Lecture Note for MAT8030: Advanced Probability

LI Liying*

October 21, 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. $\mathsf{P}(\varnothing)=0$.

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

3. **countable additivity**: for any *disjoint* $A_1, A_2, \ldots$,

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

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,

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

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

$$F = F_d + F_{ac} + F_{sc}. \tag{1.4}$$

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*

$$F(b) - F(a) = \int_a^b f(x)\,dx. \tag{1.5}$$

Remark 1.10 This is the generalized Newton–Leibniz formula. By Lebesgue Differentiability Theorem, if (1.5) 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 $P(A) = 0$.*

If $P \ll Q$, then by Radon–Nikodym Theorem, there exists a measurable function $f(\omega) \in L^1(Q)$, such that $P(A) = \int_A f(\omega) Q(d\omega)$. We write $f(\omega) = \frac{dP}{dQ}$. The measure Q is called 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 $\mu_X \perp \mathrm{Leb}$.*

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 1.17 *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 1.17, we have the following.

Proposition 1.18 *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 1.19 *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,

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

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 1.20 *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 1.20 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 1.21 *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 1.22 *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 1.21, $Y_1 = \limsup_{n\to\infty} X_n(\omega)$ 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

$$\mathsf{E}X = \int_\Omega X(\omega)\, \mathsf{P}(d\omega). \tag{1.6}$$

The definition of (1.6) 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

$$X(\omega) = \sum_{k=1}^{n} c_k \mathbb{1}_{A_k}(\omega). \tag{1.7}$$

In the case of (1.7), unquestionably we should define

$$\mathsf{E}(X) = \sum_{k=1}^{n} c_k \mathsf{P}(A_k). \tag{1.8}$$

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

$$\mathsf{E}X = \int_\Omega X(\omega)\mathsf{P}(d\omega) := \sup\Big\{\int Y(\omega)\mathsf{P}(d\omega) : Y \text{ simple},\ 0 \le Y(\omega) \le X(\omega)\Big\} \in [0, \infty]. \tag{1.9}$$

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.,

$$\mathsf{E} \lim_{n\to\infty} X_n = \lim_{n\to\infty} \mathsf{E}X_n. \tag{1.10}$$

Lemma 1.23 *Let $X_n \uparrow X$ such that $X_n \geq 0$ and X_n are simple. Then* (1.10) *holds.*

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

Proof: From the definition (1.9), we have $\mathsf{E}(X) \geq \mathsf{E}(X_n)$. It remains to establish the inequality in the other direction:

$$\mathsf{E}X \leq \lim_{n\to\infty} \mathsf{E}X_n. \tag{1.11}$$

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 $Y_\varepsilon \leq X$ and $\mathsf{E}(Y_\varepsilon) \geq \mathsf{E}(X) - \varepsilon$. For every $\delta > 0$, let $A_n = \{\omega : X_n(\omega) \leq Y_\varepsilon(\omega) - \delta\}$. Since $X_n(\omega) \uparrow X(\omega) \geq Y_\varepsilon(\omega)$, we have $A_n \uparrow \Omega$ and hence $A_n^c \downarrow \emptyset$. We have

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

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

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

Since $\varepsilon, \delta > 0$ are arbitrary, this implies (1.11).

If $\mathsf{E}X = \infty$, then by (1.9), for every $M > 0$, there exists a simple r.v. Y_M such that $Y_M \leq X$ and $\mathsf{E}Y_M \geq M$. For every $\xi > 0$, let $B_n = \{\omega : X_n(\omega) \geq Y_M(\omega) - \xi\}$. Since $X_n(\omega) \uparrow X(\omega) \geq Y_M(\omega)$, we have $B_n \uparrow \Omega$ and hence $B_n^c \downarrow \emptyset$. Therefore,

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

Letting $n \to \infty$, we obtain $\lim_{n\to\infty} \mathsf{E}X_n \geq M - \xi$. Since $M, \xi > 0$ are arbitrary, this implies (1.11). □

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 1.23 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 $a \wedge b \coloneqq \min(a, b)$ 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 1.24 (Monotone Convergence Theorem, MCT) *If $X_n \geq 0$ and $X_n \uparrow X$, then* (1.10) *holds.*

Proof: Again, it suffices to establish (1.11).

