allows to finish the proof. □

5.26 Lemma. *For a complex number z with $|z| \leq 1$, we have*

$$|e^z - (1 + z)| \leq |z|^2.$$

Proof. Using the power series expansion of e^z, we get

$$\begin{aligned}
|e^z - (1 + z)| = \left| \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots \right| \leq |z|^2 \left(\frac{1}{2!} + \frac{|z|}{3!} + \ldots \right) &\leq |z|^2 \left(\frac{1}{2!} + \frac{1}{3!} + \ldots \right) \\
&= |z|^2 (e - 2).
\end{aligned}$$

□

5.27 Lemma. *If (z_n) is a sequence of complex numbers such that $z_n \to z$ for some $z \in \mathbb{C}$, then*

$$\left(1 + \frac{z_n}{n}\right)^n \to e^z.$$

Proof. Fix $c > |z|$. Then eventually, $|z_n| < c$ and consequently, $\left|1 + \frac{z_n}{n}\right| \le 1 + \frac{c}{n} \le e^{c/n}$ and $|e^{z_n/n}| = e^{\mathrm{Re}(z_n)/n} \le e^{c/n}$, so applying Lemma 5.25 with $\theta = e^{c/n}$, for large n,

$$\left|\left(1 + \frac{z_n}{n}\right)^n - e^{z_n}\right| = \left|\prod_{j=1}^n \left(1 + \frac{z_n}{n}\right) - \prod_{j=1}^n e^{z_n/n}\right| \le \left(e^{c/n}\right)^{n-1} n \left|1 + \frac{z_n}{n} - e^{z_n/n}\right|.$$

Clearly eventually, $|z_n/n| \le 1$, so by Lemma 5.26,

$$\left|\left(1 + \frac{z_n}{n}\right)^n - e^{z_n}\right| \le \left(e^{c/n}\right)^{n-1} n \left|\frac{z_n}{n}\right|^2 \le e^c \frac{c^2}{n}.$$

It remains to use continuity, that is that $e^{z_n} \to e^z$. □

We are ready to give a complete proof of the central limit theorem.

Proof of Theorem 5.1. Let $\bar{X}_i = \frac{X_i - \mathbb{E}X_i}{\mathrm{Var}(X_1)}$. Then $\mathbb{E}\bar{X}_i = 0$, $\mathbb{E}|\bar{X}_i|^2 = 1$,

$$Z_n = \frac{X_1 + \ldots + X_n - n\mathbb{E}X_1}{\sqrt{n\,\mathrm{Var}(X_1)}} = \frac{\bar{X}_1 + \ldots + \bar{X}_n}{\sqrt{n}}$$

and by independence

$$\phi_{Z_n}(t) = \phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right) \ldots \phi_{\bar{X}_n}\left(\frac{t}{\sqrt{n}}\right) = \left[\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n.$$

We investigate pointwise convergence of ϕ_{Z_n}. By Theorem 5.10 (iv), $\phi_{\bar{X}_1}$ is twice continuously differentiable and we can compute that $\phi'_{\bar{X}_1}(0) = i\mathbb{E}\bar{X}_1 = 0$ and $\phi''_{\bar{X}_1}(0) = i^2\mathbb{E}\bar{X}_1^2 = -1$. Thus by Taylor's formula with Lagrange's remainder

$$\begin{aligned}\phi_{\bar{X}_1}(t) &= \phi_{\bar{X}_1}(0) + t\phi'_{\bar{X}_1}(0) + \frac{t^2}{2}\phi''_{\bar{X}_1}(\xi_t) \\ &= 1 + t\phi'_{\bar{X}_1}(0) + \frac{t^2}{2}\phi''_{\bar{X}_1}(0) + t^2 R(t) \\ &= 1 - \frac{t^2}{2} + t^2 R(t),\end{aligned}$$

for some ξ_t between 0 and t and $R(t) = \frac{1}{2}(\phi''_{\bar{X}_1}(\xi_t) - \phi''_{\bar{X}_1}(0))$. By the continuity of $\phi''_{\bar{X}_1}$ (at 0), $R(t) \xrightarrow[t\to 0]{} 0$. Note that $R(t)$ may be complex. By Lemma 5.27, for every $t \in \mathbb{R}$,

$$\phi_{Z_n}(t) = \left[\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 - \frac{t^2}{2n} + \frac{t^2}{n}R(t)\right]^n \xrightarrow[n\to\infty]{} e^{-t^2/2}.$$

By Theorem 5.14, Z_n converges in distribution to a random variable whose characteristic function is $e^{-t^2/2}$, that is a standard Gaussian random variable. □

5.4 Poisson limit theorem

The following result, sometimes called the law of rare events, explains how the Poisson distribution arises as a limit of the binomial distribution when the expected number of successes converges to a constant as the number of Bernoulli trials goes to infinity.

5.28 Theorem (Poisson limit theorem)**.** *Let a sequence of numbers $p_n \in [0,1]$ be such that $np_n \xrightarrow[n\to\infty]{} \lambda$ for some $\lambda > 0$. Let S_n be a binomial random variable with parameters p_n and n. Then $S_n \xrightarrow{d} X$, where X is a Poisson random variable with parameter λ.*

Proof. For nonnegative integer-valued random variables convergence in distribution is equivalent to the pointwise convergence of the probability mass functions (homework!). Thus $S_n \xrightarrow{d} X$ if and only if $\mathbb{P}(S_n = k) \xrightarrow[n\to\infty]{} \mathbb{P}(X = k)$, for every integer $k \geq 0$. Fix then such k and note that as $n \to \infty$, we have

$$\begin{aligned}\mathbb{P}(S_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)\dots(n-k+1)}{k!} p_n^k (1-p_n)^{n-k} \\ &= \frac{1+O(n^{-1})}{k!} (np_n)^k (1-p_n)^{n-k}.\end{aligned}$$

By the assumption, $np_n \to \lambda$. In particular, $p_n \to 0$. Consequently, $(1-p_n)^{-k} \to 1$ and $(1-p_n)^n \to e^{-\lambda}$, so

$$\mathbb{P}(S_n = k) \xrightarrow[n\to\infty]{} \frac{1}{k!}\lambda^k e^{-\lambda} = \mathbb{P}(X = k).$$

□

6 Quantitative versions of the central limit theorem

6.1 Berry-Esseen theorem via Stein's mehtod

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite variance. Let $Z_n = \frac{X_1+\ldots+X_n-n\mathbb{E}X_1}{\sqrt{n\,\mathrm{Var}(X_1)}}$ and let Z be a standard Gaussian random variable. The central limit theorem asserts that for every $t \in \mathbb{R}$,

$$\mathbb{P}\left(Z_n \leq t\right) \xrightarrow[n\to\infty]{} \mathbb{P}\left(Z \leq t\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} \mathrm{d}x.$$

For practical purposes, we would like to know what is the error we make when we use $\mathbb{P}\left(Z \leq t\right)$ as an approximation to $\mathbb{P}\left(Z_n \leq t\right)$ for large n. This is possible under an additional assumption (finite third moment) and is settled in the following theorem, discovered independently by Berry and Esseen.

6.1 Theorem (Berry-Esseen theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}|X_1|^3 < \infty$. Let*

$$Z_n = \frac{X_1 + \ldots + X_n - n\mathbb{E}X_1}{\sqrt{n\,\mathrm{Var}(X_1)}},$$
$$\rho = \mathbb{E}\left|\frac{X_1 - \mathbb{E}X_1}{\sqrt{\mathrm{Var}(X_1)}}\right|^3$$

and let Z be a standard Gaussian random variable. There is a universal constant C such that for every $n \geq 1$ and every $t \in \mathbb{R}$, we have

$$\left|\mathbb{P}\left(Z_n \leq t\right) - \mathbb{P}\left(Z \leq t\right)\right| \leq \frac{C\rho}{\sqrt{n}}.$$

6.2 Remark. We present a proof which will give $C = 15.2$, but this value is far from optimal. Currently, the best value is $C = 0.4774$ (estblished via Fourier analytic methods in [3]). Esseen proved a lower bound: $C \geq \frac{10+\sqrt{3}}{6\sqrt{2\pi}} = 0.4097\ldots$

6.3 Remark. The rate $1/\sqrt{n}$ of the error is optimal. Consider i.i.d. symmetric random signs $\varepsilon_1, \varepsilon_2, \ldots$ and let $Z_n = \frac{\varepsilon_1+\ldots+\varepsilon_n}{\sqrt{n}}$. For even n, by symmetry, we have

$$\mathbb{P}\left(Z_n \leq 0\right) = \frac{1 + \mathbb{P}\left(\varepsilon_1 + \ldots + \varepsilon_n = 0\right)}{2} = \frac{1}{2} + \frac{1}{2}\binom{n}{n/2}\frac{1}{2^n},$$

thus, thanks to Stirling's formula,

$$\left|\mathbb{P}\left(Z_n \leq 0\right) - \mathbb{P}\left(Z \leq 0\right)\right| = \left|\mathbb{P}\left(Z_n \leq 0\right) - \frac{1}{2}\right| = \frac{1}{2}\binom{n}{n/2}\frac{1}{2^n} \approx \frac{1}{2}\frac{\sqrt{2}}{\sqrt{\pi n}},$$

so in this case the error is of the order $1/\sqrt{n}$.

For the proof the Berry-Esseen theorem, we shall need the following elementary tail bound for the standard Gaussian distribution.

6.4 Lemma. *For $x > 0$, we have*

(i) $\int_x^\infty e^{-u^2/2}\mathrm{d}u \le \sqrt{\frac{\pi}{2}}e^{-x^2/2}$,

(ii) $\int_x^\infty e^{-u^2/2}\mathrm{d}u \le \frac{1}{x}e^{-x^2/2}$.

Proof. (i) let $f(x) = \sqrt{\frac{\pi}{2}}e^{-x^2/2} - \int_x^\infty e^{-u^2/2}\mathrm{d}u$. Since $f'(x) = \left(1 - x\sqrt{\frac{\pi}{2}}\right)e^{-x^2/2}$ is first positive, then negative, f first increases, then decreases. Combined with $f(0) = 0$ and $f(x) \xrightarrow[t\to\infty]{} 0$, this proves that $f(x) \ge 0$.

