Notes for ORF 526

Eric Qian, Filippo Palomba

November 26, 2022

Contents

1 Measure theory

We will spend the first few lectures of the course discussing measure theory. Measure theory helps unify discrete and continuous random variables and with treating mixed cases. It also helps for distributions in $\mathbb{R}^n$.

1.1 Sigma algebras

Definition (Closure)**.** *A collection $\mathcal{C}$ of subsets of E is **closed under intersection** if $\forall A, B \in \mathcal{C}$, then $A \cap B \in \mathcal{C}$.*

Definition (Algebra)**.** *Let E be a set and $\mathcal{E}$ be a non-empty collection of subsets. $\mathcal{E}$ is an **algebra** if*

1. $\emptyset \in \mathcal{E}$
2. *Closed under complements:* $A \in \mathcal{E} \implies A^c \in \mathcal{E}$
3. *Closed under finite unions:* $A, B \in \mathcal{E} \implies A \cup B \in \mathcal{E}$

Definition (Sigma-Algebra)**.** *Let E be a set and $\mathcal{E}$ be a non-empty collection of subsets. $\mathcal{E}$ is a **sigma-algebra** if*

1. $\emptyset \in \mathcal{E}$
2. *Closed under complements:* $A \in \mathcal{E} \implies A^c \in \mathcal{E}$
3. *Closed under countable unions:* $A_1, A_2, \ldots \in \mathcal{E} \implies \cup_{i=1}^{\infty} A_i \in \mathcal{E}$

Proposition (De Morgan's Law)**.** $\cap_{i \in I} A_i = \left(\cup_{i \in I} A_i^c\right)^c$.

Note. *Statements about countable unions are implied by statements about countable intersections.*

Definition (Trivial σ-algebra)**.** $\mathcal{E} = \{\emptyset, E\}$ *is called the **trivial sigma-algebra**.*

Definition (Discrete σ-algebra)**.** $\mathcal{E} = \left\{\text{collection of all subsets of } E \equiv 2^E\right\}$ *is called the **discrete sigma-algebra**.*

Property (Intersections and unions of σ-algebras)**.** *A few properties left as homework.*

1. *Intersections of an arbitrary (countable or uncountable) family of sigma-algebras on E are also sigma-algebras on E.*
2. *The union is not necessarily a sigma-algebra.*
3. *An arbitrary collection $\mathcal{C}$ of subsets of E is not necessarily a sigma-algebra.*

Definition (Generated $\sigma-$algebra)**.** *Given an arbitrary collection $\mathcal{C}$ of subsets of E. Consider all σ-algebras on E that contain $\mathcal{C}$. Take the intersection of all these. The smallest σ-algebra containing $\mathcal{C}$ is the **σ-algebra generated by** $\mathcal{C}$ denoted by $\sigma(\mathcal{C})$. Denote $\mathcal{E}_1 \vee \mathcal{E}_2 \equiv \sigma(\mathcal{E}_1 \cup \mathcal{E}_2)$.*

Definition (Topological space)**.** *A topological space is an ordered pair (X, ξ) such that:*

- *$\emptyset$ and X belong to ξ*
- *ξ is closed under (finite and infinite) union*
- *ξ is closed under finite intersection*

Members of ξ are called open sets and ξ is a topology on X.

Definition (Borel sigma-algebra)**.** *If E is a topological space, then the sigma-algebra generated by open sets is called the **Borel sigma-algebra**.*

Note. *We can generate the Borel sigma-algebra by using all sorts of intervals on the real line, i.e. closed, open, half-closed and half-open.*

Definition (Monotone class)**.** *A **monotone class** is a collection of sets $\mathcal{M}$, which is closed under countable monotone union and intersections i.e. if $A_1, \ldots \in \mathcal{M}$ and $A_1 \subseteq A_2 \subseteq \ldots$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{M}$.*

Theorem (Monotone class)**.** *Let $\mathcal{E}$ be an algebra on E. Let $\mathcal{M} = \mathcal{M}(\mathcal{E})$ be the smallest monotone class containing E. Then $\mathcal{M} = \sigma(\mathcal{E})$.*

1.2 Measurable spaces

Definition (Measurable space)**.** *A **measurable space** is $(E, \mathcal{E})$, where E is a set and $\mathcal{E}$ is a sigma-algebra on E. The elements of $\mathcal{E}$ are called measurable sets.*

Definition (Measure)**.** *Let $(E, \mathcal{E})$ be a measurable space. A **measure** on $(E, \mathcal{E})$ is a mapping $\mu : \mathcal{E} \to \overline{\mathbb{R}}_+$ such that*

1. *$\mu(\emptyset) = 0$.*
2. *$\mu\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$ for every countable collection of pairwise disjoint sets $\{A_n\}_{n=1}^{\infty}$.*

Example (Measures)**.** *Below are a few common examples.*

1. ***Dirac measure***. *Let $A \in \mathcal{E}$, $x \in E$, and $(E, \mathcal{E})$ be a measurable space. Then*
$$\delta_x (A) \equiv \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$
2. ***Counting measure***. *Fix $D \subseteq E$. Let $A \in \mathcal{E}, \nu \equiv$ # points in $A \cup D$. When D is countable, then*
$$\nu (A) = \sum_{x \in D} \delta_x (A)$$
In words, the counting measure on D assigns mass one to all those point in D that are also in A. Of course it might be that $D = E$.
3. ***Discrete measure***: *Let $D \subseteq E$ countable.*
$$\mu (A) \equiv \sum_{x \in D} m (x) \delta_x (A), \quad A \in \mathcal{E}$$
where $m (x) \geq 0$ is the "mass" at x.
4. ***Lebesgue measure***: *A measure μ on $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ is called the Lebesgue measure if for every interval A, $\mu (A)$ is the length of the interval. The Lebesgue measure on $\mathbb{R}^2$ is the* ***area measure****. The Lebesgue measure on $\mathbb{R}^3$ is the* ***volume measure****.*

Note. *The notation $A_n \nearrow A$ means that $\{A_n\}_n$ is an increasing sequence of sets in the sense that $A_1 \subseteq A_2 \subseteq A_3 \cdots$ and we define $A \equiv \cup_{j=1}^{\infty} := \lim_{n \to \infty} \cup_{j=1}^{n} A_j$. We use $A_n \searrow A$ for a decreasing sequence of sets.*

Property (Measures)**.** *Below are a few common properties.*

1. ***Finite additivity.*** $A \cap B = \emptyset \implies \mu (A \cup B) = \mu (A) + \mu (B)$

   Proof. Take the countable additivity property of measures and let $\{A, B, \emptyset, \emptyset, \ldots\}$, such that $A \cap B = \emptyset$. Then
$$\mu (A \cup B) = \mu (A \cup B \cup \emptyset \cup \cdots) = \mu (A) + \mu (B) + \mu (\emptyset) + \cdots = \mu (A) + \mu (B)$$
□

2. ***Monotonicity***. $A \subseteq B \implies \mu (A) \leq \mu (B)$.

   Proof. First, note that $B \setminus A = \{x \in E : x \in B \wedge x \notin A\} = \{x \in E : x \in B \wedge x \in A^c\} = B \cap A^c$. Note that if $A \subseteq B$, then
$$B = B \cap E = B \cap (A \cup A^c) = (B \cap A) \cup (B \cap A^c) = A \cap B \setminus A$$

and $A \cap B \setminus A$ are by definition disjoint. Thus, $\mu(B) = \mu(A) + \mu\left(B \setminus A\right) \geq \mu(A)$. □

3. ***Sequential continuity**. $A_n \nearrow A$ implies $\mu(A_n) \searrow \mu(A)$ and $A_n \searrow A$ implies $\mu(A_n) \searrow \mu(A)$*

   Proof. Let $B_1 \equiv A_1, B_2 \equiv A_2 \setminus A_1, B_3 \equiv A_3 \setminus A_2, \ldots$, hence $A_n = \cup_{i=1}^n B_i$ and $\{B_i\}_{i=1}^n$ are pairwise disjoint and all $\mathcal{E}-$measurable. Then

$$\begin{aligned} \lim_{n\to\infty} \mu(A_n) &= \lim_{n\to\infty} \left(\mu\left(\cup_{i=1}^n B_i\right)\right) \\ &= \lim_{n\to\infty} \sum_{i=1}^n \mu(B_i) && \text{(finite additivity)} \\ &= \sum_{i=1}^\infty \mu(B_i) = \mu\left(\cup_{i=1}^\infty B_i\right) = \mu(A) && \text{(sigma-additivity)} \end{aligned}$$

   □

4. ***Boole's inequality (Union bound)**. Take any sequence of sets $\{A_n\}_n$, then*

$$\mu\left(\cup_{n=1}^\infty A_n\right) \leq \sum_{n=1}^\infty \mu(A_n)$$

   Proof. For two sets A, B,

$$\begin{aligned} \mu(A \cup B) &= \mu\left(A \cup \left(B \setminus A\right)\right) \\ &= \mu(A) + \mu\left(B \setminus A\right) \leq \mu(A) + \mu(B), \end{aligned}$$

   where the second line follows from finite finite additivity. Next, use induction to show $\mu\left(\cup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mu(A_i)$. Finally take $n \to \infty$. □

5. ***Arithmetic properties**. Let $\mu, v, \mu_j, \forall j \in \mathbb{N}$ be measures. Then*
   - *(a) $c\mu$ is a measure for $c > 0$.*
   - *(b) $\mu + v$ is a measure.*
   - *(c) $\sum_{n=1}^\infty \mu_n$ is a measure.*

Definition (Measure space)**.** *A **measure space** is a triplet $\left(E, \mathcal{E}, \mu\right)$.*

Definition (Probability space)**.** *A **probability space** is a measure space with total measure 1 (i.e. $\mu(E) = 1$) and is often denoted as $(\Omega, \mathcal{F}, \mathbb{P})$.*

Definition (Borel sets)**.** *When E is a topological space and $\mathcal{E} = \mathscr{B}(E)$, then the measurable sets are called* ***Borel sets****.*

Definition (Measurable Rectangles)**.** *Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let $A \subseteq E$ and $B \subseteq F$ be measurable sets. Then*

$$A \times B \equiv \Big\{ (x, y) : x \in A, y \in B \Big\}.$$

If $A \in \mathcal{E}$ and $B \in \mathcal{F}$ (i.e. A,B are measurable), then $A \times B$ is called a ***measurable rectangle****.*

Definition (Product sigma-algebra)**.** *Given two measurable spaces $(E, \mathcal{E})$, $(F, \mathcal{F})$, the product sigma-algebra is the sigma-algebra generated by all measurable rectangles*

$$\mathcal{E} \otimes \mathcal{F} := \sigma\left(\{A \times B : A \in \mathcal{E}, B \in \mathcal{F}\}\right)$$

The measurable space $(E \times F, \mathcal{E} \otimes \mathcal{F})$ is called the ***product sigma-algebra*** *of $(E, \mathcal{E})$ and $(F, \mathcal{F})$.*

Definition (Finite)**.** *A measure μ is* ***finite*** *if $\mu(E) < \infty$.*

Example. *$\mathbb{P}$, $\mathcal{L}$ are finite and infinite respectively.*

Definition (Sigma-finite)**.** *A measure μ is $\sigma-$**finite** if there exists a measurable partition of E, $\{E_n\}_{n \geq 1}$ such that $\mu(E_n) < \infty, \forall\, n$.*

Note that finite $\implies$ σ-finite as E is a partition of itself. In words, a measure is finite if it assigns finite measure to the entire set E, whereas it is sigma-finite if we can segment E in smaller sets to which μ assigns finite measure.

Example. *$\mathcal{L}$ on $\mathbb{R}$ is not finite, as $\mathcal{L}(\mathbb{R})$ is infinity. However, it is sigma-finite. Take $E_n = [-n, n)$, so $\mu(E_n) = 2n < \infty\, \forall\, n$ and $\{E_n\}_{n \in \mathbb{N}}$ is a partition of $\mathbb{R}$.*

Theorem. *Let $(E, \mathcal{E})$ be a measurable space. Let μ, ν be two measures on $(E, \mathcal{E})$ satisfying $\mu(E) = \nu(E) < \infty$. If μ and ν agree on a collection of subsets which is closed under finite intersections and generates $\mathcal{E}$, then μ, ν are identical.*

Definition (CDF)**.** *The* ***CDF*** *of a probability measure on $\mathbb{R}$ is $F(x) := \mu\big((-\infty, x]\big)$.*

Corollary. *Two probability distributions on $\mathbb{R}$ are the same if their CDFs are the same.*

This corollary is a deep result. First, it relies on the fact that $\mathcal{C} = \big\{(-\infty, x] : x \in \mathbb{R}\big\}$ generates $\mathscr{B}_{\mathbb{R}}$ and it is closed under finite intersections. Second, it uses the fact that two probability measures assign measure 1 to the sample space, hence $1 = \mu(\mathbb{R}) = F_X(\mathbb{R}) = F_Y(\mathbb{R}) = \nu(\mathbb{R}) = 1$. Third, it relies on the theorem above saying that if the two measures μ and ν agree on $\mathcal{C}$ (rather than on the entire $\mathscr{B}_{\mathbb{R}}$) they must be equal. Finally, by applying the definition of our probability measures, we have that

$$\mu(A) = \nu(A), \forall A \in \mathcal{F} \iff F_X(x) = F_Y(x), \forall x \in \mathbb{R}$$

We can see that it is much more practical to check the right equality rather than the one on the left, but they are equivalent!

Definition (Atom, diffuse, purely atomic)**.** *Let $(E, \mathcal{E}, \mu)$ is a measure space. Assume $\{x\} \in \mathcal{E}, \forall x \in E$.*

- *A point $x \in E$ is an **atom** if $\mu(\{x\}) > 0$.*
- *A measure is **diffuse** if it has no atoms.*
- *A measure is **purely atomic** if the set of atoms D is countable and $\mu(E \setminus D) = 0$.*

Example. *A Dirac measure is purely atomic. A Lebesgue measure is diffuse.*

Theorem. *Let μ be a sigma-finite measure on $(E, \mathcal{E})$. Then we can write $\mu = \lambda + \nu$, where λ is diffuse and ν is purely atomic.*

Proof. Let μ be a sigma-finite measure on a measurable space $(E, \mathcal{E})$. Since μ is sigma-finite we know that E can contain at most countably many atoms. Denote the union of all atoms in E as A, and call $B = E \setminus A$. Note that A is countable. Let $\mu_x \equiv \mu(\{x\}), x \in A$. Now, take $S \subseteq E$, then

$$\mu(S) = \mu\Big((S \cap A) \cup (S \cap B \setminus A)\Big) = \mu(S \cap A) + \mu(S \cap B \setminus A)$$

We can see that for every set S, we can decompose μ in the trace of μ on A and the trace of μ on $B \setminus A$. The set A contains countably many atoms, thus $\mu(S \cap A) = \sum_{x \in A} \mu_x \delta_x(S)$ which is a purely atomic measure. By construction the trace of μ on $B \setminus A$ does not contain atoms, hence we decomposed μ in an atomic part $\mu(S \cap A)$ and a diffuse part $\mu(S \cap B \setminus A)$. □

Definition (Complete and negligible sets)**.** *Let $(E, \mathcal{E}, \mu)$ be a measure space.*

- *A measurable set B is **negligible** if $\mu(B) = 0$.*
- *An arbitrary subset $B \subseteq E$ is **negligible** if it is contained in a measurable set that is negligible, i.e.*

$$\exists A \in \mathcal{E} : B \subseteq A, \mu(A) = 0$$

- *A measurable space is **complete** if every negligible set is measurable.*

The first two points allow us to characterize negligible sets if they are measurable and if they are not, respectively.

Note. *If a measure space is not complete, you can enlarge to get a complete measure. Roughly, let $\mathcal{N}$ be the collection of all negligible sets of E. Then, $\overline{\mathcal{E}} := \sigma(\mathcal{E} \cup \mathcal{N})$ is called the **completion** of $(E, \mathcal{E}, \mu)$. Every $A \in \overline{\mathcal{E}}$ can be written as $A = B \cup N$ where $B \in \mathcal{E}, N \in \mathcal{N}$. Then, $\overline{\mu}(A) := \mu(B)$.*

In words, we are just adding the negligible sets to $\mathcal{E}$ to get its completion $\overline{\mathcal{E}} := \sigma(\mathcal{E} \cup \mathcal{N})$.

Definition (Lebesgue measurable sets)**.** *$E = \mathbb{R}, \mathcal{E} = \mathscr{B}_{\mathbb{R}}, \mu = Leb$. Elements of $\overline{\mathcal{E}}$ are called **Lebesgue measurable sets**.*

The Lebesgue measure on the Borel sigma-algebra is not complete, meaning that there are Borel sets of Lebesgue measure zero which contain subsets that are not Borel sets (eg. Vitali sets). However, it is complete in the completion (trivially).

Definition (Almost everywhere)**.** *If a statement holds for all x except for a negligible set of x in E, then we say that it holds **almost everywhere** (e.g. holds for $\mu-$a.e. x). Furthermore, if the underlying measure is a probability measure, we say **almost surely**. In general, an a.e. statement is about $\forall\, x \in E \setminus N$, where N is the union of negligible sets.*

1.3 Measurable functions

Example (Motivating example)**.** *Let $\Omega = \{1, 2, \ldots, 100\}, \mathcal{F} = 2^{\Omega}, \mathbb{P}$ uniform. Outcome ω. Let $X(\omega) := \omega \mod 5$ where $X : \Omega \to \{0, 1, 2, 3, 4\}$. We care about*

$$\mathbb{P}_X(X = 3) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) = 3\}\right) = \mathbb{P}\left(X^{-1}(\{3\})\right).$$

We can see that we need measurability of X^{-1}, otherwise we cannot pull back to Ω from $\{0, 1, 2, 3, 4\}$, which contains the objects we ultimately care about and to which we assigned a probability.

Definition (Function)**.** *A function f from E into F is a rule that assigns an element $f(x)$ of F to each x in E.*

We often roughly define functions as rules that assign a unique element in F to each element of E. Note that the same element in F can be assigned to different elements of E but not the viceversa.

Definition (Functions and inverse images)**.** *Let E, F be two sets and define a* ***function*** *$f : E \to F, x \in E \mapsto f(x) \in F$. The* ***inverse image of*** *B for $B \subseteq F$ is*

$$f^{-1}(B) = \left\{x \in E : f(x) \in B\right\}.$$

Properties (Functions)**.** *The following properties can be easily shown.*

1. $f^{-1}(\emptyset) = \emptyset$,
2. $f^{-1}(F) = E$,
3. $f^{-1}\left(B \setminus C\right) = f^{-1}(B) \setminus f^{-1}(C)$,
4. $f^{-1}(\cup_i B_i) = \cup_i f^{-1}(B_i)$,
5. $f^{-1}(\cap_i B_i) = \cap_i f^{-1}(B_i)$.

Proof. Follow from basic properties of interesection and union of sets. □

Definition (Measurable functions)**.** *Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces. A mapping $f : E \to F$ is* ***measurable relative to $\mathcal{E}$ and $\mathcal{F}$*** *if for every $B \in \mathcal{F}, f^{-1}(B) \in \mathcal{E}$.*

Proposition. *Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let $f : E \to F$. The mapping f is measurable relative to $\mathcal{E}$ and $\mathcal{F}$ if and only if there exists a collection of subsets $\mathcal{F}_0$ that generates $\mathcal{F}$ such that $f^{-1}(B) \in \mathcal{E}$ for every $B \in \mathcal{F}_0$.*

Proof. Suppose that f is measurable relative to $\mathcal{E}$ and $\mathcal{F}$, then $0 \subset \mathcal{E}$, so it follows that such that $f^{-1}(B) \in \mathcal{E}$ for every $B \in \mathcal{F}_0 \subset \mathcal{E}$.

Suppose there exists a collection of subsets $\mathcal{F}_0 : \mathcal{F} = \sigma(\mathcal{F}_0)$ and that $f^{-1}(B) \in \mathcal{E}$ for every $B \in \mathcal{F}_0$. Let $\mathcal{F}_1 := \left\{B \in \mathcal{F} : f^{-1}(B) \in \mathcal{E}\right\} \subset \mathcal{F}$. By assumption we have that $\mathcal{F}_0 \subset \mathcal{F}_1$. By the properties 1, 2, and 4 we have that $\mathcal{F}_1$ is a sigma-algebra. Note that for any sigma-algebra $\mathcal{A}$, we have that $\mathcal{A} = \sigma(\mathcal{A})$ as any sigma-algebra is the smallest sigma-algebra containing itself. Thus, $\mathcal{F} = \sigma(\mathcal{F}_0) \subset \sigma(\mathcal{F}_1) = \mathcal{F}_1$ which shows that $\mathcal{F} = \mathcal{F}_1$. □

Note (Checking measurability on a generating set)**.** *Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. A function $f : E \to F$ is measurable relative to $\mathcal{E}$ and $\mathcal{F}$ if and only if $\exists \mathcal{F}_0 \subseteq \mathcal{F} : \mathcal{F} = \sigma(\mathcal{F}_0)$ and $\forall B \in \mathcal{F}_0 : f^{-1}(B) \in \mathcal{E}$.*

In words, we just need to check measurability of the collection of sets that generates the sigma-algebra on $\mathcal{F}$, the codomain. We need to verify that we can pullback to $\mathcal{E}$ from those specific subsets of F that generate $\mathcal{F}$.

Definition (Composition of functions). *Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$, $(G, \mathcal{G})$ be measurable spaces. $f : E \to F, g : F \to G$. The* ***composition of f and g*** *is mapping $g \circ f$ from E to G is defined as*

$$(g \circ f)(x) \equiv g(f(x)), \forall x \in G.$$

Definition. *Let E be a set*

- *$f : E \to \mathbb{R}$ is a* ***real-valued function***
- *$f : E \to \overline{R}$ is a* ***numerical function***
- *$f : E \to \overline{R}_+$ is a* ***positive function***

Proposition. *If f is measurable relative to $\mathcal{E}$ and $\mathcal{F}$ and g is measurable relative to $\mathcal{F}$ and $\mathcal{G}$, then $g \circ f$ is measurable relative to $\mathcal{E}$ and $\mathcal{G}$.*

Proof. By definition, $\forall y \in \mathcal{G}, g^{-1}(y) \in \mathcal{F}$ and $\forall x \in \mathcal{F}, f^{-1}(x) \in \mathcal{E}$. Then, consider $(g \circ f)^{-1}(y) = f^{-1}\left(g^{-1}(y)\right), y \in \mathcal{G}$. It follows from the definitions above that $x = g^{-1}(y) \in \mathcal{F}$ and $f^{-1}(x) = f^{-1}\left(g^{-1}(y)\right)$ is in $\mathcal{E}$, which proves that $(g \circ f)$ is $\mathcal{E}$ and $\mathcal{G}$ measurable. □

Definition ($\mathcal{E}$-measurable). *A real-valued function $f : E \to \mathbb{R}$ is $\mathcal{E}-$measurable if it is measurable relative to $\mathcal{E}$ and $\mathscr{B}_{\mathbb{R}}$.*

Definition (Borel functions). *If E is a topological space and $\mathcal{E} = \mathscr{B}_{\mathbb{R}}$, then the $\mathcal{E}$-measurable functions are Borel functions.*

Proposition. *If $f : E \to \mathbb{R}$ is $\mathcal{E}-$measurable if and only if*

$$f^{-1}((-\infty, x]) \in \mathcal{E}, \forall x \in \mathbb{R}$$

Proof. The only if part is straightforward as $\mathcal{C} = \{(-\infty, x] : x \in \mathbb{R}\} \subset \mathscr{B}_{\mathbb{R}}$.

The if part follows from the fact that $\mathscr{B}_{\mathbb{R}}$ can be generated by any type of intervals on the real line and the fact that to check measurability on $\mathscr{B}_{\mathbb{R}}$ we just need to check it on the collection of sets that generates the Borel sigma-algebra. □

Definition (Positive and negative parts of functions). *Let $a \vee b := \max\{a, b\}; a \wedge b := \min\{a, b\}$, $f : E \to \mathbb{R}$. Then,*

- *$f^+ := f \vee 0 = \max\{f, 0\}$ is the* ***positive part of f***
- *$f^- := -(f \wedge 0) = -\min\{f, 0\}$ is the* ***negative part of f.***

Note that both functions are positive and $f = f^+ - f^-$.

Proposition. *f is $\mathcal{E}-$measurable $\iff$ f^+ and f^- are both $\mathcal{E}-$measurable.*

Proof. The if part is immediate. The only if part is a natural corollary of the fact that the sum of two measurable functions is measurable.

□

Note. *It suffices to focus on positive functions because any function can be decomposed in the sum of two positive functions, i.e. its positive part f^+ and its negative part f^-.*

Definition (Indicator function). *Let $A \subseteq E$. The **indicator function** of A is*

$$\mathbb{1}_A(x) := \begin{cases} 1, \text{ if } x \in A \\ 0, \text{ if } x \notin A \end{cases}$$

Proposition. *$\mathbb{1}_A$ is $\mathcal{E}$-measurable $\iff$ $A \in \mathcal{E}$.*

Definition (Simple function). *A function $f : E \to \mathbb{B}$ is **simple** if we can write it as*

$$f = \sum_{i=1}^{N} a_i \mathbb{1}_{A_i}$$

where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{E}, \forall\, i = 1, \ldots, N$. Intuitively, it is piece-wise constant with finitely many pieces, and is not unique.
*The (unique) **canonical form** for every simple function is*

$$f = \sum_{i=1}^{m} b_i \mathbb{1}_{B_i}$$

where $\{B\}_{i=1}^{m}$ forms a measurable partition of E and m is minimal (otherwise it is not canonical).
Note that **minimality** of m comes from the fact that $\{B_i\}_{i=1}^{m}$ is a partition of E.

Properties (Simple functions). *A few properties:*

1. *Every simple function is $\mathcal{E}-$measurable (sum of indicator functions)*
2. *If f is $\mathcal{E}$-measurable and takes finitely many values which are real then f is simple.*
3. *If f, g simple, then $f + g, f - g, f/g, f \vee g, f \wedge g$ are simple. For division, g must be nowhere 0.*

A few definitions and clarifications for notation.

Definition (Limits of functions)**.** *Let $\{f_n\}_{n\geq 1}$ be a sequence of real-valued/numerical functions on E.*

- *Define*
$$\liminf_{n\to\infty} f_n = \lim_{n\to\infty}\inf_{m\geq n} f_m, \qquad \limsup_{n\to\infty} f_n = \lim_{n\to\infty}\sup_{m\geq n} f_m$$
- *The functions*
$$\inf_{n\geq 1} f_n \quad \sup_{n\geq 1} f_n \quad \liminf_{n\to\infty} f_n \quad \limsup_{n\to\infty} f_n$$
are all defined pointwise in the sense that $\forall\, x \in E$ we consider the set $\{f_n(x)\}_{n\geq 1}$ and we have that all those operations are well defined.
- *If $\liminf_{n\to\infty} f_n = \limsup_{n\to\infty} f_n$, then we say $\{f_n\}_{n\geq 1}$ has a* ***pointwise limit*** *and we write $\lim_{n\to\infty} f = f; f_n \xrightarrow{n\to\infty} f$. We say that f_n converges pointwise to f if*
$$\lim_{n\to\infty} f_n(x) = f(x), \forall\, x \in E$$
- *If $\{f_n\}_{n\geq 1}$ is increasing pointwise, then $\lim_{n\to} f_n$ exists (could be ∞) and equals $\sup_{n\geq 1} f_n$. Write $f_n \nearrow f$. A similar notion can be defined for decreasing.*

Theorem (Closure of the space of measurable function with respect to taking limits)**.** *Let $\{f_n\}_{n\geq 1}$ be a sequence of $\mathcal{E}-$measurable numerical functions. The following functions are all $\mathcal{E}-$measurable:*

1. $\inf_{n\geq 1} f_n$
2. $\sup_{n\geq 1} f_n$
3. $\liminf_{n\to\infty} f_n$
4. $\limsup_{n\to\infty} f_n$.

Furthermore, if $\lim_{n\to\infty}$ exists, then it is also $\mathcal{E}$-measurable.

Proof. Suppose $\{f_n\}_n$ is a sequence of $\mathcal{E}-$measurable functions. We want to show that $f := \sup_{n\geq 1} f_n$ is $\mathcal{E}$-measurable, or that $\forall A \in \mathcal{B}_{\overline{\mathbb{R}}}$, $f^{-1}(A) \in \mathcal{E}$. It suffices to show that this property for a collection of subsets that generates $\mathscr{B}_{\mathbb{R}}$. Consider the collection of sets $\{[-\infty, r : r \in \mathbb{R}]\}$. Then

$$\begin{aligned} f^{-1}([\infty, r]) &= \{x \in E : f(x) \leq r\} \\ &= \cap_{n=1}^{\infty} \{x \in E : f_n(x) \leq r\}, \quad (f(x) \leq r \iff \forall n \geq 1, f_n(x) \leq r) \\ &= \cap_{n=1}^{\infty} \underbrace{f_n^{-1}([-\infty, r])}_{\in\mathcal{E}} \end{aligned}$$

Then, since each f_n is $\mathcal{E}$-measurable, $f_n^{-1}\left([-\infty, r]\right) \in \mathcal{E}$. The entire last line is in $\mathcal{E}$ since $\mathcal{E}$ is closed under countable intersection.

