Probability theory is the mathematical study of randomness. Today we will recall some definitions and results from measure theory. Every experiment or observation results in an outcome or sample point. Let $\Omega$ be the sample space consisting of all outcomes. We will use $\omega$ to denote an outcome from the sample space $\Omega$. Let $2^\Omega$ denote the power set of $\Omega$, consisting of all subsets of $\Omega$. First we define field and $\sigma$-field.

### 1.1 Fields and $\sigma$-fields

**Definition 1.1.** Let $\mathcal{F}$ be a collection of subsets of $\Omega$. $\mathcal{F}$ is called a field or algebra if the following holds

(i) $\Omega \in \mathcal{F}$.

(ii) $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$ and

(iii) $A, B \in \mathcal{F}$ implies that $A \cup B \in \mathcal{F}$.

**Definition 1.2.** Let $\mathcal{F}$ be a collection of subsets of $\Omega$. $\mathcal{F}$ is called a $\sigma$-field or $\sigma$-algebra if the following holds

(i) $\Omega \in \mathcal{F}$.

(ii) $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$ and

(iii) if $A_1, A_2, \ldots \in \mathcal{F}$ is a countable sequence of sets then $\bigcup_{i=1}^\infty A_i \in \mathcal{F}$.

Here and in what follows, countable means finite or countably infinite. Clearly every $\sigma$-field is a field. Since $(\bigcap_{i=1}^\infty A_i)^c = \bigcup_{i=1}^\infty A_i^c$, every $\sigma$-field is also closed under countable intersection. The tuple $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a $\sigma$-field is called a measurable space. We will use the notation $\sqcup$ to emphasize disjoint union. Now let us give some examples of field and $\sigma$-field.

**Example 1.3.** $(\Omega, \mathcal{F})$ with $\mathcal{F} = \{\emptyset, \Omega\}$. $\mathcal{F}$ is the smallest $\sigma$-field on $\Omega$, called the trivial $\sigma$-field.

**Example 1.4.** $(\Omega, \mathcal{F})$ with $\mathcal{F} = 2^\Omega$. $2^\Omega$ is the largest $\sigma$-field on $\Omega$.

**Example 1.5.** Let $\Omega$ be any sample space. Let $\mathcal{A}$ be the collection of all countable subsets and co-countable (complement is countable) subsets of $\Omega$. Check that $(\Omega, \mathcal{A})$ is a $\sigma$-field. We will call it the countable co-countable $\sigma$-field on $\Omega$.

**Example 1.6.** Let $\Omega = \{1, 2, \ldots\}$. Let $\mathcal{A}$ be the collection of all finite subsets and co-finite (complement is finite) subsets of $\Omega$. Check that $(\Omega, \mathcal{A})$ is a field. Note that $\{2\}, \{4\}, \{6\}, \ldots \in \mathcal{A}$ but $\{2, 4, 6, \ldots\} = \bigcup_{i=1}^\infty \{2i\} \notin \mathcal{A}$. Thus $\mathcal{A}$ is not a $\sigma$-field.

In general $\sigma$-field is a complicated object. But it is easy to construct one.
Definition 1.7. Given a collection $C$ of subsets of $\Omega$ (need not be a field) we define $\sigma(C)$, the $\sigma$-field generated by $C$, as the smallest $\sigma$-field containing $C$. Similarly, we define $\mathcal{A}(C)$, the field generated by $C$, as the smallest field containing $C$.

Lemma 1.8. For any $C$, $\sigma(C)$ and $\mathcal{A}(C)$ exist.

Proof. We will only prove the existence of $\sigma(C)$. The other part is similar. Let $\Lambda$ be the collection of all $\sigma$-fields containing $C$. Note that $\Lambda$ is non-empty as $2^\Omega$, the power-set of $\Omega$ is in $\Lambda$. Define

$$\mathcal{F} = \bigcap_{\mathcal{G} \in \Lambda} \mathcal{G}.$$ 

Clearly $C \subseteq \mathcal{F}$ and $\mathcal{F}$ is contained in all $\sigma$-fields containing $C$. We claim that $\mathcal{F}$ is a $\sigma$-field (Verify the axioms of $\sigma$-field!). This completes the proof. $\blacksquare$

Even if $C$ is a “nice” collection of sets, $\sigma(C)$ can become quite complicated. We give some examples.

Example 1.9. Let $\Omega$ be a sample space and $C$ be the collection of all singleton subsets of $\Omega$. We claim that $\sigma(C)$ is the countable co-countable $\sigma$-field on $\Omega$.

Definition 1.10. Let $\Omega$ be a topological space with $C$ being the collection of all open subsets of $\Omega$. The $\sigma$-field $\mathcal{B}(\Omega) = \sigma(C)$ is called the Borel $\sigma$-field on $\Omega$.

We will denote the Borel $\sigma$-field of $\mathbb{R}^n$ by $\mathcal{B}^n$. When $n = 1$ we will simply use $\mathcal{B}$.

1.2 Measure

Definition 1.11. A measure is a nonnegative countable additive set function on a measurable space $(\Omega, \mathcal{F})$, i.e., a function $\mu : \mathcal{F} \to \mathbb{R}$ such that

(i) $\mu(A) \geq 0$ for all $A \in \mathcal{F}$, and

(ii) if $A_i \in \mathcal{F}$ is a countable sequence of disjoint sets (i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$), then

$$\mu \left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mu(A_i).$$

A Probability measure $\mathbb{P}$ is a measure with $\mathbb{P}(\Omega) = 1$. A Probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathcal{F}$ is a $\sigma$-field and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.

We define a pre-measure $\nu$ as a nonnegative countable additive set function on $(\Omega, \mathcal{A})$, where $\mathcal{A}$ is only a field, i.e., a function $\mu : \mathcal{A} \to \mathbb{R}$ such that

(i) $\nu(A) \geq 0$ for all $A \in \mathcal{A}$, and

(ii) if $A_i \in \mathcal{A}$ is a countable sequence of disjoint sets (i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$) such that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$, then

$$\nu \left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \nu(A_i).$$
Abusing notation, we will also call a pre-measure as a measure (which will be clear from the context).

It is easy to check that for a measure (or pre-measure) $\mu$ we have $\mu(\emptyset) = 0$ (by condition (ii), taking $A_1 = A_2 = \cdots = \emptyset$). First we mention some properties of a measure (also true for pre-measure).

**Theorem 1.12.** Let $\mu$ be a measure on the measurable space $(\Omega, \mathcal{F})$. Then the following hold.

(a) (Monotonicity) If $A \subseteq B$ then $\mu(A) \leq \mu(B)$.

(b) (Sub-additivity) If $A \subseteq \bigcup_{i=1}^{\infty} A_i$ then $\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i)$.

(c) (Continuity from below) If $A_i \uparrow A$ (i.e., $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \bigcup_{i=1}^{\infty} A_i$) then $\mu(A_i) \uparrow \mu(A)$.

(d) (Continuity from above) If $A_i \downarrow A$ (i.e., $A_1 \supseteq A_2 \supseteq \cdots$ and $A = \bigcap_{i=1}^{\infty} A_i$) then $\mu(A_i) \downarrow \mu(A)$.

Before going to the proof, we explain an important “disjointification” operation for a sequence of sets that will be used repeatedly. Given a sequence of sets $A_1, A_2, A_3, \ldots$ (which can be arbitrary) if we define

$$B_1 = A_1, B_2 = A_2 \setminus A_1, B_3 = A_3 \setminus (A_1 \cup A_2), B_4 = A_4 \setminus (A_1 \cup A_2 \cup A_3), \ldots,$$

then $B_i$’s are disjoint and $\bigcup_{i=1}^{m} B_i = \bigcup_{i=1}^{m} A_i$ for all $m \geq 1$. Now we proceed to the proof.

**Proof.** (a) Let $C = B \setminus A$. Clearly, $C \in \mathcal{F}$ and $C \cup A = B$. Thus

$$\mu(B) = \mu(C \cup A) = \mu(C) + \mu(A) \geq \mu(A).$$

(b) We consider the sequence of sets $B_i = A \cap A_i$ and disjointify it to get $C_i$ such that $A = \bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} C_i$. Clearly $C_i \subseteq B_i \subseteq A_i$. Thus, using (ii) of the definition of measure and part (a) of this theorem, we get

$$\mu(A) = \mu(\bigcup_{i=1}^{\infty} C_i) = \sum_{i=1}^{\infty} \mu(C_i) \leq \sum_{i=1}^{\infty} \mu(A_i).$$

(c) Using the disjointification operation on the sequence of sets $A_1, A_2, A_3, \ldots$, we get the disjoint sequence $B_1 = A_1, B_i = A_i \setminus A_{i-1}$. We have $A = \bigcup_{i=1}^{\infty} B_i$ and so

$$\mu(A) = \mu(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mu(B_i) = \lim_{m \to \infty} \sum_{i=1}^{m} \mu(B_i) = \lim_{m \to \infty} \mu\left(\bigcup_{i=1}^{m} B_i\right) = \lim_{m \to \infty} \mu(A_m).$$

(d) Note that $A_1 \setminus A_n \uparrow A \setminus A$. Thus part (c) of this theorem implies that $\mu(A_1 \setminus A_n) \uparrow \mu(A \setminus A)$. Now part (a) shows that for $A \subseteq B$, $\mu(B \setminus A) = \mu(B) - \mu(A)$. It follows that $\mu(A_n) \downarrow \mu(A)$. ■

**Definition 1.13.** A measure $\mu$ is **finite** if $\mu(\Omega) < \infty$.

**Definition 1.14.** A measure $\mu$ is **$\sigma$-finite** if $\Omega = \bigcup_{i=1}^{\infty} A_i$ with $\mu(A_i) < \infty, \forall i$.

**Example 1.15.** Let $\Omega = \text{set of outcomes if we toss 2 dice} = \{1, 2, \cdots, 6\} \times \{1, 2, \cdots, 6\}$, and $\mathcal{F} = 2^\Omega$, then we can define measure $\mu(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{36}$.

**Example 1.16.** Let $\Omega$ be a countable set, $\mathcal{F} = 2^\Omega$. Take a sequence $p_w \geq 0, w \in \Omega$ s.t. $\sum_{w \in \Omega} p_w = 1$, then $\mu(A) = \sum_{w \in A} p_w$ is a probability measure.
Example 1.17. Let $\Omega = \mathbb{R}$ and $\mathcal{F}$ be a countable co-countable $\sigma$-field. Define

$$
\mu(A) = \begin{cases} 
0, & \text{if } A \text{ is countable} \\
1, & \text{if } A \text{ is co-countable}.
\end{cases}
$$

Proof. Clearly, $\mu(A) \geq 0$ and $\mu(\Omega) = 1$. Let $A = \cup_{i=1}^{\infty} A_i$. We consider two separate cases:

Case 1: $A$ is countable. Then each $A_i$ is countable and $\mu(A) = 0 = \sum_{i=1}^{\infty} \mu(A_i)$.

Case 2: $A$ is co-countable. If sets $X, Y$ are both co-countable, then $X \cap Y \neq \emptyset$. Otherwise, $X^c \cup Y^c = \Omega$ and the left side is countable while the right side is uncountable. Therefore, when $A$ is co-countable, exactly one of $A_i$ is co-countable, then $\mu(A) = 1 = 1 + 0 + \cdots = \sum_{i=1}^{\infty} \mu(A_i)$.

Example 1.18. Let $\Omega = (0, 1)$ and $\mathcal{F} = \mathcal{B}(\Omega) = \sigma(\text{open subsets of } \Omega)$. Define $\mu((a, b]) = b - a, 0 \leq a < b < 1$. How to define $\mu(A)$ in general?

Given a collection of subsets $\mathcal{C}$ one can always construct a field or $\sigma$-field containing $\mathcal{C}$. However, given a nonnegative countable additive set function on $\mathcal{C}$ (think about open sets), it might not be possible to extend it as a measure to $\sigma(\mathcal{C})$. However, this is possible under certain condition. We will mention one such condition next.

Definition 1.19. We define a semialgebra $\mathcal{S}$ as a collection of subsets such that

(i) $\Omega \in \mathcal{S}$,

(ii) $A, B \in \mathcal{S}$ implies that $A \cap B \in \mathcal{S}$ and

(iii) $A \in \mathcal{S}$ implies that $A^c$ is a finite disjoint union of sets in $\mathcal{S}$.

Note that the collection of sets

$$\mathcal{S} = \{(a, b] \mid -\infty \leq a \leq b \leq \infty\}$$

is a semialgebra.

Lemma 1.20. Let $\mathcal{S}$ be a semialgebra and $\overline{\mathcal{S}} = \mathcal{A}(\mathcal{S})$ = algebra generated by $\mathcal{S}$. Then $\overline{\mathcal{S}} = \{\text{finite disjoint union of sets from } \mathcal{S}\}$.

Proof. Let $\mathcal{C} = \{\text{finite disjoint union of sets from } \mathcal{S}\}$. Clearly, $\overline{\mathcal{S}} \supseteq \mathcal{C}$. Now we only need to check $\mathcal{C}$ is an algebra. It is easy to see that $\Omega \in \mathcal{C}$.

Claim: $\mathcal{C}$ is closed under finite intersection.
Let $A, B \in \mathcal{C}$, and $A = \cup_{i=1}^{m} A_i, B = \cup_{j=1}^{n} B_j, A_i, B_j \in \mathcal{S}$. Then $A \cap B = \cup_{i,j} (A_i \cap B_j) \in \mathcal{C}$.

Claim: $\mathcal{C}$ is closed under complement.
Let $A = \cup_{i=1}^{n} A_i, A_i \in \mathcal{S}$. Then $A^c = \cap_{i=1}^{n} A_i^c, A_i^c \in \mathcal{C}$.$\mathcal{C}$ is an algebra by previous claim, we have $A^c \in \mathcal{C}$.

Hence, $\mathcal{C}$ is an algebra.

Let $\mu$ be a set function on $\mathcal{S}$. We can extend $\mu$ to $\overline{\mu}$ on $\overline{\mathcal{S}}$ by $\overline{\mu}(A) = \sum_{i=1}^{n} \mu(A_i)$ where $A = \cup_{i=1}^{n} A_i$. 

Theorem 1.21. Let $\mathcal{S}$ be a semi-algebra and $\mu$ be a set function on $\mathcal{S}$ such that,

(i) (positive) $\mu(A) \geq 0$ for $A \in \mathcal{S}$,

(ii) (finitely additive) If $A = \bigcup_{i=1}^{n} A_i$, $A, A_i \in \mathcal{S}$, then $\mu(A) = \sum_{i=1}^{n} \mu(A_i)$.

(iii) (countably subadditive) If $A \subseteq \bigcup_{i=1}^{\infty} A_i$, $A, A_i \in \mathcal{S}$, then $\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i)$.

Then $\mu$ can be extended uniquely to a measure $\overline{\mu}$ on $\mathcal{S}$ satisfying (i), (ii), (iii). Moreover, if $\overline{\mu}$ is $\sigma$-finite, then $\overline{\mu}$ has a unique extension to a measure $\tilde{\mu}$ on $\sigma(\mathcal{S})$.

The proof can be found in the appendix of the textbook.
2.1 Measures on Real Line

Definition 2.1. A Stieltjes measure function is a function \( F : \mathbb{R} \rightarrow \mathbb{R} \) if

(i) \( F \) is non-decreasing,

(ii) \( F \) is right continuous, i.e., if \( x_n \downarrow x \), \( F(x_n) \rightarrow F(x) \).

We can define \( \mu_F((a, b]) = F(b) - F(a) \).

Application: Take a nondecreasing and right-continuous function \( F \) on \( \mathbb{R} \). Let \( S \) be the semi-algebra \( \{ (a, b] \mid -\infty \leq a \leq b \leq \infty \} \) and \( \mu((a, b]) = F(b) - F(a), -\infty < a < b < \infty \). When \( a = -\infty \) or \( b = \infty \), we take \( F(\infty) = \lim_{x \uparrow \infty} F(x) \), \( F(-\infty) = \lim_{x \downarrow -\infty} F(x) \).

(i) \( \overline{\mu}((a, b]) = F(b) - F(a) \geq 0 \),

(ii) If \( (a, b] = \bigcup_{i=1}^{n} (a_i, b_i] \), order \( a_i \) in increasing order. We claim that \( a_i = a, b_i = a_i, \cdots, b_n = b \). Then \( \mu((a, b]) = F(b) - F(a) = \sum_{i=1}^{n}(F(b_i) - F(a_i)) = \sum_{i=1}^{n} \mu((a_i, b_i]) \).

(iii) If \( (a, b] \subseteq \bigcup_{i=1}^{\infty} (a_i, b_i] \), take \( \varepsilon > 0 \). Using right-continuity of \( F \), change \( b_i \rightarrow b_i + \delta_i \) where \( \delta_i > 0, F(b_i + \delta_i) - F(b_i) \leq \frac{\varepsilon}{2^n} \), and change \( a \rightarrow a + \delta \) where \( \delta > 0, F(a + \delta) - F(a) \leq \varepsilon \). Then \( [a + \delta, b] \subseteq \bigcup_{i=1}^{\infty} (a_i, b_i + \delta_i) \). Extract a finite subcover \( [a + \delta, b] \subseteq \bigcup_{i=1}^{n} (a_{j_i}, b_{j_i} + \delta_{j_i}) \). Show that

\[
F(b) - F(a + \delta) \leq \sum_{i=1}^{n}(F(b_{j_i} + \delta_{j_i}) - F(a_{j_i})) \leq \sum_{i=1}^{\infty}(F(b_i + \delta_i) - F(a_i)).
\]

This implies \( F(b) - F(a) \leq 2\varepsilon + \sum_{i=1}^{\infty}(F(b_i) - F(a_i)) \). Since \( \varepsilon \) is arbitrary, we are done.

As an example, let \( \Omega = \mathbb{R} \), \( \mathcal{F} = \mathcal{B}(\mathbb{R}) \) and \( \mu \) be a probability measure. Define

\[
F(x) = \mu((\infty, x]), x \in \mathbb{R}.
\]

Check that \( F \) is non-decreasing and right-continuous, \( \mu((a, b]) = F(b) - F(a) \) and \( \mu(\mathbb{R}) = F(\infty) - F(-\infty) \).

Example 2.2. Let \( \Omega = \mathbb{R}^2 \) and \( \mathcal{S} = \{(a, b] \times (c, d] \mid -\infty \leq a \leq b \leq \infty, -\infty \leq c \leq d \leq \infty \} \). Show that \( \mathcal{S} \) is a semi-algebra. Find the properties needed for a function \( F : \mathbb{R}^2 \rightarrow (-\infty, \infty) \) to define a measure \( \mu_F \) on \( \mathcal{B}^2 \) such that \( \mu_F((a, b] \times (c, d]) = F(b, d) - F(a, d) - F(b, c) + F(a, c) \) for all \( a \leq b, c \leq d \).

Given a measure space \((\Omega, \mathcal{F})\) and a measure \( \mu \) on it, we define the completion of \( \mathcal{F} \) w.r.t. \( \mu \) as

\[
\bar{\mathcal{F}} := \{ A \cup B \mid A \in \mathcal{F}, B \subseteq N \in \mathcal{F} \text{ such that } \mu(N) = 0 \}.
\]

We also extend \( \mu \) to \( \bar{\mu} \) on \( \bar{\mathcal{F}} \) by \( \bar{\mu}(A \cup B) = \mu(A) \) where \( B \subseteq N \in \mathcal{F} \) with \( \mu(N) = 0 \).

Example 2.3. Show that \( \bar{\mathcal{F}} \) is a \( \sigma \)-field. Moreover, \( \bar{\mu} \) is well-defined and is a measure on \((\Omega, \bar{\mathcal{F}})\).
2.2 Random Variables

**Definition 2.4.** Given two measure spaces $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$, a function $f : \Omega \to \mathcal{S}$ is measurable if $f^{-1}(A) \in \mathcal{F}$ for any $A \in \mathcal{S}$.

**Definition 2.5.** A random variable is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel $\sigma$-field of $\mathbb{R}$. A random vector is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathcal{B}^d)$, where $\mathcal{B}^d$ is the Borel $\sigma$-field of $\mathbb{R}^d$.

Notation: We will use $X, Y, Z$ as r.v.s on $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$.

**Proposition 2.6.** (i) Composition of two measurable functions is measurable.

(ii) Let $\mathcal{C}$ be a collection of sets from $\mathcal{S}$ and $\sigma(\mathcal{C}) = \mathcal{S}$. If $f^{-1}(A) \in \mathcal{F}$ for any $A \in \mathcal{C}$, then $f$ is measurable.

(Hint: $\{A \mid f^{-1}(A) \in \mathcal{F}\}$ is a $\sigma$-field.)

**Example 2.7.** Take $\mathcal{B} = \sigma((-\infty, a] \mid a \in \mathbb{R})$. For $X : \Omega \to \mathbb{R}$, it is enough to check $X^{-1}((-\infty, a]) = \{\omega : X(\omega) \leq a\} = \{X \leq a\} \in \mathcal{F}$.

Notation: $\{X \in A\} = X^{-1}(A) = \{\omega : X(\omega) \in A\}$.

**Example 2.8.** $X : \Omega \to \mathbb{R}^d, \mathcal{B}^d = \sigma(\prod_{i=1}^{d} (-\infty, a_i] : (a_1, a_2, \ldots, a_d) \in \mathbb{R}^d)$.

**Definition 2.9.** (i) If $f : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ is measurable, then $f^{-1}(\mathcal{S}) = \{f^{-1}(A) : A \in \mathcal{S}\}$ is also a $\sigma$-field. This is called the $\sigma$-field generated by $f$.

(ii) If also, $(\Omega, \mathcal{F}, \mu)$ is a measurable space, then we can define $\gamma = \mu \circ X^{-1}$ a measure on $(S, \mathcal{S})$. Then we have $\gamma(A) = \mu(X^{-1}(A)) = \mu(x \in A), A \in \mathcal{S}$.

In particular, For random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}), \mu(A) := \mathbb{P}(x \in A)$ where $A \in \mathcal{B}$ is a probability on $(\mathbb{R}, \mathcal{B})$.

2.3 Distributions

**Definition 2.10.** The distribution function of $X$ is defined as $F(x) = \mathbb{P}(X \leq x)$ for $x \in \mathbb{R}$. Then we have $F(a) - F(b) = \mathbb{P}(X \in (b, a])$.

**Proposition 2.11.** (i) $F$ is non-decreasing.

(ii) $F$ is right-continuous.

(iii) $F(x) \downarrow 0$ as $x \downarrow -\infty, F(x) \uparrow 1$ as $x \uparrow \infty$.

(iv) $\lim_{x_n \to x^-} F(x_n)$ exists and $\lim_{x_n \to x^-} F(x_n) = \mathbb{P}(X < x)$.

**Lemma 2.12.** Given any $F$ satisfying (i)-(iii), there exists a r.v. $X$ defined on $((0,1), \mathcal{B}, \mathbb{P} = \lambda)$ with distribution function $F$. 

Proof. Define $X : ((0,1), \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ as

$$X(\omega) = \sup\{y : F(y) < \omega\}.$$ 

We need to check that i) $X$ is measurable and ii) $X$ has distribution function $F$.

We claim that, $\{\omega : X(\omega) \leq a\} = \{\omega : \omega \leq F(a)\}$. Indeed, If $\omega \leq F(a)$, then for any $y$ which satisfies $F(y) < \omega$ also satisfies $y < a$ and thus, $X(\omega) \leq a$. On the other hand, if $\omega > F(a)$, then $\exists \varepsilon > 0$ s.t. $\omega > F(a + \varepsilon)$ by right continuity of $F$. Thus, by definition of $X$, $X(\omega) \geq a + \varepsilon > a$.

In fact, $X$ behaves like the inverse function of $F$. We define $F^{-1}(x) := \sup\{y : F(y) < x\}$, $x \in (0,1)$.

Besides, $P\{\omega : X(\omega) \leq a\} = P\{\omega : \omega \leq F(a)\} = F(a)$ since $\omega \in (0,1)$. So $F$ is distribution function of $X$. \qed

Remark: If $U$ is a uniform distributed r.v. on $(0,1)$, i.e.,

$$P(U \leq x) = \begin{cases} x & 0 < x < 1 \\ 0 & x \leq 0 \\ 1 & x > 1 \end{cases}$$

then, $X = F^{-1}(U)$.

### 2.4 Properties of Random Variables

**Definition 2.13.** (i) Two random variables $X, Y$ are equal in distribution if $X$ and $Y$ has the same distribution function.

(ii) Two random variables $X, Y$ are equal a.s. if they are defined on the same measurable space $(\Omega, \mathcal{F}, P)$ and $P(X \neq Y) = 0$.

**Example 2.14.** (i) Uniform$(0,1)$ distribution with $F(a) = \int_{-\infty}^a f(x)dx$ and

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

(ii) Exponential$(\lambda)$ with distribution function

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \text{ with density } f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

(iii) The distribution for the measure $\mu\{i\} = p_i$, $i = 1, 2, 3, \ldots$ where $\sum_{i=1}^{\infty} p_i = 1$ is a step function.

(iv) Check that Cantor function gives a distribution function supported on the Cantor set.

**Proposition 2.15.** (i) Any continuous function is measurable.

(ii) If $f$ is a continuous function and $(X_1, \ldots, X_n)$ is a random vector, then $f(X_1, X_2, \ldots, X_n)$ is measurable.
(iii) \( X_i : (\Omega_i, \mathcal{F}_i) \to (\mathbb{R}, \mathcal{B}), i = 1, \ldots, n \) are r.v.'s, then \((X_1, \ldots, X_n) : (\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n) \to (\mathbb{R}^n, \mathcal{B}^n)\) is also measurable where we write \( \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n = \sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n) \).

(iv) If \( X_n \) is a sequence of r.v.'s, then \( \inf_{n \geq 1} X_n, \sup_{n \geq 1} X_n, \lim \inf X_n, \lim \sup X_n \) are r.v.'s.

Proof. Proof of (i) follows from the fact that for any open set \( A, f^{-1}(A) \) is open and hence measurable when \( f \) is continuous. Part (ii) follows from composition of two measurable functions are measurable. For (iii), note that \( \{ (X_1, \ldots, X_n) \in \prod_{i=1}^n (-\infty, a_i] \} = \prod_{i=1}^n \{ X_i \leq a_i \} \). Finally, (iv) follows from the fact that

\[
\{ \inf_{n \geq 1} X_n \geq a \} = \bigcap_{n \geq 1} \{ X_n \geq a \}, \{ \sup_{n \geq 1} X_n \leq a \} = \bigcap_{n \geq 1} \{ X_n \leq a \}
\]

and \( \mathcal{B} = \sigma([a, \infty) : a \in \mathbb{R}) = \sigma((\infty, a) : a \in \mathbb{R}) \). Finally note that,

\[
\lim \inf X_n = \sup \inf_{n \geq 1} X_m, \quad \lim \sup X_n = \inf \sup_{n \geq 1} X_m.
\]

Definition 2.16. (i) A sequence of r.v.’s \((X_n)\) is said to converge a.s. if \( \mathbb{P}(\lim X_n \text{ exists}) = 1 \).

(ii) We say that \((X_n)\) converges in probability to \( X \) if \( \lim_{n \to \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0 \) for any \( \varepsilon > 0 \).

(iii) We say that \( X_n \) converges in distribution to \( X \) if \( F_n \to F \) pointwise at all continuity points of \( F \).

Example 2.17. Indicator random variable.

\[
\mathbb{1}_A(\omega) = \begin{cases} 
1 & \text{if } \omega \in A \\
0 & \text{o.w.}
\end{cases}
\]
3.1 Expectation

Proposition 3.1. **Expectation** There exists a positive linear functional $E$ on the space of non-negative random variables on $(\Omega, \mathcal{F}, P)$ such that

(a) $E(1_A) = P(A)$ for $A \in \mathcal{F}$

(b) $E(aX) = aE(X) \forall a > 0$.

(c) $E(X + Y) = E(X) + E(Y)$.

(d) $X \geq Y \Rightarrow E(X) \geq E(Y)$.

Definition 3.2. A random variable $X$ is **integrable** if $E|X| < \infty$.

We define $X^+ := X \lor 0 = \max(X, 0)$, $X^- := (-X) \lor 0 = \max(-X, 0)$, then $X = X^+ - X^-.$

Definition 3.3. We define the expectation of an integrable random variable $X$ to be $E(X) = E(X^+) - E(X^-)$.

**Four-Step Procedure to Define Expectation:**

Step 1: **Indicator and Simple rvs:** Let $E(1_A) := P(A)$ for $A \in \mathcal{F}$. We define,

Definition 3.4 (Simple random variable). A r.v. $\phi$ is a simple RV if

$$
\phi(\omega) = \sum_{i=1}^{n} a_i 1_{A_i}(\omega),
$$

where $A_i$’s are disjoint measurable sets.

We define the expectation of a simple rv, $\phi$, as

$$
E(\phi) := \sum_{i=1}^{n} a_i P(A_i).
$$

Exercise 3.5. Check that $E$ is well-defined for simple rv’s, i.e., $E(\phi)$ is the same for different representations of $\phi$. Use the fact that

$$
\phi = \sum_{i=1}^{n} a_i 1_{A_i} = \sum_{j=1}^{m} b_j 1_{B_j} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i 1_{A_i \cap B_j}.
$$

Moreover, $E$ satisfies the conclusions in Proposition 3.1 for simple rvs.
Step 2: **Non-negative bounded rvs.** For a non-negative bounded rv $0 \leq X \leq M$, we define

$$E(X) := \sup_{\phi \in X, \phi \text{ simple}} E(\phi) = \inf_{\psi \geq X, \psi \text{ simple}} E(\psi).$$

First we check that the supremum and infimum are same, so that $E$ is well-defined. We divide the interval $[0, M]$ into multiple intervals, $J_0, J_1, \ldots, J_n$, where

$$J_i = \left( \frac{iM}{n}, \frac{(i+1)M}{n} \right], \quad i = 0, 1, \ldots, n.$$ 

We define

$$\phi_n = \sum_{i=0}^{n} \frac{iM}{n} \mathbb{1}_{X^{-1}(J_i)} \text{ and } \psi_n = \sum_{i=0}^{n} \frac{(i+1)M}{n} \mathbb{1}_{X^{-1}(J_i)}.$$ 

Clearly, $\phi_n \leq X \leq \psi_n \leq \phi_n + 1/n$. From here we have

$$\inf_{\psi \geq X, \psi \text{ simple}} E(\psi) \leq E(\psi_n) \leq E(\phi_n) + 1/n \leq \sup_{\phi \leq X, \phi \text{ simple}} E(\phi) + 1/n.$$ 

Since $n$ is arbitrary, we have

$$\sup_{\phi \leq X, \phi \text{ simple}} E(\phi) \geq \inf_{\psi \geq X, \psi \text{ simple}} E(\psi).$$ 

The other direction is easy, as $\phi \leq X \leq \psi$ implies $E(\phi) \leq E(\psi)$.

**Exercise 3.6.** $E$ satisfies the conclusions in Proposition 3.1 for non-negative bounded rvs.

Step 3: **Non-negative rvs.** For a non-negative rv $X \geq 0$, we define

$$E(X) := \sup \{E(h) | 0 \leq h \leq X, h \text{ bounded} \}.$$ 

**Lemma 3.7.** If $X \geq 0$ then $E(X \wedge n) \uparrow E(X)$ as $n \to \infty$.

**Proof.** Clearly, $E(X \wedge n)$ is non-decreasing in $n$. Take a bounded nonnegative rv $h$ such that $0 \leq h \leq X$. Let $h \leq M$ for some $M < \infty$. For $n \geq M$ we have $h \leq X \wedge n$ and thus, by definition of expectation we have

$$E(h) \leq E(X \wedge n) \text{ for all } n \geq M.$$ 

Hence, $E(h) \leq \lim_{n \to \infty} E(X \wedge n)$. Since $h$ is arbitrary, we have

$$E(X) \leq \lim_{n \to \infty} E(X \wedge n).$$ 

However, $E(X) \geq \lim_{n \to \infty} E(X \wedge n)$ as $X \wedge n$ is bounded for all $n$. Thus equality holds.

**Corollary 3.8.** The lemma above implies that $E(X + Y) = E(X) + E(Y)$.

**Proof.** To show that $E(X + Y) \leq E(X + Y)$, we note that for $0 \leq h \leq X, 0 \leq g \leq Y$, where $h, g$ are bounded rvs, we have $0 \leq h + g \leq X + Y$ and $E(h) + E(g) = E(h + g) \leq E(X + Y)$. Thus,

$$E(X) + E(Y) \leq E(X + Y).$$ 

To show the other direction, we know that $(X + Y) \wedge n \leq X \wedge n + Y \wedge n$ for all $n$. Thus $E((X + Y) \wedge n) \leq E(X \wedge n) + E(Y \wedge n)$. Taking $n \to \infty$ and using Lemma 3.7 we have

$$E(X + Y) \leq E(X) + E(Y)$$ 

d and hence the equality holds.
Exercise 3.9. \( E \) satisfies the conclusions in Proposition 3.1 for non-negative rvs.

Step 4: **Integrable rvs satisfying** \( E|X| < \infty \). For a rv \( X \) with \( E(|X|) < \infty \), we define

\[
E (X) = E(X^+) - E(X^-)
\]

where \( X^+ = X \vee 0 \), \( X^- = (-X) \vee 0 \).

**Exercise 3.10.** For any integrable rv \( E \) satisfies the following:

(i) if \( X \geq 0 \) a.s., then \( E X \geq 0 \),

(ii) for all \( a \in \mathbb{R} \), \( E(aX) = a E X \),

(iii) \( E(X + Y) = E X + E Y \),

(iv) if \( X \leq Y \) a.s., then \( E X \leq E Y \),

(v) if \( X = Y \) a.s., then \( E X = E Y \),

(vi) \( |E X| \leq E |X| \).

**Remarks:**

1. \( \{X: (\Omega, F, P) \to (\mathbb{R}, B) \mid E |X| < \infty \} = L^1(\Omega, F, P) \) is a Banach space.

2. \( \{X: (\Omega, F, P) \to (\mathbb{R}, B) \mid E |X|^p < \infty \} = L^p(\Omega, F, P) \), where \( \|X\|_p = E(|X|^p)^{1/p}, 1 \leq p \leq \infty \), is a Banach space. Here essential supremum:

\[
\|X\|_\infty = \text{ess sup}|X| = \inf\{k \geq 0 \mid P(|X| > k) = 0\}.
\]

3. \( L^2(\Omega, F, P) \) is a Hilbert space with the inner product \( \langle X, Y \rangle = E(XY) \).

**Exercise 3.11.** Show that if \( \|X\|_\infty < \infty \), then \( \|X\|_p \uparrow \|X\|_\infty \) as \( p \uparrow \infty \).

**Definition 3.12 (L^p Convergence).** We say that \( X_n \) converges in \( L^p \) to \( X \), denoted \( X_n \xrightarrow{L^p} X \), if \( X_n, X \in L^p \) and

\[
\|X_n - X\|_p \xrightarrow{n \to \infty} 0.
\]

**Lemma 3.13 (Bounded Convergence Thm (BCT)).** If \( 0 \leq X_n \leq M \) and \( X_n \rightarrow X \) in probability, then \( X_n \xrightarrow{L^1} X \) and

\[
E X_n \rightarrow E X.
\]

**Proof.** For \( \forall \varepsilon > 0 \), we have

\[
|E X_n - E X| = |E(X_n - X)| \leq E|X_n - X|
= E|X_n - X| \cdot \mathbb{1}_{|X_n - X| > \varepsilon} + E|X_n - X| \cdot \mathbb{1}_{|X_n - X| \leq \varepsilon}
\leq 2M \cdot P(|X_n - X| > \varepsilon) + \varepsilon.
\]

Now, \( P(|X_n - X| > \varepsilon) \to 0 \) as \( n \to \infty \). Thus \( \limsup_{n \to \infty} |E X_n - E X| \leq \varepsilon \). Since \( \varepsilon > 0 \) is arbitrary we have the result.
Lemma 3.14 (Fatou’s Lemma). If \( X_n \geq 0 \), then

\[
\liminf_n E(X_n) \geq E(\liminf_n X_n).
\]

Example 3.15. Let \( X_n = n \cdot 1_{(0, 1/n]} \). Then \( E(X_n) = n \cdot \frac{1}{n} = 1 \), \( \liminf_n X_n = 0 \), so \( E(\liminf_n X_n) = 0 \). Thus \( \liminf_n E(X_n) > E(\liminf_n X_n) \).

Proof of Fatou’s Lemma. Let \( Y_n := \inf_{m \geq n} X_m \). Then \( X_n \geq Y_n \) and \( \liminf_n X_n = \sup_n Y_n \). By monotonicity, we have

\[
\liminf_n E(X_n) \geq \lim inf \ E(Y_n).
\]

Thus it suffices to show that

\[
\liminf_n E(Y_n) \geq E(\sup_n Y_n).
\]

For all \( M > 0 \), \( Y_n \uparrow \sup Y_n \wedge M \). By BCT,

\[
E(Y_n \wedge M) \to E(\sup_n Y_n \wedge M).
\]

So

\[
\liminf_n E(Y_n) \geq \liminf_n E(Y_n \wedge M) = E(\sup_n Y_n \wedge M).
\]

By taking \( M \to \infty \), \( \liminf_n E(Y_n) \geq E(\sup_n Y_n) \). This completes the proof.

3.2 Properties of Expectation

We will use Fatou’s lemma to prove Monotone Convergence Theorem and Dominated Convergence Theorem.

Theorem 3.16 (Monotone Convergence Thm (MCT)). If \( X_n \geq 0 \) for all \( n \), and \( X_n \uparrow X \), then \( E(X_n) \uparrow E(X) \).

Proof. By Fatou’s Lemma,

\[
\liminf_n E(X_n) \geq E(\liminf_n X_n) = E(X).
\]

However, because \( X_n \uparrow X \), we have \( \liminf_n E(X_n) \leq E(X) \). Thus \( \liminf_n E(X_n) = E(X) \).

Theorem 3.17 (Dominated Convergence Thm (DCT)). If \( X_n \to X \) and \( 0 \leq |X_n| \leq Y \), with \( E(Y) < \infty \), then \( E(X_n) \to E(X) \).

Proof. Since \( X_n \to X \) and \( 0 \leq |X_n| \leq Y \), \( X_n + Y \geq 0 \) and converges to \( X + Y \) as \( n \to \infty \). By Fatou’s Lemma, we have

\[
\liminf_n E(X_n + Y) \geq E(X + Y)
\]

which implies that

\[
\liminf_n E(X_n) + E(Y) \geq E(X) + E(Y) \text{ and } \liminf_n E(X_n) \geq E(X).
\]
Similarly, \(-X_n + Y \geq 0\) and converges to \(-X + Y\), so
\[
\liminf_n -E(X_n) \geq -E(X) \implies \limsup_n E(X_n) \leq E(X).
\]

Thus we have \(E(X_n) \to E(X)\). \(\blacksquare\)

**Theorem 3.18 (Change of variables formula).** Let \(X : \Omega \to \mathbb{R}\) be a random variable. For any random variable \(f : \mathbb{R} \to \mathbb{R}\) such that \(E|f(X)| < \infty\),
\[
\int f(X) dP = \int f dP_X.
\]

**Sketch of the proof:** We can verify this using the 4-step procedure, checking:
1. Indicator rvs: Let \(f = 1_A\). Then \(\int f(X) dP = P(X \in A) = P_X(A) = \int f dP_X\) and simple rvs.
2. Non-negative bounded functions.
4. Any measurable functions for which both sides make sense. \(\blacksquare\)

### 3.3 Remarks about Expectation

**Exercise 3.19.** Let \(\Omega = \{1, 2, \ldots\}\), \(\mathcal{F} = 2^\Omega\). Let \(p_i \geq 0\) for all \(i \in \mathbb{N}\) and suppose \(\sum_{i=1}^{\infty} p_i = 1\). Define \(P(A) = \sum_{i \in A} p_i\) for all \(A \in \mathcal{F}\). Then for any random variable \(X : \Omega \to \mathbb{R}\), check that
\[
E_X = \sum_{i=1}^{\infty} X(i) p_i.
\]
In particular, if \(\Omega = \{1, \ldots, N\}\) is finite, and \(p_i = 1/N\) for \(i = 1, 2, \ldots, N\), then
\[
E_X = \frac{1}{N} \sum_{i=1}^{N} X(i).
\]

**Exercise 3.20.** Consider the probability space \(((0,1), \mathcal{B}(0,1), P = \text{Lebesgue measure})\). For any random variable \(X : (0,1) \to \mathbb{R}\), check that
\[
E(X) = \int_{0}^{1} X(x) dx.
\]

**Exercise 3.21.** Consider the probability space \((\Omega = \mathbb{R}, \mathcal{B}, P)\), where \(P\) has distribution function \(F\). Check that
\[
E(X) = \int_{-\infty}^{\infty} X(x) dF(x).
\]

In particular, if \(F\) has density \(f(x)\), i.e. \(F(x) = \int_{-\infty}^{x} f(x) dx\), and \(F\) is differentiable a.e., then
\[
E(X) = \int_{-\infty}^{\infty} X(x) f(x) dx.
\]

Not all distribution functions have densities. A measure is called **absolutely continuous** with respect to Lebesgue measure if it has a density and a measure is called **countable** if there exists a countable set with probability one.
3.4 Inequalities

Theorem 3.22 (Jensen’s Inequality). If $\phi$ is a convex function on the range of a random variable $X$, then
\[ \phi(E X) \leq E \phi(X) \quad (3.1) \]
assuming both sides exist.

Proof Sketch: A function is convex if it is the supremum of a set of linear functions (think of tangent lines), i.e., there exists a collection of linear functions $A = \{a_\ell x + b_\ell \mid \ell \in \Lambda\}$ such that $\phi(x) = \sup_{\ell \in \Lambda} \{a_\ell x + b_\ell\}$. Clearly $\phi(x) \geq a_\ell x + b_\ell$ for all $\ell \in \Lambda$. By monotonicity and linearity of $E$,
\[ E \phi(X) \geq a_\ell E X + b_\ell \implies E \phi(X) \geq \sup_{\ell \in \Lambda} \{a_\ell E X + b_\ell\} = \phi(E X). \]

Corollary 3.23. Let $1 \leq p < q < \infty$. For any random variable $X$, $\|X\|_p \leq \|X\|_q$, i.e.,
\[ \left( E |X|^p \right)^{\frac{1}{p}} \leq \left( E |X|^q \right)^{\frac{1}{q}}. \quad (3.2) \]

Proof Sketch: Define $\phi(x) = |x|^{q/p}$. Then $\phi(x)$ is convex since $q/p > 1$. Take $Y = |X|^p$. By Jensen’s inequality
\[ \left( E |X|^p \right)^{\frac{q}{p}} = E(\phi(Y)) \leq E(\phi(Y)) = E |X|^q. \]

Exercise 3.24. Let $X$ be a random variable and suppose $1 \leq p < \infty$. Prove that $\|X\|_p \leq \|X\|_\infty$.

