Lecture Note for MAT8030: Advanced Probability

LI Liying*

November 29, 2024

1 Measure theory preliminaries

In this section we cover some basic facts in measure theory and how they integrate into the modern probability theory, which is essential to this field. Most of the materials are still within the scope of the celebrated work, *Foundations of the theory of probability*, by Kolmogorov in 1933 ([Kol33]).

1.1 Random variables, σ-fields and measures

We start with examples of some random variables (r.v.s) that the reader should be familiar with from elementary probability. There are two types of r.v.s encountered in elementary probability: discrete and continuous.

Example 1.1 Examples of discrete r.v.s.

- **Bernoulli:** $X \sim \mathrm{Ber}(p)$, with $\mathsf{P}(X=1)=p$, $\mathsf{P}(X=0)=1-p$.
- **binomial:** $X \sim \mathrm{Binom}(n,p)$ with $\mathsf{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$, $k=0,1,\dots,n$.
- **geometry:** $X \sim \mathrm{Geo}(p)$, with $\mathsf{P}(X=k)=(1-p)^{k-1}p$, $k=1,2,\dots$.
- **Poisson:** $X \sim \mathrm{Poi}(\lambda)$, with $\mathsf{P}(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$, $k=0,1,\dots$.

Example 1.2 Examples of continuous r.v.s, described by the density function $\mathsf{P}(X\le a)=\int_{-\infty}^{a}p(x)\,dx$.

- **exponential:** $X \sim \mathrm{Exp}(\lambda)$, with $p(x)=\mathbb{1}_{[0,\infty)}(x)\cdot\lambda e^{-\lambda x}$.
- **uniform:** $X \sim \mathrm{Unif}[a,b]$, with $p(x)=\mathbb{1}_{[a,b]}(x)\cdot\frac{1}{b-a}$.
- **normal/Gaussian:** $X \sim \mathcal{N}(\mu,\sigma^2)$, with $p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$.

Recall that the distribution/law of a r.v. X is determined by its cumulative distribution function (c.d.f.). In particular, sets of the form $\{X\le a\}$ are *events* of which one can evaluate the probability, denoted by $\mathsf{P}(X\le a)$. We can say that $\mathsf{P}(\cdot)$ is a function of events, or a *set function*. A *measure* $\mathsf{P}(\cdot): A\mapsto \mathsf{P}(A)\in[0,\infty)$ is a special set function satisfying the following three properties:

1. **non-negativity**: $\mathsf{P}(A)\ge 0$, $\forall A$.
2. $\mathsf{P}(\varnothing)=0$.
3. **countable additivity**: for any *disjoint* $A_1,A_2,\dots$,

$$\mathsf{P}\Big(\bigcup_{n=1}^{\infty}A_n\Big)=\sum_{n=1}^{\infty}\mathsf{P}(A_n). \tag{1.1}$$

*With contribution from YANG Yuze who typesets some of the note.

The last property, countable additivity (a.k.a. σ-additivity) is the most important one. It is only with σ-additivity, not finite additivity, that one can get the hands on various limit theorems for integration/expectation.

Other important properties of measures can be derived from Item 1 to Item 3.

4. **finite additivity** from Items 2 and 3: let $A_{n+1} = A_{n+2} = \cdots = \varnothing$ in (1.1); then
$$\mathsf{P}\Big(\bigcup_{k=1}^{n} A_k\Big) = \sum_{k=1}^{n} \mathsf{P}(A_k).$$

5. **monotonicity** from Items 1 and 4: if $A \subset B$, then $A \cap (B \setminus A) = \varnothing$, and hence
$$\mathsf{P}(B) = \mathsf{P}(A) + \mathsf{P}(B \setminus A) \geq \mathsf{P}(A).$$

6. **sub-additivity** from Items 3 and 5: let $\tilde{A}_n = A_n \setminus \big(\bigcup_{k=1}^{n-1} A_k\big) \subset A_n$; then
$$\mathsf{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mathsf{P}(\tilde{A}_n) \leq \sum_{n=1}^{\infty} \mathsf{P}(A_n).$$

7. **continuity from above** from Items 2 and 3: if $A_n \downarrow A$ and $\mathsf{P}(A_1) < \infty$, then $\mathsf{P}(A) = \lim_{n\to\infty} \mathsf{P}(A_n)$ ($A = \bigcap_{n=1}^{\infty} A_n$). In fact, since A_1 is the disjoint union of
$$A_1 = A \cup (A_1 \setminus A_2) \cup (A_2 \setminus A_3) \cup \cdots, \tag{1.2}$$
we have
$$\mathsf{P}(A_1) = \mathsf{P}(A) + \mathsf{P}(A \setminus A_n) + \sum_{k=n}^{\infty} \mathsf{P}(A_k \setminus A_{k+1}).$$
All the terms are positive, and the left hand side is finite, so the tail of the infinite sum must converges to 0, and hence
$$\mathsf{P}(A) = \lim_{n\to\infty} \mathsf{P}(A_1) - \mathsf{P}(A \setminus A_n) - \sum_{k=n}^{\infty} \mathsf{P}(A_k \setminus A_{k+1}) = \lim_{n\to\infty} \mathsf{P}(A_1) - \mathsf{P}(A_1 \setminus A_n) = \lim_{n\to\infty} \mathsf{P}(A_n).$$
Note: the decomposition (1.2) has the following interpretation; as A_n is decreasing, any element $x \in A_1$ either appears in all A_n, and hence in A, or there is a largest n such that $x \in A_n$ but $x \notin A_{n+1}$, and hence $x \in A_n \setminus A_{n+1}$.

8. **continuity from below** from Items 2, 3, 5 and 7: if $A_n \uparrow A$, then $\mathsf{P}(A) = \lim_{n\to\infty} \mathsf{P}(A_n)$.
Noting that $\mathsf{P}(A_n)$ is increasing, by sub-additivity,
$$\mathsf{P}(A) \leq \mathsf{P}(A_1) + \sum_{n=2}^{\infty} \mathsf{P}(A_n \setminus A_{n-1}) = \lim_{n\to\infty} \mathsf{P}(A_n).$$
If $\mathsf{P}(A) = \infty$, there is nothing else to prove. Otherwise, $\mathsf{P}(A) < \infty$, and $A - A_n \downarrow \varnothing$. Then by continuity from above,
$$0 = \mathsf{P}(\varnothing) = \lim_{n\to\infty} \mathsf{P}(A \setminus A_n) = \lim_{n\to\infty} \mathsf{P}(A) - \mathsf{P}(A_n).$$

We also need to impose some conditions on the domain of the set function $\mathsf{P}(\cdot)$. The domain should behave well under countable union/intersection. This leads to the definition of σ-algebras.

Definition 1.1 *Let Ω be any non-empty set and $\mathcal{F}$ be a collection of subsets of Ω. We say that $\mathcal{F}$ is a σ-algebra (or σ-field), if*

1. $\Omega \in \mathcal{F}$,
2. $A \in \mathcal{F}$ *implies* $A^c \in \mathcal{F}$,
3. *(closure under countable union)* $A_n \in \mathcal{F}$ *implies* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Example 1.3 1. The smallest σ-algebra: $\mathcal{F} = \{\varnothing, \Omega\}$.

2. The largest σ-algebra: $\mathcal{F} = \{$ all subsets of $\Omega\}$.

A set Ω equipped with a σ-algebra $\mathcal{F}$ is called a *measurable space*, written in a pair $(\Omega, \mathcal{F})$.

Proposition 1.1 [1] *Let $\mathcal{F}$ be a σ-algebra. Then*

- $\varnothing \in \mathcal{F}$,
- $A \subset B$, $A, B \in \mathcal{F}$ *imply* $B \setminus A \in \mathcal{F}$,
- *(closure under countable intersection)* $A_n \in \mathcal{F}$ *implies* $\bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$.

Definition 1.2 *A probability space, or probability triple, $(\Omega, \mathcal{F}, \mathsf{P})$ is such that $(\Omega, \mathcal{F})$ is a measurable space and $\mathsf{P} : \mathcal{F} \to [0, 1]$ is a measure with $\mathsf{P}(\Omega) = 1$.*

If P is only σ-finite, like the Lebesgue measure on $\mathbb{R}$, then $(\Omega, \mathcal{F}, \mathsf{P})$ is called a *measure space*.

Definition 1.3 *A random variable (r.v.) $X = X(\omega) : \Omega \to \mathbb{R}$ is a map from a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ to $\mathbb{R}$, such that*

$$\{\omega : X(\omega) \leq a\} \in \mathcal{F}, \quad \forall a \in \mathbb{R},$$

or written more compactly, $X^{-1}\big((-\infty, a]\big) \in \mathcal{F}$ for all $a \in \mathbb{R}$.

Let us recall some basic facts about the pre-image map φ^{-1} for *any* map $\varphi : U \to V$. It is defined by

$$\varphi^{-1}(W) := \{u \in U : \varphi(u) \in W\}.$$

Proposition 1.2 *The map φ^{-1} commutes with most set operations, in particular:*

- $\varphi^{-1}(W_1 \cap W_2) = \varphi^{-1}(W_1) \cap \varphi^{-1}(W_2)$,
- $\varphi^{-1}(W_1 \cup W_2) = \varphi^{-1}(W_1) \cup \varphi^{-1}(W_2)$,
- $\varphi^{-1}(W^c) = \big(\varphi^{-1}(W)\big)^c$.

Let X be a r.v. on $(\Omega, \mathcal{F}, \mathsf{P})$, and let $\mathcal{B} = \{A \text{ s.t. } X^{-1}(A) \in \mathcal{F}\}$. Definition 1.3 and Proposition 1.2 imply that $\mathcal{B}$ contains all the intervals in $\mathbb{R}$. Moreover, since $\mathcal{F}$ is a σ-algebra,

$$X^{-1}(I_n) \in \mathcal{F} \quad \Rightarrow \quad X^{-1}\Big(\bigcup_{n=1}^{\infty} I_n\Big) = \bigcup_{n=1}^{\infty} X^{-1}(I_n) \in \mathcal{F}.$$

This implies that $\mathcal{B}$ is also a σ-algebra. As we will see in the next section, $\mathcal{B}$ contains the *Borel σ-algebra*, which is the most important class of σ-algebras in probability theory.

[1] In this note, readers are encouraged to work out their own proofs on propositions without proofs; they are good exercises and will be useful for understanding later materials.

1.2 Construction of σ-algebra and (probability) measures

Simply put, the Borel σ-algebra is the *smallest* σ-algebra containing by open sets. To understand what is "smallest", we start with the following observation.

Lemma 1.3 *1. If $\mathcal{F}_1$ and $\mathcal{F}_2$ are two σ-algebras on Ω, then $\mathcal{F}_1 \cap \mathcal{F}_2$ is also a σ-algebra.*

2. If $\mathcal{F}_\gamma, \gamma \in \Gamma$ are σ-algebras on Ω, where Γ is an arbitrary index set (countable or uncountable), then $\bigcap_{\gamma\in\Gamma} \mathcal{F}_\gamma$ is also a σ-algebra.

Proposition 1.4 *Let $\mathcal{A}$ be a collection of subsets in Ω. Then there exists a smallest σ-algebra containing $\mathcal{A}$, called* the σ-algebra generated by $\mathcal{A}$ *and written $\sigma(\mathcal{A})$, in the sense that if $\mathcal{G} \supset \mathcal{A}$ is a σ-algebra, then $\sigma(\mathcal{A}) \subset \mathcal{G}$.*

Proof: Take $\sigma(\mathcal{A}) = \bigcap_{\mathcal{F}\sigma\text{-algebra}:\mathcal{F}\supset\mathcal{A}} \mathcal{F}$. □

Definition 1.4 (Borel σ-algebra) *Let M be a metric space (or any topological space). The* Borel σ-algebra $\mathcal{B}(M)$ *is the σ-algebra generated by all the open sets in M.*

Example 1.4
- $\mathcal{B}(\mathbb{R}) = \sigma\big((-\infty, a], a \in \mathbb{R}\big)$.
- $\mathcal{B}(\mathbb{R}^d) = \sigma\big((-\infty, a_1] \times \cdots \times (-\infty, a_d], a_i \in \mathbb{R}\big)$.

Remark 1.5 Here, one need to first show that any open sets in $\mathbb{R}^d$ can be obtained from countable union of sets of the form $(-\infty, a_1] \times (-\infty, a_d]$. The construction requires some ideas from point-set topology, but it is elementary, and thus omitted here.

Proposition 1.5 *A map $X(\omega)$ on $(\Omega, \mathcal{F}, \mathsf{P})$ is a r.v. if and only if $X^{-1}(A) \in \mathcal{F}$ for any $A \in \mathcal{B}(\mathbb{R})$.*

Remark 1.6 In fact, this is usually taken as the definition for r.v.s.

Now let us take about the distribution of a r.v. X. One can check that $\mu = \mathsf{P} \circ X^{-1}$ defined by

$$\mu(A) = \mathsf{P}\big(\{\omega : X(\omega) \in A\}\big), \quad A \in \mathcal{B}(\mathbb{R}),$$

is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We call μ the *distribution/law* of X. Clearly, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ is a probability space. For most of the practical application, say computing expectation, variance, etc, it is enough to understand the distribution of a r.v., not the original probability measure P on some abstract space that can be potentially be very complicate. Another obvious advantage is that the distributions of all r.v.s are probability measures live on the *same* measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Note that the *cumulative distribution function (c.d.f.)* of a r.v. can be read from its distribution:

$$F_X(a) = \mathsf{P}(X \le a) = \mu\big((-\infty, a]\big), \quad a \in \mathbb{R}.$$

The central topic for this section is to understand how the c.d.f. determines μ. Along the way we will learn how to construct σ-algebras and (probability) measures. Some of the presentation here is taken from [Shi96, Chap. 2.3]. The next theorem is a fundamental and important result.

Theorem 1.6 *Every increasing, right continuous function $F : \mathbb{R} \to [0, 1]$ with $F(-\infty) = 0$ and $F(\infty) = 1$ uniquely determines a probability measure μ on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$.*

We start by introducing some notions on collections of sets.

Definition 1.5 *A collection of sets $\mathcal{S}$ is a* semi-algebra *if first, it is closed under intersection, that is, $A \cap B \in \mathcal{S}$ whenever $A, B \in \mathcal{S}$ and second, for every $A \in \mathcal{S}$, its complement A^c is disjoint union of some $A_1, A_2, \ldots, A_n$ in $\mathcal{S}$.*

A collection of sets $\mathcal{S}$ is an algebra*, or* field*, if $A, B \in \mathcal{S}$ implies $A \cap B \in \mathcal{S}$ and $A^c \in \mathcal{S}$.*

These two notions are related by the following proposition.

Proposition 1.7 *Let $\mathcal{S}$ be a semi-algebra. Then*

$$\bar{\mathcal{S}} = \{\textit{finite disjoint unions of sets in } \mathcal{S}\}$$

is an algebra.

Example 1.7 All the d-dimensional half-open, half-closed rectangles forms a semi-algebra:

$$\mathcal{S}_d = \{\varnothing, \quad (a_1, b_1] \times \cdots \times (a_d, b_d], \ -\infty \le a_i < b_i \le \infty\}.$$

Definition 1.6 *A collection of sets $\mathcal{S}$ is a* monotone class*, if $\lim_{n\to\infty} A_n \in \mathcal{S}$ for every monotone sequence of sets $A_n \in \mathcal{S}$.*

Here, for an increasing sequence $A_n \subset A_{n+1} \subset \cdots$, its limit is defined by $A := \bigcup_{n=1}^{\infty} A_n$, and for an decreasing sequence $A_n \supset A_{n+1} \supset \cdots$, its limit is defined by $A := \bigcap_{n=1}^{\infty} A_n$.

It is easy to see that any *intersection* of monotone classes is still an m-class. Therefore, it makes sense to talk about the *smallest* monotone classes containing any collection of sets $\mathcal{A}$ (c.f. Proposition 1.4). We denote this smallest monotone class by $m(\mathcal{A})$.

The monotone class condition basically bridges the difference between σ-algebras and algebras.

Proposition 1.8 *Let $\mathcal{A}$ be a collection of subsets of Ω. Then $\mathcal{A}$ is a σ-algebra if and only if $\mathcal{A}$ is both an algebra and a monotone class.*

Theorem 1.9 (Monotone Class Theorem) *Let $\mathcal{A}$ be an algebra. Then $\sigma(\mathcal{A}) = m(\mathcal{A})$.*

Proof: By Proposition 1.8, $\sigma(\mathcal{A})$ is necessarily a monotone class, and by the minimum property we have the inclusion $m(\mathcal{A}) \subset \sigma(\mathcal{A})$.

To show the other direction $\sigma(\mathcal{A}) \subset m(\mathcal{A})$, it suffices to show that $m(\mathcal{A})$ is an algebra, and hence a σ-algebra (using Proposition 1.8 again). To establish that $m(\mathcal{A})$ is an algebra, we will use the *principle of appropriate sets*.

First, $m(\mathcal{A})$ is closed under complement. Let

$$\mathcal{S} = \{A : A, A^c \in m(\mathcal{A})\} \subset m(\mathcal{A}).$$

Our goal is to show that $m(\mathcal{A}) = \mathcal{S}$. Clearly, by definition we have $\mathcal{A} \in \mathcal{S}$. Moreover, $\mathcal{S}$ is a monotone class: if $A_n \uparrow A$ and $A_n \in \mathcal{S}$, then (A_n) and (A_n^c) are both monotone sequences in $m(\mathcal{A})$, and hence their respective limits A and A^c are in $m(\mathcal{A})$; if $A_n \downarrow A$ it is similar. Therefore, $\mathcal{S}$ must contain the smallest monotone class that contains $\mathcal{A}$, which is $m(\mathcal{A})$. This shows $\mathcal{S} = m(\mathcal{A})$, and hence by the definition of $\mathcal{S}$, the collection of set $m(\mathcal{A})$ is closed under complement.

Second, $m(\mathcal{A})$ is closed under intersection. Since intersection involves two sets, the proof is slightly more complicated and we will do it in two steps. In the first step, for a fixed $A \in \mathcal{A}$, let

$$\mathcal{S}_A = \{B : B \in m(\mathcal{A}), \ A \cap B \in m(\mathcal{A})\} \subset m(\mathcal{A}).$$

It is clear that $\mathcal{A} \subset \mathcal{S}_A$ since A is an algebra and $m(\mathcal{A})$ contains $\mathcal{A}$. Also, one can check that $\mathcal{S}_A$ is a monotone class since $B_n \downarrow B$ or $B_n \uparrow B$ implies $A \cap B_n \downarrow A \cap B$ or $A \cup B_n \uparrow A \cup B$. Therefore, we have $m(\mathcal{A}) \subset \mathcal{S}_A$, and this means that $A \cap B \in m(\mathcal{A})$ whenever $A \in \mathcal{A}$ and $B \in m(\mathcal{A})$.

In the second step, let

$$\mathcal{S} = \{A \in m(\mathcal{A}) : A \cap B \in m(\mathcal{A}),\ \forall B \in m(\mathcal{A})\}.$$

The first step implies that $\mathcal{A} \subset \mathcal{S}$. Again, it is not hard to check that A is a monotone class. Hence $m(\mathcal{A}) = \mathcal{S}$ and this proves that $m(\mathcal{A})$ is closed under intersection.

In conclusion, $m(\mathcal{A})$ is an algebra and hence a σ-algebra, this completes the proof. □

A related concept is the Dynkin system (d-system, λ-class).

Definition 1.7 *Let $\mathcal{D}$ be a collection of subsets of Ω. We say that $\mathcal{D}$ is a Dynkin system if*

1. $\Omega \in \mathcal{D}$,
2. $A, B \in \mathcal{D},\ A \subset B \quad \Rightarrow \quad B \setminus A \in \mathcal{D}$,
3. $A_n \uparrow A,\ A_n \in \mathcal{D} \quad \Rightarrow \quad A \in \mathcal{D}$.

We say that $\mathcal{A}$ is a π-system if it is closed under intersection. One can check that $\mathcal{A}$ is a σ-algebra if and only if it is both a π-system and Dynkin system. Moreover, analogous to Theorem 1.9, the following is true.

Theorem 1.10 *(π-λ Theorem; Dynkin Theorem) If $\mathcal{A}$ is a π-system, then $\sigma(\mathcal{A})$ is the smallest Dynkin system containing $\mathcal{A}$.*

Proof: The proof can be done via the principle of appropriate sets. □

Given a distribution function F as in Theorem 1.6, we can introduce a (probability) measure μ_0 on the algebra

$$\bar{\mathcal{S}} = \Big\{ \bigcup_{k=1}^{n} (a_k, b_k],\ \text{disjoint union} \Big\},$$

given by

$$\mu_0(A) = \sum_{k=1}^{n} \big[F(b_k) - F(a_k)\big].$$

It is easy to check that μ_0 is finitely additive. An important step is the following.

Proposition 1.11 *The finitely additive measure μ_0 is σ-additive on $\bar{\mathcal{S}}$, i.e., if $A_n \in \bar{\mathcal{S}}$ are disjoint and $\bigcup_{n=1}^{\infty} A_n \in \bar{\mathcal{S}}$, then*

$$\mu_0\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mu_0(A_n).$$

Proof: We will use the fact that σ-additivity is equivalent to continuity at $\varnothing$, i.e., μ_0 is σ-additive if and only if $\lim_{n\to\infty} \mu_0(A_n) = \mu_0(\varnothing) = 0$ whenever $A_n \downarrow \varnothing$.

Suppose that there is some $L > 0$ such that $A_n \in [-L, L]$. Let $\varepsilon > 0$. We claim that there exists $B_n \in \bar{\mathcal{S}}$ such that $\overline{B_n} \subset A_n$ and

$$\mu_0(A_n) - \mu_0(B_n) \le \varepsilon \cdot 2^{-n}.$$

The existence of B_n is a consequence of the right continuity of F. In fact, writing $A_n = \bigcup_{i=1}^{m} (a_i^{(n)}, b_i^{(n)}]$, and $B_n = \bigcup_{i=1}^{m} (a_i^{(n)} + \delta, b_i^{(n)}]$, we have

$$\mu_0(A_n) - \mu(B_n) = \sum_{i=1}^{m} \big(F(b_i^{(n)} + \delta) - F(b_i^{(n)})\big) \to 0, \quad \delta \downarrow 0.$$

By choosing δ small enough we can make the sum less than $\varepsilon \cdot 2^{-n}$.

Since $A_n \downarrow \varnothing$ and $\overline{B_n} \subset A_n$, we have $\overline{B_n} \downarrow \varnothing$. So $C_n = [-L, L] \setminus \overline{B_n}$ forms an open cover of $[-L, L]$. By the Finite Open Cover Theorem, there exists a finite sub-cover, i.e., there exists n_0 such that

$$[-L, L] \subset \bigcup_{n=1}^{n_0} [-L, L] \setminus \overline{B_n},$$

and hence $\bigcap_{n=1}^{n_0} \overline{B_n} = \varnothing$. Therefore,

$$\mu_0(A_{n_0}) = \mu_0\Big(A_{n_0} \setminus \bigcap_{n=1}^{n_0} B_n\Big) \le \mu_0\Big(\bigcup_{n=1}^{n_0} (A_n \setminus B_n)\Big) \le \sum_{n=1}^{n_0} \mu_0(A_n \setminus B_n) \le \varepsilon \sum_{n=1}^{\infty} 2^{-n} \le \varepsilon.$$

Since $\mu_0(A_n)$ is decreasing and ε is arbitrary, we obtain $\lim_{n\to\infty} \mu_0(A_n) = 0$.

When A_n are unbounded, since $F(-\infty) = 0$ and $F(\infty) = 1$, for every $\varepsilon > 0$, we can choose L large enough so that $\mu_0(-L, L] \ge 1 - \varepsilon$. Let $\tilde{A}_n = A_n \cap (-L, L]$. Then $\tilde{A}_n \downarrow \varnothing$ and $\tilde{A}_n$ are bounded. Then $\lim_{n\to\infty} \mu_0(\tilde{A}_n) = 0$ as previously proved, and hence

$$\limsup_{n\to\infty} \mu_0(A_n) \le \limsup_{n\to\infty} \mu_0(\tilde{A}_n) + \limsup_{n\to\infty} \mu_0\big(A_n \setminus (-L, L]\big) \le 0 + \varepsilon = \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we obtain $\lim_{n\to\infty} \mu_0(A_n) = 0$ as desired. □

After establishing σ-additivity of μ_0 on $\bar{\mathcal{S}}$ using Proposition 1.11, we can extend μ_0 to a probability measure on $\sigma(\bar{\mathcal{S}}) = \mathcal{B}(\mathbb{R})$ with the help of the next theorem.

Theorem 1.12 (Carathéodory's Extension Theorem) *Let μ_0 be a σ-additive measure on an algebra $\mathcal{A}$. Then μ_0 has a unique extension to $\sigma(\mathcal{A})$.*

Here, an extension of μ_0 to $\sigma(\mathcal{A})$ is a measure μ on $\sigma(\mathcal{A})$ such that $\mu_0(A) = \mu(A)$ for every $A \in \mathcal{A}$.

Remark 1.8 We will use Theorem 1.12 in the case where μ_0 (and hence the resulting extension μ) is a *probability* measure. But the theorem also holds when μ_0 is σ-finite, which means that there exist $A_n \uparrow \Omega$ such that $\mu_0(A_n) < \infty$.

Proof of Uniqueness: Let μ, $\tilde{\mu}$ be two extensions and $\mathcal{S} = \{A : \mu(A) = \tilde{\mu}(A)\}$. We will show (i) $\mathcal{A} \subset \mathcal{S}$; (ii) $\mathcal{A}$ is a monotone class. Then, by Theorem 1.9, $\mathcal{S}$ contains $\sigma(\mathcal{A})$, so $\mu = \tilde{\mu}$ on $\sigma(\mathcal{A})$, which is the uniqueness.

The first statement $\mathcal{A} \subset \mathcal{S}$ follows from definition of the extension.

To prove the second statement, let $A_n \uparrow A$ and $A_n \in \mathcal{S}$. Since μ and $\tilde{\mu}$ are measures, and measures are continuous from below, we have $\mu(A_n) \to \mu(A)$ and $\tilde{\mu}(A_n) \to \tilde{\mu}(A)$, and thus $\mu(A) = \tilde{\mu}(A)$. Similarly, if $A_n \downarrow A$ and $A_n \in \mathcal{S}$, since μ is the continuous from above, we have $\mu(A_n) \to \mu(A)$ and $\tilde{\mu}(A_n) \to \tilde{\mu}(A)$, and thus $\mu(A) = \tilde{\mu}(A)$. This completes the proof of uniqueness. □

To prove the existence we need to use the outer measure, which is also a standard procedure in constructing the Lebesgue measure. We will only sketch the most important steps in this note.

Given a σ-additive measure μ_0 on an algebra $\mathcal{A}$, the *outer measure*, defined for *any* sets, is

$$\mu_*(A) = \inf\Big\{\sum_{n=1}^{\infty} \mu_0(A_n) : A \subset \bigcup_{n=1}^{\infty} A_n,\ A_n \in \mathcal{A}\Big\}.$$

For the Lebesgue measure, $\mathcal{A}$ consists of nice sets like intervals, rectangles, etc, and the outer measure is the generalization of length, area, volume, etc. But the outer measure cannot be measure, since the

latter is not defined for arbitrary sets. A key point is to defined what is "measurable" w.r.t. the outer measure μ_*. We say a set A is measurable, if it satisfies the *Carathéodory's condition*:

$$\mu_*(D) = \mu_*(D \cap A) + \mu_*(D \cap A^c), \quad \forall D. \tag{1.3}$$

With some more efforts, one can show:

1. every set $A \in \mathcal{A}$ satisfies (1.3) and $\mu_*(A) = \mu_0(A)$;
2. the collection of sets that satisfy (1.3), denoted by $\mathcal{F}$, forms a σ-algebra, and moreover, μ_* is a measure on $\mathcal{F}$.

The desired extension is then defined by $\mu := \mu_*\big|_{\sigma(A)}$.

Remark 1.9 Typically, $\sigma(\mathcal{A})$ is a proper subset of $\mathcal{F}$. For example, in the case of constructing *Lebesgue measure*, we have $F(x) = x$ and

$$\sigma(\mathcal{A}) = \{\text{Borel sets}\}, \quad \mathcal{F} = \{\text{Lebesgue measurable sets}\}.$$

In Proposition 1.16 we will see that there exist Lebesgue measurable sets which are not Borel.

However, if we complete $(\Omega, \sigma(\mathcal{A}), \mu)$, then the result is $(\Omega, \mathcal{F}, \mu_*\big|_{\mathcal{F}})$. Here, a *complete* measure space $(\Omega, \mathcal{F}, \mathsf{P})$ means that if $B \subset A \in \mathcal{F}$ such that $\mathsf{P}(A) = 0$, then $B \in \mathcal{F}$.

1.3 Decomposition of distribution functions

Let $F(x)$ be an increasing, right continuous function, e.g., the c.d.f. of some r.v. The goal of this section is to decompose it into the jumping (or discontinuous) part, the absolutely continuous part and the singularly continuous part, written

$$F = F_d + F_{ac} + F_{sc}.$$

First, let us look at the discontinuous part. Since F is right continuous and increasing, F only has discontinuity points of the first kind. This leads to the following definition.

Definition 1.8 *A point x is a point of jump/discontinuity of F if $F(x) - F(x-) > 0$.*

Proposition 1.13 *The points of jump for an increasing, right continuous function are countable.*

Proof: On any compact set $[-L, L]$,

$$\{x \in [-L, L] \text{ is a jump}\} = \bigcup_{n=1}^{\infty} \Big\{x \in [-L, L] : F(x) - F(x-) > \frac{1}{n}\Big\}.$$

All sets in the union are finite, since each contains at most $n\big(F(L) - F(L-)\big)$ points. The conclusion then follows. □

Let a_i, $i = 1, 2, \ldots$, be the points of jump for the function $F(x)$ and let $b_i = F(a_i) - F(a_i-)$ be the "size of jumps". Define

$$F_d(x) = \sum_{i=1}^{\infty} b_i \mathbb{1}_{[a_i, \infty)}(x).$$

We call F_d the "jumping part". The remaining part $F_c(x) = F(x) - F_d(x)$ is increasing and continuous.

Next we need to classify increasing and continuous functions.

Definition 1.9 (Absolute Continuity) *An increasing, continuous function $F(x)$ is called absolutely continuous if there exist $f \in L^1(\mathbb{R})$ such that*

$$F(b) - F(a) = \int_a^b f(x)\,dx. \tag{1.4}$$

Remark 1.10 This is the generalized Newton–Leibniz formula. By Lebesgue Differentiability Theorem, if (1.4) holds, then F' exists almost everywhere and $F' = f$.

On the other hand, using the Vitali covering theorem in real analysis, we know that an increasing functions is differentiable almost everywhere.

Proposition 1.14 *If F is increasing, then F' exists almost everywhere.*

Note that non-differentiable points in Proposition 1.14 could be points of jumps. But if we are looking at continuous, increasing functions, we have the following.

Proposition 1.15 *An increasing and continuous function F can be uniquely decomposed as*

$$F = F_{ac} + F_{sc},$$

where F_{ac} is absolutely continuous and $F_{ac} = \int_{-\infty}^x F'(x)\,dx$, and F_{sc} is increasing and continuous but $F'_{sc} \overset{a.e.}{\equiv} 0$.

Remark 1.11 The function F_{sc} appearing in Proposition 1.15 is called *singularly continuous*. One may ask if there exists non-trivial singularly continuous function. A famous example is the Cantor function, or the "Devil's staircase".

Recall that the *Cantor set*, denoted by $\mathcal{C}$, is constructed by starting with the interval $[0,1] \subset \mathbb{R}$, then dividing it into three intervals of equal length and removing the middle interval, and repeating this process of division and removal. In the end, we obtain

$$\mathcal{C} = [0,1] \setminus \bigcup_{n,k} I_n^{(k)},$$

where $I_n^{(k)}$, $1 \le k \le 2^{n-1}$, $n \ge 1$, are the intervals that are removed in the n-th steps, i.e.,

$$I_1^{(1)} = \Big(\frac{1}{3}, \frac{2}{3}\Big), \quad I_2^{(1)} = \Big(\frac{1}{9}, \frac{2}{9}\Big), \quad I_2^{(2)} = \Big(\frac{7}{9}, \frac{8}{9}\Big), \cdots.$$

Clearly, the set $\mathcal{C}$ is a closed set, and from a direct calculation of the total length of the removed intervals, one can show that $\mathcal{C}$ has Lebesgue measure 0.

The *Cantor function*, denoted by $\varphi(x)$, is an increasing function constructed as follows. Set $\varphi(x) = 0$ for $x \le 0$ and $\varphi(x) = 1$ for $x \ge 1$. When $x \in (0,1)$, set $\varphi(x) = \frac{1}{2}$ for $x \in (\frac{1}{3}, \frac{2}{3}) = I_1^{(1)}$, $\varphi(x) = \frac{1}{4}$ for $x \in (\frac{1}{9}, \frac{2}{9}) = I_2^{(1)}$, and $\varphi(x) = \frac{3}{4}$ for $x \in (\frac{7}{9}, \frac{8}{9}) = I_2^{(2)}$ and so on. Then define φ on $\mathcal{C}$ by monotonicity. It follows from the construction that φ is also continuous. See also [Dur19, Fig. 1.5].

We can use the Cantor set and the Cantor function to show the following.

Proposition 1.16 *There exists a Lebesgue measurable set which is not Borel measurable.*

Proof: We will prove the statement by contradiction.

Let $\psi(x) = \frac{1}{2}\big(x + \varphi(x)\big)$. Then $\psi(x)$ is a continuous, strictly increasing function from $[0,1]$ onto itself. Let $H = \psi^{-1}$. Then H is also continuous and strictly increasing.

It is easy to check that for any $E \subset [0,1]$,

$$\mathbb{1}_{H(E)}\big(H(x)\big) = \mathbb{1}_E(x).$$

Note that the Lebesgue measure of $\psi(\mathcal{C})$ is $1/2$. Hence, there exists a set $E \subset \psi(\mathcal{C})$ which is NOT Lebesgue measurable. On the other hand, $H(E) = \psi^{-1}(E) \subset \mathcal{C}$ is a subset of Lebesgue measure 0 set, and hence by completeness of the Lebesgue measure space (as a consequence of using outer measure in Theorem 1.12), it is also Lebesgue measurable.

Now, if all Lebesgue measurable sets are Borel, then $\mathbb{1}_{H(E)}$ will be Borel measurable as the indicator function of a Borel set. Therefore, $\mathbb{1}_E = \mathbb{1}_{H(E)} \circ H$ is the composition of two Borel measurable functions, and is also Borel measurable. But this contradicts with the fact that E is chosen to be non-measurable. □

In the first part of this section, we classify and decompose the distribution functions. In the second part, we will do similar things from the perspective of measures.

Let μ be a measure on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$.

Definition 1.10 *A point x is a point of mass if $\mu(\{x\}) > 0$.*

Let $I = \big\{x : \mu(\{x\}) > 0\big\}$ be the set of points of mass. We can define $\mu_d(A) = \sum_{x \in I} \delta_x(A) \cdot \mu(\{x\})$.

$$\delta_x(A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

is the *Dirac measure* on x. We call μ_d the discrete part of the measure μ, and this corresponds to the jumping part of the distribution functions.

The remaining part $\mu_c = \mu - \mu_d$ will not have points of mass. To further decompose it, we need to introduce the notion of absolute continuity and singularity for measures. Let P, Q are two probability measures on $(\Omega, \mathcal{F})$. For the simplest example, one can take $(\Omega, \mathcal{F}) = \big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$.

Definition 1.11 *A measure P is absolutely continuous w.r.t. Q, written $P \ll Q$, if $Q(A) = 0$ implies $P(A) = 0$.*

We recall the Radon–Nikodym derivative.

Theorem 1.17 (Radon–Nikodym Theorem) *Let ν and μ be two σ-finite measures on a measurable space $(\Omega, \mathcal{F})$ such that $\nu \ll \mu$. Then there exists a function f, measurable w.r.t. $\mathcal{F}$, such that*

$$\int_A f \, d\mu = \nu(A).$$

We call $f = \frac{d\nu}{d\mu}$ the Radon–Nikodym derivative, and μ the reference measure.

For r.v.s, the reference measure is the Lebesgue measure.

Definition 1.12 *A r.v. X is continuous if its distribution μ is absolutely continuous with respect to the Lebesgue measure. In this case, the density of X is $\frac{d\mu}{d\mathrm{Leb}}$.*

The last definition is mutual singularity.

Definition 1.13 *Two measures P, Q are mutuality singular, denoted by $P \perp Q$, if there exists A such that $P(A) = 0$ and $Q(A^c) = 0$.*

Example 1.12 Cantor set induce a distribution $\mu_{\mathcal{C}} = d\varphi$. Since

$$\mu_{\mathcal{C}}\big(\mathcal{C}^c\big) = 0, \quad \mathrm{Leb}\big(\mathcal{C}\big) = 0,$$

we have $\mu_{\mathcal{C}} \perp \mathrm{Leb}$. In fact, an increasing function F is singularly continuous if and only if $dF \perp \mathrm{Leb}$.

Definition 1.14 *A r.v. X is singular if $\mu_X \perp \text{Leb}$.*

How common are singular measures and Cantor-like sets? Surprisingly, they are ubiquitous in probability theory. They usually arise from self-similarities or fractal structures, or from infinite dimensional spaces.

Example 1.13 The example is about Brownian motion, which is a important object to study in stochastic analysis. Without getting into too many details, a Brownian motion $B_t(\omega)$ is a random continuous function.

For each $a \in \mathbb{R}$,

$$\mathcal{Z}_a(\omega) := \{t : B_t(\omega) = a\}.$$

be the level set of the Brownian motion; note the level set is also a random set. For almost every ω and every a, the level $\mathcal{Z}_a(\omega)$ is very similar to a Cantor set, in the sense that it is the complement of the union of nested open intervals, but the interval length can be very random.

To get singular measures, consider the maximal process $B_t^* = \sup_{0\le s\le t} B_s$. Since $t \mapsto B_t$ is continuous, the maximal process B_t^* is increasing and continuous. One can show that $dB_t^* \perp \text{Leb}$.

Example 1.14 Let us consider i.i.d. Bernoulli r.v.s Ber(1/3) and Ber(2/3). More precisely, let $(\Omega, \mathcal{F})$ be

$$\Omega = \big\{\omega = (\omega_1, \omega_2, \cdots),\ \omega_i \in \{0, 1\}\big\}, \quad \mathcal{F} = \mathcal{P}(\Omega).$$

We can define two probability measures on $(\Omega, \mathcal{F})$:

1. one corresponding to i.i.d. Ber(1/3): $\mathsf{P}_1(\omega_i = 1) = \frac{1}{3}$ and $\mathsf{P}_1(\omega_i = 0) = 2/3$;
2. the other one corresponding to i.i.d. Ber(2/3): $\mathsf{P}_2(\omega_i = 1) = \frac{2}{3}$ and $\mathsf{P}_2(\omega_i = 0) = 1/3$.

Let

$$A_1 = \Big\{\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \omega_k = \frac{1}{3}\Big\}, \quad A_2 = \Big\{\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \omega_k = \frac{2}{3}\Big\}.$$

Then by the Strong Law of Large Numbers, we have $\mathsf{P}_1(A_1) = 1$ and $\mathsf{P}_2(A_2) = 1$. On the other hand, we have $A_1 \cap A_2 = \emptyset$. It follows that $\mathsf{P}_1(A_2) = 0$ and $\mathsf{P}_2({A_2}^c) = 0$, so $\mathsf{P}_1 \perp \mathsf{P}_2$.

1.4 Random variables and measurable maps

Let $(S, \mathcal{S})$ be a measurable space. We say that a map $\varphi : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ is *measurable* if $\varphi^{-1}(A) \in \mathcal{F}$, $\forall A \in \mathcal{S}$. Random variables and vectors require such measurability.

Definition 1.15 *A r.v. X is a measurable map from $(\Omega, \mathcal{F})$ to $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$. A random vector $X = (X_1, \ldots, X_d)$ is a measurable map from $(\Omega, \mathcal{F})$ to $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$.*

Since the Borel σ-algebra is generated by open sets, we have a simple criterion to check whether a map defines a r.v.

Proposition 1.18 *A map X is a random variable if and only if $X^{-1}(O) \in \mathcal{F}$ for every open set O.*

Definition 1.16 *A function f is a Borel measurable if f is measurable map from $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ onto itself.*

Similar to Proposition 1.18, we have the following.

Proposition 1.19 *A function f is Borel measurable if and only if $f^{-1}(O) \subset \mathcal{B}(\mathbb{R})$ for every open set O.*

To compare with the Lebesgue measurability: f is Lebesgue measurable if and only if $f^{-1}(O)$ is Lebesgue measurable set for every open set O.

Proposition 1.20 *If f is Borel measurable and X is a random variable, then $f(X)$ is a r.v.*

Proof: Let O be a open set. Then $f^{-1}(O) \in \mathcal{B}(\mathbb{R})$ since f is Borel measurable. Hence,

$$\{\omega : f\big(X(\omega)\big) \in O\} = X^{-1}\big(f^{-1}(O)\big) \in \mathcal{F}.$$

This shows that $f(X)$ is a r.v. □

Remark 1.15 In this example, if "f is Borel measurable" is replaced by "f is Lebesgue measurable", then the conclusion is false, as seen from the proof of Proposition 1.16.

We often drop the word "measurable" and simply say "Borel sets" or "Borel functions".

Proposition 1.21 *If $f : \mathbb{R} \to \mathbb{R}^d$ is a Borel map and $X = (X_1, \ldots, X_d)$ is a random vector, then $f(X) = f(X_1, \ldots, X_d)$ is a random variable.*

Example 1.16 We can use Proposition 1.21 to create new r.v.s. For example, if X_1, X_2 are r.v.s, then $X_1 + X_2$, $\min\{X_1, X_2\}$ are also r.v.s.

Next, we need to understand the limits of r.v.s.

Proposition 1.22 *Let X_n, $n = 1, 2, \ldots$ be r.v.s. Then*

$$\sup_{n\geq 1} X_n, \quad \inf_{n\geq 1} X_n, \quad \limsup_{n\to\infty} X_n, \quad \liminf_{n\to\infty} X_n$$

are r.v.s.

Proof:

(i) Let $Y_1(\omega) = \sup_n X_n(\omega)$. We need to show that $Y_1^{-1}(-\infty, a] \in \mathcal{F}$ for every $a \in \mathbb{R}$. Indeed,

$$Y_1^{-1}(-\infty, a] = \{\omega : \sup_n X_n(\omega) \leq a\} = \bigcap_{n=1}^{\infty} \{\omega : X_n(\omega) \leq a\} \in \mathcal{F}.$$

Therefore, $Y_1(\omega) = \sup_n X_n(\omega)$ is a r.v.

(ii) Let $Y_2(\omega) = \inf_n X_n(\omega)$. We need to show that $Y_2^{-1}([a, \infty)) \in \mathcal{F}$ for every $a \in \mathbb{R}$. Indeed,

$$Y_2^{-1}[a, -\infty) = \{\omega : \inf_n X_n(\omega) \geq a\} = \bigcap_{n=1}^{\infty} \{\omega : X_n(\omega) \geq a\} \in \mathcal{F}.$$

Therefore, $Y_2(\omega) = \inf_n X_n(\omega)$ is a r.v.

(iii) By definition of lim sup, for every ω, we have

$$\limsup_{n\to\infty} X_n(\omega) = \inf_{n\geq 1} \sup_{m\geq n} X_m(\omega).$$

By part (i), for every $n \geq 1$, the map $\omega \mapsto \sup_{m\geq n} X_m(\omega)$ is measurable. Hence, for every $a \in \mathbb{R}$,

$$\{\omega : \limsup_{n\to\infty} X_n(\omega) \geq a\} = \{\omega : \inf_{n\geq 1} \sup_{m\geq n} X_m(\omega) \geq a\} = \bigcap_{n=1}^{\infty} \{\omega : \sup_{m\geq n} X_m(\omega) \geq a\} \in \mathcal{F}.$$

(iv) By definition of lim inf, for every ω, we have

$$\liminf_{n\to\infty} X_n(\omega) = \sup_{n\geq 1} \inf_{m\geq n} X_m(\omega).$$

By part (ii), for every $n \geq 1$, the map $\omega \mapsto \inf_{m\geq n} X_m(\omega)$ is measurable. Hence, for every $a \in \mathbb{R}$,

$$\{\omega : \liminf_{n\to\infty} X_n(\omega) \leq a\} = \{\omega : \sup_{n\geq 1} \inf_{m\geq n} X_m(\omega) \leq a\} = \bigcap_{n=1}^{\infty} \{\omega : \inf_{m\geq n} X_m(\omega) \leq a\} \in \mathcal{F}.$$

□

Corollary 1.23 *Let X_n, $n = 1, 2, \ldots$, be r.v.s. The set $\Omega_0 = \{\omega : \lim_{n\to\infty} X_n(\omega)$ exists$\}$ belongs to $\mathcal{F}$.*

Proof: Note that

$$\Omega_0 = \{\omega : \lim_{n\to\infty} X_n(\omega)\} = \{\omega : \limsup_{n\to\infty} X_n(\omega) - \liminf_{n\to\infty} X_n(\omega) = 0\}.$$

By Proposition 1.22, $Y_1 = \limsup_{n\to\infty} X_n(\omega)$ and $Y_2 = \liminf_{n\to\infty} X_n(\omega)$ are r.v.s, and hence $Y_1 - Y_2$ is a r.v. Therefore, $\Omega_0 = \{Y_1 - Y_2 = 0\} \in \mathcal{F}$. □

1.5 Integration and expectation

In this section, we will briefly present the theory of integration of measurable functions, or in the context of probability theory, the mathematical expectation. The main difference is that in probability theory, the probability measure has total mass 1 and is a finite measure.