We can prove 1. simply noticing that $\inf_n f_n = -\sup_n\left(-f_n\right)$.

Regarding 3. and 4., note that

$$\liminf_{n\to\infty} f_n = \sup_m \inf_{n\geq m} f_n, \quad \limsup_{n\to\infty} f_n = \inf_m \sup_{n\geq m} f_n$$

and the composition of measurable functions is measurable.

As regards the last point, if the limit of f_n exists, then it must be that

$$\limsup_{n\to\infty} f_n = \liminf_{n\to\infty} f_n = \lim_{n\to\infty} f_n$$

implying that the limit is also measurable.

□

Note. *Simple functions have nice properties, yet not all functions are simple. Fortunately, we can approximate measurable functions with simple functions.*

Definition (Identity function). *The **identity function** $f : \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}_+, x \mapsto f(x)$.*

We want to approximate the identity function as as a sum of simple functions.

Lemma. *For each $n \in \mathbb{N}$, define*

$$d_n(r) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbb{1}\left[r \in \left[\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right] + n\mathbb{1}\left[r \geq n\right].$$

Each d_n is an increasing simple function and $d_n(r) \to r, \forall r \in \overline{\mathbb{R}}_+$.

Note. *We have the following*

- *each d_n is a simple function $\forall\, n \geq 1$,*
- *$d_n \nearrow f$ pointwise, i.e. $\forall\, x \in \mathbb{R}, d_n(x) \nearrow f(x)$*

Theorem. *Consider a measure space $(E, \mathcal{E}, \mu)$. A positive function on E is $\mathcal{E}$-measurable if and only if it is the limit of an increasing sequence of positive simple functions.*

Proof. ($\Longleftarrow$) Positive simple functions are measurable, plus measurable functions are closed under limits.

($\implies$) Take a $\mathcal{E}-$measurable function f. Let $f_n := d_n \circ f$. Note that for all n, f_n is measurable (composition of measurable functions), positive (composition of positive functions), simple (takes on finitely many values). Moreover, take any point $x \in E$, it is easy to see that either $f_{n+1}(x) = f_n(x)$ or $f_{n+1}(x) > f_n(x)$) and, by the fact that $d_n(r) \to r$, we have that $\lim_{n\to\infty} d_n\left(f(x)\right) = f(x), \forall\, x \in \overline{\mathbb{R}}_+ \implies f_n \to f$. We conclude that $f_n \nearrow f$. □

The idea of the 'only if' part is that we first take f and then we chop/approximate it using d_n. As $n \to \infty$ this approximation becomes nicer.

Note. *The main idea behind this theorem is that we can approximate decently well any measurable function f in two steps. First, we decompose it in f^+ and f^- which are both positive and measurable. Second, we approximate f^+ and f^- with an increasing sequence of positive simple functions. This will lead us to define the Lebesgue integral.*

Definition (Isomorphism)**.** *Suppose $(E, \mathcal{E}), (F, \mathcal{F})$ are two measurable spaces and suppose $f : E \to F$ is a bijection. We say that f is an **isomorphism** of $(E, \mathcal{E})$ and $(F, \mathcal{F})$ if*

1. *f is a bijection.*
2. *f is measurable relative to $\mathcal{E}$ and $\mathcal{F}$.*
3. *f^{-1} is measurable relative to $\mathcal{F}$ and $\mathcal{E}$.*

Definition. *Two spaces are **isomorphic** if there exists an isomorphism between them.*

Definition (Standard measurable space)**.** *$(E, \mathcal{E})$ is a **standard measurable space** if it is isomorphic to $(F, \mathcal{B}_F)$ for some Borel subset F of $\mathbb{R}$.*

Note. *Most measurable spaces we will encounter will be standard. Deep result: every standard measurable space is isomorphic to one of the following*

- *$[0, 1]$ with the Borel sigma-algebra*
- *$\{1, \ldots, n\}$ with discrete sigma algebra*
- *$\mathbb{N}$ with discrete sigma algebra*

1.4 Integration

We have a measure space $(E, \mathcal{E}, \mu)$ and a $\mathcal{E}-$measurable function $f : E \to \mathbb{R}$. Measurability of f is key here because we want to be able to approximate its positive parts f^+ and f^- with positive simple functions!!

Note. *We need integration because later, we will be taking expected values. Given $(E, \mathcal{E}, \mu)$, $f : E \to \overline{\mathbb{R}}$, want to define integral of f with respect to μ for all reasonable f. The approach will do so using simple functions, and extend by taking limits.*

Definition (Integration notation). *We write $f \in \mathcal{E}$ to denote that f is $\mathcal{E}$-measurable. We write $f \in \mathcal{E}_+$ to denote that f is $\mathcal{E}$-measurable and positive.*
The below notation for integration are equivalent

$$\mu f \equiv \mu\left(f\right) \equiv \int_E \mu\left(dx\right) f\left(x\right) \equiv \int_E f\left(x d\mu\left(x\right)\right) \equiv \int_E f d\mu \equiv \int f d\mu.$$

Definition (Integral). *Below are notions of integration under different cases for function f.*

1. *Let $f \in \mathcal{E}_+$ be a simple function. Write f in its canonical form $f = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$. Then*

$$\mu\left(f\right) := \sum_{i=1}^{n} a_i \mu\left(A_i\right).$$

   Note that the formula remains the same even if f is not in canonical form.

2. *Let $f \in \mathcal{E}_+$. We know that positive functions can be written as the limit of an increasing sequence of positive simple functions. Let $f_n := d_n \circ f$. Define*

$$\mu\left(f\right) := \lim_{n\to\infty} \mu f_n.$$

   Note that $\mu\left(f\right)$ can be ∞.

3. *Let $f \in \mathcal{E}$, then $f = f^+ - f^-$. Define*

$$\mu\left(f\right) := \mu\left(f^+\right) - \mu\left(f^-\right) = \lim_{n\to\infty} \mu\left(d_n \circ f^+\right) - \lim_{n\to\infty} \mu\left(d_n \circ f^-\right)$$

   provided that at least one term on the RHS is finite. Otherwise the integral is not defined because $\mu\left(f^+\right) - \mu\left(f^-\right) = \infty - \infty$.

Note. *It might seem that the definition of μ using d_n is somehow arbitrary. However, the monotone class theorem ensures us that it is not arbitrary.*

Definition (Integrability). *f is **integrable with respect to** μ if $\mu\left(f\right)$ exists and is finite. This is equivalent to say that*

- *f is integrable $\iff \mu\left(f^+\right) < \infty$ and $\mu\left(f^-\right) < \infty$.*
- *f is integrable $\iff \int |f| d\mu < \infty$.*

Property (Integration). *Let $a, b \in \mathbb{R}$, $f, g, f_n \in \mathcal{E}_+$, then*

1. ***Positivity***: *$\mu f \geq 0$. If $\mu f = 0$, then $f = 0$ μ−a.e.*
2. ***Linearity***: *$\mu\left(af + bg\right) = a\mu\left(f\right) + b\mu\left(g\right)$*

   Proof. Immediately follows from definition of simple functions. In general, it follows from MCT. □

3. ***Monotonicity***: *If $f \leq g$, then $\mu\left(f\right) \leq \mu\left(g\right)$.*

   Proof. Just notice that $g - f$ is positive and apply 1) and then 2) to conclude that $\mu\left(g\right) \geq \mu\left(f\right)$. □

Example (Integration)**.** *: Below give a few simple examples in which we vary the measure the integral is defined on*

1. ***Discrete measures***: *For the Dirac measure,*

$$\delta_{x_0}\left(f\right) = \int f d\delta_{x_0} = f\left(x_0\right)$$

   and more in general for any countable set D

$$\mu = \sum_{x\in D} m\left(x\right)\delta_x, \implies \int f d\mu = \sum_{x\in D} m\left(x\right) f\left(x\right)$$

2. ***Discrete spaces***: *Suppose $\left(E, \mathcal{E}\right)$ is discrete, that is E is countable and $\mathcal{E} = 2^E$. Every μ has this form with $D = E$ and $\mu\left(x\right) = \mu\left(\{x\}\right)$. For every $f \in \mathcal{E}$,*

$$\mu f = \sum_{x\in E} \mu\left(\{x\}\right) f\left(x\right).$$

   *The notation is suggestive of an **inner product**.*

3. ***Lebesgue integrals***: *Suppose $E \subseteq \mathbb{R}^d$ is a Borel set, $\mathcal{E} = \mathscr{B}\left(E\right)$. μ is a restriction of Lebesgue measure on $\mathbb{R}^d$ to $\left(E, \mathcal{E}\right)$. Then*

$$\mu f = Leb_E f = \int_E Leb\left(dx\right) f\left(x\right) = \int_E dx f\left(x\right) = \int_E f\left(x\right) dx.$$

Note. *Consider positive functions.* ***If the Riemann integral of f exists, then so does the Lebesgue integral.*** *The two are the same. The converse is false. Indeed, Lebesgue integral exists for a larger class of functions. Consider the following example.*

Example (Valid Lebesgue integral but not Riemann). *Suppose* $E = [0,1], f(x) = \mathbb{1}_{\mathbb{Q}}(x)$. *The Riemann integral does not exist.* $Leb_E f = 0$. *Let* $\mathbb{Q} = \{q_1, q_2, \ldots\}$ *and* $A_n = \{q_1, \ldots, q_n\}$. *Then* $\mathbb{1}_{A_n} \nearrow \mathbb{1}_{\mathbb{Q}}$ *and* $\mathbb{1}_{A_n}$ *is a simple function.* $0 = Leb_E(\mathbb{1}_{A_n}) \nearrow Leb_E(\mathbb{1}_{\mathbb{Q}})$.

We can now merge what we have covered so far. Take a measurable function f and a measurable set A. Then, we can approximate $f = f^+ - f^-$ with $d_n \circ f^+$ and $d_n \circ f^-$, take the limit, and define $\mu(f)$. Moreover, being A a measurable set, we can assign it a measure. Using these two ingredients, we can define integration over a set.

Definition (Integration over a set). *Consider a measure space* $(E, \mathcal{E}, \mu)$. *Let* f *be a measurable function and* $A \in \mathcal{E}$ *be a measurable set. Then,* $f\mathbb{1}_A$ *is a measurable function and the integral of* f *over a is*

$$\mu(f\mathbb{1}_A) = \int_A f d\mu = \int_E f\mathbb{1}_A d\mu$$

1.4.1 Convergence Theorems

One of the most basic questions in integration theory is the following: If $f_n \to f$ pointwise, when can one say that

$$\int f_n d\mu \to \int f d\mu?$$

The Riemann integral is not sufficiently general to permit a satisfactory answer to this question. Perhaps the simplest condition that guarantees the convergence of the integrals is that the functions $f_n : X \to \mathbb{R}$ **converges uniformly** to $f : X \to \mathbb{R}$ and X has finite measure. In that case

$$\left| \int f_n d\mu - \int f d\mu \right| \leq \int |f_\mathrm{n} - f| \, d\mu \leq \mu(X) \sup_X |f_n - f| \to 0$$

as $n \to \infty$. The assumption of uniform convergence is **too strong** for many purposes, and the Lebesgue integral allows the formulation of simple and widely applicable theorems for the convergence of integrals. The most important of these are the monotone convergence theorem and the Lebesgue dominated convergence theorem. The utility of these results accounts, in large part, for the success of the Lebesgue integral.

Some conditions on the functions f_n in the example above are, however, necessary to ensure the convergence of the integrals, as can be seen from very simple examples. Roughly speaking, the convergence may fail because ”mass” can leak out to infinity in the limit.

Example. *Define $f_n : \mathbb{R} \to \mathbb{R}$ by*

$$f_n(x) = \begin{cases} n & \text{if} \quad 0 < x < 1/n \\ 0 & \text{otherwise} \end{cases}$$

Then $f_n \to 0$ as $n \to \infty$ pointwise on $\mathbb{R}$, but

$$\int f_n dx = 1 \quad \text{for every } n \in \mathbb{N}$$

The problem with this function is that it doesn't behave well in the limit, as it jumps back to 0.

Example. *By modifying this example we can obtain a sequence f_n that converges pointwise to zero but whose integrals diverge to infinity.*

$$f_n(x) = \begin{cases} n^2 & \text{if} \quad 0 < x < 1/n \\ 0 & \text{otherwise} \end{cases}$$

Example. *Define $f_n : \mathbb{R} \to \mathbb{R}$ by*

$$f_n(x) = \begin{cases} 1/n & \text{if} \quad 0 < x < n \\ 0 & \text{otherwise} \end{cases}$$

Then $f_n \to 0$ as $n \to \infty$ pointwise on $\mathbb{R}$, and even uniformly, but

$$\int f_n dx = 1 \quad \text{for every } n \in \mathbb{N}$$

In what follows, we will give four results that allow us, under different conditions, to interchange limits and integrals:

1. *Monotone Convergence Theorem* for increasing sequences of positive functions
2. *Fatou's Lemma* for sequences of positive functions
3. *Dominated Convergence Theorem* for dominated sequences of positive functions
4. *Bounded Convergence Theorem* for bounded sequences of positive functions

Theorem (Monotone convergence theorem)**.** *Let $\{f_n\}_{n\geq 1}$ be an increasing sequence in $\mathcal{E}_+$. Then*

$$\lim_{n\to\infty} \int f_n d\mu = \int \lim_{n\to\infty} f_n d\mu$$

Proof. $f := \lim_{n\to\infty} f_n$ is well-defined since $\{f_n\}_{n\geq 1}$ is increasing in n. $f \in \mathcal{E}_+$, so $\mu(f)$ is well-defined. From monotonicity of μ, $f \geq f_n, \forall\, n \in \mathbb{N} \implies \mu(f) \geq \mu(f_n)$. Taking the limit implies $\mu(f) \geq \lim_{n\to\infty} \mu(f_n)$, hence

$$\int \lim_{n\to\infty} f_n d\mu \geq \lim \int f_n d\mu$$

.

For the other direction, we need to use definition of integral. Show

$$\lim_{n\to\infty} \int f_n d\mu \geq \int d_k \circ f d\mu, \quad \forall\, k$$

Then, take a limit as $k \to \infty$ so that the RHS becomes $\mu(f)$. □

Note (Comments on MCT)**.** *This is the main theorem in integration.*

- *Helpful tool for interchanging limits and integration.*
- *MCT states that, if we think about μ as a mapping $f \mapsto \mu f, \mu : \mathcal{E}_+ \to \overline{\mathbb{R}}_+$ and we consider a positive measurable function, then is continuous under increasing limits.*
- *Also implies definition of integral does not depend on the choice of approximation.*
- *The monotone convergence theorem implies that failure of convergence of the integrals cannot occur in an increasing sequence of functions, even if the convergence is not uniform or the domain space does not have finite measure.*
- *MCT does not hold for the Riemann integral*

Fatou's theorem allows you to tell something about interchangeability of limits and integrals when your sequence of measurable positive functions is not necessarily increasing, thus it **might not have a well defined limit**. However, the lim sup and the lim inf do exist!

Lemma (Fatou)**.** *Let $\{f_n\}$ be a sequence in $\mathcal{E}_+$. Then*

$$\mu\left(\liminf_{n\to\infty} f_n\right) \leq \liminf_{n\to\infty} \mu(f_n).$$

Proof. Proof follows from MCT on the positive function $g_n := \inf_{k\geq n} f_k$ and is an exercise. □

It is possible to convert it in terms of lim sup.

Definition (Dominate)**.** *A function f is said to be* ***dominated*** *by the function g if $|f|\leq g$. Then, g is called the* ***dominating*** *function and f the* ***dominated*** *function.*

Definition (Bounded)**.** *If $\{f_n\}_{n\geq 1}$ can be dominated by a constant, then we say $\{f_n\}$ is* ***bounded****. That is $\exists\, c \in \mathbb{R} : |f_n|\leq c, \forall\, n \in \mathbb{N}$.*

Theorem ((Lebesgue) Dominated convergence theorem)**.** *Let $(E, \mathcal{E}, \mu)$ be a measure space and let $\{f_n\}_{n\geq 1}$ be a sequence of measurable real-valued functions. Suppose that there exists an integrable function g such that $|f_n|\leq g$. Assume $f := \lim_{n\to\infty} f_n$ exists. Then f is integrable and*

$$\mu(f) = \mu\left(\lim_{n\to\infty} f_n\right) = \lim_{n\to\infty} \mu(f_n)$$

.

Proof. Follows from MCT. □

We can also derive a corollary which requires a stronger condition than being dominated.

Corollary (Bounded convergence theorem)**.** *Let $(E, \mathcal{E}, \mu)$ be a measure space and let $\{f_n\}_{n\geq 1}$ be a sequence of measurable real-valued functions. Suppose that the measure μ is finite ($\mu(E) < \infty$), $\{f_n\}_{n\geq 1}$ is bounded, and that the limit exists $f \int f d\mu = \lim_{n\to\infty} \int f_n d\mu$. Then, f is integrable*

$$\mu\left(\lim_{n\to\infty} f_n\right) = \lim_{n\to\infty} \mu(f_n).$$

Proposition (Insensitivity of integral wrt changes over a negligible set)**.** *Below give a few simple properties.*

1. *If A is negligible, then $\mu(f\mathbb{1}_A) = 0, \forall f \in \mathcal{E}$.*
2. *If $f, g \in \mathcal{E}_+$ and $f = g$ a.e., then $\mu f = \mu g$.*
3. *If $f \in \mathcal{E}_+$ and $\mu f = 0$, then $f = 0$ a.e.*

Proof. Quick proofs of above statements.

1. True for simple function by definition. Extends to non-simple case by MCT. Then to positive and negative parts.
2. $A := \{x \in E : f(x) \neq g(x)\} = \{f \neq g\}$. By assumption, $\mu(A) = 0$ since $f = g$ $\mu-$a.e., hence A is negligible. Decompose f and g into

$$f = f\mathbb{1}_A + f\mathbb{1}_{A^c}, \quad g = g\mathbb{1}_A + g\mathbb{1}_{A^c}$$

Part (a) implies

$$\mu\left(f\mathbb{1}_A\right)=\mu\left(g\mathbb{1}_A\right)=0$$

. Since $\mu\left(f\right)=\mu\left(f\mathbb{1}_{A^c}\right),\mu\left(g\right)=\mu\left(g\mathbb{1}_{A^c}\right)$ on $A^c, f=g$.

3. Let $A_n:=\left\{x\in E:f\left(x\right)\geq\varepsilon_n\right\},\varepsilon_n\searrow 0$. $A:=\left\{x\in E,f\left(x\right)>0\right\}$. Since $A_n\nearrow A$, $\mu\left(A_n\right)\nearrow\mu\left(A\right)$ from sequential continuity. Then

$$0=\int fd\mu\geq\int_{A_n}fd\mu\geq\int_{A_n}\varepsilon_n d\mu=\varepsilon_n\mu\left(A_n\right)\implies\mu\left(A_n\right)=0,\forall n\geq 1$$

which finally implies $\mu\left(A\right)=0$.

□

Note. *Can relate all results assuming holding almost everywhere.*

Theorem (MCT - Almost Everywhere Version). *Let $\left(E,\mathcal{E},\mu\right)$ be a measure space. LEt $\left\{f_n\right\}_{n\geq 1}$ be a sequence of numerical functions on E. Suppose that*

- *$\forall\, n$ there exists g_n $\mathcal{E}-$measurable such that $g_n=f_n$ a.e.*
- *$f_n\geq 0$ a.e. $\forall\, n\geq 1$*
- *$f_{n+1}\geq f_n$ a.e. $\forall\, n\geq 1$*

Then, $\lim_{n\to\infty}f_n$ exists a.e., is non-negative a.e. and

$$\int\lim_{n\to\infty}f_n d\mu=\lim_{n\to\infty}\int f_n d\mu$$

Definition (Product space). *Let $\left(E,\mathcal{E},\mu\right),\left(F,\mathcal{F},\nu\right)$ be two measurable spaces. Then, we call product space $\left(E\times F,\mathcal{E}\otimes\mathcal{F},\mu\times\nu\right)$. Let $A\in\mathcal{E},B\in\mathcal{F}$, then we define as product measure $\left(\mu\times\nu\right)\left(A\times B\right)=\mu\left(A\right)\cdot\nu\left(B\right)$.*

Note. *Recall that the product sigma-algebra is*

$$\mathcal{E}\otimes\mathcal{F}=\sigma\left(\left\{A\times B:A\in\mathcal{E},B\in\mathcal{F}\right\}\right)$$

It's enough to specify the product measure since measurable rectangles generate the product sigma-algebra and are closed under intersection. Thus it extends uniquely to $\mathcal{E}\otimes\mathcal{F}$.

Theorem (Fubini's Theorem). *Let $\left(E\times F,\mathcal{E}\otimes\mathcal{F},\mu\times\nu\right)$ be a product space and consider $f: E\times F\to\mathbb{R}$, $\mathcal{E}\otimes\mathcal{F}-$measurable. If $\int_{E\times F}|f|d\left(\mu\times\nu\right)<\infty$, then*

$$\int_{E\times F}fd\left(\mu\times\nu\right)=\int_E\int_F fd\mu d\nu=\int_F\left(\int_E fd\mu\right)d\nu=\int_E\left(\int_F fd\nu\right)d\mu$$

***Note**: Order of integration can be interchanged.*

Theorem (Tonelli). *If $f \geq 0$, then the same conclusion holds.*

Definition (Finite product measure spaces). *Let $\left(E_i, \mathcal{E}_i, \mu_I\right), i = 1, \ldots, n$ be n measurable spaces. Then, we call product measure space $\left(\times_{i=1}^{n} E_i, \otimes_{i=1}^{n} \mathcal{E}_i, \times_{i=1}^{n} \mu_i\right)$.*

Definition (Countably infinite product measure spaces). *Let $\left(E_i, \mathcal{E}_i, \mu_i\right), i \in \mathbb{N}$ be measurable spaces. Then, we call product measure space $\left(\times_{i=1}^{\infty} E_i, \otimes_{i=1}^{\infty} \mathcal{E}_i, \times_{i=1}^{\infty} \mu_i\right)$.*

Example. *Let $\mathbb{R}^{\mathbb{N}} = \left\{x_i = \left(x_{1i}, x_{2i}, x_{3i}, \ldots\right), x_i \in \mathbb{R}\right\}$. Equip $\mathbb{R}^{\mathbb{N}}$ with the Borel sigma-algebra $\mathscr{B}_{\mathbb{R}^{\mathbb{N}}} = \sigma\left(\mathcal{R}\right)$ generated by all finite-dimensional measurable rectangles (aka cylinder sets) of the form*

$$\mathcal{R} := \left\{B_1 \times B_2 \times \cdots \times B_n \times \mathbb{R} \times \mathbb{R} \times \cdots, B_i \in \mathscr{B}_{\mathbb{R}}, n \geq 1\right\}$$

We want to build a measure on $\left(\mathbb{R}^{\mathbb{N}}, \mathscr{B}_{\mathbb{R}^{\mathbb{N}}}\right)$.

Theorem (Kolmogorov's extension theorem). *Suppose we are given a sequence of probability measures $\left\{\nu_n\right\}_{n\geq 1}$ where ν_n is a probability measure on $\left(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}\right)$ that is **consistent**, i.e.*

$$\nu_{n+1}\left(B_1 \times B_2 \times \cdots \times B_n \times \mathbb{R}\right) = \nu_n\left(B_1 \times \cdots \times B_n\right), \quad \forall n \geq 1, \forall B_i \in \mathscr{B}_{\mathbb{R}}$$

*Then, there exists a **unique** probability measure $\mathbb{P}$ on $\left(\mathbb{R}^{\mathbb{N}}, \mathscr{B}_{\mathbb{R}^{\mathbb{N}}}\right)$ such that*

$$\mathbb{P}\left(\left\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_1 \in B_1, \ldots, \omega_n \in B_n\right\}\right) = \nu_n\left(B_1 \times \cdots \times B_n\right), \quad \forall n \geq 1, \forall B_i \in \mathscr{B}_{\mathbb{R}}$$

Note. *A few facts*

- *Kolmogorov's theorem is a result about **existence** and **uniqueness***
- *Note that $\mathbb{P}$ is defined on $\mathbb{R}^{\mathbb{N}}$ but we just need to evaluate it on a finite number of elements n.*
- *In particular if $\mu_1, \mu_2, \mu_3, \ldots$ are probability measure on $\left(\mathbb{R}, \mathscr{B}_{\mathbb{R}}\right)$ and $\nu_n := \times_{i=1}^{n} \mu_i$, then $\left\{\nu_n\right\}_{n\geq 1}$ is consistent and by Kolmogorov's extension theorem there is a unique product measure on $\left(\mathbb{R}^{\mathbb{N}}, \mathscr{B}_{\mathbb{R}^{\mathbb{N}}}\right)$.*
- *Sometimes rather than indexing the sequences with $\mathbb{N}$ but with $\mathbb{R}$ (stochastic processes in continuous time). This theorem still applies.*

2 Asymptotics and the Law of Large Numbers

2.1 Probability and Measure Theory

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where Ω is the **set of possible outcomes**, $\mathcal{F}$ is the **sigma-algebra consisting of events**, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.

We care about random variables, which are $\mathcal{F}-$measurable maps $X : \Omega \to \overline{\mathbb{R}}$ and we might be interested in knowing the probability of an event $A \in \mathscr{B}_{\mathbb{R}}$, i.e.

$$\mu(A) := \mathbb{P}(X \in A) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right) = \mathbb{P}\left(X^{-1}(A)\right)$$

and μ is a probability measure on $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ that is called the **distribution** of X.

This last sentence is worth more attention as we are defining a measure via another measure.