Theorem 3.25 (Markov’s Inequality). Let $X$ be a non-negative random variable. For $a > 0$, we have
\[ P(X \geq a) \leq \frac{E X}{a} \]

Proof Sketch: Note that $1_{X \geq a} \leq X/a$. So
\[ P(X \geq a) = E(1_{X \geq a}) \leq E \left( \frac{X}{a} \right) = \frac{E X}{a}, \]
here we used monotonicity of expectation in the inequality.

Theorem 3.26 (Chebyshev’s Inequality). Let $\phi : \mathbb{R} \to \mathbb{R}$ be a nondecreasing, non-negative function. Then for any $a \geq 0$ where $\phi(a) \neq 0$,
\[ P(X \geq a) \leq \frac{E \phi(X)}{\phi(a)}. \]

Proof Sketch: By Markov’s inequality,
\[ P(X \geq a) \leq P \left( \phi(X) \geq \phi(a) \right) \leq \frac{E \phi(X)}{\phi(a)}. \]
**Definition 3.27.** Let $X$ be a random variable. The \textbf{variance} of $X$ is

$$\text{Var}(X) := \mathbb{E}(X - \mathbb{E}X)^2$$

whenever $\mathbb{E}X^2 < \infty$.

Note that $\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 \geq 0$. The original Chebyshev’s inequality is the following:

**Theorem 3.28 (Original Chebyshev Inequality).** Let $X$ be a random variable with finite variance. For any $a > 0$,

$$P(|X - \mathbb{E}X| > a) \leq \frac{\text{Var}(X)}{a^2}.$$
Product Spaces, Independence and Fubini’s Theorem

4.1 Product Spaces and Independence

Let \((\Omega_1, F_1, P_1)\) and \((\Omega_2, F_2, P_2)\) be two probability spaces. Let \(\Omega = \Omega_1 \times \Omega_2\) and \(F = F_1 \otimes F_2 = \sigma(F_1 \times F_2)\).

**Question:** Can we define a probability on \((\Omega, F)\) that is naturally a product of \(P_1\) and \(P_2\)?

The answer is yes!

**Theorem 4.1.** There exists a unique probability measure \(P\) on \((\Omega, F)\) such that for all \(A_1 \in F_1\) and \(A_2 \in F_2\),

\[
P(A_1 \times A_2) = P(A_1) P(A_2).
\]

Note that this implies \(P(A_1 \times \Omega_2) = P_1(A_1)\) and \(P(\Omega_1 \times A_2) = P_2(A_2)\).

**Proof Sketch:** First we note that \(F_1 \times F_2\) is a semi-algebra as \((A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D), (A \times B)^c = (A \times B^c) \cup (A^c \times B) \cup (A^c \times B^c)\). Define \(P\) on \(F_1 \times F_2\) by \(P(A_1 \times A_2) = P(A_1) P(A_2)\).

To prove existence and uniqueness, it suffices to check finite additivity and countable sub-additivity (by Theorem 1.1.4 in Durrett). We will prove countable additivity. Suppose \(A \times B = \cup_{i=1}^{\infty} A_i \times B_i\) is a countable disjoint union of rectangles (where \(A_i \in F_1\) and \(B_i \in F_2\)). We have \(1_A(x)1_B(y) = \sum_{i=1}^{\infty} 1_{A_i}(x)1_{B_i}(y)\) and thus

\[
1_A(x)P(B) = \sum_{i=1}^{\infty} 1_{A_i}(x)P(B_i).
\]

Taking expectation \(E_1\) w.r.t. \(x\) and using MCT we have

\[
P(A \times B) = P(A) P_2(B) = \sum_{i=1}^{\infty} P_1(A_i) P_2(B_i).
\]

So \(P\), originally defined on \(F_1 \times F_2\), can be extended uniquely to the whole \(\sigma\)-field \(\sigma(F_1 \times F_2)\). ■

Next we define the notion of independence, a property of \(\sigma\)-fields.

**Definition 4.2.** Let \(F_1\) and \(F_2\) be two sub \(\sigma\)-fields on a probability space \((\Omega, F, P)\). We say that \(F_1\) and \(F_2\) are \(P\)-**independent** if

\[
P(A_1 \cap A_2) = P(A_1) P(A_2) \text{ for all } A_1 \in F_1 \text{ and } A_2 \in F_2.
\]

**Example 4.3.** Let \((\Omega_1, F_1, P_1)\) and \((\Omega_2, F_2, P_2)\) be probability spaces. Let \(\Omega = \Omega_1 \times \Omega_2\), \(F = F_1 \otimes F_2\) and \(P = P_1 \otimes P_2\) be the product probability measure. Then the \(\sigma\)-fields \(F_1 \times F_2\) and \(\Omega_1 \times F_2\) are independent.
Definition 4.4. Two random variables $X_1$ and $X_2$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if the $\sigma$-fields $\sigma(X_1)$ and $\sigma(X_2)$ are independent, i.e.,

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2) = \mathbb{P}(X_1 \in A_1) \mathbb{P}(X_2 \in A_2)$$

(4.1)

for all $A_1, A_2 \in \mathcal{B}$.

Lemma 4.5. Two random variables $X_1$ and $X_2$ are independent iff $\mathbb{P}(X_1 \leq a_1, X_2 \leq a_2) = \mathbb{P}(X_1 \leq a) \mathbb{P}(X_2 \leq a_2)$ for all $a_1, a_2 \in \mathbb{R}$.

Proof Sketch: The forward direction is clear. The other direction follows from the observation that $\{(b_1, a_1) \times (b_2, a_2) : b_1 < a_1, b_2 < a_2\}$ is a semi-algebra.

Example 4.6. Any $\sigma$-field $\mathcal{F}$ is independent with the trivial $\sigma$-field $\mathcal{F}_{\text{triv}} = \{\emptyset, \Omega\}$.

Example 4.7. Consider the probability space $((0,1), \mathcal{B}, \mathbb{P})$ where $\mathbb{P}$ is Lebesgue measure. Let $X_1 = 1_{(1/2,1)}$ and $X_2 = 1_{(1/4,1/2) \cup (3/4,1)}$. Then $X_1$ and $X_2$ are independent.

Example 4.8. For any $\omega \in (0,1)$, let $D_i(\omega) = [2^i \omega] \mod 2 \in \{0,1\}$ for $i \geq 1$. Thus $\omega = \sum_{i \geq 1} 2^{-i} D_i(\omega)$ is the binary representation of $\omega$. Let $U_1 = \sum_{i \geq 1} 2^{-i} D_{2i-1}$ and $U_2 = \sum_{i \geq 1} 2^{-i} D_{2i}$. Then $U_1$ and $U_2$ are independent, and both have Uniform$(0,1)$ distributions.

Exercise 4.9. If $X, Y$ are independent, then for any measurable functions $\phi$ and $\psi$, $\phi(X)$ and $\psi(Y)$ are also independent.

Definition 4.10. Two events $A_1, A_2$ are independent if $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) \mathbb{P}(A_2)$. In this case we write $A_1 \perp A_2$.

Claim 4.11. $A_1, A_2$ are independent events if and only if $1_{A_1}, 1_{A_2}$ are independent random variables.

Using induction, we can take product of finitely many probability spaces. However, for infinite product we need some extra tool.

Theorem 4.12 (Kolmogorov Consistency Theorem). Let $\mu_1, \mu_2, \ldots$ be probability measures on $\mathbb{R}, \mathbb{R}^2, \ldots$, respectively, which are consistent, meaning $\mu_{n+1}(A \times \mathbb{R}) = \mu_n(A)$. Then there exists a unique probability measure $\mu$ on $\mathbb{R}^n$ such that

$$\mu(A \times \mathbb{R}^n) = \mu_n(A) \text{ for all } A \in \mathcal{B}^n, n \in \mathbb{N}.$$ 

Corollary 4.13. For each probability measure $\mu'$ on $\mathbb{R}$, there exist independent random variables $X_1, X_2, \ldots$ such that $X_i \sim \mu'$ for all $i$.

Proof Sketch: Let $\mu_n = \otimes_{i=1}^n \mu'$ and apply Kolmogorov consistency to obtain $\mu$. Let $X_i$ be the projection onto the $i^{\text{th}}$ factor from the probability space $(\mathbb{R}^n, \mathcal{B}^n, \mu)$.

Definition 4.14. Random variables $X_1, X_2, \ldots$ are i.i.d., i.e., independent and identically distributed, if they are independent and equal in distribution, $X_i \overset{d}{=} X_j$ for all $i, j$. 
4.2 Fubini’s theorem

**Theorem 4.15 (Fubini’s Theorem).** Let \((\Omega_1, \mathcal{F}_1, \mathbb{P}_1), (\Omega_2, \mathcal{F}_2, \mathbb{P}_2)\) be two probability spaces \(\Omega := \Omega_1 \times \Omega_2, \mathcal{F} := \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P} := \mathbb{P}_1 \otimes \mathbb{P}_2\). Let \(f : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})\) be a non-negative measurable function. Then

\[
\int_{\Omega} f \, d\mathbb{P} = \int_{\Omega_2} \left( \int_{\Omega_1} f(\omega_1, \omega_2) \, d\mathbb{P}_1(\omega_1) \right) \, d\mathbb{P}_2(\omega_2) = \int_{\Omega_1} \left( \int_{\Omega_2} f(\omega_1, \omega_2) \, d\mathbb{P}_2(\omega_2) \right) \, d\mathbb{P}_1(\omega_1).
\]

The same conclusion holds if \(f\) is integrable w.r.t. \(\mathbb{P}\).

First we require the following two lemmas to make sure the conclusion makes sense.

**Lemma 4.16.** If \(A\) is measurable on \(\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2\), then \(A_x = \{y \mid (x, y) \in A\} \in \mathcal{F}_2\) a.s. \(x\).

As a consequence, if \(f\) is \(\mathcal{F}/\mathcal{B}\) measurable, then a.s. \(\omega_1\) the function \(f(\omega_1, \cdot)\) is \(\mathcal{F}_2/\mathcal{B}\) measurable.

**Lemma 4.17.** If \(f \geq 0\) is \(\mathcal{F}_1 \otimes \mathcal{F}_2\) measurable, then the function

\[
x \mapsto \int_{\Omega_2} f(x, y) \, d\mathbb{P}_2(y)
\]

is \(\mathcal{F}_1/\mathcal{B}\) measurable.

The proof of Fubini’s theorem uses the standard technique: simple random variables \(\to\) non-negative bounded random variables \(\to\) non-negative random variables \(\to\) general integrable random variables. Moreover, it can be generalized to get the following result.

**Lemma 4.18.** Let \(X, Y\) be two independent random variables. Then for any non-negative measurable function \(f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}\),

\[
\mathbb{E} f(X, Y) = \int_{\mathbb{R}} \left( \int_{\mathbb{R}} f(x, y) \, d\mathbb{P}_X(dx) \right) \, d\mathbb{P}_Y(dy) = \int_{\mathbb{R}} \left( \int_{\mathbb{R}} f(x, y) \, d\mathbb{P}_Y(dy) \right) \, d\mathbb{P}_X(dx).
\]

**Proof.** Consider the probability spaces \((\mathbb{R}, \mathcal{B}, \mathbb{P}_X), (\mathbb{R}, \mathcal{B}, \mathbb{P}_Y)\) and the product probability space \((\mathbb{R}^2, \mathcal{B}^2, \mathbb{P}_X \otimes \mathbb{P}_Y)\) and work with the measurable function \(f(x, y)\) on \(\mathbb{R}^2\). The only thing to check is that \(\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y\) when \(X, Y\) are independent and will be left as an exercise. 

**Corollary 4.19.** Let \(X, Y\) be two independent integrable random variables. Then \(\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y\) whenever both sides exist.

**Proof Sketch:** If \(X, Y \geq 0\), just apply Fubini’s theorem. In general, let \(X = X^+ - X^-\) and \(Y = Y^+ - Y^-\). Then \(XY = X^+Y^+ + X^-Y^- - X^+Y^- - X^-Y^+\). So

\[
\mathbb{E}(XY) = \mathbb{E}X^+ \mathbb{E}Y^+ + \mathbb{E}X^- \mathbb{E}Y^- - \mathbb{E}X^+ \mathbb{E}Y^- - \mathbb{E}X^- \mathbb{E}Y^+ = (\mathbb{E}X^+ - \mathbb{E}X^-)(\mathbb{E}Y^+ - \mathbb{E}Y^-) = \mathbb{E} X \cdot \mathbb{E} Y.
\]
Corollary 4.20. If $X, Y$ are independent, integrable random variables with densities $f, g$, then $X + Y$ has density $f \ast g$.

Proof Sketch: For any $t \in \mathbb{R}$, we have

$$P(X + Y \leq t) = E(1_{X+Y \leq t}) = \int \int 1_{x \leq t-y} P_X(dx) P_Y(dy)$$

$$= \int P(X \leq t-y) P_Y(dy) = \int_{-\infty}^{t} f(x-y) dx P_Y(dy)$$

$$= \int_{-\infty}^{t} \int f(x-y) P_Y(dy) dx = \int_{-\infty}^{t} \int R f(x-y) g(y) dy dx = \int_{-\infty}^{t} (f \ast g)(x) dx.$$

Thus $X + Y$ has density $f \ast g$.

4.3 Dynkin’s $\pi$-$\lambda$ Theorem

Definition 4.21. A collection of events $\mathcal{C}$ is a $\pi$-system if it is closed under finite intersections.

Definition 4.22. A collection of events $\mathcal{D}$ is a $\lambda$-system if

(i) $\Omega \in \mathcal{D}$

(ii) $A \in \mathcal{D}$ implies $A^c \in \mathcal{D}$

(iii) $\{ A_i \}_{i=1}^{\infty} \subseteq \mathcal{D}$ and mutually disjoint implies $\bigcup_{i=1}^{\infty} A_i \in \mathcal{D}$.

Claim 4.23. $\mathcal{D}$ is a $\lambda$-system if and only if

(i) $\Omega \in \mathcal{D}$

(ii) $A \subseteq B$ and $A, B \in \mathcal{D}$ implies $B - A \in \mathcal{D}$

(iii) $\{ A_i \}_{i=1}^{\infty} \subseteq \mathcal{D}$ and $A_i \uparrow$ implies $\bigcup_{i=1}^{\infty} A_i \in \mathcal{D}$.

Theorem 4.24 (Dynkin’s $\pi$ – $\lambda$ theorem). If $\mathcal{C}$ is $\pi$-system, $\mathcal{D}$ is a $\lambda$-system, and $\mathcal{C} \subseteq \mathcal{D}$, then $\sigma(\mathcal{C}) \subseteq \mathcal{D}$.

Corollary 4.25. If probability measures $P_1, P_2$ agree on a $\pi$-system $\mathcal{C}$, then they agree on $\sigma(\mathcal{C})$.

Example 4.26. $\mathcal{C} = \{ (-\infty, x] \mid x \in \mathbb{R} \}$ is a $\pi$-system with $\sigma(\mathcal{C}) = \mathcal{B}$. $\mathcal{C} = \{ \prod_{i=1}^{n} (-\infty, x_i] \mid (x_i)_{i=1}^{n} \in \mathbb{R}^n \}$ is a $\pi$-system with $\sigma(\mathcal{C}) = \mathcal{B}^n$.

Lemma 4.27. If $\mathcal{F}_1, \mathcal{F}_2$ are $\sigma$-fields and $\mu_1, \mu_2$ are measures on $\mathcal{F}_1, \mathcal{F}_2$, then for all $A \in \mathcal{F}_1 \otimes \mathcal{F}_2$,

1. For all $x \in \Omega$, $A_x = \{ y \in \Omega : (x, y) \in A \} \in \mathcal{F}_2$

2. $x \mapsto \mu_2(A_x)$ is measurable

Proof sketch: Let $\mathcal{C}$ be the set of rectangular events $\mathcal{F}_1 \times \mathcal{F}_2$. Then

(i) 1 and 2 hold for all $A \in \mathcal{C}$.

(ii) $\sigma(\mathcal{C}) = \mathcal{F}_1 \otimes \mathcal{F}_2$. 
(iii) The set of all events which satisfy 1 and 2 is a $\lambda$-system.

(iv) The $\pi - \lambda$ theorem and (i),(ii), and (iii) imply that $\mathcal{F}_1 \otimes \mathcal{F}_2$ is a subset of all the events which satisfy 1 and 2.

**Definition 4.28.** For $L^2$-integrable random variables $X, Y$, the **covariance** of $X$ and $Y$ is defined as

$$\text{Cov}(X, Y) := E((X - E(X))(Y - E(Y)))$$

and the **correlation** is defined as

$$\text{Corr}(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}.$$ 

**Exercise 4.29.** Show that $L^2(\Omega, \mathcal{F}, P)$ is a Hilbert space, i.e., a vector space that is equipped with an inner product and is complete.

**Definition 4.30.** A sequence of $\sigma$-fields $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ are independent if $P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n)$ for all $A_1 \in \mathcal{F}_1, \ldots, A_n \in \mathcal{F}_n$. Random variables $X_1, \ldots, X_n$ are independent if $\sigma(X_1), \ldots, \sigma(X_n)$ are independent.

**Example 4.31.** Let $\Omega = (0, 1)$, $F = B$, $P = \lambda$, $X = 1_{(\frac{1}{2}, 1)}$, $Y = 1_{(\frac{3}{4}, \frac{1}{2})}$, and $Z = 1_{(\frac{1}{2}, \frac{3}{4})}$. Then $X \perp Y$, $Y \perp Z$, $Z \perp X$, but $X, Y, Z$ are not independent. This is because $P(X = Y = Z = 1) = 0 \neq (\frac{1}{2})^3 = P(X = 1)P(Y = 1)P(Z = 1)$.

**Lemma 4.32.** If $C_1, \ldots C_n$ are independent $\pi$-systems, then $\sigma(C_1), \ldots, \sigma(C_n)$ are independent $\sigma$-fields.

**Proof Sketch:** Apply the $\pi$-$\lambda$ theorem.

We recall that $\text{Var}(X) := E(X - E(X))^2$. Using Fubini’s theorem, we have the following.

**Theorem 4.33.** Let $X, Y$ be independent rvs with $\|X\|_2, \|Y\|_2 < \infty$. Then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

**Proof sketch:**

$$\text{Var}(X + Y) = E((X + Y) - E(X + Y))^2$$

$$= E((X - E X)^2 + 2(X - E X)(Y - E Y) + (Y - E Y)^2)$$

$$= \text{Var}(X) + 2E(X - E X) \cdot E(Y - E Y) + \text{Var}(Y)$$

$$= \text{Var}(X) + \text{Var}(Y).$$

We also know that, for all $p > q \geq 1$,

$\begin{align*}
L^p \text{ convergence} &\implies L^q \text{ convergence} \implies L^1 \text{ convergence} \\
&\implies \text{convergence in } P \implies \text{convergence in distribution.}
\end{align*}$

**Exercise 4.34.**

i) Show that if $X_n \downarrow 0$ in probability, then $X_n \downarrow 0$ almost surely.

ii) Show that, $X_n \xrightarrow{P} 0$ if and only if $X_n \xrightarrow{(d)} 0$. 
Law of Large numbers and Borel Cantelli Lemmas

5.1 Law of Large Numbers (LLN)

**Theorem 5.1 (L² – Law of Large Numbers).** Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E\, X_1 = \mu$ and $E\, X_1^2 < \infty$. Then

$$
\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{L^2} \mu \text{ as } n \to \infty.
$$

**Proof Sketch:** We have

$$
\left\| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right\|_2^2 = \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{1}{n} \text{Var}(X_1) \to 0.
$$

For the Weak Law of Large Numbers (WLLN) introduced below, no assumption on the $E\, X^2$ is needed.

**Theorem 5.2 (Weak Law of Large Numbers – WLLN).** Let $X_1, X_2, \ldots$ be i.i.d. with $E\, X_1 = \mu$ is finite. Then

$$
\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu \text{ as } n \to \infty.
$$

The convergence also holds in $L^1$.

**Proof.** Note that, it is enough to prove $L^1$-convergence, i.e., $\left\| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right\|_1 \to 0$ as $L^1$ convergence implies convergence in probability.

Fix $x > 0$. Define, $Y_i := X_i 1_{|X_i|<x}$ for $i \geq 1$. We can write $X_i = X_i 1_{|X_i| \geq x} + Y_i$. Note that, by DCT we have, $E\left(|X_1| 1_{|X_1| > x}\right) \to 0$ as $x \uparrow \infty$. The main idea behind the proof is to truncate the r.v.s, so that the bounded part converges using $L^2$ law of large numbers and the unbounded part has small expectation.

We have, for any $x > 0$

$$
\left\| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right\|_1 \leq E\left| \frac{1}{n} \sum_{i=1}^{n} X_i 1_{|X_i| \geq x} - E(X_1 1_{|X_1| \geq x}) \right| + E\left| \frac{1}{n} \sum_{i=1}^{n} Y_i - EY_1 \right|
$$

$$
\leq 2 E(|X_1| 1_{|X_1| \geq x}) + \sqrt{\text{Var} \left( \frac{1}{n} \sum_{i=1}^{n} Y_i \right)}
$$

$$
= 2 E(|X_1| 1_{|X_1| \geq x}) + \sqrt{n^{-1} \text{Var}(Y_1)} \leq 2 E(|X_1| 1_{|X_1| \geq x}) + n^{-1/2} \left\| X_1 1_{|X_1| < x} \right\|_2.
$$
Taking $x = n^{1/4}$ and letting $n \to \infty$, we get
\[
\lim_{n \to \infty} \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right|_1 = 0.
\]

Note that, to prove convergence in Probability it was enough to prove that $P(\max_{1 \leq i \leq n} |X_i| > x) \to 0$ for any $x > 0$ satisfying $\text{Var}(X_1 \mathbb{1}_{|X_1| > x}) \ll n$.

\section*{Exercise 5.3.}
Show that $X$, the set of all random variables, is metrizable under convergence in probability, that is, there exists a metric $d$ on $X$ such that $X_n \xrightarrow{P} X$ if and only if $d(X_n, X) \to 0$ and $X$ is complete under the metric.

\textbf{Hint:} There are many choices:
\begin{align*}
    d(X, Y) &= \mathbb{E}(|X - Y| \wedge 1) \\
    d(X, Y) &= \mathbb{E}(|X - Y|/(1 + |X - Y|)) \\
    d(X, Y) &= \inf \{P(|X - Y| > \varepsilon) + \varepsilon \mid \varepsilon > 0\}.
\end{align*}

We can strengthen the conclusion of WLLN without any extra assumption.

\section*{Theorem 5.4 (Strong Law of Large Numbers – SLLN).}
Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}X_1 = \mu$. Then
\[
\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu.
\]

\section*{Example 5.5.}
Let $X_1, X_2, \ldots \sim \text{Uniform}([-1, 1])$ be i.i.d. r.v. with mean zero. Note $\mathbb{E}X_i^2 = \int_{-1}^{1} x^2 \cdot \frac{1}{2} dx = \frac{1}{3}$. So by the SLLN, $\frac{1}{n} \sum_{i=1}^{n} X_i^2 \xrightarrow{a.s.} \frac{1}{3}$. Now $(X_1, \ldots, X_n) \sim \text{Uniform}([-1, 1]^n)$. So an $n$-dimensional vector with uniform coordinates approximately lies on a sphere with radius $\sqrt{\frac{n}{3}}$ with high probability.

We need certain results to obtain a.s. convergence from probability estimates.

\section*{5.2 Borel Cantelli Lemmas}

\section*{Definition 5.6.}
Let $A_1, A_2, \ldots$ be an infinite sequence of events. We define
\[
\limsup_n A_n = \cap_{m \geq 1} \cup_{n \geq m} A_n = \lim_{m \to \infty} \cup_{n \geq m} A_n = \{A_n \text{ infinitely often (i.o.)}\} = \{\omega \text{ that are in infinitely many } A_n\}.
\]

We have the following:
\begin{itemize}
    \item $\limsup_n \mathbb{1}_{A_n} = \mathbb{1}_{\limsup_n A_n}$,
    \item $\liminf_n A_n = \lim_{m \uparrow \infty} \cap_{n \geq m} A_n = \cup_{m \geq 1} \cap_{n \geq m} A_n = \{A_n \text{ eventually (ev.)}\} = \{\omega \text{ that are in all but finitely many } A_n\}$,
\end{itemize}
• \( \lim \inf_n \mathbb{1}_{A_n} = \mathbb{1}_{\lim \inf_n A_n} \),
• \( (\lim \sup_n A_n)^c = \lim \inf_n A_n^c \),
• \( \lim \inf_n A_n \subseteq \lim \sup_n A_n \).

Now we will state and prove the Borel Cantelli Lemmas:

**Lemma 5.7 (Borel–Cantelli Lemmas).** Let \( A_1, A_2, \ldots \) be a sequence of events.

(i) If \( \sum_{i=1}^{\infty} \mathbb{P}(A_i) < \infty \), then

\[ \mathbb{P}(A_n \text{ i.o.}) = 0. \]

(ii) If \( A_1, A_2, \ldots \) are independent and \( \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \infty \), then

\[ \mathbb{P}(A_n \text{ i.o.}) = 1. \]

**Proof.** (i) \( \mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}(\lim_m \cup_{n \geq m} A_n) = \lim_m \mathbb{P}(\cup_{n \geq m} A_m) \) with \( \cup_{n \geq m} A_m \) decreasing in \( m \). This implies

\[ \lim_m \mathbb{P}(\cup_{n \geq m} A_n) \leq \lim_m \sum_{n \geq m} \mathbb{P}(A_n) = 0. \]

(ii) We have

\[ \mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}(\lim_m \cap_{n \geq m} A_n^c) = \lim_m \mathbb{P}(\cap_{n \geq m} A_n^c) \]

\[ = \lim_m \prod_{n \geq m} \mathbb{P}(A_n^c) = \lim_m \prod_{n \geq m} (1 - \mathbb{P}(A_n)) \leq \lim_m e^{-\sum_{n \geq m} \mathbb{P}(A_n)} = 0, \]

since \( (1 - x) \leq e^{-x} \) and \( \sum_{n \geq m} \mathbb{P}(A_n) = \infty \) for all \( m \geq 1 \).

**Remark 5.8.** The assumption of independence in part ii) of Borel-Cantelli Lemma can be replaced by pairwise independence. See [Durrett] for the proof.

**Lemma 5.9.** The following are equivalent:

(i) \( X_n \to X \) a.s.

(ii) For every \( \varepsilon > 0 \) we have \( \mathbb{P}(|X_n - X| > \varepsilon \text{ i.o.}) = 0. \)

(iii) There exists \( \varepsilon_n \downarrow 0 \) such that \( \mathbb{P}(|X_n - X| > \varepsilon_n \text{ i.o.}) = 0. \)

(iv) For \( M_n := \max_{k \geq n} |X_k - X|, \) we have \( M_n \to 0 \) in probability.

**Proof.** i) \( \implies \) ii): \( \mathbb{P}(\lim_{n \to \infty} |X_n - X| = 0) = 1 = \mathbb{P}(\forall \varepsilon > 0, |X_n - X| \leq \varepsilon \text{ ev.}). \) This implies \( \forall \varepsilon > 0, \mathbb{P}(|X_n - X| \leq \varepsilon \text{ ev.}) = 1. \) We have the result by taking complement.

**Exercise 5.10.** Prove the equivalence in Lemma 5.9.

In v) of Lemma 5.9, we cannot replace \( \exists \) with \( \forall \), illustrated by the following example.

**Example 5.11.** Let \( X_n = U_n/n \), where \( U_1, U_2, \ldots \) are i.i.d. Uniform\((-1,1)\) rvs. Then we have

\[ \mathbb{P}(|X_n| > 1/2n \text{ i.o.}) = 1 \]

and \( X_n \xrightarrow{a.s.} 0. \) For the proof, let \( A_n = \{|X_n| > 1/2n\}, n \geq 1. \) Then \( \mathbb{P}(A_n) = \frac{1}{2} \implies \sum \mathbb{P}(A_n) = \infty \) and \( A_n \)'s are independent. Therefore, by second Borel-Cantelli Lemma, \( \mathbb{P}(A_n \text{ i.o.}) = 1 \) or \( \mathbb{P}(|X_n| > 1/2n \text{ i.o.}) = 1. \)
5.3 Random Walk on $\mathbb{Z}^d$

In this section we consider random walks in $d$-dimensions, $\mathbb{Z}^d$. Let $X_1, X_2, \ldots$ be i.i.d. $\mathbb{Z}^d$-valued random vector such that

$$P(X_1 = e_i) = P(X_1 = -e_i) = \frac{1}{2^d}, \text{ for } i = 1, 2, \ldots, d$$

where $e_i$ is the unit vector in the $i$-th coordinate direction. The simple symmetric random walk (SSRW) $S_n$ is defined as

$$S_0 = 0,$$
$$S_n = X_1 + X_2 + \cdots + X_n, \ n \geq 1.$$ 

**Theorem 5.12.** For the simple symmetric random walk on $\mathbb{Z}^d$ we have,

$$P(S_n = 0 \text{ i.o.}) = \begin{cases} 1 & \text{if } d = 1, 2 \\ 0 & \text{if } d \geq 3. \end{cases}$$

**Example 5.13 (Random Walk on $\mathbb{Z}$).** Let $X_1, X_2, \ldots$ be i.i.d. with $P(X_i = -1) = \frac{1}{2} = P(X_i = 1)$ for all $i \in \mathbb{N}$. Define

$$S_n = X_1 + X_2 + \cdots + X_n$$

for $n \in \mathbb{N}$. It can be shown that $P(S_n = 0 \text{ i.o.}) = 1$. Notice that

$$P(S_{2n} = 0) = \left(\frac{2n}{n}\right)^n \left(\frac{1}{2}\right)^n$$

for any $n$. By Stirling’s approximation: $n! \sim (2\pi)^{\frac{1}{2}} e^{-n} n^{n+\frac{1}{2}}$,

$$P(S_{2n} = 0) = \frac{2^n}{(n!)^2} \frac{1}{2^n} \sim \frac{(2\pi)^{\frac{1}{2}} e^{-2n} 2^{2n+1}}{2 \pi e^{-2n} 2^{2n+1} 2^{2n}} = \frac{1}{\sqrt{\pi n}}.$$

So $\sum_{n \geq 1} P(S_n = 0) = \infty$. However

$$P(S_{2n} = 0, S_{2n+2} = 0) = P(S_{2n} = 0) \times \frac{2}{2^2} \neq P(S_{2n} = 0) P(S_{2n+2} = 0).$$

So Borel-Cantelli Lemma cannot be applied to show $P(S_n = 0 \text{ i.o.}) = 1$.

**Example 5.14 (Another Random Walk on $\mathbb{Z}^d$).** Let $\tilde{X}_i = (x_1, x_2, \ldots, x_d)$ with probability $\frac{1}{2^d}$ where each $x_i \in \{1, -1\}$. Then, for each $n$, define

$$S_n = \tilde{X}_1 + \tilde{X}_2 + \cdots + \tilde{X}_n.$$

Analogous to random walk on $\mathbb{Z}$, we are interested in calculating $P(S_{2n} = 0)$. Write

$$S_n = (S_{n}^1, \ldots, S_{n}^d)$$

where $(S_{n}^1)_{n \geq 1}, \ldots, (S_{n}^d)_{n \geq 1}$ are i.i.d. 1-dimensional random walk. Observe that

$$P(S_{2n} = 0) = P(S_{2n}^1 = 0, \ldots, S_{2n}^d = 0) \sim \frac{1}{(n\pi)^{\frac{d}{2}}}.$$

Then $\sum_{n \geq 1} P(S_{2n} = 0) < \infty$ if and only if $d \geq 3$. In particular, for $d \geq 3$, $P(S_{2n} = 0 \text{ i.o.}) = 0.$
5.4 Proof of Strong Law of Large Number

Theorem 5.15 (Strong Law of Large number). Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with $E X_1 = \mu$ finite, then

$$n^{-1} \sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu.$$  

Proof Sketch: WLOG, assume $X_i \geq 0$. (For general random variables $X_i$, we can write $X_i = X_i^+ - X_i^-$ and apply the case when each $X_i \geq 0$ separately to $(X_i^+)_{i \geq 1}$ and $(X_i^-)_{i \geq 1}$.) We will use the following steps:

1. Truncation (inhomogeneous)
2. Error analysis for truncation
3. Subsequence argument.

Let $Y_k := X_k 1_{X_k \leq k}$.

Notice that if we set $T_n = \sum_{i=1}^{n} Y_i$, then it suffices to show that $\frac{T_n}{n} \overset{a.s.}{\to} \mu$ because of the following lemma.

Lemma 5.16. $\mathbb{P}(X_k \neq Y_k \text{ i.o.}) = 0$.

Proof. We will apply Borel-Cantelli Lemma (i) to prove this result. So it is enough to show that $\sum_k \mathbb{P}(X_k \neq Y_k) < \infty$. Notice that $\mathbb{P}(X_k \neq Y_k) = \mathbb{P}(X > k)$. So

$$\sum_k \mathbb{P}(X_k \neq Y_k) = \sum_{k \geq 1} \mathbb{P}(X > k) = E(\sum_{k \geq 1} 1_{X > k}) \leq E X < \infty$$

where the last inequality follows by $0 \leq \sum_{k \geq 1} 1_{X > k} \leq X$.

In particular, $n^{-1} \sum_{i=1}^{n} X_i - n^{-1} \sum_{i=1}^{n} Y_i \overset{a.s.}{\to} 0$ because $\left| \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i \right| \leq \sum_{i=1}^{\infty} |X_i - Y_i|$, a finite random variable.

Now recall that $T_n = \sum_{i=1}^{n} Y_i$. Take $Z_n = \frac{T_n - E T_n}{n \varepsilon^2}$. By Chebyshev’s inequality we have

$$\mathbb{P}(|Z_n| > \varepsilon) \leq \frac{\text{Var}(T_n)}{n^2 \varepsilon^4} = \frac{\sum_{i=1}^{n} \text{Var}(Y_i)}{n^2 \varepsilon^4} \leq \frac{\sum_{i=1}^{n} E Y_i^2}{n^2 \varepsilon^4} = \frac{1}{n^2 \varepsilon^2} \sum_{i=1}^{n} E X_i^2 1_{X_i \leq k} \leq \frac{1}{n^2 \varepsilon^2} E X_i^2 1_{X_i \leq n}.$$ 

In general, for any subsequences $(Z_{n_k})$, we have

$$\sum_{k \geq 1} \mathbb{P}(|Z_{n_k}| > \varepsilon) \leq \sum_{k=1}^{\infty} \frac{1}{n_k \varepsilon^2} E X_i^2 1_{X_i \leq n_k} \leq \frac{1}{\varepsilon^2} E X_i^2 \left( \sum_{k=1}^{\infty} \frac{1}{n_k} \right).$$

We want a sequence $n_k$ such that

$$\sum_{k=1}^{\infty} \frac{1}{n_k} \approx \frac{1}{\varepsilon^2}.$$
Geometric series satisfies this criteria. So we take \( n_k = \lceil \alpha^k \rceil \) for some \( \alpha > 1 \). Then, for all \( x > 0 \) we have

\[
\sum_{k \colon n_k \geq x} \frac{1}{n_k} \leq \frac{c(\alpha)}{x}
\]

for some finite constant \( c(\alpha) \) depending only on \( \alpha \). So

\[
\mathbb{E} \left( X_1^2 \cdot \sum_{k=1}^{\infty} \frac{1_{n_k \geq X_1}}{n_k} \right) \leq c(\alpha) \mathbb{E} X_1 < \infty.
\]

In particular, \( Z_{n_k} \xrightarrow{a.s.} 0 \) where \( n_k = \lceil \alpha^k \rceil \), or \( \frac{T_{n_k}}{n_k} - \frac{\mathbb{E} T_{n_k}}{n_k} \to 0 \) as \( k \to \infty \). Now,

\[
\frac{\mathbb{E} T_n}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} X_1 1_{X_1 \leq i} \to \mathbb{E} X_1 \text{ as } n \to \infty \text{ since } \mathbb{E} X_1 1_{X_1 \leq i} \to \mathbb{E} X_1 \text{ as } i \to \infty
\]

or

\[
\frac{T_{n_k}}{n_k} \xrightarrow{a.s.} \mu.
\]

Notice that \( T_{n_k} \leq T_n \leq T_{n_{k+1}} \) whenever \( n_k < n \leq n_{k+1} \), or

\[
\frac{T_{n_k}}{n_{k+1}} \leq \frac{T_n}{n} \leq \frac{T_{n_{k+1}}}{n_k}.
\]

So

\[
\frac{\mu}{\alpha} \leq \liminf_{k \to n} \frac{T_{n_k}}{n_{k+1}} \leq \limsup_{n} \frac{T_n}{n} \leq \limsup_{k} \frac{T_{n_k}}{n_k} \leq \mu \alpha.
\]

Since \( \alpha > 1 \) is arbitrary,

\[
\limsup_{n} \frac{T_n}{n} = \liminf_{n} \frac{T_n}{n} = \mu \text{ or } \frac{T_n}{n} \xrightarrow{a.s.} \mu.
\]
6.1 Kolmogorov’s maximal inequality

Theorem 6.1 (Kolmogorov’s Maximal Inequality). Let \( X_1, \ldots, X_n \) be independent random variables with \( \mathbb{E} X_i = 0, \mathbb{E} X_i^2 < \infty \) for all \( 1 \leq i \leq n \). Define \( S_n = X_1 + \cdots + X_n \). Then

\[
\mathbb{P} \left( \max_{1 \leq k \leq n} |S_k| \geq a \right) \leq \frac{\mathbb{E} S_n^2}{a^2}
\]

for any \( a > 0 \).

Proof Sketch: Let \( A_i = \{|S_1| < a, |S_2| < a, \ldots, |S_{i-1}| < a, |S_i| \geq a\}, i = 1, 2, \ldots, n \). Then \( A_i \cap A_j = \emptyset \) for all \( i \neq j \). Notice that \( \mathbb{P} \left( \max_{1 \leq k \leq n} |S_k| \geq a \right) = \sum_{i=1}^n \mathbb{P}(|S_i| \geq a, \max_{1 \leq k < i} |S_k| < a) \). Then

\[
\begin{align*}
\mathbb{E} S_n^2 &\geq \sum_{i=1}^n \mathbb{E} S_n^2 \mathbb{1}_{A_i} = \sum_{i=1}^n \mathbb{E}(S_i^2 + 2S_i(S_n - S_i) + (S_n - S_i)^2) \mathbb{1}_{A_i} \\
&\geq \sum_{i=1}^n (a^2 P(A_i) + 2 \mathbb{E} S_i \mathbb{1}_{A_i}(S_n - S_i)) = a^2 \mathbb{P} \left( \max_{1 \leq k \leq n} |S_k| \geq a \right).
\end{align*}
\]

In the last line we used the fact that \( S_i \mathbb{1}_{A_i} \) is independent of \( X_{i+1}, \ldots, X_n \).

Theorem 6.2 (Basic \( L^2 \)-convergence). Let \( X_1, X_2, \ldots \) be independent r.v.’s with \( \mathbb{E}(X_i) = 0 \) and \( \sum_{i=1}^\infty \text{Var}(X_i) < \infty \). Then \( S_n = \sum_{i=1}^n X_i \) converges in \( L^2 \) and a.s.

Proof. To see \( L^2 \)-convergence, we will prove that \( (S_n)_{n \geq 1} \) is Cauchy in \( L^2 \). It is enough to show that \( ||S_n - S_m||_2 < \varepsilon \) for all \( n, m \geq N(\varepsilon) \). Suppose \( n > m \), then we obtain

\[
||S_n - S_m||_2^2 = \mathbb{E}(S_n - S_m)^2 = \mathbb{E}(X_{m+1} + X_{m+2} + \cdots + X_n)^2
= \text{Var}(X_{m+1} + X_{m+2} + \cdots + X_n)
= \text{Var}(X_{m+1}) + \text{Var}(X_{m+2}) + \cdots + \text{Var}(X_n).
\]

For any \( \varepsilon > 0 \), there exists \( N = N(\varepsilon) \) with \( \sum_{i=N}^\infty \text{Var}(X_i) < \varepsilon^2 \), thus we have \( ||S_n - S_m||_2 < \varepsilon \) for all \( n, m \geq N(\varepsilon) \) which implies that \( (S_n)_{n \geq 1} \) is a Cauchy sequence in \( L^2 \) and therefore converges in \( L^2 \). Now, it remains to show a.s. convergence. In other words, we need to show \( \mathbb{P}(S_n \text{ converges}) = 1 \).

Define \( M_N = \max\{|S_n - S_m| : n, m \geq N\} \), then clearly \( M_N \) is decreasing. Indeed, it is enough to
show that $M_N \xrightarrow{p} 0$ as $N \to \infty$. Let $\varepsilon > 0$ be given. Then,

$$
\mathbb{P}(M_N > \varepsilon) = \mathbb{P}(\max_{n,m \geq N} |S_n - S_m| > \varepsilon) \leq \mathbb{P}(\max_{k \geq N} |S_{N+k} - S_N| > \varepsilon/2)
$$

$$
= \lim_{n \to \infty} \mathbb{P}(\max_{1 \leq k \leq n} |S_{N+k} - S_N| > \varepsilon/2)
$$

$$
\leq \lim_{n \to \infty} \frac{\text{Var}(S_{N+n} - S_N)}{(\varepsilon/2)^2}
$$

$$
= \lim_{n \to \infty} \left(\frac{2}{\varepsilon}\right)^2 \cdot \sum_{i=N+1}^{N+n} \text{Var}(X_i) = \left(\frac{2}{\varepsilon}\right)^2 \cdot \sum_{i \geq N} \text{Var}(X_i),
$$

where the second inequality follows by Kolmogorov's Maximal inequality. Hence, $\lim_{N \to \infty} \mathbb{P}(M_N > \varepsilon) = 0$ or $M_N \xrightarrow{p} 0$, thus we obtain $M_N \to 0$ a.s. Therefore, $\mathbb{P}((S_n)_{n \geq 1}$ is Cauchy$) = 1$.

Using basic $L^2$-convergence theorem, we can give a simple proof of SLLN under finite second moment. Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}X_1 = 0$ and $\text{Var}(X_1) < \infty$. By basic $L^2$-convergence theorem, we have $\sum_{k=1}^{n} X_k/k$ converges a.s as $n \to \infty$. Now, using Kronecker’s lemma (stated and proved below), we have $n^{-1} \sum_{k=1}^{n} X_k \to 0$ a.s.