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

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

Moreover, we have

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

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

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

Taking $n \to \infty$, and using that $X_n \uparrow X$, we see that $Z^{(m)} \uparrow X$. Then by Lemma 1.23, we have

$$\mathsf{E}X = \lim_{m\to\infty} \mathsf{E}Z^{(m)}. \tag{1.12}$$

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

$$\lim_{m\to\infty} \mathsf{E}Z^{(m)} \le \lim_{m\to\infty} \mathsf{E}X_m. \tag{1.13}$$

Then (1.11) follows from (1.12) and (1.13), and this completes the proof.

□

Remark 1.18 In Theorem 1.24, the condition "$X_n \ge 0$" can be replaced by

$$\text{"}X_n \ge -Y\text{, for some } Y \ge 0 \text{ with } \mathsf{E}Y < \infty\text{"}. \tag{1.14}$$

Indeed, if (1.14) holds, then $\tilde{X}_n = X_n + Y \ge 0$. 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 $0 \le \mathsf{E}Y < \infty$, 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 1.25 (Fatou's Lemma) *If $X_n \ge 0$ (or (1.14) holds), then*

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

Proof: Let

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

Clearly, $Y_n \le X_n$. By MCT (Theorem 1.24), we have

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

□

Theorem 1.26 (Dominated Convergence Theorem, DCT) *If $X_n \to X$ a.s. and $|X_n| \le Y$ for some r.v. Y with $\mathsf{E}Y < \infty$, then $\lim_{n\to\infty} \mathsf{E}X_n = \mathsf{E}X$.*

Proof: Clearly, $|X| \le Y$. Since $2Y - |X_n - X| \ge 0$, by Fatou's Lemma(Theorem 1.25), we have

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

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

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

□

Corollary 1.27 (Bounded Convergence Theorem, BCT) *If $X_n \to X$ and $|X_n| \le M$, $n \ge 1$ 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 1.28 (Jensen inequality) *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. If $\mathsf{E}|x| < \infty$, then $\mathsf{E}\varphi(x) \ge \varphi(\mathsf{E}X)$.*

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.

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

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

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

□

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

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

Proposition 1.29 (Hölder's inequality) *If $p, q \in [1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$ then*

$$\mathsf{E}|XY| \le (\mathsf{E}|X|^p)^{1/p} \cdot (\mathsf{E}|Y|^q)^{1/q}. \tag{1.15}$$

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

$$xy \le \frac{x^p}{p} + \frac{y^q}{q}, \quad x, y \ge 0. \tag{1.16}$$

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 (1.16), we have

$$\mathsf{E}\tilde{X}\tilde{Y} \le \frac{\mathsf{E}|\tilde{X}|^p}{p} + \frac{\mathsf{E}|\tilde{Y}|^q}{q} = \frac{1}{p} + \frac{1}{q} = 1 \tag{1.17}$$

This is (1.15).

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

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

Taking $M \uparrow \infty$, since $X_M \uparrow X$ and $Y_M \uparrow |Y|$, (1.15) 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 1.30 (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*

$$\mathsf{E}f(X) = \int_\Omega f\big(X(\omega)\big)\,\mathsf{P}(d\omega) = \int_{\mathbb{R}} f(y)\,\mu_X(dy), \tag{1.18}$$

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

Proof: Let

$$\mathcal{H} = \{f : f \text{ is Borel measurable s.t. (1.18) holds }\}.$$

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

   $$a_1 f_1 + \cdots + a_n f_n \in \mathcal{H},$$

   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 $f_n \ge 0$ 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, (1.18) 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

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

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

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

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

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

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 $\varepsilon > 0$,*

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

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

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

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

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

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 $|X - Y| \le |X_n - X| + |X_n - Y|$, for every $\varepsilon > 0$,

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

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

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

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

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

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 (1.10) 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 $\big(\mathsf{E}X_n\big)_{n\ge1}$, and hence (1.10) 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 $\varepsilon > 0$,*

$$\mathsf{P}\big(|X| \geq \varepsilon\big) \leq \frac{\mathsf{E}|X|}{\varepsilon}$$

Proof: Since

$$|X| = |X|\mathbb{1}_{\{|X|\geq\varepsilon\}} + |X|\mathbb{1}_{\{|X|\geq\varepsilon\}} \geq |X|\mathbb{1}_{\{|X|\geq\varepsilon\}} \geq \varepsilon\mathbb{1}_{\{|X|\geq\varepsilon\}}, \tag{2.6}$$

taking expectation on both sides, we have $\mathsf{E}|X| \geq \varepsilon\mathsf{P}\{|X| \geq \varepsilon\}$, 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