(ii) We have $\int_x^\infty xe^{-u^2/2}\mathrm{d}u \le \int_x^\infty ue^{-u^2/2}\mathrm{d}u = e^{-u^2/2}$. □

Proof of Theorem 6.1. For $t, x \in \mathbb{R}$ and $\lambda > 0$ define functions

$$h_t(x) = \mathbf{1}_{(-\infty,t]}(x),$$

and their continuous linear approximations

$$h_{t,\lambda}(x) = \begin{cases} 1, & x \le t, \\ 1 - \frac{x-t}{\lambda}, & t < x \le t+\lambda, \\ 0, & x > t+\lambda. \end{cases}$$

We will frequently use the following integral representation

$$h_{t,\lambda}(x) = \int_x^\infty \frac{1}{\lambda}\mathbf{1}_{(t,t+\lambda)}(s)\mathrm{d}s.$$

Given $\gamma \ge 1$, define the class of random variables

$$\mathscr{L}_\gamma = \{X,\ X \text{ is random variable such that } \mathbb{E}X = 0,\ EX^2 = 1,\ \mathbb{E}|X|^3 = \gamma\}$$

and for $n = 1, 2, \dots$ define two quantities

$$B_0(\gamma, n) = \sup_{X_1,\dots,X_n i.i.d., X_i \in \mathscr{L}_\gamma} \sup_{t\in\mathbb{R}} |\mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z)|,$$

$$B(\lambda, \gamma, n) = \sup_{X_1,\dots,X_n i.i.d., X_i \in \mathscr{L}_\gamma} \sup_{t\in\mathbb{R}} |\mathbb{E}h_{t,\lambda}(Z_n) - \mathbb{E}h_{t,\lambda}(Z)|.$$

Plainly, $\mathbb{P}(X \le t) = \mathbb{E}\mathbf{1}_{X\le t} = \mathbb{E}h_t(X)$, so to prove the theorem, we would like to show that

$$\frac{\sqrt{n}}{\gamma}B_0(\gamma, n) \le C, \qquad n \ge 1, \gamma \ge 1.$$

This is clear for $n = 1$ with $C = 1$ because $|\mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z)| \le 1$, so from now on we assume $n \ge 2$ and divide the rest of the proof into several steps.

Step 1: regularisation (upper bound for B_0 in terms of B). Since $h_{t-\lambda} \le h_t \le h_{t,\lambda}$, we get

$$\begin{aligned} \mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z) &\le \mathbb{E}h_{t,\lambda}(Z_n) - \mathbb{E}h_t(Z) \\ &= \mathbb{E}h_{t,\lambda}(Z_n) - \mathbb{E}h_{t,\lambda}(Z) + \mathbb{E}h_{t,\lambda}(Z) - \mathbb{E}h_t(Z) \\ &\le \mathbb{E}h_{t,\lambda}(Z_n) - \mathbb{E}h_{t,\lambda}(Z) + \mathbb{E}h_{t+\lambda}(Z) - \mathbb{E}h_t(Z). \end{aligned}$$

Observe that the first difference is upper bounded by $B(t,\lambda,n)$ by its definition. The second difference is

$$\mathbb{P}\left(t < Z \le t+\lambda\right) = \int_t^{t+\lambda} e^{-x^2/2}\frac{\mathrm{d}x}{\sqrt{2\pi}} \le \int_t^{t+\lambda}\frac{\mathrm{d}x}{\sqrt{2\pi}} = \frac{\lambda}{\sqrt{2\pi}}.$$

Altogether,

$$\mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z) \le B(t,\lambda,n) + \frac{\lambda}{\sqrt{2\pi}}.$$

Similarly,

$$\mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z) \ge -B(t,\lambda,n) - \frac{\lambda}{\sqrt{2\pi}}.$$

Thus

$$B_0(\gamma,n) \le B(t,\lambda,n) + \frac{\lambda}{\sqrt{2\pi}}.$$

Step 2: Stein's method ("encoding" $\mathbb{E}h(Z)$ into a function). Fix $t\in\mathbb{R}$, $\lambda>0$ and set $h = h_{t,\lambda}$. Our goal is to upper bound B, so to upper bound $\mathbb{E}h(Z_n) - \mathbb{E}h(Z)$. The heart of Stein's method is to rewrite this in terms of Z_n only. Let

$$f(x) = e^{x^2/2}\int_{-\infty}^{x}[h(u)-\mathbb{E}h(Z)]e^{-u^2/2}\mathrm{d}u.$$

Then

$$f'(x) - xf(x) = h(x) - \mathbb{E}h(Z),$$

so

$$\mathbb{E}h(Z_n) - \mathbb{E}h(Z) = \mathbb{E}\big[f'(Z_n) - Z_n f(Z_n)\big]. \tag{6.1}$$

Step 3: Estimates for f and f'. For every $x\in\mathbb{R}$, we have

$$|f(x) \le \sqrt{\frac{\pi}{2}}, \qquad |xf(x)|\le 1, \qquad |f'(x)|\le 2 \tag{6.2}$$

and for every $x,y\in\mathbb{R}$, we have

$$|f'(x+y) - f'(x)| \le |y|\left(\sqrt{\frac{\pi}{2}} + 2|x| + \frac{1}{\lambda}\int_0^1 \mathbf{1}_{(t,t+\lambda)}(x+vy)\mathrm{d}v\right). \tag{6.3}$$

Indeed, since h takes values in $[0,1]$, we have $|h(u)-h(v)|\le 1$ for any u and v, so for $x<0$,

$$|f(x)| \le e^{x^2/2}\int_{-\infty}^{x}|h(u)-\mathbb{E}h(Z)|e^{-u^2/2}\mathrm{d}u \le e^{x^2/2}\int_{-\infty}^{x}e^{-u^2/2}\mathrm{d}u = e^{x^2/2}\int_{-x}^{\infty}e^{-u^2/2}\mathrm{d}u,$$

which by Lemma 6.4 (i) is upper bounded by $\sqrt{\frac{\pi}{2}}$. For $x>0$, notice that $\int_{-\infty}^{\infty}[h(u) - \mathbb{E}h(Z)]e^{-u^2/2}\frac{\mathrm{d}u}{\sqrt{2\pi}} = 0$, so

$$f(x) = -e^{x^2/2}\int_x^{\infty}[h(u)-\mathbb{E}h(Z)]e^{-u^2/2}\mathrm{d}u$$

and as above we get the bound $|f(x)| \leq \sqrt{\frac{\pi}{2}}$. To bound $xf(x)$ we proceed the same way but use Lemma 6.4 (ii). Finally, since $f'(x) = xf(x) + h(x) - \mathbb{E}h(Z)$ (Step 2), we get

$$|f'(x)| \leq |xf(x)| + |h(x) - \mathbb{E}h(Z)| \leq 1 + 1 = 2.$$

This establishes (6.2). To prove (6.3), we use the formula for f' from Step 2 and write

$$\begin{aligned}
|f'(x+y) - f'(x)| &= |(x+y)f(x+y) + h(x+y) - xf(x) - h(x)| \\
&= |yf(x+y) + x(f(x+y) - f(x)) + h(x+y) - h(x)| \\
&\leq |y|\sqrt{\frac{\pi}{2}} + 2|x||y| + |h(x+y) - h(x)|,
\end{aligned}$$

where in the last inequality we used the mean value theorem writing $f(x+y) - f(x) = f'(\xi)y$ and then estimating $|f'(\xi)| \leq 2$. Finally, by the integral representation for h,

$$|h(x+y) - h(x)| = \left|\frac{1}{\lambda}\int_x^{x+y} \mathbf{1}_{(t,t+\lambda)}(u)\mathrm{d}u\right| = \left|\frac{y}{\lambda}\int_0^1 \mathbf{1}_{(t,t+\lambda)}(x+vy)\mathrm{d}v\right|$$

which after plugging back in the previous inequality finishes the proof of (6.3).

Step 4: Estimates for $B(\lambda, \gamma, n)$ via (6.1). To estimate $B(\lambda, \gamma, n)$, we need to upper bound $\mathbb{E}h(Z_n) - \mathbb{E}h(Z) = \mathbb{E}[f'(Z_n) - Z_n f(Z_n)]$ (recall (6.1) from Step 2). Here we exploit that $Z_n = \frac{X_1 + \ldots + X_n}{\sqrt{n}}$ is a sum of i.i.d. random variables. Since the X_i have the same distribution, by linearity,

$$\mathbb{E}Z_n f(Z_n) = \mathbb{E}\frac{\sum X_i}{\sqrt{n}} f(Z_n) = \sqrt{n}\mathbb{E}X_n f(Z_n).$$

Note also that $Z_n = \sqrt{\frac{n-1}{n}}Z_n + \frac{X_n}{\sqrt{n}}$ and thus

$$\begin{aligned}
\mathbb{E}[f'(Z_n) - Z_n f(Z_n)] &= \mathbb{E}[f'(Z_n) - \sqrt{n}X_n f(Z_n)] \\
&= \mathbb{E}\Big[f'(Z_n) - \sqrt{n}X_n \int_0^1 \frac{\mathrm{d}}{\mathrm{d}u} f\left(\sqrt{\frac{n-1}{n}}Z_{n-1} + u\frac{X_n}{\sqrt{n}}\right)\mathrm{d}u \\
&\qquad - \sqrt{n}X_n f\left(\sqrt{\frac{n-1}{n}}Z_{n-1}\right)\Big]
\end{aligned}$$

By independence and $\mathbb{E}X_n = 0$ the last term vanishes and after computing the derivative we get

$$\begin{aligned}
\mathbb{E}[f'(Z_n) - Z_n f(Z_n)] &= \mathbb{E}\left[f'(Z_n) - X_n^2 \int_0^1 f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1} + u\frac{X_n}{\sqrt{n}}\right)\mathrm{d}u\right] \\
&= \mathbb{E}\left[f'(Z_n) - f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1}\right)\right] \\
&\quad + \mathbb{E}\Bigg[- X_n^2 \int_0^1 \Bigg\{f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1} + u\frac{X_n}{\sqrt{n}}\right) \\
&\qquad - f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1}\right)\Bigg\}\mathrm{d}u\Bigg]
\end{aligned}$$

where in the last equality we used independence and $\mathbb{E}X_n^2 = 1$. We bound the two terms separately.

