Review of Probability Theory I

Park, Sihyung
naturale0@snu.ac.kr

Last Update: December 2020

Contents

1 Measure Theory

1.1 Basics of Measure Theory

The first chapter is devoting to the basics of measure theory. A probability space is defined and essential theorems are introduced.

1.1.1 Probability space

A probability space is a special kind of a measure space equipped with a positive finite measure. A measure space is defined as a triplet: a set, a σ-field attached to that set, and a measure function. We will define each of the components.

Definition 1 (σ-field)**.** *For a set Ω, $\mathcal{F}$, a non-empty collection of subsets of Ω, is called a σ-field(or a σ-algebra) of Ω, if the following conditions are satisfied.*
(i) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
(ii) If $A_i \in \mathcal{F},\ i = 1, 2, \cdots$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A σ-field is basically a set of sets. The first condition states that it is closed under complement and the second one states that it is closed under countable unions.

We next define a function that *measures* sizes of sets inside the σ-field.

Definition 2 (measure)**.** *$\mu : \mathcal{F} \to \mathbb{R}$ is a measure, if*
(i) $\mu(A) \geq \mu(\phi) = 0,\ \forall A \in \mathcal{F}$.
(ii) $A_i \in \mathcal{F},\ i = 1, 2, \cdots$ are disjoint. Then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

The second condition is sometimes referred to as *σ sub-additivity (countable sub-additivity)*. This is natural if we think of a common notion of a measure: if it is empty, its "size" should be zero and if we add one with another, the resulting size should be the sum. We call $\mu(A)$ the measure of a set A.

If there exists a sequence of sets $\{A_n\} \subset \mathcal{F}$ such that $\mu(A_n) < \infty$ for all n and $\cup_{n=1}^{\infty} A_n = \Omega$, then μ is called a *σ-finite* measure. If the measure of the whole set $\mu(\Omega)$ is finite, we call μ a finite measure. If $\mu(\Omega) = 1$ in addition, then we call this a *probability measure* (PM for abbreviation in the following post series). Most of the times, we name a PM with alphabet P or Q.

Finally we can define a measure space and in addition a probability space.

Definition 3 (probability space)**.** *$(\Omega, \mathcal{F})$: a pair of a set and its σ-field, is called a measurable space. A set $A \in \mathcal{F}$ is called a ($\mathcal{F}$-)measurable set. $(\Omega, \mathcal{F}, P)$: a measurable space equipped with a measure is a measure space. If P is a probability measure, we call this a probability space, Ω a sample space, and an element of $\mathcal{F}$ an event.*

If one is familiar with topology, the definition of σ-field might also be quite familiar. In fact, Borel σ-field connects a topological space with a corresponding measurable space.

Definition 4 (Borel σ-field)**.** *$\mathcal{B}(\Omega) = \cap_{\tau \subset \mathcal{F}} \mathcal{F}$ is the Borel σ-field of Ω, where τ is a topology on Ω. An element of $\mathcal{B}(\Omega)$ is called a Borel set.*

$\mathcal{B}(\Omega)$ is the smallest σ-field that contains all open sets of Ω. An important property of σ-fields related to the definition of Borel fields is that arbitrary intersections of σ-fields is a σ-algebra. (This comes directly from the definition.)

As an example of a Borel σ-field and a measure space, consider $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and the Lebesgue measure λ. It is not difficult to know that $\mathcal{B}(\mathbb{R})$ is consists of intersections and unions of sets of the form $(a, b),\ (a, b],\ [a, b)$ or $[a, b]$ where $a, b \in \mathbb{R}$. The Lebesgue measure measures their "size" as $\lambda(a, b] = b - a$.

1.1.2 Probability measures

Let's take a deeper look into the properties of measure and probability measures.

Proposition 1 (properties of probability measures)**.** *Let $P : \mathcal{F} \to [0,1]$ be a probability measure. Then the following properties hold.*
(i) (monotonicity) $A \subset B \implies P(A) \leq P(B)$.
(ii) (σ sub-additivity) $A \subset \cup_{i=1}^{\infty} A_i \implies P(A) \sum_{i=1}^{\infty} P(A_i)$.
(iii) (continuity from below) $A_1 \subset A_2 \subset \cdots$, $\bigcup_{i=1}^{\infty} A_i = A \implies \lim_n P(A_n) = P(A)$.
(iv) (continuity from above) $B_1 \supset B_2 \supset \cdots$, $\bigcap_{i=1}^{\infty} B_i = B \implies \lim_n P(B_n) = P(B)$.

Proof. (i) $B = B - A + A$, $B \neq \phi$. Thus $P(B) = P(B - A) + P(A) \geq P(A)$.
(ii) Let $A'_n = A_n \cap A$, $B_n = A'_n \setminus \cup_{i=1}^{n-1} A_i$. Then B_n's are disjoint and $\cup_{n=1}^{\infty} B_n = A$, $B_n \subset A_n$. Hence $P(A) = \sum_{n=1}^{\infty} P(B_n) \leq \sum_{n=1}^{\infty} P(A_n)$.
(iii) Let $B_n = A_n \setminus A_{n-1}$, $A_0 = \phi$ so that B_n's be disjoint. Then

$$\begin{aligned} P(\cup_n A_n) = P(\cup_n B_n) &= \sum_n P(B_n) \\ &= \lim_n P(\cup_{i=1}^{n} B_i) \\ &= \lim_n P(\cup_{i=1}^{n} A_i) \\ &= \lim_n P(A_n) \end{aligned}$$

(iv) Let $B'_n = B_1 \setminus B_n$ and use (iii). □

1.1.3 Characterization of a probability measure

Until now, we defined measures only on σ-fields. Measures in other set of subsets can be defined similarly. Furthermore, we can characterize each measure as an extension of the function similar to the measure defiend above. For this, we first define collections of sets that can be viewed as generalizations of σ-fields.

Definition 5 (semi-algebra)**.** *A collection of sets $\mathcal{S}$ is a semi-algebra, if*
(i) For all $S \in \mathcal{S}$, S^c is a finite disjoint unions of $S_i \in \mathcal{S}$.
(ii) $S, T \in \mathcal{S} \implies S \cap T \in \mathcal{S}$.

It is not necessary for a semi-algebra to contain ϕ. However in many cases it is convenient to make it do so.

Definition 6 (algebra)**.** *A collection of sets $\mathcal{A}$ is an algebra, if*
(i) $A \in \mathcal{A} \implies A^c \in \mathcal{A}$.
(ii) $S, T \in \mathcal{S} \implies S \cap T \in \mathcal{S}$.

Note that the strengthened condition on the definition of algebra allows it to be closed on both finite intersections *and* unions.

For example, $\mathcal{S}_1 = \{\phi\} \cup \{(a, b] : \ -\infty \leq a < b \leq \infty\}$ is a semi-algebra on $\mathbb{R}$. $\mathcal{A} = \{A \in \mathbb{Z} : \ A \text{ or } A^c \text{ is finite}\}$ is an algebra on $\mathbb{Z}$. It is trivial so I will leave it as exercises.

An algebra can be generated by semi-algebra.

$$\overline{\mathcal{S}} := \{\text{finite disjoint unions of sets in } \mathcal{S}\}$$

is an *algebra generated by a semi-algebra* $\mathcal{S}$. Sometimes we call $\mathcal{S}$ a generator of $\overline{\mathcal{S}}$. Like Borel σ-field, it is the smallest algebra that contains $\mathcal{S}$.

Similarly, a σ-field can be generated by (semi)algebra. $\sigma(\mathcal{S}) = \sigma(\overline{\mathcal{S}})$ is the smallest σ-field that contains $\mathcal{S}$.

Now we define "measures" on these structures.

Definition 7 (measure on algebra)**.** $\mu : \mathcal{A} \to \mathbb{R}^+ \cup \{0\}$ *is a measure on an algebra* $\mathcal{A}$ *if*
(i) $\mu(A) \geq \mu(\phi) = 0,\ \forall A \in \mathcal{A}$.
(ii) $A_i \in A,\ i = 1, 2, \cdots$ *are disjoint and* $A = \cup_{i=1}^{\infty} A_i \in \mathcal{A}$, *then* $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

σ-finiteness is defined as in the σ-field case.

We can define similar functions in semi-algebra. Let $\mu : \mathcal{S} \to \mathbb{R}^+ \cup \{0\}$ be a function on a semi-algebra $\mathcal{S}$ that satisfies (i) $\mu(\phi) = 0$ and $\mu(S) \geq 0,\ \forall S \in \mathcal{S}$. (ii) $S_i \in \mathcal{S},\ i = 1, \cdots, n \implies \mu(\cup_{i=1}^{n} S_i) = \sum_{i=1}^{n} \mu(S_i)$. (iii) $S_i \in \mathcal{S},\ i = 1, 2, \cdots \implies \mu(\cup_{i=1}^{\infty} S_i) \leq \sum_{i=1}^{\infty} \mu(S_i)$. I will call such functions *semi-measures*[1].

The following theorem states that a semi-measure can be uniquely extended to a measure on algebra. If the extended measure on algebra is σ-finite, it can be further extended to a measure on σ-field.

Theorem 1 (Caratheodory's extension)**.** $\mathcal{S}$*: a semi-algebra with* $\phi \in \mathcal{S}$.
$\mu : \mathcal{S} \to \mathbb{R}^+ \cup \{0\}$ *is a semi-measure.*
$\Rightarrow \exists!$ *a positive measure* $\overline{\mu}$ *in* $\overline{\mathcal{S}}$ *that is an extension of* μ.
In addition, if $\overline{\mu}$ *is* σ*-finite,* $\exists!$ *a measure* ν *on* $\sigma(\mathcal{S})$ *that is an extension of* $\overline{\mu}$.

Our major interest is in probability measure on $\mathbb{R}$. In undergraduate statistics, we learned that (cumulative) distribution functions uniquely determine probability distributions while densities cannot. Caratheodory's extension theorem leads us the that conclusion.

We say a real-valued function on $\mathbb{R}$ is a *Stieltjes measure function* if it is non-decreasing and right-continuous.

Theorem 2. F *is a Stieltjes measure function.* $\Rightarrow \exists!$ *a measure* μ *on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *such that* $\mu(a, b] = F(b) - F(a)$.

Proof. Let $\mathcal{S} = \{0\} \cup \{(a, b] :\ -\infty \leq a < b \leq \infty\}$ and $\nu(\phi) = 0$, $\nu(a, b] = F(b) - F(a)$. Then $\mathcal{S}$ is a semi-algebra and ν is a semi-measure. Let $S_n = (-n, n] \in \overline{\mathcal{S}}$, then it is easy to show σ-finiteness. By Caratheodory's theorem, there is a unique extension of ν. □

Since a distribution function F is a special case of Stieltjes measure function, it follows directly from the theorem that F uniquely determines a probability measure[2].

On $\mathbb{R}^d,\ d > 1$, we define functions similar to Stieltjes measure function: $F : \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$ such that F is non-decreasing, right-continuous and $\Delta_A F \geq 0$ for all $A = (a_1, b_1] \times \cdots \times (a_d, b_d]$, where $\Delta_A F = \sum_{v \in V} \text{sgn}(v) F(v)$, $V = \{a_1, b_1\} \times \cdots \times \{a_d, b_d\}$. Similar to the above, such F can be uniquely extended to a measure μ such that $\mu(A) = \Delta_A F$ for all finite rectangle A.

[1] I named this "semi-measure" just for the convenience. This might be different to the actual definition of semi-measure.

[2] Why does this mean that F uniquely determines a probability distribution? Because probability distribution is defined as a probability measure generated by a special kind of functions. This will be discussed in subsection 1.2.

1.1.4 Dynkin's π-λ theorem

I will end this subsection by stating the theorem that will be used throughout the course.

Definition 8 (π-system)**.** *A collection of sets $\mathcal{P}$ is a π-system if $A, B \in \mathcal{P}$, then $A \cap B \in \mathcal{P}$.*

Definition 9 (λ-system)**.** *A collection of sets $\mathcal{L}$ is a λ-system on Ω if the followings hold.*
(i) $\Omega \in \mathcal{L}$
(ii) $A \in \mathcal{L} \Rightarrow A^c \in \mathcal{L}$
(iii) $A_i \in \mathcal{L}, i = 1, 2, \cdots$, *where A_i's are disjoint.* $\Rightarrow \uplus_{i=1}^{\infty} A_i \in \mathcal{L}$

Theorem 3 (Dynkin's π-λ theorem)**.** *$\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system. If $\mathcal{P} \subset \mathcal{L}$, then $\sigma(\mathcal{P}) \subset \mathcal{L}$.*

The theorem implies in order to show that some property holds in a σ-field, we only need to prove that it holds in a λ-system and the generator π-system of the σ-field is contained in our λ-system. A simple but useful corollary is about equivalent probability measures.

Corollary 1. *μ_1, μ_2 are probability measures on $(\Omega, \mathcal{F})$. $\mathcal{A} \subset \mathcal{F}$ is a π-system such that $\sigma(\mathcal{A}) = \mathcal{F}$. If $\mu_1(A) = \mu_2(A)$, $\forall A \in \mathcal{A}$, then $\mu_1 \overset{A \in \mathcal{F}}{\equiv} \mu_2$.*

Proof. Let $\mathcal{L} = \{B \in \mathcal{F} : \mu_1(B) = \mu_2(B)\}$, then by construction $\mathcal{A} \subset \mathcal{L}$ and it is clear that $\mathcal{L}$ is a λ-system. By $\pi - \lambda$ theorem, $\sigma(\mathcal{A}) = \mathcal{F} \subset \mathcal{L}$ leads to the desired results. □

Other examples can be found [here](https://naturale0.github.io/probability/application-of-pi-lambda-theorem). We will get on it in one at a time.

1.2 Distributions

In this subsection, we define random variables and distribution functions.

1.2.1 Random variables

In measure theory, a function on a measurable space A onto another measurable space B is *measurable* if its inverse image of measurable sets are measurable sets. If A is a probability space and B is a Borel measurable space of $\mathbb{R}$, then we call this function a random variable.

Definition 10 (random variable)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. $X : \Omega \to \mathbb{R}$ is a random variable, if for all $B \in \mathcal{B}$, $X^{-1}(B) \in \mathcal{F}$.*

We also say X is $\mathcal{F}$- measurable or write $X \in \mathcal{F}$. That is, random variables are Borel measurable real-valued functions defined on a probability space.

Example 1. *(1) If $(\Omega, \mathcal{F}, P)$ is a discrete probability space, every function is a random variable.*
(2) An indicator function $\mathbf{1}_A$, $A \in \mathcal{F}$ is a random variable.

In undergraduate statistics, we used random variables as if they were values. Now we can understand true meaning behind these notations:

$$P(-1 \le X \le 1) = P(X^{-1}([-1, 1])) = P(\{\omega \in \Omega : \ X(\omega) \in [-1, 1]\})$$

and such. Random variables are defined as a measurable function so that their inverse images can be measured by the probability measure P.

An important fact is that every random variable *induces a probability measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. (note that it is not on $(\Omega, \mathcal{F})$.)

Definition 11 (induced probability measure)**.** $P_X(A) := P(X \in A),\ A \in \mathbb{R}$ *is a probability measure induced by a random variable* X*.* P_X *is called the distribution of* X*.*

It is not difficult to show that such P_X is a probability measure.

1.2.2 Distribution functions

A distribution function of X is defined in terms of probability.

Definition 12 (distribution function)**.** $F : \mathbb{R} \to \mathbb{R}$ *is a distribution function of* X*.*
$\Leftrightarrow F(x) := P(X \leq x) = P_X(-\infty, x],\ \forall x \in \mathbb{R}$*.*

We already saw in section 1.1 that a distribution function uniquely determines a distribution (or a random variable). The following theorem implies that every function that satisfies some conditions can be regarded as a distribution function of some random variable.

Theorem 4 (1.2.2)**.** *F is a distribution of some random variable X if and only if it is (i) non-decreasing, (ii) right-continuous, (iii)* $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$*.*

Proof. ($\Rightarrow$) is trivial.
($\Leftarrow$) we proof this by construction. Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$, $P = \lambda$ (Lebesgue measure). Let $X(\omega) = \sup\{y :\ F(y) < \omega\}$. Then X is a random variable. To show $P(X \leq x) = P(\{\omega :\ 0 \leq \omega \leq F(x)\}),\ \forall x$, it is equivalent to show that $\{\omega :\ X(\omega) \leq x\} = \{\omega : \omega \leq F(x)\},\ \forall x$.
First given x, $\omega_0 \in \{\omega :\ \omega \leq F(x)\}$, we get $x \notin \{y :\ F(y) < \omega_0\}$. Thus $X(\omega_0) \leq x$ and we get $\{\omega :\ X(\omega) \leq x\} \supset \{\omega : \omega \leq F(x)\}$. Next, suppose $\omega_0 \notin \{\omega : \omega \leq F(x)\}$, then $\omega_0 > F(x)$. Since F is right-continuous, there exists $\epsilon > 0$ such that $F(x) \leq F(x+\epsilon) < \omega_0$. Hence $x < x + \epsilon \leq X(\omega_0)$ and $\{\omega :\ X(\omega) \leq x\} \subset \{\omega : \omega \leq F(x)\}$. Finally we get $F(x) = \lambda(0, F(x)] = P(\{\omega :\ X(\omega) \leq x\}) = P(X \leq x)$. □

1.3 Random Variables

We take a closer look at random variables and random elements.

1.3.1 Random elements (measurable maps)

Random elements are generalizations of random variables. While the textbook focuses on random variables, we introduce them to provide properties that are not only related to random variables, but also to more general functions.

Definition 13 (random elements)**.** $X : (\Omega, \mathcal{F}) \to (S, \Sigma)$ *is a random element if* $X^{-1}(B) \in \mathcal{F},\ \forall B \in \Sigma$*.*

If the codomain is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ we call a random element a random variable. If the codomain is $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ we call it a random vector. If the codomain is a class of functions, we call it a random function.