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

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 $c \geq 1/p$. We have $\mathsf{E}|X_{n,k}|^p \geq 1$ 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*

$$\lim_{n\to\infty}\sup_{\alpha\in I}\mathsf{E}|X_\alpha|\mathbb{1}_{\{|X_\alpha|\geq M\}} = 0. \tag{2.7}$$

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

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

Uniform integrability can be seen as a necessary and sufficient condition for (1.10) to hold. Therefore, it will be the last resort if conditions for Theorems 1.24 to 1.26 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. *$\{X_n\}_{n\ge 1}$ 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

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

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

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

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 (2.7):

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

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

From 3 to 1. Let

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

Let $\varepsilon > 0$. We have

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

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 $\psi_M(|X|) \le \mathsf{E}|X|$, by DCT there exists $M_0 > 0$ such that

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

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

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

It follows that $(X_n)_{n\ge 1}$ 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 $\mathsf{P}(X \leq a) = F(a)$? 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 $(\Omega, \mathcal{F}, \mathsf{P}) = \big((0,1), \mathcal{B}(0,1), \text{Leb}\big)$. 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 $(\xi_n)_{n\geq 1}$ 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

$$\begin{aligned} \Omega &= \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1,\ \omega_2 \in \Omega_2\}, \\ \mathcal{F} &= \mathcal{F}_1 \otimes \mathcal{F}_2 = \sigma(A \times B : A \in \mathcal{F}_1,\ B \in \mathcal{F}_2). \end{aligned} \tag{3.6}$$

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

$$\mu_0(A \times B) = \sum_{n=1}^{\infty} \mu_0(A_n \times B_n). \tag{3.7}$$

(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

$$B = \bigcup_{n \in I(x)} B_n, \quad \forall x \in A, \tag{3.8}$$

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

$$\mathbb{1}_A(x) \cdot \mathsf{P}_2(B) = \mathbb{1}_A(x) \cdot \sum_{n \in I(x)} \mathsf{P}_2(B_n) = \sum_{n \in I(x)} \mathbb{1}_{A_n}(x)\ \mathsf{P}_2(B_n) = \sum_{n \geq 1} \mathbb{1}_{A_n}(x)\ \mathsf{P}_2(B_n). \tag{3.9}$$

The first equality holds since we have (3.8) 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 (3.9) 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 (3.9) 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

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

where we use MCT in the second line. This proves (3.7) 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*

$$f \ge 0, \tag{3.10a}$$

$$\text{or} \quad \int |f(\omega_1, \omega_2)|\, (\mathsf{P}_1 \times \mathsf{P}_2)(d\omega_1 d\omega_2) < \infty. \tag{3.10b}$$

Then the following holds.

1. *For every $\omega_1 \in \Omega$, the function $f(\omega_1, \cdot)$ is $\mathcal{F}_2$-measurable. And if (3.10b) holds,*

$$\int_{\Omega_2} |f(\omega_1, \omega_2)|\, \mathsf{P}_2(d\omega_2) < \infty, \quad \text{for almost every } \omega_1 \in \Omega. \tag{3.11}$$

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 (3.10b) holds,*

$$\int_{\Omega_1} |g(\omega_1)| \mathsf{P}_1(d\omega_1) < \infty. \tag{3.12}$$

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

$$\begin{aligned}\iint_{\Omega_1\times\Omega_2} f(\omega_1,\omega_2)\,(\mathsf{P}_1\times\mathsf{P}_2)(d\omega_1 d\omega_2) &= \int_{\Omega_1}\mathsf{P}_1(d\omega_1)\int_{\Omega_2} f(\omega_1,\omega_2)\,\mathsf{P}_2(d\omega_2)\\ &= \int_{\Omega_2}\mathsf{P}_2(d\omega_2)\int_{\Omega_1} f(\omega_1,\omega_2)\,\mathsf{P}_1(d\omega_1).\end{aligned} \tag{3.13}$$

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 (3.10a) or (3.10b) 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 (3.11) to (3.13).

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 (3.13) 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 (3.11).

   Applying Fubini's Theorem to $|f|\ge 0$, 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 (3.11).