Step 4.1: First term. Using $Z_n = \sqrt{\frac{n-1}{n}} Z_n + \frac{X_n}{\sqrt{n}}$ and (6.3),

$$\begin{aligned}
&\left| \mathbb{E}\left[f'(Z_n) - f'\left(\sqrt{\frac{n-1}{n}} Z_{n-1} \right) \right] \right| \\
&\leq \mathbb{E}\left| \frac{X_n}{\sqrt{n}} \right| \left(\sqrt{\frac{\pi}{2}} + 2\sqrt{\frac{n-1}{n}} |Z_{n-1}| + \frac{1}{\lambda} \int_0^1 \mathbf{1}_{(t,t+\lambda)} \left(\sqrt{\frac{n-1}{n}} Z_{n-1} + u \frac{X_n}{\sqrt{n}} \right) \mathrm{d}u \right)
\end{aligned}$$

Since $\mathbb{E}|X_n| \leq \sqrt{\mathbb{E}|X_n|^2} = 1$ and similarly $\mathbb{E}|Z_{n-1}| \leq 1$, as well as trivially $\sqrt{\frac{n-1}{n}} \leq 1$, we get

$$\begin{aligned}
&\left| \mathbb{E}\left[f'(Z_n) - f'\left(\sqrt{\frac{n-1}{n}} Z_{n-1} \right) \right] \right| \\
&\leq \frac{1}{\sqrt{n}} \sqrt{\frac{\pi}{2}} + 2\frac{1}{\sqrt{n}} + \frac{1}{\lambda\sqrt{n}} \mathbb{E}_{X_n} \left[|X_n| \int_0^1 \mathbb{E}_{Z_{n-1}} \mathbf{1}_{(t,t+\lambda)} \left(\sqrt{\frac{n-1}{n}} Z_{n-1} + u \frac{X_n}{\sqrt{n}} \right) \mathrm{d}u \right],
\end{aligned}$$

where in the last term we used the independence of X_n and Z_{n-1}. Note that

$$\begin{aligned}
&\mathbb{E}_{Z_{n-1}} \mathbf{1}_{(t,t+\lambda)} \left(\sqrt{\frac{n-1}{n}} Z_{n-1} + u \frac{X_n}{\sqrt{n}} \right) \\
&= \mathbb{P}_{Z_{n-1}} \left(\left(t - u\frac{X_n}{\sqrt{n}} \right) \sqrt{\frac{n}{n-1}} < Z_{n-1} < \left(t - u\frac{X_n}{\sqrt{n}} \right) \sqrt{\frac{n}{n-1}} + \lambda \sqrt{\frac{n}{n-1}} \right),
\end{aligned}$$

Denoting $a = \left(t - u\frac{X_n}{\sqrt{n}} \right) \sqrt{\frac{n}{n-1}}$ and estimating $\frac{n}{n-1} \leq 2$, we get that this probability is upper bounded by

$$\mathbb{P}\left(a < Z_{n-1} < a + \lambda\sqrt{2} \right)$$

which we rewrite in order to upper bound it in terms of B_0,

$$\begin{aligned}
\mathbb{P}\left(a < Z_{n-1} < a + \lambda\sqrt{2} \right) &= \mathbb{P}\left(Z_{n-1} < a + \lambda\sqrt{2} \right) - \mathbb{P}\left(Z < a + \lambda\sqrt{2} \right) \\
&\quad + \mathbb{P}\left(Z \leq a \right) - \mathbb{P}\left(Z_{n-1} \leq a \right) + \mathbb{P}\left(a \leq Z \leq a + \lambda\sqrt{2} \right) \\
&\leq 2B_0(\gamma, n-1) + \frac{\lambda\sqrt{2}}{\sqrt{2\pi}},
\end{aligned}$$

where the last term was crudely bounded using the maximum of standard Gaussian density. Plugging this back yields

$$\begin{aligned}
&\left| \mathbb{E}\left[f'(Z_n) - f'\left(\sqrt{\frac{n-1}{n}} Z_{n-1} \right) \right] \right| \\
&\leq \frac{1}{\sqrt{n}} \sqrt{\frac{\pi}{2}} + 2\frac{1}{\sqrt{n}} + \frac{1}{\lambda\sqrt{n}} \mathbb{E}\left[|X_n| \left(2B_0(\gamma, n-1) + \frac{\lambda}{\sqrt{\pi}} \right) \right] \\
&\leq \frac{1}{\sqrt{n}} \left(\sqrt{\frac{\pi}{2}} + 2 + \frac{2B_0(\gamma, n-1)}{\lambda} + \frac{1}{\sqrt{\pi}} \right) \\
&\leq \frac{\gamma}{\sqrt{n}} \left(\sqrt{\frac{\pi}{2}} + 2 + \frac{2B_0(\gamma, n-1)}{\lambda} + \frac{1}{\sqrt{\pi}} \right).
\end{aligned}$$

Step 4.2: Second term. Using again (6.3) and independence,

$$\begin{aligned}
&\left|\mathbb{E}\left[-X_n^2\int_0^1\left\{f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1}+u\frac{X_n}{\sqrt{n}}\right)-f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1}\right)\right\}\mathrm{d}u\right]\right| \\
&\le \mathbb{E}X_n^2\frac{|X_n|}{\sqrt{n}}\int_0^1 u\Bigg(\sqrt{\frac{\pi}{2}}+2\sqrt{\frac{n-1}{n}}|Z_{n-1}| \\
&\qquad\qquad +\frac{1}{\lambda}\int_0^1 \mathbf{1}_{(t,t+\lambda)}\left(\sqrt{\frac{n-1}{n}}Z_{n-1}+uv\frac{X_n}{\sqrt{n}}\right)\mathrm{d}v\Bigg)\mathrm{d}u \\
&\le \mathbb{E}\frac{|X_n|^3}{\sqrt{n}}\int_0^1 u\Bigg(\sqrt{\frac{\pi}{2}}+2\mathbb{E}_{Z_{n-1}}|Z_{n-1}| \\
&\qquad\qquad +\frac{1}{\lambda}\int_0^1 \mathbb{E}_{Z_{n-1}}\mathbf{1}_{(t,t+\lambda)}\left(\sqrt{\frac{n-1}{n}}Z_{n-1}+uv\frac{X_n}{\sqrt{n}}\right)\mathrm{d}v\Bigg)\mathrm{d}u \\
&\le \mathbb{E}\frac{|X_n|^3}{\sqrt{n}}\int_0^1 u\left(\sqrt{\frac{\pi}{2}}+2+\frac{1}{\lambda}\left(2B_0(\gamma,n-1)+\frac{\lambda}{\sqrt{\pi}}\right)\right)\mathrm{d}u \\
&= \frac{\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{2B_0(\gamma,n-1)}{\lambda}+\frac{1}{\sqrt{\pi}}\right).
\end{aligned}$$

Putting Steps 4.1 and 4.2 together yields

$$|\mathbb{E}[f'(Z_n)-Z_nf(Z_n)]|\le\frac{3\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{2B_0(\gamma,n-1)}{\lambda}+\frac{1}{\sqrt{\pi}}\right).$$

By Step 2, this gives

$$B(t,\lambda,n)\le\frac{3\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{2B_0(\gamma,n-1)}{\lambda}+\frac{1}{\sqrt{\pi}}\right).$$

Step 5: Optimisation of parameters and end of proof. The previous inequality and Step 1 yield

$$\begin{aligned}
B_0(\gamma,n)&\le\frac{3\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{2B_0(\gamma,n-1)}{\lambda}+\frac{1}{\sqrt{\pi}}\right)+\frac{\lambda}{\sqrt{2\pi}} \\
&=\frac{3\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{1}{\sqrt{\pi}}\right)+\frac{1}{\lambda}\frac{3\gamma B_0(\gamma,n-1)}{\sqrt{n}}+\frac{\lambda}{\sqrt{2\pi}}.
\end{aligned}$$

Set $\lambda=\alpha\frac{\gamma}{\sqrt{n}}$, $\alpha>0$ and multiply both sides by $\frac{\sqrt{n}}{\gamma}$ to get

$$B_0(\gamma,n)\frac{\sqrt{n}}{\gamma}\le\frac{3}{2}\left(\sqrt{\frac{\pi}{2}}+2+\frac{1}{\sqrt{\pi}}\right)+\frac{3}{\alpha}B_0(\gamma,n-1)\frac{\sqrt{n}}{\gamma}+\frac{\alpha}{\sqrt{2\pi}}.$$

Let

$$B=\sup_{\gamma\ge1,n\ge2}B_0(\gamma,n)\frac{\sqrt{n}}{\gamma}.$$

For $n\ge2$, we have

$$B_0(\gamma,n-1)\frac{\sqrt{n}}{\gamma}=B_0(\gamma,n-1)\frac{\sqrt{n-1}}{\gamma}\sqrt{\frac{n}{n-1}}\le\max\left\{\sqrt{2},B\sqrt{\frac{3}{2}}\right\}$$

(recall that trivially $B_0(\gamma,1)\frac{1}{\gamma}\le 1$). If $B>\frac{2}{\sqrt{3}}$, we thus obtain

$$B\le \frac{3}{2}\left(\sqrt{\frac{\pi}{2}}+2+\frac{1}{\sqrt{\pi}}\right)+\frac{3}{\alpha}B\sqrt{\frac{3}{2}}+\frac{\alpha}{\sqrt{2\pi}}.$$

For $\alpha>3\sqrt{\frac{3}{2}}$ this gives

$$B\le \frac{\alpha}{\alpha-3\sqrt{\frac{3}{2}}}\frac{3}{2}\left(\sqrt{\frac{\pi}{2}}+2+\frac{1}{\sqrt{\pi}}\right)+\frac{\alpha^2}{\alpha-3\sqrt{\frac{3}{2}}}\frac{1}{\sqrt{2\pi}}.$$

The choice of α which equates the two terms on the right hand side gives

$$B<15.4.$$

Optimising over α (which requires more computations) gives a slightly better estimate

$$B<15.2.$$

□

6.5 Remark. The proof presented here is from [1]. The heart of the argument is based on Stein's method (Step 2), introduced by Charles Stein, who developed this influential technique for teaching purposes of the central limit theorem for his course in statistics.