Proposition. *Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces, ν be a measure on $(F, \mathcal{F})$, and $h : F \to E$ be a measurable function. Then, $\nu \circ h^{-1}$ is a measure on $(E, \mathcal{E})$ and it is called the image of ν under h.*

Proof. In order for $\nu \circ h^{-1}$ to be a measure we need to check three properties:

- $\left(\nu \circ h^{-1}\right)(\emptyset) = \nu\left(h^{-1}(\emptyset)\right) = \nu(\emptyset) = 0$. This follows from the fact that h is measurable, thus h^{-1} maps back in $\mathcal{F}$ and from the fact that the pre-image of the empty set is the empty set
- for $A \in \mathcal{E}$, $\left(\nu \circ h^{-1}\right)(A) = \nu\left(h^{-1}(A)\right) = \nu(B) \geq 0$ as $B \in \mathcal{F}$ because h is measurable
- $\left(\nu \circ h^{-1}\right)\left(\cup_{i=1}^{\infty} A_i\right) = \nu\left(h^{-1}\left(\cup_{i=1}^{\infty} A_i\right)\right) = \nu\left(\cup_{i=1}^{\infty} h^{-1}(A_i)\right) = \sum_{i=1}^{\infty} \nu\left(h^{-1}(A_i)\right)$

□

Now, consider $(F, \mathcal{F}) \equiv (\Omega, \mathcal{H})$, $(E, \mathcal{E}) \equiv (\mathbb{R}, \mathscr{B}_{\mathbb{R}})$, $\nu \equiv \mathbb{P}$, and $h \equiv X$. We can see that when we define our background probability space $(\Omega, \mathcal{H}, \mathbb{P})$ and a random variable X, this gives us another measure space in which we can use mathematics and visualize things. More explicitly, we can jump back and forth from the background probability space to the forwards probability space $(\mathbb{R}, \mathscr{B}_{\mathbb{R}}, \mathbb{P}_X)$ as follows

$$\mathbb{P}(X \in A) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right) = \mathbb{P}\left(X^{-1}(A)\right) = \left(\mathbb{P} \circ X^{-1}\right)(A) = \mathbb{P}_X(A)$$

for $A \in \mathcal{H}$. Hence, the distribution $\mathbb{P}_X$ of a random variable X is the image of $\mathbb{P}$ through 'the lenses' of X, i.e. it is the transformed (through the measurable map X) version of

$\mathbb{P}$. Note that the **distribution** induces a **distribution function**, which is a distinct object. Indeed, the distribution function is defined as

$$c\left(x\right)=\mathbb{P}_X\left(\left(-\infty,x\right]\right)=\mathbb{P}\left(X\leq x\right),\quad x\in\mathbb{R}$$

If we apply a (measurable) transformation, say $f:E\rightarrow\mathcal{F}$, to our random variable X, then $Y\left(\omega\right)=f\left(X\left(\omega\right)\right)$ is going to be a random variable and $\mathbb{P}_Y$ is going to be a measure on $\left(F,\mathcal{F}\right)$, indeed

$$\mathbb{P}\left(Y\in B\right)=\mathbb{P}\left(X\in f^{-1}\left(B\right)\right)=\left(\mathbb{P}_X\circ f^{-1}\right)\left(B\right)=\mathbb{P}_Y\left(B\right),\quad B\in\mathcal{F}$$

We also care about **expectations**. Say we have a random variable $X:\Omega\rightarrow\mathbb{R}$ and a function $f:\mathbb{R}\rightarrow\mathbb{R}$, then the expectation of X is the integral of X with respect to $\mathbb{P}$, i.e.

$$\mathbb{E}\left[f\left(X\right)\right]=\int_\Omega f\left(X\left(\omega\right)\right)d\mathbb{P}\left(\omega\right)=\int_{\mathbb{R}}f\left(x\right)d\mu\left(x\right)$$

note the **change of variables** that occurs in the second equality. It exists as long as X is positive, bounded, or, more generally, integrable. Integrability usually requires $\mu\left(|f|\right)<\infty$, with random variables we denote it with $\mathbb{E}\left[|X|\right]<\infty$. Since expectations are integrals, all the results we saw hold with expectations as well, thus

Proposition. *The following are true:*

- *Positivity:* $X\geq 0\implies\mathbb{E}\left[X\right]\geq 0$
- *Monotonicity:* $X\geq Y\implies\mathbb{E}\left[X\right]\geq\mathbb{E}\left[Y\right]$
- *Linearity*
- *Insensitivity to negligible sets:* $X,Y\geq 0,X=Y\ a.s\implies\mathbb{E}\left[X\right]=\mathbb{E}\left[Y\right]$
- *Monotone convergence theorem:* $X_n\geq 0,X_n\nearrow X\implies\mathbb{E}\left[X_n\right]\nearrow\mathbb{E}\left[X\right]$
- *Fatou's lemma:* $X_n\geq 0\implies\mathbb{E}\left[\liminf X_n\right]\leq\liminf\mathbb{E}\left[X_n\right]$
- *Dominated convergence:* $|X_n|\leq Y,\mathbb{E}\left[|Y|\right]<\infty$, $\lim X_n$ *exists, then*

$$\lim_{n\rightarrow\infty}\mathbb{E}\left[X_n\right]=\mathbb{E}\left[\lim_{n\rightarrow\infty}X_n\right]$$

Definition (Sigma-algebras generated by random variables)**.** *Let $\left(F,\mathcal{F}\right)$ be a measurable space and $X:\Omega\rightarrow F$ be a random variable. Then, the sigma-algebra generated by X is defined as*

$$\sigma\left(X\right)=\left\{B\subseteq\Omega:B=X^{-1}\left(A\right),A\in\mathcal{F}\right\}$$

Note. *The sigma-algebra generated by a RV is the pre-image of measurable sets of F through X. It should be noted that it is by construction the smallest sigma-algebra $\mathcal{G}$ on Ω such that X is measurable with respect to $\mathcal{G}$.*

In general, if a random variable X is measurable with respect to a sigma-algebra $\mathcal{E}$ it will be measurable to any other sigma-algebra $\mathcal{F} \supseteq \mathcal{E}$. Intuitively, $\mathcal{F}$ is a collection of sets that contains the collection of sets $\mathcal{E}$ and something more, thus if $\mathcal{E}$ was already "precise" enough to measure X, the same must hold true for $\mathcal{F}$. This does hold, in general, for sigma-algebras such that $\mathcal{G} \subseteq \mathcal{E}$ because they are too coarse. In this sense, $\sigma(X)$ is a threshold above which it is still possible to measure X and below it is impossible.

Another important thing to note is that the sigma-algebra generated by a random variable is the mathematical equivalent of the concept of ***information****. Indeed, $\sigma(X)$ is the sigma-algebra generated by a particular collection of sets, i.e. those sets that partition Ω through the pre-image of X. It follows that if a random variable Y is measurable with respect to $\sigma(X)$. In other words, Y cannot contain more information than X, because every time that we pull back onto Ω using Y we land in an element of $\sigma(X)$, hence the partition of Ω induced by Y is by no means finer than the one induced by X. The next theorem formalizes this idea.*

Theorem. *Let X be a random variable taking values in some measurable space $(E, \mathcal{E})$. A mapping $V : \Omega \to \overline{\mathbb{R}}$ is measurable with respect to $\sigma(X)$ if and only if $V = f \circ X$ for some deterministic function $f \in \mathcal{E}$.*

Since we are dealing with random objects, we want to define an appropriate notion of convergence.

Definition (Convergence in probability)**.** *X_n* ***converges in probability*** *to m if*

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}\left(|X_n - m| \geq \varepsilon\right) = \lim_{n\to\infty} \mathbb{P}\left(\left\{\omega \in \Omega : |X_n(\omega) - m| \geq \varepsilon\right\}\right) = 0$$

It is often noted as $i.p.$ or $X_n \xrightarrow{\mathbb{P}} m$.

Note that once we fix an ε, the statement becomes deterministic and we can use standard results from limit theory.

We will now see some useful inequalities to prove convergence in proability.

Theorem (Markov's inequality)**.** *Let X be a non-negative RV and let $\lambda > 0$. Then*

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}(X)}{\lambda}$$

Proof. $\mathbb{P}(X \geq \lambda) = \mathbb{E}(\mathbb{1}(X \geq \lambda)) \leq \mathbb{E}\left(\frac{X}{\lambda}\right) = \frac{\mathbb{E}(X)}{\lambda}$. Alternatively, you could argue $\mathbb{E}(X) = \int_0^\infty xp(x)\,dx \geq \int_\lambda^\infty xp(x)\,dx \geq \lambda \int_\lambda^\infty p(x)\,dx = \lambda\mathbb{P}(X \geq \lambda)$ □

Note. *Note that $\mathbb{E}(\mathbb{1}(X \geq \lambda)) \leq \mathbb{E}\left(\frac{X}{\lambda}\right)$ comes from the fact that the indicator function is always bounded by the linear function X/λ.*

Theorem (Chebyshev's inequality)**.** *Let X be a real-valued RV with $\mathbb{E}\left(X^2\right) < \infty$ and $\lambda > 0$. Then*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \lambda) \leq \frac{\mathbb{V}(X)}{\lambda^2}$$

Proof. It follows naturally from Markov's inequality by applying it to the square of the deviation

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \lambda) = \mathbb{P}\left(|[X - \mathbb{E}(X)]^2| \geq \lambda^2\right) \leq \frac{\mathbb{V}(X)}{\lambda^2}$$

where the first equality follows from $\lambda > 0$. □

Note. *This is another way to bound the probability of an event, but rather than using a linear bound it uses a quadratic bound. Which one is better depends on the distribution of X, because the quadratic is better as long as $X < \lambda$, whereas the the linear is better if $X > \lambda$.*

Theorem (General Markov)**.** *Assume $f : \mathbb{R} \to \mathbb{R}_+$ is non-decreasing. Let X be a non-negative RV and let $\lambda > 0$. Then*

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}(f(X))}{f(\lambda)}.$$

Proof. Directly from Markov inequality

$$\mathbb{P}(X \geq \lambda) \leq \mathbb{P}(f(X) \geq f(\lambda)) \leq \frac{\mathbb{E}(f(x))}{f(\lambda)}$$

where the first inequality follows from the fact that $f(\cdot)$ is non-decreasing. □

Note. *Again, we use the same intuition that the bounding function must be above the indicator. In this case we use $f(x)/f(\lambda)$. That's why we care about f being increasing.*

Theorem (Chernoff Bound)**.** *Take $f(x) = \exp(tx)$ where $t > 0$. Then optimize bound over $t > 0$. Let $X_1, \ldots, X_n$ be mutually independent Bernoulli RVs. $p_i := \mathbb{E}(X_i)$, $S_n := \sum_{i=1}^n X_i$ and*

$\mu = \sum_i p_i$. *Then* $\forall\, \lambda > 0$

$$\mathbb{P}\left(S_n \geq \mu + \lambda\right) \leq \exp\left(-\frac{2\lambda^2}{n}\right)$$
$$\mathbb{P}\left(S_n \geq \mu - \lambda\right) \leq \exp\left(-\frac{2\lambda^2}{n}\right)$$

Note. *This bound is frequently used.*

1. *There are many other forms of this inequality, which are generally called **concentration inequalities***
2. *The proof is based on using the approximating function, optimizing wrt t and using some of the inequalities above.*
3. *Bound decays like Gaussian. Can't do much better in many situations. The non-improvability follows from CLT*
4. *This is a finite sample bound, unlike the CLT which is an asymptotic statement*

Example. *Chebyshev vs Chernoff. Let* $X_1, \ldots, X_n$ *be Bernoulli RVs.*

$$\mathbb{P}\left(|S_n - n/2| \geq \varepsilon n\right) \leq \frac{1}{4}\varepsilon^2 n \qquad \text{(Chebyshev)}$$
$$\mathbb{P}\left(|S_n - n/2| \geq \varepsilon n\right) \leq 2e^{-2\varepsilon^2 n} \qquad \text{(Chernoff)}$$

Take $n = 100$ *and* $\varepsilon = 0.2$, *then*

$$\mathbb{P}\left(|S_{100} - 50| \geq 20\right) \leq 0.00625 \qquad \text{(Chebyshev)}$$
$$\mathbb{P}\left(|S_{100} - 50| \geq 20\right) \leq 0.00067 \qquad \text{(Chernoff)}$$

Theorem (Generic Chernoff Bound). *The generic Chernoff bound for a random variable X is attained by applying Markov's inequality to* e^{tX} *For every* $t > 0$ *:*

$$\mathbb{P}\left(X \geq a\right) = \mathbb{P}\left(e^{t \cdot X} \geq e^{t \cdot a}\right) \leq \frac{\mathrm{E}\left[e^{t \cdot X}\right]}{e^{t \cdot a}}$$

Note. *When X is the sum of n random variables* $X_1, \ldots, X_n$, *we get for any* $t > 0$

$$\mathbb{P}\left(X \geq a\right) \leq e^{-ta}\mathrm{E}\left[\prod_i e^{t \cdot X_i}\right]$$

In particular, optimizing over t and assuming that X_i are independent, we obtain,

$$\mathbb{P}\left(X \geq a\right) \leq \inf_{t>0} e^{-ta} \prod_i \mathrm{E}\left[e^{tX_i}\right]$$

Similarly,

$$\mathbb{P}\left(X \leq a\right) = \mathbb{P}\left(e^{-tX} \geq e^{-ta}\right)$$

and so,

$$\mathbb{P}\left(X \leq a\right) \leq \inf_{t>0} e^{ta} \prod_i \mathrm{E}\left[e^{-tX_i}\right]$$

Specific Chernoff bounds are attained by calculating $\mathrm{E}\left[e^{-t\cdot X_i}\right]$ *for specific instances of the basic variables X_i.*

Gaussian case: *Let $X \sim N\left(\mu, \sigma^2\right)$. Using the moment-generating function for a Gaussian random variable, we conclude for all $t \geq 0$*

$$\mathbb{P}\left[X \geq \mu + t\right] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Theorem (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be an i.i.d. sequence of RVs. Assume $\mathbb{E}\left[|X_i|\right] < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then*

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}\left[X_i\right].$$

Proof. This is a first proof that relies on $\mathbb{E}\left[X_i^2\right] < \infty$ (stronger condition than the one stated above). Fix $\varepsilon > 0$. Note $\mathbb{E}\left(S_n/n\right) = \frac{1}{n}\sum_i \mathbb{E}\left(X_i\right) = m$.

$$\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right) = \mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}\left(\frac{S_n}{n}\right)\right| \geq \varepsilon\right) \leq \frac{\mathbb{V}\left(S_n/n\right)}{\varepsilon^2}.$$

Computing the variance,

$$\mathbb{V}\left(S_n/n\right) = \frac{1}{n^2}\mathbb{V}\left(S_n\right) = \frac{\sigma^2}{n}.$$

Substituting the variance into the original expression, we can take the limit as $n \to \infty$. □

Note. *A few observations about the proof strategy and assumptions.*

1. *We used independence, but same proof works for negatively correlated RVs. Assume $X_1, X_2, \ldots$ are identically distributed with Cov $\left(X_i, X_j\right) \leq 0, i \neq j$. Then*

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\left(X_i\right) + \sum_{i \neq j} Cov\left(X_i, X_j\right) \leq n\mathbb{V}\left(X_i\right)$$

2. *Can get rid of the second moment condition by a truncation argument. Will see this for SLLN.*

Definition (Almost sure convergence). *$\{X_i\}_{i\geq 1}$ sequence of RV converges to 0 **almost surely** if*

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n\left(\omega\right) = 0\right\}\right) = \mathbb{P}\left(\lim_{n\to\infty} X_n = 0\right) = \mathbb{P}\left(X_n \to 0\right) = 1.$$

We often note it as $X_n \xrightarrow{a.s.} 0$.

Note. *We convert again a probabilistic statement into a deterministic one by fixing an ω.*

We are interested in that portion of the space Ω in which the sequence $\{X_n\}_n$ converges to 0 pointwise. There can be measure-zero sets of Ω for which this does not happen, but as long as they are negligible this does not matter!

Another important thing is to note the difference between convergence in probability and almost sure convergence. The latter is indeed a stronger statement. To see this, let's rewrite the definitions of a.s. convergence and convergence i.p. and compare them

$$\forall\, \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}\left(|X_n - m| \leq \varepsilon\right) = 1 \qquad (X_n \overset{P}{\to} m)$$

$$\forall\, \varepsilon > 0, \exists\, N \in \mathbb{N} : \mathbb{P}\left(|X_n - m| \leq \varepsilon\right) = 1, \forall\, n \geq N \qquad (X_n \overset{a.s.}{\to} m)$$

We can see that convergence in probability is a statement about the probability of the sequence of RVs X_n getting closer to m that holds only in the limit. Almost sure convergence is instead a statement about the probability of the sequence getting closer to m that holds for all n larger than a certain N_ε.

Finally, there is an alternative definition of almost sure convergence which is the following

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n\left(\omega\right) = X\left(\omega\right)\right\}\right) = 1$$

which requires that the sequence $X_n\left(\omega\right)$ is convergent for almost all ω.

Proposition (a.s. $\implies$ i.p.)**.** *If $Y_n \to 0$ a.s., then $Y_n \to 0$ i.p.*

Proof. Recall that $Y_n \to 0 \iff \forall r \in \mathbb{N}_+, \exists N \in \mathbb{N} : \forall\, n \geq N, |Y_n| \leq \frac{1}{r}$. Expressed in terms of events,

$$\{Y_n \to 0\} = \cap_{r=1}^{\infty} \cup_{N=1}^{\infty} \cap_{n=N}^{\infty} \left\{ |Y_n| \leq \frac{1}{r} \right\}.$$

Then, beginning with the definition of almost sure convergence

$$\begin{aligned}
\mathbb{P}\left(Y_n \to 0\right) = 1 &\iff \mathbb{P}\left(\cap_{r=1}^{\infty} \cup_{N=1}^{\infty} \cap_{n=N}^{\infty} \left\{ |Y_n| \leq \frac{1}{r} \right\} \right) = 1 \\
&\iff \forall\, r \geq 1, \mathbb{P}\left(\cup_{N=1}^{\infty} \cap_{n=N}^{\infty} \left\{ |Y_n| \leq \frac{1}{r} \right\} \right) = 1, \quad \text{(above subset of below)} \\
&\iff \forall\, r \geq 1, \lim_{N \to \infty} \mathbb{P}\left(\cap_{n=N}^{\infty} \left\{ |Y_n| \leq \frac{1}{r} \right\} \right) = 1, \quad \text{(sequential continuity)} \\
&\implies \forall\, r \geq 1, \lim_{N \to \infty} \mathbb{P}\left(\left\{ |Y_N| \leq \frac{1}{r} \right\} \right) = 1, \quad \text{(fewer intersections)} \\
&\iff \forall\, r \geq 1, \lim_{N \to \infty} \mathbb{P}\left(\left\{ |Y_| > \frac{1}{r} \right\} \right) = 0, \quad \text{(take complement)} \\
&\iff Y_n \xrightarrow{p} 0 \quad \text{(take limit as } r \to \infty\text{)}
\end{aligned}$$

□

Note. *The converse is not true because of line 4. Indeed, note that $\mathbb{P}\left(\left\{|Y_n| \leq \frac{1}{r}\right\}\right) = 1$ implies $\cap_n \mathbb{P}\left(\left\{|Y_n| \leq \frac{1}{r}\right\}\right) \leq 1$.*

Definition (limsup and liminf)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_i \in \mathcal{F}, i \geq i$. Then **limsup** for sets is defined as*

$$\begin{aligned}
\limsup A_n &:= \cap_{j=1}^{\infty} \cup_{i=j}^{\infty} A_i \\
&= \left\{ \omega \in \Omega : \omega \in A_i \textit{ for infinitely many values of } n \right\} \\
&= \left\{ \omega \in \Omega : \omega \in A_i \textbf{\textit{ infinitely often (i.o.)}} \right\} \\
&= \left\{ \omega \in \Omega : \forall\, j \in \mathbb{N}, \exists\, i \geq j : \omega \in A_i \right\}
\end{aligned}$$

liminf *for sets is defined as*

$$\begin{aligned}\liminf A_n &:= \cup_{j=1}^{\infty} \cap_{i=j}^{\infty} A_i \\ &= \left\{\omega \in \Omega : \omega \in A_i \text{ for all but finitely many values of } n\right\} \\ &= \left\{\omega \in \Omega : \omega \in A_i \textbf{ eventually (ev)}\right\} \\ &= \left\{\omega \in \Omega : \exists\, j \in \mathbb{N}, \forall\, i \geq j : \omega \in A_i\right\}\end{aligned}$$

Note. *We should read the* $\limsup$ *as for all j there exists and $i \geq j$ such that an event holds. On the other hand, we should interpret the* $\liminf$ *as there exists a j such that for all $i \geq j$ an event holds This is because intersections are equivalent to 'for all' and unions are equivalent to 'there exists at least one'.*

From the definition we can note that $\liminf A_n \subseteq \limsup A_n$ *because the* $\liminf$ *describes a harder event to realize than* $\limsup$*. Indeed the events in* $\liminf A_n$ *imply those in* $\limsup A_n$ *but not the viceversa. To see this, note that* $\limsup A_n$ *is a subset of the sample space such that for every element of a sequence j there exists an element i further ahead that contains ω. The* $\liminf A_n$ *contains all ωs such that* ***all*** *the elements of the sequence after the j-th contain ω.*

The $\limsup$ *of A_n can be thought of as the set of outcomes that occur infinitely many times within the infinite sequence of events $\{A_n\}_n$.*

Note. *Why the terminology? Close correspondence to functions. Consider*

$$\limsup_{n \to \infty} \mathbb{1}_{A_n}(\omega) = \mathbb{1}_{\limsup A_n}(\omega).$$

Proposition. *By Fatou's Lemma,*

$$\mathbb{P}\left(\liminf A_n\right) \leq \liminf_{n \to \infty} \mathbb{P}\left(A_n\right) \leq \limsup_{n \to \infty} \mathbb{P}\left(A_n\right) \leq \mathbb{P}\left(\limsup A_n\right).$$

Proof. Start with $\mathbb{P}(A) = \mathbb{E}\left[\mathbb{1}_A\right]$ and then move on applying Fatou's lemma. □

We now discuss techniques we might want to use to prove almost sure convergence. These lemmas are called Borel-Cantelli lemmas and they are **sufficient conditions** for almost sure convergence. The first lemma is the following:

Lemma (Borel-Cantelli). *Let $\{A_n\}_n$ be a sequence of events. Then,*

$$\sum_{n=1}^{\infty} \mathbb{P}\left(A_n\right) < \infty \implies \sum_{n=1}^{\infty} \mathbb{1}\left(A_n\right) < \infty$$

This lemma tells us that if the sum of the proabilities of a sequence of events is bounded, then it must be that **only finitely many of the events hold**.

Lemma (Borel-Cantelli I). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_i \in \mathcal{F}, i \geq 1$.*
If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then

$$\mathbb{P}\left(\limsup A_n\right) = 0.$$

That is, almost surely only finitely many events $\{A_n\}_{n\geq 1}$ may occur.
In other terms, $\exists\, N \in \mathbb{N} : A_n^c, \forall\, n \geq N$.

The lemma is saying that the probability that infinitely many of the events hold is 0. This comes directly from the definition of $\limsup A_n := \left\{\omega \in \Omega : \omega \in A_i \text{ infinitely often}\right\}$.

Proof. Fix $\varepsilon > 0$. Assumption implies there exists N such that $\sum_{n=N}^{\infty} P(A_n) = \varepsilon$ (finite tail). Then

$$0 \leq \mathbb{P}\left(\limsup A_n\right) = \mathbb{P}\left(\cap_{N=1}^{\infty} \cup_{n=N}^{\infty} A_n\right) \leq \mathbb{P}\left(\cup_{n=N}^{\infty} A_n\right) \leq \sum_{n=N}^{\infty} \mathbb{P}(A_n) \leq \varepsilon.$$

The first inequality follows from $A_n \subseteq \cup_{n=N}^{\infty} A_n$ and the second inequality follows from the union bound. Finally, as ε was arbitrary, take $\varepsilon \searrow 0$. □

Lemma (Borel-Cantelli II). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_i \in \mathcal{F}, i \geq 1$.*
Suppose events $\{A_i\}_{i\geq 1}$ are mutually independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. Then

$$\mathbb{P}\left(\limsup A_n\right) = 1.$$

that is, almost surely infinitely many of the events $\{A_n\}_{n\geq 1}$ occur.
In other terms, $\exists\, N \in \mathbb{N} : A_n, \forall\, n \geq N$

Proof. Equivalently, we can show

$$\mathbb{P}\left(\left(\limsup A_n\right)^c\right) = \mathbb{P}\left(\left(\cap_{j=1}^{\infty} \cup_{i=j}^{\infty} A_i\right)^c\right) = \mathbb{P}\left(\cup_{j=1}^{\infty} \cap_{i=j}^{\infty} A_i^c\right) = 0.$$

Fix $j \geq 1, m \leq \infty$. Suffices to show $\mathbb{P}\left(\cap_{i=j}^{\infty} A_i^c\right) = 0$.

$$\begin{aligned} \mathbb{P}\left(\cap_{i=j}^{\infty} A_i^c\right) &\leq \mathbb{P}\left(\cap_{i=j}^{m} A_i^c\right) = \Pi_{i=j}^{m} P\left(A_i^c\right) \\ &= \Pi_{i=j}^{m} 1 - P(A_i) \leq \Pi_{i=j}^{m} 1 - P(A_i) \leq \Pi_{i=j}^{m} e^{-\mathbb{P}(A_i)} \\ &= \exp\left(-\sum_{i=1}^{m} P(A_i)\right). \end{aligned}$$

The first line follows from independence (if two events are independent so are the complements). The second line follows from $1 - x \leq e^{-x}$.
True for all $M \geq j$. Take $M \to \infty$, and get

$$\mathbb{P}\left(\cap_{i=j}^{\infty} A_i^c\right) \leq \exp\left(-\sum_{i=1}^{m} P\left(A_i\right)\right) \overset{M\to\infty}{\longrightarrow} 0.$$

□

Note. *Note that both lemmas can be stated in terms of* lim inf *by using the following fact*

$$\mathbb{P}\left(\limsup A_n\right) = \mathbb{P}\left(\cap_{N=1}^{\infty} \cup_{n=N}^{\infty} A_n\right) = 1 - \mathbb{P}\left(\cup_{N=1}^{\infty} \cap_{n=N}^{\infty} A_n^c\right) = 1 - \mathbb{P}\left(\liminf A_n^c\right)$$

To wrap up, we have that if $\sum_{n=1}^{\infty} \mathbb{P}\left(A_n\right) < \infty$*, then*

$$\mathbb{P}\left(\limsup A_n\right) = 0, \quad \mathbb{P}\left(\liminf A_n^c\right) = 1$$

Intuitively, Borel-Cantelli I tells us that if only finitely many of the $\{A_n\}_{n\geq 1}$ *occur, it must be that all but finitely many of the* $\{A_n^c\}_{n\geq 1}$ *occur!*
If $\sum_{n=1}^{\infty} \mathbb{P}\left(A_n\right) = \infty$ *and the events are mutually independent, then*

$$\mathbb{P}\left(\limsup A_n\right) = 1, \quad \mathbb{P}\left(\liminf A_n^c\right) = 0$$

Again, if infinitely many of the $\{A_n\}_{n\geq 1}$ *realize, then it must be that only finitely many of the* $\{A_n^c\}_{n\geq 1}$ *hold.*

2.2 Strong law of large numbers

The below results are necessary to prove the finite first moment version of the strong law of large numbers.

Theorem (Kolmogorov's inequality). *Let* $X_1, \ldots, X_n$ *be mutually independent RVs. Assume that* $\sigma_i^2 = \mathbb{E}\left(X_i^2\right) < \infty$ *and that* $\mathbb{E}\left(X_i\right) = 0$*. Then for any* $\lambda > 0$*, we have*

$$\mathbb{P}\left(\max_{1\leq i\leq n} |X_1 + \ldots + X_i| \geq \lambda\right) \leq \frac{\sum_{i=1}^{n} \sigma_i^2}{\lambda^2}.$$

Note. *Think of this result as a multivariate generalization of Chebyshev's inequality. It says that if we look at the maximum of the absolute value of all partial sums of the* X_i*s, we know that the probability of it being larger than* $\lambda > 0$ *is bounded by the variance of the total sum of the* X_i*s over* λ*. Chebyshev's inequality follows from setting* $n = 1$*.*

Proof. Let $S_k = X_1 + \ldots + X_k$. Let

$$\begin{aligned} A &:= \left\{ \max_{1 \leq i \leq n} |S_i| \geq \lambda \right\} = \cup_{i=1}^{n} \left\{ S_i \geq \lambda \right\} \\ A_k &:= \left\{ \max_{1 \leq i \leq (k-1)} |S_i| < \lambda, |S_k| \geq \lambda \right\} \end{aligned}$$

Note that $A_k \cap A_l = \emptyset$ for $k \neq l$ and $A = \cup_{k=1}^{n} A_k$. Then

$$\mathbb{P}\left(A\right) = \mathbb{E}\left(\mathbb{1}_A\right) = \mathbb{E}\left[\sum_{k}^{n} \mathbb{1}_{A_k}\right] = \sum_{k}^{n} \mathbb{E}\left[\mathbb{1}_{A_k}\right].$$

By definition, for event A_k, we have $|s_k| \geq \lambda \iff s_k^2/\lambda^2 \geq 1$, so $\mathbb{1}_{A_k} \leq \frac{S_k^2}{\lambda^2} \mathbb{1}_{A_k}$. Plugging back into the expression,

$$\mathbb{P}\left(A\right) = \sum_{k=1}^{n} \mathbb{E}\left(\mathbb{1}_{A_k}\right) \leq \sum_{k=1}^{n} \left\{ \mathbb{E}\left[\frac{s_k^2}{\lambda^2} \mathbb{1}_{A_k}\right] \right\} = \frac{1}{\lambda^2} \sum_{k=1}^{n} \mathbb{E}\left[s_k^2 \mathbb{1}_{A_k}\right]$$

We already got λ^2 at the denominator, so we just need to find the denominator. To find such a bound for the ending sum:

$$\begin{aligned} \sum_{k=1}^{n} \mathbb{E}\left[s_k^2 \mathbb{1}_{A_k}\right] &\leq \sum_{k=1}^{n} \left\{ \mathbb{E}\left[s_k^2 \mathbb{1}_{A_k}\right] + \mathbb{E}\left[\left(s_n - s_k\right)^2 \mathbb{1}_{A_k}\right] \right\} \\ &= \sum_{k=1}^{n} \mathbb{E}\left[s_n^2 \mathbb{1}_{A_k}\right] = \mathbb{E}\left[s_n^2 \sum_{k=1}^{n} \mathbb{1}_{A_k}\right] = \mathbb{E}\left[s_n^2 \mathbb{1}_{A}\right] \leq \mathbb{E}\left[s_n^2\right] = \mathbb{V}\left[s_n\right] = \sum_{i=1}^{n} \sigma_i^2. \end{aligned}$$

The first line follows from the below derivation

$$\begin{aligned} & s_k^2 + \left(s_n - s_k\right)^2 = s_n^2 - 2s_n 2s_k + 2s_k^2 = s_n^2 - 2s_k\left(s_n - s_k\right) \\ \implies & \mathbb{1}_{A_k}\left(s_k^2 + \left(s_n - s_k\right)^2\right) = \mathbb{1}_{A_k} s_n^2 - 2s_k\left(s_n - s_k\right)\mathbb{1}_{A_k} \\ \implies & \mathbb{E}\left[\mathbb{1}_{A_k} s_n^2 - 2s_k\left(s_n - s_k\right)\mathbb{1}_{A_k}\right] = \mathbb{E}\left[\mathbb{1}_{A_k} s_n^2\right] - 2\mathbb{E}\left[\mathbb{1}_{A_k} s_k\left(s_n - s_k\right)\right] \end{aligned}$$

However, in the final expectation, $s_k \mathbb{1}_{A_k} \perp \left(s_n - s_k\right)$, because the first depends on the first k summands, whilst the second on t$(X_{k+1}, \ldots, X_n)$ so the expectation factorizes and $\mathbb{E}\left(s_n - s_k\right) = 0$, producing our desired result.