□

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* (3.10a) *or* (3.10b)*. 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* (3.10b) *holds, the set N can be chosen such that for $\omega_1 \in N^c$,*

$$\int |f(\omega_1, \omega_2)| \, \mathsf{P}_2(d\omega_2) < \infty. \tag{3.14}$$

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 (3.10b) *holds, then* (3.12) *is true.*

3. (3.13) *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

$$\begin{aligned} \int q(\omega_1) \, \mathsf{P}_1(d\omega_1) &= \int [\mathsf{P}_2(D^+_{\omega_1}) - \mathsf{P}_2(D^-_{\omega_1})] \, \mathsf{P}_1(d\omega_1) \\ &= \int \left(\mathbb{1}_{D^+}(\omega) - \mathbb{1}_{D^-}(\omega) \right) (\mathsf{P}_1 \times \mathsf{P}_2)(d\omega_1 d\omega_2) = (\mathsf{P}_1 \times \mathsf{P}_2)(D^+ \setminus D^-) = 0. \end{aligned} \tag{3.15}$$

Since $q(\omega_1) \geq 0$, (3.15) 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

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

and it is an element of $\overline{\mathcal{F}_1}$ by definition. It is easy to verify (3.13).

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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_i)$, $1 \le i \le n$, 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 $\omega = (\omega_1, \dots, \omega_n)$. Let $X_i(\omega) = \omega_i$, $1 \le i \le n$. Then $\{X_i\}_{1 \le i \le n}$ 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 $\big(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty)\big)$.

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:

$$d(x^{(m)}, x^{(0)}) \to 0, \ m \to \infty \quad \Leftrightarrow \quad \pi_n x^{(m)} \to \pi_n x^{(0)}, \ \forall n \ge 1. \tag{3.16}$$

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

$$\mathcal{B}(\mathbb{R}^\infty) = \sigma\Big(O_n \times \mathbb{R}^\infty, O_n \text{ open set in } \mathbb{R}^n\Big). \tag{3.17}$$

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,

$$\mu_{n+m} \circ \pi_n^{-1} = \mu_n, \quad \forall m, n \ge 1. \tag{3.18}$$

Theorem 3.9 (Kolmogorov's Extension Theorem) *Assume* (3.18). *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* $n \geq 1$, *i.e.,*

$$\mu(A \times \mathbb{R}^\infty) = \mu_n(A), \quad \forall A \in \mathcal{B}(\mathbb{R}^n). \tag{3.19}$$

To construct an infinite sequence of independent r.v.'s, we will use Theorem 3.9 in the following way. Given $\big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda_i\big)$, $i \geq 1$, 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 (3.18) 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 (3.19) holds. Let

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

Then $(X_n)_{n\geq 1}$ 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 (3.16), 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 (3.19) specifies the measure μ on $\mathcal{C}$.

3. (3.17) implies that $\mathcal{B}(\mathbb{R}^\infty) = \sigma(\mathcal{C})$.

4. The consistency condition (3.18) implies that (3.19) 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 $\mu(D_n \setminus \hat{E}_n) = \mu_n(B_n \setminus F_n) \leq \delta 2^{-n-1}$. The sets $\hat{E}_n$ may not be decreasing, but if we set

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

then E_n are decreasing. Moreover,

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

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

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

All cylinder sets form an algebra, and a probability measure μ on this space exists, provided that its "finite-dimension distributions" $\mu \circ \pi_{t_1,\ldots,t_n}^{-1}$ 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

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

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 $X_1, X_2, \ldots$ 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

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

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 $n \geq n_0$, $|D_n| \leq n^{1+\varepsilon}$.

When the two conditions above hold for ω, by

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

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:

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

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

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

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 $|\xi_i| \leq 1$. For every k, By SLLN, for almost every $\omega \in [0,1)$,

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

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 $\vec{k} = (k_1, k_2, \ldots, k_r) \in \{0,\ldots,9\}^r$,

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

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

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

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

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

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 Strong LLN

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 (The Second Borel-Cantelli lemma) *If the events A_n are independent then $\sum \mathsf{P}(A_n) = \infty$, then $\mathsf{P}(A_n$ i.o.$) = 1$*

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

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

It follows that

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

□

Proposition 4.10 *If $X_1, X_2 \ldots$ are i.i.d. with $\mathsf{E}|X_i| = \infty$, then $\mathsf{P}(|X_n| \geq n$ i.o.$) = 1$. So if $S_n = X_1 + \cdots + X_n$ then $\mathsf{P}(\lim S_n/n$ exists $\in (-\infty, \infty)) = 0$.*