Let X be a r.v. on $(\Omega, \mathcal{F}, \mathsf{P})$. We will denote its expectation X by $\mathsf{E}(X)$, or using a more measure theory oriented notation, sometimes we also write

$$\mathsf{E}X = \int_\Omega X(\omega)\, \mathsf{P}(d\omega). \tag{1.5}$$

The definition of (1.5) is through approximation via simple r.v.s (simple functions in measure theory). To start, we say that a r.v. $X(\omega)$ is *simple*, if there exists finitely many $A_1, \ldots, A_n \in \mathcal{F}$ and $c_1, \ldots, c_n \in \mathbb{R}$ such that

$$X(\omega) = \sum_{k=1}^{n} c_k \mathbb{1}_{A_k}(\omega). \tag{1.6}$$

In the case of (1.6), unquestionably we should define

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

It is routine to verify common integral properties for expectation of simple r.v.s, e.g., linearity, monotonicity, order preserving, etc, so we omit it in this note.

For a non-negative r.v. $X(\omega)$, we define

$$\mathsf{E}X = \int_\Omega X(\omega)\mathsf{P}(d\omega) := \sup\Big\{\int Y(\omega)\mathsf{P}(d\omega) : Y \text{ simple}, \ 0 \leq Y(\omega) \leq X(\omega)\Big\} \in [0, \infty]. \tag{1.7}$$

For the general case, we write $X(\omega) = X_+(\omega) - X_-(\omega)$, where

$$X_+(\omega) = X(\omega)\mathbb{1}_{\{X>0\}}, \quad X_-(\omega) = -X(\omega)\mathbb{1}_{\{X\le 0\}}$$

are the positive and negative parts of X. If $\mathsf{E}(X_+) < \infty$ or $\mathsf{E}(X_-) < \infty$, then we define

$$\mathsf{E}(X) = \mathsf{E}(X_+) - \mathsf{E}(X_-).$$

Otherwise, $\mathsf{E}X$ is undefined since $\infty - \infty$ cannot be defined.

Next, we will discuss conditions that justifies exchanging order of limit and integration, i.e.,

$$\mathsf{E} \lim_{n\to\infty} X_n = \lim_{n\to\infty} \mathsf{E}X_n. \tag{1.8}$$

Lemma 1.24 *Let $X_n \uparrow X$ such that $X_n \ge 0$ and X_n are simple. Then (1.8) holds.*

Remark 1.17 If "$X_n \uparrow X$" is replaced by "$X_n \le X$ and $X_n \to X$", we can still an get increasing sequence by considering $Y_n = \max_{1\le k\le n} X_k$. It is easy to see that Y_n are also simple and $Y_n \uparrow X$.

Proof: From the definition (1.7), we have $\mathsf{E}(X) \ge \mathsf{E}(X_n)$. It remains to establish the inequality in the other direction:

$$\mathsf{E}X \le \lim_{n\to\infty} \mathsf{E}X_n. \tag{1.9}$$

Note that the limit on the right hand side always exists, since X_n, and hence $\mathsf{E}X_n$, are increasing in n.

If $\mathsf{E}X < \infty$, then for every $\varepsilon > 0$, by the definition of supremum, there exists a non-negative simple r.v. Y_ε such that $Y_\varepsilon \le X$ and $\mathsf{E}(Y_\varepsilon) \ge \mathsf{E}(X) - \varepsilon$. For every $\delta > 0$, let $A_n = \{\omega : X_n(\omega) \le Y_\varepsilon(\omega) - \delta\}$. Since $X_n(\omega) \uparrow X(\omega) \ge Y_\varepsilon(\omega)$, we have $A_n \uparrow \Omega$ and hence $A_n^c \downarrow \emptyset$. We have

$$\begin{aligned}
\mathsf{E}X_n &= \mathsf{E}X_n\mathbb{1}_{A_n} + \mathsf{E}X_n\mathbb{1}_{A_n^c} \ge \mathsf{E}(Y_\varepsilon - \delta)\mathbb{1}_{A_n} \\
&= \mathsf{E}Y_\varepsilon\mathbb{1}_{A_n} - \delta\mathsf{P}(A_n) \\
&= \mathsf{E}Y_\varepsilon - \mathsf{E}Y_\varepsilon\mathbb{1}_{A_n^c} - \delta\mathsf{P}(A_n) \\
&\ge \mathsf{E}X - \varepsilon - \sup_\omega Y_\varepsilon(\omega)\cdot\mathsf{P}(A_n^c) - \delta
\end{aligned}$$

Since Y_ε is simple, it is always bounded, so $\sup_\omega Y_\varepsilon(\omega) < \infty$. Letting $n \to \infty$, we obtain

$$\lim_{n\to\infty} \mathsf{E}X_n \ge \mathsf{E}X - \varepsilon - \delta.$$

Since $\varepsilon, \delta > 0$ are arbitrary, this implies (1.9).

If $\mathsf{E}X = \infty$, then by (1.7), for every $M > 0$, there exists a simple r.v. Y_M such that $Y_M \le X$ and $\mathsf{E}Y_M \ge M$. For every $\xi > 0$, let $B_n = \{\omega : X_n(\omega) \ge Y_M(\omega) - \xi\}$. Since $X_n(\omega) \uparrow X(\omega) \ge Y_M(\omega)$, we have $B_n \uparrow \Omega$ and hence $B_n^c \downarrow \emptyset$. Therefore,

$$\begin{aligned}
\mathsf{E}X_n &= \mathsf{E}X_n\mathbb{1}_{B_n} + \mathsf{E}X_n\mathbb{1}_{B_n{}^c} \ge \mathsf{E}(Y_M - \xi)\mathbb{1}_{B_n} \\
&= \mathsf{E}Y_M\mathbb{1}_{B_n} - \xi\mathsf{P}(B_n) \\
&= \mathsf{E}Y_M - \mathsf{E}Y_M\mathbb{1}_{B_n^c} - \xi\mathsf{P}(B_n) \\
&\ge M - \sup_\omega Y_M(\omega)\cdot\mathsf{P}(B_n^c) - \xi
\end{aligned}$$

Letting $n \to \infty$, we obtain $\lim_{n\to\infty} \mathsf{E}X_n \ge M - \xi$. Since $M, \xi > 0$ are arbitrary, this implies (1.9). □

Note that for any non-negative r.v. X, we can explicitly construct simple r.v.s $X_n \uparrow X$ as follows, so that Lemma 1.24 applies:

$$X_n(\omega) = \frac{[2^n X(\omega)]}{2^n} \wedge n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \mathbb{1}_{\{X(\omega)\in[\frac{k}{2^n}, \frac{k+1}{2^n})\}} + n\mathbb{1}_{\{X(\omega)\geq n\}},$$

where $a \wedge b := \min(a, b)$ and $[x]$ denotes the integer part of x. To see that $X_n \to X$, notice that

$$|X(\omega) - X_n(\omega)| \leq \frac{1}{2^n}, \quad \text{uniformly on } \{\omega : X(\omega) \leq n\}.$$

Theorem 1.25 (Monotone Convergence Theorem, MCT) *If $X_n \geq 0$ and $X_n \uparrow X$, then (1.8) holds.*

Proof: Again, it suffices to establish (1.9).

Let $Y_n^{(m)}$ be simple r.v.s such that $Y_n^{(m)} \uparrow X_n$. Let $Z^{(m)} = \max(Y_1^{(m)}, \ldots, Y_m^{(m)})$. Clearly, $Z^{(m)}$ are simple; they are also increasing in m since

$$Z^{(m)} = \max_{1\leq n\leq m} Y_n^{(m)} \leq \max_{1\leq n\leq m} Y_n^{(m+1)} \leq \max_{1\leq n\leq m+1} Y_n^{(m+1)} = Z^{(m+1)}.$$

Moreover, we have

$$Y_n^{(m)} \leq Z^{(m)} \leq X_m, \quad \forall m \geq n \geq 1.$$

Taking $m \to \infty$, we see that

$$X_n \leq \lim_{m\to\infty} Z^{(m)} \leq X, \quad \forall n \geq 1.$$

Taking $n \to \infty$, and using that $X_n \uparrow X$, we see that $Z^{(m)} \uparrow X$. Then by Lemma 1.24, we have

$$\mathsf{E}X = \lim_{m\to\infty} \mathsf{E}Z^{(m)}. \tag{1.10}$$

On the other hand, since $Y_m^{(m)} \leq Z^{(m)} \leq X_m$, we have

$$\lim_{m\to\infty} \mathsf{E}Z^{(m)} \leq \lim_{m\to\infty} \mathsf{E}X_m. \tag{1.11}$$

Then (1.9) follows from (1.10) and (1.11), and this completes the proof. □

Remark 1.18 In Theorem 1.25, the condition "$X_n \geq 0$" can be replaced by

$$\text{"}X_n \geq -Y, \text{ for some } Y \geq 0 \text{ with } \mathsf{E}Y < \infty\text{"}. \tag{1.12}$$

Indeed, if (1.12) holds, then $\tilde{X_n} = X_n + Y \geq 0$. Since $\tilde{X_n} \uparrow \tilde{X} = X + Y$, we have

$$\lim_{n\to\infty} \big(\mathsf{E}X_n + \mathsf{E}Y\big) = \lim_{n\to\infty} \mathsf{E}\tilde{X_n} = \mathsf{E}\tilde{X} = \mathsf{E}(X + Y).$$

Since $0 \leq \mathsf{E}Y < \infty$, we can subtract $\mathsf{E}Y$ from both sides to obtain $\lim_{n\to\infty} \mathsf{E}X_n = \lim_{n\to\infty} \mathsf{E}X$.

Theorem 1.26 (Fatou's Lemma) *If $X_n \geq 0$ (or (1.12) holds), then*

$$\liminf_{n\to\infty} \mathsf{E}X_n \geq \mathsf{E} \liminf_{n\to\infty} X_n.$$

Proof: Let

$$Y_n = \inf_{m \geq n} X_m \uparrow \liminf_{n\to\infty} X_n.$$

Clearly, $Y_n \leq X_n$. By MCT (Theorem 1.25), we have

$$\mathsf{E} \liminf_{n\to\infty} X_n = \lim_{n\to\infty} \mathsf{E} Y_n \leq \liminf_{n\to\infty} \mathsf{E} X_n.$$

□

Theorem 1.27 (Dominated Convergence Theorem, DCT) *If $X_n \to X$ a.s. and $|X_n| \leq Y$ for some r.v. Y with $\mathsf{E}Y < \infty$, then $\lim_{n\to\infty} \mathsf{E}X_n = \mathsf{E}X$.*

Proof: Clearly, $|X| \leq Y$. Since $2Y - |X_n - X| \geq 0$, by Fatou's Lemma (Theorem 1.26), we have

$$\liminf_{n\to\infty} \mathsf{E}(2Y - |X_n - X|) \geq \mathsf{E}(2Y).$$

Since $\mathsf{E}(2Y) < \infty$, we can subtract it from both side and obtain

$$0 \geq \limsup_{n\to\infty} \mathsf{E}|X_n - X| = 0.$$

□

Corollary 1.28 (Bounded Convergence Theorem, BCT) *If $X_n \to X$ and $|X_n| \leq M$, $n \geq 1$ for some constant M, then $\lim_{n\to\infty} \mathsf{E}X_n = \mathsf{E}X$.*

Proof: Take $Y(\omega) \equiv M$. □

Next, we will present some useful inequalities for expectation. We try to provide proofs which are fairly general, so that they can be generalized easily to other measurable maps.

Proposition 1.29 (Jensen inequality) *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. If $\mathsf{E}|x| < \infty$, then $\mathsf{E}\varphi(x) \geq \varphi(\mathsf{E}X)$.*

Proof: Let $\mathsf{E}X = a \in (-\infty, \infty)$. By convexity, there exists $k \in \mathbb{R}$ (taking $k \in [\varphi'_-(a), \varphi'_+(a)]$) s.t.

$$\varphi(t) \geq \varphi(a) + k(t - a), \quad \forall t.$$

Plugging in $t = X$ and taking expectation, we have

$$\mathsf{E}\varphi(X) \geq \mathsf{E}\varphi(a) + k\mathsf{E}(X - a) = \varphi(a) - ka + k\mathsf{E}X = \varphi(\mathsf{E}X).$$

□

Example 1.19 Let $\varphi(t) = |t|^p$, $p \geq 1$. Then for every $|X|$, we have

$$\mathsf{E}|X|^p \geq (\mathsf{E}|X|)^p.$$

Proposition 1.30 (Hölder's inequality) *If $p, q \in [1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$ then*

$$\mathsf{E}|XY| \leq (\mathsf{E}|X|^p)^{1/p} \cdot (\mathsf{E}|Y|^q)^{1/q}. \tag{1.13}$$

When $p = q = 2$, this is the Cauchy-Schwartz inequality.

Proof: We recall the *Young's inequality*: if $\frac{1}{p}+\frac{1}{q}=1$, then

$$xy \le \frac{x^p}{p}+\frac{y^q}{q}, \quad x,y\ge 0. \tag{1.14}$$

If X and Y are bounded, then we have $\mathsf{E}|X|^p,\ \mathsf{E}|Y|^q<\infty$. Let

$$\tilde{X}=\frac{|X|}{(\mathsf{E}|X|^p)^{1/p}}, \quad \tilde{Y}=\frac{|Y|}{(\mathsf{E}|Y|^q)^{1/q}}.$$

By (1.14), we have

$$\mathsf{E}\tilde{X}\tilde{Y}\le \frac{\mathsf{E}|\tilde{X}|^p}{p}+\frac{\mathsf{E}|\tilde{Y}|^q}{q}=\frac{1}{p}+\frac{1}{q}=1$$

This is (1.13).

If X and Y are not bounded, consider the truncation $X_M=|X|\wedge M$ and $Y_M=|Y|\wedge M$ where $M>0$. For every fixed M we have

$$\mathsf{E}X_MY_M\le \left(\mathsf{E}X_M^p\right)^{1/p}\cdot\left(\mathsf{E}Y_M^q\right)^{1/q}.$$

Taking $M\uparrow\infty$, since $X_M\uparrow X$ and $Y_M\uparrow|Y|$, (1.13) follows from the MCT. □

The final result in this section is about change of variables when we switch measures when performing integration. We will utilize a technique called "functional Monotone Class Theorem", which will be extremely useful in other context as well.

Theorem 1.31 (Change of variables) *Let X be a r.v. and f is a Borel function. Assume either $f\ge 0$ or $\mathsf{E}|f(X)|<\infty$. Then*

$$\mathsf{E}f(X)=\int_\Omega f\big(X(\omega)\big)\,\mathsf{P}(d\omega)=\int_{\mathbb{R}} f(y)\,\mu_X(dy), \tag{1.15}$$

where $\mu_X=\mathsf{P}\circ X^{-1}$ is the distribution of X.

Proof: Let

$$\mathcal{H}=\{f: f \text{ is Borel measurable s.t. (1.15) holds }\}.$$

We want to show that $f\in\mathcal{H}$ whenever $f\ge 0$ or $\mathsf{E}|f(X)|<\infty$. This will be done in several steps.

1. $\mathbb{1}_A\in\mathcal{H}$ for every $A\in\mathcal{B}(\mathbb{R})$.

   Indeed, by definition of the expectation and μ_X, we have

   $$\mathsf{E}\mathbb{1}_A=\int_\Omega \mathbb{1}_A\big(X(\omega)\big)\,\mathsf{P}(d\omega)=\mathsf{P}(X\in A)=\mu_X(A)=\int_{\mathbb{R}}\mathbb{1}_A(y)\,\mu_X(dy)$$

2. Let $f_1,\dots,f_n$ be functions in $\mathcal{H}$. For any $a_1,\dots,a_n\in\mathbb{R}$, we have

   $$a_1f_1+\cdots+a_nf_n\in\mathcal{H},$$

   This follows from the linearity of integrals. Combining with Item 1, $\mathcal{H}$ contains all simple functions.

3. $\mathcal{H}$ contains all non-negative functions.

   Indeed, for every nonnegative function f, there exists a sequence of simple functions f_n such that $f_n \geq 0$ and $f_n \uparrow f$. By Item 2, we have

$$\int_\Omega f_n\big(X(\omega)\big)\,\mathsf{P}(d\omega) = \int_{\mathbb{R}} f_n(y)\,\mu_X(dy)$$

   By MCT, (1.15) follows from

$$\int_\Omega f_n\big(X(\omega)\big)\,\mathsf{P}(d\omega) \to \int_\Omega f\big(X(\omega)\big)\,\mathsf{P}(d\omega), \quad \int f_n(y)\,\mu_X(dy) \to \int f(y)\,\mu_X(dy).$$

4. If $\mathsf{E}|f(x)| < \infty$, then the positive and negative parts $f_+, f_- \in \mathcal{H}$, and hence $f = f_+ - f_- \in \mathcal{H}$.

□

2 Mode of convergence for random variables

2.1 Definitions

There are four basic modes of convergence for r.v.s. We list their definitions below.

1. Almost sure convergence.

   We say that $X_n \to X$ almost surely (a.s.), if

$$\mathsf{P}(\lim_{n\to\infty} X_n = X) = 1$$

2. Convergence in probability.

   We say that $X_n \to X$ in probability (in pr.), if

$$\lim_{n\to\infty} \mathsf{P}\{|X_n - X| > \varepsilon\} = 0, \quad \forall \varepsilon > 0. \tag{2.1}$$

3. Weak convergence or convergence in distribution.

   We say that $X_n \to X$ in distribution, or in law, or weakly, or weakly-*, if for every continuous and bounded function f, have

$$\lim_{n\to\infty} \mathsf{E}f(X_n) = \mathsf{E}f(X).$$

   We also write this as $X_n \Rightarrow X$ or $X_n \Rightarrow_d X$. We will explain the origins of all these different terms in Section 2.4.

4. Convergence in L^p.

   We say that $X_n \to X$ in L^p if

$$\lim_{n\to\infty} \mathsf{E}|X_n - X|^p = 0.$$

In the next few sections, we will explore the relations between these different concepts of convergence.

2.2 Almost sure convergence and convergence in probability

Proposition 2.1 *If $X_n \to X$ a.s., then $X_n \to X$ in pr.*

Proof: If $X_n \to X$ a.s., then for every $\varepsilon > 0$, we have

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

On the other hand, since

$$\{\omega : \limsup_{n\to\infty} |X_n(\omega) - X(\omega)| > \varepsilon\} = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{\omega : |X_m(\omega) - X(\omega)| > \varepsilon\},$$

we have

$$\begin{aligned}
\mathsf{P}\{\limsup_{n\to\infty} |X_n - X| > \varepsilon\} &= \mathsf{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{\omega : |X_m(\omega) - X(\omega)| > \varepsilon\Big) \\
&= \lim_{n\to\infty} \mathsf{P}\Big(\bigcup_{m=n}^{\infty} \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\Big) \\
&\geq \limsup_{n\to\infty} \mathsf{P}\big(|X_n(\omega) - X(\omega)| > \varepsilon\big).
\end{aligned}$$

Hence, $X_n \to X$ in pr. □

Convergence in pr. does NOT imply a.s. convergence. For example, let

$$(\Omega, \mathcal{F}, \mathsf{P}) = \big([0,1], \mathcal{B}([0,1]), \mathrm{Leb}\big), \quad X_{n,k}(\omega) = \mathbb{1}_{\frac{k}{n}, \frac{k+1}{n}}(\omega), 0 \leq k \leq n-1. \tag{2.2}$$

Then $X_{n,k} \to 0$ in pr. but not a.s.

However, the other direction holds on a subsequence.

Proposition 2.2 *If $X_n \to X$ in pr., then there exists a subsequence $\{X_{n_k}\}$ such that $X_{n_k} \to X$ a.s.*

To prove this result we need some preparation. Let $A_1, A_2, \cdots \in \mathcal{F}$ be a sequence of events. We define the event where A_n happens *infinitely often* by

$$\{A_n, \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \limsup_{n\to\infty} A_n. \tag{2.3}$$

Lemma 2.3 (First Borel–Cantelli Lemma) *If $\sum_{n=1}^{\infty} \mathsf{P}(A_n) < \infty$, then $\mathsf{P}\big(\{A_n, i.o.\}\big) = 0$.*

Proof: By (2.3), we have

$$\mathsf{P}\big(\{A_n, \text{ i.o.}\}\big) \leq \mathsf{P}\Big(\bigcup_{m=n}^{\infty} A_m\Big) \leq \sum_{m=n}^{\infty} \mathsf{P}(A_n)$$

. Since $\sum_{n=1}^{\infty} \mathsf{P}(A_n) < \infty$, we have

$$\lim_{n\to\infty} \sum_{m=n}^{\infty} \mathsf{P}(A_n) = 0$$

and the conclusion follows. □

We also have Cauchy's criterion for convergence in pr.

Proposition 2.4 *There exists a r.v. X such that $X_n \to X$ in pr. if and only if for every $\varepsilon > 0$,*

$$\lim_{N\to\infty} \sup_{n,m\geq N} \mathsf{P}\{|X_n - X_m| > \varepsilon\} = 0$$

The "only if" part follows immediately from (2.1); we will use this in the proof of Proposition 2.2. The "if" part in Proposition 2.4 will use Proposition 2.2 and is left as an exercise.

Proof of Proposition 2.2: Since $X_n \to X$ in pr., by Proposition 2.4 with $\varepsilon = 2^{-k}$, there exist $N_k \uparrow \infty$ such that

$$\mathsf{P}\{|X_{N_k} - X_{N_{k+1}} \geq \frac{1}{2^k}\} \leq \frac{1}{2^k}, \quad k \geq 1.$$

Since $\sum_{k=1}^{\infty} 2^{-k} < \infty$, by Borel–Cantelli (Lemma 2.3), we have

$$\mathsf{P}\big(\{|X_{N_k} - X_{N_{k+1}}| > \frac{1}{2^k}, \text{ i.o.}\}\big) = 0,$$

i.e., for almost every ω, there exists $k_0 = k_0(\omega)$ such that

$$|X_{N_k}(\omega) - X_{N_{k+1}}(\omega)| \leq \frac{1}{2^k}, \quad \forall k \geq k_0(\omega).$$

For such ω, the infinite series

$$X_*(\omega) = X_{N_1}(\omega) + \sum_{k=1}^{\infty} \big(X_{N_{k+1}}(\omega) - X_{N_k}(\omega)\big)$$

converges absolutely. Hence, $X_{N_k}(\omega) \to X_*(\omega)$ a.s. as $k \to \infty$.

Finally, we claim that $X_* = X$ almost surely. Since $X_{N_k} \to X_*$ almost surely, we have $X_{N_k} \to X_*$ in pr. The claim then follows from Proposition 2.5 below, which asserts that the limit in pr. is unique up to a set of measure zero. □

Proposition 2.5 *If $X_n \to X$ in pr. and $X_n \to Y$ in pr., then $X = Y$ almost surely.*

Proof: Since $|X - Y| \leq |X_n - X| + |X_n - Y|$, for every $\varepsilon > 0$,

$$\mathsf{P}\big(|X - Y| \geq 2\varepsilon\big) \leq \mathsf{P}(|X_n - X| \geq \varepsilon) + \mathsf{P}(|X_n - Y| \geq \varepsilon).$$

Taking $n \to \infty$, since $X_n \to X, Y$ in pr., the left-hand side must be 0. Therefore,

$$\mathsf{P}(|X - Y| \neq 0) = \lim_{n\to\infty} \mathsf{P}\big(|X - Y| \geq 1/n\big) = 0,$$

and this completes the proof. □

As a corollary of Proposition 2.2, we have the following.

Proposition 2.6 *Almost sure convergence is not expressible via a metric.*

Proof: Assume the contrary that there exists a distance $d(\cdot,\cdot)$ such that $X_n \to X$ a.s. if and only if $d(X_n, X) \to 0$. Let $X_n \to X$ in pr. but not a.s. (such example exists by (2.2)). Then, there exists $\varepsilon_0 > 0$ and a sequence (n') such that

$$d(X_{n'}, X) \geq \varepsilon_0 > 0. \tag{2.4}$$

Clearly, as a subsequence $X_{n'}$ still converges to X in pr. By Proposition 2.2, there is a further subsequence $(n'') \subset (n')$ such that $X_{n''} \to X$ a.s. But this implies that $d(X_{n''}, X) \to 0$, which contradicts with (2.4). □

Note that convergence in pr. is expressible via a metric. For example, $X_n \to 0$ in pr. if and only if $\mathsf{E}\frac{|X_n|}{1+|X_n|} \to 0$. Therefore, a possible metric for convergence in pr. is

$$d(X, Y) = \mathsf{E}\Big[\frac{|X-Y|}{1+|X-Y|}\Big]. \tag{2.5}$$

Of course, one need to verify that (2.5) satisfies the triangle inequality and indeed defines a metric on the space of r.v.s.

We can also relax the condition of a.s. convergence in DCT to convergence in pr.

Proposition 2.7 *If $X_n \to X$ in pr. and $|X_n| \le Y$ for some Y with $\mathsf{E}Y < \infty$, then (1.8) holds.*

Proof: For every subsequence $(X_{n_k}) \subset (X_n)$, by Proposition 2.2, there exists a further subsequence $(X_{n_{k_m}}) \subset (X_{n_k})$ such that $X_{n_{k_m}} \to X$ almost surely, and hence $\mathsf{E}X_{n_{k_m}} \to \mathsf{E}X$ by DCT. This implies $\mathsf{E}X$ is the only possible limit point for the sequence $\big(\mathsf{E}X_n\big)_{n\ge 1}$, and hence (1.8) holds. □

2.3 Convergence in L^p and uniform integrability

Proposition 2.8 *If $X_n \to X$ in L^p, then $X_n \to X$ in pr.*

This proposition follows immediately from the result below.

Lemma 2.9 (Chebyshev's inequality) *For every $\varepsilon > 0$,*

$$\mathsf{P}\big(|X| \ge \varepsilon\big) \le \frac{\mathsf{E}|X|}{\varepsilon}$$

Proof: Since

$$|X| = |X|\mathbb{1}_{\{|X|\ge\varepsilon\}} + |X|\mathbb{1}_{\{|X|\ge\varepsilon\}} \ge |X|\mathbb{1}_{\{|X|\ge\varepsilon\}} \ge \varepsilon\mathbb{1}_{\{|X|\ge\varepsilon\}},$$

taking expectation on both sides, we have $\mathsf{E}|X| \ge \varepsilon\mathsf{P}\{|X| \ge \varepsilon\}$, and the conclusion follows. □

Proof of Proposition 2.8: Let $X_n \to X$ in L^p. For every $\varepsilon > 0$, by Lemma 2.9, we have

$$\mathsf{P}\big(|X_n - X| \ge \varepsilon\big) = \mathsf{P}\big(|X_n - X|^p \ge \varepsilon^p\big) \le \frac{\mathsf{E}|X_n - X|^p}{\varepsilon^p} \to 0.$$

Therefore, $X_n \to X$ in pr. □

Limits in L^p are also unique.

Proposition 2.10 *If $X_n \to X$ in L^p and $X_n \to Y$ in L^p, then $X = Y$ a.s.*

Proof: By Proposition 2.8, $X_n \to X, Y$ in pr., and hence by Proposition 2.5, $X = Y$ a.s. □

Other than Proposition 2.1 and Proposition 2.8, there are not more implications between the three modes of convergence. One counterexample is given in (2.2), counterexamples for the other implications can be obtained by modifying (2.2).

1. $X_n \to X$ in pr. does not implies $X_n \to X$ in L^p. For example, let

$$X_{n,k}(\omega) = n^c\mathbb{1}_{[\frac{k}{n},\frac{k+1}{n}]}(\omega),$$

where $c \ge 1/p$. We have $\mathsf{E}|X_{n,k}|^p \ge 1$ but $X_{n,k} \to 0$ in pr.

2. $X_n \to X$ a.s. does not implies $X_n \to X$ in L^p. For example, let

$$X_n(\omega) = n^c \mathbb{1}_{[0,\frac{1}{n})}(\omega),$$

where $c \geq 1/p$. We have $X_n(\omega) \to 0$ but $\mathsf{E}|X_n|^p \geq 1$.

3. $X_n \to X$ in L^p does not implies $X_n \to X$ a.s. For example, let

$$X_{n,k}(\omega) = n^c \mathbb{1}_{[\frac{k}{n},\frac{k+1}{n}]}(\omega),$$

where $c < 1/p$. We have $\mathsf{E}|X_{n,k}|^p \to 0$ but $X_n \not\to 0$ a.s.

Convergence in L^p and a.s. convergence are equivalent, if assuming some additional integrability condition. Without loss of generality we can restrict our discussion to $p = 1$.

Definition 2.1 (Uniform integrability) *A collection of r.v.s $(X_\alpha)_{\alpha\in I}$ is uniformly integrable (u.i.), if*

$$\lim_{n\to\infty} \sup_{\alpha\in I} \mathsf{E}|X_\alpha| \mathbb{1}_{\{|X_\alpha|\geq M\}} = 0. \tag{2.6}$$

Note that if X_α are u.i., then $\mathsf{E}|X_\alpha|$ are uniformly bounded, since

$$\sup_\alpha \mathsf{E}\ |X_\alpha| \leq M + \sup_{\alpha\in I} \mathsf{E}|X_\alpha| \mathbb{1}_{\{|X_\alpha|\geq M\}} < \infty.$$

Uniform integrability can be seen as a necessary and sufficient condition for (1.8) to hold. Therefore, it will be the last resort if conditions for Theorems 1.25 to 1.27 are not met.

Theorem 2.11 *If $\mathsf{E}|X_n| < \infty$, $\mathsf{E}|X| < \infty$ and $X_n \to X$ in pr., then the following are equivalent:*

1. *$\{X_n\}_{n\geq 1}$ are u.i.;*
2. *$X_n \to X$ in L^1;*
3. *$\mathsf{E}|X_n| \to \mathsf{E}|X|$.*

Proof: **From 1 to 2.** Let

$$\varphi_M(x) = (-M) \vee X \wedge M = \begin{cases} -M, & x \leq -M, \\ x, & x \in [-M, M], \\ M, & x \geq M. \end{cases}$$

(Here, "$\vee$" and "$\wedge$" are associative.) Clearly, we have $|X - \varphi_M(X)| \leq |X| \mathbb{1}_{\{|X|\geq M\}}$, and thus

$$\mathsf{E}|X_n - X| \leq \mathsf{E}|\varphi_M(X_n) - \varphi_M(X)| + \mathsf{E}|\varphi_M(X_n) - X_n| + \mathsf{E}|\varphi_M(X) - X|$$

Taking $n \to \infty$ and then $M \to \infty$, the first term goes to 0 by DCT, the second goes to zero since X_n are u.i., and the third goes to zero since $\mathsf{E}|X| < \infty$ which follows from Fatou's lemma and (2.6):

$$\mathsf{E}|X| \leq \liminf_{n\to\infty} \mathsf{E}|X_n| \leq \sup_n \mathsf{E}|X_n| < \infty.$$

From 2 to 3. It follows from $|\mathsf{E}X_n - X| \leq \mathsf{E}|X_n - X|$.

From 3 to 1. Let

$$\psi_M(x) = \begin{cases} x, & x \in [0, M-1], \\ 0, & x \geq M. \end{cases}$$

Let $\varepsilon > 0$. We have

$$\begin{aligned} \mathsf{E}|X_n| \mathbb{1}_{\{|X_n| \geq M\}} &\leq \mathsf{E}|X_n| - \mathsf{E}\psi_M(|X|) \\ &\leq (\mathsf{E}|X| + \varepsilon) - \big(\mathsf{E}\psi_M(|X|) - \varepsilon\big), \quad n \geq n_0, \end{aligned}$$

where such n_0 exists since $\mathsf{E}|X_n| \to \mathsf{E}X$ by the assumption and $\mathsf{E}\psi_M(|X_n|) \to \mathsf{E}\psi_M(|X|)$ by BCT. Since $\psi_M(t) \to t$ for every t and $\psi_M(|X|) \leq \mathsf{E}|X|$, by DCT there exists $M_0 > 0$ such that

$$\mathsf{E}|X| - \mathsf{E}\psi_M(|X|) \leq \varepsilon, \quad M \geq M_0,$$

Combining these we obtain that for every $\varepsilon > 0$, there exist n_0 and M_0 s.t.

$$\sup_{n \geq n_0} \mathsf{E}|X_n| \mathbb{1}_{\{|X_n| \geq M\}} \leq 3\varepsilon, \quad M \geq M_0.$$

It follows that $(X_n)_{n \geq 1}$ are u.i. □

2.4 Weak convergence

The limit of weak convergence is unique in the sense of distribution of the r.v.s.

Proposition 2.12 *If $\mathsf{E}f(X) = \mathsf{E}f(Y)$ for every bounded continuous function f, then $\mu_X = \mu_Y$ as probability measures on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$.*

Proof: For every open interval (a, b), there exist non-negative bounded continuous function f_n such that $f_n(x) \uparrow \mathbb{1}_{(a,b)}(x)$. Taking $n \to \infty$ in $\mathsf{E}f_n(X) = \mathsf{E}f_n(Y)$, by MCT, we have $\mathsf{E}\mathbb{1}_{(a,b)}(X) = \mathsf{E}\mathbb{1}_{(a,b)}(Y)$. Therefore, $\mu_X(I) = \mu_Y(I)$ for every open interval I. Since open intervals generate $\mathcal{B}(\mathbb{R})$, it follows that $\mu_X = \mu_Y$. □

As Proposition 2.12 suggests, the bounded continuous functions appearing in the definition of the weak convergence merely serve as test functions. In fact, the weak convergence $X_n \Rightarrow_d X$ can also be characterized as using only the information of μ_{X_n} and μ_X, and that is why it is also called convergence in distribution.

We also note that for weak convergence, the probability spaces on which the r.v.s X_n and X live are irrelevant; they can be completely different. This is because we concern only their distribution μ_{X_n} and μ_X, which are probability measures on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$.

Finally, it is not true that $\mu_{X_n}(A) \to \mu_X(A)$ for every $A \in \mathcal{B}(\mathbb{R})$ if $X_n \Rightarrow_d X$, even when A is an open interval. This is the reason why the convergence is *weak*. Using precise terminologies in functional analysis, this is in fact *weak-** convergence, in the sense below.

Let $\mathcal{X}$ be the Banach space of all bounded continuous functions. By Riesz's representation theorem, the *dual space*, $\mathcal{X}^*$, defined as the space of all bounded linear functional from $\mathcal{X}$ to $\mathbb{R}$, is precisely the set of bounded *signed measures* on $\mathcal{B}(\mathbb{R})$, which contains all the probability measures. For a generic Banach space $\mathcal{X}$ and its dual $\mathcal{X}^*$, we say that $u_n \to u$ weakly in $\mathcal{X}$, if

$$\ell(u_n) \to \ell(u), \quad \forall \ell \in \mathcal{X}^*,$$

and we say that $\ell_n \to \ell$ weakly-* in $\mathcal{X}^*$, if

$$\ell_n(u) \to \ell(u), \quad \forall u \in \mathcal{X}.$$

Weak and weak-* convergence are equivalent if the space $\mathcal{X}$ is reflective, i.e., $(\mathcal{X}^*)^* = \mathcal{X}$. While reflectivity holds for the most common L^p spaces ($1 \le p < \infty$), it is not the case for $\mathcal{X}^*$ being the space of bounded continuous functions. So strictly speaking, $X_n \Rightarrow_d X$ means $\mu_{X_n} \to \mu_X$ weakly-*. It is only in probability context that we drop the "*" and call it weak convergence. For weak convergence of probability measures, an excellent reference is [Bil99].

3 Independence and product measures

3.1 Definitions of independence

Recall from elementary probability that two events A and B are *independent* if and only if

$$\mathsf{P}(AB) = \mathsf{P}(A)\mathsf{P}(B).$$

We can use this to defined independence of r.v.s.

Definition 3.1 *Two r.v.s X and Y are independent if*

$$\mathsf{P}\big(X \in A, Y \in B\big) = \mathsf{P}\big(X \in A\big)\mathsf{P}\big(Y \in B\big), \quad \forall A, B \in \mathcal{B}(\mathbb{R}), \tag{3.1}$$

Using the definition of independence of evnets, Definition 3.1 is the most basic definition for independence of r.v.s. But in practice there are other more useful definitions.

Let X be a r.v. The σ-algebra generated by X, denoted by $\sigma(X)$, is the smallest σ-algebra on Ω which makes $X : \Omega \to \mathbb{R}$ measurable. It is easy to check that $\sigma(X)$ has the explicit form

$$\sigma(X) = \{X^{-1}(A), A \in \mathcal{B}(\mathbb{R})\}.$$

We may also introduce independence of σ-algebras.

Definition 3.2 *Two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ are independent, if*

$$\mathsf{P}(AB) = \mathsf{P}(A) \cdot \mathsf{P}(B), \quad \forall A \in \mathcal{F},\ B \in \mathcal{G},$$

Using the independence of σ-algebras, we can reformulate Definition 3.1 as follows.

Proposition 3.1 *Two r.v.s X and Y are independent if and only if $\sigma(X)$ and $\sigma(Y)$ are independent.*

In practice, it also useful to characterize independence via expectation.

Proposition 3.2 *Two r.v.s X and Y are independent if and only if either*

$$\mathsf{E}f(X)g(Y) = \mathsf{E}f(X)\mathsf{E}g(Y), \quad \forall f, g \text{ bounded and Borel}, \tag{3.2}$$

or,

$$\mathsf{E}f(X)g(Y) = \mathsf{E}f(X)\mathsf{E}g(Y), \quad \forall f, g \text{ bounded and continuous}. \tag{3.3}$$

Proof: (3.2) implies (3.1) since we can take $f = \mathbb{1}_A$ and $g = \mathbb{1}_B$ for any Borel sets A and B. To show the other direction, we will use the idea of "functional Monotone Class Theorem".

First, for fixed $A \in \mathcal{B}(\mathbb{R})$, let

$$\mathcal{H}_A = \big\{g : g \text{ bounded and Borel, s.t. } \mathsf{P}\{X \in A\}\mathsf{E}g(Y) = \mathsf{E}\mathbb{1}_A(X)g(Y)\big\}.$$

We claim that $\mathcal{H}_A$ contains all bounded Borel functions. The claim is proved in several steps.

1. $\mathcal{H}_A$ contains all indicator functions $\mathbb{1}_B$, $B \in \mathcal{B}(\mathbb{R})$. This follows directly from (3.1).

2. If $g_1, g_2 \in \mathcal{H}_A$, then $\alpha_1 g_1 + \alpha_2 g_2 \in \mathcal{H}_A$. That is, $\mathcal{H}_A$ is closed under linear combination. This implies that $\mathcal{H}_A$ contains all simple functions.

3. If $g_n \geq 0, g_n \in \mathcal{H}_A$ and $g_n \uparrow g$, then $g_n(Y) \uparrow g(Y)$ and $\mathbb{1}_A(X) g_n(Y) \uparrow \mathbb{1}_A(X) g(Y)$. By MCT, we have

$$\mathsf{P}(X \in A)\mathsf{E}g(Y) = \lim_{n\to\infty} \mathsf{P}(X \in A)\mathsf{E}g_n(Y) = \lim_{n\to\infty} \mathsf{E}\mathbb{1}_A(X) g_n(Y) = \mathsf{E}\mathbb{1}_A(X) g(Y).$$

   Therefore, $\mathcal{H}_A$ contains all non-negative Borel functions, and hence all bounded Borel functions by linearity.

Second, let

$$\mathcal{H} = \big\{ f : \text{ bounded and Borel s.t. } \mathsf{E}f(X) \cdot \mathsf{E}g(Y) = \mathsf{E}f(X)g(Y) \big\}.$$

Then $\mathbb{1}_A \in \mathcal{H}$ for every $A \in \mathcal{B}(\mathbb{R})$. Repeating the above argument again, we can show that $\mathcal{H}$ contains all bounded Borel functions. This establishes equivalence between (3.2) and (3.1).

Next, we show that (3.3) and (3.1) are equivalent. Clearly, (3.2) implies (3.3) since continuous functions are Borel. On the other hand, assuming (3.3), for any open intervals A and B, by choosing bounded, non-negative continuous functions f_n and g_n such that $f_n \uparrow \mathbb{1}_A$ and $g_n \uparrow \mathbb{1}_B$, MCT implies that (3.1) holds for such A and B. From open intervals to arbitrary Borel sets we need to use the monotone class theorem. Details are omitted here. □

We can also introduce the notion of a collection of r.v.s being independent.

Definition 3.3 *Let I be a countable index set. A collection of r.v.s $(X_n)_{n\in I}$ are independent, if the σ-algebras $\big(\sigma(X_n)\big)_{n\in I}$ are independent, i.e.,*

$$\mathsf{P}\Big(\bigcap_{n\in I} A_n\Big) = \prod_{n\in I} \mathsf{P}(A_n), \quad \forall A_n \in \sigma(X_n).$$

Definition 3.3 is NOT implied by "pairwise independence" of the r.v.s $(X_n)_{n\in I}$. A simplest counterexample can be given for $I = \{1, 2, 3\}$ as follows. Let $(\Omega, \mathcal{F}, \mathsf{P}) = ([0,1], \mathcal{B}([0,1]), \text{Leb})$ and

$$X_1(\omega) = \begin{cases} 1, & \omega \in [\frac{1}{2}, 1], \\ -1, & \omega \in [0, \frac{1}{2}), \end{cases} \quad X_2(\omega) = \begin{cases} 1, & \omega \in [\frac{1}{4}, \frac{1}{2}) \cup [\frac{3}{4}, 1], \\ -1, & \omega \in [0, \frac{1}{4}) \cup [\frac{1}{2}, \frac{3}{4}), \end{cases} \quad X_3(\omega) = X_1(\omega) \cdot X_2(\omega).$$

Clearly, X_1, X_2 are r.v.s since they are simple functions, and X_3 is a r.v. since it is a product of two r.v.s. It is also easy to check that X_1, X_2, X_3 are pairwise independent, but they are not independent, since

$$\mathsf{P}(X_1 = X_2 = X_3 = -1) = 0 \neq \frac{1}{8} = \mathsf{P}(X_1 = -1)\mathsf{P}(X_2 = -1)\mathsf{P}(X_3 = -1)$$

In probability theory, a fundamental model is a sequence of *independent and identically distributed (i.i.d.)* r.v.s $(X_n)_{n\geq 1}$, which, in addition to X_n being independent, requires that the distribution of X_n is the same. A natural question that we must answer first before delving into nice theories built upon i.i.d. r.v.s like the law of large numbers, central limit theorem, etc, is the existence of a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ on which there live independent r.v.s X_n with given common distribution μ.

The answer is affirmative, and its solution will be discussed in the rest of this section. It will be done in three steps.

1. The one-dimensional case: given a c.d.f. $F(x)$, how to construct a r.v. X such that $\mathsf{P}(X \le a) = F(a)$? This is done in Section 3.2.1.

2. The two/finite-dimensional case: given two probability measures μ_1 and μ_2 on $(\mathbb{R}, \mathcal{B}(R))$, how can we construct two r.v.s X, Y such that $\mathcal{L}(X) = \mu_1$, $\mathcal{L}(Y) = \mu_2$ and X and Y are independent? This is done in Section 3.2.2 with the help of product measures.

3. The infinite-dimensional case: given probability measures $(\mu_n)_{n\in I}$ on $\big(\mathbb{R}, \mathcal{B}(R)\big)$, how can we construct r.v.s $(X_n)_{n\in I}$ such that $\mathcal{L}(X_n) = \mu_n$, $n \in I$, and X_n are independent. In particular, for a sequence of i.i.d. r.v.s, we need $I = \mathbb{N}$. This is done in Section 3.3 with the help of the celebrated Kolmogorov's Extension Theorem Theorem 3.9.

On the other hand, when the common distribution μ is as simple as the Bernoulli distribution, we have very explicit construction of the probability space and r.v.s.

Example 3.1 Let $(\Omega, \mathcal{F}, \mathsf{P}) = \big((0,1), \mathcal{B}(0,1), \mathrm{Leb}\big)$. Every $\omega \in \Omega = (0,1)$ admits a dyadic expansion:

$$\omega = \sum_{n=1}^{\infty} \xi_n(\omega) \frac{1}{2^n}, \quad \xi_n(\omega) \in \{0,1\}. \tag{3.4}$$

When $\omega = \frac{k}{2^n}$ is a dyadic rational, the expansion (3.4) is non-unique; in that case, we will choose the expansion with infinitely many 1's to fix the choice. For example, we choose

$$\frac{1}{2} = 0 \cdot \frac{1}{2^1} + 1 \cdot \frac{1}{2^2} + 1 \cdot \frac{1}{2^3} + 1 \cdot \frac{1}{2^4} + \cdots = \sum_{n=2}^{\infty} \frac{1}{2^n}, \quad \text{rather than} \quad \frac{1}{2} = \frac{1}{2} + \sum_{n=2}^{\infty} 0 \cdot \frac{1}{2^n}.$$

One can verify directly that $(\xi_n)_{n\ge 1}$ are i.i.d. Bernoulli r.v.s with parameter $1/2$.

