1 Probability space and random variables

As graduate level, we inevitably need to study probability based on measure theory. It obscures some intuitions in probability, but it also supplements our intuition, and in the end hopefully it will be our new intuition.

Since measure theory by its own is a part of analysis, but not probability, we do not give proofs to measure theoretic results, and use the concepts without explanation if they are contained in standard textbooks, for example, the *Real and Complex Analysis* by W. Rudin. All the proofs of not so standard measure theoretic theorems are in our text book *Probability: Theory and Examples* by R. Durrett, unless otherwise stated.

First we review the definition of a *probability space*, which appears in undergraduate textbooks (like the *Probability and Random Processes* by G. Grimmett and D. Stirzaker), without rigorous reference to measure space.

First, the set of all possible outcomes of an experiment (not a mathematical term, but this is where the aximatic probability theory starts) is denoted by Ω. It can be very small like $\{\text{head}, \text{tail}\}$, so that no advanced measure theory is needed, while it can also be quite big like $\{\text{all Brownian motion paths}\}$, so that you would be lost without the guide of measure theory.

Some subsets of Ω are called *events*. Note that not all subsets are events, especially if Ω is quite large. There are practical reasons for that (it is impossible to single out the outcome of an experiment exactly to be $1/2 = 0.50000000\dots$ centimetre). But for us, it is due to the requirement of mathematical consistence, as we will see later.

We call the set of events $\mathcal{F}$, and require that it satisfies

- $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.
- If $A \in \mathcal{F}$, then the complement $A^{\mathrm{c}} \in \mathcal{F}$.
- If $A_1, A_2, \dots, A_n, \dots \in F$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

In measure theoretic language, it is equivalent to say that $\mathcal{F}$ is a *σ-algebra* on Ω.

To define a probability space, we need to introduce the concept of probability for each event. Let P be a function from $\mathcal{F}$ to $[0, 1]$, that satisfies

- $P(\emptyset) = 0$ and $P(\Omega) = 1$.
- $P(A^c) = 1 - P(A)$,
- If $A_1, A_2, \dots, A_n, \dots \in \mathcal{F}$ are disjoint to one another, then $P\left(\bigcup_{n=1}^{\infty} A_i\right) = \sum_{n=1}^{\infty} P(A_n)$.

The last condition is not very intuitive, and it is called the *countably additive* property of P.

Suppose $\Omega, \mathcal{F}, P$ are defined as above, we call the triple $(\Omega, \mathcal{F}, P)$ a probability space. In measure theoretic language, it is nothing but a positive measure space with total measure 1. (A measure space is a triple (X, Σ, μ), where X is a set, Σ is a σ-algebra of the subsets of X, and μ is a function from Σ to $\mathbb{R} \cup \{\pm\infty\}$, such that $\mu(\emptyset) = 0$ and for pairwise disjoint sets $E_1, \dots, E_n, \dots \in \Sigma$, $\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n)$.)

We briefly discuss the idea that a σ-algebra $\mathcal{S}$ on X is *generated* by a collection of subsets S_α of Ω. $\mathcal{S}$ is defined as the smallest σ-algebra that contains all S_α. This definition is not constructive, and the construction of $\mathcal{S}$ is not easy unless the collection of S_α is finite. If we start from the collection of open sets (assuming that Ω has a topological structure so that we can talk about the open sets there), then the generated σ-algebra is called the *Borel σ-algebra*, consisting of the *Borel sets*. We mostly encounter the Borel sets on the real line, where the open sets are unions of open intervals.

Next we define random variables on a probability space $\Omega = (\Omega, \mathcal{F}, P)$.

Definition 1. A random variable X on $(\Omega, \mathcal{F}, P)$ is a mapping $\Omega \to \mathbb{R}$ such that for each Borel set B on $\mathbb{R}$, $X^{-1}(B) \in \mathcal{F}$.

It is not hard to see (exercise) that $\mathcal{B}$ is also generated by the sets $(-\infty, x]$ where $x \in \mathbb{R}$. So a more practical definition of a random variable is

Definition 2. A random variable X on $(\Omega, \mathcal{F}, P)$ is a mapping $\Omega \to \mathbb{R}$ such that for each semi-closed set $(-\infty, x]$, $X^{-1}(-\infty, x] \in \mathcal{F}$. Then the function $F(x) = P(X^{-1}(-\infty, x])$ is a function from $\mathbb{R}$ to $[0, 1]$, and it is called the *distribution function* of X.

It is clear that for any random variable X, the distribution function F is non-decreasing, because for $a < b$,

$$F(b) - F(a) = P(X^{-1}(-\infty, b]) - P(X^{-1}(-\infty, a]) = P(X^{-1}(a, b]) \geq 0.$$

Another simple property satisfied by a distribution function is $F(\infty) = \lim_{x\to\infty} F(x) = 1$ and $F(-\infty) = \lim_{x\to-\infty} F(x) = 0$. F may not be a continuous function, but we can show that it is *right-continuous*, that is, $\lim_{x\downarrow a} F(x) = F(a)$. This is because of the countably additive property of the measure. One consequence of the countable additivity is that of $A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots$ and $\bigcap_{n=1}^\infty A_n = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$. (Exercise.) So if $x_1, x_2, \ldots$ is a decreasing sequence whose limit is a, then the sequence $X^{-1}(a, x_n]$ are nested sets whose common intersection is $\emptyset$, so

$$\lim_{n\to\infty} F(x_n) - F(a) = \lim_{n\to\infty} P(X^{-1}(a, x_n]) = 0.$$

Thus we prove the right-continuity of $F(x)$. Actually the properties above characterize distribution functions.

Theorem 1. *If a function $F : \mathbb{R} \to [0, 1]$ is non-decreasing, right-continuous, and $F(\infty) = 1$, $F(-\infty) = 0$, then it is a distribution function for a random variable.*

To prove this theorem, we need a technical result in measure theory, and we need to introduce some concepts.

We say a collection of subsets $\mathcal{A}$ of Ω an *algebra* if $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$ and $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$. It is obvious that a σ-algebra is an algebra, but not vice versa. Then let $\mu : \mathcal{A} \to [0, \infty)$ be a mapping. We say μ is a *measure* on $\mathcal{A}$ if it satisfies

1. (finitely additive) $\mu(\emptyset) = 0$, and for $A_1, \ldots, A_n \in \mathcal{A}$, $\mu(A_1 \cup \cdots \cup A_n) = \mu(A_1) + \cdots + \mu(A_n)$.
2. (countably additive) For countably disjoint $A_1, A_2, \ldots \in \mathcal{A}$, if $\bigcup_{n=1}^\infty A_n \in \mathcal{A}$, then $\mu\left(\bigcup_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty \mu(A_n)$.

We say a measure μ on an algebra $\mathcal{A}$ is *σ-finite* if there is a sequence of sets $A_n \in \mathcal{A}$ such that $\mu(A_n) < \infty$ for all n and $\bigcup_{n=1}^{\infty} = \Omega$, the whole set of the space. Then we have

Theorem 2 (Carathéodory extension)**.** *Let μ be a σ-finite measure on an algebra $\mathcal{A}$. Then μ has a unique extension to the σ-algebra generated by $\mathcal{A}$.*

Proof of Theorem 1. First we construct a measure space $(\Omega, \mathcal{F}, P)$ with $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B} = \{\text{Borel sets on } \mathbb{R}\}$, and P satisfies that for all $a < b$, $P(a, b] = F(b) - F(a)$. Then we define a random variable X on this probability space such that $X(x) = x$. It is clear that X is a well-defined random variable, and its distribution function is $F(x)$.

To justify our construction of the measure space, we need the Carathéodory extension theorem. It is clear that the collection $\mathcal{A}$ of subsets of $\mathbb{R}$ in the form of $(a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_k, b_k]$ where $a_1 < b_1 < a_2 < b_2 < \cdots < a_k < b_k$ is an algebra, and the function P defined by

$$P\Big((a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_k, b_k]\Big) = F(b_k) - F(a_k) + \ldots + F(b_2) - F(a_2) + F(b_1) - F(a_1)$$

satisfies the finitely additive condition for a measure on $\mathcal{A}$. Since as an exercise we know that $\mathcal{A}$ generates the σ-algebra $\mathcal{B}$ of Borel sets on $\mathbb{R}$, and it is also an easy exercise to show that P satisfy the σ-finite condition, we can apply the Carathéodory extension theorem to show that P is a well-defined measure on $\mathcal{B}$ as long as we show that P is countably additive, and then it is clear that P is a probability measure.

Suppose $A_1, A_2, \ldots \in \mathcal{A}$ are disjoint to each other and $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$. It is not hard to see that since P is a non-negative function,

$$\sum_{n=1}^{\infty} P(A_n) \leq P\left(\bigcup_{n=1}^{\infty} A_n\right).$$

Without loss of generality, we assume that $\bigcup_{n=1}^{\infty} A_n = (a, b]$, and it suffices to show that for any $\epsilon > 0$, there is an N such that

$$\sum_{n=1}^{N} P(A_n) > F(b) - F(a) - \epsilon.$$

By the right-continuous property of F, there is $a' > a$ such that $F(a') - F(a) < \epsilon/2$. Furthermore, for each $A_n = (a_1^{(n)}, b_1^{(n)}] \cup \cdots \cup (a_{k_n}^{(n)}, b_{k_n}^{(n)}]$, we can choose an open set $B'_n = (a_1^{(n)}, b_1^{(n),'}) \cup \cdots \cup (a_{k_n}^{(n)}, b_{k_n}^{(n),'})$ and $B_n = (a_1^{(n)}, b_1^{(n),'}] \cup \cdots \cup (a_{k_n}^{(n)}, b_{k_n}^{(n),'}] \in \mathcal{A}$, such that $b_i^{(n),'} > b_i^{(n)}$ for all $i = 1, \ldots, k_n$ and

$$\begin{aligned} P(B_n) &= F(b_{k_n}^{(n),'}) - F(a_{k_n}^{(n)}) + \ldots + F(b_1^{(n),'}) - F(a_1^{(n)}) \\ &< F(b_{k_n}^{(n)}) - F(a_{k_n}^{(n)}) + \ldots + F(b_1^{(n)}) - F(a_1^{(n)}) + \frac{\epsilon}{2^{2+i}} \\ &= P(A_n) + \frac{\epsilon}{2^{2+i}}. \end{aligned}$$

Since $B'_n \supset A_n$ and $\{A_n\}$ covers $(a, b]$, we have that $\{B'_n\}$ covers $[a', b]$, and then by a compactness argument we have that a finite subset of $\{B'_n\}$, say $\{B'_1, \ldots, B'_N\}$ without loss of generality, covers $[a', b]$. Then $\{B_1, \dots, B_N\}$ covers $(a', b]$, and by the definition of P

$$P(B_1) + P(B_2) + \cdots + P(B_N) \geq F(b) - F(a'),$$

which implies that

$$P(A_1) + P(A_2) + \cdots + P(A_N) + \left(\frac{1}{2} - \frac{1}{2^{2+N}}\right)\epsilon \geq F(b) - F(a) - \frac{\epsilon}{2},$$

and we obtain the desired result. □

By the method of the proof, if we take $F(x) = x$, then we construct the *Lebesgue measure* on $\mathbb{R}$ where the measure of an interval is its length. Although it is not a probability measure, its importance is obvious. We denote it by λ, and when we write the integration with dx without specification, it is with respect to the Lebesgue measure.

We remark that if the distribution function $F(x)$ is differentiable almost everywhere and there is an integrable function $f(x)$, which is called the *density function*, such that $\int_{-\infty}^{x} f(t)dt = F(x)$, then the construction of the probability measure P is quite straightforward:

$$P(B) = \int_B f(x)dx, \quad \text{for all Borel set } B,$$

and usually we call it a continuous distribution. If $F(x)$ is a piecewise constant function with the change from 0 to 1 purely by jumps at countable points, then the construction of the probability measure P is also simple. For example, if

$$F(x) = \begin{cases} 0 & x < 0, \\ \frac{1}{2} & 0 \leq x < 1, \\ 1 & x \geq 1, \end{cases}$$

then it defines the Bernoulli distribution on two values 0 and 1, and a random variable with this distribution attains either value with half probability. This is an example of discrete distribution, where 0 and 1 are called *point masses* or *atoms* of the probability measure. Note that there are more subtle cases, like the distribution function given by the Cantor set as follows. Recall that if we express the real numbers in $[0, 1]$ by ternary expansion, and keep all the real numbers that allow an ternary expansion with all digits 0 or 2, then we have the Cantor set. (For example, $1/3 = (0.1)_3$, but it is can also be written as $(0.02222\dots)_3$, so it is in the Cantor set, but $1/2$ can only be written as $(0.1111\dots)_3$, so it is not in the Cantor set.) Then for any real number in the Cantor set, we define for any number in the Cantor set

$$F\left(\frac{a_1}{3} + \frac{a_2}{9} + \frac{a_3}{27} + \cdots\right) = \frac{1}{2}\left(\frac{a_1}{2} + \frac{a_2}{4} + \frac{a_3}{8} + \cdots\right), \quad a_k = 0 \text{ or } 2,$$

for $x < 0$ define $F(x) = 0$, and for $x \geq 0$ not in the Cantor set

$$F(x) = \max_{t<x, \text{ and } t \text{ is in the Cantor set}} F(t).$$

Then it is not very hard to check that $F(x)$ is right-continuous and is a well defined distribution function. But it is not a continuous distribution since there is no well-defined density function whose integral is $F(x)$, and it is not a discrete distribution since there is no point mass where the distribution function has a jump.

By the Lebesgue decomposition theorem and Radom-Nikodym theorem, we do not need to consider distribution functions more exotic than the Cantor distribution. On the real line and the Borel sets, we call a σ-finite measure μ *absolutely continuous* to the Lebesgue measure, if there is a Lebesgue measurable function $f \geq 0$ such that $\mu(E) = \int_E f dx$ for all $E \in \mathcal{B}$. We say a measure ν is *singular* with respect to the Lebesgue measure, if there is a set $E \in \mathcal{B}$ such that $\nu(E) = 0$ while the Lebesgue measure of A^c is 0. In particular, we say a singular measure ν_1 is *atomic* if it is the sum of countable point masses: $\nu_1 = \sum c_n \delta_{a_n}$ where $a_n \in \mathbb{R}$ and $c_n \geq 0$ with $\sum c_n = 1$. We say a singular measure μ_2 is *singular continuous* with respect to the Lebesgue measure, if it has no point mass, that is, $\mu_2(\{a\}) = 0$ for all $a \in \mathbb{R}$. Then we have that any probability measure can be written as $\alpha\nu + \beta_1\nu_1 + \beta_2\nu_2$, where μ is absolutely continuous, ν_1 is atomic, and ν_2 is singular continuous (with respect to the Lebesgue measure) and $\alpha, \beta_1, \beta_2 \geq 0$ with $\alpha + \beta_1 + \beta_2 = 1$.

We finish the remark to Theorem 1 and its proof by noting that random variables defined on different probability spaces can have identical distribution. For example, the Bernoulli distribution can be realised on $\mathbb{R}$ with Borel sets and an atomic measure, and it can also simply be realised on the probability space $\Omega = \{0, 1\}$, with the σ-algebra $\{\emptyset, \{0\}, \{1\}, \Omega\}$, and the probability measure $P(0) = P(1) = 1/2$, by the random variable $X : \{0, 1\} \to \mathbb{R}$ such that $X(0) = 0$ and $X(1) = 1$. If two random variables, on the same probability space or not, are *equal in distribution*, we write

$$X \overset{d}{=} Y.$$

In our module, we consider the collective property of many random variables on the same probability space, especially the sum of many *independent* random variables. We say a set of random variables $\{X_\alpha\}$ on a probability space $(\Omega, \mathcal{F}, P)$ are independent, if for any finitely many of them, say $A_1, \dots, A_n$, and any Borel sets $B_1, \dots, B_n$,

$$P\left(\bigcap_{i=1}^{n} \{X_i \in B_i\}\right) = \prod_{i=1}^{n} P(X_i \in B_i),$$

where $\{X \in B\}$ means the measurable set $X^{-1}(B)$. The properties of independent random variables will be discussed later. Now we consider a theoretical question: Do there exist independent random variables with given distributions?

If we consider finitely many independent random variables, they can be constructed by the product of measure spaces.

Suppose $(\Omega_1, \mathcal{F}_1, P_1), \dots, (\Omega_n, \mathcal{F}_n, P_n)$ are probability spaces, such that $X_1, \dots, X_n$ are random variables on them respectively, with distribution functions $F_1(x), \dots, F_n(x)$ respectively. Then consider the product measure space $\Omega = \{(\omega_1, \dots, \omega_n)\} = \Omega_1 \times \cdots \times \Omega_n$ with the product σ-algebra $\mathcal{F}$ that is generated by $\{E_1 \times \cdots \times E_n\}$ where $E_i \in \mathcal{F}_i$, and the product measure P that is uniquely determined by

$$P(E_1 \times \cdots \times E_n) = P(E_1) \times \cdots \times P(E_n).$$

We define random variables $Y_1, \dots, Y_n$ on $(\Omega, \mathcal{F}, P)$ such that

$$Y_i(\omega_1, \dots, \omega_n) = X_i(\omega_i).$$

It is easy to check that the distribution function for Y_i is F_i, since, for example $i = 1$,

$$P(Y_1 \in (a,b]) = P(\{X_1 \in (a,b]\} \times \Omega_2 \times \cdots \times \Omega_n) = P_1(X_1 \in (a,b]) \times 1 \times \cdots \times 1$$
$$= F_1(b) - F_1(a),$$

and they are independent.

In later discussion, we often start with the phrase "Suppose $X_1, X_2, \ldots$ are a sequence of independent random variables ...". Is it possible to construct a probability space on which there are infinitely many independent random variables? The construction for the product of finitely many measure spaces cannot be naively used for infinite product. But in a special case, the construction is possible. To state the result, we define the set $\mathbb{R}^{\mathbb{N}}$ as

$$\mathbb{R}^{\mathbb{N}} = \{\omega = (\omega_1, \omega_2, \ldots) \mid \omega_i \in \mathbb{R}\},$$

and then define the σ-algbra $\mathcal{B}^{\mathbb{N}}$ that is generated by the so-called finite dimensional sets

$$\{(\omega_1, \omega_2, \ldots) \mid \text{there is } n \in \mathbb{N} \text{ and } B_1, \ldots, B_n \text{ are Borel sets on } \mathbb{R}$$
$$\text{such that } \omega_1 \in B_1, \ldots, \omega_n \in B_n, \text{ while } \omega_{n+1}, \omega_{n+2}, \ldots \text{ are arbitrary real numbers.}\}.$$

Note that $\mathcal{B}^{\mathbb{N}}$ is the Borel σ-algebra on $\mathbb{R}^{\mathbb{N}}$ with respect to the product topology on $\mathbb{R}^{\mathbb{N}}$. Then we have the result as follows.

Theorem 3 (Kolmogorov extension)**.** *Suppose* $(\mathbb{R}^n, \mathcal{B}^n, \mu_n)$ *are probability spaces, where* $\mathcal{B}^n$ *is the Borel* σ*-algebra on* $\mathbb{R}^n$*, and* μ_n *are consistent, that is,*

$$\mu_{n+1}((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R}) = \mu_n((a_1, b_1] \times \cdots \times (a_n, b_n]),$$

Then there is a unique probability measure P *on* $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$ *with*

$$P(\{\omega \mid \omega_1 \in (a_1, b_1], \ldots, \omega_n \in (a_n, b_n]\}) = \mu_n((a_1, b_1] \times \cdots \times (a_n, b_n]).$$

Suppose $(\mathbb{R}, \mathcal{B}, P_1), (\mathbb{R}, \mathcal{B}, P_2), \ldots$ are probability spaces, all defined on $\mathbb{R}$ with the σ-algebra consisting of the Borel sets. Then the product space of the first n of them is $(\mathbb{R}^n, \mathcal{B}^n, \mu_n)$ where μ_n is characterised by

$$\mu_n((a_1, b_1] \times \cdots \times (a_n, b_n]) = P_1(a_1, b_1) \times \cdots \times P_n(a_n, b_n].$$

It is clear that these measure spaces satisfy the consistency condition in the Kolmogorov extension theorem, so there exists a probability measure space $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, P)$ as constructed in the theorem. Now suppose X_n is a random variable on $(\mathbb{R}, \mathcal{B}, P_n)$ with distribution function F_n, then the random variable Y_n on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, P)$, defined by $Y_n(\omega) = X_n(\omega_n)$, is a random variable with distribution function F_n. It is not hard to check that $Y_1, Y_2, \ldots$ are independent.

As the conclusion of this lecture, we are pleased with ourselves that the phrase "Suppose $X_1, X_2, \ldots$ are a sequence of independent random variables ..." is meaningful, in the sense that no matter what distributions $F_1, F_2, \ldots,$ we can canstruct a probability space on which there are random variables $X_1, X_2, \ldots$ with the given distributions F_i, and they are independent.

2 Expectation and variance

(In this section and later, when we talk about a set of random variables, we assume that they are on the same probability space $(\Omega, \mathcal{F}, P)$, unless otherwise specified.)

For a random variable, the most important quantity is its *expectation*, also called *mean* or *average* in the everyday language, if it exists. Recall that a random variable X is a measurable function on a probability space $(\Omega, \mathcal{F}, P)$. The expectation of the random variable is defined by the integral of the function:

$$EX = \int X dP = \int_{\Omega} X(\omega) dP(\omega),$$

if the measurable function is also integrable. (In a more analytic language, $X(\omega)$ is an L^1 function on the measure space.) Not all measurable funcitons are integrable. If X is a non-negative random variable, its expectation either exists as a finite nonnegative number, or is $+\infty$. If X is not non-negative, then EX is well-defined as long as $E|X| < \infty$, otherwise EX may not be well-defined, even if we allow $\pm\infty$. Thus for the existence conditions involving expectation, we often consider the non-negative case and the general case separately.

