Topic 1: Weak Convergence of Probability Measures

Convergence of random variables is a central topic in modern probability theory. There are four different types of convergence that are of fundamental importance: almost sure convergence, convergence in probability, L^p-convergence, and weak convergence (also known as convergence in distribution). Weak convergence, being the weakest type of convergence among the four, is a distributional property (i.e. it is only concerned with probability laws of the random variables). It does not reflect the correlations among the random variables, and it does not rely on the probability space where the random variables are defined. As a consequence, it has larger flexibility to support finer quantitative estimates. This point will be illustrated better when we discuss the rate of convergence in the central limit theorems (Stein's method). In this topic, we develop the basic tools for the study of weak convergence. Other types of convergence will be discussed in later topics.

1 Recapturing convergence in distribution

We start by recapturing the concept of "convergence in distribution" that we have seen from elementary probability theory, when we state the classical central limit theorem. We will however use the notion of "weak convergence" instead. This is a more commonly accepted terminology in modern probability although it means the same thing as "convergence in distribution". Recall that, the *cumulative distribution function* of a real valued random variable X is the function defined by

$$F(x) \triangleq \mathbb{P}(X \leqslant x),\ x \in \mathbb{R}.$$

Definition 1.1. Let X_n $(n \geqslant 1)$ and X be random variables whose cumulative distribution functions are $F_n(x)$ and $F(x)$ respectively. We say that X_n *converges weakly to* X (as $n \to \infty$), if $F_n(x)$ converges to $F(x)$ at every continuity point x of F.

The reason why we are not allowed to replace the definition with the "seemingly more natural" requirement

$$\text{“}F_n(x) \to F(x) \quad \text{for every } x \in \mathbb{R}\text{”}$$

is best illustrated by the following simple example. Let $X_n = \frac{1}{n}$ be the deterministic random variable taking value $\frac{1}{n}$. Obviously, any useful and reasonable notion of convergence should ensure that X_n "converges to" the zero random variable $X = 0$ as $n \to \infty$. On the other hand, the cumulative distribution functions of X_n and X are given by

$$F_n(x) = \begin{cases} 0, & x < 1/n; \\ 1, & x \geqslant 1/n, \end{cases} \quad F(x) = \begin{cases} 0, & x < 0; \\ 1, & x \geqslant 0, \end{cases}$$

respectively. It is apparent that

$$F_n(0) = 0 \nrightarrow F(0) = 1$$

as $n \to \infty$. This simple example tells us that, it is generally too restrictive to require that $F_n(x)$ converges to $F(x)$ *for all* $x \in \mathbb{R}$. In this example, the issue occurs precisely when $x = 0$, which is a discontinuity point of F. One can check that, at every continuity point of F (i.e. whenever $x \neq 0$) we do have $F_n(x) \to F(x)$. In other words, X_n converges weakly to X in the sense of Definition 1.1.

Example 1.1. Let X_n be a discrete uniform random variable over $\{1, 2, \cdots, n\}$, i.e.

$$\mathbb{P}(X_n = k) = \frac{1}{n}, \quad k = 1, 2, \cdots n.$$

Let X be a continuous uniform random variable over $[0, 1]$. Then $\frac{X_n}{n} \to X$ weakly as $n \to \infty$. Indeed, the cumulative distribution functions of X_n is given by

$$\begin{aligned} F_n(x) &= \mathbb{P}\big(\frac{X_n}{n} \leqslant x\big) = \mathbb{P}(X_n \leqslant nx) \\ &= \begin{cases} 0, & x < 0; \\ \frac{[nx]}{n}, & 0 \leqslant x < 1; \\ 1, & x \geqslant 1, \end{cases} \end{aligned}$$

where $[nx]$ denotes the integer part of nx. From the simple inequality

$$\frac{[nx]}{n} \leqslant \frac{nx}{n} = x \leqslant \frac{[nx]}{n} + \frac{1}{n},$$

we know that $\frac{[nx]}{n} \to x$ as $n \to \infty$. It follows that

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 0, & x < 0; \\ x, & 0 \leqslant x < 1; \\ 1, & x \geqslant 1, \end{cases}$$

which is precisely the cumulative distribution function $F(x)$ of X. Therefore, by definition we conclude that $\frac{X_n}{n} \to X$ weakly. Note that $F(x)$ is continuous at every $x \in \mathbb{R}$.

Example 1.2. Let $\{X_n : n \geqslant 1\}$ be a sequence of independent and identically distributed random variables with finite mean and variance. Define $S_n \triangleq X_1 + \cdots + X_n$. Then the "sample average" $\frac{S_n}{n}$ converges weakly to $\mathbb{E}[X_1]$ (here we regard $\mathbb{E}[X_1]$ as a deterministic random variable taking value $\mathbb{E}[X_1]$). In addition, the normalised fluctuation $\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ converges weakly to the standard normal random variable. These are contents of law of large numbers and central limit theorem, both of which hold under greater generality. We will prove these facts in the future.

When the limiting random variable X is continuous (i.e. the cumulative distribution function of X being continuous), weak convergence does become pointwise convergence for the cumulative distribution functions at every $x \in \mathbb{R}$. A surprising fact is that, one can obtain the stronger property of uniform convergence in this context. This is a result due to Pólya.

Theorem 1.1 (Pólya's theorem). *Let X_n and X be real valued random variables with cumulative distribution functions F_n and F respectively. Suppose that F is continuous on $\mathbb{R}$. Then X_n converges weakly to X if and only if F_n converges to F uniformly on $\mathbb{R}$.*

Proof. We only need to prove necessity as the other direction is trivial. Suppose that F_n converges to F at every $x \in \mathbb{R}$. Let $k \geqslant 1$ be an arbitrary given integer. We then choose a partition

$$-\infty = x_0 < x_1 < x_2 < \cdots < x_{k-1} < x_k = \infty$$

such that

$$F(x_i) = \frac{i}{k}, \quad i = 0, 1, \cdots, k.$$

This is possible since F is continuous on $\mathbb{R}$. For any $x \in [x_{i-1}, x_i)$, according to the monotonicity of cumulative distribution functions, we have

$$F_n(x) - F(x) \leqslant F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + \frac{1}{k}.$$

Similarly,

$$F_n(x) - F(x) \geqslant F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{k}.$$

Combining the two inequalities, we obtain

$$|F_n(x) - F(x)| \leqslant \max\{|F_n(x_{i-1}) - F(x_{i-1})|, |F_n(x_i) - F(x_i)|\} + \frac{1}{k},$$

which holds whenever $x \in [x_{i-1}, x_i)$. It follows that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leqslant \max_{0\leqslant i\leqslant k} |F_n(x_i) - F(x_i)| + \frac{1}{k}$$

Since $F_n(x_i) \to F(x_i)$ at each x_i, by letting $n \to \infty$ on both sides we get

$$\overline{\lim_{n\to\infty}} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leqslant \frac{1}{k}.$$

Finally, as k is arbitrary, we conclude that

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0$$

giving the desired uniform convergence. □

If we examine the definition of weak convergence, it is only concerned with the cumulative distribution functions and has nothing to do with the underlying probability spaces where the random variables are defined. Therefore, it is sufficient to define weak convergence of cumulative distribution functions (in the same way as Definition 1.1) without referring to the actual random variables.

On the other hand, it is important to extend the notion of weak convergence to higher dimensions (i.e. for $\mathbb{R}^d$-valued random variables). In higher dimensions, the notion of cumulative distribution function is less natural. Therefore, before seeking such extension, we need to reformulate weak convergence on $\mathbb{R}$ in a more natural way. This is done through the consideration of probability laws/measures.

Recall that, the *Borel σ-algebra* on $\mathbb{R}$, denoted as $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra containing all intervals of the form $(a,b]$. Given a random variable X defined on some probability space $(\Omega,\mathcal{F},\mathbb{R})$, the *law* of X is the probability measure μ_X on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ defined by

$$\mu_X(B) \triangleq \mathbb{P}(X\in B), \quad B\in\mathcal{B}(\mathbb{R}).$$

Apparently, the law of X is related to its cumulative distribution function through the relation

$$F_X(x)=\mu_X((-\infty,x]).$$

In general, there is a one-to-one correspondence between cumulative distribution functions and probability measures on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ defined as follows. If F is a cumulative distribution function, the corresponding probability measure μ is the unique probability measure such that

$$\mu((a,b])=F(b)-F(a)$$

for all intervals $(a,b]$. Conversely, if μ is a probability measure on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$, the corresponding cumulative distribution function F is given by

$$F(x)\triangleq\mu((-\infty,x]), \quad x\in\mathbb{R}.$$

To reformulate weak convergence in terms of probability measures, we first need the following definition.

Definition 1.2. Let μ be a probability measure on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. A real number $a\in\mathbb{R}$ is called a *continuity point* of μ if $\mu(\{a\})=0$.

Proposition 1.1. *Let F_n and F be cumulative distribution functions, and let μ_n and μ be the corresponding probability measures on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. Then F_n converges weakly to F if and only if*

$$\mu_n((a,b])\to\mu((a,b]) \tag{1.1}$$

for any continuity points $a<b$ of μ.

Remark 1.1. In the context of random variables, (1.1) reads

$$\mathbb{P}(X_n\in(a,b])\to\mathbb{P}(X\in(a,b])$$

for any continuity points $a<b$ of the law of X. Note that we *cannot* replace $(a,b]$ by arbitrary Borel measurable subsets $A\in\mathcal{B}(\mathbb{R})$, even in the case when X is a continuous random variable and we know from Pólya's theorem (cf. Theorem 1.1) that F_n converges uniformly to F.

The necessity part of Proposition 1.1 is trivial. Indeed, suppose that F_n converges weakly to F, i.e. $F_n(x) \to F(x)$ at every continuity point x of F. Note that a is a continuity point of μ if and only if it is a continuity point of F. Therefore, for any continuity points $a < b$ of μ, we have

$$\mu_n((a,b]) = F_n(b) - F_n(a) \to F(b) - F(a) = \mu((a,b]).$$

The sufficiency part is not as obvious. The crucial point is a so-called *tightness* property for the sequence $\{\mu_n\}$, which is in turn based on the fact that μ_n and μ are *probability* measures. At this point, we take the result for granted as motivation. Its proof will be clear when we are more comfortable with weak convergence properties (in particular, with the tightness property).

2 Vague convergence and Helly's theorem

For the moment, it will be convenient to first relax the assumption of being probability measures, and to start by working with finite measures. A reason for this, which will be clear later on, is related to the important Helly's theorem. Recall that, a finite measure is a measure whose total mass is finite.

Let μ be a finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. In the same way as Definition 1.2, an element $a \in \mathbb{R}$ is said to be a *continuity point* of μ if $\mu(\{a\}) = 0$. The set of continuity points of μ is denoted as $\mathcal{C}(\mu)$.

Remark 2.1. The complement of $\mathcal{C}(\mu)$ is at most countable. In fact, first observe that, for each given $\varepsilon > 0$, the set $E_\varepsilon \triangleq \{a \in \mathbb{R} : \mu(\{a\}) \geqslant \varepsilon\}$ must be finite. For otherwise, say E_ε contains an infinite sequence $a_1, a_2, \cdots$, then

$$\mu(\{a_1, a_2, \cdots\}) = \sum_{i=1}^{\infty} \mu(\{a_i\}) \geqslant \sum_{i=1}^{\infty} \varepsilon = \infty,$$

contradicting the finiteness of μ. Therefore, E_ε is a finite set. It now follows that

$$\mathcal{C}(\mu)^c = \{a \in \mathbb{R} : \mu(\{a\}) > 0\} = \cup_{n=1}^{\infty} \{a \in \mathbb{R} : \mu(\{a\}) \geqslant \frac{1}{n}\}$$

is at most countable. In particular, as a consequence we also know that $\mathcal{C}(\mu)$ is dense in $\mathbb{R}$.

Inspired by Proposition 1.1, we introduce the following definition.

Definition 2.1. Let μ_n ($n \geqslant 1$) and μ be finite measures on $\mathbb{R}$. We say that μ_n *converges vaguely* to μ (as $n \to \infty$), if $\mu_n((a,b]) \to \mu((a,b])$ for any continuity points $a < b$ of μ. If in addition we further have $\mu_n(\mathbb{R}) \to \mu(\mathbb{R})$, then we say that μ_n *converges weakly* to μ.

When μ_n and μ are probability measures, vague and weak convergence are the same thing since $\mu_n(\mathbb{R}) = 1 \to 1 = \mu(\mathbb{R})$ in this case. In general, these two notions of convergence are different, as seen from the following example.

Example 2.1. Consider $\mu_n = \delta_n$ (the Dirac mass at the point $x = n$) and $\mu = 0$ (the zero measure). Every point on $\mathbb{R}$ is a continuity point of μ. For any fixed $a < b$, when n is large (precisely when $n > b$) we have $\mu_n((a,b]) = 0$. In particular, μ_n converges vaguely to μ. But

$$\mu_n(\mathbb{R}) = 1 \nrightarrow 0 = \mu(\mathbb{R}).$$

In other words, μ_n does not converge weakly to μ.

The notion of intervals is too special for practical purposes. We need to find more robust characterisations of vague and weak convergence in order to generalise these concepts. The following two results provide very useful characterisations of vague convergence and weak convergence in terms of integration against suitable test functions. They are essential for seeking generalisations of the convergence concepts to higher dimensions and to stochastic processes (infinite dimensions).

Recall that, a continuous function on $\mathbb{R}^d$ with compact support is a continuous function f which vanishes identically outside some bounded subset of $\mathbb{R}^d$. The space of continuous functions on $\mathbb{R}^d$ with compact support is denoted as $\mathcal{C}_c(\mathbb{R}^d)$. Respectively, the space of bounded continuous functions on $\mathbb{R}^d$ is denoted as $\mathcal{C}_b(\mathbb{R}^d)$. Apparently, $\mathcal{C}_c(\mathbb{R}^d) \subseteq \mathcal{C}_b(\mathbb{R}^d)$.

First of all, for vague convergence we have the following characterisation.

Theorem 2.1. *Let μ_n ($n \geqslant 1$) and μ be finite measures on $\mathbb{R}$. Then μ_n converges vaguely to μ if and only if*

$$\int_{\mathbb{R}} f(x)\mu_n(dx) \to \int_{\mathbb{R}} f(x)\mu(dx) \quad \text{for all } f \in \mathcal{C}_c(\mathbb{R}).$$

Proof. Necessity. Let $f \in \mathcal{C}_c(\mathbb{R})$. Firstly, we choose $a < b$ in $\mathcal{C}(\mu)$ (continuity points of μ) and $f(x) = 0$ outside $[a,b]$. Since f is continuous, it is uniformly continuous on $[a,b]$. In particular, given arbitrary $\varepsilon > 0$, there exists $\delta > 0$ such that whenever $x, y \in [a,b]$ with $|x-y|<\delta$, we have $|f(x)-f(y)|<\varepsilon$. For such δ, we choose a partition

$$a = x_0 < x_1 < \cdots < x_{k-1} < x_k = b$$

such that $x_i \in \mathcal{C}(\mu)$ and $|x_i - x_{i-1}| < \delta$. This is possible since $\mathcal{C}(\mu)$ is dense in $\mathbb{R}$ (cf. Remark 2.1). Now if we define the step function

$$g(x) \triangleq \sum_{i=1}^{k} f(x_{i-1}) \mathbf{1}_{(x_{i-1},x_i]}(x), \quad x \in \mathbb{R},$$

then $f(x)=g(x)=0$ when $x \notin (a,b]$ and $|f(x)-g(x)|<\varepsilon$ when $x \in [a,b]$. It follows that,

$$\begin{aligned}
&\big|\int_{\mathbb{R}} f(x)\mu_n(dx) - \int_{\mathbb{R}} f(x)\mu(dx)\big| \\
&\leqslant \big|\int_{\mathbb{R}} f(x)\mu_n(dx) - \int_{\mathbb{R}} g(x)\mu_n(dx)\big| + \big|\int_{\mathbb{R}} g(x)\mu_n(dx) - \int_{\mathbb{R}} g(x)\mu(dx)\big| \\
&\quad + \big|\int_{\mathbb{R}} g(x)\mu(dx) - \int_{\mathbb{R}} f(x)\mu(dx)\big| \\
&\leqslant \varepsilon \cdot \mu_n((a,b]) + \sum_{i=1}^{k} |f(x_{i-1})| \cdot |\mu_n((x_{i-1},x_i]) - \mu((x_{i-1},x_i])| + \varepsilon \cdot \mu((a,b]).
\end{aligned}$$

Since a, b and all those x_i's are continuity points of μ, by letting $n \to \infty$ we have

$$\mu_n((a,b]) \to \mu((a,b]),\ \mu_n((x_{i-1},x_i]) \to \mu((x_{i-1},x_i]).$$

Therefore,

$$\overline{\lim_{n\to\infty}} \big|\int_{\mathbb{R}} f(x)\mu_n(dx) - \int_{\mathbb{R}} f(x)\mu(dx)\big| \leqslant 2\varepsilon \cdot \mu((a,b]).$$

Since ε is arbitrary, we conclude that

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\mu_n(dx) = \int_{\mathbb{R}} f(x)\mu(dx).$$

Sufficiency. Let $a<b$ be two continuity points of μ, and let $g(x) \triangleq \mathbf{1}_{(a,b]}(x)$. Given an arbitrary $\delta>0$, we are going to define two "tent-shaped" functions $g_1, g_2 \in \mathcal{C}_c(\mathbb{R})$ that approximate g from above and from below. Precisely, $g_1(x) \triangleq 1$

when $x \in [a,b]$, $g_1(x) \triangleq 0$ when $x \notin [a-\delta, b+\delta]$, and $g_1(x)$ is linear when $x \in [a-\delta, a]$ and $x \in [b, b+\delta]$. Similarly, $g_2(x) \triangleq 1$ when $x \in [a+\delta, b-\delta]$, $g_1(x) \triangleq 0$ when $x \notin [a,b]$, and $g_2(x)$ is linear when $x \in [a, a+\delta]$ and $x \in [b-\delta, b]$. By the constructions, it is not hard to see that

$$g_2(x) \leqslant g(x) \leqslant g_1(x) \quad \forall x \in \mathbb{R}, \tag{2.1}$$

and

$$g_1 = g_2 \text{ on } U^c,\ 0 \leqslant g_1 - g_2 \leqslant 1 \text{ on } U \tag{2.2}$$

with $U \triangleq (a-\delta, a+\delta) \cup (b-\delta, b+\delta)$. These properties are most easily seen by draw the graphs of g_1, g, g_2 in the same picture.

By integrating (2.1) against μ_n and μ respectively, we obtain

$$\int g_2 d\mu_n \leqslant \int g d\mu_n = \mu_n((a,b]) \leqslant \int g_1 d\mu_n, \ \int g_2 d\mu \leqslant \mu((a,b]) \leqslant \int g_1 d\mu,$$

where we have omitted the region of integration and the integrating variable for simplicity. Therefore,

$$\int g_2 d\mu_n - \int g_1 d\mu \leqslant \mu_n((a,b]) - \mu((a,b]) \leqslant \int g_1 d\mu_n - \int g_2 d\mu. \tag{2.3}$$

By taking $n \to \infty$ in the first inequality and using (2.2), we obtain that

$$\begin{aligned} \varliminf_{n\to\infty} \big(\mu_n((a,b]) - \mu((a,b])\big) &\geqslant \int g_2 d\mu - \int g_1 d\mu \geqslant -\mu(U) \\ &= -\big(\mu((a-\delta, a+\delta)) + \mu((b-\delta, b+\delta))\big). \end{aligned}$$

Since a, b are continuity points of μ and δ is arbitrary, by letting $\delta \to 0$ the last term goes to zero and thus

$$\varliminf_{n\to\infty} \big(\mu_n((a,b]) - \mu((a,b])\big) \geqslant 0.$$

Exactly the same argument applied to the second inequality in (2.3) leads us to

$$\varlimsup_{n\to\infty} \big(\mu_n((a,b]) - \mu((a,b])\big) \leqslant 0.$$

Therefore, we arrive at

$$\lim_{n\to\infty} \mu_n((a,b]) = \mu((a,b]).$$

□

Respectively, for weak convergence we have the following characterisation.

Theorem 2.2. *Let μ_n ($n \geqslant 1$) and μ be finite measures on $\mathbb{R}$. Then μ_n converges weakly to μ if and only if*

$$\int_{\mathbb{R}} f(x)\mu_n(dx) \to \int_{\mathbb{R}} f(x)\mu(dx) \quad \text{for all } f \in \mathcal{C}_b(\mathbb{R}). \tag{2.4}$$

Proof. Sufficiency. The condition already implies vague convergence as a consequence of Theorem 2.1 since $\mathcal{C}_c(\mathbb{R}) \subseteq \mathcal{C}_b(\mathbb{R})$. In addition, by taking $f = 1$, we also have $\mu_n(\mathbb{R}) \to \mu(\mathbb{R})$. Therefore, μ_n converges weakly to μ.

Necessity. Let $f \in \mathcal{C}_b(\mathbb{R})$ and suppose that $|f(x)| \leqslant M$ for all x. Given an arbitrary $\varepsilon > 0$, we pick two continuity points $a < b$ of μ so that $\mu((a,b]^c) < \varepsilon$. By the weak convergence assumption, we know that

$$\mu_n((a,b]^c) = \mu_n(\mathbb{R}) - \mu_n((a,b]) \to \mu(\mathbb{R}) - \mu((a,b]) = \mu((a,b]^c).$$

In particular, $\mu_n((a,b]^c) < \varepsilon$ when n is large. It follows that,

$$\begin{aligned}
&\Big|\int_{\mathbb{R}} f(x)\mu_n(dx) - \int_{\mathbb{R}} f(x)\mu(dx)\Big| \\
&\leqslant \Big|\int_{(a,b]} f(x)\mu_n(dx) - \int_{(a,b]} f(x)\mu(dx)\Big| + \Big|\int_{(a,b]^c} f(x)\mu_n(dx) - \int_{(a,b]^c} f(x)\mu(dx)\Big| \\
&\leqslant \Big|\int_{(a,b]} f(x)\mu_n(dx) - \int_{(a,b]} f(x)\mu(dx)\Big| + 2M\varepsilon. \qquad (2.5)
\end{aligned}$$

By using the same approximation argument as in the necessity part of Theorem 2.1, we can show that the first term on the right hand side of (2.5) vanishes as $n \to \infty$. Therefore,

$$\overline{\lim_{n\to\infty}}\Big|\int_{\mathbb{R}} f(x)\mu_n(dx) - \int_{\mathbb{R}} f(x)\mu(dx)\Big| \leqslant 2M\varepsilon,$$

which further implies

$$\lim_{n\to\infty}\int_{\mathbb{R}} f(x)\mu_n(dx) = \int_{\mathbb{R}} f(x)\mu(dx)$$

since ε is arbitrary. □

Now using the characterisations given by Theorem 2.1 and Theorem 2.2, we can generalise the concepts of vague and weak convergence to higher dimensions.

Definition 2.2. Let μ_n $(n \geqslant 1)$ and μ be finite meansures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.
(i) We say that μ_n converges *vaguely* to μ if

$$\int_{\mathbb{R}^d} f(x)\mu_n(dx) \to \int_{\mathbb{R}^d} f(x)\mu(dx)$$

for every $f \in \mathcal{C}_c(\mathbb{R}^d)$ (continuous functions with compact supports).
(ii) We say that μ_n converges *weakly* to μ if

$$\int_{\mathbb{R}^d} f(x)\mu_n(dx) \to \int_{\mathbb{R}^d} f(x)\mu(dx)$$

for every $f \in \mathcal{C}_b(\mathbb{R}^d)$ (bounded continuous functions).

Remark 2.2. It can be shown that, weak convergence is equivalent to vague convergence plus the property that $\mu_n(\mathbb{R}^d) \to \mu(\mathbb{R}^d)$. In particular, when μ_n, μ are probability measures, the two notions of convergence are the same thing. In the context of $\mathbb{R}^d$-valued random variables X_n and X, X_n converges weakly to X if and only if

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$$

for every bounded continuous function f.

Recall from real analysis that a bounded sequence in $\mathbb{R}^d$ always has a convergent subsequence. The extension of this result to probability measures is the content of *Helly's theorem.* This theorem is important because it is often the first step towards proving weak convergence of probability measures. Before stating the theorem, we first introduce the following definition.

Definition 2.3. A *sub-probability measure* μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is a finite measure such that $\mu(\mathbb{R}^d) \leqslant 1$.

Theorem 2.3 (Helly's theorem). *Let $\{\mu_n : n \geqslant 1\}$ be a sequence of probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then there exists a subsequence μ_{n_k} and a sub-probability measure μ, such that μ_{n_k} converges vaguely to μ as $k \to \infty$.*

Proof. We only prove the result in one dimension. The argument can be adapted to the higher dimensions. We break down the proof into several key steps.

Step One. Consider the corresponding sequence of cumulative distribution functions $F_n(x) \triangleq \mu_n((-\infty, x])$. Let $D = \{x_j : j \geqslant 1\}$ be a countable dense subset of $\mathbb{R}^1$ (e.g. the rational numbers). We claim that, there exists a subsequence $\{F_{n_k}\}$ of $\{F_n\}$, such that $\lim_{k\to\infty} F_{n_k}(x_j)$ exists for every $x_j \in D$. To prove this, let us start with the sequence $\{F_n(x_1)\}$ of real numbers. Since this is a bounded sequence, there exists a subsequence $\{n_1(k) : k \geqslant 1\}$ of $\mathbb{N}$ and some real number denoted as $G(x_1)$, such that $F_{n_1(k)}(x_1) \to G(x_1)$. Next, for the bounded sequence $\{F_{n_1(k)}(x_2) : k \geqslant 1\}$, there exists a further subsequence $\{n_2(k)\}$ of $\{n_1(k)\}$ and some real number denoted as $G(x_2)$, such that $F_{n_2(k)}(x_2) \to G(x_2)$. If we continue this procedure, at the j-th step we find a subsequence $\{n_j(k)\}$ of the previous sequence $\{n_{j-1}(k)\}$ as well as $G(x_j) \in \mathbb{R}^1$, such that $F_{n_j(k)}(x_j) \to G(x_j)$. Now we consider the sequence $\{n_k(k) : k \geqslant 1\}$ selected diagonally. For each fixed j, by the previous construction we know that $\{n_k(k) : k \geqslant j\}$ is a subsequence of $\{n_j(k) : k \geqslant 1\}$. Therefore,

$$\lim_{k\to\infty} F_{n_k(k)}(x_j) = G(x_j),$$

which proves the desired claim.

Step Two. Using the previous numbers $\{G(x_j) : j \geqslant 1\}$, we define the function

$$F(x) \triangleq \inf\{G(x_j) : x_j > x\}.$$

It is obvious that $0 \leqslant F(x) \leqslant 1$ and $F(x)$ is increasing. Moreover, $F(x)$ is right continuous. Indeed, let $x \in \mathbb{R}^1$ and $\varepsilon > 0$. By the definition of F, there exists $x_j > x$ such that $G(x_j) < F(x) + \varepsilon$. It follows that, whenever $0 < h < x_j - x$ we have $x + h < x_j$ and thus

$$F(x+h) \leqslant G(x_j) < F(x) + \varepsilon.$$

This shows that F is right continuous at x.

Step Three. At every continuity point x of F, we have $F_{n_k(k)}(x) \to F(x)$. For simplicity we write $n_k \triangleq n_k(k)$. Given $\varepsilon > 0$, there exists $x_p > x$ such that

$$G(x_p) < F(x) + \varepsilon. \tag{2.6}$$

In addition, since x is a continuity point of F, there exists $y < x$ such that $F(x) - F(y) < \varepsilon$. Pick any $x_q \in D \cap (y, x)$. It follows that $F(y) \leqslant G(x_q)$ and thus

$$F(x) - G(x_q) \leqslant F(x) - F(y) < \varepsilon. \tag{2.7}$$

Adding (2.6) and (2.7) gives

$$G(x_p) - G(x_q) < 2\varepsilon.$$

Now we have:

$$\begin{aligned}|F_{n_k}(x)-F(x)| &\leqslant |F_{n_k}(x)-F_{n_k}(x_p)|+|F_{n_k}(x_p)-G(x_p)|+|G(x_p)-F(x)|\\ &\leqslant \big(F_{n_k}(x_p)-F_{n_k}(x_q)\big)+|F_{n_k}(x_p)-G(x_p)|+\varepsilon.\end{aligned}$$

By taking $k\to\infty$, we obtain

$$\overline{\lim_{k\to\infty}}\big|F_{n_k}(x)-F(x)\big|\leqslant G(x_p)-G(x_q)+\varepsilon\leqslant 3\varepsilon.$$

Since ε is arbitrary, we conclude that $F_{n_k}(x)\to F(x)$.

Step Four. By standard measure theory (Carathéodory's extension theorem), the function F induces a unique sub-probability μ on $(\mathbb{R}^1,\mathcal{B}(\mathbb{R}^1))$ such that $\mu((a,b])=F(b)-F(a)$ for any $a<b$. Step three shows that μ_{n_k} converges vaguely to μ, which completes the proof of Helly's theorem. □

It is important to point out that, in general one cannot strengthen the conclusion of Helly's theorem to weak convergence, as the limit point μ may fail to be a probability measure.

Example 2.2. Let μ_n be the uniform distribution over $[-n,n]$. Then μ_n converges vaguely to the zero measure (and so does any of its subsequence). Indeed, for any fixed $a<b$, when n is large we have

$$\mu_n((a,b])=\frac{b-a}{2n},$$

which converges to zero as $n\to\infty$.

The question about when a vague limit point has to be a probability measure becomes an important one. The complete answer to this question is related to the so-called *tightness* property which will be discussed in Section 4. Here let us look at a very simple but enlightening example.

Example 2.3. Let $M>0$ be a fixed number. Let $\{\mu_n:n\geqslant 1\}$ be a sequence of probability measures on $(\mathbb{R}^1,\mathcal{B}(\mathbb{R}^1))$ such that $\mu_n([-M,M])=1$ for each n. Then every vague convergent subsequence of μ_n must converge weakly to a probability measure. Indeed, let μ_{n_k} converges vaguely to some sub-probability measure μ. Pick two continuity points a,b of μ such that $a<-M$ and $b>M$. Then

$$1=\mu_n((a,b])\to\mu((a,b]),$$

showing that μ has to be a probability measure and thus μ_{n_k} converges weakly to μ. The key point behind this example is the property that masses for the sequence $\{\mu_n\}$ are uniformly concentrated on a large interval. The precise formulation of such property, known as *tightness*, will be discussed in Section 4 and is of fundamental importance in the study of weak convergence.

3 Weak convergence on metric spaces and Portmanteau's theorem

Working with probability measures over $\mathbb{R}^d$ (i.e. in finite dimensions) is not sufficient for modern probability theory. For instance, when we study distributions of stochastic processes, we are immediately led to the consideration of probability measures over infinite dimensional spaces (the space of "paths"). It is essential to extend the notion of weak convergence to the more general context of metric spaces.

Heuristically, a metric space is a set equipped with a distance function.