3.2 Product measures

3.2.1 Existence of random variables

Let F be an increasing, right continuous function with $F(-\infty) = 0$ and $F(\infty) = 1$. Theorem 1.6 and the usage of Carathéodory's Extension Theorem there gives the construction of a probability measure μ on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$, where $\mu(-\infty, a] = F(a)$. To construct a r.v. X with distribution μ, we can simply take $(\Omega, \mathcal{F}, \mathsf{P}) = \big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu\big)$ and $X(\omega) = \omega$.

Another way to construct a r.v. with given a c.d.f. $F(x)$ is to use the *generalized inverse* F^{-1}:

$$F^{-1}(x) = \sup\{y : F(y) < x\}.$$

One can check that F^{-1} is increasing and left continuous. Moreover, if F is strictly increasing and continuous, then F^{-1} is the normal inverse function of F.

Proposition 3.3 *Let $U \sim \mathrm{Unif}[0,1]$ be defined on $(\Omega, \mathcal{F}, \mathsf{P})$. Then $F^{-1}(U)$ is a r.v. on $(\Omega, \mathcal{F}, \mathsf{P})$ with c.d.f. F.*

Proof: Since F^{-1} is left continuous and increasing, it is Borel measurable. Hence, $\omega \mapsto F^{-1}\big(U(\omega)\big)$ is measurable and $F^{-1}(U)$ is a r.v. on $(\Omega, \mathcal{F}, \mathsf{P})$.

To check that the c.d.f. of $F^{-1}(U)$ is F, we will use without proof that

$$\{y : F^{-1}(y) \le x\} = \{y : y \le F(x)\}. \tag{3.5}$$

Indeed, assuming (3.5), we have

$$\mathsf{P}\big(F^{-1}(U) \le a\big) = \mathsf{P}\big(U \le F(a)\big) = F(a).$$

as desired. □

Proposition 3.3 plays an important role in computer science when it comes to stochastic simulation. On computers, one can use pseudo random number generators to produce i.i.d. uniform integers X in the set $\{1, 2, \ldots, N\}$ where N is very large. Then, X/N will approximate the uniform distribution on $[0, 1]$, and thus $F^{-1}(X/N)$ is closed to a r.v. with c.d.f. F. Of course, it is often the case where F^{-1} is costly to compute, and some other sampling methods will be efficient. But this algorithm is useful enough to generate common distributions like the exponential and Gaussian.

3.2.2 Product Measures and Fubini's Theorem

Let $(\Omega_i, \mathcal{F}_i, \mathsf{P}_i)$, $i = 1, 2$, be two probability spaces. Let

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1,\ \omega_2 \in \Omega_2\},$$
$$\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 = \sigma(A \times B : A \in \mathcal{F}_1,\ B \in \mathcal{F}_2).$$

Then $(\Omega, \mathcal{F})$ is a measurable space. A special case is $(\Omega_i, \mathcal{F}_i) = \big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ where $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2)$, where the equality is due to the fact that open rectangles

$$(a, b) \times (c, d), \quad -\infty < a < b < \infty,\ -\infty < c < d < \infty,$$

form a topological basis for open sets in $\mathbb{R}^2$.

Our goal is to construct the product measure $\mathsf{P}_1 \times \mathsf{P}_2$ on $(\Omega, \mathcal{F})$. We will need to introduce an appropriate algebra generating $\mathcal{F}$ and use Carathéodory's Extension Theorem (Theorem 1.12). Consider the collection of "rectangles"

$$\mathcal{S} = \{A \times B : A \in \mathcal{F}_1,\ B \in \mathcal{F}_2\}.$$

It is not hard to check that $\mathcal{S}$ forms a semi-algebra:

1. $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$,
2. $(A \times B)^c = (A^c \times B) \cup (A \times B^c) \cup (A^c \times B^c)$.

The semi-algebra $\mathcal{S}$ naturally generates an algebra

$$\bar{\mathcal{S}} = \Big\{ \bigcup_{i=1}^{k} I_i,\ I_i \in \mathcal{S},\ I_i \text{ disjoint} \Big\}.$$

We note that unless one of $\mathcal{F}_i$ is trivial, $\mathcal{S} \subsetneq \sigma(\mathcal{S}) = \mathcal{F}$ (actually, $\mathcal{S} \subsetneq \bar{\mathcal{S}}$ for nontrivial $\mathcal{F}_i$).

Remark 3.2 Using standard notion of Cartesian products, one may write "$\mathcal{S} = \mathcal{F}_1 \times \mathcal{F}_2$", but it may cause confusion since some authors also use "$\mathcal{F}_1 \times \mathcal{F}_2$" for the product σ-algebra. Hence, in this note we will use the tensor product notation "$\otimes$" to emphasize that the product σ-algebra is more than the usual Cartesian product of σ-algebras.

The unique measure μ defined in the next theorem is the desired product measure $\mathsf{P}_1 \times \mathsf{P}_2$.

Theorem 3.4 *There exists a unique probability measure μ on $(\Omega, \mathcal{F})$ such that*

$$\mu(A \times B) = \mathsf{P}_1(A) \cdot \mathsf{P}_2(B).$$

Proof: We can define a finitely additive probability measure μ_0 on $\bar{\mathcal{S}}$ by

$$\mu_0(D) = \sum_{i=1}^{k} \mathsf{P}_1(A_i) \cdot \mathsf{P}_2(B_i), \quad D = \text{disjoint union of } A_1 \times B_1, \ldots, A_k \times B_k.$$

The conclusion follows from Theorem 1.12, if we can show that show that μ_0 is a σ-additive on $\bar{\mathcal{S}}$. For this, it suffices to check that if $A_n \times B_n$, $n = 1, \ldots,$ are disjoint and $A \times B = \bigcup_{n=1}^{\infty}(A_n \times B_n)$, then

$$\mu_0(A \times B) = \sum_{n=1}^{\infty} \mu_0(A_n \times B_n). \tag{3.6}$$

(This is σ-additivity on $\mathcal{S}$, not on $\bar{\mathcal{S}}$, but here they are equivalent.)

For $x \in A$, let $I(x) = \{n : x \in A_n\}$. Then

$$B = \bigcup_{n \in I(x)} B_n, \quad \forall x \in A, \tag{3.7}$$

since $\{x\} \times B \subset \bigcup_{n \in I(x)}(A_n \times B_n)$. For $x \in A$, we have

$$\mathbb{1}_A(x) \cdot \mathsf{P}_2(B) = \mathbb{1}_A(x) \cdot \sum_{n \in I(x)} \mathsf{P}_2(B_n) = \sum_{n \in I(x)} \mathbb{1}_{A_n}(x)\ \mathsf{P}_2(B_n) = \sum_{n \geq 1} \mathbb{1}_{A_n}(x)\ \mathsf{P}_2(B_n). \tag{3.8}$$

The first equality holds since we have (3.7) and B_n are disjoint, the second holds since $\mathbb{1}_A(x) = \mathbb{1}_{A_n}(x) = 1$ for $n \in I(x)$, and the third holds since we are adding more zero terms.

Note that (3.8) also holds for $x \notin A$, since

$$\mathbb{1}_A(x) \cdot \mathsf{P}_2(B) = 0 = \sum_{n \geq 1} \mathbb{1}_{A_n}(x)\ \mathsf{P}_2(B_n), \quad x \notin A.$$

Integrating (3.8) over $x \in \Omega$, the left hand side becomes

$$\Big[\int_\Omega \mathbb{1}_A(x)\, P_1(dx)\Big] \cdot \mathsf{P}_2(B) = \mathsf{P}_1(A) \cdot \mathsf{P}_2(B) = \mu_0(A \times B),$$

and the right hand side becomes

$$\begin{aligned}
\int_\Omega \Big[\sum_{n \geq 1} \mathbb{1}_{A_n}(x) \mathsf{P}_2(B_n)\Big]\, \mathsf{P}_1(dx) &= \int_\Omega \Big[\lim_{N \to \infty} \sum_{n=1}^{N} \mathbb{1}_{A_n}(x) \mathsf{P}_2(B_n)\Big]\, P_1(dx) \\
&= \lim_{N \to \infty} \int_\Omega \Big[\sum_{n=1}^{N} \mathbb{1}_{A_n}(x) \mathsf{P}_2(B_n)\Big]\, \mathsf{P}_1(dx) \\
&= \sum_{n=1}^{\infty} \mathsf{P}_1(A_n) \mathsf{P}_2(B_n) = \sum_{n=1}^{\infty} \mu_0(A_n \times B_n),
\end{aligned}$$

where we use MCT in the second line. This proves (3.6) and concludes the proof. □

We can construct two independent r.v.s with given distribution using Theorem 3.4. Let X be a r.v. on $(\Omega_1, \mathcal{F}_1, \mathsf{P}_1)$ and Y a r.v. on $(\Omega_2, \mathcal{F}_2, \mathsf{P}_2)$. On $(\Omega, \mathcal{F}, \mu) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathsf{P}_1 \times \mathsf{P}_2)$, we define

$$\tilde{X}(\omega_1, \omega_2) = X(\omega_1), \quad \tilde{Y}(\omega_1, \omega_2) = Y(\omega_2).$$

Then

$$\begin{aligned}\mathsf{P}(\tilde{X}\in A,\tilde{Y}\in B) &= \mu(X^{-1}(A)\times Y^{-1}(B)) = \mathsf{P}_1(X^{-1}(A))\cdot\mathsf{P}_2(Y^{-1}(B))\\ &= \mathsf{P}_1(X\in A)\cdot\mathsf{P}_2(Y\in B) = \mathsf{P}(\tilde{X}\in A)\cdot\mathsf{P}(\tilde{Y}\in B),\end{aligned}$$

that is, $\tilde{X}$ (respectively, $\tilde{Y}$) has the same distribution as X (resp. Y), and $\tilde{X}$, $\tilde{Y}$ are independent.

Integration on the product measure space can be computed using Fubini's Theorem below. Fubini's Theorem also includes some measurability statements on jointly measurable maps.

Theorem 3.5 (Fubini's Theorem) *Let $(\Omega_i,\mathcal{F}_i,\mathsf{P}_i)$, $i=1,2$, be two measure spaces, where P_i are probability (or σ-finite) measures. Let $f:\Omega\to\mathbb{R}$ be $\mathcal{F}_1\otimes\mathcal{F}_2$-measurable where $\Omega=\Omega_1\times\Omega_2$. Assume either*

$$f\geq 0, \tag{3.9a}$$

$$\text{or}\quad \int |f(\omega_1,\omega_2)|\,(\mathsf{P}_1\times\mathsf{P}_2)(d\omega_1 d\omega_2)<\infty. \tag{3.9b}$$

Then the following holds.

1. *For every $\omega_1\in\Omega$, the function $f(\omega_1,\cdot)$ is $\mathcal{F}_2$-measurable. And if (3.9b) holds,*

$$\int_{\Omega_2}|f(\omega_1,\omega_2)|\,\mathsf{P}_2(d\omega_2)<\infty,\qquad \text{for almost every } \omega_1\in\Omega. \tag{3.10}$$

2. *The function $g(\omega_1)=\int_{\Omega_2} f(\omega_1,\omega_2)\,\mathsf{P}_2(d\omega_2)$ is $\mathcal{F}_1$-measurable. And if (3.9b) holds,*

$$\int_{\Omega_1}|g(\omega_1)|\mathsf{P}_1(d\omega_1)<\infty. \tag{3.11}$$

3. *The double integral is equal to either iterated integral, that is,*

$$\begin{aligned}\iint_{\Omega_1\times\Omega_2} f(\omega_1,\omega_2)\,(\mathsf{P}_1\times\mathsf{P}_2)(d\omega_1 d\omega_2) &= \int_{\Omega_1}\mathsf{P}_1(d\omega_1)\int_{\Omega_2} f(\omega_1,\omega_2)\,\mathsf{P}_2(d\omega_2)\\ &= \int_{\Omega_2}\mathsf{P}_2(d\omega_2)\int_{\Omega_1} f(\omega_1,\omega_2)\,\mathsf{P}_1(d\omega_1).\end{aligned} \tag{3.12}$$

Proof: Let $\mathcal{H}$ be the collection of all $\mathcal{F}_1\otimes\mathcal{F}_2$-measurable functions f such that Items 1 to 3 hold. As usual, we will show that $\mathcal{H}$ contains more and more general forms of functions, and eventually, all $\mathcal{F}_1\otimes\mathcal{F}_2$-measurable functions f such that either (3.9a) or (3.9b) holds.

1. Indicator functions of rectangles are in $\mathcal{H}$.
   Let $f(\omega_1,\omega_2)=\mathbb{1}_A(\omega_1)\mathbb{1}_B(\omega_2)$ where $A\in\mathcal{F}_1$ and $B\in\mathcal{F}_2$. We have

$$f(\omega_1,\cdot)=\begin{cases}0, & \omega_1\notin A,\\ \mathbb{1}_B(\cdot) & \omega_1\in A,\end{cases}$$

   so $f(\omega_1,\cdot)$ is $\mathcal{F}_2$-measurable for every ω_1. Moreover, direct computation gives

$$g(\omega_1)=\begin{cases}0, & \omega_2\notin A\\ \mathsf{P}_2(B), & \omega_2\in A\end{cases}=\mathbb{1}_A(\omega_1)\cdot\mathsf{P}_2(B),$$

   and hence g is $\mathcal{F}_1$-measurable. It is easy to verify (3.10) to (3.12).

2. The indicator function $\mathbb{1}_D(\omega_1, \omega_2) \in \mathcal{F}$ for every $D \in \mathcal{F}$.
   We will use the method of appropriate sets. Let

$$\mathcal{G} = \{D \in \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{1}_D \in \mathcal{H}\}.$$

   We note that $\mathcal{G}$ contains the algebra $\bar{\mathcal{S}}$ as a consequence of the first part, and that $\mathcal{G}$ is a monotone class, since the measurability conditions are preserved by taking limits, and the integral conditions are preserved by the MCT. Hence, by the monotone class theorem $\mathcal{G} = \mathcal{F}_1 \otimes \mathcal{F}_2$.

3. Simple functions of the form $\varphi(\omega) = \sum_{i=1}^n c_i \mathbb{1}_{D_i}(\omega)$ are in $\mathcal{H}$, since Items 1 to 3 are preserved by taking finite linear combination.

4. All nonnegative, $\mathcal{F}_1 \otimes \mathcal{F}_2$-measurable functions f are in $\mathcal{H}$.
   Recall that there exist simple functions $\{f_n\}$ such that $f_n(\omega) \uparrow f(\omega)$ for every ω. We have already shown that $f_n \in \mathcal{H}$.

   Since for every ω_1, the function $f_n(\omega_1, \cdot)$ is $\mathcal{F}_2$-measurable, the limit $f(\omega_1, \cdot) = \lim_{n\to\infty} f_n(\omega_1, \cdot)$ is also $\mathcal{F}_2$-measurable. By MCT,

$$g(\omega_1) = \int_{\Omega_2} f(\omega_1, \omega_2)\, \mathsf{P}_2(d\omega_2) = \lim_{n\to\infty} \int_{\Omega_2} f_n(\omega_1, \omega_2)\, \mathsf{P}_2(d\omega_2) = \lim_{n\to\infty} g_n(\omega_1).$$

   Since $g_n(\omega_1)$ are $\mathcal{F}_1$-measurable, their increasing limit $g(\omega_1)$ is also $\mathcal{F}_1$-measurable. Finally, by MCT applied to both (g_n) and (f_n),

$$\begin{aligned}\int_{\Omega_1} g(\omega_1)\, \mathsf{P}_1(d\omega_1) &= \lim_{n\to\infty} \int g_n(\omega_1)\, \mathsf{P}_1(d\omega_1) = \lim_{n\to\infty} \int_{\Omega} f_n(\omega_1, \omega_2)\, (\mathsf{P}_1 \times \mathsf{P}_2)(d\omega_1 d\omega_2) \\ &= \int_{\Omega} f(\omega_1, \omega_2)\, (\mathsf{P}_1 \times \mathsf{P}_2)(d\omega_1 d\omega_2),\end{aligned}$$

   and then by symmetry in ω_1 and ω_2,

$$\int_{\Omega} f(\omega_1, \omega_2)\, (\mathsf{P}_1 \times \mathsf{P}_2)(d\omega_1 d\omega_2) = \int_{\Omega_2} \mathsf{P}_2(d\omega_2) \int_{\Omega_1} f(\omega_1, \omega_2)\, \mathsf{P}_1(d\omega_1).$$

   This verifies (3.12) and thus $f \in \mathcal{H}$.

5. For general function f, we consider $f = f_+ - f_-$. To show that $f \in \mathcal{H}$, everything is straightforward except (3.10).

   Applying Fubini's Theorem to $|f| \geq 0$, we have

$$\int_{\Omega_1} \mathsf{P}(d\omega_1) \Big[\int_{\Omega_1} |f(\omega_1, \omega_2)\, \mathsf{P}_2(d\omega_2) \Big] = \int_{\Omega} f(\omega_1, \omega_2)\, (\mathsf{P}_1 \times \mathsf{P}_2)(d\omega_1 d\omega_2) < \infty.$$

   This implies (3.10).

□

Let $D \subset \Omega$. The cross section of D at x is defined by

$$D_x = \{y : (x, y) \in D\}.$$

As a corollary of Theorem 3.5, we obtain measurability of the cross section.

Proposition 3.6 *Let $D \in \mathcal{F}_1 \otimes \mathcal{F}_2$. Then $D_x \in \mathcal{F}_2$ for every $x \in \Omega_1$.*

Proof: Note that $y \in D_x$ if and only if $\mathbb{1}_D(x,y) > 0$. For every $x \in \Omega_1$, by Theorem 3.5, the function $\mathbb{1}_D(x,\cdot)$ is $\mathcal{F}_2$-measurable, and thus

$$D_x = \{y : \mathbb{1}_D(x,y) > 0\} \in \mathcal{F}_2.$$

□

We recall that the completion of a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ is a complete probability space $(\Omega, \bar{\mathcal{F}}, \bar{\mathsf{P}})$ such that

$$\bar{\mathcal{F}} = \{A : \exists B_1 \subset A \subset B_2,\ B_1, B_2 \in \mathcal{F} \text{ s.t. } \mathsf{P}(B_1) = \mathsf{P}(B_2),\ \mathsf{P}(B_1 \backslash B_2) = 0\},$$

and for $A \in \bar{\mathcal{F}}$, we define $\bar{\mathsf{P}}(A) = \mathsf{P}(B_1)$ where B_1 is given above. Note that $\overline{\mathcal{B}(\mathbb{R})} = \{\text{Lebesgue sets}\}$.

Proposition 3.7

$$\overline{\mathcal{B}(\mathbb{R})} \otimes \overline{\mathcal{B}(\mathbb{R})} \neq \overline{\mathcal{B}(\mathbb{R}^2)},$$

and in general,

$$\overline{\mathcal{F}_1} \otimes \overline{\mathcal{F}_2} \neq \overline{\mathcal{F}_1 \otimes \mathcal{F}_2}.$$

Proof: Let $A \subset [0,1]$ be a non-Lebesgue set and $D = A \times \{0\}$. We have $D \subset [0,1] \times \{0\}$ and

$$\text{Leb}([0,1] \times \{0\}) = \lim_{n\to\infty} \text{Leb}([0,1] \times [0,1/n]) = \lim_{n\to\infty} \frac{1}{n} = 0.$$

Hence $D \in \overline{\mathcal{B}(\mathbb{R}^2)}$ by the definition of completion. But $D \notin \overline{\mathcal{B}(\mathbb{R})} \otimes \overline{\mathcal{B}(\mathbb{R})}$, otherwise by Proposition 3.6,

$$A = \{x \in \mathbb{R} : (x,0) \in D\} = D_0 \in \overline{\mathcal{B}(\mathbb{R})},$$

which is absurd. □

Remark 3.3 In general, completion of probability spaces has to done in the final step, after the construction of product spaces.

There is a version of Fubini's Theorem stated for the completion of the σ-algebra $\overline{\mathcal{F}_1 \otimes \mathcal{F}_2}$. Although it is very technical, it also has essential applications in various scenarios. This is also the Fubini's Theorem that one learns from a real analysis course, in which Lebesgue sets rather than Borel sets are the primary interest. We include it here and sketch the additional technicalities in the proof, from which the reader can also learn how to deal with zero measure sets.

Theorem 3.8 (Fubini's Theorem for complete measure spaces) *Let $f : \Omega \to \mathbb{R}$ be $\overline{\mathcal{F}_1 \otimes \mathcal{F}_2}$-measurable. Assume either (3.9a) or (3.9b). Then*

1. *There exists a set $N \in \mathcal{F}_1$ with $\mathsf{P}(N) = 0$, such that for every $\omega_1 \in N^c$, the function $f(\omega_1, \cdot)$ is $\overline{\mathcal{F}_2}$-measurable. When (3.9b) holds, the set N can be chosen such that for $\omega_1 \in N^c$,*

$$\int |f(\omega_1, \omega_2)|\, \mathsf{P}_2(d\omega_2) < \infty.$$

2. *Let*

$$g(\omega_1) = \begin{cases} \int_{\Omega_2} f(\omega_1, \omega_2)\, \mathsf{P}_2(d\omega_2), & f(\omega_1, \cdot) \text{ is } \bar{\mathcal{F}}_2\text{-measurable}, \\ \text{undefinded}, & \text{otherwise}. \end{cases}$$

Then $g(\omega_1)$ is $\overline{\mathcal{F}_1}$-measurable. If (3.9b) holds, then (3.11) is true.

3. (3.12) *holds.*

Proof: Let $\mathcal{H}$ be the collection of $\overline{\mathcal{F}_1 \otimes \mathcal{F}_2}$-measurable functions such that the Fubini's Theorem holds.

There are two keys steps. First, we need to show that $\mathbb{1}_D \in \mathcal{H}$ for any $D \in \overline{\mathcal{F}_1 \otimes \mathcal{F}_2}$. Second, we need to show that $\mathcal{H}$ is closed under taking limit, i.e., if $f_n \in \mathcal{H}, f_n \geq 0, f_n \uparrow f$, then $f \in \mathcal{H}$.

To prove the first step, let $D \in \overline{\mathcal{F}_1 \otimes \mathcal{F}_2}$. By the definition of completion, there exists $D^{\pm} \in \mathcal{F}_1 \otimes \mathcal{F}_2$ such that

$$D^- \subset D \subset D^+, \quad (\mathsf{P}_1 \times \mathsf{P}_2)(D^+ \backslash D^-) = 0.$$

By definition of the cross section, for every $\omega_1 \in \Omega_1$, we have $D^-_{\omega_1} \subset D_{\omega_1} \subset D^+_{\omega_1}$. Moreover, by Proposition 3.6 and Theorem 3.5j, for every $\omega_1 \in \Omega_1$, we have $D^{\pm}_{\omega_1} \in \mathcal{F}_2$ and that $q(\omega_1) = \mathsf{P}_2(D^+_{\omega_1}) - \mathsf{P}_2(D^-_{\omega_1})$ is $\mathcal{F}_1$-measurable, and

$$\begin{aligned} \int q(\omega_1)\, \mathsf{P}_1(d\omega_1) &= \int [\mathsf{P}_2(D^+_{\omega_1}) - \mathsf{P}_2(D^-_{\omega_1})]\, \mathsf{P}_1(d\omega_1) \\ &= \int \left(\mathbb{1}_{D^+}(\omega) - \mathbb{1}_{D^-}(\omega)\right) (\mathsf{P}_1 \times \mathsf{P}_2)(d\omega_1 d\omega_2) = (\mathsf{P}_1 \times \mathsf{P}_2)(D^+ \setminus D^-) = 0. \end{aligned} \tag{3.13}$$

Since $q(\omega_1) \geq 0$, (3.13) implies that there exists $N \in \mathcal{F}_1$ with $\mathsf{P}_1(N) = 0$ such that

$$q(\omega_1) = \mathsf{P}_2(D^+_{\omega_1}) - \mathsf{P}_2(D^-_{\omega_1}) = 0, \quad \forall \omega_1 \notin N.$$

Hence, for $\omega_1 \notin N$, the set D_{ω_1} is $\overline{\mathcal{F}_2}$-measurable since

$$\mathsf{P}_2(D^+_{\omega_1}) = \mathsf{P}_2(D^-_{\omega_2}), \quad D^-_{\omega_1} \subset D_{\omega_1} \subset D^+_{\omega_1}.$$

Note that $g(\omega_1)$ is defined on N^c, so

$$\{\omega_1 : g(\omega_1) \text{ not definded}\} \subset N,$$

and it is an element of $\overline{\mathcal{F}_1}$ by definition. It is easy to verify (3.12).

For the second step, let $\mathcal{H} \in f_n \uparrow f$ and let $N_n \in \mathcal{F}_1$ be the corresponding zero measure sets corresponding to f_n. Let $N = \bigcup_{n=1}^{\infty} N_n$. Then $N \in \mathcal{F}_1$ and $\mathsf{P}_1(N) = 0$. If $\omega_1 \notin N$, then $\omega_1 \notin N_n$ for every n, and hence $f_n(\omega_1, \cdot)$ is $\overline{\mathcal{F}_2}$-measurable, the $f(\omega_1, \cdot)$ as the limit of $f_n(\omega_1, \cdot)$ is $\overline{\mathcal{F}_2}$-measurable, for $\omega_1 \notin N$. The rest of the conditions can be checked easily. □

3.3 Measures on $\mathbb{R}^\infty$ and Kolmogorov's Extension Theorem

The notion of product measures can be generalized to finitely many probability spaces. Hence, we can construct finitely many independent r.v.s with given distribution. More precisely, suppose that $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_i)$, $1 \leq i \leq n$, are given. Let

$$(\Omega, \mathcal{F}, \mathsf{P}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \bigtimes_{i=1}^{n} \mu_i).$$

An element of Ω is written as $\omega = (\omega_1, \ldots, \omega_n)$. Let $X_i(\omega) = \omega_i$, $1 \leq i \leq n$. Then $\{X_i\}_{1 \leq i \leq n}$ are independent and $\mathcal{L}(X_i) = \mu_i$.

In this section, we illustrate how to construct an infinite sequence of independent r.v.s. It is important to understand the structure of the measure space $\left(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty)\right)$.

The space $\mathbb{R}^\infty$ forms a metric space with the metric

$$d(x,y)=\sum_{n=1}^{\infty}2^{-n}(1\wedge|x_n-y_n|)\le 1,\quad x=(x_1,x_2,\dots)\in\mathbb{R}^\infty.$$

We say that $O\subset\mathbb{R}^\infty$ is an open set, if for every $x\in O$, there exists $\delta>0$ such that

$$\{y: d(x,y)<\delta\}\subset O.$$

It is also useful to introduce the projection: $\pi_n:\mathbb{R}^m\to\mathbb{R}^n$, $n\le m\le\infty$, where $\pi_n x$ is the first n coordinates of x. The convergence in $\mathbb{R}^\infty$ can be characterized by convergence in finite dimensional spaces:

$$d(x^{(m)},x^{(0)})\to 0,\ m\to\infty\quad\Leftrightarrow\quad \pi_n x^{(m)}\to\pi_n x^{(0)},\ \forall n\ge 1. \tag{3.14}$$

With the definition of open sets, we can define the Borel σ-algebra $\mathcal{B}(\mathbb{R}^\infty)$. It is not hard to check that, instead of open balls, $\mathcal{B}(\mathbb{R}^\infty)$ can also be generated by

$$\mathcal{B}(\mathbb{R}^\infty)=\sigma\Big(O_n\times\mathbb{R}^\infty, O_n \text{ open set in } \mathbb{R}^n\Big). \tag{3.15}$$

In general, set of the form

$$\pi_n^{-1}A=A\times\mathbb{R}^\infty,\quad A\in\mathcal{B}(\mathbb{R}^n)$$

are called *cylinder sets*.

For $n\ge 1$, let μ_n be probability measures on $\big(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n)\big)$. We say that μ_n satisfies the *consistency condition*, if

$$\mu_{n+1}\circ\pi_n^{-1}=\mu_n,$$

or, equivalently,

$$\mu_{n+1}(A\times\mathbb{R})=\mu_n(A),\quad\forall A\in\mathcal{B}(\mathbb{R}^n),$$

or,

$$\mu_{n+m}\circ\pi_n^{-1}=\mu_n,\quad\forall m,n\ge 1. \tag{3.16}$$

Theorem 3.9 (Kolmogorov's Extension Theorem) *Assume (3.16). There exists a unique measure μ on $(\mathbb{R}^\infty,\mathcal{B}(\mathbb{R}^\infty))$ such that $\mu\circ\pi_n^{-1}=\mu_n$ for every $n\ge 1$, i.e.,*

$$\mu(A\times\mathbb{R}^\infty)=\mu_n(A),\quad\forall A\in\mathcal{B}(\mathbb{R}^n). \tag{3.17}$$

To construct an infinite sequence of independent r.v.s, we will use Theorem 3.9 in the following way. Given $\big(\mathbb{R},\mathcal{B}(\mathbb{R}),\lambda_i\big)$, $i\ge 1$, let

$$\mu_n=\bigtimes_{i=1}^{n}\lambda_i$$

be probability measures on $\big(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n)\big)$. Then μ_n satisfies the consistency condition (3.16) by properties of the product measures. Then by Theorem 3.9, there exists a unique probability measure μ on $(\mathbb{R}^\infty,\mathcal{B}(\mathbb{R}^\infty))$ so that (3.17) holds. Let

$$X_n(\omega)=\omega_n,\quad n\ge 1.$$

Then $(X_n)_{n\ge1}$ are independent r.v.s on $\big(\mathbb{R}^\infty,\mathcal{B}(\mathbb{R}^\infty),\mu\big)$.

Next, we will prove Theorem 3.9. Before that, we need to understand compact sets in $\mathbb{R}^\infty$.

Proposition 3.10 *Let F_m, $m \geq 1$, be nonempty compact sets in $\mathbb{R}^m$ such that*

$$D_m = \pi_m^{-1}(F_m) = F_m \times \mathbb{R}^\infty$$

are decreasing in m. Then $\bigcap_{m=1}^\infty D_m \neq \varnothing$.

Proof: For every $m \geq 1$, pick $x^{(m)} \in D_m$. Since D_m are decreasing cylinder sets, for every $n \geq 1$, we have $(\pi_n x^{(m)})_{m \geq n} \subset \pi_n(D_n) = F_n$ is a bounded sequence in $\mathbb{R}^n$.

Bounded sequences in $\mathbb{R}^n$ have convergence subsequence. Therefore, there exists $(m_k^1)_{k\geq 1}$ so that $\pi_1 x^{(m_k^1)}$ converges in $\mathbb{R}^1$, and $(m_k^2)_{k\geq 2} \subset (m_k^1)_{k\geq 1}$ so that $\pi_2 x^{(m_k^2)}$ converges in $\mathbb{R}^2$ and so on. Let $y^{(k)} = x^{(m_k^k)}$ be the diagonal sequence. For every $n \geq 1$, the sequence $(\pi_n y^{(k)})_{k\geq 1}$ converges in $\mathbb{R}^n$ by construction. By (3.14), there exists $y^* \in \mathbb{R}^\infty$ such that $y^{(k)} \to y^*$ in $\mathbb{R}^\infty$. Noting that $\pi_n y^{(k)} \in F_n$ for $k \geq n$, we have $y^* \in D_n$ for every n, and thus $y^* \in \bigcap_{n=1}^\infty D_n$. This proves the conclusion. □

Remark 3.4 A similar argument shows that the metric we put on $\mathbb{R}^\infty$ is such that for any $L_n \in (0, \infty)$, the product set

$$\bigtimes_{n=1}^{\infty} [-L_n, L_n]$$

is sequentially compact in $\mathbb{R}^\infty$.

We also need a small lemma about the regularity of Borel sets in $\mathbb{R}^d$.

Proposition 3.11 *Let λ be a probability measure on $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^D)\big)$. Let $A \in \mathcal{B}(\mathbb{R}^d)$. For every $\varepsilon > 0$, there exists a closed set F_ε and an open set G_ε such that*

$$F_\varepsilon \subset A \subset G_\varepsilon, \quad \lambda(G_\varepsilon) - \lambda(F_\varepsilon) < \varepsilon.$$

Moreover, F_ε can be chosen to be compact since

$$\lim_{L\to\infty} \lambda(F_\varepsilon \cap [-L, L]^d) = \lambda(F_\varepsilon).$$

Proof: Let $\mathcal{S}$ be the collection of sets A that satisfy the condition. One can show that $\mathcal{S}$ forms a σ-algebra, and clearly $\mathcal{S}$ contains rectangles $(a_1, b_1) \times \cdots \times (a_d, b_d)$. Therefore, $\mathcal{S} \supset \mathcal{B}(\mathbb{R})$. □

Proof of Theorem 3.9: Let $\mathcal{C} = \{\text{cylinder sets}\}$. We have the following.

1. $\mathcal{C}$ is an algebra.
2. The condition (3.17) specifies the measure μ on $\mathcal{C}$.
3. (3.15) implies that $\mathcal{B}(\mathbb{R}^\infty) = \sigma(\mathcal{C})$.
4. The consistency condition (3.16) implies that (3.17) defines a finitely additive measure μ on $\mathcal{C}$.

Putting all these together, we can use the Carathéodory's Extension Theorem to construct the desired measure μ, provided that we verify that μ is σ-additive on $\mathcal{C}$.

To show σ-additivity, it suffices to show continuity at $\varnothing$, that is, $\mu(D_n) \to 0$ for every $\mathcal{C} \ni D_n \downarrow \varnothing$.

Without loss of generality, we can assume that $D_n = \pi_n^{-1}(B_n)$ where $B_n \in \mathcal{B}(\mathbb{R}^n)$. We will prove by contradiction.

Assume the contrary that there exists $\delta > 0$ such that $\mu(D_n) = \mu_n(B_n) \geq \delta$ for every n. By Proposition 3.11, there exist compact sets $F_n \subset B_n$ such that $\mu_n(B_n \setminus F_n) \leq \delta 2^{-n-1}$, $n \geq 1$.

Let $\hat{E}_n = \pi_n^{-1}(F_n) \in \mathcal{C}$. Then $\mu(D_n \setminus \hat{E}_n) = \mu_n(B_n \setminus F_n) \le \delta 2^{-n-1}$. The sets $\hat{E}_n$ may not be decreasing, but if we set

$$E_n = \bigcap_{m=1}^{n} \hat{E}_m, \quad n \ge 1,$$

then E_n are decreasing. Moreover,

$$\mu(D_n \setminus E_n) \le \mu\Big(\bigcup_{m=1}^{n} (D_n \setminus \hat{E}_n)\Big) \le \sum_{m=1}^{n} \frac{\delta}{2^{m+1}} \le \frac{\delta}{2}.$$

Hence, $\mu(E_n) \ge \mu(D_n) - \delta/2 \ge \delta/2$ for all $n \ge 1$. In particular, $E_n \ne \varnothing$ for all n, and hence we can apply Proposition 3.10 to conclude that $\bigcap_{n=1}^{\infty} E_n \ne \varnothing$. But

$$\bigcap_{n=1}^{\infty} E_n \subset \bigcap_{n=1}^{\infty} D_n = \varnothing,$$

and we arrive at a contradiction. □

Remark 3.5 Instead of $\big(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty)\big)$, Kolmogorov's Extension Theorem can also be stated for general measurable spaces $(\bigtimes_{n=1}^{\infty} S_n, \bigotimes_{n=1}^{\infty} \mathcal{S}_n)$. To verify the σ-additivity needed for Carathéodory's Extension Theorem, some topological information is needed for the spaces $(S_n, \mathcal{S}_n)$. A sufficient condition is that all $(S_n, \mathcal{S}_n)$ are *Borel* spaces: a measurable space $(S, \mathcal{S})$ is called Borel if there is a one-to-one map $\varphi : (S, \mathcal{S}) \to ([0,1], \mathcal{B}[0,1])$ so that φ and φ^{-1} are both measurable. In particular, all complete and separable metric spaces equipped with Borel σ-algebras are Borel.

Remark 3.6 One can also consider Kolmogorov's Extension Theorem on $\big(\mathbb{R}^T, \mathcal{B}(\mathbb{R}^T)\big)$, where T is *any* index set, and the Borel σ-algebra $\mathcal{B}(\mathbb{R}^T)$ is generated by all "(finite-dimensional) cylinder sets"

$$\pi_{t_1,t_2,\dots,t_n}^{-1}(A_n), \quad A_n \text{ open set in } \mathbb{R}^n,\ t_1, \dots, t_n \in T.$$

All cylinder sets form an algebra, and a probability measure μ on this space exists, provided that its "finite-dimension distributions" $\mu \circ \pi_{t_1,\dots,t_n}^{-1}$ satisfy the consistency condition. Every probability measure on $\big(\mathbb{R}^T, \mathcal{B}(\mathbb{R}^T)\big)$ gives rise to a *stochastic process on* T.

Unfortunately, measure spaces constructed in this way is not immediately suitable for the study of stochastic processes. For example, if $T = \mathbb{R}$, then a probability measure on $\big(\mathbb{R}^T, \mathcal{B}(\mathbb{R}^T)\big)$ will model a random function $f_\omega : \mathbb{R} \to \mathbb{R}$. However, simple events, like $\{\omega : f_\omega \text{ continuous}\}$, will not be measurable. This is the main obstacle in the construction of Brownian motions and stochastic analysis. Some discuss in this direction can be found in [Shi96, Chap. II.2.5] and [KS, Chap. 2.2].

4 Law of large numbers

The goal of this section is to establish the following strong law of large numbers (SLLN).

Theorem 4.1 (Strong law of large number) *Let $X_1, X_2, \dots$ be i.i.d. with $\mathsf{E}|X_i| < \infty$. Let $\mathsf{E}X_i = \mu$ and $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ a.s. as $n \to \infty$.*

The above theorem is called "strong" because almost sure convergence is the best that one can hope. Similar statements where the convergence holds in a weaker sense, like in L^p or in probability are called "weak" law of large numbers.

In Theorem 4.1, the first moment condition $\mathsf{E}|X_1| < \infty$ will be optimal. But we will also introduce proofs under weaker assumptions, as an opportunity to introduce useful probabilistic techniques that may be useful for other problems.

4.1 L^2-weak law of large numbers

Let X_n, $n \geq 1$, be i.i.d. r.v.s. For the discussion of law of large numbers, we assume, without loss of generality, that all X_n are *centered*, namely, $\mu := \mathsf{E}X_n = 0$. Otherwise, we can always center the r.v.s by setting $\tilde{X}_i = X_i - \mu$ and consider the centered case. For centered r.v.s, we have

$$\mathsf{E}X_i^2 = \mathrm{Var}(X_i), \quad \mathsf{E}X_iX_j = \mathrm{Cov}(X_i, X_j), \ i \neq j.$$

We also denote the partial sum by $S_n = X_1 + \cdots + X_n$.

The r.v.s $(X_i)_{i\in I}$ with $\mathsf{E}X_1^2 < \infty$ is said to be *uncorrelated* if

$$\mathsf{E}(X_iX_j) = \mathsf{E}X_i\mathsf{E}X_j \quad \text{whenever } i \neq j. \tag{4.1}$$

We note that the second moment condition $\mathsf{E}X_i^2 < \infty$ ensures that expectations in (4.1) are all defined. When $\mu = 0$, (4.1) becomes

$$\mathsf{E}(X_iX_j) = 0, \quad \forall i \neq j. \tag{4.2}$$

Let a family of random variables $(X_n)_{n\geq 1}$ with $\mathsf{E}X_1^2 < \infty$ be uncorrelated. By linearity of expectation, we have

$$\mathsf{E}S_n = \mathsf{E}X_1 + \cdots + \mathsf{E}X_n = n\mu = 0.$$

Using definition of the variance, we have

$$\mathrm{Var}(S_n) = \mathsf{E}S_n^2 = \mathsf{E}(\sum_{i=1}^n X_i)(\sum_{j=1}^n X_j) = \sum_{i=1}^n \mathsf{E}X_i^2 = n\mathsf{E}X_1^2. \tag{4.3}$$

A key observation is that the variance grows linearly in n, although it is the expectation of the sum of n^2 terms. Assuming $\mathsf{E}X_1^4 < \infty$, we can further estimate the fourth moment of S_n:

$$\begin{aligned} \mathsf{E}S_n^4 &= \sum_{i_1,i_2,i_3,i_4} \mathsf{E}X_{i_1}X_{i_2}X_{i_3}X_{i_4} = \sum_{i=1}^n \mathsf{E}X_i^4 + 6\sum_{i<j} \mathsf{E}X_i^2X_j^2 \\ &\leq n\mathsf{E}X_1^4 + 3\sum_{i<j} \mathsf{E}(X_i^4 + X_j^4) = (3n^2 - 2n)\mathsf{E}X_1^4 \leq Cn^2. \end{aligned} \tag{4.4}$$

Here, in the first line, if any index appears in i_1, i_2, i_3, i_4 only once, then by (4.2), the expectation $\mathsf{E}X_{i_1}X_{i_2}X_{i_3}X_{i_4}$ will be zero and such terms can be dropped from the summation; in the second line we use the elementary inequality $2ab \leq a^2 + b^2$. Again, we see that $\mathsf{E}S_n^4$ grows only quadratic in n, which is n^2 in order less than the number of terms, n^4. The discrepancy will get larger if we estimate higher moments of S_n. But the fourth moment is sufficient for us to use Borel–Cantelli to get the first strong law of large numbers.

Proposition 4.2 *Let $X_1, X_2, \ldots$ be i.i.d. $\mathsf{E}X_1^4 < \infty$. Then $S_n/n \to 0$ a.s.*

Proof: Since $\mathsf{E}X_1^4 < \infty$, by (4.4) and Chebyshev's inequality (Lemma 2.9), for some constant $C > 0$ we have

$$\mathsf{P}(|S_n| > n\varepsilon) = \mathsf{P}(|S_n|^4 > n^4\varepsilon^4) \leq \frac{Cn^2}{n^4\varepsilon^4} \leq \frac{C}{n^2\varepsilon^4}.$$

Since $\sum_{n=1}^\infty \frac{1}{n^2} < \infty$, by Borel–Cantelli lemma (Lemma 2.3), we have

$$\mathsf{P}(\{|S_n| > n\varepsilon \text{ i.o.}\}) = 0.$$

It follows from the ε-δ language formulation of limit

$$\Big\{\lim_{n\to\infty}\frac{S_n}{n}\neq 0\Big\}=\bigcup_{m=1}^{\infty}\Big\{|\frac{S_n}{n}|>\frac{1}{m}\text{ i.o.}\Big\}.$$

Hence, by sub-additivity,

$$\mathsf{P}\Big(\Big\{\lim_{n\to\infty}\frac{S_n}{n}\neq 0\Big\}\Big)\le\sum_{m=1}^{\infty}\mathsf{P}\Big(\Big\{|\frac{S_n}{n}|>\frac{1}{m}\text{ i.o.}\Big\}\Big)=0,$$

and this completes the proof. □

Proposition 4.2 already yields many applications, since in many practical examples r.v.s are bounded and have fourth moment. In fact, in (4.4) and Proposition 4.2, only the independence of X_n and a uniform bound $\mathsf{E}X_i^4<C$ are used. Similarly, assuming only the second moment condition, we can obtain the following *weak law of large numbers* without independence.

Theorem 4.3 (Weak law of large numbers) *Let $X_1,\ X_2,\dots$ be uncorrelated with $\mathsf{E}X_i^2\le C$ for some $C>0$. Then as $n\to\infty$, $S_n/n\to 0$ in L^2 and in pr.*

Proof: Since X_i are uncorrelated, using (4.3) we have $\mathsf{E}S_n^2\le Cn$, and hence $\mathsf{E}S_n^2/n^2\le C/n$. It follows that $S_n/n\to 0$ in L^2. By Proposition 2.8, this implies convergence in pr. □

Using the second moment condition, it is also possible to obtain almost sure convergence.

Theorem 4.4 (SLLN with $\mathsf{E}X_1^2<\infty$) *Let X_n, $n\ge 1$, be i.i.d. with $\mathsf{E}X_1^2<\infty$. Then $\frac{S_n}{n}\to 0$, a.s.*

Proof: Let $M=\mathsf{E}X_1^2$. By (4.3) and Chebyshev's inequality, we have

$$\mathsf{P}(|S_{n^2}|>n^2\varepsilon)=\mathsf{P}(|S_{n^2}|^2>n^4\varepsilon^2)\le\frac{nM}{n^4\varepsilon^2}\le\frac{M}{n^3\varepsilon^2},$$

which is summable. Hence, by Borel–Cantelli lemma, $\frac{S_{n^2}}{n^2}\to 0$ a.s. Let

$$D_n(\omega)=\max_{n^2\le k<(n+1)^2}|S_{k^2}-S_{n^2}|=\max_{1\le k\le 2n}|X_{n^2+1}+\cdots+X_{n^2+k}|.$$

For every ω, we have

$$|D_n(\omega)|^2\le\big(|X_{n^2+1}|+\cdots+|X_{n^2+2n}|\big)^2\le 2n(X_{n^2+1}^2+\cdots+X_{n^+2n}^2)$$

and hence $\mathsf{E}D_n^2\le 2nM$. Then, by Chebyshev's inequality, we have

$$\mathsf{P}(D_n\ge n^{1+\varepsilon})\le\frac{\mathsf{E}D_n^2}{n^{2+2\varepsilon}}\le\frac{2M}{n^{1+2\varepsilon}}.$$

It follows from Borel-Cantelli lemma $\mathsf{P}(\{D_n\ge n^{1+\varepsilon},\text{ i.o.}\})=0$.