**Lemma 6.3 (Kronecker’s lemma).** Let $a_n \uparrow \infty$ and $\sum_{i=1}^{n} \frac{x_i}{a_i}$ converges. Then

$$
\frac{1}{a_n} \sum_{i=1}^{n} x_i \to 0
$$

**Proof.** We have $b_n := \sum_{i=1}^{n} \frac{x_i}{a_i} \to b$, for some $b \in \mathbb{R}$, as $n \to \infty$. Moreover, $x_n = a_n(b_n - b_{n-1})$ so

$$
\frac{1}{a_n} \sum_{i=1}^{n} x_i = \frac{1}{a_n} \sum_{i=1}^{n} (a_i b_i - a_i b_{i-1}) = \frac{1}{a_n} (a_n b_n + \sum_{i=1}^{n-1} (a_i - a_{i+1})b_i) = b_n - \frac{1}{a_n} \sum_{i=1}^{n-1} (a_{i+1} - a_i)b_i.
$$

Let $d_i = a_{i+1} - a_i$. We know that $b_n \to b$ and thus

$$
\frac{\sum_{i=1}^{n-1} d_i b_i}{\sum_{i=1}^{n-1} d_i} - b = \frac{\sum_{i=1}^{n-1} d_i (b_i - b)}{\sum_{i=1}^{n-1} d_i} \to 0
$$

as $n \to \infty$, since $d_i \geq 0$ and $\sum_{i \geq 1} d_i = \infty$. So this proves that $\frac{1}{a_n} \sum_{i=1}^{n} x_i \to 0$ completing the proof.

**6.2 Applications of SLLN**

**Theorem 6.4 (Renewal Theorem).** Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.'s with $X_i \geq 0$, $\mathbb{E}X_1 = \mu > 0$. Then

$$
\frac{N_t}{t} \overset{a.s.}{\xrightarrow{t \to \infty}} \frac{1}{\mu}
$$

where $N_t = \sup\{n \mid X_1 + X_2 + \cdots + X_n \leq t\}$. 
Proof Sketch: By SLLN, if $S_n = X_1 + X_2 + \cdots X_n$, then $\frac{S_n}{n} \xrightarrow{a.s.} \mu$ as $n \to \infty$. We observe that

- $N_t \uparrow \infty$ a.s. as $t \to \infty$. ($\{N_t \geq n\} = \{X_1 + X_2 + \cdots X_n \leq t\}$).
- $\mathbb{P}\left(\frac{S_n}{n} \xrightarrow{t \to \infty} \mu, N_t \xrightarrow{t \to \infty} \infty\right) = 1$ or $\mathbb{P}\left(\frac{S_{N_t}}{N_t} \xrightarrow{t \to \infty} \mu, \frac{S_{N_t+1}}{N_t+1} \xrightarrow{t \to \infty} \mu\right) = 1$.

By definition, we have $S_{N_t} \leq t < S_{N_t+1}$ so that

$$\frac{S_{N_t}}{N_t} \leq \frac{t}{N_t} < \frac{S_{N_t+1}}{N_t+1} = \frac{S_{N_t+1}}{N_t+1} \cdot \frac{N_t+1}{N_t}.$$ 

Both upper and lower bounds converges to $\mu$ a.s. and this completes the proof. 

**Definition 6.5 (Empirical Distribution).** Let $X_1, X_2, \cdots$ be a sequence of i.i.d. r.v.’s with distribution function $F$. Define

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} \text{ for all } x \in \mathbb{R}.$$ 

$F_n$ is called the empirical distribution function. By SLLN, $F_n(x) \xrightarrow{a.s.} F(x)$, for each $x \in \mathbb{R}$.

**Theorem 6.6 (Glivenko-Cantelli Lemma).** We have

$$||F_n - F||_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$ 

Proof Sketch: Fix $k \geq 1$, define

$$x_{j,k} = F^{-1}\left(\frac{j}{k}\right) = \inf\left\{ x : F(x) \geq \frac{j}{k} \right\}, \text{ } 1 \leq j < k.$$ 

Define $x_{0,k} = -\infty, x_{k,k} = \infty$. By SLLN,

$$\mathbb{P}\left(F_n(x_{j,k}) \xrightarrow{a.s.} F(x_{j,k}) \text{ } \forall \text{ } 0 \leq j \leq k\right) = 1.$$ 

Similarly,

$$F_n(x-) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i < x} \xrightarrow{a.s.} F(x-).$$

Thus

$$\mathbb{P}\left(F_n(x_{j,k}^-) - F(x_{j,k}^-) \to 0, \text{ } F_n(x_{j,k}) - F(x_{j,k}) \to 0 \text{ } \forall \text{ } 0 \leq j \leq k\right) = 1.$$ 

Define

$$\Delta_k^{(n)} = \max_{0 \leq j \leq k} \{ |F_n(x_{j,k}) - F(x_{j,k})|, |F_n(x_{j,k}^-) - F(x_{j,k}^-)| \}.$$ 

Take $x$ in $[x_{j,k}, x_{j+1,k})$. Then,

$$F_n(x_{j,k}) \leq F_n(x) \leq F_n(x_{j+1,k}) \implies F_n(x_{j,k}) - F(x) \leq F_n(x) - F(x) \leq F_n(x_{j+1,k}^-) - F(x).$$

Since

$$|F_n(x_{j+1,k}^-) - F(x)| \leq |F_n(x_{j+1,k}^-) - F(x_{j+1,k})| + |F(x_{j+1,k}^-) - F(x)|$$

$$\leq \Delta_k^{(n)} + \frac{1}{k}.$$ 

---

**Week 6: Kolmogorov’s Maximal Inequality and Applications of SLLN**

6-3
and

\[ |F_n(x_{j,k}) - F(x)| \leq |F_n(x_{j,k}) - F(x_{j,k})| + |F(x_{j,k}) - F(x)| \leq \Delta_k^{(n)} + \frac{1}{k}, \]

we have \( \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \Delta_k^{(n)} + \frac{1}{k} \). Note that \( \Delta_k^{(n)} \xrightarrow{a.s.} 0 \). Thus

\[ P \left( \limsup_n \left\| F_n - F \right\|_\infty \leq \frac{1}{k} \right) = 1 \quad \forall k \]

\[ \implies P \left( \limsup_n \left\| F_n - F \right\|_\infty \leq \frac{1}{k} \forall k \right) = 1, \]

\[ \implies P \left( \limsup_n \left\| F_n - F \right\|_\infty = 0 \right) = 1. \]

\[ \Box \]

**Remark.** See "Bootstrap" in statistics for applications of Glivenko-Cantelli Lemma.

### 6.3 Tail events and Kolmogorov’s 3-Series Theorem

**Definition 6.7 (Tail \( \sigma \)-field).** Let \( X_1, X_2, \ldots \) be independent r.v.s. The \( \sigma \)-field

\[ T = \bigcap_{k=1}^{\infty} \sigma(X_k, X_{k+1}, \ldots) \]

is called the tail \( \sigma \)-field.

**Example 6.8.** We have \( \{ \lim X_n \text{ exists} \} \in T, \{ \limsup_n (X_1 + \ldots + X_n)/n \leq a \} \in T. \) However, the event \( \{ \limsup_n (X_1 + \ldots + X_n) \leq a \} \notin T. \)

**Theorem 6.9 (Kolmogorov’s 0-1 Law).** For all \( A \in T, P(A) \in \{0, 1\}. \)

**Proof.** We will prove that \( T \perp T \) (the tail \( \sigma \)-field is independent of itself). Then for all \( A \in T, \)

\[ P(A) = P(A \cap A) = P(A) \cdot P(A) \]

which implies \( P(A) \in \{0, 1\}. \) The proof is in 4 steps.

**Step 1.** \( \sigma(X_1, \ldots, X_{k-1}) \perp \sigma(X_{l+1}, X_{l+2}, \ldots) \) for all \( l \geq k, \) since \( X_i \)’s are independent.

**Step 2.** In particular, \( \sigma(X_1, \ldots, X_{k-1}) \perp T \) for all \( k \geq 2. \)

**Step 3.** \( \sigma(X_1, \ldots) \perp T. \)

**Step 4.** \( T \perp T. \)

**Proof of step 1:** Clearly \( \sigma(X_1, \ldots, X_{k-1}) \perp \sigma(X_l, \ldots, X_{l+j}) \) for all \( l \geq k, j \geq 0, \) which implies that

\[ \sigma(X_1, \ldots, X_{k-1}) \perp \bigcup_{j=0}^{\infty} \sigma(X_l, \ldots, X_{l+j}). \]
By \(\pi - \lambda\) theorem, \(\sigma(X_1, \ldots, X_{k-1}) \perp \sigma(\cup_{j=0}^{\infty}\sigma(X_i, \ldots, X_{i+j})) = \sigma(X_i, X_{i+1}, \ldots)\).

**Proof of step 2:** Note that \(\mathcal{F} \perp \mathcal{G}\) implies that any sub \(\sigma\)-field of \(\mathcal{F}\) is independent of \(\mathcal{G}\). (*

Clearly \(\mathcal{T} \subseteq \sigma(X_k, X_{k+1}, \ldots)\). So by (*), \(\sigma(X_1, \ldots, X_{k-1}) \perp \mathcal{T}\).

**Proof of step 3:** We have \(\bigcup_{k \geq 1} \sigma(X_1, \ldots, X_{k-1}) \perp \mathcal{T}\). Thus by \(\pi - \lambda\) theorem, \(\sigma(X_1, \ldots) \perp \mathcal{T}\).

**Proof of step 4:** Clearly \(\mathcal{T} \subseteq \sigma(X_1, X_2, \ldots)\). So by (*), we have the proof. \(\blacksquare\)

**Theorem 6.10 (Kolmogorov’s 3-series Theorem).** Let \(X_1, X_2, \ldots\) be independent. Fix \(b > 0\).

Consider the three deterministic series:

\[
\begin{align*}
(i) \sum_{i=1}^{n} \mathbb{P}(|X_i| > b), \quad (ii) \sum_{i=1}^{n} \mathbb{E}(X_i 1_{|X_i| \leq b}), \quad (iii) \sum_{i=1}^{n} \text{Var}(X_i 1_{|X_i| \leq b}).
\end{align*}
\]

Then

\[
\sum_{i=1}^{n} X_i \text{ converges a.s. if and only if } (i), (ii), (iii) \text{ converge.}
\]

**Proof.** We will first prove the if part. Write \(X_i = U_i + v_i + W_i\), where

\[
U_i := X_i 1_{|X_i| > b}, \quad v_i := \mathbb{E}X_i 1_{|X_i| \leq b}, \quad \text{and } W_i := X_i 1_{|X_i| \leq b} - v_i.
\]

By convergence of (ii), we have \(\sum_{i \geq 1} v_i < \infty\). We will prove that both \(\lim_{n \to \infty} \sum_{i=1}^{n} U_i\) and \(\lim_{n \to \infty} \sum_{i=1}^{n} W_i\) exists a.s. By Borel-Cantelli Lemma and convergence of series (i), we have \(\mathbb{P}(|X_i| > b \text{ i.o.}) = 0\) which implies that \(\lim_{n \to \infty} \sum_{i=1}^{n} U_i\) exists a.s. Convergence of (iii) and Basic \(L^2\) convergence implies \(\lim_{n \to \infty} \sum_{i=1}^{n} W_i\) exists a.s.

For the only if part, we need to use Central Limit Theorem, which states that for a sequence of iid mean zero, variance one rvs \(X_1, X_2, \ldots\), the sequence \(n^{-1/2} S_n\) converges in distribution to \(N(0, 1)\) and will be proved later. \(\blacksquare\)

Another application of Central Limit Theorem and truncation + subsequence technique is the following result, which will not be proved here. The crucial idea is that \(S_n/\sqrt{n}\) converges in distribution to \(N(0, 1)\) and for a rv \(X \sim N(0, 1)\) we have \(\mathbb{P}(X \geq t) = \mathbb{P}(X \leq -t) \leq e^{-t^2/2}, t > 0\).

**Theorem 6.11 (Law of Iterated Logarithms (LIL)).** Let \(X_1, X_2, \ldots\) be i.i.d. with mean 0, \(\text{Var}(X_1) = 1\). Define \(S_n = \sum_{i=1}^{n} X_i\). Then

\[
\limsup_{n} \frac{S_n}{\sqrt{2n \log \log n}} = 1 \text{ a.s.}
\]

\[
\liminf_{n} \frac{S_n}{\sqrt{2n \log \log n}} = -1 \text{ a.s.}
\]

A sequence \(\{t_n\}_{n \geq 1}\) is subadditive, if \(t_{m+n} \leq t_m + t_n\) for all \(m, n \geq 1\). The following lemma for a deterministic sequence can be generalize to random sequences to give another proof of SLLN.

**Theorem 6.12 (Subadditive limit theorem).** If \(\{t_n\}_{n \geq 1}\) is subadditive, then

\[
\frac{t_n}{n} \to \inf_{m \geq 1} \frac{t_m}{m} \text{ as } n \to \infty.
\]
Proof. Clearly, \( \liminf_n t_n/n \geq \inf_{m \geq 1} t_m/m \). Note that, \( t_{mn+k} \leq nt_m + tk \) for \( m, n, k \geq 1 \). Thus
\[
\frac{t_{mn+k}}{mn+k} \leq \frac{nt_m + tk}{mn+k} = \frac{\frac{tm}{m} + \frac{tk}{mn}}{1 + \frac{k}{mn}}
\]
Fix \( m \geq 1 \) and \( 0 \leq k < m \), then
\[
\limsup_n \frac{t_{mn+k}}{mn+k} \leq \frac{tm}{m}.
\]
Thus,
\[
\limsup_n \frac{tn}{n} \leq \frac{tm}{m}.
\]
Since \( m \) is arbitrary, we have
\[
\limsup_{n \to \infty} \frac{tn}{n} \leq \inf_{m \geq 1} \frac{tm}{m}.
\]

**Corollary 6.13.** Let \( X_1, X_2, \ldots \) be i.i.d. r.v.s. Define \( S_n = X_1 + X_2 + \cdots + X_n, n \geq 1 \). Then,
\[
\kappa(a) := \lim_{n \to \infty} -\frac{1}{n} \log P(S_n \geq a)
\]
exists and is non-negative (need not be finite).

**Proof.** We use the fact that \( P(S_{n+m} \geq (n+m)a) \geq P(S_m \geq ma, S_{n+m} - S_m \geq na) \geq P(S_m \geq ma) \cdot P(S_n \geq na) \) and thus the sequence \( t_n := -\log P(S_n \geq a) \geq 0 \) is subadditive.

In the next lecture, we will study conditions under which \( \kappa(a) \) is positive and finite. The random version of the subadditive limit theorem is stated below without proof.

**Theorem 6.14 (Subadditive ergodic theorem).** Let \( \{X_{m,n} \mid n > m \geq 0\} \) be a collection of rvs indexed by integers \( n > m \geq 0 \), such that
\[
X_{0,n} \leq X_{0,m} + X_{m,n} \quad \text{a.s. for } 0 < m < n.
\]
Assume that
\[
(a) \quad \text{The joint distributions of } \{X_{m+1,m+k+1} \mid k \geq 1\} \text{ are the same as those of } \{X_{m,m+k} \mid k \geq 1\} \text{ for each } m \geq 0,
\]
\[
(b) \quad \text{For each } k \geq 1, \{X_{nk,(n+1)k} \mid n \geq 1\} \text{ is a stationary process}
\]
\[
(c) \quad \text{For each } n, E|X_{0,n}| < \infty \text{ and } E X_{0,n} \geq -cn \text{ for some constant } c.
\]
Then, \( \frac{1}{n} X_{0,n} \) converges a.s. and in \( L^1 \) to a r.v. \( X \) and \( EX \in [-c, \infty) \). Moreover, if the processes in (b) are ergodic, then \( X \) is a constant a.s.

### 6.4 Large Deviation Principle

Let \( X_1, X_2, \ldots \) be i.i.d. r.v.s with mean \( \mu \) and \( S_n := X_1 + X_2 + \cdots + X_n, n \geq 1 \). By SLLN, \( S_n/n \to 0 \) a.s. and thus for any \( a > \mu \), \( \lim_{n \to \infty} P(S_n/a) = 0 \). By previous result, we know that \( \kappa(a) := -\lim_{n \to \infty} \frac{1}{n} \log P(S_n/a) \) exists and \( \kappa(a) \in [0, \infty] \). One can also prove the following.
Exercise 6.15. The following are equivalent: (a) \( \kappa(a) = \infty \), (b) \( \mathbb{P}(X_1 \geq a) = 0 \), and (c) \( \mathbb{P}(S_n \geq na) = 0 \) for all \( n \geq 1 \).

We will use Large Deviation Principle (LDP) to find \( \kappa(\cdot) \). Assume that

\[
m(\theta) := \mathbb{E} e^{\theta X_1} < \infty \text{ if } \theta \in \Theta \text{ and is infinite if } \theta \in \Theta^c,
\]

where \( \Theta \) is an open interval containing 0. Using Hölder’s inequality, we can prove that \( \log m(\theta) \) is a convex function (and thus \( \Theta \) must be an interval).

The Legendre transform of \( \log m(\cdot) \) is defined as,

\[
\ell(a) := \sup_{\theta \in \Theta} (a \theta - \log m(\theta)) \geq 0
\]

which is also convex in \( a \). If \( m(\cdot) \) is differentiable in \( \Theta \), then \( \ell(a) = a \theta^* - \log m(\theta^*) \) where \( \theta^* \) satisfies \( m'(\theta^*) = a m(\theta^*) \).

Theorem 6.16. Let \( X_1, X_2, \ldots \) be i.i.d. r.v.s with mean \( \mu \) and the function \( \ell(\cdot) \) is defined as above. Then for \( a > \mu \), we have

\[
\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n / n \geq a) = -\ell(a).
\]

Same result holds for \( \mathbb{P}(S_n / n \leq a) \) with \( a < \mu \).

Proof. We will only prove the upper bound for \( \mathbb{P}(S_n / n \geq a) \). The lower bound is bit complicated and involves change of measure. Wlog, we can assume that \( \mu = 0 \), o.w. work with \( X_1 - \mu \). By Markov’s inequality, we have

\[
\mathbb{P}(S_n \geq an) = \mathbb{P}(e^{\theta S_n} \geq e^{\theta an}) \leq e^{-a \theta n} \mathbb{E} e^{\theta S_n} = e^{-a \theta n} \prod_{i=1}^{n} \mathbb{E} e^{\theta X_i} = e^{-a \theta n} (m(\theta))^n,
\]

for \( \theta \in [0, \infty) \cap \Theta \). This implies that, \( \log \mathbb{P}(S_n / n \geq a) \leq -n(a \theta - \log m(\theta)) \) for \( \theta \in [0, \infty) \cap \Theta \). Hence

\[
\frac{1}{n} \log \mathbb{P}(S_n / n \geq a) \leq -\sup_{\theta \in [0, \infty) \cap \Theta} (a \theta - \log m(\theta))
\]

Note that, by Jensen’s inequality we have \( \log m(\theta) \geq 0 \) for all \( \theta \in \Theta \) and for \( a > 0, \theta < 0 \) we have \( \theta a - \log m(\theta) \leq 0 \). Thus, we have

\[
\sup_{\theta \in [0, \infty) \cap \Theta} (a \theta - \log m(\theta)) = \sup_{\theta \in \Theta} (a \theta - \log m(\theta)) = \ell(a).
\]

Example 6.17. Let \( X_1, X_2, \ldots \) be i.i.d. \( N(0, 1) \) r.v.s with mean 0, variance 1 and let \( S_n := X_1 + X_2 + \cdots + X_n, n \geq 1 \). Then

\[
m(\theta) = \mathbb{E} e^{\theta X_1} = e^{\theta^2 / 2} \text{ for } \theta \in \mathbb{R}.
\]

Thus,

\[
\ell(a) = \sup_{\theta \in \mathbb{R}} (a \theta - \log m(\theta)) = \sup_{\theta \in \mathbb{R}} (a \theta - \theta^2 / 2) = a^2 / 2
\]

for \( a \in \mathbb{R} \). Thus, for \( a > 0 \) we have

\[
\frac{1}{n} \log \mathbb{P}(S_n / n \geq a) \to -a^2 / 2.
\]
Example 6.18. Let $X_1, X_2, \ldots$ be i.i.d. Poisson($\lambda$) r.v.s with mean $\lambda > 0$ and let $S_n := X_1 + X_2 + \cdots + X_n, n \geq 1$. Then

$$m(\theta) = \mathbb{E} e^{\theta X_1} = e^{\lambda(e^\theta - 1)} \text{ for } \theta \in \mathbb{R}.$$ 

Thus,

$$\ell(a) = \sup_{\theta \in \mathbb{R}} (a\theta - \log m(\theta)) = \sup_{\theta \in \mathbb{R}} (a\theta - \lambda(e^\theta - 1)) = a \log(a/e\lambda) + \lambda$$

for $a > 0$. Thus, for $t > 1$ we have

$$\frac{1}{n} \log \mathbb{P}(S_n/n \geq t\lambda) \to -\lambda(t \log t - t + 1).$$
7.1 Convergence in Distribution

**Definition 7.1.** Let \((X_n)\) be a sequence of random variables. Let \(F_n(x), F(x)\) be the CDF of \(X_n\) and \(X\), respectively. Then \(X_n\) converges to \(X\) in distribution, if \(F_n(x)\) converges to \(F(x)\) at all continuity points \(x\) of \(F\), i.e.,

\[
X_n \to X \text{ in distribution} \iff F_n(x) \to F(x) \quad \forall x \text{ s.t. } F \text{ is continuous at } x,
\]

Convergence in distribution is also denoted by \(X_n \overset{(d)}{\to} X\).

**Lemma 7.2.** Almost sure convergence implies convergence in probability and convergence in probability implies convergence in distribution.

\[
a.s \implies \text{in } \mathbb{P} \implies \text{in distribution.}
\]

**Proof.** (a.s. \(\implies\) in \(\mathbb{P}\)) Proved in the homework.

(in \(\mathbb{P} \implies\) in dist.) Suppose \(X_n \overset{p}{\to} X\). Let \(F_n(X)\) denote the CDF of \(X_n\) and \(F(x)\) denote the CDF of \(X\). Then,

\[
|F_n(x) - F(x)| = |\mathbb{P}(X_n \leq x) - \mathbb{P}(X \leq x)| \leq E[1_{X_n \leq x} - 1_{X \leq x}]
\]

\[
\leq E(1_{|X_n - X| > \varepsilon} + 1_{x - \varepsilon < X \leq x + \varepsilon})
\]

\[
= \mathbb{P}(|X_n - X| > \varepsilon) + \mathbb{P}(x - \varepsilon < X \leq x + \varepsilon).
\]

As \(n \to \infty\),

\[
\limsup_{n \to \infty} |F_n(x) - F(x)| \leq 0 + \mathbb{P}(x - \varepsilon < X \leq x + \varepsilon) = F(x + \varepsilon) - F(x - \varepsilon)
\]

where convergence in \(\mathbb{P}\) is used. If \(x\) is continuous point of \(F(x)\), the RHS is zero as \(\varepsilon \to 0\).

**Example 7.3.** Consider \(X_n = \delta_{1/n}, n \geq 1\) with \(\mathbb{P}(X_n = 1/n) = 1\) and \(F_n(x) = 1_{\{x \geq 1/n\}}\). Then \(X_n \overset{a.s.}{\to} X = \delta_0, \mathbb{P}(X = 0) = 1, F(x) = 1_{\{x \geq 0\}}\). Indeed, \(F_n\) converges to \(F\) at all continuity points \(x \neq 0\) of \(F\), i.e., \(F_n(x) \to F(x)\), \(\forall x \neq 0\). But at \(x = 0\), \(F_n(0) = 0 \forall n\), but \(F(0) = 1\). In general, it is not necessarily true that \(F_n\) wouldn't converge to \(F\) at discontinuity points. For example, let \(X_n = \delta_{-1/n}\). Then \(F_n(x) \to F(x)\) for all \(x\), even \(x = 0\).

**Theorem 7.4.** Let \((X_n)\) be a sequence of r.v.s. Then \(X_n \overset{p}{\to} X\) iff,

\[
\forall \text{ sequence } (n_k)_{k \geq 1}, \exists \text{ sub-sequence } (n_{k_i})_{i \geq 1} \text{ s.t. } X_{n_{k_i}} \overset{a.s.}{\to} X \text{ as } i \to \infty.
\]
Proof. (\(\implies\)) WLOG assume \(X = 0\). By assumption, we have \(\lim_{n \to \infty} P(|X_n| > \varepsilon) = 0\) for all \(\varepsilon > 0\). Then for any given sub-sequence \((X_{n_k})\), choose \(n_{k_i}\) s.t,

\[P(|X_{n_k}| > i^{-2}) < i^{-2} \text{ for all } i \geq 1.\]

Then \(\sum_{i=1}^{\infty} P(|X_{n_k}| > i^{-2}) < \infty\) and by first Borel-Cantelli Lemma,

\[P(|X_{n_k}| > i^{-2} \text{ i.o.}) = 0,\]

which is equivalent to, \(X_{n_k} \xrightarrow{a.s.} 0\) as \(i \to \infty\).

(\(\impliedby\)) Proof by contradiction. Suppose \(X_n \not\to 0\) in Probability. Thus there exists \(\varepsilon > 0\), such that \(P(|X_n| > \varepsilon) \not\to 0\). So there exist \(\delta > 0\) and a sub-sequence \((X_{n_k})\) s.t,

\[P(|X_{n_k}| > \varepsilon) > \delta, \forall k \tag{7.1}\]

However by assumption, there exist a further sub-sequence \((X_{n_{k_i}})\) s.t \(X_{n_{k_i}} \xrightarrow{a.s.} 0 \implies X_{n_{k_i}} \xrightarrow{p} 0\) which contradicts (7.1).

**Theorem 7.5.** Let \((X_n)\) be a sequence of r.v.s. Then \(X_n \xrightarrow{d} X\) iff, there exists a sample space \((\Omega, F, P)\) and r.v.s \((Y_n)\) such that,

\[X_n \xrightarrow{d} Y_n, \quad X \xrightarrow{d} Y, \quad Y_n \xrightarrow{a.s.} Y\]

Proof. \(\Leftarrow\) is easy. For \(\Rightarrow\) consider the probability space \(((0, 1), \mathcal{B}, P = \lambda)\) where \(\lambda\) is the Lebesgue measure and the rv, \(U(\omega) = \omega, \omega \in \Omega\). Let \(F_n(x)\) be CDF of \(X_n\) and \(F(x)\) be CDF of \(X\). Then,

\[F_n^{-1}(U) \xrightarrow{a.s.} F^{-1}(U),\]

since by assumption \(F_n(x)\) converges to \(F(x)\) for all continuity points of \(F\). Define,

\[Y_n := F_n^{-1}(U), \quad Y := F^{-1}(U)\]

Then \(Y_n \xrightarrow{a.s.} Y\).

**Exercise 7.6.** Show that if \(F_n(x)\) converges to \(F(x)\) for all continuity points of \(F\), then, \(F_n^{-1}(y) \to F^{-1}(y)\) for all \(y \in (0, 1)\) but a countable number of points.

**Corollary 7.7.** Let \((X_n)_{n \geq 1}\) be a sequence of r.v.s, and \(f\) be a bounded continuous function. If \(X_n \xrightarrow{d} X\) then,

\[E(f(X_n)) \to E(f(X)). \tag{7.2}\]

Proof. Use Theorem 7.5 to construct \(Y_n \xrightarrow{a.s.} Y\). Then since \(f\) is continuous, \(f(Y_n) \xrightarrow{a.s.} f(Y)\). Then since \(f\) is bounded, use BCT to show,

\[E(f(Y_n)) \to E(f(Y))\]

which is equivalent to (7.2).
**Definition 7.8.** Let \( P, P_n, n \geq 1 \) be a sequence of probability measures. Then \( P_n \) converges weakly to \( P \), i.e., \( P_n \xrightarrow{w} P \), if

\[
\int f \, dP_n \to \int f \, dP \quad \text{for all bounded and continuous function} \ f.
\]

**Theorem 7.9.** The following are equivalent,

(i) \( X_n \xrightarrow{d} X \).

(ii) \( X_n \xrightarrow{w} X \).

(iii) \( E(f(X_n)) \to E(f(X)) \), for all \( f \in C_c^\infty(\mathbb{R}) \).

(iv) \( P(X_n \in C) \to P(X \in C) \), for all closed subsets \( C \subset \mathbb{R} \) such that \( P(X \in \partial C) = 0 \).

(v) \( P(X_n \in O) \to P(X \in O) \), for all open subsets \( O \subset \mathbb{R} \) such that \( P(X \in \partial O) = 0 \).

**Proof.** ((iii) \( \to \) (i)) Let \( F(x) = P(X \leq x) \) be CDF of \( X \). Let \( x \) be a fixed continuity point of \( F \). Fix \( \varepsilon > 0 \) small. We can approximate \( f_0(y) := 1_{y \leq x} \) by a \( C^\infty \) function,

\[
f_\varepsilon(y) = \begin{cases} 
1 & y \leq x - \varepsilon \\
g((x - y)/\varepsilon) & y \in (x - \varepsilon, x) \\
0 & y \geq x
\end{cases}
\]

where

\[
g(y) = \frac{\int_0^y e^{-1/t(1-t)} dt}{\int_0^1 e^{-1/t(1-t)} dt} \in [0, 1], \quad y \in [0, 1].
\]

Then, \( 0 \leq f_0(y) - f_\varepsilon(y) \leq 1_{y \in [x - \varepsilon, x]} \) and

\[
F_n(x) - F(x) = E( f_0(X_n) - f_0(X) ) \geq E( f_\varepsilon(X_n) - f_0(X) ) \\
\geq E( f_\varepsilon(X_n) - f_\varepsilon(X) + f_\varepsilon(X) - f_0(X) ) \\
\geq E( f_\varepsilon(X_n) - f_\varepsilon(X) ) - P(X \in [x - \varepsilon, x]).
\]

Note that, even though \( f_\varepsilon \in C^\infty(\mathbb{R}) \), it is not \( C_c^\infty \). We choose \( K > 0 \) large and define

\[
f_{\varepsilon, K}(y) = \begin{cases} 
1 & x - K < y \leq x - \varepsilon \\
g((x - y)/\varepsilon) & y \in (x - \varepsilon, x) \\
g(y - x + K + 1) & y \in (x - K - 1, x - K) \\
0 & \text{otherwise}
\end{cases}
\]

Then \( 0 \leq f_\varepsilon(y) - f_{\varepsilon, K}(y) \leq 1_{y \leq x - K} \) and \( f_\varepsilon \in C_c^\infty \). Thus

\[
F_n(x) - F(x) \geq E( f_{\varepsilon, K}(X_n) - f_{\varepsilon, K}(X) ) - P(X \in [x - \varepsilon, x]) \\
\geq E( f_{\varepsilon, K}(X_n) - f_{\varepsilon, K}(X) ) - P(X \in [x - \varepsilon, x]) - P(X \leq x - K).
\]

Taking the limit \( n \to \infty \),

\[
\liminf_{n \to \infty} (F_n(x) - F(x)) \geq -P(X \in [x - \varepsilon, x]) - P(X \leq x - K) = F(x - \varepsilon) - F(x) - F(x - K)
\]
where the assumption about the convergence of expectation of $C^\infty_c$ functions is used. Taking the limit $\varepsilon \to 0, K \to \infty$, we get
\[
\liminf_{n \to \infty} (F_n(x) - F(x)) \geq 0 \tag{7.3}
\]
where the fact that $F$ is continuous at $x$ is used. To obtain an upper bound, we use the fact that $F_n(x) - F(x) = \mathbb{E}(1_{X>x} - 1_{X_n>x})$ and approximate $1_{y>x}$ by $f_{\varepsilon,K}(x-y)$ to get
\[
\limsup_{n \to \infty} (F_n(x) - F(x)) \leq 0.
\]
This combined with (7.3) would yield the desired result. $\lim_{n \to \infty} F_n(x) = F(x)$. $\blacksquare$

### 7.2 Central Limit Theorems

**Definition 7.10 (Normal distribution).** $X \sim N(0,1)$ if density function of $X$ is,
\[
\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}
\]
and the distribution function is $\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$ and $X \sim N(\mu, \sigma^2)$ if $\frac{X-\mu}{\sigma} \sim N(0,1)$.

**Proposition 7.11.** Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. If $X_1 \perp X_2$, then
\[
X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).
\]
In particular if $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, then $X_1 + \ldots + X_n \sim N(n\mu, n\sigma^2)$ or,
\[
\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1).
\]

**Theorem 7.12 (Basic Central Limit Theorem).** If $X_1, \ldots, X_n$ are i.i.d. random variables with $\mathbb{E}(X_1) = \mu$ and $\text{Var}(X_1) = \sigma^2$. Then,
\[
\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n\sigma^2}} \overset{d}{\to} Z \sim N(0,1).
\]

**Theorem 7.13 (Lindeberg’s Central Limit Theorem).** Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$. Define $s_n^2 := \sum_{i=1}^{n} \sigma_i^2$. Then,
\[
\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \overset{(d)}{\Rightarrow} N(0,1),
\]
if $s_n \to \infty$ and $\forall \varepsilon > 0$,
\[
\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left(|X_i - \mu_i|^2 1_{|X_i - \mu_i| \geq \varepsilon s_n}\right) \to 0.
\]
8.1 Central Limit Theorems

Definition 8.1 (Normal distribution). \(X \sim N(0,1)\) if density function of \(X\) is,
\[
\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}
\]
and the distribution function is \(\Phi(x) = \int_{-\infty}^{x} \phi(t) dt\).

Proposition 8.2. Let \(X_1 \sim N(\mu_1, \sigma_1^2)\) and \(X_2 \sim N(\mu_2, \sigma_2^2)\). If \(X_1 \perp X_2\), then
\[
X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).
\]
In particular if \(X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)\), then \(X_1 + \ldots + X_n \sim N(n\mu, n\sigma^2)\) or,
\[
\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1).
\]

Theorem 8.3 (Basic Central Limit Theorem). If \(X_1, \ldots, X_n\) are i.i.d. random variables with \(E(X_1) = \mu\) and \(\text{Var}(X_1) = \sigma^2\), then,
\[
\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} Z \sim N(0,1).
\]

Theorem 8.4 (Lindeberg’s Central Limit Theorem). Let \(X_1, \ldots, X_n\) be independent random variables with \(E(X_i) = \mu_i\) and \(\text{Var}(X_i) = \sigma_i^2\). Define \(s_n^2 := \sum_{i=1}^{n} \sigma_i^2\). Then,
\[
\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{(d)} N(0,1),
\]
if \(s_n \to \infty\) and \(\forall \varepsilon > 0\),
\[
\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E} \left( |X_i - \mu_i|^2 1_{|X_i - \mu_i| \geq \varepsilon s_n} \right) \to 0.
\]

8.1.1 Triangular Arrays

Roughly speaking, a sum of many small independent random variables will be approximately normally distributed. To formulate such a limit theorem, we must consider a sequence of sums of more and more, smaller and smaller random variables. Therefore, throughout this section we shall study the sequence of sums
\[
S = \sum_{j} X_{ij}
\]
obtained by summing the rows of a **triangular array** of random variables

\[
\begin{align*}
X_{11}, X_{12}, \ldots, X_{1n_1} \\
X_{21}, X_{22}, \ldots, X_{2n_2} \\
X_{31}, X_{32}, \ldots, X_{3n_3} \\
\vdots & \vdots & \vdots & \vdots
\end{align*}
\]

It will be assumed throughout that the triangular arrays we consider satisfy three **Triangular Array Conditions** (here \(i\) ranges over \(\{1, 2, \ldots\}\), and \(j\) ranges over \(\{1, 2, \ldots, n_i\}\)):

1. For each \(i\), the \(n_i\) random variables \(X_{i1}, X_{i2}, \ldots, X_{im_i}\) in the \(i\)th row are mutually independent.
2. \(\mathbb{E} X_{ij} = 0\) for all \(i, j\), and
3. \(\sum_j \mathbb{E} X_{ij}^2 = 1\) for all \(i\).

We have some remarks for these conditions:

- It is **not** assumed that random variables in each row are identically distributed.
- It is **not** assumed that different rows are independent. In fact, a common application of triangular arrays is sums \(X_1 + X_2 + \cdots + X_n\) obtained from a sequence of independent random variables \(X_1, X_2, \ldots\).
- It will usually be the case that \(n_i \to \infty \) as \(i \to \infty\). And according to the nature of our problem, we should have the variables in each row tend to be smaller and smaller as \(i\) increases. Both of these two conditions are implied by the Lindeberg Condition which we will discuss below.

### 8.1.2 The Lindeberg Condition and Some Consequences

**Theorem 8.5 (Lindeberg’s Theorem).** Suppose that in addition to the Triangular Array Conditions, the triangular array satisfies Lindeberg’s condition:

\[
\forall \varepsilon > 0, \quad \lim_{i \to \infty} \sum_{j=1}^{n_i} \mathbb{E}[X_{ij}^2 \mathbbm{1}(|X_{ij}| > \varepsilon)] = 0. \tag{8.1}
\]

Then \(S_i \xrightarrow{(d)} N(0, 1)\).

The Lindeberg condition makes precise the sense in which the random variables must be smaller and smaller. It says that for arbitrarily small \(\varepsilon > 0\), the contribution to the total row variance from the terms with absolute value greater than \(\varepsilon\) becomes negligible as you go down the rows. We see this as follows:

\[
X_{ij}^2 \leq \varepsilon^2 + X_{ij}^2 \mathbbm{1}(|X_{ij}| > \varepsilon) \\
\mathbb{E} X_{ij}^2 \leq \varepsilon^2 + \mathbb{E} X_{ij}^2 \mathbbm{1}(|X_{ij}| > \varepsilon) \leq \varepsilon^2 + \sum_{k \geq 1} \mathbb{E} X_{ik}^2 \mathbbm{1}(|X_{ik}| > \varepsilon).
\]

---

\(^1\)This is not standard terminology, but is used here as a simple referent for these conditions.
This last inequality is true for all \( j \), so we have:

\[
\max_j E X_{ij}^2 \leq \varepsilon^2 + \sum_j E X_{ij}^2 \mathbb{1}(|X_{ij}| > \varepsilon) \tag{8.2}
\]

The Lindeberg condition says that, as \( i \to \infty \), the summation on the RHS of (8.2) tends to zero. Since (8.2) holds for all \( \varepsilon > 0 \), we get

\[
\lim_{i \to \infty} \max_j E X_{ij}^2 = 0, \tag{8.3}
\]

which implies \( n_i \to \infty \) as \( i \to \infty \), since we assume in Triangular Array Condition that \( \sum_j E X_{ij}^2 = 1 \) for all \( i \). Another consequence follows from (8.3) and Chebyshev’s inequality: since we have

\[
P(|X_{ij}| > \varepsilon) \leq \frac{E X_{ij}^2}{\varepsilon^2} \quad \text{for all } \varepsilon > 0,
\]

taking the maximum over \( j \) and \( i \to \infty \), we get that \( X_{ij} \xrightarrow{p} 0 \), uniformly in \( j \):

\[
\forall \varepsilon > 0, \lim_{i \to \infty} \max_j P(|X_{ij}| > \varepsilon) = 0. \tag{8.4}
\]

An array with property (8.4) is said to be Uniformly Asymptotically Negligible (UAN), and there is a striking converse to Lindeberg’s Theorem:

**Theorem 8.6** (Feller’s Theorem). If a triangular array satisfies the Triangular Array Conditions and is UAN, then \( S_i \xrightarrow{d} N(0,1) \) (if and only if Lindeberg’s condition (8.1) holds.

**Proof.** See Billingsley, Theorem 27.4, or Kallenberg, 5.12.

**Proof of Basic Central Limit Theorem.** We show that the Lindeberg condition holds by taking

\[
X_{nj} = \frac{X_j - \mu}{\sqrt{n\sigma^2}}, \quad j = 1, 2, \ldots, n.
\]

Then

\[
\sum_{j=1}^{n} E[X_{nj}^2 \mathbb{1}(|X_{nj}| > \varepsilon)] = \frac{n}{n\sigma^2} \sigma^2 \mathbb{1}(|X_1 - \mu| > \varepsilon \sqrt{n})
\]

which converges to 0 by DCT.

**8.1.3 Rate of Convergence for CLT**

**Theorem 8.7** (Berry-Esseen Theorem). Let \( X_1, X_2, \ldots \) be i.i.d. r.v.s with \( E(X_1) = 0, \text{Var}(X_1) = \sigma^2 \) and \( E|X_1|^3 < \infty \). Define \( S_n := (X_1 + X_2 + \cdots + X_n)/\sqrt{n\sigma^2} \). Then

\[
\sup_{x \in \mathbb{R}} |P(S_n \leq x) - \Phi(x)| \leq \frac{3 E|X_1|^3}{\sqrt{n\sigma^3}}.
\]
8.2 Preliminaries to the proof of Lindeberg’s Theorem

There are several ways of proving Central Limit Theorems:

1. Use characteristic or moment generating functions or some distributional transform, or
2. Use moment method to show that the $k$-th moment converges to the $k$-th moment of standard Normal for all $k \geq 1$, or
3. Use Fixed Point method (e.g., maximizing entropy given fixed mean and variance, zero bias transformation etc.) or
4. Replacement or exchange techniques.
5. Stein’s method (to be discussed later in the course).

We introduce two preliminaries to the proof.

**Lemma 8.8.** If $X \sim N(0, \sigma^2)$, $Y \sim N(0, \tau^2)$ are independent, then $X + Y \sim N(0, \sigma^2 + \tau^2)$.

**Lemma 8.9.** $S_i \overset{(d)}{\to} Z$ if and only if $\lim_{i \to \infty} \mathbb{E} f(S_x) = \mathbb{E} f(Z)$ for all $f \in C_b^3(\mathbb{R})$, the set of functions from $\mathbb{R}$ to $\mathbb{R}$ with three bounded, continuous derivatives.

**Proof.** See Durrett, Theorem 2.2, and use that $C_b^3(\mathbb{R})$ is dense in $C_b(\mathbb{R})$. $lacksquare$

8.3 Proof of Lindeberg’s Theorem

**Proof.** First, we will work with a fixed row. To simplify things we will avoid writing the subscript $i$, so that $X_{ij}$ will be denoted by $X_j$ and $n_i$ will be denoted by $n$.

Let $X_1, X_2, \ldots, X_n$ be independent random variables, not necessarily identically distributed. Suppose $\mathbb{E} X_j = 0$ and let $\sigma_j^2 = \mathbb{E} (X_j^2) < \infty$. Then for $S = \sum_{j=1}^{n} X_j$ we have $1 = \text{Var} S = \sum_{j=1}^{n} \sigma_j^2$.

Note that,

1. If $\forall j, X_j \sim N(0, \sigma_j^2)$, then $S \sim N(0, 1)$.

2. Given independent random variables $X_1, X_2, \ldots, X_n$ with arbitrary distributions, we can always construct a new sequence $Z_1, Z_2, \ldots, Z_n$ of Normal random variables with matching means and variances so that all of $Z_i$ and $X_i$ are mutually independent. This may involve changing the basic probability space, but that does not matter because the distribution of $S$ is determined by the joint distribution of $(X_1, X_2, \ldots, X_n)$, which remains the same.