Definition 3.1. Let S be a non-empty set. A *metric* on S is a non-negative function $\rho : S \times S \to [0, \infty)$ which satisfies the following three properties:

(i) Positive definiteness: $\rho(x, y) = 0$ if and only if $x = y$;
(ii) Symmetry: $\rho(x, y) = \rho(y, x)$;
(iii) Triangle inequality: $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z)$.

When a set S is equipped with a metric ρ, we say that (S, ρ) is a *metric space.*

Example 3.1. An obvious metric on $\mathbb{R}^d$ is the Euclidean metric:

$$\rho(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_d - y_d)^2}.$$

But there are other choices of metrics, such as

$$\rho'(x, y) = |x_1 - y_1| + \cdots + |x_d - y_d| \quad \text{(the } l^1 \text{ metric)}$$

or

$$\rho''(x, y) = \max_{1 \leqslant i \leqslant d} |x_i - y_i| \quad \text{(the } l^\infty \text{ metric).}$$

Example 3.2. One important infinite dimensional example of a metric space is the space of paths. More precisely, let $W = C[0, 1]$ be the set of all continuous functions $w : [0, 1] \to \mathbb{R}^1$. Define $\rho : W \times W \to [0, \infty)$ by

$$\rho(w_1, w_2) \triangleq \sup_{0 \leqslant t \leqslant 1} |w_1(t) - w_2(t)|, \quad w_1, w_2 \in W.$$

It is a simple exercise to check that ρ is a metric on W (it is called the *uniform metric*). We will frequently encounter this metric space (W, ρ) when we work with continuous stochastic processes such as the Brownian motion and the related stochastic calculus.

Let (S,ρ) be a given metric space. We can describe some basic classes of subsets. Unlike the usual $\mathbb{R}^d$, there are no analogues of intervals on S. However, we have the natural notion of *open balls*

$$B(x,r)\triangleq\{y\in S: d(y,x)<r\}$$

and similarly of closed balls. Given a subset $A\subseteq S$, a point $x\in A$ is called an *interior point* of A if there exists $r>0$ such that $B(x,r)\subseteq A$. A subset $G\subseteq S$ is said to be *open* if every point in G is an interior point. A subset $F\subseteq S$ is said to be *closed* if its complement F^c is open. A subset $K\subseteq S$ is said to be *compact* if any open cover of K contains a finite subcover, namely whenever K is contained in the union of a family of open sets, one can always choose finitely many members in that family whose union still contains K.

The above concepts are better illustrated along with the notion of convergence. Let x_n $(n\geqslant 1)$ and x be points in S. We say that x_n *converges to* x, denoted as $x_n\rightarrow x$, if $\rho(x_n,x)\rightarrow 0$ as $n\rightarrow\infty$. One can show that, a subset F is closed if and only if

$$x_n\in F,\ x_n\rightarrow x\implies x\in F.$$

In addition, a subset K is compact if and only if it is closed and any sequence in K admits a convergent subsequence.

Let A be a subset of S. The *closure* of A, denoted as $\bar{A}$, is the smallest closed subset containing A. Equivalently, $\bar{A}$ consists of all limit points of A. The *interior* of A, denoted as $\mathring{A}$, is the largest open subset contained in A. Equivalently, $\mathring{A}$ is the set of interior points of A. The *boundary* of A is defined to be $\partial A\triangleq\bar{A}\backslash\mathring{A}$.

Continuous functions and uniformly continuous functions are defined in the usual way. A function $f:S\rightarrow\mathbb{R}$ is *continuous* at x, if

$$x_n\in S,\ x_n\rightarrow x\implies f(x_n)\rightarrow f(x).$$

A *continuous function* on S is a function that is continuous at every point in S. A function is *uniformly continuous*, if for any $\varepsilon>0$, there exists $\delta>0$ such that

$$x,y\in S,\ d(x,y)<\delta\implies|f(y)-f(x)|<\varepsilon.$$

If $f:S\rightarrow\mathbb{R}$ is continuous, then

$$\begin{aligned}U\subseteq\mathbb{R},\ U\text{ is open}&\implies f^{-1}U\text{ is open in }S;\\ C\subseteq\mathbb{R},\ C\text{ is closed}&\implies f^{-1}C\text{ is closed in }S.\end{aligned}$$

The space of bounded and continuous functions on S is denoted as $\mathcal{C}_b(S)$.

Remark 3.1. All the above concepts are natural generalisations of the $\mathbb{R}^d$ case, and they are best visualised when one refers back to the example of $\mathbb{R}^d$. One major difference from the $\mathbb{R}^d$ case is notion of compact sets. In $\mathbb{R}^d$, we know that a subset is compact if and only if it is bounded and closed. This will not be true in general metric spaces–compactness can be much subtler and more luxurious to expect. In the appendix, we provide the description of compact subsets in the space $C[0,1]$ of Example 3.2.

Now we can consider the notion of probability measures and weak convergence on metric spaces. The crucial missing object is a natural σ-algebra (the class of events). Let (S,ρ) be a given metric space.

Definition 3.2. The *Borel σ-algebra* over S, denoted as $\mathcal{B}(S)$, is the smallest σ-algebra containing all open subsets.

Example 3.3. Open balls, closed balls, open sets, closed sets, compact sets and any countable unions/intersections of these sets are all in $\mathcal{B}(S)$.

Remark 3.2. In the case of $\mathbb{R}^1$ (or $\mathbb{R}^d$), the Borel σ-algebra is generated by the class of open intervals (a,b). For general metric spaces, the Borel σ-algebra may not necessarily be generated by open balls. Nevertheless, this will be the case if the metric space (S,ρ) is *separable*, namely if there exists a countable subset $D \subseteq S$ such that $\bar{D} = S$.

We will always work with the Borel σ-algebra $\mathcal{B}(S)$, and probability measures are all assumed to be defined on $\mathcal{B}(S)$. A natural way of generalising the notion of weak convergence to metric spaces is through the characterisation given by Theorem 2.2 for the $\mathbb{R}^d$ case.

Definition 3.3. Let μ_n $(n \geqslant 1)$ and μ be probability measures on $(S,\mathcal{B}(S),\rho)$. We say that μ_n *converges weakly to* μ, if

$$\int_S f(x)\mu_n(dx) \to \int_S f(x)\mu(dx)$$

for all bounded and continuous functions $f \in \mathcal{C}_b(S)$.

One immediate question is: how can we understand weak convergence through testing against "sets", namely through understanding the convergence of $\mu_n(A)$ for $A \in \mathcal{B}(S)$? One cannot expect that $\mu_n(A) \to \mu(A)$ for all $A \in \mathcal{B}(S)$, and just like the $\mathbb{R}^d$ case, we need some sort of continuity for the set A with respect to the limiting measure μ.

Definition 3.4. Let μ be a probability measure on $(S, \mathcal{B}(S), \rho)$. A subset $A \in \mathcal{B}(S)$ is said to be *μ-continuous*, if $\mu(\partial A) = 0$.

Example 3.4. In the case of $S = \mathbb{R}^1$, it is obvious that an interval $(a, b]$ is μ-continuous if and only if a, b are both continuity points of μ (cf. Definition 1.2).

Example 3.5. Let $S = \{(x, y) : 0 \leqslant x, y \leqslant 1\}$ be the unit square in $\mathbb{R}^2$, and let μ be the uniform probability measure, i.e. $\mu(A) \triangleq \text{Area}(A)$ for each $A \in \mathcal{B}(S)$. Note that μ is the law of a uniform random point (X, Y) taking values in S. Then any region in S enclosed by a smooth curve is μ-continuous, as its boundary is the enclosing curve which has zero area.

The following core result in this section, known as the *Portmanteau theorem*, provides a set of equivalent characterisations for weak convergence.

Theorem 3.1. *Let μ_n ($n \geqslant 1$) and μ be probability measures on $(S, \mathcal{B}(S), \rho)$. The following statements are equivalent:*

(i) μ_n converges weakly to μ;
(ii) for any bounded and uniformly continuous function f on S, we have

$$\int_S f(x)\mu_n(dx) \to \int_S f(x)\mu(dx);$$

(iii) for any closed subset $F \subseteq S$, we have

$$\overline{\lim_{n\to\infty}} \mu_n(F) \leqslant \mu(F);$$

(iv) for any open subsets $G \subseteq S$, we have

$$\underline{\lim}_{n\to\infty} \mu_n(G) \geqslant \mu(G);$$

(v) for any Borel measurable subset $A \in \mathcal{B}(S)$ that is μ-continuous, we have

$$\lim_{n\to\infty} \mu_n(A) = \mu(A).$$

Proof. (i) $\Longrightarrow$ (ii) is trivial.

(ii) $\Longrightarrow$ (iii). Let F be a closed subset of S. For $k \geqslant 1$, define

$$f_k(x) = \left(\frac{1}{1 + \rho(x, F)}\right)^k, \ x \in S,$$

where $\rho(x, F)$ is the distance between x and F. Then f_k is bounded and uniformly continuous. In addition,

$$\mathbf{1}_F(x) \leqslant f_k(x) \leqslant 1, \tag{3.1}$$

and $f_k(x) \downarrow \mathbf{1}_F(x)$ as $k \to \infty$, where $\mathbf{1}_F(x)$ denotes the indicator function of F. It follows from (3.1) and the assumption that

$$\overline{\lim_{n\to\infty}} \mu_n(F) \leqslant \lim_{n\to\infty} \int_S f_k(x)\mu_n(dx) = \int_S f_k(x)\mu(dx)$$

for every $k \geqslant 1$. By taking $k \to \infty$ and using the dominated convergence theorem, we conclude that

$$\overline{\lim_{n\to\infty}} \mu_n(F) \leqslant \mu(F).$$

(iii)$\Longleftrightarrow$(iv) is obvious as they are complement to each other.

(iii)+(iv) $\Longrightarrow$ (v). Let $A \in \mathcal{B}(S)$ be such that $\mu(\partial A) = 0$. Then

$$\mu(\mathring{A}) = \mu(A) = \mu(\bar{A}).$$

By the assumptions of (iii) and (iv), we have

$$\begin{aligned} \overline{\lim_{n\to\infty}} \mu_n(A) &\leqslant \overline{\lim_{n\to\infty}} \mu_n(\bar{A}) \\ &\leqslant \mu(\bar{A}) = \mu(A) = \mu(\mathring{A}) \\ &\leqslant \underline{\lim_{n\to\infty}} \mu_n(\mathring{A}) \\ &\leqslant \underline{\lim_{n\to\infty}} \mu_n(A). \end{aligned}$$

Therefore, $\mu_n(A) \to \mu(A)$.

(v) $\Longrightarrow$ (i). Let $f \in \mathcal{C}_b(S)$ be a bounded continuous function. The idea is to approximate f by linear combinations of indicator functions of μ-continuous sets.

We first assume that $0 < f < 1$. Since μ is a probability measure, for each $n \geqslant 1$ the set $\{a \in \mathbb{R}^1 : \mu(f = a) \geqslant 1/n\}$ must be finite, and thus the set $\{a \in \mathbb{R}^1 : \mu(f = a) > 0\}$ is at most countable. Given $k \geqslant 1$, for each $1 \leqslant i \leqslant k$ we can then choose some $a_i \in ((i-1)/k, i/k)$ such that $\mu(f = a_i) = 0$. Set $a_0 \triangleq 0$, $a_{k+1} \triangleq 1$. Note that $|a_i - a_{i-1}| < 2/k$ for all i. Next, define the subsets

$$B_i \triangleq \{x \in S : a_{i-1} \leqslant f(x) < a_i\}, \quad 1 \leqslant i \leqslant k+1.$$

The B_i's are disjoint and $S = \cup_{i=1}^{k+1} B_i$ since $0 < f < 1$. In addition, from the continuity of f, we see that

$$\overline{B_i} \subseteq \{a_{i-1} \leqslant f \leqslant a_i\}, \ \{a_{i-1} < f < a_i\} \subseteq \mathring{B}_i.$$

Therefore,

$$\partial B_i \subseteq \{f = a_{i-1}\} \cup \{f = a_i\},$$

showing that $\mu(\partial B_i) = 0$. We consider the step function

$$g(x) \triangleq \sum_{i=1}^{k+1} a_{i-1}\mathbf{1}_{B_i}(x).$$

The function g approximates f in the sense that

$$|f(x) - g(x)| \leqslant \frac{2}{k} \quad \text{for any } x \in S,$$

which is easily seen from the construction of the B_i's and g.

It follows that,

$$\begin{aligned}
&\left|\int_S f d\mu_n - \int_S f d\mu\right| \\
&\leqslant \int_S |f(x) - g(x)| d\mu_n + \int_S |f(x) - g(x)| d\mu + \Big| \int_S g d\mu_n - \int_S g d\mu \Big| \\
&\leqslant \frac{4}{k} + \sum_{i=1}^{k+1} a_{i-1} \cdot \big|\mu_n(B_i) - \mu(B_i)\big|.
\end{aligned}$$

Since $\mu(\partial B_i) = 0$, by taking $n \to \infty$ we have

$$\overline{\lim_{n\to\infty}} \left|\int_S f d\mu_n - \int_S f d\mu\right| \leqslant \frac{4}{k}.$$

Since k is arbitrary, we conclude that $\int_S f d\mu_n \to \int_S f d\mu$.

Finally, if f is a general bounded continuous function, say $a < f(x) < b$, by considering the function

$$0 < \bar{f}(x) \triangleq \frac{f(x) - a}{b - a} < 1,$$

we are led to the previous case. The proof is now complete. □

4 Tightness and Prohorov's theorem

Knowing the existence of a weakly convergent subsequence is an important first step for many deeper problems in probability theory. Helly's theorem provides a partial answer to this question as it guarantees the existence of a vaguely convergent subsequence (though it is only true in finite dimensions). The key to ensuring that vague limit points are always probability measures is through a tightness property.

Definition 4.1. A family $\{\mu : \mu \in \Lambda\}$ of probability measures on a metric space $(S, \mathcal{B}(S), \rho)$ is said to be *tight*, if for any $\varepsilon > 0$ there exists a compact subset $K \subseteq S$, such that

$$\mu(K) \geqslant 1 - \varepsilon \quad \text{for every } \mu \in \Lambda. \tag{4.1}$$

In the context of random variables, we say that a family of real valued random variables is *tight* if the induced family of probability laws on $S = \mathbb{R}^1$ is tight.

Note that when $S = \mathbb{R}^1$, the condition (4.1) means that, for any $\varepsilon > 0$ there exists $M > 0$, such that

$$\mu([-M, M]) \geqslant 1 - \varepsilon \quad \text{for every } \mu.$$

The following result, known as *Prokhorov's theorem*, is fundamental in the study of weak convergence. We only prove the finite dimensional version, which gives the precise condition under which vague limit points are always probability measures, thus enhancing Helly's theorem to the level of weak convergence.

Theorem 4.1 (Prokhorov's theorem in $\mathbb{R}^d$)**.** *Let $\{\mu : \mu \in \Lambda\}$ be a family of probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The the following two statements are equivalent:*

(i) The family $\{\mu : \mu \in \Lambda\}$ is tight;
(ii) Every sequence in the family $\{\mu : \mu \in \Lambda\}$ admits a weakly convergent subsequence.

Proof. For simplicity we only consider the one dimensional case, i.e. when $d = 1$.

(i) $\Longrightarrow$ (ii). Let $\mu_n \in \Lambda$ be a given sequence in the family. According to Helly's theorem (cf. Theorem 2.3), there exists a subsequence μ_{n_k} and a sub-probability measure μ, such that μ_{n_k} converges vaguely to μ. We need to show that μ is a probability measure. Since the family is tight by assumption, for given $m \geqslant 1$ there exists a closed interval K_m such that

$$\mu_{n_k}(K_m) \geqslant 1 - \frac{1}{m} \quad \text{for all } k.$$

We may assume that K_m is contained in $(a_m, b_m]$, where $a_m < b_m$ are continuity points of μ such that $a_m \downarrow -\infty$ and $b_m \uparrow \infty$ (as $m \to \infty$). It follows that

$$\mu_{n_k}((a_m, b_m]) \geqslant 1 - \frac{1}{m} \quad \text{for all } k.$$

Letting $k \to \infty$, we obtain that $\mu((a_m, b_m]) \geqslant 1 - 1/m$, and by further sending $m \to \infty$ we conclude that $\mu(\mathbb{R}^1) \geqslant 1$. Therefore, μ must be a probability measure.

(ii) $\Longrightarrow$ (i). Suppose on the contrary that the family is not tight. Then there exists $\varepsilon > 0$, such that for each closed interval $[-n, n]$ one can find $\mu_n \in \Lambda$ with

$$\mu_n([-n, n]) < 1 - \varepsilon. \tag{4.2}$$

On the other hand, by the assumption of (ii), μ_n has a weakly convergent subsequence, say μ_{n_k} converging weakly to some probability measure μ. The property (4.2) implies that, for each fixed n, when k is large we have

$$\mu_{n_k}([-n, n]) \leqslant \mu_{n_k}([-n_k, n_k]) < 1 - \varepsilon.$$

It follows from Portmanteau's theorem (cf. Theorem 3.1 (iv)) that

$$\mu((-n, n)) \leqslant \varliminf_{k\to\infty} \mu_{n_k}((-n, n)) \leqslant 1 - \varepsilon$$

for every fixed n. Letting $n \to \infty$, we obtain that $\mu(\mathbb{R}^1) \leqslant 1 - \varepsilon$ which is a contradiction to the fact that μ is a probability measure. Therefore, the family $\{\mu : \mu \in \Lambda\}$ is tight. □

Example 4.1. Let $\{X_n : n \geqslant 1\}$ be a sequence of random variables such that

$$L \triangleq \sup_n \mathbb{E}[|X_n|] < \infty.$$

Then this family is tight. Indeed, let μ_n be the law of X_n. Then for each $M > 0$, we have

$$\mu_n([-M, M]^c) = \mathbb{P}(|X_n| > M) \leqslant \frac{\mathbb{E}[|X_n|]}{M} \leqslant \frac{L}{M}$$

for all $n \geqslant 1$. When M is large enough, the right hand side can be made arbitrarily small uniformly in n. This gives the tightness property.

We must point out the remarkable fact that Prokhorov's theorem holds in the general context of metric spaces. We only state the result as its proof is beyond the scope of the subject. A metric space (S, ρ) is said to be *complete*, if every Cauchy sequence is convergent to some point. Examples 3.1 and 3.2 are both complete (and separable) metric spaces. The general Prokhorov's theorem is stated as follows.

Theorem 4.2 (Prokhorov's theorem in metric spaces). *Let $\{\mu : \mu \in \Lambda\}$ be a family of probability measures defined on a separable metric space $(S, \mathcal{B}(S), \rho)$.*

(i) If the family $\{\mu : \mu \in \Lambda\}$ is tight, then every sequence in the family admits a weakly convergent subsequence.
(ii) Suppose further that S is complete. If every sequence in the family $\{\mu : \mu \in \Lambda\}$ admits a weakly convergent subsequence, then the family is tight.

5 An important example: $C[0,1]$

We conclude this topic by presenting a useful tightness criterion in the Example 3.2 of the path space $C[0,1]$. This result is important for studying convergence of stochastic processes such as functional central limit theorems.

Definition 5.1. A *stochastic process* on $[0,1]$ is a family of random variables $\{X(t) : t \in [0,1]\}$ defined over some common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Since the $X(t)$'s are random variables, there is a hidden dependence on sample points $\omega \in \Omega$. It is therefore more precise to write $X(t,\omega)$ to indicate such dependence. Instead of regarding a stochastic process as a bunch of random variables, an important perspective is that, for each fixed $\omega \in \Omega$, the function $[0,1] \ni t \mapsto X(t,\omega)$ defines a real valued path on $[0,1]$ (called a *sample path*). In this way, a stochastic process on $[0,1]$ can be equivalently viewed as a mapping from Ω to "the space of paths".

Recall that $W = C[0,1]$ is the space of continuous functions (paths) $x : [0,1] \to \mathbb{R}$ equipped with the uniform metric (cf. Example 3.2). Let $\{X(t) : t \in [0,1]\}$ be a stochastic process defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The process is said to be *continuous*, if every sample path is continuous, i.e. for every $\omega \in \Omega$, the function $[0,1] \ni t \mapsto X(t,\omega)$ is continuous. Using the sample path viewpoint, a continuous stochastic process can be defined as a mapping from Ω to W.

Definition 5.2. Let $X = \{X(t) : t \in [0,1]\}$ be a continuous stochastic process defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, viewed as a measurable mapping $X : (\Omega, \mathcal{F}) \to (W, \mathcal{B}(W))$. The *law* of X is the probability measure μ_X on $(W, \mathcal{B}(W))$ defined by

$$\mu_X(\Gamma) \triangleq \mathbb{P}(X \in \Gamma), \quad \Gamma \in \mathcal{B}(W).$$

We are often interested in the weak convergence of a sequence of stochastic processes $X_n(t)$. The following result provides a convenient criterion for proving tightness, which is usually an important ingredient in this kind of problems. Its proof, which is quite enlightening but also involved, is put in the appendix.

Theorem 5.1. *Let $X_n = \{X_n(t) : t \in [0,1]\}$ ($n \geqslant 1$) be a sequence of continuous stochastic processes defined over some common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that:*

(i) there exists $r > 0$ such that

$$\sup_{n \geqslant 1} \mathbb{E}[|X_n(0)|^r] < \infty;$$

(ii) there exists $\alpha, \beta, C > 0$ such that

$$\mathbb{E}[|X_n(t) - X_n(s)|^\alpha] \leqslant C|t-s|^{1+\beta}$$

for all $s, t \in [0,1]$ and $n \geqslant 1$.

Let μ_n be the law of X_n on $(W, \mathcal{B}(W))$. Then the sequence of probability measures $\{\mu_n : n \geqslant 1\}$ is tight.

6 Appendix: Compactness in $C[0,1]$ and proof of Theorem 5.1

The characterisation of compact subsets in $W = C[0,1]$ is given by the renowned Arzelà-Ascoli's theorem in functional analysis.

Theorem 6.1. *A subset $F \subseteq W$ is pre-compact (i.e. the closure of F is compact), if and only if the following two conditions hold:*

(i) F is bounded at $t = 0$, in the sense that there exists $M > 0$ such that

$$|w(0)| \leqslant M \quad \text{for all } w \in F.$$

(ii) F is uniformly equicontinuous, in the sense that for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|w(t) - w(s)| < \varepsilon$$

for all $w \in F$ and $s, t \in [0,1]$ with $|t-s| < \delta$.

In particular, F is compact if and only if it is closed and Conditions (i),(ii) hold.

Remark 6.1. Conditions (i) and (ii) can be equivalently reformulated in the following more concise forms:

$$\sup_{w \in F} |w(0)| < \infty$$

and

$$\lim_{\delta \downarrow 0} \sup_{w \in F} \Delta(\delta; w) = 0$$

respectively, where $\Delta(\delta; w)$ is the modulus of continuity for w defined by

$$\Delta(\delta; w) \triangleq \sup_{|t-s|<\delta} |w(t) - w(s)|. \tag{6.1}$$

Remark 6.2. In its more common form, Condition (i) is often replaced by the following uniform boundedness condition: there exists $M > 0$ such that

$$|w(t)| \leqslant M \quad \text{for all } w \in F \text{ and } t \in [0,1].$$

With the extra Condition (ii), we leave the reader to see that the above uniform boundedness condition is equivalent to Condition (i).

Remark 6.3. Let $L > 0$ be a fixed number. Define F to be the set of paths $w \in W$ such that $w(0) = 0$ and

$$|w(t) - w(s)| \leqslant L|t-s| \quad \text{for all } s,t \in [0,1].$$

Then F is compact.

Since the definition of tightness for probability measures is closely related to compact sets, it is natural to expect that tightness over W can be characterised in terms of suitable probabilistic versions of Conditions (i) and (ii) appearing in Arzelà-Ascoli's theorem. This is the content of the following result.

Theorem 6.2. *Let $\{\mu : \mu \in \Lambda\}$ be a family of probability measures on $(W, \mathcal{B}(W))$. Suppose that:*

(i) we have

$$\lim_{M\to\infty} \sup_{\mu\in\Lambda} \mu(\{w : |w(0)| > M\}) = 0;$$

(ii) for any $\varepsilon > 0$, we have

$$\lim_{\delta\downarrow 0} \sup_{\mu\in\Lambda} \mu(\{w : |\Delta(\delta; w)| > \varepsilon\}) = 0. \tag{6.2}$$

Then the family $\{\mu : \mu \in \Lambda\}$ is tight.

Proof. Let $\varepsilon > 0$. We wish to find a compact subset $K \subseteq W$ such that $\mu(K^c) < \varepsilon$ for all $\mu \in \Lambda$. To this end, by Assumption (i) we know that there exists $M > 0$, such that

$$\mu(\{w : |w(0)| > M\}) < \frac{\varepsilon}{2} \quad \text{for all } \mu \in \Lambda.$$

In addition, by Assumption (ii), for each $n > 0$ there exists $\delta_n > 0$ such that

$$\mu\big(\{w : |\Delta(\delta_n; w)| > \frac{1}{n}\}\big) < \frac{\varepsilon}{2^{n+1}} \quad \text{for all } \mu \in \Lambda.$$

Now we define

$$\Gamma_\varepsilon \triangleq \{w : |w(0)| \leqslant M\} \cap \big(\cap_{n=1}^\infty \{w : |\Delta(\delta_n; w)| \leqslant \frac{1}{n}\}\big).$$

It is easy to check that Γ_ε satisfies the two conditions in Arzelà-Ascoli's theorem and is thus pre-compact. In other words, its closure $\overline{\Gamma_\varepsilon}$ is compact. On the other hand, we also have

$$\overline{\Gamma_\varepsilon}^c \subseteq \Gamma_\varepsilon^c = \{w : |w(0)| > M\} \cup \big(\cup_{n=1}^\infty \{w : |\Delta(\delta_n; w)| > \frac{1}{n}\}\big),$$

and thus

$$\begin{aligned}\mu(\overline{\Gamma_\varepsilon}^c) &\leqslant \mu(\{w : |w(0)| > M) + \sum_{n=1}^\infty \mu\big(\{w : |\Delta(\delta_n; w)| > \frac{1}{n}\}\big) \\ &< \frac{\varepsilon}{2} + \sum_{n=1}^\infty \frac{\varepsilon}{2^{n+1}} \\ &< \varepsilon\end{aligned}$$

for any $\mu \in \Lambda$. This establishes the tightness property. □

Now we use the tightness criterion given by Theorem 6.2 to prove Theorem 5.1.

We first recall an elementary fact about real numbers that will be needed in the proof. We will make use of dyadic partitions of $[0, 1]$. For $m \geqslant 0$, define

$$D_m = \{k/2^m : 0 \leqslant k \leqslant 2^m\}$$

to be the m-th dyadic partition of $[0, 1]$. Let $D \triangleq \cup_{m=0}^\infty D_m$. D is the collection of dyadic points on $[0, 1]$. Every real number $t \in [0, 1]$ admits a unique dyadic expansion

$$t = \sum_{i=0}^\infty a_i(t) 2^{-i}$$

where $a_i(t) = 0$ or 1 for each i. If $t \in D$, then the expansion is a finite sum (i.e. there are at most finitely many 1's among the $a_i(t)$'s). For instance,

$$D \ni \frac{11}{16} = 0 \cdot 2^{-0} + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 0 + 0 + \cdots .$$

Proof of Theorem 5.1. We need to check the two conditions in Theorem 6.2 for the laws of the sequence $\{X_n : n \geqslant 1\}$ of stochastic processes.

Condition (i) is a simple consequence of Chebyshev's inequality:

$$\mathbb{P}(|X_n(0)| > M) \leqslant \frac{\mathbb{E}[|X_n(0)|^r]}{M^r} \leqslant \frac{L}{M^r}$$

where

$$L \triangleq \sup_{n \geqslant 1} \mathbb{E}[|X_n(0)|^r] < \infty.$$

Therefore,

$$\lim_{M \to \infty} \sup_{n \geqslant 1} \mathbb{P}(|X_n(0)| > M) = 0$$

and Condition (i) holds.

Checking Condition (ii) is much more challenging, which involves enlightening probabilistic ideas. Since the following argument is uniform in n, to simplify notation we write $Y(t) = X_n(t)$.

Let $\gamma \in (0, \beta/\alpha)$ be a fixed number (recall the occurrence of α, β in the second assumption of the theorem). According to Chebyshev's inequality, we have

$$\begin{aligned}
&\mathbb{P}\big(|Y(k/2^m) - Y((k-1)/2^m)| > 2^{-\gamma m}\big) \\
&\leqslant 2^{\alpha\gamma m} \cdot \mathbb{E}\big[|Y(k/2^m) - Y((k-1)/2^m)|^{\alpha}\big] \\
&\leqslant C \cdot 2^{-m(1+\beta-\alpha\gamma)},
\end{aligned}$$

for all $1 \leqslant k \leqslant 2^m$. It follows that,

$$\begin{aligned}
&\mathbb{P}\big(\max_{1 \leqslant k \leqslant 2^m} |Y(k/2^m) - Y((k-1)/2^m)| > 2^{-\gamma m}\big) \\
&\leqslant \mathbb{P}\big(\cup_{k=1}^{2^m} \big\{|Y(k/2^m) - Y((k-1)/2^m)| > 2^{-\gamma m}\big\}\big) \\
&\leqslant \sum_{k=1}^{2^m} \mathbb{P}\big(|Y(k/2^m) - Y((k-1)/2^m)| > 2^{-\gamma m}\big) \\
&\leqslant C \cdot 2^{-m(\beta-\alpha\gamma)}.
\end{aligned}$$

Note that since $\gamma < \beta/\alpha$, the right hand side is summable in m. In particular, for given $\eta > 0$, there exists $p \geqslant 1$, such that if we define

$$\Omega_p \triangleq \cup_{m=p}^{\infty} \big\{\max_{1 \leqslant k \leqslant 2^m} |Y(k/2^m) - Y((k-1)/2^m)| > 2^{-\gamma m}\big\},$$

then

$$\mathbb{P}(\Omega_p) \leqslant C \cdot \sum_{m=p}^{\infty} 2^{-m(\beta-\alpha\gamma)} < \eta.$$

We wish to show that,

$$\big\{\Delta(\delta; Y) > \varepsilon\big\} \subseteq \Omega_p \text{ or equivalently } \Omega_p^c \subseteq \big\{\Delta(\delta; Y) \leqslant \varepsilon\big\},$$

when δ is small enough, where we recall that $\Delta(\delta; w)$ is the modulus of continuity for Y defined by (6.1) and ε is a given fixed number appearing in (6.2). To this end, suppose that Ω_p^c happens, i.e.

$$|Y(k/2^m) - Y((k-1)/2^m)| \leqslant 2^{-\gamma m} \quad \text{for all } m \geqslant p \text{ and } 1 \leqslant k \leqslant 2^m.$$

Now let $s, t \in D$ (the set of dyadic points on $[0, 1]$) be such that

$$0 < |t - s| < \delta \triangleq 2^{-p}.$$

For each l we use the notation s_l (respectively, t_l) to be the largest l-th dyadic point in D_l such that $s_l \leqslant s$ (respectively, $t_l \leqslant t$). Let $m \geqslant p$ be the unique integer such that

$$2^{-(m+1)} < |t - s| < 2^{-m}.$$

Note that either $s_m = t_m$ or $t_m - s_m = 2^{-m}$. It follows that,

$$\begin{aligned}
&|Y(t) - Y(s)| \\
&\leqslant |Y(t_m) - Y(s_m)| + \sum_{l=m}^{\infty} |Y(t_{l+1}) - Y(t_l)| + \sum_{l=m}^{\infty} |Y(s_{l+1}) - Y(s_l)| \\
&\leqslant 2^{-\gamma m} + 2\sum_{l=m}^{\infty} 2^{-\gamma(l+1)} \\
&= \big(1 + \frac{2}{2^\gamma - 1}\big) \cdot 2^{-\gamma m} \\
&\leqslant 2^\gamma\big(1 + \frac{2}{2^\gamma - 1}\big)|t - s|^\gamma \\
&< 2^\gamma\big(1 + \frac{2}{2^\gamma - 1}\big) \cdot 2^{-p\gamma}.
\end{aligned}$$

If we further assume that p satisfies

$$2^\gamma\big(1 + \frac{2}{2^\gamma - 1}\big) \cdot 2^{-p\gamma} < \varepsilon$$

at the beginning, then we will have

$$|Y(t) - Y(s)| < \varepsilon.$$

Since this is true from all $s, t \in D$ with $|t - s| < \delta$ and D is dense in $[0, 1]$, by continuity we conclude that $\Delta(\delta; Y) < \varepsilon$.