To summarize, for almost every ω, we have

1. $\lim_{n\to\infty}\frac{S_n^2}{n^2}=0$.
2. There exists $n_0=n_0(\omega)$, for every $n\ge n_0$, $|D_n|\le n^{1+\varepsilon}$.

When the two conditions above hold for ω, by

$$\frac{S_{n^2} - D_n}{(n+1)^2} \le \frac{S_k}{k} \le \frac{S_{n^2} + D_n}{n^2}, \quad n^2 \le k < (n+1)^2,$$

and the Squeeze Theorem, we have $S_k/k \to 0$. This completes the proof. □

Remark 4.1 1. We only need $\mathsf{E}X_iX_j = 0$, $i \neq j$ (uncorrelated) and $\sup_n \mathsf{E}X_n^2 < \infty$.

2. The above condition can be further weaken to allow some finite-range correlation:

$$|\mathsf{E}X_iX_j| \le M \cdot \mathbb{1}_{\{|i-j|\le L\}}$$

for some $L > 0$ and $M > 0$.

Example 4.2 (Normal number) Every $\omega \in [0, 1)$ admits a decimal expansion

$$\omega = 0.x_1x_2x_3x_4\cdots, \quad x_i = x_i(\omega) \in \{0, 1, \ldots, 9\}.$$

Let

$$\nu_k^{(n)}(\omega) = \big|\{1 \le i \le n : x_i = k\}\big| = \sum_{i=1}^{n} \mathbb{1}_{\{x_i(\omega)=k\}}$$

be the number of occurrence of number k in the first n digits. It is clear that $x_i(\omega)$ are i.i.d., uniformly on $\{0, 1, \ldots, 9\}$. Then $\xi_i = \mathbb{1}_{\{X_i(\omega)=k\}}$ are i.i.d. Ber(1/10). Clearly, we have $|\xi_i| \le 1$. For every k, By SLLN, for almost every $\omega \in [0, 1)$,

$$\frac{\nu_k^{(n)}(\omega)}{n} = \frac{\sum_{i=1}^n \xi_i}{n} \to \mathsf{E}\xi_i = \frac{1}{10}, \quad k \in \{0, \ldots, 9\}. \tag{4.5}$$

A number ω is called a *normal number* (Borel, 1909) if for its fractional part, the limit (4.5) holds. As a consequence of the SLLN, almost every number in $[0, 1)$ is normal. However, we do not know whether common transcendental numbers like π or ε are normal.

We can also strengthen the definition slightly. A number $\omega \in [0, 1)$ is *completely normal*, if for every pattern $\vec{k} = (k_1, k_2, \ldots, k_r) \in \{0, \ldots, 9\}^r$,

$$\frac{\nu_{\vec{k}}^{(n)}(\omega)}{n} := \frac{\big|\{1 \le i \le n : (x_i, \ldots, x_{i+r-1}) = \vec{k}\}\big|}{n} \to \frac{1}{10^r}, \quad n \to \infty.$$

Using the remark after Theorem 4.4 with $L = r$ and $M = 1$, almost every $\omega \in [0, 1)$ is also completely normal.

As an illustration, if a monkey sits before a typewriter randomly typing, then eventually it will produce all Shakespeare's works (more than once), as any pattern $\vec{k}$, even as long as all Shakespeare's works, has a small but positive probability of occurrence. This seems paradoxical, but note that the waiting time will be much longer than the age of the universe in this case, so it is not practically possible.

Example 4.3 (Empirical distribution function) Let $X_1, X_2, \ldots$ be i.i.d. samples with c.d.f. F and let

$$F_n(x) = n^{-1} \sum_{m=1}^{n} \mathbb{1}_{\{X_m \le x\}}, \quad \forall x \in \mathbb{R}$$

be the *empirical distribution function* from n samples. For every x, the indicators $\xi_n(\omega) := \mathbb{1}_{X_n(\omega)\le x}$ are i.i.d. r.v.s since they are Borel functions of X_n. By SLLN, we have

$$F_n(x) = \frac{\sum_{m=1}^n \xi_m}{n} \xrightarrow{\text{a.s.}} \mathsf{E}\xi_n = \mathsf{P}(\{X_n \le x\}) = F(x).$$

Theorem 4.5 (Glivenko–Cantelli theorem) *As $n \to \infty$, $\sup_x |F_n(x) - F(x)| \to 0$ a.s.*

(To fill in the proof.)

Example 4.4 (Waiting time Paradox) This is example is related to the renewal theory.

Let $X_1, X_2, \dots$ are i.i.d. Suppose that the n-th bus from the bus terminal at time S_n, where $S_n = X_1 + \cdots + X_n$. For simplicity assume that $\mathsf{P}\{X_n = a\} = \mathsf{P}\{X_n = b\} = \frac{1}{2}$ for some $a < b$. We are trying to compute the "average waiting time" for a person randomly arriving at the terminal before departure.

We first compute how many buses departing in the time interval $[0, T]$. Let

$$N = N_T(\omega) = \text{the number of buses departing in } [0,T] = \max\{n : S_n(\omega) \le T\}.$$

Since

$$\frac{X_1 + \cdots + X_{N_T}}{N_T} < \frac{T}{N_T} < \frac{X_1 + \cdots + X_{N_T+1}}{N_T + 1} \cdot \frac{N_T + 1}{N_T}$$

it follows from the Squeeze Theorem and SLLN that

$$\frac{T}{N_T} \to \mathsf{E}X_1 = \frac{a+b}{2}, \quad \text{a.s.},$$

and hence

$$\frac{N_T}{T} \to \frac{1}{\mathsf{E}X_1} = \frac{2}{a+b}, \quad \text{a.s.} \tag{4.6}$$

We interpret the "average waiting time" as follows. Let a person arrive at the bus stop at time $\xi \sim U[0,1]$, where ξ is independent of $(X_n)_{n\ge1}$ (we can realize this by accommodate ξ and $(X_n)_{n\ge1}$ on a bigger product probability space). The average waiting time Q is given by

$$Q = \frac{1}{T}\int_0^T (S_{n_\xi} - \xi)\, d\xi,$$

where $n_\xi = \min\{m : S_m > \xi\}$ is the departure time of the next bus after time ξ. Noting that $n_\xi = n$ if $\xi \in [S_{n-1}, S_n)$, we have

$$Q = \frac{1}{T}\sum_{n=1}^{N_T}\int_{S_{n-1}}^{S_n} (S_n - \xi)\, d\xi = \frac{1}{T}\sum_{n=1}^{N_T}\frac{(S_n^2 - S_{n-1}^2)}{2} = \frac{1}{T}\sum_{n=1}^{N_T}\frac{X_n^2}{2}.$$

it follows from the SLLN for X_i^2 and (4.6)

$$Q = \frac{1}{T}\sum_{n=1}^{N_T}\frac{X_n^2}{2} = \frac{X_1^2 + \cdots + X_{N_T}^2}{X_T} \cdot \frac{N_T}{2T} \overset{\text{a.s.}}{\to} \mathsf{E}X_I^2 \cdot \frac{1}{a+b} = \frac{a^2+b^2}{2(a+b)} = \frac{1}{2}\Big(a \cdot \frac{a}{a+b} + b \cdot \frac{b}{a+b}\Big). \tag{4.7}$$

How to understand (4.7)? If the time for the next departure is τ, then for a person arriving at a random time the average waiting time should be $\tau/2$. One would think naively that since τ takes the value a and b with probability $1/2$, then the average waiting time should be $(a+b)/2$. But this is WRONG. Indeed, the number of intervals with length a and b are roughly 50%, but since their lengths are different, the random arrival time hitting these two types of intervals are also different, or more precisely, proportional to their lengths. Therefore, the probability of the arrival time hitting $[S_{n-1}, S_n)$ with $X_{n-1} = a$ is asymptotically $\frac{a}{a+b}$, and $\frac{b}{a+b}$ otherwise. This explains the rightmost decomposition in (4.7).

4.2 Weak law for triangular arrays

Many classical limit theorems in probability concern arrays $X_{n,k}$, $1 \le k \le n$, of random variables and investigate the limiting behavior of their row sums $S_n = X_{n,1} + \cdots + X_{n,n}$.

Proposition 4.6 *Let $(X_{n,k})_{k=1}^n$ be independent and $\mu_n = \mathsf{E}S_n$, $\sigma_n^2 = \mathrm{Var}(S_n)$. If $\sigma_n^2/b_n^2 \to 0$, then*

$$\frac{S_n - \mu_n}{b_n} \to 0, \quad \textit{in probability.}$$

Proof: Chebyshev's inequality gives that for every $\varepsilon > 0$,

$$\mathsf{P}\Big(\Big|\frac{S_n - \mu_n}{b_n}\Big| \geq \varepsilon\Big) \leq \frac{\mathrm{Var}(S_n)}{\varepsilon^2 b_n^2} = \frac{\sigma_n^2}{\varepsilon^2 b_n^2} \to 0.$$

□

Example 4.5 (Coupon collector) Let $\xi_1, \xi_2, \ldots$ be i.i.d. uniform on $\{1, 2, \ldots, n\}$. The numbers $1, \ldots, n$ are thought of as "coupons" while ξ_m is the m-th coupon that one collects. Let

$$\tau_k^n = \min\{m : m \geq 0,\ |\{\xi_1, \ldots, \xi_m\}| \geq k\}$$

be the first time that one collects k different coupons. For example, we always have $\tau_1^n = 1$. We Set $\tau_0^n = 0$ for consistency of notation.

For $1 \leq k \leq n$, let $X_{n,k} = \tau_k^n - \tau_{k-1}^n$ represent the time spent to collect the k-th coupon. We claim the following two facts without proof:

1. $X_{n,k}$ is independent of $X_{n,1}, \ldots, X_{n,k-1}$;
2. $X_{n,k}$ has a geometric distribution with parameter $1 - (k-1)/n$.

Let $S_n = X_{n,1} + X_{n,2} + \cdots + X_{n,n} = \tau_n^n$. We want to understand the asymptotic behavior of S_n, the time spent to collect all coupons.

To use the result from Proposition 4.6, we need to compute $\mathsf{E}S_n$ and $\mathrm{Var}(S_n)$. Note that if $Y \sim \mathrm{Geo}(p)$, then $\mathsf{E}Y = 1/p$ and $\mathsf{E}Y^2 \leq 1/p^2$. We have

$$\mathsf{E}S_n = \sum_{k=1}^n X_{n,k} = \sum_{k=1}^n (1 - \frac{k-1}{n})^{-1} = n \sum_{m=1}^n m^{-1} \sim n \log n, \tag{4.8}$$

and

$$\mathrm{Var}(S_n) = \sum_{k=1}^n \mathrm{Var}(X_{n,k}) \leq n^2 \sum_{m=1}^n \frac{1}{m^2} \leq Cn^2.$$

Hence, for $b_n = n \log n$, $\sigma_n / b_n \to 0$, and it follows from Proposition 4.6

$$\frac{S_n - \mathsf{E}S_n}{b_n} \to 0 \text{ in probability.}$$

Noting (4.8), we have $\frac{S_n}{b_n} \to 1$ in probability.

Let $\mathsf{E}|X| < \infty$ and $(X_{n,k})_{k=1}^n$, $1 \leq k \leq n$ be independent. Let $b_n > 0$ with $b_n \to \infty$. We introduce the "truncation" of $X_{n,k}$ as follows:

$$\bar{X}_{n,k} = X_{n,k} \mathbb{1}_{(|X_{n,k}| \geq b_n)} = \begin{cases} X_{n,k}, & \text{if } |X_{n,k}| \leq b_n \\ 0, & \text{if } |X_{n,k}| > b_n. \end{cases} \tag{4.9}$$

The truncation will help us to obtain the weak law to random variables without a finite second moment.

Theorem 4.7 (Weak LLN for triangular arrays) *Let $X_{n,k}$, $1 \leq k \leq n$, be independent. Let $b_n > 0$ with $b_n \to \infty$ and $X_{n,k}$ be defined in (4.9). Suppose that as $n \to \infty$,*

1. $\sum_{k=1}^n \mathsf{P}(|X_{n,k}| > b_n) \to 0$, *and*
2. $b_n^{-2} \sum_{k=1}^n \mathsf{E}\bar{X}_{n,k}^2 \to 0$.

Then

$$(S_n - a_n)/b_n \to 0 \quad \textit{in probability,}$$

where $S_n = X_{n,1} + \cdots + X_{n,n}$ and $a_n = \sum_{k=1}^n \mathsf{E}\bar{X}_{n,k}$.

Proof: For every $\varepsilon > 0$, we have

$$\mathsf{P}(|\frac{S_n - a_n}{b_n}| > \varepsilon) \le \mathsf{P}(S_n \neq \bar{S}_n) + \mathsf{P}(|\frac{\bar{S}_n - a_n}{b_n}| > \varepsilon)$$

To estimate the first term, we note that

$$\mathsf{P}(S_n \neq \bar{S}_n) \le \mathsf{P}(\bigcup_{k=1}^{n} \{\bar{X}_{n,k} \neq X_{n,k}\}) \le \sum_{k=1}^{n} \mathsf{P}(|X_{n,k}| > b_n) \to 0$$

by the first condition. For the second term, we use Chebyshev's inequality to obtain

$$\begin{aligned}\mathsf{P}(|\frac{\bar{S}_n - a_n}{b_n}| > \varepsilon) &\le \frac{1}{\varepsilon^2}\mathsf{E}|\frac{\bar{S}_n - a_n}{b_n}|^2 = \frac{\operatorname{Var}(\bar{S}_n)}{\varepsilon^2 b_n^2} \\ &= \frac{\sum_{k=1}^{n} \operatorname{Var}(\bar{X}_{n,k})}{\varepsilon^2 b_n^2} \le \frac{\sum_{k=1}^{n} \mathsf{E}(\bar{X}_{n,k})^2}{\varepsilon^2 b_n^2} \to 0\end{aligned}$$

by the second condition, and the proof is complete. □

Theorem 4.8 *Let $X_1, X_2, \ldots$ be i.i.d. with $\mathsf{E}|X_i| < \infty$. Let $S_n = X_1 + \cdots + X_n$ and let $\mu = \mathsf{E}X_1$. Then $S_n/n \to \mu$ in probability.*

Proof: Let $X_{n,k} = X_k$ and $b_n = n$. We need to check the two conditions of Theorem 4.7.

For the first condition, by DCT, we have

$$\sum_{k=1}^{n} \mathsf{P}(|X_k| > n) = n\mathsf{P}(|X_1| > n) \le \mathsf{E}|X_1|\mathbb{1}_{\{|X_1| \ge n\}} \to 0, \tag{4.10}$$

since $\mathbb{1}_{\{|X_1| \ge n\}}|X_1| \overset{\text{a.s.}}{\to} 0$ and $\mathsf{E}|X_1| < \infty$.

For the second condition, we have

$$\frac{1}{n^2}\sum_{k=1}^{n} \mathsf{E}|X_k|^2\mathbb{1}_{\{|X_k| \le n\}} = \frac{1}{n}\mathsf{E}|X_1|^2\mathbb{1}_{\{|X_1| \le n\}}$$

and

$$\begin{aligned}\mathsf{E}|X_1|^2\mathbb{1}_{\{|X_1| \le n\}} &= \sum_{k=1}^{n} \mathsf{E}|X_1|^2\mathbb{1}_{\{|X_1| \in [k-1,k]\}} \\ &\le \sum_{k=1}^{n} k^2\mathsf{P}(|X_1| \in [k-1,k]) \\ &= \mathsf{P}(|X_1| \in [0,1]) + \sum_{k=1}^{n}((k+1)^2 - k^2)\mathsf{P}(|X_1| \in [1,n]) \\ &\le \mathsf{P}(|X_1| \in [0,1]) + \sum_{k=1}^{n} 3k\mathsf{P}(|X_1| \ge k)\end{aligned}$$

By Stolz's theorem, we have $\frac{1}{n}\sum_{k=1}^{n} 3k\mathsf{P}(|X_1| \ge k) \to \lim_{n\to\infty} n\mathsf{P}(|X_1| \ge n) = 0$, again by (4.10).

Note that $a_n = n\mu_n$ where $\mu_n = \mathsf{E}X_1\mathbb{1}_{\{|X_1| \le n\}} \neq \mu$ due to the truncation. But by DCT,

$$\mu_n = \mathsf{E}X_1\mathbb{1}_{(|X_1| \le n)} \to \mathsf{E}X_1 = \mu.$$

□

Example 4.6 (St. Petersburg's game) Let $X_1, X_2, \ldots$ be independent random variables with

$$\mathsf{P}(X_i = 2^j) = 2^{-j} \quad \text{for } j \geq 1. \tag{4.11}$$

Imagine you are playing a game continuously tossing a coin. You win 2^j dollars if it takes $j+1$ tosses to get a head, but if you can a head the first toss you leave without any reward. Now we want to determine what is the "fair" entry fee to play this game. Since $\mathsf{E}X_1 = \infty$, the LLN is useless, as it is not reasonable to ask ∞ dollars for the entry fee!

Now we will try to use Theorem 4.7 to find out how much we should ask for the entry fee. The answer will depend on the total number of games to be played. Indeed, we are trying to find c_n where $S_n/nc_n \to 1$.

In the setting of Theorem 4.7, let $X_{n,k} = X_k$. We need to determine $b_n = nc_n$. We observe that if m is an integer

$$\mathsf{P}(X_1 \geq 2^m) = \sum_{j=m}^{\infty} 2^{-j} = 2^{-m+1}$$

Let $m(n) = \log_2 n + K(n)$ where $K(n) \to \infty$ and is chosen so that $m(n)$ is an integer (and hence the displayed formula is valid). Letting $b_n = 2^{m(n)}$, we have

$$\mathsf{E}\bar{X}_{n,k}^2 = \sum_{j=1}^{m(n)} 2^{2j} \cdot 2^{-j} \leq 2^{m(n)} \sum_{k=0}^{\infty} 2^{-k} = 2b_n$$

The last two steps are to evaluate a_n and to apply the theorem.

$$\mathsf{E}\bar{X}_{n,k} = \sum_{j=1}^{m(n)} 2^j 2^{-j} = m(n)$$

so $a_n = nm(n)$. We have $m(n) = \log_2 n + K(n)$, so if we pick $K(n)/\log_2 n \to 0$ then $a_n/n\log_2 n \to 1$ as $n \to \infty$. Now we have

$$\frac{S_n - a_n}{n2^{K(n)}} \to 0 \quad \text{in probability}$$

If we suppose that $K(N \leq \log_2 \log_2 n)$ for large n then the last conclusion holds with the denominator replaces by $n\log_2 n$, and it follows that $S_n/(n\log_2 n) \to 1$ in probability.

4.3 First proof of SLLN

4.3.1 Some preparation

We recall the (first) Borel-Cantelli Lemma: if $\sum_{n=1}^{\infty} \mathsf{P}(A_n) < \infty$ then $\mathsf{P}(A_n \text{ i.o.}) = 0$. For the other direction, we have the following.

Theorem 4.9 (Second Borel–Cantelli lemma) *If the events A_n are independent and $\sum_{n=1}^{\infty} \mathsf{P}(A_n) = \infty$, then $\mathsf{P}(\{A_n,\ i.o.\}) = 1$*

Proof: By definition of the i.o. sets, we have

$$\{A_n \text{ i.o.}\}^c = \bigcup_{n=1}^{\infty} \bigcap_{n=m}^{\infty} A_n^c.$$

Using independence, it follows that

$$\mathsf{P}\Big(\bigcap_{n=m}^{\infty} A_n^c\Big) = \lim_{M\to\infty} \prod_{n=m}^{M} \mathsf{P}(A_n^c) = \lim_{M\to\infty} \prod_{n=m}^{M} \big(1 - \mathsf{P}(A_n)\big) = 0,$$

where the last limit is due to $\sum_{n=1}^{\infty} \mathsf{P}(A_n) = \infty$. □

The following proposition states that the $\mathsf{E}|X_1| < \infty$ is also necessary for the existence of $\lim_{n\to\infty} S_n/n$.

Proposition 4.10 *If $X_1, X_2 \dots$ are i.i.d. and $\mathsf{E}|X_i| = \infty$, then $\mathsf{P}\big(\lim S_n/n \text{ exists } \in (-\infty, \infty)\big) = 0$.*

Proof: Let $A_n = \{|X_n| \geq n\}$. We claim that on the event $\{A_n, \text{ i.o.}\}$, a *finite* limit $\lim_{n\to\infty} S_n/n$ cannot exist. Indeed, by Cauchy criterion, if such limit exists, for $\varepsilon_0 = \frac{1}{2}$, there exists $n_0 = n_0(\omega)$ such that $|X_n/n| = |\frac{S_n}{n} - \frac{S_{n-1}}{n}| < \frac{1}{2}$ for every $n > n_0$. This contradicts with $|X_n| \geq n$ for infinitely many n's.

By Theorem 4.9, since

$$\infty = \mathsf{E}|X_1| \leq \sum_{n=0}^{\infty} \mathsf{P}(|X_1| > n) = \sum_{n=1}^{\infty} \mathsf{P}(|X_1| \geq n)$$

and $X_1, X_2 \dots$ are i.i.d., it follows that $\mathsf{P}(\{A_n, \text{ i.o.}\}) = 1$. This completes the proof. □

Example 4.7 (St. Petersburg's game (continued)) Let X_n, $n \geq 1$, be i.i.d. with distribution given by (4.11). By Proposition 4.10, since $\mathsf{E}X_1 = \infty$, we know that S_n/n does not have a limit. But if we are more careful about the estimate, we have

$$\mathsf{P}(|X_n| \geq n \log_2 n) = \sum_{j \geq j_0 = [\log_2(n \log_2 n)]} 2^{-j} \sim 2^{-\log_2(n \log_2 n)} = \frac{1}{n \log_2 n}$$

which is not summable (one can compare this with $\int_1^\infty \frac{dx}{x \log_2 x}$). Hence, almost surely, for infinitely many n's, it happens that $(S_{n+1} - S_n)/n \log_2 n \geq 1$, and hence $S_n/n \log_2 n \not\to 1$.

With only the finite first moment assumption, we need to truncate the r.v.s first. Let $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$ and $T_n = Y_1 + \cdots + Y_n$ be the partial sum of $(Y_n)_{n\geq 1}$. We still have the independence of Y_n, but they are no longer identically distributed. With the truncation we can estimate the second moment of Y_n. The following proposition show that the limits of T_n/n and S_n/n are the same,

Proposition 4.11 *$T_n/n \to \mu$ a.s. if and only if $S_n/n \to \mu$ a.s.*

Proof: We have

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathsf{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathsf{P}(|X_n| \geq n) &= \sum_{n=1}^{\infty} \int_0^\infty \mathbb{1}_{\{y \geq n\}} \mu_{|x|}\, dy \\
&= \int_0^\infty \sum_{n=1}^{\infty} \mathbb{1}_{(y \geq n)} \mu_{|x|}\, dy \\
&\leq \int_0^\infty [y] \mu_{|x|}\, dy = \mathsf{E}[|X_1|] \leq \mathsf{E}|X_1| + 1 < \infty.
\end{aligned}$$

By Borel–Cantelli lemma, we have $\mathsf{P}(\{X_n \neq Y_n, \text{ i.o.}\}) = 0$. Therefore, almost surely, there exists $n_0 = n_0(\omega)$ such that $X_n = Y_n$ for all $n \geq n_0$, and when this happens, we have $\lim_{n\to\infty} \frac{T_n}{n} = \lim_{n\to\infty} \frac{S_n}{n}$, provided either of the two limits exists. This completes the proof. □

The next lemma is technical but very useful in our discussion of the SLLN.

Proposition 4.12 $\sum_{k=1}^{\infty} \frac{1}{n^2} \operatorname{Var}(Y_k) \leq 4\mathsf{E}|X_1| < \infty$.

Proof: We start with

$$\operatorname{Var}(Y_n) \leq \mathsf{E}|Y_n|^2 = \int_0^\infty 2y \mathsf{P}(|Y_n| > y)\, dy \leq \int_0^n 2y \mathsf{P}(|X_1| > y)\, dy.$$

Treating the sum as integration w.r.t. to the counting measure on $\mathbb{N}$ and using Fubini's theorem (since everything is non-negative), we have

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{1}{n^2} \mathsf{E} {Y_n}^2 &\le \sum_{n=1}^{\infty} \frac{1}{n^2} \int_0^{\infty} \mathbb{1}_{\{y \le n\}} 2y \mathsf{P}(|X_1| > y)\, dy \\ &= \int_0^{\infty} \Big[\sum_{k=1}^{\infty} \frac{1}{n^2} \mathbb{1}_{\{y \le n\}} \Big] \cdot 2y \mathsf{P}(|X_1| > y)\, dy.\end{aligned}$$

Since $\mathsf{E}|X_1| = \int_0^{\infty} \mathsf{P}(|X_1| > y)\, dy$, it suffices to show

$$2y \sum_{n \ge y} \frac{1}{n^2} \le 4, \quad \forall y > 0. \tag{4.12}$$

Noting that

$$\frac{1}{n^2} \le \frac{1}{n(n-1)} = \frac{1}{n-1} - \frac{1}{n},$$

for $y \ge 2$, we have

$$2y \sum_{n \ge y} \frac{1}{n^2} \le 2y \sum_{n \ge y} (\frac{1}{n-1} - \frac{1}{n}) \le \frac{2y}{y-1} \le 4,$$

and for $1 < y < 2$, we have

$$2y \sum_{n \ge y} \frac{1}{n^2} = 2y \sum_{n=2}^{\infty} (\frac{1}{n-1} - \frac{1}{n}) \le 2y \le 4.$$

If $0 < y \le 1$, then

$$2y \sum_{n \ge y} \frac{1}{n^2} \le 2 \sum_{n=1}^{\infty} \frac{1}{n^2} \le 2 \cdot \frac{\pi^2}{6} \le 4.$$

These establish (4.12) and complete the proof. □

4.3.2 Etemadi's argument

The argument presented in this section was due to Etemadi (1981).

We have seen in the proof of Theorem 4.4, it is useful to first consider almost sure convergence along a subsequence (n_k), then use other means to control what happens for $n \in (n_k, n_{k+1})$. Etemadi's idea is to use monotonicity of the partial sum when the summands are non-negative to control the intermediate terms.

Let us assume first that $X_n \ge 0$ (and hence $Y_n \ge 0$), and that for some subsequence (n_k),

$$T_{n_k}/n_k \to \mu, \quad \text{a.s.}, \tag{4.13}$$

and see how far we can get. Since Y_n are non-negative, for $n \in (n_k, n_{k+1})$ we have

$$\frac{T_{n_k}}{n_{k+1}} = \frac{T_{n_k}}{n_k} \cdot \frac{n_k}{n_{k+1}} \le \frac{T_n}{n} \le \frac{T_{n_{k+1}}}{n_k} = \frac{T_{n_{k+1}}}{n_{k+1}} \cdot \frac{n_{k+1}}{n_k}. \tag{4.14}$$

Taking the limit $k \to \infty$, we have

$$\mu \cdot \liminf_{k\to\infty} \frac{n_k}{n_{k+1}} \le \liminf_{n\to\infty} \frac{T_n}{n} \le \limsup_{n\to\infty} \frac{T_n}{n} \le \mu \cdot \limsup_{k\to\infty} \frac{n_{k+1}}{n_k}. \tag{4.15}$$

Intuitively, the condition $n_{k+1}/n_k \approx 1$ will bring lim inf and lim sup of T_n/n very close, similar to the argument of the squeeze theorem. In fact, for any polynomial growth $n_k = k^p$, $p \geq 1$, the limit $\lim_{k\to\infty} n_{k+1}/n_k$ is 1, and (4.14) implies $\lim_{n\to\infty} T_n/n = \mu$. Such n_k is used in the proof of Theorem 4.4; as we will see, however, such growth cannot guarantee (4.14) when only finite first moment is assumed.

Let us go through our usual argument of combining Chebyshev's inequality and Borel–Cantelli lemma to see what is needed for the subsequence of (n_k) to guarantee (4.13). For every $\varepsilon > 0$, by Chebyshev's inequality, we have

$$\begin{aligned}\sum_{n=1}^{\infty} \mathsf{P}(|T_{n_k} - \mathsf{E}T_{n_k}| > \varepsilon n_k) &\leq \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \operatorname{Var}(T_{n_k})/n_k^2 = \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{n_k^2} \sum_{m=1}^{n_k} \operatorname{Var}(Y_m) \\ &= \frac{1}{\varepsilon^2} \sum_{m=1}^{\infty} \operatorname{Var}(Y_m) \sum_{n_k \geq m} \frac{1}{n_k^2}\end{aligned} \tag{4.16}$$

where we have used Fubini's theorem to interchange the two summations of nonnegative terms. Now, in light of Proposition 4.12, we are hoping for

$$\sum_{n_k \geq m} \frac{1}{n_k^2} \leq \frac{C}{m^2}. \tag{4.17}$$

It can be easily checked that (4.17) cannot happen if n_k only grows linearly, i.e., $n_k \sim k^p$ for some $p > 0$. For (4.17) to holds, the sum needs to be controlled by the first and largest summand, which happens only if n_k grows *exponentially*. That means $\alpha = \liminf_{k\to\infty} n_{k+1}/n_k > 1$. Using (4.14), we can close the argument by taking α arbitrarily close to 1.

First proof of Theorem 4.1: Assume first $X_n \geq 0$.

Take $n_k = [\alpha^k]$, $k \geq 1$, where $\alpha > 1$ and $[\cdot]$ denotes the integer part. Let k_0 be the smallest k such that $n_k \geq m$. Then $k_0 \geq \log_\alpha m$. Since the sum of a geometric series is dominated by the largest term, we have

$$\sum_{n_k \geq m} \frac{1}{n_k^2} \leq \sum_{k=k_0}^{\infty} \frac{1}{[\alpha^k - 1]^2} \leq C_\alpha \frac{1}{\alpha^{2k_0}} \leq \frac{C_\alpha}{m^2}.$$

for some constant $C_\alpha > 0$. By (4.16) and Borel–Cantelli lemma, we have

$$\frac{T_{n_k} - \mathsf{E}T_{n_k}}{n_k} \to 0, \quad \text{a.s.}$$

By Stolz lemma and DCT,

$$\lim_{n\to\infty} \frac{\mathsf{E}T_n}{n} = \lim_{n\to\infty} \mathsf{E}Y_n = \lim_{n\to\infty} \mathsf{E}X_1 \mathbb{1}_{\{X_1 \leq 0\}} = \mathsf{E}X_1 = \mu.$$

Recalling $n_k = [\alpha^n]$, from (4.15) we have

$$\frac{\mu}{\alpha} \leq \liminf_{n\to\infty} \frac{T_n}{n} \leq \limsup_{n\to\infty} \frac{T_n}{n} \leq \alpha\mu. \tag{4.18}$$

Since (4.18) holds for arbitrary $\alpha > 1$, by letting $\alpha \downarrow 1$ we see that $\lim_{n\to\infty} T_n/n = \mu$ a.s., and by Proposition 4.11 $\lim_{n\to\infty} S_n/n = \mu$ a.s. as desired.

For general X_n, let $X_n = X_n^+ - X_n^-$ be the decomposition into positive and negative parts, and let $S_n^\pm$ be the partial sums of $X_n^\pm$. Then

$$\lim_{n\to\infty} \frac{S_n}{n} = \lim_{n\to\infty} \frac{S_n^+ - S_n^-}{n} = \mathsf{E}X_1^+ - \mathsf{E}X_1^- = \mathsf{E}X_1, \quad \text{a.s.}$$

The proof is complete. □

As a corollary, we can also treat the case when $\mathsf{E}X_1 = \pm\infty$.

Corollary 4.13 *Let $X_1, X_2, \ldots$ be i.i.d. with $\mathsf{E}X_i^+ = \infty$ and $\mathsf{E}X_i^- < \infty$. Then $S_n/n \to \infty$ a.s.*

Proof: Let $M > 0$ and $X_i^M = X_i \wedge M$. The X_i^M are i.i.d. with $\mathsf{E}|X_i^M| < \infty$. Let S_n^M be the partial sum of X_i^M. Using Theorem 4.1 and $X_i \geq X_i^M$, we have

$$\liminf_{n\to\infty} S_n/n \geq \lim_{n\to\infty} S_n^M/n = \mathsf{E}X_1^M, \quad \text{a.s.}$$

The MCT implies $\mathsf{E}(X_1^M)^+ \uparrow \mathsf{E}X_1^+ = \infty$ as $M \uparrow \infty$, so $\mathsf{E}X_i^M = \mathsf{E}(X_i^M)^+ - \mathsf{E}(X_i^M)^- \uparrow \infty$. Hence $\liminf_{n\to\infty} S_n/n \geq \infty$ a.s., and the conclusion follows. □

4.4 Second proof of SLLN

In the section we follow Komolgorov's treatment of the SLLN.

4.4.1 Tail σ-algebras and zero-one law

In this section we will have a small detour to introduce the tail σ-algebras and Kolmogorov's zero-one law. These results are not directly used in the proof of SLLN, but they give useful intuition.

We first do some measure theory.

Let $(Y_n)_{n\in I}$ be r.v.s on $(\Omega, \mathcal{F}, \mathsf{P})$ where I is a countable index set. We introduce the "smallest σ-algebra" with respect to which all Y_n are measurable. It is defined by

$$\sigma(Y_n,\ n \in I) = \sigma(Y_n^{-1}(A),\ A \in \mathcal{B}(\mathbb{R}),\ n \in I). \tag{4.19}$$

The σ-algebra in (4.19) is generated by the semi-algebra

$$\mathcal{S} = \Big\{ \bigcap_{n\in I_1} Y_n^{-1}(A_n) : A_n \in \mathcal{B}(\mathbb{R}),\ I_1 \subset I \text{ finite} \Big\}. \tag{4.20}$$

We check that $\mathcal{S}$ is a semi-algebra using the following lemma.

Lemma 4.14 *Let I be a countable index set and $\mathcal{F}_n$, $n \in I$, be σ-algebras. Then*

$$\mathcal{S}_1 = \{ \bigcap_{n\in I_1} B_n,\ B_n \in \mathcal{F}_n,\ I_1 \subset I \text{ finite}\}$$

is a semi-algebra.

In particular, when $\mathcal{F}_n = \sigma(Y_n)$, $\mathcal{S}$ defined in (4.20) is a semi-algebra.

Proof: Let

$$A = \bigcap_{n\in I_1} A_n, \quad \tilde{A} = \bigcap_{m\in I_2} \tilde{A_m}, \tag{4.21}$$

be two sets in $\mathcal{S}_1$. We can replace I_1 and I_2 by their union $J = I_1 \cup I_2$ by adding Ω in both of the intersection in (4.21) if necessary. Note that J is the union of two finite sets I_1 and I_2 and thus is also finite. We have

$$A \cap \tilde{A} = \bigcap_{n\in J} (A_n \cap \tilde{A_n}) \in \mathcal{S}_1.$$

This shows that $\mathcal{S}_1$ is closed under intersections.

It remains to show that the complement of every set in $\mathcal{S}_1$ can be written as a finite disjoin union of sets in $\mathcal{S}_1$. This follows from

$$(A_{n_1} \cap \cdots \cap A_{n_k})^c = (A_{n_1}^c) \cup (A_{n_1} \cap A_{n_2}^c) \cup \cdots \cup (A_{n_1} \cap \cdots \cap A_{n_k}^c).$$

□

Proposition 4.15 *If $\mathcal{F}_n$, $n \in I$ and $\mathcal{G}_m$, $m \in J$ are independent where I, J are countable, then*

$$\sigma(\mathcal{F}_n,\ n \in I) \quad \text{and} \quad \sigma(\mathcal{G}_m,\ m \in J)$$

are independent.

In particular, if X_n, $n \in I$ and Y_m, $m \in J$ are all independent, then $\sigma(X_n,\ n \in I)$ and $\sigma(Y_m,\ m \in J)$ are independent.

Proof: Recall that two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ are independent if $\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)$ for every $A \in \mathcal{F}$ and $B \in \mathcal{G}$. By Lemma 4.14, $\mathcal{F}$ and $\mathcal{G}$ are generated by the semi-algebras

$$\mathcal{S}_1 = \{\bigcap_{n \in I_1} A_n,\ A_n \in \mathcal{F}_n,\ I_1 \subset I \text{ finite}\}, \quad \mathcal{S}_2 = \{\bigcap_{m \in J_1} B_m,\ B_m \in \mathcal{G}_m,\ J_1 \subset J \text{ finite}\}.$$

To show independent of $\mathcal{F}$ and $\mathcal{G}$, it suffices to show that

$$\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B), \quad \forall A \in \mathcal{S}_1,\ B \in \mathcal{S}_2. \tag{4.22}$$

Extending (4.22) to arbitrary A and B can be done by the usual appropriate set arguments and continuity of probability measures.

For $A \in \mathcal{S}_1$ and $B \in \mathcal{S}_2$, the independence of $\mathcal{F}_n$ and $\mathcal{G}_m$ implies

$$\begin{aligned}\mathsf{P}(A \cap B) &= \mathsf{P}(\bigcap_{n \in I_1} A_n \cap \bigcap_{m \in J_1} B_m) = \prod_{n \in I_1} \mathsf{P}(A_n) \prod_{m \in J_1} \mathsf{P}(B_m) \\ &= \mathsf{P}(\bigcap_{n \in I_1} A_n)\mathsf{P}(\bigcap_{m \in J_1} B_m) = \mathsf{P}(A)\mathsf{P}(B).\end{aligned}$$

This establishes (4.22) and completes the proof. □

Let $(X_n)_{n \geq 1}$ be independent on $(\Omega, \mathcal{F}, \mathsf{P})$. Let us introduce

$$\mathcal{F}_n = \sigma(X_1, \ldots, X_n), \quad \mathcal{F}_{>n} = \sigma(X_{n+1}, X_{n+2}, \cdots) = \sigma(X_m,\ m > n).$$

The σ-algebra $\mathcal{F}_n$, containing information before time n, should be regarded as the "past", while $\mathcal{F}_{>n}$ should be regarded as the "future". By Proposition 4.15, $\mathcal{F}_n$ and $\mathcal{F}_{>n}$ are independent for all $n \geq 1$, which agrees with our intuition.

Definition 4.1 (Tail σ-algebra) *The tail σ-algebra is $\mathcal{T} = \bigcap_{n=0}^{\infty} \mathcal{F}_{>n}$.*

The tail σ-algebra should be regarded as the "remote future", as it does not concern anything happening in finite time.

Example 4.8 (Examples of sets in $\mathcal{T}$) 1. $\{\lim_{n \to \infty} S_n \text{ exists}\} \in \mathcal{T}$.

2. $\{\limsup_{n \to \infty} \frac{S_n}{n} > x\} \in \mathcal{T}$ for any x.

We will only verify for the first set. Since $\lim_{m \to \infty} S_m$ exists if and only if $\lim_{m \to \infty}(S_{n+m} - S_n)$ exists, for all $n \geq 0$, we have

$$\{\lim_{m \to \infty}(S_{n+m} - S_n) \text{ exists}\} = \{\lim_{m \to \infty}(X_{n+1} + X_{n+2} + \cdots + X_{n+m}) \text{ exists}\} \in \mathcal{F}_{>n}.$$

Hence, $\{\lim_{m \to \infty} S_m\} \in \bigcap_{n \geq 0} \mathcal{F}_{>n} = \mathcal{T}$.

An important observation of Kolmogorov was that the tail σ-algebra is trivial, and thus it makes sense to study the almost sure convergence of random series.

Theorem 4.16 (Kolmogorov's $0-1$ law) *If $X_1, X_2, \dots$ are independent and $A \in \mathcal{T}$ then $\mathsf{P}(A) = 0$ or 1.*

Proof: For all $m \geq n$, since $\mathcal{F}_n$ and $\mathcal{F}_{>m}$ are independent, so we have $\mathcal{F}_n \perp \bigcap_{m \geq n} \mathcal{F}_{>m}$, i.e., for all $n \geq 1$, $\mathcal{F}_n \perp \mathcal{T}$. Let $\mathcal{F}_\infty = \sigma(X_1, X_2, \dots)$. For all $n \geq 1$, since $\mathcal{T}$ and $\sigma(X_n)$ are independent, so we have $\mathcal{T} \perp \sigma(X_1, X_2, \dots) = \mathcal{F}_\infty$, and thus $\mathcal{T} = \bigcap_{n=0}^{\infty} \mathcal{F}_{>n} \subset \mathcal{F}_\infty$.It follows that $\mathcal{T} \perp \mathcal{T}$. For all $A \in \mathcal{T}$, A is independent to itself, i.e.,

$$\mathsf{P}(A \cap A) = \mathsf{P}(A) = \mathsf{P}(A)^2$$

it follows that $\mathsf{P}(A) = 1$ or 0. □

4.4.2 Kolmogorov's proof

Proposition 4.17 (Kolmogorov's maximal inequality) *Let $X_1, \dots, X_n$ be independent with $\mathsf{E}X_i = 0$ and $\mathrm{Var}(X_i) < \infty$. Then*

$$\mathsf{P}(\max_{1 \leq k \leq n} |S_k| \geq x) \leq \frac{\mathrm{Var}(S_n)}{x^2}. \tag{4.23}$$

Remark 4.9

1. Note that Chebyshev's inequality can only give (4.23) *without* the maximum inside the probability, so Proposition 4.17 is highly non-trivial.
2. In fact, $(S_n)_{n \geq 1}$ forms a martingale and Proposition 4.17 is a special case of the *Doob's maximal inequality* for martingales. In the proof, we will also use the idea of "stopping time", which is common in martingale analysis

Proof: Let $T(\omega) = \min\{k : k \geq 1, |S_k(\omega)| \geq x\}$ to be the first time that $|S_k|$ exceeds x. More precisely,

$$\{T(\omega) = k\} = \{|S_1|, |S_2|, \dots, |S_{k-1}| < x, |S_k| \geq x\},$$

for $k \in \{1, \dots, n\}$ and $T = \infty$ if the event in (4.23) does not happen. Clearly, $\{T = k\} \in \sigma(X_1, \dots, X_k) =: \mathcal{F}_k$ for $k\{1, \dots, n\}$.

We have

$$\mathsf{E}S_n^2 \geq \sum_{k=1}^{n} \mathsf{E}S_n^2 \mathbb{1}_{\{T=k\}} = \sum_{k=1}^{n} \mathsf{E}S_n^2 \mathbb{1}_{\{T=k\}}.$$

For $k \leq n$, we have

$$\begin{aligned}
\mathsf{E}S_n^2 \mathbb{1}_{\{T=k\}} &= \mathsf{E}\big(S_k + (S_n - S_k)\big)^2 \mathbb{1}_{\{T=k\}} \\
&\geq \mathsf{E}S_k^2 \mathbb{1}_{\{T=k\}} + 2\mathsf{E}S_k \mathbb{1}_{\{T=k\}} \cdot (S_n - S_k) + \mathsf{E}(S_n - S_k)^2 \mathbb{1}_{\{T=k\}} \\
&\geq \mathsf{E}S_k^2 \mathbb{1}_{\{T=k\}} + 2\mathsf{E}S_k \mathbb{1}_{\{T=k\}} \cdot \mathsf{E}(S_n - S_k) \\
&= \mathsf{E}S_k^2 \mathbb{1}_{\{T=k\}} \geq x^2 \mathsf{P}(T = k).
\end{aligned}$$

Here, in the third line we use the independence of $S_k \mathbb{1}_{\{T=k\}} \in \sigma(X_1, \dots, X_k)$ and $S_n - S_k \in \sigma(X_{k+1}, \dots)$. Summing over $k \in \{1, \dots, n\}$, we have

$$\mathsf{E}S_n^2 \geq x^2 \mathsf{P}(T \leq n),$$

and this gives (4.23). □

Proposition 4.18 (Kolmogorov's one-series theorem) *Let $X_1, X_2, \dots$ be independent with $\mathsf{E}X_n = 0$. If*

$$\sum_{n=1}^{\infty} \mathsf{E}{X_n}^2 < \infty, \tag{4.24}$$

then $\sum_{n=1}^{\infty} X_n(\omega)$ converges a.s.

Proof: For every $\varepsilon > 0$, by Proposition 4.17, for all $M, N > 0$, we have

$$\mathsf{P}(\max_{M \le n \le N} |S_n - S_m| \ge \varepsilon) \le \frac{\mathrm{Var}(S_n - S_m)}{\varepsilon^2}.$$

Letting $N \to \infty$, by MCT, we have

$$\mathsf{P}(u_M := \sup_{n \ge M} |S_n - S_M| \ge \varepsilon) \le \frac{\sum_{n=M}^{\infty} \mathsf{E}X_n^2}{\varepsilon^2},$$

which goes to 0 by (4.24). Let $\tilde{u}_M = \sup_{m,n \ge M} |S_n - S_m|$. Then for every $\varepsilon > 0$,

$$\lim_{m \to \infty} \mathsf{P}(\tilde{u}_M \ge \varepsilon) \le 2 \lim_{m \to \infty} \mathsf{P}(u_M \ge \frac{\varepsilon}{2}) = 0.$$

Hence, for every $\varepsilon > 0$, we have $\mathsf{P}(\lim_{n \to \infty} \tilde{u}_M \ge \varepsilon) = 0$ since $\tilde{u}_M$ is decreasing. Therefore, $\tilde{u}_M \downarrow 0$ as $M \to \infty$ almost surely, and consequently $\lim_{n \to \infty} S_n$ exists by Cauchy criterion. □

Example 4.10 It is well known that alternating harmonic series $\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$ converges only conditionally. What if we put random $\pm$ signs before the harmonic series?