Let

\[
S := W_0 := X_1 + X_2 + X_3 + \ldots + X_n,
W_1 := Z_1 + X_2 + X_3 + \ldots + X_n,
W_2 := Z_1 + Z_2 + X_3 + \ldots + X_n,
\vdots
\vdots
\vdots
T := W_n := Z_1 + Z_2 + Z_3 + \ldots + Z_n,
\]
We want to show that $S$ is “close” in distribution to $T$, i.e., that $E f(S)$ is close to $E f(T)$ for all $f \in C^3_b(\mathbb{R})$ with uniform bound $K$ on $f$ and its first three derivatives: $|f^{(i)}|$, $i = 1, 2, 3$.

By the triangle inequality,

$$|E f(S) - E f(T)| \leq \sum_{j=1}^{n} |E f(W_j) - E f(W_{j-1})|.$$  \hspace{1cm} (8.5)

Let $W_j^- = Z_1 + Z_2 + \cdots + Z_{j-1} + X_{j+1} + \cdots + X_n$ be the sum of the common terms in $W_{j-1}$ and $W_j$. Then

$$W_{j-1} = W_j^- + X_j \text{ and } W_j = W_j^- + Z_j.$$

Note that by construction, $W_j^-$ and $X_j$ are independent, as are $W_j^-$ and $Z_j$. We need to compare $E f(W_j^- + X_j)$ and $E f(W_j^- + Z_j)$. By the Taylor series expansion up to the third term,

$$f(W_j^- + X_j) = f(W_j^-) + X_j f^{(1)}(W_j^-) + \frac{X_j^2}{2!} f^{(2)}(W_j^-) + Err_f(W_j^-, X_j)$$

$$f(W_j^- + Z_j) = f(W_j^-) + Z_j f^{(1)}(W_j^-) + \frac{Z_j^2}{2!} f^{(2)}(W_j^-) + Err_f(W_j^-, Z_j)$$

where

$$Err_f(a, x) := f(a + x) - f(a) - xf'(a) - \frac{x^2}{2} f''(a)$$

satisfies

$$|Err_f(a, x)| \leq \min \left\{ \frac{|x|^3}{6} \cdot ||f^{(3)}||_{\infty}, x^2 \cdot ||f^{(2)}||_{\infty} \right\} \leq K x^2 \min\{1, |x|\}.$$  

Here we used the minimum bound, so that upper bound stays integrable for $x = X_j$.

We can take expectations in each of these identities and subtract the resulting equations. Using independence and the fact that $X_j$ and $Z_j$ agree on their first and second moments, we see that everything up to the second order cancels. Therefore,

$$|E f(W_j) - E f(W_{j-1})| = |E f(W_j^- + X_j) - E f(W_j^- + Z_j)|$$

$$\leq E(|Err_f(W_j^-, X_j)| + |Err_f(W_j^-, Z_j)|)$$

$$\leq K E \left( X_j^2 \min\{1, |X_j|\} + |Z_j|^3 \right). \hspace{1cm} (8.6)$$

Let $c = \sqrt{8/\pi}$ be the third absolute moment of a standard normal random variable as,

$$c := 2 \int_{0}^{\infty} x^3 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 2 \cdot \frac{2}{\sqrt{2\pi}} < \infty$$

Therefore, $E |Z_j|^3 = c \sigma_j^3$. Moreover, we have $\min\{1, |x|\} \leq \varepsilon + 1_{|x|>\varepsilon}$ for all $\varepsilon > 0$. Applying this to (8.6), we get

$$|E f(W_j) - E f(W_{j-1})| \leq K \left( \varepsilon E X_j^2 + E X_j^2 1_{|X_j|>\varepsilon} + c \sigma_j^3 \right).$$

Combining, from (8.5), we get

$$|E f(S) - E f(T)| \leq K \left( \varepsilon + \sum_j E X_j^2 1_{|X_j|>\varepsilon} + c \sum_j \sigma_j^3 \right). \hspace{1cm} (8.7)$$
So far we have only considered one row of the array, but (8.7) is in fact true for every row with \(K\) and \(c\) unchanged and \(T \sim N(0, 1)\). Thus, for each \(i\) we have,

\[
|\mathbb{E} f(S_i) - \mathbb{E} f(T)| \leq K (\varepsilon + \sum_j \mathbb{E} X_{ij}^2 |X_{ij}| > \varepsilon + c \sum_j \sigma_{ij}^3)
\]

\[
\leq K (\varepsilon + \sum_j \mathbb{E} X_{ij}^2 |X_{ij}| > \varepsilon + c \max_j \sigma_{ij}).
\] (8.8)

Under Lindeberg’s condition, the RHS of (8.8) goes to \(K\varepsilon\) as \(i \to \infty\). Since \(\varepsilon > 0\) is arbitrary, the LHS converges to zero as \(i \to \infty\). By Lemma 10.6, \(S_i \overset{(d)}{\rightarrow} N(0, 1)\) as \(i \to \infty\). 

Here, we proved that for \(f \in C^3_b(\mathbb{R})\) : \(|f''|, |f'''| < \infty\), we have

\[
|\mathbb{E} f(S_i) - \mathbb{E} f(T)| = \left| \sum_{j=1}^n \mathbb{E}(\text{Err}_f(W_{ij}, X_{ij}) - \text{Err}_f(W_{ij}, Z_{ij})) \right|
\]

\[
\leq \frac{1}{6} \sum_{j=1}^n \mathbb{E}(X_{ij}^2 \min\{6|f''|_\infty, |f'''|_\infty \cdot |X_{ij}| \} + c \cdot |f'''|_\infty \cdot \sigma_{ij}^3)
\]

In fact, we proved a stronger result than convergence in distribution. We proved that

\[
\sup_{f \in C^3_b(\mathbb{R}) : |f''|, |f'''| \leq K} |\mathbb{E} f(S_i) - \mathbb{E} f(T)| \leq K \sum_{j=1}^n (\mathbb{E}(X_{ij}^2 \min\{6|X_{ij}| \}) + c \cdot \sigma_{ij}^3)
\]

which goes to zero under Lindeberg’s condition. If we have \(\mathbb{E}|X_{ij}|^3 < \infty\), we get

\[
\sup_{f \in C^3_b(\mathbb{R}) : |f''|, |f'''| \leq K} |\mathbb{E} f(S_i) - \mathbb{E} f(T)| \leq \frac{K}{6} \sum_{j=1}^n (\mathbb{E}|X_{ij}|^3 + c \cdot \sigma_{ij}^3)
\]

\[
\leq \frac{K}{6} (1 + c) \sum_{j=1}^n \mathbb{E}|X_{ij}|^3
\]

using the fact that \(\sigma_{ij}^3 = ||X_{ij}||^3_2 \leq ||X_{ij}||^3 \leq \mathbb{E}|X_{ij}|^3\). This is similar to the Berry-Esseen bound, but here the functions \(f\) are much smoother.

**Example 8.10. (Least Square estimate in Linear Regression Model)** Let \(Y_i = \beta x_i + \eta_i\) for \(i = 1, 2, \ldots, n\) where \((x_i)_{i \geq 1}\) is a sequence of real numbers, \(\eta_i\’s\) are independent random variable with \(\mathbb{E}(\eta_i) = 0\) and \(\text{Var}(\eta_i) = \sigma^2\) for all \(i\). Here, \(\beta\) is unknown and \(\sigma^2 > 0\) is known. The least square estimate of \(\beta\) is given by

\[
\beta_{LS} := \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.
\]

We have

\[
\sqrt{\frac{\sum_{i=1}^n x_i^2}{\sigma^2}} \cdot (\beta_{LS} - \beta) \overset{(d)}{\rightarrow} N(0, 1)
\]

if for some \(\varepsilon > 0\)

\[
\max_{i \geq 1} \mathbb{E}|\eta_i|^{2+\varepsilon} < \infty\quad\text{and}\quad \frac{\max_{1 \leq i \leq n} x_i^2}{\sum_{i=1}^n x_i^2} \to 0\quad\text{as}\quad n \to \infty.
\]
8.4 Poisson Convergence and Law of Small Numbers

By Central Limit Theorem,
\[
\frac{S_n - np}{\sqrt{np(1-p)}} \overset{(d)}{\to} N(0, 1)
\]
as \(n \to \infty\), where \(X_1, X_2, \ldots\) are iid Bernoulli\((p)\) rvs and \(S_n = X_1 + X_2 + \cdots + X_n\), \(n \geq 1\). In general, if \(p = p_n \to 0\) and \(np_n(1-p_n) \to \infty\) or \(\frac{1}{n} \ll p_n \ll 1\), the convergence still holds.

If \(np_n \to 0\), we have
\[
E S_n^2 = (np_n)^2 + np_n(1-p_n) \to 0 \text{ and hence } S_n \overset{1^2}{\to} 0.
\]

What if \(np_n \to \lambda > 0\)?

**Definition 8.11.** A r.v. \(X\) is Poisson\((\lambda)\) distributed if
\[
\Pr(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots.
\]

**Lemma 8.12.** If \(np_n \to \lambda \in (0, \infty)\), then \(S_n \overset{(d)}{\to} \text{Poisson}(\lambda)\).

**Proof.** Suffices to show that, \(\Pr(S_n = k) \to e^{-\lambda} \frac{\lambda^k}{k!}\) for all \(k = 0, 1, 2, \ldots\) We can compute this quantity exactly:
\[
\Pr(S_n = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \frac{1}{k!} \prod_{i=0}^{k-1} \left(1 - \frac{i}{n}\right) (np)^k (1 - \frac{np}{n})^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}.
\]

**Theorem 8.13 (Law of Small Numbers).** Let \(\{X_{ij}, j = 1, 2, \ldots, n_i\}\) be independent, integer-valued random variables for each \(i \geq 1\). Define \(S_i := \sum_{j=1}^{n_i} X_{ij}\) for \(i \geq 1\). Assume that

(i) \(\Pr(X_{ij} = 1) = p_{ij}\) satisfies \(\sum_{j=1}^{n_i} p_{ij} \to \lambda \in (0, \infty)\),

(ii) \(\sum_{j=1}^{n_i} \Pr(X_{ij} \notin \{0, 1\}) \to 0\) and

(iii) \(\max_{1 \leq j \leq n_i} p_{ij} \to 0\)

as \(i \to \infty\). Then \(S_i \overset{(d)}{\to} \text{Poisson}(\lambda)\).

There is a more general version of the LSN relaxing the independence assumption. The **Poisson Universality Class** contains sequences of rvs satisfying the three conditions above.

**Proof Sketch:** We’ll show the total variation distance
\[
d_{TV}(S_i, N) := \sup_{A \in B} |\Pr(S_i \in A) - \Pr(N \in A)| \to 0
\]
where \(N \sim \text{Poisson}(\lambda)\) (this is stronger than convergence in distribution).

**Exercise 8.14.** (a) Prove that \(d_{TV}(\cdot, \cdot)\) is a metric on the set of probability measures on \(\Omega\).
(b) Prove that \(d_{TV}(\mu, \nu) = \frac{1}{2} \sum_k |\mu\{k\} - \nu\{k\}|\) if \(\mu, \nu\) are discrete distributions on \(\mathbb{Z}\).

Again, we fixed \(i\) and remove the subscript \(i\) for easy writing. Let
\[
\hat{X}_j = 1_{X_j = 1}, \quad S = \sum_{j=1}^{n} X_j \quad \text{and} \quad \hat{S} = \sum_{j=1}^{n} \hat{X}_j.
\]
We have
\[ d_{TV}(S, N) \leq d_{TV}(\tilde{S}, N) + d_{TV}(S, \tilde{S}) \]
and
\[ d_{TV}(S, \tilde{S}) \leq \mathbb{P}(S \neq \tilde{S}) \leq \mathbb{P}(\cup_{j=1}^{n}\{X_j \neq \tilde{X}_j\}) \leq \sum_{j=1}^{n} \mathbb{P}(X_j \notin \{0, 1\}). \]

We have \( \mathbb{P}(\tilde{X}_j = 1) = p_j \) and \( \mathbb{P}(\tilde{X} = 0) = 1 - p_j \). We proceed as follows

1. Sum of independent Poisson rvs is Poisson: \( N_j \sim \text{Poisson}(\lambda_j) \) for \( j = 1, 2, \ldots, n \) and independent, then \( \sum_{j=1}^{n} N_j \sim \text{Poisson}(\sum_{j=1}^{n} \lambda_j) \).

2. Now assume that \( \tilde{\lambda} := \sum_{j=1}^{n} p_j \). Let \( Y_j \sim \text{Poisson}(p_j) \), independent, for \( i = 1, 2, \ldots, n \). Then \( \sum_{j=1}^{n} Y_j \sim \text{Poisson}(\tilde{\lambda}) \).

3. We have
\[
\begin{align*}
d_{TV}(\tilde{X}_j, Y_j) &= \frac{1}{2} \sum_{k=0}^{\infty} | \mathbb{P}(\tilde{X}_j = k) - \mathbb{P}(Y_j = k) | \\
&= \frac{1}{2} |(1 - p_j - e^{-p_j}) + |p_j - p_je^{-p_j}| + (1 - e^{-p_j} - p_je^{-p_j})| = p_j(1 - e^{-p_j}) \leq p_j^2.
\end{align*}
\]

4. Finally, if suffices to show \( d_{TV}(X_1 + X_2, Y_1 + Y_2) \leq d(X_1, Y_1) + d(X_2, Y_2) \) when \( X_1 \perp X_2 \) and \( Y_1 \perp Y_2 \). We have,
\[
\begin{align*}
&\sum_k | \mathbb{P}(X_1 + X_2 = k) - \mathbb{P}(Y_1 + Y_2 = k) | \\
&= \sum_k | \sum_{\ell} (\mathbb{P}(X_1 = \ell) \mathbb{P}(X_2 = k - \ell) - \mathbb{P}(Y_1 = \ell) \mathbb{P}(Y_2 = k - \ell) ) | \\
&\leq \sum_{k, \ell} |(\mathbb{P}(X_1 = \ell) - \mathbb{P}(Y_1 = \ell)) \cdot (\mathbb{P}(X_2 = k - \ell) \mathbb{P}(Y_2 = k - \ell))| \\
&\leq \sum_{\ell} | \mathbb{P}(X_1 = \ell) - \mathbb{P}(Y_1 = \ell) | + \sum_{\ell} | \mathbb{P}(X_2 = \ell) - \mathbb{P}(Y_2 = \ell) |
\end{align*}
\]

So
\[
d_{TV}(\sum_{j=1}^{n} \tilde{X}_j, \sum_{j=1}^{n} Y_j) \leq \sum_{j=1}^{n} d_{TV}(\tilde{X}_j, Y_j) \leq \sum_{j=1}^{n} p_j^2.
\]

5. We have \( d_{TV}(\text{Poisson}(\tilde{\lambda}), \text{Poisson}(\lambda)) \leq \mathbb{P}(\text{Poisson}(|\tilde{\lambda} - \lambda|) > 0) \leq |\tilde{\lambda} - \lambda| \).

Combining we get
\[
\begin{align*}
d_{TV}(S, N) &\leq \sum_{j=1}^{n} \mathbb{P}(X_j \notin \{0, 1\}) + \sum_{j=1}^{n} p_j^2 + \left| \sum_{j=1}^{n} p_j - \lambda \right|.
\end{align*}
\]

Finally putting back \( i \) and using the three assumptions we have the result. \( \square \)
Remark 8.15. It follows from the previous analysis that, if $X_j \sim \text{Bernoulli}(p_j)$ for $j = 1, 2, \ldots, n$ and are independent, then

$$d_{TV}\left(\sum_j X_j, \text{Poisson}(\lambda)\right) \leq \sum_j p_j^2,$$

where $\lambda := \sum_{j=1}^n p_j$. However, the bound is not optimal. One can prove using Stein-Chen method that

$$d_{TV}\left(\sum_j X_j, \text{Poisson}(\lambda)\right) \leq \frac{\sum_j p_j^2}{\max\{1, \lambda\}} \leq \max_j p_j.$$

Moreover, if $X_j$’s are positively correlated, we have

$$d_{TV}\left(\sum_j X_j, \text{Poisson}(\lambda)\right) \leq \frac{\sum_j p_j^2 + \sum_{i \neq j} \text{Cov}(X_i, X_j)}{\max\{1, \lambda\}}$$

and if $X_j$’s are negatively correlated, we have

$$d_{TV}\left(\sum_j X_j, \text{Poisson}(\lambda)\right) \leq \frac{\sum_j p_j^2 - \sum_{i \neq j} \text{Cov}(X_i, X_j)}{\max\{1, \lambda\}}$$
9.1 Helly’s Selection Theorem

9.1.1 Extended Random Variables

Definition 9.1. An extended random variable is a measurable function $X: (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{R}^*, \mathcal{B}^*)$ where $\mathbb{R}^* = \mathbb{R} \cup \{\pm \infty\}, \mathcal{B}^* = \sigma(\mathcal{B}, \{\infty\}, \{-\infty\})$ such that

- $\mathbb{P}(X = \infty) = \lim_{x \uparrow \infty} \mathbb{P}(X > x)$ and
- $\mathbb{P}(X = -\infty) = \lim_{x \downarrow -\infty} \mathbb{P}(X \leq x)$.

Definition 9.2. An extended distribution function (EDF) is a non-decreasing, right-continuous function from $\mathbb{R}$ to $[0, 1]$.

Note that for an EDF $F$, it is a CDF iff $1 - F(x) + F(-x) \rightarrow 0$ as $x \uparrow \infty$. Convergence in distribution generalizes naturally to extended random variables:

Definition 9.3. Let $X_n, X$ be extended random variables with EDF’s $F_n$ and $F$, respectively. We say $X_n$ converges in distribution to $X$ and write $X_n \xrightarrow{d} X$, if

$$F_n(x) \rightarrow F(x)$$

as $n \rightarrow \infty$ for every continuity point $x$ of $F$.

9.1.2 Helly’s Selection Theorem

Theorem 9.4 (Helly Bray Selection theorem). Given a sequence of EDF’s $F_1, F_2, \ldots$ there exists a subsequence $(n_k)$ such that $F_{n_k} \xrightarrow{d} F$ for some EDF $F$.

To prove this theorem, we need the following lemma:

Lemma 9.5. Let $(F_n)_{n \geq 1}$ be a sequence of EDFs such that for a dense subset $D$, $\lim_{n \rightarrow \infty} F_n(d) = G(d)$ exists for all $d \in D$. Define $F_*(x) = \inf_{d \in x} G(d)$, then $F_n(x) \xrightarrow{d} F_*$.

Proof. It is easy to check that $F_*$ is an EDF. Then, for any continuity point $x$ of $F_*$, there exists $d_1, d_2 \in D$ such that $d_1 < x < d_2$ and $F_*(x) - \varepsilon < F_n(d_1) \leq F_n(x) < F_n(d_2) < F_*(x) + \varepsilon$ for $n$ large enough.

Proof of Theorem 9.4. Use Cantor’s Diagonal Argument to find a sequence $F_{n_k}$ which is convergent at any point within $Q$, then by Lemma 9.5 we are done.
Definition 9.6. A collection of distributions \( \{ P_{\lambda} : \lambda \in \Lambda \} \) on \( \mathbb{R} \) is tight if

\[
\lim_{x \to \infty} \sup_{\lambda \in \Lambda} P_{\lambda}(-x, x) = 0.
\]

Equivalently, \( \forall \varepsilon > 0, \exists \) a compact set \( B_{\varepsilon} \) such that \( P_{\lambda}(B_{\varepsilon}^c) < \varepsilon, \forall \lambda \in \Lambda \).

Theorem 9.7 (Helly’s selection theorem). Let \( (F_n)_{n \geq 1} \) be a sequence of CDFs which are tight, then there exists a subsequence \( (n_k) \) such that \( F_{n_k} \xrightarrow{(d)} F \) for some CDF \( F \).

Proof. By Helly Bray Selection Theorem, \( F_{n_k} \xrightarrow{(d)} F_* \) for some EDF \( F_* \). Given \( \varepsilon > 0 \), find two continuity points of \( F_* \), \( \pm d \), such that \( \sup_k |1 - F_{n_k}(d) + F_{n_k}(-d)| \leq \varepsilon \), then \( |1 - F_*(d) + F_*(-d)| \leq \varepsilon \). Thus \( F_* \) is a CDF.

9.2 Metric on the space of Probability Measures

Take \( \{ f_n, n \geq 1 \} \) a countable sequence of continuous functions bounded by 1, which is dense in \( C_b(\mathbb{R}) \). Define

\[
d(\mu, \gamma) := \sum_{k \leq 1} \frac{1}{2^k} \left| \int f_k d\mu - \int f_k d\gamma \right|,
\]

then \( d(\mu_n, \mu) \to 0 \) iff \( \mu_n \xrightarrow{(d)} \mu \).

Definition 9.8 (Lévy metric). Let \( F \) and \( G \) be two CDFs, define

\[
d_L(F, G) := \inf \{ \varepsilon : F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon \},
\]

then \( d_L(F, G) \to 0 \) iff \( F_n \xrightarrow{(d)} F \).

Thus the space of probability measures on \( \mathbb{R} \) with convergence in distribution is metrizable and is complete by Helly’s selection theorem.

Definition 9.9 (Determining Class). Let \( \mathcal{C} \) be a collection of bounded continuous functions. \( \mathcal{C} \) is called determining class if \( \int f d\mu = \int f d\gamma \ \forall \ f \in \mathcal{C} \implies \mu = \gamma \).

Example 9.10. Define

\[
f_{x, \varepsilon}(y) = \begin{cases} 
1 & y \leq x \\
0 & y > x + \varepsilon \\
1 - (y - x)/\varepsilon & x < y < x + \varepsilon 
\end{cases}
\]

Then \( \mathcal{C} := \{ f_{x, 1/k} : x \in \mathbb{Q}, k \leq 1 \} \) is a determining class, since for any CDF \( F, x \in \mathbb{Q} \), we have \( F(x) = \lim_{k \to \infty} \int f_{x, 1/k}(y) dF(y) \).

Example 9.11. \( \{ x \mapsto \sin(\theta x), x \mapsto \cos(\theta x), \theta \in \mathbb{R} \} \) is a determining class.

Theorem 9.12. Let \( \mathcal{C} \) be a determining class and \( (F_n)_{n \geq 1} \) be a tight sequence of CDFs. If \( \int g(x) dF_n(x) \) converges for \( \forall g \in \mathcal{C} \), then \( F_n \xrightarrow{(d)} F \) for some CDF \( F \) and

\[
\lim_{n \to \infty} \int g(x) dF_n(x) = \int g(x) dF(x).
\]
To prove Theorem 9.12, we need the following lemma.

**Lemma 9.13.** Let \((S, d)\) be a metric space and \(x_n\) be a sequence in \(S\), then \(x_n \rightarrow x\) iff \(∀\) subsequence \((n_k)\), \(∃\) a further subsequence \((m_k)\) such that \(x_{m_k} \rightarrow x\).

**Proof.** Only if part is easy. For the if part, assume that \(x_n \not\rightarrow x\), then \(∃\varepsilon > 0, \exists\) a subsequence such that \(d(x_{n_k}, x) > \varepsilon, ∀k\). Contradiction. \(\blacksquare\)

**Proof of Theorem 9.12.** Now, we apply Lemma 9.13 to prove Theorem 9.12. By Helly Selection Theorem, \(∃\) a subsequence \((n_k)\) such that \(F_{n_k} \overset{(d)}{\rightarrow} F\) for some CDF \(F\) and thus

\[
\int g(x)dF(x) = \lim_{n \rightarrow \infty} \int g(x)dF_n(x), \forall g \in C.
\]

If \(F_{m_k} \overset{(d)}{\rightarrow} G\) for some subsequence \(m_k\), then \(\int g(x)dG(x) = \lim_{n \rightarrow \infty} \int g(x)dF_n(x), \forall g \in C\). Then \(F = G\) since \(C\) is a determining class. By Lemma 9.13, \(F_n \overset{(d)}{\rightarrow} F\) and \(\lim_{n \rightarrow \infty} \int g(x)dF_n(x) = \int g(x)dF(x)\). \(\blacksquare\)

Consider all random variables on the probability space \((Ω, F, P)\), namely, \(X := \{X : (Ω, F, P) → (R, B) | X \text{ measurable}\}\). We have metrizable spaces \(L^p(Ω, F, P) \subseteq X\) for \(p \geq 1\), which guarantee the convergence of r.v.s in probability, i.e., convergence w.r.t. the norm of the spaces implies the convergence in probability. We can also consider the metric \(d_p(X, Y) := E\min\{1, |X - Y|\}\) which induces convergence in probability. However, in general almost sure convergence is not metrizable. Using Lemma 9.13 one can prove that, if almost sure convergence is metrizable, it will be equivalent to convergence in probability. However, Egorov’s theorem states that a.s. convergence implies almost uniform convergence.

For convergence in distribution, we have the similar metrizable spaces guaranteeing the convergence. Let \(M(Ω, F) = \{P | P\text{ is a probability measure on } (Ω, F)\}\). We have seen Lévy metric, Total Variation metric, Kolmogorov-Smirnov metric and Kantarovich-Wasserstein metric for convergence in distribution on \((R, B)\). In general, let \(D\) be a determining class of bounded continuous functions, then we can define distance

\[
d_D(\mu, ν) = \sup_{f ∈ D} |\int f d\mu - \int f dν|.
\]

Then, the metric mentioned above could be defined with corresponding determining classes:

- Total Variation metric: \(D = \{f ∈ C_b(R) | |f|_∞ ≤ 1\}\),
- Kolmogorov-Smirnov metric: \(D = \{1_{(-∞, x]} | x ∈ R\}\),
- Kantarovich-Wasserstein metric: \(D = \{f ∈ C_b^1 | |f'|_∞ ≤ 1\}\).

Recall that \(M = M(R, B)\) with convergence in distribution is metrizable under Lévy metric and is complete by Helly’s selection theorem. In general, all the other metrics mentioned are stronger than Lévy metric.
10.1 The Characteristic Function

**Definition 10.1.** For a random variable $X$, define the characteristic function of $X$, $\varphi_X : \mathbb{R} \to \mathbb{C}$, by

$$\varphi_X(t) := \mathbb{E}(e^{itX}) = \mathbb{E}(\cos tX) + i \mathbb{E}(\sin tX), \quad t \in \mathbb{R}.$$ 

**Proposition 10.2.** Let $X, Y$ be random variables. Then

(i) $\varphi_X(0) = 1$.

(ii) $|\varphi_X| \leq 1$.

(iii) $\forall n \geq 0$, if $\mathbb{E}|X|^n < \infty$, then $\varphi_X \in C^n(\mathbb{R})$.

(iv) If $X \perp Y$, $\varphi_{X+Y} = \varphi_X \cdot \varphi_Y$.

(v) $\forall t \in \mathbb{R}$, $\varphi_X(t) = \varphi_X(-t)$.

(vi) $\forall a, b, t \in \mathbb{R}$, $\varphi_{aX+b}(t) = e^{ibt}\varphi_X(at)$.

(vii) $\mathbb{E}(\varphi_X(Y)) = \mathbb{E}(\varphi_Y(X))$, and if $X \perp Y$, they are same as $\mathbb{E}(\varphi_{XY}(1))$.

**Example 10.3.**

(i) $X \sim \text{Uniform}(0,1)$. $\varphi_X(t) = \int_0^1 e^{itx} dx = \frac{e^{it} - 1}{it}$.

(ii) $X \sim \text{Exponential}(\lambda)$. $\varphi_X(t) = \int_0^\infty e^{itx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - it}$.

(iii) $X = X_1 - X_2$, $X_1 \perp X_2$, $X_1, X_2 \sim \text{Exponential}(1)$.

$$\varphi_X(t) = \varphi_{X_1-X_2}(t) = \varphi_{X_1}(t)\varphi_{X_2}(t) = \frac{1}{1 - it} \frac{1}{1 + it} = \frac{1}{1 + t^2}$$

The density of $X$ is given by $f(x) = \frac{1}{2} e^{-|x|}$, known as the “Laplace density”.

(iv) $X \sim \mathcal{N}(0,1)$. $\varphi_X(t) = e^{-\frac{t^2}{2}}$.

1st Proof of (iv): For each $\theta \in \mathbb{R}$,

$$\mathbb{E}(e^{\theta X}) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{x^2}{2} + \theta x} dx = \frac{e^{\frac{\theta^2}{2}}}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(x-\theta)^2}{2}} dx = e^{\frac{\theta^2}{2}}.$$ 

Now, it can be checked that the functions from $\mathbb{C}$ to $\mathbb{C}$ defined $z \mapsto \mathbb{E}(e^{zX})$ and $z \mapsto e^{\frac{z^2}{2}}$ are analytic, and the previous calculation shows that they agree on $\mathbb{R}$. The identity principle then implies that the must agree everywhere, so in particular $\varphi_X(t) = \mathbb{E}(e^{itX}) = e^{\frac{(it)^2}{2}} = e^{-\frac{t^2}{2}}$. 

10-1
2\textsuperscript{nd} Proof of (iv): \( \phi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx = e^{-t^2/2} \) by a contour integral argument. ■

3\textsuperscript{rd} Proof of (iv): By Gaussian integration by parts, \( \mathbb{E}(Xg(X)) = \mathbb{E}(g'(X)) \) for all differentiable \( g \) with \( \|g\|_\infty + \|g'\|_\infty < \infty \). Applying this to \( g(x) = e^{itx} \) yields \( \mathbb{E}(Xe^{itX}) = \mathbb{E}(itX) \). Then \( \varphi'_X(t) = \mathbb{E}(iXe^{itX}) = \mathbb{E}(-te^{itX}) = -t\varphi_X(t) \). \( \varphi_X \) also has initial data \( \varphi_X(0) = 1 \), and so the unique solution to this initial value problem is \( \varphi_X(t) = e^{-t^2/2} \). ■

4\textsuperscript{th} Proof of (iv): Let \( (X_i)_{i=1}^\infty \) be iid with distribution \( \mathbb{P}(X_1 = \pm 1) = \frac{1}{2} \). Then \( \mathbb{E}(X_1) = 0 \), \( \text{Var}(X_1) = 1 \), and so \( \frac{X_1 + \cdots + X_n}{\sqrt{n}} \Rightarrow X \) by CLT. This implies \( \varphi_{X_1+\cdots+X_n}(t) \rightarrow \varphi_X(t) \) for all \( t \in \mathbb{R} \).

Next, \( \varphi_{X_1+\cdots+X_n} = \varphi_{X_1} \cdots \varphi_{X_n} \left( \frac{t}{\sqrt{n}} \right) = \varphi_{X_1} \left( \frac{t}{\sqrt{n}} \right)^n \). It’s an easy check that \( \varphi_{X_1} = \cos \), so for all \( t \in \mathbb{R} \),

\[
\varphi_X(t) = \lim_{n \to \infty} \varphi_{X_1+\cdots+X_n}(t) = \lim_{n \to \infty} \left( \cos \left( \frac{t}{\sqrt{n}} \right) \right)^n = \lim_{n \to \infty} \left( 1 - \frac{t^2}{2n} + O \left( \frac{t^4}{n^4} \right) \right)^n = e^{-t^2/2}.
\]

Example 10.4. \( X \sim N(\mu, \sigma^2) \). \( \varphi_X(t) = e^{it\mu - \frac{t^2\sigma^2}{2}} \).

Proposition 10.5. If \( (X_i)_{i=1}^\infty \) is a sequence of random variables with \( X_i \sim N(0, \sigma_i^2) \) for some \( (\sigma_i)_{i=1}^\infty \) bounded away from 0 and \( \infty \), and \( X_i \Rightarrow X \) for some \( X \), then \( \sigma_i \to \sigma \) for some \( \sigma \) and \( X \sim N(0, \sigma^2) \).

Proof: Notice that \( e^{-\frac{t^2}{2}} = \varphi_{X_n}(1) \to \varphi_X(1) \). Since \( (\sigma_i)_i \) is real and bounded away from 0 and \( \infty \), \( \varphi_X(1) \) is real and bounded away from 0 and 1. Thus, there exists \( \sigma \in (0, \infty) \) such that \( \phi_X(1) = e^{-\frac{\sigma^2}{2}} \). This implies \( \sigma_i \to \sigma \). Let \( Z \sim N(0, \sigma) \). Then for every \( t \in \mathbb{R} \),

\[
\varphi_X(t) = \lim_{n \to \infty} \varphi_{X_n}(t) = \lim_{n \to \infty} e^{-\frac{t^2\sigma_i^2}{2}} = e^{-\frac{t^2\sigma^2}{2}} = \varphi_Z(t)
\]

implying \( X \sim Z \). ■

10.2 Inversion of the Characteristic Function

For any random variable \( X \) with absolutely continuous CDF, let \( \rho_X \) denote its density.

**Lemma 10.6.** Let \( Z \) be a random variable with \( Z \sim N(0, 1) \), and let \( \sigma > 0 \). Then for any random variable \( X \), \( \rho_{X + \sigma Z}(y) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(z) e^{-izy} e^{-\frac{z^2}{2}} dz \).

**Proof.** By convolution formula, for any random variable \( X \), the density of \( X + \sigma Z \) at \( x \) is

\[
\rho_{X + \sigma Z}(x) = \int_{\mathbb{R}} \sigma^{-1} \phi((x-z)/\sigma) \rho_X(z) dz.
\]

Thus we have,

\[
\rho_{X + \sigma Z}(0) = \frac{1}{\sqrt{2\pi\sigma}} \int_{\mathbb{R}} e^{-\frac{z^2}{2\sigma^2}} \rho_X(z) dz = \frac{1}{\sqrt{2\pi\sigma}} \mathbb{E} \left( \varphi_{\sigma Z}(X) \right) = \frac{1}{\sqrt{2\pi\sigma}} \mathbb{E}(\varphi_X(Z/\sigma)) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(z) e^{-\frac{z^2}{2}} dz.
\]
The desired equality holds for \( y = 0 \). The general case follows by applying this result to the random variable \( X - y \):

\[
\rho_{X+\sigma Z}(y) = \rho_{(X-y)+\sigma Z}(0) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_{X-y}(z)e^{-\frac{y^2z^2}{2}}dz = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(z)e^{-izy}e^{-\frac{z^2y^2}{2}}dz.
\]

\[\Box\]

**Lemma 10.7.** Let \( X \) be a random variable. If \( \int_{\mathbb{R}} |\varphi_X(t)|dt < \infty \), then \( X \) has a density \( \rho_X \) and \( \rho_X(y) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(t)e^{-ity}dt \).

**Proof.** Let \( Z \) be a rv independent from \( X \) with \( Z \sim N(0,1) \). By the preceding lemma, \( \rho_{X+\sigma Z}(y) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(t)e^{-ity}e^{-\frac{z^2y^2}{2}}dt \). Then we note that, for any \( g \in C_b(\mathbb{R}) \),

\[
E(g(X+\sigma Z)) = \int_{\mathbb{R}} g(y)\rho_{X+\sigma Z}(y)dy = \frac{1}{2\pi} \int_{\mathbb{R}} g(y) \int_{\mathbb{R}} \varphi_X(t)e^{-ity}e^{-\frac{z^2y^2}{2}}dydt = \frac{1}{2\pi} \int_{\mathbb{R}} g(y) \int_{\mathbb{R}} \varphi_X(t)e^{-ity}dydt.
\]

We also note the fact that as \( \sigma \to 0 \), \( X + \sigma Z \xrightarrow{(d)} X \). Then

\[
E(g(X)) = \lim_{\sigma \to 0} E(g(X+\sigma Z)) = \lim_{\sigma \to 0} \frac{1}{2\pi} \int_{\mathbb{R}} g(y) \int_{\mathbb{R}} \varphi_X(t)e^{-ity}dydt = \frac{1}{2\pi} \int_{\mathbb{R}} g(y) \int_{\mathbb{R}} \varphi_X(t)e^{-ity}dydt = \frac{1}{2\pi} \int_{\mathbb{R}} g(y) \int_{\mathbb{R}} \varphi_X(t)e^{-ity}dt dy,
\]

which implies \( \rho_X(y) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(t)e^{-ity}dt \).

\[\Box\]

**Theorem 10.8 (Inversion Lemma).** For any random variable \( X \) and \( a < b \),

\[
\lim_{T \to \infty} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t)dt = P(a < X < b) + \frac{1}{2} P(X \in \{a, b\}).
\]

10.2.1 Tightness

**Lemma 10.9.** \( \frac{1}{u} \int_{-u}^{u} (1 - \varphi_X(t))dt \geq P(|X| \geq \frac{2}{u}) \).

**Proof.** We have

\[
\frac{1}{u} \int_{-u}^{u} (1 - \varphi_X(t))dt = \frac{1}{u} \int_{-u}^{u} (1 - E(e^{itX}))dt = E \left( \frac{1}{u} \int_{-u}^{u} (1 - e^{itX})dt \right)
\]

\[
= E \left( \frac{2 - e^{iuX} - e^{-iuX}}{iuX} \right)
\]

\[
= E \left( \frac{2 - 2\sin(uX)}{uX} \right)
\]

\[
\geq 2 \left( 1 - \frac{1}{uX} \right) \geq P(|uX| \geq 2) = P(|X| \geq 2/u).
\]

\[\Box\]
Theorem 10.10 (Lévy’s Continuity Theorem). Let \((X_n)_n\) be a sequence of random variables and \(\varphi : \mathbb{R} \to \mathbb{C}\) a function continuous at 0 such that \(\varphi_{X_n}(t) \to \varphi(t)\) for all \(t \in \mathbb{R}\). Then there exists a random variable \(X\) such that \(\varphi_X = \varphi\) and \(X_n \Rightarrow X\).

Proof: It is necessary and sufficient to show that \(|X_n|\) is tight, or equivalently, that 
\[
\sup_n P(|X_n| \geq \frac{2}{u}) \to 0 \text{ as } u \to 0.
\]
By the lemma above, it suffices to show \(\sup_n \frac{1}{u} \int_{-u}^{u} (1 - \varphi_{X_n}(t))dt \to 0\).
By DCT we have \(\frac{1}{u} \int_{-u}^{u} (1 - \varphi_{X_n}(t))dt \to \frac{1}{u} \int_{-u}^{u} (1 - \varphi(t))dt\). Since \(\varphi\) is continuous at 0, \(\frac{1}{u} \int_{-u}^{u} (1 - \varphi(t))dt \to 0\) as \(u \to 0\). 

\[\blacklozenge\]


10.3 Characteristic function and Cauchy Distribution

We recall that for a r.v. \(X\) the characteristic function is defined as
\[
\varphi_X(t) = E(e^{itX}), \; t \in \mathbb{R}.
\]
We have \(\varphi_X(t) = \varphi_Y(t), \forall t\) implies \(X \overset{d}{=} Y\) and if \(X\) has density \(f\), then its characteristic function \(\varphi_X(t) = \int_{\mathbb{R}} f(x)e^{ixt}dx\) is the Fourier transform of \(f\).

Lemma 10.11. If \(\int_{\mathbb{R}} |\varphi_X(t)|dt < \infty\), then \(X\) has a density given by \(f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(t)e^{-ixt}dt\) for \(x \in \mathbb{R}\).

For example if \(X = X_1 - X_2\) \(\overset{i.i.d.}{\sim} \exp(1),\) then \(f(x) = \frac{1}{2}e^{-|x|}, \; x \in \mathbb{R},\) and \(\varphi_X(t) = \frac{1}{1+it}, \; t \in \mathbb{R}\).
By the lemma above, \(\frac{1}{2}e^{-|x|} = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{1}{1+it}e^{-ixt}dt\) or \(\int_{\mathbb{R}} \frac{1}{\pi(1+t^2)}e^{itx}dt = e^{-|x|}\). Then, for the density \(\rho(x) = \frac{1}{\pi(1+t^2)}, \; x \in \mathbb{R},\) the c.f. is \(e^{-|t|}\), \(t \in \mathbb{R}\).

Definition 10.12 (Cauchy(\(\lambda\)) Distribution). A r.v. \(X\) is Cauchy distribution with parameter \(\lambda\) if it has the following as its density and c.f. ,
\[
\rho(x) = \frac{\lambda}{\pi(\lambda^2 + t^2)}, \; x \in \mathbb{R} \text{ and } \varphi_X(t) = e^{-\lambda|t|}, \; t \in \mathbb{R}.
\]

One can easily check that \(E|X| = \infty\) for Cauchy distribution.

Lemma 10.13. If \(X_i \overset{i.i.d.}{\sim} \text{Cauchy}(\lambda_i), i = 1, 2, \ldots, n,\) then \(\frac{X_1 + X_2 + \ldots + X_n}{\lambda_1 + \lambda_2 + \ldots + \lambda_n} \overset{d}{\sim} \text{Cauchy}(1)\).

Proof. Let \(Y = \frac{X_1 + X_2 + \ldots + X_n}{\lambda_1 + \lambda_2 + \ldots + \lambda_n},\) then \(\varphi_Y(t) = E(\prod_{i=1}^{n} e^{itX_i/\sum \lambda_j}) = \prod_{i=1}^{n} e^{-\lambda_i|t|/\sum \lambda_j} = e^{-|t|}.\) In particular \(X_i \overset{i.i.d.}{\sim} \text{Cauchy}(1),\) then \(\frac{X_1 + \ldots + X_n}{n} \overset{d}{\sim} \text{Cauchy}(1).\) 

\[\blacklozenge\]
10.4 Conditional Expectation

10.4.1 Definition, existence and uniqueness

**Definition 10.14.** Given σ-fields $G \subseteq F$ and a r.v $X \in L^1(\Omega, F, \mathbb{P})$, we define $E(X \mid G)$ as a r.v $Y$ s.t.

1. $Y$ is in $L^1(\Omega, G, \mathbb{P})$.
2. $E(X 1_A) = E(Y 1_A) \forall A \in G \iff E((X - Y)Z) = 0$ for all bounded $G$-measurable function $Z$.

**Lemma 10.15.** Conditional expectation, if exists, is unique a.s.

**Proof.** Suppose $Y_1, Y_2$ are conditional expectations of $X$ given $G$. Take $W := Y_1 - Y_2$ which is $G$ measurable. Then $E(W 1_A) = 0$ for all $A \in G$. Take $A = \{W > \varepsilon\} \in G$. Then,

$$0 = E(W 1_{W > \varepsilon}) > \varepsilon P(W > \varepsilon) \implies P(W > \varepsilon) = 0$$

Similarly, taking $A = \{W < -\varepsilon\} \in G$ we get

$$0 = E(W 1_{W < -\varepsilon}) < -\varepsilon P(W < -\varepsilon) \implies P(W < -\varepsilon) = 0$$

Therefore $P(W \in [-\varepsilon, \varepsilon]) = 1$ for all $\varepsilon > 0$. Hence $P(W = 0) = 1$ and $Y_1 = Y_2$ a.s.

We will prove that

**Theorem 10.16.** Conditional expectation exists.

We will prove this later.