As we can see in the definition of random variables (measurable functions), functions are closely related to sets. To be specific, properties of a class of functions is closely related to its domain and image. Similar to π-λ theorem, we can check if a function is a random element by just checking inverse images of elements of the generator of σ-field of its codomain.

Theorem 5. *Let $X : \Omega \to S$ be a function and $\mathcal{A}$ be a collection of sets such that $\sigma(\mathcal{A}) = \Sigma$. If $X^{-1}(A) \in \mathcal{F},\ \forall A \in \mathcal{A}$, then X is a random element.*

Proof. Let $\mathcal{S} = \{B :\ X^{-1}(B) \in \mathcal{F}\}$,then $\mathcal{S}$ is a σ-field and $\mathcal{A} \subset \mathcal{S}$. Hence by definition $\sigma(\mathcal{A}) = \Sigma \subset \mathcal{S}$ and the desired result follows. □

In the proof of the theorem, observe that if $\mathcal{A}$ is a σ-field on the codomain S of a random element X, then $\{A :\ X^{-1}(A) \in \mathcal{F}\}$ is also a σ-field on S and $\{X^{-1}(A) :\ A \in \mathcal{A}\}$ is a σ-field on Ω. In fact, the latter is the smallest σ-field that makes X a random element. We call it a σ-field generated by a measurable map X.

1.3.2 Closure properties of random variables

In this subsection, we are interested in operations on random elements/variables that preserves its measurability.

The first two theorems are about a composition of two measurable functions.

Theorem 6 (1.3.4)**.** *$X : (\Omega, \mathcal{F}) \to (S, \Sigma)$ is a random element.*
$f : (S, \Sigma) \to (T, \mathcal{T})$ is a measurable function.
$\Rightarrow f \circ X : (\Omega, \mathcal{F}) \to (T, \mathcal{T})$ is a random element.

Proof. Given $B \in \mathcal{T}$,

$$(f \circ X)^{-1}(B) = X^{-1}(f^{-1}(B)) = X^{-1}(S) \in \mathcal{F}$$

where $S = f^{-1}(B) \in \Sigma$. □

Theorem 7 (1.3.5)**.** *$X_1, \cdots, X_n : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ are random variables.*
$f : \mathbb{R}^d \to \mathbb{R}$ is a Borel measurable function.
$\Rightarrow f(X_1, \cdots, X_n)$ is a random variable.

Proof. Let $Y = (X_1, \cdots, X_n)$, $\mathcal{A} = \{A_1 \times \cdots \times A_n :\ A_i \in \mathcal{B}(\mathbb{R}),\ i = 1, \cdots, n\}$ then use the fact that $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R}^d)$ and theorem 1 to show that Y is a random vector. By theorem 2, $f(Y)$ is a random variable. □

The next two theorems are about minimum/maximum and limiting behaviors of random variables.

Theorem 8 (1.3.6)**.** *Let X_n, $n = 1, 2, \cdots$ be random variables. The followings are random variables.*
(i) $\inf_n X_n$
(ii) $\sup_n X_n$
(iii) $\liminf_n X_n$
(iii) $\limsup_n X_n$

Proof. Let $\mathcal{A} = \{(-\infty, x] :\ x \in \mathbb{R}\}$ so that $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R})$.
(i) Let $Y = \inf_n X_n$. Given $A = (-\infty, x] \in \mathcal{A}$. $Y^{-1}(A) = \{\inf_n X_n \le x\} = \cup_{i=1}^{\infty}\{X_i \le x\} \in \mathcal{F}$.
(iii) $\liminf_n X_n = \sup_n \inf_{m \ge n} X_m$. Use (i), (ii).
(ii), (iv) Use $-X_n$ and (i), (iii). □

By (iii) and (iv), if X_n is a random variable and its limit X_∞ exists, then X_∞ is also a random variable.

Corollary 2.

$$\begin{aligned}\Omega_0 &:= \{\omega : \lim_n X_n(\omega) = X_\infty(\omega)\} \\ &= \{\omega : \limsup_n X_n(\omega) - \liminf_n X_n(\omega) = 0\} \in \mathcal{F}.\end{aligned}$$

i.e. Ω_0 *is a measurable set.*

We used $\limsup_n X_n(\omega) - \liminf_n X_n(\omega) = 0$ instead of $\limsup_n X_n(\omega) = \liminf_n X_n(\omega)$ to cover the case where X_∞ is either $\pm\infty$.

I will finish this section by defining a kind of convergence used in probability theory.

Definition 14 (almost sure convergence)**.** *We say that* X_n *converges almost surely to* X_∞ *or write* $X_n \to X_\infty$ *a.s., if* $P(\{\lim_n X_n = X_\infty\}) = 1$.

Here, a.s. stands for "almost surely". In general measure theory, this term is usually replaced with a.e. or "almost everywhere".

Almost sure convergence implies that in probability space, in most of the times only the subsets with positive measures are important: it might be enough to have the fact that X_n converges a.s. Convergence theorems in Lebesgue integration are the examples. We will cover this in detail later.

1.4 Lebesgue integral

In this section, we define the expectation of a random variable as the Lebesgue integral with respect to the probability measure. First, I will introduce the standard machine in measure theory and use it to define the Lebesgue integral. Next, the definition of expectation will be discussed in terms of the Lebesgue integral.

1.4.1 Standard machine

The standard machine is a proof scheme in measure theory that is especially useful when proving properties of a general measurable function. The proof starts from the simplest form of functions, indicator functions, and repeatedly use the result from simpler functions to prove properties in more general functions. To be specific, the standard machine procedes in the following steps:

1. indicator function 2. simple function 3. non-negative measurable function 4. general measurable function

Some textbooks (including PTE) drops proof for indicator functions and intead add proof for bounded functions in between simple functions and non-negative functions. However in my experience starting with indicator function was much less cumbersome.

We define the Lebesgue integral using the standard machine, starting from the one for indicator functions.

1.4.2 Lebesgue integral

Unlike the Riemann integration which was defined as the limit of sums of areas of rectangles that partitions the domain of a function (Riemann sum), the Lebesgue integral is defined as those that partitions the image of a measurable function. While the "height" of each rectangle was of interest in the Riemann integral, it is the "width" of it that is important when defining the Lebesgue integral.

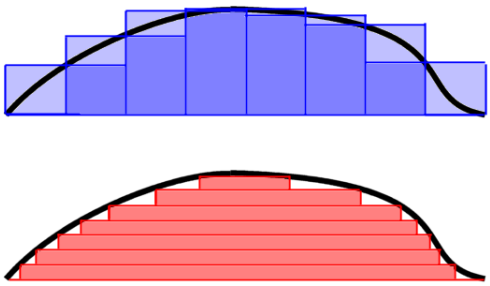

Figure 1: (Blue) Riemann integral. (Red) Lebesgue integral. (source)

We use the measure of inverse image of a partition of a function's codomain as the width. (This is why we named such functions "measurable")

The following definitions and statements are all assuming a measure space $(\Omega, \mathcal{F}, \mu)$ where μ is σ-finite unless otherwise noted.

1. indicator function

Definition 15 (indicator function)**.** $\mathbf{1}_A(x) := \begin{cases} 1 & , \text{ if } x \in A \\ 0 & , \text{ otherwise} \end{cases}$
where A is measurable, is an indicator function.

An indicator function is trivially a measurable function since its inverse image is either ϕ or Ω.

Definition 16 (Lebesgue integral of an indicator function)**.** $\int_E \mathbf{1}_A d\mu := \mu(A \cap E),\ E \in \mathcal{B}(\mathbb{R})$ *is the integral of* $\mathbf{1}_A$ *over* E*.*

2. simple function A simple function is a finite weighted sum of indicator functions.

Definition 17 (simple function)**.** $s(x) := \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i}$*, where* A_i*'s are measurable and* $\alpha_i \in \mathbb{R},\ i = 1, \cdots, n$*.*

Since the sum of measurable functions is also measurable, simple functions are measurable. Naturally, the Lebesgue integral of a simple function is defined in a similar way.

Definition 18 (Lebesgue integral of a simple function)**.** $\int_E s d\mu := \sum_{i=1}^n \alpha_i \mu(A_i \cap E),\ E \in \mathcal{B}(\mathbb{R})$.

3. non-negative measurable function Things get a little interesting for non-negative measurable functions. For a non-negative function, its integral is defined as sup (or inf) of simple functions.

Definition 19 (Lebesgue integral of a non-negative measurable function)**.** *Let* $f : \Omega \to \mathbb{R}^+ \cup \{0\}$ *be measurable.*
Let φ *be a simple function.*
$\int_E f d\mu := \sup_{0 \le \varphi \le f} \int_E \varphi d\mu,\ E \in \mathcal{B}(\mathbb{R})$.

Consider a sequence of simple functions $\{\varphi_n\}_{n\in\mathbb{N}}$ that $\varphi_n(x) := (\lfloor 2^n f(x)\rfloor/2^n) \wedge n$. Then φ_n monotonically increases as n increases and approaches f from below. i.e. $\varphi_n \uparrow f$ as $n \uparrow \infty$. Thus such simple function in the definition exists and it is well-defined.

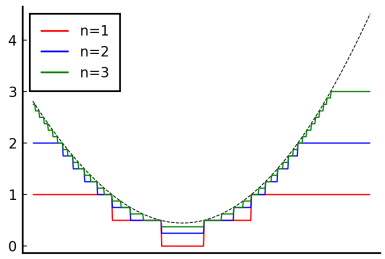

Figure 2: Shape of φ_n with varying n. f: black dashed line.

4. general measurable function Now that we defined the integral of non-negative functions, the remaining part is easy. Let f be a measurable function. Define its positive part $f^+ := f \vee 0$ and the negative part $f^- := -(f \wedge 0)$. Then $f = f^+ - f^-$ and f^+, f^- are non-negative measurable functions. The integral of f on a measurable set E is defined as $\int_E f d\mu := \int_E f^+ d\mu - \int_E f^- d\mu$.

If $E = \Omega$, then we omit Ω for simplicity. That is, $\int f d\mu = \int_\Omega f d\mu$. We call f is (Lebesgue) integrable if $\int |f| d\mu = \int f^+ d\mu + \int f^- d\mu < \infty$ and write $f \in L^1(\mu)$ or just $f \in L^1$ if the measure is clear.

There are several properties of the integral naturally arises from the definition.

Proposition 2 (properties of the integral)**.** *Let f, g be integrable, A, B, E be measurable.*
(i) $f \le g \implies \int_E f d\mu \le \int_E g d\mu$.
(ii) $A \subset B \implies \int_A f d\mu \le \int_B f d\mu$.
(iii) $\int_E cf d\mu = c\int_E f d\mu,\ c \in \mathbb{R}$.
(iv) $f = 0$ *on* $E \implies \int_E f d\mu = 0$.
(v) $\mu(E) = 0 \implies \int_E f d\mu = 0,\ \forall f$.
(vi) $\int_E f d\mu = \int f \mathbf{1}_E d\mu$.

Since it is not difficult to show, I will leave it as an exercise. To prove somewhat straightforward-looking property $\int_E f + g d\mu = \int_E f d\mu + \int_E g d\mu$, we need help of monotone convergence theorem, which will be covered soon.

1.4.3 Expectation

I would like to finish this part by defining expectation of random variables. Recall that a random variable X is a measurable function on a probability space $(\Omega, \mathcal{F}, P)$. The expectation is merely the integral of X on Ω with respect to P. That is, $EX := \int X dP$.

As a side note, the Lebesgue integration can be viewed as a generalization of the Riemann integration. It can be shown that any Riemann integrable functions are Lebesgue integrable. Thus it is still viable to interpret expectation as a weight sum of probabilities for many random variables with Riemann integrable densities.

1.5 Convergence Theorems

In the previous section, we defined the Lebesgue integral and the expectation of random variables and showed basic properties. However the additive property of integrals is yet to be proved. In addition, since our major interest throughout the textbook is convergence of random variables and its rate, we need our toolbox for it.

In this section we assume a measure space $(\Omega, \mathcal{F}, \mu)$ with finite measure μ. It is enough to state and prove theorems only on the finite measure case since our interest is in the probability space.

1.5.1 Convergence theorems

Earlier, I mentioned that to prove the additive property of the Lebesgue integrals, we need the monotone convergence theorem. Monotone convergence theorem (or MCT) is for a sequence of non-negative functions that increases monotonically to the limiting function. In fact, there are other convergence theorems - bounded and dominated one - in addition to the MCT. Fatou's lemma, a corollary of the MCT, is another useful tool for our journey through probability theory.

Monotone convergence theorem (MCT)

Theorem 9 (Monotone convergence)**.** *$\{f_n\}_{n\in\mathbb{N}} : \Omega \to [0,\infty]$: a sequence of measurable functions.*
$f : \Omega \to [0,\infty]$: the limiting function of $\{f_n\}$.
$f_n \uparrow f$ a.s. $\implies$ (i) f is measurable, (ii) $\int f_n d\mu \uparrow \int f d\mu$.

To prove this, we need two lemmas.

Lemma 1 (1.25(a) from RCA)**.** *$s : \Omega \to [0,\infty]$: a simple function.*
$\varphi(E) = \int_E s d\mu$, $\forall E \in \mathcal{F} \implies \varphi$ is a measure on $(\Omega, \mathcal{F})$.

Lemma 2. *$s : \Omega \to [0,\infty]$: a simple function.*
$E_n \uparrow E$: E_n, E are measurable.
$\implies \lim_n \int_{E_n} s d\mu = \int_E s d\mu$.

The first lemma is easy to check. The second one is trivial by continuity of measure ϕ defined as in the first one. The latter one will be used in the proof of the MCT.

MCT. Let $\alpha_n = \int f_n d\mu$ so that $\alpha_1 \le \alpha_2 \le \cdots \le \int f d\mu$ and $\lim_n \alpha_n = \alpha$ for some α. We need to show that $\alpha = \int f d\mu$.
(1) It is trivial that $\alpha \le \int f d\mu$.
(2) Now, let s be a simple function such that $0 \le s \le f$ and a constant $0 < c < 1$. Define $E_n = \{\omega \in \Omega : \ f_n(\omega) \ge c \cdot s(\omega)\}$. Then since E_n increases monotonically as f_n increases monotonically and $\cup_{n=1}^{\infty} E_n = \Omega$, $E_n \uparrow \Omega$. By the lemma 3, this implies $\int_{E_n} s d\mu \uparrow \int s d\mu$.

$$\begin{aligned}
&\int f_n d\mu \ge \int_{E_n} f_n d\mu \ge \int_{E_n} c \cdot s d\mu \\
&\overset{\lim_n}{\implies} \alpha \ge c \int s d\mu \\
&\overset{c\to 1}{\implies} \alpha \ge \int s d\mu \\
&\overset{\sup_{0\le s\le f}}{\implies} \alpha \ge \int f d\mu
\end{aligned}$$

By (1) and (2), the desired result follows. □

The MCT allows us to prove the yet to be shown property.

Proposition 3 (additivity of the integrals)**.** *$f, g : \Omega \to [0, \infty]$: measurable.*
$\implies \int f + g d\mu = \int f d\mu + \int g d\mu$.

Proof. Let $s_n = (\lfloor 2^n f \rfloor / 2^n) \wedge n$ and $t_n = (\lfloor 2^n g \rfloor / 2^n) \wedge n$. Then $s_n + t_n \uparrow f + g$ and $\int s_n + t_n d\mu = \int s_n d\mu + \int t_n d\mu$. By the MCT, the result follows. □

Fatou's lemma Fatou's lemma, another important convergence theorem can be directly derived by the MCT.

Theorem 10 (Fatou)**.** *$\{f_n\} : \Omega \to [0, \infty]$: a sequence of measurable functions.*
$\implies \int \liminf_n f_n d\mu \leq \liminf_n \int f_n d\mu$.

Proof. Let $g_n = \inf_{k \geq n} f_k$ so that $g_n \uparrow \liminf_n f_n$. By MCT, $\lim_n \int g_n d\mu = \int \liminf_n f_n d\mu$. Since g_n is monotone and $g_n \leq f_n$, we get $\int g_n d\mu \leq \int f_n d\mu,\ \forall n$ thus $\lim_n \int g_n d\mu \leq \liminf_n \int f_n d\mu$. □

It is worth noting that Fatou's lemma does not require convergence. Thus it can be applied to any sequence of non-negative measurable functions. In many cases however, the lemma is used in the form of $\int X dP \leq \liminf_n \int X_n dP$ where $X_n \to X$ a.s..

Dominated convergence theorem (DCT) While the MCT is very useful, it can only be applied to a sequence of functions that monotonically converges. Lebesgue's dominated convergence theorem (DCT) provides a tool for not only monotonically convergent, but general convergent functions that are uniformly dominated by some integrable function.

Theorem 11 (Dominated convergence)**.** *$\{f_n\} : \Omega \to \mathbb{R}$: a sequence of measurable functions.*
$f : \Omega \to \mathbb{R}$: a measurable functions.
$f_n \to f$ a.s. and $|f_n| \leq g$, g is integrable.

$$\begin{aligned} \implies & \textit{(i)}\ f \in L^1(\mu). \\ & \textit{(ii)}\ \int |f_n - f| d\mu \to 0. \\ & \textit{(iii)}\ \int f_n d\mu \to \int f d\mu. \end{aligned}$$

The usefulness of the DCT is that it not only shows convergence of the integral, but also integrability of the limiting function and L^1 convergence[3] to it.