To model it, let $(\xi_n)_{n \ge 1}$ be i.i.d. with $\mathsf{P}\{\xi_n = \pm 1\} = \frac{1}{2}$. By Proposition 4.18, since $\sum_{n=1}^{\infty} \mathsf{E}\frac{|\xi_n|^2}{n^2} = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$, we have $\sum_{n=1}^{\infty} \frac{\xi_n}{n}$ converges a.s. In fact, the conclusion holds for $\sum_{n=1}^{\infty} \xi_n / n^p$ with $p \in (1/2, 1]$.

The next result immediately follows from Proposition 4.18 and treat the case of non-centered r.v.s.

Proposition 4.19 (Kolmogorov's two-series theorem) *Let $X_1, X_2, \dots$ are independent with $\mathsf{E}|X_n| < \infty$. If*

$$\sum_{n=1}^{\infty} \mathsf{E}X_n \text{ exists}, \quad \sum_{n=1}^{\infty} \mathsf{E}X_n^2 < \infty,$$

then $\sum_{n=1}^{\infty} X_n(\omega)$ converges a.s.

For the almost sure convergence of random series, the final theorem provides necessary and sufficient conditions.

Theorem 4.20 (Kolmogorov's three-series theorem) *Let $A > 0$. Let $X_1, X_2, \dots$ be independent and $Y_i = X_i \mathbb{1}_{(|X_i| \le A)}$. Then, $\sum_{n=1}^{\infty} X_n$ converges a.s. if and only if all of the following conditions hold:*

$$\sum_{n=1}^{\infty} \mathsf{P}(|X_n| \ge A) < \infty, \tag{4.25a}$$

$$\sum_{n=1}^{\infty} \mathsf{E}Y_n \text{ exists}, \tag{4.25b}$$

$$\sum_{n=1}^{\infty} \mathrm{Var}(Y_n) < \infty. \tag{4.25c}$$

Proof: **The "if" part.** By Borel-Cantelli lemma, (4.25a) implies that $\mathsf{P}(\{|X_n| > A,\ i.o.\}) = 0$. On the event $\{|X_n| > A,\ i.o.\}^c$, there exists $n_0 = n_0(\omega)$ such that $Y_n = X_n$ for every $n > n_0$, and hence $\sum_{n=1}^\infty X_n$ converges if and only if $\sum_{n=1}^\infty Y_n$ does; on the other hand, the latter random series converges a.s. by Proposition 4.19, (4.25b) and (4.25c).

The "only if" part. Assume now that $\sum_{n=1}^\infty X_n$ converges a.s. If (4.25a) fails, by the second Borel-Cantelli lemma (Theorem 4.9), we have $\mathsf{P}(\{|X_n| \geq A,\ \text{i.o.}\}) = 1$. But on $\{|X_n| \geq A,\ i.o.\}^c$, the series $\sum_{n=1}^\infty X_n(\omega)$ cannot converge as the Cauchy criterion is violated. Hence, (4.25a) must holds. Then, as has been proven, (4.25a) implies that $\sum_{n=1}^\infty Y_n$ also converges a.s.

Now we have $\big|Y_n - \mathsf{E}Y_n\big| \leq 2A$. By Lemma 4.21 proven below, we have

$$\mathsf{P}\Big(\max_{M\leq n\leq N}\Big|\sum_{k=M}^n Y_k\Big| \leq 1\Big) \leq \frac{(2A+1)^2}{\sum_{n=M}^{N-1}\mathrm{Var}(Y_n)}. \tag{4.26}$$

If (4.25c) fails and $\sum_{n=1}^\infty \mathrm{Var}(Y_n) = \infty$, then (4.26) implies

$$\mathsf{P}\Big(\sup_{n\geq M}\Big|\sum_{k=M}^n Y_k\Big| \leq 1\Big) = 0, \quad \forall M \geq 1,$$

which implies that $\sum_{n=1}^\infty Y_n$ diverges a.s. and leads to a contradiction. Thence, (4.25c) also holds.

Finally we need to show (4.25c). By Proposition 4.18, (4.25c) implies that $\sum_{n=1}^\infty \big(Y_n - \mathsf{E}Y_n\big)$ converges a.s., and hence

$$\sum_{n=1}^\infty \mathsf{E}Y_n = \sum_{n=1}^\infty Y_n - \sum_{n=1}^\infty \big(Y_n - \mathsf{E}Y_n\big)$$

also converges. This completes the proof of the "only if" part. □

For (4.26) we use the following results, which says if a random walk have large variance in each step, then it is unlikely that it will stay in a small region; this is the opposite direction of Proposition 4.17.

Lemma 4.21 *Let $Z_1, \ldots, Z_n$ be independent with $\mathsf{E}Z_i = 0$ and $|Z_i| \leq L$ for some $L > 0$. Let $S_n = Z_1 + \cdots + Z_n$. Then for every $\varepsilon > 0$,*

$$\mathsf{P}\big(\max_{1\leq k\leq n}|S_k| \leq \varepsilon\big) \leq \frac{(L+\varepsilon)^2}{\mathrm{Var}(S_n)}.$$

Proof: Let

$$T = \min\{k : |S_k| > \varepsilon\} \in \{1, 2, \ldots, n, \infty\},$$

with the convention $T = \infty$ if $\max_{1\leq k\leq n}|S_k| \leq \varepsilon$. We note that $\{T = k\}, \{T \geq k+1\} \in \sigma(X_1, \ldots, X_k)$ for every $1 \leq k \leq n$.

Let $a_k = \mathsf{E}S_k^2 \mathbb{1}_{\{T\geq k+1\}}$, $0 \leq k \leq n$. Since $\{T \geq k+2\} = \{T \geq k+1\} \setminus \{T = k+1\}$, we have

$$a_{k+1} = \mathsf{E}S_{k+1}^2 \mathbb{1}_{\{T\geq k+1\}} - \mathsf{E}S_{k+1}^2 \mathbb{1}_{\{T=k+1\}}. \tag{4.27}$$

Using independence of X_{k+1} and $S_k, \mathbb{1}_{\{T\geq k+1\}}$ (both measurable w.r.t. $\sigma(X_1, \ldots, X_k)$), the first term in (4.27) is bounded below by

$$\begin{aligned}\mathsf{E}S_{k+1}^2 \mathbb{1}_{\{T\geq k+1\}} &= \mathsf{E}S_k^2 \mathbb{1}_{\{T\geq k+1\}} + 2\mathsf{E}X_{k+1}\cdot \mathsf{E}S_k \mathbb{1}_{\{T\geq k+1\}} + \mathsf{E}X_{k+1}^2 \cdot \mathsf{P}(T \geq k+1)\\ &\geq a_k + \mathsf{E}X_{k+1}^2 \cdot \mathsf{P}(T \geq k+1).\end{aligned} \tag{4.28}$$

For the second term in (4.27), since

$$\{T = k+1\} = \{|S_1| \leq \varepsilon, \ldots, |S_k| \leq \varepsilon,\ |S_{k+1}| > \varepsilon\},$$

when $T = k+1$, we have $|S_{k+1}| \le |S_k| + |X_{k+1}| \le L + \varepsilon$, and hence

$$\mathsf{E}S_{k+1}^2 \mathbb{1}_{\{T=k+1\}} \le (L+\varepsilon)^2 \mathsf{P}(T = k+1). \tag{4.29}$$

Combining (4.28) and (4.29), we have

$$(a_{k+1} - a_k) + (L+\varepsilon)^2 \mathsf{P}(T = k+1) \ge \mathsf{E}X_{k+1}^2 \cdot \mathsf{P}(T \ge k+1), \quad k = 0, \dots, n-1.$$

Summing the above inequality over k, and using that $\mathsf{P}(T \ge k+1) \ge \mathsf{P}(T = \infty)$, we have

$$\mathrm{Var}(S_n) \cdot \mathsf{P}(T = \infty) \le (L+\varepsilon)^2 \mathsf{P}(T \le n) + \mathsf{E}S_n^2 \mathbb{1}_{\{T \ge n+1\}}.$$

Noting that when $T \ge n+1$, $S_n^2 \le \varepsilon^2$, the last term in the last display is bounded by $\varepsilon^2 \cdot \mathsf{P}(T \ge n+1)$, so we obtain

$$\mathrm{Var}(S_n) \cdot \mathsf{P}(T = \infty) \le (L+\varepsilon)^2.$$

The desired inequality follows. □

Proposition 4.22 (Kronecker's lemma) *If $a_n \uparrow \infty$ and $\sum_{n=1}^{\infty} x_n/a_n$ converges, then*

$$a_n^{-1} \sum_{m=1}^{n} x_m \to 0$$

Proof: Let $S_n := \sum_{m=1}^{n} x_m/a_m$ and $S := \lim_{n\to\infty} S_n$ Using Summation by parts (a.k.a. Abel's transformation) we have

$$\frac{1}{a_n} \sum_{m=1}^{n} a_m (S_m - S_{m-1}) = S_n - \sum_{m=1}^{n} \frac{(a_m - a_{m-1})}{a_n} S_{m-1}$$

By using Generalized Stolz's Lemma (Lemma 4.23) below with $\rho_{n,k} = (a_k - a_{k-1})/a_n$, we have

$$\lim_{n\to\infty} \frac{1}{a_n} \sum_{m=1}^{n} a_m (S_m - S_{m-1}) = \lim_{n\to\infty} S_n - \lim_{n\to\infty} \sum_{m=1}^{n} \frac{(a_m - a_{m-1})}{a_n} S_{m-1} = S - S = 0.$$

□

Lemma 4.23 (Generalized Stolz) *Let $\rho_{n,k} \ge 0$, $1 \le k \le n$, be such that*

$$\lim_{n\to\infty} \sum_{k=K}^{n} \rho_{n,k} = 1, \quad \lim_{n\to\infty} \sum_{k=1}^{K-1} \rho_{n,k} = 0,$$

for every $K > 0$. Then

$$\lim_{n\to\infty} \sum_{k=1}^{n} \rho_{n,k} x_k = \lim_{n\to\infty} x_n$$

provided that second limit exists.

In particular, when $\rho_{n,k} = 1/n$, this is the Stolz's Lemma.

Proof: Let $L = \lim_{n\to\infty} x_n$. For simplicity we assume $|L| < \infty$, but the lemma also holds for $L = \infty$ or $-\infty$.

For every $\varepsilon > 0$, there exists $K > 0$ such that $x_k \geq L - \varepsilon$ for $k \geq K$. Hence, we have

$$\sum_{k=1}^{n} \rho_{n,k} x_k \geq -(\sup_m |x_m|) \sum_{k=1}^{K-1} \rho_{n,k} + (L-\varepsilon) \sum_{k=K}^{n} \rho_{n,k}.$$

Taking $n \to \infty$ and using the assumption on $\rho_{n,k}$ we obtain $\liminf_{n\to\infty} \sum_{k=1}^{n} \rho_{n,k} x_k \geq L - \varepsilon$, and since $\varepsilon > 0$ is arbitrary we have $\liminf_{n\to\infty} \sum_{k=1}^{n} \rho_{n,k} x_k \geq L$. Similarly, we can show $\limsup_{n\to\infty} \sum_{k=1}^{n} \rho_{n,k} x_k \leq L$. The conclusion follows. □

We can now give the proof of Theorem 4.1.

Proof of Theorem 4.1: By Proposition 4.10, it is equivalent to show that $(T_n - \mu_n)/n \to 0$, a.s., where $T_n = \sum_{k=1}^{n} Y_n$ and $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$. By Proposition 4.22, it suffices to show that $\sum_{n=1}^{\infty} \frac{Y_n}{n}$ converges a.s., and this follows from Proposition 4.18 and Proposition 4.12. □

The above proof also gives us a way to estimate the rate of convergence, as the next result shows.

Proposition 4.24 *Let $X_1, X_2, \ldots$ are i.i.d. r.v.s with $\mathsf{E}X_i = 0$ and $\mathsf{E}X_i^2 = \sigma^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then, for every $\varepsilon > 0$,*

$$\frac{S_n}{\sqrt{n}(\log n)^{1/2+\varepsilon}} \to 0 \quad a.s.$$

Proof: Let $a_n = n^{\frac{1}{2}} (\log n)^{\frac{1}{2}+\varepsilon}$, $n \geq 2$. We have

$$\sum_{n=2}^{\infty} \mathrm{Var}(\frac{x_n}{a_n}) = \sigma^2 \sum_{n=2}^{\infty} \frac{1}{a_n^2} = \sigma^2 \sum_{n=2}^{\infty} \frac{1}{n(\log n)^{1+2\varepsilon}} < \infty.$$

By Proposition 4.18, the series $\sum_{n=1}^{\infty} \frac{x_n}{a_n}$ converges a.s., and it follows from Proposition 4.22 that $\frac{1}{a_n} \sum_{k=1}^{n} x_n \to 0$ a.s. □

To conclude our discussion of the SLLN, we cite a result given by Feller (see also [Dur19, Thm 2.5.13]), which says when the r.v.s are not integrable, it is not possible to obtain something like SLLN even by changing the rate.

Proposition 4.25 *Let $X_1, X_2, \ldots$ are i.i.d. r.v.s with $\mathsf{E}|X_1| = \infty$. Let a_n be a sequence of positive numbers with a_n/n increasing. Then $\limsup_{n\to\infty} |S_n|/a_n = 0$ or ∞ according as $\sum_{n=1}^{\infty} \mathsf{P}(|X_1| \geq a_n) < \infty$ or $= \infty$.*

5 Weak convergence and central limit theorem

Let $\mathcal{P}(\mathbb{R})$ denote the set of all probability measures on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$. To goal of this section is to study the weak convergence of probability measure. Before starting the subject, we first mention the *total variation distance*, a very natural metric on $\mathcal{P}(\mathbb{R})$.

Definition 5.1 *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$. The total variation distance between μ and ν is given by*

$$\|\mu - \nu\|_{TV} := 2 \cdot \sup_{A \in \mathcal{B}(\mathbb{R})} |\mu(A) - \nu(A)| \in [0, 2]. \tag{5.1}$$

It is not hard to check that (5.1) defines a metric: (a) it is positive definite: $\|\mu-\nu\|_{TV} \geq 0$, with $\|\mu-\nu\|_{TV} = 0$ if and only if $\mu(A) = \nu(A)$ for all A, which means $\mu = \nu$; (b) it is symmetric: $\|\mu - v\|_{TV} = \|\nu - \mu\|_{TV}$; (c) it satisfies the triangle inequality by taking supremum of $|\mu(A)-\nu(A)| \leq |\mu(A) - \lambda(A)| + |\lambda(A) - \mu(A)|$.

Example 5.1 Let μ and ν are mutually singular (see Definition 1.13) and let $A \in \mathcal{B}(\mathbb{R})$ be such that $\mu(A) = 0$ and $\nu(A) = 1$. Then $|\mu(A) - \nu(A)| = 1$, and hence $\|\mu - \nu\|_{TV} = 2$ since 2 is the maximum for total variation distance.

Example 5.2 Let μ and ν are absolutely continuous and $f, g \in L^1(\mathbb{R})$ be their densities. Then

$$\|\mu - \nu\|_{TV} = \|f - g\|_{L^1(\mathbb{R})},$$

where the supremum in (5.1) is achieved by $A = \{x : f(x) \geq g(x)\}$ in (5.1). More generally, let F and G be the c.d.f.s of μ and ν; then

$$\|\mu - \nu\|_{TV} = \sup_{t_1<\cdots<t_n} \sum_{i=1}^{n-1} \Big|(F - G)(t_{i+1}) - (F - G)(t_i)\Big|,$$

which is the *total variation* of $F - G$ over $\mathbb{R}$.

It also immediately follows from (5.1) that convergence in total variational distance implies convergence of the set function.

Proposition 5.1 *If $\|\mu_n - \mu\|_{TV} \to 0$, then*

$$\lim_{n\to\infty} \mu_n(A) = \mu(A), \quad \forall A \in \mathcal{B}(\mathbb{R}).$$

In particular, by taking $A = (-\infty, x]$,

$$\lim_{n\to\infty} F_{\mu_n}(x) = F_\mu(x), \quad \forall x. \tag{5.2}$$

However, as can be seen from the next two examples, the convergence in total variation distance is too restrictive.

Example 5.3 Let $\mu_n = \delta_{1/n}$ and $\mu = \delta_0$. We expect $\mu_n \to \mu$ since $1/n \to 0$. On the other hand, $\|\mu_n - \mu\|_{TV} = 2$ so no convergence in total variation distance.

Example 5.4 Let $\mu_n = \text{Unif}\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\}$. We expect $\mu_n \to \mu = \text{Unif}[0, 1]$, which will justify the standard procedure to generate Unif$[0, 1]$ r.v.s on computers mentioned at the end of Section 3.2.1. On the other hand, we have $\mu_n(\mathbb{Q}) = 1$ while $\mu(\mathbb{Q}) = 0$, so μ_n and μ are mutually singular and thus $\|\mu_n - \mu\|_{TV} \equiv 2$.

Weak convergence is one way to give a more relaxed mode of convergence for measures.

5.1 Definition of weak convergence

Throughout this section, μ_n, μ will be probability measures and F_n, F be their c.d.f.s.

Definition 5.2 (Weak convergence) *We say that μ_n converge to μ weakly, written $\mu_n \Rightarrow \mu$, if*

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \textit{almost every } x. \tag{5.3}$$

With abuse of notation we also write $F_n \Rightarrow F$ for (5.3).

Let X_n, X be r.v.s. We say that X_n converge to X in distribution/law, or weakly, written $X_n \Rightarrow X$, if $\mu_{X_n} \Rightarrow \mu_X$.

By directly computing the c.d.f's, one can verify that

$$\delta_{1/n} \Rightarrow \delta_0, \quad \text{Unif}\Big\{0, \frac{1}{n}, \dots, \frac{n-1}{n}\Big\} \Rightarrow \text{Unif}[0,1].$$

(5.3) is *weaker* than (5.2) since it allows an exceptional set of measure zero. The first question we ask is the uniqueness of such limit.

Proposition 5.2 *If $F_n \Rightarrow F$ and $F_n \Rightarrow G$, then $F = G$.*

Proof: Implicitly here, we require that both F and G are c.d.f.s, that is, right continuous and increasing functions. We know that such functions are determined by its value on a dense set. By the assumption, there exist measure zero sets N_1 and N_2 such that $F_n(x) \to F(x)$ for $x \notin N_1$ and $F_n(x) \to G(x)$ for $x \notin N_2$, and hence $F(x) = G(x)$ for $x \notin N_1 \cup N_2$. The measure of $N_1 \cup N_2$ is 0, and the complement of any measure zero set is dense, so $F = G$ as desired. □

In fact, we know precisely what is the exceptional set in (5.3).

Proposition 5.3 (Also an alternative definition for $F_n \Rightarrow F$) *The condition (5.3) is equivalent to*

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \forall \text{ continuous point } x \text{ of } F. \tag{5.4}$$

Proof: (5.3) follows from (5.4) easily since discontinuous point of c.d.f. is at most countable, and a countable set has measure 0.

For the other direction, let x_0 be a continuous point of F. For every $\varepsilon > 0$, there exists $\delta > 0$ such that $|F(x) - F(x_0)| < \varepsilon$ for $|x - x_0| < \delta$. Then, we can find $y_1 \in (x_0 - \delta, x_0)$ and $y_2 \in (x_0, x_0 + \delta)$ such that y_1 and y_2 are not in the exceptional set in (5.3). We also have, by the monotonicity of F_n,

$$\begin{aligned} F(x_0) - \varepsilon < F(y_1) = \lim_{n\to\infty} F_n(y_1) &\le \liminf_{n\to\infty} F_n(x_0) \\ &\le \limsup_{n\to\infty} F_n(x_0) \le \lim_{n\to\infty} F_n(y_2) = F(y_2) < F(x_0) + \varepsilon. \end{aligned} \tag{5.5}$$

By letting $\varepsilon \downarrow 0$, we obtain $\lim_{n\to\infty} F_n(x_0) = F(x_0)$. □

The real power of weak convergence is the extraction of convergence subsequence (so-called *sequential pre-compactness*) under minimum assumption.

Definition 5.3 *We say that $(\mu_n)_{n\in I}$ is tight, if for every $\varepsilon > 0$, there exists a compact set $K = K_\varepsilon$ such that*

$$\mu_n(K^c) \le \varepsilon, \quad \forall n \in I. \tag{5.6}$$

Compact sets in $\mathbb{R}$ are bounded closed sets, and the above condition can be reformulated as the existence of $L > 0$ such that

$$\mu_n[-L, L]^c < \varepsilon, \quad \forall n \in I. \tag{5.7}$$

Note that using (5.6), the notion of tightness can be generalized to arbitrary metric spaces. [We will take up this again in XXX.]

Theorem 5.4 (Helly selection theorem) *If $(\mu_n)_{n\ge1}$ is tight, then there exists a subsequence $(\mu_{n_k})_{k\ge1}$ and a probability measure μ_∞ such that $\mu_{n_k} \Rightarrow \mu_\infty$.*

Proof: For every $q \in \mathbb{Q}$, the sequence $\big(F_n(q)\big)_{n\geq 1} \subset [0,1]$ has a convergent subsequence. Such subsequence depends on $q \in \mathbb{Q}$, but since $\mathbb{Q}$ is countable, by a standard diagonal sequence argument, there exists a common subsequence $(F_{n_k})_{k\geq 1}$ such that

$$\lim_{k\to\infty} F_{n_k}(q) =: \bar{F}(q), \quad \forall q \in \mathbb{Q}.$$

The limiting function $\bar{F}$ is increasing on $\mathbb{Q}$, so we can use it to define a right continuous, increasing function by

$$F(x) := \inf\{\bar{F}(q) : q \geq x\}.$$

We first show that $\lim_{k\to\infty} F_{n_k}(x) = F(x)$ for every continuous point of F. Let x_0 be a continuous point of F. Then for every $\varepsilon > 0$, there exists $\delta > 0$ such that $|\bar{F}(q) - F(x_0)| < \varepsilon$ for all $|q - x_0| < \delta$. Pick any $q_1 \in (x_0 - \delta, x_0) \cap \mathbb{Q}$ and $q_2 \in (x_0, x_0 + \delta)$. Similar to (5.5), we have

$$F(x_0) - \varepsilon < \bar{F}(q_1) \leq \liminf_{k\to\infty} F(x_0) \leq \limsup_{k\to\infty} F(x_0) \leq \bar{F}(q_2) < F(x_0) + \varepsilon.$$

By sending $\varepsilon \to 0$ we prove the desired limit.

Second, we need to show that F is a c.d.f. It suffices to verify $\bar{F}(-\infty) = 0$ and $\bar{F}(\infty) = 1$, for which we will use tightness. Indeed, since $(\mu_n)_{n\geq 1}$ is tight, for every ε, there exists $L \in Q$ such that $F_n(L) - F_n(-L) \geq 1 - \varepsilon$ (see (5.7)). Letting $n = n_k \to \infty$, we have

$$\bar{F}(L) - \bar{F}(-L) \geq 1 - \varepsilon.$$

Therefore,

$$\lim_{q\to-\infty} \bar{F}(q) = -\infty, \quad \lim_{q\to\infty} \bar{F}(q) = \infty,$$

and this show that F is a c.d.f. □

Remark 5.5 Tightness is necessary to prevent the "escape of mass to infinity". Consider $\mu_n = \text{Unif}[n, n+1]$, then $F_n(x) \to 0$ for every x, but the limiting function is 0, and cannot be a distribution function.

Proposition 5.5 (necessity of tightness) *If $\mu_n \Rightarrow \mu$, then $(\mu_n)_{n\geq 1}$ is tight.*

Proof: This will be an easy consequence of Theorem 5.7, but we give another proof here by analyzing the c.d.f.s.

The first observation is that a single probability measure is tight. Therefore, for every $\varepsilon > 0$, there exists $L > 0$ such that $\pm L$ are both continuous points of F and $F(-L) \leq \varepsilon/4$, $F(L) \geq 1 - \varepsilon/4$. Since $\lim_{n\to\infty} F_n(\pm L) = F(\pm L)$, there exists n_0 such that $F_n(-L) \leq \varepsilon/2$ and $F_n(L) \geq 1 - \varepsilon/2$. Also, for each $1 \leq k \leq n_0$, there exist L_k such that $F_k(-L_k) \leq \varepsilon/2$, $F_k(L_k) \geq 1 - \varepsilon/2$. Let $\bar{L} = \max\{L, L_1, \ldots, L_{n_0}\}$. Then $\mu_n[-\bar{L}, \bar{L}]^c \leq \varepsilon$ for all $n \geq 1$, and this proves the tightness. □

Without the tightness, the convergence in the first part in the proof of Theorem 5.4 is sometimes called *vague convergence*.

Definition 5.4 *We say that $F_n \to F$ vaguely if $\lim_{n\to\infty} F_n(x) = F_n(x)$ for almost every x. Here, F may only define a sub-probability measure.*

We can reformulate the previous results.

Proposition 5.6 *Every sequence of probability measures $(\mu_n)_{n\geq 1}$ has a vaguely convergent subsequence. The limit is a probability measure if and only if the subsequence is tight.*

5.2 Other characterizations of weak convergence

The concept of weak convergence can be generalized to arbitrary metric spaces. An excellent reference in this account is [Bil99]. We only present a topological way to define weak convergence, which is applicable to more general settings.

Theorem 5.7 *Let μ_n, μ be probability measures on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$. The following statements are equivalent.*

1. *$\mu_n \Rightarrow \mu$, that is, (5.4) holds.*
2. *For every bounded continuous function g,*
$$\lim_{n\to\infty} \int g \, d\mu_n = \int g \, d\mu. \tag{5.8}$$
3. *For every open set G,*
$$\liminf_{n\to\infty} \mu_n(G) \geq \mu(G). \tag{5.9}$$
4. *For every closed set K,*
$$\limsup_{n\to\infty} \mu_n(K) \leq \mu(K). \tag{5.10}$$
5. *For every A with $\mu(\partial A) = 0$, $\lim_{n\to\infty} \mu_n(A) = \mu(A)$.*

Proof: **From Item 1 to Item 2.** We will start from simplest forms of g.

First, consider
$$g(x) = \begin{cases} 0, & x < a \\ (b-a), & x > b \\ \text{linear interpolation}, & x \in [a,b]. \end{cases} \tag{5.11}$$

Then $g(x) = \int_{-\infty}^{x} \mathbb{1}_{(a,b)}(y)\, dy$. Using Fubini's theorem we have
$$\begin{aligned} \int g \, d\mu_n &= \int \Big[\int_{-\infty}^{\infty} \mathbb{1}_{(a,b)}(y) \mathbb{1}_{\{y<x\}} \, dy \Big] \, d\mu_n(x) \\ &= \int_{-\infty}^{\infty} \mathbb{1}_{(a,b)}(y)\, dy \cdot \int_{-\infty}^{\infty} \mathbb{1}_{\{y<x\}} \, d\mu_n(x) \\ &= \int_a^b [1 - F_n(y)] \, dy \\ &\to \int_a^b [1 - F(y)] \, dy = \int g \, d\mu, \end{aligned}$$
where the last line is due to $F_n(y) \to F(y)$ for a.e. y and BCT on the finite interval $[a,b]$.

Second, we consider g being a piecewise linear function with compact support. Then such g can be written as a linear combination of functions in the form (5.11), so (5.8) holds.

Third, let K be a compact set and consider
$$g \in \mathcal{C}_K = \{g : \text{ continuous}, \operatorname{supp} g \subset K\}.$$
Then there exist $g_m \in \mathcal{C}_K$ piecewise linear with compact support such that $g_m \to g$ uniformly on K, by uniform continuity of g. To estimate the difference of the terms in (5.8), we replace g by g_m, with error controlled by the triangle inequality. We have for every m,
$$\limsup_{n\to\infty} \Big| \int g \, d\mu_n - \int g \, d\mu \Big| \leq \limsup_{n\to\infty} \Big| \int g_m \, d\mu_n - \int g_m \, d\mu \Big| + \int |g_m - g| \, (d\mu_m + d\mu) \leq 2 \cdot \sup |g_m - g|. \tag{5.12}$$

Letting $g_m \to g$ we see that the LHS must be zero.

Finally, let g be bounded and continuous. For every compact set K, there exists $g_K \in \mathcal{C}_K$ such that g_K has compact support, $g_K \equiv g$ on K and $\sup|g_K| \le \sup|g|$. By tightness, for every $\varepsilon > 0$, there exists a compact set K such that $\mu_n(K^c), \mu(K^c) < \varepsilon$. Similar to (5.12), and using that $g_K = g$ on K, we have

$$\limsup_{n\to\infty} \Big| \int g\, d\mu_n - \int g\, d\mu \Big| \le \sup|g_K - g| \cdot \limsup_{n\to\infty} \big(\mu(K^c) + \mu_n(K^c)\big) \le 4 \sup|g| \varepsilon.$$

Letting $\varepsilon \downarrow 0$, the limit on the LHS is 0. This proves (5.8).

From Item 2 to Item 3. For every G open, there exists $g_m \ge 0$, bounded and continuous such that $g_n \uparrow \mathbb{1}_G$. For every g_m, by (5.8) we have

$$\liminf_{n\to\infty} \mu_n(G) \ge \liminf_{n\to\infty} \int g_m\, d\mu_n = \int g_m\, d\mu.$$

The right hand side increases to $\int \mathbb{1}_G\, d\mu = \mu(G)$ by MCT, so (5.9) holds.

Note that Item 3 and Item 4 are equivalent since K is closed if and only if $G = K^c$ is open, and $\mu(K) = 1 - \mu(K^c)$.

From Items 3 and 4 to Item 5. For any set A, we use $\operatorname{int} A$ to denote the *interior* of A, defined by

$$\operatorname{int} A = \{x \in A : \exists r > 0 \text{ s.t. } B_r(x) \subset A\},$$

and $\bar{A}$ the *closure* of A. Then $\operatorname{int} A \subset A \subset \bar{A}$, and $\partial A = \bar{A} \setminus \operatorname{int} A$. By (5.9) and (5.10),

$$\mu(\operatorname{int} A) \le \liminf_{n\to\infty} \mu_n(A) \le \limsup_{n\to\infty} \mu_n(A) \le \mu(\bar{A}).$$

But $\mu(\partial A) = \mu(\bar{A}) - \mu(\operatorname{int} A) = 0$, so $\mu(\operatorname{int} A) = \mu(\bar{A}) = \mu(A)$, and the conclusion follows.

From Item 5 to Item 1. If x_0 is a continuous point of F, then $\mu\{x_0\} = 0 = \mu\big(\partial(-\infty, x_0)\big)$. Hence,

$$\lim_{n\to\infty} F_n(x_0) = \lim_{n\to\infty} \mu_n(-\infty, x_0] = \mu(-\infty, x_0] = F(x_0).$$

□

Weak convergence can also be characterized using metrics on $\mathcal{P}(\mathbb{R})$.

From the proof of Theorem 5.7, (5.8) holds if and only if it holds for all compactly supported continuous functions, denoted by $\mathcal{C}_c(\mathbb{R})$. There is a countable dense subset $(g_m)_{m\ge1} \subset \mathcal{C}_c(\mathbb{R})$, in the sense that for every $\varepsilon > 0$ and every $g \in \mathcal{C}_c(\mathbb{R})$, there exists g_{m_0} such that $\sup|g_{m_0} - g| < \varepsilon$. One such subset is all the piecewise linear continuous functions, whose graphs are obtained by connecting points $\big(x_i, g(x_i)\big) \in \mathbb{Q}^2$. Fixing such a dense subset $(g_m)_{m\ge1}$, we have $\mu_n \Rightarrow \mu$ if and only if

$$\int g_m\, d\mu_n \to \int g_m\, d\mu, \quad \forall g_m.$$

This means that weak convergence is characterized by the following metric:

$$d(\mu, \nu) = \sum_{m=1}^{\infty} 2^{-m} \Big(1 \wedge \Big| \int g_m\, d\mu - \int g_m\, d\nu \Big|\Big),$$

that is, $\mu_n \Rightarrow \mu$ if and only if $d(\mu_n, \mu) \to 0$.

Another useful metric is called the *Lévy distance* between monotone function. To motivate it, let us consider the graph of any increasing function F, defined by

$$\Gamma_F = \{(x, y) : x \in \mathbb{R},\ F(x-) \le y \le F(x+)\} \subset \mathbb{R}^2.$$

The distinction of continuous and discontinuous points in (5.4) is caused by the potential discontinuous point in F; but from the point of view of the graphs, (5.4) just means that the graphs of F_n converge to that of F. To make this idea precise, we can use the *Hausdorff distance* to measure closeness between any $A, B \subset \mathbb{R}^2$ ($\mathbb{R}^2$ could be any metric space):

$$\begin{aligned} d_H(A, B) &= \inf\{\varepsilon > 0 : A \subset B_\varepsilon,\ B \subset A_\varepsilon\}, \quad D_\varepsilon = \bigcup_{x \in D} \{y : |y - x| \le \varepsilon\}, \\ &= \inf\{\varepsilon > 0 : d(a, B) \le \varepsilon,\ \forall a \in A,\ d(b, A) \le \varepsilon,\ \forall b \in B\}, \end{aligned} \tag{5.13}$$

where D_ε is also known as the ε-neighborhood of D. The Lévy distance can be defined by

$$d_L(F, G) = d_H(\Gamma_F, \Gamma_G).$$

The more well-known form of Lévy distance is

$$d_L(F, G) = \inf\{\varepsilon > 0 : F(x - \varepsilon) - \varepsilon \le G(x) \le F(x + \varepsilon) + \varepsilon\}.$$

The two definitions are exactly the same if we use ℓ^∞ distance in $\mathbb{R}^2$ in (5.13).

We conclude this section by two simple properties of weak convergence.

Proposition 5.8 *If $X_n \to X$ in probability, then $X_n \Rightarrow X$.*

Proof: It suffices to show that $\mathsf{E}g(X_n) \to \mathsf{E}g(X)$ for any bounded continuous function g, which follows from DCT. □

The converse is not true, unless the limit is a δ-measure.

Proposition 5.9 *If $X_n \Rightarrow X$ where $\mathsf{P}(X = c) = 1$ for some $c \in \mathbb{R}$, then $X_n \to c$ in probability.*

Proof: Since $\mu_X\{c - \varepsilon, c + \varepsilon\} = 0$, by Item 5 in Theorem 5.7, we have

$$\lim_{n\to\infty} \mathsf{P}(|X_n - c| \ge \varepsilon) = \lim_{n\to\infty} \mu_{X_n}(c - \varepsilon, c + \varepsilon)^c = \mu_X(c - \varepsilon, c + \varepsilon)^c = 0.$$

□

5.3 Characteristic functions

In this section we introduce the *characteristic function (ch.f.)* of a r.v. X, defined by

$$\varphi_X(\xi) = \mathsf{E}e^{i\xi X} = \mathsf{E}\cos(\xi X) + i\mathsf{E}\sin(\xi X).$$

The integration of the complex-valued r.v. $e^{i\xi X}$ can be done by integrating the real and imaginary parts separately, that is, $\mathsf{E}X := \mathsf{E}X_1 + i\mathsf{E}X_2$ if X_1 and X_2 are real and complex parts of X. We also recall the norm of a complex number $z = a + bi$ is $|z| = \sqrt{a^2 + b^2}$. Like absolute values we have the following inequality for complex norms.

Lemma 5.10 *Let $X \in \mathbb{C}$ be a complex-valued r.v. Then $|\mathsf{E}X| \le \mathsf{E}|X|$.*

Proof: Write $X = X_1 + iX_2$. Since $\phi(a,b) = \sqrt{a^2+b^2}$ is convex, by Jensen's inequality we have

$$|\mathsf{E}X| = \sqrt{(\mathsf{E}X_1)^2 + (\mathsf{E}X_2)^2} = \phi(\mathsf{E}X_1, \mathsf{E}X_2) \le \mathsf{E}\phi(X_1, X_2) =\le \mathsf{E}\sqrt{X_1^2 + X_2^2} = \mathsf{E}|X|.$$

□

The ch.f. of a r.v. X is just the *Fourier transform* (up to some constants and signs) of this distribution μ_X. Unsurprisingly, a probability measure is uniquely determined by its Fourier transform, and we will assume this fact without proof in this section.

We have some basic properties of the ch.f.s.

Proposition 5.11 *Let $\varphi(\xi) = \mathsf{E}e^{i\xi X}$. Then*

1. *$\varphi(-\xi) = \overline{\varphi(\xi)}$, where $\bar{\cdot}$ denotes complex conjugate.*
2. *$\mathsf{E}e^{i\xi(aX+b)} = e^{ib\xi}\varphi(a\xi)$.*
3. *$\varphi(0) = 1$ and $|\varphi(\xi)| \le 1$.*
4. *$\xi \mapsto \varphi(\xi)$ is uniformly continuous.*

Proof: We will only prove the last one as the others are very straightforward.

For any ξ_1 and ξ_2, we have

$$|\varphi(\xi_1) - \varphi(\xi_2)| \le \mathsf{E}|e^{i\xi_1 X} - e^{i\xi_2 X}| = \mathsf{E}|e^{i(\xi_1-\xi_2)X} - 1|.$$

Since $|e^{iy} - 1| \le 2$ for any $y \in \mathbb{R}$, by BCT, we have

$$\lim_{h\to 0} \mathsf{E}|e^{ihX} - 1| = \mathsf{E}\lim_{h\to 0}|e^{ihX} - 1| = \mathsf{E}|e^0 - 1| = 0.$$

Since the upper bound only depends on $\xi_1 - \xi_2$, the conclusion follows. □

When studying sum of independent r.v.s, characteristic functions are very useful, since the Fourier transform turns convolutions into products.

Proposition 5.12 *Let X, Y be independent. Then $\varphi_{X+Y} = \varphi_X \cdot \varphi_Y$.*

Proof: For every fixed ξ, the function $h(x) = e^{i\xi x}$ is bounded and continuous. Proposition 3.2 also holds for complex-valued functions, and hence

$$\varphi_{X+Y}(\xi) = \mathsf{E}h(X)h(Y) = \mathsf{E}h(X) \cdot \mathsf{E}h(Y) = \varphi_X(\xi)\varphi_Y(\xi).$$

□

Since $h(x) = e^{i\xi x}$ is a bounded continuous function for every ξ, we know $\varphi_{X_n} \to \varphi_X$ pointwise if $X_n \Rightarrow X$. The converse is true, with an additional condition guaranteeing the tightness of $(\mu_{X_n})_{n\ge 1}$. This is the next result.

Theorem 5.13 (continuity theorem) *If $\varphi_{X_n}(\xi) \to \varphi(\xi)$ for every ξ, and φ is continuous at $\xi = 0$, then there exists a r.v. X such that $X_n \Rightarrow X$ and $\varphi_X = \varphi$.*

Proof: We first show that the continuity of φ at 0 implies the tightness of $(\mu_{X_n})_{n\ge 1}$. We will use Lemma 5.14 proven below, which gives

$$\mu_{X_n}\{x : |x| \ge 2/u\} \le \frac{1}{u}\int_{-u}^{u} \big(1 - \varphi_{X_n}(\xi)\big)\, d\xi. \tag{5.14}$$

Note that $\varphi(0) = \lim_{n\to\infty} \varphi_{X_n}(0) = 1$. By continuity, for every $\varepsilon > 0$, there exists $\delta > 0$ such that $|1 - \varphi(\xi)| \le \varepsilon$ when $|\xi| \le \delta$. Taking $u = \delta$ in (5.14), since $|1 - \varphi_{X_n}| \le 2$, by BCT we have

$$\begin{aligned}\limsup_{n\to\infty} \mu_{X_n}\{x : |x| \ge 2/\delta\} &\le \lim_{n\to\infty} \frac{1}{\delta}\int_{-\delta}^{\delta} |1 - \varphi_{X_n}(\xi)|\, d\xi \\ &= \frac{1}{\delta}\int_{-\delta}^{\delta} \lim_{n\to\infty} |1 - \varphi_{X_n}(\xi)|\, d\xi \\ &= \frac{1}{\delta}\int_{-\delta}^{\delta} |1 - \varphi(\xi)|\, d\xi \le 2\varepsilon.\end{aligned}$$

This implies the tightness of $(\mu_{X_n})_{n\ge1}$.

Since $(\mu_{X_n})_{n\ge1}$ is tight, by Theorem 5.4 there exists a subsequence (X_{n_k}) such that $X_{n_k} \Rightarrow X$ for some r.v. X. Then $\mathsf{E}e^{i\xi X_{n_k}} \to \mathsf{E}e^{i\xi X}$ and hence $\varphi_X \equiv \varphi$. Next we will show that $X_n \Rightarrow X$ along the full sequence. If not, then there exists f bounded, continuous and $\varepsilon_0 > 0$ such that

$$|\mathsf{E}f(X_{m_k}) - f(X)| \ge \varepsilon_0, \quad \text{along some subsequence } (X_{m_k}). \tag{5.15}$$

Since $\mu_{X_{m_k}}$ are also tight, there exists a further subsequence $\mu_{X_{m'_k}}$ such that $X_{m'_k} \Rightarrow Y$ for some Y. But then we have $\varphi_Y = \varphi = \varphi_X$, this contradicts with (5.15). □

Lemma 5.14 *Let ν be a probability measure on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ and φ be its ch.f. Then*

$$\nu\{x : |x| \ge 2/u\} \le \frac{1}{u}\int_{-u}^{u} \big(1 - \varphi(\xi)\big)\, d\xi.$$

Here, since $\varphi(-\xi) = \overline{\varphi(\xi)}$ and the domain is symmetric, the integral on the right side is real.

Proof: Using Fubuni's Theorem, we have

$$\begin{aligned}\int_{-u}^{u} \big(1 - \varphi(\xi)\big)\, d\xi &= \int \nu(dx) \int_{-u}^{u} (1 - e^{i\xi x})\, d\xi \\ &= \int \nu(dx) \int_{-u}^{u} (1 - \cos \xi x)\, d\xi \\ &= \int \Big(2u - \frac{2\sin ux}{x}\Big)\, \nu(dx) \\ &= 2u \int \Big(1 - \frac{\sin ux}{ux}\Big)\, \nu(dx) \\ &\ge u \int_{\{x:|ux|\ge2\}} \frac{1}{2}\, \nu(dx) = u \cdot \nu\{x : |ux| \ge 2\}\end{aligned}$$

Here, we use $1 - \frac{\sin ux}{ux} \ge 1 - |ux|^{-1} \ge 1/2$ if $|ux| \ge 2$. □

5.4 *Notes on Fourier transform

This section will give a brief introduction to the Fourier transform. The goal is to help the readers to understand characteristic functions in a more general context.

Fourier transform is first defined for functions. The *Fourier transform* of a function $g \in L^1(\mathbb{R})$ is defined by

$$(\mathbb{F}g)(\xi) := \int e^{i\xi x} g(x)\, dx. \tag{5.16}$$

The integrability condition $g \in L^1(\mathbb{R})$ is to ensure the integral in (5.16) to be defined.

Remark 5.6 In general, one needs to decide where to put constants and plus/minus signs in defining the Fourier transform; for example, more common definitions in harmonic analysis are

$$(\mathbb{F}g)(\xi) = \frac{1}{\sqrt{2\pi}} \int e^{-i\xi x} g(x)\, dx, \quad \text{or} \quad (\mathbb{F}g)(\xi) = \int e^{-2\pi i \xi x} g(x)\, dx.$$

But (5.16) agrees with the form of characteristic functions used in the probability theory so we will stick to it.

One can also define the *inverse Fourier transform* by

$$(\mathbb{F}^{-1}h)(x) := \frac{1}{2\pi} \int e^{-i\xi x} h(\xi)\, d\xi. \tag{5.17}$$

Note that like $\mathbb{F}$, the natural domain for $\mathbb{F}^{-1}$ are functions in $L^1(\mathbb{R})$. However, if $g \in L^1(\mathbb{R})$, then in general we merely have $\mathbb{F}g \in L^\infty(\mathbb{R})$, so $\mathbb{F}^{-1}$ is not a true "inverse" (but it will be after a proper adjustment). When it happens that $\mathbb{F}g \in L^1(\mathbb{R})$, the map $\mathbb{F}^{-1}$ indeed takes $\mathbb{F}g$ back to g. Here, the form of $\mathbb{F}^{-1}$ in (5.17) depends on the choice we made in (5.16) to define $\mathbb{F}$.

Proposition 5.15 *If $g \in L^1(\mathbb{R})$ and $\mathbb{F}g \in L^1(\mathbb{R})$, then $(\mathbb{F}^{-1} \circ \mathbb{F})g = g$.*

The proof usually involves some integration tricks, and can be found in most analysis/PDE textbooks that present the Fourier transform. We skip the proof here since the most important thing for us is to know that the Fourier transform does have an inverse, at least in some sense.