10.4.2 Properties of Conditional Expectation

Let $G \subseteq F$ be σ-fields, then $E(\cdot \mid G) : L^1(F) \to L^1(G)$ where $L^p(H) := L^p(\Omega, H, \mathbb{P}), p \geq 1$ for a sub σ-field $H$ of $F$.

(i) **Positive:** $X \geq 0$ a.s. $\implies \hat{X} := E(X \mid G) \geq 0$ a.s.

Proof: Using $E(X 1_A) = E(\hat{X} 1_A) \forall A \in G$; for $A = \{\hat{X} < 0\}$, we get $E(\hat{X} 1_{\hat{X} < 0}) = 0$ which implies $\hat{X} \geq 0$ a.s.

(ii) **Linear:** $E(X + Y \mid G) = E(X \mid G) + E(Y \mid G)$ a.s. and $E(cX \mid G) = c E(X \mid G)$ a.s. for $c \in \mathbb{R}$.

(iii) **Contractive on $L^p(F)$ → $L^p(G)$:** For $X \in L^p(F), p \geq 1$ we have $E(X \mid G) \in L^p(G)$ and $\|E(X \mid G)\|_p \leq \|X\|_p$.

Proof for $p = 1$: Let $\hat{X} = E(X \mid G)$, we have $E(X 1_A) = E(\hat{X} 1_A)$ for all $A \in G$. Taking $A_1 = \{\hat{X} \geq 0\}, A_2 = \{\hat{X} < 0\}$, we have,

$$E|\hat{X}| = E(\hat{X} 1_{\hat{X} \geq 0}) - E(\hat{X} 1_{\hat{X} < 0}) = E(X 1_{\hat{X} \geq 0}) - E(X 1_{\hat{X} < 0}) \leq E|X|.$$
For general $p > 1$, use $E(XZ) = E(\hat{X}Z)$ for any bounded $\mathcal{G}$-measurable $Z$. The r.v. $Z_n = |\hat{X}|^{p-1} \cdot (1_{0 < \hat{X} \leq n} - 1_{0 < -\hat{X} \leq n})$ is bounded and $\mathcal{G}$-measurable. Thus $E(\hat{X}Z_n) = E(XZ_n)$ implies that
\[
E|\hat{X}1_{|\hat{X}|\leq n}|^p = E(XZ_n) \leq E(|X| \cdot |\hat{X}1_{|\hat{X}|\leq n}|^{p-1}) \leq (E|X|^p)^{1/p} (E|\hat{X}1_{|\hat{X}|\leq n}|^p)^{1-1/p}
\]
where we used Hölder inequality in the last line. Thus we have
\[
E|\hat{X}1_{|\hat{X}|\leq n}|^p \leq E|X|^p \text{ for all } n \geq 1.
\]

Taking limit $n \to \infty$ we get the results.

(iv) **William’s Tower Property**: Let $\mathcal{G} \subseteq \mathcal{H}$. Suppose $E(\cdot \mid \mathcal{G})$ and $E(\cdot \mid \mathcal{H})$ are well defined, then $E(E(X \mid \mathcal{H}) \mid \mathcal{G}) = E(X \mid \mathcal{G})$.

(v) If $X$ is $\mathcal{G}$-measurable then $E(X \mid \mathcal{G}) = X$ a.s.

(vi) **Projection**: $E(E(X \mid \mathcal{G}) \mid \mathcal{G}) = E(X \mid \mathcal{G})$, which follows from Tower property.

(vii) **Monotone**: $X \geq Y$ a.s. $\implies E(X \mid \mathcal{G}) \geq E(Y \mid \mathcal{G})$ a.s.

(viii) **Conditional MCT**: $X_n \geq 0$, $X_n \uparrow X$ $\implies E(X_n \mid \mathcal{G}) \uparrow E(X \mid \mathcal{G})$ a.s.

(ix) $E(X \mid \{\emptyset, \Omega\}) = E X$.

(x) **Jensen’s Inequality**: If $\phi$ is convex and $E|\phi(X)| < \infty$, $\phi(E(X \mid \mathcal{G})) \leq E(\phi(X) \mid \mathcal{G})$ a.s.

(xi) **Cauchy-Schwartz and Hölder Inequality**: $|E(XY \mid \mathcal{G})| \leq (E(|X|^p \mid \mathcal{G}))^{1/p} (E(|Y|^q \mid \mathcal{G}))^{1/q}$ a.s. for $1/p + 1/q = 1$, $p, q \geq 1$ and $X \in L^p(\mathcal{F}), Y \in L^q(\mathcal{F})$.

(xii) Suppose $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_\infty = \sigma(\cup_{n \geq 0} \mathcal{F}_n) \subseteq \mathcal{F}$, and $X \in L^p(\mathcal{F})$, $p \geq 1$. Then $E(X|\mathcal{F}_n) \xrightarrow{a.s./L^p} E(X|\mathcal{F}_\infty)$. In particular, if $\mathcal{F}_\infty = \mathcal{F}$ then $E(X \mid \mathcal{F}_n) \xrightarrow{a.s./L^p} X$.

(xiii) $(\hat{X}_n = E(X|\mathcal{F}_n))_{n \geq 0}$ is adapted to $(\mathcal{F}_n)_{n \geq 0}$ and $E(\hat{X}_n \mid \mathcal{F}_{n-1}) = \hat{X}_{n-1}$.

### 10.4.3 Proof of Theorem 10.16: Construction of $E(\cdot \mid \mathcal{G})$

Here are a number of proofs for Theorem 10.16.

*(Measure theoretic proof)*. We will use the following theorem from measure theory.

**Theorem 10.17 (Lebesgue-Radon-Nikodym derivative)**. Let $\mu$ and $\lambda$ be two finite positive measures on $(\Omega, \mathcal{F})$ such that $\mu \ll \lambda$ ($\mu$ is absolutely continuous w.r.t. $\lambda$, i.e., $\lambda(A) = 0 \implies \mu(A) = 0$). Then there exists a measurable function $f \in L^1(\Omega, \mathcal{F}, \lambda)$ s.t.,
\[
f = \frac{d\mu}{d\lambda} \iff \mu(A) = \int_A f d\lambda \\forall \ A \in \mathcal{F}.
\]

Assume that $X \geq 0$. Consider Theorem 10.17 with $\lambda := P$ and $\mu(A) := \int_A X dP$ for all $A \in \mathcal{G}$. $\lambda, \mu$ are positive finite measures on $(\Omega, \mathcal{G})$ and $\mu \ll \lambda$. Then by Theorem 10.17 there exists $f \in L^1(\Omega, \mathcal{G}, P)$ s.t.
\[
\mu(A) = \int_A f dP \implies E(X 1_A) = E(f 1_A) \\forall A \in \mathcal{G}.
\]
So $f$ is the conditional expectation. In general $X = X^+ - X^-$ and we define

$$E(X \mid \mathcal{G}) = E(X^+ \mid \mathcal{G}) - E(X^- \mid \mathcal{G}).$$

* (**Functional analysis proof**). We will use the following lemma on orthogonal projection in Hilbert spaces.

**Lemma 10.18.** Let $K \subseteq H$ be a close subspace of Hilbert space $H$. Then for all $X \in H$, there exists a unique decomposition $X = Y + Z$ s.t $Y \in K$ and $Z \in K^\perp$.

Assume $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ which is a Hilbert space. We will use shorthand notation $L^2(\mathcal{F})$ for $L^2(\Omega, \mathcal{F}, \mathbb{P})$. By Lemma 10.18 there exists a unique decomposition $X = Y + E$ such that $Y \in L^2(\mathcal{G})$ and $E \in L^2(\mathcal{G})^\perp$. So,

$$\forall \ Z \in L^2(\mathcal{G}), \ E((X - Y)Z) = 0 \implies Y = E(X \mid \mathcal{G})$$

This proof can be generalized to $X \in L^1$.

* (**Hands on proof**).

(i) Assume that $|\mathcal{G}| < \infty$. Then $\mathcal{G} = \sigma(C_1, \ldots, C_k)$ such that $\{C_1, C_2, \ldots, C_k\}$ is a disjoint partition of $\Omega$. Now any $\mathcal{G}$-measurable r.v. $Z$ is of the form, $Z = \sum_{i=1}^{k} a_i 1_{C_i}$ for some $a_1, a_2, \ldots, a_k \in \mathbb{R}$. Then the conditional expectation is,

$$E(X \mid \mathcal{G}) = Y := \sum_{i=1}^{k} \frac{E(X 1_{C_i})}{E(1_{C_i})} 1_{C_i},$$

since for all $\mathcal{G}$-measurable r.v. $Z$,

$$E(Y Z) = \sum_{i=1}^{k} a_i \sum_{j=1}^{k} \frac{E(X 1_{C_i})}{E(1_{C_j})} E(1_{C_j} 1_{C_i}) = \sum_{i=1}^{k} a_i E(X 1_{C_i}) = E(XZ).$$

(ii) Suppose $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \cdots \subseteq \mathcal{G} = \sigma(\cup_{i \geq 1} \mathcal{G}_i)$ and $X \in L^2(\mathcal{F})$. Then, let

$$\hat{X}_n = E(X \mid \mathcal{G}_n).$$

We have $\|\hat{X}_n\|_2 \leq \|X\|_2$. Let $\Delta_n = \hat{X}_n - \hat{X}_{n-1}, n \geq 1$. By William’s Tower property we have, $E(\hat{X}_n \mid \mathcal{G}_{n-1}) = \hat{X}_{n-1}$. Thus for any $L^2(\mathcal{G}_{n-1})$ r.v. $Y$, $E((\hat{X}_n - \hat{X}_{n-1})Y) = 0$. Thus $\{\hat{X}_1, \Delta_2, \Delta_3, \ldots\}$ are uncorrelated. By definition of $\Delta_n$, we have $\hat{X}_n = \hat{X}_1 + \Delta_2 + \Delta_3 + \cdots + \Delta_n$. This implies $S_n^2 := \|\hat{X}_n\|_2^2 = \|\hat{X}_1\|_2^2 + \|\Delta_2\|_2^2 + \cdots + \|\Delta_n\|_2^2 \leq \|X\|_2^2$. Thus $\hat{X}_n$ is $L^2$-Cauchy, since $S_n^2 \uparrow S_\infty^2 \leq E(X^2)$. Therefore $\hat{X}_n \xrightarrow{L^2} \hat{X}_\infty$.

**Claim:** $\hat{X}_\infty$ is $\mathcal{G}$-measurable.

**Claim:** $E(X 1_A) = E(\hat{X}_\infty 1_A)$ for all $A \in \cup_{n \geq 1} \mathcal{G}_n$. To verify the claim, note that: $\exists n$ s.t. $A \in \mathcal{G}_m$ for $m \geq n$. Also $E(X 1_A) = E(\hat{X}_m 1_A), m \geq n$, and as $m \to \infty$, it converges to $E(\hat{X}_\infty 1_A)$.

Now, $\{A : E(X 1_A) = E(\hat{X}_\infty 1_A)\}$ is a $\lambda$ system and $\cup_{n \geq 1} \mathcal{G}_n$ is a $\pi$-system. By $\pi - \lambda$ theorem, $E(X 1_A) = E(\hat{X}_\infty 1_A)$ for all $A \in \sigma(\cup_{n \geq 1} \mathcal{G}_n) = \mathcal{G}$. Therefore, we have, $\hat{X}_\infty = E(X \mid \mathcal{G})$ a.s.

(iii) If $\mathcal{G} = \sigma(C_1, C_2, \ldots)$ then take $\mathcal{G}_n = \sigma(C_1, C_2, \ldots, C_n), n \geq 1$. In general, given $X \in L^2(\mathcal{F})$, $\mathcal{G} \subseteq \mathcal{F}$, $S_n := \sup_{|\mathcal{H}| < \infty, \mathcal{H} \subseteq \mathcal{G}} \|E(X \mid \mathcal{H})\|_2$ exists. Take $\mathcal{H}_i$ s.t. $\|E(X \mid \mathcal{H}_i)\|_2 \uparrow S_i$ and work with $\mathcal{G}_i = \sigma(\mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots \cup \mathcal{H}_i), i \geq 1$. 

Week 10: Characteristic function and Conditional Expectation 10-7
Finally, from $X \in L^2(\mathcal{F})$, one can generalize to $X \in L^1(\mathcal{F})$ by approximating a non-negative r.v. by a bounded r.v. monotonically.

10.4.4 Lévy’s 0-1 Law

We have $P(A|\mathcal{F}_n) \overset{n \to \infty}{\to} 1_A$ whenever $\mathcal{F}_n \uparrow \mathcal{F}$.

**Exercise 10.19.** Let $X_1, X_2, \ldots$ be independent r.v.s and $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, $A \in \cap_{n \geq 1} \sigma(X_n, X_{n+1}, \ldots)$ then $P(A) \in \{0, 1\}$. 
11.1 Regular Conditional Probability (RCP)

Given two measure spaces \((\Omega, \mathcal{F})\) and \((S, \mathcal{G})\), a Markov kernel \(Q(\omega, A) : \Omega \times \mathcal{G} \to [0, 1]\) is a function such that

1. \(Q(\cdot, A)\) is \(\mathcal{F}\) measurable for fixed \(A \in \mathcal{G}\),
2. \(Q(\omega, \cdot)\) is a probability measure on \((S, \mathcal{G})\).

For a measurable function \(X : (\Omega, \mathcal{F}) \to (S, \mathcal{G})\) and a sub σ-field \(\mathcal{G} \subseteq \mathcal{F}\), Regular Conditional Probability (RCP) is a Markov kernel \(Q(\omega, A)\) on \(\Omega \times \mathcal{G}\) which is a version of \(P(X \in A \mid \mathcal{G})\), \(A \in \mathcal{G}\), i.e.,

\[
P(Q(\cdot, A) = P(X \in A \mid \mathcal{G})) = 1 \text{ for all } A \in \mathcal{G}.
\]

**Theorem 11.1.** RCP exists for real-valued r.v. \(X\).

**Proof.** For \(r \in \mathbb{Q}\), \(P(X \leq r \mid \mathcal{G})\) is defined a.s. on \(\omega \in A_r, P(A_r) = 1\). Therefore, \(A = \cap_{r \in \mathbb{Q}} A_r\) has \(P(A) = 1\) and \(F(\omega, r) := P(X \leq r \mid \mathcal{G})\) exists and is monotone in \(r\) for all \(r \in \mathbb{Q}, \omega \in A\). For \(\omega \in A\), define \(F(\omega, x) = \inf_{r > x, r \in \mathbb{Q}} F(\omega, r), x \in \mathbb{R}\). Define \(Q(\omega, C)\) as follows:

\[
Q(\omega, C) = \begin{cases} 
\int_C dF(\omega, x) & \text{if } \omega \in A \\
1_{\{0 \in C\}} & \text{if } \omega \not\in A.
\end{cases}
\]

Then \(Q(\cdot, A)\) is a RCP for \(P(X \in A \mid \mathcal{G})\), \(A \in \mathcal{B}\).

**Lemma 11.2.** \(X : (\Omega, \mathcal{F}) \to (S, \mathcal{G}), \mathcal{G} \subseteq \mathcal{F}\). RCP exists if \((S, \mathcal{G})\) is Borel isomorphic to \((\mathbb{R}, \mathcal{B})\). \(\exists\) bimeasurable bijection \((S, \mathcal{G}) \to (\mathbb{R}, \mathcal{B}))\).

**Example 11.3.** Let \(X \perp \mathcal{G}, X, Y : (\sigma, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{R})\) and \(Y\) is \(\mathcal{G}\)-measurable, \(\mathcal{G} \subseteq \mathcal{F}\). Find RCP for \(X + Y\mid \mathcal{G}\). We want to define \(P(X + Y \in A \mid \mathcal{G})\) for \(\omega \in \Omega\) and \(A \in \mathcal{R}\). Since \(Y\) is \(\mathcal{G}\)-measurable, given \(\mathcal{G}\), \(Y\) can be think of as a constant. Let \(\mu_X(A) = P(X \in A)\), then RCP for \(X + Y\mid \mathcal{G}\) is \(\mu_X(A - Y(\omega))\).

**Proof.**

1. \(\forall A \in \mathcal{R}, \mu_X(A - Y(\omega))\) is a function of \(Y\), so it is \(\mathcal{G}\)-measurable.
2. \(\forall \omega \in \Omega, \mu_X(A - Y(\omega))\) is a probability measure.
3. \(\forall Z\) bounded and \(\mathcal{G}\)-measurable, want to show that \(E(Z \mu_X(A - Y)) = E(Z 1_{X+Y \in A})\). We have

\[
E(Z \mu_X(A - Y(\omega))) = E(Z \int 1_{x+Y(\omega) \in A} P_X(dx)) = \int E(Z 1_{x+Y(\omega)}) P(dx) = E(Z 1_{X+Y \in A})
\]

where the last equality follows by Fubini.

---

11-1
Example 11.4. Suppose \((X,Y)\) has the joint density \(f(x,y)\,dx\,dy\) and \(\mathcal{G} = \sigma(Y)\), then \(X|Y = X|\sigma(Y)\) has RCP defined as the following. Marginal density of \(Y\) is \(f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx\). Conditional density \(g_Y(x) = \frac{f(x,y)}{f_Y(y)}\). We claim that RCP of \(X \mid Y\) is

\[
P(X \in A \mid Y) = \left. \int_A \frac{f(x,y)}{f_Y(y)} \, dx \right|_{y=Y(\omega)} \quad \forall \omega \in \Omega \text{ and } A \in \mathcal{B}.
\]

**Proof.** We have \(\int_A \frac{f(x,y)}{f_Y(y)} \, dx \big|_{y=Y(\omega)}\) satisfies

1. \(Y\)-measurable (for fixed \(A\)).
2. a probability measure (for fixed \(\omega\)).
3. for all \(Z, \sigma(Y)\)-measurable and bounded we have, \(E(Z \mathbb{1}_{X \in A}) = E(Z \int_A \frac{f(x,y)}{f_Y(y)} \, dx \big|_{y=Y(\omega)})\). We use the fact that, any \(Z\) which is \(\sigma(Y)\) measurable must be \(\eta(Y)\) for some measurable \(\eta\). Thus we have

\[
E\left( \int_A \frac{f(x,y)}{f_Y(y)} \, dx \big|_{y=Y}\right) = \int \left[ \int_A \frac{f(x,y)}{f_Y(y)} \eta(y) \, dx \right] f_Y(y) \, dy = E\left( \eta(Y) \mathbb{1}_{X \in A}\right).
\]

**Note:** Suppose \(\mathcal{G} = \sigma(Y)\) and the RCP of \(X \mid Y\) is \(Q(\omega, A) = P(X \in A \mid \sigma(Y))\), then this RCP is a function of \(Y\), it must be of the form \(\hat{Q}(Y(\omega), A)\). Since RCP is a measurable function of \(\mathcal{G}\), and thus it is a function of \(Y\). We will use the notation, \(P(X \in A \mid Y = y)\) for \(Q(y, A)\).

![Diagram](image)

Example 11.5. Suppose that a point with Cartesian coordinate \((X,Y)\) is uniformly distributed on the half disk \(\{ (x,y) : y \geq 0, x^2 + y^2 \leq 1 \}\), then \(P(Y \leq 1/2 \mid X = 0) = 1/2\). In fact, the RCP of \(Y \mid X\) is given by

\[
P(Y \in A \mid X) = \left| A \cap [0, \sqrt{1-x^2}] \right|_{x=X} / \sqrt{1-x^2}
\]

and then evaluate this at \(x = 0, A = [0, 1/2]\).

If \(P(Y \in A \mid X) = \hat{Q}(X(\omega), A)\), then

\[
P(Y \in A \mid X \in (x \pm \varepsilon)) = \frac{P(Y \in A, X \in (x \pm \varepsilon))}{P(X \in (x \pm \varepsilon))} = \frac{E(\hat{Q}(X, A) \mathbb{1}_{X \in (x\pm\varepsilon)})}{P(X \in (x \pm \varepsilon))}.
\]

Thus if RCP exists and \(x\) is in the support of \(X\), \(\lim_{\varepsilon \to 0} P(Y \in A \mid X \in (x \pm \varepsilon)) = \hat{Q}(x, A)\).

**Note:** Rigorously speaking, for set \(A\) and set \(B\), \(P(A \mid B)\) is not defined. We only define a conditional probability given a \(\sigma\)-field. Intuitively \(P(A \mid B) \equiv P(A \mid \sigma(B) = \{\emptyset, B, B^c, \Omega\})\)

Example 11.6. Think of the previous example in a polar coordinate. Let a point with polar coordinate \((R, \Theta)\) be uniformly distributed on a half disk, then \(P(R \leq 1/2 \mid \Theta = \pi/2) = (1/2)^2 = 1/4\). The existence of RCP of \(R \mid \Theta\) implies that \(\lim_{\varepsilon \to 0} P(R \leq 1/2 \mid \Theta \in (\pi/2 \pm \varepsilon))\) exists and equals \(P(R \leq 1/2 \mid \Theta = \pi/2)\).
11.2 Stopping time

Definition 11.7. A (discrete) filtration is a sequence of increasing σ-fields,
\[ \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots. \]
e.g., Let \( X_1, X_2, \ldots \) be i.i.d r.v.s. Then \( \mathcal{F}_n = \sigma(X_1, \ldots , X_n), n \geq 0 \) is a filtration.

Definition 11.8. A sequence of events \( (A_n)_{n \geq 0} \) is adapted to the filtration \( (\mathcal{F}_n)_{n \geq 0} \) if,
\[ A_n \in \mathcal{F}_n, \quad \forall \ n \geq 0. \]
A sequence of r.v.s \( (X_n)_{n \geq 0} \) is adapted to the filtration \( (\mathcal{F}_n)_{n \geq 0} \) if,
\[ \sigma(X_n) \subseteq \mathcal{F}_n, \quad \forall \ n \geq 0. \]

Definition 11.9. A sequence of events \( (A_n)_{n \geq 0} \) is predictable w.r.t. the filtration \( (\mathcal{F}_n)_{n \geq 0} \) if,
\[ A_{n+1} \in \mathcal{F}_n, \quad \forall \ n \geq 0. \]
A sequence of r.v.s \( (X_n)_{n \geq 1} \) is predictable w.r.t. to the filtration \( (\mathcal{F}_n)_{n \geq 1} \) if,
\[ \sigma(X_{n+1}) \subseteq \mathcal{F}_n, \quad \forall \ n \geq 0. \]

Definition 11.10. Stopping time w.r.t. a filtration \( (\mathcal{F}_n)_{n \geq 0} \) is a r.v. \( T : \Omega \to \{0,1,2,\ldots\} \) s.t.,
\[ \{T = n\} \in \mathcal{F}_n \quad \forall \ n \geq 0. \]

Example 11.11. Let \( X_1, X_2, \ldots \) be a sequence of i.i.d r.v.s and \( S_n = X_1 + X_2 + \cdots + X_n, n \geq 0. \)
Then
\[ T_A = \inf\{n | S_n \in A\} \]
is a stopping time (hitting time) with respect to the standard filtration \( \mathcal{F}_n := \sigma(X_1, \ldots , X_n). \)

Proof. We have \( \{T = n\} = \{S_n \in A; S_i \notin A, i = 1, \ldots , n - 1\} \in \sigma(X_1, \ldots , X_n) = \mathcal{F}_n. \)

Example 11.12. Let \( X_1, \ldots , X_N \) be i.i.d. r.v.s. Then
\[ T = \min\{i | X_i = \max(X_1, \ldots , X_N)\} \]
is not a stopping time w.r.t. \( \mathcal{F}_n := \sigma(X_1, \ldots , X_n). \)

Proof. We have \( \{T = n\} = \{X_1, X_2, \ldots , X_{n-1} < X_n, X_{n+1}, \ldots , X_N \leq X_n\} \notin \sigma(X_1, \ldots , X_n). \)

Example 11.13. Consider Example 11.11 with \( \mathcal{F}_n := \sigma(X_1, \ldots , X_{n+1}). \) Then \( \{T = n\} \) is predictable.

Lemma 11.14. \( T \) is stopping time w.r.t filtration \( \mathcal{F}_n \) iff
\[ \{T \leq n\} \in \mathcal{F}_n, \quad \text{or} \quad \{T > n\} \in \mathcal{F}_n, \quad \forall \ n. \]

Lemma 11.15. If \( T \) and \( S \) are two stopping time w.r.t the same filtration \( (\mathcal{F}_n) \), then \( S + T, S \lor T \) and \( S \land T \) are stopping times.

Proof. For \( S + T \), we have \( \{S + T = n\} = \bigcup_{k=0}^n \{S = k\} \cap \{T = n - k\} \in \mathcal{F}_n. \)

Example 11.16. Consider the setup in Example 11.11. Then,
\[ T_A \land T_B = T_{A \cup B} \quad \text{and} \quad T_A \lor T_B = T_{A \cap B}. \]
11.3 Martingale

After studying random variables and random vectors, a natural next step is a countably infinite vector of random variables \((X_n)_{n \geq 0}\). We can think of the sequence as a time indexed sequence where the time is in \( \mathbb{Z}_+ = \{0, 1, 2, \ldots\}\). As we measure information by \( \sigma \)-fields, we can define \( \mathcal{F}_n \) as the amount of information available at time \( n \geq 0 \) and naturally we have \( \mathcal{F}_n \subseteq \mathcal{F}_{n+1} \) for all \( n \), which is the definition of a filtration. Also, \( X_n \) should be \( \mathcal{F}_n \) measurable for all \( n \). By abstract result (Ionescu-Tulcea Extension theorem) it follows that, if we know the conditional distribution of \( X_{n+1} \) given \( \sigma(X_i, 0 \leq i \leq n) \) for all \( n \), we can reconstruct the whole infinite process. There are mainly two direction one can take: a) assume that the dependence on the past is weak, say the conditional distribution of \( X_{n+1} \) given \( \sigma(X_i, 0 \leq i \leq n) \) depends only on \( X_n \) (or \( n, X_n \)) for all \( n \), which gives rise to Markov Chains (or time-inhomogeneous Markov chain), b) instead of compromising on the dependence on the past, we assume the conditional expectation (or some other statistic) of \( X_{n+1} \) given \( \sigma(X_i, 0 \leq i \leq n) \) has a fixed sign (simplest assumption). The second direction gives rise to Martingales and sub/super-martingales.

The word “Martingale” has the dictionary meaning “a horse strap to keep the horse head steady”. It also means a system of gambling in which the stakes are doubled or otherwise raised after each loss. In general, for us it will mean conditional zero mean process at every time point. Formally, we define the following.

**Definition 11.17.** Given a filtration \( \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \), a sequence of r.v.s \((M_n)_{n \geq 0}\) is a martingale w.r.t. \((\mathcal{F}_n)_{n \geq 0}\) if 1) \( M_n \in L^1(\mathcal{F}_n) \), 2) \( \mathbb{E}(M_{n+1}|\mathcal{F}_n) = M_n \ \forall \ n \geq 0 \).

**Definition 11.18.** Given a filtration \( \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \), a sequence of r.v.s \((M_n)_{n \geq 0}\) is a sub-martingale w.r.t. \((\mathcal{F}_n)_{n \geq 0}\) if 1) \( M_n \in L^1(\mathcal{F}_n) \), 2) \( \mathbb{E}(M_{n+1}|\mathcal{F}_n) \geq M_n \ \forall \ n \geq 0 \).

**Definition 11.19.** Given a filtration \( \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \), a sequence of r.v.s \((M_n)_{n \geq 0}\) is a super-martingale w.r.t. \((\mathcal{F}_n)_{n \geq 0}\) if 1) \( M_n \in L^1(\mathcal{F}_n) \), 2) \( \mathbb{E}(M_{n+1}|\mathcal{F}_n) \leq M_n \ \forall \ n \geq 0 \).

If a sequence of r.v.s is a martingale, it is both sub-martingale and super-martingale. Moreover, for a martingale \( \mathbb{E}(M_n) \) is constant (increasing or decreasing for sub and super-martingales, respectively).

**Example 11.20.** Given a filtration \( \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \) and \( X \in L^1(\mathcal{F}) \), define \( M_n = \mathbb{E}(X | \mathcal{F}_n) \) for \( n \geq 0 \). Then we have:

1. \( M_n \in L^1(\mathcal{F}_n) \ \forall \ n \geq 0 \).
2. \( \mathbb{E}(M_{n+1} | \mathcal{F}_n) = \mathbb{E}(\mathbb{E}(X | \mathcal{F}_{n+1}) | \mathcal{F}_n) = \mathbb{E}(X | \mathcal{F}_n) = M_n \), where the middle equality follows by Tower property.

**Example 11.21.** Suppose \( X_1, X_2, \ldots \) are independent r.v.s with mean 0; and \( \mathcal{F}_n = \sigma(X_1, \ldots, X_n) \), \( M_n = X_1 + \cdots + X_n \).

1. \( M_n \in L^1(\mathcal{F}_n) \), \( \forall n \geq 0 \).
2. \( \mathbb{E}(M_{n+1} | \mathcal{F}_n) = \mathbb{E}(M_n + X_{n+1} | \mathcal{F}_n) = M_n + \mathbb{E}(X_{n+1}) = M_n \). Here we used that, 1) \( X \in \mathcal{G} \implies \mathbb{E}(X|\mathcal{G}) = X \); 2) \( X \perp \mathcal{G} \implies \mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X) \).

**Example 11.22.** Given \((M_n, \mathcal{F}_n)_{n \geq 0}\) is a martingale, and \( \phi \) is a convex function with \( \mathbb{E}|\phi(M_n)| < \infty, \forall n \). Then \( (\phi(M_n), \mathcal{F}_n) \) is a sub-martingale.
1. \( \phi(M_n) \in L^1(F_n), \forall n \geq 0. \)

2. \( \mathbb{E}(\phi(M_{n+1}) | F_n) \geq \phi(\mathbb{E}(M_{n+1} | F_n)) = \phi(M_n). \) (Conditional Jensen's inequality)

**Example 11.23.** Given a random vector \( X \), uniformly distributed on a \( d \)-dimensional unit ball \( B(0,1) \), and a sequence of random vectors \( X_1, X_2, \ldots, X_n, \ldots \overset{i.i.d.}{\sim} X \), we consider the random walk

\[
S_n = \begin{cases} 
 x_0 \in \mathbb{R}^d & \text{if } n = 0, \\
 x_0 + X_1 + \cdots + X_n & \text{if } n \geq 1.
\end{cases}
\]

Let \( f \) be a super-harmonic function on \( \mathbb{R}^d \), i.e., \( f(x) \geq |B(0,1)|^{-1} \int_{y \in B(x,1)} f(y) dy \) for all \( x \).

Then \( (M_n = f(S_n))_{n \geq 0} \) is a super-martingale w.r.t. the filtration \( F_n = \sigma(X_1, \ldots, X_n), n \geq 0 \), if \( \mathbb{E}|M_n| < \infty \) for all \( n \geq 0 \).

1. \( \phi(M_n) \in L^1(F_n), \forall n \geq 0. \)

2. \( \mathbb{E}(f(S_{n+1}) | F_n) = \mathbb{E}(f(S_n + X_{n+1}) | F_n) = \mathbb{E}(f(x + X_{n+1}) | x = S_n) \leq f(S_n) = M_n. \)

**Example 11.24** (Galton-Watson martingale). Let \( X \) be a non-negative integer-valued random variable with \( \mathbb{E}X = \mu. \) Let \( (X_i^{(j)}) \) be a sequence of iid sequence with \( X_i^{(j)} \sim X. \) Then set

\[
Z_0 = 1, \quad Z_1 = X_1^{(1)}, \quad Z_n = \sum_{i=1}^{Z_{n-1}} X_i^{(n)} \text{ for } n \geq 1.
\]

Define

\[
W_n = \frac{Z_n}{\mu^n}, \quad F_n = \sigma(X_i^{(j)}, j \leq n) \text{ for } n \geq 0.
\]

Notice that \( W_n \in L^1(F_n) \) and

\[
\mathbb{E}(W_{n+1} | F_n) = \frac{1}{\mu^{n+1}} \mathbb{E} \left( \sum_{i=1}^{Z_n} X_i^{(n+1)} | F_n \right) = \frac{1}{\mu^{n+1}} \sum_{i=1}^{Z_n} \mathbb{E}(X_i^{(n+1)} | F_n) = \frac{1}{\mu^{n+1}} \sum_{i=1}^{Z_n} \mathbb{E}(X_i^{(n+1)}) = \frac{Z_n}{\mu^n} = W_n.
\]

So \( (W_n) \) is a \((F_n)-martingale.\)

**Exercise 11.25.** Suppose there exists \( \theta > 0 \), such that \( \mathbb{E}(\theta^X) = \theta \). Let \( M_n = \theta^{Z_n}, n \geq 0. \) Then \((M_n)\) is a \((F_n)-martingale.\)

**Example 11.26** (Second moment martingale). Let \((X_n)\) be an iid sequence of random variables with \( \mathbb{E}(X_1) = 0, \mathbb{E}(X_1^2) < \infty. \) Define

\[
M_n = (X_1 + \cdots + X_n)^2 - \sum_{i=1}^{n} \text{Var}(X_i), \quad F_n = \sigma(X_1, X_2, ..., X_n)
\]
for \( n \geq 1 \). Then \((M_n)\) is a \((\mathcal{F}_n)\)-martingale. It is clear that \(M_n \in L^1(\mathcal{F}_n)\) for any \(n\). Notice that

\[
E(M_{n+1}\mid \mathcal{F}_n) = E((s_n + X_{n+1})^2 - \sum_{i=1}^{n+1} \text{Var}(X_i) \mid \mathcal{F}_n)
\]

\[
= s_n^2 + 2s_n E(X_{n+1} \mid \mathcal{F}_n) + E(X_{n+1}^2 \mid \mathcal{F}_n) - \sum_{i=1}^{n+1} \text{Var}(X_i) \quad (\ast)
\]

\[
= s_n^2 - \sum_{i=1}^{n} \text{Var}(X_i) = M_n
\]

where \(s_n := \sum_{i=1}^{n} X_i\), and notice that at \((\ast)\) the following facts are applied: 1) if \(X\) is \(\mathcal{G}\) measurable, then \(E(XY \mid \mathcal{G}) = X E(Y \mid \mathcal{G})\), 2) if \(X\) is independent of \(\mathcal{G}\), then \(E(X \mid \mathcal{G}) = E(X)\).

**Example 11.27** (Likelihood Ratio martingale). Let \(X_n, n \geq 1\) be an i.i.d. sequence of random variables with \(\phi(\theta) := E e^{i\theta X_1} < \infty\). Define

\[
M_n = \phi(\theta)^{-n} \cdot e^{\theta S_n}, \quad S_n := X_1 + X_2 + \cdots + X_n, \quad \mathcal{F}_n = \sigma(X_1, X_2, \ldots, X_n)
\]

for \(n \geq 1\). It is clear that \(M_n \in L^1(\mathcal{F}_n)\). Since \(M_n \geq 0\) and

\[
E(M_{n+1} \mid \mathcal{F}_n) = \phi(\theta)^{-n-1} \cdot e^{\theta S_n} E(e^{\theta X_{n+1}} \mid \mathcal{F}_n) = \phi(\theta)^{-n} \cdot e^{\theta S_n} = M_n
\]

Thus \((M_n)\) is a \((\mathcal{F}_n)\)-martingale.
Wald’s Identities

12.1 Doob’s martingale transform

Definition 12.1 (Martingale difference sequence). Let \((\mathcal{F}_n)\) be a filtration and \((\Delta_n)\) be an adapted sequence of random variables to \((\mathcal{F}_n)\). Then \((\Delta_n)\) is a Martingale difference sequence if it satisfies 1) \(\Delta_n \in L^1(\mathcal{F}_n)\), 2) \(\mathbb{E}(\Delta_{n+1} | \mathcal{F}_n) = 0\) for all \(n \geq 0\).

Observe that it is possible to give an equivalent definition of Martingale in terms of Martingale difference sequence as follows: Let \((\mathcal{F}_n)\) be a filtration and \((\Delta_n)\) be a Martingale difference sequence. Then the sequence \((M_n)\) defined by

\[
M_n = \sum_{i=0}^{n} \Delta_i, n \geq 1
\]

is a Martingale. Notice that if 2) is changed to \(\mathbb{E}(\Delta_{n+1} | \mathcal{F}_n) \geq 0\) \((\mathbb{E}(\Delta_{n+1} | \mathcal{F}_n) \leq 0)\), then \((M_n)\) is a sub-martingale (super-martingale).

Theorem 12.2 (Doob’s martingale transform). Suppose \((M_n, \mathcal{F}_n)_{n \geq 0}\) is a martingale and \((H_n, \mathcal{F}_n)_{n \geq 0}\) is predictable. Define

\[
\Delta_n = M_n - M_{n-1}, \quad (H \cdot M)_n = \sum_{i=1}^{n} H_i \Delta_i
\]

for \(n \geq 1\). Then \(((H \cdot M)_n, \mathcal{F}_n)\) is a martingale whenever \(\mathbb{E}|(H \cdot M)_n| < \infty\) for all \(n\).

Proof. Since \((H \cdot M)_{n+1} - (H \cdot M)_n = H_{n+1} \cdot \Delta_{n+1}\), we have for \(n \geq 0\), \(\mathbb{E}(H_{n+1} \Delta_{n+1} | \mathcal{F}_n) = H_{n+1} \mathbb{E}(\Delta_{n+1} | \mathcal{F}_n) = 0\).

Observe that if \(H_n \geq 0\), then \((M_n)\) is a sub-martingale (super-martingale) implies \((H \cdot M)_n\) is a sub-martingale (super-martingale).

Theorem 12.3. If \((M_n, \mathcal{F}_n)\) is a martingale (sub/super-martingale) and \(T\) is a stopping time with respect to a filtration \((\mathcal{F}_n)\), then

\[
W_n = M_{n \wedge T}, n \geq 0
\]

is a \((\mathcal{F}_n)\)-martingale (sub/super-martingale).

Proof. Notice that

\[
M_{n \wedge T} = \sum_{i=1}^{n \wedge T} \Delta_i = \sum_{i=1}^{n} \Delta_i 1_{T \geq i}.
\]

In particular, \(\mathbb{E}(M_{n \wedge T}) = \mathbb{E}(M_0)\) for all \(n \geq 0\).
Note that, $M_{n,T} \xrightarrow{a.s.} M_T$ as $n \to \infty$ when $M_T$ is well-defined a.s.

**Example 12.4.** Let $(X_i)$ be an iid sequence of random variable which is identically distributed to simple random walk on $\mathbb{Z}$. Let $M_n = S_n = \sum_{i=1}^{n} X_i$. Then $E(M_0) = 0$. Let $T = \inf\{n \mid S_n \geq 1\}$. Then
\[ M_T = S_T = 1 \text{ yields } E(M_T) = 1 \neq E(M_0). \]
This example shows that it is not always true that $E(M_0) = \lim E(M_{n,T}) = E(\lim M_{n,T}) = E(M_T)$.

Notice that if $T$ is bounded by some $k$, then $n \wedge T = T$ for $n \geq k$ and $M_{n,T} = M_T, n \geq k$.

### 12.2 Wald’s Identities

**Theorem 12.5 (Wald’s first identity).** Let $(X_n, n \geq 1)$ be a sequence of mean zero independent r.v.s with $\sup_{i \geq 1} E|X_i| < \infty$. Consider the martingale
\[ M_n = S_n := X_1 + X_2 + \cdots + X_n, \quad F_n := \sigma(X_1, \ldots, X_n), \quad n \geq 0. \]
Let $T$ be a $(F_n, n \geq 0)$ stopping time with $E(T) < \infty$. Then
\[ E(M_T) = E(M_0) = 0. \]
In particular, if $X_i$’s are i.i.d. r.v.s with $E X_1 = \mu$ and $T$ is a $(\sigma(X_1, \ldots, X_n), n \geq 0)$ stopping time with $E(T) < \infty$, then
\[ E S_T = \mu E T. \]

**Proof.** We have $M_{T \wedge n} = \sum_{i=1}^{n} X_i 1_{T \geq i} \rightarrow M_T$ a.s. as $n \rightarrow \infty$. Moreover $E(M_{T \wedge n}) = 0$ for $n \geq 0$. By the fact that $X_i \perp F_{i-1}, 1_{T \geq i} \in F_{i-1}$ we have
\[ E \sum_{i=1}^{\infty} |X_i| 1_{T \geq i} = \sum_{i=1}^{\infty} E(|X_i| 1_{T \geq i}) = \sum_{i=1}^{\infty} E|X_i| \cdot E(1_{T \geq i}) \leq \sup_{i \geq 1} E|X_i| \cdot E(T) < \infty. \]
Using DCT with $|M_{T \wedge n}| \leq \sum_{i=1}^{\infty} |X_i| 1_{T \geq i}$ for all $n \geq 0$ we get the result. \hfill \blacksquare

**Theorem 12.6 (Wald’s second identity).** Let $(X_n, n \geq 1)$ be a sequence of mean zero independent r.v.s with $\sup_{i \geq 1} E X_i^2 < \infty$. Consider the martingale
\[ M_n = S_n^2 - \sum_{i=1}^{n} E X_i^2, \quad F_n := \sigma(X_1, \ldots, X_n), \quad n \geq 0 \]
where $S_n := X_1 + X_2 + \cdots + X_n$. Let $T$ be a $(F_n, n \geq 0)$ stopping time with $E(T) < \infty$. Then
\[ E(M_T) = E(M_0) = 0. \]
In particular, if $X_i$’s are i.i.d. r.v.s with $E X_1 = 0, \text{Var}(X_1) = \sigma^2$ and $T$ is a $(\sigma(X_1, \ldots, X_n), n \geq 0)$ stopping time with $E(T) < \infty$, then
\[ E S_T^2 = \sigma^2 E(T). \]
Proof. Let $\sigma_i^2 = \mathbb{E} X_i^2$ for $i \geq 1$. We have
\[
\mathbb{E} M_{T \wedge n} = \mathbb{E} S_{T \wedge n}^2 - \mathbb{E} \sum_{i=1}^{T \wedge n} \sigma_i^2 = 0 \text{ for all } n \geq 0.
\]
Now
\[
S_{T \wedge n} = \sum_{i=1}^{n} X_i 1_{T \geq i} \rightarrow S_T = \sum_{i=1}^{\infty} X_i 1_{T \geq i} \text{ a.s. as } n \rightarrow \infty,
\]
$X_i$ is orthogonal to $\mathcal{F}_{i-1}$ and $1_{T \geq i} \in \mathcal{F}_{i-1}$ for $i \geq 1$. Thus
\[
\mathbb{E}(X_i 1_{T \geq i} \cdot X_j 1_{T \geq j}) = 0
\]
for all $i \neq j$ and for $0 \leq n < \infty$
\[
\sup_{m \geq n} ||S_{T \wedge m} - S_{T \wedge n}||_2 = \sup_{m \geq n} \sum_{i=n+1}^{m} ||X_i 1_{T \geq i}||_2
\]
\[
= \sup_{m \geq n} \sum_{i=n}^{m} \mathbb{E}(X_i^2 1_{T \geq i}) = \sum_{i=n}^{\infty} \mathbb{E}(X_i^2) \cdot \mathbb{E}(1_{T \geq i}) \leq \sup_{i \geq 1} \mathbb{E}(X_i^2) \cdot \sum_{i=n}^{\infty} \mathbb{E}(1_{T \geq i}).
\]
Thus, $(S_{T \wedge n})$ is $L^2$-Cauchy and $\mathbb{E} S_{T \wedge n}^2 \rightarrow \mathbb{E} S_T^2$. Similar argument and DCT shows that $\mathbb{E} \sum_{i=1}^{T \wedge n} \sigma_i^2 \rightarrow \mathbb{E} \sum_{i=1}^{T} \sigma_i^2$. Thus we have the result. $lacksquare$

Theorem 12.7 (Wald’s third identity). Let $(X_n, n \geq 1)$ be a sequence of i.i.d. r.v.s with $\mathbb{E}(e^{\theta X_1}) = \phi(\theta) < \infty$. Let $T$ be a $(\sigma(X_1, X_2, \ldots, X_n))$-stopping time. If $T$ is a.s. bounded or $\phi(\theta)^{-n} \cdot e^{\theta S_n} \cdot 1_{T \geq n} \leq K$ for all $n \geq 0$,
then
\[
\mathbb{E}\left(\phi(\theta)^{-T} \cdot e^{\theta S_T}\right) = 1.
\]

Proof. Since $M_n := \phi(\theta)^{-n} \cdot e^{\theta S_n}$ is a martingale w.r.t. the filtration $\mathcal{F}_n := \sigma(X_1, \ldots, X_n), n \geq 0$, we have $\mathbb{E}(M_{T \wedge n}) = \mathbb{E}(M_{T \wedge 0}) = 1$. The results follows in the first case by the fact that $T \leq K$ a.s. implies $T \wedge K = T$ and in the second case by DCT. $lacksquare$

We provide some examples to illustrate how to use Wald’s identities.