Now the proof of Theorem 5.1 is complete.

□

Topic 2: The Law of Large Numbers

One of the most important results in probability theory is the law of large numbers. Heuristically, it says that if we sample from a given distribution independently, the sample average will eventually stabilise at the theoretical mean as the sample size is getting larger. The goal of this topic is to make this fundamental fact mathematically precise, and to explore some of its implications.

1 Almost sure convergence and convergence in probability

The notion of weak convergence is only concerned with the distributions of X_n and X. On the other hand, there are two stronger notions of convergence, *almost sure convergence* and *convergence in probability*, that do rely on the correlations between X_n and X as well as the common probability space on which the random variables are defined. These notions of convergence are important in the study of the law of large numbers.

In what follows, $(\Omega, \mathcal{F}, \mathbb{P})$ is a given fixed probability space, and all random variables are assumed to be defined over Ω. To discuss almost sure convergence, we first need the following definition.

Definition 1.1. A *null event* is an event $N \in \mathcal{F}$ with zero probability, i.e. $\mathbb{P}(N) = 0$. A property E, which is given by an event $E \in \mathcal{F}$, is said to hold *almost surely*, or *with probability one*, if it holds outside a null event. Equivalently, this is saying that $\mathbb{P}(E) = 1$ or $\mathbb{P}(E^c) = 0$.

Example 1.1. In the random experiment of tossing a fair coin repeatedly in a sequence, we consider the property E that a "head" appears eventually. This is an almost sure event seen as follows. For $n \geqslant 1$, let E_n be the event that a "head" first appears in the n-th toss. Then these E_n's are disjoint, and $E = \cup_{n=1}^{\infty} E_n$.

Therefore,

$$\mathbb{P}(E)=\sum_{n=1}^{\infty}\mathbb{P}(E_n)=\sum_{n=1}^{\infty}\frac{1}{2^{n-1}}\cdot\frac{1}{2}=1.$$

Now we define the notion of almost sure convergence.

Definition 1.2. Let X_n $(n\geqslant 1)$ and X be random variables defined on the probability space $(\Omega,\mathcal{F},\mathbb{P})$. We say that X_n converges to X *almost surely* or *with probability one*, if there exists a null event N, such that for any $\omega\notin N$ we have

$$\lim_{n\to\infty}X_n(\omega)=X(\omega).$$

Equivalently,

$$\mathbb{P}\big(\{\omega\in\Omega:\lim_{n\to\infty}X_n(\omega)=X(\omega)\big)=1.$$

We often use the short-handed notation "$X_n\to X$ a.s." to denote almost sure convergence.

In contrast to the notion of almost sure convergence, we have the following weaker notion of convergence: convergence in probability.

Definition 1.3. Let X_n $(n\geqslant 1)$ and X be random variables defined on the probability space $(\Omega,\mathcal{F},\mathbb{P})$. We say that X_n converges to X *in probability*, if for any $\varepsilon>0$, we have

$$\lim_{n\to\infty}\mathbb{P}(|X_n-X|>\varepsilon)=0.$$

We often use the short-handed notation "$X_n\to X$ in prob." to denote convergence in probability.

The first natural question one can ask is the relation among the three types of convergence. In fact, we have the following result.

Theorem 1.1. *Almost sure convergence $\Longrightarrow$ Convergence in probability $\Longrightarrow$ Weak convergence.*

Proof. Firstly, suppose that X_n converges to X almost surely. Let $\varepsilon>0$ be fixed. Since

$$\big\{\lim_{n\to\infty}X_n=X\big\}\subseteq\cup_{n=1}^{\infty}\cap_{m=n}^{\infty}\big\{|X_m-X|\leqslant\varepsilon\big\},$$

we know that

$$\begin{aligned}1 &= \mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) \\ &\leqslant \lim_{n\to\infty} \mathbb{P}\big(\cap_{m=n}^{\infty}\left\{|X_m - X| \leqslant \varepsilon\right\}\big) \\ &\leqslant \lim_{n\to\infty} \mathbb{P}\big(|X_n - X| \leqslant \varepsilon\big).\end{aligned}$$

Therefore, $\mathbb{P}(|X_n - X| \leqslant \varepsilon)] \to 1$, or equivalently $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$. This gives convergence in probability.

Now suppose that X_n converges to X in probability. We use the second charaterisation in the Portmanteau theorem to show that X_n converges weakly to X. To this end, let f be a bounded and uniformly continuous function on $\mathbb{R}$. Given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|x - y| \leqslant \delta \implies |f(x) - f(y)| \leqslant \varepsilon.$$

It follows that

$$\begin{aligned}&\big|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]\big| \\ &\leqslant \mathbb{E}[|f(X_n) - f(X)|] \\ &= \mathbb{E}[|f(X_n) - f(X)|; |X_n - X| \leqslant \delta] + \mathbb{E}[|f(X_n) - f(X)|; |X_n - X| > \delta] \\ &\leqslant \varepsilon + 2\|f\|_\infty \mathbb{P}(|X_n - X| > \delta),\end{aligned}$$

where $\|f\|_\infty \triangleq \sup_{x\in\mathbb{R}} |f(x)|$. Since $X_n \to X$ in probability, by letting $n \to \infty$ we see that

$$\overline{\lim_{n\to\infty}} \big|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]\big| \leqslant \varepsilon.$$

As ε is arbitrary, we conclude that $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$, yielding the desired weak convergence. □

We may not be surprised by the fact that none of the reverse directions in Theorem 1.1 is true. This is illustrated by the following example.

Example 1.2. (i) *Convergence in probability does not imply almost sure convergence.* Consider the random experiment of choosing a point $\omega \in \Omega = [0, 1]$ uniformly at random. We construct a sequence $\{Y_n : n \geqslant 1\}$ of random variables as follows. Firstly, divide $[0, 1]$ into two sub-intervals, and define $Y_1 \triangleq \mathbf{1}_{[0,1/2]}$ and $Y_2 \triangleq \mathbf{1}_{[1/2,1]}$. Next, divide $[0, 1]$ into three sub-intervals, and define $Y_3 \triangleq \mathbf{1}_{[0,1/3]}$,

$Y_4 \triangleq \mathbf{1}_{[1/3,2/3]}$ and $Y_5 \triangleq \mathbf{1}_{[2/3,1]}$. Now the procedure continues in the obvious way to define the whole sequence $\{Y_n\}$. Since the event

$$\{\omega \in [0,1] : |Y_n(\omega)| > \varepsilon\} = \{\omega : Y_n(\omega) = 1\}$$

is given by a particular sub-interval whose length tends to zero, we conclude that Y_n converges to zero in probability. However, $Y_n(\omega)$ does not converge to zero at any $\omega \in [0,1]$. Indeed, for each ω, by the construction there must exist a subsequence n_k such that $Y_{n_k}(\omega) = 1$ for all k.
(ii) *Weak convergence does not imply convergence in probability.* Let W be a Bernoulli random variable with parameter $1/2$. Define $Z_n \triangleq W$ for all n and $Z \triangleq 1 - W$. Since Z_n and Z are both Bernoulli random variables with parameter $1/2$, it is trivial that Z_n converges weakly to Z. However, for any $0 < \varepsilon < 1$, we have

$$\mathbb{P}\big(|Z_n - Z| > \varepsilon\big) = \mathbb{P}(|2W - 1| > \varepsilon) = 1.$$

Therefore, Z_n does not converge to Z in probability.

2 Independence and Borel-Cantelli's lemma

In general, obtaining almost sure convergence is much more challenging than proving convergence in probability. However, there is a rather power tool, which allows us to establish almost sure properties fairly easily in many situations. This is known as *Borel-Cantelli's lemma.* Before discussing it, let us first recall the basic notion of independence. Throughout the rest, we are always given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Events and random variables are assumed to be defined on it.

Two events $A, B \in \mathcal{F}$ are said to be *independent*, if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Two random variables $X, Y : \Omega \to \mathbb{R}$ are said to be *independent*, if

$$\mathbb{P}(X \in E, Y \in F) = \mathbb{P}(X \in E) \cdot \mathbb{P}(Y \in F) \quad \text{for all } E, F \in \mathcal{B}(\mathbb{R}).$$

Using standard measure-theoretic arguments, one can show that two random variables X, Y are independent if and only if

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]$$

for all bounded Borel-measurable functions f, g on $\mathbb{R}$. It is also equivalent to the condition that the joint cumulative distribution function of (X, Y) equals the product of the marginal ones, i.e.

$$\mathbb{P}(X \leqslant x, Y \leqslant y) = \mathbb{P}(X \leqslant x) \cdot \mathbb{P}(Y \leqslant y) \quad \text{for all } x, y \in \mathbb{R}.$$

Given an event A, one can define an associated random variable (the indicator random variable of A) by

$$X_A(\omega) \triangleq \begin{cases} 1, & \omega \in A; \\ 0, & \omega \notin A. \end{cases}$$

In this way, the independence for two events A, B is equivalent to the independence for the associated indicator random variables X_A and X_B. Therefore, it is enough to consider independence for random variables.

To study convergence of random variables, we need to extend the notion of independence to sequences of random variables.

Definition 2.1. A sequence $\{X_n : n \geqslant 1\}$ of random variables are said to be *independent*, if for any $n \geqslant 1$ and any $E_1, \cdots, E_n \in \mathcal{B}(\mathbb{R})$, we have

$$\mathbb{P}(X_1 \in E_1, \cdots, X_n \in E_n) = \mathbb{P}(X_1 \in E_1) \cdots \mathbb{P}(X_n \in E_n).$$

A sequence $\{A_n : n \geqslant 1\}$ of events are said to be *independent*, if the associated sequence $\{X_{A_n} : n \geqslant 1\}$ of indicator random variables are independent.

Now we present Borel-Cantelli's lemma, which is an extremely powerful tool in many probabilistic applications. Recall that, given a sequence $\{A_n : n \geqslant 1\}$ of events,

$$\overline{\lim_{n\to\infty}} A_n \triangleq \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m$$

defines the event that "A_n happens for infinitely many n's", or equivalently "A_n happens infinitely often". Sometimes we simply write this event as "A_n i.o." Respectively,

$$\underline{\lim_{n\to\infty}} A_n \triangleq \cup_{n=1} \cap_{m=n} A_m$$

defines the event that "from some point on every A_n happens", or equivalently "A_n happens for all but finitely many n's". Sometimes we simply write this event as "A_n happens eventually." It is obvious that

$$\big(\overline{\lim_{n\to\infty}} A_n\big)^c = \underline{\lim_{n\to\infty}} A_n^c, \ \big(\underline{\lim_{n\to\infty}} A_n\big)^c = \overline{\lim_{n\to\infty}} A_n^c.$$

Theorem 2.1. *Let $\{A_n : n \geqslant 1\}$ be a sequence of events.*

(i) [The first Borel-Cantelli's lemma] If

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$

then with probability zero A_n happens infinitely often, namely

$$\mathbb{P}\big(\overline{\lim_{n\to\infty}} A_n\big) = 0.$$

(ii) [The second Borel-Cantelli's lemma] Suppose that the sequence $\{A_n : n \geqslant 1\}$ are independent. If

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty,$$

then with probability one A_n happens infinitely often, namely

$$\mathbb{P}\big(\overline{\lim_{n\to\infty}} A_n\big) = 1.$$

Proof. (i) By assumption, we have

$$\begin{aligned}\mathbb{P}\big(\overline{\lim_{n\to\infty}} A_n\big) &= \mathbb{P}\big(\cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m\big) = \lim_{n\to\infty} \mathbb{P}(\cup_{m=n}^{\infty} A_m) \\ &\leqslant \lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbb{P}(A_m) = 0.\end{aligned}$$

(ii) We look at the complement:

$$\begin{aligned}\mathbb{P}\big(\big(\overline{\lim_{n\to\infty}} A_n\big)^c\big) &= \mathbb{P}\big(\cup_{n=1}^{\infty} \cap_{m=n}^{\infty} A_m^c\big) = \lim_{n\to\infty} \mathbb{P}\big(\cap_{m=n}^{\infty} A_m^c\big) \\ &= \lim_{n\to\infty} \lim_{N\to\infty} \mathbb{P}\big(\cap_{m=n}^{N} A_m^c\big).\end{aligned}$$

Now we analyse the above limit. First of all, by independence we know that

$$\begin{aligned}\mathbb{P}\big(\cap_{m=n}^{N} A_m^c\big) &= (1 - \mathbb{P}(A_n)) \cdots (1 - \mathbb{P}(A_N)) \\ &= \exp\big(\sum_{m=n}^{N} \log(1 - \mathbb{P}(A_m)\big) \\ &\leqslant \exp\big(-\sum_{m=n}^{N} \mathbb{P}(A_m))\big),\end{aligned}$$

where we have used the simple fact that $\log(1-x) \leqslant -x$. Since $\sum_{n=1}^{\infty}\mathbb{P}(A_n)=\infty$, by letting $N\to\infty$ we have,

$$\lim_{N\to\infty}\mathbb{P}\big(\cap_{m=n}^{N}A_m^c\big)\leqslant \exp\big(-\sum_{m=n}^{\infty}\mathbb{P}(A_m)\big)=\exp(-\infty)=0.$$

This is true for every n. Therefore,

$$\mathbb{P}\big(\big(\overline{\lim_{n\to\infty}}A_n\big)^c\big)=\lim_{n\to\infty}\lim_{N\to\infty}\mathbb{P}\big(\cap_{m=n}^{N}A_m^c\big)=0,$$

and the result follows. □

The independence assumption is essential for the second Borel-Cantelli's lemma to hold, as illustrated by the following example.

Example 2.1. Let X be a uniform random variable over $[0,1]$. Define $A_n \triangleq \{X \leqslant 1/n\}$ $(n \geqslant 1)$. Then $\mathbb{P}(A_n)=\frac{1}{n}$ and thus $\sum_n \mathbb{P}(A_n)=\infty$. However,

$$\overline{\lim_{n\to\infty}}A_n=\{X=0\}$$

which is an event of zero probability. Note that the A_n's are apparently not independent.

Remark 2.1. The second Borel-Cantelli's lemma remains true if the independence assumption is weakened as *pairwise independence*, i.e. only assuming that A_n and A_m are independent for each pair of (n,m).

Example 2.2. Suppose we toss a fair coin independently in a sequence. Let A_1 be the event that the first 10^{10} consecutive tosses all end up being "head", let A_2 be next 10^{10} consecutive tosses all end up being "head", and so forth. It is obvious that these events A_n are independent, and each one has a rather small probability:

$$\mathbb{P}(A_n)=\big(\frac{1}{2}\big)^{10^{10}}>0.$$

However, we have $\sum_{n=1}^{\infty}\mathbb{P}(A_n)=\infty$. According to the second Borel-Cantelli's lemma, we conclude that

$$\mathbb{P}\big(\overline{\lim_{n\to\infty}}A_n\big)=1.$$

In other words, with probability one A_n happens infinitely often. This implies that, with probability one, we will see infinitely many intervals of length 10^{10} that contain only "head"! There is another interesting way of describing this phenomenon. If a monkey randomly types one letter at each time, then with probability one it will eventually produces an exact copy of Shakespeare's "Hamlet" (in fact infinitely many copies!). Now the next question is: how long does it on average for the monkey to first produce such a copy?

3 The weak law of large numbers

We demonstrate an important application of Borel-Cantelli's lemma to the proof of the *weak law of large numbers*. We first prove a simple property for the expectation that will be used later on. Recall that if $X \geqslant 0$, then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)dx.$$

Lemma 3.1. *Let X be non-negative random variable with finite mean. Then*

$$\sum_{n=1}^\infty \mathbb{P}(X > n) < \infty.$$

Proof. We have

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty \mathbb{P}(X > x)dx = \sum_{n=1}^\infty \int_{n-1}^n \mathbb{P}(X > x)dx \\ &\geqslant \sum_{n=1}^\infty \int_{n-1}^n \mathbb{P}(X > n)dx = \sum_{n=1}^\infty \mathbb{P}(X > n),\end{aligned}$$

and the result follows. □

The weak law of large numbers is stated as follows.

Theorem 3.1. *Let $\{X_n : n \geqslant 1\}$ be a sequence of pairwise independent, identically distributed random variables with finite mean m. Define $S_n \triangleq X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to m \quad \text{in prob.} \tag{3.1}$$

as $n \to \infty$.

Before developing the proof, we first examine a rather simple but enlightening situation. For the moment, let us further assume that all the X_n's have finite variance σ^2. By Chebyshev's inequality, in this case we have

$$\mathbb{P}\big(\big|\frac{S_n}{n} - m\big| > \varepsilon\big) \leqslant \frac{1}{\varepsilon^2}\text{Var}\big[\frac{S_n}{n}\big] = \frac{1}{\varepsilon^2 n^2}\text{Var}[S_n] = \frac{\sigma^2}{n\varepsilon^2}. \tag{3.2}$$

This trivially gives the convergence (3.1). If we think about this argument in a deeper way, the key point is that $\text{Var}[S_n] = o(n^2)$ as $n \to \infty$.

The main idea to treat the general case is truncating X_n to a bounded random variable. This is a very common technique in the study of probabilistic limit theorems.

Proof of Theorem 3.1. We divide the proof into several steps. Let $F(x)$ be the cumulative distribution function of X_1 (equivalently, of any X_n).

Step one: truncation. We define

$$Y_n \triangleq \begin{cases} X_n, & \text{if } |X_n| \leqslant n; \\ 0, & \text{otherwise.} \end{cases}$$

Observe that $\{X_n \neq Y_n\} = \{|X_n| > n\}$. Since all the X_n's are identically distributed, we know that

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1| > n) < \infty,$$

where the last part follows from Lemma 3.1. According to the first Borel-Cantelli's lemma, we have

$$\mathbb{P}\big(X_n \neq Y_n \ \text{ for infinitely many } n\big) = 0.$$

In other words, with probability one, $X_n = Y_n$ for all n sufficiently large.

Step two: the weak law of large numbers for $\{Y_n\}$. Define $T_n \triangleq Y_1 + \cdots + Y_n$. Inspired by the argument for (3.2), let us estimate $\mathrm{Var}[T_n]$. Since $Y_1, \cdots, Y_n$ are independent, we have

$$\mathrm{Var}[T_n] = \sum_{j=1}^{n} \mathrm{Var}[Y_j] \leqslant \sum_{j=1}^{n} \mathbb{E}[Y_j^2].$$

Our goal is to show that the above quantity is of $o(n^2)$. By the construction of Y_n, we have

$$\begin{aligned} \sum_{j=1}^{n} \mathbb{E}[Y_j^2] &= \sum_{j=1}^{n} \mathbb{E}[X_j^2 \mathbf{1}_{\{|X_j| \leqslant j\}}] = \sum_{j=1}^{n} \int_{\{|x| \leqslant j\}} x^2 dF(x) \\ &= \sum_{j \leqslant \sqrt{n}} \int_{\{|x| \leqslant j\}} x^2 dF(x) + \sum_{\sqrt{n} < j \leqslant n} \int_{\{|x| \leqslant j\}} x^2 dF(x). \end{aligned} \tag{3.3}$$

We estimate the above two sums separately. For the first one,

$$\begin{aligned} \sum_{j \leqslant \sqrt{n}} \int_{\{|x| \leqslant j\}} x^2 dF(x) &\leqslant \sum_{j \leqslant \sqrt{n}} \int_{\{|x| \leqslant j\}} \sqrt{n} \cdot |x| dF(x) \\ &\leqslant \sum_{j \leqslant \sqrt{n}} \sqrt{n} \int_{-\infty}^{\infty} |x| dF(x) \\ &= n \cdot \mathbb{E}[|X_1|]. \end{aligned}$$

For the second one,

$$\begin{aligned}\sum_{\sqrt{n}<j\leqslant n}\int_{\{|x|\leqslant j\}}x^2dF(x) &= \sum_{\sqrt{n}<j\leqslant n}\big(\int_{\{|x|\leqslant\sqrt{n}\}}x^2dF(x)+\int_{\{\sqrt{n}<|x|\leqslant j\}}x^2dF(x)\big)\\ &\leqslant n\sqrt{n}\cdot\int_{-\infty}^{\infty}|x|dF(x)+n^2\int_{\{|x|>\sqrt{n}\}}|x|dF(x)\\ &= n\sqrt{n}\cdot\mathbb{E}[|X_1|]+n^2\mathbb{E}[|X_1|\cdot\mathbf{1}_{\{|X_1|>\sqrt{n}\}}].\end{aligned}$$

Note that

$$\lim_{n\to\infty}\mathbb{E}[|X_1|\cdot\mathbf{1}_{\{|X_1|>\sqrt{n}\}}]=0$$

since $\mathbb{E}[|X_1|]<\infty$ and $\mathbb{P}(|X_1|>\sqrt{n})\to 0$. Therefore, we see that both sums on the right hand side of (3.3) is of $o(n^2)$, and thus $\mathrm{Var}[T_n]=o(n^2)$. It follows in the same way as in (3.2) that

$$\frac{T_n-\mathbb{E}[T_n]}{n}\to 0 \quad \text{in prob.}$$

as $n\to\infty$.

Step three: relating back to the sequence $\{X_n\}$. To complete the proof, let us compare $\frac{S_n}{n}-m$ with $\frac{T_n-\mathbb{E}[T_n]}{n}$. We first observe that

$$\big|\big(\frac{S_n}{n}-m\big)-\big(\frac{T_n-\mathbb{E}[T_n]}{n}\big)\big|\leqslant\frac{|S_n-T_n|}{n}+\big|\frac{\mathbb{E}[T_n]}{n}-m\big|.$$

In Step One, we have seen that with probability one, $X_n=Y_n$ for all sufficiently large n. This implies that, with probability one,

$$S_n-T_n=(X_1-Y_1)+\cdots+(X_n-Y_n)$$

stops depending on n after some point and thus

$$\frac{|S_n-T_n|}{n}\to 0 \quad \text{as } n\to\infty.$$

In addition, it is apparent that

$$\mathbb{E}[Y_n]=\int_{\{|x|\leqslant n\}}xdF(x)\to\int_{-\infty}^{\infty}xdF(x)=m$$

as $n\to\infty$. This implies that,

$$\frac{\mathbb{E}[T_n]}{n}=\frac{\mathbb{E}[Y_1]+\cdots+\mathbb{E}[Y_n]}{n}\to m,$$

where we have used the elementary analytic fact that

$$a_n \to a \implies \frac{a_1 + \cdots + a_n}{n} \to a.$$

To summarise, we conclude that with probability one,

$$\lim_{n\to\infty} \big|\big(\frac{S_n}{n} - m\big) - \big(\frac{T_n - \mathbb{E}[T_n]}{n}\big)\big| = 0.$$

Combining with Step Two, the result follows.

□

We discuss an interesting application of the weak law of large numbers to the approximation of continuous functions. Constructing polynomial approximations of continuous functions is an important question in practice. The following result, known as *Bernstein's approximation theorem*, provides an elegant solution to this question.

Theorem 3.2. *Let $f(x)$ be a continuous function on $[0,1]$. For each $n \geqslant 1$, define the polynomial*

$$p_n(x) \triangleq \sum_{k=0}^{n} f\big(\frac{k}{n}\big)\Big(\begin{array}{c} n \\ k \end{array}\Big)x^k(1-x)^{n-k}, \quad x \in [0,1].$$

Then p_n converges to f uniformly on $[0,1]$ as $n \to \infty$.

Proof. Fix $x \in [0,1]$. Let $\{X_n : n \geqslant 1\}$ be a sequence of independent and identically distributed random variables, each following the Bernoulli distribution with parameter x, i.e.

$$\mathbb{P}(X_n = 1) = x, \ \mathbb{P}(X_n = 0) = 1 - x.$$

Define $S_n \triangleq X_1 + \cdots + X_n$. It is straight forward to see that $p_n(x) = \mathbb{E}\big[f\big(\frac{S_n}{n}\big)\big]$. According to the weak law of large numbers (cf. Theorem 3.1), $\frac{S_n}{n} \to \mathbb{E}[X_1] = x$ in probability. In particular, $\frac{S_n}{n} \to x$ weakly. Since f is bounded continuous, this already implies that

$$p_n(x) = \mathbb{E}\big[f\big(\frac{S_n}{n}\big)\big] \to \mathbb{E}[f(x)] = f(x),$$

for every given $x \in [0,1]$.

Proving uniform convergence requires a little bit of extra effort. First of all, since f is uniformly continuous on $[0,1]$, for any given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$x, y \in [0,1],\ |x-y| \leqslant \delta \implies |f(x)-f(y)| \leqslant \varepsilon.$$

Exactly the same argument as the proof of Theorem 1.1 (the second part) gives

$$\big|p_n(x) - f(x)\big| \leqslant \varepsilon + 2\|f\|_\infty \cdot \mathbb{P}\big(\big|\frac{S_n}{n} - x\big| > \delta\big),$$

where $\|f\|_\infty \triangleq \sup_{x\in[0,1]} |f(x)|$ denotes the supremum norm of f. In addition, from Chebyshev's inequality we know that

$$\mathbb{P}\big(\big|\frac{S_n}{n} - x\big| > \delta\big) \leqslant \frac{1}{\delta^2}\mathrm{Var}\big[\frac{S_n}{n}\big] = \frac{x(1-x)}{n\delta^2} \leqslant \frac{1}{4n\delta^2},$$

where we have used the elementary inequality that $x(1-x) \leqslant \frac{1}{4}$. Therefore, we arrive at

$$\big|p_n(x) - f(x)\big| \leqslant \varepsilon + \frac{\|f\|_\infty}{2n\delta^2}.$$

When n is large, the right hand side can be made smaller than 2ε, uniformly in $x \in [0,1]$. This concludes the desired uniform convergence. □

A remarkable fact is that, the conclusion of Theorem 3.1 can be strengthened to almost sure convergence under the same assumptions, hence yielding a strong law of large numbers. In the next section, we will prove a version of such result under the stronger assumption of total independence.

4 The strong law of large numbers

Strong laws of large numbers concern with convergence in the almost sure sense. Establishing this type of laws is more challenging than proving weak laws. From the viewpoint of real analysis, laws of large numbers are essentially related to the following type of convergence properties:

$$\frac{1}{a_n}\sum_{j=1}^n x_j \to 0 \tag{4.1}$$

where $0 < a_n \uparrow \infty$ and $x_n \in \mathbb{R}$. It is often the case that $a_n = n$ and $x_n = X_n(\omega) - \mathbb{E}[X_n]$. The property (4.1) can be obtained by means of the so-called *Kronecker's lemma.*

Lemma 4.1. *Let $\{x_n : n \geqslant 1\}$ be a real sequence and $\{a_n : n \geqslant 1\}$ be a positive sequence increasing to infinity. If the series $\sum_{n=1}^{\infty} \frac{x_n}{a_n}$ is convergent, then (4.1) holds.*

Kronecker's lemma will be proved in the appendix. This lemma inspires us that, we can study strong laws of large numbers through the convergence of certain random series.

4.1 Kolmogorov's two-series theorem

We start by taking some effort to discuss random series. Let $\{X_n : n \geqslant 1\}$ be a sequence of random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that the random series $\sum_{n=1}^{\infty} X_n$ is *convergent almost surely* (a.s.), if

$$\mathbb{P}\big(\{\omega : \sum_{n=1}^{\infty} X_n(\omega) \text{ is convergent}\}\big) = 1.$$

Recall from the Cauchy criterion that a real series $\sum_{n=1}^{\infty} x_n$ is convergent if and only if for any $\varepsilon > 0$, there exists $n \geqslant 1$, such that for any $i, j > n$ we have

$$\big|s_i - s_j\big| < \varepsilon,$$

where $s_n \triangleq x_1 + \cdots + x_n$ is the partial sum sequence. A natural probabilistic analogue of the Cauchy criterion gives us the following characterisation for the almost sure convergence of random series. Its proof is however quite technical, and we leave it in the appendix.

Proposition 4.1. *Let $\{X_n : n \geqslant 1\}$ be a sequence of random variables and set $S_n \triangleq X_1 + \cdots + X_n$. The random series $\sum_{n=1}^{\infty} X_n$ is convergent a.s. if and only if for any $\varepsilon > 0$, we have*

$$\lim_{n\to\infty} \lim_{N\to\infty} \mathbb{P}\big(\max_{n\leqslant l\leqslant N} |S_l - S_n| \geqslant \varepsilon\big) = 0 \tag{4.2}$$

The probabilistic Cauchy criterion (4.2) is in general quite difficult to verify. However, there is rather simple criterion due to Kolmogorov in the context of independent random variables. This is known as *Kolmogorov's two-series theorem.*

Theorem 4.1 (Kolmogorov's two-series theorem)**.** *Let $\{X_n : n \geqslant 1\}$ be a sequence of independent random variables. Suppose that each X_n has finite variance. If both of the real series $\sum_n \mathbb{E}[X_n]$ and $\sum_n \mathrm{Var}[X_n]$ are convergent, then the random series $\sum_n X_n$ is convergent a.s.*

This theorem is an immediate consequence of the following inequality also due to Kolmogorov, the proof of which is ingenious.

Lemma 4.2. *Let $X_1, \cdots, X_n$ be independent random variables. Suppose that $\mathbb{E}[X_k] = 0$ and $\mathrm{Var}[X_k] < \infty$ for each k. Then for any $\varepsilon > 0$, we have*

$$\mathbb{P}\big(\max_{1\leqslant k\leqslant n} |S_k| \geqslant \varepsilon\big) \leqslant \frac{1}{\varepsilon^2}\sum_{k=1}^{n} \mathrm{Var}[X_k], \tag{4.3}$$

where $S_k \triangleq X_1 + \cdots + X_k$.