6.6 Example. Let us apply the Berry-Esseen theorem to i.i.d. Bernoulli random variables $X_1,\dots,X_n$ with parameter $0<p<1$. We have $\mathbb{E}X_i=p$, $\mathrm{Var}(X_i)=p(1-p)$ and we obtain for every real t and every integer $n\ge 1$

$$\left|\mathbb{P}\left(\frac{X_1+\ldots+X_n-np}{\sqrt{np(1-p)}}\le t\right)-\mathbb{P}\left(Z\le t\right)\right|\le C\frac{\rho}{\sqrt{n}},$$

where

$$\rho=\mathbb{E}\left|\frac{X_1-p}{\sqrt{p(1-p)}}\right|^3=\frac{p(1-p)^3+(1-p)p^3}{\sqrt{p(1-p)}^3}=\frac{1-2p(1-p)}{\sqrt{p(1-p)}}.$$

In particular, when np is of the constant order for large n, the Berry-Esseen theorem is not useful at all because the bound of the error, $C\frac{\rho}{\sqrt{n}}$ is of the order $\frac{C}{\sqrt{np(1-p)}}$ which is constant. This might suggest that the Gaussian approximation is not valid in this case, which is in fact true in view of the Poisson limit theorem (Theorem 5.28).

6.2 Local central limit theorem

In applications we often need to address the following: suppose $X_1,X_2,\dots$ are i.i.d. discrete, say integer-valued random variables and we would like to know for large n what is the approximate value of $\mathbb{P}\left(X_1+\ldots+X_n=x_n\right)$ for some $x_n\in\mathbb{Z}$. If $\mathbb{E}X_i^2<\infty$,

$\mu = \mathbb{E}X_1$, $\sigma^2 = \mathrm{Var}(X_1)$ and $\frac{x_n - n\mu}{\sqrt{n}} \approx y$ is of constant order for large n, by the central limit theorem,

$$\begin{aligned}
&\mathbb{P}\left(X_1 + \ldots + X_n = x_n\right) \\
&= \mathbb{P}\left(x_n - \frac{1}{2} < X_1 + \ldots + X_n < x_n + \frac{1}{2}\right) \\
&= \mathbb{P}\left(\frac{x_n - n\mu}{\sqrt{n}} - \frac{1}{2\sqrt{n}} < \frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}} < \frac{x_n - n\mu}{\sqrt{n}} + \frac{1}{2\sqrt{n}}\right) \\
&\approx \frac{1}{\sqrt{2\pi}\sigma} \int_{y - \frac{1}{2\sqrt{n}}}^{y + \frac{1}{2\sqrt{n}}} e^{-\frac{t^2}{2\sigma^2}} \mathrm{d}t \\
&\approx \frac{1}{\sqrt{n}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}},
\end{aligned}$$

obtaining the approximation for $\mathbb{P}\left(X_1 + \ldots + X_n = x_n\right)$ by the Gaussian density. To control the error in this approximation, we cannot simply use the Berry-Esseen theorem here because its error bound $O(\frac{1}{\sqrt{n}})$ is of the same order as the value of our approximation $\frac{1}{\sqrt{n}} \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2}$. The local central limit theorem addresses this deficiency. We only discuss the discrete case. There are also versions which give approximations to densities of sums of i.i.d. continuous random variables.

We shall use the common notation $a + b\mathbb{Z}$ for the set $\{a + bx,\ x \in \mathbb{Z}\}$.

6.7 Theorem (Local central limit theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. integer-valued random variables such that $\mathbb{E}X_1^2 < \infty$. Suppose X_i is not supported on any proper subprogression of $\mathbb{Z}$, that is there are no $r > 1$, $a \in \mathbb{R}$ such that $\mathbb{P}\left(X_i \in a + r\mathbb{Z}\right) = 1$. Denote $\mu = \mathbb{E}X_1$, $\sigma = \sqrt{\mathrm{Var}(X_1)}$ and*

$$p_n(x) = \mathbb{P}\left(\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}} = x\right), \qquad x \in \frac{\mathbb{Z} - n\mu}{\sqrt{n}}.$$

Then

$$\sup_{x \in \frac{\mathbb{Z} - n\mu}{\sqrt{n}}} \left| \sqrt{n} p_n(x) - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \right| \xrightarrow[n \to \infty]{} 0.$$

6.8 Lemma. *For an integer-valued random variable X and an integer k, we have*

$$\mathbb{P}\left(X = k\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \phi_X(t) \mathrm{d}t.$$

Proof. Note that for two integers k and l, we have

$$\mathbf{1}_{\{l=k\}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{it(l-k)} \mathrm{d}t.$$

Thus

$$\begin{aligned}
\mathbb{P}\left(X = k\right) = \mathbb{E}\mathbf{1}_{\{X=k\}} = \mathbb{E}\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{it(X-k)} \mathrm{d}t &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \mathbb{E}e^{itX} \mathrm{d}t \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \phi_X(t) \mathrm{d}t.
\end{aligned}$$

□

Proof of Theorem 6.7. Applying Lemma 6.8 to $X_1+\ldots+X_n$ and changing the variables yields

$$\begin{aligned} p_n(x) &= \mathbb{P}\left(X_1+\ldots+X_n = x\sqrt{n}+n\mu\right) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-it(x\sqrt{n}+n\mu)}\phi_{X_1+\ldots+X_n}(t)\mathrm{d}t \\ &= \frac{1}{\sqrt{n}}\frac{1}{2\pi}\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} e^{-itx}\left[e^{-i\frac{t}{\sqrt{n}}\mu}\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n \mathrm{d}t. \end{aligned}$$

Using that the characteristic function of a centred Gaussian random variable with variance $1/\sigma^2$ is $e^{-\frac{x^2}{2\sigma^2}}$, we have

$$e^{-\frac{x^2}{2\sigma^2}} = \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx}e^{-t^2\sigma^2/2}\mathrm{d}t,$$

which gives (by symmetry, we can write e^{-itx} instead of e^{itx})

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}e^{-t^2\sigma^2/2}\mathrm{d}t.$$

Therefore,

$$\begin{aligned} \left|\sqrt{n}p_n(x) - \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}\right| \le{}& \frac{1}{2\pi}\int_{|t|\le\pi\sqrt{n}}\left|\phi_{X_1-\mu}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2\sigma^2/2}\right|\mathrm{d}t \\ &+ \frac{1}{2\pi}\int_{|t|\ge\pi\sqrt{n}} e^{-t^2\sigma^2/2}\mathrm{d}t. \end{aligned}$$

Since the right hand side does not depend on x, we need to show that it converges to 0 as $n\to\infty$. The second integral clearly does. To deal with the first integral, we change the variables

$$\int_{|t|\le\pi\sqrt{n}}\left|\phi_{X_1-\mu}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2\sigma^2/2}\right|\mathrm{d}t = \frac{1}{\sigma}\int_{|t|\le\pi\sigma\sqrt{n}}\left|\phi_{\frac{X_1-\mu}{\sigma}}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right|\mathrm{d}t,$$

let $\bar{X}_1 = \frac{X_1-\mu}{\sigma}$ (which has mean 0 and variance 1) and break it into two pieces

$$\int_{|t|\le\varepsilon\sqrt{n}}\left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right|\mathrm{d}t + \int_{\varepsilon\sqrt{n}\le|t|\le\pi\sigma\sqrt{n}}\left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right|\mathrm{d}t. \quad (6.4)$$

Recall from the proof of the central limit theorem that

$$\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n \to e^{-t^2/2}$$

and by Taylor's formula,

$$\phi_{\bar{X}_1}(t) = 1 - \frac{t^2}{2} + t^2R(t),$$

for some (complex-valued) function R such that $R(t)\to 0$ as $t\to 0$. Choose $\varepsilon<1$ such that $|R(t)|<\frac{1}{4}$ for all $|t|<\varepsilon$. Then for $|t|\le\varepsilon\sqrt{n}$,

$$\left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)\right| \le \left|1-\frac{t^2}{2n}\right| + \frac{t^2}{4n} = 1 - \frac{t^2}{4n} \le e^{-\frac{t^2}{4n}},$$

so

$$\left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right| \leq e^{-t^2/4} + e^{-t^2/2}.$$

By Lebesgue's dominated convergence theorem, the first piece in (6.4) converges to 0 as $n \to \infty$. Finally, to handle the second piece, we claim that: $|\phi_{\bar{X}_1}(t)| < c_\varepsilon$ for all $\varepsilon \leq |t| \leq \pi\sigma$ for some constant $c_\varepsilon < 1$. This suffices because then

$$\int_{\varepsilon\sqrt{n}\leq|t|\leq\pi\sigma\sqrt{n}} \left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right| \mathrm{d}t \leq \int_{\varepsilon\sqrt{n}\leq|t|\leq\pi\sigma\sqrt{n}} (c_\varepsilon^n + e^{-t^2/2})\mathrm{d}t$$

and the right hand side clearly goes to 0 as $n \to \infty$. Now we use that X_1 is integer-valued, not concentrated on any proper subprogreesion to show the claim. Since X_1 is integer-valued, ϕ_{X_1} is 2π-periodic and in particular $\phi_{X_1}(2\pi) = 1$. Moreover, $|\phi_{X_1}(t)| < 1$ for all $0 < t < 2\pi$. Otherwise, if $|\phi_{X_1}(t_0)| = 1$ for some $0 < t_0 < 2\pi$, then $e^{it_0X_1}$ is constant, say equal to e^{ia}. Consequently, $X_1 \in \frac{a}{t_0} + \frac{2\pi}{t_0}\mathbb{Z}$, which contradicts the assumption. By periodicity and continuity, there is $c_\varepsilon < 1$ such that $|\phi_{X_1}(t)| < c_\varepsilon$ for all $\varepsilon < |t| \leq \pi$. Since $\phi_{\bar{X}_1}(t) = e^{-i\frac{\mu}{\sigma}}\phi_{X_1}(\frac{t}{\sigma})$, the claim follows. □

Of course, in the proof it was not important that the X_i are integer-valued because by rescaling we could assume that they take values in $a+r\mathbb{Z}$ for some $a, r \in \mathbb{R}$. Such random variables are said to have a **lattice distribution**. We finish this section by summarising periodicity properties of their characteristic functions, which played a crucial role in the proof of the local central limit theorem.