□

Theorem. *Let $X_1, \dots, X_n$ be mutually independent. Assume $\sigma_i^2 := \mathbb{E}\left[X_i^2\right] < \infty$ and $\mathbb{E}\left[X_i\right] = 0$. Assume variances $\sum_{i=1}^{\infty} \sigma_i^2 < \infty$. Then $\lim_{n\to\infty} \sum_{i=1}^{n} X_i$ exists and is finite a.s.*

Proof. Let $S_n := \sum_{i=1}^{n} X_i$. We don't know what the limit of S_n will be, so we'll show that $\{S_n\}_n$ is a Cauchy sequence a.s. Let

$$A_{N,r} := \left\{ \exists\, i, j \geq N : |S_i - S_j| \geq \frac{1}{r} \right\}$$

Then $A_{N,r}$ is increasing in r and decreasing in N. Consider the event

$$\left\{ \{S_n\}_{n\geq 1} \text{ is not Cauchy } \right\} = \cup_{r=1}^{\infty} \cap_{N=1}^{\infty} A_{N,r}$$

Then,

$$\mathbb{P}\left(\left\{ \{S_n\}_{n\geq 1} \text{ is not Cauchy } \right\} \right) = \mathbb{P}\left(\cup_{r=1}^{\infty} \cap_{N=1}^{\infty} A_{N,r} \right) = \lim_{r\to\infty} \lim_{N\to\infty} \mathbb{P}\left(A_{N,r}\right)$$

where the second equality follows from sequential continuity of limits. We will now show that $\forall\, r \geq 1 : \lim_{N\to\infty} \mathbb{P}\left(A_{N,r}\right) = 0$. Define

$$B_{N,r} := \left\{ \exists\, i \geq N : |S_i - S_N| \geq \frac{1}{2r} \right\}$$

Note that $A_{N,r} \subseteq B_{N,r}$ by the triangle inequality, therefore $\mathbb{P}\left(A_{N,r}\right) \leq \mathbb{P}\left(B_{N,r}\right)$ so we now have to deal with an event that has only one index within it. Then

$$\mathbb{P}\left(B_{N,r}\right) = \lim_{N_i\to\infty} \mathbb{P}\left(\exists\, i \in [N, N_i] : |S_i - S_N| \geq \frac{1}{2r} \right)$$

where the equality follows from sequential continuity of $\mathbb{P}$ (we are just taking the limit of a finite union of events). Note that $S_i - S_N =_{N+1} + \cdots + N_i$ is a partial sum of mutually independent random variable, so we can apply Kolmogorov's inequality and get

$$\mathbb{P}\left(\exists\, i \in [N, N_i] : |S_i - S_N| \geq \frac{1}{2r} \right) \leq \frac{\sum_{i=N}^{N_i} \sigma_i^2}{1/4r^2}$$

The object on the RHS might not have a limit but does have a lim sup, thus

$$\mathbb{P}\left(B_{N,r}\right) \leq \limsup_{N_i\to\infty} 4r^2 \sum_{i=N+1}^{N_i} \sigma_i^2 = 4r^2 \sum_{i=N+1}^{\infty} \sigma_i^2$$

Hence

$$\limsup_{N\to\infty} \mathbb{P}\left(B_{N,r}\right) \leq 4r^2 \limsup_{N\to\infty} \sum_{i=N+1}^{\infty} \sigma_i^2 = 0$$

which allows us to conclude that the probability that $\{S_n\}_{n\geq 1}$ is not Cauchy is 0 a.s. □

Proposition (Kronecker Lemma). *If $\{a_k\}_{k\geq 1}$ is such that $\sum_{k=1}^{\infty} \frac{a_k}{k}$ is convergent, then $\frac{1}{n}\sum_{k=1}^{n} a_k \to 0$ as $n \to \infty$.*

Theorem. *Let $X_1, X_2, \ldots$ be mutually independent with $\mu_i := \mathbb{E}\left[X_i\right]$ and $\sigma_i^2 := \mathbb{V}\left(X_i\right) < \infty$. Suppose that $\sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} < \infty$. Then*

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu_i\right) \xrightarrow[n\to\infty]{a.s.} 0.$$

Proof. Let $Y_i := \frac{X_i - \mu_i}{i}$. Note that $\mathbb{E}\left[Y_i\right] = 0$ and $\mathbb{V}\left[Y_i\right] = \frac{1}{i^2}\mathbb{V}\left[X_i\right] = \frac{\sigma_i^2}{i^2}$. By assumption $\sum_{i=1}^{\infty} \mathbb{V}\left(Y_i\right) = \sum_{i=1}^{\infty} \frac{\sigma_i^2}{i_2} < \infty$. By the theorem above we know that $\lim \sum_{i=1}^{n} Y_i = \lim_{n\to\infty} \sum_{i=1} \frac{X_i \mu_i}{i}$ exists and is finite. By Kronecker's lemma

$$\left\{\omega \in \Omega : \sum_{i=1}^{n} \frac{X_i\left(\omega\right) - \mu_i}{i} \text{ converges}\right\} \subseteq \left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^{n}\left(X_i\left(\omega\right) - \mu_i\right) \to 0\right\}$$

To conclude, note that

$$1 = \mathbb{P}\left(\left\{\omega \in \Omega : \sum_{i=1}^{n} \frac{X_i\left(\omega\right) - \mu_i}{i} \text{ converges}\right\}\right) \leq \mathbb{P}\left(\left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^{n}\left(X_i\left(\omega\right) - \mu_i\right) \to 0\right\}\right)$$

□

Note. *This statement require finite second moments so its stronger than the SLLN.*

Theorem (SLLN). *Let $X_1, X_2, \ldots$ be i.i.d. RVs with $\mathbb{E}\left[\left|X_i\right|\right] < \infty$ and let $m := \mathbb{E}\left[X_i\right]$. Let $S_n := \sum_{i=1}^{n} X_i$. Then*

$$\frac{S_n}{n} \xrightarrow{a.s.} m.$$

Proof. This is a sketch. See class notes for details. Proceed in steps.

1. Kolmogorov's inequality places a bound on $\sum_{i=1}^{n} X_i$.
2. Apply the theorem on convergence. After applying Kolmogorov's inequality, summable variances imply almost sure convergence.
3. Apply Kronecker's Lemma to show that if $\sum_{k=1}^{\infty} a_k/k$ is convergent, then

$$\frac{1}{n}\sum_{k=1}^{n} a_k \to 0 \quad a.s., n \to \infty.$$

Moreover, that for mutually independent X_i with $\mu_i := \mathbb{E}\left[X_i\right]$ and $\sigma_i^2 := Var\left(X_i\right) < \infty$. Suppose that $\sum_{k=1}^{\infty} \sigma_k^2 / k^2 < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu_i\right) \xrightarrow[n\to\infty]{a.s.} 0.$$

4. Up to this point we proved the SLLN for finite second moments, however we'd like to prove it without this assumption. Assume that $\mathbb{E}\left[X_i\right] = 0$ (centered iid rvs). Then, use truncation argument by defining two variables

$$Y_k := x_k \mathbb{1}_{|X_k| \le k}, \qquad Z_k := x_k \mathbb{1}_{|X_k| > k}$$

so $X_k = Y_k + Z_k$. We will show that $\frac{1}{n}\sum_{i=1}^{n} Y_i \to 0$ a.s. and $\frac{1}{n}\sum_{i=1}^{n} Z_i \to 0$ a.s., so that also $\frac{1}{n}\sum_{i=1}^{n} X_i \to 0$ a.s.

We start with the Z_ks. We will show something stronger, i.e. that a.s. only finitely many of the Z_ks are non-zero. We use Borel-Cantelli I. Define the event $A_k = \{Z_k \neq 0\}$. We want to show that $\sum_{k=1}^{\infty} \mathbb{P}\left(A_k\right) < \infty$.

$$\sum_{k=1}^{\infty} \mathbb{P}\left(A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}\left(|X_k| > k\right) \le \mathbb{E}\left[|X_k|\right] < \infty$$

Let's turn to the X_ks. Let $a_n := \int_{n-1}^{n} x dF\left(x\right) - \int_{-n}^{-n+1} x dF\left(x\right)$. Note that $\sum_{n=1}^{\infty} a_n = \mathbb{E}\left[|X_1|\right] < \infty$. Note that

$$\int_{n-1}^{n} x^2 dF\left(x\right) - \int_{-n}^{-n+1} x^2 dF\left(x\right) \le a_n := \int_{n-1}^{n} nx dF\left(x\right) - \int_{-n}^{-n+1} nx dF\left(x\right) = na_n$$

Now

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{\mathbb{V}\left(Y_k\right)}{k^2} &\le \sum_{k=1}^{\infty} \frac{1}{k^2}\mathbb{E}\left[Y_k^2\right] \\
&= \sum_{k=1}^{\infty} \frac{1}{k^2} \int_{-k}^{k} x^2 dF\left(x\right) \\
&= \sum_{k=1}^{\infty} \frac{1}{k^2} \sum_{\ell=1}^{k} \left(\int_{\ell-1}^{\ell} x^2 dF\left(x\right) + \int_{-\ell}^{-\ell+1} x^2 dF\left(x\right) \right) \\
&\le \sum_{k=1}^{\infty} \frac{1}{k^2} \sum_{\ell=1}^{k} \ell a_\ell \\
&= \sum_{\ell=1}^{\infty} \ell a_\ell \sum_{k=\ell}^{\infty} \frac{1}{k^2} \\
&\le \sum_{\ell=1}^{\infty} \ell a_\ell \frac{4}{\ell} = 4\sum_{\ell=1}^{\infty} a_\ell = \mathbb{E}\left[|X_1|\right] < \infty
\end{aligned}$$

By step 3 we showed that

$$\frac{1}{n}\sum_{k=1}^{\infty}\left(Y_k - \mathbb{E}\left[Y_k\right]\right) \overset{a.s.}{\to} 0$$

but we wanted to show that $\frac{1}{n}\sum_{k=1}^{n} Y_k \overset{a.s.}{\to} 0$. We need $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_k\right] \to 0$

$$\mathbb{E}\left[Y_k\right] = \int_{-k}^{k} xdF\left(x\right) \overset{k\to\infty}{\to} \int_{-\infty}^{\infty} xdF\left(x\right) = \mathbb{E}\left[X_1\right] = 0$$

□

Note. *In the truncation argument we used the fact that $\sum_{k=1}^{\infty}\mathbb{P}\left(|X_k| > k\right) \leq \mathbb{E}\left[|X_k|\right]$. Let $F\left(x\right) := \mathbb{P}\left(X \leq x\right)$, then*

$$\begin{aligned}\mathbb{P}\left(|X_k| > k\right) &= \mathbb{P}\left(|X_1| > k\right)\\ &= \mathbb{P}\left(X_1 < -k\right) + \mathbb{P}\left(X_1 > k\right)\\ &= F\left(-k-0\right) + \left(1 - F\left(k\right)\right)\\ &\leq \int_{-k}^{-k+1} F\left(y\right)dy + \int_{k-1}^{k}\left(1 - F\left(y\right)\right)dy\end{aligned}$$

Then

$$\begin{aligned}\sum_{k=1}^{\infty}\mathbb{P}\left(|X_k| > k\right) &\leq \int_{-\infty}^{0} F\left(y\right)dy + \int_{0}^{\infty}\left(1 - F\left(y\right)\right)dy\\ &= -\int_{-\infty}^{0} xdF\left(x\right) + \int_{0}^{\infty} xdF\left(x\right)\\ &= \int |x|dF\left(x\right) = \mathbb{E}\left[|X_1|\right]\end{aligned}$$

Recap. As a recap, we learnt three different ways of proving almost sure convergence ($X_n \overset{a.s.}{\to} m$):

1. using the definition, i.e. $\mathbb{P}\left(\lim_{n\to\infty} X_n = m\right) = 1$. This is quite hard in general and also requires knowledge of the background probability space Ω. If you use this approach you should check that $X_n\left(\omega\right) \to m$ pointwise $\forall, \omega \in \Omega \setminus \mathcal{N}$.
2. using Borel-Cantelli lemma. To use this approach we need to define an auxiliary sequence of events $\left\{A_n\right\}_{n\geq 1}$ and show that

$$\sum_{n=1}^{\infty}\mathbb{P}\left(A_n\right) < \infty \quad \text{or} \quad \sum_{n=1}^{\infty}\mathbb{P}\left(A_n\right) = \infty$$

If we are in the first case we can say that A_n realizes finitely many times in the sequence, hence

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 0$$

If we are in the second case, assuming that the events are mutually independent, then we know that the event A_n realizes infinitely many times, thus

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1$$

Sometimes it is required to use both lemmas to bound the value of the lim sup. This lemma is also useful for the lim inf if we are able to construct a lower bound for this quantity, as we can conclude by saying $c \leq \liminf_{n\to\infty} A_n \leq \limsup_{n\to\infty} A_n \leq c$. Another common trick is to compute the probability $\mathbb{P}\left(A_n \geq \varepsilon\right), \forall\, \varepsilon > 0$ to get something tractable and convergent as an upper bound of such quantity. Then, since it holds for all ε, we take the limit as $\varepsilon \to 0$. Finally, when we want to use Borel-Cantelli I, we want to upper bound our quantity with something more tractable that we are able to show it is finite, whilst with Borel-Cantelli II we use want to show that a lower bound is divergent.

3. Using the SLLN, once we know that $\mathbb{E}\left[|X_i|\right] < \infty$, we can claim that $S_n/n \overset{a.s.}{\to} \mathbb{E}\left[X_i\right]$

3 Central Limit Theorem and Characteristic Functions

LLNs give us information on the first order behavior of sums of RVs. However, we might be interested in second order behavior as well. The CLT tells us that

$$\frac{S_n - m \cdot n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

3.1 Central Limit Theorem

Definition (Convergence in distribution/Weak Convergence)**.** *Let F denote the CDF of X where $F(x) := \mathbb{P}(X \leq x)$. Let $\{X_n\}_{n\geq 1}$ be a sequence of real-valued RVs such that X_n has a CDF F_n. Let X be a RV with cdf F. If*

$$F_n(x) \xrightarrow[n\to\infty]{} F(x)$$

for every continuity point x of F, then we say that $\{X_n\}_{n\geq 1}$ converges in distribution to X and that $\{F_n\}_{n\geq 1}$ converges weakly to F.

Note. *Again, we fix an x and then the statement we want to verify becomes deterministic.*

A continuity point of F is a point where F is continuous. Thus we just require that F is non-decreasing, but it can be discontinuous.

The deep result behind this is that we can use F to fully characterize $\mathbb{P}$.

We usually denote it with

$$X_n \xrightarrow{d} X, \quad X_n \xrightarrow{D} X, \quad X_n \xrightarrow{w} X, \quad X_n \implies X, \quad X_n \xrightarrow{L} X, \quad \mathcal{L}(X_n) \to \mathcal{L}(X)$$

or simply using F_n and F in place of X_n and X. Also sometimes we abuse notation and write $X_n \implies F$.

Example (Why continuity points?)**.** *Consider $X_n \sim U(0, 1/n)$. Consider*

$$F_n(x) = \begin{cases} 0, & x < 1/n \\ 1, & x \geq 0 \end{cases}, \quad F_n(x) \xrightarrow{n\to\infty} \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}.$$

However, we usually define CDFs as being right continuous, thus

$$F(x) \xrightarrow{n\to\infty} \begin{cases} 0, & x < 0 \\ 1, & x \geq 0. \end{cases}$$

which is the CDF of δ_0. We have $F_n \Rightarrow F, \forall x \neq 0$, where we can remove 0 because it is a discontinuity point of F, thus we care about pointwise convergence in all other points.

Note. *The following expressions are equivalent*

$$F_n(x) \to F(x) \iff \mathbb{P}(X_n \le x) \to \mathbb{P}(X \le x) \iff \mathbb{E}\left[\mathbb{1}_{(-\infty,x)}(X_n)\right] \to \mathbb{E}\left[\mathbb{1}_{(-\infty,x)}(X)\right].$$

Using the definition of weak convergence might not be immediate, thus we need a characterization. In particular, we will see that to show weak convergence it is sufficient to show that some particular expectations of our random variables need to coincide. In general, it is easier to deal with expectations.

Theorem (Characterization of weak convergence). $X_n \Rightarrow X$ *if and only if* $\mathbb{E}\left(g(X_n)\right) \to \mathbb{E}\left(g(X)\right)$ *for every continuous and bounded function* g.

Proof. Will only show $(\Longleftarrow)$. In general, half-line indicator functions can be approximated by continuous and bounded functions. The intuition is that you can "bound" the jump with linear functions. Define function $g_{x,\varepsilon}$ that is 1 until x, declining until reaching 0 at $x+\varepsilon$ and staying there. Then,

$$g_{x-\varepsilon,\varepsilon}(\cdot) \le \mathbb{1}_{(-\infty,x)}(\cdot) \le g_{x,\varepsilon}(\cdot) \le \mathbb{1}_{(-\infty,x+\varepsilon)}(\cdot)$$

and

$$F_n(x) = \mathbb{E}\left[\mathbb{1}_{(-\infty,x)}(X_n)\right] \le \mathbb{E}\left[g_{x,\varepsilon}(X_n)\right] \le \mathbb{E}\left[\mathbb{1}_{(-\infty,x+\varepsilon]}(X_n)\right] \le F_n(x+\varepsilon).$$

Also, $F(x) \le \mathbb{E}\left[g_{x,\varepsilon}(X)\right] \le F(x+\varepsilon)$. Now, want to squeeze F by taking the limits

$$\begin{aligned} F(x-\varepsilon) &\le \mathbb{E}\left[g_{x-\varepsilon,\varepsilon}(X)\right] = \lim_{n\to\infty} \mathbb{E}\left[g_{x-\varepsilon,\varepsilon}(X_n)\right] \le \liminf_{n\to\infty} F_n(x) \\ &\le \limsup F_n(x) \le \lim_{n\to\infty} \mathbb{E}\left[g_{x,\varepsilon}(X_n)\right] = \mathbb{E}\left[g_{x,\varepsilon}(X)\right] \le F(x+\varepsilon) \end{aligned}$$

Then, $F(x-\varepsilon) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x+\varepsilon)$.
Taking $\varepsilon \searrow 0$, if x is a continuity point, then LHS, RHS converge to $F(x)$ and $\lim_{n\to} F_n(x) = F(x)$. □

Definition (Converges weakly). *Let S be a complete metric space with its Borel sigma-algebra. Let $\{\mu_n\}_{n\ge1}$ and μ be measures on S. We say μ_n* ***converges weakly to*** *μ denoted $\mu_n \Rightarrow \mu$ if*

$$\int g d\mu_n \to \int g d\mu$$

for every continuous bounded $g : S \to \mathbb{R}$.

Note (Proving weak convergence). *Which functions to check? C_b is a rather broad class of functions. Will later show that it suffices to check* ***characteristic functions****. Only need to check trigonometric functions, eg.* $\sin(tx), \cos(tx), t \in \mathbb{R}$.

Theorem (Central limit theorem). *Let $X_1, X_2, \ldots$ be a sequence of iid real-valued random variables. Assume $\mathbb{E}\left(X_i^2\right) < \infty$. Let $S_n := X_1 + \ldots + X_n$. Then*

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}(S_n)}} \Rightarrow \mathcal{N}(0,1).$$

That is, for every $x \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathbb{V}(S_n)}} \leq x\right) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-1/2z^2} dz.$$

Note, $\mathbb{E}[S_n] = n\mathbb{E}(X_1)$ and $\mathbb{V}(S_n) = n\mathbb{V}(X_1) \iff$

$$\frac{S_n - n\mathbb{E}[X_n]}{\sqrt{n\mathbb{V}(X_n)}}.$$

Note. *Note that the numerator of is $O_p(n)$, whereas the denominator is $O_p(\sqrt{n})$, thus the fraction is $O_p(\sqrt{n})$. The fluctuations of this fraction are distributed as a Gaussian random variable.*

Proof (Lindeberg). To make the proof easier, assume there exists $C \in \mathbb{R}_{++}$ such that $\forall\, i \in \mathbb{N}$, $\mathbb{E}\left(|X_i|^3\right) \leq C$ (note that this is a uniform bound on all the X_is). On the other hand, we will not assume X_i have the same distribution. Only need that they are mutually independent. May assume without loss of generality that X_i is mean zero and variance one. Otherwise, can normalize and define $Y_i := \frac{X_i - \mathbb{E}[X_i]}{\sqrt{\mathbb{V}[X_i]}}$. Now, it sufficient to show $S_n/\sqrt{n} \Rightarrow \mathcal{N}(0,1)$.

We will use the characterization above for a particular dense class of functions. Let $g \in \mathcal{C}_b^3$, that is let g be three times differentiable with g, g', g'', g''' be continuous and bounded. The space of these functions is dense in C_b. By the characterization theorem, it suffices to show that

$$\mathbb{E}\left[g\left(S_n/\sqrt{n}\right)\right] \to \mathbb{E}\left[g(Z)\right], \quad \forall\, g \in \mathcal{C}_b^3 \tag{$\star$}$$

where $Z \sim \mathcal{N}(0,1)$.

Before proceeding with the proof, let's make a few notes.

1. **Observation 1:** If $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, then
$$\frac{Z_1 + Z_2 + \ldots + Z_n}{\sqrt{n}}$$
is also a standard Gaussian RV. Let $T_n := Z_1 + \ldots + Z_n$. Thus
$$\frac{T_n}{\sqrt{n}} \Rightarrow Z.$$
CLT **holds exactly** for the Gaussian case without taking the limit.
2. **Observation 2:** Compare the summands to Gaussians. We can reframe our goal ($\star$) as
$$\mathbb{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g\left(\frac{T_n}{\sqrt{n}}\right)\right] \xrightarrow[n\to\infty]{} 0.$$
3. **Lindeberg's swapping trick:** Exchange X_i to Z_i one-by-one.

Assume that X_is and Z_i live in the same probability space and they are all mutually independent. Define the following random variables

$$\begin{aligned}
S_n^{(0)} &= X_1 + X_2 + \cdots + X_n = S_n \\
S_n^{(1)} &= Z_1 + X_2 + \cdots + X_n \\
&\vdots \\
S_n^{(j)} &= Z_1 + Z_2 + \cdots + Z_j + X_{j+1} + \cdots + X_n \\
&\vdots \\
S_n^{(n)} &= Z_1 + Z_2 + \cdots + Z_n = T_n
\end{aligned}$$

Then we can write our difference of functions as

$$g\left(\frac{S_n}{\sqrt{n}}\right) - g\left(\frac{T_n}{\sqrt{n}}\right) = \sum_{i=1}^{n}\left[g\left(\frac{S_n^{(j-1)}}{\sqrt{n}}\right) - g\left(\frac{S_n^{(j)}}{\sqrt{n}}\right)\right]$$

by simply using the telescoping sum and the fact that $S_n^{(0)} = S_n$ and $S_n^{(n)} = T_n$.

Then, note that

$$\begin{aligned}
S_n^{(j-1)} &= R_j + X_j \\
S_n^{(j)} &= R_j + Z_j
\end{aligned}$$

where

$$R_j = Z_1 + .. + Z_{j-1} + X_{j+1} + \ldots + X_n.$$

Finally, R_j, X_j, Z_j are mutually independent. Key idea is to take a third order Taylor-Young expansion of g around r

$$g\left(r+x\right) = g\left(r\right) + xg'\left(r\right) + \frac{x^2}{2}g''\left(r\right) + \frac{x^3}{3!}g'''\left(r'\right)$$

where $r' \in [r, r+x]$. Applying this to R and X, which are independent

$$\begin{aligned}\mathbb{E}\left[g\left(R+X\right)\right] &= \mathbb{E}\left[g\left(R\right)\right] + \mathbb{E}\left[Xg'\left(R\right)\right] + \mathbb{E}\left[\frac{X^2}{2}g''\left(R\right)\right] + \mathbb{E}\left[\frac{X^3}{3!}g'''\left(R'\right)\right] \\ &= \mathbb{E}\left[g\left(R\right)\right] + \mathbb{E}\left[X\right]\mathbb{E}\left[g'\left(R\right)\right] + \mathbb{E}\left[\frac{X^2}{2}\right]\mathbb{E}\left[g''\left(R\right)\right] + \mathbb{E}\left[\frac{X^3}{3!}g'''\left(R'\right)\right]\end{aligned}$$

Next, apply this to $R = R_j/\sqrt{n}$ and $X = X_j/\sqrt{n}$.

$$\begin{aligned}&\mathbb{E}\left[g\left(\frac{S_n^{(j-1)}}{\sqrt{n}}\right)\right] = \mathbb{E}\left[g\left(\frac{R_j}{\sqrt{n}} + \frac{X_j}{\sqrt{n}}\right)\right] \\ &= \mathbb{E}\left[g\left(\frac{R_j}{\sqrt{n}}\right)\right] + \mathbb{E}\left[\frac{X_j}{\sqrt{n}}\right]\mathbb{E}\left[g'\left(\frac{R_j}{\sqrt{n}}\right)\right] + \frac{1}{2}\mathbb{E}\left[\left(\frac{X_j}{\sqrt{n}}\right)^2\right]\mathbb{E}\left[g''\left(\frac{R_j}{\sqrt{n}}\right)\right] \\ &\quad + \frac{1}{3!}\mathbb{E}\left[\left(\frac{X_j}{\sqrt{n}}\right)^3 g'''\left(\frac{R_j'}{\sqrt{n}}\right)\right] \\ &= \mathbb{E}\left[g\left(\frac{R_j}{\sqrt{n}}\right)\right] + 0\mathbb{E}\left[g'\left(\frac{R_j}{\sqrt{n}}\right)\right] + \frac{1}{2n}\mathbb{E}\left[g''\left(\frac{R_j}{\sqrt{n}}\right)\right] + \frac{1}{6n^{3/2}}\mathbb{E}\left[X_j^3 g'''\left(\frac{R_j'}{\sqrt{n}}\right)\right]\end{aligned}$$

In the same way,

$$\begin{aligned}\mathbb{E}\left[g\left(\frac{S_n^{(j)}}{\sqrt{n}}\right)\right] &= \mathbb{E}\left[g\left(\frac{R_j}{\sqrt{n}}+\frac{Z_j}{\sqrt{n}}\right)\right]\\ &= \mathbb{E}\left[g\left(\frac{R_j}{\sqrt{n}}\right)\right]+0\cdot\mathbb{E}\left[g'\left(\frac{R_j}{\sqrt{n}}\right)\right]+\frac{1}{2n}\mathbb{E}\left[g''\left(\frac{R_j}{\sqrt{n}}\right)\right]\\ &\quad+\frac{1}{6n^{3/2}}\mathbb{E}\left[Z_j^3 g'''\left(\frac{\tilde{R}_j}{\sqrt{n}}\right)\right]\end{aligned}$$

Taking the difference between the two expressions, the third order terms remain

$$\mathbb{E}\left[g\left(\frac{S^{(j-1)}}{\sqrt{n}}\right)\right]-\mathbb{E}\left[g\left(\frac{S^{(j)}}{\sqrt{n}}\right)\right]=\frac{1}{6n^{3/2}}\left\{\mathbb{E}\left[X_j^3 g'''\left(\frac{R_j'}{\sqrt{n}}\right)\right]-\mathbb{E}\left[Z_j^3 g'''\left(\frac{\tilde{R}_j}{\sqrt{n}}\right)\right]\right\}$$

Bounding the third derivative by some absolute constant,

$$\left|\mathbb{E}\left[g\left(\frac{S^{(j-1)}}{\sqrt{n}}\right)\right]-\mathbb{E}\left[g\left(\frac{S^{(j)}}{\sqrt{n}}\right)\right]\right|\le\frac{c\left(g\right)}{n^{3/2}}$$

after assuming a finite third absolute moment.
Finally, after plugging back into the telescoping sum,

$$\begin{aligned}\mathbb{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right)\right]-\mathbb{E}\left[g\left(\frac{T_n}{\sqrt{n}}\right)\right] &\le\sum_{j=1}^{n}\left|\mathbb{E}\left[g\left(\frac{S_n^{(j-1)}}{\sqrt{n}}\right)\right]-\mathbb{E}\left[g\left(\frac{S_n^{(j)}}{\sqrt{n}}\right)\right]\right|\\ &\le\sum_{j=1}^{n}\frac{\tilde{C}}{n^{3/2}}=\frac{\tilde{C}}{\sqrt{n}}.\end{aligned}$$

□

Theorem (Lindeberg's CLT for triangular arrays)**.** *Let $X_{n,1},X_{n,2},\dots,X_{n,n}$ be mutually independent random variables. Let row sum*

$$\tilde{S}_n:=\sum_{k=1}^{n}X_{n,k}.$$

Assume that $\mathbb{E}\left[X_{n,k}\right] = 0, \forall n, k$. *Also assume that*

$$v_n := \sum_{k=1}^{n} \mathbb{E}\left[X_{n,k}^2\right] \to 1, \quad n \to \infty$$
$$g_n(\varepsilon) := \sum_{k=1}^{n} \mathbb{E}\left[X_{n,k}^2 \cdot \mathbb{1}_{|X_{n,k} \geq \varepsilon|}\right] \to 0, \quad n \to \infty.$$

Then

$$\widetilde{S_n} \Rightarrow \mathcal{N}(0,1).$$

Proof. Proof is basically the same as above, but requires a more clever argument about the remainder term in the Taylor expansion. Use truncation argument for $|X_{n,k}| \leq \varepsilon$ (from third order expansion) and $|X_{n,k}| > \varepsilon$ use second order expansion. □

3.2 Characteristic functions

Definition (Characteristic function)**.** *Given a RV X, its* ***characteristic function*** $\phi_X : \mathbb{R} \to \mathbb{C}$ *is defined by*

$$\phi_X(t) := \mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[\cos(tX)\right] + i\mathbb{E}\left[\sin(tX)\right].$$

This is, in other words, a Fourier transform.