Proof. We decompose the event

$$A \triangleq \big\{\max_{1\leqslant k\leqslant n} |S_k| \geqslant \varepsilon\big\}$$

according to the first k such that $|S_k| \geqslant \varepsilon$. More precisely, for each $1 \leqslant k \leqslant n$ we introduce the event

$$A_k \triangleq \big\{|S_1| < \varepsilon, \cdots, |S_{k-1}| < \varepsilon,\ |S_k| \geqslant \varepsilon\big\}.$$

It is obvious that $A_1, \cdots, A_n$ are disjoint and

$$A = \cup_{k=1}^{n} A_k.$$

Therefore,

$$\mathbb{P}\big(A\big) = \sum_{k=1}^{n} \mathbb{P}(A_k) \leqslant \frac{1}{\varepsilon^2}\sum_{k=1}^{n} \mathbb{E}[S_k^2 \mathbf{1}_{A_k}], \tag{4.4}$$

where the last inequality follows from the fact that $|S_k| \geqslant \varepsilon$ on A_k.

Here comes the crucial point. We claim that

$$\mathbb{E}[S_k^2 \mathbf{1}_{A_k}] \leqslant \mathbb{E}[S_n^2 \mathbf{1}_{A_k}] \tag{4.5}$$

for every k. Coming up with such an observation is much harder than its proof, which requires some insight from the viewpoint of martingales. Let us just verify this property directly. Note that

$$\begin{aligned}\mathbb{E}[S_n^2 \mathbf{1}_{A_k}] &= \mathbb{E}[(S_n - S_k + S_k)^2 \mathbf{1}_{A_k}] \\ &= \mathbb{E}[(S_n - S_k)^2 \mathbf{1}_{A_k}] + 2\mathbb{E}[(S_n - S_k)S_k \mathbf{1}_{A_k}] + \mathbb{E}[S_k^2 \mathbf{1}_{A_k}]. \end{aligned} \tag{4.6}$$

Since $X_1, \cdots, X_n$ are independent, we have

$$\begin{aligned}\mathbb{E}[(S_n - S_k)S_k\mathbf{1}_{A_k}] &= \mathbb{E}[(X_{k+1} + \cdots + X_n)S_k\mathbf{1}_{A_k}] \\ &= \mathbb{E}[X_{k+1} + \cdots + X_n] \cdot \mathbb{E}[S_k\mathbf{1}_{A_k}] \\ &= 0.\end{aligned}$$

In addition, the first term on the right hand side of (4.6) is non-negative. Therefore, the property (4.5) holds.

It follows from (4.4) and (4.5) that

$$\mathbb{P}\big(A\big) \leqslant \frac{1}{\varepsilon^2}\sum_{k=1}^{n}\mathbb{E}[S_n^2\mathbf{1}_{A_k}] = \frac{1}{\varepsilon^2}\mathbb{E}[S_n^2\mathbf{1}_A] \leqslant \frac{1}{\varepsilon^2}\mathbb{E}[S_n^2].$$

On the other hand, we also have

$$\begin{aligned}\mathbb{E}[S_n^2] &= \mathbb{E}[(X_1 + \cdots + X_n)^2] \\ &= \sum_{k=1}^{n}\mathbb{E}[X_k^2] + \sum_{i\neq j}\mathbb{E}[X_iX_j] \\ &= \sum_{k=1}^{n}\mathbb{E}[X_k^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j] \\ &= \sum_{k=1}^{n}\mathrm{Var}[X_k].\end{aligned}$$

Therefore, the result follows. □

Using Kolmogorov's inequality (4.3), the proof Theorem 4.1 is almost trivial.

Proof of Theorem 4.1. We will verify the Cauchy criterion (4.2). Without loss of generality, we may assume that $\mathbb{E}[X_n] = 0$, for otherwise we can consider the sequence $X_n - \mathbb{E}[X_n]$ instead. In this case, according to inequality (4.3), for any $\varepsilon > 0$ we have

$$\mathbb{P}\big(\max_{n\leqslant l\leqslant N}|S_l - S_n| \geqslant \varepsilon\big) \leqslant \frac{1}{\varepsilon^2}\big(\mathrm{Var}[X_{n+1}] + \cdots + \mathrm{Var}[X_N]\big),$$

where $S_n \triangleq X_1 + \cdots + X_n$ is the partial sum sequence. It follows that

$$\lim_{N\to\infty}\mathbb{P}\big(\max_{n\leqslant l\leqslant N}|S_l - S_n| \geqslant \varepsilon\big) \leqslant \frac{1}{\varepsilon^2}\sum_{k=n+1}^{\infty}\mathrm{Var}[X_k].$$

Since $\sum_n \mathrm{Var}[X_n] < \infty$, we obtain

$$\lim_{n\to\infty}\lim_{N\to\infty}\mathbb{P}\big(\max_{n\leqslant l\leqslant N}|S_l - S_n| \geqslant \varepsilon\big) \leqslant \frac{1}{\varepsilon^2}\lim_{n\to\infty}\sum_{k=n+1}^{\infty}\mathrm{Var}[X_k] = 0.$$

Therefore, (4.2) holds and we conclude that the random series $\sum_n X_n$ converges a.s.

□

4.2 The strong laws of large numbers

We now use Kronecker's lemma and Kolmogorov's two-series theorem to establish the strong law of large numbers. It strengthens the weak law (cf. Theorem 3.1) but under the stronger assumption of total independence.

Theorem 4.2. *Let $\{X_n : n \geqslant 1\}$ be a sequence of independent and identically distributed random variables.*

(i) If $\mathbb{E}[|X_1|] < \infty$, then

$$\frac{S_n}{n} \to \mathbb{E}[X_1] \quad \text{a.s.}$$

as $n \to \infty$.
(ii) If $\mathbb{E}[|X_1|] = \infty$, then

$$\overline{\lim_{n\to\infty}}\frac{|S_n|}{n} = \infty \quad \text{a.s.}$$

as $n \to \infty$.

Proof. (i) *Step one.* We use the same truncation idea as in the proof of the weak law. Define $Y_n \triangleq X_n\mathbf{1}_{\{|X_n|\leqslant n\}}$. According to the first step in that proof, we know that with probability one, $X_n = Y_n$ for all n sufficiently large. In particular, this implies that

$$\frac{1}{n}\sum_{j=1}^{n}(X_j - Y_j) \to 0 \quad \text{a.s.} \tag{4.7}$$

as $n \to \infty$, since a.s. the sum stops depending on n after some point.

Step two. We try to apply Kolmogorov's two-series theorem to the random series $\sum_n Z_n$ where $Z_n \triangleq \frac{Y_n - \mathbb{E}[Y_n]}{n}$. Since Z_n has mean zero, we only need to check that $\sum_n \mathrm{Var}[Z_n] < \infty$. For this purpose, note that

$$\mathrm{Var}[Z_n] \leqslant \frac{1}{n^2}\mathbb{E}[Y_n^2] = \frac{1}{n^2}\int_{\{|x|\leqslant n\}} x^2 dF(x)$$

where $F(x)$ is the cumulative distribution function of X_1. It follows that

$$\begin{aligned}
&\sum_{n=1}^{\infty}\mathrm{Var}[Z_n]\\
&\leqslant \sum_{n=1}^{\infty}\frac{1}{n^2}\int_{\{|x|\leqslant n\}}x^2dF(x)\\
&=\sum_{n=1}^{\infty}\frac{1}{n^2}\sum_{j=1}^{n}\int_{\{j-1<|x|\leqslant j\}}x^2dF(x)\\
&=\sum_{j=1}^{\infty}\big(\int_{\{j-1<|x|\leqslant j\}}x^2dF(x)\big)\sum_{n=j}^{\infty}\frac{1}{n^2}\quad \text{(exchange of summation).}
\end{aligned}$$

In the first place, we have

$$\int_{\{j-1<|x|\leqslant j\}}x^2dF(x)\leqslant j\cdot\int_{\{j-1<|x|\leqslant j\}}|x|dF(x).$$

In addition, we also know that

$$\sum_{n=j}^{\infty}\frac{1}{n^2}\leqslant\sum_{n=j}^{\infty}\frac{1}{(n-1)n}=\sum_{n=j}^{\infty}\big(\frac{1}{n-1}-\frac{1}{n}\big)=\frac{1}{j-1}\leqslant\frac{2}{j}$$

when $j\geqslant 2$. This inequality is also true when $j=1$ since

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}<2.$$

Therefore,

$$\begin{aligned}
\sum_{n=1}^{\infty}\mathrm{Var}[Z_n]&\leqslant\sum_{j=1}^{\infty}j\cdot\big(\int_{\{j-1<|x|\leqslant j\}}|x|dF(x)\big)\cdot\frac{2}{j}\\
&=2\sum_{j=1}^{\infty}\int_{\{j-1<|x|\leqslant j\}}|x|dF(x)\\
&=2\mathbb{E}[|X_1|]<\infty.
\end{aligned}$$

By Kolmogorov's two-series theorem, we know that $\sum_n Z_n$ converges a.s. According to Kronecker's lemma (cf. Lemma 4.1) with $a_n=n$ and $x_n=Y_n-\mathbb{E}[Y_n]$, this implies that

$$\frac{1}{n}\sum_{j=1}^{n}(Y_j-\mathbb{E}[Y_j])\to 0\quad \text{a.s.}$$

as $n \to \infty$.

Step three. We have

$$\mathbb{E}[Y_n] = \int_{\{|x| \leqslant n\}} x dF(x) \to \mathbb{E}[X_1].$$

Therefore,

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[Y_j] \to \mathbb{E}[X_1],$$

and thus

$$\frac{1}{n}\sum_{j=1}^{n} Y_j \to \mathbb{E}[X_1] \quad \text{a.s.}$$

as $n \to \infty$. Now the assertion follows from (4.7) in Step One.

(ii) Suppose that $\mathbb{E}[|X_1|] = \infty$. A simple adaptation of the proof of Lemma 3.1 implies that for any given $A > 0$, we have $\sum_{n=1}^{\infty} \mathbb{P}(|X_1| > An) = \infty$. Since the X_n's are independent and identically distributed, we see that

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > An) = \infty,$$

which further implies by the second Borel-Cantelli's lemma that

$$\mathbb{P}(|X_n| > An \quad \text{for infinitely many } n) = 1.$$

Observe that

$$\{X_n > An\} \subseteq \big\{|S_n| > \frac{An}{2}\big\} \cup \big\{|S_{n-1}| > \frac{A(n-1)}{2}\big\}.$$

Therefore, we have

$$\mathbb{P}\big(|S_n| > \frac{An}{2} \quad \text{for infinitely many } n\big) = 1.$$

Now if we define

$$\Omega_m \triangleq \big\{|S_n| > mn \quad \text{for infinitely many } n\big\}, \quad m \geqslant 1,$$

and set $\Omega \triangleq \cap_{m=1}^{\infty} \Omega_m$, then $\mathbb{P}(\Omega) = 1$. But we know that

$$\Omega \subseteq \big\{\overline{\lim_{n\to\infty}} \frac{|S_n|}{n} \geqslant m \quad \text{for all } m\big\} = \big\{\overline{\lim_{n\to\infty}} \frac{|S_n|}{n} = \infty\big\}.$$

Consequently, the result follows. □

Remark 4.1. It was a remarkable result of N. Etemadi (cf. N. Etemadi, An elementary proof of the strong law of large numbers, *Z. Wahrscheinlichkeitstheorie* 55 (1981) 119–122) that the assertion of Theorem 4.2 (i) remains true when the assumption of total independence is weakened to pairwise independence.

We conclude this topic by an interesting application of Theorem 4.2 to number theory. Recall that, every real number $x \in (0,1)$ admits an expansion

$$x = 0.x_1 x_2 \cdots x_n \cdots$$

in the usual decimal system where each $x_n = 0, 1, \cdots, 9$. Except for countably many points in $(0,1)$ (points of the form $x = m/10^n$ where m, n are positive integers) whose expansions terminate in finitely many steps, such expansion is unique and infinite.

Given $x \in (0,1)$ and $0 \leqslant k \leqslant 9$, let $\nu_n^{(k)}(x)$ be the number of digits among the first n digits of x that are equal to k. Apparently, $\frac{\nu_n^{(k)}(x)}{n}$ is the relative frequency of the digit k in the first n places. It is reasonable to believe that, for most of the points $x \in (0,1)$, this frequency should be close to $\frac{1}{10}$ as $n \to \infty$. Probabilistically, all the ten digits should occur equally likely in the decimal expansion of x if x is chosen randomly.

Definition 4.1. A real number $x \in (0,1)$ is said to be *simply normal* (in base 10) if

$$\lim_{n \to \infty} \frac{\nu_n^{(k)}(x)}{n} = \frac{1}{10} \quad \text{for every } k = 0, 1, \cdots, 9.$$

The following result, which was due to Borel, asserts that almost every real number in $(0,1)$ is simply normal.

Theorem 4.3. *Let X be a point in $(0,1)$ chosen uniformly at random (i.e. $X \overset{d}{=} U(0,1)$). Then with probability one, X is a simply normal number.*

Proof. We write X in its decimal expansion: $X = 0.X_1 X_2 \cdots X_n \cdots$. The crucial point is that, the sequence $\{X_n : n \geqslant 1\}$ of digits are independent and identically distributed, each following the distribution

$$\mathbb{P}(X_n = k) = \frac{1}{10}, \quad k = 0, 1, \cdots, 9. \tag{4.8}$$

We first show that (4.8) holds. To understand the event $\{X_n = k\}$, let $A_1, \cdots, A_m$ ($m = 10^{n-1}$) be the partition of $(0,1)$ into 10^{n-1} sub-intervals of equal length. For each j we further evenly sub-divide A_j into 10 sub-intervals and let $B_{j,k}$ be the k-th one of these sub-intervals. Then

$$\{X_n = k\} = \cup_{j=1}^m \{X \in B_{j,k}\},$$

and thus

$$\mathbb{P}(X_n = k) = \sum_{j=1}^m \mathbb{P}\big(X \in B_{j,k}\big) = 10^{n-1} \cdot \frac{1}{10^n} = \frac{1}{10}.$$

The intuition behind the above argument is best seen when one considers base 2 instead of 10 and draw a picture for the cases $n = 1, 2, 3$. If one understands the geometric intuition behind these digits $X_1, X_2, \cdots$, it is immediate that for any given $n \geqslant 1$ and $0 \leqslant k_1, \cdots k_n \leqslant 9$, the event

$$\{X_1 = k_1, X_2 = k_2, \cdots, X_n = k_n\}$$

simply means X falls in one particular sub-interval (depending on $k_1, \cdots, k_n$) in the even partition of $(0, 1)$ into 10^n sub-intervals. In particular,

$$\mathbb{P}\big(X_1 = k_1, \cdots, X_n = k_n\big) = \frac{1}{10^n} = \mathbb{P}(X_1 = k_1) \cdots \mathbb{P}(X_n = k_n).$$

This gives the independence among $X_1, \cdots, X_n$.

To prove the theorem, let $0 \leqslant k \leqslant 10$, and consider the Bernoulli sequence

$$Y_n = \begin{cases} 1, & X_n = k; \\ 0, & \text{otherwise.} \end{cases}, \quad n \geqslant 1.$$

Then $\nu_n^{(k)}(X) = Y_1 + \cdots + Y_n$. According to the strong law of large numbers (cf. Theorem 4.2), we conclude that with probability one,

$$\frac{\nu_n^{(k)}(X)}{n} \to \mathbb{E}[Y_1 = 1] = \mathbb{E}[X_1 = k] = \frac{1}{10} \tag{4.9}$$

as $n \to \infty$. In other words, with $\Omega_k \triangleq \big\{\frac{\nu_n^{(k)}(X)}{n} \to \frac{1}{10}\big\}$ we have $\mathbb{P}(\Omega_k) = 1$. The conclusion of the theorem follows by observing that

$$\mathbb{P}\big(\frac{\nu_n^{(k)}(X)}{n} \to \frac{1}{10} \text{ for every } k\big) = \mathbb{P}\big(\cap_{k=0}^9 \Omega_k\big) = 1.$$

□

Remark 4.2. Although Theorem 4.3 tells us that almost every real number in $(0, 1)$, it does not explicitly give us a single one! In fact, one can easily come up with numbers that are not simply normal. For instance, $x = 0.111\cdots$. However, it is more challenging to explicitly construct numbers that are simply normal. Can you give one?

5 Appendix: Proofs of Kronecker's lemma and the probabilistic Cauchy criterion for random series

Now we prove the two key lemmas that we have used in studying the strong law of large numbers.

The first one is Kronecker's lemma which is purely analytic.

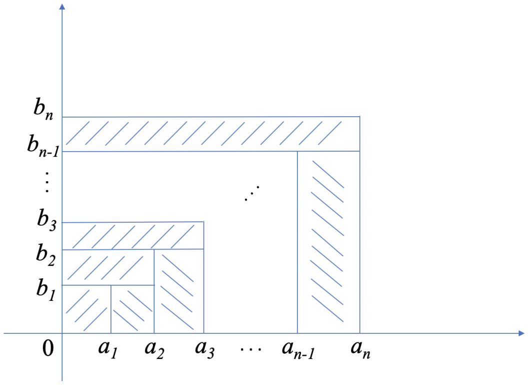
Proof of Lemma 4.1. Define $b_n \triangleq \sum_{j=1}^n \frac{x_j}{a_j}$ and set $a_0 = b_0 \triangleq 0$. Then $x_n = a_n(b_n - b_{n-1})$, and thus

$$\frac{1}{a_n}\sum_{j=1}^n x_j = \frac{1}{a_n}\sum_{j=1}^n a_j(b_j - b_{j-1}).$$

The crucial step is to write

$$\sum_{j=1}^n a_j(b_j - b_{j-1}) = a_n b_n - \sum_{j=0}^{n-1} b_j(a_{j+1} - a_j).$$

This is a discrete version of integration by parts, and the intuition behind this formula is best illustrated by the following figure.



Therefore,

$$\frac{1}{a_n}\sum_{j=1}^n x_j = b_n - \sum_{j=0}^{n-1} \frac{a_{j+1} - a_j}{a_n} \cdot b_j.$$

Since b_n is convergent by assumption, let us assume that $b_n \to b \in \mathbb{R}$. We claim that

$$\lim_{n\to\infty} \sum_{j=0}^{n-1} \frac{a_{j+1}-a_j}{a_n} \cdot b_j = b.$$

Indeed, given $\varepsilon > 0$, there exists $N \geqslant 1$ such that for all $n > N$, $|b_n - b| < \varepsilon$. It follows that, for $n > N$ we have

$$\begin{aligned} \Big|\sum_{j=0}^{n-1} \frac{(a_{j+1}-a_j)b_j}{a_n} - b\Big| &= \Big|\sum_{j=0}^{n-1} \frac{(a_{j+1}-a_j)(b_j-b)}{a_n}\Big| \\ &= \Big|\Big(\sum_{j\leqslant N} + \sum_{N<j\leqslant n-1}\Big)\frac{(a_{j+1}-a_j)(b_j-b)}{a_n}\Big| \\ &\leqslant \frac{a_{N+1}}{a_0}\cdot 2M + \varepsilon \cdot \sum_{N<j\leqslant n-1} \frac{a_{j+1}-a_j}{a_n} \\ &\leqslant \frac{2Ma_{N+1}}{a_n} + \varepsilon, \end{aligned}$$

where $M > 0$ is a constant such that $|b_n| \leqslant M$ for all n. By letting $n \to \infty$, we obtain

$$\overline{\lim_{n\to\infty}} \Big|\sum_{j=0}^{n-1} \frac{(a_{j+1}-a_j)b_j}{a_n} - b\Big| \leqslant \varepsilon.$$

The result follows since ε is arbitrary.

□

The second one is the probabilistic Cauchy criterion for the almost sure convergence of random series.

Proof of Proposition 4.1. The argument is a tedious unwinding of the statement in the usual Cauchy criterion. We first recall that, $\sum_n X_n$ is convergent if and only if for any $\varepsilon > 0$, there exists $n \geqslant 1$, such that whenever $i, j \geqslant n$ we have $|S_i - S_j| < \varepsilon$. Equivalently, $\sum_n X_n$ is divergent if and only if,

$$\exists \varepsilon > 0,\ \forall n \geqslant 1,\ \exists i, j \geqslant n,\ \text{s.t. } |S_i - S_j| \geqslant \varepsilon.$$

The statement $\exists i, j \geqslant n,\ \text{s.t. } |S_i - S_j| \geqslant \varepsilon$ can obviously be replaced by

$$\exists N \geqslant n,\ \text{s.t. } \max_{n\leqslant i,j\leqslant N} |S_i - S_j| \geqslant \varepsilon.$$

To summarise, we conclude that (assuming ε is rational)

$$\begin{aligned}
&\mathbb{P}\big(\sum_{n=1}^{\infty} X_n \text{ is divergent}\big) = 0 \\
&\iff \mathbb{P}\big(\cup_{\varepsilon>0} \cap_{n\geqslant 1} \cup_{N\geqslant n} \big\{\max_{n\leqslant i,j\leqslant N} |S_i - S_j| \geqslant \varepsilon\big\}\big) = 0 \\
&\iff \text{for all } \varepsilon > 0,\ \mathbb{P}\big(\cap_{n\geqslant 1} \cup_{N\geqslant n}\big\{\max_{n\leqslant i,j\leqslant N} |S_i - S_j| \geqslant \varepsilon\big\}\big) = 0 \\
&\iff \text{for all } \varepsilon > 0,\ \lim_{n\to\infty} \lim_{N\to\infty} \mathbb{P}\big(\big\{\max_{n\leqslant i,j\leqslant N} |S_i - S_j| \geqslant \varepsilon\big\}\big) = 0.
\end{aligned}$$

Now the result follows from the observation that

$$\big\{\max_{n\leqslant l\leqslant N} |S_l - S_n| \geqslant \varepsilon\big\} \subseteq \big\{\max_{n\leqslant i,j\leqslant N} |S_i - S_j| \geqslant \varepsilon\big\} \subseteq \big\{\max_{n\leqslant l\leqslant N} |S_l - S_n| \geqslant \frac{\varepsilon}{2}\big\}.$$

□

Topic 3: Characteristic Functions

In elementary probability theory, we have seen the notion of moment generating functions. There are many important reasons for introducing the moment generating function. For instance, it uniquely determines the law of a random variable, it can be used to study convergence in distribution and to compute moments effectively etc. One disadvantage of the moment generating function is that it is not always well defined (consider the Cauchy distribution as an example). Even when it is defined, it comes with its intrinsic domain of definition making the analysis cumbersome.

On the other hand, the characteristic function is always well defined for any random variable and achieves its greater power in probability theory. Analytic properties of the characteristic function is nicer and more robust than the moment generating function, although a price to pay is that one needs to work with complex numbers (mostly in the obvious manner). In this topic, we develop the basic theory of characteristic functions. The characteristic function is also a fundamental tool for proving the central limit theorem, as we will see in the next topic.

1 Definition of the characteristic function and its basic properties

The characteristic function will take complex values in general. To begin with, we first recall that, for $z = x + iy \in \mathbb{C}$, e^z is the complex number given by

$$e^z = e^x(\cos y + i \sin y).$$

In particular, we have the *Euler formula*:

$$e^{iy} = \cos y + i \sin y, \quad y \in \mathbb{R}. \tag{1.1}$$

Definition 1.1. Let X be a (real-valued) random variable. The *characteristic function* of X is the complex-valued function given by

$$f_X(t) \triangleq \mathbb{E}[e^{itX}], \quad t \in \mathbb{R}. \tag{1.2}$$

Remark 1.1. Using the Euler formula (1.1), equation (1.2) is interpreted as

$$f_X(t) = \mathbb{E}[\cos tX] + i\mathbb{E}[\sin tX].$$

But in many circumstances there is no need to treat the real and imaginary parts separately. It is more efficient to work with complex numbers.

The characteristic function is defined in terms of the distribution of X, and the underlying probability space is of no importance. In fact, we have

$$f_X(t) = \int_{-\infty}^{\infty} e^{itx} dF_X(x)$$

where $F_X(x)$ is the cumulative distribution function of X. Equivalently, we can simply define the *characteristic function of a probability measure* μ on $\mathbb{R}$ as

$$f_\mu(t) \triangleq \int_{\mathbb{R}} e^{itx} \mu(dx)$$

without referring to any random variables.

Remark 1.2. When X (or μ) admits a density function $\rho(x)$, the characteristic function is given by

$$f(t) = \int_{-\infty}^{\infty} e^{itx} \rho(x) dx,$$

which is commonly known as the *Fourier transform* of the function $\rho(x)$.

The characteristic function is defined for all $t \in \mathbb{R}$. Indeed, by the triangle inequality we have

$$|f_X(t)| \leqslant \mathbb{E}[|e^{itX}|] = \mathbb{E}[1] = 1.$$

It is obvious that $f_X(0) = 1$ and

$$\overline{f_X(t)} = \overline{\mathbb{E}[e^{itX}]} = \mathbb{E}[e^{-itX}] = f_X(-t).$$

Heuristically, the characteristic function is related to the moment generating function $M_X(t)$ by the simple relation

$$f_X(t) = M_X(it),$$

and like the moment generating function case it has the following elementary properties.

Proposition 1.1. *Let $a, b \in \mathbb{R}$ and X, Y be random variables.*
(i) $f_{aX+b}(t) = e^{itb} \cdot f_X(at)$ and $f_{-X}(t) = \overline{f_X(t)}$.
(ii) If X and Y are independent, then $f_{X+Y}(t) = f_X(t) \cdot f_Y(t)$.

Proof. (i) By definition,

$$f_{aX+b}(t) = \mathbb{E}[e^{it(aX+b)}] = \mathbb{E}[e^{itaX} \cdot e^{itb}] = e^{itb} \cdot f_X(at),$$

and

$$f_{-X}(t) = \mathbb{E}[e^{it\cdot(-X)}] = f_X(-t) = \overline{f_X(t)}.$$

(ii) Recall that, X and Y are independent if and only if

$$\mathbb{E}[\varphi(X)\psi(Y)] = \mathbb{E}[\varphi(X)] \cdot \mathbb{E}[\psi(Y)]$$

for any bounded Borel-measurable functions φ, ψ. Applying this to the bounded function $\varphi(x) = \psi(x) = e^{itx}$ (for fixed t), we get

$$f_{X+Y}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX} \cdot e^{itY}] = \mathbb{E}[e^{itX}] \cdot \mathbb{E}[e^{itY}] = f_X(t)f_Y(t).$$

□

The first important analytic property which the moment generating function may fail to have is the following.

Proposition 1.2. *$f_X(t)$ is uniformly continuous on $\mathbb{R}$.*

Proof. By definition, for any $t, h \in \mathbb{R}$ we have

$$\begin{aligned} f_X(t+h) - f_X(t) &= \int_{-\infty}^{\infty} (e^{i(t+h)x} - e^{itx}) dF_X(x) \\ &= \int_{-\infty}^{\infty} e^{itx}(e^{ihx} - 1) dF_X(x). \end{aligned}$$

According to the triangle inequality,

$$|f_X(t+h) - f_X(t)| \leqslant \int_{-\infty}^{\infty} |e^{ihx} - 1| dF_X(x). \tag{1.3}$$

Note that the right hand side is independent of t, and the integrand $|e^{ihx} - 1| \to 0$ as $h \to 0$ for each fixed x. By the dominated convergence theorem, the right hand side of (1.3) converges to zero as $h \to 0$. This implies the uniform continuity of $f_X(t)$. □

We do not list the explicit formulae for the characteristic functions of those special distributions we encounter in elementary probability theory. Some of them are straight forward while some can be quite challenging. Here we give one important example: the standard normal distribution.

Example 1.1. The characteristic function of $X \overset{d}{=} \mathcal{N}(0,1)$ is given by $f(t) = e^{-t^2/2}$. The following is an enlightening but semi-rigorous argument for verifying this fact. We start with the definition

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx.$$

By differentiation and integration by parts,

$$\begin{aligned} f'(t) &= \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{itx} e^{-x^2/2} dx \\ &= -\frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} d\big(e^{-x^2/2}\big) \\ &= \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} d(e^{itx}) \\ &= -\frac{t}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx \\ &= -tf(t). \end{aligned}$$

This is a first order ODE that can be solved uniquely with the obvious initial condition $f(0) = 1$. The solution gives $f(t) = e^{-t^2/2}$.

We conclude this section with an elementary inequality for the complex exponential which will be used frequently later on.

Lemma 1.1. *For any $a, b \in \mathbb{R}$, we have*

$$|e^{ib} - e^{ia}| \leqslant |b - a|. \tag{1.4}$$

Proof. Let us assume that $a < b$. A simple use of the triangle inequality yields:

$$|e^{ib} - e^{ia}| = \Big| \int_a^b ie^{it} dt \Big| \leqslant \int_a^b \big|ie^{it}\big| dt = \int_a^b 1 dt = b - a.$$

□

2 The uniqueness theorem and the inversion formula

One important reason for using the characteristic function is that it uniquely determines the distribution of the random variable. In addition, one can recover the distribution and learn many of its properties from the characteristic function in a fairly explicit way.

The central theorem for this part is the following *inversion formula*, which almost trivially implies the uniqueness property.

Theorem 2.1. *Let μ be a probability measure on $\mathbb{R}$ and let $f(t)$ be its characteristic function. Then for any real numbers $x_1 < x_2$, we have*

$$\mu((x_1,x_2)) + \frac{1}{2}\mu(\{x_1\}) + \frac{1}{2}\mu(\{x_2\}) = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t)dt. \quad (2.1)$$

Remark 2.1. The function $\frac{e^{-itx_1}-e^{-itx_2}}{it}$ at $t=0$ is defined in the limiting sense as $x_2 - x_1$. It should be pointed out that the right hand side of (2.1) cannot simply be understood as the integral

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t)dt,$$

which may not be well-defined unless $f(t)$ is integrable over $\mathbb{R}$.

We postpone the proof of Theorem 2.1 to the end of this section and discuss some of its consequences. First of all, it implies the following uniqueness result, which asserts that a probability measure is uniquely determined by its characteristic function.

Corollary 2.1. *Let μ_1 and μ_2 be two probability measures. If they have the same characteristic function, then $\mu_1 = \mu_2$.*

Proof. Let $D_i \triangleq \{x \in \mathbb{R}^1 : \mu_i(\{x\}) > 0\}$ denote the set of atoms (discontinuity points) for μ_i $(i = 1, 2)$ and $D \triangleq D_1 \cup D_2$. Since μ_1 and μ_2 have the same characteristic function, by the inversion formula (2.1) we have

$$\mu_1((x_1,x_2)) = \mu_2((x_1,x_2)), \quad \text{for all } x_1 < x_2 \text{ in } D^c. \quad (2.2)$$

On the other hand, D_1, D_2 are both countable and so is D. In particular, D^c is dense in $\mathbb{R}$. By a standard approximation argument, the relation (2.2) is enough to conclude that $\mu_1((a,b]) = \mu_2((a,b])$ for all real numbers $a < b$. This in turns implies $\mu_1 = \mu_2$ by Dynkin's π-λ theorem in measure theory. □

Another nice consequence of the uniqueness theorem is that many properties of the original distribution can be detected from its characteristic function. We give two examples of this kind. The first one only uses the uniqueness property while the second one requires application of the inversion formula.