The expectation satisfies some well known identities and inequalities for integrations:

Theorem 4. *Suppose the expectations for random variables X and Y exist. Then*

- *$E(X + Y) = EX + EY$,*
- *$E(aX + b) = aEX + b$, and*
- *if $X \geq Y$, that is, $X(\omega) \geq Y(\omega)$ for all $\omega \in \Omega$, then $EX \geq EY$.*

Theorem 5 (Hölder's inequality)**.** *Suppose $p, q > 0$ and $1/p + 1/q = 1$, and random variables X and Y are L^p-integrable and L^q-integrable respectively, that is, $E|X|^p$ and $E|Y|^q$ exist. Then $E(XY)$ exists and*

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

(The $p = q = 2$ special case of Hölder's theorem, the Cauchy-Schwarz theorem, is most useful.)

The following theorem is not in all real analysis textbooks, because it is valid only if the measure space is a probability space. But it is in Rudin's book and we omit the proof.

Theorem 6 (Jensen's inequality)**.** *Suppose function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, that is, for all $x < y \in \mathbb{R}$ and $a \in (0, 1)$, $a\varphi(x) + (1 - a)\varphi(y) \geq \varphi(ax + (1 - a)y)$. Then*

$$E(\varphi(X)) \geq \varphi(EX),$$

provided that both EX and $E(\varphi(X))$ exist.

The next theorem is not commonly seen in real analysis textbooks, so we include the proof. To sate the theorem, we first introduce a notation: For a random variable X and a measurable set $A \in \mathcal{F}$,

$$E(X; A) = \int_A X dP.$$

Theorem 7 (Chebyshev's inequality). *Suppose φ is a non-negative function on $\mathbb{R}$, and $B \in \mathcal{B}$ is a Borel set on $\mathbb{R}$, then*

$$\inf_{x\in B}(\varphi(x))P(X \in B) \leq E(\varphi(X); X \in B) \leq E(\varphi(X)).$$

Proof. The second inequality is a direct consequence of the non-negativity of φ:

$$E(\varphi(X)) - E(\varphi(X); X \in B) = \int_{\Omega\setminus X^{-1}(B)} \varphi(X)dP \geq 0.$$

For the first inequality, we note that for all ω such that $X(\omega) \in B$, $\varphi(\omega) \geq \inf_{x\in B}(\varphi(x))$, so

$$\begin{aligned} E(\varphi(X); X \in B) = \int_{X^{-1}(B)} \varphi(X(\omega))dP(\omega) &\geq \int_{X^{-1}(B)} \inf_{x\in B}(\varphi(x))dP(\omega) \\ = \inf_{x\in B}(\varphi(x)) \int_{X^{-1}(B)} 1dP &= \inf_{x\in B}(\varphi(x))P(X \in B). \end{aligned}$$

□

Since the expectation of a random variable is an integral, the convergence theorems we have learnt in real analysis can be used. We recall the most well known ones:

Lemma 8 (Fatou). *If $X_n \geq 0$, then*

$$\inf_{n\to\infty} EX_n \geq E\left(\liminf_{n\to\infty} X_n\right).$$

Theorem 9 (monotone convergence). *If $X_1, X_2, \ldots$ are non-negative random variables such that $X_n \uparrow X$, that is, $X_1(\omega) \leq X_2(\omega) \leq \cdots$ for all $\omega \in \Omega$, and $X_n \to X$ a.s., then $EX_n \uparrow EX$. (Here EX and EX_n are allowed to be $+\infty$.)*

Theorem 10 (dominated convergence). *If $X_n \to X$ a.s., $|X| \leq Y$ for all n and $EY < +\infty$, then EX_n and EX exist and $EX_n \to EX$.*

Theorem 11. *Suppose $X_n \to X$ a.s. Let g, h be continuous functions on $\mathbb{R}$ such that*

- *$g(x) \geq 0$ for all x and $g(x) > 0$ for large enough x,*
- *$|h(x)|/g(x) \to 0$ as $|x| \to 0$, and*
- *$E(g(X_n)) \leq K < \infty$ for all n.*

Then $E(h(X_n)) \to E(h(X))$.

Proof. We use the method of truncation, which we will use again several times in this module. Let M be a large enough real number, such that $g(x) > 0$ for all $|x| \geq M$, and M satisfies some other conditions to be specified later. For X_n and X, we denote (the random variable Y stands for either X_n or X)

$$Y^{(M)}(\omega) = \begin{cases} Y(\omega) & \text{if } |Y(\omega)| \leq M, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have $X_n^{(M)} \to X^{(M)}$ a.s. as long as $P(|X| = M) = 0$. Since there can be at most countably many $x \in \mathbb{R}$ such that $P(|X| = x) > 0)$, it is easy to choose M to satisfy this condition. Using the dominated convergence theorem and that $|X_n| \leq \sup_{|X| \leq M} h(x)$, we have

$$Eh(X_n^{(M)}) \to Eh(X^{(M)}).$$

Next, we have

$$\begin{aligned}
|Eh(X_n) - Eh(X_n^{(M)})| &= \left| \int_{|x|>M} h(X_n) dP \right| \leq \int_{|x|>M} |h(X_n)| dP \\
&\leq \epsilon_M \int_{|X|>M} g(X_n) dP \leq \epsilon_M \int_{\Omega} g(X_n) dP \\
&= Eg(X_n) \leq \epsilon_M K.
\end{aligned}$$

On the other hand, using the argument as above together with the Fatou lemma, we have

$$|Eh(X) - Eh(X^{(M)})| \leq \epsilon_M Eg(X) = \epsilon_M E\left(\liminf_{n\to\infty} g(X_n)\right) \leq \epsilon_M \liminf_{n\to\infty} E(g(X_n)) \leq \epsilon_M K.$$

Combining the limit identity and the two inequalities above, we have

$$\begin{aligned}
\limsup_{n\to\infty} |Eh(X_n) - Eh(X)| \leq & \limsup_{n\to\infty} |Eh(X_n^{(M)}) - Eh(X^{(M)})| \\
& + \limsup_{n\to\infty} |Eh(X_n) - Eh(X_n^{(M)})| \\
& + \limsup_{n\to\infty} |Eh(X) - Eh(X^{(M)})| \\
\leq & \, 2\epsilon_M K.
\end{aligned}$$

Since the right-hand side can be arbitrarily small, we prove that $\lim_{n\to\infty} |Eh(X_n) - Eh(X)| = 0$. □

After the discussion of the theoretical properties of expectation, we turn to the computation of expectation, if the distribution of the random variable is known. The next theorem shows that the integral on the (possibly very large) probability space can be transformed into an integral on the real line.

For a random variable X, we call a measure μ defined on $(\mathbb{R}, \mathcal{B})$ as its *distribution*, if for any Borel set $B \in \mathcal{B}$, $P(X \in B) = \mu(B)$. Recall that in Section 1, we defined the distribution function $F(x)$ for a random variable X. It is clear that given μ, F is determined by μ simply as $F(b) - F(a) = \mu(a, b]$, while we proved that given any distribution function F, the distribution μ can be constructed by Carathéodory extension theorem. Hence we have

Theorem 12. *Let f be a measurable function from $(\mathbb{R}, \mathcal{B})$ to $(\mathbb{R}, \mathcal{B})$. Under the condition either (a) $f \geq 0$, or (b) $E|f(X)| < \infty$, we have*

$$Ef(X) = \int_{\Omega} f(X) dP = \int_{\mathbb{R}} f(y) \mu(dy).$$

The proof of the theorem is measure theoretic, and we give the idea of the proof. You can fill in the detail. First, if f is an indicator function such that $f(x) = 1$ if $x \in B$ and $f(x) = 0$ if $x \in B^c$, then the right-hand side is simply $\mu(B)$ and the left-hand side is $P(X \in B)$ that is equal to $\mu(B)$ by the definition of distribution. Next, if f is a simple function, that is, a linear combination of indicator functions, the identity holds due to linearity. The next step is to use a sequence of simple function functions to approximate a non-negative function, and prove the theorem in the case $f \geq 0$. The last step is to consider f_+ and f_- separately and prove the theorem for signed function f under the condition that $E|f(X)| < \infty$. Note that the 4-step routine: indicator function — simple function — nonnegative function — general signed function is a standard trick for measure-theoretic proofs.

If the distribution μ is absolutely continuous with respect to the Lesbegue measure, the integral with respect to $\mu(dy)$ can be done easily. If μ is a discrete measure, X is a a discrete random variable, and you know how to deal with it. (Examples are random variables normal distribution and Poisson distribution. Please compute $E(X^k)$ with X having these distributions.) Now we consider another example.

Example 1. Let X be a random variable with the *Cantor distribution* that is defined by the Cantor set in Section 1. Compute EX and EX^2.

First we compute EX. By definition, $\mu(a,b] = F(b) - F(a)$, where $F((0.a_1a_2\dots)_3) = (0.\frac{a_1}{2}\frac{a_2}{2}\dots)_2$ if all $a_1, a_2, \dots$ are all 0 or 2. Also we have that $\mu(-\infty, 0] = 0$ and $\mu(1,\infty) = 0$. So

$$EX = \int_{\mathbb{R}} y\mu(dy) = \int_0^1 y\mu(dy).$$

Now we divide $(0,1]$ into 3^n equal intervals: $I_k = ((k-1)/3^n, k/3^n]$, where $k = 1, \dots, 3^n$. Then

$$\sum_{k=1}^{3^n} \frac{k-1}{3^n}\mu(I_k) \leq EX \leq \sum_{k=1}^{3^n} \frac{k}{3^n}\mu(I_k).$$

We have that $\mu(I_k) = 1/2^n$ if $(k-1)/3^n = (0.a_1a_2\dots a_n)_3$ if $a_1, \dots, a_n$ are 0 or 2, and $\mu(I_k) = 1$ otherwise. Then the inequality above can be simplified as

$$\sum_{a_1=0,2}\sum_{a_2=0,2}\cdots\sum_{a_n=0,2}\left(\frac{a_1}{3} + \frac{a_2}{9} + \cdots + \frac{a_n}{3^n}\right)\frac{1}{2^n} \leq EX$$
$$\leq \sum_{a_1=0,2}\sum_{a_2=0,2}\cdots\sum_{a_n=0,2}\left(\frac{a_1}{3} + \frac{a_2}{9} + \cdots + \frac{a_n}{3^n} + \frac{1}{3^n}\right)\frac{1}{2^n}.$$

Taking the limit $n \to \infty$, we derive that $EX = 1/2$. Similarly, we have

$$\sum_{a_1=0,2}\sum_{a_2=0,2}\cdots\sum_{a_n=0,2}\left(\frac{a_1}{3} + \frac{a_2}{9} + \cdots + \frac{a_n}{3^n}\right)^2\frac{1}{2^n} \leq EX^2$$
$$\leq \sum_{a_1=0,2}\sum_{a_2=0,2}\cdots\sum_{a_n=0,2}\left(\frac{a_1}{3} + \frac{a_2}{9} + \cdots + \frac{a_n}{3^n} + \frac{1}{3^n}\right)^2\frac{1}{2^n},$$

and derive that $EX^2 = 3/32$ by letting $n \to \infty$. (Please check it.)

The expectation of X^k of random variable X, if exists, is called the *k-th monent* of X, and is important, especially for $k = 1$ (the expectation, usually denoted by μ) and $k = 2$. We then define the *variance* of random variable X by

$$\operatorname{var}(X) = E(X - \mu)^2 = EX^2 - 2\mu EX + \mu^2 = EX^2 - \mu^2.$$

The variance has the property that it is invariant if the random variable is added by a constant, and it changes quadratically if the random variable is multiplied by a constant. To be precise,

$$\begin{aligned}\operatorname{var}(aX + b) &= E(aX + b)^2 - (E(aX + b))^2 = E(a^2X^2 + 2abX + b^2) - (aEX + b)^2 \\ &= a^2EX^2 + 2ab\mu + b^2 = a^2\mu^2 = 2ab\mu - b^2 \\ &= a^2(EX^2 - \mu^2) = a^2 \operatorname{var}(X).\end{aligned}$$

So the variance is not a linear functional on X, and generally we cannot expect that $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y)$. However, if X and Y are independent, we have this identity. To prove it rigorously, we need to learn more properties of independence.

Let $X_1, \dots, X_n$ be random variables. They together form a *random vector* $(X_1, \dots, X_n)$ that is a mapping from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^n, \mathcal{B}_n)$ such that the inverse of $B \in \mathcal{B}_n$ is a measurable set in $\mathcal{F}$. (Think why?) We call a probability measure μ on $(\mathbb{R}^n, \mathcal{B}_n)$ a *distribution* for $(X_1, \dots, X_n)$ if $P((X_1, \dots, X_n) \in B) = \mu(B)$. (So the distribution for a single random variable is a special case.) For any random vector, the distribution exists and is a probability measure. To see it, we note that μ is the *induced measure* from the measure P on $(\Omega, \mathcal{F}, P)$ by the measurable mapping f.

Theorem 13. *Suppose $X_1, \dots, X_n$ are independent random variables and X_i has distribution μ_i. Then $(X_1, \dots, X_n)$ has distribution $\mu_1 \times \mu_2 \times \cdots \times \mu_n$, the product measure of $\mu_1, \dots, \mu_n$ on $(\mathbb{R}^n, \mathcal{B}_n)$.*

For the proof of the theorem, we need to introduce some more notations and concepts. We call a collection $\mathcal{A}$ of subsets of Ω a *π-system*, if it is closed under intersection, that is, if $A, B \in \mathcal{A}$, then $A \cap B \in \mathcal{A}$.

Then we have the measure-theoretic result

Theorem 14. *Let P be a π-system. If ν_1 and ν_2 are measures that agree on P and there is a sequence $A_n \in P$ with $A_n \uparrow \Omega$ and $\nu_i(A_n) < \infty$, then ν_1 and ν_2 agree on $\sigma(P)$.*

The proof of this theorem is given in [Durrett, Theorem A.1.5]. It depends on the $\pi - \lambda$ theorem, which we do not introduce in this module.

Now we can continue the proof to Theorem 13.

Proof to Theorem 13. We want to show that for any $B \in \mathcal{B}_n$, $P((X_1, \dots, X_n) \in B) = \mu_1 \times \mu_2 \times \cdots \times \mu_n(B)$. In the special case that $B = B_1 \times \cdots \times B_n$ where $B_1, \dots, B_n$ are Borel sets on $\mathbb{R}$, we have by the independence

$$\begin{aligned}P((X_1, \dots, X_n) \in B_1 \times \cdots \times B_n) &= P(X_1 \in B_1, \dots, X_n \in B_n) \\ &= P(X_1 \in B_1) \times \cdots \times P(X_n \in B_n) \\ &= \mu_1(B_1) \times \cdots \times \mu_n(B_n) \\ &= \mu_1 \times \cdots \times \mu_n(B_1 \times \cdots \times B_n).\end{aligned}$$

Now we note that the collection of the "cube-like" subsets of $\mathbb{R}^n$, $\{B_1 \times \cdots \times B_n\}$, is a π-system. To see it, we note that $(A_1 \times \cdots \times A_n) \cap (B_1 \times \cdots \times B_n) = (A_1 \cap B_1) \times \cdots \times (A_n \cap B_n)$. Since both the distribution for $(X_1, \ldots, X_n)$ and the product measure $\mu_1 \times \cdots \times \mu_n$ are probability measures on $(\mathbb{R}^n, \mathcal{B}_n)$, and they agree on the π-system $\{B_1 \times \cdots \times B_n\}$, we derive by Theorem 14, they agree on $\sigma(\{B_1 \times \cdots \times B_n\}) = B_n$, and they are the same. □

Similar to the expectation formula in Theorem 12, we have the following result.

Theorem 15. *Suppose $X_1, \ldots, X_n$ are random variables, and the distribution for the random vector $(X_1, \ldots, X_n)$ is μ. If $f : (\mathbb{R}^n, \mathcal{B}_n) \to (\mathbb{R}, \mathcal{B})$ is a measurable mapping, then under the condition either (a) $f \geq 0$, or (b) $E|f(X_1, \ldots, X_n)| < \infty$, we have*

$$Ef(X) = \int_\Omega f(X_1, \ldots, X_n) dP = \int_{\mathbb{R}^n} f(y) \mu(dy).$$

The proof is the same as the one-dimensional case and we omit it. In the special case that X_1 and X_2 are independent, with distributions μ_1 and μ_2 respectively, we have

$$E(f(X_1, X_2)) = \iint f(y_1, y_2) \mu_1 \times \mu_2(dy)$$

(if either of the two conditions in Theorem 15 is satisfied). We can use Fubini's theorem to compute it. Recall:

Theorem 16 (Fubini). *If $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ are two measure spaces, $\Omega = \Omega_1 \times \Omega_2$ is the product set, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$ is the product σ-algebra, and $\mu = \mu_1 \times \mu_2$ is the product measure. Suppose $h : \Omega \to \mathbb{R}$ is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$. Under the condition either (a) $h \geq 0$, or (b) $\int |h| d\mu < \infty$, we have that*

$$\int_{\Omega_1} \left(\int_{\Omega_2} f(x, y) \mu_2(dy) \right) \mu_1(dx) = \int_\Omega f d\mu = \int_{\Omega_2} \left(\int_{\Omega_1} f(x, y) \mu_1(dx) \right) \mu_2(dy).$$

Now suppose the independent random variables X_1 and X_2 are both non-negative. Then $X_1 X_2 = |X_1 X_2|$, and we have (μ_1, μ_2 are distributions for X_1, X_2 respectively)

$$\begin{aligned} E(X_1 X_2) = E(|X_1 X_2|) = \int_{\mathbb{R}^2} |y_1 y_2| \mu_1 \times \mu_2(dy) &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} |y_1 y_2| \mu_1(dy_1) \right) \mu_2(dy_2) \\ &= \left(\int_{\mathbb{R}} |y_1| \mu_1(dy_1) \right) \left(\int_{\mathbb{R}} |y_2| \mu_2(dy_2) \right) \\ &= E|X_1| E|X_2| = EX_1 EX_2. \end{aligned}$$

On the other hand, if X_1 and X_2 satisfy $E|X_1| < \infty$, $E|X_2|X_2 < \infty$, then we have $E|X_1 X_2| = E(|X_1||X_2|) = E|X_1| E|X_2| < \infty$. Then the condition $\int |h| d\mu < \infty$ for Fubini's theorem is satisfied, where $h = y_1 y_2$ and $\mu = \mu_1 \times \mu_2$, and we still have the result

$$E(X_1 X_2) = \int_{\mathbb{R}^2} y_1 y_2 \mu_1 \times \mu_2(dy) = \left(\int_{\mathbb{R}} |y_1| \mu_1(dy_1) \right) \left(\int_{\mathbb{R}} |y_2| \mu_2(dy_2) \right) = EX_1 EX_2.$$

The final result in this section is:

Theorem 17. *Suppose random variables $X_1, \dots, X_n$ are independent. Under the condition either (a) $X_i \geq 0$, or (b) $E|X_i| < \infty$, for all $i = 1, \dots, n$, then* $\mathrm{var}(X_1 + \cdots + X_n) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)$.

Proof. Under either condition,

$$E(X_1 + \cdots + X_n)^2 = \sum_{i=1}^{n} EX_i^2 + 2 \sum_{1 \leq i < j \leq n} EX_iX_j = \sum_{i=1}^{n} EX_i^2 + 2 \sum_{1 \leq i < j \leq n} EX_iEX_j.$$

Then it is easy to derive the formula for $\mathrm{var}(X_1 + \cdots + X_n)$. □

3 More on independence, and weak laws of large numbers

For the independence of random variables, we still do not have an effective way to check if a collection of random variables are independent. The definition of independence of random variables involves arbitrary Borel sets, and it is not practical. Even for theoretical questions, the definition may not be directly applicable. For example, if we know that X_1, X_2, X_3 are independent, are the two random variables X_1 and X_2X_3 independent? It should be true, but if we want to verify it by definition, the condition $X_2X_3 \in B$ cannot be simply expressed by conditions like $X_2 \in B'$ and $X_3 \in B''$. To solve the question, as usual we need to introduce more concepts and notations.

Definition 3. We say events $A_1, A_2, \dots, A_n$ are independent if for any $m_1, \dots, m_k \in \{1, 2, \dots, n\}$, $P(A_{m_1} \cap \cdots \cap A_{m_k}) = P(A_{m_1}) \cdots P(A_{m_k})$.

Definition 4. Let $\mathcal{A}_1, \dots, \mathcal{A}_n$ be subsets of $\mathcal{F}$ on the probability space $(\Omega, \mathcal{F}, P)$. We say $\mathcal{A}_1, \dots, \mathcal{A}_n$ are independent if for any $A_i \in \mathcal{A}_i$, $A_1, \dots, A_n$ are independent.

A random variable X defines a σ-algebra $\sigma(X)$, which consists of sets $\{X^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\}$. It is clear that $X_1, \dots, X_n$ are independent if and only if the σ-algebras $\sigma(X_1), \dots, \sigma(X_n)$ are independent. Then the following theorem can reduce our task of checking independence of σ-algebras.