6.9 Lemma. *For a random variable X with characteristic function ϕ_X the following are equivalent*

(i) $\phi_X(s) = 1$ *for some* $s \neq 0$,

(ii) $\mathbb{P}\left(X \in \frac{2\pi}{s}\mathbb{Z}\right) = 1$,

(iii) ϕ_X *is* $|s|$ *periodic.*

Proof. $(i) \Rightarrow (ii)$. Since $1 = \phi_X(s) = \mathbb{E}\cos(sX) + i\mathbb{E}\sin(sX)$, we have $0 = \mathbb{E}(1 - \cos(sX))$. Since $1 - \cos(sX)$ is a nonnegative random variable whose expectation is 0, we have $\mathbb{P}(\cos(sX) = 1) = 1$ (see Theorem 1.2 (c)), equivalently $\mathbb{P}(sX \in 2\pi\mathbb{Z}) = 1$.

$(ii) \Rightarrow (iii)$. We have

$$\phi_X(t + 2\pi|s|) = \mathbb{E}e^{i(t+|s|)X} = \sum_{k\in\mathbb{Z}} e^{i(t+|s|)\frac{2\pi}{|s|}k}\mathbb{P}\left(X = \frac{2\pi}{|s|}k\right) = \sum_{k\in\mathbb{Z}} e^{it\frac{2\pi}{|s|}k}\mathbb{P}\left(X = \frac{2\pi}{|s|}k\right)$$
$$= \phi_X(t).$$

$(iii) \Rightarrow (i)$. Plainly, $\phi_X(s) = \phi_X(0) = 1$. □

6.10 Lemma. *Let X be a random variable with characteristic function ϕ_X. There are only 3 possibilities*

(i) $|\phi_X(t)| < 1$ *for every* $t \neq 0$,

(ii) $|\phi_X(s)| = 1$ *for some* $s > 0$ *and* $|\phi_X(t)| < 1$ *for all* $0 < t < s$ *and then* ϕ_X *is* s*-periodic and* $X \in a + \frac{2\pi}{s}\mathbb{Z}$ *a.s. for some* $a \in \mathbb{R}$,

(iii) $|\phi_X(t)| = 1$ *for every* $t \in \mathbb{R}$ *and then we have that* $\phi_X(t) = e^{ita}$ *for some* $a \in \mathbb{R}$, *that is* $X = a$ *a.s.*

If (ii) holds, X has a lattice distribution and since $|\phi_X(t)| < 1$ for all $0 < t < s$, by Lemma 6.9, s is the largest $r > 0$ such that $\mathbb{P}(X \in a + r\mathbb{Z}) = 1$. We sometimes call s the **span** of the distribution of X.

Proof. Let us first explain the implication in (ii). Suppose $|\phi_X(s)| = 1$ for some $s > 0$. Then $\phi_X(s) = e^{ia}$ for some $a \in \mathbb{R}$. Since $1 = e^{-ia}\phi_X(s) = \phi_{X-a}(s)$, by Lemma 6.9 applied to $X - a$, we get that $X - a \in \frac{2\pi}{s}\mathbb{Z}$ a.s. and ϕ_{X-a} is s-periodic, so $\phi_X = e^{ia}\phi_{X-a}$ is s-periodic.

To prove the trichotomy, suppose (i) and (ii) do not hold. Then there is a positive sequence $t_n \to 0$ such that $|\phi_X(t_n)| = 1$. Consequently, by what we just proved, there are $a_n \in \mathbb{R}$ such that $X \in a_n + \frac{2\pi}{t_n}\mathbb{Z}$ a.s. and ϕ_X is t_n-periodic. Without loss of generality, we can pick $a_n \in (-\frac{\pi}{t_n}, \frac{\pi}{t_n}]$. Since $t_n \to 0$, we have $\mathbb{P}\left(X \in (-\frac{\pi}{t_n}, \frac{\pi}{t_n})\right) \to 1$, which combined with $X \in a_n + \frac{2\pi}{t_n}\mathbb{Z}$ and $a_n \in (-\frac{\pi}{t_n}, \frac{\pi}{t_n}]$ gives $\mathbb{P}(X = a_n) \to 1$. Consequently, there is n_0 such that for all $n \geq n_0$, $\mathbb{P}(X = a_n) > 3/4$, but then all a_n, $n \geq n_0$ have to be equal, say $a_n = a$ and $\mathbb{P}(X = a_n) \to 1$ finally gives $\mathbb{P}(X = a) = 1$. Then $\phi_X(t) = e^{ita}$, consequently (iii) holds. □

7 Simple random walk

Let $0 < p < 1$ and let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = -1) = 1 - p$. Let S_0 be a random variable independent of the X_i. For $n = 1, 2, \ldots$ define

$$S_n = S_0 + X_1 + \ldots + X_n.$$

The sequence $(S_n)_{n \geq 0}$ is an example of a stochastic process (a collection of random variables), called a **simple random walk** (in dimension 1, i.e. on $\mathbb{Z}$) starting at S_0. We have

$$S_{n+1} = \begin{cases} S_n + 1, & \text{with probability } p, \\ S_n - 1, & \text{with probability } 1 - p. \end{cases}$$

("the future depends only on the present, not past", which is called the **Markov property** and hence the simple random walk is an example of a Markov process). For example,

$$\mathbb{P}(S_n = S_0) = \mathbb{P}(X_1 + \ldots + X_n = 0) = \begin{cases} 0, & \text{if } n \text{ is odd}, \\ \binom{n}{n/2} p^{n/2}(1-p)^{n/2}, & \text{if } n \text{ is even}. \end{cases}$$

The main question we would like to address is: what is the chance that the walk revisits its starting point, that is what is

$$\beta = \mathbb{P}(\exists n \geq 1 \; S_n = S_0)?$$

By the strong law of large numbers, $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}X_1 = 2p - 1$. Consequently, for $p \neq \frac{1}{2}$, S_n converges almost surely to $+\infty$ or $-\infty$ (depending whether $p > \frac{1}{2}$). This suggests that if $p \neq \frac{1}{2}$, then $\beta < 1$. What is β in the symmetric case $p = \frac{1}{2}$?

We shall say that a walk is **recurrent** if $\beta = 1$ and **transient** if $\beta < 1$.

7.1 Dimension 1

In dimension 1, we can obtain an explicit and very simple expression for β.

7.1 Theorem. *For a simple random walk $(S_n)_{n \geq 0}$ starting at 0 (that is $S_0 = 0$), we have*

$$\beta = \mathbb{P}(S_n = 0 \text{ for some } n \geq 1) = 1 - |2p - 1|.$$

7.2 Corollary. *A simple random walk in dimension 1 is recurrent if and only if $p = \frac{1}{2}$. Moreover, then $\mathbb{P}(S_n$ revisits 0 infinitely many times $\mid S_0 = 0) = 1$.*

Proof. The main idea is that by the Markov property,

$$\mathbb{P}(S_{n+m} = 0 \text{ for some } m \geq 1 \mid S_n = 0) = \mathbb{P}(S_n = 0 \text{ for some } m \geq 1 \mid S_0 = 0) = 1$$

(the last equality following from Theorem 7.1). We leave the details as an exercise. □

Proof of Theorem 7.1. For $n \geq 0$ define

$$A_n = \{S_n = 0\}$$

(a revisit to the origin at time n) and for $n \geq 1$ define

$$B_n = \{S_n = 0,\ S_k \neq 0, 1 \leq k \leq n-1\}$$

(a first revisit to the origin at time n). Note that $A_0 = \Omega$ and $B_1 = \varnothing$. Since $A_n \subset \bigcup_{k=1}^n B_k$ and the B_k are disjoint, we have for $n \geq 1$,

$$\mathbb{P}(A_n) = \mathbb{P}\left(A_n \cap \bigcup_{k=1}^n B_k\right) = \sum_{k=1}^n \mathbb{P}(A_n \cap B_k).$$

The key observation is that by the Markov property, $\mathbb{P}(A_n \cap B_k) = \mathbb{P}(B_k)\,\mathbb{P}(A_{n-k})$, so

$$\mathbb{P}(A_n) = \sum_{k=1}^n \mathbb{P}(B_k)\,\mathbb{P}(A_{n-k}).$$

Introducing the sequences,

$$u_k = \mathbb{P}(A_k) \qquad \text{and} \qquad f_k = \mathbb{P}(B_k),$$

it becomes

$$u_n = \sum_{k=1}^n f_k u_{n-k}, \qquad n \geq 1.$$

Our goal is to find

$$\beta = \mathbb{P}\left(\bigcup_{n\geq 1}\{S_n = 0\}\right) = \mathbb{P}\left(\bigcup_{n\geq 1} B_n\right) = \sum_{n\geq 1}\mathbb{P}(B_n) = \sum_{n\geq 1} f_n.$$

We use generating functions. Set

$$U(s) = \sum_{n=0}^\infty u_n s^n \qquad \text{and} \qquad F(s) = \sum_{n=0}^\infty f_n s^n$$

(with $f_0 = 0$) which are well defined for $|s| < 1$. By the recurrence relation derived above, we obtain

$$U(s) - 1 = \sum_{n=1}^\infty u_n s^n = \sum_{n=1}^\infty \left(\sum_{k=1}^n f_k u_{n-k}\right) s^n = \sum_{k=1}^\infty f_k s^k \sum_{n=k}^\infty u_{n-k} s^{n-k} = F(s)U(s),$$

thus

$$F(s) = 1 - \frac{1}{U(s)}, \qquad |s| < 1.$$

By virtue of Abel's theorem, we get

$$\beta = \sum_{n=1}^\infty f_n = \lim_{s\to 1-} F(s) = 1 - \frac{1}{\lim_{s\to 1-} U(s)}. \tag{7.1}$$

(the series $\sum_{n=1}^{\infty} f_n s^n$ for $s > 0$ converges or equals $+\infty$ as having nonnegative terms; similarly for $\sum_{n=1}^{\infty} u_n s^n$). Using the explicit expression for u_n, we find that

$$U(s) = \sum_{n=0}^{\infty} u_n s^n = \sum_{m=0}^{\infty} \binom{2m}{m} p^m (1-p)^m (s^2)^m = \left[1 - 4p(1-p)s^2\right]^{-1/2},$$

where the last sum is evaluated using the (infinite) binomial theorem (the Taylor expansion of $(1+x)^\alpha$). Plainly, $U(s) \xrightarrow[s\to 1-]{} (1 - 4p(1-p))^{-1/2}$, so finally

$$\beta = 1 - \sqrt{1 - 4p(1-p)} = 1 - |2p - 1|.$$

□

It is now easy to compute the expected time to return to the starting point. It turns out that it is infinity in the symmetric case (in words, a symmetric random walk revisits its starting point almost surely infinitely many times but the waiting time for a return is infinitely long in expectation).