DCT. (i) Trivial since $|f| \leq g$.
(ii) $|f_n - f|$ are integrable since $\leq 2g$. $0 \leq 2g - |f_n - f| \leq 2g$. By fatou's lemma,

$$\int 2g d\mu \leq \int 2g d\mu - \limsup_n \int |f_n - f| d\mu \limsup_n \int |f_n - f| d\mu \leq 0 \therefore \lim_n \int |f_n - f| d\mu = 0$$

(iii) $|\int f_n d\mu - \int f d\mu| \leq \int |f_n - f| d\mu \to 0$. □

[3] $f_n \to f$ in $L^1(\mu)$ is equivalent to $\int |f_n - f| d\mu \to 0$ and $f \in L^1(\mu)$. It will be covered in the next section.

Bounded convergence theorem (BCT) A special case of the DCT is where the sequence $\{f_n\}$ is uniformly bounded almost surely (i.e. $Y = c \in \mathbb{R}$). In this case we call it the bounded convergence theorem.

1.5.2 Inequalities

Along with convergence theorems, these integral inequalities will be used intensely throughout the probability theory. In the following theorems, assume f, g are measurable. For $p \geq 1$, if $|f|^p$ is integrable, $\|f\|_p := (\int |f|^p d\mu)^{\frac{1}{p}}$ is the $L^p(\mu)$-norm of f.

Theorem 12 (Jensen's inequality)**.** $\varphi : \mathbb{R} \to \mathbb{R}$*: convex function.*
$\int |f| d\mu < \infty$, $\int |\varphi(f)| d\mu < \infty$.
$\implies \varphi(\int f d\mu) \leq \int \varphi(f) d\mu$.

Proof. Let $t = \int f d\mu$, then there exists $\beta = \sup_{s<t} \frac{\varphi(t)-\varphi(s)}{t-s} \in \mathbb{R}$ such that $\varphi(x) \geq \beta(x-t) + \varphi(t)$. Take integral to both sides and we get the result. □

Theorem 13 (Hölder's inequality)**.** $p, q \in (1, \infty)$ *such that* $\frac{1}{p} + \frac{1}{q} = 1$.

$$\implies \int |fg| d\mu \leq (\int |f|^p d\mu)^{\frac{1}{p}} (\int |g|^q d\mu)^{\frac{1}{q}}.$$
$$\textit{i.e. } \|fg\|_1 \leq \|f\|_p \|g\|_q.$$

Proof. Let $A = \|f\|_p$ and $B = \|g\|_q$, $F = f/A$ and $G = g/B$. Then $\int F^p d\mu = \int G^q d\mu = 1$.
Our claim is that $x, y \geq 0 \Rightarrow xy \leq \frac{x^p}{p} + \frac{y^q}{q}$. Let $h(x) = xy - x^p/p$, then $h'(x) = y - x^{p-1} = y - x^{p/q}$ and h achieves maximum at $x = y^{q/p}$.
By the claim, $\int FG d\mu \leq \int F^p/p + G^q/q d\mu = 1$ and the desired result follows. □

A special case of Hölder's inequality is the Cauchy-Schwarz inequality: $\|fg\|_1 \leq \|f\|_2 \|g\|_2$.

Theorem 14 (Minkowski's inequality)**.** $p \geq 1$, $\int |f|^p d\mu < \infty$, $\int |g|^p d\mu < \infty$
$\implies \|f + g\|_p \leq \|f\|_p + \|g\|_p$.

Proof.

$$\int (f+g)^p d\mu = \int f(f+g)^{p-1} d\mu + \int g(f+g)^{p-1} d\mu$$

By Hölder's inequality,

$$\begin{aligned}\int f(f+g)^{p-1} d\mu &\leq \left(\int f^p d\mu\right)^{1/p} \left(\int (f+g)^{(p-1)q} d\mu\right)^{1/q} \\ &= \left(\int f^p d\mu\right)^{1/p} \left(\int (f+g)^p d\mu\right)^{1/q}\end{aligned}$$

Similarly, $\int g(f+g)^{p-1} d\mu \leq \left(\int g^p d\mu\right)^{1/p} \left(\int (f+g)^p d\mu\right)^{1/q}$. Thus

$$\int (f+g)^p d\mu \le \left(\int (f^p + g^p) d\mu\right)^{1/p} \left(\int (f+g)^p d\mu\right)^{1/q}$$
$$\therefore \|f+q\|_p \le \|f\|_p + \|g\|_p$$

□

Finally, we state Markov-Chebyshev inequality. Assume a probability space $(\Omega, \mathcal{F}, P)$ and a random variable X on it.

Theorem 15 (Markov-Chebyshev's inequality)**.** $\varphi : \mathbb{R} \to \mathbb{R}$, $\varphi \ge 0$.
$A \in \mathcal{B}(\mathbb{R})$, $i_A := \inf\{\varphi(y) : \ y \in A\}$.
$\implies i_A \cdot P(X \in A) \le \int_A \varphi(X) dP \le E\varphi(X)$.

Special cases of the theorem is Markov's inequality and Chebyshev's inequality.

Corollary 3 (Markov's inequality)**.** $X \ge 0$ *a.s.*, $a > 0 \implies P(X \ge a) \le EX/a$.

Corollary 4 (Chebyshev's inequality)**.** $a > 0 \implies P(X \ge a) \le EX^2/a^2$.

1.5.3 Concluding remarks

Since the expectation EX is defined as a mere integral, all of the above theorems can be applied. For instance, if $X_n \to X$ a.s. and $|X_n| \le Y$ for some Y such that $E|Y| < \infty$, then by DCT $EX_n \to EX$ as $n \to \infty$.

1.6 L^p Space

In the previous section, we defined Lebesgue integrability and write $f \in L^1$ for such function f. We also defined the L^p-norm. L in these notations stands for "Lebesgue" and L^p for $p \ge 1$ becomes a space of functions that is integrable in the order of p. We name it L^p space.

1.6.1 L^p space on probability space

Definition 20 (L^p space)**.** $L^p = L^p(\Omega, \mathcal{F}, P) := \{X : E|X|^p < \infty\}$.
$\|X\|_p := (E|X|^p)^{1/p}$, $X \in L^p$ *is the* L^p*-norm.*

L^p-norms are monotone. That is, $\|X\|_p \le \|X\|_q$, $1 \le p \le q$. Thus if $X \in L^q$, then $X \in L^p$ for $1 \le p < q$.

1.6.2 Convergence in L^p

Definition 21 (L^p convergence)**.** *Let* X_n, X *be random variables and* $X_n \in L^p$, $\forall n$. *If* $X \in L^p$ *and* $\|X_n - X\|_p \to 0$*, then we say* X_n *converges to* X *in* L^p *and write* $X_n \xrightarrow{L^p} X$.

Or we could just say $X_n \to X$ in L^p. It is known that for $p \ge 1$, L^p are closed (Banach) spaces. If $p = 2$, it becomes a Hilbert space since we can define the inner product as $\langle X, Y \rangle = EXY$.

1.7 Product Space

Let X, Y be random variables on $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ respectively. In order to well-define sets such as $\{X + Y \leq 0\}$, we should consider a random vector (X, Y) on a product space $\mathbb{R} \times \mathbb{R}$ since $+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined this way. In addition, to measure the probability of such sets, we also need to define another product probability space $(\Omega, \mathcal{F}, P)$ where $\Omega = \Omega_1 \times \Omega_2$. Main question to answer in this subsection is: how can we define a proper σ-field $\mathcal{F}$ and a product measure P on such Ω?

1.7.1 Product measure space

Definition 22 (product measure space)**.** *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be measure spaces with σ-finite measures μ_1, μ_2.*
(i) $\Omega = \Omega_1 \times \Omega_2 := \{(\omega_1, \omega_2) : \ \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$ is the product space.
(ii) $\mathcal{S} := \{E_1 \times E_2 : \ E_1 \in \mathcal{F}_1, E_2 \in \mathcal{F}_2\}$. $S \in \mathcal{S}$ is called a (measurable) rectangle.
(iii) $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 := \sigma(\mathcal{S})$ is the σ-field of the product space.

Note that instead of directly product the two σ-fields, we define $\mathcal{F}$ with measurable rectangles. In fact, direct product of the σ-field does not yield a σ-field.

We now define a product measure on $(\Omega, \mathcal{F})$. Uniqueness of such measure is as important as existence of it. Even if we define a proper measure, it would be useless unless it is uniquely determined. Thankfully, Caratheodory's theorem ensures uniqueness.

Theorem 16 (uniqueness of the product measure)**.** *There uniquely exists a measure μ on $(\Omega, \mathcal{F})$ such that $\mu(E_1 \times E_2) = \mu_1(E_1)\mu_2(E_2), \ \forall E_1 \in \mathcal{F}_1, E_2 \in \mathcal{F}$.*

Proof. $\mathcal{S}$ is a semi-algebra. Show (i) $\mu(A \times B) = \sum_{i=1}^{\infty} \mu(A_i \times B_i) = \sum_{i=1}^{\infty} \mu_1(A_i)\mu_2(B_i)$ for $A \times B = \cup_{i=1}^{\infty}(A_i \times B_i)$ and (ii) σ-finiteness of μ. Use Caratheodory's theorem to get the result. □

1.7.2 Fubini's theorem

Measure of a measurable rectangle is comprehensive. However it is cumbersome to compute and intuitively understand measurable sets that are not rectangles. We *could* understand it as the limit of measure of rectangles, but it is still a problem to calculate it - we do not know anything more than the existence of the measure. Fubini's theorem is here for the rescue.

Theorem 17 (Fubini)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be the product space of $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$. If $f \geq 0$ a.s. ¡b¿or¡/b¿ $f \in L^1(\mu)$, then*

$$\begin{aligned}\int_{\Omega} f d\mu &= \int_{\Omega_2} \int_{\Omega_1} f(x, y) d\mu_1(x) d\mu_2(y) \\ &= \int_{\Omega_1} \int_{\Omega_2} f(x, y) d\mu_2(y) d\mu_1(x).\end{aligned}$$

Fubini's theorem implies that for almost surely non-negative or integrable functions, we can calculate the integral with respect to the product measure as a double integral, and its order is negligible.

Finishing the section, I would like to remark some points.

1. We can even define n-dimensional product spaces to extend the notion even further. 2. If (X, Y) is a random element on a metric space $S \times S$, then X, Y are random elements on S. However

the converse is not always the truth. In fact, it is known that the converse holds if S is *separable*. Since $(\mathbb{R}, d)$ is a separable space, we can be sure that (X, Y) is a random vector if and only if X, Y are random variables.

2 Laws of Large Numbers

2.1 Independence

First chapter was about the essential of measure theory. We especially focused on important result for finite measure or at least σ-finite measures. We defined a probability space as a measure space and a random variable as a measurable function in it.

Following chapters will cover two fundamental theory in convergence of random variables: the strong law of large numbers (Chapter 2) and the central limit theorem (Chapter 3). We start by assuming nice but in many real cases inadequate condition - mutual independence of random variables - and modify the result to achieve the stronger one. As a starting point, this subsection covers the notion of independence. The rest of the chapter is about law of large numbers.

2.1.1 Independence of random variables

As I pointed out earlier, it is natural to define a property of a function as a property of the related set (its domain). We do the same here: we define independence of σ-fields first.

Definition 23 (independence 1)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{F}_1, \cdots, \mathcal{F}_n \subset \mathcal{F}$ be sub σ-fields, $E_1, \cdots, E_n$, $E_i \in \mathcal{F}_i$ be the events.*
(i) $E_1, \cdots, E_n$ are independent if $P(\bigcap_{i=1}^{n} E_i) = \prod_{i=1}^{n} P(E_i)$.
(ii) $\mathcal{F}_1, \cdots, \mathcal{F}_n$ are independent if $P(\bigcap_{i=1}^{n} E_i) = \prod_{i=1}^{n} P(E_i)$ for all $E_i \in \mathcal{F}_i$.

In fact to be extra specific we say it is P-mutually independent if the above condition is met. It gives us extra information about regarding probability measure (P) and that it is mutual. If we just write "independence" it means mutual independence. We drop P if the measure is clear without confusion.

Independence of random variables are defined as independence of generated σ-fields.

Definition 24 (independence 2)**.** *Let $X_1, \cdots, X_n$ be random variables in $(\Omega, \mathcal{F}, P)$. $X_1, \cdots, X_n$ are independent if $\sigma(X_i)$, $i = 1, \cdots, n$ are independent.*

We sometimes write $X \perp Y$ for independence between X and Y.

It is not necessary to check all possible products of the events just to check the independence of σ-fields. We can use π-λ theorem.

Theorem 18 (2.1.7)**.** *$\mathcal{A}, \mathcal{B}$ are π-systems and are subsets of $\mathcal{F}$. If $P(A \cap B) = P(A)P(B)$ for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$, then $\sigma(\mathcal{A}), \sigma(\mathcal{B})$ are independent.*

Proof. (1) For a fixed $A \in \mathcal{A}$, let $\mathcal{L}_A = \{B \in \mathcal{F} : P(A \cap B) = P(A)P(B)\}$ be a λ-system containing $\mathcal{B}$. By π-λ theorem, $\sigma(\mathcal{B}) \subset \mathcal{L}_A$.
(2) Now for fixed $B \in \sigma(\mathcal{B})$ let $\mathcal{L}_B = \{A \in \mathcal{F} : \mu(A \cap B) = \mu(A)\mu(B)\}$. Then similar to the above, $\mathcal{L}_B$ is a λ-system that contains $\mathcal{A}$ and $\sigma(\mathcal{A}) \subset \mathcal{L}_B$ follows. □

It is clear that $\mathcal{P} = \{X^{-1}(-\infty, x] : \ x \in (-\infty, \infty]\}$ is a π-system and $\sigma(\mathcal{P}) = \sigma(X)$, so the corollary directly follows.

Corollary 5 (2.1.8)**.** *$X_1, \cdots, X_n$ are independent if and only if $P(X_1 \leq x_1, \cdots, X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i)$ for all $x_i \in (-\infty, \infty]$.*

2.1.2 Existence of a sequence of independent random variables

In the following chapters, we will construct a sequence of independent random variables to state and prove limiting laws. It is important to mention that such sequence exists.

For a finite number $n \in \mathbb{N}$, we can construct n independent random variables using product space. Given distribution functions F_i, $i = 1, \cdots, n$, let X_i with $P(X_i \leq x) = F_i(x)$ and $X_i(\omega_1, \cdots, \omega_n) = \omega_i$. (i.e. X_i is the coordinate-wise projection.) Let $(\Omega, \mathcal{F}, P) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}), P)$ where $P((a_1, b_1] \times \cdots \times (a_n, b_n]) = \prod_{i=1}^{n} (F_i(b_i) - F_i(a_i))$ then X_i's are independent.

Now we need infinite number of independent random variables. Consider $\mathbb{R}^\infty := \{(x_1, x_2, \cdots) : \ x_i \in \mathbb{R}\}$, an infinite-dimensional product space of $\mathbb{R}$ and corresponding product σ-field $\mathcal{R}^\infty$. Kolmogorov's extension theorem states that we can construct a unique probability measure on this space.

Theorem 19 (Kolmogorov's extension)**.** *Given probability measures P_n on $(\mathbb{R}^n, \mathcal{R}^n)$ that satisfies*

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

there uniquely exists a probability measure P on $(\mathbb{R}^\infty, \mathcal{R}^\infty)$ such that

$$P((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R}^{\infty - n}) = P_n((a_1, b_1] \times \cdots \times (a_n, b_n]).$$

Furthermore, if a measurable space $(S, \mathcal{S})$ is *nice*, that is, there is a one-to-one map $\varphi : S \to \mathbb{R}$ so that φ, φ^{-1} are both measurable, then we can also construct a sequence of random elements $\{X_n\}_{n \in \mathbb{N}} : \Omega \to S$.

2.2 Weak Laws of Large Numbers

We say a random variable X_n converges in probability (or P-converges) to another random variable X and write $X_n \xrightarrow{P} X$ if $\lim_n P(|X_n - X| > \epsilon) = 0$ for all $\epsilon > 0$. We can also define convergence in probability to a constant by letting $X = c \in \mathbb{R}$. It is easy yet useful to know that $X_n \xrightarrow{L^p} X$ with $p > 0$ implies $X_n \xrightarrow{P} X$ by [Markov-Chebyshev inequality](https://naturale0.github.io/probability/PTE-1.5-Convergence-theorems-and-inequalities#inequalities) with $\varphi(x) = |x|^p$.[4]

Week law of large numbers is about convergence of the sample mean of independent random variables to their expectation in probability. In this section, we cover variants of the week law of large numbers (WLLN), from the simplest L^2-WLLN to the most familiar form which only requires L^1 condition.

2.2.1 L^2 weak law

Definition 25 (uncorrelatedness)**.** *X_1, X_2 are uncorrelated if $EX_1X_2 = EX_1EX_2$.*

[4] Relationship between L^p convergence, convergence in probability, almost sure convergence will be covered later in detail.

If X_n's are independent then it is uncorrelated but the converse does not hold. It is also clear that if X_i's are uncorrelated then $\mathrm{Var}(X_1+\cdots+X_n) = \sum_{i=1}^n \mathrm{Var}(X_i)$. Using Chebyshev's inequality we can easier get our first WLLN for uncorrelated random variables.

Theorem 20 ($\mathbf{L^2}$ weak law)**.** *Let X_i's be uncorrelated with $EX_i = \mu$ and $Var(X_i) \le C < \infty$ for all i. Let $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ in probability and in L^2.*

Proof. By chebyshev's inequality, given $\epsilon > 0$,

$$P(|S_n/n - \mu| > \epsilon) \le \mathrm{Var}(S_n/n)/\epsilon^2 \le C/\epsilon^2 n \to 0$$

and by definition of variance,

$$E|S_n/n - \mu|^2 = \mathrm{Var}(S_n/n) \le C/n \to 0.$$

□

An example of L^2 weak law is in [high-dimensional problem](https://naturale0.github.io/probability/high-dimensional-box).

Theorem 1 requires strong conditions that X_i's should have the same expectation and their variance should be uniformly bounded. Theorem 2 relieves these so that the result could be more useful.