Example 12.8 (SSRW (Simple Symmetric Random Walk) on Z). Let $(X_n)$ be a sequence of i.i.d. r.v.s with $\mathbb{P}(X_1 = +1) = \mathbb{P}(X_1 = -1) = 1/2, \mathbb{E} X_1 = 0, \mathbb{E} X_1^2 = 1$. Define $S_0 = 0, S_n := S_{n-1} + X_n, n \geq 1$. Let $a, b \in \mathbb{Z}$ with $a, b > 0$. Define
\[
T = \inf\{n \geq 0 \mid S_n = -a \text{ or } S_n = b\}.
\]
Then $T < \infty$, by using second moment martingale to get $\mathbb{E}(T \wedge n) = \mathbb{E} S_{T \wedge n}^2 \leq a^2 + b^2$ for all $n \geq 0$.
Moreover, by Wald’s first identity we have $\mathbb{E}(S_T) = 0$, i.e.,
\[
a \mathbb{P}(S_T = -a) = b \mathbb{P}(S_T = b) = b(1 - \mathbb{P}(S_T = a)).
\]
So
\[
\mathbb{P}(S_T = -a) = \frac{b}{a+b}, \mathbb{P}(S_T = b) = \frac{a}{a+b}.
\]
Example 12.9 (SRW (Simple Random Walk) with drift). Fix \( p \in (0,1), p \neq 1/2 \). Let \((X_n)\) be a sequence of i.i.d. r.v.s with \( \mathbb{P}(X_1 = +1) = p, \mathbb{P}(X_1 = -1) = q = 1 - p \), \( \mu := \mathbb{E}X_1 = 2p - 1, \mathbb{E}X_1^2 = 1 \). Define \( S_0 = 0, S_n := S_{n-1} + X_n, n \geq 1 \). Let \( a, b \in \mathbb{Z} \) with \( a, b > 0 \). Define

\[
T = \inf\{n \geq 0 \mid S_n = -a \text{ or } S_n = b\}.
\]

By Wald’s first identity, we have

\[
\mathbb{E}(T \wedge n) = \mu^{-1} \mathbb{E}S_{T \wedge n} \leq |\mu|^{-1} \cdot (a + b) \text{ for all } n \geq 0
\]

and thus \( \mathbb{E}T < \infty \). Choose \( \theta = \log(q/p) \in \mathbb{R} \) so that \( e^\theta = q/p \) and \( \mathbb{E} e^{\theta X_1} = 1 \). We also have \( e^{\theta S_n} \mathbb{1}_{\{T \geq n\}} \leq e^{|\theta| \cdot (a+b)} \) for all \( n \geq 0 \). By Wald’s Third Identity we have

\[
1 = \left(\frac{q}{p}\right)^{-a} \mathbb{P}(S_T = -a) + \left(\frac{q}{p}\right)^b (1 - \mathbb{P}(S_T = -a)).
\]

Thus, we have

\[
\mathbb{P}(S_T = -a) = \frac{p^b q^a - q^{a+b}}{p^{a+b} - q^{a+b}} \quad \text{and} \quad \mathbb{P}(S_T = b) = \frac{p^a + b q^a - p q^{a+b}}{p^{a+b} - q^{a+b}}.
\]

Using Wald’s third identity with general \( \theta \in \mathbb{R} \), one can calculate the probabilities \( \mathbb{P}(T = k), k \geq 0 \) (see the homework exercise). The first identity can be generalized under the assumption

\[
\sup_{n \geq 1} \mathbb{E}(|M_n - M_{n-1}| \mid \mathcal{F}_{n-1}) \leq K < \infty \text{ a.s.}
\]
13.1 Martingale Convergence Theorem

Let \((M_n, F_n)_{n \geq 0}\) be a submartingale and \(a < b\). We define \(N_0 = -1\) and
\[
N_1 = \inf \{i > N_0 \mid M_i \leq a\}, \quad N_2 = \inf \{i > N_1 \mid M_i \geq b\},
\]
\[
N_3 = \inf \{i > N_2 \mid M_i \leq a\}, \quad N_4 = \inf \{i > N_3 \mid M_i \geq b\}.
\]
In general, \(N_{2k-1} = \inf \{i > N_{2k-2} \mid M_i \leq a\}\), \(N_{2k} = \inf \{i > N_{2k-1} \mid M_i \geq b\}\) for \(k \geq 1\).

![Diagram]

First we claim that.

**Lemma 13.1.** For \(i \geq 1\), \(N_i\) is a stopping time w.r.t. the filtration \((F_n)_{n \geq 0}\).

**Proof.** The proof is by induction. Clearly, \(N_1\) is a stopping time. Suppose \(N_1, \ldots, N_{i-1}\) are stopping time. If \(i\) is even, then
\[
\{N_i \leq n\} = \bigcup_{j=1}^{n} \{N_{i-1} = j - 1\} \cap \{j \leq N_i \leq n\} = \bigcup_{j=1}^{n-1} \{N_{i-1} = j - 1\} \cap \bigcup_{k=j}^{n} \{M_k \geq b\} \in F_n.
\]
Similarly, when \(i\) is odd,
\[
\{N_i \leq n\} = \bigcup_{j=1}^{n} \{N_{i-1} = j - 1\} \cap \{j \leq N_i \leq n\} = \bigcup_{j=1}^{n-1} \{N_{i-1} = j - 1\} \cap \bigcup_{k=j}^{n} \{M_k \leq a\} \in F_n.
\]

Define the upcrossing random variable
\[
U_n(a, b) = \sup \{k \mid N_{2k} \leq n\}.
\]
For fixed \(m \geq 1\) and any \(k \geq 1\),
\[
\{N_{2k-1} < m \leq N_{2k}\} = \{N_{2k-1} \leq m - 1 \text{ and } N_{2k} > m - 1\} \in F_{m-1}.
\]
Thus \(H_n = \bigcup_{k \geq 1} \{N_{2k-1} < n \leq N_{2k}\}, n \geq 0\) is a predictable sequence and
\[
(H \cdot M)_n \geq (b - a)U_n(a, b).
\]
Theorem 13.2 (Upcrossing Inequality). For any $a < b$ and any submartingale $(M_n, \mathcal{F}_n)_{n \geq 0}$, we have,
\[(b - a) \mathbb{E}(U_n(a, b)) \leq \mathbb{E}(M_n - a)^+ - \mathbb{E}(M_0 - a)^+ \text{ for all } n \geq 1.\]

Proof. Define $Y_n := \phi(M_n) = (M_n - a)^+ + a, n \geq 0$ where $\phi(x) = (x - a)^+ + a$ is a non-decreasing convex function. By conditional Jensen’s inequality, we have
\[\mathbb{E}(Y_{n+1} | \mathcal{F}_n) = \mathbb{E}(\phi(M_{n+1}) | \mathcal{F}_n) \geq \phi(\mathbb{E}(M_{n+1} | \mathcal{F}_n)) \geq \phi(M_n) = Y_n.\]
Thus $(Y_n, \mathcal{F}_n)_{n \geq 0}$ is a submartingale. Moreover, upcrossings for $M_n$ and $Y_n$ over the interval $[a, b]$ are the same. Thus $(H \cdot Y)_n \geq (b - a)U_n(a, b)$ implies that
\[(b - a) \mathbb{E}(U_n(a, b)) \leq \mathbb{E}(H \cdot Y)_n.\]
Now, we have $Y_n - Y_0 = (H \cdot Y)_n + ((1 - H) \cdot Y)_n$ and $\mathbb{E}((1 - H) \cdot Y)_n \geq 0$. Thus,
\[\mathbb{E}(H \cdot Y)_n \leq \mathbb{E}(Y_n - Y_0).\]

Theorem 13.3 (Martingale Convergence Theorem (MGCT)). If $(M_n, \mathcal{F}_n)_{n \geq 0}$ is a submartingale with $\sup_n \mathbb{E}M_n^+ < \infty$, then
\[M_n \to M_\infty \text{ a.s.}\]
for some $M_\infty \in L^1(\mathcal{F})$.

Proof. Let $K = \sup_n \mathbb{E}M_n^+$. Fix $a < b$. We have
\[(b - a) \mathbb{E}U_n(a, b) \leq \mathbb{E}(M_n - a)^+ \leq \mathbb{E}M_n^+ + a \leq K + a.\]
Note that, $U_n(a, b) \uparrow U(a, b)$ as $n \to \infty$ where $U(a, b)$ is the total number of upcrossings of the interval $[a, b]$. Thus, $\mathbb{E}U(a, b) < \infty$ and in particular, $U(a, b) < \infty$ a.s. Thus,
\[\mathbb{P}(\cup_{a < b, a, b \in Q} \{U(a, b) = \infty\}) = 0,\]
which implies that
\[\mathbb{P}(\liminf M_n = \limsup M_n) = 1 - \mathbb{P}(\liminf M_n < \limsup M_n)\]
\[= 1 - \mathbb{P}(\cup_{a < b, a, b \in Q} \{\liminf M_n < a < b < \limsup M_n\})\]
\[= 1 - \mathbb{P}(\cup_{a < b, a, b \in Q} \{U(a, b) = \infty\}) = 1.\]
In particular,
\[M_n \to M_\infty \text{ a.s.}\]
for some r.v. $M_\infty$. Now $M_n^+ \to M_\infty^+$ and $M_n^- \to M_\infty^-$ a.s. By Fatou’s Lemma, we have
\[\mathbb{E}M_\infty^+ = \mathbb{E}(\liminf M_n^+) \leq \liminf \mathbb{E}M_n^+, \quad \mathbb{E}M_\infty^- = \mathbb{E}(\liminf M_n^-) \leq \liminf \mathbb{E}M_n^-.
\]
Moreover,
\[\mathbb{E}M_n^- = \mathbb{E}(M_n^+ - M_n) = \mathbb{E}M_n^+ - \mathbb{E}M_n \leq \sup \mathbb{E}M_n^+ - \mathbb{E}M_0\]
which implies that $\mathbb{E}M_\infty^+, \mathbb{E}M_\infty^- < \infty$. \[\blacksquare\]
Note that, for a submartingale \((X_n, \mathcal{F}_n)_{n \geq 1}\)
\[
\sup_n \mathbb{E} X_n^+ < \infty \iff \sup_n \mathbb{E} |X_n| < \infty.
\]

**Corollary 13.4.** A positive super-martingale converges a.s. to an integrable r.v.

**Proof.** If \((X_n)_{n \geq 0}\) is a positive super-martingale, then \((-X_n)_{n \geq 0}\) is a negative sub-martingale, i.e., \(\sup_n \mathbb{E}(-X_n)^+) = 0 < \infty\) and thus we can apply MGCT to \(-X_n\) to get the desired result. 

**Example 13.5 (Galton-Watson Process).** Let \(X\) be a non-negative integer valued random variable with \(\mathbb{E} X = \mu\). Let \((X_i^{(j)})\) be a sequence of iid r.v.s with \(X_i^{(j)} \overset{d}{=} X\).

(a) If \(Z_0 = 1, Z_1 = X_1^{(1)}\) and
\[
Z_n = \sum_{i=1}^{Z_{n-1}} X_i^{(n)} \text{ for } n \geq 1,
\]
then
\[
(M_n = \mu^{-n} \cdot Z_n, \mathcal{F}_n = \sigma(X_i^{(j)}, j \leq n))
\]
is a non-negative martingale. Thus, MGCT implies that
\[
M_n = \mu^{-n} \cdot Z_n \xrightarrow{a.s.} M_\infty.
\]
If \(\mu < 1\), we get \(Z_n \overset{a.s.}{\rightarrow} 0\) and so \(\mathbb{P}\text{(finite time extinction)} = 1\).

(b) Suppose \(\mathbb{E} X = \mu > 1\). Then there exists a unique \(\theta \in (0, 1)\) such that \(\mathbb{E} \theta^X = \theta\) and
\[
(Q_n = \theta^{Z_n}, \mathcal{F}_n = \sigma(X_i^{(j)}, j \leq n))
\]
is a bounded non-negative martingale. By MGCT \(Q_n = \theta^{Z_n} \xrightarrow{a.s.} Q_\infty\) and by DCT \(\mathbb{E} Q_\infty = \theta\). Since we know that \(\mu^{-n} Z_n \xrightarrow{a.s.} M_\infty\) and \(\mu > 1\), we have \(\lim_{n \to \infty} Z_n \in \{0, \infty\}\) and \(\mathbb{P}(Z_n \to 0) = \theta\).

**Example 13.6.** For \(X \in L^1, X_n := \mathbb{E}(X \mid \mathcal{F}_n)\) is a martingale for any filtration \(\mathcal{F}_n\). Moreover, \(\mathbb{E}|X_n| \leq \mathbb{E}|X| < \infty\) implies that \(X_n \to X_\infty\) a.s., where
\[
X_\infty := \mathbb{E}(X \mid \mathcal{F}_\infty), \quad \mathcal{F}_\infty = \sigma(\cup_{n=1}^\infty \mathcal{F}_n).
\]

**Example 13.7 (MGCT does not imply \(L^1\) convergence).** Consider \(\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)\) with i.i.d. r.v.s \(x_i = \pm 1\) with probability \(\frac{1}{2}\) and
\[
S_0 = 1, \quad S_n = S_{n-1} + \xi_n \text{ for } n \geq 1 \text{ and the stopping time } N = \inf\{n \geq 0 \mid S_n = 0\}.
\]
Then \((M_n = S_{N\wedge n})\) is a non-negative martingale with respect to \((\mathcal{F}_n)\). By MGCT, \(M_n \to M_\infty\) a.s. But, \(M_\infty = 0\) and \(\mathbb{E}(M_n) = \mathbb{E}(M_0) = 1\) for \(n \geq 1\). Thus \(M_n\) does not converge to \(M_\infty\) in \(L^1\).

**Exercise 13.8.** In the previous example, show that \(\mathbb{P}(\sup_n S_{N\wedge n} \geq k) = \frac{1}{k}, k \geq 1\).

**Exercise 13.9.** If \((M_n, \mathcal{F}_n)_{n \geq 1}\) is a martingale such that \(M_n \to M_\infty\) a.s. and in \(L^1\), then \(M_n = \mathbb{E}(M_\infty \mid \mathcal{F}_n)\) for all \(n \geq 1\).

**Hint:** \(M_n = \mathbb{E}(M_m \mid \mathcal{F}_n)\) for any \(m \geq n\).

From the previous example, we see that we need additional conditions to get \(L^1\) convergence.
13.2 Doob’s Decomposition

From the definition of a sub-Martingale \((X_n, F_n)_{n \geq 0}\), it is easy to see that if at every time point \(n\) we subtract a positive \(F_n\)-measurable r.v. \(E(X_{n+1} - X_n \mid F_n)\) we can keep the conditional mean zero. Thus, the process \(X_n - \sum_{i=0}^{n-1} E(X_{i+1} - X_i \mid F_i)\) will be a Martingale. Formally, we have the following.

**Theorem 13.10 (Doob’s Decomposition).** Let \((X_n, F_n)_{n \geq 0}\) be a sub-martingale. Then there exists a unique decomposition \(X_n = M_n + A_n\), where \((M_n, F_n)_{n \geq 0}\) is a martingale with \(M_0 \equiv 0\) and \((A_n)_{n \geq 0}\) is predictable and non-decreasing.

*Proof.* If \(X_n = M_n + A_n\), then \(\Delta X_n = X_n - X_{n-1} = \Delta M_n + \Delta A_n\). Since \(\Delta A_n\) is \(F_{n-1}\) measurable and \(E(M_n - M_{n-1} \mid F_{n-1}) = 0\), this implies that

\[
E(\Delta X_n \mid F_{n-1}) = \Delta A_n, \ \forall n \geq 0.
\]

Thus, \(A_n - A_0 = \sum_{i=0}^{n} E(X_i - X_{i-1} \mid F_{i-1})\), \(A_0 = X_0\) and \(M_n = X_n - A_n = \sum_{i=1}^{n} (X_i - E(X_i \mid F_{i-1}))\). Now, \((X_n)\) is a sub-martingale, hence \(E(X_{n+1} \mid F_n) - X_n \geq 0\), which implies that \(A_n\) is non-decreasing. We can easily check that \(A_n\) is predictable and \(M_n\) is a mean zero martingale.

13.3 Martingale Central Limit Theorem

**Theorem 13.11.** Let \((M_n, F_n)_{n \geq 1}\) be a mean zero martingale with \(s_n^2 := E(M_n^2) < \infty, n \geq 1\) and \(\sigma_n^2 \to \infty\) as \(n \to \infty\). Then,

\[
\frac{M_n}{s_n} \xrightarrow{(d)} N(0, 1),
\]

if,

(i) \(s_n^{-2} \sum_{i=1}^{n} E(\Delta_i^2 I_{|\Delta_i| \geq \varepsilon s_n}) \overset{n \to \infty}{\to} 0\) for all \(\varepsilon > 0\), where \(\Delta_i = M_i - M_{i-1}\).

(ii) \(\frac{1}{s_n^2} \sum_{i=1}^{n} E(\Delta_i^2 \mid F_{i-1}) \overset{p}{\to} 1\).

13.4 Maximal inequalities

**Theorem 13.12 (Doob’s Maximal Inequality).** If \((X_n, F_n)_{n \geq 0}\) is a non-negative sub-martingale, and \(\bar{X}_n := \max_{1 \leq i \leq n} X_i\), then for all \(\lambda > 0\), \(A = \{\bar{X}_n \geq \lambda\}\),

\[
P(A) \leq \frac{1}{\lambda} E(X_n 1_A) \leq \frac{1}{\lambda} E(X_n).
\]

*Proof.* Define the stopping time \(N := \inf\{i \mid X_i \geq \lambda\} \wedge n\). Then for \(A = \{\bar{X}_n \geq \lambda\} = \{X_N \geq \lambda\}\) we have

\[
P(A) \leq \frac{1}{\lambda} E(X_N 1_A) = \frac{1}{\lambda} (E(X_N - E(X_N 1_{A^c})) \leq \frac{1}{\lambda} (E(X_n) - E(X_n 1_{A^c})) = \frac{1}{\lambda} E(X_n 1_A),
\]

since \(N = n\) if \(\bar{X}_n < \lambda\).
Note that

- If $X_n$ is a sub-martingale, then $X_n^+$ is a non-negative sub-martingale.
- If $X_n$ is a martingale, then $|X_n|$ is a non-negative sub-martingale.

The above maximal inequality holds for all sub/super/martingales if we replace the inequality by

$$
P(\max_{1 \leq i \leq n} |X_i| \geq \lambda) \leq \frac{3}{\lambda} \sup_{1 \leq i \leq n} E(|X_i|).
$$

But this is not enough for the next result.

**Theorem 13.13 (L^p Maximal Inequality).** Let $(X_n, \mathcal{F}_n)_{n \geq 0}$ be a non-negative sub-martingale, with $E|X_n|^p < \infty$ for some $p > 1$. Then,

$$
\|\bar{X}_n\|_p \leq \frac{p}{p-1} \|X_n\|_p.
$$

If $p = 1$, then

$$
\|\bar{X}_n\|_1 \leq 1 + E(|X_n| \log^+ X_n) \frac{1}{1-e^{-1}}, \text{ provided that } E(|X_n| \log^+ X_n) < \infty.
$$

Here we mention that the constant $p/(p-1)$ is optimal for $p > 1$. If we do not care about the constant, one can directly prove that $\|\bar{X}_n\|_p \leq c \|X_n\|_p$ for some $c \in (0, \infty)$.

**Proof.** First we claim that

$$
E(\bar{X}_n^p) = \int_0^\infty py^{p-1} P(\bar{X}_n \geq y) dy.
$$

Applying Fubini’s, we get

$$
\int_0^\infty py^{p-1} P(\bar{X}_n \geq y) dy = \int_0^\infty \int_\Omega py^{p-1} \mathbb{1}_{\bar{X}_n(\omega) \geq y} dy P(d\omega) = \int_\Omega \bar{X}_n(\omega)^p P(d\omega) = E(\bar{X}_n^p).
$$

Now, by Doob’s Maximal Inequality (taking $\lambda = y$), for $p > 1$,

$$
E|\bar{X}_n|^p \leq \int_0^\infty py^{p-1} \cdot \frac{1}{y} E(\bar{X}_n \mathbb{1}_{\bar{X}_n \geq y}) dy = E(\bar{X}_n \int_0^{\bar{X}_n} py^{p-2} dy) = \frac{p}{p-1} E(\bar{X}_n \bar{X}_n^{p-1}).
$$

Take $q = \frac{p}{p-1}$, so that $\frac{1}{p} + \frac{1}{q} = 1$. By Hölder’s Inequality,

$$
\frac{p}{p-1} E(\bar{X}_n \bar{X}_n^{p-1}) \leq \frac{p}{p-1} \|X_n\|_p \|\bar{X}_n^{p-1}\|_q,
$$

and

$$
\|\bar{X}_n^{p-1}\|_q = (E(\bar{X}_n^{p-1} \bar{X}_n^{p-1}))^{\frac{p-1}{p}} = (E(\bar{X}_n^p))^{1-\frac{1}{p}}.
$$

Thus,

$$
E|\bar{X}_n|^p \leq \frac{p}{p-1} \|X_n\|_p (E(\bar{X}_n^p))^{1-\frac{1}{p}}
$$

$$
\|\bar{X}_n\|_p = (E|\bar{X}_n|^p)^{\frac{1}{p}} \leq \frac{p}{p-1} \|X_n\|_p.
$$

Here we assumed that $\bar{X}_n^{p-1} \in L^q$. For a more rigorous proof, work with $\bar{X}_n \wedge M$, show that $\|\bar{X}_n \wedge M\|_p \leq \frac{p}{p-1} \|X_n\|_p$, and take $M \uparrow \infty$. The proof for $p = 1$ is left as an exercise. \qed
Exercise 13.14. Show that

1. For $x > 0, y > 0$, $xy \leq e^{x-1} + y \log^+ y$. (Young's Inequality)

2. $\|X_n\| \leq \frac{1 + E(X_n \log^+ X_n)}{1 - e^{-1}}$.

13.5 $L^p$ Convergence Theorem

Theorem 13.15 ($L^p$ Convergence Theorem). Let $(X_n, F_n)_{n \geq 0}$ be a non-negative sub-martingale, with

$$\sup_{n \geq 1} E|X_n|^p < \infty \text{ for some } p > 1$$

Then, $X_n \overset{L^p}{\to} X_\infty$, in other words, $E|X_\infty|^p < \infty$, and $E|X_n - X_\infty|^p \to 0$.

Proof. By Martingale Convergence Theorem, $X_n \overset{a.s.}{\to} X_\infty$. Thus to show that $X_n \overset{L^p}{\to} X_\infty$, we need to show two things: 1) $X_\infty \in L^p$ and 2) $X_n \overset{L^p}{\to} X_\infty$.

By $L^p$ Maximal Inequality, we have $\sup_{n \geq 1} \|X_n\|_p \leq \frac{p}{p-1} \sup_{n \geq 1} \|X_n\|_p < \infty$. By Monotone Convergence Theorem, $X_n \uparrow X_\infty$, so $E X_n^p \to E X_\infty^p$. Now $X_n \leq X_\infty$ implies that $X_\infty \leq X_\infty$ and $|X_n - X_\infty| \leq 2X_\infty$. Thus, $X_\infty \in L^p$, and by DCT we have, $E|X_n - X_\infty|^p \to 0$.

Corollary 13.16. If $(M_n, F_n)_{n \geq 1}$ is a $L^p$ bounded martingale, i.e. $\sup_{n \geq 0} E|M_n|^p < \infty$, then $M_n \overset{a.s.}{\to} M_\infty$.

Proof. By Martingale Convergence Theorem, $M_n \overset{a.s.}{\to} M_\infty$. Take $X_n = |M_n|$, then $X_n$ is a non-negative sub-martingale, and $X_\infty \in L^p$. $X_\infty$ is also a bound for $M_\infty$.

Corollary 13.17. Let $X \in L^p$, and $F_0 \subseteq F_1 \subseteq \ldots$ be a filtration. Then $E(\cdot | F_n) \overset{a.s.}{\to} E(\cdot | F_\infty)$.

Proof. By MGCT the almost sure part holds, and $E(|E(X | F_n)|^p) \leq E|X|^p < \infty$, thus the $L^p$ convergence follows.

13.6 $L^1$ Convergence Theorem

The $L^p$ convergence theorem holds for $p > 1$. What happens when $p = 1$?

Theorem 13.18. If $(M_n, F_n)$ is a martingale, and $M_n \to M_\infty$ in $L^1$, then $M_n = E(M_\infty | F_n)$.

Proof. For $m \geq n$, $M_n = E(M_m | F_n)$ for all $m \geq n$. Thus $E|E(M_m | F_n) - E(M_\infty | F_n)| \leq E|M_m - M_\infty | \overset{m \to \infty}{\to} 0$, so $E|M_n - E(M_\infty | F_n)| = 0$. Thus, $M_n = E(M_\infty | F_n)$.

Definition 13.19 (Uniform Integrability). A collection of random variables, $\{X_\lambda : \lambda \in \Lambda\}$, is uniformly integrable if $\forall \varepsilon > 0, \exists k > 0$, such that

$$\sup_{\lambda \in \Lambda} E(|X_\lambda| 1_{\{X_\lambda \geq k\}}) \leq \varepsilon.$$
Corollary 13.20. If \((X_\lambda, \lambda \in \Lambda)\) is u.i., then \(\sup_{\lambda \in \Lambda} E|X_\lambda| < \infty\).

Proof. Fix \(\varepsilon = 1\), find \(k\) so that \((X_\lambda, \lambda \in \Lambda)\) is u.i.. Then

\[
E|X_\lambda| = E|X_\lambda|1_{|X_\lambda| < k} + E|X_\lambda|1_{|X_\lambda| \geq k} \leq k + \varepsilon < \infty.
\]

Exercise 13.21. (i) If \(|X_n| \leq Y\) for all \(n\) for some \(Y \in L^1(\mathcal{F})\), then \((X_n)\) is u.i.

(ii) If \((X_n)\) and \((Y_n)\) are both u.i., then so is \((X_n + Y_n)\).

(iii) Show that \(X_n \xrightarrow{L^1} x\) implies that \((X_n)\) is u.i.

One can check that uniform integrability is just the right condition to go from convergence in probability to \(L^1\)-convergence. So that, convergence in probability + uniform integrability = \(L^1\) convergence.

Theorem 13.22. Suppose \(X_n \to X_\infty\) in probability and \((X_n)_{n \geq 0}\) is u.i., then \(X_\infty \in L^1\) and \(X_n \xrightarrow{L^1} X_\infty\).

Proof. Fix \(\varepsilon > 0\), choose \(k\) so that \(E(|X_n|1_{\{|X_n| \geq k\}}) \leq \varepsilon\) for all \(n\). We know that \(X_\infty \in L^1\) by Fatou’s Lemma \((E\liminf |X_n| \leq \liminf E|X_n| \leq \infty)\). So, we can choose \(k\) large so that \(E(|X_\infty|1_{\{|X_\infty| \geq k\}}) \leq \varepsilon\).

Consider

\[
\phi_k(x) = \begin{cases} 
  x & |x| \leq k; \\
  k & x > k; \\
  -k & x < -k.
\end{cases}
\]

Then \(\phi_k(X_n) \to \phi_k(X_\infty)\) in probability and is bounded by \(k\) uniformly. By BCT, \(E|\phi_k(X_n) - \phi_k(X_\infty)| \xrightarrow{n \to \infty} 0\). Partitioning \(E|X_n - X_\infty|\) to two parts, one with either \(|X_n|\) or \(|X_\infty| \geq k\), the other with both \(|X_n|\) and \(|X_\infty|\) bounded by \(k\), we have

\[
E|X_n - X_\infty| \leq E(|X_n|1_{\{|X_n| \geq k\}}) + E(|X_\infty|1_{\{|X_\infty| \geq k\}}) + E|\phi_k(X_n) - \phi_k(X_\infty)|
\]

\[
\leq 2\varepsilon + E|\phi_k(X_n) - \phi_k(X_\infty)|
\]

\[
\implies \limsup E|X_n - X_\infty| \leq 2\varepsilon.
\]

Since \(\varepsilon\) is arbitrary, taking \(\varepsilon \to 0\), \(E|X_n - X_\infty| \to 0\). So \(X_n \xrightarrow{L^1} X_\infty\).

Note that

- If \((X_n)\) u.i., then take \(\phi \geq 0\), \(\phi(x) \uparrow \infty\) as \(x \uparrow \infty\), then

\[
\sup E(|X_n|1_{|X_n| \geq \lambda}) = \sup E(|X_n| \cdot \frac{\phi(|X_n|)}{\phi(|X_n|)}1_{|X_n| \geq \lambda}) \leq \frac{1}{\phi(\lambda)} \sup E|X_n|\phi(|X_n|).
\]

- \((X_n)\) u.i. \(\iff\) \(\lim_{k \to \infty} \sup_n E(|X_n|1_{|X_n| \geq k}) = 0\).

Theorem 13.23 (\(L^1\) Convergence Theorem.). Let \((M_n, \mathcal{F}_n)\) be a martingale. Then the following are equivalent:

**Part 1:**

**Part 2:**
(i) \( M_n \xrightarrow{a.s.} M_\infty \).

(ii) \( M_n = \mathbb{E}(M_\infty \mid \mathcal{F}_n), \forall n \geq 0. \)

(iii) \( (M_n)_{n \geq 0} \) is u.i.

**Proof.** i) \( \implies \) ii) follows from Theorem 23.6. iii) \( \implies \) i) follows from MGCT and Theorem 23.9. We’ll show ii) \( \implies \) iii):

\[
\mathbb{E}(|M_n| \mathbb{1}_{|M_n| \geq k}) = \mathbb{E}(|M_\infty| \mathbb{1}_{|M_\infty| \geq k}) \leq \mathbb{E}(\mathbb{E}(|M_\infty| \mathbb{1}_{|M_\infty| \geq k} \mid \mathcal{F}_n)) = \mathbb{E}(\mathbb{E}(|M_\infty| \mathbb{1}_{|M_\infty| \geq k} \mid \mathcal{F}_n)).
\]

For \( \forall n \geq 0, \mathbb{P}(|M_n| \geq k) \leq \frac{1}{k} \mathbb{E}|M_n| \leq \frac{1}{k} \mathbb{E}|M_\infty|. \) We claim that \( \forall \varepsilon > 0, \exists \delta \) s.t. \( \mathbb{P}(A) \leq \delta \) implies \( \mathbb{E}|X| \mathbb{1}_A < \varepsilon. \) Otherwise, \( \exists \varepsilon > 0, \delta_n \downarrow 0, \) and \( \mathbb{P}(A_n) \leq \delta_n, \) such that \( \mathbb{E}|X| \mathbb{1}_{A_n} \geq \varepsilon. \) (contradiction)

By the claim, \( \forall \varepsilon > 0, \) choose \( \delta \) satisfying the claim. Choose \( k > 0 \) such that \( \frac{\mathbb{E}|M_\infty|}{k} < \delta, \) then

\[
\mathbb{E}|M_n| \mathbb{1}_{|M_n| \geq k} \leq \mathbb{E}|M_\infty| \mathbb{1}_{|M_\infty| \geq k} \leq \varepsilon.
\]

Thus \( (M_n) \) is u.i.. \( \blacksquare \)
14.1 Reverse Filtration and Reverse Martingale

**Definition 14.1 (Reverse Filtration).** We define a reverse filtration as a sequence of decreasing σ-fields \( \mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \mathcal{G}_2 \supseteq \cdots \). For consistency, we will use negative index with \( \mathcal{F}_n = \mathcal{G}_{-n}, n \geq 0 \) so that \( \cdots \subseteq \mathcal{F}_{-n} \subseteq \cdots \mathcal{F}_{-2} \subseteq \mathcal{F}_{-1} \subseteq \mathcal{F}_0 \) is an increasing sequence of σ-fields.

**Definition 14.2 (Reverse Martingale).** \((M_{-n})_{n \geq 0}\) is a reverse martingale w.r.t. the reverse filtration \((\mathcal{F}_{-n})_{n \geq 0}\) if

(i) \( M_{-n} \in L^1(\mathcal{F}_{-n}), n \geq 0 \),

(ii) \( E(M_{-n} | \mathcal{F}_{-n-1}) = M_{-n-1}, n \geq 0 \)

**Remark:** Clearly \( M_{-n} = E(M_0 | \mathcal{F}_{-n}) \). Also the definition of sub/super reverse martingale extends in the same way as it does for martingales.

**Theorem 14.3.** Let \((M_{-n}, \mathcal{F}_{-n})_{n \geq 0}\) be a reverse martingale. Then

\[ M_{-n} \overset{a.s.}{\rightarrow} L^1 M_{-\infty}. \]

**Proof.** (i) \((M_{-n})_{n \geq 0}\) is u.i.

(ii) Let \( U_n(a, b) = \) number of up-crossings of \((M_{-n}, M_{-n+1}, \cdots, M_0)\) of the interval \((a, b)\). By Doob’s upcrossing inequality we have \( E(U_n(a, b) \leq \frac{E|M_0|^2 + a}{b-a} \). Taking \( n \to \infty \), \( U_n(a, b) < \infty \) a.e. \( \forall a < b \). From here we conclude that \( M_{-n} \) converges a.s. to some random variable \( M_{-\infty} \).

(iii) \( M_{-n} \overset{a.s.}{\rightarrow} M_{-\infty} \) and is u.i. This implies \( M_{-n} \overset{L^1}{\rightarrow} M_{-\infty} \).

This completes the proof.

**Corollary 14.4.** If \( M_0 \in L^p \) then \( M_{-n} \overset{L^p}{\rightarrow} M_{-\infty} \).

**Corollary 14.5.** \( M_{-\infty} = E(M_0 | \mathcal{F}_{-\infty}) \) where \( \mathcal{F}_{-\infty} = \cap_{n \geq 0} \mathcal{F}_{-n} \)

### 14.1.1 SLLN (using reverse martingale)

**Theorem 14.6.** Let \( X_1, X_2, \cdots \) be i.i.d. with mean \( \mu \). Then

\[ \frac{S_n}{n} \overset{a.s.}{\rightarrow} \mu \ \text{where} \ S_n = X_1 + X_2 + \cdots + X_n. \]
Proof. Observe that $\mathcal{F}_n = \sigma(S_n, S_{n+1}, \ldots) = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots), n \geq 1$ is a reverse filtration.

Claim: $\mathbb{E}(X_1 \mid \mathcal{F}_n) = \frac{S_n}{n}$.

From symmetry, we have $\mathbb{E}(X_1 \mid \mathcal{F}_n) = \mathbb{E}(X_i \mid \mathcal{F}_n)$ for $i = 1, 2, \ldots, n$. To this we need to show that for any $A \in \mathcal{F}_n$,

$$E(X_1 1_A) = E(X_i 1_A).$$

Observe that $X_1$ and $X_i$ are exchangeable conditioned on filtration $\mathcal{F}_n$, i.e.,

$$(X_1, S_n, X_{n+1}, \ldots) \overset{d}{=} (X_i, S_n, X_{n+1}, \ldots).$$

Thus, $\mathbb{E}(X_1 \mid \mathcal{F}_n) = \frac{1}{n} \mathbb{E}(\sum_{i=1}^{n} X_i \mid \mathcal{F}_n) = \frac{S_n}{n}$. Thus $\frac{S_n}{n} \overset{a.s.}{\rightarrow} \overset{L^1}{\rightarrow} M_{-\infty}$ for some $M_{-\infty}$. $M_{-\infty}$ is a constant since it is tail $\sigma$-field measurable ($\{\lim \frac{S_n}{n} \in A\}$ is a tail event for all $A$). Since $L^1$ convergence preserves mean, $M_{-\infty} = \mu$. \hfill \qed

### 14.2 Exchangeability and Hewitt-Savage 0-1 Law

**Definition 14.7.** Let $X_1, X_2, \ldots$ be a sequence of r.v.’s. Define,

$$\mathcal{E}_n = \sigma(f(X_1, X_2, \ldots, X_n, \ldots, X_k), k \geq n),$$

where $f$ is symmetric under co-ordinate permutation in the first $n$-coordinates. The $\sigma$-field

$$\mathcal{E}_\infty = \cap_{n \geq 1} \mathcal{E}_n$$

is called the exchangeable $\sigma$-field.

It can be shown that $\mathcal{E}_n = \sigma(X_1^{(n)}, \ldots, X_n^{(n)}, X_{n+1}, \ldots)$, where $X_1^{(n)} \leq X_2^{(n)} \leq \cdots \leq X_n^{(n)}$ are the order statistic of $X_1, \ldots, X_n$.

**Lemma 14.8 (Hewitt-Savage 0-1 Law).** For a sequence of i.i.d. r.v.’s and $A \in \mathcal{E}_\infty$, $\mathbb{P}(A) = 0$ or 1.

**Proof.** We will show that $\mathcal{E}_\infty \perp \sigma(X_1, X_2, \ldots, X_n)$ which will imply that $A \perp A$ and thus $\mathbb{P}(A) = 0$ or 1.

**Claim:** If $\mathbb{E}(1_A \mid \mathcal{E}_\infty) = \mathbb{P}(A)$ then $A \perp \mathcal{E}_\infty$.

For $B \in \mathcal{E}_\infty$, $\mathbb{P}(AB) = \mathbb{E}(1_A 1_B) = \mathbb{E}(\mathbb{E}(1_A \mid \mathcal{E}_\infty) 1_B) = \mathbb{E}(\mathbb{P}(A) 1_B) = \mathbb{P}(A) \mathbb{P}(B)$. This implies $A \perp B$.

Note that by reverse martingale, $\mathbb{E}(1_A \mid \mathcal{E}_n) \overset{a.s.}{\rightarrow} \overset{L^1}{\rightarrow} E(1_A \mid \mathcal{E}_\infty)$.

**Claim:** For $Y = \phi(X_1, \ldots, X_k)$,

$$E(Y \mid \mathcal{E}_n) = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \phi(X_{i_1}, \ldots, X_{i_k}).$$

**Proof:** $E(\phi(X_1, \ldots, X_k) \mid \mathcal{E}_n) = E(\phi(X_{i_1}, \ldots, X_{i_k}) \mid \mathcal{E}_n)$, for $1 \leq i_1 < i_2 < \cdots < i_k \leq n$. 


Claim: Take \( Y = \phi(X_1, \ldots, X_k) \), then \( \mathbb{E}(Y \mid \mathcal{E}_\infty) \perp X_1 \).

Proof: \( \mathbb{E}(Y \mid \mathcal{E}_n) = \frac{1}{k} \) (terms that depend on \( X_1 \)) + \( \frac{1}{k} \) (functions of \( X_2, \ldots, X_n \)). Now number of terms that depends on \( X_1 \) is \( \leq n^{k-1} \). Moreover, we have \( \mathbb{E}(Y \mid \mathcal{E}_n) \overset{a.s.}{\rightarrow} \mathbb{E}(Y \mid \mathcal{E}_\infty) \) and \( |\frac{1}{k} \) (terms that depend on \( X_1 \))| \( \leq \frac{n^{k-1}}{k} \rightarrow 0 \). Thus \( \mathbb{E}(Y \mid \mathcal{E}_\infty) \in \sigma(X_2, X_3, \ldots) \). Same argument implies, for any \( X_\ell, \ell \geq 1 \), we have \( \mathbb{E}(Y \mid \mathcal{E}_\infty) \perp \sigma(X_1, \ldots, X_\ell) \), and thus \( \mathbb{E}(Y \mid \mathcal{E}_\infty) \perp \sigma(X_1, X_2, \ldots) \), which implies

\[
\mathbb{E}(Y \mid \mathcal{E}_\infty) \perp \mathbb{E}(Y \mid \mathcal{E}_\infty), \quad \text{and} \quad \mathbb{E}(Y \mid \mathcal{E}_\infty) = Y.
\]

**Definition 14.9.** We say that \((X_1, X_2, \cdots)\) a sequence of random variables is exchangeable if

\[
(X_1, \ldots, X_n) \overset{d}{=} (X_{\pi_1}, \ldots, X_{\pi_n}),
\]

where \( \pi \in \mathfrak{S}_n = \{ \tau : \{1, \ldots, n\} \to \{1, \ldots, n\} \text{ is a bijection} \} \).

**Theorem 14.10** (de Finetti’s Theorem). Let \( X_1, X_2, \ldots \) be an exchangeable sequence of random variables. Then

\[
\mathbb{P}(X_i \in A_i, i = 1, \ldots, n \mid \mathcal{E}_\infty) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i \mid \mathcal{E}_\infty),
\]

where \( \mathcal{E}_\infty = \cap_{n \geq 1} \sigma(X_1^{(n)}, \ldots, X_n^{(n)}, X_{n+1}, \ldots) \) is the exchangeable \( \sigma \)-field.