7.3 Theorem. *For a symmetric simple random walk $(S_n)_{n\geq 0}$ starting at the origin ($S_0 = 0$), define $T = \min\{n \geq 1,\ S_n = 0\}$ (the waiting time to return to the starting point). Then $\mathbb{E}T = \infty$.*

Proof. Using the notation from the previous proof that f_n is the probability of the first revisit happening at time n, by Abel's theorem,

$$\mathbb{E}T = \sum_{n=1}^{\infty} n f_n = \lim_{s\to 1-} \sum_{n=1}^{\infty} n f_n s^{n-1} = \lim_{s\to 1-} F'(s).$$

It was established that $F(s) = 1 - \frac{1}{U(s)} = 1 - \sqrt{1-s^2}$ when $p = \frac{1}{2}$, so $F'(s) = \frac{s}{\sqrt{1-s^2}}$ and $\lim_{s\to 1-} F'(s) = \infty$. □

7.2 Dimension 2 and higher

Let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ (1 at ith coordinate), $i = 1, \ldots, d$, be the standard basis vectors in $\mathbb{R}^d$. We define a symmetric simple random walk $(S_n)_{n\geq 0}$ in dimension d starting at $0 = (0, 0, \ldots, 0)$ by $S_0 = (0, \ldots, 0)$ and $S_n = X_1 + \ldots + X_n$, where $X_1, X_2, \ldots$ are i.i.d. random vectors, each uniformly distributed on the set $\{e_1, -e_1, e_2, -e_2, \ldots, e_d, -e_d\}$, that is

$$S_{n+1} = \Big\{ S_n \pm e_i, \quad \text{with probability } \frac{1}{2d}.$$

In words, at each step, the walk chooses uniformly at random and independently of the past one of the $2d$ directions and moves to the neighbouring integer lattice site along this direction.

We are interested whether the walk is recurrent or transient.

7.4 Remark. From (7.1), it follows that a walk is recurrent ($\beta = 1$) if and only if $\sum_{n=1}^{\infty} u_n = \infty$. Note that to derive (7.1), we only used the Markov property of the walk. In particular, this conclusion remains valid for walks in all dimensions.

7.5 Theorem. *A symmetric simple random walk $(S_n)_{n\geq 0}$ in dimension d starting at 0 is recurrent if $d = 1, 2$ and transient if $d \geq 3$.*

Proof. The case $d = 1$ was done in Theorem 7.1. Consider now $d \geq 2$. We shall use Remark 7.4.

Let $d = 2$. For odd n, clearly $u_n = 0$ and for even n, say $n = 2m$, we have

$$u_n = \mathbb{P}(S_n = 0) = \sum_{k=0}^{m} \binom{2m}{k}\binom{2m-k}{k}\binom{2m-2k}{m-k}\binom{m-k}{m-k}\left(\frac{1}{4}\right)^{2m}$$

(the walk takes k steps which are $+e_1$, k steps which are $-e_1$, $m-k$ steps which are $+e_2$ and $m-k$ steps which are $-e_2$). Since

$$\binom{2m}{k}\binom{2m-k}{k}\binom{2m-2k}{m-k}\binom{m-k}{m-k} = \frac{(2m)!}{(k!)^2((m-k)!)^2} = \binom{2m}{m}\binom{m}{k}^2,$$

this simplifies to

$$u_{2m} = \binom{2m}{m}\left(\frac{1}{4}\right)^{2m}\sum_{k=0}^{m}\binom{m}{k}^2 = \binom{2m}{m}^2\left(\frac{1}{4}\right)^{2m}.$$

By Stirling's formula, $u_{2m} \approx \frac{1}{\pi m}$, so $\sum_{n=1}^{\infty} u_n = +\infty$. Consequently, the walk is recurrent.

Let $d \geq 3$. We proceed identically as in the case $d = 2$. For $n = 2m$, we have

$$\begin{aligned} u_{2m} &= \sum_{k_1+\ldots+k_d=m}\binom{2m}{k_1}\binom{2m-k_1}{k_1}\cdots\binom{2m-2k_1-\ldots-2k_{d-1}}{k_d}\left(\frac{1}{2d}\right)^{2m} \\ &= \sum_{k_1+\ldots+k_d=m}\frac{(2m)!}{(k_1!)^2\ldots(k_d!)^2}\left(\frac{1}{2d}\right)^{2m} \\ &= \binom{2m}{m}\left(\frac{1}{2d}\right)^{2m}\sum_{k_1+\ldots+k_d=m}\frac{(m!)^2}{(k_1!)^2\ldots(k_d!)^2}. \end{aligned}$$

The last sum can be estimated as follows: the coefficient $\frac{m!}{k_1!\ldots k_d!}$ is maximised over nonnegative integers $k_1+\ldots+k_d = m$ when they are all equal, which gives

$$\frac{m!}{k_1!\ldots k_d!} \leq \frac{m!}{\left(\left(\frac{m}{d}\right)!\right)^d}$$

(if m/d is not integral, we mean $(m/d)! = \Gamma(m/d-1)$; this inequality easily follows from the log-convexity of the Gamma function which in turn can be shown using Hölder's inequality). The multinomial theorem gives

$$\sum_{k_1+\ldots+k_d=m}\frac{m!}{k_1!\ldots k_d!} = (\underbrace{1+\ldots+1}_{d \text{ times}})^m = d^m.$$

Putting these two observations together yields

$$u_{2m} \leq \binom{2m}{m} \left(\frac{1}{2d}\right)^{2m} \frac{m!}{\left(\left(\frac{m}{d}\right)!\right)^d} d^m = \binom{2m}{m} \frac{1}{2^{2m}} \frac{m!}{\left(\left(\frac{m}{d}\right)!\right)^d}.$$

By Stirling's formula, for large m (d is fixed!) the right hand side is asymptotic to $\frac{c_d}{m^{d/2}}$. Consequently, $\sum u_n < \infty$ and the walk is transient. $\square$

8 Some concentration inequalities

Sums of independent random variables, say $X_1, \ldots, X_n$ tend to "concentrate" around the mean, i.e. $\mathbb{P}\left(|X_1 + \ldots + X_n - \mathbb{E}(X_1 + \ldots + X_n)| > t\right)$ is usually (exponentially) small for $t > 0$. This is a consequence of independence, but of course exact quantitative statements depend on additional assumptions on the X_i and t. We shall discuss two such inequalities for bounded random variables.

There is an elegant generalisation of Bernstein's inequality (Theorem 2.7) to any bounded random variables, which as Bernstein's inequality also provides a Gaussian tail.

8.1 Theorem (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that for each i, $X_i \in [a_i, b_i]$ with some reals $a_i < b_i$. For $S = X_1 + \ldots + X_n$ and $t > 0$, we have*

$$\mathbb{P}\left(S - \mathbb{E}S > t\right) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}. \tag{8.1}$$

8.2 Lemma. *Let X be a random variable such that $\mathbb{E}X = 0$ and $X \in [a, b]$ for some reals $a < b$. Then for every t,*

$$\mathbb{E}e^{tX} \leq \exp\left\{\frac{(b-a)^2}{8}t^2\right\}.$$

Proof. For $x \in [a, b]$, writing tx as a convex combination of ta and tb, that is $tx = \frac{b-x}{b-a}ta + \frac{x-a}{b-a}tb$, we get

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}.$$

Taking the expectation and using $\mathbb{E}X = 0$ gives

$$\mathbb{E}e^{tX} \leq pe^{ta} + (1-p)e^{tb},$$

where we denote $p = \frac{b}{b-a}$, $1 - p = \frac{-a}{b-a}$ which are both between 0 and 1. It suffices to show that

$$pe^{ta} + (1-p)e^{tb} \leq \exp\left\{\frac{(b-a)^2}{8}t^2\right\}.$$

Let

$$h(t) = \log(pe^{ta} + (1-p)e^{tb}).$$

We have, $h(0) = 0$,

$$h'(t) = \frac{pae^{ta} + (1-p)be^{tb}}{pe^{ta} + (1-p)e^{tb}},$$

so $h'(0) = 0$ and

$$\begin{aligned} h''(t) &= \frac{(pa^2e^{ta} + (1-p)b^2e^{tb})(pe^{ta} + (1-p)e^{tb}) - (pae^{ta} + (1-p)be^{tb})^2}{(pe^{ta} + (1-p)e^{tb})^2} \\ &= \frac{p(1-p)(b-a)^2e^{t(a+b)}}{(pe^{ta} + (1-p)e^{tb})^2} = \frac{(b-a)^2}{4}\left(\frac{2\sqrt{pe^{ta}(1-p)e^{tb}}}{pe^{ta} + (1-p)e^{tb}}\right)^2, \end{aligned}$$

so $h''(t) \leq \frac{(b-a)^2}{4}$. By Taylor's formula with the Lagrange remainder,

$$h(t) = h(0) + h'(0)t + \frac{1}{2}h''(\xi)t^2 = \frac{1}{2}h''(\xi)t^2 \leq \frac{(b-a)^2}{8}t^2.$$

□

Proof of Theorem 8.1. Considering $\tilde{X}_i = X_i - \mathbb{E}X_i$, we can assume that $\mathbb{E}X_i = 0$. Then $\mathbb{E}S = 0$. The main idea is to take advantage of independence by considering the exponential moments. For $\lambda > 0$, by Chebyshev's inequality, we have

$$\mathbb{P}(S > t) = \mathbb{P}\left(e^{\lambda S} > e^{\lambda t}\right) \leq e^{-\lambda t}\mathbb{E}e^{\lambda S} = e^{-\lambda t}\prod \mathbb{E}e^{\lambda X_i}.$$

Lemma 8.2 yields

$$\mathbb{P}(S > t) \leq e^{-\lambda t}\prod \mathbb{E}e^{\lambda^2 \frac{(b_i-a_i)^2}{8}} = \exp\left\{-\lambda t + \lambda^2 \frac{\sum(b_i-a_i)^2}{8}\right\}.$$

Choosing λ which minimises the right hand side, that is $\lambda = \frac{4t}{\sum(b_i-a_i)^2}$ finishes the proof. □

8.3 Example. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables with parameter p. Then $X_i \in [0,1]$, $S = X_1 + \ldots + X_n$ is binomial and with $t = \delta np$, we get

$$\mathbb{P}(S > (1+\delta)np) \leq \exp\{-2\delta^2 p^2 n\}.$$

This gives a Gaussian decay of the tail for large δ, provided that p is of constant order. If, say $p = 1/n$, we get

$$\mathbb{P}(S > 1 + \delta) \leq \exp\left\{-2\delta^2 \frac{1}{n}\right\},$$

which is not good because as $n \to \infty$, S converges in distribution to a Poisson random variable with parameter 1, so its tail should decay like the one of the Poisson distribution.