Proposition (Characteristic functions)**.** *Below are a few simple properties.*

1. $\phi_X(0) = 1$.
2. $\phi_X(-t) = \overline{\phi_X(t)}$. *Therefore,* ϕ_X *is real* $\iff$ *the distribution of X is symmetric around zero i.e.* $X \overset{d}{=} -X$.
3. $|\phi_X(t)| \leq 1$.
4. $t \mapsto \phi_X(t)$ *is uniformly continuous on* $\mathbb{R}$.

   Proof.

$$\begin{aligned} |\phi(t) - \phi(s)| &= |\mathbb{E}\left[e^{itX} - e^{isX}\right]| \leq \mathbb{E}|e^{itX} - e^{isX}| && \text{(Jensen's inequality)} \\ &= \mathbb{E}\left[|e^{itX} - e^{isX}|\mathbb{1}_{|X| \leq M}\right] + \mathbb{E}\left[|e^{itX} - e^{isX}|\mathbb{1}_{|X| > M}\right] \\ &\leq |t-s| \cdot M + 2\mathbb{P}\left(|X| > M\right) \end{aligned}$$

   Given $\varepsilon > 0$, choose M such that $\mathbb{P}\left(|X| > M\right) \leq \frac{\varepsilon}{4}$. The, let $\delta := \varepsilon/2M$. □

5. *ϕ is positive definite in the following sense: $\forall n, \forall t_1, \dots, t_n \in \mathbb{R}$, the matrix $\left\{\phi\left(t_i - t_j\right)\right\}_{i,j=1}^{N}$*

$$\sum_{i,j=1}^{n} z_i \phi\left(t_i - t_j\right) \overline{z}_j \geq 0.$$

Proof.

$$\begin{aligned}\sum_{i,j=1}^{n} z_i \phi\left(t_i - t_j\right) \overline{z_j} &= \sum_{i,j=1}^{n} z_i \overline{z_j} \mathbb{E}\left[e^{i\left(t_i - t_j\right)X}\right] \\ &= \mathbb{E}\left[\sum_{i,j=1}^{n} z_i \overline{z_j} e^{it_iX} \overline{e^{it_jX}}\right] && \text{(Property 2)} \\ &= \mathbb{E}\left[|\sum_{k=1}^{n} z_k e^{itX}|^2\right] \geq 0\end{aligned}$$

□

Theorem (Bochner's)**.** *Let $\phi : \mathbb{R} \to \mathbb{C}$. If*

1. *$\phi(0) = 1$,*
2. *ϕ is continuous at $t = 0$,*
3. *ϕ is positive definite,*

then ϕ is a positive definite function for some RV. That is there exists a CDF F such that

$$\phi(t) = \int e^{itX} dF(x).$$

Note. *This is good to know but rarely used in practice.*

Property (Additional properties of characteristic functions)**.** *Below are a few useful properties.*

1. *Linear transformation of RVs. For constants a, b and random variable X,*

$$\phi_{aX+b}(t) = e^{itb} \phi_X(at)$$

Proof. Follows by definition $\mathbb{E}\left[e^{it(aX+b)}\right] = e^{itb} \mathbb{E}\left[e^{i(at)X}\right] = e^{itb} \phi_X(at)$. □

2. *If X, Y are independent, then $X + Y$*

$$\phi_{X+Y}(t) = \phi_X(t) \phi_Y(t).$$

Proof.

$$\phi_{X+Y}(t) = \mathbb{E}\left[e^{it(X+Y)}\right] = \mathbb{E}\left[e^{itX}\right]\mathbb{E}\left[e^{itY}\right] = \phi_X(t)\,\phi_Y(t)$$

□

3. *Fourier transform of density: If $F' = f$, then*

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx = \hat{f}(t)\,.$$

Example (Characteristic functions)**.** *Proofs of the below examples are left as exercises.*

1. ***Bernoulli***: $X \sim Be(p) \iff \phi_X(t) = e^{it} + 1 - p$.
2. ***Binomial***: $X \sim Bin(n,p) \iff \phi_X(t) = \left[pe^{it} + 1 - p\right]^n$ *from above and independence.*
3. ***Rademacher***: $\phi_X(t) = \frac{1}{2}\left[e^{it} + e^{-it}\right] = \cos(t)$.
4. ***Uniform***: $X \sim Uni(a,b) \iff \phi_X(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}$. *As a common example,* $X \sim Uni(-1,1) \iff \phi_X(t) = \frac{\sin(t)}{t}$.
5. ***Normal***: $X \sim \mathcal{N}\left(\mu, \sigma^2\right) \iff \phi_X(t) = e^{it\mu} - t^2\sigma^2/2$.

Proof. Suffices to show for standard Gaussian because of result on translation and scaling. Before beginning, note that

$$\begin{aligned}\mathbb{E}\left(iXe^{itX}\right) &= \mathbb{E}\left[iX\left(\cos(tX) + i\sin(tX)\right)\right] = i\mathbb{E}\left(X\cos(tX)\right) - \mathbb{E}\left[X\sin(tX)\right] \\ &= -\mathbb{E}\left[X\sin(tX)\right].\end{aligned}$$

Since a Gaussian is symmetric and the cosine function is odd, the cosine term must equal zero.

Proceeding with the proof, take the derivative with respect to t of the characteristic function

$$\phi_X'(t) = \mathbb{E}\left[iXe^{itX}\right] = -\int_{\mathbb{R}} x\sin(tx)\,f_X(x)\,dx.$$

Then, from the Fourier transform property and the result above, $f_X'(x) = -xf_X(x)$. Plugging this back into the expression and performing integration by parts in the second equality,

$$\phi_X'(t) = \int_{\mathbb{R}} \sin(tx)\,f_X'(x)\,dx = -\int_{\mathbb{R}} t\cos(tx)\,f_X(x)\,dx = -t\phi_X(t)\,.$$

Then, $\phi'(t) = -t\phi(t), \phi(0) = 1$. Thus $\phi(t) = e^{-t^2/2}$.

□

6. ***Exponential:*** *$X \sim Exp(\lambda) \iff \phi_X(t) = \frac{\lambda}{\lambda - it}$.*
7. ***Cauchy:*** *$\phi_X(t) = e^{-|t|}$.*

Proposition. *If $\mathbb{E}\left[|X|^k\right] < \infty$, then $\phi_X \in C^k$, i.e. it is continuous and k-times differentiable, and the derivatives are given by*

$$\phi_X^{(k)}(t) = \mathbb{E}\left[(iX)^k e^{itX}\right].$$

In particular, for $t = 0$,

$$\phi_X^{(k)}(0) = i^k \mathbb{E}\left(X^k\right).$$

Proof. Consider the case for $k = 1$. Iterate forward for larger k.

$$\begin{aligned}
\phi'(t) &= \frac{d}{dt}\phi(t) = \frac{d}{dt}\mathbb{E}[itX] \\
&= \lim_{h\to 0}\frac{1}{h}\left(\mathbb{E}\left[e^{i(t+h)X}\right] - \mathbb{E}\left[e^{itX}\right]\right) && \text{(Definition of derivative)} \\
&= \lim_{h\to 0}\mathbb{E}\left[\frac{1}{h}\left(e^{i(t+h)X} - e^{itX}\right)\right] && \text{(Linearity of expectation)} \\
&= \mathbb{E}\left[\lim_{h\to 0}\frac{1}{h}\left(e^{i(t+h)X} - e^{itX}\right)\right] && \text{(DCT)} \\
&= \mathbb{E}\left[iXe^{itX}\right] && \text{(Definition of derivative } \tfrac{d}{dt}\phi(t)\text{)}
\end{aligned}$$

Note that the order of the limit and expectation can be interchanged because of the Dominated Convergence Theorem with dominating function $2|x|+1$. We need to bound

$$\left|\frac{1}{h}\left(e^{i(t+h)X} - e^{itX}\right)\right|$$

as a sequence of h. □

Proposition (Taylor expansion of characteristic function about $t = 0$)**.** *If $\mathbb{E}\left[|X|^m\right] < \infty$, then*

$$\phi_X(t) = \sum_{k=0}^{m}\mathbb{E}\left[(iX)^k\right]\frac{t^k}{k!} + o\left(t^m\right).$$

Proof. Consider the Taylor expansion of e^{ax} about $x = 0$

$$e^{ax} = 1 + ax + a^2\frac{x^2}{2} + a^3\frac{x^3}{3!} + \cdots a^n\frac{x^n}{n!} + o\left(x^n\right)$$

Hence, regrouping, substituting a with iX, and taking expectations on both sides

$$\mathbb{E}\left[e^{itX}\right] = \sum_{k=0}^{n} \mathbb{E}\left[(iX)^k\right] \frac{t^k}{k!} + o\left(t^n\right)$$

□

Theorem (CLT and characteristic functions)**.** *$X_1, X_2, \ldots$ are iid with mean 0 and variance 1 (i.e. we are assuming finite second moments). Let $S_n = X_1 + \ldots + X_n$. Then*

$$\frac{S_n}{\sqrt{n}} \Rightarrow \mathcal{N}(0,1).$$

Proof. Below is a sketch of the proof. Consider the characteristic function of $\frac{S_n}{\sqrt{n}}$

$$\phi_{S_n/\sqrt{n}}(t) = \phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) = \prod_{i=1}^{n} \phi_{X_i}\left(\frac{t}{\sqrt{n}}\right) = \left[\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n.$$

The first equality follows from scaling. The last two equalities hold from i. and i.d., respectively. Fix $t \in \mathbb{R}$ and take a Taylor expansion of ϕ_{X_1} about $t = 0$. Assuming finite first/second moments we get

$$\phi_{X_1}(s) = 1 + i\mathbb{E}\left[X_1\right]s - \frac{1}{2}\mathbb{E}\left[X_1^2\right]s^2 + o\left(s^2\right) = 1 - s^2/2 + o\left(s^2\right), \quad s \to 0.$$

where the minus sign after the first equality comes from $i^2 = -1$ and we used the mean zero and variance one assumptions in the second equality.
If we think of $t/\sqrt{n}$ with t fixed and as $n \to \infty$, then $\phi_{X_1}\left(t/\sqrt{n}\right) = 1 - \frac{t^2}{2n} + o\left(t^2/n\right)$. Plugging into the previous expression,

$$\phi_{S_n/\sqrt{n}}(t) = \left(\phi_{X_1}\left(t/\sqrt{n}\right)\right)^n = \left(1 - \frac{t^2}{2n} + o\left(1/n\right)\right)^n \to e^{-t^2/2},$$

which is the characteristic function of a standard Gaussian. □

Note. *We did not show uniqueness of characteristic functions or convergence properties. Neither are trivial.*
Note that we still need to show that characteristic functions characterize distributions and convergence of characteristic functions implies convergence in distribution.

The following theorem is a technical one, but its corollary implies that if two random variables have the same characteristic function, then they are equal.

Theorem (Lèvy's inversion). *Suppose that X has CDF F_X and characteristic function ϕ_X. For every $a < b$ and t, define*

$$\psi_{a,b}(t) := \frac{1}{2\pi}\int_a^b e^{-itu}du = \frac{e^{-itb} - e^{-ita}}{-2\pi it}.$$

Then

$$\lim_{T\to\infty}\int_{-T}^{T}\psi_{a,b}(t)\,\phi_X(t) = \frac{1}{2}\left[F_X(b) + F_X(b-)\right] - \frac{1}{2}\left[F_X(a) + F_X(a-)\right].$$

where $F_X(x-) := \lim_{s\to x^-} F(s)$. When a, b are continuity points of F, then the limit is $F_X(b) - F_X(a) = \mathbb{P}(a < X \leq b)$. Furthermore, if $\int |\phi_X(t)| < \infty$, then X has a bounded continuous probability density function f_X and

$$f_X(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\phi_X(t)\,dt.$$

Note. *A few points on the theorem:*

- *This is a special case of Fourier inversion formula (integrated). Holds even in the absence of a density function.*
- *If the characteristic function of a random variable is integrable, then it admits pdf and there is a way to derive the pdf from the characteristic function*

Corollary. *If X, Y have the same characteristic functions (that is $\phi_X(t) = \phi_Y(t)$) them $X \overset{d}{=} Y$.*

Proof. Exercise using the formula above. □

Theorem (Lèvy's continuity). *Let $\{F_n\}_{n\geq 1}$ be a sequence of CDFs with characteristic functions $\{\phi_n\}_{n\geq 1}$.*

1. *If $F_n \Rightarrow F$ then $\phi_n(t) \to \phi(t) = \int_{\mathbb{R}} e^{itx}dF(x), \forall t \in \mathbb{R}$.*
2. *Suppose that $\forall t \in \mathbb{R}$, $\lim_{n\to\infty}\phi_n(t)$ exists and denote it by $\phi(t)$. Suppose that ϕ is continuous at $t = 0$. Then there exists CDF F such that*

$$\phi(t) = \int_{\mathbb{R}} e^{itx}dF(x), \quad \forall t \in \mathbb{R}$$

and $F_n \Rightarrow F$.

Note. *The theorem is not an if and only if because for 2. we require the stronger condition that ϕ is continuous at $t = 0$. More importantly, we have now a new tool to show weak convergence of random variables.*

Definition (Tightness in $\mathbb{R}$). *A sequence of probability measures on $\mathbb{R}$ $\{F_n\}_{n\geq 1}$ is* ***uniformly tight*** *if for every $\varepsilon > 0$, there exists $K_\varepsilon < \infty$ such that*

$$F_n\left([-K_\varepsilon, K_\varepsilon]\right) \geq 1 - \varepsilon, \forall n \geq 1.$$

Note. *Uniform tight random variables are often called* ***bounded in probability****. The "uniform" bit comes from the fact that the same K_ε can be chosen for all elements of the sequence.*

Example (Tightness). *Tightness is about requiring that most of the mass does not escape to infinity*

1. *$U(-n, n)$ is not tight because there is no finite interval containing most of the mass. If you visualize the pdf of a continuous uniform it always has 'fat tails'.*
2. *$\{\delta_n\}_{n\geq 1}$ is not tight for the same reason. Indeed, the pmf is a unique point that is pushed to infinity.*
3. *$\left\{\delta_{(-1)^n}\right\}$ is tight, simply take $K = 1$. The pmf is a unique point (either 1 or -1).*

Note in the last case the series of random variable does not converge weakly, however it is tight. So tightness is a weaker requirement than weak convergence.

Definition (Tightness (general)). *Let S be a complete metric space. A sequence of measures $\{\mu_n\}_{n\geq 1}$ is* ***tight*** *if $\forall\, \varepsilon > 0$, there exists compact set $K \subset S$ such that*

$$\mu_n\left(K^c\right) \leq \varepsilon, \forall n \geq 1.$$

Proposition. *If $F_n \Rightarrow F$, then $\{F_n\}_{n\geq 1}$ is tight.*

In general, the converse is not true.

Theorem (Helly). *If $\{F_n\}_{n\geq 1}$ is tight, there exists a subsequence $\{n_k\}_{k\geq 1}$ and F such that $F_{n,k} \Rightarrow F$ as $k \to \infty$.*

Theorem (Prohorov). *Let S be a complete separable metric space. If $\{\mu_n\}_{n\geq 1}$ is tight, then there exists a weakly convergent subsequence $\{\mu_{n_k}\}_{k\geq 1}$.*

Note (Framework for showing weak convergence (Prohorov)). *Below are a few general steps, broadly applicable for many problems.*

1. *Show tightness.*
2. *Identify the limit.*
3. *Show uniqueness of the limit.*

4 Stochastic processes

4.1 Introduction and overview

Setup

- $(\Omega, \mathcal{F}, \mathbb{P})$ probability space
- $\mathbb{T}$ index set representing time (eg. $\mathbb{R}, \mathbb{R}_+, [a,b], \mathbb{Z}, \mathbb{Z}_+, \ldots$).
- S state space, usually a locally compact, complete metric space.

Definition (Stochastic process)**.** *A stochastic process is a measurable function $t \mapsto X_t$, $X : \mathbb{T} \times \Omega \to S$ or, alternatively, $\forall t \in \mathbb{T}$, $X(t, \cdot) : \Omega \to S$. Consider the following characterizations*

1. *$t \in \mathbb{T}$ fixed, then $X_t(t, \cdot) : \Omega \to S$ is an $S-$valued **random variable** (marginal).*
2. *$\omega \in \Omega$ fixed. Then $X(\cdot, \omega) : \mathbb{T} \to S$ is a **random trajectory**.*
3. *$X : \Omega \to S^T = \{\text{space of } S\text{-valued functions with domain } T\}$*

Example (Stochastic processes)**.** *Below are a few motivating examples.*

1. *$\{X_n\}_{n \geq 0}$ iid RVs, where the index set is $\mathbb{N}$.*
2. *$X_n = \sum_{i=1}^{n} Y_i$ where Y_i are iid. Such process is called a **random walk**.*
   - *RW on $\mathbb{Z}, \mathbb{Z}^d$.*
   - ***Simple random walk** defines $Y_i \in \{\pm 1\}$.*
   - ***Simple symmetric random walk** assigns SRW with probabilities 1/2 and 1/2.*
   - ***Biased random walk** assigns SRW different probabilities (eg. gambler's ruin)*
3. *Can also define random walks on graphs, like Google's PageRank algorithm.*
4. *Epidemics (or their spread) can be modeled as a stochastic process.*
5. *Relatedly, stories on social media can be modeled as stochastic processes.*
6. *Queuing theory.*
7. *Card shuffling.*
8. *Brownian motion.*
9. *Extreme value theory: modeling rare events $e.g.M_n := \max_{1 \leq 1 \leq n} X_i$.*
10. *MCMC.*

4.2 Markov processes

Intuitively, the key defining property is that conditioning on the past and present is the same as conditioning on just the present. In the below discussion, will focus on the simplest case (discrete time/space).

Definition (Markov chain)**.** *Let $\{X_n\}_{n\geq 0}$ be a stochastic process taking values in some state space S. We say that $\{X_n\}_{n\geq 0}$ is a **Markov chain** on a discrete state space S if*

$$\mathbb{P}\left(X_{n+1} = a_{n+1} | X_n = a_n, X_{n-1} = a_{n-1}, \ldots, X_0 = a_0\right) = \mathbb{P}\left(X_{n+1} = a_{n+1} | X_n = a_n\right)$$

*for all $n \geq 0$ and for all $a_0, a_1, \ldots, a_{n+1} \in S$. This is what is usually denoted as **Markov property**. Will note $\underline{a}_0^n := (a_0, \ldots, a_n)$.*

Note. *The Markov property tells us that to learn about the future the only thing we need is the past. Once we condition on the present, the past and the future become independent of each other.*

Property (Decomposition of Markov Processes)**.** *Markov chains can be decomposed over time as*

$$\begin{aligned}
\mathbb{P}\left(\underline{X}_0^n = \boldsymbol{a}_0^n\right) &= \mathbb{P}\left(X_n = a_n \mid \underline{X}_0^{n-1} = \underline{a}_0^{n-1}\right) \mathbb{P}\left(\underline{X}_0^{n-1} = \underline{a}_0^{n-1}\right) && \text{(intersection)}\\
&= \ldots && \text{(Iterate)}\\
&= \mathbb{P}\left(X_n = a_n \mid \underline{X}_0^{n-1} = \underline{a}_0^{n-1}\right) \mathbb{P}\left(X_{n-1} = a_{n-1} \mid \underline{X}_0^{n-2} = \underline{a}_0^{n-2}\right)\\
&\quad \cdot \ldots \cdot \mathbb{P}\left(X_1 = a_1 \mid X_0 = a_0\right) \mathbb{P}\left(X_0 = a_0\right)\\
&= \mathbb{P}\left(\underline{X}_0^n = \boldsymbol{a}_0^n\right) = \mathbb{P}\left(X_0 = a_0\right) \prod_{j=1}^{n} \mathbb{P}\left(X_j = a_j \mid X_{j-1} = \underline{a}_0^{j-1}\right) && \text{(true in general)}\\
&= \mathbb{P}\left(\underline{X}_0^n = \boldsymbol{a}_o^n\right) = \mathbb{P}\left(X_0 = a_0\right) \prod_{j=1}^{n} \mathbb{P}\left(X_j = a_j \mid X_{j-1} = a_{j-1}\right)\\
& && \text{(Markov property)}
\end{aligned}$$

Definition (Transition/Stochastic matrix)**.** *A **transition** or **stochastic matrix** is defined as $P_{xy}(j) := \mathbb{P}\left(X_j = y | X_{j-1} = x\right)$, where $P_{xy}(j)$ is termed **transition probability**. In many cases, transition probabilities do not depend on time. These are called **time homogenous Markov chains**.*

Note. *$P_{xy}(j)$ is the probability that the MC switches from x to y at time j. If the MC is time homogeneous the same $P_{xy}(j)$ just depends on the time shift j and not on the period t.*

Definition (Stochastic matrix)**.** *Matrix P is a **stochastic matrix** if*

1. *$P_{xy} \geq 0$, i.e. each element is non-negative*
2. *$\sum_{y\in S} P_{xy} = 1$ rows sum to one.*

Note. *The two requirements imply that $P_{xy}\left(j\right) \in [0,1]$. Moreover, each row represents a probability mass function that depends on x (and j of course). Each row hence represents the probability we reach any state starting from x in j steps.*

Property. *Time homogenous Markov chains are entirely summarized by*

1. *an initial distribution $\mathbb{P}\left(X_0 = a_0\right)$*
2. *a transition matrix P*

Indeed, for time-homogenous MCs

$$\mathbb{P}\left(\underline{X}_0^n = \underline{a}_0^n\right) = \mathbb{P}\left(X_0 = a_0\right) \prod_{j=1}^{n} P_{a_{j-1},a_j}$$

Proposition. $\mathbb{P}\left(X_n = y \mid X_0 = x\right) = \left(P^n\right)_{xy}$.

Proof. Use induction. The base step $n = 1$ follows from definition, indeed

$$\mathbb{P}\left(X_1 = y \mid X_0 = x\right) := P_{xy}$$

. Proceed to the inductive step. Suppose we know the statement is true for n, i.e.

$$\mathbb{P}\left(X_n = y \mid X_0 = x\right) = \left(P\right)^n_{xy}$$

Then

$$\begin{aligned}
\mathbb{P}\left(X_{n+1} = y \mid X_0 = x\right) &= \sum_{z \in S} \mathbb{P}\left(X_{n+1} = y, X_n = z \mid X_0 = x\right) && \text{(law of total probability)} \\
&= \sum_{z \in S} \mathbb{P}\left(X_{n+1} = y \mid X_n = z, X_0 = x\right) \mathbb{P}\left(X_n = z \mid X_0 = x\right) \\
&= \sum_{z \in S} \mathbb{P}\left(X_{n+1} = y \mid X_n = z\right) \mathbb{P}\left(X_n = z \mid X_0 = x\right) && \text{(Markov property)} \\
&= \sum_{z \in S} \left(P^n\right)_{xz} P_{zy} && \text{(Inductive hypothesis)} \\
&= \left(P^{n+1}\right)_{xy}.
\end{aligned}$$

□

Note (Interpretation of transition matrix). *Note two things*

- *Column vectors → think as functions. The transition matrix acts forward on column vectors*
- *Row vectors → think as measures. The transition acts backward on row vectors*

1. ***Probabilistic interpretation**: Let $f : S \to \mathbb{R}$ be a column vector. Then*

$$\left(P^n f\right)(x) = \mathbb{E}\left[f\left(X_n\right) \mid X_0 = x\right].$$

Proof.

$$\begin{aligned}
\left(P^n f\right)(x) &= \sum_{y \in S} \left(P^n\right)_{xy} f\left(y\right) \\
&= \sum_{y \in S} \mathbb{P}\left(X_n = y \mid X_0 = x\right) f\left(y\right) && \text{(From previous claim)} \\
&= \mathbb{E}\left[f\left(X_n\right) | X_0 = x\right].
\end{aligned}$$

□

Therefore, if we are interested in the expectation of any transformation $f(\cdot)$ of a Markov Chain, we just need to compute $P^n f$

2. *$\mu : S \to \mathbb{R}_+$; $\sum_{x \in S} \mu(x) = 1$; μ is a probability measure on S. This is a row vector. If $\mathbb{P}\left(X_0 = y\right) = \mu\left(y\right)$, then $\left(\mu P^n\right) = \mathbb{P}\left(X_n = x\right)$.*

Proof.

$$\begin{aligned}
\left(\mu P^n\right)(x) &= \sum_{y \in S} \mu\left(y\right) \left(P^n\right)_{yx} \\
&= \sum_{y \in S} \mu\left(y\right) \mathbb{P}\left(X_n = x \mid X_0 = y\right) \\
&= \sum_{y \in S} \mathbb{P}\left(X_0 = y\right) \mathbb{P}\left(X_n = x \mid X_0 = y\right) \\
&= \mathbb{P}\left(X_n = x\right).
\end{aligned}$$

□

Proposition. *If the initial distribution is μ, then the distribution at time n is μP^n.*

Definition (Stationary distribution)**.** *If $\pi P = \pi$, then $\pi P^n = \pi$, so if the initial distribution is π, then the distribution at every time n is π. We call such π a **stationary distribution**.*

4.3 Classification of states

Definition (Closed)**.** *$A \subset S$ is **closed** if* $\mathbb{P}\left(A \to A^c\right) = 0$.

Example. *The sets $\emptyset$ and S are closed. If A and B are closed, so are $A \cup B$ and $A \cap B$*

Definition (Closure)**.** *Let $A \subset S$. The **closure** of A is*

$$\overline{A} := \cap_{A \subseteq B, B \text{ is closed}} B.$$

In words, the closure of A is the smallest closed set containing A.

Definition (Irreducible)**.** *Suppose $A = \overline{A}$ and A non-empty. A is **irreducible** if*

$$\forall\, B \subseteq A, \overline{B} = A, B \neq \emptyset$$

Another definition is that P is irreducible if and only if $\forall\, x, y \in S, \exists\, n_0 = n_0\left(x, y\right)$ such that $\left(P^{n_0}\right)_{xy} > 0$.

Note. *There can be closed sets that are not irreducible, for example the set composed by two closed sets is not irreducible because you can always split it in two irreducible smaller sets. If we add the property of aperiodicity, then n_0 is uniform for all points in the state space.*

Definition (Absorbing state)**.** *$x \in S$ is an **absorbing state** if $\{x\}$ is closed.*

Definition (Inessential state)**.** *$x \in S$ is an **inessential state** if it is not part of any irreducible component. If x is not inessential, it is **essential**. An alternative terminology is **transient** and **recurrent**. Recurrent states can be either **positive recurrent** (finite number of steps to come back) or **null recurrent** (the markov chain will return in the state with probability 1 but in an infinite number of steps).*

Definition (Irreducible Markov Chain)**.** *A Markov Chain is irreducible if S is irreducible.*

Proposition (State space decomposition)**.** *If S is finite, then*

$$S = C_1 \cup \cdots \cup C_r \cup D$$

for irreducible components C_j and inessential states D.