Proof: Let $A_n = \{|X_n| \geq n\}$. 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 from the second Borel-Cantelli lemma that $\mathsf{P}(|X_n| \geq n \text{ i.o.}) = 1$. If there exist $a = \lim_{n\to\infty} \frac{S_n}{n} \in (-\infty, \infty)$, by Cauchy criterion, for $\varepsilon_0 = \frac{1}{2}$, there exists $n_0 = n_0(\omega)$ such that $|\frac{S_n}{n} - \frac{S_{n-1}}{n-1}| < \frac{1}{2}$ for every $n > n_0$. So

$$\{\text{exists } \lim_{n\to\infty} \frac{S_n}{n} \in (-\infty, \infty)\} \subset \{|X_n| \geq n \text{ i.o.}\}^c$$

□

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 $S_n/n \log_2 n \not\to 1$.

5 Notations

5.1 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

5.2 Sets

$\mathbb{Z}$	set of integers
$\mathbb{N}$	set of natural numbers $\{0, 1, 2, \dots\}$
$\mathbb{Q}$	set of rational numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$ (resp. $\mathbb{R}_-$)	set of non-negative (resp. non-positive) real numbers

5.3 Relations

$\Rightarrow_d$ or $\Rightarrow$	convergence in distribution/law
$\stackrel{\mathrm{d}}{=}$	equal in law

5.4 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

5.5 Operations

$a \wedge b$	$\min(a, b)$
$a \vee b$	$\max(a, b)$
$\langle a, b \rangle$	inner product in a Euclidean space/Hilbert space (or) a linear functional a in the dual space $\mathcal{X}^*$ acting on an element b in a Banach space $\mathcal{X}$
$A \Delta B = (A \setminus B) \cup (B \setminus A)$	the difference set.

5.6 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

additivity
  countable, σ-, 2
  finite, 2
  sub-, 2
algebra, 5
  σ-, or σ-field, 3
  Borel σ-, 4
  semi-, 5

Borel
  σ-algebra, 4
  (measurable) function, 12
  (measurable) set, 12
  space, 34
Borel–Cantelli Lemma
  First, 19
Brownian motion, 11, 34

Cantor
  function, 9
  set, 9
Carathédory's
  condition, 8
  Extension Theorem, 7
Chebyshev's inequality, 21
consistency condition, 32, 34
continuity
  absolute (for functions), 9
  absolute (for measures), 10
  at $\varnothing$, 6
  from above (fom measures), 2
  from below (for measures), 2
continuous
  absolutely, 9, 10
  singularly, 9
convergence
  almost sure, 18
  in L^p, 18
  in distribution, 18
  in probability, 18
  weak, weak-*, 23
Convergence Theorem
  Bounded (BCT), 16
  Dominated (DCT), 15
  Monotone (MCT), 14
cumulative distribution function/c.d.f., 4

distribution
  of a r.v., 4
distribution function
  cumulative (c.d.f.), 1
  empirical, 37

Fatou's Lemma, 15
finite-dimensional distribution, 34
Fubini's Theorem, 28
  for complete measure spaces, 31

generalized inverse, 26

Hölder's inequality, 16

independence
  for σ-algebras, 24
  for a collection of r.v.'s, 25
  for events/sets, 23
  for r.v.'s, 23
  pairwise, 25
independent and identically distributed, 25
inequality
  Chebyshev's, 21
  Hölder's, 16
  Jensen, 16
  Young's, 16
infinitely often, 19

Jensen inequality, 16

Kolmogorov's
  Extension Theorem, 33
  Extension Theorm for general spaces, 34

law of large nubmers
  weak, 36
law of large numbers
  strong (SLLN), 35

measurable
  Borel-, 11
  map, 11
  space, 3
measure, 1
  Dirac, 10
  Lebesgue, 8
  outer, 7

product, 27
reference, 10
signed, 23
monotone class, 5
Monotone Class Theorem, 5
functional, 17

normal number, 37

principle of appropriate sets, 5

random variable, 3
continuous, 10
simple, 13
singular, 10
random vector, 11

set
Cantor, 9
cylinder, 32
i.o., 19
simple r.v./function, 13
singularity (for measures), 10
space
Borel, 34
complete measure, 8
dual, 23
measurable, 3
measure, 3
probability, 3
stochastic process, 34
system
π-, 6
Dynkin, d-system, λ-class, 6

uncorrelated r.v.'s, 35
uniform integrability, 21

Young's inequality, 16

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.