Theorem 21 (2.2.6)**.** *Let $\mu_n = ES_n$, $\sigma_n^2 = Var(S_n)$.*
$\frac{\sigma_n^2}{b_n^2} \to_n 0 \implies \frac{S_n - \mu_n}{b_n} \xrightarrow{P} 0$.

Proof. $\sigma_n^2/b_n^2 = E\left(\frac{S_n-\mu_n}{b_n}\right)^2 \to 0.$ □

This can be used in any kind of sequence of random variable $(S_n)_{n\in\mathbb{N}}$. (Even the uncorrelatedness is not required!)

2.2.2 Weak law for triangular arrays

Many topics in probability theory concerns triangular arrays $(X\_nk)\_1 \le k \le n$ such that X_{nk}, $k = 1, \cdots, n$ are independent (row-wise independent). We are interested in limiting behaviors of row-wise sum $S_n := X_{n1} + \cdots + X_{nn}$.

To deal with the case where the finite second moment condition is not ensured, we introduce truncated random variables.

Definition 26 (truncation)**.** $\bar{X} := X\mathbf{1}_{(|X|\le M)} = \begin{cases} X & , \text{ if } |X| \le M \\ 0 & , \text{ otherwise} \end{cases}$

By ignoring the tail part of X, truncated random variable $\bar{X}$ always has finite moments.

Theorem 22 (Weak law for triangular arrays)**.** *For each n, X_{nk}, $k = 1 \cdots, n$ are independent.*
Let $b_n > 0$, $b_n \to_n \infty$.
Define $\bar{X}_{nk} := X_{nk}\mathbf{1}_{(|X_{nk}|\le b_n)}$.
(i) $\sum_{k=1}^n P(|X_{nk}| > b_n) \to_n 0$.
(ii) $\frac{1}{b_n^2}\sum_{k=1}^n E\bar{X}_{nk}^2 \to_n 0$.
$\implies \frac{S_n - a_n}{b_n} \xrightarrow{P} 0$, where $a_n = \sum_{k=1}^n E\bar{X}_{nk}$.

Proof. Let $\bar{S}_n = \bar{X}_{n1} + \cdots + \bar{X}_{nn}$ so that $a_n = E\bar{S}_n$. Then for an arbitrary $\epsilon > 0$, $P(|\frac{S_n - a_n}{b_n}| > \epsilon) \leq \underbrace{P(S_n \neq \bar{S}_n)}_{\text{(i)}} + \underbrace{P(|\frac{\bar{S}_n - a_n}{b_n}| > \epsilon)}_{\text{(ii)}}$.

For the first part,

$$\begin{aligned}\text{(i)} &\leq P(\bigcup_{k=1}^{n}\{X_{nk} \neq \bar{X}_{nk}\}) \leq \sum_{k=1}^{n} P(X_{nk} \neq \bar{X}_{nk}) \\ &= \sum_{k=1}^{n} P(|X_{nk}| > b_n) \to_n 0\end{aligned}$$

and for the latter,

$$\begin{aligned}\text{(ii)} &\leq E(\frac{\bar{S}_n - a_n}{b_n})^2/\epsilon^2 \leq \mathrm{Var}(\bar{S}_n)/\epsilon^2 b_n^2 \\ &= \sum_{k=1}^{n} \mathrm{Var}(\bar{S}_n)/\epsilon^2 b_n^2 \\ &= \sum_{k=1}^{n} E\bar{X}_{nk}^2/\epsilon^2 b_n^2 \to_n 0\end{aligned}$$

$\therefore P(|\frac{S_n - a_n}{b_n}| > \epsilon) \to_n 0$ for all $\epsilon > 0$. □

2.2.3 Weak law of large numbers

Result for triangular arrays paves the way for general weak law of large numbers. Before we get to it, let's take a look at a highly useful lemma.

Lemma 3. *Let $Y \geq 0$ be a random variable and $p > 0$.*
$\implies EY^p = \int_0^\infty py^{p-1}P(Y > y)dy$.

Proof.

$$\begin{aligned}\int_0^\infty py^{p-1}P(Y > y)dy &= \int_0^\infty \int_{(Y>y)} py^{p-1}dPdy \\ &= \int_\Omega \int_0^Y py^{p-1}dydP \\ &= \int_\Omega Y^p dP = EY^p.\end{aligned}$$

The second equality comes from the Fubini's theorem. □

Theorem 23 (2.2.12)**.** *$X_1, X_2, \cdots$ are i.i.d. $S_n = \sum_{i=1}^{n} X_i$.*
$xP(|X_i| > x) \to 0$ as $x \to \infty$.
$\implies S_n/n - \mu_n \xrightarrow{P} 0$ where $\mu_n = EX_1\mathbf{1}_{(|X_1| \leq n)}$.

Proof. Use the weak law for triangular arrays. Let $X_{nk} = X_k$ and $b_n = n$.
(i) $\sum_{i=1}^{n} P(|X_i| > n) = nP(|X_1| > n) \to_n 0$.
(ii) Let $\bar{X}_k = X_k \mathbf{1}_{(|X_k| \le n)}$. Then $\frac{1}{n^2} \sum_{k=1}^{n} E\bar{X}_k^2 = \frac{1}{n} E\bar{X}_1^2$. By the lemma, $E\bar{X}_1^2 = \int_0^n 2xP(X_1 > x)dx$ so $\frac{1}{n} E\bar{X}_1^2 \to 0$ as $x \to \infty$.
By (i) and (ii) and theorem 3, $S_n/n - \mu_n \xrightarrow{P} 0$. □

Use this to get the weak law of large numbers.

Theorem 24 (weak law of large numbers)**.** *$X_1, X_2, \cdots$ are i.i.d. with $E|X_i| < \infty$.*
Let $S_n = X_1 + \cdots X_n$, $\mu = EX_1$.
$\implies S_n/n \xrightarrow{P} \mu$.

Proof. (i) $xP(|X_1| > x) = xE\mathbf{1}_{(|X_1|>x)} \le E|X_1|\mathbf{1}_{(|X_1|>x)}$. Since $|X_1|\mathbf{1}_{(|X_1|>x)} \le |X_1|$ and $E|X_1| < \infty$ by DCT $E|X_1|\mathbf{1}_{(|X_1|>x)} \to 0$ as $x \to \infty$.
(ii) Let $\mu_n = E|X_1|\mathbf{1}_{(|X_1|\le n)}$. Since $|X_1|\mathbf{1}_{(|X_1|\le n)} \le |X_1|$ and $E|X_1| < \infty$ by DCT $E|X_1|\mathbf{1}_{(|X_1|\le n)} \to_n \mu$.
By theorem 5, $S_n/n - \mu_n \xrightarrow{P} 0$. Since $\mu_n \xrightarrow{P} \mu$, the desired result follows. □

Note that we do not need finite second moment condition.

While I will leave other examples as further readings, I would like to discuss the example of coupon collector's problem (example 2.2.7).

Example 2 (coupon collector's problem)**.** *$X_1, X_2, \cdots \overset{iid}{\sim} \mathcal{U}\{1, 2, \cdots, n\}$. Define a ¡i¿stopping time¡/i¿ $\tau_k^n = \inf\{m : \ |\{X_1, \cdots, X_m\}| = k\}$ and let $T_n = \tau_n^n$. Then $\frac{T_n}{n \log n} \xrightarrow{P} 1$.*

Here, $\mathcal{U}\{1, 2, \cdots, n\}$ is a discrete uniform distribution from one to n. In this example it can be interpreted that n is the total number of different coupons and τ_k^n represents the number of trials until k different kinds of coupons are collected, T_n is the number of trials until all of the coupons are collected. This example says that asymptotically $n \log n$ trials are required for all n coupons to be collected. Likewise, asymptotic rate of a stochastic process can be calculated using the weak laws.

example 1. Let $X_{nk} = \tau_k^n - \tau_{k-1}^n$, $k = 1, \cdots, n$. Then $X_{nk} \sim \text{Geo}(p)$ where $p = 1 - \frac{k-1}{n}$. The memoryless property of Benoulli process implies $X_{nk} \perp X_{nj}$, $j = 1, \cdots, k-1$ and $T_n = X_{n1} + \cdots + X_{nn}$.
Let $b_n = n \log n$ then it satisfies the condition of theorem 2. □

2.3 Borel-Cantelli Lemmas

In this section I would like to cover the Borel-Cantelli lemmas, or B-C lemmas for short. Borel-Cantelli lemmas are quintessential tools for analysis of tail events and deriving almost sure convergence from P-convergence.

2.3.1 Limit of a sequence of sets

Borel-Cantelli lemmas are statements about probability of "limit" of sets. To state and prove the lemma, we first define lim sup and lim inf of a sequence of sets.

Definition 27. *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of sets.*
$\limsup_n A_n = \bigcap_n \bigcup_{k\geq n} A_k = \{A_n \text{ i.o.}\}$.
$\liminf_n A_n = \bigcup_n \bigcap_{k\geq n} A_k = \{A_n \text{ eventually}\}$.

i.o. stands for "infinitely often". We named these $\limsup, \liminf$ since there is a connection with those of functions. It is not difficult to show that $\limsup\_n \mathbf{1}\_A_n = \mathbf{1}\_\limsup_n A_n$ and $\liminf\_n \mathbf{1}\_A_n = \mathbf{1}\_\liminf_n A_n$. As a remark, by de Morgan's law, $\{A_n \text{ i.o.}\}^c = \{A_n^c \text{ eventually}\}$ thus $P(A_n \text{ i.o.}) = 0$ is equivalent to $P(A_n^c \text{ eventually}) = 1$.

2.3.2 Borel-Cantelli lemmas

Theorem 25 (Borel-Cantelli)**.** *(i)* $\sum_{n=1}^{\infty} P(A_n) < \infty \implies P(A_n \text{ i.o.}) = 0$.
(ii) If A_n's are independent, $\sum_{n=1}^{\infty} P(A_n) = \infty \implies P(A_n \text{ i.o.}) = 1$.

Proof. (i) $\sum_{k=1}^{n} \mathbf{1}_{A_k} \uparrow \sum_{n=1}^{\infty} \mathbf{1}_{A_n}$. By MCT $E\sum_{n=1}^{\infty} \mathbf{1}_{A_n} = \sum_{n=1}^{\infty} P(A_n) < \infty$ which leads to $\sum_{n=1}^{\infty} \mathbf{1}_{A_n} < \infty$ a.s.
(ii) Using the inequality $1 - x \leq e^{-x}$,

$$P(\bigcap_{k\geq m} {A_k}^c) = \prod_{k\geq m} \big(1 - P(A_k)\big)$$
$$\leq \prod_{k\geq m} e^{-P(A_k)} = e^{-\sum_{k\geq m} P(A_k)} = 0, \ \forall m > 0$$

$\therefore P(\bigcup_{k\geq m} A_k) = 1$ and $P(\limsup_n A_n) = P(A_n \text{ i.o.}) = 1$. □

Using Borel-Cantelli lemmas, we can convert the sum of probabilities to probability of tail events.

Tail events and the 0-1 law I mentioned tail events several times but never actually defined it.

Definition 28 (tail event)**.** *Let (X_n) be a sequence of random variables. Let $\mathcal{G}_n = \sigma(X_n, X_{n+1}, \cdots)$. We call $\mathcal{T} = \bigcap_n \mathcal{G}_n$ a ¡b¿tail σ-field¡/b¿ of (X_n), $A \in \mathcal{T}$ a ¡b¿tail event¡/b¿.*

Simply put, a tail event is an event that does not depend on finite number of random variables. For instance, $\{\limsup_n S_n/C_n > x\} \in \mathcal{T}$ for $C_n, x \in \mathbb{R}$, $S_n = X_1 + \cdots + X_n$ if $X_n \to_n \infty$ while $\{\limsup_n S_n > 0\} \notin \mathcal{T}$.

Definition 29 (P-trivial event)**.** *An event A is P-trivial if $P(A) = 0$ or 1.*

The next theorem implies that Borel-Cantelli lemmas are converses of each other.

Theorem 26 (Kolmogorov's 0-1 law)**.** *If $X_1, X_2, \cdots$ are independent, $A \in \mathcal{T}$, where $\mathcal{T} := \bigcap_n \sigma(X_n, X_{n+1}, \cdots)$, then A is a P-trivial event.*

Proof. Let $A \in \sigma(X_1, \cdots, X_k)$ and $B \in \sigma(X_{k+1}, \cdots, X_{k+j})$ for some j. Because X_i's are independent, $A \perp B$. Thus $\sigma(X_1, \cdots, X_k) \perp \bigcup_j \sigma(X_{k+1}, \cdots, X_{k+j})$. Since both are π-systems, by Dynkin's π-λ theorem,

$$\sigma(X_1, \cdots, X_k) \perp \sigma(X_{k+1}, \cdots) \tag{1}$$

Now, let $A \in \sigma(X_1, \cdots, X_j)$ for some j and $B \in \mathcal{T} \subset \sigma(X_{j+1}, \cdots)$. By (1) and similar process,

$$\sigma(X_1, \cdots) \perp \mathcal{T} \tag{2}$$

By (2), since $\mathcal{T} \subset \sigma(X_1, \cdots)$, $\mathcal{T}$ is independent of itself.
$\therefore A \in \mathcal{T} \to P(A) = P(A \cap A) = P(A)P(A)$. □

2.3.3 Application of Borel-Cantelli lemmas

Application of the first lemma The first application connects convergence in probability to almost sure convergence.

Theorem 27. *$X_n \xrightarrow{P} X$ if and only if for all subsequence (X_{n_k}), there exists a further subsequence $(X_{n_{k(m)}})$ such that $X_{n_{k(m)}} \to X$ a.s.*

Proof. ($\Rightarrow$) Without loss of generality, we only have to prove that there exists a.s. convergent subsequence. Let $\epsilon_k > 0$, $\epsilon_k \downarrow 0$. There exists $n_k \in \mathbb{N}$ such that if $n \geq n_k$, $P(|X_n - X| > \epsilon_k) < \frac{1}{2^k}$ and $n_{k+1} > n_k$. Hence $\sum_{k=1}^{\infty} P(|X_{n_k} - X| \geq \epsilon_k) \leq 1 < \infty$ and by the first Borel-Cantelli lemma, $P(|X_{n_k} - X| \geq \epsilon_k \text{ i.o.}) = 0$.
($\Leftarrow$) It is easy to show that for a sequence (y_n) on a topological space, if for all subsequence there exists a further subsequence that converges to y, then $y_n \to y$. Given $\delta > 0$, let $y_n = P(|X_n - X| > \delta)$ then it is a sequence on topological space $\mathbb{R}$ that converges to 0. By the condition, the result directly follows. □

Another application is in the form of strong law of large numbers.

Theorem 28 (2.3.5)**.** *$X_1, \cdots, X_n$ are i.i.d. with $EX_1 = \mu$ and $EX_1^4 \leq C < \infty$. $\implies$ $S_n/n \to \mu$ a.s.*

Proof.

$$\begin{aligned} P(|S_n/n - \mu| > \epsilon) &= P(|S_n - n\mu| > n\epsilon) \\ &\leq E(S_n - n\mu)^4/n^4\epsilon^4 \leq C/n^2\epsilon^4. \end{aligned}$$

$\sum_{n=1}^{\infty} P(|S_n/n - \mu| > \epsilon) \leq \sum_{n=1}^{\infty} C_1/n^2 < \infty$ for some C_1. By the first Borel-Cantelli lemma, the result follows. □

Application of the second lemma As a sneak peek of the next section, the second Borel-Cantelli lemma leads to a necessary condition of the strong law of large numbers.

Theorem 29 (necessity of the strong law)**.** *$X_1, X_2, \cdots$ are i.i.d. with $E|X_1| = \infty$. $S_n = X_1 + \cdots + X_n$.*

$$\begin{aligned} \implies &\textit{(i) } P(|X_1| \geq n \textit{ i.o.}) = 1. \\ &\textit{(ii) } P(\lim_n S_n/n \textit{ exists and finite}) = 0. \end{aligned}$$

Note that this implies $E|X_1| < \infty$ is required for the strong law.

Proof. (i) $E|X_1| = \int_0^\infty P(|X_1| > x)dx \leq \sum_{n=0}^{\infty} P(|X_1| \geq n) = \infty$. By the second B-C lemma, $P(|X_1| \geq n \text{ i.o.}) = 1$.
(ii) $\frac{S_n}{n} - \frac{S_{n+1}}{n+1} = \frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1}$. Let $C = \{\lim_n S_n/n \text{ exists and finite}\}$, then on C, $\lim_n \frac{S_n}{n(n+1)} = 0$. On $C \cap (|X_i| \geq n \text{ i.o.})$,

$$\begin{aligned}\limsup_n |\frac{S_n}{n} - \frac{S_{n+1}}{n+1}| &= \limsup_n |\frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1}| \\ &\geq \limsup_n |\frac{X_{n+1}}{n+1}| - \cancelto{0}{\lim_n |\frac{S_n}{n(n+1)}|} \geq 1.\end{aligned}$$

By (i), the desired result follows. □

Generalization of the second lemma Another important result is a generalization of the lemma.