Example (Classification example)**.** *Consider Figure **??**, which shows a connected graph of some Markov process.*

- ***Closed:*** $\{C, D\}, \{E, F, G\}$
- $\overline{\{C\}} = \{C, D\}$

- *$\{E, F, G\}$, $\{C, D\}$ are **irreducible***
- *$\{H\}$ is an **absorbing state***
- *$\{C, D\}$, $\{H\}$, $\{E, F, G\}$ are irreducible components. $\{A, B\}$ give inessential states*

In the long-run, inessential states do not matter. Will focus primarily on irreducible Markov chains. Can use **first step analysis** (conditioning on first step of Markov chain), possible to compute

- Probabilities of ending up in different irreducible parts of the a Markov chain.
- Expected time to reach an irreducible component.

Proposition. *$C \subseteq S$ is irreducible if and only if $C = \overline{C}$ (C is minimal closed) and $\forall x, y \in C$, there exists a path from x to y.*

Proposition. *If S is finite, then there always exists a non-empty irreducible component.*

Proposition. *If S is not finite, then there does not need to exist an irreducible component.*

Definition (Period)**.** *$x \in S$. **Period** $per(x) :=$ GCD of lengths of walks returning to x. In the example, $per(H) = 1, per(D) = GCD(2,3) = 1, per(E) = GCD(3,6,9,\ldots) = 3$.*

Proposition. *If $x, y \in S$ are in the same irreducible component, then $per(x) = per(y)$.*

Corollary. *The period of all states within an irreducible component is the same.*

Definition. *The period of an irreducible component is the period of any of its components. The period of an irreducible Markov Chain is the period of any of its components.*

Definition (Aperiodic)**.** *A MC is **aperiodic** if the period is 1.*

Example. *The period of a simple random walk on $\mathbb{Z}$ is 2. Even times are associated with even locations. Odd times are associated with odd locations.*

Definition. *A state is **ergodic** if it is positive recurrent and aperiodic.*

Definition (Ergodic)**.** *P is **ergodic** if it is irreducible and aperiodic.*

Definition (Stationary distribution)**.** *π is a **stationary distribution** of the Markov chain with transition matrix P if $\pi P = \pi$, where π is a probability distribution on S (i.e. $\pi(x) \geq 0, \sum_{x \in S} \pi(x) = 1$).*

Natural questions about the stationary distribution of a Markov Chain regard: existence, uniqueness, and convergence in law.

Example. *Below are a few simple examples.*

1. *RW on a connected undirected graph $G = (V, E)$.*
2. *RW on a disconnected graph. Suppose nodes A, B are disconnected from nodes C, D. Then $\pi^1 = \left(1/2, 1/2, 0, 0\right)$ and $\pi^{(2)} = \left(0, 0, 1/2, 1/2\right)$ are stationary distributions for $(\pi_A, \pi_B, \pi_C, \pi_D)$. There is actually an entire one-parameter family of stationary distributions $\left(\alpha, \alpha, \alpha - 1/2, \alpha - 1/2\right)$, for $\alpha \in \left[0, 1/2\right]$*
3. *Suppose P is **doubly stochastic**, that is also its column sums are one. Then π is uniform, that is*
$$\pi_x = \frac{1}{|S|}, \quad \forall\, x \in S$$
There is not necessarily a unique stationary distribution (look at point 2)
4. *Suppose that P, π satisfy*
$$\pi_x P_{xy} = \pi_y P_{yx}, \quad \forall x, y \in S.$$
*Above are known as the **detailed balance equations**. Then we say that P is reversible with respect to π. In this case, π is a stationary distribution of P.*

   Note. *Can interpret $\pi_x P_{xy}$ as a probability/mass flow of x to y in stationarity (and vice versa). At equilibrium, there is "microscopic" reversibility. Can't distinguish the direction of time, hence the name.*
5. *It is often useful to interpret the action of the matrix P when left or right multiplied. If it is left multiplied by a vector μ_0 whose components are the probability that the chain is in a certain state, i.e.*
$$\mu_0 = \begin{pmatrix} \mathbb{P}\left(X_0 = 1\right) \\ \mathbb{P}\left(X_0 = 2\right) \\ \vdots \\ \mathbb{P}\left(X_0 = |S|\right) \end{pmatrix}$$
then the resulting element of $\mu_0 P$ is
$$\left(\mu_0 P\right)_s = \sum_{i=1}^{|S|} \mu_{0i} P_{is} = \sum_{i=1}^{|S|} \mathbb{P}\left(X_1 = s \mid X_0 = i\right) \mathbb{P}\left(X_0 = i\right) = \sum_{i=1}^{|S|} \mathbb{P}\left(X_1 = s, X_0 = i\right) = \mathbb{P}\left(X_1 = s\right)$$

What happens if instead the transition matrix P is right multiplied by a vector X? Simply we compute the expected value

$$(PX)_i = \sum_{s=1}^{|S|} P_{is} X_s = \sum_{s=1}^{|S|} \mathbb{P}\left(X_1 = s \mid X_0 = i\right) X_s = \mathbb{E}\left[X_s \mid X_0 = i\right]$$

therefore PX is a vector containing the expected value of X conditional on a specific past value of X_0

4.4 Existence

Property (P acting forward on functions)**.** *Natural space to consider is*

$$l^\infty(s) = \left\{ f : S \to \mathbb{R}, \|f\|_\infty < \infty \right\}.$$

1. *P keeps positivity. If $f \geq 0$ (coordinate-wise), then $Pf \geq 0$.*
2. *P is a contraction $\|P\|_{\infty,\infty} \leq 1$, where $||P||_{\infty,\infty} = \sup_{f \neq 0} \frac{||Pf||_\infty}{||f||_\infty}$.*

   Proof.

   $$\begin{aligned} \|Pf\|_\infty &= \max_x |Pf(x)| = \max_x |\sum_y P_{xy} f(y)| \\ &\leq \max_x \sum_y P_{xy} |f(y)| \leq \max_x \|f\|_\infty \sum_y P_{xy} = \|f\|_\infty. \end{aligned}$$

   The first equality follows from the definition of the infinity norm and the definition of $Pf(x)$. □

3. *$P\mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ is a vector of ones. Equivalently, 1 is an eigenvalue of P.*

Property (P acting backwards on measures)**.** *The natural space here is the one of finite measures, i.e. $l^1(s) = \left\{ \mu : S \to R : \|\mu\|_1 < \infty \right\}$.*

1. *Positivity: if $\mu \geq 0$, then $\mu P \geq 0$.*
2. *P is a contraction, then $||P||_{1,1} \leq 1$*

   Proof.

   $$\begin{aligned} \|\mu P\|_1 &= \sum_x |\sum_y \mu(y) P_{yx}| \\ &\leq \sum_x \sum_y |\mu(y)| P_{yx} && \text{(Triangle inequality)} \\ &= \sum_y |\mu(y)| \sum_x P_{xy} = \|\mu\|_1. \end{aligned}$$

□

3. *There exists μ such that $\mu P = \mu$. We saw above that 1 is an eigenvalue when P acts on the right, so it is also an eigenvalue when P acts on the left.*

Lemma. *If $\mu P = \mu$, then $|\mu|P = |\mu|$.*

Proof. Let $|\mu|(x) := |\mu\left(x\right)|$. Then,

$$\left(|\mu|P\right)\left(x\right) = \sum_y |\mu\left(y\right)| P_{yx} \geq |\sum_y \mu\left(y\right) P_{yx}| = |\mu\left(x\right)|.$$

The inequality follows from the triangle inequality. Then $\left(|\mu|P\right)\left(x\right) \geq |\mu|(x), \forall x$. If there exists an x such that the inequality holds strictly, then $\left\||\mu|P\right\|_1 > \left\||\mu|\right\|_1$, which is a contradiction since P is a contraction. Thus, the statement must hold with equality. □

Corollary. *$\mu P = \mu$ has a nonnegative solution. Therefore, we can always take the absolute vaue and normalize the solution by its L^1 norm, thus what we obtain is a distribution because the solution has only positive entries and L^1 norm equal to 1.*

Note. *We used that the state space is finite. Indeed, we just consider the space of functions with finite L^1 norm so that we are able to normalize the solution at the end. If we consider a RW on the integers then there is no stationary distribution because we cannot normalize at the end (the L^1 norm is not finite).*

4.5 Uniqueness

Proposition. *The dimension of the eigenspace is the number of irreducible components of the transition matrix P, i.e.*

$$\dim\left\{f : Pf = f\right\} = \textit{number of irreducible components of } P$$

.

Proof. Will first consider the case for irreducible P. Proof for general P is left as an exercise.

1. Suppose P is irreducible. Want to show that the dimension of the eigenspace is 1. We know that if f is a constant vector

$$f = c \cdot \mathbb{1},$$

then $Pf = f$. Assume that $\dim \left\{ f : Pf = f \right\} \geq 2$. Then there must exist f such that f is nonconstant and $Pf = f$. Define

$$A := \left\{ x \in S : f\left(x\right) = \max_{y \in S} f\left(y\right) \right\}.$$

Since S is finite, $A \neq \emptyset$. f nonconstant implies $A \neq S$.
Since P is irreducible, there must exist $x \in A, y \in A^c$ such that $P_{xy} > 0$, since P is irreducible. Now we know that $f\left(x\right) = \max_{z \in S} f\left(z\right) \equiv M$, and $Pf = f$, so

$$f\left(x\right) = \sum_{x \in S} P_{xz} f\left(z\right) = P_{xy} f\left(y\right) + \sum_{z \in S, z \neq y} P_{xz} f\left(z\right) \leq P_{xy} f\left(y\right) + \left(1 - P_{xy}\right) M < M$$

which is a contradiction.

2. More generally, $\geq$ direction is easy. On each irreducible component C, define $f = \alpha 1 c$.

□

Corollary. *The stationary distribution is unique if and only if P is irreducible.*

4.6 Convergence of Markov chains

Definition (Coupling). *If X and Y are two random variables, then a coupling $\left(X', Y'\right)$ is a pair of RVs defined on the same probability space such that $X' =^d X$ and $Y' =^d Y$.*

Note. *A few brief comments on coupling.*

1. *A trivial but useless coupling is taking X' and Y' independent with appropriate marginals.*
2. *Often define coupling with one of two goals in mind.*
   (a) *$X' \leq Y'$ a.s. (**stochastic domination**)*
   (b) *Minimize $\mathbb{P}\left(X' \neq Y'\right)$.*

Example. *Suppose $X \sim Bin\left(n, p\right)$ and $Y \sim Bin\left(m, p\right)$ with $m > n$. We have that for all $z \in \mathbb{R}$,*

$$\mathbb{P}\left(X \geq z\right) \leq \mathbb{P}\left(Y \geq z\right).$$

We can show the above inequality more formally using coupling.

Proof. Let $Z_1, Z_2, \ldots, Z_m$ be i.i.d. Be(p) RVs. Let $X' := \sum_{i=1}^n Z_i$ and $Y' := \sum_{i=1}^m Z_i$. Clearly the marginals are consistent. Since $Y' = X' + \sum_{i=n+1}^m Z_i \geq X'$,

$$\mathbb{P}\left(Y \geq z\right) = \mathbb{P}\left(Y' \geq z\right) \geq \mathbb{P}\left(X' \geq z\right) = \mathbb{P}\left(X \geq z\right).$$

□

Theorem (Markov chain convergence)**.** *Let S be a finite state space. Let P be an ergodic transition matrix (i.e. irreducible and aperiodic). Let π denote the stationary distribution of P. For every $x \in S$, if $X_0, X_1, \ldots$ is a Markov chain with $X_0 = x$ and transition matrix P, then*

$$X_n \Rightarrow \pi.$$

Moreover, $\exists\, \varepsilon > 0$, $C < \infty$ such that for all n,

$$\max_{x \in S} \sum_{y \in S} |\mathbb{P}\left(X_n = y \mid X_0 = x\right) - \pi\left(y\right)| \leq C\left(1 - \varepsilon\right)^n.$$

Proof. Let $\{X_n\}_{n \geq 0}$ be a Markov chain with transition matrix P and $X_0 = x$. Let Y_n be a Markov chain with transition matrix P and $Y_0 \sim \pi$ for stationary distribution π. Then marginally,

$$Y_n \sim \pi, \forall n.$$

Let $T := \min\{n \geq 0 : X_n = Y_n\}$ be the first time X_n and Y_n meet. Define a third stochastic process

$$Z_n = \begin{cases} X_n & \text{if } n \leq T \\ Y_n & \text{if } n > T \end{cases}$$

Since X_n and Y_n share the sample transition matrix P and Z_n begins at the same initial condition as X_n, Z_n shares the same distribution as X_n. Moreover, from coupling Z_n is not independent of Y_n. Then, we want to prove that

$$|\mathbb{P}\left(X_n = y\right) - \pi\left(y\right)| \leq \mathbb{P}\left(T > n\right)$$

To do so, note the following

$$\begin{aligned} \mathbb{P}\left(X_n = y\right) &= \mathbb{P}\left(Z_n = y\right) && \text{(coupling)} \\ &= \mathbb{P}\left(Z_n = y \mid T > n\right)\mathbb{P}\left(T > n\right) + \mathbb{P}\left(Z_n = y \mid T \leq n\right)\mathbb{P}\left(T \leq n\right) \end{aligned}$$

Recall the definition of Z_n: if $n > T$ it means that X_n and Y_n have already met, thus $Z_n = Y_n$. Moreover, the event $\{Z_n = y, T \leq n\}$ is equivalent to the event $\{Y_n = y, T \leq n\}$

This implies that

$$\begin{aligned}\mathbb{P}\left(Z_n = y, T \leq n\right) &= \mathbb{P}\left(Y_n = y, T \leq n\right)\\ &= \mathbb{P}\left(Y_n = y\right) - \mathbb{P}\left(Y_n = y, T > n\right)\\ &= \pi\left(y\right) - \mathbb{P}\left(Y_n = y, T > n\right)\\ &= \pi\left(y\right) - \mathbb{P}\left(Y_n = y \mid T > n\right)\mathbb{P}\left(T > n\right)\end{aligned}$$

and similarly

$$\mathbb{P}\left(Z_n = y, T > n\right) = \mathbb{P}\left(Z_n = y \mid T > n\right)\mathbb{P}\left(T > n\right)$$

Therefore, putting everything together

$$\mathbb{P}\left(X_n = y\right) = \pi\left(y\right) + \mathbb{P}\left(T > n\right)\Big(\mathbb{P}\left(Z_n = y \mid T > n\right) - P\left(Y_n = y \mid T > n\right)\Big)$$

which, because the term in brackets is between 1 and -1, implies

$$|\mathbb{P}\left(X_n = y\right) - \pi\left(y\right)| \leq \mathbb{P}\left(T > n\right)$$

The above inequality gives a bound on the tail probability that we want to squeeze to 0. Since P is ergodic, there exists $M \geq 1$ such that $\left(P^M\right)_{x,y} > 0$ for all $x, y \in S$. For simplicity, assume that $M = 1$. Let $\delta := \min_{x',y'} P_{x',y'} > 0$ give the smallest positive entry in the transition matrix. Want to check each time step to see whether or not the two chains have met

$$\begin{aligned}\mathbb{P}\left(T > n+1 \mid T > n\right) &\leq \max_{x' \neq y'} \mathbb{P}\left(X_{n+1} \neq Y_{n+1} \mid X_n = x', Y_n = y'\right) && \text{(Markov+haven't met)}\\ &= 1 - \min_{x' \neq y'} \mathbb{P}\left(X_{n+1} = Y_{n+1} \mid X_n = x', Y_n = y'\right) && \text{(Complement)}\\ &\leq 1 - \min_{x' \neq y'} \mathbb{P}\left(X_{n+1} = Y_{n+1} = x' \mid X_n = x', Y_n = y'\right) && \text{(Smaller set)}\\ &= 1 - \min_{x' \neq y'} P_{x'x'}P_{y'x'} && \text{(independence of } Y_n \text{ and } X_n + \text{Markov)}\\ &\leq 1 - \delta^2.\end{aligned}$$

This implies that $\mathbb{P}\left(T > n\right) \leq \left(1 - \delta^2\right)^n \xrightarrow{n \to \infty} 0$. Explicitly,

$$\begin{aligned}\mathbb{P}\left(T > n\right) &= \mathbb{P}\left(T > n \mid T > n-1\right)\mathbb{P}\left(T > n-1\right)\\ &\leq \left(1 - \delta^2\right)\mathbb{P}\left(T > n-1\right)\\ &\vdots\\ &\leq \left(1 - \delta^2\right)^n \overset{n \to \infty}{\to} 0\end{aligned}$$

Finally

$$\max_{x\in S}\sum_{y\in S}|\mathbb{P}\left(X_n=y\mid X_0=x\right)-\pi\left(y\right)|\leq|S|\left(1-\delta^2\right)^n$$

□

Note. *Above indicates exponential decay to the stationary distribution, which could be fast. However, ε may depend on $|S|$, so could actually be slow for big $|S|$ i.e. $e^{-|S|}$.*

Note that for uniqueness of the stationary distribution π we just need the transition matrix to be irreducible, still we need aperiodicity.

Now we consider the same theorem but removing aperiodicity.

Theorem. *Let S be a finite state space, P irreducible, with stationary distribution π. $\{X_n\}_{n\geq 1}$ is a Markov chain on S with transition matrix P. Let k be the period of the Markov chain. Then*

$$\frac{1}{k}\left\{\mathcal{L}\left(X_n\right)+\mathcal{L}\left(X_{n+1}\right)+\ldots+\mathcal{L}\left(X_{n+k-1}\right)\right\}\Rightarrow\pi.$$

Note. *Note that the stationary distribution in this case is a mixture of distributions. The number of distributions used to do this is given by the period of the MC.*

The next result is the equivalent of the SLLN for Markov Chains.

Theorem (Ergodic theorem for Markov chains)**.** *Suppose S is a finite state space, P is irreducible, and π is a unique stationary distribution. Let $f : S\to\mathbb{R}$ and define the quantity $m := \sum_{x\in S}f\left(x\right)\pi\left(x\right)=\mathbb{E}_\pi\left[f\right]$. Let $\left\{X_j\right\}_{j=0}^{n-1}$ be a Markov chain with transition matrix P. Then,*

$$\frac{1}{n}\sum_{i=0}^{n-1}f\left(X_i\right)\xrightarrow[n\to\infty]{a.s.}m.$$

In words, the time average converges to the space average.

Proof (Sketch). Any function $f : S\to\mathbb{R}$ can be written as $f\left(x\right)=\sum_{y\in S}f\left(y\right)\mathbb{1}_{x=y}$. Hence, by linearity, it suffices to prove this for indicator functions $f\left(x\right) := \mathbb{1}_{X=X_0}$. Fix $x_0\in S$. Want to show

$$\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{1}_{X_i=x_0}\xrightarrow[n\to\infty]{a.s.}\pi\left(x_0\right).$$

The term on the left hand side is the fraction of time spent in state x_0 during periods $\{0,1,\ldots,n-1\}$. Central idea is to break the trajectories of the Markov chains according to excursions from x_0 to x_0. Define the following random times

$$\begin{aligned} T_0 &:= \min\left\{s : X_s = x_0\right\}, \\ T_{j+1} &:= \min\left\{s > T_j : X_s = x_0\right\}, \\ \tau_j &:= T_j - T_{j-1}, j \geq 1. \end{aligned}$$

In words, T_j gives the time index of the jth time X_s hits x_0. τ_j gives the length of excursion j, which is also known as **return times to** x_0. From the Markov property, $\left\{\tau_j\right\}_{j\geq 1}$ are i.i.d. Let

$$N_t := \sum_{s=0}^{t-1} \mathbb{1}_{\left\{X_s = X_0\right\}}.$$

Key observation is that the events $\left\{\lim_{t\to\infty} N_t/t = A\right\} = \left\{\lim_{k\to\infty} T_k/k = 1/A\right\}$ are the same. If $T_k \leq t < T_{k+1}$ (in the middle of an excursion), then $N_t = k+1$. If $T_k \approx \frac{k}{A} \approx t \implies k \approx tA \implies N_t \approx tA$. Note that

$$T_k = T_0 + (T_1 - T_0) + (T_2 + T_1) + \cdots + (T_k - T_{k-1}) = T_0 + \tau_1 + \tau_2 + \cdots + \tau_k$$

where T_0 in principle can be infinite but this happens wp 0, thus it is bounded almost surely. In particular, it can be stochastically dominated by a geometric random variable, hence it admits finite expectation. Dividing by k

$$\frac{T_k}{k} = \frac{T_0}{k} + \frac{1}{k}\sum_{j=1}^{k} \tau_j \xrightarrow{a.s} \mathbb{E}\left[\tau_1\right]$$

by the SLLN, as τ_1 admits finite expectation because it is stochastically dominated by a geometric random variable. Now, define

$$T_{x_0} := \min\left\{n \geq 1 : X_n = x_0\right\}$$

then $\mathbb{E}\left[\tau_1\right] = \mathbb{E}\left[T_{x_0} \mid X_n = x_0\right]$. We want to show that $\mathbb{E}\left[T_{x_0} \mid X_0 = x_0\right] = 1/\pi\left(x_0\right)$. Define

$$r\left(x \mid x_0\right) = \mathbb{E}\left[\sum_{i=0}^{T_{x_0}-1} \mathbb{1}_{X_i = x} \mid X_n = x_0\right]$$

then if I sum over all $x \in S$

$$\sum_{x\in S} r\left(x \mid x_0\right) = \mathbb{E}\left[T_{x_0} \mid X_n = x_0\right]$$

Therefore

$$\frac{r\left(x \mid x_0\right)}{\mathbb{E}\left[T_{x_0} \mid X_n = x_0\right]}, \quad x \in S$$

is a probability distribution. More it's true, it is the stationary distribution. Indeed, note that $r\left(x_0 \mid x_0\right) = 1$. □

4.7 Recurrence and transience

Let S be countably infinite. Many things extend from the finite case (like irreducibility, periodicity) but many things are different. In particular a stationary distribution does not always exist (eg. SSRW on $\mathbb{Z}$).

Definition (First hitting time)**.** *Let $x \in S$, we say that*

$$T_x := \inf\left\{n \geq 1 : X_n = x\right\}, \quad x \in S$$

*is the **first hitting time** for X_n.*

It represents the amount of time (number of steps) until the chain returns to state x given that it started in state x. Note how "never returning" is allowed in the definition by defining $T_x = \infty$ if $X_n \neq x, \forall\, n \geq 1$.

Definition (Recurrent and transient states)**.** *We say that $x \in S$ is a **recurrent state** if*

$$p_x := \mathbb{P}\left(T_x < \infty \mid X_0 = x\right) = 1.$$

*We say that $x \in S$ is **transient** if*

$$p_x := \mathbb{P}_x\left(T_x < \infty \mid X_0 = x\right) < 1.$$

Note. *By the Markov property, once the chain revisits state x, the future is independent of the past, and it is as if the chain is starting all over again in state x for the first time: Each time state x is visited, it will be revisited with the same probability p_x independent of the past. In particular, if $p_x = 1$, then the chain will return to state x over and over again, an infinite number of times. That is why the word recurrent is used. If state x is transient $\left(p_x < 1\right)$, then it will only be visited a finite (random) number of times, after which only the remaining states $y \neq x$ can be visited by the chain.*

Definition (Recurrent)**.** *If $x \in S$ is **recurrent**, we say that:*

- *$x \in S$ is **positive recurrent** if $\mathbb{E}_x\left[T_x\right] < \infty$*
- *$x \in S$ is **null recurrent** if $\mathbb{E}_x\left[T_x\right] = \infty$.*

Definition (Number of returns)**.** *The number of times a MC returns to state $x \in S$ is given by*

$$N_x = \sum_{n=1}^{\infty} \mathbb{1}\left(X_n = x \mid X_0 = x\right)$$

Proposition. *A Markov chain is recurrent if and only if*

$$\mathbb{E}\left[N_x\right] = \sum_{n=1}^{\infty} \mathbb{P}\left(X_n = x \mid X_0 = x\right) = \infty, \quad \forall\, x \in S$$

A state is transient if and only if

$$\mathbb{E}\left[N_x\right] = \sum_{n=1}^{\infty} \mathbb{P}\left(X_n = x \mid X_0 = x\right) < \infty,$$

Proof. If $x \in S$ is recurrent, then we return with probability one, thus $N = \infty$. If $x \in S$ is transient, then $N \sim Geo\left(1 - p_x\right) \implies \mathbb{E}\left[N_x\right] < \infty$. □

Note. *Note that if $x \in S$ is transient $p_x < 1$, that is we have a probability smaller than 1 to come back to x. This leads to the interpretation of N as drawing from a Geometric distribution in which the "success" is defined as not coming back anymore to x. If we count the first visit as $X_0 = x$, then $\mathbb{P}\left(N = n\right) = 1 \cdot p_x^{n-1}\left(1 - p_x\right)$, the CDF of a $Geo\left(1 - p_x\right)$.*

Theorem (Polya's recurrence theorem)**.** *Let $\left\{X_n\right\}_{n \geq 0}$ be a simple symmetric random walk on $\mathbb{Z}^d$. Then $\left\{X_n\right\}_{n \geq 0}$ is recurrent for $d = 1, 2$ and transient for $d \geq 3$. "Drunk man will always return home while a drunk bird may not."*

Proof. Use the previous claim. Start at the origin $X_0 = 0$. Let N be the number of returns to the origin. Consider the $d = 1$ case. We want to know if $\mathbb{E}\left[N\right]$ is finite or not. First, note that $\mathbb{E}\left[N\right] = \mathbb{E}\left[\sum_{n=0}^{\infty} \mathbb{1}\left(X_n = 0\right)\right] = \sum_{n=0}^{\infty} \mathbb{P}\left(X_n = 0\right) = \sum_{n=0}^{\infty} \mathbb{P}\left(X_{2n} = 0\right)$ as the process can go back to the origin only in even times since the period of the process is 2.
Let W_n be the number of jumps to the right. Then the number of jumps to the left is $n - W_n$. Current position X_n is

$$X_n = W_n - \left(n - W_n\right) = 2W_n - n.$$

Notice that $X_n = 0 \iff 2W_n = n \iff W_n = n/2$. Thus, $X_{2n} = 0 \iff W_{2n} = n$. Note that $W_n \sim Bin\left(n, 1/2\right)$ (think of it as the sum of all jumps - aka Bernoulli - to the right). Therefore

$$\mathbb{P}\left(X_{2n} = 0\right) = \mathbb{P}\left(W_{2n} = n\right) = \binom{2n}{n}\left(\frac{1}{2}\right)^n\left(\frac{1}{2}\right)^n = \frac{(2n)!}{n!\,n!}2^{-2n} = \left(1 + o\left(1\right)\right)\frac{1}{\sqrt{\pi n}}$$

The last equality follows from Stirling's approximation of $n! \approx \left(1 + o\left(1\right)\right)\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$. Then,

$$\mathbb{E}\left[N\right] = \sum_{n=1}^{\infty}\mathbb{P}\left(X_{2n} = 0\right) = \sum_{n=1}^{\infty}\mathbb{P}\left(W_{2n} = n\right) \geq \sum_{n=1}^{\infty}\frac{c}{\sqrt{n}} = \infty.$$

Thus the SSRW on the integer line is recurrent.

If $d = 2$, then $\mathbb{P}\left(X_{2n} = 0\right) \geq c/n$ (imagine it as being back to the origin means that you have to be back in both dimensions). If $d \geq 3$, $\mathbb{P}\left(X_n = 0\right) \leq \frac{c}{n^{d/2}} + de^{-cn}$. The details are left as an exercise. □

4.8 Other Results

Proposition. *For any communicating class C, all states in C are either recurrent or all states in C are transient. Thus, if i and j communicate and i is recurrent, then so is j. Equivalently if i and j communicate and i is transient, then so is j. In particular, for an irreducible Markov chain, either all states are recurrent or all states are transient.*

Proposition. *All states in a communicating class C are all together either positive recurrent, null recurrent or transient*

Proposition. *An irreducible Markov chain with a finite state space S is always recurrent.*

Note. *Clearly if the state space is finite for a given Markov chain, then not all the states can be transient. For, otherwise after a finite number of steps (time) the chain would leave every state never to return; where would it go?*

Note (Interpretation of stationary distribution). *When π exists, let π_j denote the long run proportion of the time that the chain spends in state j when the chain starts from state i*

$$\pi_j = \lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n}\mathbb{1}\left(X_m = j \mid X_0 = i\right), \quad \forall i \in S$$

Taking expectations on both sides and applying the bounded convergence theorem to swap limits and integrals we have

$$\pi_j = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \mathbb{P}\left(X_m = j \mid X_0 = i\right), \quad \forall\, i \in S$$

or, alternatively,

$$\lim_{n\to\infty} \sum_{m=1}^{n} P^m = \begin{pmatrix} \pi \\ \pi \\ \vdots \end{pmatrix} = \begin{pmatrix} \pi_0, \pi_1, \dots \\ \pi_0, \pi_1, \dots \\ \vdots \end{pmatrix}$$

That is, when we average the m-step transition matrices, each row converges to the vector of stationary probabilities π. The $i-$ th row refers to the initial condition $X0 = i$, and for each such fixed row i, the $j-$ th element of the averages converges to π_j. A nice way of interpreting π: If you observe the state of the Markov chain at some random time way out in the future, then π_j is the probability that the state is j.

