Lecture Note for MAT8030: Advanced Probability

LI Liying*

September 28, 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-field. 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 $\mathsf{P}(\Omega) = 1$.*

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*

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

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

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

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 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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*

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, i.e., $A, B \in \mathcal{S} \Rightarrow A \cap B \in \mathcal{S}$ and second, for every $A \in \mathcal{S}$, A^c is disjoint union of $A_1, A_2, \ldots, A_n$ 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 (m-class)*, if for every monotone sequence $A_n \in \mathcal{S}$, $A = \lim_{n\to\infty} A_n \in \mathcal{S}$.*

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 m-classes is still an m-class. Therefore, it makes sense to talk about the *smallest* m-classes containing any collection of sets $\mathcal{A}$ (c.f. Proposition 1.4). We denote this smallest m-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 an m-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 an m-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 an m-class: if $A_n \uparrow A$, $A_n \in \mathcal{S}$, then A_n, A_n^c are both monotone sequence in $m(\mathcal{A})$, and hence their limits $A, A^c \in m(\mathcal{A})$; if $A_n \downarrow A$ it is similar. Therefore, $\mathcal{S}$ must contain the smallest m-class that contains $\mathcal{A}$, which is $m(\mathcal{A})$. This shows $\mathcal{S} = m(\mathcal{A})$, so by the definition of $\mathcal{S}$, $m(\mathcal{A})$ 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 fixed $A \in \mathcal{A}$, let

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

Clearly, $\mathcal{A} \subset \mathcal{S}_A$ since A is an algebra and $m(\mathcal{A})$ contains $\mathcal{A}$. Also, $\mathcal{S}_A$ is an m-class as $B_n \downarrow B$ or $B_n \uparrow B$ implies $A \cap B_n \downarrow A \cap B$ or $A \cup B_n \uparrow A \cup B$. Therefore, $m(\mathcal{A}) \subset \mathcal{S}_A$, and we have shown that $A \cap B \in m(\mathcal{A})$ whenever $A \in \mathcal{A}$ and $B \in m(\mathcal{A})$.

In the second step, let

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

By the first step, $\mathcal{A} \subset \mathcal{S}$. Again, it is not hard to check that A is an m-class. Therefore $m(\mathcal{A}) = \mathcal{S}$, 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

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

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*

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

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 for every $A_n \downarrow \varnothing$, $\lim_{n\to\infty} \mu_0(A_n) = \mu_0(\varnothing) = 0$.

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 mostly a consequence of the right continuity of F. In fact, let $A_n = \bigcup_{i=1}^{m} (a_i^{(n)}, b_i^{(n)}]$, and $B_n = \bigcup_{i=1}^{m} (a_i^{(n)} + \delta, b_i^{(n)}]$. Then

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

We just need to choose δ small enough so that the sum is less that $\varepsilon \cdot 2^{-n}$.

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., $\exists n_0$ s.t.

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

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

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

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

For unbounded A_n, since $F(-\infty) = 0$ and $F(\infty) = 1$, for every $\varepsilon > 0$, we can choose L s.t. $\mu_0((-L, L]) \ge 1 - \varepsilon$. Let $\tilde{A}_n = A_n \cap (-L, L]$. Then $\tilde{A}_n \downarrow \varnothing$ and $\tilde{A}_n$ are bounded. Hence, $\lim_{n\to\infty} \mu_0(\tilde{A}_n) = 0$. Therefore,

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

Since $\varepsilon > 0$ is arbitrary, we see that $\lim_{n\to\infty} \mu_0(A_n) = 0$, 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, $\tilde{\mu}$ 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 $A_n \in \mathcal{S}$, using the continuity 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 part 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

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

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

seen from that it is 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, $F(x) = x$ 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* probability 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's 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.,

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

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.

Then the Cantor function, denoted by $\varphi(x)$, is 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 $x \in (\frac{7}{9}, \frac{8}{9}) = I_2^{(2)}$, ... Then define φ on $\mathcal{C}$ by continuity. 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 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 measurable 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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

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

Let $I = \{x : \mu(\{x\}) > 0\}$ 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 absolute 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

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

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

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

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,

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

This shows that $f(X)$ is a r.v. □

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

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

Proposition 1.20 *If $f : \mathbb{R} \to \mathbb{R}^d$ is a Borel map and $X = (X_1, \cdots, X_d)$ is a random vector, then $f(X) = f(X_1, \cdots, X_d)$ 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 random variables.

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 random variables.

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) Since for every ω,

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

we have for every $a \in \mathbb{R}$,

$$\begin{aligned}
\{\omega : \limsup_{n\to\infty} X_n(\omega) > a\} &= \{\omega : \inf_{n\geq 1} \sup_{m\geq n} X_m(\omega) > a\} \\
&= \bigcap_{n=1}^{\infty} \{\omega : \sup_{m\geq n} X_m(\omega) > a\} \\
&= \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{\omega : X_m(\omega) > a\}
\end{aligned}$$

(iv) Since for every ω,

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

we have for every $a \in \mathbb{R}$,

$$\begin{aligned}\{\omega : \liminf_{n\to\infty} X_n(\omega) < a\} &= \{\omega : \sup_{n\geq 1} \inf_{m\geq n} X_m(\omega) < a\} \\ &= \bigcap_{n=1}^{\infty} \{\omega : \inf_{m\geq n} X_m(\omega) < a\} \\ &= \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{\omega : X_m(\omega) < a\}.\end{aligned}$$

□

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 a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. 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 random variables (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

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

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 we cannot define $\infty - \infty$.