Theorem 30 (2.3.9)**.** *Let $A_1, A_2, \cdots$ be pairwise independent events.*
$\sum_{n=1}^{\infty} P(A_n) = \infty \implies \frac{\sum_{m=1}^n \mathbf{1}_{A_m}}{\sum_{m=1}^n P(A_m)} \to 1$ *a.s.*

Proof. Let $X_m = 1_{A_m}$, $S_n = X_1 + \cdots + X_n$. Note that $0 \leq X_m \leq 1$.
(Step 1: $\frac{\sum_{m=1}^n \mathbf{1}_{A_m}}{\sum_{m=1}^n P(A_m)} = \frac{S_n}{ES_n} \xrightarrow{P} 1$.)
$EX_n^2 = P(\mathbf{1}_{A_n}) < \infty$. Thus $ES_n^2 = \sum_{i=1}^n EX_n^2 + 2\sum_{1 \leq i < j \leq n} EX_iEX_j < \infty$.

$$\begin{aligned}P(|\frac{S_n - ES_n}{ES_n}| > \epsilon) &\leq \frac{\mathrm{Var}(S_n)}{\epsilon^2(ES_n)^2} \\ &\leq \frac{\sum_{i=1}^n EX_i^2}{\epsilon^2(ES_n)^2} \\ &= \frac{\sum_{i=1}^n EX_i}{\epsilon^2(ES_n)^2} \\ &= \frac{1}{\epsilon^2 ES_n} \to 0.\end{aligned}$$

(Step 2: $\exists n_k$ s.t. $\frac{S_{n_k}}{ES_{n_k}} \to 1$ a.s.)
Let $n_k = \inf\{n : ES_n > k^2\}$ and $T_k = S_{n_k}$, then $k^2 \leq ET_k \leq (k+1)^2$. Since $P(|\frac{T_k - ET_k}{ET_k}| > \delta) \leq \frac{1}{\delta^2 ET_k} \leq \frac{1}{\delta^2 k^2}$, $\sum_k P(|\frac{T_k - ET_k}{ET_k}| > \delta) \leq \sum_k \frac{1}{k^2}/\delta^2 < \infty$. By the first Borel-Cantelli lemma, $P(|\frac{T_k - ET_k}{ET_k}| > \delta \text{ i.o.}) = 0$, hence $\frac{T_k}{ET_k} \to 1$ a.s.
(Step 3: sandwich)
Let $\Omega_0 = \{\lim_k \frac{T_k}{ET_k} = 1\}$, then $P(\Omega_0) = 1$. Pick $\omega \in \Omega_0$. Because S_n is non-decreasing, for all n there exists k such that $T_k \leq S_n \leq T_{k+1}$ and $ET_k \leq ES_n \leq ET_{k+1}$ and thus $\frac{T_k(\omega)}{ET_{k+1}} \leq \frac{S_n(\omega)}{ES_n} \leq \frac{T_{k+1}(\omega)}{ET_k}$. Since $\frac{T_k(\omega)}{ET_k} \to 1$ a.s. and $k^2 \leq ET_k \leq ET_{k+1} \leq (k+2)^2$, by sandwich theorem $\frac{S_n(\omega)}{ES_n} \to 1$ and we get the result. □

As well as its implication, the proof scheme of it is important as it is used to prove almost sure convergence of bounded random variables in the form of fractions.

(TBD: example - head runs)

2.4 Strong Law of Large Numbers

Putting together all the topics we have covered so far, we now move on to one of the most impactful theorem in probability theory: the strong law of large numbers (SLLN).

2.4.1 Strong law of large numbers

Theorem 31 (strong law of large numbers)**.** *Let $X_1, X_2, \cdots$ be pairwise independent, identically distributed. $S_n = \sum_{i=1}^n X_i$.*
$E|X_1| = \mu < \infty \implies \frac{S_n}{n} \to \mu$ a.s.

The strong law not only requires less conditions (pairwise rather than mutual independence) than the weak law but provides convergence in almost sure sense. We use the proof scheme similar to that of (2.3.9). In addition, the Cesàro's mean will be used: if $\lim_k Y_k = a$ then $\lim_n \frac{1}{n}\sum_{k=1}^n Y_k = a$.

the strong law. Without loss of generality, we only need to prove the case where $X_i \geq 0$. Let $Y_k = X_k \mathbf{1}_{X_k \leq k}$ and $T_n = \sum_{i=1}^n Y_i$.

¡i¿(Step 1. ¡b¿Claim¡/b¿: $|\frac{T_n}{n} - \frac{ET_n}{n}| \to 0$ a.s. $\implies \frac{S_n}{n} \to \mu$ a.s.)¡/i¿
(i) $\sum_{k=1}^\infty P(X_k > k) \leq EX_1 < \infty$. By the first Borel-Cantelli lemma, $P(X_k > k \text{ i.o.}) = P(X_k \neq Y_k \text{ i.o.}) = 0$ and $X_k \neq Y_k$ for at most finitely many k's. Thus there exists R such that $|T_n - S_n| \leq R < \infty$ almost surely for all n and $|\frac{T_n}{n} - \frac{S_n}{n}| \to 0$ a.s. as $n \to \infty$.
(ii) $\frac{ET_n}{n} = \frac{1}{n}\sum_{k=1}^n EY_k$ and $\lim_k EY_k = EX_1 = \mu$ by MCT. By the Cesàro mean, $\frac{ET_n}{n} \to \mu$.
Combine (i) and (ii) and the claim is proved.

¡i¿(Step 2. $\exists k(n)$ s.t. $|\frac{T_{k(n)}}{k(n)} - \frac{ET_{k(n)}}{k(n)}| \to 0$ a.s.)¡/i¿

$$\begin{aligned} & P\left(|\frac{T_{k(n)}}{k(n)} - \frac{ET_{k(n)}}{k(n)}| > \epsilon\right) \\ \leq & \frac{1}{\epsilon^2 k(n)^2}\mathrm{Var}(T_{k(n)}) \\ = & \frac{1}{\epsilon^2 k(n)^2}\sum_{m=1}^{k(n)}\mathrm{Var}(Y_m) \end{aligned}$$

We want the sum of probabilities

$$\begin{aligned} & \sum_{n=1}^\infty P\left(|\frac{T_{k(n)}}{k(n)} - \frac{ET_{k(n)}}{k(n)}| > \epsilon\right) \\ \leq & \frac{1}{\epsilon^2}\sum_{n=1}^\infty \frac{1}{k(n)^2}\sum_{m=1}^{k(n)}\mathrm{Var}(Y_m) \\ = & \frac{1}{\epsilon^2}\sum_{m=1}^\infty \mathrm{Var}(Y_m)\sum_{n:k(n)\geq m}\frac{1}{k(n)^2} \end{aligned}$$

to be finite. Let $k(n) = \lfloor \alpha^n \rfloor$ for $\alpha > 1$ then since $\lfloor \alpha^n \rfloor \geq \frac{1}{2}\alpha^n$,

$$\begin{aligned}
&\sum_{n:k(n)\geq m} \frac{1}{k(n)^2} \\
&= \sum_{n:k(n)\geq m} \frac{1}{\lfloor \alpha^n \rfloor^2} \\
&\leq 4 \sum_{n:k(n)\geq m} \alpha^{-2n} \\
&\leq \frac{4}{m^2(1-\alpha^{-2})} \approx \frac{c}{m^2}.
\end{aligned}$$

In order to show finiteness of the sum, we need to show $\sum_{k=1}^{\infty} \frac{1}{k^2}\mathrm{Var}(Y_k) < \infty$. This comes from

$$\begin{aligned}
\mathrm{Var}(Y_k) \leq EY_k^2 &= \int_0^k 2yP(X_1 > y)dy. \\
\sum_{k=1}^{\infty} \frac{1}{k^2}\mathrm{Var}(Y_k) &\leq \sum_{k=1}^{\infty} \frac{1}{k^2} \int_0^k 2yP(X_1 > y)dy \\
&= \int_0^{\infty} \left(\sum_{k=1}^{\infty} \frac{1}{k^2} \mathbf{1}_{(y<k)} \right) 2yP(X_1 > y)dy \\
&\leq 4EX_1 < \infty.
\end{aligned}$$

The last inequality holds because

$$\begin{aligned}
\sum_{k:k>y} \frac{1}{k^2} 2y &= 2y \sum_{k:k\geq \lfloor y \rfloor + 1} \frac{1}{k^2} \\
&\leq 2y \int_{\lfloor y \rfloor}^{\infty} \frac{1}{x^2} dx \\
&= \frac{2y}{\lfloor y \rfloor} \leq 4.
\end{aligned}$$

Hence $|\frac{T_{k(n)}}{k(n)} - \frac{ET_{k(n)}}{k(n)}| \to 0$ a.s.

¡i¿(Step 3. sandwich)¡/i¿
$T_{k(n)} \leq T_m \leq T_{k(n+1)}$ for $m: \ k(n) \leq m \leq k(n+1)$. Since $k(n) \to \infty$, every $m \in \mathbb{N}$ has a $k(n)$ and $k(n+1)$ that sandwiches m. As $n \to \infty$, $\frac{k(n)}{k(n+1)} = \frac{\lfloor \alpha^n \rfloor}{\lfloor \alpha^{n+1} \rfloor} \to \frac{1}{\alpha}$. Take lim inf and lim sup to each sides of

$$\frac{T_{k(n)}}{k(n+1)} \leq \frac{T_m}{m} \leq \frac{T_{k(n+1)}}{k(n)}$$

and we get

$$\frac{1}{\alpha}\mu \leq \liminf_m \frac{T_m}{m} \leq \limsup_m \frac{T_m}{m} \leq \alpha\mu \ \text{ a.s., } \forall \alpha > 1.$$

Since α was arbitrary, $\alpha \to 1$ then the proof is done. □

2.4.2 Application of the strong law

I would like to cover three important results that follow the strong law. The first example is about extending the theorem for the case where $EX_1 = \infty$. The second will be about a simple form of renewal theory. The last will be the foundation of the theory of empirical processes: the Glivenko-Cantelli theorem. In addition to the results, I will mention the law of iterated logarithm, which is not named in the textbook, but quite useful to know for the following section of martingales.

1. Extension of the strong law

Theorem 32 (2.4.5)**.** *$X_1, X_2, \cdots$ are i.i.d. with $EX_1^+ = \infty$, $EX_1^- < \infty$.*
$\implies \frac{S_n}{n} \to \infty$ a.s.

This implies the strong law holds even if EX_1 is infinite.

Proof. It suffices to show that $\sum_{i=1}^n X_i^+/n \to \infty$ a.s. Let $Y_i^m = X_i^+ \mathbf{1}_{(X_i^+ \le m)}$ so that Y_i^m's be i.i.d. and integrable. By SLLN, $\sum_{i=1}^n X_i^+/n \ge \sum_{i=1}^n Y_i^m \to EY_1^m$ a.s. By MCT, $EY_1^m \uparrow EX_1^+ = \infty$ as $m \uparrow \infty$. □

2. Renewal theory The following theorem is about a simple version of Renewal theory arose from the question: How frequently should the light bulbs be exchanged to a new one? We are interested in the rate of the number of light bulbs go out until the time t versus t. Intuitively we can guess that it would be one over the mean lifetime of a bulb. The theorem confirms it.

Theorem 33 (renewal theory)**.** *$X_1, X_2, \cdots$ are i.i.d. with $0 < X_i < \infty$ and $EX_i = \mu \le \infty$ for all $i = 1, 2, \cdots$.*
Let $T_n = X_1 + \cdots + X_n$, $N_t = \sup\{n : \ T_n \le t\}$.
$\implies \frac{N_t}{t} \to \frac{1}{\mu}$ a.s.

Here, X_i can be thought of as the lifetime of ith light bulb and N_t as the number of bulbs went out until the time t. We regard $1/\mu = \infty$ if $\mu = 0$.

Proof. $T_n/n \to \mu$ a.s. by SLLN. Observe that $T_{N_t} \le t \le T_{N_t+1}$. As $t \to \infty$, $N_t \to \infty$ a.s. and $\frac{N_t+1}{N_t} \to 1$ a.s. Taking $\lim_{t\to\infty}$ on each terms of

$$\frac{T_{N_t}}{N_t + 1} \le \frac{t}{N_t} \le \frac{T_{N_t+1}}{N_t}$$

we get the result we want. □

3. Glivenko-Cantelli theorem

Theorem 34 (Glivenko-Cantelli)**.** *Let $X_1, X_2, \cdots \overset{iid}{\sim} F$.*
Let $F_n(x) = \frac{1}{n}\sum_{k=1}^n \mathbf{1}_{(X_k \le x)}$.
$\implies \sup_x |F_n(x) - F(x)| \to 0$ a.s. as $n \to \infty$

The theorem states that as the number of independent observations grows, the empirical distribution approaches to the (unknown) population distribution almost surely.

Proof. ¡i¿(1) pointwise convergence¡/i¿
For a fixed $x \in \mathbb{R}$, $F_x(x) \to F(x)$ a.s. Let $Z_n = \mathbf{1}_{(X_n<x)}$ then $F_n(x^-) = \sum_{k=1} nZ_k \to F(x^-)$ a.s.

¡i¿(2) uniform convergence at grid points¡/i¿
Let $x_{jk} = \inf\{y : \ F(y) \geq j/k\}$, $1 \leq j \leq k-1$, $x_{0k} = -\infty$, $x_{kk} = \infty$. By (1), for each k there exists $N_k \in \mathbb{N}$ such that if $n \geq N_k$ then $|F_n(x_{jk}) - F(x_{jk})| < 1/k$ and $|F_n(x_{jk}^-) - F(x_{jk}^-)| < 1/k$ for all $1 \leq j \leq k-1$.

¡i¿(3) uniform convergence¡/i¿
For all $x \in (x_{j-1,k}, x_{jk})$ and $n \geq N_k$,

$$\begin{aligned} F_n(x) &\leq F_n(x_{jk}^-) \leq F(x_{jk}^-) + \frac{1}{k} \\ &\leq F(x_{j-1,k}) + \frac{2}{k} \leq F(x) + \frac{2}{k}. \end{aligned}$$

Similarly, $F_n(x) \geq F(x) - \frac{2}{k}$. Hence the desired result follows. □

4. Law of iterated logarithm I will finish this section by briefly stating the law.

Remark 1. *Let $X_1, X_2, \cdots$ be i.i.d. with $EX_1 = 0$ and $Var(X_1) = \sigma^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\limsup_n \frac{S_n}{\sigma\sqrt{2n\log\log n}} = 1 \ \text{a.s.and} \ \liminf_n \frac{S_n}{\sigma\sqrt{2n\log\log n}} = -1 \ \text{a.s.}$$

This gives us not only the asymptotic rate of S_n but also the intuition that such random walk, regardless of its distribution, fluctuates between the values $\pm\sqrt{2n\log\log n}$. As noted in Wikipedia, it bridges the gap between the law of large numbers (which is about $\frac{S_n}{n}$) and the central limit theorem ($\frac{S_n}{\sqrt{n}}$).

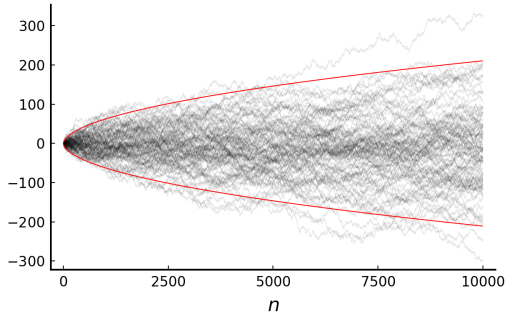

Figure 3: Black: 100 trials of (S_n). Red: $\pm\sqrt{2n\log\log n}$.

2.5 Convergence of Random Series

As the last section in chapter 2, we cover convergence of random series. Especially, since I already explained what tail σ-fields and tail events are, our focus will be on Kolmogorov's maximal inequality and the three series theorem.

2.5.1 Kolmogorov's maximal inequality

Theorem 35 (Kolmogorov's maximal inequality)**.** *If X_i's are independent, $EX_i = 0$, $\mathrm{Var}(X_i) < \infty$, $S_n := \sum_{k=1}^{n} X_k$, then the following inequality holds.*

$$P(\max_{1\le k\le n} |S_k| > x) \le \frac{1}{x^2}\mathrm{Var}(S_n)$$

Proof. Let $A_k = \{S_k > x, S_j \le x, 1 \le j \le n\}$. A_k's are disjoint and $\bigcup_{k=1}^{n} A_k = \{\max_{1\le k\le n} |S_k| > x\}$.

$$\begin{aligned}
\mathrm{Var}(S_n) &= \mathrm{E}S_n^2 \ge \mathrm{E}(S_n^2; \cup_{k=1}^n A_k) \\
&= \sum_{k=1}^{n} \int_{A_k} S_n^2 \, dP \\
&= \sum_{k=1}^{n} \int_{A_k} (S_n - S_k + S_k)^2 \, dP \\
&= \sum_{k=1}^{n} \int_{A_k} (S_n - S_k)^2 + S_k^2 + 2S_k(S_n - S_k) \, dP \\
&\ge \sum_{k=1}^{n} \Big\{ \int_{A_k} S_k^2 + 2 \int_{\Omega} S_k \mathbf{1}_{A_k} \, dP \int_{\Omega} (S_n - S_k) \, dP \Big\} \\
&= \sum_{k=1}^{n} 2 \int \mathbf{1}_{A_k} S_k dP \cdot \int S_n - S_k dP + \sum_{k=1}^{n} \int_{A_k} S_k^2 \, dP \\
&\ge \sum_{k=1}^{n} \int_{A_k} x^2 \, dP \\
&= x^2 P(\cup_{k=1}^n A_k) \\
&= x^2 P(\max_{1\le k\le n} |S_k| > x)
\end{aligned}$$

□

As a result we get the result similar to that of Chebyshev's inequality for maximum of a random series. In the previous post, I also proved that the same results holds with series of shifted starting points.

2.5.2 Kolmogorov's three series theorem

We need two more theorems to prove convergence of random series with certain conditions.

Lemma 4. *If (X_n) decreases almost surely, then $X_n \xrightarrow{P} 0$ if and only if $X_n \to 0$ a.s.*

Theorem 36 (2.5.6)**.** *$X_1, X_2, \cdots$ are independent with $EX_i = 0$, $\sum_{n=1}^{\infty} Var(X_n) < \infty$.*
$\implies$ S_n converges a.s.