Proposition 2.1. *Let X be a random variable with characteristic function $f_X(t)$. Then $f_X(t)$ is real-valued if and only if X and $-X$ have the same distribution.*

Proof. Note that $f_{-X}(t) = f_X(-t) = \overline{f_X(t)}$. Therefore, $f_X(t)$ is real-valued if and only if $f_{-X}(t) = f_X(t)$, which according to the uniqueness theorem is equivalent to saying that $X \overset{\mathrm{d}}{=} -X$. □

Proposition 2.2. *Let X be a random variable with cumulative distribution function $F(x)$ and characteristic function $f(t)$ respectively. Suppose that $f(t)$ is integrable over $(-\infty, \infty)$. Then $F(x)$ is continuously differentiable on $\mathbb{R}$, and its derivative (the probability density function) is given by the formula*

$$\rho(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx} f(t)dt. \tag{2.3}$$

Proof. We can actually express the inversion formula (2.1) as

$$\mathbb{P}(x_1 < X < x_2) + \frac{1}{2}\mathbb{P}(X = x_1) + \frac{1}{2}\mathbb{P}(X = x_2) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t)dt. \tag{2.4}$$

Indeed, according to (1.4) we have

$$\left|\frac{e^{-itx_1} - e^{-itx_2}}{it}\right| \leqslant |x_1 - x_2|.$$

It follows from the assumption that the function $t \mapsto \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t)$ is integrable over $(-\infty, \infty)$.

We first show that F is left continuous (and thus continuous). Let $x \in \mathbb{R}$ and $h > 0$. Using the relations

$$\begin{aligned}\mathbb{P}(x-h < X < x) &= F(x-) - F(x-h),\\ \mathbb{P}(X = x-h) &= F(x-h) - F((x-h)-),\\ \mathbb{P}(X = x) &= F(x) - F(x-),\end{aligned}$$

the inversion formula (2.4) applied to the case when $x_1 = x - h$ and $x_2 = x$ simplifies to

$$\frac{1}{2}(F(x) - F(x-h)) + \frac{1}{2}(F(x-) - F((x-h)-)) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-it(x-h)} - e^{-itx}}{it} f(t)dt. \tag{2.5}$$

Note that we always have

$$\lim_{h\downarrow 0} F((x-h)-) = F(x-) \quad \text{(why?)}.$$

In addition, since

$$\lim_{h\downarrow 0} \frac{e^{-it(x-h)} - e^{-itx}}{it} = 0$$

for every fixed t, by the dominated convergence theorem we know that the right hand side of (2.5) tends to zero as $h \downarrow 0$. Therefore, $F(x-h) \to F(x)$ as $h \downarrow 0$ which shows that F is left continuous at x.

Since F is continuous, by applying the inversion formula to the case when $x_1 = x$ and $x_2 = x + h$ and dividing it by h, we obtain

$$\frac{F(x+h) - F(x)}{h} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx} - e^{-it(x+h)}}{ith} f(t)dt.$$

By the dominated convergence theorem, the right hand side tends to $\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx} f(t)dt$ as $h \to 0$. Therefore, F is differentiable at x with derivative

$$F'(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx} f(t)dt.$$

The continuity of $F'(x)$ follows from the continuity of the integral $x \mapsto \int_{-\infty}^{\infty} e^{-itx} f(t)dt$ which is again a simple consequence of the dominated convergence theorem. □

Proof of the inversion formula (2.1)

The proof of (2.1) (as well as many other analytic aspects of the characteristic function) relies on the following basic identity (the *Dirichlet integral*):

$$\int_0^{\infty} \frac{\sin u}{u} du = \frac{\pi}{2}. \tag{2.6}$$

Note that this integral must be understood as an improper integral $\lim_{R\to\infty} \int_0^R \frac{\sin u}{u} du$ with $\frac{\sin 0}{0} \triangleq 1$.

To prove the inversion formula we begin with its right hand side. We fix $x_1 < x_2$ throughout the discussion. By the definition of the characteristic function, for each $T > 0$ we have

$$\begin{aligned}\int_{-T}^{T}\frac{e^{-itx_1}-e^{-itx_2}}{it}f(t)dt &= \int_{-T}^{T}\frac{e^{-itx_1}-e^{-itx_2}}{it}\big(\int_{-\infty}^{\infty}e^{itx}\mu(dx)\big)dt \\ &= \int_{-\infty}^{\infty}\big(\int_{-T}^{T}\frac{e^{-it(x_1-x)}-e^{-it(x_2-x)}}{it}dt\big)\mu(dx). \qquad (2.7)\end{aligned}$$

Here we have used Fubini's theorem to change the order of integration. This is legal since

$$\left|\frac{e^{-itx_1}-e^{-itx_2}}{it}\cdot e^{itx}\right| \leqslant |x_2 - x_1|$$

which is integrable over $[-T, T] \times \mathbb{R}$ with respect to the product measure $dt \times \mu$. The reader who is not familiar with measure theory can take the exchange of double integral as granted.

Next, we denote the integrand in the μ-integral on the right hand side of (2.7) as

$$I_T(x; x_1, x_2) \triangleq \int_{-T}^{T}\frac{e^{-it(x_1-x)}-e^{-it(x_2-x)}}{it}dt.$$

By writing out the real and imaginary parts we obtain

$$\begin{aligned}&I_T(x; x_1, x_2) \\ &= \int_{-T}^{T}\frac{\big(\cos t(x_1-x)-\cos t(x_2-x)\big)+i\big(\sin t(x_2-x)-\sin t(x_1-x)\big)}{it}dt \\ &= 2\big(\int_0^T\frac{\sin t(x_2-x)}{t}dt-\int_0^T\frac{\sin t(x_1-x)}{t}dt\big),\end{aligned}$$

where the cosine part is gone since the cosine function is even. We apply a change of variables and discuss according to different scenarios of x to get

$$I_T(x; x_1, x_2) = \begin{cases} 2\big(\int_0^{T(x_2-x)}\frac{\sin u}{u}du - \int_0^{T(x_1-x)}\frac{\sin u}{u}du\big), & x < x_1; \\ 2\int_0^{T(x_2-x)}\frac{\sin u}{u}du, & x = x_1; \\ 2\big(\int_0^{T(x_2-x)}\frac{\sin u}{u}du + \int_0^{T(x-x_1)}\frac{\sin u}{u}du\big), & x_1 < x < x_2; \\ 2\int_0^{T(x-x_1)}\frac{\sin u}{u}du, & x = x_2; \\ 2\big(-\int_0^{T(x-x_2)}\frac{\sin u}{u}du + \int_0^{T(x-x_1)}\frac{\sin u}{u}du\big), & x > x_2. \end{cases} \qquad (2.8)$$

By sending $T \to \infty$ and using the Dirichlet integral (2.6), we obtain

$$\lim_{T\to\infty} I_T(x;x_1,x_2) = \begin{cases} 0, & x < x_1; \\ \pi, & x = x_1; \\ 2\pi, & x_1 < x < x_2; \\ \pi, & x = x_2; \\ 0, & x > x_2. \end{cases} \tag{2.9}$$

Note that we have already expressed the right hand side of the inversion formula (2.1) as

$$\lim_{T\to\infty} \frac{1}{2\pi}\int_{-\infty}^{\infty} I_T(x;x_1,x_2)\mu(dx).$$

The equation (2.9) urges us to take limit under the integral sign. This is indeed legal as a result of the following elementary fact.

Lemma 2.1. *We have*

$$0 \leqslant \int_0^y \frac{\sin u}{u}du \leqslant \int_0^\pi \frac{\sin u}{u}du \quad \text{for all } y \geqslant 0.$$

In view of the expression (2.8) of $I_T(x;x_1,x_2)$, Lemma 2.1 tells us that

$$|I_T(x;x_1,x_2)| \leqslant 4\int_0^\pi \frac{\sin u}{u}du < \infty \quad \text{for all } x \text{ and } T.$$

According to the dominated convergence theorem and (2.9), we obtain that

$$\lim_{T\to\infty} \frac{1}{2\pi}\int_{-\infty}^{\infty} I_T(x;x_1,x_2)\mu(dx) = \frac{1}{2}\mu(\{x_1\}) + \mu((x_1,x_2)) + \frac{1}{2}\mu(\{x_2\})$$

which concludes the desired inversion formula.

As the last piece of the puzzle, it remains to prove Lemma 2.1 which is an interesting calculus exercise.

Proof of Lemma 2.1. We first show that $D(y) \triangleq \int_0^y \frac{\sin u}{u}du$ is non-negative for all $y \geqslant 0$. This is obvious when $y \in [0,\pi]$. When $y \in [(2k-1)\pi,(2k+1)\pi]$ with any

$k \geqslant 1$, we have

$$\begin{aligned}\int_0^y \frac{\sin u}{u} du \geqslant \int_0^{2k\pi} \frac{\sin u}{u} du &= \sum_{l=1}^{k} \int_{2(l-1)\pi}^{2l\pi} \frac{\sin u}{u} du \\ &= \sum_{l=1}^{k} \big(\int_{(2l-2)\pi}^{(2l-1)\pi} \frac{\sin u}{u} du + \int_{(2l-1)\pi}^{2l\pi} \frac{\sin u}{u} du \big) \\ &= \sum_{l=1}^{k} \int_{(2l-2)\pi}^{(2l-1)} (\sin u) \cdot \big(\frac{1}{u} - \frac{1}{u+\pi} \big) du \\ &\geqslant 0,\end{aligned}$$

where to reach the last equality we have applied a change of variables to the integral $\int_{(2l-1)\pi}^{2l\pi} \frac{\sin u}{u} du$.

Next we show that $D(y)$ is maximised at $y = \pi$. Due to the sign pattern of $\sin u$, it is enough to show that (why?)

$$\int_{\pi}^{(2k+1)\pi} \frac{\sin u}{u} du \leqslant 0 \quad \text{for all } k \geqslant 0.$$

This can be proved in a similar way as the positivity part:

$$\begin{aligned}\int_{\pi}^{(2k+1)\pi} \frac{\sin u}{u} du &= \sum_{l=1}^{k} \big(\int_{(2l-1)\pi}^{2l\pi} \frac{\sin u}{u} du + \int_{2l\pi}^{(2l+1)\pi} \frac{\sin u}{u} du \big) \\ &= \sum_{l=1}^{k} \int_{(2l-1)\pi}^{2l\pi} (\sin u) \cdot \big(\frac{1}{u} - \frac{1}{u+\pi} \big) du \\ &\leqslant 0.\end{aligned}$$

□

Remark 2.2. The proof of the inversion formula (2.1) we give here is not very enlightening, since we have started from the right hand side of the formula pretending that it was known in advance. In the context of Fourier transform, in fact it took mathematician a long journey to understand why the simple inversion formula

$$\rho(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} f(t) dt$$

recovers the original function $\rho(x)$ from its Fourier transform $f(t)$. One needs to go into Fourier analysis to understand in a deeper way how the inversion formula arises naturally.

3 Lévy-Cramér's continuity theorem

The most important and useful result of the characteristic function is that convergence in distribution for random variables is equivalent to pointwise convergence for their characteristic functions. This is the content of *Lévy-Cramér's continuity theorem*, which is sometimes referred to as the *convergence theorem.* As we will see, it provides a useful tool for proving central limit theorems.

We start with the easier part of the theorem.

Theorem 3.1. *Let μ_n $(n \geqslant 1)$ and μ be probability measures on $\mathbb{R}$, with characteristic functions f_n $(n \geqslant 1)$ and f respectively. If μ_n converges weakly to μ, then f_n converges to f uniformly on every finite interval of $\mathbb{R}$.*

Proof. For each fixed t, the function $x \mapsto e^{itx}$ is a bounded continuous function. Therefore, the convergence of $f_n(t)$ to $f(t)$ (for fixed t) is a trivial consequence of the weak convergence of μ_n to μ. Here there is no difficulty with e^{itx} being complex-valued: just work with the real and imaginary parts separately.

The uniformity assertion requires more effort than the simple pointwise convergence. We first claim that, under the assumption, the family of functions $\{f_n : n \geqslant 1\}$ is uniformly equicontinuous on $\mathbb{R}$. Recall that *uniform equicontinuity* means, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|f_n(t) - f_n(s)| < \varepsilon$$

for all $n \geqslant 1$ and all s, t with $|t - s| < \delta$. To prove the uniform equicontinuity of $\{f_n\}$, first note that the family of probability measures $\{\mu_n : n \geqslant 1\}$ is tight, as a consequence of weak convergence. In particular, for given $\varepsilon > 0$, there exists $A = A(\varepsilon) > 0$ such that

$$\mu_n([-A, A]^c) < \varepsilon \quad \text{for all } n.$$

Next, for any real numbers t and h and $n \geqslant 1$, we have

$$\begin{aligned} |f_n(t+h) - f_n(t)| &= \Big| \int_{-\infty}^{\infty} e^{i(t+h)x} \mu_n(dx) - \int_{-\infty}^{\infty} e^{itx} \mu_n(dx) \Big| \\ &\leqslant \int_{-\infty}^{\infty} |e^{ihx} - 1| \mu_n(dx) \\ &= \int_{\{x:|x| \leqslant A\}} |hx| \mu_n(dx) + \int_{\{x:|x|>A\}} 2\mu_n(dx) \\ &\leqslant |h|A + 2\mu_n([-A, A]^c) \\ &< |h|A + 2\varepsilon. \end{aligned}$$

When $|h|$ is small enough (in a way that is independent of t), the right hand side can be made less than 3ε. This proves the uniform equicontinuity property.

Now we establish the desired uniform convergence property. Let $I = [a,b]$ be an arbitrary finite interval. First of all, for an given $\varepsilon > 0$, by uniform equicontinuity there exists $\delta > 0$, such that whenever $|t-s| < \delta$ we have

$$|f_n(t) - f_n(s)| < \varepsilon.$$

We may also assume that for the same δ we have $|f(t) - f(s)| < \varepsilon$, since f is uniformly continuous (cf. Proposition 1.2). Next, we fix a finite partition

$$\mathcal{P} : a = t_0 < t_1 < \cdots < t_{r-1} < t_r = b$$

of $[a,b]$ such that $|t_i - t_{i-1}| < \delta$ for all $1 \leqslant i \leqslant r$. Since at each partition point t_i we have the pointwise convergence $f_n(t_i) \to f(t_i)$ (as $n \to \infty$) and there are finitely many of them, one can find $N \geqslant 1$, such that

$$|f_n(t_i) - f(t_i)| < \varepsilon \text{ for all } n > N \text{ and } 0 \leqslant i \leqslant r.$$

It follows that for each $n > N$ and $t \in [a,b]$, with $t_i \in \mathcal{P}$ being the partition point such that $t \in [t_i, t_{i+1}]$, we have

$$\begin{aligned} |f_n(t) - f(t)| &\leqslant |f_n(t) - f_n(t_i)| + |f_n(t_i) - f(t_i)| + |f(t_i) - f(t)| \\ &< \varepsilon + \varepsilon + \varepsilon \\ &= 3\varepsilon. \end{aligned}$$

This concludes the uniform convergence of f_n to f on $[a,b]$. □

The harder (and more useful) part of the theorem is the other direction asserting that weak convergence can be established through pointwise convergence of the characteristic function.

Theorem 3.2. *Let $\{\mu_n : n \geqslant 1\}$ be a sequence of probability measures on $\mathbb{R}$ with characteristic functions $\{f_n : n \geqslant 1\}$ respectively. Suppose that:*

(i) $f_n(t)$ converges pointwisely to some limiting function $f(t)$;
(ii) $f(t)$ is continuous at $t = 0$.

Then there exists a probability measure μ, such that μ_n converges weakly to μ. In addition, f is the characteristic function of μ.

We postpone its proof to the end of this section. There are two important remarks concerning the assumptions in the above two theorems. On the one hand, in Theorem 3.1, it is crucial to assume weak convergence of μ_n. As illustrated by the following example, f_n may fail to converge if only vague convergence is assumed.

Example 3.1. Let $\mu_n = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_n$ be the two-point distribution at 0 and n with equal probabilities. It is a simple exercise that μ_n converges vaguely to $\frac{1}{2}\delta_0$ (which is not a probability measure). The characteristic function of μ_n is given by $f_n(t) = \frac{1}{2} + \frac{1}{2}e^{int}$, which fails to converge at any $t \notin 2\pi\mathbb{Z}$.

On the other hand, the following example illustrates that, in Theorem 3.2, the continuity assumption of the limiting function at $t = 0$ is crucial. As we will see in the proof, this assumption guarantees the tightness property which is crucial for expecting weak convergence.

Example 3.2. Let μ_n be the normal distribution with mean zero and variance n. Then

$$f_n(t) = e^{-\frac{1}{2}nt^2} \overset{n\to\infty}{\longrightarrow} f(t) = \begin{cases} 0, & t \neq 0; \\ 1, & t = 0. \end{cases}$$

Note that although f_n converges pointwisely, the limiting function is not continuous at $t = 0$. The sequence μ_n converges vaguely to the zero measure and thus fails to be weakly convergent.

Combining the two theorems, we obtain the following elegant but weaker formulation.

Corollary 3.1. *Let μ_n $(n \geqslant 1)$ and μ be probability measures on $\mathbb{R}$, with characteristic functions f_n $(n \geqslant 1)$ and f respectively. Then μ_n converges weakly to μ if and only if f_n converges pointwisely to f.*

Proof. Necessity is trivial. For sufficiency, since we know that f is a characteristic function it must be continuous at $t = 0$. In particular, the two conditions of Theorem 3.2 are both verified. Therefore, there exists a probability measure ν such that μ_n converges weakly to ν and f is the characteristic function of ν. Since f is assumed to be the characteristic function of μ, by the uniqueness theorem we have $\nu = \mu$, showing that μ_n converges weakly to μ. □

Proof of Theorem 3.2

Before proving Theorem 3.2, we first derive a general estimate for the characteristic function which is also of independent interest.

Lemma 3.1. *Let μ be a probability measure on $\mathbb{R}$ with characteristic function f. Then for any $\delta > 0$, we have*

$$\mu([-2\delta^{-1}, 2\delta^{-1}]) \geqslant \frac{1}{\delta}\Big|\int_{-\delta}^{\delta} f(t)dt\Big| - 1. \tag{3.1}$$

Proof. By definition, we have

$$\begin{aligned}\int_{-\delta}^{\delta} f(t)dt &= \int_{-\delta}^{\delta}\int_{-\infty}^{\infty} e^{itx}\mu(dx)dt \\ &= \int_{-\infty}^{\infty}\mu(dx)\int_{-\delta}^{\delta}(\cos tx + i\sin tx)dt \\ &= \int_{-\infty}^{\infty}\frac{2\sin\delta x}{x}\mu(dx).\end{aligned}$$

Since $|\frac{\sin x}{x}| \leqslant 1$, it follows that

$$\begin{aligned}\frac{1}{2\delta}\Big|\int_{-\delta}^{\delta} f(t)dt\Big| &= \Big|\int_{-\infty}^{\infty}\frac{\sin\delta x}{\delta x}\mu(dx)\Big| \\ &\leqslant \int_{\{x:|\delta x|\leqslant 2\}}\mu(dx) + \int_{\{x:|\delta x|>2\}}\frac{1}{|\delta x|}\mu(dx) \\ &\leqslant \mu([-2\delta^{-1}, 2\delta^{-1}]) + \frac{1}{2}\mu([-2\delta^{-1}, 2\delta^{-1}]^c) \\ &= \frac{1}{2} + \frac{1}{2}\mu([-2\delta^{-1}, 2\delta^{-1}]).\end{aligned}$$

Reorganising the terms gives the desired inequality. □

The significance of Lemma 3.1 lies in telling us that the continuity of $f(t)$ at $t = 0$ controls the speed that μ loses its mass at infinity. Indeed, a re-arrangement of (3.1) yields

$$\begin{aligned}\mu([-2\delta^{-1},2\delta^{-1}]^c) &\leqslant 2-\frac{1}{\delta}\Big|\int_{-\delta}^{\delta}f(t)dt\Big| \\ &= \frac{\big|\int_{-\delta}^{\delta}f(0)dt\big|-\big|\int_{-\delta}^{\delta}f(t)dt\big|}{\delta} \quad \text{(since } f(0)=1\text{)} \\ &\leqslant \frac{1}{\delta}\int_{-\delta}^{\delta}\big|f(t)-f(0)\big|dt. \end{aligned} \tag{3.2}$$

This inequality shows that the speed that $\mu([-2\delta^{-1},2\delta^{-1}]^c)\to 0$ as $\delta\downarrow 0$ is controlled by the speed of convergence to zero for the right hand side, which is in turn controlled by the (modulus of) continuity of $f(t)$ at $t=0$.

Remark 3.1. In the language of analysis, the study of the precise relationship between the tail behaviour of a function and the behaviour near the origin of its Fourier transform is the content of Tauberian theory.

The key step for proving Theorem 3.2 is to show that the family $\{\mu_n : n\geqslant 1\}$ is tight, which guarantees the existence of a weak convergent subsequence, and the remaining steps are easy. The assumptions in the theorem play an essential role for establishing tightness through the estimate (3.2).

Proof of Theorem 3.2. Step one: tightness of $\{\mu_n\}$. According to (3.2), for every $\delta>0$ we have

$$\begin{aligned}\mu_n([-2\delta^{-1},2\delta^{-1}]^c) &\leqslant \frac{1}{\delta}\int_{-\delta}^{\delta}\big|f_n(t)-f_n(0)\big|dt \\ &\leqslant \frac{1}{\delta}\int_{-\delta}^{\delta}\big|f_n(t)-f(t)\big|dt+\frac{1}{\delta}\int_{-\delta}^{\delta}\big|f(t)-f(0)\big|dt\end{aligned}$$

where we have also used the obvious fact that $f_n(0)=f(0)=1$. Now given $\varepsilon>0$, by the continuity assumption for $f(t)$ at $t=0$, there exists $\delta=\delta(\varepsilon)>0$ such that

$$\frac{1}{\delta}\int_{-\delta}^{\delta}\big|f(t)-f(0)\big|dt<\varepsilon.$$

Since $f_n(t)\to f(t)$ for every t and $|f_n(t)-f(t)|\leqslant 2$, by the dominated convergence theorem (for such fixed δ) we know that

$$\lim_{n\to\infty}\int_{-\delta}^{\delta}|f_n(t)-f(t)|dt=0.$$

In particular, there exists $N = N(\varepsilon) \geqslant 1$, such that

$$\frac{1}{\delta}\int_{-\delta}^{\delta}\big|f_n(t) - f(t)\big|dt < \varepsilon \quad \text{for all } n > N.$$

It follows that

$$\mu_n([-2\delta^{-1}, 2\delta^{-1}]^c) < 2\varepsilon \quad \text{for all } n > N. \tag{3.3}$$

By further shrinking δ, we can ensure that (3.3) holds for $\mu_1, \cdots, \mu_N$ as well and thus for all n. This gives the tightness property.

Step two: there is precisely one weak limit point of μ_n. Since the family $\{\mu_n\}$ is tight, we know that there exists a subsequence μ_{n_k} converging weakly to some probability measure μ. Let μ_{m_j} another subsequence which converges weakly to another probability measure ν. According to Theorem 3.1, we have

$$f_{n_k}(t) \to f_\mu(t),\ f_{m_j}(t) \to f_\nu(t)$$

for the corresponding characteristic functions. But from assumption we know that $f_n(t)$ converges pointwisely. Therefore, we conclude that $f_\nu(t) = f_\mu(t)$, which implies $\nu = \mu$ by the uniqueness theorem. Therefore, the sequence has one and only one weak limit point μ.

Step three: μ_n *converges weakly to* μ. This is a very natural consequence of Step Two. Let $f \in C_b(\mathbb{R})$ and denote $c_n \triangleq \int_{-\infty}^{\infty} f(x)\mu_n(dx)$. Suppose that c is a limit point of c_n, say along a subsequence c_{m_j}. By tightness, there is a further weakly convergent subsequence $\mu_{m_{j_l}}$, whose weak limit has to be μ by Step Two. Therefore,

$$c_{m_{j_l}} = \int_{-\infty}^{\infty} f(x)\mu_{m_{j_l}}(dx) \to \int_{-\infty}^{\infty} f(x)\mu(dx)$$

as $l \to \infty$. This shows that $c = \int_{-\infty}^{\infty} f(x)\mu(dx)$. In other words, c_n has precisely one limit point c. Therefore,

$$c_n = \int_{-\infty}^{\infty} f(x)\mu_n(dx) \to c = \int_{-\infty}^{\infty} f(x)\mu(dx)$$

as $n \to \infty$. This proves the weak convergence of μ_n to μ.

□

4 Some applications of the characteristic function

We discuss a few simple applications of the characteristic function. The more powerful applications to central limit theorems will be discussed in the next topic.

If we take the k-th derivative of the expression $f(t) = \mathbb{E}[e^{itX}]$ at $t = 0$, we obtain (formally) that $f^{(k)}(0) = i^k\mathbb{E}[X^k]$. This tells us that we can use the characteristic function to compute moments. The following result makes this fact precise.

Theorem 4.1. *Suppose that the random variable X has absolute moments up to order n. Then its characteristic function $f(t)$ has bounded continuous derivatives up to order n, given by*

$$f^{(k)}(t) = i^k\mathbb{E}[X^k e^{itX}], \quad 1 \leqslant k \leqslant n.$$

In particular, $\mathbb{E}[X^k] = \frac{f^{(k)}(0)}{i^k}$ for each $1 \leqslant k \leqslant n$.

Proof. We only consider the case when $n = 1$, as the general case will follow by induction. First of all, for any real numbers t and h we have

$$\frac{f(t+h) - f(t)}{h} = \mathbb{E}\big[\frac{e^{i(t+h)X} - e^{itX}}{h}\big].$$

Note that

$$\frac{e^{i(t+h)X} - e^{itX}}{h} \to iXe^{itX} \quad \text{as } h \to 0,$$

and $\big|\frac{e^{i(t+h)X} - e^{itX}}{h}\big| \leqslant |X|$ which is integrable by assumption. According to the dominated convergence theorem, we conclude that

$$\frac{f(t+h) - f(t)}{h} \to \mathbb{E}[iXe^{itX}] \quad \text{as } h \to 0,$$

which is the derivative of $f(t)$. Its continuity is an obvious consequence of the dominated convergence theorem again. □

The following result is a direct corollary of Theorem 4.1 and the Taylor approximation theorem in calculus.

Corollary 4.1. *Under the same assumption as in Theorem 4.1, we have*

$$f(t) = \sum_{k=0}^{n} \frac{i^k\mathbb{E}[X^k]}{k!}t^k + o(|t|^n),$$

where $o(|t|^n)$ denotes a function such that $\frac{o(|t|^n)}{|t|^n} \to 0$ as $t \to 0$.

As another application, we reproduce the weak law of large numbers in the i.i.d. case, using the theory of characteristic functions.

Theorem 4.2. *Let $\{X_n : n \geqslant 1\}$ be a sequence of i.i.d. random variables with finite mean $m \triangleq \mathbb{E}[X_1]$. Then*

$$\frac{X_1 + \cdots + X_n}{n} \to m \quad \text{in prob.}$$

as $n \to \infty$.

Proof. Since the asserted limit is a deterministic constant, it is equivalent to proving convergence in distribution. Let $f(t)$ be the characteristic function of X_1 (and thus of X_n for every n). Then, with $S_n \triangleq X_1 + \cdots + X_n$ we have

$$f_{S_n/n}(t) = \mathbb{E}\big[e^{it(X_1+\cdots+X_n)/n}\big] = \big(f\big(\frac{t}{n}\big)\big)^n.$$

Since X_1 has finite mean, by Corollary 4.1 we can write

$$f_{S_n/n}(t) = \big(1 + \frac{imt}{n} + o(1/n)\big)^n = (1 + q_n)^{\frac{1}{q_n} \cdot nq_n},$$

where $q_n \triangleq \frac{imt}{n} + o(1/n)$. Note that $q_n \to 0$ and $nq_n \to imt$ as $n \to \infty$. Therefore, $(1 + q_n)^{1/q_n} \to e$ and

$$f_{S_n/n}(t) \to e^{imt}$$

as $n \to \infty$. Since e^{imt} is the characteristic function of the constant random variable $X = m$, we conclude from Lévy-Cramér's continuity theorem that $\frac{S_n}{n}$ converges to m in distribution. □

5 Pólya's criterion for characteristic functions

In this section, we consider the following natural question. Suppose that $f(t)$ is a given function. How can we know if it is the characteristic function of some random variable/probability measure? There is a general theorem, due to Bochner, which provides a necessary and sufficient condition for a function to be a characteristic function. Bochner's criterion is not very easy to verify in practice. On the other hand, there is a rather useful criterion (a sufficient condition) due to Pólya. In many situations, Pólya's criterion can be checked explicitly and be used to construct a rich class of characteristic functions. In what follows, we discuss this elegant and important result of Pólya.

Theorem 5.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a real valued function which satisfies the following properties:*

(i) $f(0) = 1$ and $f(t) = f(-t)$ for all t;
(ii) $f(t)$ is decreasing and convex on $(0, \infty)$;
(iii) $f(t)$ is continuous at the origin, and $\lim_{t\to\infty} f(t) = 0$.

Then $f(t)$ is the characteristic function of some random variable/probability measure.

The generic shape of functions that satisfy Pólya's criterion is sketched in the figure below.

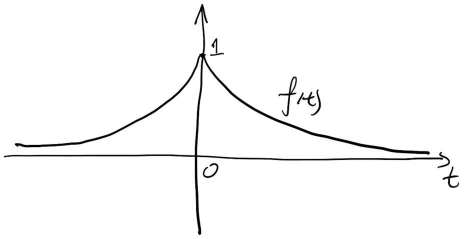


Figure 1: Pólya's Criterion

Remark 5.1. Note that the conditions imply that $f(t)$ is non-negative. The condition that $f(t)$ is continuous at $t = 0$ is important. Indeed, the function

$$f(t) \triangleq \begin{cases} 1, & t = 0; \\ 0, & t \neq 0, \end{cases}$$

satisfies all conditions of the theorem except for continuity at the origin. This function is apparently not a characteristic function. The condition that $\lim_{t\to\infty} f(t) = 0$ is not important, and can be replaced by $\lim_{t\to\infty} f(t) = c > 0$ for some $c \in (0, 1)$. Indeed, in the latter case, we consider

$$g(t) \triangleq \frac{f(t) - c}{1 - c}.$$

Then $g(t)$ satisfies the conditions of the theorem and is thus a characteristic function. But we can write

$$f(t) = (1 - c) \cdot g(t) + c \cdot 1,$$

which is a convex combination of two characteristic functions ($g(t)$ and 1). Therefore, $f(t)$ is also a characteristic function.

Before proving Theorem 5.1, we first look at a simple but enlightening example.