Theorem 18. *Suppose $\mathcal{A}_1, \dots, \mathcal{A}_n$ are independent subsets of $\mathcal{F}$, and each $\mathcal{A}_i$ is a π-system. Then $\sigma(\mathcal{A}_1), \dots, \sigma(\mathcal{A}_n)$ are independent.*

The proof of the theorem requires the "$\pi - \lambda$ theorem" and you can find the proof, together with the proof of the $\pi - \lambda$ theorem, in our textbook. Here we note an important case: The semi-infinite sets $(-\infty, a]$ form a π-system, and they generate the Borel σ-algebra on $\mathbb{R}$. Then for any random variable X, the sets $\{X \leq a\} = \{\omega \mid X(\omega) \leq a\}$ form a π-system and they generate the σ-algebra $\sigma(X)$. Hence we have the consequence of last theorem:

Corollary 19. *$X_1, \dots, X_n$ are indepdent if and only if for all $m_1, \dots, m_k \in \{1, 2, \dots, n\}$ and $x_{m_1}, \dots, x_{m_k}$, $P(X_{m_1} \leq x_{m_1}, \dots, X_{m_k} \leq x_{m_k}) = \prod_{i=1}^{k} P(X_{m_i} \leq x_{m_i})$.*

Now we can go back to the question that how to show X_1 and X_2X_3 are independent, given that X_1, X_2, X_3 are independent. We need to show that $\sigma(X_1)$ and $\sigma(X_2X_3)$ are independent. To describe $\sigma(X_2X_3)$, we introduce the mapping $f : \Omega \to \mathbb{R}^2$ by $f(\omega) = (X_2(\omega), X_3(\omega))$, and the mapping $g : \mathbb{R}^2 \to \mathbb{R}$ by $g(x, y) = xy$. Then $\sigma(X_2X_3) = \{(g \circ f)^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\}$, and it is generated by $\{(g \circ f)^{-1}(-\infty, a]\} = \{f^{-1}(A_a)\}$, where $A_a = \{(x, y) \mid xy \leq a\}$. It is clear that $A_a \in \mathcal{B}(\mathbb{R}^2)$, and then $\sigma(X_2X_3) \subseteq \mathcal{A} = \{f^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R}^2)\}$. Then it suffices to show that $\sigma(X_1)$ and $\mathcal{A}$ are independent. Since $\mathcal{B}(\mathbb{R}^2)$ is generated by $\{(-\infty, x_2] \times (-\infty, x_3]\}$, $\mathcal{A}$ is generated by $\{f^{-1}(-\infty, x_2] \times (-\infty, x_3]\} = \{X_2 \leq x_2\} \cap \{X_3 \leq x_3\}$. Since $\sigma(X_1)$ is generated by $\{X_1 \leq x_1\}$, we need only to check that

$$P\left(\{X_1 \leq x_1\} \cap (\{X_2 \leq x_2\} \cap \{X_3 \leq x_3\})\right) = P(X_1 \leq x_1)P(X_2 \leq x_2, X_3 \leq x_3),$$

and this is a direct consequence of the independence of X_1, X_2, X_3.

The argument above can be generalised to prove the following result:

Corollary 20. *If for $1 \leq i \leq n$, $1 \leq j \leq m(i)$, $X_{i,j}$ are independent, and $f_i : \mathbb{R}^{m(i)} \to \mathbb{R}$ are measuable, then $f_i(X_{i,1}, \dots, X_{i,m(i)})$ are independent.*

We prove the special case with $n = 2$, $m(1) = 1$, $m(2) = 2$, $f_1(x) = x$ and $f_2(x, y) = xy$ above, and leave the proof for the general case to you.

Now we generalise a result in last section:

Corollary 21. *Suppose random variables $X_1, \dots, X_n$ are either all non-negative or $E|X_i| < \infty$ for all $i = 1, \dots, n$. Then*

$$E(X_1 \cdots X_n) = E(X_1) \cdots E(X_n).$$

Proof. The $n = 2$ case is already proved. If $n > 2$, we use induction, and denote $Y = X_1 \cdots X_{n-1}$. We have that Y and X_n are independent. If all $X_i \geq 0$, then $Y \geq 0$. If all $E|X_i| < \infty$, then by the induction hypothesis, $E|Y| = E(|X_1| \cdots |X_{n-1}|) = E|X_1| \cdots E|X_{n-1}| < \infty$. Thus in either case,

$$E(X_1 \cdots X_n) = E(YX_n) = EYEX_n = EX_1 \cdots EX_{n-1}EX_n,$$

and finish the proof. □

Now we start to introduce the first of the two most important topics in this module: the Law of Large Numbers (LLN), (while the other is the Central Limit Theorem, (CLT)). Basically, a law of large numbers is that a sequence of random variables $\{Y_n\}$ converge to a fixed number. The problem is: In what sense do we talk about the convergence? Recall that a random variable is a function on the probability space. In calculus we learn the pointwise convergence and the uniform convergence, and they are not equivalent. In the further study of real analysis we learn about the L^1 convergence and L^2 convergence (for L^1/L^2 integrable functions), and the weak* convergence (if we view the space of integrable functions as a Banach/Hilbert space). First we consider weak laws of large numbers, which involve some weak form of convergence, in contrast to the strong laws of large numbers to be introduced later.

Theorem 22. *Let $X_1, X_2, \dots$ be independent random variables with $EX_i = \mu$ and $\operatorname{var}(X_i) \leq C < \infty$. If $S_n = X_1 + X_2 + \cdots + X_n$, then $S_n/n \to \mu$ in L^2.*

Proof. We need to show that

$$\lim_{n\to\infty} \int \left(\frac{S_n}{n} - \mu \right)^2 dP = \lim_{n\to\infty} E \left(\frac{S_n}{n} - \mu \right)^2 = 0.$$

Noting that $E(S_n/n) = n^{-1}E(X_1 + \cdots + X_n) = n^{-1}(E(X_1) + \cdots + E(X_n)) = \mu$, we only need to show that $\lim_{n\to\infty} \operatorname{var}(S_n/n) \to 0$. Using the independence of $X_1, X_2, \dots$, we have

$$\lim_{n\to\infty} \operatorname{var} \left(\frac{S_n}{n} \right) = \lim_{n\to\infty} \frac{S_n}{n^2} = \lim_{n\to\infty} \frac{\operatorname{var}(X_2) + \cdots + \operatorname{var}(X_n)}{n^2} \leq \lim_{n\to\infty} \frac{nC}{n^2} = 0,$$

and finish the proof. □

Remark 1. Here we only need the consequence of the independence of $X_1, X_2, \dots$ that $\text{var}(X_1 + \ldots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n)$, and this kind of identities hold as long as $E(X_i X_j) = E(X_i)E(X_j)$ for all $i \neq j$, which is the *uncorrelation* of $X_1, X_2, \dots$. So Theorem 22 holds if the independence condition is replaced by the weaker condition that $X_1, X_2, \dots$ are uncorrelated.

Remark 2. Theorem 22, and other laws of large numbers, are mostly applied in the setting that $X_1, X_2, \dots$ are "independent and identically distributed" (i.i.d. for short).

The L^2 convergence is not the commonly used convergence in probability theory, since it does not sound "probabilistic". One important convergence is the convergence in probability, as defined below:

Definition 5. We say a sequence of random variables $\{Y_n\}$ converges to Y in probability if for all $\epsilon > 0$, $P(|Y_n - Y| > \epsilon) \to 0$ as $n \to \infty$.

A simple result is

Lemma 23. *If $p > 0$ and $E|Y_n|^p \to 0$, then $Y_n \to 0$ in probability.*

Proof. Given any $\epsilon, \delta > 0$, there is N such that for all $n > N$,

$$\int |Y_n|^p dP < \delta \epsilon^p.$$

Then for $n > N$, $P(|Y_n| > \epsilon) < \delta$. Thus we prove the lemma. □

Remark 3. This lemma is a consequence of the Chebyshev inequality.

Then as a direct consequence of Lemma 23 with $p = 2$, we have that Theorem 22 implies

Theorem 24. *Let $X_1, X_2, \dots$ satisfy the conditions in Theorem 22, and μ and S_n be defined as in Theorem 22. Then $S_n/n \to \mu$ in probability.*

It turns out that for the average of i.i.d. random variables to converge to their expectation in probability, the requirement that the variance is finite is unnecessary. We have the following result:

Theorem 25. *Let $X_1, X_2, \dots$ be i.i.d. with $E|X_i| < \infty$ and $EX_i = \mu$. Let $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ in probability.*

The proof of this theorem is more involved, and we need to establish some technical lemmas.

Lemma 26. *For each n, let $X_{n,1}, \dots, X_{n,n}$ be independent random variables. Let $b_n > 0$ be positive numbers with $b_n \to \infty$ as $n \to \infty$, and let $\bar{X}_{n,k} = X_{n,k} 1_{|X_{n,k}| \leq b_n}$, that is,*

$$\bar{X}_{n,k}(\omega) = \begin{cases} X_{n,k}(\omega) & \textit{if } |X_{n,k}(\omega)| \leq b_n, \\ 0 & \textit{otherwise.} \end{cases}$$

Suppose that as $n \to \infty$,

$$\sum_{k=1}^{n} P\left(|X_{n,k}| > b_n\right) \to 0, \quad \textit{and} \quad \frac{1}{b_n^2}\sum_{k=1}^{n} E\bar{X}_{n,k}^2 \to 0.$$

If we let $S_n = X_{n,1} + \cdots + X_{n,n}$ and $a_n = \sum_{k=1}^n E\bar{X}_{n,k}$, then $(S_n - a_n)/b_n \to 0$ in probability.

Before giving the proof to Lemma 26, we state a lemma that is similar to Lemma 23, whose proof is left to you.

Lemma 27. *Let $S_1, S_2, \ldots$ be random variables such that $ES_n = \mu_n$ and $\operatorname{var}(S_n) = \sigma_n^2$. Suppose $\{b_n\}$ are positive numbers and $\sigma_n^2/b_n^2 \to 0$ as $n \to \infty$, then $(S_n - \mu_n)/b_n \to 0$ in probability.*

Proof of Lemma 26. First consider $\bar{S}_n = \bar{X}_{n,1} + \cdots + \bar{X}_{n,n}$ instead of S_n. $\bar{S}_n$ has the advantage that its variance is finite. Furthermore,

$$\operatorname{var}(\bar{S}_n) = \sum_{k=1}^{n} var(\bar{X}_{n,k}) \le \sum_{k=1}^{n} E|\bar{X}_{n,k}|^2.$$

(Here we use that $\bar{X}_{n,1}, \ldots, \bar{X}_{n,n}$ are independent. Why?) Thus by Lemma 27, we have that $(\bar{S}_n - a_n)/b_n \to 0$ in probability, or equivalently, for any $\epsilon, \delta > 0$, there is N such that for all $n > N$, $P(|(\bar{S}_n - a_n)/b_n| > \epsilon) < \delta$.

Next we use the property that $X_{n,k}$ and $\bar{X}_{n,k}$ are similar. We have that for any $\delta' > 0$, there is N' such that for all $n > N'$,

$$P(S_n \ne \bar{S}_n) \le \sum_{k=1}^{n} P(X_{n,k} \ne \bar{X}_{n,k}) = \sum_{k=1}^{n} P(|X_{n,k}| > b_n) < \delta'.$$

Therefore for $n > \max(N, N')$, $P(|(S_n - a_n)/b_n| > \epsilon) \le P(|(\bar{S}_n - a_n)/b_n| > \epsilon) + P(S_n \ne \bar{S}_n) < \delta + \delta'$, and we prove the lemma. □

The lemma above for arrays of random variables imply the following result for a sequence of random variables, and it is called the weak law of large numbers.

Theorem 28 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with*

$$xP(|X_i| > x) \to 0, \quad \textit{as } x \to \infty.$$

Let $S_n = X_1 + \cdots + X_n$ and let $\mu_n = E(X_1 1_{|X_1| \le n})$. Then $S_n/n - \mu_n \to 0$ in probability.

Proof. We use the result of Lemma 26. Let $X_{n,k} = X_k$ and $b_n = n$. Then

$$\lim_{n\to\infty} \sum_{k=1}^{n} P(|X_{n,k} > b_n) = \lim_{n\to\infty} \sum_{k=1}^{n} P(|X_k| > n) = \lim_{n\to\infty} nP(|X_1| > n) = 0.$$

On the other hand,

$$\lim_{n\to\infty} \frac{1}{b_n^2} \sum_{k=1}^{n} E\bar{X}_{n,k}^2 = \lim_{n\to\infty} \frac{1}{n^2} \sum_{k=1}^{n} E(X_k 1_{|X_k| \le n})^2 = \lim_{n\to\infty} \frac{1}{n} E(X_1 1_{|X_1| \le n})^2.$$

We denote $|X_1|1_{|X_1|\leq n} = Y_n$. Then

$$\begin{aligned} E(Y_n^2) &= \int_\Omega Y_n^2 dP = \int_\Omega \left(\int_0^{Y_n} 2y dy \right) dP = \int_\Omega \left(\int_0^\infty 2y 1_{Y_n > y} dy \right) dP \\ &= \int_0^\infty \left(\int_\Omega 2y 1_{Y_n > y} dP \right) dy = \int_0^\infty 2y \left(\int_\Omega 1_{Y_n > y} dP \right) dy \\ &= \int_0^\infty 2y P(Y_n > y) dy. \end{aligned}$$

Using that $0 \leq Y_n \leq n$ and for all $y \in [0, n]$, $P(Y_n > y) \leq P(|X_1| > y)$, we have

$$\frac{1}{n} E(Y^2) \leq \int_0^n 2y P(|X_1| > y) dy = 2 \int_0^1 nx P(|X_1| > nx) dx.$$

Since for all $x > 0$, $nxP(|X_1| > nx) \to 0$, we have (exercise: justify the argument)

$$\lim_{n\to\infty} \frac{1}{b_n^2} \sum_{k=1}^n E\bar{X}_{n,k}^2 = \lim_{n\to\infty} \frac{1}{n} E(Y^2) = 0.$$

Thus Lemma 26 yields the theorem. □

An intermediate step in the proof can ge generalised to the following result:

Lemma 29. *If $Y \geq 0$ and $p > 0$, then $E(Y^p) = \int_0^\infty py^{p-1} P(Y > y) dy$.*

The proof is left as an exercise.

At last, we can prove Theorem 25, the practically most convenient form of the weak law of large numbers.

Proof of Theorem 25. Since $E|X_1| < \infty$, by the dominanted convergence theorem, we have

$$\lim_{x\to\infty} xP(|X_1| > x) = 0 \quad \text{and} \quad \lim_{n\to\infty} E(X_1 1_{|X_1|\leq n}) = EX_1.$$

Hence Theorem 28 implies Theorem 25. □

4 Borel-Cantelli lemmas and strong law of large numbers

In this section we introduce the *strong* laws of large numbers, that is, the convergence of the average of random variables to their expectation, *almost surely.* Recall that we say a sequence of random variables $\{X_n\}$ converges to X a.s. if for all $\omega \in \Omega \setminus E$, $X_n(\omega) \to X(\omega)$ as $n \to \infty$, where $E \in \mathcal{F}$ and $P(E) = 0$. We call this kind of laws of large strong, because the almost sure convergence implies the convergence in probability, but the converse is not true. To see it, suppose $X_n \to X$ a.s. we define the random variable $Y_n = \sup_{k \geq n} |X_k - X|$. They are non-negative and decreases as n increases. Thus EY_n are non-negative and decreasing. Furthermore, we have $\liminf_{n\to\infty} Y_n = 0$ a.s.. By Fatou's lemma,

$$\liminf_{n\to\infty} EY_n \leq E\left(\liminf_{n\to\infty} Y_n\right) = 0.$$

So for any $\epsilon, \delta > 0$, there is N such that for all $n > N$, $E|X_n - X| \leq EY_n < \epsilon\delta$, and then $P(|X_n - X| > \epsilon) < \delta$.

On the other hand, we have examples that $X_n \to X$ in probability but not almost surely. To construct an example, we define random variables $\{X_{2,1}, X_{2,2}, X_{4,1}, X_{4,2}, X_{4,3}, X_{4,4}, X_{8,1}, \dots, X_{8,8}, X_{16,1}, \dots\}$ on the probability space $([0,1], \mathcal{B}, \lambda)$, where λ is the Lebesgue measure, such that

$$X_{2^n,k}(\omega) = \begin{cases} 1 & \text{if } (k-1)/2^n \leq \omega \leq k/2^n, \\ 0 & \text{otherwise.} \end{cases}$$

Then the sequence converges to 0 in probability, but does not converge to any limit almost surely.

The tool to prove strong laws of large numbers is the Borel-Cantelli lemma, and the second Borel-Cantelli lemma. They are about the probability that infinitely many events occurs, given the probability of each event. To be precise, we consider a sequence of events $A_1, A_2, \dots \in \mathcal{F}$ on the probability space $(\Omega, \mathcal{F}, P)$. Then the event $\{$at least one A_n occurs$\}$ is simply $A_1 \cup A_2 \cup \cdots$, the event $\{$at least k of A_n occur$\}$ is $\bigcup_{n_1=1}^{\infty} \bigcup_{n_2=n_1+1}^{\infty} \cdots \bigcup_{n_k=n_{k-1}+1}^{\infty} (A_{n_1} \cap A_{n_2} \cap \cdots \cap A_{n_k}$, and the event $\{$at least infinitely many A_n occur$\}$ is

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \left(\bigcup_{k=n}^{\infty} A_k \right),$$

and we denote it as A_n i. o. where i. o. means "infinitely often".

Lemma 30 (Borel-Cantelli)**.** *If* $\sum_{n=1}^{\infty} P(A_n) < \infty$*, then* $P(A_n \text{ i. o.}) = 0$.

The intuitive interpretation of of this lemma is simple. Think each A_n as a partial cover of Ω. If the total area of the covers is finite, then the area of the region that is covered infinitely many times has to be zero.

Proof. To show that $P(A_n \text{ i. o.}) = P(\limsup_{n\to\infty} A_n = 0$, it suffices to show that for all $\epsilon > 0$, there is N such that $P(\bigcup_{n=N}^{\infty} A_n) < \epsilon$. Since $P(\bigcup_{n=N}^{\infty} A_n) \leq \sum_{n=N}^{\infty} P(A_n)$, we can take N to be large enough such that $\sum_{n=N}^{\infty} P(A_n) < \epsilon$, and it is clear that such N exists. □

The Borel-Cantelli theorem implies that if a sequence of random variables converges in probability, then there is a subsequence that converges almost surely. Actually we have a stronger result:

Theorem 31. *The sequence of random variables $X_n \to X$ in probability, if and only if for any subsequence $X_{n(m)}$, there is a further subsequence $X_{n(m_k)}$ that converges almost surely to X.*

Proof. First suppose $X_n \to X$ in probability. Without loss of generality, we assume that $\{X_{(n(m)}\} = \{X_n\}$, and it suffices to show that there is a subsequence X_{n_k} that converges to X a.s.. We choose X_{n_k} such that

$$P\left(|X_{n_k} - X| > \frac{1}{k}\right) < \frac{1}{2^k}.$$

Denoting $A_k = \{|X_{n_k} - X| > 1/k\}$, we have that $\sum_{k=1}^{\infty} P(A_k) < 1$, and then $P(A_k \,\text{i.o.}) = 0$ by the Borel-Cantelli lemma. For all $\omega \notin A_k \,\text{i.o.}$, we have that there is N such that $\omega \notin A_k$ for all $k > N$, that is, $|X_{n_k}(\omega) - X(\omega)| \leq 1/k$ for all $k > N$, and then $X_{n_k}(\omega) \to X(\omega)$. Thus we prove that $X_{n_k} \to X$ a.s..

On the other hand, if $\{X_n\}$ does not converge to X, then there exist $\epsilon, \delta > 0$ and a subsequence $\{X_{n(m)}\}$ such that

$$P\left(|X_{n(m)} - X| > \epsilon\right) > \delta \quad \text{for all } n(m).$$

It is clear that any subsequence of $\{X_{n(m)}\}$ does not converge to X in probability. Suppose $\{X_{n(m)}\}$ has a subsequence that converges to X a.s., then the subsequence also converge to X in probability, and it is a contradiction. Thus we finish the proof. □

Theorem 31 connects the two kinds of convergence. As an application, we consider the convergence of $\{f(X_n)\}$, where $\{X_n\}$ converges and f is a continuous function. In the setting of almost sure convergence, it is straightforward. $X_n(\omega) \to X(\omega)$ implies that $f(X_n(\omega)) \to f(X(\omega))$, so if $X_n \to X$ a.s., then $f(X_n) \to f(X)$ a.s.. Furthermore, if f is bounded, that is, $|f(x)| < M$ for all $x \in \mathbb{R}$, then by the dominated convergence theorem, since $|f(X_n)| < M$, we have $Ef(X_n) \to Ef(X)$. The following corollary show that the results above are also valid if the convergence is in probability.

Corollary 32. *If f is a continuous function and $X_n \to X$ in probability, then $f(X_n) \to f(X)$ in probability. In addition, if f is bounded, then $Ef(X_n) \to Ef(X)$.*

Proof. Suppose $X_n \to X$ in probability, then using Theorem 31, we have that any subsequence $\{X_{n(m)}\}$ has a further subsequence $\{X_{n(m_k)}\}$ that converges a.s. to X. Thus any subsequence $\{f(X_{n(m)})\}$ has a further subsequence $\{f(X_{n(m_k)})\}$ that converges a.s. to $f(X)$. Using Theorem 31 conversely, we have that the sequence $\{f(X_n)\}$ converges to $f(X)$ in probability.

To prove the remaining part of the theorem, we note that for any subsequence $\{Ef(X_{n(m)})\}$ of $\{Ef(X_n)\}$, it has a further subsequence $\{Ef(X_{n(m_k)})\}$ that converges to $Ef(X)$, since we can take the further subsequence $f(X_{n(m_k)})$ to converge a.s.to $f(X)$. Hence we finish the proof by the simple fact: If any subsequence of $\{x_n\} \subseteq \mathbb{R}$ has a further subsequence that converges to x, then $x_n \to x$. □

The converse of the Borel-Cantelli lemma is not true, and it is an exercise for you to find a counterexample. However, with the independence of events, we have the following result.