Proof. Let $W_m = \sup_{n,k\geq m} |S_n - S_k|$ be a sequence monotonically decreasing to 0. By the maximal inequality,

$$P(\max_{m\leq k\leq N} |S_k - S_m| > \epsilon) \leq \frac{1}{\epsilon^2} \sum_{k=m+1}^{N} \mathrm{Var}(X_k).$$

Take $N \to \infty$ to both sides then as $m \to \infty$,

$$P(\sup_{k\geq m} |S_k - S_m| > \epsilon) \leq \frac{1}{\epsilon^2} \sum_{k=m+1}^{\infty} \mathrm{Var}(X_k) \to 0.$$

This implies $W_m \xrightarrow{P} 0$. By the lemma $W_m \to 0$ a.s. This means for any $\omega \in \Omega_0 = \{W_m \text{ converges}\}$, $P(\Omega_0) = 1$ and $(S_n(\omega))$ is a Cauchy sequence on $\mathbb{R}$. Hence S_n converges a.s. □

Now we state the main theorem.

Theorem 37 (Kolmogorov's three series)**.** *$X_1, X_2, \cdots$ are independent. Let $Y_i = X_i \mathbf{1}_{|X_i|\leq A}$ for $A > 0$. Then, $S_n = X_1 + \cdots + X_n$ converges a.s. if and only if the followings hold.*
(i) $\sum_{n=1}^{\infty} P(|X_n| > A) < \infty$.
(ii) $\sum_{n=1}^{\infty} EY_n$ converges.
(iii) $\sum_{n=1}^{\infty} Var(Y_n) < \infty$.

Intuitively, (i) assures that $X_n = Y_n$ eventually. (ii) and (iii) makes the series S_n to be convergent with the help of the theorem 2.5.6. We will only prove the $\Leftarrow$ direction and leave the other direction for the next chapter.

Proof. ($\Leftarrow$) Let $Z_n = Y_n - EY_n$. Then Z_n's are independent with mean 0 and $\sum_{n=1}^{\infty} \mathrm{Var}(Z_n) < \infty$. By 2.5.6, $\sum_{n=1}^{\infty} Z_n$ converges a.s. and thus by (ii), $\sum_{n=1}^{\infty} Y_n$ also converges a.s. Since the first B-C lemma and (i) implies that $X_n = Y_n$ for all but finite n's, it yields $\sum_{n=1}^{\infty} X_n$ to be almost surely convergent. □

Application The three series theorem can be used to determine convergence of series. It is especially useful to determine aymptotic rate of series, as in conjunction with the Kronecker's lemma, it can be applied to the form $\frac{S_n}{f(n)}$ (just like LIL!). The next theorem is one example.

Lemma 5 (Kronecker)**.** *$a_n \uparrow \infty$ and $\sum_{n=1}^{\infty} \frac{b_n}{a_n}$ converges. Then $\frac{1}{a_n} \sum_{k=1}^{n} b_k \to 0$ as $n \to \infty$.*

Theorem 38 (2.5.12)**.** *$X_1, X_2, \cdots$ are i.i.d. with $EX_1 = 0$, $E|X_1|^p < \infty$ for $1 < p < 2$.*
$\implies \frac{S_n}{n^{1/p}} \to 0$ a.s.

Proof. Let $Y_k = X_k \mathbf{1}_{|X_k|\leq k^{1/p}}$ then we can check it satisfies the conditions of the three series theorem. □

2.6 Convergence Concepts

This is a quick review of convergence concepts and their relations.

2.6.1 Relationship between convergence concepts

Theorem 39 (L^p and P (1))**.** $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{P} X$.

[I already mentioned the proof of it](http://naturale0.github.io/probability/PTE-2.2-Weak-laws-of-large-numbers): Use Markov-Chebyshev inequality.

Theorem 40 (L^p and P (2))**.** *Let $X_n, X \in L^p$ be random variables that $X_n \xrightarrow{P} X$. If there exists $Y \in L^p$ such that $|X_n - X| \leq Y$ for all n, then $X_n \xrightarrow{L^p} X$.*

The result directly follows from the dominated convergence theorem.

Theorem 41. *For $1 \leq q < p$, $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{L^q} X$.*

This is from the Jensen's inequality. Hölder's inequality can also yield the result.

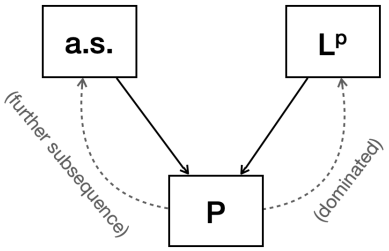

Figure 4: Relationship between convergence concepts.

2.6.2 Counterexamples

Example 3. *Let $(\Omega, \mathcal{F}, P) = ((0,1], \mathcal{B}((0,1]), \lambda)$ where λ is the Lebesgue measure.*
Let $X_{n,m}(\omega) = \begin{cases} 1 & , \ \frac{m-1}{2^n} < \omega \leq \frac{m}{2^n} \\ 0 & , \ otherwise \end{cases}$
Let $\{Y_n : \ n \in \mathbb{N}\} = \{X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2}, \cdots\}$. Then $Y_n \xrightarrow{P} 0$ and $Y_n \xrightarrow{L^p} 0$, but $Y_n \not\to 0$ a.s.

Example 4. *Let $(\Omega, \mathcal{F}, P) = ((0,1], \mathcal{B}((0,1]), \lambda)$ where λ is the Lebesgue measure.*
$X_n(\omega) := \begin{cases} 2^n & , \ 0 < \omega \leq \frac{1}{2^n} \\ 0 & , \ otherwise \end{cases}$
Then $X_n \xrightarrow{P} 0$ but not in L^p nor a.s. In fact, it does not L^p-converge to a constant at all if $p > 1$.

3 Central Limit Theorem

3.1 Weak Convergence

Now that we covered convergence of point estimates (specifically, the sample mean) our next interest is in weaker concept of convergence where convergence in probability is not guaranteed. In undergraduate statistics, we call it convergence in distribution. Here we prefer borrowing terminology from measure theory and call it weak convergence and write $X_n \xrightarrow{w} X$.

3.1.1 Definition

Definition 30 (weak convergence 1)**.** *Let (F_n) be a sequence of distribution functions. F_n converges weakly to a distribution F, if $F_n(x) \to F(x)$ for all x that is a continuity point of F. Likewise, a sequence of random variables X_n converges weakly to a random variable X if the corresponding distribution functions weakly converges.*

The name "weak convergence" was named after weak topology generated by bounded continuous functions: Convergence in weak topology is weak convergence. Some textbook separates the usage of terms weak convergence and convergence in distribution by assigning the former to distribution functions and the latter to random variables. Durrett used the former for distribution and both for random variables.

While Durrett define the concept as pointwise convergence at continuity points, Billingsley define it differently.

Definition 31 (weak convergence 2)**.** *A sequence of probability measures P_n converges weakly to a probability measure P, if $\int f dP_n \to \int f dP_n$ for all bounded continuous function f. $F_n \xrightarrow{w} F$ if their corresponding probability measures weakly converges. $X_n \xrightarrow{w} X$ if $Ef(X_n) \to Ef(X)$ for all bounded continuous f.*

Equivalence of both definitions is not difficult to show. $1 \to 2$ is trivial by the Skorohod's representation theorem (see below) and BCT. To show $2 \to 1$, consider a function

$$g_{x,\epsilon}(y) = \begin{cases} 1 & , \ y \le x \\ 0 & , \ y \ge x + \epsilon \\ \text{linear} & , \ x < y < x + \epsilon \end{cases}$$

for x: a continuity point of F and an arbitrary $\epsilon > 0$. Then $g_{x,\epsilon}$ is continuous and bounded with

$$g_{x-\epsilon,\epsilon} \le \mathbf{1}_{y \le x} \le g_{x,\epsilon}$$

and we get

$$F(x - \epsilon) \le \liminf_n F_n(x) \le \limsup_n F_n(x) \le F(x + \epsilon).$$

Since x was a continuity point and $\epsilon > 0$ was arbitrary, letting $\epsilon \to 0$ yields the result we want.

3.1.2 Converging together lemmas

Lemma 6. *(i) $X_n \xrightarrow{w} X$ and $Y_n \xrightarrow{w} c$ for $c \in \mathbb{R} \implies X_n + Y_n \xrightarrow{w} X + c$.*
(ii) $X_n \xrightarrow{w} X$ and $Y_n \xrightarrow{w} c$ for $c \in \mathbb{R} \implies X_n Y_n \xrightarrow{w} cX$.
(iii) $X_n \xrightarrow{P} X \implies X_n \xrightarrow{w} X$.
(iv) $X_n \xrightarrow{w} c$ for $c \in \mathbb{R} \implies X_n \xrightarrow{P} c$.

The proof is not difficult so I would like to leave it as an exercise.

3.1.3 Tools

Skorohod's representation theorem Even though the weak convergence is weak, we can relate it with much stronger almost sure convergence. Skorohod's representation theorem allows us to do so.

Theorem 42 (Skorohod's representation)**.** *Suppose $F_n \xrightarrow{w} F_\infty$. Then there exists random variables $Y_n \sim F_n$ and $Y_\infty \sim F_\infty$ such that $Y_n \to Y_\infty$ a.s.*

The proof is similar to that of theorem 1.2.2.

Proof. Let $(\Omega, \mathcal{F}, P) = ((0,1), \mathcal{B}(0,1), \lambda)$ where λ is the Lebesgue measure. Define $Y_n(\omega) = \sup\{y : F_n(y) < \omega\}$ so that $Y_n \sim F_n$ for $1 \le n \le \infty$ which implies

$$F(y) < \omega \iff y < Y_\infty(\omega),$$
$$F(z) > \omega \iff z > Y_\infty(\omega).$$

We now need almost sure convergence. Let $a_\omega = \sup\{y : F_\infty(y) < \omega\}$, $b_\omega = \inf\{y : F_\infty(y) > \omega\}$, $\Omega_0 = \{\omega : (a_\omega, b_\omega) = \phi\}$. Observe that $\omega \notin \Omega_0$ are at most countable since (a_ω, b_ω)'s are disjoint non-empty intervals containing distinct rational numbers. Thus $P(\Omega_0) = 1$. Pick an arbitrary $\omega, \omega' \in \Omega_0$ such that $\omega < \omega'$. Then for given $\epsilon > 0$, $Y_n(\omega) \to Y_\infty(\omega)$ (ω is a continuity point of Y_∞) implies for all large n,

$$Y_\infty(\omega) - \epsilon < Y_n(\omega) < Y_\infty(\omega') + \epsilon.$$

So it follows that

$$Y_\infty(\omega) \le \liminf_n Y_n(\omega) \le \limsup_n Y_n(\omega) \le Y_\infty(\omega').$$

Let $\omega' \to \omega$ then for all $\omega \in \Omega_0$, $\lim_n Y_n(\omega) = Y_\infty(\omega)$. □

Note that Ω_0 in the proof is the set where F_∞ is strictly increasing. Hence F_∞^{-1} can be defined on Ω_0. We defined Y_n, Y to be F_n^{-1}, F_∞^{-1} on Ω_0 respectively.

Convergence theorems The theorem allows us to use our favorites tools such as convergence theorems directly with minimum effort of proving it.

Theorem 43 (Fatou's lemma)**.** *$g \ge 0$, g is continuous, $X_n \xrightarrow{w} X_\infty$.* $\implies$ $\liminf_n Eg(X_n) \ge Eg(X_\infty)$.

Proof. Let $Y_n \sim F_n$ for $1 \le n \le \infty$ be random variables such that $Y_n \to Y$ a.s. By Fatou's lemma for almost surely convergent random variables, $\liminf_n Eg(Y_n) \ge Eg(Y_\infty)$. Since $X_n \stackrel{d}{=} Y_n$ for all $1 \le n \le \infty$ the result follows. □

Theorem 44 (DCT)**.** *g, h are continuous, $g > 0$, $\frac{|h(x)|}{g(x)} \to 0$ as $|x| \to \infty$, $\int g dF_n < \infty$. If $F_n \xrightarrow{w} F$, then $\int h dF_n \to \int h dF$.*

Continuous mapping theorem

Theorem 45 (continuous mapping)**.** *g is a measurable function. $D_g = \{x : g \text{ is not continuous at} x\}$. If $X_n \xrightarrow{w} X_\infty$ and $P(X_\infty \in D_g) = 0$, then $g(X_n) \xrightarrow{w} g(X_\infty)$.*

Such D_g is called the **discontinuity set** of g. The theorem extends the possible choice of g from continuous to measurable functions.

Portmanteau theorem The following theorem is so important that it is named portmanteau. It is actually equivalent definitions of weak convergence. I found that almost every theorem regarding weak convergence can be proved with this to some extent.

Theorem 46 (Portmanteau theorem)**.** *For probability measures P_n, P on a measurable space $(\Omega, \mathcal{F})$ where Ω is a metric space, the followings are equivalent.*
(i) $P_n \xrightarrow{w} P$.
(ii) $\limsup_n P_n(F) \le P(F),\ \forall$*closed* F.
(iii) $\liminf_n P_n(G) \ge P(G),\ \forall$*open* G.
(iv) $P_n(A) \to P(A),\ \forall A :\ P(\partial A) = 0$.

Proof. ((i)⇒(ii)) For given $\epsilon > 0$, let $f^\epsilon(x) = (1 - \inf_{y \in F} |x - y|/\epsilon)^+$ so that $\mathbf{1}_F \le f^\epsilon \le \mathbf{1}_{F^\epsilon}$ where $F^\epsilon = \{x : \inf_{y \in F} |x - y| < \epsilon\}$. Then f^ϵ is continuous and bounded so

$$\limsup_n P_n(F) \le \limsup_n \int f^\epsilon dP_n = \int f^\epsilon dP \le P(F^\epsilon).$$

Letting $\epsilon \to 0$ leads to the result.
((ii)⇔(iii)) is trivial.
((ii)&(iii)⇒(iv))

$$\limsup_n P_n(A) \le \limsup_n P_n(\overline{A}) \le P(\overline{A}) = P(A). \liminf_n P_n(A) \ge \liminf_n P_n(A^\circ) \ge P(A^\circ) = P(A).$$

((iv)⇒(i)) Without loss of generality, let f be a continuous function with $0 \le f \le 1$. Note that $P(f = t) > 0$ for at most countably many t's so that $P(\partial\{f > t\}) = 0$. Hence by (iv),

$$\begin{aligned} \int f dP_n &= \int_0^\infty P_n(f > t) dt \\ &\to \int_0^\infty P(f > t) dt = \int f dP. \end{aligned}$$

is the desired result. □

3.1.4 Metric for weak convergence

Definition 32 (Lévy metric)**.** $\rho(F, G) = \inf\{\epsilon :\ F(x - \epsilon) - \epsilon \le G(x) \le F(x + \epsilon) + \epsilon, \forall x\}$ *is the Lévy metric.*

The Lévy metric is a metric defined on a space of distribution functions that metrizes weak convergence. That is, $\rho(F_n, F_\infty) \to 0$ if and only if $F_n \xrightarrow{w} F_\infty$. With this measure we can regard a space of probability measures as a metric space, which is well studied. Our next interest is in whether a sequence of distribution functions converges weakly. To be more specific, subsequential convergence of distribution functions are is the topic of this subsection. Helly's selection theorem shows there always exists a vaguely convergent subsequence. Uniform tightness of a sequence strengthen this result to be weakly convergent.

3.1.5 Helly's selection theorem

Definition 33 (vague convergence)**.** *A sequence of distribution (F_n) vaguely converges to a function F if $F_n(x) \to F(x)$ for all x that is a continuity point of F and write $F_n \to_{v.} F$.*

Notice that unlike weak convergence, vague convergence does not guarantee that the limiting function F is a distribution function.

Theorem 47 (Helly's selection)**.** *For a sequence of distribution functions (F_n), there always exists a subsequence (F_{n_k}) and a decreasing right-continuous function F such that $F_{n_k} \to_{v.} F$.*

Proof. (Step 1. Diagonal argument) Let $q_1, q_2, \cdots \in \mathbb{Q}$ be enumeration of rationals. For $k = 1$, $\{F_1(q_1), F_2(q_1), \cdots\}$ is a bounded real sequence so has a limit point. Pick a subsequence (n_{1k}) such that $F_{1k}(q_1) \to G(q_1)$ for a limit point $G(q_1)$. For $k = 2$, we do the similar work. $\{F_{n_{11}}(q_2), F_{n_{12}}(q_2), \cdots\}$ is a bounded real sequence so has a limit point $G(q_2)$. Picka further subsequence $(n_{2k}) \subset (n_{1k})$ such that $F_{2k}(q_2) \to G(q_2)$. Repeat this process for all consecutive $k \in \mathbb{N}$ and let $m_k = n_{kk}$ so that $F_{mk} \to G(q)$ for all $q \in \mathbb{Q}$. Then by construction G is non-decreasing. Let $F(x) = \inf\{G(q) : \ q \in \mathbb{Q}, q > x\}$.

(Step 2. $F \leftarrow_{v.} F_{m_k}$ is right-continuous)
First we prove that F is right-continuous.

$$\begin{aligned}\lim_{x_n \downarrow x} F(x_n) &= \lim_{x_n \downarrow x} \inf\{G(q) : \ q \in \mathbb{Q}, q > x_n\} \\ &= \inf\{G(q) : \ q \in \mathbb{Q}, q > x_n \text{ for some } n\} \\ &= \inf\{G(q) : \ q \in \mathbb{Q}, q > x\} = F(x).\end{aligned}$$

Next we show that $F_{m_k} \to_{v.} F$. Given x a continuity point of F, pick $r_1, r_2, s \in \mathbb{Q}$ so that $r_1 < r_2 < x < s$ and $F(x) - \epsilon < F(r_1) \le F(r_2) \le F(x) \le F(s) < F(x) + \epsilon$. Then

$$F_{m_k}(r_2) \to G(r_2) \ge F(r_1) = \inf\{G(q) : \ q \in \mathbb{Q}, q > r_1\}, F_{m_k}(s) \to G(s) \le F(s) = \inf\{G(q) : \ q \in \mathbb{Q}, q > s\}.$$

Thus for all large k,

$$F(x) - \epsilon < F_{m_k}(x) < F(x) + \epsilon.$$

so $F_{m_k} \to_{v.} F$. □

3.1.6 Tightness theorem

While Helly's thereom provides valuable result, vague convergence is not enough. We want additional condition that allows us to ensure weak convergence. Uniform tightness[5] is the one.