**Corollary 14.11.** Let \((X_1, \cdots, X_n, \cdots)\) be exchangeable and \( X_i \in \{0, 1\} \ \forall i \), then there exists a r.v. \( \Theta \in [0, 1] \) which is \( \sim \mathcal{E}_\infty \) measurable, s.t.

\[
\mathbb{P}(X_1 = i_1, \ldots, X_k = i_k \mid \mathcal{E}_\infty) = \Theta^{\#(i = 1)}(1 - \Theta)^{n - \#(i = 1)},
\]

where \( \Theta = \mathbb{P}(X_1 = 1 \mid \mathcal{E}_\infty) \).

### 14.3 Optional Stopping Theorem

**Lemma 14.12.** If \((\mathcal{F}_n)_{n \geq 0}\) is a filtration and \( N \) is a stopping time, then

\[
\mathcal{F}_N = \{ A \in \mathcal{F} \mid A \cap \{ N = i \} \in \mathcal{F}_i \ \text{for all} \ i \geq 0 \}
\]

is a \( \sigma \)-field.

**Remark 14.13.** If \( N = k \) a.s., then \( \mathcal{F}_N = \mathcal{F}_k \). Moreover, \( N \leq M \) implies \( \mathcal{F}_N \subseteq \mathcal{F}_M \).

**Lemma 14.14.** Let \( N \) be a stopping time with respect to \((\mathcal{F}_n)\) such that \( 0 \leq T \leq N \leq l \) a.s. Let \((X_n, \mathcal{F}_n)_{n \geq 1}\) be a sub-martingale. Then,

(i) \( X_N \leq \mathbb{E}(X_T \mid \mathcal{F}_N) \).

(ii) \( \mathbb{E}(X_T) \leq \mathbb{E}(X_N) \leq \mathbb{E}(X_I) \).
Proof. Take $A \in \mathcal{F}_N$. We want to prove that $\mathbb{E}(X_i \mathbb{1}_A) \geq \mathbb{E}(X_N \mathbb{1}_A)$. Since $N \leq i$, it is enough to show that $\mathbb{E}(X_i \mathbb{1}_{A \cap \{N=i\}}) \geq \mathbb{E}(X_i \mathbb{1}_{A \cap \{N=i\}})$ for $i = 0, 1, \ldots, l$. Now, result (i) follows by the fact that $A \cap \{N = i\} \in \mathcal{F}_i$ for all $i$.

Taking expectation in (i), we get $\mathbb{E}(X_N) \leq \mathbb{E}(X_l)$. For the other inequality, note that $X_{N \wedge n} - X_{T \wedge n} = \sum_{i=1}^n \Delta X_i \mathbb{1}_{T < i \leq N}$. Now, $X_n$ is a sub-martingale and $\mathbb{1}_{T < i \leq N} = \mathbb{1}_{T < i-1 \leq N}, i \geq 1$ is predictable, hence $X_{N \wedge n} - X_{T \wedge n}$ is a sub-martingale, which implies that $\mathbb{E}(X_{N \wedge n} - X_{T \wedge n}) \geq 0$ and take $n = l$ to get the other inequality in (ii).

If we assume uniform integrability, so that $X_\infty = \lim_{n \to \infty} X_n$ exists in a.s. and $L^1$ sense, we can prove more.

**Theorem 14.15.** If $(X_n, \mathcal{F}_n)_{n \geq 0}$ is a uniformly integrable submartingale, then for any stopping time $N$, $(X_{N \wedge n})_{n \geq 0}$ is uniformly integrable.

**Proof.** The proof involves showing that $\mathbb{E}|X_N| < \infty$, since $|X_{N \wedge n}| = |X_N| \mathbb{1}_{\{N \leq n\}} + |X_n| \mathbb{1}_{\{N > n\}}$ combined with u.i. of $X_n$ proves the result. We note that $\mathbb{E}X^+_{N \wedge n} \leq \mathbb{E}X^+_n$ for $n \geq 0$ to complete the proof.

**Theorem 14.16 (Optional Stopping Theorem).** If $L \leq M$ are stopping times w.r.t. the filtration $(\mathcal{F}_n)_{n \geq 0}$ and $(X_{M \wedge n}, \mathcal{F}_n)_{n \geq 0}$ is a uniformly integrable submartingale, then

$$X_L \leq \mathbb{E}(X_M | \mathcal{F}_L) \text{ and } \mathbb{E}X_L \leq \mathbb{E}X_M.$$  

**Proof.** Enough to show that $\mathbb{E}((X_M - X_L) \mathbb{1}_A) \geq 0$ for all $A \in \mathcal{F}_L \subseteq \mathcal{F}_M$. Fix $A \in \mathcal{F}_L$ and define $N = M \mathbb{1}_{A^c} + L \mathbb{1}_A$.

We claim that $N$ is a stopping time and $\mathbb{E}(X_M - X_N) \geq 0$. The first claim follows since $\{N = n\} = (A \cap \{L = n\}) \cup (\{M = n\} \cap A^c) \in \mathcal{F}_n$.

The second claim follows since $(X_{M \wedge n})_{n \geq 0}$ is a uniformly integrable submartingale and $\mathbb{E}(X_{M \wedge n} - X_{N \wedge n}) \geq 0$ for all $n \geq 0$. Now note that $X_M - X_N = (X_M - X_L) \mathbb{1}_A$ and we have the proof.

**Corollary 14.17 (Generalization of Wald’s First Identity).** Suppose $(X_n)_{n \geq 0}$ is a submartingale and $\sup_{n \geq 0} \mathbb{E}(|X_{n+1} - X_n| | \mathcal{F}_n) \leq B < \infty$ a.s. If $N$ is a stopping time with $\mathbb{E}N < \infty$, then $(X_{N \wedge n}, n \geq 0)$ is uniformly integrable and hence $\mathbb{E}X_N \geq \mathbb{E}X_0$. Moreover, if $(X_n)_{n \geq 0}$ is also a martingale, then $\mathbb{E}X_N = \mathbb{E}X_0$.

### 14.4 Azuma-Hoeffding Inequality

**Theorem 14.18 (Azuma-Hoeffding Inequality).** Let $(M_n, \mathcal{F}_n)_{n \geq 1}$ be a super-martingale with the Martingale Difference Sequence $\Delta_n := M_n - M_{n-1}$ satisfying

$$|\Delta_n| \leq c_n \text{ for all } n \geq 1.$$  

Then for all $t > 0$ we have

$$\mathbb{P}(M_n - M_0 \geq t) \leq e^{-t^2/2s_n^2}.$$
where
\[ s_n^2 := \sum_{i=1}^{n} c_i^2, \quad n \geq 1. \]

In particular, if \((M_n, \mathcal{F}_n)_{n \geq 1}\) is a martingale, we have
\[
P(|M_n - M_0| \geq t) \leq 2e^{-t^2/2s_n^2}, \quad t > 0.
\]

Proof. We will use the following result: For \(0 \leq |x| \leq c\), we have
\[
e^x \leq \frac{1}{2}(e^c + e^{-c}) + \frac{1}{2c}(e^c - e^{-c}) \cdot x \leq e^{c^2/2} + \frac{\sinh(c)}{c} \cdot x.
\]
Note that \(E(\Delta_n \mid \mathcal{F}_{n-1}) \leq 0\) for all \(n \geq 1\). Thus for any \(\theta > 0, n \geq 1\), we have
\[
E(e^{\theta \Delta_n} \mid \mathcal{F}_{n-1}) \leq e^{\theta^2 c_n^2/2}.
\]
Using induction, we have
\[
E(e^{\theta(M_n-M_0)} \leq \prod_{i=1}^{n} e^{\theta^2 c_i^2/2} = e^{\theta^2 s_n^2/2}.
\]
In particular, using exponential Markov’s inequality we have
\[
P(M_n - M_0 \geq t) \leq \inf_{\theta > 0} e^{-\theta t} E(e^{\theta(M_n-M_0)}) \leq \inf_{\theta > 0} e^{-\theta t + \theta^2 s_n^2/2} = e^{-t^2/2s_n^2}.
\]
15.1 Weak convergence and the classical CLT

One of the most celebrated results of probability theory says that if $S_n = \xi_1 + \xi_2 + \ldots + \xi_n$ is a sum of i.i.d. random variables with mean 0 and variance 1, then for large $n$ the distribution of $n^{-1/2}S_n$ can roughly be approximated by a standard Gaussian distribution on the real line. In other words, as $n \to \infty$, the distribution of $n^{-1/2}S_n$ approaches to standard normal distribution with density $(2\pi)^{-1/2}e^{-x^2/2}$ for $x \in \mathbb{R}$.

We have proved this result using Characteristic function approach and Lindeberg technique. However, the proof breaks down for sums of “weakly dependent” random variables. Martingale central limit theorem gives one example where CLT holds for dependent sums. Stein’s method, introduced by Charles Stein (1972) and developed by later researchers, is a way of proving CLT for dependent sums that also gives explicit rates of convergence.

We recall that,

**Definition 15.1.** For $P_n, P$ be two probability measures on $\mathbb{R}$, we say that $P_n$ converges weakly to $P$, which we write by $P_n \Rightarrow P$, if

$$\int f dP_n \to \int f dP$$

for all continuous bounded functions $f : \mathbb{R} \to \mathbb{R}$.

For a random variable $X : (\Omega, \mathcal{F}, \mu) \to \mathbb{R}$, by $\mathcal{L}(X)$, we mean the probability measure induced by $X$ on $\mathbb{R}$, i.e., $\mu \circ X^{-1}$. We say that $X \Rightarrow Y$ if $\mathcal{L}(X) \Rightarrow \mathcal{L}(Y)$.

15.2 Distances between probability measures

The space of probability is metrizable under the weak convergence. One such metric is known as Prohorov metric which is given below.

$$d(P, Q) = \inf\{\varepsilon > 0 : P(A) \leq Q(A^\varepsilon) + \varepsilon, Q(A) \leq P(A^\varepsilon) + \varepsilon \forall A \in \mathcal{B}(\mathbb{R})\},$$

where $A^\varepsilon = \{x : d(x, A) < \varepsilon\}$.

As it turns out the above metric is not very useful for calculations. There are several other distances between two probability measures which we can work with and are stronger than the metric of weak convergence. A typical distance between probability measures can be of the following type

$$d(P, Q) = \sup \left\{\left|\int f dP - \int f dQ\right| : f \in \mathcal{D}\right\},$$

15-1
where $\mathcal{D}$ is some class of $\mathbb{R}$-valued functions. Note that $d(\cdot, \cdot)$ defined above is always a proper metric as long as $\mathcal{D}$ is a separating class, i.e.

$$\int f dP = \int f dQ \ \forall f \in \mathcal{D} \implies P = Q.$$  

15.2.1 Total variation distance

Let us take $\mathcal{D}$ to be the class of indicator functions of Borel sets of $\mathbb{R}$ in $(??)$. The resulting metric is known as the total variation distance. Thus the total variation distance between two probability measures $P$ and $Q$ on $S$ is given by

$$d_{TV}(P, Q) = \sup_{A \in \mathcal{B}(\mathbb{R})} |P(A) - Q(A)| = \sup_{h: \|h\|_{\infty} \leq 1} \frac{1}{2} \left| \int h dP - \int h dQ \right|.$$  

Note that this ranges in $[0, 1]$.

15.2.2 Wasserstein distance

This is also known as the Kantorovich-Monge-Rubinstein metric. Here $\mathcal{D} =$ set of all 1-Lipschitz functions on $\mathbb{R}$.

$$d_{W}(P, Q) := \sup \left\{ \left| \int f dP - \int f dQ \right| : f \text{ is 1-Lipschitz} \right\}.$$  

The Wasserstein distance can range in $[0, \infty]$.

15.2.3 Kolmogorov-Smirnov distance

It is only defined for probability measures on $\mathbb{R}$.

$$d_{K}(P, Q) := \sup_{x \in \mathbb{R}} |P((-\infty, x]) - Q((-\infty, x])|.$$  

(15.1)  

It also falls under the same general framework of (??) when $\mathcal{D} = \{1_{(-\infty, x]} : x \in \mathbb{R}\}$.

Given any metric $d$ on the set of all probability measures on $\mathbb{R}$ and random variables $X, Y$, we will abuse notation by writing $d(X, Y)$ for $d(\mathcal{L}(X), \mathcal{L}(Y))$.

Remark 15.2.  

- All the distances defined above are stronger than weak convergence. That is, if any of these distances go to zero as $n \to \infty$, then we have weak convergence. But converse is not true.

- Total variation is a very strong notion, often too strong to be useful. Suppose $X_1, X_2, \ldots$ i.i.d. $\pm 1$ with probability $1/2$ each. $S_n = \sum_1^n X_i$. Then

$$\frac{S_n}{\sqrt{n}} \implies N(0, 1).$$

But $d_{TV}(\frac{S_n}{\sqrt{n}}, Z) = 1$ for all $n$. But both Wasserstein and Kolmogorov distances go to 0 at rate $n^{-1/2}$.  

15.3 Stein’s method for normal approximation

Stein’s method was introduced by Charles Stein in the early 70’s to prove central limit theorem for dependent random variables and more importantly to find explicit estimates of the accuracy of the approximation. Instead of using characteristic functions he used, what is now called Stein’s Characterizing operator, to convert the ‘global problem’ of analyzing the whole random variable into more tractable ‘local problems’.

Suppose we want to show that the random variable $X$ has a distribution which is approximately equal to that of another random variable $Z$. As noted in the previous section, the distributional proximity of two random variables is measured by $\sup_{g \in \mathcal{D}} |Eg(X) - Eg(Z)|$ for a large family of test functions $\mathcal{D}$. Various choices of family $\mathcal{D}$ lead to different notions of distances between two probability measures. Given $\mathcal{D}$ and the target distribution $Z$, the main goal of Stein’s method is to compute upper bounds of the complicated looking object $\sup_{g \in \mathcal{D}} E|g(X) - g(Z)|$. The basic theme behind the Stein’s method will be as follows

1. Identify the “Stein Characterizing operator” for $Z$, that is, find a suitable operator $\mathcal{A}$ such that for any random variable $W$, $E\mathcal{A}f(W) = 0$ for all functions $f$ belonging to a suitably large class $\mathcal{D}$ if and only if $W \overset{d}{=} Z$. For example, if $Z$ has a standard normal distribution, then
   \[ (\mathcal{A}f)(x) = f'(x) - xf(x) \quad \text{for } f \in \mathcal{D} \]
   where $\mathcal{D}$ = set of all locally absolutely continuous functions, is a Stein’s characterizing operator.

2. Invert the Stein’s operator, that is, for a given function $g \in \mathcal{D}$, find $f$ such that $\mathcal{A}f(x) = g(x) - Eg(Z)$. Establish the smoothness properties of $f$ in terms of $g$. It is often the case that $\mathcal{D}'$ has more smoothness properties than $\mathcal{D}$. Let $\mathcal{D}'$ be a class of (‘smooth’) functions so that for each $g \in \mathcal{D}$, there exists $f \in \mathcal{D}'$. For instance, $\mathcal{D} = \{ g : g \ 1\text{-Lipschitz} \}$ will give rise to Wasserstein metric. We will show that if $Z$ is standard normal, we may take $\mathcal{D}' = \{ f : |f|_\infty \leq 1, |f'|_\infty \leq \sqrt{\frac{2}{\pi}} \text{ and } |f''|_\infty \leq 2 \}$. By that, we mean that given $g \in \mathcal{D}$ with $Eg(Z) = 0$, we can get $f \in \mathcal{D}'$ so that $\mathcal{A}f = g$.

3. $|Eg(X) - Eg(Z)| \leq |E\mathcal{A}f(X)|$. So, $\sup_{g \in \mathcal{D}} |Eg(X) - Eg(Z)| \leq \sup_{f \in \mathcal{D}'} |E\mathcal{A}f(X)|$. Since $\mathcal{A}$ is a characterizing operator, $E\mathcal{A}f(Z) = 0$. But if distribution of $X$ is close to that of $Z$, we should hope that $E\mathcal{A}f(X) \approx 0$. As it turns out in practice that bounding $\sup_{f \in \mathcal{D}'} |E\mathcal{A}f(X)|$ is easier than bounding $\sup_{g \in \mathcal{D}} |Eg(X) - Eg(Z)|$ itself. In normal example, we have an upper bound in Wasserstein metric as follows
   \[ d_W(X, Z) \leq \sup_{f \in \mathcal{D}'} |E(f'(X) - Xf(X))|. \]

Let us now analyze the case of normal distribution. The following lemma is due to Charles Stein.

Lemma 15.3 (Stein’s Lemma). A random variable $W$ has a standard normal distribution iff for every piecewise continuously differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$ with $E|f'(Z)| < \infty$, $Z \sim N(0, 1)$ we have $EWf(W) = Ef'(W)$.

Remark 15.4. If we define the operator $T : \mathcal{F} \rightarrow \mathcal{C}$ where $\mathcal{F}$ is the set of all piecewise continuously differentiable function $f$ with $E|f'(Z)| < \infty$, $Z \sim N(0, 1)$ and $\mathcal{C}$ is the set of all continuous functions, as $Tf(x) = f'(x) - xf(x)$, then the above lemma says that $W \sim N(0, 1)$ iff $ETf(W) = 0$ for all $f \in \mathcal{F}$. We will call $T$ the characterizing operator for standard normal distribution.
Proof. Only if part: First let us assume that \( f \) has compact support contained in \((a,b), -\infty < a < b < \infty\). Then the result follows from integration by parts

\[
\int_a^b x f(x) e^{-x^2/2} dx = \left[ f(x) e^{-x^2/2} \right]_a^b + \int_a^b f'(x) e^{-x^2/2} dx.
\]

Now take any \( f \) s.t. \( \mathbb{E}|Z f(Z)| < \infty, \mathbb{E}|f'(Z)| < \infty, \mathbb{E}|f(Z)| < \infty \). Take a piecewise linear function \( g \) that takes value 1 in \([-1,1]\), 0 outside \([-2,2]\), and between 0 and 1 elsewhere. Let

\[
\mathbf{f}_n(x) := f(x)g(x/n).
\]

Then clearly,

\[
|\mathbf{f}_n(x)| \leq |f(x)| \quad \text{for all } x \text{ and } \mathbf{f}_n(x) \to f(x) \text{ pointwise}.
\]

Similarly, \( f'_n \to f' \) pointwise. Note that the support of \( f_n \) is contained in \([-n,n]\). Now the rest follows by DCT. The last step is to show that the finiteness of \( \mathbb{E}|f'(Z)| \) implies the finiteness of the other two expectations.

Suppose \( \mathbb{E}|f'(Z)| < \infty \). Then

\[
\int_0^\infty |x f(x)| e^{-x^2/2} dx \leq \int_0^\infty x \int_0^x |f'(y)| dy e^{-x^2/2} dx
\]

\[
= \int_0^\infty |f'(y)| \int_y^\infty xe^{-x^2/2} dx dy = \int_0^\infty |f'(y)| e^{-y^2/2} dy < \infty
\]

and finiteness of \( \mathbb{E}|f(Z)| \) follows from the inequality

\[
|f(x)| \leq \sup_{|t| \leq 1} |f(t)| + |x f(x)|.
\]

If part: Fix any real number \( t \in \mathbb{R} \). It is enough to prove that \( \mathbb{P}(W \leq t) = \phi(t) \) where \( \phi(t) = \int_{-\infty}^t (2\pi)^{-1/2} \exp(-x^2/2) dx \) is the standard normal cdf. Let \( f \) be the particular solution of the differential equation

\[
f'(x) - xf(x) = 1 \{ x \leq t \} - \phi(t)
\]

given by

\[
f(x) = \sqrt{2\pi} e^{x^2/2} (\phi(x \wedge t) - \phi(x) \phi(t)).
\]

\( f \) is piecewise continuously differentiable and is bounded using the fact that \( e^{x^2/2} \phi(-|x|) \to 0 \) as \( x \to \pm \infty \). Hence we have \( \mathbb{E}|f'(Z)| < \infty, Z \sim N(0,1) \). Now from the hypothesis we have \( \mathbb{E}(f'(W) - W f(W)) = 0 \) and this implies that \( \mathbb{P}(W \leq t) = \phi(t) \).

\[
15.4 \quad \text{Inversion of the Stein operator}
\]

In this section we will study the solution of the differential equation

\[
f'(x) - xf(x) = g(x) - \mathbb{E} g(Z), \quad Z \sim N(0,1).
\]
Lemma 15.5. Given function $g : \mathbb{R} \to \mathbb{R}$ such that $E |g(Z)| < \infty$ where $Z \sim N(0,1)$,

$$f(x) = e^{x^2/2} \int_{-\infty}^{x} e^{-y^2/2}(g(y) - E g(Z))dy$$  \hspace{1cm} (15.3)$$
is an absolutely continuous solution of (15.2). Moreover, any a.c. solution $\tilde{f}$ of (15.2) is of the form

$$\tilde{f}(x) = f(x) + ce^{x^2/2}, \quad c \in \mathbb{R}.$$  

Finally, $f$ is the only solution that satisfies $\lim_{|x| \to \infty} f(x)e^{-x^2/2} = 0$.

Proof. By the method of integrating factors, we have that if $f$ is a solution to (15.2), then

$$\frac{d}{dx} (e^{-x^2/2} f(x)) = e^{-x^2/2} (f'(x) - xf(x)) = e^{-x^2/2} (g(x) - E g(Z)).$$

So, (15.3) is a reasonable candidate as a solution of (15.2). And it is easy to verify directly that (15.3) indeed satisfies (15.2). If $\tilde{f}$ is any other solution of (15.2), then

$$\frac{d}{dx} \left( e^{-x^2/2} (f(x) - \tilde{f}(x)) \right) = 0.$$

Hence, $\tilde{f}(x) = f(x) + ce^{x^2/2}$ for some $c \in \mathbb{R}$. Clearly, from definition

$$\lim_{x \to -\infty} f(x)e^{-x^2/2} = 0 \quad \text{(by DCT)}.$$

Note that since $Z \sim N(0,1)$, we have

$$\int_{-\infty}^{\infty} e^{-y^2/2}(g(y) - E g(Z))dy = 0.$$

So, $f$ can also be written as follows

$$f(x) = -e^{x^2/2} \int_{x}^{\infty} e^{-y^2/2}(g(y) - E g(Z))dy.$$  \hspace{1cm} (15.4)$$

Therefore, by DCT, $\lim_{x \to +\infty} f(x)e^{-x^2/2} = 0$.

When the function $g$ is Lipschitz we can write the solution in a different way, which will help us to find optimal bounds.

Lemma 15.6. Assume $g$ is Lipschitz. Then

$$f(x) = -\int_{0}^{1} \frac{1}{2\sqrt{t}(1-t)} E \left[ Zg(\sqrt{t}x + \sqrt{1-t}Z) \right] dt, \quad Z \sim N(0,1)$$  \hspace{1cm} (15.5)$$
is a solution of (15.2). In fact, it is the same as (15.3).
Proof. Let $g$ is $C$-Lipschitz. Then $|g'|_{\infty} \leq C$. On differentiating $f$ and carrying the derivative inside the integral and expectation which can be justified using DCT, we have

$$f'(x) = -\int_0^1 \frac{1}{2\sqrt{1-t}} \mathbb{E} \left[ g'(\sqrt{1-t}x + \sqrt{1-t}Z) \right] dt.$$  \hfill (15.6)

On the other hand, the Stein identity gives us

$$\mathbb{E} \left[ Zg(\sqrt{1-t}x + \sqrt{1-t}Z) \right] = \sqrt{1-t} \mathbb{E} \left[ g'(\sqrt{1-t}x + \sqrt{1-t}Z) \right].$$

Thus,

$$f'(x) - xf(x) = \int_0^1 \mathbb{E} \left[ \left(-\frac{Z}{2\sqrt{1-t}} + \frac{x}{2\sqrt{t}} \right) g'(\sqrt{1-t}x + \sqrt{1-t}Z) \right] dt$$

$$= \int_0^1 \mathbb{E} \left[ \frac{d}{dt} g'(\sqrt{1-t}x + \sqrt{1-t}Z) \right] dt$$

$$= \mathbb{E} \left[ \int_0^1 \frac{d}{dt} g'(\sqrt{1-t}x + \sqrt{1-t}Z) dt \right] = g(x) - \mathbb{E} g(Z).$$

Now that we know the form of the solution to (15.2) the next obvious question is what can we say about the solution when $g$ has some nice properties like boundedness or differentiability. Our next theorem says that indeed the solution $f$ inherits the ‘niceness’ of the function $g$. Recall the notation that $Ng := \mathbb{E} g(Z)$.

**Lemma 15.7.** If $g : \mathbb{R} \to \mathbb{R}$ is bounded,

I. $|f_{\infty}| \leq \sqrt{\frac{\pi}{2}} |g - Ng|_{\infty}$ and

II. $|f'|_{\infty} \leq 2 |g - Ng|_{\infty}$.

and if $g$ is Lipschitz, but not necessarily bounded, then

III. $|f_{\infty}| \leq |g'|_{\infty}$,

IV. $|f'|_{\infty} \leq \sqrt{\frac{2}{\pi}} |g'|_{\infty}$ and

V. $|f''|_{\infty} \leq 2 |g'|_{\infty}$.

Here for a function $f$ we define $|f|_{\infty} := \sup\{|f(x)| : x \in \mathbb{R}\}$. Moreover, all the above bounds are tight. The bounds (I), (II) and (V) were obtained by Stein.

### 15.4.1 Example: Ordinary CLT in the Wasserstein metric

Suppose $X_1, X_2, \ldots, X_n$ are independent, mean 0, variance 1, $\mathbb{E}|X_i|^3 < \infty$. Let $S_n = \sum_1^n X_i$ and $W = n^{-1/2}S_n$. Take any $f \in C^1$ with $f'$ absolutely continuous, and satisfying $|f| \leq 1$, $|f'| \leq \sqrt{2/\pi}$ and $|f''| \leq 2$. First, note that

$$\mathbb{E} W f(W) = n^{-1/2} \sum \mathbb{E} (X_i f(W)).$$

---

1 Any Lipschitz function $g$ is absolutely continuous. Hence, it is (Lebesgue) almost surely differentiable. Define $g'$ to be derivative of $g$ at the points where it exists and 0 elsewhere.
Now define
\[ W_i = W - n^{-1/2} X_i = n^{-1/2} \sum_{j \neq i} X_j, \quad i = 1, 2, \ldots, n. \]
Observe that \( X_i, W_i \) are independent. Thus
\[ \mathbb{E} X_i f(W_i) = \mathbb{E} X_i \mathbb{E} f(W_i) = 0 \]
and so, we can write
\[ \mathbb{E} (X_i f(W)) = \mathbb{E} (X_i (f(W) - f(W_i))) = \mathbb{E} (X_i (f(W) - f(W_i) - (W - W_i)f'(W_i))) + \mathbb{E} [X_i(W - W_i)f'(W_i)]. \]
Note that
\[ |f(b) - f(a) - (b - a)f'(a)| \leq \frac{1}{2} (b - a)^2 |f''|_{\infty} \]
and that \( W - W_i = n^{-1/2} X_i \). Thus
\[ \left| \mathbb{E} \left[ X_i \left( f(W) - f(W_i) - n^{-1/2} X_i f'(W_i) \right) \right] \right| \leq \frac{1}{2} \mathbb{E} \left| X_i (n^{-1/2} X_i)^2 \right| \cdot |f''|_{\infty} \leq n^{-1} \mathbb{E} |X_i|^3. \]
Again,
\[ \mathbb{E} [X_i(W - W_i)f'(W_i)] = n^{-1/2} \mathbb{E} X_i^2 f'(W_i) = n^{-1/2} \mathbb{E} f'(W_i) \]
since \( \mathbb{E} X_i^2 = 1 \) and \( X_i \) is independent of \( W_i \).
From (??) and the above calculation we see that
\[ \left| \mathbb{E} W f(W) - n^{-1} \sum_{i=1}^{n} \mathbb{E} f'(W_i) \right| \leq n^{-3/2} \sum_{i=1}^{n} \mathbb{E} |X_i|^3. \]
Finally, note that
\[ \left| n^{-1} \sum_{i=1}^{n} \mathbb{E} f'(W_i) - \mathbb{E} f'(W) \right| \leq n^{-1} |f''|_{\infty} \sum_{i=1}^{n} \mathbb{E} |W - W_i| \]
\[ = n^{-3/2} |f''|_{\infty} \sum_{i=1}^{n} \mathbb{E} |X_i| \leq 2n^{-3/2} \sum_{i=1}^{n} \mathbb{E} |X_i|. \]
Combining, we have
\[ \left| \mathbb{E} f(W) W - \mathbb{E} f'(W) \right| \leq n^{-3/2} \sum_{i=1}^{n} \mathbb{E} |X_i|^3 + 2n^{-3/2} \sum_{i=1}^{n} \mathbb{E} |X_i|. \]
Since \( \mathbb{E} X_i^2 = 1 \) we can conclude that \( \mathbb{E} |X_i|^3 \geq 1 \) and hence \( \mathbb{E} |X_i| \leq (\mathbb{E} |X_i|^3)^{1/3} \leq \mathbb{E} |X_i|. \)
We have now arrived at a ‘Berry-Essén bound’ for the Wasserstein metric:
**Theorem 15.8.** Suppose \( X_1, \ldots, X_n \) are independent with mean 0, variance 1, and finite third moments. Then
\[ d_W(n^{-1/2} \sum_{i=1}^{n} X_i, Z) \leq \frac{3}{n^{3/2}} \sum_{i=1}^{n} \mathbb{E} |X_i|^3, \]
where \( Z \sim N(0, 1) \).
15.5 Dependency graph approach

There are plenty of examples in the literature that suggest that the CLT for the sums of the random variables goes well beyond the regime of independent random variables. As a natural generalization to i.i.d. random variables, we can think of $m$-dependent sequences for which it is well known that CLT holds as well. In some sense the fact that for a $m$-dependent sequence the variables that are ‘far apart’ are independent makes us believe that the CLT should hold and it is indeed the case. In the section, we try to formalize this vague notion of ‘local dependence structure’ for a general family of random variables and develop procedures to extract CLT out of it, using the example of i.i.d. sum of r.v.s.

Let’s now a define what we mean by a dependency graph.

**Definition 15.9.** Let $\{X_v\}_{v \in V}$ be a family of random variables. A dependency graph for $\{X_v\}$ is any graph $G$ with vertex set $V$ such that if $A$ and $B$ are two disjoint subsets of $V$ so that there is no edge of $G$ between any vertex in $A$ to any vertex in $B$, then the families $\{X_v\}_{v \in A}$ and $\{X_v\}_{v \in B}$ are mutually independent.

For a given family of dependent random variables $\{X_v\}_{v \in V}$, suppose we want to find a CLT for $\sum_{v \in V} X_v$. The very basic idea behind dependency graph is to compactly encode the informations about the dependence between the random variables. If we can find a dependency graph which is sparse, then that ensures that there is a lot of independence in the family $\{X_v\}$ which can be exploited by Stein’s method to prove the CLT. Of course, the choice of the dependency graph is not unique, for the complete graph serves as a dependency graph for all families of random variables. But in practical examples, it often turns out that there is a natural choice of dependency graph which is optimal in a sense that it can not be made any sparser by deleting edges.

Given $u \in V$, let $\overline{N}(u)$ denote the set of all vertices consisting of all the neighbors of $u$ in $G$ and $u$ itself. Let $D := 1 + \max$ degree $= \max_{u \in G} |\overline{N}(u)|$. We have the following theorem.

**Theorem 15.10.** Let $\{X_v\}_{v \in V}$ be a family of random variables with dependency graph $G$. Let $W = \sigma^{-1} \sum_{v \in V} X_v$, where $\sigma^2 = \text{Var}(\sum_{v \in V} X_v)$. Also assume that $E X_v = 0, v \in V$. Then

$$d_W(W, Z) \leq \sigma^{-3} \sum_{v \in V} \sum_{u, w \in \overline{N}(v)} \left( 3E|X_v X_u X_w| + 4E|X_v X_u|E|X_w| \right),$$

(15.7)

where $Z \sim N(0, 1)$.

**Corollary 15.11.** If, in Theorem 15.10, we additionally assume $E|X_v|^3 < \infty$ for all $v \in V$ then

$$d_W(W, Z) \leq \frac{7D^2}{3\sigma^3} \sum_{v \in V} E|X_v|^3.$$  

(15.8)

**Proof of Theorem 15.10.** Let $Z_v = \sigma^{-1} \sum_{u \in \overline{N}(v)} X_u$ and $W_v = \sigma^{-1} \sum_{u \notin \overline{N}(v)} X_u$ so that $W = W_v + Z_v$ for each $v \in V$. We can further decompose $W_v$ as $W_v = W_{vu} + Y_{vu}, v \in V, u \in \overline{N}(v)$, where $W_{vu} = \sigma^{-1} \sum_{w \notin \overline{N}(v) \cup \overline{N}(u)} X_w$ and $Y_{vu} = \sigma^{-1} \sum_{w \in \overline{N}(u) \setminus \overline{N}(v)} X_w$. Note that $W_v$ is independent of $X_v$ and $W_{vu}$ is independent of $(X_v, X_u)$. 


Take any function $f$ with $|f''|_\infty \leq 2$. We can write

$$E\, Wf(W) - E\, f'(W) = \left\{ E\, Wf(W) - \sigma^{-1} \sum_{v \in V} E\, X_v Z_v f'(W_v) \right\}$$

$$+ \left\{ \sigma^{-1} \sum_{v \in V} E\, X_v Z_v f'(W_v) - \sigma^{-2} \sum_{v \in V} \sum_{u \in N(v)} E(X_v X_u) E\, f'(W_{vu}) \right\}$$

$$+ \sigma^{-2} \sum_{v \in V} \sum_{u \in N(v)} \left\{ E(X_v X_u)|E\, f'(W_{vu}) - E\, f'(W)|\right\}$$

where we have used the fact that

$$\sigma^{-2} \sum_{v \in V} \sum_{u \in N(v)} E(X_v X_u) = \sigma^{-1} \sum_{v \in V} E\, X_v Z_v = \sigma^{-1} \sum_{v \in V} E\, X_v W = E\, W^2 = 1.$$  

We are now going to bound each of the three term using Taylor’s expansion. For the first term, we have

$$Wf(W) = \sigma^{-1} \sum_{v \in V} X_v f(W) = \sigma^{-1} \sum_{v \in V} \left\{ X_v f(W_v) + X_v Z_v f'(W_v) + \frac{1}{2} X_v Z_v^2 f''(W_v^*) \right\},$$

for some random variable $W_v^*$ between $W$ and $W_v$. Thus, by using that $E\, X_v f(W_v) = E\, X_v E\, f(W_v) = 0$, we obtain

$$\left| E\, Wf(W) - \sigma^{-1} \sum_{v \in V} E\, X_v Z_v f'(W_v) \right| \leq \sigma^{-1} \sum_{v \in V} E(|X_v| Z_v^2)$$

$$\leq \sigma^{-3} \sum_{v \in V} \sum_{u \in N(v)} E |X_v X_u X_w|$$

Again by Taylor’s expansion,

$$X_v Z_v f'(W_v) = \sigma^{-1} \sum_{u \in N(v)} X_v X_u f'(W_v)$$

$$= \sigma^{-1} \sum_{u \in N(v)} X_v X_u \{ f'(W_{vu}) + Y_{vu} f''(W_{vu}^*) \},$$

where $W_{vu}^*$ is some random variable between $W_v$ and $W_{vu}$. This gives

$$\left| \sigma^{-1} \sum_{v \in V} E\, X_v Z_v f'(W_v) - \sigma^{-2} \sum_{v \in V} \sum_{u \in N(v)} E(X_v X_u) E\, f'(W_{vu}) \right|$$

$$\leq 2\sigma^{-2} \sum_{v \in V} \sum_{u \in N(v)} E |X_v X_u Y_{vu}| = 2\sigma^{-2} \sum_{u \in N(v)} \sum_{v \in V} E |X_v X_u Y_{vu}|$$

$$\leq 2\sigma^{-3} \sum_{v \in V} \sum_{u \in N(v)} \sum_{w \in N(u)} E |X_v X_u X_w|.$$  

Now since $W_{vu} = W_v - Y_{vu} = W - (Z_v + Y_{vu})$, we have

$$f'(W_{vu}) = f'(W) - (Z_v + Y_{vu}) f'(W_{vu}^*)$$
for some $W^*_{vu}$ between $W_{vu}$ and $W$. Therefore,

$$|E f'(W_{vu}) - E f'(W)| \leq 2 E |Z_0 + Y_{vu}| \leq 2 \sigma^{-1} \sum_{w \in N(v) \cup N(u)} E |X_w|.$$  

Putting the ingredients together, we get the result.

**Remark 15.12.** Though the above theorem gives a relatively tight error bound, it is hard to believe that it is optimal as it only uses the fact $|f''|_{\infty}$ is bounded but it never used the boundedness of $|f'|_{\infty}$.

### 15.5.1 Sum of independent random variables

Let’s start with the simplest of an example when the family $\{X_i\}_{1 \leq i \leq n}$ consists of independent random variables with $E X_i = 0$, $\text{Var}(X_i) = 1$ and $E |X_i|^3 < \infty$. Define $W = n^{-1/2} \sum_{i=1}^{n} X_i$. In this case, the graph $G$ with the vertex set $V = \{1, 2, \ldots, n\}$ and with no edges is a straightforward choice for the dependency graph. For this graph $G$, $N(i)$ is simply $\{i\}$ itself and $D = 1$. Also $\sigma^2 = \text{Var}(\sum_{i=1}^{n} X_i) = \sqrt{n}$. To prove CLT, we can now directly apply Corollary 15.11 to obtain the following bound on the Wasserstein distance between $W$ and $Z$,

$$\text{Wass}(W, Z) \leq \frac{7 \max_i E |X_i|^3}{\sqrt{n}}.$$  

### 15.5.2 $m$-dependent sequence

The $m$-dependent sequence is one of the most natural examples one can think of where the method dependency graph technique can be applied.

**Definition 15.13.** A sequence of a random variables $\{X_i : i \geq 0\}$ is called $m$-dependent if $\{X_i : i \leq t\}$ and $\{X_i : i \geq s\}$ are independent whenever $s - t > m$.

**Definition 15.14.** A sequence $\{X_i : i \geq 0\}$ is said to be weakly stationary if

1. $E X_i^2 < \infty \quad \forall i \geq 0$.
2. $E X_i = \mu \quad \forall i \geq 0$ for some $\mu \in \mathbb{R}$.
3. $\text{Cov}(X_{r+t}, X_{s+t}) = \text{Cov}(X_r, X_s) \quad \forall r, s, t \geq 0$.

For a weakly stationary sequence $\{X_i : i \geq 0\}$ let us define the autocovariance function $\gamma$ as

$$\gamma(h) := \text{Cov}(X_h, X_0) \quad \text{for } h \geq 0.$$  

Note that the weak-stationarity implies that $\gamma(h) = \text{Cov}(X_{r+h}, X_r)$ for all $r \geq 0$.

Next we will illustrate an application of Theorem 15.10 to give an easy prove of the central limit theorem for the sum of weakly stationary $m$-dependent process.
**Theorem 15.15.** Fix $m \geq 0$. Let $\{X_i : i \geq 0\}$ be a weakly stationary $m$-dependent sequence. Assume that $\nu_m := \gamma(0) + 2 \sum_{i=1}^m \gamma(i) \neq 0$ and $\sup_{i \geq 0} \mathbb{E}|X_i|^3 < \infty$. Then

$$d_W\left(\frac{\sum_{i=0}^{n-1} X_i - n\mu}{\sqrt{\text{Var}(\sum_{i=0}^{n-1} X_i)}}, Z\right) \leq \frac{7(2m + 1)^2 \cdot n \cdot \max_i \mathbb{E}|X_i|^3}{\sigma^3},$$

where $\mu$ is the common mean of $X_i$ and $Z$, as usual, denote $N(0,1)$ random variable and

$$\sigma^2 = n\nu_m - 2 \sum_{i=1}^m i\gamma(i).$$

**Proof.** Without loss of generality, we can assume $\mu = 0$. From the definition of $m$-dependent sequence, it is almost immediate how to construct a dependency graph for $\{X_i : 0 \leq i \leq n-1\}$. The vertex set is $V = \{0, 1, 2, \ldots, n-1\}$ and vertex $i$ and vertex $j$ is connected by an edge if and only if $1 \leq |i - j| \leq m$. For $i \in V$, we have

$$\mathcal{N}(i) = \{i - m, i - m + 1, \ldots, i - 1, i, i + 1, \ldots, i + m\} \cap V,$$

so that $D = 2m + 1$.

From the weak-stationarity of $\{X_i : i \geq 0\}$

$$\sigma^2 := \text{Var}(\sum_{i=0}^{n-1} X_i) = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \mathbb{E}X_iX_j = \sum_{i=1}^{n-1} \sum_{j \in \mathcal{N}(i)} \mathbb{E}X_iX_j$$

$$= n\nu_m - 2 \sum_{i=1}^m i\gamma(i). \tag{15.9}$$

Note that as $\nu_m \neq 0$, we have $\sigma = \Omega(n^{1/2})$. Define $W = \sigma^{-1}(\sum_{i=0}^{n-1} X_i)$. We can again apply Corollary 15.11 to conclude

$$d_W(W, Z) \leq \frac{7(2m + 1)^2 n \max_i \mathbb{E}|X_i|^3}{\sigma^3}.$$ 

Hence the proof is complete. \(\blacksquare\)

**Remark 15.16.** Actually the central limit theorem for the $m$-dependent sequence still holds without any third moment assumption. But in order to obtain an error bound for the Berry-Esseen or the Wasserstein distance we need the assumption like uniformly bounded third moments.

### 15.5.3 Number of triangles in $\mathbb{G}(n, p)$

Let $\mathbb{G}(n, p)$ be the Erdős Rényi random graph on $n$ vertices where each of the $\binom{n}{3}$ possible edges is included in the graph independently with probability $p$. Let $T = T_n$ denote the number of triangles in $\mathbb{G}(n, p)$. Define $W = \sigma^{-1}(T - \mathbb{E}T)$ where $\sigma^2 = \text{Var}(T)$. We will show that the number of triangles in $\mathbb{G}(n, p)$ approximately follows the normal distribution.

**Theorem 15.17.** Let $W$ be as above and $p = p(n) \leq 1/2$ be such that $np \to \infty$. Then

$$\text{Wass}(W, Z) \to 0 \quad \text{as } n \to \infty,$$

where $Z \sim N(0,1)$. 

Proof. Let \( \mathcal{T} \) be the set of all triangles in the complete graph on \( \{1, 2, \ldots, n\} \). We can write \( T \) as a sum of indicator random variables

\[
T = \sum_{\alpha \in \mathcal{T}} 1(\alpha \text{ is present in } G(n,p)).
\]

Define \( X_\alpha := 1(\alpha \text{ is present in } G(n,p)) \). Then \( T - E T = \sum_{\alpha \in \mathcal{T}} X_\alpha \). For \( \alpha, \beta \in \mathcal{T} \), we will use the notation \( e(\alpha \cap \beta) \) to denote the number of common edges shared by the triangles \( \alpha \) and \( \beta \). Similarly, for three triangles \( \alpha, \beta \) and \( \gamma \), we define \( e(\alpha, \beta, \gamma) \) is the number of edges that are present in each of the triangles.