The next question is that we need to define the Fourier transform for objects other than L^1 functions, like the probability measures. One can say that probability measures are like L^1 functions, but we will see below that the Fourier transform can even be defined for unbounded functions/measures. The key are the "Schwartz space" and its dual space, the "tempered distributions".

The *Schwartz space* contains smooth functions that decays very fast at ∞; more precisely,

$$\mathcal{S} = \{g \in \mathcal{C}^\infty(\mathbb{R}) : \lim_{|x|\to\infty} |x^k| |g^{(m)}(x)| = 0,\ \forall k, m \geq 0\}.$$

The functions in $\mathcal{S}$ are called *Schwartz functions*. We can talk about convergence in $\mathcal{S}$: $g_n \to g$ in $\mathcal{S}$ if for every $k, m \geq 0$, $\sup_x |x|^k |g_n^{(m)}(x) - g^{(m)}(x)| \to 0$. The convergence can also characterized by the metric

$$d(f,g) = \sum_{k,m=0}^{\infty} \frac{|f-g|_{k,m} \wedge 1}{2^{m+k}}, \quad |h|_{k,m} := \sup_x |x|^k |h^{(m)}(x)|.$$

A nice thing about the Fourier transform is that it turns differentiation ∂_x^k into multiplication $(-i\xi)^k$ and vise versa.

Proposition 5.16 *Let $g \in \mathcal{S}$. Then for $k \geq 1$,*

$$(\mathbb{F}g^{(k)})(\xi) = (-i\xi)^k (\mathbb{F}g)(\xi), \quad \mathbb{F}\big((-ix)^k g\big) = \mathbb{F}g^{(k)}.$$

Hence, the Schwartz space $\mathcal{S}$ is invariant under $\mathbb{F}$. In light of Proposition 5.15, it is a bijection on $\mathcal{S}$.

Proposition 5.17 *The Fourier transform $\mathbb{F} : \mathcal{S} \to \mathcal{S}$ is a bijection.*

Another obvious fact is that $\mathbb{F}$ is linear: $\mathbb{F}(f+g) = \mathbb{F}f + \mathbb{F}g$. It is natural to consider the action of $\mathbb{F}$ on the dual of $\mathcal{S}$, called the *tempered distribution*, defined by

$$\begin{aligned} \mathcal{S}' &:= \{\text{continuous, linear functional on } \mathcal{S}\} \\ &= \{\ell \text{ linear} : \mathcal{S} \to \mathbb{R},\ |\ell(g)| \leq C_{m,k} |g|_{k,m}, \forall k, m \geq 0\}. \end{aligned}$$

The space $\mathcal{S}'$ contains all probability measures μ, identified with the linear functional

$$\ell_\mu(g) := \int g(x)\, d\mu(x).$$

It also contains $\mathcal{S}$ itself, identified with the linear functionals defined by taking L^2 inner product:

$$\ell_h(g) := \int g(x)h(x)\, dx, \quad h \in \mathcal{S}.$$

The Fourier transform can be defined on $\mathcal{S}'$ by duality:

$$(\mathbb{F}\ell)(g) := \ell(\mathbb{F}g).$$

For example, if μ is a probability measure on $\mathbb{R}$, then by Fubini's Theorem,

$$(\mathbb{F}\mu)(g) = \mu(\mathbb{F}g) = \int \Big[\int e^{i\xi x}\, dx\Big]\, d\mu(\xi) = \int \Big[\int e^{i\xi x}\, d\mu(\xi)\Big] g(x)\, dx = \int \varphi_\mu(x)g(x)\, dx, \quad \forall g \in \mathcal{S},$$

where φ_μ is the ch.f. of μ. Hence, the ch.f. φ_μ is $\mathbb{F}(\mu)$, when μ is treated as an element in $\mathcal{S}'$. Since $\mathbb{F} : \mathcal{S} \to \mathcal{S}$ is a bijection, it is also a bijection on $\mathcal{S}'$. Therefore, a probability measure is *uniquely* determined by its ch.f.

If one needs more information, the inversion from ch.f.s to probability measures can also be done via the *inversion formula*, which is more or less equivalent to Proposition 5.15.

Theorem 5.18 (Inversion formula) *Let μ be a probability measure on $\mathbb{R}$ and φ its ch.f. Then for every $a < b$,*

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

5.5 Central limit Theorem

5.5.1 CLT for i.i.d. random variables

We will use ch.f.s to prove central limit theorems. An important fact is that the derivatives of the ch.f. is related to the moments of the r.v.; there is a more general result for the Fourier transform, see Proposition 5.16.

Proposition 5.19 *If $\mathsf{E}|X|^n < \infty$, then $\varphi^{(n)}(\xi) = \mathsf{E}(iX)^n e^{i\xi X}$.*

Remark 5.7 Let $g(x,\xi) = e^{i\xi x}$. Then $\frac{\partial^n}{\partial \xi^n} g = (ix)^n e^{i\xi x}$, so Proposition 5.19 gives conditions to guarantee the "exchange of differentiation and integral (expectation)"

$$\frac{d^n}{d\xi^n} \mathsf{E} g(X,\xi) = \mathsf{E} \frac{\partial^n}{\partial \xi^n} g(X,\xi).$$

Proof: We will only prove the case $n = 1$. For $n \geq 2$, the proof uses induction and a similar argument.

Since $\mathsf{E}|X| < \infty$, we can define $h_1(\xi) = \mathsf{E}(iX)e^{i\xi X}$ as $|(iX)e^{i\xi X}| \leq |X|$. Also, $|h_1(\xi)| \leq \mathsf{E}|X|$, and similar to Proposition 5.11, one can show that $\xi \mapsto h_1(\xi)$ is (uniformly) continuous.

By Fubini's Theorem, for every $a < b$, we have

$$\int_a^b h_1(\xi)\, d\xi = \int_a^b \mathsf{E}(iX)e^{i\xi X}\, d\xi = \mathsf{E}\int_a^b (iX)e^{i\xi X}\, d\xi = \mathsf{E}(e^{ibX} - e^{-iaX}) = \varphi(b) - \varphi(a).$$

Since h_1 is continuous, φ is the anti-derivative of h_1 by the fundamental theorem of calculus, i.e., $\varphi' = h_1$. □

To deal with complex logarithm we frequently use the following lemma.

Lemma 5.20 *If $|z| \le 1/2$, then $|\log(1+z) - z| \le |z|^2$.*

Proof: The function $\log(1+z)$ is analytic for $|z| < 1$, and hence we have the Taylor expansion

$$\log(1+z) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} z^k.$$

Therefore,

$$|\log(1+z) - z| \le \sum_{k=2}^{\infty} \frac{|z^2|}{k} \le |z|^2 \sum_{k=2}^{\infty} \frac{1}{2^{k-2}k} \le |z|^2.$$

□

We say that X has normal distribution $\mathcal{N}(\mu, \sigma^2)$ if X is a continuous r.v. with density

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

Let us compute the ch.f. for normal distributions.

Proposition 5.21 *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then*

$$\varphi_X(\xi) = e^{i\mu\xi - \frac{1}{2}\sigma^2\xi^2}.$$

Proof: By Proposition 5.11, without loss of generality we can assume $\mu = 0$ and $\sigma = 1$. We need to show

$$\int_{\mathbb{R}} e^{i\xi x} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = e^{-\frac{\xi^2}{2}}.$$

Completing the square, the left hand side is

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{\frac{(x-i\xi)^2}{2}} \cdot e^{-\frac{\xi^2}{2}} \, dx,$$

so it suffices to show

$$1 = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-i\xi)^2}{2}} \, dx. \tag{5.18}$$

If $i\xi$ is a real number, then (5.18) is trivial by a change of variables, but it is not. We need some contour integral trick from complex variables.

Assume $\xi \ge 0$. Let $h(z) = \frac{1}{2\pi} e^{-z^2/2}$, $z \in \mathbb{C}$. Then $h(z)$ is an entire function (since the exponential function is nice), and by Cauchy integral theorem, $\int_\Gamma h(z)\, dz = 0$ for any closed contour Γ. Consider the contour $\Gamma = \bigcup_{k=1}^4 \Gamma_k$, where

$$\Gamma_1 = [-L, L], \quad \Gamma_2 = \{L + iy : y \in [0, \xi]\}, \quad \Gamma_3 = \{x + i\xi : x \in [-L, L]\}, \quad \Gamma_4 = \{-L + iy : y \in [0, \xi]\}$$

with proper orientation (counter-clockwise). Then

$$\Big| \int_{\Gamma_2} h(z)\, dz \Big| \le \int_0^\xi |h(L+iy)|\, dy \le \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(L^2 - \xi^2)} \cdot \xi \to 0, \quad L \to \infty,$$

and similar limit holds for Γ_4. Hence,

$$\lim_{L\to\infty} \int_{\Gamma_1} h(z)\, dz = -\lim_{L\to\infty} \int_{\Gamma_3} h(z)\, dz. \tag{5.19}$$

The right hand side of (5.19) is right hand side of (5.18), while the left hand side of (5.19) is the integration of the density of $\mathcal{N}(0,1)$, which is 1. This completes the proof. □

Theorem 5.22 *Let $X_1, X_2, \dots$ be i.i.d. with $\mathsf{E}X_1 = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Rightarrow \mathcal{N}(0,1).$$

Proof: By Theorem 5.13 and Proposition 5.21, it suffices to show that for every $\xi \in \mathbb{R}$,

$$\mathsf{E}e^{i\xi\frac{S_n-n\mu}{\sigma\sqrt{n}}} \to e^{-\frac{1}{2}\xi^2}, \quad n \to \infty.$$

Rearranging, the LHS becomes

$$\mathsf{E}e^{i\frac{\xi}{\sqrt{n}}\sum_{m=1}^n \frac{X_m-\mu}{\sigma}} = \mathsf{E}e^{i\frac{\xi}{\sqrt{n}}\sum_{m=1}^n Y_m},$$

where $Y_m = \frac{X_m-\mu}{\sigma}$ is the normalized r.v.s, with $\mathsf{E}Y_m = 0$, $\mathsf{E}Y_m^2 = 1$. The r.v.s Y_m are i.i.d. Let φ be the ch.f. of Y_1. Then by independence and Proposition 5.12, we need to show

$$\left[\varphi(\xi/\sqrt{n})\right]^n \to e^{-\frac{1}{2}\xi^2},$$

or equivalently, since the limit is positive and exponential function is continuous,

$$n\log\varphi(\xi/\sqrt{n}) \to -\frac{1}{2}\xi^2.$$

By Proposition 5.19, since Y_1 has second moment, its ch.f. φ is twice differentiable, and $\varphi(0) = 1$, $\varphi'(0) = 1$, $\varphi''(0) = -1$. In particular, we have Taylor expansion for φ at 0 with Peano remainder:

$$\varphi(\eta) = 1 - \frac{1}{2}\eta^2 + \eta^2\alpha(\eta), \quad \lim_{\eta\to 0}|\alpha(\eta)| = 0. \tag{5.20}$$

Note that the $o(1)$ term $\alpha(\eta)$ is complex.

For n large enough, $|\varphi(\xi/\sqrt{n}) - 1| \le 1/2$, and hence by Lemma 5.20 and (5.20), we have

$$\begin{aligned}\left|n\log\varphi(\xi/\sqrt{n}) + \frac{1}{2}\xi^2\right| &\le \left|n\log\varphi(\xi/\sqrt{n}) - n\big(\varphi(\xi/\sqrt{n}) - 1\big)\right| + \left|n\big(\varphi(\xi/\sqrt{n}) - 1\big) + \frac{1}{2}\xi^2\right| \\ &\le n\left|\varphi(\sqrt{\xi}/\sqrt{n}) - 1\right|^2 + \xi^2|\alpha(\xi/\sqrt{n})|.\end{aligned}$$

The second term obviously converges to 0; the first term is

$$\left|-\frac{1}{2}\xi^2 + \xi^2\alpha(\xi/\sqrt{n})\right| \cdot \left|\varphi(\xi/\sqrt{n}) - 1\right| \le C\left|\varphi(\xi/\sqrt{n}) - 1\right|$$

and also converges to 0. This completes the proof. □

5.5.2 CLT for triangular arrays

The motivation to study CLT for triangular arrays is that normal distributions in real life, such as height, weights etc, are results of many independent, yet *not identically distributed* small factors. When the r.v.s are not i.i.d., we need more delicate control of the ch.f.s.

The following result is useful.

Proposition 5.23 *Let $h(z) = e^{iz}$ and $P_k(z)$ be the k-th order Taylor polynomials of $h(z)$ at $z = 0$. Then*

$$|\varphi(\xi) - \mathsf{E}P_k(\xi X)| \le \mathsf{E}\Big(\frac{|\xi X|^{k+1}}{(k+1)!} \wedge \frac{2|\xi X|^k}{k!}\Big).$$

In particular, when $k = 2$, we have $P_k(z) = 1 + iz - \frac{z^2}{2}$ and

$$|\varphi(\xi) - (1 + i\xi\mathsf{E}X - \frac{\xi^2}{2}\mathsf{E}X^2)| \le \xi^2\mathsf{E}\Big(\frac{|\xi X^3|}{6} \wedge |X|^2\Big).$$

Proof: Since $|h^{(k+1)}| \le 1$, we have

$$|h(\xi X) - P_k(\xi X)| = \Big|\int_0^\xi \frac{(iX)^{k+1}\theta^k h^{(k+1)}(\theta X)}{k!}\, d\theta\Big| \le \frac{|\xi X|^{k+1}}{(k+1)!}. \tag{5.21}$$

The bound is bad if $|\xi X|$ is large. Using (5.21), we also have

$$|h(\xi X) - P_k(\xi X)| \le |h(\xi X) - P_{k-1}(\xi X)| + |\frac{(i\xi X)^k}{k!}| \le 2\frac{|\xi X|^k}{k!}. \tag{5.22}$$

The conclusion follows from combining (5.21) and (5.22), and then taking expectation. □

Theorem 5.24 (Linderburg-Feller) *Let $(X_{n,m})_{m=1}^n$ be independent with $\mathsf{E}X_{n,m} = 0$. Assume that*

$$\sum_{m=1}^n \mathsf{E}X_{n,m}^2 \to \sigma^2, \quad n \to \infty, \tag{5.23}$$

and the so-called "Linderburg's condition":

$$\forall \varepsilon > 0, \quad M_n := \sum_{m=1}^n \mathsf{E}X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}| \ge \varepsilon\}} \to 0, \quad n \to \infty. \tag{5.24}$$

Then $S_n = X_{n,1} + \cdots + X_{n,n} \Rightarrow \mathcal{N}(0, \sigma^2)$.

Proof: By independence,

$$\mathsf{E}e^{i\xi S_n} = \prod_{m=1}^n \mathsf{E}e^{i\xi X_{n,m}} =: \prod_{m=1}^n \varphi_{n,m}(\xi).$$

By Theorem 5.13 and Proposition 5.21, it suffices to show that for every $\xi \in \mathbb{R}$,

$$\sum_{m=1}^n \log \varphi_{n,m}(\xi) \to -\frac{1}{2}\sigma^2\xi^2. \tag{5.25}$$

The idea is to use the approximations $\log \varphi_{n,m}(\xi) \approx \varphi_{n,m}(\xi) - 1 \approx -\mathsf{E}X_{n,m}^2$ and sum over m. To control the accumulated error after all these approximations, we need to use use the Linderburg's condition (5.24) and Lemma 5.20.

By Proposition 5.23, we have

$$|\varphi_{n,m}(\xi) - 1 + \frac{\xi^2}{2}\mathsf{E}X_{n,m}^2| \le \xi^2\mathsf{E}(|\xi X_{n,m}^3|) \wedge |X_{n,m}|^2 \le \xi^2(\varepsilon\mathsf{E}|\xi X_{n,m}^2| + \mathsf{E}X_{n,m}^2\mathbb{1}_{\{|X_{n,m}| \ge \varepsilon\}})$$

We also have

$$\mathsf{E}X_{n,m}^2 \le \varepsilon^2 + \mathsf{E}X_{n,m}^2\mathbb{1}_{\{|X_{n,m}| \ge \varepsilon\}}.$$

Therefore, for some constant $C = C(\xi)$,

$$|\varphi_{n,m}(\xi) - 1|, |\varphi_{n,m}(\xi) - 1 + \frac{\xi^2}{2}\mathsf{E}X_{n,m}^2| \le C(\varepsilon + \mathsf{E}X_{n,,m}^2 \mathbb{1}_{\{|X_{n,m}| \ge \varepsilon\}}) \le C(\varepsilon + M_n).$$

Therefore, by first choosing ε small enough, then n large enough, we can ensure that $|\varphi_{n,m}(\xi) - 1| \le 1/2$ for all m.

Using Lemma 5.20, we have

$$\begin{aligned}\sum_{m=1}^n \left| \log \varphi_{n,m}(\xi) + \frac{\xi^2}{2}\mathsf{E}X_{n,m}^2 \right| &\le \sum_{m=1}^n \left| \log \varphi_{n,m}(\xi) - \varphi_{n,m}(\xi) + 1 \right| + \sum_{m=1}^n \left| \varphi_{n,m}(\xi) - 1 + \frac{\xi^2}{2}\mathsf{E}X_{n,m}^2 \right| \\ &\le \sum_{m=1}^n \left| \varphi_{n,m}(\xi) - 1 \right|^2 + C_1 \sum_{m=1}^n (\varepsilon \mathsf{E}X_{n,m}^2 + \mathsf{E}X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}| \ge \varepsilon\}}).\end{aligned}$$

By (5.23) and (5.24), the lim sup of second term is bounded by $C_2\varepsilon$ as $n \to \infty$. For the first term, we have

$$\sum_{m=1}^n \left| \varphi_{n,m}(\xi) - 1 \right|^2 \le \max_{1 \le m \le n} \left| \varphi_{n,m}(\xi) - 1 \right| \cdot \sum_{m=1}^n \left| \varphi_{n,m}(\xi) - 1 \right| \le C(\varepsilon + M_n) \cdot \sum_{m=1}^n \left| \varphi_{n,m}(\xi) - 1 \right|,$$

which is bounded by $C_3\varepsilon$ since the summation is bounded by

$$\sum_{m=1}^n \left[\frac{\xi^2}{2}\mathsf{E}X_{n,m}^2 + \left| \varphi_{n,m}(\xi) - 1 + \frac{\xi^2}{2}\mathsf{E}X_{n,m}^2 \right| \right]$$

Since $\varepsilon > 0$ is arbitrary, we have

$$\lim_{n \to \infty} \sum_{m=1}^n \left| \log \varphi_{n,m}(\xi) + \frac{\xi^2}{2}\mathsf{E}X_{n,m}^2 \right| = 0.$$

Then (5.25) follows from this and (5.23). □

Example 5.8 We can recover Theorem 5.22 from Theorem 5.24.

Let Y_n be i.i.d. with $\mathsf{E}Y_n = 0$ and $\mathsf{E}Y_n^2 = 1$. Let $X_{n,m} = \frac{Y_m}{\sqrt{n}}$. Then (5.23) is satisfied. For (5.24), we have

$$\lim_{n \to \infty} \sum_{m=1}^n \mathsf{E}X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}| \ge \varepsilon\}} = \lim_{n \to \infty} \sum_{m=1}^n \mathsf{E}\frac{Y_m^2}{n} \mathbb{1}_{\{|Y_m| \ge \sqrt{n}\varepsilon\}} = \lim_{n \to \infty} \mathsf{E}Y_1^2 \mathbb{1}_{\{|Y_1| \ge \sqrt{n}\varepsilon\}} = \mathsf{E} \lim_{n \to \infty} \mathsf{E}Y_1^2 \mathbb{1}_{\{|Y_1| \ge \sqrt{n}\varepsilon\}} = 0,$$

where the exchange of limit and expectation is due to $\mathsf{E}Y_1^2 < \infty$ and DCT.

Example 5.9 Theorem 5.24 can treat the case where the r.v.s are not identically distributed. Note that $\mathsf{E}Y_n^2 \le C$ along cannot guarantee that $X_{n,m} = \frac{Y_m}{\sqrt{n}}$ satisfies the Linderburg's condition (5.24). A sufficient condition is that Y_n has uniform $(2+\delta)$-moment for any $\delta > 0$, i.e., $\mathsf{E}|Y_n|^{2+\delta} \le C$ for some $C > 0$ and $\delta > 0$.

Indeed, for such Y_n, we have

$$\sum_{m=1}^n \mathsf{E}\frac{Y_m^2}{n} \mathbb{1}_{\{|Y_m| \ge \sqrt{n}\varepsilon\}} \le \sum_{m=1}^n \mathsf{E}\frac{|Y_m|^{2+\delta}}{n \cdot (\sqrt{n}\varepsilon)^\delta} \mathbb{1}_{\{|Y_m| \ge \sqrt{n}\varepsilon\}} \le n \cdot \frac{C}{n \cdot (\sqrt{n}\varepsilon)^\delta} = \frac{C}{(\sqrt{n}\varepsilon)^\delta} \to 0, \quad n \to \infty.$$

Next we use Theorem 5.24 to derive a CLT-type limit theorem for i.i.d. r.v.s with infinite variance. We should mention a result due to Lévy.

Theorem 5.25 *Let X_n be i.i.d. and S_n be its partial sum. Then there exist a_n, b_n such that $\frac{S_n - a_n}{b_n} \Rightarrow \mathcal{N}(0,1)$ if and only if*

$$\frac{y^2 \mathsf{P}(|X_1| \geq y)}{\mathsf{E}|X_1|^2 \mathbb{1}_{\{|X_1| \leq y\}}} \to 0. \tag{5.26}$$

The idea behind (5.26) is that to have normal distribution as the limit, each term in the partial sum cannot be too large, while here the "largeness" is measured by $y^2\mathsf{P}(|X_1| \geq y)$, compared to the truncated second moment of X_1.

Let us consider i.i.d. r.v.s X_n where $\mathsf{P}(X_1 < -x) = \mathsf{P}(X_1 > x) = x^{-2}$, $x \geq 1$. We will show that

$$\frac{X_1 + \cdots + X_n}{\sqrt{n \log n}} \Rightarrow \mathcal{N}(0,1).$$

To apply Theorem 5.25, we need to compute variance. Since $\mathsf{E}X_1^2 = \infty$, we need to apply truncation first. Let $Y_{n,m} = X_m \mathbb{1}_{\{|X_m| \leq c_n\}}$ and $\tilde{S}_n = Y_{n,1} + \cdots + Y_{n,n}$. We first prove a simple result, saying that if the truncation does not affect S_n much, then it will not affect the weak convergence.

Proposition 5.26 *If $\mathsf{P}(S_n \neq \tilde{S}_n) \to 0$ and $\frac{\tilde{S}_n}{b_n} \Rightarrow \mathcal{N}(0,1)$, then $\frac{S_n}{\sqrt{n}} \Rightarrow \mathcal{N}(0,1)$.*

Proof: Let g be a bounded continuous function and $N \sim \mathcal{N}(0,1)$. We have

$$\begin{aligned}\left|g(\frac{S_n}{b_n}) - g(N)\right| &\leq \left|g(S_n/b_n) - g(\tilde{S}_n b_n)\right| + \left|g(\tilde{S}_n/n) - g(N)\right| \\ &\leq 2\sup|g| \cdot \mathsf{P}(S_n \neq \tilde{S}_n) + \left|g(\tilde{S}_n/n) - g(N)\right| \to 0.\end{aligned}$$

□

To have $\mathsf{P}(\tilde{S}_n \neq S_n) \to 0$, a sufficient condition is

$$\sum_{m=1}^{n} \mathsf{P}(Y_{n,m} \neq X_m) = n\mathsf{P}\big(|X_1| \geq c_n\big) = \frac{n}{c_n^2} \to 0.$$

We will choose $c_n = n^{1/2} \log\log n$. The reason for double logarithm will be clear in a moment.

Now let us verify the two conditions (5.23) and (5.24).

For (5.23), we have

$$\begin{aligned}\sum_{m=1}^{n} \mathsf{E}Y_{n,m}^2 &= n \int_1^{c_n} 2y\mathsf{P}(|X_1| \geq y)\, dy = n \int_1^{c_n} \frac{2dy}{y} \\ &= 2n \log(n^{1/2} \log\log n) = n \log n + 2n \log\log\log n.\end{aligned}$$

So

$$\sum_{m=1}^{n} \mathsf{E}\Big|\frac{Y_{n,m}}{\sqrt{n \log n}}\Big|^2 \to 1.$$

For (5.24), we have

$$\sum_{m=1}^{n} \mathsf{E}\Big|\frac{Y_{n,m}}{\sqrt{n \log n}}\Big|^2 \mathbb{1}_{\{|Y_{n,m}| \geq \varepsilon\sqrt{n \log n}\}} = 0$$

for large n, since $|Y_m| \leq c_n = n^{1/2} \log\log n \ll \sqrt{n \log n}$.

5.5.3 Multidimensional CLT

In this section we discuss how to generalize the CLT to $\mathbb{R}^d$. First we need to introduce the weak convergence and the ch.f. in $\mathbb{R}^d$.

We write $X_n = (X_{n,1}, \dots, X_{n,d}) \in \mathbb{R}^d$ for i.i.d. random vectors in $\mathbb{R}^d$. We say that $X_n \Rightarrow X$ if $\mathsf{E}g(X_n) \to \mathsf{E}g(X)$ for every bounded continuous $g : \mathbb{R}^d \to \mathbb{R}$.. A family of probability measures on $\mathbb{R}^d$, $(\mu_n)_{n\geq 1}$, are *tight*, if for every $\varepsilon > 0$, there exists a compact set K, or more specifically, $K = [-L, L]^d$ (since compacts sets in $\mathbb{R}^d$ are bounded closed) so that $\mu_n(K^c) \leq \varepsilon$ for all $n \geq 1$.

Let $X \in \mathbb{R}^d$ be a random vector. Its characteristic function φ_X is defined by

$$\varphi_X(\xi) = \mathsf{E}e^{i\xi\cdot X}, \quad \xi \in \mathbb{R}^d.$$

Here, $\cdot$ denotes the inner/dot product in $\mathbb{R}^d$: $\xi\cdot x := \xi_1 x_1 + \cdots + \xi_d x_d$. We have a version of Theorem 5.13 in $\mathbb{R}^d$.

Theorem 5.27 *Let X_n, X_∞ be random vectors in $\mathbb{R}^d$. Then $X_n \Rightarrow X_\infty$ if and only if $\varphi_{X_n}(\xi) \to \varphi_{X_\infty}(\xi)$ for every $\xi \in \mathbb{R}^d$.*

Proof: The "only if" part follows from the definition of weak convergence and that $x \mapsto e^{i\xi\cdot x}$ is bounded continuous.

For the "if" part, similar to the proof of Theorem 5.13, it suffices to show that $(\mu_n = \mu_{X_n})$ is tight, and we can use a subsequence argument to finish the proof.

Let e_k be the unit vector in the k-th direction. Then $(e_k \cdot X_n)_{n\geq 1}$ is a family of r.v.s, with ch.f.s

$$\varphi_n^{(k)}(\eta) = \mathsf{E}e^{i\eta e_k\cdot X_n} = \varphi_{X_n}(\eta e_k) \to \varphi_{X_\infty}(\eta e_k),$$

where the limit is the ch.f. of $e_k\cdot X_\infty$ and hence continuous at $\eta = 0$. By Theorem 5.13, the distribution of $e_k \cdot X_n$ is tight, namely, for $\varepsilon/d > 0$, there exists $L_k > 0$ such that

$$\mu_n\{e_k \cdot X_n \notin [-L_k, L_k]\} \leq \frac{\varepsilon}{d}, \quad \forall n \geq 1.$$

Do this for every $k \in \{1, \dots, d\}$, and let $L = \max\{L_1, \dots, L_d\}$, we have

$$\mu_n\{X_n \notin [-L, L]^d\} \leq \sum_{k=1}^{d} \mu_n\{e_k \cdot X_n \notin [-L_k, L_k]\} \leq d \cdot \frac{\varepsilon}{d} = \varepsilon,$$

and hence $(\mu_n)_{n\geq 1}$ is tight. □

In fact in the proof we have characterize weak convergence in $\mathbb{R}^d$ via weak convergence in $\mathbb{R}$.

Proposition 5.28 *The weak convergence $X_n \Rightarrow X$ holds for random vectors in $\mathbb{R}^d$ if and only if $\theta \cdot X_n \Rightarrow \theta \cdot X$ as r.v.s for every $\theta \in \mathbb{R}^d$.*

We are ready to give a multidimensional version of CLT.

Theorem 5.29 (CLT in $\mathbb{R}^d$) *Let X_n be i.i.d. random vectors in $\mathbb{R}^d$, with $\mathsf{E}X_1 = \mu \in \mathbb{R}^d$, and covariance matrix $\Gamma = \mathsf{E}(X_1 - \mu)\cdot(X_1^T - \mu)$, that is,*

$$\Gamma_{jk} = \mathrm{Cov}(X_{1,j}, X_{1,k}), \quad 1 \leq j, k \leq d.$$

Then

$$\mathsf{E}e^{i\frac{S_n - n\mu}{\sqrt{n}}\cdot\xi} \to e^{-\frac{1}{2}\xi^T\Gamma\xi}, \quad \xi \in \mathbb{R}^d,$$

where $e^{-\frac{1}{2}\xi^T\Gamma\xi}$ is the ch.f. of the multi-variate normal distribution $\mathcal{N}(0, \Gamma)$.

Proof: Assume $\mu = 0$. Let $N \sim \mathcal{N}(0, \Gamma)$. Then $\theta \cdot N \sim \mathcal{N}(0, \theta^T \Gamma \theta)$. We have

$$e^{i\frac{S_n}{\sqrt{n}}\cdot\theta\eta} = e^{i\frac{\eta}{\sqrt{n}}\sum_{m=1}^n (X_m\cdot\theta)} \to e^{-\frac{\eta^2}{2}\cdot\theta^T\Gamma\theta}$$

by Theorem 5.22 and

$$\mathsf{E}(\theta \cdot X_1)^2 = \mathsf{E}\theta^T X_1 X_1^T \theta = \theta^T \Gamma \theta.$$

This and Proposition 5.28 prove the theorem. □

To allow possible degeneracy, the most convenient way to define the multi-variate normal $\mathcal{N}(\mu, \Gamma)$ is to use the ch.f.:

$$\varphi_{\mathcal{N}(\mu,\Gamma)}(\xi) = e^{i\xi\cdot\mu - \frac{1}{2}\xi^T\Gamma\xi}.$$

Note that the covariance matrix Γ is always semi-positive definite symmetric. If all eigenvalues of Γ are positive, then $\mathcal{N}(0, \Gamma)$ has a density given by

$$\frac{1}{(\sqrt{2\pi})^d\sqrt{\det(\Gamma)}} e^{-\frac{(x-\mu)^T\Gamma^{-1}(x-\mu)}{2}}.$$

In general, one can diagonalize Γ to get $\Gamma = \sum_{k=1}^d \lambda_k v_k^T v_k$ where $\vec{v}_k$ form an orthonormal basis in $\mathbb{R}^d$ and $\lambda_k \geq 0$. Then $\mathcal{N}(0, \Gamma)$ can be realized as

$$N = \sum_{k=1}^d \lambda_k \varepsilon_k \vec{v}_k,$$

where ε_k are i.i.d. $\mathcal{N}(0, 1)$ r.v.s. Another way to define multi-variate normal is by projection: we say that $N \sim \mathcal{N}(\mu, \Gamma)$ if $\theta \cdot N \sim \mathcal{N}(\theta \cdot \mu, \theta^T \Gamma \theta)$ for every $\theta \in \mathbb{R}^d$.

6 Weak convergence on general spaces and functional CLT

6.1 Preliminaries for probability measures on metric spaces

Let (M, d) be a generic metric space, where $d : M \times M \to [0, \infty)$ is the metric. Examples include:

- $M = \mathbb{R}^d$, the d-dimensional Euclidean space, with $d(x, y) = |x - y|_p$, $p \in [1, \infty]$.
  Here, all ℓ_p norms are *equivalent*, that is, they generate the same open sets.
- $M = \mathcal{C}[0, 1]$, the space of continuous function on $[0, 1]$, with $d(x, y) = \sup_{t\in[0,1]} |x(t) - y(t)|$.
- $M = L^p(\Omega, \mathcal{F}, \mu)$, the space of L^p-functions, with $d_{L^p}(f - g) = \int |f(x) - g(x)|^p \, \mu(dx)$.

We first recall some basic concepts for metric spaces.

- **Open sets.** A set $G \subset M$ is *open* if G contains an ε-ball around every $x \in G$, that is, $\forall x \in G$, $\exists \varepsilon > 0$ s.t. $y \in G$ whenever $d(y, x) < \varepsilon$.
- **Closed set.** A set $F \subset M$ is *closed* if F^c is open.
- **Completeness.** The space M is said to be *complete* if every Cauchy sequence $(x_n) \subset M$ has a limit point in M, that is, if $\lim_{N\to\infty} \sup_{n,m\geq N} d(x_n, x_m) = 0$, then there exists $x_\infty \in M$ such that $\lim_{n\to\infty} x_n = x_\infty$.
  Most metric spaces are complete, like $\mathbb{R}^d$, L^p and $\mathcal{C}[a, b]$. If starting from a non-complete metric space, one can always *complete* it using Cauchy sequences, as one did in completing $\mathbb{Q}$ to get $\mathbb{R}$.

- **Compact sets.** A set $K \subset M$ is *compact*, if every open cover $\bigcup_{i\in I} G_i \supset K$ contains a finite subcover $G_{i_1} \cup \cdots \cup G_{i_n} \supset K$.

- **Separable.** The space M is called separable if there exists a countable dense subset D. We say that D is dense in M, if for every $\varepsilon > 0$ and $x \in M$, there exists $y \in D$ such that $d(x, y) < \varepsilon$.

  The Euclidean space $\mathbb{R}^d$ is separable by taking $D = \mathbb{Q}^d$.

  The space $\mathcal{C}[0, 1]$ is separable by taking $D = \bigcup_{n=1}^{\infty} D_n$, where

  $$D_n = \Big\{ x \in \mathcal{C}[0,1] : x(t) \in \frac{1}{n}\mathbb{Z},\ t \in \frac{1}{n}\mathbb{Z},\ \text{and linear on } [i/n, (i+1)/n] \Big\}.$$

  The space $L^p(\mathbb{R}^d)$ is separable when $p \neq \infty$.

  Another class of non-separable spaces are linear operators on Banach spaces. For example, all the bounded linear map from $L^p(\mathbb{R})$ ($p \in [1, \infty)$) into itself, equipped with the operator norm, is non-separable, even if $L^p(\mathbb{R})$ is separable. To see this, the translation operators $(\tau_s f)(x) = f(s + x)$ are bounded and linear on $L^p(\mathbb{R})$, while $\|\tau_s - \tau_{s'}\| = 1$ whenever $s \neq s'$, so $(\tau_s)_{s\in\mathbb{R}}$ cannot be close to a countable set.

  One way to use compactness and separability is as follows: if M is a complete, separable, compact metric space, then $\mathcal{C}(M)$ is compact.

- **Sequentially compact.** A set K is *sequentially compact*, if every sequence $(x_n) \subset K$ has a subsequence (x_{n_k}) such that $x_{n_k} \to x_*$ for some $x_* \in K$.

  On separable metric spaces, compactness is equivalent to sequentially compactness, so we do not distinguish between them hereafter.

  On $\mathbb{R}^d$, compact sets are bounded, closed sets.

  On $\mathcal{C}[0, 1]$, by Arzelà–Ascoli, $(x_n)_{n\in I}$ are sequentially compact if and only if they are *uniformly bounded*,

  $$\exists C > 0,\ \text{s.t. } \sup_{n\in I} \sup_{t\in[0,1]} |x_n(t)| \le C, \tag{6.1}$$

  and *equi-continuous*,

  $$\forall \varepsilon > 0,\ \exists \delta > 0,\ \text{s.t. } |x_n(t_1) - x_n(t_2)| \le \varepsilon, \quad \forall n \in I,\ \forall |t_1 - t_2| < \delta. \tag{6.2}$$

  Assuming (6.2), (6.1) can also be replaced by

  $$\exists C > 0,\ \text{s.t. } \sup_{n\in I} |x_n(0)| \le C, \tag{6.3}$$

Let μ_n, μ_∞ be probability measures on (M, d). We say that μ_n converges to μ_∞ weakly, denoted by $\mu_n \Rightarrow \mu_\infty$, if for every bounded continuous function $g : M \to \mathbb{R}$,

$$\int_M g d\mu_n \to \int_M g d\mu.$$

We say that $(\mu_n)_{n\in I}$ is *tight*, if for every $\varepsilon > 0$, there exists a compact set $K \subset M$ such that

$$\mu_n(K^c) \le \varepsilon, \quad n \in I.$$

We have seen these definitions for $M = \mathbb{R}$. For general metric spaces we have the following result.

Theorem 6.1 (Prohorov Theorem) *Let (M, d) be a separable and complete metric space (a.k.a. Polish space). Let $(\mu_n)_{n\ge1}$ be probability measures on (M, d). If (μ_n) is tight, then there exists a subseqeucne (μ_{n_k}) and μ_∞ such that $\mu_{n_k} \Rightarrow \mu_\infty$.*

6.2 Donsker invariance principle

Let X_n be i.i.d. r.v.s with $\mathsf{E}X_1 = 0$ and $\mathsf{E}X_1^2 = 1$. We can extend the partial sum $S_n = X_1 + \cdots + X_n$ to a continuous function on $[0,1]$ by defining

$$\tilde{S}_n(t) = \begin{cases} S_m, & t = \frac{m}{n},\ m = 0, 1, \ldots, n, \\ \text{linear}, & \frac{m}{n} < t < \frac{m+1}{n}. \end{cases}$$

Then $\frac{\tilde{S}_n}{\sqrt{n}}$ is a random element in $\mathcal{C}[0,1]$.

Theorem 6.2 (Donsker's Invariance Principle/functional CLT) *The law of $\frac{\tilde{S}_n}{\sqrt{n}}$, as a probility distribution on $\mathcal{C}[0,1]$, converges weakly to the Wiener measure, the law of the Brownian motion.*

In this section we will not rigorously define what is the Brownian motion, or the Wiener measure, as it is a large subject studied in details in stochastic analysis. We will be content with establishing the tightness of the law of $\frac{\tilde{S}_n}{\sqrt{n}}$, and have a better understanding of the central limit theorem.

First we want to reformulate the equi-continuity condition (6.2). For $x \in \mathcal{C}[0,1]$ and $\delta > 0$, we define the *modules of continuity* of x at δ by

$$\omega(x;\delta) = \sup_{|t_1 - t_2| < \delta} |x(t_1) - x(t_2)|.$$

Then (6.2) is equivalent to

$$\lim_{\delta \to 0} \sup_n \omega(x_n, \delta) = 0. \tag{6.4}$$

Proposition 6.3 *Let μ_n be the law of random elements $x_n \in \mathcal{C}[0,1]$. The $(\mu_n)_{n\geq 1}$ is tight if and only if for every $\varepsilon > 0$, there exists $C > 0$ such that*

$$\mu_n\{|x_n(0)| > C\} < \varepsilon, \quad \forall n \geq 1, \tag{6.5}$$

and for every $\eta > 0$, there exists $\delta > 0$ such that

$$\mu_n\{\omega(x_n, \delta) > \eta\} < \varepsilon, \quad \forall n \geq 1. \tag{6.6}$$

Proof: The "only if" part is simple. We will prove the "if" part.

Let $\varepsilon > 0$. By (6.5), there exists $C > 0$ such that

$$\mu_n(A_0) \geq 1 - \varepsilon/2, \quad n \geq 1, \quad A_0 = \{x : |x(0)| < C\}.$$

By (6.6), there exist δ_k for all $k \geq 1$ such that

$$\mu_n(A_k) \geq 1 - \varepsilon/2^{k+1}, \quad n \geq 1, \quad A_k = \{x : \omega(x, \delta_k) < \frac{1}{k}\}.$$

Now let $A = \bigcap_{k=0}^{\infty} A_k$ and let $\bar{A}$ be its closure. By subadditivity, for all $n \geq 1$ we have

$$\mu_n(\bar{A}) \geq \mu_n(A) \geq 1 - \sum_{k=0}^{\infty} \mu_n(A_k^c) \geq 1 - \varepsilon.$$

We will estabilish the tightness, by showing that $\bar{A}$ is a compact set in $\mathcal{C}[0,1]$.

It suffices to check the two conditions (6.3) and (6.4) on A. Let $x \in A$. We have $|x(0)| \le C$ since $x \in A_0$. Since $x \in A_k$, $\omega(x, \delta_k) \le \frac{1}{k}$ for all k. Since $\omega(x, \delta)$ is decreasing in δ, we have

$$\lim_{\delta \to 0} \sup_{x \in A} \omega(x, \delta) \le \limsup_{k \to \infty} \frac{1}{k} = 0.$$

□

Now we will use Proposition 6.3 to show that $\mu_n = \mathcal{L}(\frac{\tilde{S}_n(\cdot)}{\sqrt{n}})$ is tight.

(6.5) holds since $\tilde{S}_n(0) \equiv 0$.

For (6.6), let $\eta, \delta > 0$, and let m be such that $\frac{m}{2n} < \delta \le \frac{m}{n}$. We claim that

$$\mu_n\{\omega(x, \delta) \ge \eta\} \le \sum_{k=0}^{[n/m]} \mu_n\Big(\max_{km \le \ell \le (k+1)m} \Big| \frac{S_\ell - S_{km}}{\sqrt{n}} \Big| \ge \eta/3 \Big). \tag{6.7}$$

Indeed, let us decompose $[0, 1]$ into intervals $[0, m/n]$, $[m/n, 2m/n], \dots, [(n-1)/n, 1]$. Then for $|t_1 - t_2| < \delta \le m/n$, the points t_1 and t_2 either fall into the same interval, or into two adjacent intervals, or otherwise their distance will be larger than m/n which is impossible. On the union of the events at the RHS of (6.7), if $t_1, t_2 \in [km, (k+1)/n]$, then

$$|\tilde{S}_n(t_1) - \tilde{S}(t_2)| \le |\tilde{S}_n(t_1) - S_k| + |\tilde{S}_n(t_2) - S_k| \le \frac{2\sqrt{n}\eta}{3},$$

or if $km \le t_1 < (k+1)m \le t_2 \le (k+2)m$,

$$|\tilde{S}_n(t_1) - \tilde{S}(t_2)| \le |\tilde{S}_n(t_1) - S_k| + |S_{k+1} - S_k| + |\tilde{S}_n(t_2) - S_{k+1}| \le \sqrt{n}\eta.$$

Now let us continue (6.7). Since X_i are i.i.d., all the summmands in the sum are the same and equal to the first one. We have

$$\begin{aligned}
\mu_n\{\omega(x, \delta) \ge \eta\} &\le \frac{2}{\delta} \mu_n\Big(\max_{\ell \le m} \frac{|S_\ell|}{\sqrt{n}} \le \eta/3 \Big) \\
&\le \frac{2}{\delta} \mu_n\Big(\max_{\ell \le m} \frac{|S_\ell|}{\sqrt{m}} \ge \eta/3\sqrt{\delta} \Big) \\
&= \frac{C}{\lambda^2} \mathsf{P}\Big(\max_{\ell \le m} |S_\ell| \ge \lambda\sqrt{m} \Big),
\end{aligned}$$

where $\lambda = \eta/3\sqrt{\delta}$. It remains to show for every $\varepsilon > 0$, there exists $\lambda > 0$ such that

$$\limsup_{m \to \infty} \lambda^2 \mathsf{P}\Big(\max_{\ell \le m} |S_\ell| \ge \lambda\sqrt{m} \Big) \le \varepsilon. \tag{6.8}$$

Note that by Kolmogorov's maximal inequality Proposition 4.17, at best we have

$$\lambda^2 \mathsf{P}\Big(\max_{\ell \le m} |S_\ell| \ge \lambda\sqrt{m} \Big) \le \mathrm{Var}(S_m)/m = \mathsf{E}X_1^2,$$

which cannot be made arbitrarily small.

We will use the following improvement of Proposition 4.17.

Lemma 6.4 *Let X_m be independent with $\mathsf{E}X_m = 0$ and $D_n^2 = \mathrm{Var}(S_n)$. Then*

$$\mathsf{P}\Big(\max_{1 \le k \le n} |S_k| \ge \lambda D_n \Big) \le 2\mathsf{P}\Big(|S_n| \ge (\lambda - \sqrt{2}) D_n \Big). \tag{6.9}$$

Let us postpone the proof of Lemma 6.4 and see first why it is helpful. If applying Chebyshev's inequality on the RHS of (6.9), it is not better than Kolmogorov's inequality. However, if X_m are i.i.d., then by the central limit thoerem, $S_n/\sqrt{n} \Rightarrow \mathcal{N}(0,1)$, and hence

$$\limsup_{n\to\infty} \frac{1}{\lambda^2}\mathsf{P}\Big(|S_n| \ge (\lambda-\sqrt{2})\sqrt{n}\Big) = \frac{2}{\lambda^2}\int_{(\lambda-\sqrt{2})}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx \le C\frac{1}{\lambda^2}e^{-\frac{(\lambda-\sqrt{2})^2}{2}}.$$

And (6.8) follows.