Lemma 33 (Second Borel-Cantelli)**.** *If the events A_n are independent, then $\sum_{n=1}^{\infty} P(A_n) = \infty$ implies that $P(A_n \text{ i. o.}) = 1$.*

Proof. It suffices to show that for all n, $P(\bigcup_{k=n}^{\infty} A_k) = 1$, or equivalently, $P(\bigcap_{k=n}^{\infty} A_k^c) = 0$. Since $A_n, A_{n+1}, \dots$ are independent, $A_n^c, A_{n+1}^c, \dots$ are also independent, and for any $N \geq n$, we have

$$\begin{aligned} P\left(\bigcap_{k=n}^{\infty} A_k^c\right) \leq P\left(\bigcap_{k=n}^{N} A_k^c\right) = \prod_{k=n}^{N} P(A_k^c) = \exp\left(\sum_{k=n}^{N} \log(1 - P(A_k))\right) \\ \leq \exp\left(-\sum_{k=n}^{N} P(A_k)\right). \end{aligned}$$

Here we use the inequality that $\log(1-x) \leq -x$ for all $x \in [0,1]$. Since for any $\epsilon > 0$, we can let N large enough such that $\sum_{k=n}^{N} P(A_k) > -\log \epsilon$, we can make the right-hand side of the inequality above less than ϵ, and have $P(\bigcap_{k=n}^{\infty} A_k^c) < \epsilon$. Since ϵ is arbitrary, we derive that $P(\bigcap_{k=n}^{\infty} A_k^c) = 0$ and finish the proof. □

An application of the second Borel-Cantelli lemma is the following negative result for the strong law of large numbers.

Theorem 34. *If $X_1, X_2, \dots$ are i.i.d. with $E|X_i| = \infty$, then $P(|X_n| \geq n \text{ i. o.}) = 1$. So if $S_n = X_1 + \cdots + X_n$, then $P(\lim_{n\to\infty} S_n/n \text{ exists } \in (-\infty, \infty)) = 0$.*

Proof. Let μ be the distribution of X_1. Then $E|X_1| = \int |x| \mu(dx)$ and $P(|X_n| \geq n) = P(|X_1| \geq n) = \int 1_{|x| \geq n} \mu(dx)$. We have

$$P(|X_1| \geq 1) + P(|X_2| \geq 2) + \cdots = \int f(x)\mu(dx), \quad \text{where } f(x) = k \text{ for all } k \leq |x| < k+1.$$

It is clear that

$$\int f(x)\mu(dx) \leq \int |x|\mu(dx) \leq \int (f(x)+1)\mu(dx) = \int f(x)\mu(dx) + 1,$$

and so $P(|X_1| \geq 1) + P(|X_2| \geq 2) + \cdots = \infty$. Using the second Borel-Cantelli lemma, we have that $P(|X_n| \geq n \text{ i. o.}) = 1$.

Next, denote the set $A_k \subseteq \Omega$ as the set $\{\omega \mid \lim_{n\to\infty} S_n(\omega)/n \text{ exists } \in [-k,k]\}$. We can check that $A_k \in \mathcal{F}$. Below we show that $A_k \subseteq \Omega \setminus \{|X_n| \geq n \text{ i. o.}\}$, and so $P(A_k) = 0$. Hence we derive that $P(\lim_{n\to\infty} S_n/n \text{ exists } \in (-\infty, \infty)) = P(A_1 \cup A_2 \cup \cdots) = 0$.

Suppose $\omega \in A_k$. Then there exists $c \in [-k,k]$ and N such that for all $n > N$,

$$\left(c - \frac{1}{3}\right) n < S_n(\omega) = X_1(\omega) + \cdots + X_n(\omega) < \left(c + \frac{1}{3}\right) n.$$

We have

$$|X_{n+1}(\omega)| = |S_{n+1}(\omega) - S_n(\omega)| < \left(c + \frac{1}{3}\right)(n+1) - \left(c - \frac{1}{3}\right) n = \frac{2}{3} n + c + \frac{1}{3}$$
$$\leq \frac{2}{3} n + k + \frac{1}{3}.$$

Suppose without loss of generality that $N > 3k$, then $|X_{n+1}(\omega)| < n + 1$ for all $n > N$, which means that $\omega \notin \{|X_n| \geq n \,\text{i. o.}\}$. □

The theorem above implies that the condition $E|X_i| < \infty$ is necessary for a reasonable strong law of large numbers, in contrast to the weak law of large numbers where we only require $nP(|X_n| \geq n) \to 0$ as $n \to \infty$ in Theorem 28. (To be fair, we need that μ_n converges to a limit in Theorem 28 to make the result comparable to Theorem 34. But $nP(|X_n| \geq n) \to 0$ together with the convergence of $\{\mu_n\}$ is still weaker than $E|X_i| < \infty$.

Finally we give the proof of the strong law of large numbers, which is slightly stronger than the converse of Theorem 34.

Theorem 35. *Let $X_1, X_2, \dots$ be pairwise independent identically distributed random variables with $E|X_i| < \infty$. Let $EX_i = \mu$ and $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ a.s. as $n \to \infty$.*

Before giving the proof to Theorem 35, we remark that the *pairwise* independence of random variables $X_1, X_2, \dots$ means that any pair of random variables X_i, X_j are independent, but the independence of three or more random variables may fail. So this condition is weaker than the independence of $\{X_n\}$.

The basic idea of the proof of Theorem 35 is again the truncation.

Lemma 36. *Let $Y_k = X_k 1_{|X_k| \leq k}$ and $T_n = Y_1 + \ldots + Y_n$. Then Theorem 35 is equivalent to that $T_n/n \to \mu$ a.s..*

Proof. If we can show that $X_k = Y_k$ almost surely for all large enough k, then almost surely $(S_n/n - T_n/n) \to 0$, and the equivalence is proved. Next, $X_k(\omega) = Y_k(\omega)$ for all large enough k if and only if $\omega \in \Omega \setminus \{|X_k| > k \,\text{i. o.}\}$. By the assumption that $E|X_i| < \infty$, we can show that $P(|X_1| > 1) + P(|X_2| > 2) + \cdots < \infty$, see the proof of Theorem 34. Thus the applicaiton of the Borel-Cantelli lemma implies that $P\{|X_k| > k \,\text{i. o.}\} = 0$ and we finish the proof. □

Below we prove that $T_n/n \to \mu$ a.s.. First we derive a technical lemma.

Lemma 37. *For the random variables Y_k defined in Lemma 36, we have*

$$\sum_{k=1}^{\infty} \frac{1}{k^2} EY_k^2 < \infty.$$

Proof. Let μ be the distribution of X_1. Then

$$EY_k^2 = \int x^2 1_{|x| \leq k} \mu(dx), \quad \text{and} \quad EY_1^2 + EY_2^2 + \cdots = \int x^2 g(x) \mu(dx),$$

where

$$g(x) = \begin{cases} \sum_{n=k+1}^{\infty} \frac{1}{n^2} & \text{for } |x| \in (k, k+1], \\ \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} < 1.645 & \text{for } x \in [-1,1]. \end{cases}$$

Note that for $|x| > 1$,

$$g(x) = g(|x|) < \int_{|x|}^{\infty} \frac{1}{t^2} dt = \frac{1}{|x|},$$

and then

$$\begin{aligned} \int x^2 g(x)\mu(dx) &= \int_{-1}^{1} x^2 g(x)\mu(dx) + \int_{\mathbb{R}\setminus[-1,1]} x^2 g(x)\mu(dx) \\ &\le \int_{-1}^{1} \frac{\pi^2}{6}\mu(dx) + \int_{\mathbb{R}\setminus[-1,1]} |x|\mu(dx) \\ &\le 1.645 + E|X_1| < \infty. \end{aligned}$$

□

The next lemma is left as an exercise.

Lemma 38. *If $X'_n \to \mu'$ a.s., and $X''_n \to \mu''$ a.s., then $\{X_n = X'_n \pm X''_n\}$ converges to $\mu = \mu' \pm \mu''$ a.s..*

We are going to use the lemma above in the special case that $X_n = X_n^+ - X_n^-$, where $X_n^{\pm}$ is the positive/negative part of X_n. If $E|X_i| < \infty$, then $EX^+ < \infty$ and $EX^- < \infty$. Thus we only need to prove Theorem 35 in the case that X_n are non-negative.

Proof of Theorem 35. First we show that a subsequence of $\{T_n\}$ converges to μ a.s.. Let $\alpha > 1$, and define $k(n) = [\alpha^n]$. We take the subsequence as $\{T_{k(n)}\}$.

For all $\epsilon > 0$, we have

$$\begin{aligned} P\left(\left|\frac{T_{k(n)}}{k(n)} - E\frac{T_{k(n)}}{k(n)}\right| > \epsilon\right) &\le \epsilon^{-2} E\left(\frac{T_{k(n)}}{k(n)} - E\frac{T_{k(n)}}{k(n)}\right)^2 = \frac{\epsilon^{-2}}{k(n)^2}\operatorname{var}(T_{k(n)}) \\ &= \frac{\epsilon^{-2}}{k(n)^2}\sum_{m=1}^{k(n)} \operatorname{var}(Y_m). \end{aligned}$$

Here we use Chebyshev's inequality and that $Y_1, \dots, Y_m$ are pairwise independent. Then

$$\begin{aligned} \sum_{n=1}^{\infty} P\left(\left|\frac{T_{k(n)}}{k(n)} - E\frac{T_{k(n)}}{k(n)}\right| > \epsilon\right) &= \epsilon^{-2}\sum_{n=1}^{\infty}\frac{1}{k(n)^2}\sum_{m=1}^{k(n)}\operatorname{var}(Y_m) \\ &= \epsilon^{-2}\sum_{m=1}^{\infty}\operatorname{var}(Y_m)\sum_{n:k(n)\ge m}\frac{1}{k(n)^2}. \end{aligned}$$

Using the inequality (exercise)

$$\sum_{n:\alpha^n \ge m} \frac{1}{[\alpha^n]^2} \le \frac{4}{(1-\alpha^{-2})m^2},$$

we have

$$\sum_{n=1}^{\infty} P\left(\left|\frac{T_{k(n)}}{k(n)} - E\frac{T_{k(n)}}{k(n)}\right| > \epsilon\right) \le \frac{4\epsilon^{-2}}{(1-\alpha^{-2})} \sum_{m=1}^{\infty} E(Y_m^2)\frac{1}{m^2} < \infty.$$

Thus by the Borel-Cantelli lemma, $T_{k(n)}/k(n) - E(T_{k(n)}/k(n))$ converges to 0 a.s.. Since $EY_k \to \mu = EX_1$ by the dominated convergence theorem (also by the monotone convergence theorem, since we assume X_1 is non-negative,) we have $E(T_{k(n)}/k(n)) \to \mu$, and then we prove that the subsequence $\{T_{k(n)}\}$ converges to μ.

To extend the convergence from the subsequence to the whole sequence, we note that for $k(n) \le m < k(n+1)$, by the non-negativity of Y_m, we have

$$\frac{k(n)}{k(n+1)}\frac{T_{k(n)}}{k(n)} = \frac{T_{k(n)}}{k(n+1)} \le \frac{T_m}{m} \le \frac{T_{k(n+1)}}{k(n)} = \frac{k(n+1)}{k(n)} \le \frac{T_m}{m} \le \frac{T_{k(n+1)}}{k(n)+1}.$$

Using the property that $k(n+1)/k(n) \to \alpha$ as $n \to \infty$, we derive that

$$\frac{1}{\alpha}\mu \le \liminf_{m\to\infty} \frac{T_m}{m} \le \limsup_{m\to\infty} \frac{T_m}{m} \le \alpha\mu.$$

Since $\alpha > 1$ can be arbitrarily close to 1, we derive the desired almost sure convergence for T_m/m. □

5 Weak convergence

We have learnt the convergence in probability and the almost sure convergence. Although they are defined as $X_n \to X$ where X is a random variable, in previous applications we took X to be a constant number. The constant "random" variable is the only random variable that can be determined by its distribution function. Other random variables cannot. For example, in the simplest case that the probability space is $(\Omega = (\text{head}, \text{tail}), \mathcal{F} = \{\emptyset, \Omega, \{\text{head}\}, \{\text{tail}\}\}, P(\text{head}) = P(\text{tail}) = 1/2)$, the random variables X and X', defined as

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = \text{head}, \\ 0 & \text{if } \omega = \text{tail}, \end{cases} \quad X'(\omega) = \begin{cases} 0 & \text{if } \omega = \text{head}, \\ 1 & \text{if } \omega = \text{tail}. \end{cases}$$

Both have the distribution function

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1/2 & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x \geq 0, \end{cases}$$

and they are both Bernoulli random variables. Actually, in many cases we do not need the information of the random variable other than its distribution function. (X and X' are equally useful in practice.) Recall that for random variable whose distribution functions are exactly the same, like X and X' above, we say they are equal in distribution. But how to understand the statement that two random variables are approximately equal in distribution? More importantly, how to describe that a sequence of random variables X_n converge to X in distribution? One obvious way to describe the convergence in distribution is by the convergence of their distribution functions. As an example, we let X_n be the random variables on the $\{\text{head}, \text{tail}\}$ probability space just described, and let

$$X_n = \begin{cases} 1 & \text{if } \omega = \text{head}, \\ 1/n & \text{if } \omega = \text{tail}. \end{cases}$$

Then X_n converges to X a.s. and then in probability. It would be unreasonable if $\{X_n\}$ fails to converge to X in probability. But the distribution function of X_n is

$$F_n(x) = \begin{cases} 0 & \text{if } x < 1/n, \\ 1/2 & \text{if } 1/n \leq x \leq 1, \\ 1 & \text{if } x \geq 0. \end{cases}$$

Although the graph of F_n approaches that of F in an obvious way, we have that if we measure the distance between F_n and F by the maximal norm,

$$\|F_n - F\|_\infty \geq |F_n(0) - F(0)| = 1/2.$$

So in this sense, $\{F_n\}$ does not converge to F.

Definition 6. A sequence of random variables X_n, whose distribution functions are F_n, converges to a random variable X, whose distribution function is F, if $F_n(x) \to F(x)$ at all continuous points of F. In this case, we also say the sequence of distirbution functions $\{F_n\}$ converges to F.

We denote $X_n \Rightarrow X$ ($F_n \Rightarrow F$ respectively) if $X_n \to X$ ($F_n \to F$ respectively) in distribution.

The different between this definition and the heuristic understanding of the convergence in distribution is not as huge as it seems to be.

Lemma 39. *A non-decreasing function $f : \mathbb{R} \to \mathbb{R}$ can be discontinuous at countably many points at most.*

Proof. At any point x that f is not continuous, we have $a_x := \lim_{y \uparrow x} f(y) < \lim_{y \downarrow x} f(y) =: b_x$, and these intervals (a_x, b_x) for all discontinuous points are disjoint. There can be at most countably many disjoint open intervals on the real line, since each such interval contains a rational number, and then the number of disjoint open intervals is no more than the number of rational numbers. Thus the number of discontinuous points of f is at most countable. □

Remark 4. The convergence in distribution is also called *weak* convergence. The term weak has two meanings. One is intuitive: It is weaker than the convergence in probability, and then weaker than the almost sure convergence. The other meaning is that it is related to the weak* convergence in functional analysis. In the following theorem, if we interpret the bounded continuous functions as linear functionals, then the convergence of the sequence of random variables is the weak* convergence in the Banach space that we have not defined yet.

Theorem 40. *If random variables $\{X_n\}$ converge to X in probability, then they converge to X weakly.*

Proof. Let $F(x)$ be the distribution function of X, and x be a continuous point of F, that is, for all $\epsilon > 0$, there is $\delta > 0$ such that $0 \le F(x + \delta) - F(x) < \epsilon$ and $0 \le F(x) - F(x - \delta) < \epsilon$. Then since $X_n \to X$ in probability, there exists N such that for all $n > N$, $P(|X_n - X| > \delta) < \epsilon$. Then for $n > N$

$$\begin{aligned}|F_n(x) - F(x)| &\le P(X_n \le x, X > x) + P(X_n > x, X \le x) \\ &\le P(X_n \le x, X > x + \delta) + P(x \le X \le x + \delta) \\ &\quad + P(X_n > x, X \le x - \delta) + P(x - \delta < X \le x) \\ &\le \epsilon + \epsilon + \epsilon + \epsilon = 4\epsilon,\end{aligned}$$

and we prove the desired convergence. □

A direct consequence of this result is that if $\{X_n\} \to X$ a.s., then $X_n \Rightarrow X$. We have the following result which is in a sense the converse of the statement above.

Theorem 41. *If distribution functions $F_n \Rightarrow F_\infty$, then there are random variables $\{X_n\}$ and X_∞ with distribution functions F_n and F, such that $X_n \to X$ a.s..*

Proof. We construct X_*, $* = n$ or ∞, on the same probability space $(\Omega, \mathcal{F}, P)$ as follows. Let $\Omega = (0, 1)$, the interval on $\mathbb{R}$, $\mathcal{F} =$ Borel sets on $(0, 1)$, and $P = \lambda$, the Lebesgue measure. Let

$$X_*(x) = \sup\{y \mid F_*(y) < x\}.$$

So $X_*(x)$ is a non-decreasing function, and then it is a Lebesgue measurable function and a well-defined random variable. On the other hand,

$$P(X_* \leq x) = \int_0^1 1_{X_*(t)\leq x} dt = \int_0^1 1_{\sup\{y|F_*(y)<t\}\leq x} dt = \int_0^1 1_{F(x)\geq t} dt = F(x).$$

Note that in one step we used the argument "If for all y that $F_*(y) < t$ we have $y \leq x$, then $F_*(x) \geq t$", and it is based on the right-continuity of F_*.

For any $t \in (0,1)$, it is in one of the following three cases:

1. There is a unique x such that $F_\infty(x) = t$.

2. t is not in the range of F_∞.

3. There are $x_1 < x_2$ such that $F_\infty(x_1) = F_\infty(x_2) = t$.

In Case 1, we have $X_\infty(t) = \sup\{y \mid F(y) < t\} = x$. Also we have that for any $\epsilon > 0$, there is $\delta > 0$ such that

$$F_\infty(x+\epsilon) > F_\infty(x) + \delta = t + \delta, \quad F_\infty(x-\epsilon) < F_\infty(x) - \delta = t - \delta.$$

By Lemma 39, we have that there exist $x_1 \in (x - 2\epsilon, x-\epsilon)$ and $x_2 \in (x+\epsilon, x+2\epsilon)$ such that F_∞ is continuous at x_1 and x_2. Then by the convergence in probability, $F_n(x_1) \to F_\infty(x_1) < F_\infty(x) - \delta$ and $F_n(x_2) \to F_\infty(x_2) > F_\infty(x) + \delta$ as $n \to \infty$. We have that there exists N such that for all $n > N$,

$$F_n(x_1) \leq F_\infty(x) - \delta = t - \delta, \quad F_n(x_2) \geq F_\infty(x) + \delta = t + \delta.$$

Thus we have that for $n > N$

$$X_n(t-\delta) \geq x_1 > x - \epsilon, \quad X_n(t+\delta) \leq x_2 < x + 2\epsilon,$$

which imply that $x - 2\epsilon < X_n(t) < x + 2\epsilon$. Since ϵ is arbitrary, we conclude that $X_n(t) \to X_\infty(t)$.

In Case 2, let $x = \inf\{y \mid F(y) \geq t\}$. By the right continuity of F_∞, we have that $F(x) = t' > t$ and then $X_\infty(t) = x$. Again we have that for all $\epsilon > 0$, there is $\delta > 0$ such that $F(x+\epsilon) > t + \delta$ and $F(x-\epsilon) < t - \delta$. So we repeat the argument for Case 1 and derive that $X_n(t) \to X_\infty(t)$.

In Case 3, we cannot show that $X_n(t) \to X(t)$. But in this case, $F_\infty(x)$ is a constant on the open interval (x_1, x_2). Similar to the proof of Lemma 39, we can show that for a non-decreasing function f, the inverse image $f^{-1}(t)$ can contain an open interval for at most countably many t. Thus the set of t in Case 3 is at most countable, and these t does not affect the a.s. convergence of $\{X_n\}$ to X_∞. □

As an application of Theorem 41, we can prove the alternative characterization of weak convergence.

Theorem 42. *The sequence of random variables $X_n \Rightarrow X_\infty$ if and only if for every bounded continuous function g, we have $Eg(X_n) \to Eg(X_\infty)$.*

Proof. First we prove that the convergence in distribution implies the convergence of expectation of $g(X_n)$. Recall that $Eg(X_n) = \int g(x)\mu_n(dx)$ where μ_n is the distribution of the random variable X_n, which is determined by the distribution function F_n of X_n. So $Eg(X_n)$ ($Eg(X_\infty$ respectively) is determined by the distribution function F_n (F_∞ respectively). Therefore if $Y_n \overset{d}{=} X_n$ and $Y_\infty \overset{d}{=} X_\infty$, and we can show that $Eg(Y_n) \to Eg(Y_\infty)$, it implies that $Eg(X_n) \to Eg(X_\infty)$.

Hence we can take Y_n and Y_∞ on the same probability space and $Y_n \to Y_\infty$ a.s., by Theorem 41. In this special case, the result is a direct consequence of dominated convergence theorem.