Define a graph \( G \) on the vertex set \( \mathcal{T} \) so that two distinct triangles \( \alpha \) and \( \beta \) are connected in \( G \) if and only if \( e(\alpha \cap \beta) > 0 \). Clearly, \( G \) is a dependency graph for \( \{ X_\alpha \}_{\alpha \in \mathcal{T}} \). Indeed, \( X_\alpha \) and \( X_\beta \) are independent if \( \alpha \) and \( \beta \) do not share any common edge. Note that in \( G \),

\[
\bar{N}(\alpha) = \{ \beta \in \mathcal{T} : e(\alpha \cap \beta) = 1 \text{ or } 3 \},
\]

whose size is bounded by \( 3n \). Note that

\[
E 1(\alpha \text{ is present in } G(n,p)) = p^3, \quad E T = \binom{n}{3} p^3
\]

and

\[
\sigma^2 = \sum_{\alpha \in \mathcal{T}, \beta \in \bar{N}(\alpha)} E X_\alpha X_\beta = \sum_{\alpha \in \mathcal{T}} E X_\alpha^2 + \sum_{\alpha, \beta \in \mathcal{T}, e(\alpha, \beta) = 1} E X_\alpha X_\beta
\]

\[
= \binom{n}{3} p^3 (1 - p^3) + \binom{n}{3} (n - 3) p^5 (1 - p) . \tag{15.10}
\]

Now the proof follows by using Theorem 15.10.

15.5.4 Volume of a random polytope in unit ball in \( \mathbb{R}^d \)

Let \( B \) be the (closed) unit ball centered at the origin and let \( \mathcal{P}_n \) be a Poisson point process with intensity \( n \) in \( \mathbb{R}^d, d \geq 2 \). Let \( X_1, X_2, \ldots, X_N \) be the random number of points of \( \mathcal{P}_n \) that fall within the ball \( B \). Define the random polytope \( \Pi_n \) as the convex hull \( [X_1, X_2, \ldots, X_N] = [B \cap \mathcal{P}_n] \) of these random points. Let Volume(\( \Pi_n \)) denote the volume of \( \Pi_n \). In this section we will illustrate how to prove central limit theorem for Volume(\( \Pi_n \)) using the dependency graph techniques. We will also establish upper bound for the rate of convergence in Berry-Esséen distance. Let us now state the main theorem.

**Theorem 15.18.** With notations as above,

\[
\left| P \left( \frac{\text{Volume}(\Pi_n) - E \text{ Volume}(\Pi_n)}{\sqrt{\text{Var}(\text{Volume}(\Pi_n))}} \leq x \right) - \phi(x) \right| \leq b_1 n^{-\frac{1}{2} + \frac{1}{d+1}} \ln^{2 + \frac{2}{d+2}} n,\]

where \( \phi(\cdot) \) is the cumulative distribution function of the standard Normal distribution and \( b_1 \) is a constant which only depends on \( d \).

The proof of the theorem is long, but the main ingredient is dependency graph structure for the volume.
15.6 Exchangeable pairs approach

In this section we will look at the method of exchangeable pair for Stein’s method and apply it to a variety of problems. Method of exchangeable pair is originally due to Charles Stein (1986) himself and it is probably the easiest approach for Stein’s method to apply on a concrete problem. The usefulness comes from the fact that in many examples there is a natural way to perturb the random variable by a small amount without changing the distribution. The basic idea for the exchangeable pair approach is the following. Suppose that $W$ is a random variable with mean zero and variance one. Also suppose that it is possible to perturb $W$ by a ‘small’ amount to get another random variable $W'$ having same marginal as $W$. Assume that the pair $(W',W)$ is exchangeable, that is $(W,W')$ and $(W',W)$ have the same distribution. Now for any ‘nice’ function $f$, by the exchangeability condition and ‘smallness’ of $|W - W'|$ we have

$$E(W' - W)(f(W) + f(W')) = 0$$

or

$$E(W' - W)(f(W') - f(W)) = -2E(W' - W)f(W)$$

or

$$E(W' - W)^2 E(W') - f(W) = -2E(W' - W)f(W)$$

or

$$E[E((W' - W)^2|W) \cdot f'(W)] \approx E[-2E(W' - W|W) \cdot f(W)]$$

Now if we have $E(W' - W|W) \approx -\lambda W$ for some $\lambda \in (0,1)$, exchangeability yields that

$$E(W' - W)^2 = 2E(W - W') \approx 2\lambda E W^2 = 2\lambda.$$

So if we further assume that $\frac{1}{2\lambda} E((W' - W)^2|W)$ is concentrated around its mean $\frac{1}{2\lambda} E(W' - W)^2 \approx 1$, then for any ‘nice’ function $f$ we have

$$E[f'(W)] \approx \frac{1}{2\lambda} E[E((W' - W)^2|W) \cdot f'(W)]$$

$$\approx E\left[-\frac{2E(W' - W|W)}{2\lambda} \cdot f(W)\right] \approx E[Wf(W)]$$

that is, $W$ approximately satisfies the characterizing equation for standard normal distribution. Before going to concrete examples let us state and prove the result precisely. Though exchangeability condition was used in the original theorem, in the proof it suffices to consider equidistributed pair.

15.7 Bound on Wasserstein Metric

**Theorem 15.19.** Let $(W,W')$ be a pair of random variables defined on the same probability space having same marginal distributions, i.e. $W \overset{d}{=} W'$. Suppose $E W = 0, E W^2 = 1$ and

$$E(W' - W|W) = -\lambda W \text{ a.e.}$$

(15.12)

for some constant $\lambda \in (0,1)$. Then, we have

$$d_W(W,Z) \leq \sqrt{\frac{2}{\pi} \text{Var} \left( E\left( \frac{1}{2\lambda}(W' - W)^2|W\right) \right)} + \frac{1}{3\lambda} E|W' - W|^3$$

(15.13)

where $Z \sim N(0,1)$. 

**Corollary 15.20.** Let \((W, W')\) be a pair of random variables defined on the same probability space having same marginal distributions. Suppose \(\mathbb{E}W = 0, \mathbb{E}W^2 = 1\) Define the random variable \(R = R(W)\) by

\[
\mathbb{E}(W' - W|W) = -\lambda W + R \text{ a.e.}
\]  

(15.14)

where \(\lambda\) is a number satisfying \(0 < \lambda < 1\). Then, we have

\[
d_W(W, Z) \leq \sqrt{\frac{2}{\pi}} \text{Var} \left( \mathbb{E} \left( \frac{1}{2\lambda} (W' - W)^2|W \right) \right) + \frac{1}{3\lambda} \mathbb{E} |W' - W|^3 + 2 \frac{\sqrt{\mathbb{E} R^2}}{\lambda}
\]

(15.15)

where \(Z \sim N(0, 1)\).

**Remark 15.21.** Condition (15.12) is natural in many cases. If \((W, W')\) is close to bivariate normal distribution, then the linearity of conditional expectation as a function of \(W\) should hold approximately and one may expect the remainder \(R\) to be small. Also, when \((W, W')\) is bivariate normal, it is easy to check that \(\frac{1}{2\lambda}(W' - W)^2|W\) = \(\lambda W^2 + 1 - \lambda/2\). This indicates that \(\lambda\) should be small and then, if \(W\) is close to normal, one may expect \(\text{Var}(\frac{1}{2\lambda}(W' - W)^2|W))\) to be small.

**Proof.** (proof of Theorem 15.19) Using the condition \(\mathbb{E}(W' - W|W) = -\lambda W\) and \(\mathbb{E}W^2 = 1\) we have

\[
\mathbb{E}(W' - W)^2 = \mathbb{E}[W'^2 + W^2 - 2W'] = \mathbb{E}[2W(W - W')]
\]

\[
= \mathbb{E}[2W^2 - \mathbb{E}(W - W')|W] = 2 \lambda.
\]

Now take any function \(f\) with \(|f|_\infty \leq 1, |f'|_\infty \leq \sqrt{2/\pi}, |f''|_\infty \leq 2\). Define a new function \(F\) by

\[
F(x) = \int_0^x f(y)dy.
\]

(15.16)

Note that for \(x < 0\) we define the integral \(\int_0^x\) as \(-\int_x^0\). Clearly \(F\) is three times differentiable and \(|F(x)| \leq |x|\). Hence \(\mathbb{E}|F(W)| < \infty\). Now from exchangeability we have

\[
0 = \mathbb{E}[F(W') - F(W)]
\]

\[
= \mathbb{E} \left( (W' - W) f(W) + \frac{1}{2} (W' - W)^2 f'(W) + \text{Rem} \right)
\]

(15.17)

where \(|\text{Rem}| \leq \frac{1}{8} |W - W'|^3 |f''|_\infty \leq \frac{1}{3} |W - W'|^3\). In fact using Taylor’s formula we can write down the remainder term explicitly as

\[
\text{Rem} = \frac{V^3}{2} \int_0^1 (1 - s)^2 f''(W + sV)ds.
\]

where \(V = W' - W\). This form of the remainder will be used in the Berry-Esseen bound. From equation (15.17), using properties of conditional expectation we have

\[
\mathbb{E} \left[ -2\lambda W f(W) + \mathbb{E}(W' - W)^2|W) f'(W) + 2\text{Rem} \right] = 0
\]

or \(\mathbb{E}[W f(W)] = \frac{1}{2\lambda} \mathbb{E} \left[ \mathbb{E}(W' - W)^2|W) f'(W) + 2\text{Rem} \right] \).
Hence
\[
|E f'(W) - EW f(W)| 
\leq |E \left[ f'(W) \left( \frac{1}{2\lambda} E((W' - W)^2|W) - 1 \right) \right]| + \frac{1}{\lambda} |E \text{Rem}|
\]
\[
\leq |f'|_{\infty} E \left( \frac{1}{2\lambda} E((W' - W)^2|W) - 1 \right)| + \frac{1}{\lambda} |E \text{Rem}|
\]
\[
\leq \sqrt{\frac{2}{\pi}} \text{Var} \left( \frac{1}{2\lambda} E((W' - W)^2|W) \right) + \frac{1}{3\lambda} E |W - W'|^3.
\]

From the above calculations it follows that
\[
d_W(W, Z) 
= \sup \{ |E g(W) - E g(Z)| : |g'|_{\infty} \leq 1 \}
\leq \sup \{ |E f'(W) - EW f(W)| : |f|_{\infty} \leq 1, |f'|_{\infty} \leq \sqrt{2/\pi}, |f''|_{\infty} \leq 2 \}
\leq \sqrt{\frac{2}{\pi}} \text{Var} \left( \frac{1}{2\lambda} E((W' - W)^2|W) \right) + \frac{1}{3\lambda} E |W - W'|^3
\]
where \( Z \sim N(0,1). \)

Now let us apply the exchangeable pair method to the simplest possible example, sums of independent random variable.

**15.7.1 Example: Sum of i.i.d. random variables**

Suppose \( X_1, X_2, \ldots \) are i.i.d. random variable with zero mean and unit variance. We want to prove CLT for \( W = n^{-1/2} \sum_{i=1}^{n} X_i. \) To construct an exchangeable pair we proceed as follows. Suppose \( X'_1, X'_2, \ldots \) are independent random variables with \( X'_i \) having the same distribution as \( X_1. \) Choose an index \( I \) uniformly at random from \( \{1, 2, \ldots, n\}. \) Replace \( X_I \) in \( W \) by \( X'_I. \) Let
\[
W' = \frac{1}{\sqrt{n}} \sum_{j \neq I} X_j + \frac{X'_I}{\sqrt{n}} = W + \frac{X'_I - X_I}{\sqrt{n}}.
\]

Clearly \((W, W')\) is an exchangeable pair. Now we have
\[
E[W' - W|W] = \frac{1}{\sqrt{n}} E[X'_I - X_I|W] = \frac{1}{\sqrt{n}} \frac{1}{n} \sum_{i=1}^{n} E[X'_I - X_i|W]
\]
\[
= -\frac{1}{n^{3/2}} E \left[ \sum_{i=1}^{n} X_i|W \right] = -\frac{1}{n} W.
\]

Hence the condition of Theorem 15.19 is satisfied with \( \lambda = n^{-1}. \) Note that
\[
\frac{1}{3\lambda} E |W' - W|^3 = \frac{n}{3n^{3/2}} E |X'_I - X_I|^3
\]
\[
= \frac{1}{3n^{3/2}} \sum_{i=1}^{n} E |X'_I - X_i|^3 \leq \frac{8}{3\sqrt{n}} E |X_1|^3
\]
and
\[ E \left[ \frac{1}{2\lambda} (W' - W)^2 \middle| W \right] = \frac{n}{2n} E((X'_i - X_i)^2 \middle| W) \]
\[ = \frac{1}{2n} \sum_{i=1}^{n} E((X'_i - X_i)^2 \middle| W) = \frac{1}{2} + E \left[ \frac{1}{2n} \sum_{i=1}^{n} X_i^2 \middle| W \right] . \]

Hence we have
\[ d_W(W, Z) \leq \sqrt{\frac{2}{\pi} \text{Var} \left( E \left[ \frac{1}{2n} \sum_{i=1}^{n} X_i^2 \middle| W \right] \right) + \frac{8}{3n^{3/2}} \sum_{i=1}^{n} E |X_i|^3} \]
\[ \leq \sqrt{\frac{1}{2\pi n^2} \text{Var} \left( \sum_{i=1}^{n} X_i^2 \right) + \frac{8}{3n} E |X_1|^3} \]
\[ \leq \left( \frac{\sqrt{\text{Var} \left( X_1^2 \right)}}{2\pi} + \frac{8}{3} \frac{E |X_1|^3}{n} \right) \cdot \frac{1}{\sqrt{n}} . \]

And this gives another proof of the famous Central Limit theorem. \( \square \)

### 15.7.2 Hoeffding’s combinatorial central limit theorem

Suppose \( A = ((a_{ij}))_{i,j=1}^{n} \) is an array of real numbers. Let \( \pi \) be a uniform random permutation of \( \{1, 2, ..., n\} \). Define the random variable \( W \) by \( W = W_A = \sum_{i=1}^{n} a_{i\pi(i)} \). Define \( \mu_A = EW_A \) and \( \sigma_A^2 = \text{Var}(W_A) \). This type of statistics arises in nonparametric tests and sampling from finite population. If we look at the normalized random variable \( \frac{1}{n} \sum_{i=1}^{n} X_i \), the question then arises is under what conditions on \( A = ((a_{ij}))_{i,j=1}^{n} \) we can approximate the random variable by standard normal distribution. Hoeffding, in 1951, proved that for a sequence of matrices \( A_n, \sigma_A^{-1}(W_{A_n} - \mu_{A_n}) \) is approximately the standard Gaussian distribution under certain regularity condition on \( A_n \). His original proof gave conditions for convergence to normality using the method of moments. In this section we will use the exchangeable pair approach to prove the combinatorial central limit theorem. We will bound the Wasserstein distance using theorem (15.19).

Define
\[ a_i = \frac{1}{n} \sum_{j=1}^{n} a_{ij}, \quad a_j = \frac{1}{n} \sum_{i=1}^{n} a_{ij}, \quad a_\pi = \frac{1}{n^2} \sum_{i,j=1}^{n} a_{ij} \]  \hspace{1cm} (15.18)

and
\[ \tilde{a}_{ij} = \frac{a_{ij} - a_i - a_j + a_\pi}{\sigma_A} . \]

Define the matrix \( \tilde{A} = ((\tilde{a}_{ij}))_{i,j=1}^{n} \). Clearly \( \mu_A = na_\pi \) and we have \( \tilde{a}_{i\pi} = \tilde{a}_{j\pi} = 0 \) for all \( 1 \leq i, j \leq n \). Define
\[ \tilde{W} := W_{\tilde{A}} = \frac{1}{\sigma_A} \sum_{i=1}^{n} \tilde{a}_{i\pi(i)} \]
\[ = \frac{1}{\sigma_A} \left( \sum_{i=1}^{n} a_{i\pi(i)} - \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} a_{\pi(i)} + na_\pi \right) = \frac{W - \mu_A}{\sigma_A} . \]
Hence $E \hat{W} = 0$ and $\text{Var}(\hat{W}) = 1$. Now we have $E \hat{a}_{i\pi(i)} = n^{-1} \sum_{j=1}^{n} \hat{a}_{ij} = 0$ for all $1 \leq i \leq n$. Hence we can also write the variance as

$$
\text{Var}(\hat{W}) = \sum_{i=1}^{n} E(\hat{a}_{i\pi(i)}^2) + \sum_{i \neq j} E(\hat{a}_{i\pi(i)} \hat{a}_{j\pi(j)}) = \frac{1}{n} \sum_{i,j=1}^{n} \hat{a}_{ij}^2 + \frac{1}{n(n-1)} \sum_{i,j \neq k,l} \hat{a}_{ik} \hat{a}_{jl} = \frac{1}{n(n-1)} \sum_{i,k=1}^{n} \hat{a}_{ik} \hat{a}_{ik} = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} \hat{a}_{ij}^2
$$

$$
= \frac{1}{(n-1)\sigma_A^2} \sum_{i,j=1}^{n} (a_{ij} - a_{i.} - a_{.j} + a_{..})^2.
$$

Here we used the fact that for any $i, k$ we have $\sum_{t \neq k} \hat{a}_{it} = -\hat{a}_{ik}$ and $\sum_{j \neq i} \hat{a}_{jk} = -\hat{a}_{ik}$. This implies that

$$
\sigma_A^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij} - a_{i.} - a_{.j} + a_{..})^2
$$

$$
= \frac{1}{(n-1)} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 - n \sum_{i=1}^{n} a_{i.}^2 - n \sum_{j=1}^{n} a_{.j}^2 + n^2 a_{..}^2 \right). \tag{15.19}
$$

Hence without loss of generality we assume the following,

$$
\sum_{j=1}^{n} a_{ij} = 0, \quad \sum_{i=1}^{n} a_{ij} = 0 \quad \text{and} \quad \frac{1}{n-1} \sum_{i,j=1}^{n} a_{ij}^2 = 1.
$$

Under condition (15.20) we have $E W = 0$ and $E W^2 = 1$. Now we will define a random variable $W'$ so that $(W, W')$ satisfies the condition (15.12). Define $\pi' = \pi \circ (I, J)$ where $(I, J)$ is a uniformly chosen random transposition. Clearly $(\pi, \pi')$ is an exchangeable pair. Let $W' = \sum_{i=1}^{n} a_{i\pi'(i)}$. So $(W, W')$ is an exchangeable pair. Note that

$$
W' - W = a_{I\pi(I)} + a_{J\pi'(J)} - a_{I\pi(I)} - a_{J\pi(J)} = a_{I\pi(I)} + a_{J\pi(J)} - a_{I\pi(I)} - a_{J\pi(J)}.
$$

So, by summing over the choices for $(I, J)$,

$$
E(W' - W | \pi) = \frac{1}{n(n-1)} \sum_{i,j \neq i} (a_{i\pi(j)} + a_{j\pi(i)} - a_{i\pi(i)} - a_{j\pi(j)})
$$

$$
= \frac{1}{n(n-1)} \left( - \sum_{i} a_{i\pi(i)} - \sum_{i} a_{i\pi(i)} - 2(n-1) \sum_{i} a_{i\pi(i)} \right)
$$

$$
= - \frac{2}{n-1} W.
$$

Hence condition (15.12) is satisfied with $\lambda = 2(n-1)^{-1}$. Note that

$$
E(|W' - W|^3 | \pi) = \frac{1}{n(n-1)} \sum_{i,j \neq i} |a_{i\pi(i)} + a_{j\pi(j)} - a_{i\pi(i)} - a_{j\pi(j)}|^3
$$

$$
\leq \frac{16}{n(n-1)} \sum_{i,j \neq i} (|a_{i\pi(i)}|^3 + |a_{j\pi(j)}|^3 + |a_{i\pi(i)}|^3 + |a_{j\pi(j)}|^3).
$$
So we have
\[ E(|W' - W|^3) \leq \frac{64}{n^2} \sum_{i,j} |a_{ij}|^3. \]

Now we will prove concentration for the conditional variance \( \text{Var}((W' - W)^2|\pi) \). We have
\[
E((W' - W)^2|\pi) = \frac{1}{n(n-1)} \sum_{i,j} (a_{i\pi(i)} + a_{j\pi(j)} - a_{i\pi(j)} - a_{j\pi(i)})^2
\]
\[
= \frac{1}{n(n-1)} \left( 2n \sum_{i=1}^{n} a_{i\pi(i)}^2 + 2 \sum_{i,j} a_{i\pi(i)}^2 + 2 \sum_{i} a_{i\pi(i)}^2 + 2 \sum_{i,j} a_{i\pi(j)} a_{j\pi(i)} \right)
\]
\[
= \frac{2(n+1)}{n(n-1)} \sum_{i=1}^{n} a_{i\pi(i)}^2 + 2 \frac{2}{n(n-1)} W^2 + 2 \frac{2}{n(n-1)} \sum_{i,j \neq i} a_{i\pi(j)} a_{j\pi(i)}.
\]

So it is enough to bound
\[
\frac{1}{2\lambda} \left| E((W' - W)^2|\pi) - E(W' - W)^2 \right|
\]
\[
= \frac{n-1}{4} \left( \frac{2(n+1)}{n(n-1)} \left| \sum_{i=1}^{n} a_{i\pi(i)}^2 - \sum_{i=1}^{n} a_{i\pi(i)}^2 \right| + \frac{2}{n(n-1)} E|W^2 - 1| + \frac{2}{n(n-1)} \sum_{i,j \neq i} a_{i\pi(j)} a_{j\pi(i)} \right)
\]
\[
\leq \sqrt{\text{Var}\left( \sum_{i=1}^{n} a_{i\pi(i)}^2 \right)} + \frac{1}{n} + \frac{1}{2n} \sqrt{\text{Var}\left( \sum_{i,j \neq i} a_{i\pi(j)} a_{j\pi(i)} \right)}.
\]

Defining \( b_{ij} = a_{ij}^2 \) and using (15.19) we have
\[
\text{Var}\left( \sum_{i=1}^{n} a_{i\pi(i)}^2 \right) = \text{Var}\left( \sum_{i=1}^{n} b_{i\pi(i)} \right) \leq \frac{1}{n-1} \left( \sum_{i,j} b_{ij}^2 + n^2 b^2. \right)
\]
\[
= \frac{1}{n-1} \sum_{i,j} a_{ij}^4 + \frac{n-1}{n^2}.
\]

Also we have
\[
\text{Var}\left( \sum_{i,j \neq i} a_{i\pi(j)} a_{j\pi(i)} \right) \leq E\left( \sum_{i,j \neq i} a_{i\pi(j)} a_{j\pi(i)} \right)^2 = \frac{1}{n(n-1)} \sum_{k,l \neq k} a_{ik} a_{jl}^2
\]
\[
= \frac{1}{n(n-1)} \sum_{k,l \neq k} \left( \sum_{i} a_{ik} a_{il} \right)^2 \leq \frac{1}{n(n-1)} \sum_{k,l} \left( \sum_{i} a_{ik} a_{il} \right)^2
\]
\[
\leq \frac{1}{n(n-1)} \sum_{k,l} \left( \sum_{i} a_{ik}^2 \right)^2 \left( \sum_{i} a_{il}^2 \right) = \frac{n-1}{n} \leq 1.
\]

Hence combining everything we have the following theorem.
Theorem 15.22. Let \((a_{ij})_{i,j=1}^{n}\) be an array of real numbers and \(\pi\) be a uniform random permutation of \(\{1, 2, \ldots, n\}\). Let \(W\) be the random variable \(\sum_{i=1}^{n} a_{i\pi(i)}\). Define the matrix \(((\tilde{a}_{ij})_{i,j=1}^{n})\) by
\[
\tilde{a}_{ij} = \frac{a_{ij} - a_{i.} - a_{.j} + a_{..}}{\sqrt{(n-1)^{-1} \sum_{k=1}^{n} (a_{kl} - a_{k.} - a_{.l} + a_{..})^2}}
\] (15.20)
for all \(1 \leq i, j \leq n\). Then we have
\[
d_W \left( \frac{W - EW}{\sqrt{\text{Var}(W)}} , Z \right) \leq \sqrt{\frac{2}{\pi}} \cdot \frac{1}{n-1} \sum_{i,j} \tilde{a}_{ij}^4 + \frac{2}{\sqrt{n}} + \frac{11}{n} \sum_{i,j} |\tilde{a}_{ij}|^3
\] (15.21)
where \(Z\) follows standard Gaussian distribution.

Note that if \(a_{ij}\)'s are of constant order, then \(\tilde{a}_{ij}\)'s are of order \(n^{-1/2}\). And hence the above theorem gives a rate of convergence of \(n^{-1/2}\). Note that since
\[
\frac{1}{n} \sum_{i,j} |\tilde{a}_{ij}|^3 \leq \frac{1}{n^{1/4}} \left( \frac{1}{n} \sum_{i,j} \tilde{a}_{ij}^4 \right)^{3/4}
\]
we have the following corollary.

Corollary 15.23. Let \(A_n = ((a_{ij}^{(n)})_{i,j=1}^{n})\) be a sequence of real matrices. Let \(W_n := W_{A_n}\) be the random variable defined as above. Let \(\mu_n = EW_n\) and \(\sigma_n^2 = \text{Var}(W_n)\) for \(n \geq 1\). Then
\[
\frac{W_n - \mu_n}{\sigma_n} \to N(0, 1) \text{ in distribution as } n \to \infty
\]
if
\[
\frac{1}{n} \sum_{i,j} (\tilde{a}_{ij}^{(n)})^4 \to 0 \text{ as } n \to \infty.
\] (15.22)

15.7.3 CLT for magnetization in Ising model

Consider a graph \(G = (V, E)\) with vertex set \(V\) of size \(n\) and edge set \(E\). The ferromagnetic Ising model on \(G\) is defined as follows. At each vertex \(v \in V\) of \(G\) there is a particle with spin \(\sigma_v\) taking values in the set \(\{+1, -1\}\). Denote by \(\sigma = (\sigma_v)_{v \in V}\) the spin configuration of the whole system. The spins \(\sigma_v\) interact in pairs. In particular the negative Hamiltonian\(^2\) of the system with configuration \(\sigma\) is defined by
\[
H(\sigma) = \sum_{u \sim v} \sigma_u \sigma_v
\] (15.23)
where \(u \sim v\) means \(u, v\) are neighbors in \(G\). At inverse temperature \(\beta\) the probability of a configuration \(\sigma\) is proportional to \(e^{\beta H(\sigma)}\). In particular, the probability of a configuration \(\sigma\) is
\[
P(\sigma) = \frac{1}{Z_\beta} e^{\beta H(\sigma)}
\]
\(^2\)Physicists use the formula \(H(\sigma) = -\sum_{u \sim v} \sigma_u \sigma_v\) for defining Hamiltonian and they define the probability of a configuration \(\sigma\) as \(P(\sigma) \propto e^{-\beta H(\sigma)}\). For simplicity, instead of taking two negative sign, we take the definition of the Hamiltonian having positive sign.
where \( Z = \sum_\sigma \exp(\beta H(\sigma)) \) is the partition function. The magnetization corresponding to the configuration \( \sigma \) is defined as the average spin

\[
m(\sigma) = \frac{1}{n} \sum_{v \in V} \sigma_v.
\]

(15.24)

Also for a given vertex \( v \in V \) we define

\[
N_v(\sigma) = \sum_{u \sim v} \sigma_u
\]

as the sum of spins of the neighbors of \( v \). Note that from the fact that \( \mathbb{P}(\sigma) = \mathbb{P}(\sigma) \) for all configurations \( \sigma \) we have \( \mathbb{E}(\sigma_v) = 0 \) for all \( v \in V \). Also note that conditional on the neighbors of a vertex \( v \) the spin \( \sigma_v \) is independent of the spins of remaining vertices. In particular we have

\[
\mathbb{P}(\sigma_v = \varepsilon | \sigma_u : u \neq v) = \mathbb{P}(\sigma_v = \varepsilon | \sigma_u : u \sim v) = \frac{\exp(\varepsilon \beta N_v(\sigma))}{\exp(\beta N_v(\sigma)) + \exp(-\beta N_v(\sigma))}
\]

for \( \varepsilon \in \{+1, -1\} \) and

\[
\mathbb{E}(\sigma_v | \sigma_u : u \neq v) = \frac{\exp(\beta N_v(\sigma)) - \exp(-\beta N_v(\sigma))}{\exp(\beta N_v(\sigma)) + \exp(-\beta N_v(\sigma))} = \tanh(\beta N_v(\sigma)).
\]

(15.26)

Define \( \nu^2 = \text{Var}(\sum_{v \in V} \sigma_v) \). Define the random variable \( W := W_n = n^{-1} m(\sigma) = \nu^{-1} \sum_{v \in V} \sigma_v \) where \( n \) is the size of the vertex set \( V \). Clearly \( \mathbb{E}(W) = 0 \) and \( \text{Var}(W) = 1 \). In this section we will prove that for certain sequences of graphs and values of the inverse temperature \( \beta \), \( W_n \) is approximately standard Gaussian. In particular we will prove CLT for magnetization in Ising model in \( n \)-cycles and complete graphs under appropriate conditions. We will use the exchangeable pair theorem. To define the exchangeable pair we will use the Glauber dynamics, which is defined as follows.

Given \( \sigma \) from the Gibbs distribution \( \mathbb{P} \), construct \( \sigma' \) as follows. Choose an vertex \( I \) uniformly at random from \( V \). Replace \( \sigma_I \) by \( \sigma'_I \) drawn from the conditional distribution of \( \sigma_I \) given \( \{\sigma_u : u \neq I\} \). Keep all other \( \sigma'_u, u \neq I \) same as \( \sigma_u, u \neq I \). If we define \( W' = \nu^{-1} \sum_{v \in V} \sigma'_v \) we have \( W' \overset{d}{=} W \) and

\[
W' - W = \nu^{-1}(\sigma'_I - \sigma_I).
\]

Hence \( |W' - W| \leq 2\nu^{-1} \). Also note that

\[
\mathbb{E}(W' - W | \sigma) = \frac{1}{\nu} \mathbb{E}(\sigma'_I - \sigma_I | \sigma) = \frac{1}{n\nu} \sum_{v \in V} (\tanh(\beta N_v(\sigma)) - \sigma_v).
\]

Now we will go into the details.

### 15.7.3.1 Ising model on complete graph: Curie-Weiss model

Here we consider the complete graph \( G = G_n \) with vertex set \( V = [n] \) and egde set \( E = \{(i, j) : 1 \leq i < j \leq n\} \). Note that here every vertex has degree \( n - 1 \). So instead of \( \beta \) we generally take \( \beta/n \) as the inverse temperature parameter. This model is also known as Curie-Weiss
model. So here the probability of a configuration $\sigma \in \{+1,-1\}^n$ is $P(\sigma) = Z_\beta^{-1} \exp(\beta H(\sigma)/n)$ where $H(\sigma) = \sum_{i<j} \sigma_i \sigma_j$ is the negative Hamiltonian and $Z_\beta = \sum_{\sigma \in \{+1,-1\}^n} \exp(\beta H(\sigma)/n)$ is the partition function. We will prove CLT for the magnetization $m(\sigma) = n^{-1} \sum_{i=1}^n \sigma_i$ when $\beta < 1$.

Again let $\sigma, \sigma', W, W', \nu$ be as defined before. Note that here for $i \in [n]$ we have $N_i(\sigma) = \sum_{j \neq i} \sigma_j$. Hence using (15.26) we have

$$E(\sigma_i | \sigma_j, j \neq i) = \tanh(\beta m_i(\sigma))$$

where $m_i(\sigma) = n^{-1} N_i(\sigma) = n^{-1} \sum_{j \neq i} \sigma_j$. Now using the conditional expectation formula we can write

$$E(W' - W | \sigma) = -\frac{1}{n} \left( W - \frac{1}{\nu} \sum_{i=1}^n \tanh(\beta m_i(\sigma)) \right)$$

$$= -\frac{1 - \beta}{n} W + \frac{1}{n\nu} \left( \sum_{i=1}^n \tanh(\beta m_i(\sigma)) - n\beta m(\sigma) \right)$$

$$= -\frac{1 - \beta}{n} (W - \nu^{-1} D)$$

where $D := D(\sigma) = (1 - \beta)^{-1} (\sum_{i=1}^n \tanh(\beta m_i(\sigma)) - n\beta m(\sigma))$. Hence the conditions in corollary (15.20) is satisfied with

$$\lambda = \frac{1 - \beta}{n} \quad \text{and} \quad \frac{R}{\lambda} = \frac{D}{\nu}.$$  

Note that here we have

$$(1 - \beta)\nu \cdot \frac{|R|}{\lambda} \leq \sum_{i=1}^n |\tanh(\beta m_i(\sigma)) - \beta m(\sigma)|$$

$$\leq \sum_{i=1}^n |\tanh(\beta m_i(\sigma)) - \tanh(\beta m(\sigma))| + n |\tanh(\beta m(\sigma)) - \beta m(\sigma)|$$

$$\leq \beta + \frac{n\beta^3 |m(\sigma)|^3}{3}.$$  

Here we use the fact that

$$|\tanh(x) - \tanh(y)| \leq |x - y|$$

and $|m(\sigma) - m_i(\sigma)| \leq n^{-1}$. We will prove that $\nu^2$ is asymptotically $n/(1 - \beta)$ and $m(\sigma)$ is concentrated around its mean 0. Note that $E(m(\sigma))^2 = \nu^2/n^2 \approx (n(1 - \beta)^{-1}$. Hence indeed $\lambda^{-1} E[R]$ is small compared to $W$. Expanding the squares and using the conditional expectation formula again we have

$$E((W' - W)^2 | \sigma) = \frac{1}{n
u^2} \sum_{i=1}^n E((\sigma'_i - \sigma_i)^2 | \sigma)$$

$$= \frac{2}{n\nu^2} \left( n - \sum_{i=1}^n \sigma_i \tanh(\beta m_i(\sigma)) \right) = \frac{2}{n\nu^2} (n - Y)$$

where $Y := Y(\sigma) = \sum_{i=1}^n \sigma_i \tanh(\beta m_i(\sigma))$. So

$$\frac{1}{2\lambda} E((W' - W)^2 | \sigma) = \frac{n - Y}{(1 - \beta)\nu^2} \approx 1 - \frac{Y}{n}.$$  

(15.29)
Note that
\[ Y \approx n m(\sigma) \tanh(\beta m(\sigma)) \approx n \beta(m(\sigma))^2 \approx \beta \nu^2/n \approx \beta/(1 - \beta). \]
So indeed the conditional expectation \( \frac{1}{2}\nu \mathbb{E}((W' - W)^2|\sigma) \) is concentrated around its mean 1. Hence from the result of corollary (15.20), we have
\[
d_W(W, Z) \leq \sqrt{\frac{2}{n}} \mathbb{E} \left| \frac{n - Y}{(1 - \beta)\nu^2} \right| + \frac{8n}{3(1 - \beta)\nu^3} + \frac{2E|R|}{\lambda}. \]

One can prove that
\[
|f(x)| \leq |g - Ng|_\infty \left( e^{x^2/2} \int_x^{\infty} e^{-y^2/2} dy \right).
\]
Recall the Mill’s ratio inequality on \( \Phi(x) \) for \( x > 0 \):
\[
\frac{xe^{x^2/2}}{\sqrt{2\pi}(1 + x^2)} \leq 1 - \Phi(x) \leq \frac{e^{x^2/2}}{x\sqrt{2\pi}}.
\]

Now, \( \frac{d}{dx} e^{x^2/2} \int_x^{\infty} e^{-y^2/2} dy = -1 + xe^{x^2/2} \int_x^{\infty} e^{-y^2/2} dy \leq 0 \forall x > 0 \) by Mill’s ratio inequality (15.31). So, \( e^{x^2/2} \int_x^{\infty} e^{-y^2/2} dy \) is maximized at \( x = 0 \) on \([0, \infty)\) where its value is \( \sqrt{\frac{\pi}{2}} \). Hence,
\[
|f(x)| \leq \sqrt{\frac{\pi}{2}} |g - Ng|_\infty \forall x > 0.
\]
For \( x < 0 \), use the form (15.3) and proceed in the similar manner.

Proof of bound (II) : Again, we will only consider \( x > 0 \) case. The other case will be similar. Note that
\[
f'(x) = g(x) - Ng + xf(x) = g(x) - Ng - xe^{x^2/2} \int_x^{\infty} e^{-y^2/2}(g(y) - Ng)dy.
\]

Therefore,
\[
|f'(x)| \leq |g - Ng|_\infty \left( 1 + xe^{x^2/2} \int_x^{\infty} e^{-y^2/2} dy \right)
\leq 2|g - Ng|_\infty \text{ by Mill’s ratio inequality (15.31).}
\]

Proof of bound (III) : Applying Stein’s identity on (15.5), we have
\[
f(x) = -\int_0^1 \frac{1}{2\sqrt{t}} \mathbb{E} \left[ g'((\sqrt{t}x + \sqrt{1-t})^2) \right] dt.
\]
Hence,
\[ |f|_\infty \leq |g'|_\infty \int_0^1 \frac{1}{2\sqrt{t}} = |g'|_\infty. \]

**Proof of bound (IV):** From (15.6), it follows that
\[ |f|_\infty \leq (E|Z|)|g'|_\infty \int_0^1 \frac{1}{2\sqrt{1-t}} = \sqrt{\frac{2}{\pi}} |g'|_\infty. \]

**Proof of bound (V):** On differentiating (15.2) and rearranging
\[ f''(x) = g'(x) + f(x) + x f'(x) \]
\[ = g'(x) + f(x) + x(g(x) - Ng + x f(x)) \]
\[ = g'(x) + x(g(x) - Ng) + (1 + x^2) f(x). \] (15.33)

We can write \( g(x) - Ng \) in terms of \( g' \) as follows,
\[
g(x) - Ng = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-y^2/2}(g(x) - g(y))dy \\
= \frac{1}{\sqrt{2\pi}} \left[ \int_{-\infty}^x \int_y^\infty g'(z)e^{-y^2/2}dzdy - \int_{-\infty}^x \int_y^\infty g'(z)e^{-y^2/2}dzdy \right] \\
= \frac{1}{\sqrt{2\pi}} \left[ \int_{-\infty}^x g'(z) \int_{-\infty}^z e^{-y^2/2}dydz - \int_{-\infty}^x g'(z) \int_{z}^\infty e^{-y^2/2}dydz \right] \\
= \int_{-\infty}^x g'(z)\Phi(z)dz - \int_x^\infty g'(z)\Phi(z)dz
\]
where \( \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-y^2/2}dy \) is the distribution function for standard normal and \( \Phi(z) = 1 - \Phi(z) \). Similarly,
\[
f(x) = e^{x^2/2} \int_{-\infty}^\infty e^{-y^2/2}(g(y) - Eg(Z))dy \\
= e^{x^2/2} \int_{-\infty}^x e^{-y^2/2} \left( \int_y^\infty g'(z)\Phi(z)dz - \int_y^\infty g'(z)\Phi(z)dz \right) dzdy \\
= e^{x^2/2} \left( \int_{-\infty}^x g'(z)\Phi(z) \int_z^\infty e^{-y^2/2}dydz - \int_{-\infty}^x g'(z)\Phi(z) \int_{-\infty}^{z\vee x} e^{-y^2/2}dydz \right) \\
= \sqrt{2\pi} e^{x^2/2} \left( \int_{-\infty}^x g'(z)\Phi(z)(\Phi(z) - \Phi(x))dz \\
- \int_{-\infty}^x g'(z)\Phi(z)\Phi(z)dz - \int_x^\infty g'(z)\Phi(z)\Phi(x)dz \right) \\
= -\sqrt{2\pi} e^{x^2/2} \left[ \Phi(x) \int_{-\infty}^x g'(z)\Phi(z)dz + \Phi(x) \int_x^\infty g'(z)\Phi(z)dz \right].
\]
Substituting the above expressions for \( g - Ng \) and \( f \) in (15.33), we get
\[
f''(x) = g'(x) + (x - \sqrt{2\pi}(1 + x^2)e^{x^2/2}\Phi(x)) \int_{-\infty}^x g'(z)\Phi(z)dz \\
+ (-x - \sqrt{2\pi}(1 + x^2)e^{x^2/2}\Phi(x)) \int_x^\infty g'(z)\Phi(z)dz.
\]
This gives
\[
|f''(x)| \leq |g'\bigg|_{\infty} \left[ 1 + \left| x - \sqrt{2}\pi(1 + x^2)e^{x^2/2}(1 - \Phi(x)) \right| \int_{-\infty}^{x} \Phi(z) dz \\
+ \left| -x - \sqrt{2}\pi(1 + x^2)e^{x^2/2}\Phi(x) \right| \int_{x}^{\infty} (1 - \Phi(z)) dz \right].
\] (15.34)

There is a similar bound for \( x \leq 0 \). To proceed, we wish to remove the absolute values in equation (15.34), by determining the sign of the expressions within the absolute value. From the Mill’s ratio (15.31) we have
\[
x + \sqrt{2}\pi(1 + x^2)e^{x^2/2}\Phi(x) > 0
\] (15.35)
and
\[
-x + \sqrt{2}\pi(1 + x^2)e^{x^2/2}(1 - \Phi(x)) > 0.
\] (15.36)
One can check (15.36) by noting that for \( x < 0 \) the inequality is obvious, and for \( x > 0 \) use the lower Mill’s ratio inequality; (15.35) follows similarly. Hence both expressions within the absolute values in equation (15.34) are negative.

To finish the simplification, observe that integration by parts gives
\[
\int_{-\infty}^{x} \Phi(z) dz = x\Phi(x) + \frac{e^{-x^2/2}}{\sqrt{2}\pi}
\]
and
\[
\int_{x}^{\infty} (1 - \Phi(z)) dz = -x(1 - \Phi(x)) + \frac{e^{-x^2/2}}{\sqrt{2}\pi}.
\]
Combining, we get
\[
|f''(x)| \leq |g'\bigg|_{\infty} \left[ 1 + \left( x + \sqrt{2}\pi(1 + x^2)e^{x^2/2}(1 - \Phi(x)) \right) \left( x\Phi(x) + \frac{e^{-x^2/2}}{\sqrt{2}\pi} \right) \\
+ \left( -x - \sqrt{2}\pi(1 + x^2)e^{x^2/2}\Phi(x) \right) \left( -x(1 - \Phi(x)) + \frac{e^{-x^2/2}}{\sqrt{2}\pi} \right) \right] \\
= 2|g'\bigg|_{\infty}.
\]
This proves the desired bound.