[5]Many textbooks including Durrett and Billingsley use the term *tightness* to describe this condition. However I want to separate the terms for clarity: *uniform tightness* for a collection of measures, and *tightness* for a measure.

Definition 34 (uniform tightness)**.** *A set of measures $\{P_n\}_{n\in\mathbb{N}}$ is uniformly tight, if for all $\epsilon > 0$ there exists a compact set K such that $\inf_n P_n(K) \geq 1-\epsilon$. Likewise, We say a set of distribution functions $\{F_n\}_{n\in\mathbb{N}}$ is uniformly tight, if for all $\epsilon > 0$ there exists $M \in \mathbb{R}$ such that $\limsup_n(1 - F_n(M) + F_n(-M)) \leq \epsilon$.*

Uniform tightness implies the measure of any set can be approximated from below by some compact set. This condition leads to a stronger result.

Theorem 48 (tightness theorem)**.** *Every subsequential limit of a sequence of distribution functions (F_n) is a distribution function if and only if (F_n) is uniformly tight.*

Proof. ($\Leftarrow$) We need to show that for $F_{n_k} \to_{v.} F$, F is a distribution function. It suffices to show $\lim_{x\to\infty}(1 - F(x) + F(-x)) = 0$. Given $\epsilon > 0$, there exists $M > 0$ such that $\limsup_n(1 - F_n(M) + F_n(-M)) \leq \epsilon$. Let $r < -M$ and $s > M$ be continuity points of F.

$$\begin{aligned} 1 - F(s) + F(r) &= \lim_k (1 - F_{n_k}(s) + F_{n_k}(r)) \\ &\leq \limsup_n (1 - F_n(M) + F_n(-M)) \leq \epsilon. \end{aligned}$$

Thus

$$\limsup_{x\to\infty}(1 - F(x) + F(-x)) \leq \epsilon.$$

($\Rightarrow$) Suppose (F_n) is not tight and show that F is not a distribution function in that case. There exists $\epsilon > 0$ such that $\limsup_n(1 - F_n(M) + F_n(M)) > \epsilon$ for all M. Let $r < 0 < s$ be continuity points of F, then

$$\begin{aligned} 1 - F(s) + F(r) &= \lim_j (1 - F_{n(k_j)}(s) + F_{n(k_j)}(r)) \\ &\geq \liminf_j (1 - F_{n(k_j)}(k_j) + F_{n(k_j)}(-k_j)) \geq \epsilon. \end{aligned}$$

As $s \to \infty$ and $r \to -\infty$, $\lim_{x\to\infty}(1 - F(x) + F(-x)) > 0$ so F is not a distribution function. □

There is a criterion for checking whether a sequence is uniformly tight.

Theorem 49 (sufficiency of tightness)**.** *For a sequence (F_n), if there exists a function $\varphi \geq 0$ such that*

$$\begin{cases} \varphi(x) \to \infty \text{ as } |x| \to \infty. \\ \sup_n \int \varphi dF_n < \infty. \end{cases}$$

then (F_n) is uniformly tight.

Proof.

$$\begin{aligned} & 1 - F_n(M) + F_n(-M) \\ &= \int_{|x|>M} 1 dF_n(x) \\ &\leq \frac{\sup_n \int \varphi dF_n}{\inf_{|x|\geq M} \varphi(x)}. \end{aligned}$$

As $M \to \infty$, $\inf_{|x|\geq M} \varphi(x) \to \infty$ and we get the result. □

3.2 Characteristic Functions

In undergraduate statistics, we learned moment generating functions and that it uniquely determines the distribution if exists. However moment generating function does not always exists for all distributions. Characteristic functions always exists for all real-valued random variables and it provides alternative approach to working with distributions.

3.2.1 Characteristic function

Definition 35 (characteristic function)**.** *For a random variable X,*

$$\varphi_X(t) = Ee^{itX}$$

is the characteristic function of X, where i is the complex unit.

Or we could just notate it **ch.f.** for short. Every characteristic functions have the following properties regardless of the distribution.

Proposition 4 (properties of ch.f.)**.** *(i)* $\varphi_X(0) = 1$.
(ii) $\varphi_X(-t) = \overline{\varphi_X(t)}$.
(iii) $|\varphi_X(t)| \leq 1$.
(iv) φ_X *is uniformly continuous.*
(v) $\varphi_{aX+b} = e^{itb}\varphi_X(at)$.
(vi) $X_1 \perp X_2$ *then* $\varphi_{X_1+X_2} = \varphi_{X_1}\varphi_{X_2}$.

Proof. (iv)

$$\begin{aligned}|\varphi(t+h) - \varphi(t)| &= |E(e^{i(t+h)X - e^{itX}})| \\ &\leq E|\cancel{e^{itX}}^{1}||e^{ihX-1}|\end{aligned}$$

holds for all $t \in \mathbb{R}$. By BCT, $E|e^{ihX} - 1| \to 0$ as $h \to \infty$. Thus for all $\epsilon > 0$, there exists h such that $|\varphi(t+h) - \varphi(t)| < \epsilon$ for all t. □

other proofs are direct from the definition. Another property comes from the additivity of lebesgue integral.

Lemma 7 (3.3.9)**.** *Let $F_1, \cdots, F_n$ be distributions that have $\varphi_1, \cdots, \varphi_n$ as ch.f.s respectively. Let* $\lambda_i \geq 0$, $\sum_{i=1}^n \lambda_i = 1$.
$\implies$ $\sum_{i=1}^n \lambda_i F_i$ *has a ch.f.* $\sum_{i=1}^n \lambda_i \varphi_i$.

This simply says that the mixture of distributions have the ch.f. which is also the mixture.

Proof.

$$\begin{aligned}\varphi_{\sum_{i=1}^n \lambda_i F_i} &= \int e^{itx} d\left(\sum_{i=1}^n \lambda_i F_i(x)\right) \\ &= \sum_{i=1}^n \lambda_i \int e^{itx} dF_n(x) = \sum_{i=1}^n \lambda_i \varphi_i.\end{aligned}$$

□

Using the lemma we can show still another property.

Lemma 8. *φ is a ch.f. $\implies$ Reφ and $|\varphi|^2$ are also ch.f.s.*

Proof. Let $X \sim F$ and $(-X) \sim F'$ be independent.
(i) Let Y be a random variable with distribution $F_Y(x) = 0.5F(x) + 0.5F'(x)$. By the lemma, $\varphi_Y = \mathrm{Re}\varphi_X$.
(ii) $\varphi_{X+X'} = \varphi_X\overline{\varphi_X} = |\varphi_X|^2$. □

Here are some of the ch.f.s of well-known distributions.

Example 5. *(i) (Poisson) $X \sim \mathcal{P}(\lambda)$. $\varphi_X(t) = e^{\lambda(e^{it}-1)}$.*
(ii) (Normal) $X \sim \mathcal{N}(0,1)$. $\varphi_X(t) = e^{-t^2/2}$.
(iii) (Exponential) $X \sim Exp(1)$. $\varphi_X(t) = \frac{1}{1-it}$.
(iv) (Double exponential) $X \sim DE(1)$. $\varphi_X(t) = \frac{1}{1+t^2}$.

Note that the ch.f. of DE(1) is a density of the Cauchy distribution.

For a random variable that has a density with respect to the Lebesgue measure, Characteristic function is in fact the Fourier transform of the distribution function of X. So we can naturally deduce that a proper inversion might fully recover the distribution from ch.f.

3.2.2 Inversion formula

Theorem 50 (inversion formula)**.** *Let $\varphi(t) = \int e^{itx}d\mu(x)$ be the ch.f. of a probability measure μ. If $a < b$, then*

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

Notice that it is not $\int_{-\infty}^{\infty}$ but $\lim_T \int_{-T}^{T}$. We will find out why during the proof.

Proof. Let I_T be the integrand of the left hand side.

$$\begin{aligned}
I_T &= \int_{-T}^{T}\int \frac{e^{-ita} - e^{-itb}}{it}e^{itx}d\mu(x)dt \\
&= \int\int_{-T}^{T} \frac{1}{it}\left(e^{it(x-a)} - e^{it(x-b)}\right)dtd\mu(x) \\
&= \int\int_{-T}^{T} \frac{1}{t}\left(\sin t(x-a) = \sin t(x-b)\right)dtd\mu(x) \\
&= \int R(x-a,T) - R(x-b,T)d\mu(x)
\end{aligned}$$

where $R(\theta,T) = \int_{-T}^{T}\frac{\sin\theta t}{t}dt$. The first equality comes from Fubini's theorem since

$$\frac{e^{-ita} - e^{-itb}}{it} = \int_a^b e^{-itx}dx \le b - a$$

so it is integrable. Now observe that if we let $S(T) = \int_0^T \frac{\sin x}{x} dx$,

$$R(\theta, T) = \begin{cases} \int_{-T\theta}^{T\theta} \frac{\sin x}{x} dx = 2\int_0^{T\theta} \frac{\sin x}{x} dx = 2S(T\theta) & , \ \theta \geq 0 \\ -R(-\theta, T) = -2S(-T\theta) & , \ \theta < 0 \end{cases}$$

We get $R(\theta, T) = 2\text{sgn}(\theta) S(T|\theta|)$. By the result of exercise 1.7.5 of the textbook, $S(T) \to \frac{\pi}{2}$ as $T \to \infty$. Thus

$$R(x-a, T) - T(x-b, T) \to \begin{cases} 2\pi & , \ a < x < b \\ \pi & , \ x = a \text{ or } x = b \\ 0 & , \ x < a \text{ or } x > b \end{cases}$$

Since $|R(\theta, T)| \leq 2 \sup_y \int_0^y \frac{\sin t}{t} dt < \infty$, with the help of BCT we get

$$\begin{aligned} &\lim_T \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \\ &= \lim_T \frac{1}{2\pi} \int R(x-a, T) - R(x-b, T) d\mu(x) \\ &= \frac{1}{2\pi} \int \lim_T \left(R(x-a, T) - R(x-b, T) \right) d\mu(x) \\ &= \frac{1}{2\pi} \left(2\pi\mu(a, b) + \pi\mu\{a, b\} \right) \\ &= \mu(a, b) + \frac{1}{2}\mu\{a, b\}. \end{aligned}$$

□

By inversion formula, φ_X is real if and only if X is symmetric.

If we use $\int_{-\infty}^{\infty}$ instead then $\int_{-\infty}^{\infty} \frac{\sin \theta t}{t} dt$ would not be integrable. If we have $\int |\varphi(t)| dt < \infty$, then it is possible to use $\int_{-\infty}^{\infty}$. The next theorem is about this case.

Theorem 51 (inversion formula for density)**.** *If $\int |\varphi(t)| dt < \infty$, then μ has a bounded continuous density f such that*

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

Observe that it is exactly the same to the inverse Fourier transform. In fact, μ is absolutely continuous to the Lebesgue measure and $f = \frac{d\mu}{d\lambda}$.

Proof. From the proof the inversion formula, $|\frac{e^{-ita} - e^{-itb}}{it}| \leq |b - a|$. Since $\int |\varphi(t)| dt < \infty$,

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \leq \frac{|b-a|}{2\pi} \int_{-\infty}^{\infty} |\varphi(t)| dt < \infty.$$

Thus the inversion formula converges absolutely in this case and $\mu(a, b) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt$. Letting $b \to a$ and it follows that μ has no point masses.
Using the above and Fubini's theorem we get

$$\begin{aligned}\mu(x, x+h) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-itx}-e^{-it(x+h)}}{it}\varphi(t)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{x}^{x+h}e^{-ity}dy\varphi(t)dt \\ &= \int_{x}^{x+h}\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-ity}\varphi(t)dtdy\end{aligned}$$

so $\frac{1}{2\pi}\int e^{-ity}\varphi(t)dt$ is a density of μ.
$|f(y)| \leq \frac{1}{2\pi}\int|\varphi(t)|dt < \infty$ and f is bounded.
$|f(y+h)-f(y)| \leq \frac{1}{2\pi}\int|e^{-ity}(e^{-ith}-1)\varphi(t)|dt \leq \frac{1}{\pi}\int|\varphi(t)|dt < \infty$. By DCT $|f(y+h)-f(y)| \to 0$ as $h \to 0$ and f is continuous. □

The formula indicates that if the ch.f. is integrable, then the distribution has a continuous density and it can be directly calculated. As an example of the inversion, remember the ch.f. of double exponential being a density of Cauchy distribution? In fact, Cauchy distribution has the ch.f. that is the same to a density of double exponential distribution.

3.2.3 Continuity theorem

We now relate pointwise convergence of the characteristic function to weak convergence of distributions.

Theorem 52 (continuity theorem)**.** *Let μ_n, $1 \leq n \leq \infty$ be probability measures with ch.f. φ_n. Then*
(i) $\mu_n \xrightarrow{w} \mu_\infty \implies \varphi_n \to \varphi_\infty$.
(ii) If there exists φ such that $\varphi_n \to \varphi$ and φ is continuous at 0, then $\mu_n \xrightarrow{w} \mu$ where μ is a probability measure with φ as its ch.f.

Proof. (i) Trivial since e^{itx} is bounded and continuous.
(ii) First, show that (μ_n) is uniformly tight. Since φ is continuous at 0, given $\epsilon > 0$, there exists $\delta > 0$ such that $\frac{1}{\delta}\int_0^\delta(1-\varphi(t))dt = \int 2(1-\frac{\sin\delta x}{\delta x})d\mu(x) < \epsilon$. So for large n, $\frac{1}{\delta}(1-\varphi_n(t))dt = \int 2(1-\frac{\sin\delta x}{\delta x})d\mu_n(x) < 2\epsilon$. The left hand side is bounded below by

$$\begin{aligned}&\int 2(1-\frac{\sin\delta x}{\delta x})d\mu_n(x) \\ &\geq \int 2(1-\frac{1}{\delta|x|})d\mu_n(x) \\ &\geq \int_{x:\frac{2}{\delta|x|}\leq 1} 2(1-\frac{1}{\delta|x|})d\mu_n(x) \\ &\geq \int_{x:\frac{2}{\delta|x|}\leq 1} 1d\mu_n(x).\end{aligned}$$

So $1-F_n(2/\delta)+F_n(-2/\delta) < 2\epsilon$ and μ_n is uniformly tight.
By ¡a href="http://naturale0.github.io/probability/PTE-3.2.2-vague-convergencetightness-theorem]tightness

theorem¡/a¿, there exists a subsequence μ_{n_k} that converges weakly to some probability measure μ. We need to show that every subsequence has a further subsequence that weakly converges to the same μ. Given a subsequence (μ_{n_k}), again by tightness theorem there exists a further subsequence that converges weakly to some probability measure ν. By (i), $\varphi_{n_k} \to \varphi_\nu$ and since $\varphi_n \to \varphi$, ϕ_ν should be identical to φ. Now given a bounded continuous function f, $(\int f d\mu_n)$ is a sequence on $\mathbb{R}$. By the last result every subsequence of it has a further subsequence that converges to $\int f d\mu$, so $\mu_n \xrightarrow{w} \mu$. □

3.2.4 Moment generating property

The result during the proof of the continuity theorem implies that smoothness of ch.f. φ is related to the tail probability of underlying measure μ. To be specific, for $\delta > 0$, the inequality holds:

$$\mu(\{x : \ |x| > \frac{2}{\delta}\}) \leq \frac{1}{\delta} \int_{-\delta}^{\delta} (1 - \varphi(t)) dt.$$

This leads to another relationship between the ch.f. and its distribution: differentiability and derivative of the characteristic function is directly related to moments of its distribution.

Theorem 53 (3.3.18)**.** *If $E|X|^n = \int |x|^n d\mu(x) < \infty$, then φ is n times differentiable and $\varphi^{(n)}(t) = \int (ix)^n e^{itx} d\mu(x)$.*

The obvious corollary is that $\varphi^{(n)}(0) = i^n EX^n$ for $1 \leq n \leq p$ if $X \in L^p(\mu)$. The next theorem implies the converse in L^2 case.

Theorem 54 (3.3.21)**.** $\limsup_{h \downarrow 0} \frac{\varphi(h) + \varphi(-h) - 2\varphi(0)}{h^2} > -\infty \implies EX^2 < \infty$.

Proof. For necessity of Fubini's theorem see $\frac{e^{ihx} + e^{-ihx} - 2}{h^2} = \frac{2(\cos hx - 1)}{h^2} \leq 0$. In addition, $\lim_{h \to 0} \frac{2(\cos hx - 1)}{h^2} = \lim_{h \to 0} -x^2 \cos hx = -x^2$.

$$\begin{aligned} EX^2 &= \int x^2 dF(x) \\ &\leq \liminf_{h \downarrow 0} \int \frac{2(1 - \cos hx)}{h^2} dF(x) \\ &= -\limsup_{h \downarrow 0} \frac{\varphi(h) + \varphi(-h) - 2\varphi(0)}{h^2} \\ &< \infty. \end{aligned}$$

The first equality is from the Fubini's theorem and the first inequality is from the Fatou's lemma. □

3.2.5 Estimation of the error term

Before ending the section, I would like to cover the estimation of ch.f.s with Taylor series about 0. This will be used when proving the central limit theorem (the next section) and the existence of canonical representation of infinitely divisible distributions (section 3.9).