Property. *Let $\{X_n\}_{n\geq 0}$ be an irreducible MC with finite state space. Then a unique stationary distribution π exists and it is given by*

$$\pi_x = \frac{1}{\mathbb{E}_x\left[T_x\right]}, \quad \forall\, x \in S$$

In words, on average the chain visits state x every $\mathbb{E}_x\left[T_x\right]$ periods, therefore the probability that it is in that state is $\mathbb{E}_x\left[T_x\right]^{-1}$.

Note. *This property is useful to switch back and forth from the stationary distribution to the expected time to return in a state. Moreover, it shows us that recurrence is sufficient if S is finite, whereas positive recurrence is necessary is S is not finite to have a stationary distribution (recall that if a state is null recurrent then $\mathbb{E}_x\left[T_x\right] = \infty$*

5 Martingales

The Markov Property– that conditioned on the present, the past and future are independent– was central to our study of Markov chains. In this section, we will study martingales and the martingale property. Roughly, in expectation, the process does not change. Want to begin by introducing the machinery of conditional expectations. Then introduce martingales and study their properties.

5.1 Elementary conditional expectations

To begin, below are a few non-measure theoretic characterizations of conditional expectations. These will help motivate later technical discussion.

Definition (Conditional expectations)**.** *Suppose that X, Y are two discrete RVs on $(\Omega, \mathcal{F}, \mathbb{P})$. For simplicity, assume that X, Y take on finitely many values. For every y in the support of Y*

$$\mathbb{P}\left(X = x \mid Y = y\right) = \frac{\mathbb{P}\left(X = x, Y = y\right)}{\mathbb{P}\left(Y = y\right)}.$$

Then the ***conditional expectation of*** X ***given*** $Y = y$ *is*

$$\mathbb{E}\left[X \mid Y = y\right] = \sum x\mathbb{P}\left(X = x \mid Y = y\right) = \sum_x x\frac{\mathbb{P}\left(X = x, Y = y\right)}{\mathbb{P}\left(Y = y\right)} = \frac{\mathbb{E}\left[X\mathbb{1}_{Y=y}\right]}{\mathbb{P}\left(Y = y\right)}.$$

Note that this is a real number that is a function of y. For convenience, write

$$f\left(y\right) := \mathbb{E}\left(X \mid Y = y\right).$$

Then the ***conditional expectation of*** $\mathbb{E}\left(X \mid Y\right)$ *can be defined as*

$$\mathbb{E}\left(X \mid Y\right) := f\left(Y\right).$$

Note that $\mathbb{E}\left(X \mid Y\right)$ is a RV which is also defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $Z := \mathbb{E}\left(X \mid Y\right)$. Then we have that $Z\left(\omega\right) = \mathbb{E}\left(X \mid Y = y\right)$ whenever ω is such that $Y\left(\omega\right) = y$.

Property (Conditional expectations)**.** *Below are a few basic properties.*

1. ***Linearity***: $\mathbb{E}\left(a_1X_1 + a_2X_2 \mid Y\right) = a_1\mathbb{E}\left(X_1 \mid Y\right) + a_2\mathbb{E}\left(X_2 \mid Y\right)$.

   Proof. Follows from linearity of expectation operator. □

2. ***Tower rule (law of total expectation)**: $\mathbb{E}\left[\mathbb{E}\left(X \mid Y\right)\right] = \mathbb{E}\left(X\right)$.*

   Proof.

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left(X \mid Y\right)\right] &= \sum_y \mathbb{E}\left(X \mid Y = y\right)\mathbb{P}\left(Y = y\right) = \sum_y\sum_x x\mathbb{P}\left(X = x \mid Y = y\right)\mathbb{P}\left(Y = y\right) \\ &= \sum_y\sum_x x\frac{\mathbb{P}\left(X = x, Y = y\right)}{\mathbb{P}\left(Y = y\right)}\mathbb{P}\left(Y = y\right) \\ &= \sum_x x\sum_y \mathbb{P}\left(X = x, Y = y\right) = \sum_x x\mathbb{P}\left(X = x\right) = \mathbb{E}\left(X\right).\end{aligned}$$

□

3. ***Independence**: If X, Y are independent RVs, then*

$$\mathbb{E}\left(Y \mid X\right) = \mathbb{E}\left(Y\right).$$

4. ***Taking out what is known**: For any measurable function $g : Supp\left(Y\right) \to \mathbb{R}$,*

$$\mathbb{E}\left(Xg\left(Y\right) \mid Y\right) = g\left(y\right)\mathbb{E}\left(Y \mid X\right).$$

We want to extend these elementary conditional expectation definitions to the context of martingales.

Property. *Let $Y : \mathcal{F} \to \mathbb{R}$ be an $\mathcal{F}-$measurable random variable and let $G_y := \left\{\omega \in \Omega : Y\left(\omega\right) = y\right\}$ be the partition of the sample space Ω induced by the random variable Y. Call the elements of the partition **Y-atoms**. $Z := \mathbb{E}\left(X \mid Y\right)$ is by construction constant for each G_y. Let $\mathcal{G} := \sigma\left(Y\right) := \sigma\left\{Y^{-1}\left(B\right) : B \in \mathscr{B}_{\mathbb{R}}\right\} \subseteq \mathcal{F}$ be the sigma-algebra generated by Y. Claim that $\mathcal{G}$ consists of all possible unions of G_y's, hence Z is measurable with respect to $\mathcal{G}$.*
Any $G \in \mathcal{G}$ is of the form $G = \cup_{y \in \mathcal{I}} G_y$, so

$$\begin{aligned}\mathbb{E}\left(Z\mathbb{1}_G\right) &= \sum_{y \in \mathcal{I}} \mathbb{E}\left[Z\mathbb{1}_{G_y}\right] = \sum_{y \in \mathcal{I}} \mathbb{E}\left(X \mid Y = y\right)\mathbb{P}\left(Y = y\right) \\ &= \sum_{y \in \mathcal{I}}\sum_x x\mathbb{P}\left(X = x \mid Y = y\right)\mathbb{P}\left(Y = y\right) = \mathbb{E}\left(X\mathbb{1}_G\right)\end{aligned}$$

The above equality hints at a way of defining conditional expectations more generally.

5.2 General conditional expectations

Before giving the definition of conditional expectation, a quick preamble on sigma-algebras is needed. A sigma-algebra is the mathematical equivalent of the loose concept of **information**. Indeed, a sigma-algebra $\mathcal{F}$ is nothing else than a collection of subsets of Ω and if we take $A \in \mathcal{F}$ we know whether it occurred or not. The expectation of a RV Y that is $\mathcal{G}-$measurable with $\mathcal{F} \subseteq \mathcal{G}$, conditioning on $\mathcal{F}$ means that we know that a certain $A \in \mathcal{F}$ has occurred, thus we want to focus on the probability of X in the region of Ω where A occurs. Note that the finer the partition induced by $\mathcal{F}$ is, the more the information we get, because the set A of candidate events gets smaller and smaller.

This is even more telling if we think about $\sigma(X)$. The sigma-algebra generated by X is nothing else than the sigma-algebra generated using a special collection of sets of Ω: the partition of Ω induced by the pre-image of X. If I observe X, then I know in which piece of the partition of Ω the occurred event is. Thus $\mathbb{E}\left[Y \mid \sigma(X)\right]$ will compute the average value of Y in that portion of the sample space.

Definition (Conditional expectation (general))**.** *Suppose that X is an integrable RV ($\mathbb{E}|X|<\infty$) on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a $\sigma-$algebra. Then the* ***conditional expectation*** *$\mathbb{E}\left(X \mid \mathcal{G}\right)$ is a random variable that is measurable with respect to $\mathcal{G}$ such that*

$$\forall G \in \mathcal{G}, \quad \mathbb{E}\left(X \mathbb{1}_G\right) = \mathbb{E}\left[\mathbb{E}\left(X \mid \mathcal{G}\right) \mathbb{1}_G\right].$$

Definition (Absolutely continuous)**.** *Let μ, ν be two measures on a measurable space $(E, \mathcal{E})$. ν is* ***absolutely continuous with respect to*** *μ denoted as $\nu << \mu$ if $\forall A \in \mathcal{E}, \mu(A) = 0$ implies that $\nu(A) = 0$.*

Note. *There is another theorem that characterizes absolutely continuous measures. It states that given a sequence of sets $\{A_n\}_{n\geq 1}$ such that $\nu(A_n) \to 0$, then $\mu(A_n) \to 0$ as well. If we imagine A_n as being the difference of two sets, then we see the relationship with continuity in the traditional sense. This is because as the two sets becomes closer and closer under ν, they are such also under the dominated measure μ.*

Theorem. *Given two measures ν and μ on the same measurable space $(E, \mathcal{E})$, if there exists a random variable $f : E \to \mathbb{R}$ such that*

$$\mu(A) = \int_A f(\omega)\, d\nu(\omega), \forall A \in \mathcal{E}$$

then μ is absolutely continuous with respect to ν, i.e. $\mu << \nu$.

Note. *The converse is also true and it is the Radon-Nikodym theorem.*

Example (Absolutely continuous)**.** *A few brief examples.*

1. *Standard Gaussian measure on $\mathbb{R}$ is absolutely continuous with respect to the Lebesgue measure. One way to think of it is that if a set has Lebesgue measure zero, then also the probability that the Gaussian RV takes values in that set is zero. This is easy to see once we approach the problem from the right angle. Call ν the Lebesgue measure, and μ the Gaussian probability measure. Then, it is possible to 'link' them through a function $f(\cdot)$. Such a function is simply the density of a Gaussian random variable!*
2. *Discrete measures with the same support are absolutely continuous respect to each other (like Geometric and Poisson).*

Theorem (Radon-Nikodym)**.** *Suppose that μ and ν are two σ-finite measures on $(E, \mathcal{E})$ and that $\nu << \mu$. Then there exists a non-negative $\mathcal{E}-$measurable function g such that $\forall f \in \mathcal{E}_+$*

$$\int_E f(x)\, d\nu(x) = \int_E g(x) f(x)\, d\mu(x),$$

where g is essentially unique. If the above holds for $\widetilde{g}$ too, then $g = \widetilde{g}$ $\mu-$a.e.

Definition (Radon-Nikodym derivative)**.** *The function g defined above is called a **Radon-Nikodym derivative** of μ with respect to ν. Often denoted as*

$$g = \frac{d\nu}{d\mu}, g(x) = \frac{d\nu(x)}{d\mu(x)}$$

write as $\nu = g\mu$.

Radon-Nikodym is useful for going between different probability distributions. Implicitly, probability distributions are often defined using Radon-Nikodym derivatives.

Definition (Singular)**.** *A measure μ is **singular with respect to** ν if there exists a set $D \in \mathcal{E}$ such that $\mu(D) = 0$ and $\nu\left(E \setminus D\right) = 0$.*

Example. *A simple example is to compare purely atomic with diffuse measures, like $D = \mathbb{Z}$ and μ be the Gaussian measure and ν the Poisson measure.*

Definition. *Let $(E, \mathcal{E})$ be a measurable space and μ a measure on it. Let $D \in \mathcal{E}$.*

- *Let $\nu(A) = \mu(A \cap D)$ for every $A \in \mathcal{E}$. Then ν is a measure on $(E, \mathcal{E})$; and it is called the **trace** of μ on D.*

- *Let $\mathcal{D}$ be the trace of $\mathcal{E}$ on D, that is, $\mathcal{D} := \mathcal{E} \cap D = \{A \cap D : A \in \mathcal{E}\}$. Let $\nu(A) = \mu(A)$ for every $A \in \mathcal{D}$, then ν is a measure on $(D, \mathcal{D})$ and it is called the* ***restriction*** *of μ to D.*

Note. *The first thing to realize is that the trace $\mathcal{D}$ of a sigma-algebra $\mathcal{E}$ is a sigma-algebra as well. It is a specific sigma-algebra though. Indeed, it is projecting down the larger sigma-algebra $\mathcal{E}$ defined on Ω onto the set $D \subset \Omega$. If we take the original measure μ on $(\Omega, \mathcal{E})$ and we restrict its domain to $(D, \mathcal{D})$, we get the restriction of μ on D.*

Theorem (Existence and uniqueness of conditional expectations)**.** *Suppose we have an integrable RV X on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sigma-algebra $\mathcal{G} \subseteq \mathcal{F}$. There exists a $\mathcal{G}-$measurable RV Z such that for every $G \in \mathcal{G}$*

$$\mathbb{E}[X\mathbb{1}_G] = \mathbb{E}[Z\mathbb{1}_G]$$

Moreover if the condition above is satisfied by some other $\mathcal{G}-$measurable random variable $\widetilde{Z}$ on $(\Omega, \mathcal{G}, \mathbb{P})$ then $\mathbb{P}\left(Z = \widetilde{Z}\right) = 1$.

Proof. Consider case when X is non-negative[1] and integrable. Let μ denote the restriction of $\mathbb{P}$ to the measurable space $(\Omega, \mathcal{G})$ so that $\forall g \in \mathcal{G}, \mu(G) = \mathbb{P}(G)$. Define ν as

$$\forall G \in \mathcal{G}, \quad \nu(g) := \int_G X d\mathbb{P} = \mathbb{E}[X\mathbb{1}_G].$$

Note that ν is a finite measure since $\nu(g) = \mathbb{E}(X) < \infty$ because of integrability. Then $\nu << \mu$ because if we take a set that has measure 0 according to μ then it has measure 0 according to ν by the definition of ν. By the Radon-Nikodym theorem, there exists $Z \in \mathcal{G}_+$ such that $\nu = Z\mu$. Then, by definition,

$$\mathbb{E}[X\mathbb{1}_G] = \nu(G) = (Z\mu)(G) = \mathbb{E}[Z\mathbb{1}_G]$$

Z is $\mathcal{G}-$measurable and equality holds, so Z is a conditional expectation $\mathbb{E}\left[X \mid \mathcal{G}\right]$. For general integrable X, use $X = X_+ - X_-$. By above, there exists $Z_+ \equiv \mathbb{E}\left[X_+ \mid \mathcal{G}\right]$ and $Z_- := \mathbb{E}\left[X_- \mid \mathcal{G}\right]$ and $Z := Z_+ - Z_-$.

For uniqueness, assume $Z, \widetilde{Z}$ are both conditional expectations. Then, taking the difference,

$$\mathbb{E}\left[\left(Z - \widetilde{Z}\right)\mathbb{1}_G\right] = 0, \forall G \in \mathcal{G}$$

[1] In this setting the expectation of X always exists (it might be infinite though).

Let $G_+ := \left\{Z - \widetilde{Z} > 0\right\}$ and $G_- := \left\{Z - \widetilde{Z} < 0\right\}$. Then

$$0 = \mathbb{E}\left[\left(Z - \widetilde{Z}\right)\mathbb{1}_{G_+} - \mathbb{E}\left[\left(Z - \widetilde{Z}\right)\right]\mathbb{1}_{G_-}\right] = \mathbb{E}\left[\left|Z - \widetilde{Z}\right|\right]$$

Therefore $Z = \widetilde{Z}$ almost surely.

□

Note (Conditional expectations)**.** *A few important observations and special cases.*

1. *Important special case is when the $\sigma-$algebra $\mathcal{G}$ is the sigma-algebra generated by some RV Y defined on $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathcal{G} = \sigma(Y)$. In this case, we simply write $\mathbb{E}\left(X \mid Y\right)$ instead of*

$$\mathbb{E}\left(X \mid \sigma(Y)\right) \equiv \mathbb{E}\left(X \mid \mathcal{G}\right).$$

2. *If X is $\mathcal{G}-$measurable, then $\mathbb{E}\left(X \mid \mathcal{G}\right) = X$. Follows immediately from the definition. Theorem implies uniqueness. Note that this is not true in general as $\mathcal{G} \subseteq \mathcal{F}$, thus since X is $\mathcal{F}-$measurable it is not necessarily true that it is also $\mathcal{G}-$measurable.*
3. *If $\mathcal{G} = \{\emptyset, \Omega\}$ is the trivial sigma-algebra, then $\mathbb{E}\left[X \mid \mathcal{G}\right]$ is a constant RV which is equal to $\mathbb{E}(X)$. For the empty set case, $\mathbb{1}_G = 0$. For Ω, $\mathbb{1}_G = 1 \implies \mathbb{E}(X) = \mathbb{E}\left[\mathbb{E}\left(X \mid Y\right)\right]$.*

Note (Conditional expectations)**.** *Below properties are extensions to the elementary characterization.*

1. *Linearity: $\mathbb{E}\left[a_1X_1 + a_2X_2 \mid \mathcal{G}\right] = a_1\mathbb{E}\left[X_1 \mid \mathcal{G}\right] + a_2\mathbb{E}\left[X_2 \mid \mathcal{G}\right]$.*
2. *Tower rule. Suppose $\mathcal{G}_1 \subseteq \mathcal{G}_2$. Then*

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}_2\right] \mid \mathcal{G}_1\right] = \mathbb{E}\left(X \mid \mathcal{G}_1\right).$$

Note that the richer sigma-algebra is always in the inner expectation.

3. *Independence: Let $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ be sigma-algebras. They are independent if $\forall\, A \in \mathcal{G}, B \in \mathcal{H}$, we have that $\mathbb{P}(A \cup B) = \mathbb{P}(A)\,\mathbb{P}(B)$. If the sigma-algebras $\mathcal{G}$ and $\sigma(X)$ are independent, then*

$$\mathbb{E}\left(X \mid \mathcal{G}\right) = \mathbb{E}(X).$$

4. *If Y is $\mathcal{G}$-measurable, X is integrable, and XY is also integrable, then*

$$\mathbb{E}\left[XY \mid \mathcal{G}\right] = Y\mathbb{E}\left[X \mid \mathcal{G}\right].$$

Proposition (Conditional expectation as orthogonal projection)**.** *Suppose that X is square-integrable, that is $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Note $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space with inner product $\langle X, Y \rangle = \mathbb{E}(XY)$, and hence with norm*

$$\|X\|_{L^2} = \sqrt{\langle X, X \rangle} = \sqrt{\mathbb{E}\left(X^2\right)}.$$

Let $K := L^2(\Omega, \mathcal{G}, \mathbb{P})$ denote the subspace of $\mathcal{G}$-measurable square integrable random variables. Claim that $\mathbb{E}\left[X \mid \mathcal{G}\right]$ is the orthogonal projection of X onto K.

Proof. Let $Y = \mathbb{E}\left[X \mid \mathcal{G}\right]$. Want to show that for all $Z \in K$, $\langle X - Y, Z - Y \rangle = 0$.

$$\begin{aligned} \langle X - Y, Z - Y \rangle &= \mathbb{E}\left[(X - Y)(Z - Y)\right] = \mathbb{E}\left[\mathbb{E}\left((X - Y)(Z - Y) \mid \mathcal{G}\right)\right] \\ &= \mathbb{E}\left[(Z - Y)\,\mathbb{E}\left[X - Y \mid \mathcal{G}\right]\right] && \text{(Tower property)} \\ &= \mathbb{E}\left[(Z - Y)(Y - Y)\right] = 0 && \text{(Definition of conditional expectation)} \end{aligned}$$

□

The intuition is that the conditional expectation is the $\mathcal{G}$-measurable random variable that is closest to X.

5.3 Martingales

Definition (Filtration)**.** *Given a totally ordered set $\mathcal{T}$, a filtration is a collection of sigma-algebras $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$ such that $\mathcal{F}_s \subseteq \mathcal{F}_t$ for all $s < t$.*

Think of $\mathcal{T}$ as indexing over time and each $\mathcal{F}_t$ representing information sets.

Definition. *Given a stochastic process $\{X_n\}_{n \geq 0}$ its* ***natural filtration*** *is given by*

$$\mathcal{F}_n := \sigma\left(X_0, X_1, \ldots, X_n\right).$$

Definition (Adapted)**.** *We say that a stochastic process $\{X_t\}_{t \in \mathcal{T}}$ is* ***adapted*** *to filtration $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$ if X_t is $\mathcal{F}_t$-measurable for all $t \in \mathcal{T}$.*

Note that a stochastic process is always adapted to its natural filtration.

Definition (Martingale)**.** *A stochastic process $\{X_t\}_{t \in \mathcal{T}}$ is a Martingale with respect to a filtration $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$ if the following three conditions hold*

1. *$\{X_t\}_{t \in \mathcal{T}}$ is adapted to $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$.*

2. *$\forall\, t \in \mathcal{T}$, X_t is integrable. That is $\mathbb{E}\left[|X_t|\right] < \infty$.*
3. *$\forall\, s, t \in \mathcal{T}$ such that $s < t$,*

$$\mathbb{E}\left[X_t \mid \mathcal{F}_s\right] = X_s. \qquad \text{(martingale property)}$$

Note. *A few points on martingales*

- *If you show the martingale property for $n = 1$ the rest follows by induction*
- *Often $\mathcal{F}_n$ is the natural filtration of $\{X_n\}_{n\geq 0}$, that is $\mathcal{F}_n = \sigma\left(X_0, \ldots, X_n\right)$*
- *By the tower rule*

  $$X_n = \mathbb{E}\left[X_{n+m} \mid \mathcal{F}_n\right] \implies \mathbb{E}\left[X_n\right] = \mathbb{E}\left[\mathbb{E}\left[X_{n+m} \mid \mathcal{F}_n\right]\right] = \mathbb{E}\left[X_{n+m}\right]$$

  In particular

  $$\mathbb{E}\left[X_n\right] = \mathbb{E}\left[X_0\right], \forall\, n \geq 0$$

- *When you have two sigma-algebras $\mathcal{G}_n \subseteq \mathcal{F}_n$ and a stochastic process Y_n that is a martingale with respect to $\mathcal{F}_n$, then by the tower property it is a martingale with respect to $\mathcal{G}_n$ as well. We still need to check that Y_n is adapted to $\mathcal{G}_n$ as well. Recall indeed that measurability is preserved from the coarser partition to the finer, but not the viceversa!!*

Example (Martingales)**.** *Consider a few simple examples.*

1. ***Sum of independent mean zero summands**. Let $\{X_i\}_{i\geq 0}$ be i.n.i.d. with mean zero. Let $S_n := \sum_{i=1}^{n} X_i$, $S_0 = 0$. Note that we have two natural and equivalent natural filtrations $\mathcal{F}_n := \sigma\left(S_0, \ldots, S_n\right) = \sigma\left(X_0, X_1, \ldots, X_n\right)$. S_n is $\mathcal{F}_n$-measurable, thus we have adaptivity (this is always the case when we consider the natural filtration). Regarding integrability*

   $$\mathbb{E}\left[|S_n|\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[|X_i|\right] < \infty$$

   We need to verify the martingale property now. Note that

   $$\mathbb{E}\left(S_{n+1} \mid \mathcal{F}_n\right) = \mathbb{E}\left(S_n + X_{n+1} \mid \mathcal{F}_n\right) = S_n + \mathbb{E}\left(X_{n+1} \mid \mathcal{F}_n\right) = S_n + \mathbb{E}\left(X_{n+1}\right) = S_n,$$

   Hence we conclude that S_n is a martingale with respect to the natural filtration.
2. ***Product of independent positive mean zero random variables**. Let $\{X_n\}_{n\geq 0}$ be mutually independent positive random variables with mean 1. Let $Z_n := \prod_i X_i$, $Z_0 = 1$, and consider the natural filtration $\mathcal{F}_n := \sigma\left(Z_0, Z_1, \ldots, Z_n\right) = \sigma\left(X_1, \ldots, X_n\right)$. Adaptivity is clear since we are considering the natural filtration, whereas integrability*

   $$\mathbb{E}\left[|Z_n|\right] = \mathbb{E}\left[Z_n\right] = \mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}\left[X_i\right] = 1 < \infty$$

Then

$$\mathbb{E}\left(Z_{n+1} \mid \mathcal{F}_n\right) = Z_n \mathbb{E}\left(X_{n+1} \mid \mathcal{F}_n\right) = Z_n \mathbb{E}\left(X_{n+1}\right) = Z_n.$$

The first equality follows because Z_n is $\mathcal{F}_n$ measurable. The second equality follows from independence.

3. ***Symmetric Random walk**. Let $\{X_1, X_2, \ldots\}$ be iid and symmetric, that is $\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = -1) = 1/2$ and define the random walk as $S_n := \sum_{i=1}^n X_i$, $S_0 = 0$. It is a martingale from example 1. Consider now the process defined as $Y_n := S_n^2 - n$. Y_n is a martingale with respect to its natural filtration defined as usual as $\mathcal{G}_n := \sigma(Y_0, Y_1, \ldots, Y_n)$ and with respect to $\{\mathcal{F}_n\}_{n\geq 0}$, where $\mathcal{F}_n := \sigma(S_0, S_1, \ldots, S_n)$. Note that $\mathcal{G}_n \subseteq \mathcal{F}_n$. The intuition is that we lose information when squaring S_n, this is why $\mathcal{G}_n$ is coarser than $\mathcal{F}_n$.*

   Consider now the proof that Y_n is a martingale with respect to $\mathcal{F}_n$. Adaptivity follows from the fact that $\mathcal{G}_n \subseteq \mathcal{F}_n$, and integrability from the fact that each S_n is bounded in probability. Regarding the martingale property

$$\begin{aligned}\mathbb{E}\left[Y_{n+1} \mid \mathcal{F}_n\right] &= \mathbb{E}\left[(S_n + X_{n+1})^2 - (n+1) \mid \mathcal{F}_n\right] = \mathbb{E}\left[S_n^2 + 2S_nX_{n+1} + X_{n+1}^2 - (n+1) \mid \mathcal{F}_n\right] \\ &= \mathbb{E}\left[S_n^2 - n \mid \mathcal{F}_n\right] + 2\mathbb{E}\left[S_nX_{n+1} \mid \mathcal{F}_n\right] + \mathbb{E}\left[X_{n+1}^2 - 1 \mid \mathcal{F}_n\right] \\ &= Y_n + 2S_n\mathbb{E}\left[X_{n+1} \mid \mathcal{F}_n\right] + 0 = Y_n + 2S_n\mathbb{E}\left[X_{n+1}\right] + 0 = Y_n\end{aligned}$$

   The final line follows from X_{n+1} being independent of $\mathcal{F}_n$. Next, want to show Y_n is a martingale with respect to $\mathcal{G}_n$. Apply the Tower rule

$$\mathbb{E}\left[Y_{n+1} \mid \mathcal{G}_n\right] = \mathbb{E}\left[\mathbb{E}\left[Y_{n+1} \mid \mathcal{F}_n\right] \mid \mathcal{G}_n\right] = \mathbb{E}\left[Y_n \mid \mathcal{G}_n\right] = Y_n.$$

   The last equality follows because Y_n if $\mathcal{G}_n$ measurable.

4. ***Betting**: Start with M_0 dollars. At time n, can bet B_n dollars based on what happened previously. You get $B_n \cdot W_n$, where $\{W_i\}_{i\geq 1}$ are iid mean zero random variables. Assume that at time $n-1$ you decide the bet at B_n, this makes B_n being $\mathcal{F}_{n-1}$ measurable. Such a process is called a **predictable process**. Now define*

$$M_n = M_{n-1} + B_nW_n, \quad \mathbb{E}(W_1) = 0.$$

   Then $\{M_n\}_{n\geq 0}$ is a martingale.

$$\mathbb{E}\left[M_n \mid \mathcal{F}_{n-1}\right] = \mathbb{E}\left[M_{n-1} + B_nW_n \mid \mathcal{F}_{n-1}\right] = M_{n-1} + B_n\mathbb{E}\left[W_n\right] = M_{n-1}.$$

Definition (Supermartingale)**.** *A **supermartingale** satisfies the same conditions as a martingale except in (iii) we have that*

$$\mathbb{E}\left(X_{n+m} \mid \mathcal{F}_n\right) \leq X_n, \quad \forall\, m.$$

As an example, we would not expect wealth to increase at a casino.