Example 5.1. The simplest example that satisfies Pólya's criterion is the following function:

$$f(t) = (1-|t|)^+ \triangleq \begin{cases} 1-|t|, & |t| \leqslant 1; \\ 0, & \text{otherwise.} \end{cases}$$

For this example, there is no need to use Theorem 5.1 to see that it is a characteristic function. By evaluating the inversion formula (2.3) explicitly, one easily finds that $f(t)$ is the characteristic function of the distribution whose probability density function is given by

$$\rho(x) = \frac{1-\cos x}{\pi x^2}, \quad x \in \mathbb{R}.$$

Example 5.2. Another interesting example that satisfies the conditions of the theorem is the function $f_\alpha(t) \triangleq e^{-|t|^\alpha}$ ($\alpha \in (0,1]$). In particular, this covers the case of the Cauchy distribution (when $\alpha = 1$). When $\alpha \in (1,2)$, f_α is still a characteristic function, however, Theorem 5.1 does not apply since f_α is no longer convex. The treatment of this case will be given in the next topic (by a different approach) when we study the central limit theorem.

The starting point for proving Theorem 5.1 is the following observation: if f_1, f_2 are characteristic functions and $\lambda \in (0,1)$, then

$$\lambda_1 f_1 + \lambda_2 f_2$$

is also a characteristic function (cf. Week 6 Practice Problem 2-ii). This property is easily generalised to the case of more than two members: if $f_1, \cdots, f_n$ are characteristic functions and $\lambda_1, \cdots, \lambda_n$ are positive numbers such that $\lambda_1 + \cdots + \lambda_n = 1$, then

$$\lambda_1 f_1 + \cdots + \lambda_n f_n$$

is also a characteristic function. Without surprise, this fact can be further generalised to the case of the convex combination of a continuous family of characteristic functions. To be precise, let ν be a probability measure on $(0,\infty)$, and for each $r \in (0,\infty)$ let $t \mapsto f_r(t)$ be a characteristic function. Under some measurability property of $r \mapsto f_r$, one can show that the function

$$t \mapsto \int_{(0,\infty)} f_r(t)\nu(dr)$$

is also a characteristic function. The assumption that ν is a probability measure on $(0,\infty)$ guarantees that this is a convex combination of the family $\{f_r : r \in (0,\infty)\}$ of characteristic functions, weighted by the measure ν.

The key idea of the proof of Theorem 5.1 is to express $f(t)$ as a convex combination of a (continuous) family of characteristic functions, more precisely, as

$$f(t) = \int_{(0,\infty)} f_r(t)\nu(dr) \tag{5.1}$$

where f_r is some classical characteristic function (for each $r > 0$) and ν is a probability measure on $(0,\infty)$. The above discussion then shows that f must also be a characteristic function. We now carry out this scheme.

Proof of Theorem 5.1. Step one. We first collect some standard properties arising from the convexity of $f(t)$ as well as the other assumptions in the theorem. For each $t > 0$, we define the *right derivative* of $f(t)$ as

$$f'_+(t) \triangleq \lim_{h\downarrow 0} \frac{f(t+h) - f(t)}{h}.$$

(i) f'_+ is well defined and we have $-\infty < f'_+(t) \leqslant 0$ for every $t > 0$.
(ii) f'_+ is increasing and right continuous on $(0,\infty)$.
(iii) For each given $t > 0$, f is Lipschitz (and absolutely continuous) on $[t,\infty)$.
(iv) Since $\lim_{t\to\infty} f(t) = 0$, we have

$$\lim_{t\to\infty} f'_+(t) = 0.$$

Step two. Since f'_+ is increasing and right continuous, we can define a measure μ on $(0,\infty)$ by using the relation

$$\mu((a,b]) \triangleq f'_+(b) - f'_+(a), \quad 0 < a < b.$$

The Carathéodory extension theorem in measure theory ensures the construction of μ. Using μ and the density function $\rho(r) = r$, we introduce another measure ν on $(0,\infty)$ by

$$\nu(dr) \triangleq r\mu(dr).$$

The definition of ν is understood as $\frac{d\nu}{d\mu}(r) = r$ or

$$\nu(A) \triangleq \int_A r\nu(dr),\ A \in \mathcal{B}((0,\infty)).$$

Step three. We are going to express $f(t)$ as an integral with respect to ν in the form (5.1). To this end, first note that, by the definition of μ, we have

$$\begin{aligned} -f'_+(s) &= 0 - f'_+(s) = f'_+(\infty) - f'_+(s) \\ &= \int_s^\infty \mu(dr) = \int_s^\infty r^{-1}\nu(dr) \end{aligned}$$

for every $s > 0$. In addition, by the fundamental theorem of calculus, we have

$$\begin{aligned} f(t) &= -(f(\infty) - f(t)) = -\int_t^\infty f'_+(s)ds \\ &= \int_t^\infty \int_s^\infty r^{-1}\nu(dr)ds \end{aligned}$$

for every $t > 0$. Using Fubini's theorem, we obtain that

$$\begin{aligned} f(t) &= \int_t^\infty \big(\int_t^r ds\big)r^{-1}\nu(dr) = \int_t^\infty \big(1 - \frac{t}{r}\big)\nu(dr) \\ &= \int_{(0,\infty)} \big(1 - \frac{t}{r}\big)^+\nu(dr), \quad \text{for all } t > 0. \end{aligned}$$

Since $f(t)$ is an even function, we arrive at

$$f(t) = \int_{(0,\infty)} \big(1 - \frac{|t|}{r}\big)^+\nu(dr), \quad \text{for all } t \in \mathbb{R}\backslash\{0\}. \tag{5.2}$$

Step four. For each given $r > 0$, the function

$$f_r(t) \triangleq \big(1 - \frac{|t|}{r}\big)^+, \quad t \in \mathbb{R}$$

is a characteristic function. This is a direct consequence of Example 5.1 and the scaling property of characteristic functions.

Step five. It remains to show that ν is a probability measure on $(0,\infty)$, which then recognises (5.2) as a convex combination of the family $\{f_r : r > 0\}$ of characteristic functions. To see this, we let $t \downarrow 0$ in the equation (5.2). By the assumption, the left hand side converges to $f(0) = 1$. For the right hand side, note that for each fixed r,

$$\big(1 - \frac{|t|}{r}\big)^+ \uparrow 1 \quad \text{as } t \downarrow 0.$$

By the monotone convergence theorem, we conclude that

$$1=\lim_{t\downarrow 0}\int_{(0,\infty)}\big(1-\frac{|t|}{r}\big)^{+}\nu(dr)=\int_{(0,\infty)}1\nu(dr)=\nu(0,\infty).$$

Therefore, ν is a probability measure on $(0,\infty)$, finishing the proof of Theorem 5.1.

□

We conclude this topic by two interesting applications of Pólya's theorem.

Corollary 5.1. *Let $c>0$. There exist two different characteristic functions f_1, f_2 such that*

$$f_1(t)=f_2(t)\quad \text{for } t\in(-c,c).$$

Proof. Let $f_1(t)=e^{-|t|}$ be the characteristic function of the Cauchy distribution. We draw the tangent line of $f_1(t)$ at the point $A=(c,f_1(c))$ and let this line intersect the positive t-aixs at the point B. We define f_2 to be the function whose graph on $(0,\infty)$ is given by

(i) the graph of f_1 on the part of $(0,c)$;
(ii) the line segment $\overline{AB}$ on the part from A to B;
(iii) the zero function from B to infinitely.

The construction is mostly clear when one draws a picture. $f_2(t)$ is assumed to be extended to the negative axis by symmetry. It is readily checked that f_2 satisfies Pólya's criterion and is thus a characteristic function. The functions f_1, f_2 satisfy the desired property. □

Corollary 5.2. *There exist three characteristic functions f_1, f_2, f_3 such that $f_1\neq f_2$ but $f_1f_3=f_2f_3$.*

Proof. Let f_1, f_2 be given as in Corollary 5.1. Let

$$f_3(t)\triangleq(1-\frac{|t|}{c'})^{+}$$

where $c'\in(0,c)$ is a fixed constant. Then f_1, f_2, f_3 are desired. □

Remark 5.2. Corollary 5.2 tells us that the cancellation law does not hold for characteristic functions, i.e.

$$f_1f_3=f_2f_3\nRightarrow f_1=f_2.$$

6 Appendix: The uniqueness theorem without inversion

Using the inversion formula to prove the uniqueness result (as we did) is quite heavy and unnatural. There is a more direct argument which gives us better insight into the uniqueness property. Suppose that μ_1 and μ_2 have the same characteristic function, i.e.

$$\int_{-\infty}^{\infty} e^{itx}\mu_1(dx) = \int_{-\infty}^{\infty} e^{itx}\mu_2(dx) \quad \text{for all } t \in \mathbb{R}.$$

We want to show that $\mu_1 = \mu_2$. The general idea is the following.

(i) it is enough to show that

$$\int_{-\infty}^{\infty} f(x)\mu_1(dx) = \int_{-\infty}^{\infty} f(x)\mu_2(dx) \tag{6.1}$$

for a sufficiently large class of functions f,
(ii) and this class of functions can be approximated by linear combinations of functions from the family $\{e^{itx} : t \in \mathbb{R}\}$.

The first point is natural to expect. The fact that the family $\{e^{itx} : t \in \mathbb{R}\}$ generates a wide class of functions is also natural from the view of Fourier series: any continuous periodic function $f(x)$ with period $T = 1$ (i.e. $f(x+1) = f(x)$) admits a Fourier series expansion

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{2\pi inx}, \quad x \in [0,1],$$

where $c_n = \int_0^1 f(x)e^{2\pi inx}dx$ is the Fourier coefficient.

Instead of using Fourier series, we rely on a rather power theorem of Stone-Weierstrass, stated in the context of periodic functions as follows.

Theorem 6.1 (The Stone-Weierstrass Theorem for period functions)**.** *Let $T > 0$. Define $\mathcal{C}_T$ to be the space of continuous periodic functions $f : \mathbb{R} \to \mathbb{C}$ with period T. Let $\mathcal{A}$ be a subset of $\mathcal{C}_T$ satisfying the following properties:*

(i) $\mathcal{A}$ is an algebra: $f, g \in \mathcal{A}, a, b \in \mathbb{R} \implies af + bg, f \cdot g \in \mathcal{A}$;
(ii) $\mathcal{A}$ vanishes at no point: for any $x \in [0, T)$, there exists $f \in \mathcal{A}$ such that $f(x) \neq 0$;

(iii) $\mathcal{A}$ separates points: for any $x \neq y \in [0,T)$, there exists $f \in \mathcal{A}$ such that $f(x) \neq f(y)$.

Then $\mathcal{A}$ is dense in $\mathcal{C}_T$ with respect to uniform convergence on $[0,T]$. More precisely, for any periodic function $f \in \mathcal{C}_T$ and $\varepsilon > 0$, there exists $g \in \mathcal{A}$ such that

$$\sup_{t \in [0,T]} |f(t) - g(t)| < \varepsilon.$$

Now we prove the uniqueness result for the characteristic function by using the Stone-Weierstrass theorem.

Another proof of Corollary 2.1. Let μ_1, μ_2 be two probability measures having the same characteristic function.

We first claim that (6.1) holds for any continuous periodic function f. Indeed, let $T > 0$ be an arbitrary positive number and define $\mathcal{C}_T$ to be the space of periodic functions $f : \mathbb{R} \to \mathbb{C}$ with period T. Let $\mathcal{A}_T \subseteq \mathcal{C}_T$ be the vector space spanned by the family $\{e^{2\pi i n x/T} : n \in \mathbb{Z}\}$ of functions. It is tedious to check that $\mathcal{A}_T$ satisfies all the assumptions in Theorem 6.1. Therefore, $\mathcal{A}_T$ is dense in $\mathcal{C}_T$ with respect to uniform convergence on $[0,T]$. On the other hand, by assumption we know that (6.1) holds for every $f \in \mathcal{A}_T$. It follows from a simple approximation argument that (6.1) holds for every $f \in \mathcal{C}_T$.

Next, we claim that (6.1) holds for every bounded continuous function f. The idea is to replace f by a periodic function with large period. Given an arbitrary $\varepsilon > 0$, there exists $M > 0$ such that

$$\mu_i([-M,M]^c) < \varepsilon \quad \text{for } i = 1, 2.$$

Let $g : [-M-1, M+1] \to \mathbb{R}$ be the continuous function given by

$$g(x) \triangleq \begin{cases} f(x), & x \in [-M, M]; \\ 0, & x \in (-\infty, -M-1) \cup (M+1, \infty); \\ \text{linear}, & x \in [-M-1, -M] \text{ or } x \in [M, M+1]. \end{cases}$$

By definition we have $g(-M-1) = g(M+1)$ and

$$|g(x)| \leqslant \|f\|_\infty \triangleq \sup_{y \in \mathbb{R}} |f(y)| \quad \text{for all } x \in [-M-1, M+1]$$

Let $\bar{g} : \mathbb{R} \to \mathbb{R}$ be the periodic extension of g to $\mathbb{R}$ with period $T = 2M + 2$. From the previous step we know that (6.1) holds for $\bar{g}$. Since we also have $f = \bar{g}$ on

$[-M, M]$, it follows that

$$\begin{aligned}
&|\int f d\mu_1 - \int f d\mu_2| \\
&\leqslant |\int f d\mu_1 - \int \bar{g} d\mu_1| + |\int \bar{g} d\mu_1 - \int \bar{g} d\mu_2| + |\int \bar{g} d\mu_2 - \int f d\mu_2| \\
&= |\int f d\mu_1 - \int \bar{g} d\mu_1| + |\int \bar{g} d\mu_2 - \int f d\mu_2| \\
&\leqslant 2\|f\|_\infty \cdot \big(\mu_1([-M, M]^c) + \mu_2([-M, M]^c)\big) \\
&< 4\|f\|_\infty \varepsilon.
\end{aligned}$$

Since ε is arbitrary, we conclude that (6.1) holds for f.

Finally, if (6.1) holds for all bounded continuous functions, we must have $\mu_1 = \mu_2$. The verification of this point is left to the reader as an exercise.

□

Topic 4: The Central Limit Theorem

The classical central limit theorem describes the phenomenon that the fluctuation of the partial sum of an i.i.d. sequence around its mean is asymptotically Gaussian. This behaviour is universal as the particular distribution of the sequence is of little relevance and one ends up with a canonical Gaussian limit. The mathematics behind the appearance of this Gaussian nature is rather deep. In vague terms, the reason lies in two aspects: the dependence structure in the sequence is weak and the contribution of each individual term in the sum is negligible in some sense. In this topic, we develop some insights into the hidden mechanism of this fundamental phenomenon from several perspectives. The next topic, which deals with the rate of convergence, will uncover the secrets in an even deeper way.

1 The classical central limit theorem

We start by recapturing the classical central limit theorem in the context of i.i.d. random variables. This fundamental result was due to Lindeberg and Lévy.

Theorem 1.1. *Let $\{X_n : n \geqslant 1\}$ be a sequence of independent and identically distributed random variables. Suppose that X_1 has finite mean and variance. Then $\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ converges weakly to the standard normal distribution as $n \to \infty$, where $S_n \triangleq X_1 + \cdots + X_n$.*

Proof. We may assume that $\mathbb{E}[X_1] = 0$, for otherwise we can consider the sequence $X_n - \mathbb{E}[X_n]$ instead, which does not change the original claim. Let $f(t)$ be the characteristic function of X_1. Since X_1 has finite second moment, according to Topic 3, Corollary 4.1, we have

$$f(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2),$$

where $\sigma^2 \triangleq \mathrm{Var}[X_1]$. Since the sequence $\{X_n\}$ is i.i.d., the characteristic function of $\frac{S_n-\mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ is easily seen to be given by

$$\begin{aligned} f_n(t) &= \big(f\big(\frac{t}{\sigma\sqrt{n}}\big)\big)^n \\ &= \big(1-\frac{t^2}{2n}+o\big(\frac{t^2}{n\sigma^2}\big)\big)^n. \end{aligned}$$

Note that t is fixed and the infinitesimal term $o(t^2/n\sigma^2)$ is understood as $n\to\infty$. If we write

$$c_n \triangleq -\frac{t^2}{2n}+o\big(\frac{t^2}{n\sigma^2}\big),$$

then

$$f_n(t) = (1+c_n)^{\frac{1}{c_n}\cdot nc_n} \to e^{-t^2/2}.$$

The limit is precisely the characteristic function of the standard normal distribution. According to the Lévy-Cramér theorem, we conclude that

$$\frac{S_n-\mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}} \to N(0,1), \quad \text{weakly.}$$

□

The above proof, as the most standard one, is so simple that it has unfortunately concealed most of the deeper insights into this fundamental theorem. The use of characteristic functions is somehow like a piece of magic, leaving the audience in shock after the play is over without telling the deeper truth of why. On the other hand, the following argument perhaps provides us with a little bit more clues towards the matter.

It is often true (and is natural to believe) that most of the common distributions are uniquely determined by the sequence of moments. The normal distribution is one such example. Recall that, the moments of $Z \overset{d}{=} N(0,1)$ are given by

$$\mathbb{E}[Z^{2m-1}]=0,\ \mathbb{E}[Z^{2m}]=(2m-1)\cdot(2m-3)\cdot\cdots\cdot 3\cdot 1$$

for each $m \geqslant 1$.

Let us compute moments of the quantity $\frac{S_n-\mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$. We assume that $\mathbb{E}[X_1]=0$ and $\mathrm{Var}[X_1]=1$, so that $\frac{S_n-\mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ becomes $\frac{S_n}{\sqrt{n}}$. The general case can always be reduced to this standardised one. To make use of the idea of moments, let us further assume that X_1 has finite moments of all orders. The following result is the crucial point why we expect that $\frac{S_n}{\sqrt{n}}$ converges weakly to $N(0,1)$.

Lemma 1.1. *For each $m \geqslant 1$, we have*

$$\lim_{n\to\infty}\mathbb{E}\big[\big(\frac{S_n}{\sqrt{n}}\big)^m\big] = L_m,$$

where $L_m \triangleq \mathbb{E}[Z^m]$ is the m-th moment of the standard normal distribution.

Proof. We prove the claim by induction on m. The case when $m = 1$ is trivial. When $m = 2$, we have

$$\mathbb{E}[\big(\frac{S_n}{\sqrt{n}}\big)^2] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[X_j^2] = 1 = L_2.$$

Now suppose that the claim is true for a general m. To examine the $(m+1)$-case, first observe that

$$\begin{aligned}\mathbb{E}[S_n^{m+1}] &= \mathbb{E}[(X_1 + \cdots + X_n)S_n^m] \\ &= n\cdot\mathbb{E}[X_nS_n^m] \quad \text{(since } \{X_n\} \text{ are i.i.d.)} \\ &= n\cdot\mathbb{E}[X_n(X_n + S_{n-1})^m] \\ &= n\cdot\sum_{j=0}^{m}\Big(\begin{array}{c} m \\ j \end{array}\Big)\mathbb{E}[X_n^{j+1}]\mathbb{E}[S_{n-1}^{m-j}] \\ &= nm\cdot\mathbb{E}[S_{n-1}^{m-1}] + n\cdot\sum_{j=2}^{m}\Big(\begin{array}{c} m \\ j \end{array}\Big)\mathbb{E}[X_n^{j+1}]\mathbb{E}[S_{n-1}^{m-j}], \end{aligned} \tag{1.1}$$

where to reach the last equality we have used the fact that $\mathbb{E}[X_n] = 0$ and $\mathbb{E}[X_n^2] = 1$.

Now let us take into account the $\sqrt{n}$-normalisation. To simplify the notation, we set

$$L_m(n) \triangleq \mathbb{E}\big[\big(\frac{S_n}{\sqrt{n}}\big)^m\big]$$

and $C_j \triangleq \mathbb{E}[X_n^{j+1}]$. It follows from (1.1) that

$$\begin{aligned}L_{m+1}(n) = {} & mL_{m-1}(n-1)\cdot\big(\frac{n-1}{n}\big)^{\frac{m-1}{2}} \\ & + \sum_{j=2}^{m}\Big(\begin{array}{c} m \\ j \end{array}\Big)C_jL_{m-j}(n-1)\cdot\frac{(n-1)^{(m-j)/2}}{n^{(m-1)/2}}.\end{aligned}$$

According to the induction hypothesis and the simple observation that (for $j \geqslant 2$)

$$\frac{(n-1)^{(m-j)/2}}{n^{(m-1)/2}} \to 0 \text{ as } n \to \infty,$$

we conclude that

$$L_{m+1}(n) \to mL_{m-1}$$

as $n \to \infty$. This not only shows the convergence of $L_{m+1}(n)$, but more importantly its convergence to the correct limit

$$mL_{m-1} = L_{m+1},$$

which is precisely the relation that the moments of $N(0,1)$ satisfy. This completes the proof of the lemma. □

Based on Lemma 1.1, it is now reasonable to expect that the central limit theorem holds true, i.e. $\frac{S_n}{\sqrt{n}}$ converges weakly to $N(0,1)$. Technically there is still a missing component in the proof, namely why the convergence of each moment implies the weak convergence. This is the content of the more general *moment problem* in probability theory. We will not delve into this property which is not so surprising to believe at the heuristic level.

An application of the classical central limit theorem

We give an enlightening application of the classical central limit to prove the famous Stirling's formula:

$$n! \sim \sqrt{2\pi n}\big(\frac{n}{e}\big)^n,$$

where the notation $a_n \sim b_n$ means $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$.

We fist provide a heuristic but semi-rigorous argument. Let $\{X_n : n \geqslant 1\}$ be a sequence of independent and Poisson distributed random variables with parameter 1. Define $S_n \triangleq X_1 + \cdots + X_n$. We can then write

$$\begin{aligned}\mathbb{P}(S_n = n) &= \mathbb{P}(n-1 < S_n \leqslant n) \\ &= \mathbb{P}\big(-\frac{1}{\sqrt{n}} < \frac{S_n - n}{\sqrt{n}} \leqslant 0\big).\end{aligned}$$

By the central limit theorem, we know that $\frac{S_n - n}{\sqrt{n}} \to N(0,1)$ weakly. In particular,

$$\mathbb{P}(S_n = n) \approx \frac{1}{\sqrt{2\pi}} \int_{-1/\sqrt{n}}^{0} e^{-x^2/2} dx.$$

Note that

$$\mathbb{P}(S_n = n) = \frac{n^n e^{-n}}{n!}$$

since $S_n \overset{d}{=} \text{Poisson}(n)$, and

$$\int_{-1/\sqrt{n}}^{0} e^{-x^2/2}dx \approx \frac{1}{\sqrt{n}}.$$

It follows that

$$\frac{n^n e^{-n}}{n!} \approx \frac{1}{\sqrt{2\pi n}}$$

which is precisely the Stirling approximation. This argument is not rigorous since the step

$$\mathbb{P}\big(-\frac{1}{\sqrt{n}} < \frac{S_n - n}{\sqrt{n}} \leqslant 0\big) \approx \frac{1}{\sqrt{2\pi}} \int_{-1/\sqrt{n}}^{0} e^{-x^2/2}dx.$$

is by no mean a simple consequence of the central limit theorem, as we are also varying the end point of the interval.

To give a rigorous treatment, let us instead use a sequence $\{X_n : n \geqslant 1\}$ of independent and exponential distributed random variables with parameter 1. It is well known that $S_n \triangleq X_1 + \cdots + X_n$ now follows a Gamma distribution with parameter n and 1. Using the explicit formula for the density function of a Gamma distribution, it is plain to check that

$$\begin{aligned} &\mathbb{P}\big(0 \leqslant \frac{S_{n+1} - (n+1)}{\sqrt{n+1}} \leqslant 1\big) \\ &= \frac{\sqrt{n+1}}{n!} \int_0^1 (\sqrt{n+1} \cdot (x + \sqrt{n+1}))^n e^{-\sqrt{n+1}(x+\sqrt{n+1})}dx. \end{aligned} \tag{1.2}$$

In the first place, according to the central limit theorem, we know that

$$\lim_{n\to\infty} \mathbb{P}\big(0 \leqslant \frac{S_{n+1} - (n+1)}{\sqrt{n+1}} \leqslant 1\big) = \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-x^2/2}dx. \tag{1.3}$$

On the other hand, if we apply two steps of change of variables:

$$y = \sqrt{n+1}(x + \sqrt{n+1}),\ z = \frac{y-n}{\sqrt{n}}$$

and on the right hand side of (1.2), it leads us to

$$\begin{aligned}&\mathbb{P}\big(0 \leqslant \frac{S_{n+1}-(n+1)}{\sqrt{n+1}} \leqslant 1\big)\\ &= \frac{1}{n!}\int_{1+n}^{1+n+\sqrt{n+1}} y^n e^{-y} dy\\ &= \frac{\sqrt{n}n^n e^{-n}}{n!}\int_{\frac{1}{\sqrt{n}}}^{\frac{1}{\sqrt{n}}+\sqrt{1+\frac{1}{n}}} \big(1+\frac{z}{\sqrt{n}}\big)^n e^{-\sqrt{n}z} dz.\end{aligned}$$

Note that

$$\begin{aligned}\big(1+\frac{z}{\sqrt{n}}\big)^n &= \exp\big(n\log\big(1+\frac{z}{\sqrt{n}}\big)\big)\\ &= \exp\big(n\cdot\big(\frac{z}{\sqrt{n}}-\frac{z^2}{2n}+o(\frac{1}{n})\big)\big),\end{aligned}$$

and thus

$$\lim_{n\to\infty}\big(1+\frac{z}{\sqrt{n}}\big)^n e^{-\sqrt{n}z} = e^{-z^2/2}.$$

It follows that

$$\mathbb{P}\big(0 \leqslant \frac{S_{n+1}-(n+1)}{\sqrt{n+1}} \leqslant 1\big) \sim \frac{\sqrt{n}n^n e^{-n}}{n!}\int_0^1 e^{-z^2/2}dz. \tag{1.4}$$

Now Stirling's formula follows from comparing (1.3) and (1.4).

2 Lindeberg's central limit theorem

There are at least two reasons why we still wish to push the matter further. *The first reason* is that, in the classical central limit theorem we have assumed that the sequence of random variables $\{X_n\}$ are identically distributed. The two proofs given in the last section make use of this condition in a crucial way. However, this condition is not an essential point at all for the central limit theorem. We need to understand the deeper reason that has led to this phenomenon. *The second reason* is that, the previous proofs are only qualitative, as it tells us nothing about how close the distribution of $\frac{S_n-\mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ is to the standard normal one for each given n. For practical purposes it is necessary and important to develop robust tools for studying the rate of convergence in the central limit theorem.

Lindeberg's central limit theorem provides essential insights towards the above two aspects. For the first aspect, it suggests that some sort of "uniform negligibility of each summand X_m $(1 \leqslant m \leqslant n)$ with respect to S_n" is crucial to for the central limit theorem to hold. For the second aspect, recall that probability measures μ_n on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ converge weakly to μ if and only if

$$\int_{\mathbb{R}} f(x)\mu_n(dx) \to \int_{\mathbb{R}} f(x)\mu(dx) \quad \text{for all } f \in C_b(\mathbb{R}).$$

In this spirit, a natural way of comparing the "distance" between μ_n and μ is to quantitatively estimate the distance $\left| \int_{\mathbb{R}} f d\mu_n - \int_{\mathbb{R}} f d\mu \right|$ for each f in a suitable class of functions. In the context of random variables and the central limit theorem, this is about estimating the distance

$$\left|\mathbb{E}\left[f\left(\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}\right)\right] - \mathbb{E}[f(Z)]\right| \quad \text{for suitable class of functions } f,$$

where $Z \overset{d}{=} N(0,1)$. Lindeberg's central limit theorem precisely gives an answer to the question of such kind.

Before stating the theorem, we first present the basic set-up. We are again considering a sequence $\{X_n : n \geqslant 1\}$ of independent (but not necessarily identically distributed!) random variables with finite mean and variance. We assume that $\mathbb{E}[X_n] = 0$, for otherwise we can always centralised the sequence to have mean zero. For each $n \geqslant 1$, let

$$\sigma_n \triangleq \sqrt{\mathrm{Var}[X_n]}, \ \Sigma_n \triangleq \sqrt{\mathrm{Var}[S_n]}, \ \hat{S}_n \triangleq \frac{S_n}{\Sigma_n}.$$

We introduce two key quantities that will appear in the rate of convergence estimate:

$$r_n \triangleq \max_{1 \leqslant m \leqslant n} \frac{\sigma_m}{\Sigma_n} \tag{2.1}$$

and

$$g_n(\varepsilon) \triangleq \frac{1}{\Sigma_n^2} \sum_{m=1}^{n} \mathbb{E}[X_m^2; |X_m| \geqslant \varepsilon \Sigma_n], \quad \varepsilon > 0.$$

Vaguely speaking, these two quantities reflect the relative magnitude of each summand X_m $(1 \leqslant m \leqslant n)$ with respect to S_n. We also recall the notation $\|f\|_\infty \triangleq \sup_{x \in \mathbb{R}} |f(x)|$ for a given function $f : \mathbb{R} \to \mathbb{R}$.

Now we are able to state Lindeberg's central limit theorem. In many ways it is deeper and more fundamental than the classical central limit theorem in the last section. We denote $\mathcal{C}_b^3(\mathbb{R})$ as the space of functions $f : \mathbb{R} \to \mathbb{R}$ that are continuously differentiable with bounded derivatives up to order three.

Theorem 2.1. *Under the aforementioned set-up, let $f \in \mathcal{C}_b^3(\mathbb{R})$. Then for each $\varepsilon > 0$ and $n \geqslant 1$, we have*

$$\big|\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(Z)]\big| \leqslant \big(\frac{\varepsilon}{6} + \frac{\gamma \cdot r_n}{6}\big)\|f'''\|_\infty + g_n(\varepsilon) \cdot \|f''\|_\infty, \tag{2.2}$$

where $Z \overset{d}{=} N(0,1)$ and $\gamma \triangleq \mathbb{E}[|Z|^3] = \sqrt{\frac{8}{\pi}}$ is the third absolute moment of Z. In addition, if

$$\lim_{n\to\infty} g_n(\varepsilon) = 0 \quad \text{for every } \varepsilon > 0, \tag{2.3}$$

then

$$\hat{S}_n \to Z \quad \text{weakly}$$

as $n \to \infty$, giving the central limit theorem for $\{X_n\}$.

The condition (2.3) is known as *Lindeberg's condition.* Theorem 2.1 therefore tells us that Lindeberg's condition implies a central limit theorem in the context of independent random variables with finite mean and variance. As a direct corollary, we can recover the classical limit theorem. Indeed, if $\{X_n : n \geqslant 1\}$ is i.i.d., then $\Sigma_n = \sqrt{n}\sigma$ ($\sigma^2 \triangleq \mathrm{Var}[X_1]$), and thus

$$\begin{aligned} g_n(\varepsilon) &= \frac{1}{n\sigma^2}\sum_{m=1}^n \mathbb{E}[X_m^2; |X_m| \geqslant \varepsilon\sqrt{n}\sigma] \\ &= \frac{1}{\sigma^2}\mathbb{E}[X_1^2 : |X_1| \geqslant \varepsilon\sqrt{n}\sigma] \end{aligned}$$

which goes to zero as $n \to \infty$. In particular, Lindeberg's condition holds. A more interesting corollary of Lindeberg's theorem is the following *Lyapunov's central limit theorem.*

Corollary 2.1. *Let $\{X_n : n \geqslant 1\}$ be a sequence of independent random variables with mean zero and finite third moments. Define S_n, Σ_n as before, and we also set*

$$\Gamma_n \triangleq \sum_{m=1}^n \mathbb{E}[|X_m|^3].$$

If $\frac{\Gamma_n}{\Sigma_n^3} \to 0$, then $\frac{S_n}{\Sigma_n}$ converges weakly to $N(0,1)$.