Now prove the other part of the theorem, that is, if $Eg(X_n) \to Eg(X_\infty)$ for all bounded continuous g, then at any x where F_∞ is continuous, $F_n(x) \to F_\infty(x)$. Due to the continuity of F_∞ at x, for all $\epsilon > 0$, there is $\delta > 0$ such that $F_\infty(x) - \epsilon < F_\infty(x-\delta) \leq F_\infty(x) \leq F_\infty(x+\delta) < F_\infty(x) + \epsilon$, or equivalently,

$$F_\infty(x) - \epsilon < P(X_\infty \leq x - \delta) \leq P(X_\infty \leq x) \leq P(X_\infty \leq x + \delta) < F(x) + \epsilon.$$

Now consider the function f_δ and g_δ that are defined as

$$f_\delta(t) = \begin{cases} 1 & \text{if } t \leq x - \delta, \\ 0 & \text{if } t \geq x, \\ \frac{t-(x-\delta)}{\delta} & \text{if } x - \delta < t < x, \end{cases} \qquad g_\delta(t) = \begin{cases} 1 & \text{if } t \leq x, \\ 0 & \text{if } t \geq x + \delta, \\ \frac{t-x}{\delta} & \text{if } x < t < x + \delta. \end{cases}$$

The assumption of the theorem implies that

$$\limsup_{n\to\infty} F_n(x) = \limsup_{n\to\infty} P(X_n \leq x) \geq \lim_{n\to\infty} Ef_\delta(X_n) = Ef_\delta(X_\infty) \geq P(X_\infty \leq x - \delta) > F(x) - \epsilon,$$
$$\liminf_{n\to\infty} F_n(x) = \liminf_{n\to\infty} P(X_n \leq x) \leq \lim_{n\to\infty} Eg_\delta(X_n) = Eg_\delta(X_\infty) \leq P(X_\infty \leq x + \delta) > F(x) + \epsilon.$$

By the arbitrariness of ϵ, we obtain that $F_n(x) \to F_\infty(x)$. □

The following theorem is another application of Theorem 41. It is called "continuous mapping" theorem, since it shows that if $X_n \Rightarrow X_\infty$ and g is continuous, then $g(X_n) \Rightarrow g(X_\infty)$. The continuous mapping theorem has its counterparts with the "convergence in distribution" replace by "convergence in probability" and "almost sure convergence". The statements and proofs for the other two versions are left to you.

Theorem 43 (Continuous mapping). *Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function and $D_g = \{x \mid g$ is discontinuous at $x\}$. If $X_n \Rightarrow X_\infty$ and $P(X_\infty \in D_g) = 0$, then $g(X_n) \Rightarrow g(X_\infty)$. If in addition g is bounded, then $Eg(X_n) \to Eg(X_\infty)$.*

Proof. Like in the proof of "only if" part of Theorem 42, without loss of generality we assume that $X_n \to X_\infty$ a.s.. We denote $E_1 = \{\omega \mid X_n(\omega) \not\to X_\infty(\omega)\}$, and $E_2 = X_\infty^{-1}(D_g)$. Both E_1 and E_2 are of probability 0, and if $\omega \in \Omega \setminus (E_1 \cup E_2)$, then $X_n(\omega) \to X_\infty(\omega)$, and by the continuity of g at $X_\infty(\omega)$ we have $g(X_n(\omega)) \to g(X_\infty(\omega))$. Thus we show that almost surely (except for $E_1 \cup E_2$) $g(X_n) \to g(X_\infty)$, and so $g(X_n) \Rightarrow g(X_\infty)$.

The other part of the theorem is straightforward and is left for you. □

Next we consider more equivalent forms of the weak convergence condition. They together are called the *portmanteau* theorem.

Theorem 44 (Portmanteau). *The following statements are equivalent:*

(a) $X_n \Rightarrow X_\infty$.

(b) For all open sets $G \subseteq \mathbb{R}$, $\liminf_{n\to\infty} P(X_n \in G) \geq P(X_\infty \in G)$.

(c) For all closed sets $K \subseteq \mathbb{R}$, $\limsup_{n\to\infty} P(X_n \in K) \leq P(X_\infty \in K)$.

(d) For all sets A *with* $P(X_\infty \in \partial A) = 0$, $\lim_{n\to\infty} P(X_n \in A) = P(X_\infty \in A)$, *where* $\partial A = \bar{A} \setminus \operatorname{int} A$, *the sets of points that are in the closure of* A *but not the interior of* A.

Proof.

(a) ⇒ (b) By the same reason as in the proof of Theorem 42, we assume that $X_n \to X_\infty$ a.s. without loss of generality. Consider the indicator functions

$$f_*(\omega) = 1_{X_*(\omega)\in G}, \quad * = n \text{ or } \infty.$$

If $X_n(\omega) \to X_\infty(\omega)$ and $X_\infty(\omega) \in G$, then eventually $X_n(\omega) \in G$ and then $f_n(\omega) = f_\infty(\omega) = 1$ for large enough n. If $X_n(\omega) \to X_\infty(\omega)$ and $X_\infty(\omega) \notin G$, then we do not have $X_n(\omega) \to X_\infty(\omega)$ generally, but nevertheless $f_n(\omega) \geq f_\infty(\omega) = 0$ for all n. Hence we have $\liminf_{n\to\infty} f_n(\omega) \geq f_\infty(\omega)$ a.s., since $X_n \to X_\infty$ a.s.. By Fatou's lemma,

$$\begin{aligned}\liminf_{n\to\infty} P(X_n \in G) &= \liminf_{n\to\infty} \int_\Omega f_n(\omega)dP(\omega) \geq \int_\Omega \liminf_{n\to\infty} f_n(\omega)dP(\omega) \\ &\geq \int_\Omega f_\infty(\omega)dP(\omega) = P(X_\infty \in G).\end{aligned}$$

(b) ⇐ (c) Let $G = K^c$, and use that $P(X_* \in K) = 1 - P(X_* \in G)$ where $* = n$ or ∞.

(b) + (c) ⇒ (d) For any subset $A \subseteq \mathbb{R}$, we have $\operatorname{int} A \subseteq A \subseteq \bar{A}$, $\operatorname{int} A$ is open, and $\bar{A}$ is closed. Then we have

$$P(X_\infty \in \operatorname{int} A) \leq \liminf_{n\to\infty} P(X_n \in \operatorname{int} A) \quad \text{and} \quad \limsup_{n\to\infty} P(X_n \in \bar{A}) \leq P(X_\infty \in \bar{A}).$$

The difference between $P(X_\infty \in \bar{A})$ and $P(X_\infty \in \operatorname{int} A)$ is $P(X_\infty \in \partial A) = 0$, so we have $P(X_\infty \in \bar{A}) = P(X_\infty \in \operatorname{int} A) = P(X_\infty \in A)$. Hence

$$\begin{aligned}P(X_\infty \in A) \leq \liminf_{n\to\infty} P(X_n \in \operatorname{int} A) &\leq \liminf_{n\to\infty} P(X_n \in A) \\ &\leq \limsup_{n\to\infty} P(X_n \in A) \leq \limsup_{n\to\infty} P(X_n \in \bar{A}) \leq P(X_\infty \in A),\end{aligned}$$

and we have that $\lim_{n\to\infty} P(X_n \in A) = P(X_\infty \in A)$, the desired result.

(d) $\Rightarrow$ (a) For any $x \in \mathbb{R}$ such that F_∞ is continuous at x, we take $A = (-\infty, x]$ and then $\partial A = \{x\}$ and $P(X_\infty \in \partial A) = 0$. Thus we have $\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P(X_n \in A) = P(X_\infty \in A) = F_\infty(x)$, and check that $X_n \Rightarrow X_\infty$ by Definition 6. □

The weak* topology has an important property, the Banach-Alouglu theorem, such that the closed ball of the dual space of a normed space is compact. If you find the terminologies in the statement above arcane, do not worry, we are not going to use it anywhere in our module, but state the following result in an intelligible way.

Theorem 45 (Helly's selection)**.** *For every sequence F_n of distribution functions, there is a subsequence, and a right-continuous non-decreasing function F, so that at any point x where F is continuous, the value of the functions in the subsequence converges to $F(x)$.*

Proof. First we construct the subsequence and F, and then we prove that they satisfy the conditions. Let $\{r_1, r_2, \dots\}$ be an ordering of all rational numbers.

1. Let $\{F_{n_1(1)}, F_{n_1(2)}, F_{n_1(3)}, \dots\}$ be the subsequence of the original sequence $\{F_n\}$ such that $F_{n_1(k)}(r_1)$ converges, and denote the limit $f(r_1)$.

2. Let $\{F_{n_2(1)}, F_{n_2(2)}, F_{n_2(3)}, \dots\}$ be the subsequence of the sequence $\{F_{n_1(k)}\}$ such that $F_{n_2(k)}(r_2)$ converges, and denote the limit $f(r_2)$.

3. Let $\{F_{n_3(1)}, F_{n_3(2)}, F_{n_3(3)}, \dots\}$ be the subsequence of the sequence $\{F_{n_2(k)}\}$ such that $F_{n_3(k)}(r_3)$ converges, and denote the limit $f(r_3)$.

4. ....

At last, we choose the "diagonal" subsequence $\{F_{n_1(1)}, F_{n_2(2)}, \dots, F_{n_k(k)}, \dots\}$ as the desired subsequence. The limit function F is constructed by f as

$$F(x) = \inf_{y\in\mathbb{Q}, y\geq x} f(y).$$

It is clear that for all r_l, $f(r_l) \in [0,1]$, and as a mapping from $\mathbb{Q}$ to $[0,1]$, f is non-decreasing. So $F(x)$ is well-defined for all $x \in \mathbb{R}$, and it is easy to check that F is non-decreasing and right-continuous, and $F(y) = f(y)$ if $y \in \mathbb{Q}$.

Let x be a continuity point of F. For any $\epsilon > 0$, there are $y_1, y_2 \in \mathbb{Q}$ such that $y_1 < x$, $y_2 > x$, and $F(y_1) > F(x) - \epsilon$, $F(y_2) < F(x) + \epsilon$. If k is large enough, we have, by the convergence of $F_{n_k(k)}$ to f at rational points, $F_{n_k(k)}(y_1) < F(x) - \epsilon$ and $F_{n_k(k)}(y_2) < F(x) + \epsilon$. Then

$$F_{n_k(k)}(x) \in [F_{n_k(k)}(y_1), F_{n_k(k)}(y_2)] \subseteq (F(x) - \epsilon, F(x) + \epsilon).$$

By the arbitrariness of ϵ, we prove that $F_{n_k(k)}(x) \to F(x)$ as $k \to \infty$. □

Remark 5. As a caveat, we should stress that the limit function F in Theorem 45 may not be a distribution function, since it may not satisfy $F(-\infty) = 0$ and $F(+\infty) = 1$. For an example, let $F_n(x) = 0$ for $x < n$ and $F_n(x) = 1$ for $x \geq n$. Then no matter what subsequence we choose, the limit function F is always the constant function $F(x) = 0$. We call the convergence in the sense of Theorem 45 the *vague* convergence, and write

$$F_{n_k(k)} \overset{v}{\Rightarrow} F.$$

Theorem 46. *For a sequence of distribution functions F_n, every sub-sequential limit is the distribution function of a probability measure if and only if the sequence is* tight, *that is, for all $\epsilon > 0$, there is an $M_\epsilon > 0$, such that $1 - F_n(M_\epsilon) < \epsilon$ and $F_n(-M_\epsilon) < \epsilon$ for all large enough n.*

Proof. If the tightness condition is satisfied, then for any $\epsilon > 0$, if a subsequence $F_{n_k} \overset{v}{\Rightarrow} F$, then $F_{n_k}(M_\epsilon) > 1 - \epsilon$ for all large enough k, and then for any point $x > M_\epsilon$ where F is continuous, (by Lemma 39, such x exists,) we have $F(x) = \lim_{k\to\infty} F_{n_k}(x) \geq \limsup_{k\to\infty} F_{n_k}(M_\epsilon) \geq 1 - \epsilon$, and so $F(+\infty) \geq F(x) > 1 - \epsilon$. By the arbitrariness of ϵ, we conclude that $F(+\infty) = 1$. Similarly, $F(-\infty) = 0$.

On the other hand, if the tightness condition is not satisfied, without loss of generality, there are $\{F_{n_k}\}$ and $\epsilon > 0$ such that $F_{n_k}(k) < 1-\epsilon$ for all k. A subsequence of $\{F_{n_k}\} \overset{v}{\Rightarrow} F$, and without loss of generality we assume that $F_{n_k} \overset{v}{\Rightarrow} F$. Then for any x on which F is continuous, we have $F(x) = \lim_{k\to\infty} F_{n_k}(x) \leq 1 - \epsilon$. By Lemma 39, such x is almost everywhere, and so

$$F(+\infty) = \limsup_{x\to\infty, F \text{ is continuous at } x} F(x) \leq 1 - \epsilon,$$

and this subsequence has a vague limit that is not a distribution function. □

In the end of this section, we remark that although we use the weak* convergence in functional analysis as an inspiration of the weak convergence in probability, we have not defined the exact relation between these two concepts, since we are not going to use this relation in future.

6 Characteristic functions

In analysis, a powerful tool is Fourier transform. If we want to consider the property of a function but cannot do it directly, a natural strategy is to consider its Fourier transform. For example, given two (square-integrable and continuous) functions f and g, we want to compute the *convolution*

$$f * g(x) = \int_{-\infty}^{\infty} f(y)g(x-y)dy.$$

Besides the direct computation, the most common method is to compute the Fourier transforms of f and g

$$\check{f}(t) = \int_{-\infty}^{\infty} f(x)e^{itx}dx, \quad \check{g}(t) = \int_{-\infty}^{\infty} g(x)e^{itx}dx.$$

Then use the property that the Fourier transform of the convolution is the product of the Fourier transforms, that is,

$$(f * g)^{\vee}(t) = \check{f}(t)\check{g}(t),$$

and take the inverse Fourier transform

$$f * g(x) = (fg)^{\vee}\widehat{}= (\check{f}\check{g})\widehat{}= \frac{1}{2\pi}\int_{-\infty}^{\infty} \check{f}(t)\check{g}(t)e^{-itx}dt.$$

(Here the introduction to Fourier transform is flawed. Usually the transform $f \to \check{f}$ is called the *inverse Fourier transform*, and the transform $f : \hat{f}$ is called the *Fourier transform*. Our choice of the prefactors is also not the common one.)

The Fourier transform is applied in probability theory under the name of characteristic function.

Definition 7. Let X be a random variable, we define its *characteristic function* by

$$\varphi(t) = E(e^{itX}) = E(\cos(tX)) + iE(\sin(tX)).$$

Suppose the distribution of X is μ, which is a probability measure on $\mathbb{R}$, then

$$\varphi(g) = \int e^{itx}\mu(dx).$$

Since the characteristic function of a random variable is expressed in its distribution, we can also say the characteristic function is associated to the distribution (function). If X is a continuous random variable with density function $f(x)$, that is, $\mu(dx) = f(x)dx$, we have

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

the (inverse) Fourier transform of f.

Since e^{itx} is a bounded function in x if $t \in \mathbb{R}$, we have that $\varphi(t)$ is well defined for all real t, and by definition

$$\varphi(0) = E(e^0) = 1, \quad \textit{and} \quad |\varphi(t)| = |E(e^{itX})| \leq E|e^{itX}| = E(1) = 1 \quad \text{for all } t.$$

Furthermore, we also have that $\varphi(t)$ is a uniformly continuous function in t, since

$$|\varphi(t+\epsilon)-\varphi(t)| = |E(e^{i(t+\epsilon)X}-e^{itX})| \leq E|e^{i(t+\epsilon)X}-e^{itX}| = E|(e^{i\epsilon X}-1)e^{itX}| = E|e^{i\epsilon X}-1|,$$

and as $\epsilon \downarrow 0$, $e^{i\epsilon X} - 1 \to 0$ a.s., we derive that $E|e^{i\epsilon X} - 1| \to 0$ as $\epsilon \downarrow 0$ by the dominated convergence theorem.

Recall that in an exercise we derived that if independent random variables X, Y have density functions f, g respectively, then the density function of $X + Y$ is the convolution of f and g. Inspired by the Fourier transform formula for convolutions, we can state, if not prove, the following result:

Theorem 47. *If X and Y are independent and have characteristic functions $\varphi(t)$ and $\psi(t)$ respectively, then $X + Y$ has characteristic function $\varphi(t)\psi(t)$.*

Proof. By the independence,

$$E(e^{it(X+Y)}) = E(e^{itX}e^{itY}) = E(e^{itX})E(e^{itY}) = \varphi(t)\psi(t).$$

□

One simple example of characteristic function is for a random variable X in Bernoulli distribution, such that $P(X = 0) = P(X = 1) = 1/2$. Then $\varphi(t) = \frac{1}{2}e^{it\cdot 0} + \frac{1}{2}e^{it\cdot 1} = cos(t/2)e^{it/2}$. The most important example of characteristic function is for a random variable in normal distribution $N(\mu, \sigma^2)$, where μ is the expectation and σ^2 is the variance, so that the density function is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

In the simplest case $\mu = 0$ and $\sigma^2 = 1$, we have

$$\varphi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}+itx}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(x-it)^2}{2}}e^{-\frac{t^2}{2}}dx = e^{-\frac{t^2}{2}}\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty-it}^{\infty-it} e^{-\frac{t^2}{2}}dz\right).$$

where the integral of z is on a contour in the complex plane that is parallel to the real axis. It is a standard fact that the expression in the parenthesis is 1 if $t = 0$. By a standard application of the residue theorem in complex analysis, the expression is independent of t. Thus we conclude that $\varphi(t) = e^{-t^2/2}$ in this case. In the general case, we can repeat the argument, but a faster way is to recorganise that if X is in $N(0, 1)$ distribution, then $\sigma X + \mu$ is in $N(\mu, \sigma^2)$ distribution (exercise), and then use the following result

$$E(e^{it(aX+b)}) = e^{itb}E(e^{i(ta)X}).$$

Hence the characteristic function for $\sigma X + \mu$ is $e^{i\mu t}e^{-(\sigma t)^2/2} = \exp(-\frac{\sigma^2}{2}t^2 + i\mu t)$.

Now let's consider the question: If we know the characteristic function, can we recover the distribution (function) of the random variable? If we know in advance that the density function exists, then the (inverse) Fourier transform gives the density function from the characteristic function. But we know that the density function only exists for continuous random variables. We have a more complete and more sophisticated result:

Theorem 48. *Let $\varphi(t)$ be the characteristic function of a random variable whose distribution is μ. Then for any $a < b$,*

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt = \mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\}.$$

We note that the integral domain $[-T, T]$ cannot be replaced by $(-\infty, \infty)$ in general, otherwise the convergence is not guaranteed.

Proof of Theorem 48. We denote

$$I_T = \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt = \frac{1}{2\pi} \int_{-T}^{T} \left(\int_a^b e^{-ity} dy \right) \left(\int_{\mathbb{R}} e^{itx} \mu(dx) \right) dt.$$

Since $|e^{-ity} e^{itx}| \leq 1$ and

$$\int_{-T}^{T} \int_a^b \int_{\mathbb{R}} 1\mu(dx) \times dy \times dt = 2T(b-a) < \infty,$$

we can apply Fubini's theorem and write

$$\begin{aligned} I_T &= \int_{\mathbb{R}} \mu(dx) \left(\frac{1}{2\pi} \int_a^b dy \int_{-T}^{T} dt e^{it(x-y)} \right) = \int_{\mathbb{R}} \mu(dx) \int_a^b dy \frac{e^{iT(x-y)} - e^{-iT(x-y)}}{2\pi i (x-y)} \\ &= \int_{\mathbb{R}} \mu(dx) \int_a^b dy \frac{\sin(T(x-y))}{\pi(x-y)} \\ &= \int_{\mathbb{R}} \mu(dx) (R(T(x-a)) - R(T(x-b))), \end{aligned}$$

where

$$R(x) = \int_0^x \frac{\sin s}{\pi s} ds.$$

It is a tricky result in calculus that

$$\lim_{x\to\infty} R(x) = \int_0^\infty \frac{\sin s}{\pi s} ds = \frac{1}{2}, \quad \lim_{x\to-\infty} R(x) = -\int_0^\infty \frac{\sin s}{\pi s} ds = -\frac{1}{2}.$$

So as $T \to \infty$,

$$\lim_{T\to\infty} R(T(x-a)) - R(T(x-b)) = \begin{cases} 0 & \text{if } x < a \text{ or } x > b, \\ 1 & \text{if } a < x < b, \\ \frac{1}{2} & \text{if } x = a \text{ or } x = b. \end{cases}$$

On the other hand, it is not hard to see that there exists $C > 0$ such that $-C < R(x) < C$ for all x, and then $|R(T(x-a)) - R(T(x-b))| < 2C$ for all x, T. Thus by the dominated convergence theorem,

$$\lim_{T\to\infty} I_T = \int_{\mathbb{R}} \mu(dx) (\chi_{a<x<b} + \frac{1}{2}\chi_{x=a} + \frac{1}{2}\chi_{x=b}) = \mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\}.$$

□

We remark again that the cumbersome notation $\lim_{T\to\infty}\int_{-T}^{T}$ cannot be replaced by $\int_{-\infty}^{\infty}$ in Theorem 48. (Actually the former notation is the Cauchy principal value of that generalises the latter.) But if the characteristic function is integrable, that is, $\int_{-\infty}^{\infty}|\varphi(t)|dt < \infty$, then we can replace the Cauchy principal value simply by $\int_{-\infty}^{\infty}$. Furthermore, we have the following result.

Theorem 49. *If $\int|\varphi(t)|dt < \infty$, then μ has bounded continuous density*

$$f(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\varphi(t)dy.$$

Proof. First, $f(y)$ is well defined and bounded, since $|e^{-ity}\varphi(t)| = |\varphi(t)|$, and so $|f(y)| \le \frac{1}{2\pi}\int|\varphi(t)|dt$. Next, $f(y)$ is continuous. To see it, we consider

$$|f(y+h) - f(y)| = \frac{1}{2\pi}\left|\int \left(e^{-it(y+h)} - e^{-ity}\right)\varphi(t)dt\right| \le \frac{1}{2\pi}\int|1 - e^{-ith}||\varphi(t)|dt.$$

As $h \to 0$, the factor $|1 - e^{-ith}| \to 0$. Since the integrand on the right-hand side of the formula above is dominated by $|\varphi(t)|$, by the dominated convergence theorem, we have $|f(y+h) - f(y)| \to 0$ as $h \to 0$.

Next, we have that for all $a < b$

$$\begin{aligned}\mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\} &= \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_a^b e^{-ity}dy\right)\varphi(t)dt \\ &= \int_a^b\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-ity}\varphi(t)dt\right)dy \\ &= \int_a^b f(y)dy.\end{aligned}$$

It is clear that μ has no pointmass and $f(y)$ is the density function of μ. □

The next theorem shows that characteristic functions are useful tool to analyse weak convergence. Here and later, when we say the weak convergence of distributions, we mean the weak convergence of the corresponding distribution functions.