Proof of Lemma 6.4: Let $T = \inf\{T : |S_k| \ge \lambda\sqrt{n}D_n\}$. Then we have

$$\begin{aligned}\mathsf{P}(T \le n) &\le \mathsf{P}(S_n > (\lambda-\sqrt{2})D_n) + \sum_{k=1}^{n-1} \mathsf{P}(T=k,\ S_n < (\lambda-\sqrt{2})D_n)\\ &\le \mathsf{P}(S_n > (\lambda-\sqrt{2})D_n) + \sum_{k=1}^{n} \mathsf{P}(T=k,\ |S_n - S_k| > \sqrt{2}D_n).\end{aligned}$$

Since $\{T=k\} \in \sigma(X_1,\dots,X_k)$ and $|S_n - S_k| \in \sigma(X_{k+1},\dots,X_n)$, they are independent, so we can continue to get

$$\begin{aligned}\mathsf{P}(T \le n) &\le \mathsf{P}(S_n > (\lambda-\sqrt{2})D_n) + \sum_{k=1}^{n} \mathsf{P}(T=k)\mathsf{P}(|S_n - S_k| > \sqrt{2}D_n)\\ &\le \mathsf{P}(S_n > (\lambda-\sqrt{2})D_n) + \sum_{k=1}^{n} \mathsf{P}(T=k)\frac{1}{2D_n^2}\operatorname{Var}(S_n - S_k)\\ &\le \mathsf{P}(S_n > (\lambda-\sqrt{2})D_n) + \sum_{k=1}^{n} \mathsf{P}(T=k)\frac{1}{2}\\ &\le \mathsf{P}(S_n > (\lambda-\sqrt{2})D_n) + \frac{1}{2}\mathsf{P}(T \le n).\end{aligned}$$

Rearraging the terms, we obtain (6.9). □

Finally, let us mention that any sequential limit of μ_n has the same finite dimensional distribution, and hence the limit point is unique. Indeed, consider the bounded, continous function

$$F(x) := \exp\Big(i\big(\xi_1 x(t_1) + \xi_2(x(t_2)-x(t_1)) + \cdots + \xi_m(x(t_m)-x(t_{m-1}))\big)\Big),$$

where $\xi_k \in \mathbb{R}$ and $0 \le t_1 < \cdots < t_m \le 1$. Then by the functional CLT, $\int F(x)\,d\mu_n$ converge. On the other hand, by CLT, $\int F(x)\,d\mu_n$ as the ch.f. of the random vector $\Big(x_n(t_1), x_n(t_2)-x_n(t_1), \dots, x_n(t_m)-x_n(t_{m-1})\Big)$ will converge to the ch.f. of $\mathcal{N}(0, \operatorname{diag}\{t_1, t_2-t_1, \dots, t_m - t_{m-1}\})$. The charaterizes the f.d.d. of the Brownian motion.

6.3 *Tightness and weak-* convergence

(to be completed)

7 Poisson limit theorem and stable laws

After studying the CLT, one may wonder why is the normal distribution is so special, and what will happen if the i.i.d. r.v.s do not have second moment. In this section, we will investigate this problem.

As a prototype, we consider X_n i.i.d. with $\mathsf{P}(|X_1| > x) \sim x^{-\alpha}$ for some $\alpha < 2$. Note that $\alpha < 2$ implies that $\mathsf{E}|X_1|^2 = \infty$ since

$$\mathsf{E}|X_1|^2 = \int_0^\infty 2y\mathsf{P}(|X_1| \geq y)\, dy.$$

We want to study the weak limit of $\frac{S_n - b_n}{a_n}$, where a_n, b_n are properly chosen so that the limiting distribution is non-degenerate. Motivated by the functional CLT, we can be more ambitious by asking what is the function limit

$$L(t) = \lim_{n\to\infty} \frac{S_{[nt]} - b_n}{a_n}$$

as a random function of t.

The answer is that when $\mathsf{E}|X_1|^2 = \infty$, the function $L(t)$ is no longer continuous. If we record the locations, t, and the sizes, ℓ, of all the jumps, then we get a random point process in the (t, ℓ)-plane. This point process is a so-called *Poisson Point Process*, which enjoys the maximal degree of indepedence of all point processes.

7.1 Poisson limit theorem and Poisson point processes

Recall that $X \sim \mathrm{Poi}(\lambda)$ if $\mathsf{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$, $k \geq 0$. Here, the probabilities sum up to one thanks to the Taylar expansion $e^\lambda = \sum_{k=0}^\infty \frac{\lambda^k}{k!}$. The ch.f. of X is then given by

$$\mathsf{E}e^{i\xi X} = \sum_{k=0}^\infty (e^{i\xi})^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda(e^{i\xi}-1)}. \tag{7.1}$$

The Poisson distributin models the cummulative effect of many rare events, as the following result shows.

Theorem 7.1 (Poisson Limit Theorem) *Let $X_{n,m} \sim \mathrm{Ber}(p_{n,m})$ be independent. Assume that*

$$\sum_{m=1}^n p_{n,m} \to \lambda, \quad n \to \infty,$$

and

$$\max_{1\leq m\leq n} |p_{n,m}| \to 0, \quad n \to \infty. \tag{7.2}$$

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

Proof: We will prove this by computing the ch.f.

By independence, we have

$$\mathsf{E}e^{i\xi S_n} = \prod_{m=1}^n \left[(1 - p_{n,m}) + p_{n,m}e^{i\xi}\right] = \prod_{m=1}^n \left[1 + p_{n,m}(e^{i\xi} - 1)\right].$$

By (7.2), $|p_{n,m}(e^{i\xi} - 1)| \leq 1/2$ for large n. Using Lemma 5.20, we have

$$\left|\log \mathsf{E}e^{i\xi S_n} - \sum_{m=1}^n p_{n,m}(e^{i\xi} - 1)\right| \leq \sum_{m=1}^n \left|p_{n,m}(e^{i\xi} - 1)\right|^2 \leq \max|p_{n,m}| \cdot \sum_{m=1}^n p_{n,m} \to 0.$$

This completes the proof. □

Next we define the Poisson point process on the measurable space $(H, \mathcal{H}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. In the sequal the cases $d = 1$ or 2 are most relevant to us.

A *Poisson point process (PPP)* on $(H, \mathcal{H})$ is a random counting measure ν on H, that is, for every $C \in \mathcal{H}$, $\nu(C)$ is a r.v. taking values in $\{0, 1, 2, \dots\} \cup \{\infty\}$. The quantity $\nu(C)$ counts how many points fall into the set C, since the location of points are random, $\nu(C)$ is also random. Moreover, the distribution of all $\nu(C)$ is characterized by the following.

1. For every $C \in \mathcal{H}$, either $\mathsf{E}\nu(C) = \infty$, or $\nu(C) \sim \mathrm{Poi}(\mathsf{E}\nu(C))$. We define $\lambda(C) \coloneqq \mathsf{E}\nu(C)$. Then $\lambda(C)$ is a deterministic measure. We call λ the *intensity* of the PPP ν.

2. For disjoint $C_1, \dots, C_n \in \mathcal{H}$, the r.v.s $\nu(C_1), \dots, \nu(C_n)$ are independent.

Example 7.1 (Poisson process) Let $H = \mathbb{R}$ and λ be the Lebesgue measure on $\mathbb{R}$. The PPP can be charaterized by $N_t = \nu(0, t]$. The process N_t is the *Poisson process*. The point process models the arrival times of customers, where the waiting time for the next customer are i.i.d. Exp(1) r.v.s.

Example 7.2 (Compound Poisson) Let Y_n be i.i.d. and N_t be the Poisson process, independent of all Y_n. The r.v.

$$Z_t = \sum_{m=1}^{N_t} Y_m$$

is called a *compound Poisson*. Note that Z_t can be represented as an integral against a PPP on $\mathbb{R}^2$:

$$Z_t = \int_{[0,t]\times\mathbb{R}} \ell\, \nu(dtd\ell),$$

where ν is a PPP on $\mathbb{R}^2$ with intensity $\lambda = \mathrm{Leb} \otimes \mu_Y$.

Example 7.3 We can further generalize the compound Poisson as follows. Let ν be a PPP with intensity λ, and let f be a measurable function such that $\int_H |f(z)|\, \lambda(dz) < \infty$. Then we can study the r.v.

$$Z = \int f(z)\, \nu(dz). \tag{7.3}$$

7.2 stable law limit theorem

Let $X_n \geq 0$ be i.i.d. with $\mathsf{P}(X_1 > x) \sim x^{-\alpha}$ where $\alpha < 2$.

Proposition 7.2 *Let $0 < a < b$. Then*

$$\#\{1 \leq m \leq n : X_m/n^{1/\alpha} \in (a, b)\} \Rightarrow \mathrm{Poi}(a^{-\alpha} - b^{-\alpha}). \tag{7.4}$$

Proof: The LHS of (7.4) can be written as sum of i.i.d. Bernoulli random variables, $\sum_{m=1}^n \xi_{n,m}$, where

$$\xi_{n,m} = \mathbb{1}_{\{X_m/n^{1/\alpha} \in (a,b)\}} \sim \mathrm{Ber}(p_n),$$

and

$$p_n = \mathsf{P}(X_1/n^{1/\alpha} \in (a, b)) \sim (a^{-\alpha} - b^{-\alpha})n^{-1}.$$

The conclusion then follows from Theorem 7.1. □

In fact, a much stronger statement holds. Consider the rescaled point process given by

$$\nu_n = \sum_{m=1}^{n} \delta_{(\frac{m}{n}, \frac{X_m}{n^{1/\alpha}})}. \tag{7.5}$$

Proposition 7.2 says that for every rectangle $R = (a,b) \times [0,1]$, $\nu_n(R) \Rightarrow \mathrm{Poi}\big(\lambda(R)\big)$, where

$$\lambda(dtd\ell) = \mathrm{Leb} \otimes (\mathbb{1}_{\ell>0} \alpha \ell^{-\alpha+1}\, d\ell). \tag{7.6}$$

One can show that the $\nu_n \Rightarrow$ a PPP with intensity λ given in (7.6).

With the representation (7.5), we can express $\frac{S_{[nt]}}{n^{1/\alpha}}$ as

$$\frac{S_{[nt]}}{n^{1/\alpha}} = \int_{[0,t]\times\mathbb{R}} \ell\, \nu_n(dtd\ell).$$

In particular, the limit of $\frac{S_n}{\sqrt{n^{1/\alpha}}}$ should be related to

$$\int_{[0,1]\times\mathbb{R}} \ell\, \nu(dtd\ell), \tag{7.7}$$

where ν is the PPP with intensity (7.6). This is a r.v. taking the form (7.3).

If we consider a more general tail condition

$$\mathsf{P}(X_1 > x) \sim \theta x^{-\alpha}, \quad \mathsf{P}(X_1 < -x) \sim (1-\theta) x^{-\alpha}, \tag{7.8}$$

then the intensity of the corresponding PPP will be $\lambda = \mathrm{Leb} \otimes \lambda_{\alpha,\theta}$ where

$$\lambda_{\alpha,\theta} := \Big(\mathbb{1}_{\{\ell<0\}} \alpha(1-\theta)\ell^{-\alpha-1} + \mathbb{1}_{\{\ell>0\}} \alpha\theta \ell^{-\alpha-1} \Big)\, d\ell. \tag{7.9}$$

Theorem 7.3 (Stable law limit theorem) *Let X_n be i.i.d. that satisfy the tail condition (7.8). Let*

$$a_n = \inf\{x : \mathsf{P}(|X_1| > x) \le n^{-1}\} \sim n^{1/\alpha}, \quad b_n = n\mathsf{E}X_1 \mathbb{1}_{\{|X_1|<a_n\}}.$$

Then $\frac{S_n - b_n}{a_n} \Rightarrow Y$, where Y is a non-degenerate r.v. with ch.f.

$$\mathsf{E}e^{i\xi Y} = \exp\Big(i\xi c + \int_{-\infty}^{\infty} (e^{i\xi x} - 1 - \frac{i\xi x}{1+x^2}) \lambda_{\alpha,\theta}(dx) \Big) \tag{7.10}$$

where $c \in \mathbb{R}$ and $\lambda_{\alpha,\theta}$ is given by (7.9).

The proof of Theorem 7.3 is quite technical and we will try to understand the statement rather than prove it. The connection between (7.10) and (7.7) can be seen through the following computation.

Proposition 7.4 *Let ν be a PPP with intensity λ, and let f be a measurable function such that $\int |f|\, d\lambda < \infty$. Then*

$$\mathsf{E}e^{i\xi \int f\, d\nu} = \exp\Big(\int (e^{i\xi f(x)} - 1)\, \lambda(dx) \Big). \tag{7.11}$$

Proof: We will prove (7.11) for f an indicator function, a simple function, and then a general function in $L^1(d\lambda)$.

First, let $f = \mathbb{1}_A$. Then $\int f\, d\nu = \nu(A) \sim \mathrm{Poi}(\lambda(A))$ by the definition of PPP. Hence, by (7.1),

$$\mathsf{E}e^{i\xi \int f\, d\nu} = \exp\big(\lambda(A)(e^{i\xi} - 1)\big) = \exp\Big(\int (e^{i\xi \mathbb{1}_A(z)} - 1)\, \lambda(dz) \Big),$$

since $e^{i\xi \mathbb{1}_A(z)} - 1 = 0 \Leftrightarrow \mathbb{1}_A(z) = 0$.

Next, let $f(x) = \sum_{k=1}^n c_k \mathbb{1}_{A_k}(x)$ to be a simple function. We can assume that A_k are disjoint. By definition of PPP, $\nu(A_k)$ are independent $\mathrm{Poi}(\lambda(A_k))$ r.v.s, so we have

$$\mathsf{E} e^{i\xi \sum_{k=1} c_k \nu(A_k)} = \prod_{k=1}^n e^{\lambda(A_k)(e^{i\xi c_k}-1)}.$$

It is easy to check that

$$\sum_{k=1}^n \lambda(A_k)(e^{i\xi c_k} - 1) = \int_H (e^{i\xi f(z)} - 1)\, \lambda(dz).$$

Finally, for a general function f, we can approximate it by f_n simple, with $|f_n| \le |f|$ and $f_n(z) \to f(z)$ for every z. It suffices to show that we can pass the limit on both sides of (7.11).

Note that $\int |f|\, d\lambda < \infty$ implies that

$$\mathsf{E} \int |f|\, d\nu = \int |f|\, d\lambda < \infty,$$

so $\int |f|\, d\nu < \infty$ for almost every ν, and hence by DCT on such ν,

$$\lim_{n\to\infty} \int f_n(z)\, \nu(dz) = \int f(z)\, \nu(dz).$$

Then by BCT,

$$\lim_{n\to\infty} \mathsf{E} e^{i\xi \int f_n(z)\, \nu(dz)} = \mathsf{E} e^{i\xi \int f(z)\, \nu(dz)}.$$

For the right hand side, since

$$|e^{i\xi f_n(z)} - 1| \le |\xi| \cdot |f_n(z)| \le |\xi| \cdot |f_n(z)|,$$

it follows from the DCT that

$$\lim_{n\to\infty} \int (e^{i\xi f_n(z)} - 1)\, \lambda(dz) = \int (e^{i\xi f(z)} - 1)\, \lambda(dz).$$

□

From Proposition 7.2, in (7.10) the term

$$\int_{-\infty}^{\infty} (e^{i\xi x} - 1)\, \lambda_{\alpha,\theta}(dx)$$

will correspond to

$$\int_{[0,1]\times\mathbb{R}} \ell \nu(dtd\ell) \approx S_n / n^{1/\alpha},$$

where ν is PPP with intensity $\lambda_{\alpha,\theta}$. The extra term $\frac{i\xi x}{1+x^2}$ is to compesate the asmmetry of the tail of X_1; in the symmetric case $\theta = 1/2$, the term can be dropped, since it is odd and $\lambda_{\alpha,\theta}$ is even.

How do we understand the index α? In the proof of Theorem 7.3, one needs to the contribution from small jumps and from big jumps. In (7.10) only contribution from big jumps matters; they appear in the limiting PPP. We can compare the contribution from small jumps and big jumps from the form of $\lambda_{\alpha,\theta}$. For simplicity, let us say the law of X_1 is symmetric, and thus $\theta = 1/2$. When $\alpha < 1$, the sum of small jumps are negligible compared to large jumps, since

$$\mathsf{E} \Big| \int_{-\varepsilon}^{\varepsilon} \ell\, \nu(d\ell) \Big| \le \int_0^{\varepsilon} \ell \cdot \frac{\alpha}{\ell^{-\alpha-1}}\, d\ell < \infty, \tag{7.12}$$

while the large jumps contribute much more since

$$\int_{\varepsilon}^{\infty} \ell^{-\alpha}\, d\ell = \infty.$$

When $\alpha \in (1,2]$, the RHS of (7.12) is also ∞, but if we consider the cancellation of positive and negative jumps, as the Kolmogorov's one-series theorem suggest, we should integrate ℓ^2 rather than ℓ. Since

$$\int_{0}^{\varepsilon} \ell^2 \ell^{-\alpha-1}\, d\ell < \infty, \tag{7.13}$$

the sum of small jumps still converges, while the sum of large jumps diverges since

$$\int_{\varepsilon}^{\infty} \ell^2 \ell^{-\alpha-1}\, d\ell = \infty. \tag{7.14}$$

But when $\alpha > 2$, the integral (7.13) will diverge, while the integral (7.14) is finite, so the main contribution to the sum S_n is from small jumps, and since all jumps are small, we see a continuous distribution as the limit; this intuition agrees with the Linderburg's condition (5.24) and the equi-continuity estimate (6.2) when we derive the functional CLT.

7.3 Stable laws and infinite divisible laws

Random variables with ch.f.s of the form (7.10) are called *stable laws*. Besides the ch.f., an "official" definition for the stable law is the following.

Definition 7.1 *A r.v. Y has stable law if for every k, there exist a_k, b_k such that*

$$\frac{Y_1 + \cdots + Y_k - b_k}{a_k} \overset{\mathrm{d}}{=} Y,$$

where Y_j are i.i.d. and $Y_j \overset{\mathrm{d}}{=} Y$. Here, $X_1 \operatorname{dist} X_2$ means that X_1 and X_2 have the same distribution.

The next theorem explains the word "stable".

Theorem 7.5 *A r.v. Y has stable law if and only if there exist i.i.d. r.v. X_n, constants a_n and b_n such that*

$$\frac{X_1 + \cdots + X_n - b_n}{a_n} \Rightarrow Y.$$

Proof: The "only if" part follows from the definition with $X_n \overset{\mathrm{d}}{=} Y$.

For the "if" part, we only explain the intuition.

If X_n, a_n and b_n exist, then for each k,

$$\begin{aligned} Y &\overset{\mathrm{d}}{=} \lim_{n\to\infty} \frac{X_1 + \cdots + X_{kn} - b_{kn}}{a_{kn}} \\ &= \lim_{n\to\infty} \frac{\frac{X_1+\cdots+X_n-b_n}{a_n} + \cdots + \frac{X_{(k-1)n+1}+\cdots+X_{kn}-b_n}{a_n} + \frac{kb_n-b_{kn}}{a_{kn}}}{a_{kn}/a_n} \\ &\overset{\mathrm{d}}{=} \frac{Y_1 + \cdots + Y_k - \tilde{a}_k}{\tilde{b}_k}, \end{aligned}$$

where

$$\tilde{a}_k = \lim_{n\to\infty} a_{kn}/a_n, \quad \tilde{b}_k = \lim_{n\to\infty} \frac{kb_n - b_{kn}}{a_{kn}}. \tag{7.15}$$

So after we show that the two limits in (7.15) indeed exist, we know that Y has stable law. □

A closely related concept is *infinitely divisible law*. A r.v. Y has infinitely divisible law if for every $n \geq 1$, there exists $X_{n,1}, \dots, X_{n,n}$ i.i.d. such that $Y \stackrel{\mathrm{d}}{=} X_{n,1} + \cdots + X_{n,n}$. Clearly, a stable law is infinitely divisible, by taking $X_{n,k} = \frac{Y_k - b_n/n}{a_n}$. Also, Y is infinitely divisible if and only if for every $n \geq 1$, the n-th root of its ch.f. is also a ch.f. for some r.v. $(X_{n,1})$. There is a characterization of infinitely divisible laws similar to Theorem 7.5.

Theorem 7.6 *A r.v. Y has infinitely divisible law if and only if there exist i.i.d. $X_{n,k}$ such that $X_{n,1} + \cdots + X_{n,n} \Rightarrow Y$.*

The celebrated Levy–Khinchin Theorem completely characterized the ch.f.s for infinitely divisible law.

Theorem 7.7 *[Levy–Khinchin] A r.v. Y has infinitely divisible law if and only if its ch.f. takes the form*

$$\log \varphi(\xi) = ic\xi - \frac{\sigma^2}{2}\xi + \int (e^{i\xi x} - 1 - \frac{i\xi x}{1+x^2})\, \mu(dx),$$

where $c, \sigma \in \mathbb{R}$ and μ is a measure with $\mu\{0\} = 0$ and $\int \frac{x^2}{1+x^2}\, \mu(dx) < \infty$.

We conclude by some examples of infinite divisible laws.

Example 7.4
1. Normal distribution. The ch.f. is $e^{i\mu\xi - \frac{1}{2}\sigma^2\xi^2}$.
2. Stable laws. The ch.f. is given in (7.10).
3. Poisson distribution. The ch.f. is $e^{\lambda(e^{i\xi}-1)}$.
4. Compound Poisson or integral against a PPP. The ch.f. is given in Proposition 7.4.

8 Martingales

8.1 Conditional expectation

8.1.1 Definition

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and $\mathcal{G} \subset \mathcal{F}$ be a sub–σ-algebra. Let X be a r.v. with $\mathsf{E}|X| < \infty$. The *conditional expectation* $Y = \mathsf{E}[X \,|\, \mathcal{G}]$ is a r.v. that satisfies the following two properties:

$$Y \text{ is } \mathcal{G}\text{-measurable},$$

$$\mathsf{E}Y\mathbb{1}_A = \mathsf{E}X\mathbb{1}_A, \quad \forall A \in \mathcal{G}. \tag{8.1a}$$

Such r.v.s Y are not unique. Each of them is called a *version* of $\mathsf{E}[X \,|\, \mathcal{G}]$.

Let us look at a very simple σ-algebra $\mathcal{G} = \{\varnothing, B, B^c, \Omega\}$ and take $X = \mathbb{1}_A$. It is not hard to show that any $\mathcal{G}$-measurable map can be written as a linear combination of $\mathbb{1}_B$ and $\mathbb{1}_{B^c}$. Therefore,

$$\mathsf{E}[\mathbb{1}_A \,|\, \mathcal{G}](\omega) = c_1\mathbb{1}_B(\omega) + c_2\mathbb{1}_B(\omega),$$

And we need to determine c_1 and c_2. Since B and B^c are in $\mathcal{G}$, by (8.1a) we have

$$\mathsf{E}\mathbb{1}_A\mathbb{1}_B = \mathsf{E}\Big(c_1\mathbb{1}_B + c_2\mathbb{1}_{B^c}\Big)\mathbb{1}_B = c_1\mathsf{P}(B), \quad \mathsf{E}\mathbb{1}_A\mathbb{1}_{B^c} = \mathsf{E}\Big(c_1\mathbb{1}_B + c_2\mathbb{1}_{B^c}\Big)\mathbb{1}_{B^c} = c_2\mathsf{P}(B^c).$$

So

$$c_1\mathsf{P}(B) = \mathsf{P}(AB), \quad c_2\mathsf{P}(B^c) = \mathsf{P}(AB^c).$$

There are three cases.

1. If $\mathsf{P}(B), \mathsf{P}(B^c) \neq 0$, then $c_1 = \frac{\mathsf{P}(AB)}{\mathsf{P}(B)} = \mathsf{P}[A \,|\, B]$, $c_2 = \frac{\mathsf{P}(AB^c)}{\mathsf{P}(B^c)} = \mathsf{P}[A \,|\, B^c]$ are the classical conditional probabilties.

2. If $\mathsf{P}(B) = 0$ (and necessarily $\mathsf{P}(B^c) = 1$), then c_1 can be arbitrary and $c_2 = \mathsf{P}(A)$, but the conditional expectation is only undetermined on a zero measure set B.

3. If $\mathsf{P}(B^c) = 0$ (and necessarily $\mathsf{P}(B) = 1$), then c_2 can be arbitrary and $c_1 = \mathsf{P}(A)$, and this is similar to the previous case.

Note that from elementary probability, we also avoid $\mathsf{P}[A \,|\, B]$ if $\mathsf{P}(B) = 0$.

As a generalization of the previous example, suppose we can partition the sample space Ω into disjoint union of at most countably many sets $\Omega = \bigcup_{n=1}^{N} \Omega_n$, where $\mathsf{P}(\Omega_n) > 0$ and $1 \leq N \leq \infty$. Let $\mathcal{G} = \sigma(\Omega_n,\ n \geq 1)$. Then we have

$$\mathsf{E}[X \,|\, \mathcal{G}](\omega) = \frac{\mathsf{E}X\mathbb{1}_{\Omega_n}}{\mathsf{P}(\Omega_n)}, \quad \omega \in \Omega_n.$$

In particular, when $X = \mathbb{1}_A$, the *conditional probability* of A w.r.t. G is defined by

$$\mathsf{P}[A \,|\, \mathcal{G}](\omega) \coloneqq \mathsf{P}[A \,|\, \Omega_n] = \frac{\mathsf{P}(A\Omega_n)}{\mathsf{P}(\Omega_n)}, \quad \omega \in \Omega_n.$$

8.1.2 Uniqueness and Existence

Since r.v.s are defined up to measure zero set, our best hope is that conditional expection is unique in the almost sure sense. We start with a simple lemma.

Lemma 8.1 *Let Y be a version of $\mathsf{E}[X \,|\, \mathcal{G}]$. Then $\mathsf{E}|Y| \leq \mathsf{E}|X|$.*

Proof: Since $Y \in \mathcal{G}$, we have $A = \{Y \geq 0\} \in \mathcal{G}$. By (8.1a), we have

$$\mathsf{E}Y^+ = \mathsf{E}Y\mathbb{1}_A = \mathsf{E}X\mathbb{1}_A \leq \mathsf{E}|X|\mathbb{1}_A.$$

Similarly, $A^c \in \mathcal{G}$ and we have

$$\mathsf{E}Y^- = -\mathsf{E}Y\mathbb{1}_{A^c} = -\mathsf{E}X\mathbb{1}_{A^c} \leq \mathsf{E}|X|\mathbb{1}_{A^c}.$$

□

Proof of a.s. uniqueness of conditional expectation: Let Y and Y' be two versions of $\mathsf{E}[X \,|\, \mathcal{G}]$. For every $\varepsilon > 0$, let $A_\varepsilon = \{Y - Y' \geq \varepsilon\} \in \mathcal{G}$. By (8.1a), we have

$$\mathsf{P}(A_\varepsilon) \leq \mathsf{E}(Y - Y')\mathbb{1}_{A_\varepsilon} = \mathsf{E}X\mathbb{1}_{A_\varepsilon} - \mathsf{E}X\mathbb{1}_{A_\varepsilon} = 0.$$

Therefore,

$$\mathsf{P}(Y - Y' > 0) \leq \sum_{n=1}^{\infty} \mathsf{P}(A_{1/n}) = 0.$$

By symmetry we also have $\mathsf{P}(Y' - Y > 0) = 0$. Hence, $\mathsf{P}(Y' = Y) = 1$, as desired. □

For the existence of conditional expectation, we need the Radon–Nikodym Theorem, Theorem 1.17. Let

$$\nu_\pm(A) = \mathsf{E}X^\pm\mathbb{1}_A, \quad A \in \mathcal{G}.$$

For any disjoint $A_n \in \mathcal{G}$, since $\mathsf{E}|X| < \infty$ and $|X| \geq \sum_{n=1}^N X^{\pm}\mathbb{1}_{A_n}$, by DCT, we have

$$\nu_{\pm}(\bigcup_{n=1}^{\infty} A_n) = \mathsf{E}\lim_{N\to\infty}\sum_{n=1}^{N} X^{\pm}\mathbb{1}_{A_n} = \lim_{N\to\infty}\sum_{n=1}^{N}\mathsf{E}X^{\pm}\mathbb{1}_{A_n} = \sum_{n=1}^{\infty}\nu_{\pm}(A_n).$$

Also, $\nu_{\pm}(\Omega) \leq \mathsf{E}|X| < \infty$. So $\nu_{\pm}$ are finite measures. Clearly, $\nu_{\pm} \ll \mathsf{P}$. By Theorem 1.17, there exist r.v.s $Y^{\pm} \in \mathcal{G}$ such that $\nu_{\pm}(A) = \mathsf{E}Y^{\pm}\mathbb{1}_A$. Let $Y = Y^+ - Y^-$. Then Y is a version of $\mathsf{E}[X\,|\,\mathcal{G}]$.

Let $X \in L^1$ and Y be an arbitrary r.v. We use $\mathsf{E}[X\,|\,Y]$ to denote the conditional expectation $\mathsf{E}\big[X\,\big|\,\sigma(Y)\big]$, since any $\sigma(Y)$-measurable map can be written in the form $h(Y)$ where h is a Borel measurable function. As another example, let (X, Y) be 2d random vector with density $f(x, y)$, that is, for any $B \in \mathcal{B}(\mathbb{R}^2)$,

$$\mathsf{P}\big((X, Y) \in B\big) = \int_B f(x, y)\,dxdy.$$

Proposition 8.2 *Let g be bounded measurable. Then $\mathsf{E}\big[g(X)\,\big|\,Y\big] = h(Y)$, where*

$$h(y) = \begin{cases} \frac{\int g(x)f(x,y)\,dx}{\int f(x',y)\,dx'}, & \int f(x, y)\,dx \neq 0, \\ c, & \int f(x, y)\,dx = 0, \end{cases} \tag{8.2}$$

where $c \in \mathbb{R}$ is arbitrary.

This means that the conditional law "$\mathcal{L}[\cdot\,|\,Y]$" has density

$$\rho_{X|Y}(x|y) = \frac{f(x, y)}{\int f(x', y)\,dx'}.$$

Proof: Since $\sigma(Y) = \{Y^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^2)\}$, for any $A \in \sigma(Y)$, there exists some B_0 such that $A = Y^{-1}(B_0)$, and $\mathbb{1}_A(\omega) = \mathbb{1}_{B_0}\big(Y(\omega)\big)$. We have

$$\begin{aligned}
\int h\big(Y(\omega)\big)\mathbb{1}_A(\omega)\,\mathsf{P}(d\omega) &= \int h\big(Y(\omega)\big)\mathbb{1}_{B_0}\big(Y(\omega)\big)\,\mathsf{P}(d\omega) \\
&= \int h(y)\mathbb{1}_{B_0}(y)f(x, y)\,dxdy \\
&= \int \mathbb{1}_{B_0}(y)\,dy\Big[\int h(y)f(x, y)\,dx\Big] \\
&= \int \mathbb{1}_{B_0}(y)\,dy\int g(x)f(x, y)\,dx \\
&= \int g(x)\mathbb{1}_{B_0}(y)f(x, y)\,dxdy = \int g\big(X(\omega)\big)\mathbb{1}_A\big(Y(\omega)\big)\,\mathsf{P}(d\omega).
\end{aligned}$$

That is, $\mathsf{E}h(Y)\mathbb{1}_A = \mathsf{E}g(X)\mathbb{1}_A$. In the fouth line, we have used

$$\int h(y)f(x, y)\,dx = \int g(x)f(x, y)\,dx, \quad \forall y,$$

which follows from (8.2). Therefore, $h(Y) = \mathsf{E}[X\,|\,Y]$. This completes the proof. □

8.1.3 Properties of conditional expectation

Proposition 8.3 *Let* $\mathsf{E}|X|, \mathsf{E}|Y| < \infty$.

1. *(linearity) For all* $a, b \in \mathbb{R}$

$$\mathsf{E}[aX + bY \,|\, \mathcal{G}] = a\mathsf{E}[X \,|\, \mathcal{G}] + b\mathsf{E}[Y \,|\, \mathcal{G}].$$

2. *(order preserving) If* $X \le Y$, *then*

$$\mathsf{E}[X \,|\, \mathcal{G}] \le \mathsf{E}[Y \,|\, \mathcal{G}], \quad a.s.$$

3. *(conditional MCT) Let* $X_n \ge 0$, $X_n \uparrow X$ *and* $\mathsf{E}X < \infty$. *Then*

$$\mathsf{E}[X_n \,|\, \mathcal{G}] \uparrow \mathsf{E}[X \,|\, \mathcal{G}], \quad a.s.$$

Proof: For Item 1, we have for every $A \in \mathcal{G}$,

$$\begin{aligned}\mathsf{E}\big(\mathbb{1}_A \cdot \mathsf{E}[aX + bY \,|\, \mathcal{G}]\big) &= \mathsf{E}\mathbb{1}_A(aX + bY) = a\mathsf{E}\mathbb{1}_A X + b\mathsf{E}\mathbb{1}_A Y \\ &= a\mathsf{E}\big(\mathbb{1}_A \cdot \mathsf{E}[X \,|\, \mathcal{G}]\big) + b\mathsf{E}\big(\mathbb{1}_A \cdot \mathsf{E}[Y \,|\, \mathcal{G}]\big) = \mathsf{E}\Big(\mathbb{1}_A \cdot \big(a\mathsf{E}[X \,|\, \mathcal{G}] + b\mathsf{E}[Y \,|\, \mathcal{G}]\big)\Big).\end{aligned}$$

For Item 2, consider $A_\varepsilon = \{\mathsf{E}[X \,|\, \mathcal{G}] - \mathsf{E}[Y \,|\, \mathcal{G}] > \varepsilon\}$ and proceed as in the proof of uniqueness.
For Item 3, let $Y_n = \mathsf{E}[X_n \,|\, \mathcal{G}]$. Then by Item 2, $Y_n \uparrow$ almost surely. Let Y be the a.s. limit of Y_n. For any $A \in \mathcal{G}$, since $Y_n\mathbb{1}_A \uparrow Y\mathbb{1}_A$ a.s., by MCT we have

$$\lim_{n\to\infty} \mathsf{E}Y_n\mathbb{1}_A = \mathsf{E}Y\mathbb{1}_A.$$

Since $X_n\mathbb{1}_A \uparrow X\mathbb{1}_A$, by MCT we have

$$\lim_{n\to\infty} \mathsf{E}X_n\mathbb{1}_A = \mathsf{E}X\mathbb{1}_A.$$

But $\mathsf{E}X_n\mathbb{1}_A = \mathsf{E}Y_n\mathbb{1}_A$ by definition. Therefore, $\mathsf{E}X\mathbb{1}_A = \mathsf{E}Y\mathbb{1}_A$ for every $A \in \mathcal{G}$, so $Y = \mathsf{E}[X \,|\, \mathcal{G}]$. □

Using Item 3, it is not hard to establish Fatou's lemma and DCT for conditional expectations.

Proposition 8.4 (Jensen inequality for conditional expectation) *Let* φ *be convex. Suppose that* $\mathsf{E}|X| < \infty$ *and* $\mathsf{E}\varphi(X) < \infty$. *Then*

$$\mathsf{E}\big[\varphi(X) \,\big|\, \mathcal{G}\big] \ge \varphi\big(\mathsf{E}[X \,|\, \mathcal{G}]\big), \quad a.s.$$

Proof: Recall that in the proof of the unconditional version Proposition 1.29, we take expectation of $\varphi(x) \ge ax + b$ where the equality is achieved by $x = \mathsf{E}X$; the coefficients will then depend on $\mathsf{E}X$. Here, however, $\mathsf{E}[X \,|\, \mathcal{G}]$ potentially can take any value in $\mathbb{R}$, and a and b will change correspondingly. What is worse, it is impossible to take conditional expectation of the inequality $\varphi(x) \ge ax + b$, since a and b vary with every choice of $\mathsf{E}[X \,|\, \mathcal{G}]$, and the union of these uncountably many measure zero exceptional sets can cease to be measure zero.

Here is the fix. A convex function φ can be characterized by all the straight line below it, and we can take a countable family of them to determine φ. More precisely,

$$\varphi(x) = \sup\{ax + b : a, b \in \mathbb{Q},\ \varphi(t) \ge at + b,\ \forall t\}. \tag{8.3}$$

For every (a, b) in (8.3), by Proposition 8.3 there is a $N_{a,b}$ with $\mathsf{P}(N_{a,b}) = 0$ such that

$$\mathsf{E}\big[\varphi(X) \,\big|\, \mathcal{G}\big](\omega) \ge \mathsf{E}[aX + b \,|\, \mathcal{G}] = a\mathsf{E}[X \,|\, \mathcal{G}](\omega) + b, \quad \forall \omega \in N_{a,b}^c. \tag{8.4}$$

Hence, when $x \notin N = \bigcup_{a,b} N_{a,b}$, by (8.3) and (8.4) we have

$$\mathsf{E}\big[\varphi(X) \,\big|\, \mathcal{G}\big](\omega) \geq \varphi\big(\mathsf{E}[X \,|\, \mathcal{G}](\omega)\big). \tag{8.5}$$

On the other hand, by σ-subadditiviy, $\mathsf{P}(N) \leq \sum_{a,b} \mathsf{P}(N_{a,b}) = 0$, and hence (8.5) holds a.s. □

Below we prove some other useful properties for conditional expectation.

Proposition 8.5 $\mathsf{E}\big(\mathsf{E}[X \,|\, \mathcal{G}]\big) = \mathsf{E}X$.

Proof: It follows from (8.1a) by taking $A = \Omega$. □

Proposition 8.6 *Let $p \geq 1$. Then $\mathsf{E}|X|^p \geq \mathsf{E}\big|\mathsf{E}[X \,|\, \mathcal{G}]\big|^p$.*

When $p = 1$, this is contained in the proof of Lemma 8.1.

Proof: Since $x \mapsto |x|^p$ is convex for $p \geq 1$, by Proposition 8.4 we have

$$\mathsf{E}[|X|^p \,|\, \mathcal{G}] \geq \big|\mathsf{E}[X \,|\, \mathcal{G}]\big|^p.$$

Taking expectation of both sides and using Proposition 8.5, we obtain the statement. □

Proposition 8.7 *If $Y \in \mathcal{G}$, then*

$$\mathsf{E}[XY \,|\, \mathcal{G}] = Y\mathsf{E}[X \,|\, \mathcal{G}]. \tag{8.6}$$

Proof: Let $\mathcal{H}$ be the collection of Y's such that (8.6) holds. Since we have linearity and MCT for conditional expectation (Proposition 8.3), it suffices to show that $\mathcal{H}$ contains all indicator functions.

Let $Y = \mathbb{1}_B$ where $B \in \mathcal{G}$. Let $A \in \mathcal{G}$ and we need to check (8.1a). Indeed,

$$\mathsf{E}\big(\mathbb{1}_A \cdot \mathsf{E}[XY \,|\, \mathcal{G}]\big) = \mathsf{E}XY\mathbb{1}_A = \mathsf{E}X\mathbb{1}_{A\cap B} = \mathsf{E}\big(\mathbb{1}_{A\cap B}\mathsf{E}[X \,|\, \mathcal{G}]\big) = \mathsf{E}\Big(\mathbb{1}_A \cdot \big(Y\mathsf{E}[X \,|\, \mathcal{G}]\big)\Big).$$

This completes the proof. □

Proposition 8.8 *If X is independent of $\mathcal{G}$, then $\mathsf{E}[X \,|\, \mathcal{G}] = \mathsf{E}X$, and if $X \in \mathcal{G}$, then $\mathsf{E}[X \,|\, \mathcal{G}] = X$.*

Proof: Let $A \in \mathcal{G}$. If X and $\mathcal{G}$ are indepedent, then

$$\mathsf{E}\big(\mathbb{1}_A \cdot \mathsf{E}[X \,|\, \mathcal{G}]\big) = \mathsf{E}X\mathbb{1}_A = \mathsf{E}X \cdot \mathsf{E}\mathbb{1}_A = \mathsf{E}(\mathbb{1}_A \cdot \mathsf{E}X).$$

This proves the first statement. The second statme is obvious.

□

Proposition 8.9 *Let $\mathcal{G}_1 \subset \mathcal{G}_2$. Then*

$$\mathsf{E}\big[\mathsf{E}[X \,|\, \mathcal{G}_1] \,\big|\, \mathcal{G}_2\big] = \mathsf{E}\big[\mathsf{E}[X \,|\, \mathcal{G}_2] \,\big|\, \mathcal{G}_1\big] = \mathsf{E}[X \,|\, \mathcal{G}_1].$$

Proof: Since $\mathsf{E}[X \,|\, \mathcal{G}_1] \in \mathcal{G}_1 \subset \mathcal{G}_2$, we have $\mathsf{E}\big[\mathsf{E}[X \,|\, \mathcal{G}_1] \,\big|\, \mathcal{G}_2\big] = \mathsf{E}[X \,|\, \mathcal{G}_1]$.

For the second one, let $A \in \mathcal{G}_1$, and we have

$$\mathsf{E}\big(\mathbb{1}_A \cdot \mathsf{E}[X \,|\, \mathcal{G}_2]\big) = \mathsf{E}\mathsf{E}[\mathbb{1}_A X \,|\, \mathcal{G}_2] = \mathsf{E}\mathbb{1}_A X = \mathsf{E}\big(\mathbb{1}_A \cdot \mathsf{E}[X \,|\, \mathcal{G}_1]\big),$$

and hence $\mathsf{E}\big[\mathsf{E}[X \,|\, \mathcal{G}_2] \,\big|\, \mathcal{G}_1\big] = \mathsf{E}[X \,|\, \mathcal{G}_1]$. □

The conditional expectation can also be understood as a projection in a Hilbert space. This also leads to a proof of Theorem 1.17 using Hilbert space theory. See LAX.

Proposition 8.10 *If* $\mathsf{E}X^2 < \infty$, *then*

$$\mathsf{E}\big(X - \mathsf{E}[X \,|\, \mathcal{G}]\big)^2 = \min_{Y \in \mathcal{G}} \mathsf{E}(X - Y)^2. \tag{8.7}$$

Proof: The space

$$H = \{Z : \mathsf{E}Z^2 < \infty\}$$

is a Hilbert space, with the inner product $\mathsf{E}X \cdot Y$. The space $H_1 = \{Z \in H : Z \in \mathcal{G}\}$ is a linear subspace of H. By standard Hilbert space theory, the orthogonal projection $Y = \pi_{H_1}(X)$ exists and achieves the minimum in (8.7). Moreover, the projection is characterized by

$$\mathsf{E}(X - Y)Z = 0, \quad \forall Z \in H_1.$$

In particular, taking $Z = \mathbb{1}_A$, $A \in \mathcal{G}$, we see that $Y = \mathsf{E}[X \,|\, \mathcal{G}]$. □

8.1.4 Regular conditional expectation

Let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ be a measurable map. One can take $(S, \mathcal{S}) = \big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$, but we allow more generality here. Let $\mathcal{G} \subset \mathcal{F}$ be a sub–σ-algebra.

We note that for every set $A \in \mathcal{F}$, the conditional probability $\mathsf{P}[X \in A \,|\, \mathcal{G}]$ exists a.s. We also know the for any disjoint A_n,

$$\mathsf{P}\Big[X \in \bigcup_{n=1}^{\infty} A_n \,\Big|\, \mathcal{G}\Big] = \sum_{n=1}^{\infty} \mathsf{P}[X \in A_n \,|\, \mathcal{G}], \quad \text{a.s.}, \tag{8.8}$$

where the zero measure exceptional set in (8.8) will depend on $A_1, A_2, \ldots$.

It is tempting to say that $\mathsf{P}[X \in \cdot \,|\, \mathcal{G}](\omega)$ defines a probability measure on $\mathcal{F}$. However, the σ-additivity may not hold, since there are uncountably many ways of choosing $A_1, A_2, \ldots$, so the union of all exceptional sets in (8.8), may not be negligible. If one succeeds in finding a common negligible set, it is called the *regular conditional probability*.

Definition 8.1 (regular conditional probability) *A map* $\mu : \Omega \times \mathcal{S} \to [0, 1]$ *is called a regular conditional probability of* $\mathsf{P}(X \in \cdot)$ *with respect to* $\mathcal{G}$, *if*

1. *for a.e.* ω, $\mu(\omega, \cdot)$ *is a probability measure on* $(S, \mathcal{S})$;
2. *for every* $A \in \mathcal{S}$, $\mu(\omega, A)$ *is a version of* $\mathsf{P}[X \in A \,|\, \mathcal{G}]$.

Example 8.1 Let (X, Y) have density $f(x, y)$ and $\mathcal{G} = \sigma(Y)$. Then

$$\mu(\omega, A) = \begin{cases} \frac{\int_A f\big(x, Y(\omega)\big)\, dx}{\int_{\mathbb{R}} f\big(x, Y(\omega)\big)\, dx}, & \int_{\mathbb{R}} f\big(x, Y(\omega)\big)\, dx \neq 0, \\ 0, & \text{else}, \end{cases}$$

is a regular contional probabilty of μ_X w.r.t. $\sigma(Y)$.

For existence of the regular conditional probability, the key is to find a common negligible set. This is possible when, say, a measure can be determined by its value on countably many sets A, by more generally, by countable many test functions. For example, a measure μ on $\mathbb{R}$ is uniqueness determined by $\mu(-\infty, q]$ where $q \in \mathbb{Q}$; A measure μ on a complete separable metric space M is uniquely determined by $\int f\, d\mu$ where $f \in \mathcal{C}_c(M)$, the space of continuous functions on M which have compact support. Note that $\mathcal{C}_c(M)$ is also separable, so this imposes countably many conditions.