8.4 Theorem (Chernoff's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [0,1]$ for each i. For $S = X_1 + \ldots + X_n$, $\mu = \mathbb{E}S$ and $t > 0$, we have*

$$\mathbb{P}(S \geq \mu + t) \leq e^t \left(\frac{\mu}{\mu + t}\right)^{\mu+t}.$$

Proof. For $\lambda > 0$, by Chebyshev's inequality, we have

$$\mathbb{P}(S > \mu + t) = \mathbb{P}\left(e^{\lambda S} > e^{\lambda(\mu+t)}\right) \leq e^{-\lambda(\mu+t)}\mathbb{E}e^{\lambda S} = e^{-\lambda(\mu+t)}\prod \mathbb{E}e^{\lambda X_i}.$$

By convexity, for every $x \in [0,1]$, $\frac{e^{\lambda x}-1}{\lambda x} \leq \frac{e^{\lambda}-1}{\lambda}$, so $e^{\lambda x} \leq 1 + x(e^{\lambda} - 1)$, which after taking the expectation gives

$$\mathbb{E}e^{\lambda X_i} \leq 1 + (e^{\lambda} - 1)\mathbb{E}X_i.$$

Using the AM-GM inequality yields

$$\prod \mathbb{E}e^{\lambda X_i} \le \left(\frac{\sum[1+(e^\lambda-1)\mathbb{E}X_i]}{n}\right)^n = \left(1+\frac{e^\lambda-1}{n}\mu\right)^n.$$

Consequently,

$$\mathbb{P}\left(S > \mu+t\right) \le e^{-\lambda(\mu+t)}\left(1+\frac{e^\lambda-1}{n}\mu\right)^n.$$

To obtain the assertion, we choose λ such that $e^{-\lambda} = \frac{\mu}{\mu+t}$ and get

$$\mathbb{P}\left(S > \mu+t\right) \le \left(\frac{\mu}{\mu+t}\right)^{\mu+t}\left(1+\frac{t}{n}\right)^n \le \left(\frac{\mu}{\mu+t}\right)^{\mu+t} e^t.$$

(the optimal choice for λ gives a more complicated expression). □

8.5 Remark. The same arguments give bounds for lower tails, that is for $0 < t < \mu$, we have

$$\mathbb{P}\left(S \le \mu-t\right) \le e^{-t}\left(\frac{\mu}{\mu-t}\right)^{\mu-t}$$

(use $\mathbb{P}\left(S \le \mu-t\right) \le e^{-\lambda(\mu-t)}\mathbb{E}e^{\lambda S}$ for $\lambda < 0$).

8.6 Remark. By the inequality $x-(1+x)\log(1+x) \le \frac{x^2}{2(1+\frac{x}{3})}$, $x > 0$, we also get

$$\mathbb{P}\left(S > \mu+t\right) \le \exp\left\{-\frac{t^2}{2(\mu+\frac{t}{3})}\right\}.$$

8.7 Remark. The Chernoff bound is usually written with $t = \delta\mu$ as

$$\mathbb{P}\left(S > (1+\delta)\mu\right) \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu = \exp\left\{\mu[\delta-(1+\delta)\log(1+\delta)]\right\}.$$

For small δ, we get a Gaussian tail because $\delta-(1+\delta)\log(1+\delta) \le -\frac{\delta^2}{3}$ for $0 \le \delta \le \frac{3}{2}$, so

$$\mathbb{P}\left(S > (1+\delta)\mu\right) \le \exp\left\{-\frac{\delta^2\mu}{3}\right\}, \qquad 0 \le \delta \le \frac{3}{2}.$$

For large δ, we get a Poisson tail because $\delta-(1+\delta)\log(1+\delta) \le -\frac{1}{3}(1+\delta)\log(1+\delta)$ for $\delta \ge \frac{3}{2}$, so

$$\mathbb{P}\left(S > (1+\delta)\mu\right) \le \exp\left\{-\frac{\mu}{3}(1+\delta)\log(1+\delta)\right\}, \qquad \delta \ge \frac{3}{2}.$$

8.8 Example. When S is binomial with parameters n and $p = \frac{1}{p}$ as in Example 8.3, we have $\mu = 1$ and Theorem 8.4 gives

$$\mathbb{P}\left(S \ge 1+t\right) \le e^t\left(\frac{1}{1+t}\right)^{1+t} = e^{-1}\left(\frac{e}{1+t}\right)^{1+t}.$$

For large n, by the Poisson limit theorem, S tends to a Poisson random variable X with parameter 1 and

$$\mathbb{P}\left(X \ge 1+t\right) \ge \mathbb{P}\left(X = 1+t\right) = e^{-1}\frac{1}{(1+t)!} \approx e^{-1}\left(\frac{e}{1+t}\right)^{1+t}\frac{1}{\sqrt{2\pi(1+t)}}.$$

Thus Chernoff's inequality removes the inefficiency of Hoeffding's inequality from Example 8.3.

If we can control the variance, sometimes the following concentration inequality gives some improvements on the previous two.

8.9 Theorem (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that for each i, $\mathbb{E}X_i = 0$ and $X_i \in [-1, 1]$. Let $S = X_1 + \ldots + X_n$ and $\sigma^2 = \mathrm{Var}(S)$. Then, for $t > 0$,*

$$\mathbb{P}\left(S > t\right) \leq \exp\left\{-\frac{t^2}{2(\sigma^2 + \frac{t}{3})}\right\}.$$

Proof. We begin in the same way as in the proof of Hoeffding's and Chernoff's inequalities: for $\lambda > 0$, we have

$$\mathbb{P}\left(S > t\right) \leq e^{-\lambda t} \prod \mathbb{E}e^{\lambda X_i}.$$

By Taylor's expansion and the assumption $\mathbb{E}X_i = 0$,

$$\mathbb{E}e^{\lambda X_i} = \mathbb{E}\left(1 + \lambda X_i + \frac{1}{2}(\lambda X_i)^2 + \ldots\right) \leq 1 + \sum_{k \geq 2} \frac{\lambda^k}{k!} \mathbb{E}|X_i|^k.$$

Since $|X_i| \leq 1$, for $k \geq 2$, we have $\mathbb{E}|X_i|^k = \mathbb{E}|X_i|^{k-2}|X_i|^2 \leq \mathbb{E}|X_i|^2$ and thus

$$\mathbb{E}e^{\lambda X_i} \leq 1 + \left(\sum_{k \geq 2} \frac{\lambda^k}{k!}\right) \mathbb{E}X_i^2 = 1 + \left(e^\lambda - \lambda - 1\right) \mathbb{E}X_i^2 \leq \exp\left\{\left(e^\lambda - \lambda - 1\right) \mathbb{E}X_i^2\right\}.$$

Consequently,

$$\mathbb{P}\left(S > t\right) \leq e^{-\lambda t} \exp\left\{\left(e^\lambda - \lambda - 1\right) \sigma^2\right\} = \exp\left\{-\lambda(t + \sigma^2) + e^\lambda \sigma^2 - \sigma^2\right\}.$$

Choosing λ such that $e^\lambda = 1 + \frac{t}{\sigma^2}$ yields

$$\mathbb{P}\left(S > t\right) \leq \exp\left\{-(t + \sigma^2) \log\left(1 + \frac{t}{\sigma^2}\right) + t\right\} = e^t \left(\frac{\sigma^2}{\sigma^2 + t}\right)^{t + \sigma^2}.$$

We finish as in Remark 8.6. □

A Appendix: Stirling's formula

Our goal here is to give a complete and self-contained proof of indispensable Stirling's approximation for factorials (with error bounds), stated in the following theorem.

A.1 Theorem (Stirling's formula)**.** *For every integer $n \geq 1$, we have*

$$\sqrt{2\pi}\frac{n^{n+1/2}}{e^n}e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi}\frac{n^{n+1/2}}{e^n}e^{\frac{1}{12n}}, \tag{A.1}$$

or equivalently, there is $\theta_n \in (0,1)$ such that

$$n! = \sqrt{2\pi}\frac{n^{n+1/2}}{e^n}e^{\frac{1}{12n+\theta_n}}. \tag{A.2}$$

In particular, as $n \to \infty$

$$n! \approx \sqrt{2\pi}\frac{n^{n+1/2}}{e^n}. \tag{A.3}$$

As usual, here $a_n \approx b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$.

Why $\frac{n^{n+1/2}}{e^n}$?

Before giving a proper proof of Stirling's formula, let us try to explain why $\frac{n^{n+1/2}}{e^n}$ appears. Clearly, $\log(n!) = \sum_{k=1}^{n} \log k$ and because the logarithm is an increasing function, $\int_{k-1}^{k} \log x \, \mathrm{d}x < \log k < \int_{k}^{k+1} \log x \, \mathrm{d}x$. Adding these inequalities yields

$$\int_0^n \log x \, \mathrm{d}x < \log(n!) < \int_1^{n+1} \log x \, \mathrm{d}x,$$

that is ($\int \log x = x \log x - x$)

$$n \log n - n < \log(n!) < (n+1)\log(n+1) - n.$$

This suggests that $\log(n!)$ can be approximated by some sort of mean of $n \log n - n$ and $(n+1)\log(n+1) - n$. We take $(n + \frac{1}{2}) \log n - n$.