Theorem 50. *Let $\mu_1, \mu_2, \ldots$ be distributions, and $\varphi_1(t), \varphi_2(t), \ldots$ be characteristic functions associated to them.*

(a) If $\mu_n \Rightarrow \mu_\infty$ where μ_∞ is a distribution with characteristic function $\varphi_\infty(t)$, then $\varphi_n(t) \to \varphi_\infty(t)$ pointwise.

(b) If $\varphi_n(t) \to \varphi_\infty(t)$ pointwise, and the limit function $\varphi_\infty(t)$ is continuous at point 0, then $\varphi_\infty(t)$ is the characteristic function for a distribution, say μ_∞, and $\mu_n \Rightarrow \mu_\infty$.

Proof. Part (a) is a direct consequence of Theorem 42 where the bounded continuous function is e^{itx}.

To prove part (b), we denote the distribution function for μ_n by F_n. By Helly's selection theorem, we have that a subsequence $\{F_{n_k}\}$ converges to a limit F_∞ *vaguely. Suppose we have that the sequence* $\{F_n\}$ *is tight.* Then F_∞ is a distribution function, corresponding to a distribution μ_∞, and then $F_n \Rightarrow F_\infty$, or equivalently, $\mu_n \Rightarrow \mu_\infty$, and by part (a) we have that the characteristic function of μ_∞ is $\varphi_\infty(t)$. Although we have just proved the result for a subsequence $\{\mu_{n_k}\}$, the result can be extended to the sequence $\{m_n\}$. Suppose not, then there exists a bounded continuous function f and another subsequence $\{m_{m_k}\}$ such that

$$\left|\int f(x)m_{m_k}(dx) - \int f(x)m_\infty(dx)\right| > \epsilon$$

for some $\epsilon > 0$. Using Helly's selection theorem and the tightness of $\{F_n\}$, we have that a further subsequence $\{m_{m_{k(l)}}\}$ converge weakly to $m'_\infty \neq m_\infty$. Then by part (a), $\varphi_{m_{k(l)}}(t) \to \varphi'_\infty(t)$, the characteristic function of m'_∞, and it is different from $\varphi_\infty(t)$. Then we derive a contradiction.

Thus the remaining part of the proof is to show that $\{F_n\}$ is tight, that is, for any $\epsilon > 0$, there is an M such that $\mu_n(-M, M) > 1-\epsilon$, or equivalently, $\mu_n(\mathbb{R} \setminus (-M, M)) < \epsilon$ for all μ_n. Actually we only need to prove it for large enough n.

Consider the integral

$$I_n(M) = \int \left(1 - \frac{\sin(M^{-1}x)}{M^{-1}x}\right) \mu_n(dx).$$

Since $1 - \sin(x)/x \geq 0$ for all x, and for $|x| \geq 1$, we have $1 - \sin(x)/x > c > 0$ where c is a positive constant (say 1/10), we derive the inequality

$$I_n(M) \geq \int_{\mathbb{R}\setminus(-M,M)} \left(1 - \frac{\sin(M^{-1}x)}{M^{-1}x}\right) \mu_n(dx) > c\mu_n(\mathbb{R} \setminus (-M, M)).$$

On the other hand,

$$1 - \frac{\sin(M^{-1}x)}{M^{-1}x} = 1 - \frac{e^{iM^{-1}x} - e^{-iM^{-1}x}}{2iM^{-1}x} = 1 - \frac{M}{2}\int_{-M^{-1}}^{M^{-1}} e^{ixt}dt = \frac{M}{2}\int_{-M^{-1}}^{M^{-1}} (e^{ix\cdot 0} - e^{ixt})dt,$$

and then by Fubini's theorem

$$\begin{aligned} I_n(M) &= \frac{M}{2}\int \mu_n(dx) \int_{-M^{-1}}^{M^{-1}} (e^{ix\cdot 0} - e^{ixt})dt \\ &= \frac{M}{2}\int_{-M^{-1}}^{M^{-1}} \left(\int e^{ix\cdot 0}\mu_n(dx) - \int e^{ixt}\mu_n(dx)\right) dt \\ &= \frac{M}{2}\int_{-M^{-1}}^{M^{-1}} (\varphi_n(0) - \varphi_n(t))dt. \end{aligned}$$

Since we assume that $\varphi_\infty(t)$ is continuous at 0, for large enough M, we have

$$\frac{M}{2}\int_{-M^{-1}}^{M^{-1}} (\varphi_\infty(0) - \varphi_\infty(t))dt < c^{-1}\epsilon.$$

Then using the pointwise convergence of $\{\varphi_n(t)$ to $\varphi_\infty(t)$ and the dominated convergence theorem, for the same M, if n is large enough,

$$I_n(M) = \frac{M}{2}\int_{-M^{-1}}^{M^{-1}} (\varphi_n(0) - \varphi_n(t))dt < c^{-1}\epsilon,$$

and we conclude that $\mu_n(\mathbb{R} \setminus (-M, M)) < \epsilon$ for such M if n is large enough. Thus we prove the tightness. □

Characteristic functions are handy tools to analyse moments of random variables. Formally,

$$\varphi(t) = \int e^{itx}\mu(dx) = \int \sum_{k=0}^{\infty} \frac{t^k}{k!} x^k \mu(dx) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \int x^k \mu(dx) = 1 + itEX - \frac{t^2}{2}EX^2 + \cdots.$$

But this is only a formal argument. One problem is the change of order of the infinite sum and integral, which keeps pestering us since we began studying calculus. Another trouble is that EX^k may not always exist. So the following result is non-trivial:

Theorem 51. *If $EX^2 < \infty$, then the characteristic function $\varphi(t)$ for X satisfies*

$$\lim_{t\to 0} \frac{1}{t^2}\left(\varphi(t) - \left(1 + itEX - \frac{t^2}{2}EX^2\right)\right) = 0.$$

or in other words, $\varphi(t) = 1 + itEX - (t^2/2)EX^2 + o(t^2)$.

Proof. We need to estimate

$$\begin{aligned}\frac{1}{t^2}\left|\varphi(t) - \left(1 + itEX - \frac{t^2}{2}EX^2\right)\right| &= \frac{1}{t^2}\left|\int \left(e^{itx} - \left(1 + itx + \frac{(itx)^2}{2}\right)\right)\mu(dx)\right| \\ &\leq \int \frac{1}{t^2}|R(tx)|\mu(dx),\end{aligned}$$

where $R(y)$ is the remainder term of the second order Taylor expansion for the function e^{iy} at 0. By l'Hôpital's rule, $t^{-2}R(tx) \to 0$ as $t \to 0$ for all x. So if we can show that there exists an integrable function that dominates $t^{-2}R(tx)$ for all t, then the proof is done by the dominated convergence theorem. Actually we can take the desired integrable function as x^2. The integrability of x^2 is equivalent to $EX^2 = \int x^2\mu(dx) < \infty$, and the dominance is the consequence of the following lemma. □

Lemma 52. *For all $y \in \mathbb{R}$, $|R(y)| \leq y^2$.*

Proof. First, for $|y| \geq 4$, we have

$$|R(y)| = \left|e^{iy} - \left(1 + iy + \frac{(iy)^2}{2}\right)\right| \leq |e^{iy}| + 1 + |y| + \frac{y^2}{2} = 2 + y + \frac{y^2}{2} \leq \frac{y^2}{8} + \frac{y^2}{4} + \frac{y^2}{2} \leq y^2.$$

Next, for $|y| < 4$, we recall that the remainder term in Taylor expansion $R(y)$ has the integral form

$$|R(y)| = \left|\int_0^y \frac{(e^{it})'''}{2!}(y-t)^2 dt\right| \leq \pm\int_0^y \frac{1}{2}(y-t)^2 dt = \frac{|y|^3}{6} \leq y^2,$$

where ± is the sign of y. Hence we prove the lemma. □

The converse to Theorem 51 is also true in a certain form. To be precise, we have the following theorem:

Theorem 53. *For a random variable X, $EX^2 < \infty$ under the condition on its characteristic function $\varphi(t)$:*

$$\liminf_{h\downarrow 0} \frac{1}{h^2}(2-\varphi(h)-\varphi(-h)) < +\infty.$$

Proof. Noting that

$$\frac{1}{h^2}(2-\varphi(h)-\varphi(-h)) = \int \frac{1}{h^2}(2-e^{ihx}-e^{-ihx})\mu(dx) = \int \frac{2-2\cos(hx)}{h^2}\mu(dx),$$

and that $h^{-2}(2-2\cos(hx))$ is non-negative and $h^{-2}(2-2\cos(hx)) \to x^2$ as $h \downarrow 0$, we have by Fatou's lemma

$$EX^2 = \int x^2\mu(dx) \le \liminf_{h\downarrow 0} \int \frac{2-2\cos(hx)}{h^2}\mu(dx) = \liminf_{h\downarrow 0} \frac{1}{h^2}(2-\varphi(h)-\varphi(-h)) < +\infty,$$

and prove the theorem. □

7 Central limit theorems

The hard work on characteristic functions is rewarded when we find that the proof of the celebrated central limit theorem is an easy application of the properties of the characteristic functions.

Theorem 54. *Let $X_1, X_2, \dots$ be i.i.d. with $EX_i = \mu$ and $\operatorname{var}(X_i) = \sigma^2 \in (0, \infty)$. If $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n - n\mu}{\sigma n^{1/2}} \Rightarrow \chi,$$

where χ has the standard normal distribution $N(0,1)$.

Proof. By Theorem 50, we only need to show that the characteristic function of $(S_n - n\mu)/(\sigma n^{1/2})$, which we denoted as $\varphi_n(t)$, converges pointwise at $e^{-t^2/2}$, the characteristic function of χ. We denote $Y_n = X_n - \mu$, which are i.i.d. random variables, and denote their common characteristic function by $\varphi(t)$. Then

$$\begin{aligned}\varphi_n(t) = E\left(\exp\left(it\frac{S_n - n\mu}{\sigma n^{1/2}}\right)\right) &= E\left(\prod_{k=1}^{n} \exp\left(it\frac{Y_k}{\sigma n^{1/2}}\right)\right) = \prod_{k=1}^{n} E\left(\exp\left(it\frac{Y_k}{\sigma n^{1/2}}\right)\right) \\ &= \varphi\left(\frac{t}{\sigma n^{1/2}}\right)^n.\end{aligned}$$

Since Y_i has the first moment 0 and second moment σ^2, we have, by Theorem 51,

$$\varphi\left(\frac{t}{\sigma n^{1/2}}\right) = 1 - \frac{EY^2}{2}\left(\frac{t}{\sigma n^{1/2}}\right)^2 + o\left(\left(\frac{t}{\sigma n^{1/2}}\right)^2\right) = 1 - \frac{t^2}{2n} + o(n^{-1}).$$

It is a basic fact in calculus that as $x \to +\infty$, $(1 + x^{-1})^x \to e$, and so for any fixed t, as $n \to \infty$, $(1 - t^2/(2n))^n \to e^{-t^2/2}$. The $o(n^{-1})$ term in the formula above can be ignored. For any $\epsilon > 0$, we have $|o(n^{-1})| < \epsilon/n$ for large enough n, and then

$$\begin{aligned}\limsup_{n\to\infty}\left(1 - \frac{t^2}{2n} + o(n^{-1})\right)^n &\le \limsup_{n\to\infty}\left(1 - \left(\frac{t^2}{2} + \epsilon\right)\frac{1}{n}\right)^n = e^{-t^2/2}e^{\epsilon}, \\ \liminf_{n\to\infty}\left(1 - \frac{t^2}{2n} + o(n^{-1})\right)^n &\le \liminf_{n\to\infty}\left(1 - \left(\frac{t^2}{2} - \epsilon\right)\frac{1}{n}\right)^n = e^{-t^2/2}e^{-\epsilon}.\end{aligned}$$

By the arbitrariness of ϵ, we prove

$$\lim_{n\to\infty} \varphi_n(t) = \varphi\left(\frac{t}{\sigma n^{1/2}}\right)^n = e^{-t^2/2},$$

the characteristic function of χ, and finish the proof. □

A generalization of the central limit theorem above is the weak convergence to normal distribution for a triangular array of random variables.

Theorem 55 (Lindeberg-Feller). *For each n, let $X_{n,m}$ $(1 \le m \le n)$ be independent random variables with $EX_{n,m} = 0$. Suppose* *(i)* $\sum_{m=1}^{n} \sigma_{n,m}^2 \to \sigma^2 > 0$, *where* $\sigma_{n,m}^2 = EX_{n,m}^2$.

(ii) For all $\epsilon > 0$, $\lim_{n\to\infty} \sum_{m=1}^{n} \sigma_{n,m}^2(\epsilon) = 0$, *where* $\sigma_{n,m}^2(\epsilon) = E\left(|X_{n,m}|^2; |X_{n,m}| > \epsilon\right)$.

Then $S_n = X_{n,1} + \cdots + X_{n,n} \Rightarrow \sigma\chi = N(0, \sigma^2)$ *as* $n \to \infty$.

Remark 6. Before giving the proof, we remark that Condition (ii) implies that as $n \to \infty$, all $X_{n,m}$ converge to 0 (in distribution/probability). To be precise, we can see that given any ϵ, if n is large enough, then $\sigma_{n,m}^2 < \epsilon$ for all $m = 1, \ldots, n$ (Exercise). The central limit theorem is about the collective behaviour of many random variables. If any one of them is so big that it alone affects the whole in an non-negligible way, then the central limit theorem does not apply.

Proof. In the ideal case that all random variables $X_{n,m} = W_{n,m}$ which are independent and with normal distribution $N(0, \sigma_{n,m}^2)$. Then it is obvious that the distribution of $W_{n,1} + \cdots + W_{n,n}$ converges to $N(0, \sigma^2)$ in distribution. To see it, we have that the characteristic function of $W_{n,m}$ is $\exp(-\sigma_{n,m}^2 t^2/2)$, and the characteristic function of their sum is $\exp(-(\sigma_{n,1}^2 + \cdots + \sigma_{n,n}^2)t^2/2) \to \exp(-\sigma^2 t^2/2)$.

To prove the general result, our strategy is to compare the characteristic function of $X_{n,1} + \cdots + X_{n,n}$, which we denote as $\varphi_n(t)$, with that of $W_{n,1} + \cdots + W_{n,n}$, and show that the difference is small. Denote the characteristic function of $X_{n,m}$ by $\varphi_{n,m}(t)$, we have

$$
\begin{aligned}
&\varphi_{n,m}(t) - e^{-\sigma_{n,m}^2 t^2/2} = E(e^{itX_{n,m}}) - E(e^{itW_{n,m}}) \\
= {}& E\left(1 + itX_{n,m} - \frac{t^2}{2}X_{n,m}^2 + R(tX_{n,m})\right) - E\left(1 + itW_{n,m} - \frac{t^2}{2}W_{n,m}^2 + R(tW_{n,m})\right) \\
= {}& \left(1 + itEX_{n,m} - \frac{t^2}{2}EX_{n,m}^2 + E(R(tX_{n,m}))\right) \\
& - \left(1 + itEW_{n,m} - \frac{t^2}{2}EW_{n,m}^2 + E(R(tW_{n,m}))\right) \\
= {}& E(R(tX_{n,m})) - E(R(tW_{n,m})),
\end{aligned}
$$

where $R(x) = e^{ix} - (1 + ix - x^2/2)$ is the remainder of the Taylor expansion of e^{ix} of degree 2.

Note that

$$E(R(tX_{n,m})) = E(R(tX_{n,m}); |X_{n,m}| \leq \epsilon) + E(R(tX_{n,m}); |X_{n,m}| > \epsilon).$$

By Lemma 52, we have

$$|E(R(tX_{n,m}); |X_{n,m}| \leq \epsilon)| \leq E((tX_{n,n})^2; |X_{n,m}| \leq \epsilon) = t^2\sigma_{n,m}^2(\epsilon),$$

and by Lemma 56(a) below,

$$
\begin{aligned}
|E(R(tX_{n,m}); |X_{n,m}| > \epsilon)| &\leq E(|tX_{n,m}|^3; |X_{n,m}| > \epsilon) \leq E(t^3\epsilon X_{n,m}^2; |X_{n,m}| > \epsilon) \\
&\leq t^3\epsilon E(X_{n,m}^2) = t^3\epsilon\sigma_{n,m}^2.
\end{aligned}
$$

On the other hand, by Lemma 56(b) below,

$$|E(R(tW_{n,m}))| = |e^{-\sigma_{n,m}^2 t^2/2} - (1 - \sigma_{n,m}^2 t^2/2)| \leq (\sigma_{n,m}^2 t^2/2)^2 = \frac{t^4}{4}\sigma_{n,m}^4.$$

If n is large enough, as discussed in Remark 6, we have $\sigma_{n,m}^2 < \epsilon$, and then

$$|\varphi_{n,m}(t) - e^{-\sigma_{n,m}^2 t^2/2}| < A_{n,m},$$

where

$$A_{n,m} = t^2\sigma_{n,m}^2(\epsilon) + t^3\epsilon\sigma_{n,m}^2 + \frac{t^4}{4}\sigma_{n,m}^4 \leq t^2\sigma_{n,m}^2(\epsilon) + \left(t^3 + \frac{t^4}{4}\right)\epsilon\sigma_{n,m}^2.$$

It is easy to see that

$$\lim_{n\to\infty}\sum_{m=1}^{n} A_{n,m} = \left(t^3 + \frac{t^4}{4}\right)\epsilon\sigma^2.$$

Then we have

$$\left|\frac{\varphi_{n,m}(t)}{e^{-\sigma_{n,m}^2 t^2/2}} - 1\right| \leq e^{\sigma_{n,m}^2 t^2/2} A_{n,m} \leq e^{\epsilon t^2/2} A_{n,m},$$

and with

$$\sum_{m=1}^{n}\log(1 - e^{\epsilon t^2/2}A_{n,m}) \leq \log\left(\frac{\varphi_n(t)}{e^{-(\sigma_{n,1}^2+\cdots+\sigma_{n,n}^2)t^2/2}}\right)$$
$$= \sum_{m=1}^{n}\log\left(\frac{\varphi_{n,m}(t)}{e^{-\sigma_{n,m}^2 t^2/2}}\right) \geq \sum_{m=1}^{n}\log(1 + e^{\epsilon t^2/2}A_{n,m}).$$

By Lemma 56(c), for ϵ small enough, we have

$$\log(1 + e^{\epsilon t^2/2}A_{n,m}) \leq e^{\epsilon t^2/2}A_{n,m}, \quad \log(1 - e^{\epsilon t^2/2}A_{n,m}) \geq -2e^{\epsilon t^2/2}A_{n,m}.$$

So

$$\limsup_{n\to\infty}\log\left(\frac{\varphi_n(t)}{e^{-(\sigma_{n,1}^2+\cdots+\sigma_{n,n}^2)t^2/2}}\right) \leq \limsup_{n\to\infty}\sum e^{\epsilon t^2/2}A_{n,m} = e^{\epsilon t^2/2}\left(t^3 + \frac{t^4}{4}\right)\epsilon\sigma^2,$$
$$\liminf_{n\to\infty}\log\left(\frac{\varphi_n(t)}{e^{-(\sigma_{n,1}^2+\cdots+\sigma_{n,n}^2)t^2/2}}\right) \leq \limsup_{n\to\infty}\sum e^{\epsilon t^2/2}A_{n,m} = -2e^{\epsilon t^2/2}\left(t^3 + \frac{t^4}{4}\right)\epsilon\sigma^2.$$

Since ϵ is arbitrary, we can conclude that

$$\lim_{n\to\infty}\log\left(\frac{\varphi_n(t)}{e^{-(\sigma_{n,1}^2+\cdots+\sigma_{n,n}^2)t^2/2}}\right) = 1,$$

and then conclude that $\lim_{n\to\infty}\varphi_n(t) \to e^{-\sigma^2 t^2/2}$. Thus we prove that theorem. □

The technical results we need in the proof of Theorem 55 is collected in the following lemma.

Lemma 56. *(a) $|R(y)| \leq |y|^3$ for all $y \in \mathbb{R}$, where $R(y)$ is the same as in Lemma 52.*

(b) $|e^{-x} - (1-x)| \leq x^2$ *for all* $x \geq 0$.

(c)

$$\log(1+x) \begin{cases} \leq x & \textit{for } x \geq 0, \\ \geq 2x & \textit{for } x \in (-1/2, 0). \end{cases}$$

We only prove part (a) The other two parts are easier and are left for exercise.

Proof of Lemma 56(a). Similar to the proof of Lemma 52, we have

$$|R(y)| = \left| \int_0^y \frac{(e^{it})'''}{2!}(y-t)^2 dt \right| \leq \pm \int_0^y \frac{1}{2}(y-t)^2 dt = \frac{|y|^3}{6},$$

where ± is the sign of y. Hence we prove the lemma. □

8 Poisson convergence

We consider a weak convergence result analogous to the central limit theorem, which is the "law of rare events" where we sum up many discrete random variables which are closed to 0 individually. The weak limit of the sum is the Poisson distribution.

Definition 8. A random variable X is in *Poisson distribution with mean* λ (denoted as $\mathrm{Poisson}(\lambda)$), if the values of X are non-negative and

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

It is not hard to find that the characteristic function of $X = \mathrm{Poisson}(\lambda)$ is $\exp(\lambda(e^{it} - 1))$.

Theorem 57. *For each n, let $X_{n,m}$ ($1 \leq m \leq n$) be independent Bernoulli random variables with $P(X_{n,m} = 1) = p_{n,m}$ and $P(X_{n,m} = 0) = 1 - p_{n,m}$. Suppose*

1. $\displaystyle\sum_{m=1}^{n} p_{n,m} \to \lambda \in (0, \infty)$.
2. $\displaystyle\max_{1\leq m\leq n} p_{n,m} \to 0$.