Proof. We verify Lindeberg's condition by using Chebyshev's inequality:

$$g_n(\varepsilon) = \frac{1}{\Sigma_n^2}\sum_{m=1}^n \mathbb{E}[X_m^2; |X_m| \geqslant \varepsilon\Sigma_n] \leqslant \frac{1}{\varepsilon\Sigma_n^3}\sum_{m=1}^n \mathbb{E}[|X_m|^3] = \frac{\Gamma_n}{\varepsilon\Sigma_n^3} \to 0.$$

□

Remark 2.1. Lyapunov's central limit theorem can be derived by using the method of characteristic functions, by looking at third order Taylor expansions for the characteristic function. We leave it as a good exercise to the reader.

The rest of this section is devoted to the proof of Lindeberg's central limit theorem.

Proof of Theorem 2.1

We first establish the quantitative estimate (2.2), and then show how leads to the weak convergence property for the central limit theorem.

The quantitative estimate.

Fix $n \geqslant 1$. For each $1 \leqslant m \leqslant n$, we define $\hat{X}_m \triangleq \frac{X_m}{\Sigma_n}$ so that

$$\hat{S}_n = \hat{X}_1 + \cdots + \hat{X}_n.$$

The main idea of the proof is to swap each $\hat{X}_m$ to a reference normal random variable $\hat{Y}_m$ (one flip at each step) in a way that after n swaps the accumulated error between $\hat{S}_n$ and $\hat{Y}_1 + \cdots + \hat{Y}_n \stackrel{d}{=} N(0,1)$ is controllable.

Step one: Introducing the reference normal random variables. To implement this idea mathematically, let us first assume that, there are n standard normal random variables $Y_1, \cdots, Y_n$ defined on the same probability space as $X_1, \cdots, X_n$ are, and

$$X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_n$$

are all independent. This is always possible by enlarging the original probability space (a standard measure-theoretic construction). We set

$$\hat{Y}_m \triangleq \frac{\sigma_m Y_m}{\Sigma_n}, \quad 1 \leqslant m \leqslant n,$$

and

$$\hat{T}_n \triangleq \hat{Y}_1 + \cdots + \hat{Y}_n.$$

Observe that $\hat{Y}_m$ is a normal random variable that has mean zero and the same variance as $\hat{X}_m$ does. In addition, $\hat{T}_n \stackrel{d}{=} N(0,1)$. The problem is now essentially about estimating

$$\left|\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(\hat{T}_n)]\right|$$

where $f \in \mathcal{C}^3(\mathbb{R};\mathbb{R})$ is the given fixed test function.

Step two: Forming the telescoping sum. Let us form a telescoping sum to estimate the above quantity by flipping $\hat{X}_m$ to $\hat{Y}_m$, one at each step. More precisely, we write

$$\begin{aligned}
&\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(\hat{T}_n)] \\
&= \mathbb{E}[f(\hat{X}_1 + \hat{X}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] - \mathbb{E}[f(\hat{Y}_1 + \hat{X}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] \\
&\quad + \mathbb{E}[f(\hat{Y}_1 + \hat{X}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] - \mathbb{E}[f(\hat{Y}_1 + \hat{Y}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] \\
&\quad + \mathbb{E}[f(\hat{Y}_1 + \hat{Y}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] - \mathbb{E}[f(\hat{Y}_1 + \hat{Y}_2 + \hat{Y}_3 + \cdots + \hat{X}_n)] \\
&\quad \cdots \\
&\quad + \mathbb{E}[f(\hat{Y}_1 + \cdots \hat{Y}_{n-1} + \hat{X}_n)] - \mathbb{E}[f(\hat{Y}_1 + \cdots + \hat{Y}_{n-1} + \hat{Y}_n)]. \qquad (2.4)
\end{aligned}$$

To rewrite the expression in a more enlightening form, let us introduce for $1 \leqslant m \leqslant n$,

$$U_m \triangleq \hat{Y}_1 + \cdots + \hat{Y}_{m-1} + \hat{X}_{m+1} + \cdots + \hat{X}_n.$$

Then (2.4) can be written as

$$\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(\hat{T}_n)] = \sum_{m=1}^{n} \big(\mathbb{E}[f(U_m + \hat{X}_m)] - \mathbb{E}[f(U_m + \hat{Y}_m)] \big).$$

Step three: Introducing the Taylor approximation. Now we use Taylor's approximation for the function f to estimate

$$\big| \mathbb{E}[f(U_m + \hat{X}_m)] - \mathbb{E}[f(U_m + \hat{Y}_m)] \big|.$$

For this purpose, define

$$R_m(\xi) \triangleq f(U_m + \xi) - f(U_m) - f'(U_m)\xi - \frac{f''(U_m)}{2}\xi^2, \quad \xi \in \mathbb{R}.$$

This is the remainder for the second order Taylor expansion of f around U_m. Since $U_m, \hat{X}_m, \hat{Y}_m$ are independent, and $\hat{X}_m, \hat{Y}_m$ have the same mean and variance, we see that

$$\mathbb{E}[f(U_m + \hat{X}_m)] - \mathbb{E}[f(U_m + \hat{Y}_m)] = \mathbb{E}[R_m(\hat{X}_m)] - \mathbb{E}[R_m(\hat{Y}_m)].$$

This is a very important observation. It follows that

$$\begin{aligned}
&\big| \mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(\hat{T}_n)] \big| \\
&\leqslant \sum_{m=1}^{n} \big| \mathbb{E}[R_m(\hat{X}_m)] \big| + \sum_{m=1}^{n} \big| \mathbb{E}[R_m(\hat{Y}_m)] \big|. \qquad (2.5)
\end{aligned}$$

Step four: Estimating the $\mathbb{E}[R_m(\hat{X}_m)]$ and $\mathbb{E}[R_m(\hat{Y}_m)]$ sums separately. Next we estimate the right hand side of (2.5). First of all, by using a third order Taylor expansion of f, we have

$$|R_m(\xi)| \leqslant \frac{1}{3!}\|f'''\|_\infty |\xi|^3, \tag{2.6}$$

In addition, the second order Taylor expansion gives

$$|f(U_m+\xi) - f(U_m) - f'(U_m)\xi| \leqslant \frac{1}{2}\|f''\|_\infty |\xi|^2,$$

and thus we also have

$$\begin{aligned} |R_m(\xi)| &\leqslant \frac{1}{2}\|f''\|_\infty|\xi|^2 + \frac{1}{2}|f''(U_m)|\cdot|\xi|^2 \\ &\leqslant \|f''\|_\infty|\xi|^2. \end{aligned} \tag{2.7}$$

We use (2.6) to estimate the $\mathbb{E}[R_m(\hat{Y}_m)]$-sum as follows:

$$\begin{aligned} \sum_{m=1}^n \left|\mathbb{E}[R_m(\hat{Y}_m)]\right| &\leqslant \frac{1}{6}\|f'''\|_\infty \sum_{m=1}^n \mathbb{E}[|\hat{Y}_m|^3] \\ &= \frac{\gamma}{6}\|f'''\|_\infty \sum_{m=1}^n \frac{\sigma_m^3}{\Sigma_n^3} \\ &\leqslant \frac{\gamma}{6}\|f'''\|_\infty \cdot \frac{\max_{1\leqslant m\leqslant n}\sigma_m}{\Sigma_n} \cdot \sum_{m=1}^n \frac{\sigma_m^2}{\Sigma_n^2} \\ &= \frac{\gamma}{6}\|f'''\|_\infty \cdot r_n, \end{aligned} \tag{2.8}$$

where we recall that r_n is defined in (2.1) and $\gamma \triangleq \mathbb{E}[|Y_1|^3] = \sqrt{\frac{8}{\pi}}$ is the third absolute moment of the standard normal distribution.

The estimation of the $\mathbb{E}[R_m(\hat{X}_m)]$-sum is a bit more complicated, and we need to split the region of integration into two parts:

$$\begin{aligned}
&\sum_{m=1}^{n} \big|\mathbb{E}[R_m(\hat{X}_m)]\big| \\
&= \sum_{m=1}^{n} \big|\mathbb{E}[R_m(\hat{X}_m); |\hat{X}_m| < \varepsilon]\big| + \sum_{m=1}^{n} \big|\mathbb{E}[R_m(\hat{X}_m); |\hat{X}_m| \geqslant \varepsilon]\big| \\
&\leqslant \frac{\|f'''\|_\infty}{6} \sum_{m=1}^{n} \mathbb{E}[|\hat{X}_m|^3; |\hat{X}_m| < \varepsilon] + \|f''\|_\infty \sum_{m=1}^{n} \mathbb{E}[|\hat{X}_m|^2; |\hat{X}_m| \geqslant \varepsilon] \\
&\leqslant \frac{\|f'''\|_\infty \varepsilon}{6} \sum_{m=1}^{n} \frac{\sigma_m^2}{\Sigma_n^2} + \|f''\|_\infty g_n(\varepsilon) \\
&= \frac{\varepsilon}{6} \|f'''\|_\infty + g_n(\varepsilon)\|f''\|_\infty, \qquad (2.9)
\end{aligned}$$

where we have used (2.6) and (2.7) to estimate the two parts respectively.

The desired estimate (2.2) is now a consequence of (2.8) and (2.9).

Obtaining the central limit theorem.

Now we show that, if Lindeberg's condition (2.3) holds, then we have the central limit theorem

$$\hat{S}_n \to N(0,1) \quad \text{weakly.}$$

To this end, we first recall the definition of r_n given by (2.1). Let m be the integer at which the maximum in (2.1) is attained, i.e. $r_n = \frac{\sigma_m}{\Sigma_n}$. It follows that

$$\begin{aligned}
r_n^2 = \frac{\sigma_m^2}{\Sigma_n^2} &= \mathbb{E}[\hat{X}_m^2] \\
&= \mathbb{E}[\hat{X}_m^2; |\hat{X}_m| < \varepsilon] + \mathbb{E}[\hat{X}_m^2 : |\hat{X}_m| \geqslant \varepsilon] \\
&\leqslant \varepsilon^2 + g_n(\varepsilon),
\end{aligned}$$

for every $\varepsilon > 0$. In particular, if Lindeberg's condition (2.3) holds, then $r_n \to 0$. According to (2.2), we have

$$\mathbb{E}[f(\hat{S}_n)] \to \mathbb{E}[f(Z)] \quad \text{for every } f \in \mathcal{C}^3(\mathbb{R}),$$

where $Z \overset{d}{=} N(0,1)$.

In order to show weak convergence, using the second characterisation in the Portmanteau theorem, we have to strengthen the class $\mathcal{C}^3(\mathbb{R})$ of test functions to the class of bounded, uniformly continuous functions. This is possible due to a standard technique of molification in analysis.

Lemma 2.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a bounded and uniformly continuous function. Then there exists a sequence $f_n \in \mathcal{C}_b^3(\mathbb{R})$ such that f_n converges uniformly to f.*

Proof. The idea is to apply convolution of f with some "nice" function. One possible choice is the following. For each $\eta > 0$, define

$$\rho_\eta(x) = \frac{1}{\sqrt{2\pi\eta}} e^{-\frac{x^2}{2\eta}}, \quad x \in \mathbb{R}$$

to be the density function of $N(0, \eta)$. Let

$$f_\eta(x) \triangleq (\rho_\eta * f)(x) \triangleq \int_{\mathbb{R}} \rho_\eta(x - y) f(y) dy.$$

Since $\int_{\mathbb{R}} \rho_\eta(x)dx = 1$ and f is bounded, we know that f_η is well defined. Indeed, f_η is smooth and its k-th derivative is given by

$$f_\eta^{(k)}(x) = \int_{\mathbb{R}} \rho_\eta^{(k)}(x - y) f(y) dy$$

which is easily seen to be bounded on $\mathbb{R}$.

We now show that f_η converges uniformly to f as $\eta \to 0$. First of all, since f is uniformly continuous, given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|y - x| < \delta \implies |f(y) - f(x)| < \varepsilon.$$

It follows that

$$\begin{aligned} \big|f_\eta(x) - f(x)\big| &= \big|\int_{\mathbb{R}} \rho_\eta(x - y)(f(y) - f(x))dy\big| \\ &\leqslant \big|\int_{\{y:|y-x|<\delta\}} \rho_\eta(x - y)(f(y) - f(x))dy\big| \\ &\quad + \big|\int_{\{y:|y-x|\geqslant\delta\}} \rho_\eta(x - y)(f(y) - f(x))dy\big| \\ &\leqslant \varepsilon + 2\|f\|_\infty \cdot \int_{\{y:|y-x|\geqslant\delta\}} \rho_\eta(x - y)dy \\ &= \varepsilon + 2\|f\|_\infty \cdot \mathbb{E}[|X_\eta| \geqslant \delta] \end{aligned}$$

where $X_\eta \overset{d}{=} N(0, \eta)$. Note that

$$\mathbb{E}[|X_\eta| \geqslant \delta] = \mathbb{E}\big[Z \geqslant \frac{\delta}{\sqrt{\eta}}\big] \to 0 \quad \text{as } \eta \to \infty,$$

where $Z \overset{d}{=} N(0,1)$. Therefore,

$$\varlimsup_{\eta\to 0}\|f_\eta - f\|_\infty \leqslant \varepsilon,$$

and the result follows as ε is arbitrary. □

To complete the proof of the central limit theorem, let f be a bounded and uniformly continuous function on $\mathbb{R}$. Given $\varepsilon > 0$, let $g \in \mathcal{C}_b^3(\mathbb{R})$ be such that

$$\|g - f\|_\infty \triangleq \sup_{x\in\mathbb{R}} |g(x) - f(x)| < \varepsilon.$$

The existence of g is guaranteed by Lemma 2.1. It follows that

$$\begin{aligned}
&\big|\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(Z)]\big| \\
&\leqslant \big|\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[g(\hat{S}_n)]\big| + \big|\mathbb{E}[g(\hat{S}_n)] - \mathbb{E}[g(Z)]\big| \\
&\quad + \big|\mathbb{E}[g(Z)] - \mathbb{E}[f(Z)]\big| \\
&\leqslant 2\varepsilon + \big|\mathbb{E}[g(\hat{S}_n)] - \mathbb{E}[g(Z)]\big|.
\end{aligned}$$

According to (2.2), the second term tends to zero as $n \to \infty$. Since ε is arbitrary, we conclude that

$$\mathbb{E}[f(\hat{S}_n)] \to \mathbb{E}[f(Z)].$$

This yields the desired weak convergence.

Remark 2.2. We have seen that Lindeberg's condition (2.3) implies that

$$\text{(i) } \frac{S_n}{\Sigma_n} \overset{\text{weakly}}{\to} N(0,1) \text{ and (ii) } r_n \to 0.$$

Later on, Feller proved that Lindeberg's condition is also necessary for (i) and (ii) to hold. This result together with Theorem 2.1 is known as the *Lindeberg-Feller* theorem.

3 Non-Gaussian central limit theorems: an example

In the i.i.d. context, if the random variables have finite mean and variance, the limiting distribution for the normalised partial sum sequence is Gaussian. However, if the random variables have heavy tails leading to less integrability, the limiting distribution (if exists) may no longer be Gaussian. We give one example to illustrate this.

Let $0 < \alpha < 2$ be fixed. Define F_α to be the distribution whose probability density function is given by

$$p_\alpha(x) \triangleq \begin{cases} \frac{\alpha}{2|x|^{1+\alpha}}, & |x| \geqslant 1; \\ 0, & \text{otherwise.} \end{cases}$$

Let $\{X_n : n \geqslant 1\}$ be an i.i.d. sequence with distribution F_α. We are interested in the behaviour of $\frac{X_1+\cdots+X_n}{a_n}$ with some normalising sequence a_n. Note that here X_1 does not have finite variance and we are not in the setting of the classical central limit theorem.

Let $f_\alpha(t)$ be the characteristic function of X_1. The crucial point for understanding this situation is to figure out the behaviour of $f_\alpha(t)$ near $t = 0$. Since $f_\alpha(0) = 1$, let us write

$$\begin{aligned} 1 - f_\alpha(t) &= \int_{-\infty}^{\infty} \big(1 - e^{itx}\big) p_\alpha(x) dx \\ &= \alpha \int_1^\infty \frac{1 - \cos tx}{x^{1+\alpha}} dx \\ &= \alpha |t|^\alpha \int_{|t|}^\infty \frac{1 - \cos u}{u^{1+\alpha}} du \\ &= \alpha |t|^\alpha \Big(\int_0^\infty \frac{1 - \cos u}{u^{1+\alpha}} du - \int_0^{|t|} \frac{1 - \cos u}{u^{1+\alpha}} du \Big). \end{aligned}$$

Since $1 - \cos u = \frac{1}{2}u^2 + o(u^2)$, we know that first integral on the right hand side is finite and

$$\int_0^{|t|} \frac{1 - \cos u}{u^{1+\alpha}} du = \int_0^{|t|} \frac{\frac{1}{2}u^2 + o(u^2)}{u^{1+\alpha}} du = O(|t|^{2-\alpha}).$$

Therefore, we see that

$$1 - f_\alpha(t) = C_\alpha |t|^\alpha + O(|t|^2) \tag{3.1}$$

when t is small, where $C_\alpha > 0$ is a constant depending only on α.

The relation (3.1) will give rise to the correct normalisation in the corresponding central limit theorem. In fact, the characteristic function of $\frac{S_n}{n^{1/\alpha}}$ ($S_n \triangleq X_1 + \cdots + X_n$) is given by

$$f_{\frac{S_n}{n^{1/\alpha}}}(t) = \big(f_\alpha\big(\frac{t}{n^{1/\alpha}}\big)\big)^n = \big(1 - \frac{C_\alpha |t|^\alpha}{n} + O\big(\frac{t^2}{n^{2/\alpha}}\big)\big)^n.$$

In the above equation, t is fixed and the term $O(\frac{t^2}{n^{2/\alpha}})$ is understood as $n \to \infty$. It follows that

$$\lim_{n\to\infty} f_{\frac{S_n}{n^{1/\alpha}}}(t) = e^{-C_\alpha|t|^\alpha}.$$

According to Lévy-Cramér's theorem, the function $g_\alpha(t) \triangleq e^{-C_\alpha|t|^\alpha}$ must be a characteristic function (of some distribution G_α) and

$$\frac{S_n}{n^{1/\alpha}} \to G_\alpha \quad \text{weakly}$$

as $n \to \infty$.

Remark 3.1. It is plain to check that when $\alpha = 1$, G_α is a Cauchy distribution.

Remark 3.2. When $\alpha > 2$, we are in the setting of the classical central limit theorem and thus $\frac{S_n}{\sqrt{n}}$ converges weakly to a normal distribution. What happens if $\alpha = 2$?

To understand in a deeper way what limiting distributions can arise from partial sums of independent random variables, we will be led to the theory of infinitely divisible distributions which is beyond the scope of our current study.

Topic 5: Stein's Method for Gaussian Approximations

To have a deeper understanding about the central limit theorem, we need to find effective ways to analyse certain "distance" between probability distributions/measures. A powerful modern technique is known as *Stein's method.* In this topic, we develop the basic ideas behind this method for Gaussian approximations, and use it to derive quantitative error estimates for the central limit theorem.

1 The general picture

Recall that, the central limit theorem asserts that $\hat{S}_n \to Z$ weakly, where $\hat{S}_n$ is a suitably normalised random variable and $Z \overset{d}{=} N(0,1)$. To understand the rate of convergence in the central limit theorem, we first need to have a natural notion of "distance" between two distribution functions (or equivalently, between two probability measures).

To get the essential idea, suppose that W and Z are two random variables with distribution functions F and G respectively. Among others, there are at least two apparent notions of "distance" between F and G :

(i) *The uniform distance*:

$$\|F-G\|_\infty \triangleq \sup_{x\in\mathbb{R}} |F(x)-G(x)|. \tag{1.1}$$

(ii) *The L^1-distance*:

$$\|F-G\|_{L^1} \triangleq \int_{\mathbb{R}} |F(x)-G(x)|dx. \tag{1.2}$$

There is a unified viewpoint to look at these two distances. Let μ,ν be the probability laws of W,Z respectively. We have seen in the definition of weak convergence and the proof of the central limit theorem that, the quantity

$$\big|\mathbb{E}[\varphi(W)] - \mathbb{E}[\varphi(Z)]\big| = \big|\int_{\mathbb{R}} \varphi d\mu - \int_{\mathbb{R}} \varphi d\nu\big|$$

when φ ranges over certain class of test functions, gives a natural sense of "closeness" between the two distributions. In fact, if one fixes a suitable class $\mathcal{H}$ of test functions on $\mathbb{R}$, there is an associated notion of distance defined by

$$d_{\mathcal{H}}(\mu, \nu) \triangleq \sup\big\{\big|\int_{\mathbb{R}} \varphi d\mu - \int_{\mathbb{R}} \varphi d\nu\big| : \varphi \in \mathcal{H}\big\}. \tag{1.3}$$

Apparently, this notion of distance depends crucially on what class of test functions we are taking.

(i)If $\mathcal{H}$ is the class of *indicator functions for half intervals*, i.e.

$$\mathcal{H} \triangleq \{\mathbf{1}_{(-\infty,a]}(x) : a \in \mathbb{R}\},$$

then $d_{\mathcal{H}}(\mu, \nu)$ recovers the uniform distance between F and G defined in (1.1). The uniform distance is often known as the *Kolmogorov distance.*

(ii) Now assume further that W and Z both have finite mean. If we take $\mathcal{H}$ to be the class of 1-*Lipschitz functions*, i.e. the class of functions $\varphi : \mathbb{R} \to \mathbb{R}$ such that

$$|\varphi(x) - \varphi(y)| \leqslant |x - y| \quad \text{for all } x, y \in \mathbb{R},$$

then it can be shown that $d_{\mathcal{H}}(\mu, \nu)$ recovers the L^1-distance between F and G defined in (1.2). This fact, which is not entirely obvious, will be clear from the appendix. This distance is often known as the 1-*Wasserstein distance.*

(iii) There is another natural distance associated with the class of test functions taken to be all *indicator functions of Borel subsets*, i.e. $\mathcal{H} \triangleq \{\mathbf{1}_A(x) : A \in \mathcal{B}(\mathbb{R})\}$. The associated distance, given by

$$d_{\mathcal{H}}(\mu, \nu) = \sup_{A \in \mathcal{B}(\mathbb{R})} \big|\int_{\mathbb{R}} \mathbf{1}_A d\mu - \int_{\mathbb{R}} \mathbf{1}_A d\nu\big| = \sup_{A \in \mathcal{B}(\mathbb{R})} \big|\mu(A) - \nu(A)\big|,$$

is known as the *total variation distance.* This distance is commonly used in the context of discrete random variables, in particular in the study of Poisson approximations.

From the above discussion, we see that in order to estimate the "distance" between the distributions of W and Z, a crucial ingredient is to find an effective way to estimate the quantity

$$|\mathbb{E}[\varphi(W)] - \mathbb{E}[\varphi(Z)]| \tag{1.4}$$

in terms of suitable "norms" of the test function φ. For instance, from Topic 4, Theorem 2.1 (Lindeberg's central limit theorem) we have seen such type of estimate in terms of the third derivative of φ. But this is not sufficient for many applications, and we need to strengthen the estimate to other norms of φ (e.g. in terms of the first derivative of φ).

In the 1960s, C. Stein developed a powerful method, now known as *Stein's method*, to estimate distributional distances defined through quantities like (1.4). The scope of Stein's method goes way beyond the central limit theorem and Gaussian approximations. However, we will only discuss the Gaussian case in the most classical set-up. The analysis we develop here contains the essential ideas behind this method.

As our main goal of this topic, we will use Stein's method to estimate the L^1-distance between the distributions of $W = \hat{S}_n$ and $Z \stackrel{d}{=} N(0,1)$ in the context of independent random variables. This estimate is known as the *L^1-Berry-Esseen estimate.* The uniform Berry-Esseen estimate (i.e. the corresponding estimate for the uniform distance) is much harder to obtain, but is still achievable along the lines of Stein's method.

The philosophy and basic ingredients of Stein's method for Gaussian approximation

Recall that, in the context of the central limit theorem, Z is a standard normal random variable and $W = \hat{S}_n$. The starting point of Stein's method is the following simple calculation. Let f be some nice test function. By applying integration by parts (assuming the boundary term goes away), we have

$$\begin{aligned}\mathbb{E}[f'(Z)] &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f'(z) e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} z f(z) e^{-z^2/2} dz \\ &= \mathbb{E}[Z f(Z)].\end{aligned}$$

A key observation is that, the above property indeed characterises the standard normal distribution. Namely, a random variable Z is $N(0,1)$-distributed if and only if

$$\mathbb{E}[f'(Z)] - \mathbb{E}[Z f(Z)] = 0 \tag{1.5}$$

for a wide class of test functions f. This will be the content of Stein's lemma in Section 2. From this point, one naturally expects that, if the distribution of W is "close to" $N(0,1)$, then the quantity $\mathbb{E}[f'(W)] - \mathbb{E}[W f(W)]$ should be "small".

To quantify this property, recall that we wish to estimate (1.4) for some given test function φ, where W is a general random variable and $Z \overset{d}{=} N(0,1)$. The next key step, is to write down a so-called *Stein's equation* associated with the given function φ:

$$f'(x) - xf(x) = \varphi(x) - c_\varphi, \tag{1.6}$$

where

$$c_\varphi \triangleq \mathbb{E}[\varphi(Z)] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \varphi(z) e^{-z^2/2} dz$$

is the mean of φ with respect to the standard normal distribution. The form of this equation is naturally motivated from the characterisation (1.5). Stein's equation (1.6) is a first order linear ODE, whose solution f can be written down easily. It follows that

$$f'(W) - Wf(W) = \varphi(W) - c_\varphi.$$

Now if we take expectation on both sides, we arrive at

$$\mathbb{E}[f'(W)] - \mathbb{E}[Wf(W)] = \mathbb{E}[\varphi(W)] - \mathbb{E}[\varphi(Z)].$$

In particular, the original task of estimating (1.4) is magically transferred to the estimation of the quantity

$$\mathbb{E}[f'(W)] - \mathbb{E}[Wf(W)]. \tag{1.7}$$

Note that if $W = Z$, this quantity is zero which is consistent with the characterisation (1.5). In general, this quantity can be estimated in terms of certain derivatives of the function f (the development of this part is the last step in Stein's method). Since our original goal is to estimate (1.4) in terms of φ, we must find a way to estimate derivatives of the solution f in terms of suitable norms of φ. This part corresponds to the analysis of Stein's equation, which will be developed in Section 3.

The last step, is to estimate the quantity (1.7). There is no universal approach to this step. The analysis of this part depends heavily on the specific problem we are considering (i.e. the specific assumption on the random variable W). To illustrate the essential idea, we will only develop this step in Section 4 in the context of independent random variables, i.e. when $W = \hat{S}_n$ with $\{X_n : n \geqslant 1\}$ being an independent sequence. Nonetheless, we must point out that, this step can be developed extensively in the dependent context, which makes Stein's method robust and powerful.

To summarise, there are three main steps for developing Stein's method.

Step one. Establish the characterising property for the standard normal distribution. Abstractly, this characterising property takes the form

$$\mathbb{E}[\mathcal{A}f(Z)] = 0 \quad \text{for all suitable test functions } f.$$

For the standard normal distribution, we have seen that $(\mathcal{A}f)(x) = f'(x) - xf(x)$.

Step two. Write down Stein's equation associated with a given test function φ. This equation takes the form

$$\mathcal{A}f = \varphi - c_{\varphi}.$$

For the standard normal distribution, this equation is given by (1.6). Estimate the solution f in terms of the given function φ.

Step three. Using the specific structure of the random variable W to estimate the quantity $\mathbb{E}[\mathcal{A}f(W)]$ in terms of f. In our Gaussian context, this quantity is given by (1.7).

Remark 1.1. Although we are only considering Gaussian approximations, the formulation of the previous three steps is robust enough to be applied to other types of distributional approximations, i.e. when the limiting random variable Z has other distributions (e.g. the Poisson distribution).

In the following sections, we develop the analysis for each step carefully with our ultimate goal towards the L^1-Berry-Esseen estimate in the independent case.

2 Step one: Stein's lemma for the standard normal distribution

We start by establishing the characterising property (1.5) of $N(0,1)$ rigorously. This is known as *Stein's lemma* for the normal distribution.

Lemma 2.1. *Let Z be a random variable. Then the following two statements are equivalent:*

(i) $Z \overset{d}{=} N(0,1)$;
(ii) for any piecewise differentiable function $f : \mathbb{R} \to \mathbb{R}$ that is integrable with respect to the standard Gaussian density, we have both of $\mathbb{E}[f'(Z)]$ and $\mathbb{E}[Zf(Z)]$ being finite, and

$$\mathbb{E}[f'(Z)] = \mathbb{E}[Zf(Z)].$$

Proof. (i) $\Longrightarrow$ (ii). Suppose that $Z \overset{d}{=} N(0,1)$. Given such f, we have

$$\begin{aligned}\mathbb{E}[f'(Z)] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(z) e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} f'(z) \big(\int_{-\infty}^{z} (-x) e^{-x^2/2} dx \big) dz \\ &\quad + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} f'(z) \big(\int_{z}^{\infty} x e^{-x^2/2} dx \big) dz,\end{aligned}$$

where we have written

$$e^{-z^2/2} = \int_{-\infty}^{z} (-x) e^{-x^2/2} dx = \int_{z}^{\infty} x e^{-x^2/2} dx.$$

By using Fubini's theorem (exchanging the order of integration), we have

$$\begin{aligned}&\int_{-\infty}^{0} f'(z) \big(\int_{-\infty}^{z} (-x) e^{-x^2/2} dx \big) dz \\ &= \int_{-\infty}^{0} (-x) e^{-x^2/2} dx \int_{x}^{0} f'(z) dz \\ &= \int_{-\infty}^{0} (-x) e^{-x^2/2} \big(f(0) - f(x) \big) dx \\ &= \int_{-\infty}^{0} x \big(f(x) - f(0) \big) e^{-x^2/2} dx.\end{aligned}$$

Similarly,

$$\int_{0}^{\infty} f'(z) \big(\int_{z}^{\infty} x e^{-x^2/2} dx \big) dz = \int_{0}^{\infty} x \big(f(x) - f(0) \big) e^{-x^2/2} dx.$$

Therefore,

$$\begin{aligned}\mathbb{E}[f'(Z)] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} x \big(f(x) - f(0) \big) e^{-x^2/2} dx \\ &\quad + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} x \big(f(x) - f(0) \big) e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x f(x) e^{-x^2/2} dx \\ &= \mathbb{E}[Z f(Z)].\end{aligned}$$

where we have used the fact that

$$\int_{-\infty}^{\infty} xe^{-x^2/2}dx = 0.$$

(ii) $\Longrightarrow$ (i). Let $\varphi(t) \triangleq \mathbb{E}[e^{itZ}]$ be the characteristic function of Z. Taking $f = 1$ in the assumption, we know that $\mathbb{E}[Z]$ is finite. By Topic 3, Theorem 4.1, $\varphi(t)$ is differentiable and

$$\varphi'(t) = i\mathbb{E}[Ze^{itZ}].$$

On the other hand, if we choose $f(x) = e^{itx}$ (with t fixed), we have

$$\mathbb{E}[f'(Z)] = it\mathbb{E}[e^{itZ}] = it\varphi(t),$$

and

$$\mathbb{E}[Zf(Z)] = \mathbb{E}[Ze^{itZ}] = -i\varphi'(t).$$

The assumption implies that $it\varphi(t) = -i\varphi'(t)$, or equivalently

$$\varphi'(t) = -t\varphi(t).$$

Since $\varphi(0) = 1$, the above first order linear ODE has a unique solution $\varphi(t) = e^{-t^2/2}$ which is precisely the characteristic function of $N(0,1)$. Therefore, we conclude that $Z \overset{d}{=} N(0,1)$. □

3 Step two: Analysing Stein's equation

As the next step, for a given function φ, we wish to estimate the solution f to Stein's equation

$$f'(x) - xf(x) = \varphi(x) - c_\varphi$$

in terms of φ. To do so, we first need the following important lemma regarding Gaussian tail estimates.