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 $X_n \ge 0$ and X_n are simple. Then* (1.10) *holds.*

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

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

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

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

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

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

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

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

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

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

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

Note that for any non-negative r.v. X, one can find simple r.v.'s $X_n \uparrow X$ so that Lemma 1.23 applies. A simple construction is

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

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

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

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

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

Moreover, we have

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

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

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

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

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

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

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

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

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

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

Indeed, if (1.14) holds, then $\tilde{X}_n = X_n + Y \geq 0$. Since $\tilde{X}_n \uparrow \tilde{X} = X + Y$, we have

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

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

Theorem 1.25 (Fatou's Lemma) *If $X_n \geq 0$ (or (1.14) holds), then*

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

Proof: Let

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

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

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

□

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

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

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

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

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

□

Corollary 1.27 (Bounded Convergence Theorem, BCT) *If $X_n \to X$ and $|X_n| \leq M$, $n \geq 1$ for some constant M, then $\lim_{n\to\infty} \mathsf{E}X_n = \mathsf{E}X$.*

Proof: Take $Y(\omega) \equiv M$. □

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

Proposition 1.28 (Jensen's Inequality) *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. If $\mathsf{E}|x| < \infty$, then $\mathsf{E}\varphi(x) \geq \varphi(\mathsf{E}X)$.*

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

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

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

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

□

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

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

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

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

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

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

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

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

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

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

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

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

   By MCT, (1.18) 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

$$\begin{aligned} \mathsf{P}\{\limsup_{n\to\infty} |X_n - X| > \varepsilon\} &= \mathsf{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{\omega : |X_m(\omega) - X(\omega)| > \varepsilon\Big) \\ &= \lim_{n\to\infty} \mathsf{P}\Big(\bigcup_{m=n}^{\infty} \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\Big) \\ &\geq \limsup_{n\to\infty} \mathsf{P}\big(|X_n(\omega) - X(\omega)| > \varepsilon\big) \end{aligned} \tag{2.2}$$

Hence, $X_n \to X$ in pr. □

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

$$(\Omega, \mathcal{F}, \mathsf{P}) = ([0,1], \mathcal{B}([0,1]), \text{Leb}), \quad X_{n,k}(\omega) = \mathbb{1}_{\frac{k}{n}, \frac{k+1}{n}}(\omega), 0 \leq k \leq n-1. \tag{2.3}$$

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

$$\{A_n, \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \limsup_{n\to\infty} A_n. \tag{2.4}$$

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 (2.4), we have

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

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

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

and the conclusion follows. □

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

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

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

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

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

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

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

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

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

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

For such ω, the infinite series

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

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

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

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

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

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

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

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

and this completes the proof. □

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

Proposition 2.6 *a.s. 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 (2.3)). Then, there exists $\varepsilon_0 > 0$ and a sequence (n') such that

$$d(X_{n'}, X) \geq \varepsilon_0 > 0. \tag{2.5}$$

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 (2.5). □

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

$$d(X, Y) = \mathsf{E}\Big[\frac{|X-Y|}{1+|X-Y|}\Big]. \tag{2.6}$$

Of course, one need to verify that (2.6) 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 $\big(\mathsf{E}X_n\big)_{n\ge 1}$, 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 $\varepsilon > 0$,*

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

Proof: Since

$$|X| = |X|\mathbb{1}_{\{|X|\ge\varepsilon\}} + |X|\mathbb{1}_{\{|X|\ge\varepsilon\}} \ge |X|\mathbb{1}_{\{|X|\ge\varepsilon\}} \ge \varepsilon\mathbb{1}_{\{|X|\ge\varepsilon\}}, \tag{2.7}$$

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

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

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

Therefore, $X_n \to X$ in pr. □

Limits in L^p are also unique.

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

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

Other than Proposition 2.1 and Proposition 2.8, there are not more implications between the three modes of convergence. One counterexample is given in (2.3), counterexamples for the other implications can be obtained by modifying (2.3).

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

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

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

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

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

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

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

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

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

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

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

$$\lim_{n\to\infty} \sup_{\alpha\in I} \mathsf{E}|X_\alpha| \mathbb{1}\{|X_\alpha| \geq M\} = 0. \tag{2.8}$$

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

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

Uniform integrability can be seen as a necessary and sufficient condition for (1.10) 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. *$\{X_n\}_{n\geq 1}$ are u.i.;*
2. *$X_n \to X$ in L^1;*
3. *$\mathsf{E}|X_n| \to \mathsf{E}|X|$.*

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

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

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

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

Taking $n \to \infty$ and then $M \to \infty$, the first term goes to 0 by DCT, the second goes to zero since X_n are u.i., and the third goes to zero since $\mathsf{E}|X| < \infty$ which follows from Fatou's lemma and (2.8):

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

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

From 3 to 1. Let

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

Let $\varepsilon > 0$. We have

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

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

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

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

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

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

2.4 Weak convergence

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

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

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

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

We also note that for weak convergence, the probability spaces on which the r.v.'s X_n, 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

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

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

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

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. □

3.2 Product measures

4 Notations

4.1 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

4.2 Sets

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

4.3 Relations

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

4.4 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

4.5 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.

4.6 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

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.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer New York, 1996.