Then the sum $S_n := X_{n,1} + \cdots + X_{n,n} \Rightarrow \mathrm{Poisson}(\lambda)$.

Proof. We only need to show that the characteristic function of S_n, which is

$$\varphi_n(t) = \prod_{m=1}^{n} \left((1 - p_{n,m})e^{it\cdot 0} + p_{n,m}e^{it\cdot 1}\right) = \prod_{m=1}^{n} \left((1 - p_{n,m}(1 - e^{it\cdot 0})\right)$$

converges to $\exp(\lambda(e^{it} - 1))$ for all t. Since the characteristic functions are all complex, we consider the absolute value and the argument separately. We need to show that the log of the absolute value of $\varphi_n(t)$,

$$\log\left(\prod_{m=1}^{n} \left|(1 - p_{n,m}(1 - e^{it\cdot 0})\right|\right) = \frac{1}{2}\sum_{m=1}^{n} \log\left(1 - 2(1 - \cos t)p_{n,m}(1 - p_{n,m})\right)$$

converge to $\log|\exp(\lambda(e^{it} - 1))| = -\lambda(1 - \cos t)$, and the argument of $\varphi_n(t)$,

$$\sum_{m=1}^{n} \arctan\frac{p_{n,m}\sin t}{1 - p_{n,m}(1 - \cos t)}$$

converges to $\lambda \sin t$.

For the limit of absolute value, we apply the estimate

$$|\log(1 + x) - x| \leq x^2 \quad \text{for } x \in (-1/2, 1/2),$$

a result similar to Lemma 56(c). If $p_{n,m}$ are close to 0, then

$$\begin{aligned}&\left|\left(\sum_{m=1}^{n}\log\left(1-2(1-\cos t)p_{n,m}(1-p_{n,m})\right)\right)-\left(\sum_{m=1}^{n}-2(1-\cos t)p_{n,m}(1-p_{n,m})\right)\right|\\ &\leq\sum_{m=1}^{n}4(1-\cos t)p_{n,m}^2(1-p_{n,m})^2\\ &\leq 4(1-\cos t)\left(\max_{1\leq m\leq n}p_{n,m}\right)\sum_{m=1}^{n}p_{n,m},\end{aligned}$$

and we have that it vanishes as $n\to\infty$. Thus

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{2}\sum_{m=1}^{n}\log\left(1-2(1-\cos t)p_{n,m}(1-p_{n,m})\right)\\ =&\lim_{n\to\infty}\sum_{m=1}^{n}-(1-\cos t)p_{n,m}(1-p_{n,m})\\ =&-(1-\cos t)\left(\lim_{n\to\infty}\sum p_{n,m}-\lim_{n\to\infty}\sum p_{n,m}^2\right)\\ =&-(1-\cos t)(\lambda-0),\end{aligned}$$

and we prove the convergence of absolute values.

On the other hand, we have

$$|\arctan(x)-x|\leq x^2\quad\text{for all } x\in\mathbb{R}.$$

So

$$\begin{aligned}&\left|\left(\sum_{m=1}^{n}\arctan\frac{p_{n,m}\sin t}{1-p_{n,m}(1-\cos t)}\right)-\left(\sum_{m=1}^{n}\frac{p_{n,m}\sin t}{1-p_{n,m}(1-\cos t)}\right)\right|\\ &\leq\sin^2 t\sum_{m=1}^{n}\frac{p_{n,m}^2}{(1-p_{n,m}(1-\cos t))^2}\\ &\leq\sin^2 t\left(\max_{1\leq m\leq n}p_{n,m}\right)\sum_{m=1}^{n}\frac{p_{n,m}}{(1-(\max_{1\leq m\leq n}p_{n,m})(1-\cos t))^2},\end{aligned}$$

and it vanishes as $n\to\infty$. We then have

$$\lim_{n\to\infty}\sum_{m=1}^{n}\arctan\frac{p_{n,m}\sin t}{1-p_{n,m}(1-\cos t)}=\lim_{n\to\infty}\sum_{m=1}^{n}\frac{p_{n,m}\sin t}{1-p_{n,m}(1-\cos t)}.$$

Since

$$\begin{aligned}\limsup_{n\to\infty}\sum_{m=1}^{n}\frac{p_{n,m}\sin t}{1-p_{n,m}(1-\cos t)}&\leq\limsup_{n\to\infty}\sum_{m=1}^{n}p_{n,m}\sin t=\lambda\sin t,\\ \liminf_{n\to\infty}\sum_{m=1}^{n}\frac{p_{n,m}\sin t}{1-p_{n,m}(1-\cos t)}&\leq\liminf_{n\to\infty}\sum_{m=1}^{n}\frac{p_{n,m}\sin t}{1-(\max_{1\leq m\leq n}p_{n,m})(1-\cos t)}=\lambda\sin t,\end{aligned}$$

we get the desired convergence of the argument. □

The basic form of Poisson convergence, which is only for Bernoulli random variables, has a direct generalization.

Theorem 58. *Let $X_{n,m}$, $1 \leq m \leq n$ be independent non-negative integer valued random variables, with $P(X_{n,m} = 1) = p_{n,m}$ and $P(X_{n,m} \geq 2) = \epsilon_{n,m}$, such that*

1. $\displaystyle\sum_{m=1}^{n} p_{n,m} \to \lambda \in (0, \infty)$.
2. $\displaystyle\max_{1 \leq m \leq n} p_{n,m} \to 0$.
3. $\displaystyle\sum_{m=1}^{n} \epsilon_{n,m} \to 0$.

Then the sum $S_n := X_{n,1} + \cdots + X_{n,n} \Rightarrow \text{Poisson}(\lambda)$.

Proof. Let $X'_{n,m} = X_{n,m} 1_{X_{n,m}=1}$. Then Theorem 57 implies that $S'_n := X_{n,1} + \cdots + X_{n,n} \Rightarrow \text{Poisson}(\lambda)$. On the other hand, $S_n - S'_n \to 0$ in probability, since

$$P(S_n \neq S'_n) \leq \sum_{m=1}^{n} P(X_{n,m} \neq X'_{n,m}) = \sum_{m=1}^{n} \epsilon_{n,m}.$$

By the "converging together lemma" (an exercise), we finish the proof. □

The next theorem is a corollary of the theorem above, and it shows how the Poisson distribution occurs in applications.

Theorem 59. *Suppose random positive points $x_1 < x_2 < \cdots$, which are called "arrivals", are placed on $(0, \infty)$, and let $N(s, t)$ be the number of arrivals in the interval $(s, t]$ if $0 \leq s < t$. Suppose the following assumptions hold:*

1. *The number of arrivals in disjoint intervals are independent.*
2. *The distribution of $N(s, t)$ only depends on $t - s$.*
3. *$P(N(0, h) = 1) = \lambda h + o(h)$ as $h \downarrow 0$.*
4. *$P(N(0, h) \geq 2) = o(h)$ as $h \downarrow 0$.*

Then $N(0, t) = \text{Poisson}(\lambda t)$ for all $t > 0$.

Proof. Note that $N(0, t) = \lim_{n\to\infty} \sum_{m=1}^{n} X_{n,m}$ where $X_{n,m} = N((m-1)/n, m/n)$, and apply Theorem 58. □

We can interpret the theorem by a real life example. Let x_k be the time that the k-th customer comes to a bank after the opening time $t = 0$ (so explained the term "arrival"), then we assume the ideal conditions:

1. In non-overlapping different time intervals, the numbers of incoming customers are independent.

2. The number of incoming customers in any time interval from s to t depends only on $t - s$.

3. At any infinitesimal time, the rate for a customer to come is λ.

4. The case that more than one customers come together is practically impossible.

Then the distribution of the number of customers in total time t is Poisson(λt).

Inspired by the bank customer example, we define the following Poisson process.

Definition 9. A family of random variables N_t, $t \geq 0$, satisfies

(a) If $0 = t_0 < t_1 < \cdots < t_n$, then $N(t_k) - N(t_{k-1})$ $(1 \leq k \leq n)$ are independent.

(b) $N(t) - N(s) = \text{Poisson}(\lambda(t-s))$.

Then $\{N_t\}$ is called a *Poisson process with rate* λ.

We claim that N_t can be defined as follows. Let $\xi_1, \xi_2, \dots$ be i.i.d. positive random variables with exponential distribution with rate λ, that is, $P(\xi_i > t) = e^{-\lambda t}$. Let $T_n = \xi_1 + \cdots + \xi_n$ and define $N_t = \sup\{n \mid T_n \leq t\}$. Then $\{N_t\}$ satisfies the two requirements for a Poisson process with rate λ.

Below we check the two conditions for $\{N_t\}$ constructed above. Since this is a topic covered by MA3236, we do not give all the calculational details. First compute the density of T_n as

$$f_{T_n}(s) = \frac{\lambda^n s^{n-1}}{(n-1)!} e^{-\lambda s}.$$

So

$$P(N_t = 0) = P(T_1 > t) = e^{-\lambda t}$$

and for $n \geq 1$

$$P(N_t = n) = P(T_n \leq t) - P(T_{n+1} \leq t) = \int_0^t f_{T_n}(s) - f_{T_{n+1}}(s) ds = e^{-\lambda t}\frac{(\lambda t)^n}{n!}.$$

We then verify Condition (b) with $s = 0$. Below we show Condition (a). After that, Condition (b) follows. To see it, we note that if $X = N(t) - N(s)$ is independent to $Y = N(s)$, and we have $Y = \text{Poisson}(s)$ and $X + Y = \text{Poisson}(t)$, we can compute that $X = \text{Poisson}(t - s)$ (exercise).

To check the independence condition (a), we compute the conditional probability that for $u \geq t > 0$

$$P(T_{n+1} > u \mid N_t = n) = \frac{P(T_{n+1} > u, T_n \leq n)}{P(N_t = n)} = e^{-\lambda(u-t)}.$$

This computation shows that if we denote $\xi'_1 = T_{N(t)+1} - t$, then ξ'_1 is independent of $N(t)$ and its distribution is the same as ξ_i: exponential distribution with rate λ. Similarly, if we denote $\xi'_k = T_{N(t)+k} - T_{N(t)+k-1}$ for $k \geq 2$, we have that all the $\xi'_1, \xi'_2, \dots$ are i.i.d. with exponential distribution with rate λ, and they are independent of $N(t)$.

Since $\xi'_1, \xi'_2, \dots$ have the same distribution as $\xi_1, \xi_2, \dots$, with $t = t_1$, we have that the distributions of $N(t_2) - N(t_1)$, $N(t_3) - N(t_2)$, $\dots$, $N(t_n) - N(t_{n-1})$ have the same distribution as $N(t_2 - t_1) - N(0)$, $N(t_3 - t_1) - N(t_2 - t_1)$, $\dots$, $N(t_n - t_1) - N(t_{n-1} - t_1)$, and they are independent by inductive assumption. Also since all the $N(t_k) - N(t_{k-1})$ $(k \geq 2)$ are derived from $\xi'_1, \xi'_2, \dots$, they are independent to $N(t_1) = N(t_1) - N(t_0)$.

9 Stable laws

Now we consider the complement of the central limit theorem, in the following sense: Let $X_1, X_2, \ldots$ be i.i.d. random variables with $EX_i^2 = \infty$. Then the central limit theorem cannot hold, but we may ask whether a non-degenerate weak limit $(S_n = b_n)/a_n$ exists, where $S_n = X_1 + \cdots + X_n$ as usual, and a_n, b_n are properly chosen real numbers? (We say non-degenerate, because it is easy to let the weak limit to be 0 if we let $a_n \to \infty$ fast enough.)

Our first example is the sum of i.i.d. random variables $X_1, X_2, \ldots$ in *Cauchy distribution*, that is,

$$F(x) := P(X_i \leq x) = \frac{1}{\pi} \int_{-\infty}^{x} \frac{dy}{1+y^2} = \frac{\arctan x}{\pi} + 1/2.$$

We have that the characteristic function of X_i is

$$\varphi(t) = \int_{-\infty}^{\infty} \frac{e^{ity}}{\pi(1+y^2)} dy = \exp(-|t|).$$

Then the characteristic function of S_n is $\exp(-n|t|)$, and we have that S_n/n has Cauchy distribution. Hence in this example, we may take $a_n = n$ and $b_n = 0$, and the limit is the Cauchy distribution.

Here we see that the Cauchy distribution has a special property that it is a *stable law.*

Definition 10. A distribution μ is called a *stable law* if for any n and $X_1, \ldots, X_n$ are i.i.d. random variables with distribution μ, there exist a_n and b_n such that $(S_n - b_n)/a_n$ also has distribution μ, where $S_n = X_1 + \cdots + X_n$ and a_n, b_n are properly chosen real numbers.

It is clear that normal distributions are also stable laws. Actually stable laws are closely related to the central limit theorem and its infinite variance counterpart. Below we prove the following general result:

Theorem 60. *Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with a distribution that satisfies*

(i) $\displaystyle\lim_{x \to +\infty} \frac{P(X_i > x)}{P(|X_i| > x)} = \theta \in [0, 1].$

(ii) $P(|X_i| > x) = x^{-\alpha} L(x)$ *for all* $x > 0$*, where* $\alpha \in (0, 2)$*, and* $L(x)$ *is a* slow varying *function, such that* $\displaystyle\lim_{x \to +\infty} L(tx)/L(x) \to 1$ *for all* $t > 0$*.*

Then with $S_n = X_1 + \cdots + X_n$*,* $a_n = \inf\{x \mid P(|X_i| > 0) \leq n^{-1}\}$ *and* $b_n = nE(X_i 1_{|X_i| \leq a_n})$*, we have that* $(S_n - b_n)/a_n \Rightarrow Y$ *where* Y *has a non-degenerate distribution.*

Before giving the proof, we clarify some definitions that are not very straightforward. First, the slow varying condition implies that if $L(x)$ grows, then it grows very slow, and if $L(x)$ vanishes, then it vanishes very slow, in the sense that given $\epsilon > 0$, $x^{-\epsilon} < L(x) < x^{\epsilon}$

for all large enough x. To see it, we assume that for all $x > M$, $L(2x)/L(x) < 2^{\epsilon/2}$. Then we take $c = \inf_{x \in (M,2M]} L(x)$ and $C = \sup_{x \in (M,2M]} L(x)$, and have inductively

$$2^{\frac{-\epsilon}{2}k} c < \inf_{x \in (2^k M, 2^{k+1} M]} L(x) \leq \sup_{x \in (2^k M, 2^{k+1} M]} L(x) < 2^{\frac{\epsilon}{2}k} C.$$

Thus for large enough x, $x^{-\epsilon} L(x) < x^{\epsilon}$. Similarly, we can show that for any $\epsilon > 0$, there exist $M, C, c > 0$ such that for all $y > x > M$,

$$c \left(\frac{y}{x} \right)^{-\epsilon} < \frac{L(y)}{L(x)} < C \left(\frac{y}{x} \right)^{\epsilon}.$$

The proof is left as an exercise.

The slow growth property of $L(x)$ implies that $a_n \to \infty$ faster than $\sqrt{n}$. To see it, we note that for large enough x, $L(x) > x^{\alpha/2-1}$, and then $P(|X_i| > x) > x^{-\alpha/2-1}$. We conclude that if n is large enough, then $P(|X_i| > n^{1/(\alpha/2+1)} > n^{-1}$, and then $a_n \leq n^{1/(\alpha/2+1)}$.

By the definition of a_n, we can derive that $P(|X_i| > a_n) \leq n^{-1}$, but do not have that the equal sign holds, unless the distribution function is continuous at $\pm a_n$. But we have the limiting result

$$nP(|X_i| > a_n) \to 1 \quad \text{as } n \to \infty.$$

To check it, we argue by contradiction, and assume that for any $\epsilon > 0$, there is a subsequence $\{n(k)\}$ such that $P(|X_i| > a_{n(k)}) \leq (1-\epsilon) n(k)^{-1}$. On the other hand, by the definition of a_n, we have that $P(|X_i| > (1-\epsilon)^{1/(2\alpha)} a_{n(k)}) > n(k)^{-1}$. Equivalently, we have

$$a_{n(k)}^{-\alpha} L(a_{n(k)}) \leq (1-\epsilon) n^{-1} \quad \text{and} \quad (1-\epsilon)^{-1/2} a_{n(k)}^{-\alpha} L((1-\epsilon) a_{n(k)}) > n^{-1}.$$

It implies the inequality

$$\frac{L((1-\epsilon) a_{n(k)})}{L(a_{n(k)})} > (1-\epsilon)^{-1/2}.$$

Since $a_{n(k)} \to \infty$, it contradicts the slow varying condition.

Proof of Theorem 60. For any $\epsilon > 0$, we define the triangular arrays of random variables $\bar{X}_{n,m}(\epsilon)$ and $\hat{X}_{n,m}(\epsilon)$, where $1 \leq m \leq n$, as

$$\bar{X}_{n,m}(\epsilon) = X_m 1_{|X_m| \leq \epsilon a_n}, \quad \hat{X}_{n,m}(\epsilon) = X_m 1_{|X_m| > \epsilon a_n},$$
$$\bar{\mu}_n(\epsilon) = E\bar{X}_{n,1}(\epsilon), \quad \hat{\mu}_n(\epsilon) = E(\hat{X}_{n,1}(\epsilon); |\hat{X}_{n,1}| \leq a_n) = E(X_m 1_{\epsilon a_n < X_m \leq a_n}),$$

and then let

$$\bar{S}_n(\epsilon) = \sum_{m=1}^{n} \bar{X}_{n,m}(\epsilon), \quad \hat{S}_n(\epsilon) = \sum_{m=1}^{n} \hat{X}_{n,m}(\epsilon).$$

It is clear that $S_n = \bar{S}_n(\epsilon) + \hat{S}_n(\epsilon)$, and we also have

$$\frac{S_n - b_n}{a_n} = \frac{\bar{S}_n(\epsilon) - n\bar{\mu}_n(\epsilon)}{a_n} + \frac{\hat{S}_n(\epsilon) - n\hat{\mu}_n(\epsilon)}{a_n}.$$

We will show that $(\bar{S}_n(\epsilon) - n\bar{\mu}_n(\epsilon))/a_n$ is small (comparable to ϵ), and then compute the weak limit of $(\hat{S}_n(\epsilon) - n\hat{\mu}_n(\epsilon))/a_n$.

We estimate the $\mathrm{var}(\bar{S}_n(\epsilon)$ as follows. First,

$$\mathrm{var}(\bar{S}_n(\epsilon)) = n\,\mathrm{var}(\bar{X}_{n,m}(\epsilon)) \leq nE(\bar{X}^2_{n,m}(\epsilon)).$$

Suppose $\bar{\mu}$ is the distribution of $|X_i|$ and $\bar{F}$ is the distribution function. Also suppose that

$$\frac{L(x)}{L(y)} < C\left(\frac{x}{y}\right)^{\alpha/2-1}, \quad \text{or equivalently,} \quad \frac{1-\bar{F}(x)}{1-\bar{F}(y)} < C\left(\frac{x}{y}\right)^{-1-\alpha/2} \quad \text{for } x > M \text{ and } t > 1.$$

We have

$$nE\bar{X}^2_{n,m}(\epsilon) = n\int_0^{\epsilon a_n} x^2\bar{\mu}(dx) = n\int_0^M x^2\bar{\mu}(dx) + n\int_M^{\epsilon a_n} x^2\bar{\mu}(dx) \leq nM^2 + n\int_M^{\epsilon a_n} x^2\bar{\mu}(dx).$$

On the other hand,

$$\begin{aligned} n\int_M^{\epsilon a_n} x^2\bar{\mu}(dx) &= n\int_M^{\epsilon a_n}\left(\int_0^\infty 2t\cdot 1_{t\leq x}dt\right)\bar{\mu}(dx) \\ &= n\int_0^\infty 2t\left(\int_M^{\epsilon a_n} 1_{t\leq x}\bar{\mu}(dx)\right)dt \\ &= n\int_0^M 2t(\bar{F}(\epsilon a_n) - \bar{F}(C))dt + n\int_M^{\epsilon a_n} 2t(\bar{F}(\epsilon a_n) - F(t))dt \\ &\leq n\int_0^M 2tdt + \int_M^{\epsilon a_n} 2tn(1-\bar{F}(t))dt \\ &\leq nM^2 + 2\int_M^{\epsilon a_n} n(1-\bar{F}(a_n))Ct\left(\frac{t}{a_n}\right)^{-1-\alpha/2}dt. \end{aligned}$$

Since we have that $n(1-F(a_n)) = nP(|X_i| > a_n) \to 1$ and

$$\int_M^{\epsilon a_n} t\left(\frac{t}{a_n}\right)^{-1-\alpha/2}dt \leq \int_0^{\epsilon a_n} t\left(\frac{t}{a_n}\right)^{-1-\alpha/2}dt = a_n^2\int_0^\epsilon y^{-\alpha/2}dy = \frac{\epsilon^{1-\alpha/2}}{1-\alpha/2}a_n^2,$$

we conclude that

$$\limsup_{n\to\infty}\mathrm{var}(\bar{S}_n(\epsilon)) \leq 2nM^2 + \frac{\epsilon^{1-\alpha/2}}{1-\alpha/2}a_n^2.$$

Finally, since $a_n \leq n^{1/(\alpha/2+1)}$, we have

$$\limsup_{n\to\infty} E((\bar{S}_n(\epsilon) - \bar{\mu}_n(\epsilon))/a_n) = \limsup_{n\to\infty}\mathrm{var}(\bar{S}_n(\epsilon)/a_n) \leq \frac{\epsilon^{1-\alpha/2}}{1-\alpha/2}.$$

Now we consider the distribution of $\hat{S}_n/a_n$. Note that since $nP(|X_i| > a_n) \to n$, we have that for all $t > 0$, $nP(|X_i| > ta_n) \to t^{-\alpha}$, by condition (ii), and furthermore, $nP(X_i > ta_n) \to \theta t^{-\alpha}$ and $nP(X_i < -ta_n) \to (1-\theta)t^{-\alpha}$, as $n \to \infty$.