Lemma 3.1. *For any $x \in \mathbb{R}$, we have*

$$|x|e^{x^2/2}\int_{|x|}^{\infty} e^{-t^2/2}dt \leqslant 1, \quad e^{x^2/2}\int_{|x|}^{\infty} e^{-t^2/2}dt \leqslant \sqrt{\frac{\pi}{2}}.$$

Proof. Apparently we can assume that $x \geqslant 0$. The first claim follows from

$$xe^{x^2/2}\int_{x}^{\infty} e^{-t^2/2}dt \leqslant e^{x^2/2}\int_{x}^{\infty} te^{-t^2/2}dt = 1.$$

For the second claim, we consider the function

$$q(x) \triangleq e^{x^2/2}\int_x^\infty e^{-t^2/2}dt, \quad x \geqslant 0.$$

Using the first part, we see that

$$q'(x) = xe^{x^2/2}\int_x^\infty e^{-t^2/2}dt - 1 \leqslant 0,$$

and thus

$$q(x) \leqslant q(0) = \int_0^\infty e^{-t^2/2}dt = \sqrt{\frac{\pi}{2}}.$$

□

Remark 3.1. Heuristically, Lemma 3.1 quantifies the fact that $\mathbb{P}(|Z| > r)$ ($Z \overset{d}{=} N(0,1)$) decays like $e^{-r^2/2}$ as $r \to \infty$.

The key estimates for the solution of Stein's equation is contained in the following result.

Proposition 3.1. *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be continuously differentiable with bounded derivative. Set*

$$\tilde{\varphi}(x) \triangleq \varphi(x) - c_\varphi,$$

where recall that $c_\varphi \triangleq \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \varphi(x)e^{-x^2/2}dx$ is the mean of φ with respect to $N(0,1)$. Then

$$f(x) \triangleq e^{x^2/2}\int_{-\infty}^x \tilde{\varphi}(t)e^{-t^2/2}dt, \quad x \in \mathbb{R}, \tag{3.1}$$

is the unique bounded solution to Stein's equation

$$f'(x) - xf(x) = \tilde{\varphi}(x) \tag{3.2}$$

associated with φ. In addition, f has bounded, continuous derivatives up to order two, and the following estimates hold:

$$\|f\|_\infty \leqslant 2\|\varphi'\|_\infty, \ \|f'\|_\infty \leqslant 3\sqrt{\frac{\pi}{2}}\|\varphi'\|_\infty, \ \|f''\|_\infty \leqslant 6\|\varphi'\|_\infty.$$

Proof. From standard ODE theory, the general solution to the linear ODE (3.2) is found to be

$$f_c(x) = c \cdot e^{x^2/2} + f(x),$$

where $f(x)$ is the function defined by (3.1) and c is an arbitrary constant. In what follows we will prove that $f(x)$ is bounded. From this, it is clear that $f(x)$ is the unique bounded solution, since any choice of $c \neq 0$ will lead to an unbounded solution due to the unboundedness of $e^{x^2/2}$.

(i) Estimating f.

Let us assume that $\varphi(0) = 0$, as subtracting a constant to φ does not change $\tilde{\varphi}$ and the ODE (3.2). In this case, we have

$$|\varphi(t)| = |\varphi(t) - \varphi(0)| \leqslant \|\varphi'\|_\infty \cdot |t| \tag{3.3}$$

and

$$|c_\varphi| \leqslant \|\varphi'\|_\infty \cdot \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |t| e^{-t^2/2} dt = \|\varphi'\|_\infty \cdot \sqrt{\frac{2}{\pi}}, \tag{3.4}$$

where we have used the explicit expression for the first absolute moment of $N(0,1)$.

To estimate $f(x)$, we first consider the case when $x \leqslant 0$. By using (3.3) and (3.4), we have

$$\begin{aligned} |f(x)| &\leqslant e^{x^2/2} \int_{|x|}^\infty \big(\|\varphi'\|_\infty \cdot t + \|\varphi'\|_\infty \cdot \sqrt{\frac{2}{\pi}}\big) e^{-t^2/2} dt \\ &= \|\varphi'\|_\infty \cdot e^{x^2/2} \int_{|x|}^\infty t e^{-t^2/2} dt + \sqrt{\frac{2}{\pi}} \|\varphi'\|_\infty \cdot e^{x^2/2} \int_{|x|}^\infty e^{-t^2/2} dt \\ &= \|\varphi'\|_\infty + \sqrt{\frac{2}{\pi}} \|\varphi'\|_\infty \cdot e^{x^2/2} \int_{|x|}^\infty e^{-t^2/2} dt. \end{aligned}$$

According to Lemma 3.1, we see that

$$|f(x)| \leqslant 2\|\varphi'\|_\infty. \tag{3.5}$$

If $x \geqslant 0$, we use the alternative expression for f given by

$$f(x) = -e^{x^2/2} \int_x^\infty \tilde{\varphi}(t) e^{-t^2/2} dt, \tag{3.6}$$

which follows from the observation that

$$\int_{-\infty}^\infty \tilde{\varphi}(t) e^{-t^2/2} dt = 0.$$

The same argument applied to (3.6) gives the same estimate as (3.5) in this case. Therefore, we conclude that

$$\|f\|_\infty \leqslant 2\|\varphi'\|_\infty.$$

(ii) Estimating f'.

First note that, since φ is differentiable, by differentiating the ODE (3.2) we have

$$f''(x) - xf'(x) = f(x) + \varphi'(x). \tag{3.7}$$

Inspired by the previous argument, in order to estimate f', we may wish to express f' as the product of $e^{x^2/2}$ and another function (an $\int_{-\infty}^{x}$-integral), just like the case for f. For this purpose, we compute

$$\begin{aligned}\frac{d}{dx}\big(e^{-x^2/2}f'(x)\big) &= e^{-x^2/2}f''(x) - xe^{-x^2/2}f'(x) \\ &= e^{-x^2/2}\big(f(x)+\varphi'(x)\big).\end{aligned}$$

Therefore,

$$f'(x) = e^{x^2/2}\cdot\int_{-\infty}^{x}\big(f(t)+\varphi'(t)\big)e^{-t^2/2}dt. \tag{3.8}$$

Similar to Part (i), we first consider $x \leqslant 0$. In this case, using the estimate on f we just obtained as well as Lemma 3.1, we have

$$|f'(x)| \leqslant 3\|\varphi'\|_\infty e^{x^2/2}\int_{|x|}^{\infty} e^{-t^2/2}dt \leqslant 3\sqrt{\frac{\pi}{2}}\|\varphi'\|_\infty. \tag{3.9}$$

If $x \geqslant 0$, we resort to the alternative expression that

$$f'(x) = -e^{x^2/2}\int_{x}^{\infty}\big(f(t)+\varphi'(t)\big)e^{-t^2/2}dt. \tag{3.10}$$

This is legal since

$$\int_{-\infty}^{\infty}\big(f(t)+\varphi'(t)\big)e^{-t^2/2}dt = \int_{-\infty}^{\infty}\big(f''(t)-tf'(t)\big)e^{-t^2/2}dt = 0,$$

where the second equality is a simple consequence of integration by parts. The same argument applied to (3.10) again gives (3.9) in this case. Therefore, we have

$$\|f'\|_\infty \leqslant 3\sqrt{\frac{\pi}{2}}\|\varphi'\|_\infty.$$

(iii) Estimating f''.

According to the equation (3.7) for f'' and the expression (3.8) for f', we have

$$f''(x) = xe^{x^2/2} \int_{-\infty}^{x} \big(f(t) + \varphi'(t)\big) e^{-t^2/2} dt + \big(f(x) + \varphi'(x)\big).$$

We have already got all the needed ingredients to estimate the above terms. To be precise, again by considering the cases $x \leqslant 0$ and $x \geqslant 0$ separately, we have

$$\begin{aligned}|f''(x)| &\leqslant \big(\|f\|_\infty + \|\varphi'\|_\infty\big) \cdot |x| e^{x^2/2} \int_{|x|}^{\infty} e^{-t^2/2} dt + \big(\|f\|_\infty + \|\varphi'\|_\infty\big) \\ &\leqslant 2\big(\|f\|_\infty + \|\varphi'\|_\infty\big) \\ &\leqslant 6\|\varphi'\|_\infty,\end{aligned}$$

where we have used Lemma 3.1 and the estimate on f obtained in Part (i).

Now the proof of the Proposition is complete. □

4 Step 3: Establishing the L^1-Berry-Esseen estimate

The previous two steps on Stein's method are entirely general. To develop the last step, we restrict ourselves to the independent case. In its complete form, the main theorem is stated as follows.

Theorem 4.1. *Let $\{X_n : n \geqslant 1\}$ be a sequence of independent random variables, each having mean zero and finite third moment. For each n, set*

$$\Sigma_n \triangleq \sqrt{\mathrm{Var}[S_n]},\ \tau_n \triangleq \big(\mathbb{E}[|X_n|^3]\big)^{1/3},\ \hat{S}_n \triangleq \frac{S_n}{\Sigma_n},$$

where $S_n \triangleq X_1 + \cdots + X_n$. Then for any continuously differentiable function $\varphi : \mathbb{R} \to \mathbb{R}$ with bounded derivative, we have

$$\big|\mathbb{E}[\varphi(\hat{S}_n)] - \mathbb{E}[\varphi(Z)]\big| \leqslant 9\|\varphi'\|_\infty \cdot \frac{\sum_{m=1}^{n} \tau_m^3}{\Sigma_n^3}$$

for every $n \geqslant 1$, where $Z \overset{d}{=} N(0,1)$. In particular, if $\{X_n : n \geqslant 1\}$ is an i.i.d. sequence with mean zero, unit variance and $\tau \triangleq \big(\mathbb{E}[|X_1|^3]\big)^{1/3} < \infty$, we have

$$\big|\mathbb{E}[\varphi(\hat{S}_n)] - \mathbb{E}[\varphi(Z)]\big| \leqslant 9\|\varphi'\|_\infty \cdot \frac{\tau^3}{\sqrt{n}}.$$

Proof. Let f be the unique bounded solution to Stein's equation (3.2) corresponding to φ. Then

$$\mathbb{E}[\varphi\big(\hat{S}_n\big)] - \mathbb{E}[\varphi(Z)] = \mathbb{E}[f'(\hat{S}_n)] - \mathbb{E}[\hat{S}_n f(\hat{S}_n)].$$

As the last step in Stein's method, our goal is to estimate the right hand side of the above equation. For this purpose, we first introduce the following notation:

$$\hat{X}_m \triangleq \frac{X_m}{\Sigma_n},\ \hat{\sigma}_m \triangleq \frac{\sigma_m}{\Sigma_n}, \quad 1 \leqslant m \leqslant n.$$

Note that $\hat{\sigma}_m^2 = \mathbb{E}[\hat{X}_m^2]$, and

$$\sum_{m=1}^n \hat{X}_m = \hat{S}_n,\ \sum_{m=1}^n \hat{\sigma}_m^2 = 1.$$

We can now rewrite

$$\mathbb{E}[f'(\hat{S}_n)] - \mathbb{E}[\hat{S}_n f(\hat{S}_n)] = \sum_{m=1}^n \mathbb{E}[\hat{\sigma}_m^2 f'(\hat{S}_n)] - \sum_{m=1}^n \mathbb{E}[\hat{X}_m f(\hat{S}_n)].$$

The next crucial point is to relate $f(\hat{S}_n)$ with $f(\hat{S}_n - \hat{X}_m)$ through f' (this is beneficial since $\hat{S}_n - \hat{X}_m$ and $\hat{X}_m$ are independent). To this end, recall from calculus that

$$f(y) = f(x) + \int_0^1 f'\big((1-t)x + ty\big) \cdot (y-x)dt.$$

Taking $x = \hat{S}_n - \hat{X}_m$ and $y = \hat{S}_n$, we can write

$$f(\hat{S}_n) = f(\hat{S}_n - \hat{X}_m) + \int_0^1 f'(T_{n,m}(t))\hat{X}_m dt,$$

where to simplify notation we have set

$$T_{n,m}(t) \triangleq (1-t)(\hat{S}_n - \hat{X}_m) + t\hat{S}_n.$$

It follows that

$$\begin{aligned}\mathbb{E}[\hat{X}_m f(\hat{S}_n)] &= \mathbb{E}[\hat{X}_m f(\hat{S}_n - \hat{X}_m)] + \mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 f'(T_{n,m}(t))dt\big] \\ &= \mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 f'(T_{n,m}(t))dt\big].\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[f'(\hat{S}_n)] - \mathbb{E}[\hat{S}_n f(\hat{S}_n)] \\
&= \sum_{m=1}^{n} \mathbb{E}[\hat{\sigma}_m^2 f'(\hat{S}_n)] - \sum_{m=1}^{n} \mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 f'(T_{n,m}(t))dt\big] \\
&= \sum_{m=1}^{n} \mathbb{E}\big[\hat{\sigma}_m^2 \cdot \big(f'(\hat{S}_n) - f'(T_{n,m}(0))\big)\big] \\
&\quad - \sum_{m=1}^{n} \mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 \big(f'(T_{n,m}(t)) - f'(T_{n,m}(0))\big)dt\big], \qquad (4.1)
\end{aligned}$$

where to reach the last equality we have used the observation that $T_{n,m}(0) = \hat{S}_n - \hat{X}_m$ and thus

$$\mathbb{E}[\hat{\sigma}_m^2 f'(T_{n,m}(0))] = \mathbb{E}[\hat{X}_m^2 f'(T_{n,m}(0))].$$

To estimate the first summation on the right hand side of (4.1), we use

$$\big|f'(\hat{S}_n) - f'(T_{n,m}(0))\big| \leqslant \|f''\|_\infty \cdot |\hat{X}_m|.$$

This gives

$$\big|\mathbb{E}\big[\hat{\sigma}_m^2 \cdot \big(f'(\hat{S}_n) - f'(T_{n,m}(0))\big)\big]\big| \leqslant \|f''\|_\infty \hat{\sigma}_m^2 \cdot \mathbb{E}[|\hat{X}_m|] \leqslant \|f''\|_\infty \cdot \frac{\tau_m^3}{\Sigma_n^3},$$

where we have used the fact that $[1, \infty) \ni p \mapsto (\mathbb{E}[|X|^p])^{1/p}$ is an increasing function (as seen from Hölder's inequality), and in particular,

$$\mathbb{E}[|X_m|] \leqslant \tau_m, \ \sigma_m \leqslant \tau_m.$$

To estimate the second summation on the right hand side of (4.1), note that

$$\big|f'(T_{n,m}(t)) - f'(T_{n,m}(0))\big| \leqslant t\|f''\|_\infty \cdot |\hat{X}_m|.$$

This gives

$$\big|\mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 \big(f'(T_{n,m}(t)) - f'(T_{n,m}(0))\big)dt\big]\big| \leqslant \frac{1}{2}\|f''\|_\infty \cdot \frac{\tau_m^3}{\Sigma_n^3}.$$

Finally, using the estimate $\|f''\|_\infty \leqslant 6\|\varphi'\|_\infty$ given by Proposition 3.1, we arrive at

$$\big|\mathbb{E}[f'(\hat{S}_n)] - \mathbb{E}[\hat{S}_n f(\hat{S}_n)]\big| \leqslant 9\|\varphi'\|_\infty \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3},$$

which is the desired estimate. □

We have mentioned earlier that we wish to obtain the L^1-Berry-Esseen estimate, i.e. the L^1-distance between the distribution functions of $\hat{S}_n$ and $N(0,1)$. We give a heuristic argument to see how Theorem 4.1 gives rise to such an L^1-estimate. Making this argument rigorous is not a trivial task, which will be deferred to the appendix.

Let F_n be the distribution function of $\hat{S}_n$, and let Φ be the distribution function of $N(0,1)$. First of all, a naive integration by parts gives

$$\int_{\mathbb{R}} \varphi(x)dF_n(x) - \int_{\mathbb{R}} \varphi(x)d\Phi(x) = \int_{\mathbb{R}} \varphi'(x)(\Phi(x) - F_n(x))dx. \tag{4.2}$$

Therefore, Theorem 4.1 tells us that

$$\Big| \int_{\mathbb{R}} \varphi'(x)\big(\Phi(x) - F_n(x)\big)\Big| \leqslant C_n \cdot \|\varphi'\|_\infty \tag{4.3}$$

for any φ with $\varphi' \in \mathcal{C}_b(\mathbb{R})$, where

$$C_n \triangleq 9 \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3}$$

is the constant giving the rate of convergence. Of course symbolically we can just write $\psi \triangleq \varphi'$ to indicate that

$$\Big| \int_{\mathbb{R}} \psi(x)(F_n(x) - \Phi(x))dx\Big| \leqslant C_n \|\psi\|_\infty$$

is true *for any bounded continuous function* ψ. From this point, it is not too surprising to expect that,

$$\|F_n - \Phi\|_{L^1} = \int_{\mathbb{R}} |F_n(x) - \Phi(x)| \leqslant C_n.$$

In fact, if we were allowed to choose $\psi(x) = \mathrm{sgn}(F_n(x) - G(x))$ where $\mathrm{sgn}(x)$ is the function defined by

$$\mathrm{sgn}(x) \triangleq \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ 0, & x = 0, \end{cases}$$

then we have $\|\psi\|_\infty \leqslant 1$ and

$$\int_{\mathbb{R}} \psi(x)(F_n(x) - \Phi(x))dx = \int_{\mathbb{R}} |F_n(x) - \Phi(x)|dx,$$

yielding the desired L^1-estimate. The main difficulty here is that $\psi(x)$ is not a continuous function. Getting around this difficulty requires some analysis.

To summarise, the L^1-Berry-Esseen estimate is given by the following result.

Corollary 4.1. *Under the same set-up as in Theorem 4.1, we have*

$$\int_{-\infty}^{\infty} |F_n(x) - \Phi(x)| dx \leqslant 9 \frac{\sum_{m=1}^{n} \tau_m^3}{\Sigma_n^3}. \tag{4.4}$$

In particular, in the i.i.d. context we have

$$\int_{-\infty}^{\infty} |F_n(x) - \Phi(x)| dx \leqslant 9 \frac{\tau^3}{\sqrt{n}}.$$

5 Some further remarks and scopes

We conclude this topic and the subject by giving a few further comments on the development of Stein's method.

(i) Let us take a second look at our previous heuristic argument about obtaining the L^1-Berry-Esseen estimate. The fact that the right hand side of (4.3) is the uniform norm of φ' leads us to the L^1-estimate for the distribution functions. Through a naive duality viewpoint, if we were able to replace the right hand side of (4.3) by the L^1-norm of φ' ($\|\varphi'\|_{L^1} \triangleq \int_{\mathbb{R}} |\varphi'(x)| dx$), we should be able to deduce the uniform Berry-Esseen estimate. This requires strengthening the analysis of Stein's equation (cf. Proposition 3.1) to estimating the uniform norms of f, f', f'' in terms of $\|\varphi'\|_{L^1}$. It can be done for f, f' but not for f''! This is what makes the uniform Berry-Esseen estimate much harder than the L^1-estimate. The result is stated as follows.

Theorem 5.1 (The Uniform Berry-Esseen Estimate)**.** *Under the same notation as in Theorem 4.1 and Corollary 4.1, we have*

$$\|F_n - \Phi\|_\infty \leqslant 10 \cdot \frac{\sum_{m=1}^{n} \tau_m^3}{\Sigma_n^3}.$$

In particular, in the i.i.d. context we have

$$\|F_n - \Phi\|_\infty \leqslant 10 \cdot \frac{\tau^3}{\sqrt{n}}.$$

A proof that follows the current line of argument is contained in Reference [3].

(ii) As we have mentioned earlier, Step 3 in Stein's method can be developed in the more general context of dependent random variables. In addition, the philosophy of this method is robust enough to be applied to other types of distributional approximations. One important topic is about Poisson approximations. Reference [1] contains the discussion of such extensions.

(iii) There are extensions of the one-dimensional theory we developed here to multivariate Gaussian approximations. A natural way of performing the analysis in the multidimensional context is to combine modern tools from Gaussian analysis. Reference [2] contains a nice modern introduction to this topic.

(iv) There is a modern viewpoint of Stein's method, known as the *generator approach*, that leads to more profound applications such as distributional approximations for stochastic processes. Suppose that μ is the target distribution that we wish to approximate. μ can be defined on $\mathbb{R}$, $\mathbb{R}^n$ or even an infinite dimensional space S (e.g. the space of continuous paths if we are in the context of stochastic process approximations). As the first step in Stein's method, we need to identify the Stein operator, say $\mathcal{A}$, which is an operator acting on a space of functions on S, so that the property

$$\mathbb{E}[\mathcal{A}f(Z)] = 0 \quad \forall f$$

uniquely characterises the distribution μ. The key idea behind the generator approach is to regard μ as the invariant measure of some S-valued Markov process. The Stein operator $\mathcal{A}$ will then be given by the generator of this Markov process, and the associated Stein's equation can be studied through the structure of this Markov process. Reference [1] contains an introduction to this approach.

References

[1] A.D. Barbour and L.H. Chen, *An introduction to Stein's method*, World Scientific, 2005.

[2] I. Nourdin and G. Peccati, *Normal approximations with Malliavin calculus*, Cambridge University Press, 2012.

[3] D.W. Stroock, *Probability theory-an analytic view*, Second Edition, Cambridge University Press, 2011.

6 Appendix: A functional analytic lemma for obtaining the L^1-Berry-Esseen estimate

We now provide the precise details which allow us to obtain the L^1-Berry-Esseen estimate (4.4) from Theorem 4.1. As the first main ingredient, we prove (4.2) rigorously.

Lemma 6.1. *Let $F, G : \mathbb{R} \to \mathbb{R}$ be distribution functions with finite first moment, i.e. $\int_{\mathbb{R}} |x| dF(x)$ and $\int_{\mathbb{R}} |x| dG(x)$ are both finite. Let ψ be a bounded Borel-measurable function and define $\varphi(x) \triangleq \int_0^x \psi(u) du$. Then*

$$\int_{\mathbb{R}} \varphi(x) dF(x) - \int_{\mathbb{R}} \varphi(x) dG(x) = \int_{\mathbb{R}} \psi(x) \big(G(x) - F(x) \big) dx. \tag{6.1}$$

Proof. We can write

$$\begin{aligned}
&\int_{\mathbb{R}} \varphi(x) dF(x) \\
&= \int_{\mathbb{R}} \big(\int_0^x \psi(u) du \big) dF(x) \\
&= -\int_{-\infty}^0 \int_x^0 \psi(u) du dF(x) + \int_0^\infty \int_0^x \psi(u) du dF(x) \\
&= -\int_{-\infty}^0 \psi(u) F(u) du + \int_0^\infty \psi(u)(1 - F(u)) du,
\end{aligned}$$

where in the last equality we have used Fubini's theorem to exchange the order of integration. Similarly,

$$\int_{\mathbb{R}} \varphi(x) dG(x) = -\int_{-\infty}^0 \psi(u) G(u) du + \int_0^\infty \psi(u)(1 - G(u)) du.$$

By subtracting the two results, we obtain (6.1). □

The other main ingredient is to see why restricting to the class $\mathcal{C}_b(\mathbb{R})$ of test functions will allow us to recover the L^1-norm. This is the content of the following lemma.

Lemma 6.2. *Let $Q : \mathbb{R} \to \mathbb{R}$ be a function which contains at most countably many discontinuity points and suppose that $\int_{\mathbb{R}} |Q(x)| dx < \infty$. Then*

$$\int_{\mathbb{R}} |Q(x)| dx = \sup \big\{ \big| \int_{\mathbb{R}} \varphi(x) Q(x) dx \big| : \varphi \in C_b(\mathbb{R}), \|\varphi\|_\infty \leqslant 1 \big\}. \tag{6.2}$$

Proof. For any φ with $\|\varphi\|_\infty \leqslant 1$, we have

$$\big|\int_{\mathbb{R}} \varphi(x)Q(x)dx\big| \leqslant \|\varphi\|_\infty \cdot \int_{\mathbb{R}} |Q(x)|dx \leqslant \int_{\mathbb{R}} |Q(x)|dx.$$

Therefore, the right hand side of (6.2) is not greater than the left hand side. To prove the other direction, first note that

$$\int_{\mathbb{R}} |Q(x)|dx = \int_{\mathbb{R}} \mathrm{sgn}(Q(x)) \cdot Q(x)dx,$$

where $\mathrm{sgn}(x)$ is the function defined by

$$\mathrm{sgn}(x) \triangleq \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ 0, & x = 0. \end{cases}$$

We set $\psi(x) \triangleq \mathrm{sgn}(Q(x))$. The main difficulty here is that $\psi(x)$ is not a continuous function, and thus we need to construct $C_b(\mathbb{R})$-approximations.

For this purpose, for each $\varepsilon > 0$, let us choose a continuous function $\rho_\varepsilon : \mathbb{R} \to \mathbb{R}$ such that

$$\rho_\varepsilon \geqslant 0, \quad \int_{\mathbb{R}} \rho_\varepsilon(x)dx = 1$$

and $\rho_\varepsilon(x) = 0$ for any $|x| > \varepsilon$. Define ψ_ε to be the convolution of ψ and ρ_ε, i.e.

$$\psi_\varepsilon(x) \triangleq \int_{\mathbb{R}} \psi(x-y)\rho_\varepsilon(y)dy = \int_{\mathbb{R}} \rho_\varepsilon(x-y)\psi(y)dy. \tag{6.3}$$

Using the latter expression, one can check that ψ_ε is continuous. Since $|\psi| \leqslant 1$, we also know that

$$|\psi_\varepsilon(x)| \leqslant \int_{\mathbb{R}} |\psi(x-y)| \cdot \rho_\varepsilon(y)dy \leqslant \int_{\mathbb{R}} \rho_\varepsilon(y)dy = 1.$$

Therefore, $\psi_\varepsilon \in \mathcal{C}_b(\mathbb{R})$. It may not be true that $\psi_\varepsilon(x) \to \psi(x)$ for every $x \in \mathbb{R}$ as $\varepsilon \to 0$, however, we claim that

$$\lim_{\varepsilon \to 0} \int_{\mathbb{R}} \psi_\varepsilon(x)Q(x)dx = \int_{\mathbb{R}} \psi(x)Q(x)dx. \tag{6.4}$$

If we can prove this, it is then immediate that the left hand side of (6.2) is not greater than the right hand side, and the proof of (6.2) will be finished.

To prove (6.4), let $\mathcal{C}_Q$ be the set of continuity points of Q. The crucial observation is that,

$$\psi_\varepsilon(x)\mathbf{1}_{\{x:Q(x)\neq 0\}\cap\mathcal{C}_Q}(x) \to \psi(x)\mathbf{1}_{\{x:Q(x)\neq 0\}\cap\mathcal{C}_Q}(x) \tag{6.5}$$

as $\varepsilon \to 0$. Indeed, if x is a continuity point of Q and $Q(x) \neq 0$, we know by continuity that $Q(x)$ does not change sign in a small neighbourhood of x. Suppose that $Q(x) > 0$ (so that $\psi(x) = 1$). Then there exists $\delta > 0$ such that $Q(x-y) > 0$ for any $y \in (-\delta, \delta)$. In particular,

$$\psi(x-y) = \operatorname{sgn}(Q(x-y)) = 1, \quad y \in (-\delta, \delta).$$

According to the constructions of ρ_ε and ψ_ε, for any $\varepsilon < \delta$ we have

$$\psi_\varepsilon(x) = \int_{\mathbb{R}} \psi(x-y)\rho_\varepsilon(y)dy = \int_{(-\varepsilon,\varepsilon)} 1 \cdot \rho_\varepsilon(y)dy = 1 = \psi(x),$$

which trivially implies that $\psi_\varepsilon(x) \to \psi(x)$ as $\varepsilon \to 0$. Therefore, (6.5) holds. The dominated convergence theorem then implies that

$$\int_{\{x:Q(x)\neq 0\}\cap\mathcal{C}_Q} \psi_\varepsilon(x)Q(x)dx \to \int_{\{x:Q(x)\neq 0\}\cap\mathcal{C}_Q} \psi(x)Q(x)dx.$$

On the other hand, since $\mathcal{C}_Q^c$ is at most countable (and thus has zero Lebesgue measure), we know that

$$\begin{aligned}\int_{\mathbb{R}} \psi_\varepsilon(x)Q(x)dx &= \int_{\{x:Q(x)\neq 0\}} \psi_\varepsilon(x)Q(x)dx \\ &= \int_{\{x:Q(x)\neq 0\}\cap\mathcal{C}_Q} \psi_\varepsilon(x)Q(x)dx.\end{aligned}$$

The same property holds for $\psi(x)$. Therefore, we conclude that (6.4) holds. □

The above two lemmas enable us to make our previous heuristic argument of obtaining the L^1-Berry-Esseen estimate from Theorem 4.1 rigorous.

Proof of Corollary 4.1. In Theorem 4.1, we have shown that

$$\Big|\int_{\mathbb{R}} \varphi(x)dF_n(x) - \int_{\mathbb{R}} \varphi(x)d\Phi(x)\Big| \leqslant 9\|\varphi'\|_\infty \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3}$$

for any φ with $\varphi' \in \mathcal{C}_b(\mathbb{R})$. Using Lemma 6.1 and setting $\psi \triangleq \varphi'$, we conclude that

$$\Big| \int_{\mathbb{R}} \psi(x)\big(F_n(x) - \Phi(x)\big)dx\Big| \leqslant 9\|\psi\|_\infty \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3}$$

for any $\psi \in \mathcal{C}_b(\mathbb{R})$. The L^1-estimate (4.4) then follows from Lemma 6.2 with $Q \triangleq F_n - \Phi$.

□