Theorem 55 (3.3.19)**.**

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} \right| \leq \frac{|x|^{n+1}}{(n+1)!} \wedge \frac{2|x|^n}{n!}.$$

Proof. Let $R_n(x) = e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!}$. We will prove by induction. For $n = 0$, $R_0(x) = e^{ix} - 1 = \int_0^x ie^{iy}dy$. Thus

$$|R_0(x)| = \begin{cases} |e^{ix} - 1| \le |e^{ix}| + 1 = 2 \\ |\int_0^x ie^{iy}dy| \le \int_0^x |ie^{iy}|dy = x \end{cases}$$

Now suppose $|R_n(x)| \le \frac{|x|^{n+1}}{(n+1)!} \wedge \frac{2|x|^n}{n!}$. $R_{n+1} = R_n(x) - \frac{(ix)^{n+1}}{(n+1)!} = \int_0^x iR_n(y)dy$. This is from term-by-term integration of iR_n.

$$\begin{aligned} |R_{n+1}| &\le \int_0^x |iR_n(y)|dy \\ &\le \int_0^x (\frac{|y|^{n+1}}{(n+1)!} \wedge \frac{2|y|^n}{n!})dy \\ &= \frac{|x|^{n+2}}{(n+2)!} \wedge \frac{2|x|^{n+1}}{(n+1)!}. \end{aligned}$$

□

Using this, we can reduce the order of error to 2 (from 3) for second-order Taylor series. We write $g(x) = o(f(x))$ if $g(x)/f(x) \to 0$ as $x \to 0$.

Theorem 56 (upper bound of the error)**.**

$$EX^2 < \infty \implies \varphi(t) = 1 + itEX - \frac{t^2EX^2}{2} + o(t^2).$$

Proof. By the previous theorem we get

$$\begin{aligned} |ER_2(tX)| &\le E|R_2(tX)| \\ &\le E(\frac{|tX|^3}{3!} \wedge \frac{2|tX|^2}{2!}) \\ &\le t^2E(|t||X|^3 \wedge X^2). \end{aligned}$$

As $t \to 0$, $|t||X|^3 \wedge X^2 \to 0$ and $|t||X|^3 \wedge X^2 \le X^2$ with $EX^2 < \infty$. By DCT, $ER_2(tX)/t^2 \to 0$ as $t \to 0$ so $ER_2(tX) = o(t^2)$. □

3.3 Central Limit Theorem

Now that we have all the right tools, we state and prove the central limit theorem (CLT for short), starting from the simplest form for i.i.d. cases and to Lindeberg and Lyapounov conditions.

3.3.1 Central limit theorem for i.i.d. sequences

We need three short lemmas on complex numbers for the proof of our first CLT.

Lemma 9 (3.4.3)**.** *Let $\xi_1, \cdots, \xi_n$, $\omega_1, \cdots, \omega_n \in \mathcal{C}$, $|xi_i|, |\omega_i| \le \theta$. Then*

$$\left| \prod_{i=1}^{n} \xi_i - \prod_{i=1}^{n} \omega_i \right| \le \theta^{n-1} \sum_{i=1}^{n} |\xi_i - \omega_i|.$$

The proof is clear from induction.

Lemma 10 (3.4.4)**.** *For $b \in \mathbb{C}$ such that $|b| \leq 1$, $|e^b - (1+b)| \leq |b|^2$.*

Proof. Taylor expansion gives

$$e^b - (1+b) = \frac{b^2}{2!} + \frac{b^3}{3!} + \cdots .$$

Thus

$$|e^b - (1+b)| \leq \frac{|b|^2}{2}(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots) = |b|^2.$$

□

Lemma 11 (3.4.2)**.** *For $c_n \to c \in \mathbb{C}$, $(1 + \frac{c_n}{n})^n \to e^c$.*

Proof. Let $\gamma > |c|$ then $\gamma > |c_n|$ for large n. By the two lemmas above,

$$\begin{aligned} |e^{c_n} - (1 + \frac{c_n}{n})^n| &\leq |(e^{\frac{c_n}{n}})^n - (1 + \frac{c_n}{n})^n| \\ &\leq (e^{\frac{\gamma}{n}})^{n-1} |e^{\frac{c_n}{n}} - (1 + \frac{c_n}{n})| \\ &\leq e^{\frac{(n-1)\gamma}{n}} n |\frac{c_n}{n}|^2 \to 0 \end{aligned}$$

as $n \to \infty$. □

Now we prove the main theorem.

Theorem 57 (CLT for iid sequence)**.** *Let $X_1, \cdots, X_n$ be i.i.d. with $EX_1 = \mu$ and $Var(X_1) = \sigma^2 < \infty$. $\implies \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{w} \mathcal{N}(0,1)$ as $n \to \infty$.*

Proof. Without loss of generality, let $\mu = 0$. Let $S = \frac{S_n}{\sigma\sqrt{n}}$ and φ be ch.f. of X. By ¡a href="https://naturale0.github.io/p
3.3-characteristic-functionscontinuity-theorem]continuity theorem¡/a¿, showing $\varphi_S \to e^{-\frac{t^2}{2}}$ is enough.
Since $EX^2 < \infty$, ¡a href="https://naturale0.github.io/probability/PTE-3.3-characteristic-functionsestimation-of-the-error-term]the upper bound of the error¡/a¿ gives us the nice approximation $\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2)$.

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

This with the lemmas gives us

$$\begin{aligned} &\left|\varphi_S(t) - \left(1 - \frac{t^2}{2n}\right)^n\right| \\ &= \left|\left(\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n - \left(1 - \frac{t^2}{2n}\right)^n\right| \\ &\leq n\left|\varphi\left(\frac{t}{\sigma\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right)\right| \\ &= n \cdot o\left(\frac{1}{n}\right) \to 0 \end{aligned}$$

as $n \to \infty$ and the proof is done. □

The next example shows that unlike the strong law of large numbers, pairwise independence is not enough for the central limit theorem.

3.4 Poisson Convergence

I would like to finish reviewing Probability theory I by briefly mentioning the Poisson convergence (section 3.6) and limit theorems in $\mathbb{R}^d$ (3.10).

Poisson convergence is about limiting laws of a sequence of independent discrete Bernoulli-like random variables with ES_n converges to a constant λ. I used the term "Bernoulli-like" because for large enough n, such X_n should behave similar to Bernoulli random variables. For those sequences, the limiting law is not normal, but actually Poisson.

3.4.1 Basic Poisson convergence

Theorem 58 (3.6.1)**.** *Let $(X_{nk})_{k=1}^n$ be a rowwise independent triangular array with $P(X_{nk}=1)=p_{nk}$ and $P(X_{nk}=0)=1-p_{nk}$. As $n\to\infty$, if*
(i) $\sum_{k=1}^n p_{nk}\to\lambda\in(0,\infty)$.
(ii) $\max_{1\le k\le n} p_{nk}\to 0$.
Then $S_n=\sum_{k=1}^n X_{nk}\xrightarrow{w}\mathcal{P}(\lambda)$.

Proof. As in the proof of the CLT, it suffices to show $\varphi_{S_n}(t)\to\exp(\lambda(e^{it}-1))$.

$$\begin{aligned}\varphi_{S_n}(t)&=\prod_{k=1}^n\varphi_{nk}(t)\\&=\prod_{k=1}^n\left(p_{nk}(e^{it}-1)+1\right).\end{aligned}$$

Thus

$$\begin{aligned}&\left|\varphi_{S_n}-e^{\lambda(e^{it}-1)}\right|\\&=\left|\prod_{k=1}^n\left(p_{nk}(e^{it}-1)+1\right)-\prod_{k=1}^n e^{p_{nk}(e^{it}-1)}\right|\\&=\sum_{k=1}^n\left|p_{nk}(e^{it}-1)+1-e^{p_{nk}(e^{it}-1)}\right|\\&\quad(b=p_{nk}(e^{it}-1),\ |b|\le 2p_{nk}\le 1\text{ for large n.})\\&\le\sum_{k=1}^n p_{nk}^2|e^{it}-1|^2\le 4\sum_{k=1}^n p_{nk}^2\\&\le 4\max_{1\le k\le n}p_{nk}\sum_{k=1}^n p_{nk}\to 0.\end{aligned}$$

The first inequality is from the lemmas we covered before

$$\begin{aligned}&|p_{nk}(e^{it}-1)+1|\le 1-p_{nk}+p_{nk}|e^{it}|\le 1.\\&|e^{p_{nk}(e^{it}-1)}|\le e^{p_{nk}|e^{it}-1|}\le e^{2p_{nk}}\le 1\text{ for large }n.\end{aligned}$$

□

For a special case, consider $X_k\overset{\text{iid}}{\sim}\mathcal{B}(p_n)$ for $k=1,\cdots,n$ where $np_n\to\mu$. Then $S_n\xrightarrow{w}\mathcal{P}(\mu)$.

3.4.2 General Poisson convergence

Theorem 59 (Poisson convergence)**.** *Let $(X_{nk})_{k=1}^n$ be rowwise independent, non-negative, integer-valued triangular array with $P(X_{nk}=1)=p_{nk}$ and $P(X_{nk}\geq 2)=\epsilon_{nk}$. As $n\to\infty$ if*
(i) $\sum_{k=1}^n p_{nk}\to\lambda\in(0,\infty)$.
(ii) $\max_{1\leq k\leq n} p_{nk}\to 0$.
(iii) $\sum_{k=1}^n \epsilon_{nk}\to 0$.
then $S_n=\sum_{k=1}^n X_{nk}\xrightarrow{w}\mathcal{P}(\lambda)$.

Proof. Let $X'_{nk}=X_{nk}$ if $X_{nk}=1$ and 0 otherwise. Then

$$P(S_n\neq S'_n)\leq\sum_{k=1}^n P(X_{nk}\geq 2)=\sum_{k=1}^n\epsilon_{nk}\to 0.$$

By basic Poisson convergence theorem, $S'_n\xrightarrow{w}\mathcal{P}(\lambda)$. By the ¡a href="https://naturale0.github.io/probability/PTE-3.2.1-weak-convergenceconverging-together-lemmas]converging together lemma¡/a¿, we get the desired result. □

3.4.3 Total variation and weak convergence

Lastly, I will introduce total variation, a metric on a space of discrete probability measures. Like Levy metric, convergence in total variation distance is equivalent to weak convergence on such space.

Definition 36 (total variation)**.** *Let μ,ν be measures on countable space $\mathcal{S}$.*

$$\begin{aligned}\|\mu-\nu\| &:= \frac{1}{2}\sum_{z\in\mathcal{S}}|\mu(z)-\nu(z)|\\ &= \sup_{A\subset\mathcal{S}}|\mu(A)-\nu(A)|\end{aligned}$$

is the total variation distance between μ and ν.

While the first formula is the definition, the second one is a derived property. Note that

$$\begin{aligned}\sum_{z\in\mathcal{S}}|\mu(z)-\nu(z)| &\geq |\mu(A)-\nu(A)|+|\mu(A^c)-\nu(A^c)|\\ &= 2|\mu(A)-\nu(A)|.\end{aligned}$$

Let $A=\{z:\ \mu(z)\geq\nu(z)\}$ then the equality holds.

Lemma 12 (total variation and weak convergence)**.** *(i) $d(\mu,\nu)=\|\mu-\nu\|$ is a metric on the space of probability measures on $\mathbb{Z}$.*
(ii) $\|\mu_n-\mu\|\to 0$ if and only if $\mu_n(z)\to\mu(z)$ for all $z\in\mathbb{Z}$. (i.e. $\mu_n\xrightarrow{w}\mu$.)

Proof. (ii)
($\Rightarrow$) Given $z\in\mathbb{Z}$,

$$|\mu_n(z)-\mu(z)|\leq\sup_{A\subset\mathbb{Z}}|\mu_n(A)-\mu(A)|\to 0.$$

($\Leftarrow$) $\mu_n\xrightarrow{w}\mu$ then

$$\frac{1}{2}\sum_{z\in\mathbb{Z}}|\mu_n(z)-\mu(z)|=\sum_{z\in\mathbb{Z}}(\mu_(z)-\mu(z))^+\to 0$$

by DCT. □

3.5 Limit Theorems in R^d

This part covers limit theorems regarding random vectors $\mathbf{X} = (X_1, \cdots, X_d)' \in \mathbb{R}^d$.

3.5.1 Definitions

Definition 37 (distribution function). *$F : \mathbb{R}^d \to \mathbb{R}$ is a distribution function if it is*
(i) non-decreasing.
(ii) right-continuous.
(iii) $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$. *Where* $x \to \infty$ *means* $x_i \to \infty$ *for all* $i = 1, \cdots, d$ *and* $x \downarrow x$ *means* $y_i \downarrow x_i$ *for all* i.
(iv) $\Delta_A F := P(X \in A) \geq 0$ *for all rectangle* A, *where* $P(X \in A) = \sum_{v \in V} sgn(v) F(v)$, $A = (a_1, b_1] \times \cdots \times (a_d, b_d]$, $V = \{a_1, b_1\} \times \cdots \times \{a_d, b_d\}$, $sgn(v) = (-1)^{(\#\ of\ a_i\text{'}s\ in\ v)}$.

3.5.2 Limit theorems

Definition 38 (Lipschitz continuity). *f is Lipschitz continuous on a metric space (S, ρ) if there exists $c \in \mathbb{R}$ such that $|f(x) - f(y)| \leq c \cdot \rho(x, y)$ for all $x, y \in S$.*

We arrive at another portmanteau theorem.

Theorem 60 (Portmanteau theorem). *The followings are equivalent.*
(i) $Eg(X_n) \to Eg(X_\infty)$, $\forall g$: *bounded, continuous.*
(ii) $Eg(X_n) \to Eg(X_\infty)$, $\forall g$: *bounded, Lipschitz continuous.*
(iii) $\limsup_n P(X_n \in F) \leq P(X_\infty \in F)$, $\forall F$: *closed.*
(iv) $\liminf_n P(X_n \in G) \geq P(X_\infty \in G)$, $\forall G$: *open.*
(v) $\lim_n P(X_n \in A) = P(X_\infty \in A)$, $\forall A$: $P(X_\infty \in \partial A) = 0$.
(vi) $Ef(X_n) \to Ef(X_\infty)$, $\forall f$: *bounded and* $P(X_\infty \in D_f) = 0$.

In addition, we get multi-dimensional version of tightness theorem.

Definition 39 (uniform tightness). *(μ_n) is uniformly tight if for all $\epsilon > 0$, there exists M such that* $\liminf_n \mu_n([-M, M]^d) \geq 1 - \epsilon$.

Theorem 61 (Prohorov). *If (μ_n) is tight, there exists a weakly convergent subsequence.*

The proof uses both Helly's selection theorem and uniform tightness.

3.5.3 Characteristic functions and inversion

Definition 40 (characteristic function).

$$\varphi(\mathbf{t}) := Ee^{i\mathbf{t} \cdot \mathbf{X}} = Ee^{i(t_1 X_2 + \cdots + t_d X_d)}$$

where $\mathbf{t} = (t_1, \cdots, t_d)$ *and* $\mathbf{X} = (X_1, \cdots, X_d)$.

Theorem 62 (inversion formula). *Let* $A = [a_1, b_1] \times \cdots \times [a_d, b_d]$ *such that* $\mu(\partial A) = 0$. *Let φ be a ch.f. of μ. Define*

$$\psi_j(t_j) = \frac{1}{it_j} \left(e^{-it_j a_j} - e^{-it_j b_j} \right)$$

then

$$\mu(A) = \lim_T \frac{1}{(2\pi)^d} \int_{[-T,T]^d} \prod_{j=1}^{d} \psi_j(t_j)\varphi(\mathbf{t})d\mathbf{t}.$$

We achieve the result by Fubini's theorem and the inversion formula for random variables.

3.5.4 Central limit theorem in $\mathbb{R}^d$

Theorem 63 (Levy's convergence)**.** *Let* $\mathbf{X}_n$, $1 \le n \le \infty$ *be random vectors with ch.f.* φ_n. *then* $\mathbf{X}_n \xrightarrow{w} \mathbf{X}_\infty$ *if and only if* $\varphi_n(\mathbf{t}) \to \varphi_\infty(\mathbf{t})$.

Corollary 6 (Cramer-Wold device)**.**

$$\theta \cdot \mathbf{X}_n \xrightarrow{w} \theta \cdot \mathbf{X}_\infty, \ \forall \theta \in \mathbb{R}^d \implies \mathbf{X}_n \xrightarrow{w} \mathbf{X}_\infty.$$

Cramer-Wold device implies if every linear combination converges weakly to the same random variable, then the random vector itself weakly converges. In fact for normal random vector we can prove that $\mathbf{X} \sim \mathcal{N}_d(0, \mathbf{\Gamma})$ if and only if $\mathbf{t} \cdot \mathbf{X} \sim \mathcal{N}(0, \mathbf{t}'\mathbf{\Gamma}\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^d$.

Theorem 64 (central limit theorem)**.** $\mathbf{X}_i$*'s are i.i.d. random vectors with* $E\mathbf{X}_i = \mu$, $\Gamma_{ij} = E(X_{ni} - \mu_i)(X_{nj} - \mu_j)$.
Let $\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n$, *then* $\frac{\mathbf{S}_n - n\mu}{\sqrt{n}} \xrightarrow{w} \mathcal{N}_d(\mathbf{0}, \mathbf{\Gamma})$.

Proof. Without loss of generality, let $\mu = 0$. Given $\mathbf{t} \in \mathbb{R}^d$,

$$\frac{1}{\sqrt{n}}\mathbf{t} \cdot \mathbf{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{t} \cdot \mathbf{X}_i \xrightarrow{w} \mathcal{N}(0, \sigma_t^2)$$

where $\sigma_t^2 = \mathrm{Var}(\mathbf{t} \cdot \mathbf{X}_i) = \mathbf{t}'\mathbf{\Gamma}\mathbf{t}$. By Cramer-Wold device and the fact, the desired result follows. □