Then for any $\hat{X}_{n,m}(\epsilon)$, we have $nP(\hat{X}_{n,m}(\epsilon) \neq 0) \to \epsilon^{-\alpha}$. So by the Poisson convergence result, we have that the number of nonzero $X_{n,m}(\epsilon)$ with $m = 1, \ldots, n$ has the Poisson distribution with mean $\epsilon^{-\alpha}$ as the limit as $n \to \infty$. To be precise,

$$p_k(n) := P(k \text{ out of } n\ \hat{X}_{n,m}(\epsilon) \text{ are nonzero}) \to e^{-\epsilon^{-\alpha}}\frac{\epsilon^{-\alpha k}}{k!}, \quad \text{as } n \to \infty.$$

Under the condition that k out of n $\hat{X}_{n,m}(\epsilon)$ are nonzero, these k nonzero ones are i.i.d., and they have the same distribution as i.i.d. random variables $Y_{n,1}(\epsilon), \ldots, Y_{n,k}(\epsilon)$ such that

$$\lim_{n\to\infty} P(Y_{n,1}(\epsilon)/a_n > t) = \theta t^{-\alpha}/\epsilon^{-\alpha}, \quad \lim_{n\to\infty} P(Y_{n,1}(\epsilon)/a_n < -t) = (1-\theta)t^{-\alpha}/\epsilon^{-\alpha},$$

or equivalently, their limiting distribution of $Y_{n,1}(\epsilon)/a_n)$ is given by the density function

$$f(x) = \begin{cases} \theta\alpha\epsilon^{\alpha}t^{-\alpha-1} & \text{if } t \geq \epsilon, \\ (1-\theta)\alpha\epsilon^{\alpha}(-t)^{-\alpha-1} & \text{if } t \leq -\epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

So we have that the characteristic function $\varphi_n(t;\epsilon)$ of $\hat{S}_n(\epsilon)/a_n$ is expressed as

$$\varphi_n(x;s) = E\left(\prod_{m=1}^{n} e^{it\hat{X}_{n,m}(\epsilon)}\right) = \sum_{k=1}^{n} p_k(n) E\left(e^{itY_{n,1}(\epsilon)}\right)^k.$$

By the limit of $p_k(n)$ as $n \to \infty$, and the limit

$$\begin{aligned}\lim_{n\to\infty} E\left(e^{itY_{n,1}(\epsilon)}\right) &= \int^{\infty} f(x)dx + \int_{-\infty}^{-\epsilon} f(x)dx \\ &= \alpha\epsilon^{\alpha}\int_{\epsilon}^{\infty}(\cos(tx) + i(2\theta-1)\sin(tx))x^{-\alpha-1}dx,\end{aligned}$$

we have that

$$\begin{aligned}\lim_{n\to\infty}\varphi_n(t;\epsilon) &= e^{-\epsilon^{-\alpha}}\sum_{k=0}^{\infty}\frac{e^{-\alpha k}}{k!}\left(\alpha\epsilon^{\alpha}\int_{\epsilon}^{\infty}(\cos(tx) + i(2\theta-1)\sin(tx))x^{-\alpha-1}dx\right)^k \\ &= e^{-\epsilon^{-\alpha}}\exp\left(\alpha\int_{\epsilon}^{\infty}(\cos(tx) + i(2\theta-1)\sin(tx))x^{-\alpha-1}dx\right) \\ &= \exp\left(\alpha\int_{\epsilon}^{\infty}(\cos(tx) - 1 + i(2\theta-1)\sin(tx))x^{-\alpha-1}dx\right).\end{aligned}$$

At last we consider the characteristic function of $(\hat{S}_n(\epsilon) - n\hat{\mu}_n(\epsilon))/a_n$. We have

$$\begin{aligned}\frac{n\hat{\mu}_n(\epsilon)}{a_n} &= nE\left(X_1/a_n; \epsilon < |X_1|/a_n \leq 1\right) \\ &= nP(|X_i| > \epsilon a_n)E\left(Y_{n,1}(\epsilon)/a_n; \epsilon < |Y_{n,1}(\epsilon)|/a_n \leq 1\right),\end{aligned}$$

and then the limit

$$\lim_{n\to\infty}\frac{n\hat{\mu}_n(\epsilon)}{a_n} = \alpha\int_{\epsilon}^{1}(2\theta-1)x \cdot x^{-\alpha-1}dx.$$

Although further simplification is possible, the form above is the most suitable one for our purpose, because then we have that the limit of the characteristic function of $(\hat{S}_n(\epsilon)-$

$n\hat{\mu}_n(\epsilon))/a_n$ is

$$\begin{aligned}\lim_{n\to\infty}\varphi_n(t;\epsilon)e^{itn\hat{\mu}_n(\epsilon)/a_n} = &\exp\left(\alpha\int_\epsilon^\infty(\cos(tx)-1)x^{-\alpha-1}dx\right)\\ &\times\exp\left(\alpha\int_1^\infty i(2\theta-1)\sin(tx))x^{-\alpha-1}dx\right)\\ &\times\exp\left(\alpha\int_\epsilon^1 i(2\theta-1)(\sin(tx)-tx)x^{-\alpha-1}dx\right).\end{aligned}$$

We note that all the three terms are well defined, and the first and third terms have well defined limits as $\epsilon\downarrow 0$.

As ϵ becomes small, we have that $(\bar{S}_n(\epsilon)-\bar{\mu}_n(\epsilon))/a_n$ has a very small second moment as n is large, and the weak limit of $(\hat{S}_n(\epsilon)-n\hat{\mu}_n(\epsilon))/a_n$ exists and has the characteristic function close to

$$\begin{aligned}\varphi(t) = \exp\left(\alpha\int_0^\infty(\cos(tx)-1)x^{-\alpha-1}dx\right)&\exp\left(\alpha\int_1^\infty i(2\theta-1)\sin(tx))x^{-\alpha-1}dx\right)\\ &\times\exp\left(\alpha\int_0^1 i(2\theta-1)(\sin(tx)-tx)x^{-\alpha-1}dx\right).\end{aligned}$$

It is not hard to see that by letting $\epsilon\downarrow 0$ the sum of them, $(S_n-b_n)/a_n$, has a weak limit whose characteristic function is $\varphi(t)$. □

From the proof, we see that the limiting behaviour of the sum S_n is determined largely by a few large random variable X_m, whose values are bigger than ϵa_n. This is different from the central limit theorem where the a few largest random variables do not affect the limiting distribution. Also, Poisson distribution occurs in the argument when we deal with the rare and big random variables.

You may wonder: What is the relation between the theorem and the concept "stable laws"? Actually, the limiting distributions in Theorem 60 are stable laws, and these distributions, together with normal distributions, exhaust stable laws. But we may not have time to discuss more on it.

10 Convergence of random series

In this semester, we studied the laws of large numbers, central limit theorem, and some related results. All the results are in this form: A series, or a triangular array, of random variables, converges to a limit (in some sense), and we can compute the limit. But it is normal in mathematics that the existence of a limit is already a challenge, while it is exceptional that the limit can be computed explicitly.

The last topic of our module is convergence theorems of random series, while the computation of the limit is no longer our main interest.

First we introduce an important and rather abstract result, Kolmogorov's 0-1 law. It depends on measure theory in a subtle way.

Let $X_1, X_2, \dots$ be a sequence of random variables on the probability space $(\Omega, \mathcal{F}, P)$. Recall that $\sigma(X_i)$ is a σ-algebra on Ω, consisting of the subsets $X_i^{-1}(B)$ where $B \in \mathcal{B}$ is a Borel set on $\mathbb{R}$. For several $X_{n_1}, X_{n_2}, \dots, X_{n_k}$, we define the σ-algebra $\sigma(X_{n_1}, \dots, X_{n_k})$ as the σ-algebra generated by all $X_{n_i}^{-1}(B)$, $B \in \mathcal{B}$. The definition is valid if $k = \infty$, and we denote $\mathcal{F}'_n = \sigma(X_n, X_{n+1}, \dots)$. It is clear that $\mathcal{F}'_n \subseteq F'_m$ if $n > m$. Then we define

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{F}'_n.$$

To understand the meaning of $\mathcal{T}$, we think each $\sigma(X_n)$ as the information carried by X_n and interpret n as time. Then $\mathcal{F}'_n$ is the information related to time $\geq n$, and $\mathcal{T}$ is the information in remote future. To make sense of the definitions, we consider the following examples.

Example 2. 1. Let B_n be a Borel set on $\mathbb{R}$ for each n, and then $A = \{X_n \in B_n \text{ i. o.}\}$ is a subset of Ω. We have

$$\begin{aligned} A = \{\omega \in \Omega \mid X_n(\omega) \in B_n \text{ for infinitely many } n\} &= \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{X_k \in B_k\} \\ = \bigcap_{n=m}^{\infty} \bigcup_{k=n}^{\infty} \{X_k \in B_k\} \quad \text{for all } m \\ \in \mathcal{F}'_m \quad \text{for all } m. \end{aligned}$$

Therefore $A \in \mathcal{T}$.

2. Let $S_n = X_1 + \cdots + X_n$, and then $B = \{\lim_{n\to\infty} S_n \text{ exists}\}$ is a subset of Ω. We have

$$\begin{aligned} B &= \{\omega \in \Omega \mid \sum_{n=1}^{\infty} X_n(\omega) \text{ converges}\} \\ &= \{\omega \in \Omega \mid \sum_{n=m}^{\infty} X_n(\omega) \text{ converges}\} \quad \text{for all } m \\ &\in F'_m \quad \text{for all } m. \end{aligned}$$

Therefore $B \in \mathcal{T}$.

3. Continued from last example, let $C = \{\limsup_{n\to\infty} S_n > 0\}$. C is not in $\mathcal{T}$ in general, and even $C \notin \mathcal{F}'_2$ in general. (In special cases, like $X_1 = X_2 = \cdots$, $C \in \mathcal{T}$.) Because even if we know the values of $X_2, X_3, \dots$, we still cannot tell if $\limsup_{n\to\infty} S_n > 0$, without the knowledge of the value of X_1.

Now we can state Kolmogorov's 0-1 law, and it implies that $S_n = X_1 + \cdots + X_n$ converges a.s., or diverges a.s..

Theorem 61 (Kolmogorov's 0-1 law)**.** *If $X_1, X_2, \dots$ are independent, and $A \in \mathcal{T}$, then $P(A) = 0$ or 1.*

Before giving the proof, we review the concept of independence. We say two collections $\mathcal{A}$ and $\mathcal{B}$ of measurable sets, which may be σ-algebras or may not, are independent, if

$$P(A \cap B) = P(A)P(B), \quad \text{for all } A \in \mathcal{A} \text{ and } B \in \mathcal{B}.$$

If $Y_1, \dots, Y_m, Z_1, \dots, Z_n$ are independent random variables, we have $\sigma(Y_i)$ and $\sigma(Z_j)$ are independent, by definition. We also have that $\sigma(Y_1, \dots, Y_m)$ and $\sigma(Z_1, \dots, Z_n)$ are independent. To check it, we can start with the definition, but a faster way is to apply Theorem 18. Since $\sigma(Y_1) \cup \cdots \cup \sigma(Y_m)$ and $\sigma(Z_1) \cup \cdots \cup \sigma(Z_n)$ are independent, and both of them are "π-systems", we conclude that $\sigma(\sigma(Y_1) \cup \cdots \cup \sigma(Y_m)) = \sigma(Y_1, \dots, Y_m)$ and $\sigma(\sigma(Z_1) \cup \cdots \cup \sigma(Z_n)) = \sigma(Z_1, \dots, Z_n)$ are independent, by Theorem 18.

Proof of Theorem 61. We show that A is independent to itself. Then $P(A) = P(A \cup A) = P(A)P(A)$, and we conclude that $P(A) = 0$ or 1.

First, we show that for any n, $\mathcal{H}_n = \sigma(X_1, X_2, \dots, X_n)$ and $F'_{n+1} = \sigma(X_{n+1}, X_{n+2}, \dots)$ are independent. It is already known that $\mathcal{H}_n$ and $\sigma(X_{n+1}, \dots, X_{n+k})$ are independent, for all k. Thus $\mathcal{H}_n$ and $\bigcup_{k=1}^\infty \sigma(X_{n+1}, \dots, X_{n+k})$ are independent. Since $\bigcup_{k=1}^\infty \sigma(X_{n+1}, \dots, X_{n+k})$ is a π-system, and $\sigma(\bigcup_{k=1}^\infty \sigma(X_{n+1}, \dots, X_{n+k})) = \mathcal{F}'_{n+1}$, we get the desired result by Theorem 18.

Next, we show that $\mathcal{F}'_1 = \sigma(X_1, X_2, \dots)$ and $\mathcal{T}$ are independent. To see it, we first note that for all n, $\mathcal{H}_n$ and $\mathcal{T}$ are independent, since $T \subseteq \mathcal{F}'_{n+1}$. Then $\bigcup_{n=1}^\infty \mathcal{H}_n$ and $\mathcal{T}$ are independent. Since $\bigcup_{n=1}^\infty \mathcal{H}_n$ is a π-system, and it generates $\mathcal{F}'_1$, we derive the result by Theorem 18.

Then since $A \in \mathcal{F}'_1$ and $A \in \mathcal{T}$, we prove the theorem. □

Below we derive some method to tell if $\{S_n\}$ converges a.s.. A basic technical result is the following lemma.

Lemma 62 (Kolmogorov's maximal inequality)**.** *Suppose $X_1, X_2, \dots, X_n$ are independent with $EX_i = 0$ and $\operatorname{var}(X_i) < \infty$. If $S_k = X_1 + \ldots + X_k$, then*

$$P\left(\max_{1\le k\le n} |S_k| \ge x\right) \le x^{-2} \operatorname{var}(S_n).$$

Proof. We divide the event $\max_{1\le k\le n}|S_k| \ge x$ into disjoint subevents

$$A_k = \{|S_k| \ge x, \text{ but } |S_j| < x \text{ for all } j < k\}, \quad \text{for all } k = 1, \dots, n.$$

Consider the inequality

$$\begin{aligned}
\mathrm{var}(S_n) = ES_n^2 &\geq \sum_{k=1}^n \int_{A_k} S_n^2 dP \\
&= \sum_{k=1}^n \int_{A_k} S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2 dP \\
&\geq \sum_{k=1}^n \int_{A_k} S_k^2 dP + 2\int_{A_k} S_k(S_n - S_k)dP \\
&= \left(\sum_{k=1}^n \int_{A_k} S_k^2 dP\right) + 2\left(\sum_{k=1}^n \int_\Omega S_k 1_{A_k}(S_n - S_k)dP\right).
\end{aligned}$$

Here $S_n - S_k$ is a random variable, and it is a function of $X_{k+1}, \dots, X_n$: $S_n = X_{k+1} + \cdots + X_n$. On the other hand, $S_k 1_{A_k}$ is also a random variable, and it is a function of $X_1, \dots, X_k$:

$$S_k 1_{A_k}(\omega) = (X_1(\omega) + \cdots + X_k(\omega))1_{|X_1(\omega)+\cdots+X_k(\omega)|\geq x \text{ but } |X_1(\omega)+\cdots+X_j(\omega)|<x \text{ for all } j<k}.$$

So the random variables $S_n - S_k$ and $S_k 1_{A_k}$ are independent, and then

$$\int_\Omega S_k 1_{A_k}(S_n - S_k)dP = E((S_k 1_{A_k})(S_n - S_k)) = E(S_k 1_{A_k})E(S_n - S_k) = E(S_k 1_{A_k}) \cdot 0 = 0.$$

Thus we conclude that

$$x^2 P\left(\max_{1\leq k\leq n} |S_k| \geq x\right) = \sum_{k=1}^n x^2 P(A_k) \leq \sum_{k=1}^n \int_{A_k} S_k^2 dP \leq \mathrm{var}(S_n),$$

and prove the inequality. □

Remark 7. Kolmogorov's maximal inequality can be easily extended to the infinite random variable case, that is, for $X_1, X_2, \dots$ with $EX_i = 0$ and $\sum_{i=1}^\infty \mathrm{var}(X_i) = C < \infty$,

$$P\left(\max_{k=1}^\infty |S_k| \geq x\right) \leq x^{-2}C.$$

As an application of Kolmogorov's maximal inequality, we prove the following result:

Theorem 63 (Kolmogorov's two-series)**.** *Suppose $X_1, X_2, \dots$ are independent with $EX_n = 0$. If $\sum_{n=1}^\infty \mathrm{var}(X_n) < \infty$, then with probability 1, $\{S_n\}$ converges, where $S_n = X_1 + \cdots + X_n$.*

Proof. To show the almost sure convergence, it suffices to show that for any $\epsilon > 0$, there is a set A such that $P(A) \geq 1 - \epsilon$, and there are $m_1, m_2, \dots$ such that for all $\omega \in A$, $|S_n - S_{m_k}| \leq 2^{-k}$ if $n > m_k$.

Given $\epsilon > 0$, we define m_k as

$$m_k = \min\{n \in \mathbb{Z}_+ \mid \sum_{j=n}^\infty \mathrm{var}(X_j) \leq 2^{-3k}\epsilon\}.$$

Then by the infinite extension of Kolmogorov's maximal inequality,

$$P\left(\max_{n=m_k+1}^{\infty}|S_n - S_{m_k}| \geq 2^{-k}\right) \leq 2^{2k}\sum_{j=m_k}^{\infty}\operatorname{var}(X_j) \leq 2^{-k}\epsilon.$$

Let the events $B_k = \{\max_{n=m_k+1}^{\infty}|S_n - S_{m_k}| \geq 2^{-k}\}$, and $A = \Omega \setminus (B_1 \cup B_2 \cup \cdots)$. We check that this set A satisfies the requirement. □

Next we prove a similar result, for general random variables.

Theorem 64 (Kolmogorov's three-series)**.** *Let $X_1, X_2, \dots$ be independent. In order that $\sum_{n=1}^{\infty} X_n$ converges a.s., it is sufficient that for some $A > 0$, all the following conditions hold, where $Y_i = X_i 1_{|X_i|\leq A}$:*

(i) $\sum_{n=1}^{\infty} P(|X_n| > A) < \infty$.

(ii) $\sum_{n=1}^{\infty} EY_n$ *converges.*

(iii) $\sum_{n=1}^{\infty} \operatorname{var}(Y_n) < \infty$.

Conversely, if $\sum_{n=1}^{\infty} X_n$ converges a.s., then for all $A > 0$, the three conditions above hold.

Proof. First we prove the sufficiency. Let $Z_i = Y_i - EY_i$. By Kolmogorov's two-series theorem and Condition (iii), $\sum_{n=1}^{\infty} Z_n(\omega)$ converges a.s.. Then by Condition (ii) $\sum_{n=1}^{\infty} Y_n(\omega) = \sum_{n=1}^{\infty}(Z_n(\omega) + EY_i)$ converges wherever $\sum_{n=1}^{\infty} Z_n(\omega)$ does. At last, Borel-Cantelli lemma and Condition (i) imply that $\{X_n \neq Y_n \text{ i. o.}\}$ is of probability 0. Thus we prove that $\sum_{n=1}^{\infty} X_n(\omega)$ converges a.s..

Next we prove the necessity. Suppose $\sum_{n=1}^{\infty} X_n$ converges a.s.. If $\sum_{n=1}^{\infty} X_n(\omega)$ converges, then for all but finitely many n, $|X_n(\omega)| < A$ for all $A > 0$, and hence $\{|X_n| > A \text{ i. o.}\}$ has probability 0. But if condition (i) does not hold for some $A > 0$, we have that $P(\{|X_n| > A \text{ i. o.}\}) = 1$ by the second Borel-Cantelli lemma, and we derive a contradiction.

Now we admit that Condition (i) holds, but assume that Condition (iii) fails. Then we define the triangular array of random variables $W_{n,m}$ $(1 \leq m \leq n)$

$$W_{n,m} = \frac{1}{\sqrt{C_n}}(Y_n - EY_n), \quad \text{where} \quad C_n = \sum_{i=1}^{n}\operatorname{var}(X_i).$$

Then we can check that $W_{n,m}$ satisfies the assumptions for the Lindeberg–Fellor central limit theorem, and we have that $\sum_{m=1}^{n} W_{n,m} \Rightarrow N(0,1)$, and the characteristic function of $\sum_{m=1}^{n} W_{n,m}$ converges pointwise to $e^{-t^2/2}$. Now consider $C_n^{-1/2}\sum_{m=1}^{n} Y_m = (\sum_{m=1}^{n} W_{n,m}) + \sum_{m=1}^{n} EY_m$. The characteristic function of his random variable is the characteristic function of $\sum_{m=1}^{n} W_{n,m}$ times $\exp(it\sum_{m=1}^{n} EY_m)$, so we have that the absolute value of the characteristic function of $C_n^{-1/2}\sum_{m=1}^{n} Y_m$ converges to $e^{-t^2/2}$ pointwise.

But, if $\sum_{m=1}^n X_m$ converges a.s., we have that $\sum_m^n Y_m$ converges a.s. too, by the Borel-Cantelli lemma, and we call the limit S'_∞. Since $C_n^{-1/2} \to 0$ as $n \to \infty$, we have that $C_n^{-1/2} \sum_{m=1}^n Y_m$ converges to 0 a.s.. Thus we conclude that the (absolute value of) the characteristic function of $C_n^{-1/2} \sum_{m=1}^n Y_m$ converges to 1 pointwise. Thus we derive a contradiction.

So we need to admit both Conditions (i) and (iii), for all $A > 0$. At last we show that if both Conditions (i) and (iii) hold, then Condition (ii) holds too. To see it, we use the result that $\sum_{n=1}^\infty Y_n$ converges a.s., and by Kolmogorov's two series theorem and Condition (iii), we have that $\sum_{n=1}^\infty (Y_n - EY_n)$ converges a.s.. So their difference, $\sum_{n=1}^\infty EY_n$ converges (a.s.). □