Another explanation: start with $n! = \int_0^\infty x^n e^{-x} \mathrm{d}x$ and change the variables $x = ny$, to get $n! = n^{n+1} \int_0^\infty (ye^{-y})^n \mathrm{d}y$. The function $y \mapsto ye^{-y}$ is maximal at $y = 1$ with maximum equal to e^{-1}. Rewriting, so that the maximum is at $y = 0$ and equals 1 gives $n! = n^{n+1} e^{-n} \int_{-1}^\infty ((1+y)e^{-y})^n \mathrm{d}y$. The Taylor expansion of $(1+y)e^{-y}$ at $y = 0$ is $1 - \frac{y^2}{2}$ and $\int_{-1}^\infty ((1+y)e^{-y})^n \mathrm{d}y$, heuristically, is approximately $\int_{-\infty}^\infty e^{-ny^2/2} \mathrm{d}y = \sqrt{\frac{2\pi}{n}}$. This heuristics in fact gives exactly (A.3) and can be turned into a rigorous proof, which is a particular instance of the so-called Laplace method (the drawback is that we would not get precise error estimates as in (A.2)).

Proof of Stirling's formula

Now we prove (A.1). As argued above, it makes sense to expect that $\log(n!)$ can be approximated well by $(n+\frac{1}{2})\log n - n$, therefore we consider

$$d_n = \log(n!) - \left[\left(n+\frac{1}{2}\right)\log n - n\right].$$

We have,

$$d_n - d_{n+1} = \left(n+\frac{1}{2}\right)\log\frac{n+1}{n} - 1.$$

Note that $\frac{n+1}{n} = \frac{1+\frac{1}{2n+1}}{1-\frac{1}{2n+1}}$ and for $x \in (0,1)$,

$$\begin{aligned}\log\frac{1+x}{1-x} = \log(1+x) - \log(1-x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots - \left(-x - \frac{x^2}{2} - \frac{x^3}{3} - \ldots\right) \\ &= 2\left(x + \frac{x^3}{3} + \ldots\right),\end{aligned}$$

thus

$$d_n - d_{n+1} = \frac{2n+1}{2}\log\frac{1+\frac{1}{2n+1}}{1-\frac{1}{2n+1}} - 1 = \frac{1}{3(2n+1)^2} + \frac{1}{5(2n+1)^4} + \ldots.$$

Estimating crudely $\frac{1}{5} < \frac{1}{3}$, $\frac{1}{7} < \frac{1}{3}\ldots$, we get

$$d_n - d_{n+1} < \frac{1}{3}\sum_{k=1}^{\infty}\frac{1}{(2n+1)^{2k}} = \frac{1}{3}\frac{1}{(2n+1)^2-1} = \frac{1}{3}\frac{1}{2n(2n+2)} = \frac{1}{12n} - \frac{1}{12(n+1)},$$

which means that the sequence

$$a_n = d_n - \frac{1}{12n}$$

is increasing.

On the other hand, estimating below by the first term gives

$$d_n - d_{n+1} > \frac{1}{3(2n+1)^2} > \frac{1}{12n+1} - \frac{1}{12(n+1)+1}$$

(the second inequality is equivalent to $(12n+1)(12(n+1)+1) > 12\cdot 3(2n+1)^2$, that is $24n+13>36$). This means that the sequence

$$b_n = d_n - \frac{1}{2n+1}$$

is decreasing. Since $a_n < b_n$, (a_n) increases, (b_n) decreases, (a_n) is bounded above (by b_1) and (b_n) is bounded below (by a_1). Thus both (a_n) and (b_n) converge and because $b_n - a_n \to 0$, they converge to the same limit, say c. Moreover, for every $n \geq 1$,

$$a_n < c < b_n,$$

that is

$$\log(n!) - \left[\left(n+\frac{1}{2}\right)\log n - n\right] - \frac{1}{12n} < c < \log(n!) - \left[\left(n+\frac{1}{2}\right)\log n - n\right] - \frac{1}{12n+1}$$

which can be rewritten as

$$e^c \frac{n^{n+1/2}}{e^n} e^{\frac{1}{12n}} < n! < e^c \frac{n^{n+1/2}}{e^n} e^{\frac{1}{12n+1}}.$$

In other words, if we write

$$n! = e^c \frac{n^{n+1/2}}{e^n} e^{\frac{1}{12n+\theta_n}}, \tag{A.4}$$

we have $\theta_n \in (0,1)$. To get (A.1) and (A.2), it remains to show that $e^c = \sqrt{2\pi}$.

Wallis' formula

To show that $e^c = \sqrt{2\pi}$, we shall use Wallis' formula which asserts that

$$\frac{\pi}{2} = \lim_{n\to\infty} \frac{1}{2n+1} \left[\frac{2\cdot 4\cdot \ldots \cdot (2n)}{1\cdot 3\cdot \ldots \cdot (2n-1)}\right]^2. \tag{A.5}$$

We give a short proof (it is Wallis' original elementary proof based on evaluating $\int_0^{\pi/2} \sin^n x \mathrm{d}x$). Integrating by parts gives

$$\int_0^{\pi/2} \sin^n x \ \mathrm{d}x = \int_0^{\pi/2} \sin^{n-1} x(-\cos x)' \mathrm{d}x = (n-1)\int_0^{\pi/2} \sin^{n-2} x \cos^2 x \ \mathrm{d}x,$$

which implies

$$I_n = \frac{n-1}{n} I_{n-2},$$

where

$$I_n = \int_0^{\pi/2} \sin^n x \ \mathrm{d}x.$$

Plainly, $I_0 = \frac{\pi}{2}$ and $I_1 = 1$, so iterating gives

$$I_{2n} = \frac{2n-1}{2n} \cdot \frac{2n-3}{2n-2} \cdot \ldots \cdot \frac{1}{2} \cdot \frac{\pi}{2} \tag{A.6}$$

and

$$I_{2n+1} = \frac{2n}{2n+1} \cdot \frac{2n-2}{2n-1} \cdot \ldots \cdot \frac{2}{3}. \tag{A.7}$$

Since $\sin x \in (0,1)$ for $x \in (0, \pi/2)$, we have

$$\int_0^{\pi/2} \sin^{2n+1} x \ \mathrm{d}x < \int_0^{\pi/2} \sin^{2n} x \ \mathrm{d}x < \int_0^{\pi/2} \sin^{2n-1} x \ \mathrm{d}x,$$

that is

$$\frac{2n}{2n+1} \cdot \frac{2n-2}{2n-1} \cdot \ldots \cdot \frac{2}{3} < \frac{2n-1}{2n} \cdot \frac{2n-3}{2n-2} \cdot \ldots \cdot \frac{1}{2} \cdot \frac{\pi}{2} < \frac{2n-2}{2n-1} \cdot \frac{2n-4}{2n-3} \cdot \ldots \cdot \frac{2}{3}$$

The left inequality is equivalent to

$$\frac{1}{2n+1} \left[\frac{2\cdot 4\cdot \ldots \cdot (2n)}{1\cdot 3\cdot \ldots \cdot (2n-1)}\right]^2 < \frac{\pi}{2},$$

whereas the right one, to

$$\frac{2n}{2n+1} \cdot \frac{\pi}{2} < \frac{1}{2n+1} \left[\frac{2\cdot 4\cdot \ldots \cdot (2n)}{1\cdot 3\cdot \ldots \cdot (2n-1)}\right]^2.$$

By the sandwich theorem, we obtain (A.5).

Evaluation of the constant

Multiplying the numerator and denominator by $2 \cdot 4 \cdot (2n)$, we get

$$\frac{2 \cdot 4 \cdot \ldots \cdot (2n)}{1 \cdot 3 \cdot \ldots \cdot (2n-1)} = \frac{2^{2n}(n!)^2}{(2n)!}.$$

Using Wallis' formula (A.5),

$$\sqrt{\pi} = \lim_{n\to\infty} \sqrt{\frac{2}{2n+1}} \frac{2 \cdot 4 \cdot \ldots \cdot (2n)}{1 \cdot 3 \cdot \ldots \cdot (2n-1)} = \lim_{n\to\infty} \sqrt{\frac{2n}{2n+1}} \frac{1}{\sqrt{n}} \frac{2^{2n}(n!)^2}{(2n)!},$$

so

$$\frac{1}{\sqrt{n}} \frac{2^{2n}(n!)^2}{(2n)!} \xrightarrow[n\to\infty]{} \sqrt{\pi}.$$

On the other hand, by (A.4), we have

$$\frac{1}{\sqrt{n}} \frac{2^{2n}(n!)^2}{(2n)!} = \frac{1}{\sqrt{n}} 2^{2n} \frac{e^{2c} n^{2n+1} e^{-2n} e^{\frac{1}{12n+\theta_n}}}{e^c (2n)^{2n+1/2} e^{-2n} e^{\frac{1}{24n+\theta_{2n}}}} = \frac{e^c}{\sqrt{2}} e^{\frac{1}{12n+\theta_n} - \frac{1}{24n+\theta_{2n}}} \xrightarrow[n\to\infty]{} \frac{e^c}{\sqrt{2}}$$

(because $\theta_n \in (0,1)$ for every n). Thus $\frac{e^c}{\sqrt{2}} = \sqrt{\pi}$, which finishes the evaluation of c and the proof of Theorem A.1.

References

[1] Bolthausen, E., An estimate of the remainder in a combinatorial central limit theorem. *Z. Wahrsch. Verw. Gebiete* 66 (1984), no. 3, 379–386.

[2] Durrett, R., Probability: theory and examples. Fourth edition. Cambridge Series in Statistical and Probabilistic Mathematics, 31. *Cambridge University Press, Cambridge*, 2010.

[3] Tyurin, I. S., Improvement of the remainder in the Lyapunov theorem. (Russian) *Teor. Veroyatn. Primen.* 56 (2011), no. 4, 808–811; translation in *Theory Probab. Appl.* 56 (2012), no. 4, 693–696.

[4] Feller, W., An introduction to probability theory and its applications. Vol. I and II *John Wiley & Sons, Inc.*, 1968.

[5] Grimmett, G., Welsh, D., Probability – an introduction. Second edition. *Oxford University Press*, Oxford, 2014.