The technical requirement for existence of regular conditional probability is for the space $(S, \mathcal{S})$ to be *Borel*, namely, there exists a map $\varphi : (S, \mathcal{S}) \to \big([0, 1], \mathcal{B}[0, 1]\big)$ so that φ is 1-1 and both φ, φ^{-1} are measurable. Complete, separable metric spaces, like $\mathbb{R}$ and $\mathcal{C}[a, b]$, are Borel spaces.

8.2 Basic martingale theory

A *filtration* $(\mathcal{F}_n)_{n\geq 0}$ on $(\Omega, \mathcal{F}, \mathsf{P})$ is an increasing sub–σ-algebra of $\mathcal{F}$.

Definition 8.2 *A family of random variables $(M_n)_{n\geq 1}$ is a $(\mathcal{F}_n)$-martingale if $M_n \in \mathcal{F}_n$ and $\mathsf{E}|M_n| < \infty$ for all n, and*

$$\mathsf{E}\big(M_{n+1} \mid \mathcal{F}_n\big) = M_n, \quad n \geq 1. \tag{8.9}$$

If "$=$" in (8.9) is replaced by "$\leq$" or "$\geq$", then (M_n) is called a super-martingale or a sub-martingale.

If the filtration is not specified, we take the *natural filtration* $\mathcal{F}_n = \mathcal{F}_n^M := \sigma(M_1, \ldots, M_n)$. If $X_n \in \mathcal{F}_n$ for all $n \geq 1$, we say that (X_n) is *adapted to* the filtration $(\mathcal{F}_n)$. We also note that (8.9) can be replaced by

$$\mathsf{E}[M_{n+m} \,|\, \mathcal{F}_n] = M_n, \quad n \geq 1,\ m \geq 1,$$

since by Proposition 8.9 and the increasing property of $(\mathcal{F}_n)$,

$$\mathsf{E}[M_{n+n} \,|\, \mathcal{F}_n] = \mathsf{E}\big[\mathsf{E}[M_{n+m} \,|\, \mathcal{F}_{n+m-1}] \,\big|\, \mathcal{F}_n\big] = \mathsf{E}[M_{n+m-1} \,|\, \mathcal{F}_n] = \cdots = \mathsf{E}[M_{n+1} \,|\, \mathcal{F}_n].$$

Example 8.2 Let X_i be independent random variables with $\mathsf{E}X_i = 0$. Then the partial sum $S_n = X_1 + \cdots + X_n$ forms a martingale with respect to $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, since by independence,

$$\mathsf{E}[S_{n+m} \,|\, \mathcal{F}_n] = X_1 + \cdots + X_n + \mathsf{E}(X_{n+1} + \cdots + X_m) = S_n.$$

The process S_n is also said to have "mean zero independent increment".

Proposition 8.11 *Let $(X_n)_{n\geq 0}$ has mean zero independent increments. Then*

1. *$(X_n)_{n\geq 0}$ is a martingale.*
2. *If $X_n \in L^2$ for all $n \geq 0$, then $(X_n^2 - \mathsf{E}X_n^2)_{n\geq 0}$ is a martingale.*
3. *If for some $\lambda \in \mathbb{R}$, $\mathsf{E}e^{\lambda X_n} < \infty$ for all $n \geq 0$, then $\left(\frac{e^{\lambda X_n}}{\mathsf{E}e^{\lambda X - t}}\right)_{n\geq 0}$ is a martingale.*

Proof:

1. This is obvious.
2. We have for all $n \geq 1$,

$$\begin{aligned}
\mathsf{E}[X_{n+1}^2 - X_n^2 \,|\, \mathcal{F}_n] &= \mathsf{E}\big[(X_{n+1} - X_n + X_n)^2 - X_n^2 \,\big|\, \mathcal{F}_n\big] \\
&= \mathsf{E}\big[(X_{n+1} - X_n)^2 \,\big|\, \mathcal{F}_n\big] + 2X_n \mathsf{E}[X_{n+1} - X_n \,|\, \mathcal{F}_n] \\
&= \mathsf{E}(X_{n+1} - X_n)^2 = \mathsf{E}(X_{n+1} - X_n)(X_{n+1} + X_n) - 2\mathsf{E}X_n(X_{n+1} - X_n) \\
&= \mathsf{E}X_{n+1}^2 - \mathsf{E}X_n^2.
\end{aligned}$$

3. We have for $n \geq 1$,

$$\mathsf{E}[e^{\lambda X_{n+1}} \,|\, \mathcal{F}_n] = e^{\lambda X_n} \mathsf{E}[e^{\lambda(X_{n+1} - X_n)} \,|\, \mathcal{F}_n] = e^{\lambda X_n} \mathsf{E}e^{\lambda(X_{n+1} - X_n)} = e^{\lambda X_n} \frac{\mathsf{E}e^{\lambda X_{n+1}}}{\mathsf{E}e^{\lambda X_n}}.$$

□

We can use convex/concave function to generate new super- or sup-martingales.

Proposition 8.12 *If $(M_n)_{n\geq 1}$ is a martingale, and $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function, then $\big(\varphi(M_n)\big)_{n\geq 1}$ is a sub-martingale.*

Proof: By Proposition 8.4, we have for all $n \geq 1$ and $m \geq 1$,

$$\mathsf{E}\big[\varphi(M_{n+m}) \,\big|\, \mathcal{F}_n\big] \geq \varphi\Big(\mathsf{E}[X_{n+m} \,|\, \mathcal{F}_n]\Big) = \varphi(X_n). \tag{8.10}$$

□

The function $|x|^p$ $(p \geq 1)$ is convex. So if (M_n) is a martingale, then $|M_n|^p$ is a sub-martingale.

Proposition 8.13 *If $(M_n)_{n\geq 1}$ is a sub-martingale and $\varphi : \mathbb{R} \to \mathbb{R}$ is convex and increasing, then $\big(\varphi(M_n)\big)_{n\geq 1}$ is also a sub-martingale.*

Proof: Since φ is increasing and $(M_n)_{n\geq 1}$ is a sub-martingale, the last equality in (8.10) will become

$$\varphi\Big(\mathsf{E}[X_{n+m} \,|\, \mathcal{F}_n]\Big) \geq \varphi(X_n),$$

and this completes the proof. □

The functions $x \vee a$ $(a \in \mathbb{R})$ and in particular $x^+ = x \vee 0$ are convex and increasing. So if (M_n) is a sub-martingale, then $M_n \vee a$ and M_n^+ are also sub-martingales.

Another way to create new smartingales is to use stopping times.

Definition 8.3 (stopping time) *We say that a r.v. $T \in \{0, 1, \dots\} \cup \{\infty\}$ is a stopping time w.r.t. $(\mathcal{F}_n)$ if $\{T \leq n\} \in \mathcal{F}_n$ for every $n \geq 1$.*

Proposition 8.14 *If $(M_n)_{n\geq 1}$ is a sub-martingale and T is a stopping time which is almost surely finite, then $(M_{n\wedge T})_{n\geq 1}$ is a also sub-martingale.*

In particular, if $(M_n)_{n\geq 1}$ is a martingale, then $(M_{n\wedge T})_{n\geq 1}$ is also a martingale.

Proof: By definition we have

$$M_{n\wedge T} = \sum_{k=0}^{n} \mathbb{1}_{\{T=k\}} M_k + \mathbb{1}_{\{T\geq n+1\}} M_n.$$

Therefore,

$$\begin{aligned} \mathsf{E}[M_{(n+1)\wedge T} - M_{n\wedge T} \,|\, \mathcal{F}_n] &= \mathsf{E}[\mathbb{1}_{\{T\geq n+2\}} M_{n+1} + \mathbb{1}_{\{T=n+1\}} M_{n+1} - \mathbb{1}_{\{T\geq n+1\}} M_n \,|\, \mathcal{F}_n] \\ &= \mathsf{E}[\mathbb{1}_{\{T\geq n+1\}} M_{n+1} - \mathbb{1}_{\{T\geq n+1\}} M_n \,|\, \mathcal{F}_n]. \end{aligned} \tag{8.11}$$

But $\{T \geq n+1\} = \{T \leq n\}^c \in \mathcal{F}_n$, so the last line of (8.11) is

$$\mathbb{1}_{\{T\geq n+1\}} \mathsf{E}[M_{n+1} - M_n \,|\, \mathcal{F}_n] \geq \mathbb{1}_{\{T\geq n+1\}} \cdot 0 = 0.$$

For the last statement, note that if (M_n) if and only if (M_n) and $(-M_n)$ are both sub-martingales. □

A more general way to generate new sub-martingales is to use a (discrete) martingale integration. We say that a process (H_n) is *predictable* if $H_n \in \mathcal{F}_{n-1}$ for all n. We define

$$(H \cdot X)_n = \sum_{m=1}^{n} H_m (X_m - X_{m-1}), \tag{8.12}$$

which can be thought of as the discrete Riemann–Steiljes integration $\int_0^t H_s \, dX_s$. Clearly, $(H \cdot X)_n \in \mathcal{F}_n$ for all n.

Proposition 8.15 *If (H_n) is predictable and non-negative, and (X_n) is a sub-martingale, then $(H\cdot X)_n$ is a sub-martingale.*

If (H_n) is predictable and (X_n) is a martingale, then $(H\cdot X)_n$ is a martingale.

Proof: By (8.12), we have

$$\mathsf{E}\big[(H\cdot X)_{n+1}-(H\cdot X)_n\,\big|\,\mathcal{F}_n\big]=\mathsf{E}\big[H_{n+1}(X_{n+1}-X_n)\,\big|\,\mathcal{F}_n\big]=H_n\mathsf{E}[X_{n+1}-X_n\,|\,\mathcal{F}_n]\geq 0, \tag{8.13}$$

if $H_n\geq 0$ and (X_n) is a sub-martingale. If instead (X_n) is martingale, then the RHS of (8.13) is 0 regardless of the sign of H_n, so $(H\cdot X)_n$ is a marginale. □

Example 8.3 If T is a stopping time, then Let $H_n=\mathbb{1}_{\{T\geq n\}}$ where T is a stopping time. Since $\{T\geq n\}=\{T\leq n-1\}^c\in\mathcal{F}_{n-1}$, the process (H_n) is predictable. Now

$$(H\cdot X)_n=\sum_{m=1}^{n}\mathbb{1}_{\{T\geq m\}}(X_m-X_{m-1})=X_{n\wedge T}-X_0,$$

so by Proposition 8.15, the process $(H\cdot X_n)$ is a sub-martingale if (X_n) is a sub-martingale, and we recover Proposition 8.14.

8.3 Convergence of martingales

In this section we discuss the almost sure and L^1-limits of martingales. The main tools are *Doob's Up-crossing Theorem* and uniform integrability.

Let (X_n) be $(\mathcal{F}_n)$-adapted and $a<b$. Consider the following stopping times: $T_b^{(0)}=-\infty$,

$$T_a^{(\ell)}=\inf\{t\geq T_b^{(\ell-1)}:X_n\leq a\},\quad T_b^{(\ell)}=\inf\{t\geq T_a^{(\ell)}:X_n\geq b\},\quad \ell\geq 1. \tag{8.14}$$

In every interval $[T_a^{(\ell)},T_b^{(\ell)}]$, the process (X_n) completes an up-crossing of $[a,b]$. The total number of up-crossing in a given time interval $[0,n]$ is defined by

$$U_{ab}^X[0,n]=\max\{k:T_b^{(k)}\leq n\}.$$

Theorem 8.16 (Doob's up-crossing inequality) *Let $(X_n)_{n\geq 1}$ be a sub-martingale, then*

$$\mathsf{E}U_{ab}^X[0,n]\leq\frac{1}{b-a}\Big(\mathsf{E}(X_n-a)_+-\mathsf{E}(X_0-a)_+\Big).$$

Proof: Let $Y_n=(X_n-a)_+$ If (X_n) is a sub-martingale, then (Y_n) is also a sub-martingale since $x\mapsto(x-a)_+$ is convex and increasing. Moreover, X_n and (Y_n) have the same number of up-crossing, that is

$$U_{ab}^X[0,n]=U_{ab}^Y[0,n],\quad \forall n\geq 1,\ \forall a<b. \tag{8.15}$$

Let us define $T_a^{(\ell)}$ and $T_b^{(\ell)}$ using (Y_n) in (8.14), and estimate $U_{ab}^Y[0,n]$.

Let

$$H_m=\begin{cases}1, & T_a^{(\ell)}<m\leq T_b^{(\ell)},\\ 0, & \text{else.}\end{cases} \tag{8.16}$$

Since all $T_a^{(\ell)}$ and $T_b^{(\ell)}$ are stopping times, we have

$$\{H_m=1\}=\bigcup_{\ell=1}^{\infty}\{T_a^{(\ell)}<m\}\cap\{m\leq T_b^{(\ell)}\}=\bigcup_{\ell=1}^{\infty}\{T_a^{(\ell)}\leq m-1\}\cap\{m-1\geq T_b^{(\ell)}\}^c\in\mathcal{F}_{m-1}.$$

Hence (H_m) is predictable.

Let $K_m = 1 - H_m$. By Proposition 8.15, both $(H \cdot Y)_n$ and $(K \cdot Y)_n$ are sub-martingales, so $\mathsf{E}(K \cdot Y)_n \geq \mathsf{E}(K \cdot Y)_0 = 0$. On the other hand, $H_n + K_n \equiv 1$. Combining these, we have

$$\mathsf{E}(H \cdot Y)_n \leq \mathsf{E}(H \cdot Y)_n + \mathsf{E}(K \cdot Y)_n = \sum_{m=1}^{n} \mathsf{E}(Y_m - Y_{m-1}) = \mathsf{E}(X_n - a)_+ - \mathsf{E}(X_0 - a)_+. \tag{8.17}$$

Note that by (8.16),

$$\mathsf{E}(H \cdot Y)_n \geq U_{ab}^{Y}[0, n] \cdot (b - a). \tag{8.18}$$

The conclusion follows from (8.15), (8.17) and (8.18). □

An important observation is that there must be infinitely many up-crossing for a divergent sequence.

Proposition 8.17 *If (X_n) is a sub-martingale, and $\sup_n \mathsf{E}X_n^+ < \infty$. Then there exists X such that $X_n \to X$ a.s.*

Proof: The up-crossing number is increasing in n, and hence by assumption and Theorem 8.16,

$$\mathsf{E}U_{ab}^{X}[0, \infty) = \lim_{n\to\infty} \mathsf{E}U_{ab}^{X}[0, n] \leq \frac{\sup_n \mathsf{E}X_n^+ + |a|}{b - a} < \infty.$$

This implies that $U_{ab}^{X}[0, \infty)$ is a.s. finite r.v., with probability one, any interval $[a, b]$ is being up-crossed by at most finitely many times. As a consequence, for any fixed a and b, there exists $N_{a,b}$ with $\mathsf{P}(N_{a,b}) = 0$ such that

$$\liminf_{n\to\infty} X_n(\omega) < a < b < \limsup_{n\to\infty} X_n(\omega), \tag{8.19}$$

cannot happen on $N_{a,b}^c$.

Then, on N^c where $N = \bigcup_{a,b\in\mathbb{Q}} N_{a,b}$, (8.19) does not happen for all $a, b \in \mathbb{Q}$, and hence

$$\limsup_{n\to\infty} X_n(\omega) = \liminf_{n\to\infty} X_n(\omega), \quad \forall \omega \in N^c.$$

Note that $\mathsf{P}(N) \leq \sum_{a,b\in\mathbb{Q}} \mathsf{P}(N_{a,b}) = 0$. This shows that $\lim_{n\to\infty} X_n$ exists a.s. □

Example 8.4 If a martingale $(X_n)_{n\geq 0}$ is non-negative, then $\mathsf{E}X_n^+ = \mathsf{E}X_n = \mathsf{E}X_0$, and hence $\lim_{n\to\infty} X_n$ exists by Proposition 8.17.

Next we will discuss the L^1-convergence of smartingales. Recall the definition of uniform integrability Definition 2.1, which gives a sufficient and necessary condition of L^1-convergence for a.s. convergence r.v.s (Theorem 2.11).

Proposition 8.18 *Let $Z \in L^1(\Omega, \mathcal{F}, \mathsf{P})$. Then the collection of r.v.s*

$$\mathsf{E}[Z \,|\, \mathcal{G}], \quad \mathcal{G} \text{ is a sub-}\sigma\text{-field of } \mathcal{F},$$

is uniformly integrable.

Proof: Since $Z \in L^1(\Omega, \mathcal{F}, \mathsf{P})$, for every $\varepsilon > 0$, there is $\delta > 0$ such that whenever $\mathsf{P}(A) < \delta$, $\mathsf{E}|Z|\mathbb{1}_A < \varepsilon$.

By Jensen inequality, for $A = \{|\mathsf{E}[Z \,|\, \mathcal{G}]| \geq M\} \in \mathcal{G}$, we have

$$\mathsf{E}\Big(\mathbb{1}_A \cdot |\mathsf{E}[Z \,|\, \mathcal{G}]|\Big) \leq \mathsf{E}\big(\mathbb{1}_A \cdot \mathsf{E}[|Z| \,|\, \mathcal{G}]\big) = \mathsf{E}|Z|\mathbb{1}_A.$$

When $A = \Omega$, the above inequality gives $\big|\mathsf{E}[Z\,|\,\mathcal{G}]\big| \le \mathsf{E}|Z|$ (or this is Proposition 8.6). Then by Chebyshev's inequality,

$$\mathsf{P}(A) \le \frac{\mathsf{E}|Z|}{M},$$

uniformly for all sub–σ-field $\mathcal{G}$. Combining all these together we prove the statement. □

Proposition 8.19 *A martingale (X_n) is uniformly integrable, if and only if there exists $X_\infty \in L^1$ such that $X_n = \mathsf{E}[X_\infty\,|\,\mathcal{F}_n]$.*

Proof: The "⇒" direction. Uniform integrability implies that $\sup_n \mathsf{E}|X_n| < \infty$, hence Proposition 8.17 implies that there exists X_∞ such that $X_n \to X_\infty$ a.s. But (X_n) is also uniformly integrable, so the limit is also in L^1 by Theorem 2.11.

For any $A \in \mathcal{F}_n$, since $\mathsf{E}[X_\infty\,|\,\mathcal{F}_n] \in \mathcal{F}_n$ and $X_{n+m}\mathbb{1}_A \to X_\infty \mathbb{1}_A$ in L^1, we have

$$\mathsf{E}\big(\mathsf{E}[X_\infty\,|\,\mathcal{F}_n]\mathbb{1}_A\big) = \mathsf{E}X_\infty\mathbb{1}_A = \lim_{m\to\infty} \mathsf{E}X_{n+m}\mathbb{1}_A = \lim_{m\to\infty} \mathsf{E}\big(\mathsf{E}[X_{n+m}\,|\,\mathcal{F}_n]\mathbb{1}_A\big) = \mathsf{E}X_n\mathbb{1}_A.$$

Since $X_n \in \mathcal{F}_n$, by the definition of the conditional expectation, we have

$$\mathsf{E}[X_\infty\,|\,\mathcal{F}_n] = X_n, \quad \text{a.s.}$$

The "⇐" direction. It follows from Proposition 8.18. □

8.4 Optional Sampling Theorem

In this section, we assume all stopping times are a.s. finite.

Recall from Proposition 8.14 that if (X_n) is a martingale and T is a stopping time, then $(X_{n\wedge T})_{n\ge 1}$ is also a martingale. In particular,

$$\mathsf{E}X_{n\wedge T} = \mathsf{E}X_{0\wedge T} = \mathsf{E}X_0. \tag{8.20}$$

Assume that X_n is bounded, then as $n \to \infty$, by BCT the LHS of (8.20) converges to $\mathsf{E}X_T$. Thus we obtain the simplest form of the *optional sampling theorem*

Theorem 8.20 (optional sampling theorem) *Let X_n be a $(\mathcal{F}_n)$-martingale and T an $(\mathcal{F}_n)$-stopping time. Assume that X_n is bounded,*

$$\mathsf{E}X_T = \mathsf{E}X_0.$$

The optional sampling theorem says that no strategy can safely gain you any profit in a fair game.

To prove a more general form of optional sampling theorem, let us introduce the *stopping σ-algebra*.

Definition 8.4 *Let T be a stopping time. The stopping σ-algebra is*

$$\mathcal{F}_T = \{A \in \mathcal{F} : A \cap \{T \le n\} \in \mathcal{F}_n,\ \forall n\}.$$

Intuitively, $\mathcal{F}_T$ contains the information before a stopping time T.

Example 8.5 Let $m \ge 0$ and consider $T = m$ (a constant time). Then T is a stopping time and $\mathcal{F}_T = \mathcal{F}_m$.

We can compare the stopping σ-algebras for different stopping time, or extract information from the stopping σ-algebra.

Proposition 8.21 *If $S \le T$ are two stopping times, then $\mathcal{F}_S \subset \mathcal{F}_T$.*

Remark 8.6 Since $S \le T$, "information before S" is less than "information before T".

Proof: If $A \subset \mathcal{F}_S$, then for every n,

$$A \cap \{T \le n\} = \Big(A \cap \{S \le n\}\Big) \cap \{T \le n\} \in \mathcal{F}_n.$$

So $A \subset \mathcal{F}_T$. This completes the proof. □

Proposition 8.22 *If T is a stopping time and $S \ge T$ is random time such that S is $\mathcal{F}_T$-measurable, then S is also a stopping time.*

Proof: For each $n \ge 0$, since $\{S \le n\} \in \mathcal{F}_T$,

$$\{S \le n\} = \{S \le n\} \cap \{T \le n\} \in \mathcal{F}_n.$$

This completes the proof. □

Remark 8.7 The stopping time S will take the form $S = f(T)$ for some function $f : \mathbb{N} \to \mathbb{N}$ with $f(m) \ge m$.

Theorem 8.23 *Let $(X_n)_{n\ge 0}$ be a martingale, and $S \le T$ be two stopping times. Suppose that either*

1. *S, T are bounded, i.e., there is a constant $N > 0$ such that $S, T \le N$, or*
2. *$(X_n)_{n\ge 1}$ is uniformly integrable.*

Then

$$X_S = \mathsf{E}[X_T \,|\, \mathcal{F}_S].$$

In particular, $\mathsf{E}X_S = \mathsf{E}X_T = \mathsf{E}X_0$.

Remark 8.8 The first condition implies that $X_n = \mathsf{E}[X_N \,|\, \mathcal{F}_n]$, and the second condition by Proposition 8.19 implies that

$$X_n = \begin{cases} \mathsf{E}[X_\infty \,|\, \mathcal{F}_n], & n < \infty \\ X_\infty, & n = \infty. \end{cases} \tag{8.21}$$

So both conditions implies that there is a r.v. $Z \in L^1$ such that $X_n = \mathsf{E}[Z \,|\, \mathcal{F}_n]$ for all n that we care about.

Proof: Let $Z = X_N$ if the first condition holds and $Z = X_\infty$ if the second condition holds. Then (8.21) holds with $X_\infty = Z$. It suffices to show

$$X_T = \mathsf{E}[Z \,|\, \mathcal{F}_T]. \tag{8.22}$$

Indeed, if (8.22) holds, since $\mathcal{F}_S \subset \mathcal{F}_T$, we have

$$\mathsf{E}[X_T \,|\, \mathcal{F}_S] = \mathsf{E}\big[\mathsf{E}[Z \,|\, \mathcal{F}_T] \,\big|\, \mathcal{F}_S\big] = \mathsf{E}[Z \,|\, \mathcal{F}_S] = X_S.$$

Now let us prove (8.22). For all $A \in \mathcal{F}_S$, we have

$$\begin{aligned} \mathsf{E}\big(\mathsf{E}[Z \,|\, \mathcal{F}_S]\mathbb{1}_A\big) = \mathsf{E}Z\mathbb{1}_A &= \sum_{n=1}^{\infty} \mathsf{E}Z\mathbb{1}_{A\cap\{T=n\}} + \mathsf{E}Z\mathbb{1}_{A\cap\{T=\infty\}} \\ &= \sum_{n=1}^{\infty} \mathsf{E}\big(\mathbb{1}_{A\cap\{T=n\}} \cdot \mathsf{E}[Z \,|\, \mathcal{F}_n]\big) + \mathsf{E}Z\mathbb{1}_{A\cap\{T=\infty\}} \\ &= \sum_{n=1}^{\infty} \mathsf{E}\mathbb{1}_{A\cap\{T=n\}} X_n + \mathsf{E}Z\mathbb{1}_{A\cap\{T=\infty\}} \\ &= \mathsf{E}X_T\mathbb{1}_A, \end{aligned}$$

where in the second line we use that $A \cap \{T = n\} \in \mathcal{F}_n$ since T is a stopping time. □

Example 8.9 We can recover Proposition 8.14. If T is a stopping time, $(M_n)_{n\geq 1}$ is a martingale, then $(M_{n\wedge T})_{n\geq 1}$ is also a martingale, since

$$\mathsf{E}[M_{n\wedge T}|\mathcal{F}_{m\wedge T}] = M_{m\wedge T}, \quad \forall n > m,$$

by Theorem 8.23 and the boundedness of the stopping time $n \wedge T$, $m \wedge T$.

Note: convert this example to simple random walk.

Example 8.10 Let $(B_n)_{n\geq 1}$ be Brownian motion, and T_a, T_b be the first hitting time of $a > 0 > b$. Applying Theorem 8.23 to the bounded stopping time $T_a \wedge T_b \wedge n$ gives

$$\mathsf{E}B_{T_a\wedge T_b\wedge n} = \mathsf{E}B_0 = 0. \tag{8.23}$$

Since $|B_{T_a\wedge T_b\wedge n}| \leq |a| \vee |b|$ and $\mathsf{P}(T_a \wedge T_b < \infty) = 1$ (one can easily show for some $\rho < 1$, $\mathsf{P}(T_a \wedge T_b \geq k) \leq \rho^k$), we can take $n \to \infty$ in (8.23) and get

$$0 = \mathsf{E}B_{T_a\wedge T_b} = a\mathsf{P}(T_a < T_b) + b\mathsf{P}(T_a > T_b).$$

Also $\mathsf{P}(T_a < T_b) + \mathsf{P}(T_a > T_b) = 1$. Hence, we have

$$\mathsf{P}(T_a < T_b) = \frac{-b}{a-b}, \quad \mathsf{P}(T_a > T_b) = \frac{a}{a-b}. \tag{8.24}$$

In particular, letting $b \downarrow -\infty$ and $T_b \uparrow \infty$, we obtain $\mathsf{P}(T_a < \infty) = 1$.

Example 8.11 Apply Theorem 8.23 to the martingale $(B_n - t^2)_{n\geq 1}$ and the stopping time $T_a \wedge T_b \wedge n$, we have

$$\mathsf{E}B^2_{T_a\wedge T_b\wedge n} - (T_a \wedge T_b \wedge n) = 0.$$

In the limit $n \to \infty$, the first term is bounded by $|a|^2 \vee |b|^2$, the second term is increasing in n, so by Bounded Convergence Theorem and Monotone Convergence Theorem, we have

$$\mathsf{E}B_{T_a\wedge T_b} - (T_a \wedge T_b).$$

Combining with (8.24) we have $\mathsf{E}T_a \wedge T_b = |ab|$. Letting $b \downarrow -\infty$ and obtain $\mathsf{E}T_a = \infty$.

We will also mention the Optional Sampling Theorem for sub-/super-martingales.

Definition 8.5 *A smartingale $(X_n)_{n\geq 1}$ has a last element/is closed by X_∞, if $\exists X_\infty \in L^1$ such that $(X_n)_{0\leq t\leq\infty}$ forms a smartingale.*

Example 8.12 If $(M_n)_{n\geq 1}$ is a martingale, then by Proposition 8.19, it has a last element if and only if it is uniformly integrable. Moreover, M_∞ is the a.s. and L^1 limit of M_n.

Example 8.13 If $(X_n)_{n\geq 1}$ is a non-negative super-martingale, then it always has a last element $X_\infty = 0$, since it is trivially true that

$$X_n \geq 0 = \mathsf{E}[X_\infty|\mathcal{F}_n], \quad \forall n \geq 1.$$

But having a last element is weaker than uniform integrability. Consider $X_n = 1 + S_{n\wedge T_{-1}}$ which is a martingale and hence super-martingale. It is non-negative. It is easy to see that

$$X_\infty = \lim_{t\to\infty} X_n = 1 + S_{T_{-1}} = 0,$$

but $1 = \lim_{t\to\infty} \mathsf{E}X_n \neq \mathsf{E}X_\infty = 0$, so it cannot be uniformly integrable.

Theorem 8.24 *Let $(X_n)_{n\geq 1}$ is a sub-martingale and $S \leq T$ be two stopping times. If either*

1. *S, T are bounded, or*
2. *$(X_n)_{n\geq 1}$ has a last element $X_\infty \in L^1$,*

then

$$\mathsf{E}[X_T \,|\, \mathcal{F}_S] \geq X_S. \tag{8.25}$$

A similar statement also holds for super-martingale.

Remark 8.14 The argument in Theorem 8.23 no longer works since the conclusion of the theorem cannot be derived from $\mathsf{E}[X_\infty \mid \mathcal{F}_T] \geq X_T$.

Proof: Let $A \in \mathcal{F}_S$ and

$$H_n = \mathbb{1}_A \cdot \mathbb{1}_{\{S<n\leq T\}}.$$

Then (H_n) is predictable since

$$\{H_n = 1\} = \big(A \cap \{S \leq n-1\}\big) \cap \{T \geq n-1\}^c \in \mathcal{F}_{n-1}, \quad \forall n \geq 1.$$

By Proposition 8.15,

$$(H \cdot X)_0 = 0, \quad (H \cdot X)_n = \sum_{k=1}^{n} H_k(X_k - X_{k-1}),\ n \geq 1,$$

is a sub-martingale. In particular, for all $n \geq 1$

$$0 = \mathsf{E}(H \cdot X)_0 \leq \mathsf{E}(H \cdot X)_n = \mathsf{E}(X_{T\wedge n} - X_{S\wedge n})\mathbb{1}_{A\cap\{S<T\}}.$$

But $X_{T\wedge n} = X_{S\wedge n}$ on $\{S = T\}$, so we obtain

$$\mathsf{E}(X_{T\wedge n} - X_{S\wedge n})\mathbb{1}_A \geq 0, \quad \forall n \geq 1,\ A \in \mathcal{F}_S. \tag{8.26}$$

If T is bounded, then there exists N such that $T \leq N$ a.s., and taking $n = N+1$ in (8.26) gives

$$\mathsf{E}(X_T - X_S)\mathbb{1}_A \geq 0.$$

and this proves (8.25).

If (X_n) has a last element, without loss of generality we can assume $X_\infty = 0$, otherwise we can consider $X'_n = X_n - \mathsf{E}[X_\infty \,|\, \mathcal{F}_n]$, and (8.25) is equivalent to

$$\mathsf{E}[X'_T \,|\, \mathcal{F}_S] \geq X'_S,$$

by Theorem 8.23, and (X'_n) has a last element 0.

If T is unbounded but $S \leq N$ is bounded, then since $-X_n \geq 0$, by Fatou we have

$$-\mathsf{E}X_S\mathbb{1}_A = -\mathsf{E}X_{S\wedge(N+1)}\mathbb{1}_A \geq \liminf_{n\to\infty} \mathsf{E}(-X_{T\wedge n})\mathbb{1}_A \geq \mathsf{E}(-X_T)\mathbb{1}_{A\cap\{T<\infty\}}.$$

By adding $0 = \mathsf{E} - X_\infty\mathbb{1}_{A\cap\{T=\infty\}}$ to both sides, we obtain

$$-\mathsf{E}X_S\mathbb{1}_A \geq -\mathsf{E}X_T\mathbb{1}_A. \tag{8.27}$$

This proves (8.25).

Now we need to treat the case where S is unbounded. For every $m \geq 1$, $S \wedge m$ is a bounded stopping time, and moreover, $\{S = m\} \cap A \in \mathcal{F}_{S\wedge m}$ since

$$\big(\{S = m\} \cap A\big) \cap \{S \wedge m \leq n\} \begin{cases} = \{S = m\} \cap A \in \mathcal{F}_m \subset \mathcal{F}_n, & m \leq n, \\ = \varnothing \in \mathcal{F}_n, & m \geq n+1. \end{cases}$$

From what we have proven, we have

$$-\mathsf{E}X_S\mathbb{1}_{\{S=m\}\cap A} = -\mathsf{E}X_{S\wedge m}\mathbb{1}_{\{S=m\}\cap A} \geq -\mathsf{E}X_T\mathbb{1}_{\{S=m\}\cap A}. \tag{8.28}$$

Summing (8.28) over $m \in \{0, 1, \dots\}$, we have

$$-\mathsf{E}X_S\mathbb{1}_{\{S<\infty\}\cap A} \geq -\mathsf{E}X_T\mathbb{1}_{\{S<\infty\}\cap A}.$$

Noting that $X_\infty = 0$, and $S = \infty$ implies $T = \infty$, we can remove $\{S < \infty\}$ on both sides to obtain (8.27). This completes the proof.

□

8.5 Doob's Maximal inequality

We will state the maximal inequality for sub-martingales. Similar statements also hold for super-martingales.

Theorem 8.25 *Let $(X_n)_{n\geq 1}$ be a sub-martingale and $\lambda > 0$. Then*

$$\lambda\mathsf{P}\Big(\max_{0\leq m\leq n} X_m > \lambda\Big) \leq \mathsf{E}X_n^+, \tag{8.29}$$

$$\lambda\mathsf{P}\Big(\inf_{0\leq m\leq n} X_m < -\lambda\Big) \leq \mathsf{E}X_n^+ - \mathsf{E}X_0. \tag{8.30}$$

Proof: Denote the event in (8.29) as A and let $T = \min\{m : X_m \geq \lambda\}$. Then $A = \{T \leq n\}$. Since X is a sub-martingale, X^+ is also a sub-martingale. By Theorem 8.24 we have

$$\mathsf{E}X_n^+ \geq \mathsf{E}X_{n\wedge T}^+ \geq \mathsf{E}X_{n\wedge T}^+\mathbb{1}_{\{T\leq n\}} = \lambda\mathsf{P}(A).$$

This proves (8.29).

Denote the event in (8.30) by B and let $S = \min\{m : X_m \leq \lambda\}$. Then $B = \{S \leq n\}$. Again by Theorem 8.24, we have

$$\begin{aligned}\mathsf{E}X_0 &\leq \mathsf{E}X_{n\wedge S} = \mathsf{E}X_n\mathbb{1}_{\{T>n\}} + \mathsf{E}X_T\mathbb{1}_{\{T\leq n\}} \\ &\leq \mathsf{E}X_n\mathbb{1}_{\{T>n\}} - \lambda\mathsf{P}(B) \leq \mathsf{E}X_n^+ - \lambda\mathsf{P}(B),\end{aligned}$$

and (8.30) follows. □

Proposition 8.26 *Let $(M_n)_{n\geq 1}$ be a continuous martingale. Then for every $\lambda > 0$,*

$$\lambda\mathsf{P}\Big(\max_{0\leq m\leq n} |M_m| \geq \lambda\Big) \leq \mathsf{E}|X_n|.$$

Proof: We apply (8.29) in Theorem 8.25 to the sub-martingale $(|M_n|)_{n\geq 1}$. □

For martingales, we also have the control on the maximum of L^p norm.

Theorem 8.27 *Let $(M_n)_{n\geq 1}$ be a martingale. Then for every $p > 1$,*

$$\mathsf{E}\max_{0\leq m\leq n} |M_m|^p \leq \Big(\frac{p}{p-1}\Big)^p \mathsf{E}|X_n|^p.$$

Proof: Let $Y = \max_{0\le m\le n}|M_m|$. Since $(|M_n|)_{n\ge 1}$ is a sub-martingale, we have

$$\lambda \mathsf{P}(Y \ge \lambda) + \mathsf{E}|M_n|\mathbb{1}_{\{Y<\lambda\}} \le \mathsf{E}|M_n|,$$

and hence

$$\mathsf{P}(Y \ge \lambda) \le \frac{1}{\lambda}\mathsf{E}|M_n|\mathbb{1}_{\{Y\ge\lambda\}}.$$

Now

$$\begin{aligned}
\mathsf{E}Y^p &= p\int_0^\infty \lambda^{p-1}\mathsf{P}(Y\ge\lambda)\,d\lambda \\
&\le p\int_0^\infty \lambda^{p-2}\mathsf{E}\big(|M_n|\mathbb{1}_{\{Y\ge\lambda\}}\big) \\
&= \mathsf{E}\Big(|M_n|\int_0^Y p\lambda^{p-2}\,d\lambda\Big) \\
&= \frac{p}{p-1}\cdot\mathsf{E}\big(|M_n|\cdot Y^{p-1}\big) \\
&\le \frac{p}{p-1}\big(\mathsf{E}|M_n|^p\big)^{1/p}\big(\mathsf{E}Y^p\big)^{p/(p-1)}.
\end{aligned}$$

The last inequality is just Hölder's inequality. Hence, if $\mathsf{E}Y^p < \infty$, then we can divide both sides by $\big(\mathsf{E}Y^p\big)^{p/(p-1)}$ and then take the p-th power to get $\mathsf{E}Y^p \le \Big(\frac{p}{p-1}\Big)^p E|M_n|^p$. To treat the general case where $\mathsf{E}Y^p < \infty$ is not known, we use truncation, that is, we first get the estimate

$$\mathsf{E}(Y\wedge m)^p \le \big(\frac{p}{p-1}\big)^p\mathsf{E}|M_n|^p$$

for the bounded r.v $(Y\wedge m)$ with any $m>0$. Then we let $m\to\infty$ and get the desired conclusion. □

9 Some applications of martingales

9.1 Radon–Nikodym derivatives

Proposition 9.1 *Let $T\in(0,\infty]$*

1. *Let P and $\tilde{\mathsf{P}}$ be two probability measures on a filtered probability space $(\Omega,\mathcal{F}_T,(\mathcal{F}_t)_{0\le t\le T})$.Let P_t and $\tilde{\mathsf{P}}_t$ be the restriction of P and $\tilde{\mathsf{P}}$ on the smaller σ-field $\mathcal{F}_t\subset\mathcal{F}_T$. Suppose that $\tilde{\mathsf{P}}\ll\mathsf{P}$. Then $\tilde{\mathsf{P}}_t\ll\mathsf{P}_t$ for all t, and*

$$Z_t = \frac{d\tilde{\mathsf{P}}_t}{d\mathsf{P}_t} = \mathsf{E}\Big[\frac{d\tilde{\mathsf{P}}}{d\mathsf{P}}\mid\mathcal{F}_t\Big],\quad 0\le t\le T \tag{9.1}$$

   is a martingale.

2. *Let $(Z_t)_{0\le t<T}$ be a P-martingale. Then*

$$\tilde{\mathsf{P}}(A) = \mathsf{E}\mathbb{1}_A(\omega)Z_t(\omega),\quad \forall A\in\mathcal{F}_t,\ 0\le t<T, \tag{9.2}$$

   defines a probability measure $\tilde{\mathsf{P}}$.

   Moreover, if $(Z_t)_{0\le t<T}$ is u.i., that is, $Z_T=\lim_{t\to T}Z_t$ exists in L^1 and a.s., then $\tilde{\mathsf{P}}\ll\mathsf{P}$ and $Z_T=\frac{d\tilde{\mathsf{P}}}{d\mathsf{P}}$.

Proof:

1. Let $A \in \mathcal{F}_t$. Then

$$\mathsf{P}_t(A) = 0 \quad \Rightarrow \quad \mathsf{P}(A) = 0 \quad \Rightarrow \quad \tilde{\mathsf{P}}(A) = 0 \quad \Rightarrow \quad \tilde{\mathsf{P}}_t(A) = 0.$$

   Hence, $\tilde{\mathsf{P}} \ll \mathsf{P}$ implies that $\tilde{\mathsf{P}}_t \ll \mathsf{P}_t$.

   To show that Z_t is a martingale, it suffices to show the second equality in (9.1). Let $A \in \mathcal{F}_t$. Then by the definition of Radon–Nikodym derivatives,

$$\tilde{\mathsf{P}}_t(A) = \mathsf{E}\mathbb{1}_A \frac{d\tilde{\mathsf{P}}_t}{d\mathsf{P}_t}, \quad \tilde{\mathsf{P}}(A) = \mathsf{E}\mathbb{1}_A \frac{d\tilde{\mathsf{P}}}{d\mathsf{P}}.$$

   Therefore,

$$\mathsf{E}\mathbb{1}_A \frac{d\tilde{\mathsf{P}}_t}{d\mathsf{P}_t} = \mathsf{E}\mathbb{1}_A \frac{d\tilde{\mathsf{P}}}{d\mathsf{P}}$$

   for all $A \in \mathcal{F}_t$. Hence the second equality in (9.1) holds by the definition of conditional expectation.

2. We need to check that (9.2) gives a consistent definition of a probability measure, since if $A \in \mathcal{F}_s \subset \mathcal{F}_t$, there are two definitions for $\tilde{\mathsf{P}}(A)$:

$$\tilde{\mathsf{P}}(A) = \mathsf{E}\mathbb{1}_A Z_t, \quad \tilde{\mathsf{P}}(A) = \mathsf{E}\mathbb{1}_A Z_s.$$

   But $\mathsf{E}\mathbb{1}_A Z_t = \mathsf{E}\mathbb{1}_A Z_s$ just follows from Z_t being martingale.

   Suppose now that Z_T exists. For any $A \in \mathcal{F}_t$, $\{\mathbb{1}_A Z_r,\ r \geq t\}$ is u.i. since Z_r are u.i. Then,

$$\tilde{\mathsf{P}}(A) = \lim_{r \to T} \mathsf{E}\mathbb{1}_A Z_r = \mathsf{E}\mathbb{1}_A Z_T.$$

   Since $\tilde{\mathsf{P}}(A) = \mathsf{E}\mathbb{1}_A Z_T$ holds for any $A \in \mathcal{F}_t$, $t \geq 0$, it holds for any $A \in \mathcal{F}_T$. Therefore, $\tilde{\mathsf{P}} \ll \mathsf{P}$ and Z_T is the Radon–Nikodym derivative.

□

An analog in the case of product measures is the *Kakutani's dichotomy* (see also [Dur19, Example 4.3.7, Theorem 4.3.8]). Let $(\Omega, \mathcal{F}) = (\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$, and consider two product measures

$$\mathsf{P} = G_1 \otimes G_2 \otimes G_3 \otimes \cdots, \quad \tilde{\mathsf{P}} = F_1 \otimes F_2 \otimes F_3 \otimes \cdots$$

Assume that $P_n \ll G_n$, and $q_n = \frac{dF_n}{dG_n} > 0$, G_n-a.s. Then, $X_n = \frac{d\mathsf{P}_n}{d\tilde{\mathsf{P}}_n}$ is a $\mathcal{F}_n$-martingale. Note that by the nature of the product measure, X_n are independent random variables. Since

$$\{\lim_{n\to\infty} X_n = 0\} = \{\sum_n \log q_n > -\infty\}$$

belongs to the tail σ-algebra, the zero-one law guarantees that $X_n \to X$ P-a.s. for some X. We have either $\mathsf{P} \perp \tilde{\mathsf{P}}$ if $X = 0$, or $\tilde{\mathsf{P}} \ll \mathsf{P}$ if $X > 0$.

[Or rephrase this for infinite dimension Gaussian] Interestingly, this is not too far from our Brownian motion case. Recall the Gaussian white noise construction of Brownian motion

$$B_t(\omega) = \sum_{n=1}^{\infty} \xi_n(\omega) \langle \mathbb{1}_{[0,t]}, e_n \rangle_{L^2},$$

where $\{e_n\}$ is an ONB and ξ_n are i.i.d. $\mathcal{N}(0,1)$ r.v.'s. So Brownian motion also have some product measure structure.

10 Notation

10.1 Abbreviations

i.i.d.	independent, identically distributed
r.v.	random variable
p.m.	probability measure
c.d.f.	cumulative distribution function
f.d.d.	finite-dimensional distribution
ch.f.	characteristic function
u.i.	uniformly integrable

10.2 Relations

$\Rightarrow_d$ or $\Rightarrow$	convergence in distribution/law
$\overset{\mathrm{d}}{=}$	equal in law

10.3 Functional spaces

$\mathcal{C}[a,b]$	continuous function defined on the interval $[a,b]$
$\mathcal{C}^\alpha[a,b]$	α-Hölder continuous function defined on the interval $[a,b]$
$\mathcal{M}(E)$	probability measures on a metric space E

10.4 Miscellaneous

$\mathcal{L}(X)$	distribution/law of a random variable/element X
$\mathcal{N}(\mu,\sigma^2)$	normal distribution
$\mathrm{Exp}(\lambda)$	exponential distribution
$\mathrm{Poi}(\lambda)$	Poisson distribution

Index

References

[Bil99] Patrick Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Probability and Statistics. Wiley, 2nd ed edition, 1999.

[Dur19] Richard Durrett. *Probability: Theory and Examples*. Number 49 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, fifth edition edition, 2019.

[Kol33] A.N. Kolmogorov. *Foundations of the Theory of Probability (English Translation)*. 1933.

[KS] Ioannis Karatzas and Steven Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer-Verlag, 2 edition.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer New York, 1996.