Definition (Submartingale)**.** *A **submartingale** satisfies the same conditions as a martingale except in (iii) we have that*

$$\mathbb{E}\left(X_{n+m} \mid \mathcal{F}_n\right) \geq X_n, \quad \forall\, m.$$

As an example, if we were the casino, we would expect wealth to increase.

5.4 Stopping times

Let's start with a motivating example. It sometimes is a useful exercise to separate the random from the non-random pieces of the puzzle. Let's build up a stopping time, starting without randomness, along the lines of your intuition. Suppose that the observations X_j take values in the space S, and let $S^{\mathbb{N}}$ be the space of S-valued sequences. For any strategy or stopping policy, and any $0 \leq n < \infty$ we may define a two-valued map $\phi_n : S^{\mathbb{N}} \to \{\text{GO}, \text{STOP}\}$ which tells me what to do at time n if I were to observe $s = (s_0, s_1, \ldots)$. We require that $\phi_n(s)$ only depends on the first part of the sequence $(s_0, s_1, \ldots, s_n)$. That is, the decision to stop at time n must only depend on the observations up to time n. No peeking into the future! Now define

$$\phi(s) := \inf\left(n \geq 0 : \phi_n(s) = \text{STOP}\right)$$

This gives a map $\phi : S^{\mathbb{N}} \to \mathbb{N} \cup \{\infty\}$ which expresses our policy, by telling us when to stop. Finally we can put probability back into the picture by defining $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ by

$$\tau(\omega) = \phi\left(\mathrm{X}_0(\omega), \mathrm{X}_1(\omega), \mathrm{X}_2(\omega), \ldots\right)$$

This random variable is the stopping strategy applied to the random sequence

$$\left(\mathrm{X}_0(\omega), \mathrm{X}_1(\omega), \mathrm{X}_2(\omega), \ldots\right)$$

Definition (Stopping time)**.** *let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration. A **stopping time***

$$T : \Omega \to \mathbb{N}_+ \cup \{\infty\}$$

is an $\mathcal{F}-$measurable random variable where $\forall\, n \in \mathbb{N}, \left\{\omega \in \Omega : T(\omega) \leq n\right\} \in \mathcal{F}_n$.

Note. *Note that T is a random time. The intuition behind the definition is that T is a stopping time if at any given fixed time n, we can determine whether it has occurred or not, that is whether that event can be in the information set we have. The key is that we cannot use future information to determine it!*

There is an alternative definition

Definition (Stopping time)**.** *T is an **equivalent stopping time** if*

$$\forall n \in \mathbb{N}, \{\omega \in \Omega : T(\omega) = n\} \in \mathcal{F}_n.$$

Example. *Below are a few simple examples.*

1. ***First hitting time**: Let $\{X_n\}_{n\geq 0}$ be a stochastic process with $\{\mathcal{F}_n\}_{n\geq 0}$ as its natural filtration. S is the state space. $A \subseteq S$. Then*

$$T_a := \inf\{m \geq 0 : X_m \in A\}$$

is a first hitting time of A. Recall that if an event never happens, we have the infimum of an empty set which is ∞.

2. ***Return times**: Return times are a special case of a first hitting time*

$$R := \inf\{m > 0 : X_m = X_0\}.$$

Below are a few instructive non-examples.

1. ***Last hitting times**: Let* $\sup\{m : X_m \in A\}$ *give a last hitting time. This is not a stopping time because we would need to observe whether the event has occurred in the future, i.e. it would require seeing ahead. Technically* $\sup\{m : X_m \in A\}$ *is not $\mathcal{F}_n-$measurable.*
2. ***Selling before the market drops**: Can formally describe the process as*

$$\inf\{m : Y_{m+1} - Y_m < 0\}.$$

3. ***Constant time before hitting time**: Can formally describe the process as*

$$\inf\{m : Y_m \in A\} - 5.$$

Would need to see 5 steps ahead. Think gambling and bankruptcy and a gambler that would like to stop 5 periods before the ruin.

Proposition. *If T_1 and T_2 are stopping times, then*

$$\min\{T_1, T_2\} \equiv T_1 \wedge T_2$$
$$\max\{T_1, T_2\} \equiv T_1 \vee T_2$$

are stopping times as well.

Definition (Stopped martingale)**.** $(\Omega, \mathcal{F}, \mathbb{P})$ *is a probability space. Let* $\{\mathcal{F}_n\}_{n\geq 0}$ *be a filtration. Let* $\{X_n\}$ *be a martingale/submartingale/supermartingale. We define a* ***stopped process*** *as*

$$X_n^T := X_{T\wedge n}$$

Theorem. *If* $\{X_n\}_{n\geq 0}$ *is a martingale/submartingale/supermartingale, then* $\left\{X_n^T\right\}_{n\geq 0}$ *is also a martingale/submartingale/supermartingale.*

Proof. Proceed with the martingale case. Proofs for submartingale and supermartingale are nearly identical. Show the three conditions.

1. **Adapted**: We can decompose the stopped process as $X_n^T = \sum_{k=0}^{n-1} \mathbb{1}_{T=k}X_k + \mathbb{1}_{T\geq n}X_n$. This is because the first n terms take care of the possibility that the stopping time has already occurred and the process has stopped. The last term considers the possibility the process has not stopped yet. Then, the first term is $\mathcal{F}_k$ measurable because all indicators in the sum are $\mathcal{F}_k$ measurable and X_k is adapted by assumption. Therefore the first term is $\mathcal{F}_{n-1}-$measurable. Regarding the second term, the indicator function can be rewritten as $1 - \mathbb{1}_{T\leq n-1}$ so it is $\mathcal{F}_{n-1}$ measurable, whilst X_n is $\mathcal{F}_n-$measurable. Therefore all terms are $\mathcal{F}_n-$ measurable, which shows that X_n^T is adapted.
2. **Integrability**: $|X_n^T|\leq \sum_{k=0}^{n}|X_k|$. Sum of finitely many integrable RVs is also integrable.
3. **Martingale property**: The equality in the third line can be changed to weak inequal-

ities for the submartingale and supermartingale cases.

$$\begin{aligned}\mathbb{E}\left[X_n^T \mid \mathcal{F}_{n-1}\right] &= \mathbb{E}\left[\sum_{k=0}^{n-1} \mathbb{1}_{T=k} X_k + \mathbb{1}_{T\geq n} X_n \mid \mathcal{F}_{n-1}\right] \\ &= \sum_{k=0}^{n-1} \mathbb{1}_{T=k} X_k + \mathbb{1}_{T\geq n} \mathbb{E}\left[X_n \mid \mathcal{F}_{n-1}\right] \\ &= \sum_{k=0}^{n-1} \mathbb{1}_{T=k} X_k + \mathbb{1}_{T\geq n} X_{n-1} && \text{(Martingale property)} \\ &= \sum_{k=0}^{n-2} \mathbb{1}_{T=k} X_k + \mathbb{1}_{T\geq n-1} X_{n-1} && \text{(grouping terms)} \\ &= X_{n-1}^T\end{aligned}$$

□

Corollary. *If $\{X_n\}_{n\geq 0}$ is a martingale and T is a stopping time, then then $\left\{X_n^T\right\}_{n\geq 0}$ is also a martingale, hence*

$$\mathbb{E}\left[X_{T\wedge n}\right] = \mathbb{E}\left[X_{T\wedge 0}\right] = \mathbb{E}\left[X_0\right].$$

Thus also $\lim_{n\to\infty} \mathbb{E}\left[X_{T\wedge n}\right] = \mathbb{E}\left[X_0\right]$.

Note (Switching order of limit and expectation). *Suppose $\mathbb{P}(T<\infty)=1$. Then* $\lim_{n\to\infty} X_n^T = X_T$ *a.s.. Not always true that $\mathbb{E}(X_T) = \mathbb{E}(X_0)$. Consider X_n being a SSRW on $\mathbb{Z}$ with first hitting time*

$$T := \inf\{n \geq 0, X_n = 1\}.$$

SSRW on $\mathbb{Z}$ is recurrent, so indeed $\mathbb{P}(T<\infty) = 1$, that is $T<\infty$ a.s. However note that $\mathbb{E}(X_0) = 0$, simply because the initial condition is given, and $X_T = 1$ a.s., thus $\mathbb{E}(X_T) = 1$.
A few observations are needed. The following holds for any martingale

$$\mathbb{E}\left[X_0\right] = \mathbb{E}\left[X_{T\wedge n}\right] = \lim_{n\to\infty} \mathbb{E}\left[X_{T\wedge n}\right]$$

whereas

$$\mathbb{E}\left[X_T\right] = \mathbb{E}\left[\lim_{n\to\infty} X_{T\wedge n}\right]$$

does not hold in general. What went wrong? To apply the corollary, need to be able to exchange the order of the limit and expectation. For example we need the dominated convergence theorem to hold.

The following theorem states when the following chain of equalities holds

$$\mathbb{E}\left[X_0\right] = \mathbb{E}\left[X_{T\wedge n}\right] = \lim_{n\to\infty} \mathbb{E}\left[X_{T\wedge n}\right] = \mathbb{E}\left[X_T\right]$$

The first and second equalities are true in general for Martingales. Indeed, the first one follows from the fact that the stopped process is a Martingale and $T \wedge 0 = 0$, whilst the second one follows from the fact that the first equality is true $\forall\, n \in \mathbb{N}$. The third equality is the tricky one, because we need to be able to swap limits and integrals!

Theorem (Doob's Optional stopping theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration, and $\{X_n\}_{n\geq 0}$ a martingale with respect to the filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Define T as the stopping time and let $T < \infty$ a.s. If any of the following conditions hold, then*

$$\mathbb{E}\left(X_T\right) = \mathbb{E}\left(X_0\right).$$

The conditions are

1. ***Bounded stopping time*** *- $\exists\, N < \infty$ such that $T \leq N$ almost surely.*
2. ***Bounded until stopping time*** *- $\exists\, K < \infty$ such that $\mathbb{P}\left(|X_{T\wedge n}| \leq K\right) = 1, \forall n \geq 0$*
3. ***Bounded increments and finite expected stopping time*** *- $\mathbb{E}\left(T\right) < \infty$ and $\exists\, K < \infty$ such that*

$$\mathbb{P}\left(|X_{T\wedge(n+1)} - X_{T\wedge(n)}| \leq K\right) = 1, \quad \forall\, n \geq 0$$

Proof. Will show that in each case, we can apply the dominated convergence theorem to conclude that

$$\lim_{n\to\infty} \mathbb{E}\left[X_{T\wedge n}\right] = \mathbb{E}\left[\lim_{n\to\infty} X_{T\wedge n}\right].$$

Continue case-by-case.

1. $\mathbb{E}\left[X_T\right] = \mathbb{E}\left[X_{T\wedge N}\right] = \mathbb{E}\left[X_0\right]$. The first equality follows from $T \leq N$ almost surely. The last equality follows from the fact that the stopped martingale is a martingale.
2. $|X_{T\wedge n}| \leq K$ almost surely so K is an integrable dominating function.
3. Decompose the process into the sum of its increments

$$X_{T\wedge n} = X_0 + \sum_{k=1}^{T\wedge n}\left(X_{T\wedge(k)} - X_{T\wedge(k-1)}\right),$$

so from the triangle inequality,

$$|X_{T\wedge n}| \leq |X_0| + \sum_{k=1}^{T\wedge n} |X_k - X_{k-1}| \leq |X_0| + K\left(T \wedge n\right) \leq |X_0| + KT.$$

Therefore the dominating integrable function is $|X_0|+KT$. It is integrable because we know that $\mathbb{E}\left[|X_0|\right] < \infty$ and we assumed $\mathbb{E}\left[|T|\right] < \infty$.

□

5.5 Applications

Example (Gambler's ruin)**.** *Suppose a gambler enters a casino with b dollars. Say losing b dollars is $-b$ and goal winning amount is a. Model as a simple random walk*

$$X_{n+1} = X_n + \xi_{n+1}$$

where $\{\xi_i\}_{i\geq 1}$ are i.i.d. Rademacher random variables. From early, showed that $\{X_n\}_{n\geq 1}$ is a martingale. Let

$$T_x := \inf\{n \geq 0 : X_n = x\}$$

be the first hitting time of x. Then T_{-b} is the first time going bankrupt and T_a is the goal. ***What is the probability of hitting*** *b* ***before*** *a**?*** *Note that $T := T_a \wedge T_b$ is a stopping time.*

Proof. We check the conditions of the optional stopping theorem for $\{X_n\}_{n\geq 0}$ and T. The first condition does not hold (you cannot always stay between, say, 0 and 1). The second condition does hold because the stopped process lives in $\max\{a, b\}$. (It also follows from Polya's recurrence theorem). We can also check the third requirement using coupling. Divide time t into intervals of size $(a+b)$. Note that no matter where you are in the interval $[-a, b]$, the probability that you stop in the next $a+b$ steps is bounded from below by $2^{-(a+b)}$. Define Z as the index of the first block where all jumps are to the right. Note that $Z \sim Geo\left(2^{-(a+b)}\right)$. Then $T \preceq (a+b)\, Z$ (stochastic domination), because Z always implies T, indeed it is always true that $T \leq (a+b)\, Z$. Implies

$$\mathbb{E}\,(T) \leq (a+b)\,\mathbb{E}\,(Z) = (a+b)\,2^{a+b} < \infty.$$

By the optional stopping theorem, $\mathbb{E}\,[X_T] = \mathbb{E}\,[X_0] = 0$. On the other hand, we know that $X_T \in \{-b, a\}$ so

$$\mathbb{E}\,(X_T) = -b\mathbb{P}\,(X_T = -b) + a\mathbb{P}\,(X_T = a) = -b\mathbb{P}\,(X_T = -b) + a\left(1 - \mathbb{P}\,(X_T = -b)\right).$$

Putting the two equalities together, $\mathbb{P}\,(X_T = -b) = \frac{a}{a+b}$.
What is the expected stopping time $\mathbb{E}\,(T)$? Recall $\left\{X_n^2 - n\right\}_{n\geq 0}$ is a martingale. Note that the first two conditions of the optional stopping theorem do not hold. Indeed, we can

always oscillate between 0 and 1 and both would be violated. Let's check condition 3. From before, $\mathbb{E}(T) < \infty$. Moreover, increments are bounded

$$\begin{aligned}&|X^2_{T\wedge(n+1)} - (n+1) - \left(X^2_{T\wedge n} - n\right)| = |X^2_{T\wedge(n+1)} - X^2_{T\wedge n} - 1|\\ &\leq |X_{T\wedge(n+1)}|^2 + |X_{T\wedge(n)}|^2 + 1 \leq 2\max\{a,b\}^2 + 1 < \infty.\end{aligned}$$

Then, after applying the optional stopping theorem, we obtain

$$\mathbb{E}\left[X^2_T - T\right] = \mathbb{E}\left[X^2_0 - 0\right] = 0 \implies \mathbb{E}(T) = \mathbb{E}\left(X^2_T\right) = a^2\frac{b}{a+b} + (-b)^2\frac{a}{a+b} = ab.$$

□

Example (ABRACADABRA problem). *How long does it take, in expectation, for a monkey typing randomly on a typewriter to write "ABRACADABRA"? More precisely, let T be the first time the monkey spells ABRACADABRA. Want to find $\mathbb{E}(T)$. Can create a simple bound using independent blocks (otherwise we would have to think about a "rolling window" of 11 blocks which are not independent). The probability that the monkey writes ABRACADABRA in a given block is 26^{-11}. Then, let $Z \sim Geo\left(\frac{1}{26^{-11}}\right)$. We have that $T \leq 11Z$, (it is not an equality because it can happen that the monkey writes down ABRACADABRA across blocks), so we conclude that $\mathbb{E}[T] \leq \mathbb{E}(Z) = 11 \cdot 26^{11}$. We can now prove the result using the optimal stopping theorem. Introduce a casino, At each time $n = 1, 2, \ldots$ a new gambler comes in with \$1 in their pocket. Each gambler does the following*

1. *First they bet on the letter A. If they lose, they are out. If they win, they get 26 dollars. Note that this is a fair bet.*
2. *If the gambler is still in, they bet 26 dollars on B. Again, if they lose, then they are out. If they win, they get 26^2 dollars. Note that this is still a fair game.*

After time t, t dollars have been bet (only money is coming from the gamblers). Then let X_n gives the wealth increase of the casino until time n. We can think of

$$X_n = n - \left(\textit{money of the people at the casino}\right)$$

Since all bets are fair, $\{X_n\}_{n\geq 0}$ is a martingale. However, it is not bounded nor it is bounded before the stopping time (think about the monkey typing A indefinitely). However, it has bounded increments and finite expected stopping time. To see this note that the martingale cannot increment by more than $11 \cdot 26^{11}$. Claim that we can use the optional stopping theorem. If we can, then

$$\mathbb{E}(X_T) = \mathbb{E}(X_0) = 0.$$

On the other hand,

$$X_T = T - \left(26^{11} + 26^4 + 26\right) \approx 3.67 \cdot 10^{15}$$

because the bettor who gambles on the first A of ABRACADABRA receives 26^{11} *dollars, third A gets* 26^4 *dollars, and final A gets* 26 *dollars. Therefore,*

$$\mathbb{E}\left[T\right] = 26^{11} + 26^4 + 26.$$

In short, defining a martingale allows us to exactly compute the expected time without having to rely on independent blocks.

5.6 Martingale convergence

Example (Polya urns)**.** *Suppose an urn initially has one blue ball and one red ball. Draw a ball uniformly at random, look at its color, put it back together with a ball of the same color. This is a self-reinforcing process, which is often hidden in more complex processes. How do the number of balls in each color evolve?*

Let X_n *give the number of blue balls in the urn when there are* n *total balls. Start with* $X_1 = 1$. *Observe that*

$$\mathbb{E}\left[X_{n+1} \mid X_n\right] = X_n/n\left(X_n + 1\right) + \left(1 - X_n/n\right)X_n = \frac{n+1}{n}X_n.$$

The first term corresponds to a draw of a blue ball. The second term corresponds to a draw of a red ball. We immediately see that after normalizing by $n+1$, *we obtain a martingale. In particular, define* $x_n = X_n/n$

$$\mathbb{E}\left[x_{n+1} \mid \mathcal{F}_n\right] = \mathbb{E}\left[X_{n+1}/(n+1) \mid \mathcal{F}_n\right] = \frac{1}{n+1}\frac{n+1}{n}X_n = x_n.$$

What is the distribution at large times? Consider a simple four-step example

$$\mathbb{P}\left(\textit{Obtaining this specific sequence}\right) = \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{1}{4}\cdot\frac{3}{5}.$$

There are however many ways to obtain 4 blue balls after 6 periods, but they turn out to have the same probability (you basically just swap the numerators of the fractions above)! Indeed, observe that the denominators form a factorial and the numerators are 1 *at some point (from red) and* $1, 2, 3$ *for the blue. Immediately see that the probability of any sequence leading to* $X_6 = 4$ *is the same.*

Processes with this feature are known as ***exchangeable****, because the probability of a given sequence is invariant to permutations of the elements in the sequence. In this case,*

$$\mathbb{P}\left(X_6 = 4\right) = \binom{4}{1} \frac{1 \cdot 1 \cdot 2 \cdot 3}{5!}.$$

More generally,

$$\mathbb{P}\left(X_{n+2} = 1 + k\right) = \binom{n}{k} \frac{\{1 \cdot 2 \cdot \ldots \cdot k\} \{1 \cdot 2 \cdot \ldots n - k\}}{(n+1)\,!} = \frac{1}{n+1}.$$

Wee see that X_{n+1} is uniformly distributed on its support $\{1, 2, \ldots, n\}$ and the specific probability is independent of k. After normalizing, x_n is uniformly distributed on $\{1/n, 2/n, \ldots, (n-1)/n\}$ and $x_n \xrightarrow{d} Uni\,[0,1]$. More is true. Indeed, $\{x_n\}_{n \geq 2}$ converges ***almost surely*** *to a random variable uniformly distributed on $[0, 1]$. This phenomenon is very general for martingales.*

Theorem (Martingale convergence)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_n\}_{n \geq 0}$ be a filtration, $\mathcal{F}_\infty := \sigma\left(\cup_{k \geq 0}, \mathcal{F}_k\right)$, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_\infty \subseteq \mathcal{F}$. Suppose $\{X_n\}_{n \geq 0}$ is a martingale/submartingale/supermartingale with respect to $\{\mathcal{F}_n\}_{n \geq 0}$. Assume $\exists\, K < \infty$ such that $\mathbb{E}\left[|X_n|\right] \leq K$ for all $n \geq 0$ (uniform boundedness in $\mathcal{L}_1$). Then there exists a random variable X_∞ which is $\mathcal{F}_\infty$-measurable such that*

$$X_n \xrightarrow{a.s.} X_\infty.$$

Moreover $\mathbb{E}\left[|X_\infty|\right] \leq K$.

In the previous case x_n, being a fraction, was clearly uniformly bounded and it was also a nice example in which we were able to compute X_∞, which is not necessarily easy to do.

Proof. Assume $\{X_n\}_{n \geq 0}$ is a supermartingale. The following reasoning directly applies to martingales and submartingales with $\{-X_n\}_{n \geq 0}$. Fix $-\infty < a < b < \infty$. Want to understand oscillations of the process over $[a, b]$. The idea is that if the process crosses the interval infinitely many times, then it cannot converge.

Define $U_n^{[a,b]}$ as the number of upcrossings of the interval until time n. Formally,

$$U_n^{[a,b]} = \max\left\{r : \exists\, 0 \leq s_1 < t_1 < s_2 < t_2 < \cdots < s_r < t_r \leq n, x_{s_j} < a \text{ and } x_{t_j} \geq b, \forall j = 1, \ldots, r\right\}.$$

Let's make a casino. Repeat the following steps

1. Wait until $X_\cdot \leq a$.

2. Bet 1 unit of money every time until $X_\cdot \geq b$.

Formally, define the bets as

$$\begin{aligned} B_1 &:= \mathbb{1}_{B_0 \leq a} \\ B_k &:= \mathbb{1}_{B_{k-1}=1, X_{k-1}<b} + \mathbb{1}_{B_{k=1}=0, X_{k-1}\leq a}, \quad k > 1 \end{aligned}$$

The first term corresponds to the case where a bet was placed in the previous round and the process is less than b. The second term corresponds to the case where a bet was not placed in the previous round but the process is less than or equal to a. Note that $\{B_n\}_n$ is a predictable process since it is $\mathcal{F}_n-$measurable.

Let wealth $Y_0 := 0$ and $Y_n := \sum_{k=1}^n B_k \left(X_k - X_{k-1}\right)$. In words, the wealth at time n is the sum of the wealth changes. Claim that $\{Y_n\}_{n\geq 0}$ is a supermartingale because $\{X_n\}_{n\geq 1}$ is a supermartingale:

$$\begin{aligned} \mathbb{E}\left[Y_n \mid \mathcal{F}_{n-1}\right] &= \mathbb{E}\left[Y_{n-1} + B_n\left(X_n - X_{n-1}\right) \mid \mathcal{F}_{n-1}\right] \\ &= Y_{n-1} + B_n \mathbb{E}\left[X_n - X_{n-1} \mid \mathcal{F}_{n-1}\right] \leq Y_{n-1}. \end{aligned}$$

because $\mathbb{E}\left[X_n - X_{n-1} \mid \mathcal{F}_{n-1}\right] \leq 0$ because X_n is a supermartingale. In particular, $\mathbb{E}\left(Y_n\right) \leq \mathbb{E}\left(Y_0\right) = 0$. What are the gains and losses in betting? Gains are at least $(b-a)$ when they occur. Can only lose at the end, because it means that the process never upcrossed the interval anymore.

$$\begin{aligned} Y_n &\geq (b-a)\, U_n^{[a,b]} - \left|X_n - a\right| \\ \mathbb{E}\left(Y_n\right) &\geq (b-a)\,\mathbb{E}\left(U_n^{[a,b]}\right) - \left(|a| + \mathbb{E}\left[\left|X_n\right|\right]\right) \end{aligned}$$

Using that $\mathbb{E}\left(Y_n\right) \leq \mathbb{E}\left(Y_0\right) = 0$ and rearranging terms,

$$\mathbb{E}\left(U_n^{[a,b]}\right) \leq \frac{|a| + \mathbb{E}\left[\left|X_n\right|\right]}{b-a} \leq \frac{|a| + K}{b-a} < \infty.$$

Above is called **Doob's upcrossing inequality**. We know that $\left\{U_n^{[a,b]}\right\}_{n\geq 0}$ is nondecreasing

$$0 \leq U_0 \leq U_1 \leq \ldots \leq U_n \leq U_{n+1}.$$

Thus $U_n \nearrow U_\infty$ almost surely, because it is a non-decreasing and non-negative process.

By the monotone convergence theorem,

$$\mathbb{E}\left[U_\infty\right] = \mathbb{E}\left[\lim_{n\to\infty} U_n\right] = \lim_{n\to\infty} \mathbb{E}\left(U_n\right) \leq \frac{|a|+K}{b-a} < \infty$$

but this can only happen if $U_\infty < \infty$ almost surely, or, equivalently, if $\mathbb{P}\left(U_\infty < \infty\right) = 1$ (otherwise the expectation would blow up).

Upcrossings happen finitely many times. First define the event

$$\left\{\liminf_{n\to\infty} X_n < \limsup_{n\to\infty} X_n\right\} = \left\{\exists\, a, b \in \mathbb{Q} : U_\infty^{[a,b]} = \infty\right\}$$

this is so because because if the limit does not exist it means that the limsup an liminf must differ. Given any two distinct real numbers it is always possible to find two rationals in the middle. They define an interval in which the process must oscillate indefinitely because it must be bounded by its limsup and liminf. Then

$$\mathbb{P}\left(\left\{\liminf_{n\to\infty} X_n < \limsup_{n\to\infty} X_n\right\}\right) \leq \sum \mathbb{P}\left(U_\infty^{[a,b]} = \infty\right) = 0$$

Therefore a limit exists almost surely. Finally,

$$\mathbb{E}\left[\left|X_\infty\right|\right] = \mathbb{E}\left[\lim_{n\to\infty} \left|X_n\right|\right] \leq \limsup \mathbb{E}\left(\left|X_n\right|\right) \leq K.$$

The second-to-last inequality follows from Fatou's lemma. Final equality follows from assumption. □

Corollary. *If $\{X_n\}_{n\geq 0}$ is a nonnegative supermartingale, then $X_n \xrightarrow{a.s.} X_\infty$.*

Proof.

$$\mathbb{E}\left[\left|X_n\right|\right] = \mathbb{E}\left[X_n\right] \leq \mathbb{E}\left(X_0\right) := K < \infty.$$

The first equality follows from non-negativity. The inequality follows from being a supermartingale. Thus the above expression gives bounds. □

Corollary. *If $\{X_n\}_{n\geq 0}$ is a submartingale such that $X_n \leq K$ for some finite K, then $X_n \xrightarrow{a.s.} X_\infty$.*

Proof. Apply the previous corollary with $Y_n := K - X_n$ which is a non-negative supermartingale. □

Theorem (Doob decomposition)**.** *Given an integrable stochastic process $\{X_n\}_{n\geq 1}$ adapted to the filtration $\{\mathcal{F}_n\}_{n\geq 0}$, there exists a decomposition*

$$X_n = M_n + A_n$$

where $\{M_n\}_{n\geq 0}$ is a martingale with respect to filtration $\{\mathcal{F}_n\}_{n\geq 0}$ and $\{A_n\}_{n\geq 0}$ is a $\{\mathcal{F}_n\}_{n\geq 0}$ predictable process, i.e. it is $\mathcal{F}_{n-1}-$measurable. Moreover, this decomposition is unique up to the value of M_0.

The uniqueness part comes from the fact that you can add something to M_0 and subtract it from A_0. Usually $\{A_n\}_{n\geq 0}$ is called a *drift*.

Proof. $A_0 := 0$. $A_n := A_{n-1} + \mathbb{E}\left[X_n - X_{n-1} \mid \mathcal{F}_{n-1}\right]$. By construction, A_n is $\mathcal{F}_{n-1}$-measurable. Let $M_n := X_n - A_n$. Let's check the martingale condition

$$\begin{aligned}\mathbb{E}\left[M_n - M_{n-1} \mid \mathcal{F}_{n-1}\right] &= \mathbb{E}\left[(X_n - X_{n-1}) - (A_n - A_{n-1}) \mid \mathcal{F}_{n-1}\right] \\ &= \mathbb{E}\left[X_n - X_{n-1} \mid \mathcal{F}_{n-1}\right] - (A_n - A_{n-1}) = 0.\end{aligned}$$

The second-to-last equality follows from A_n, A_{n-1} being $\mathcal{F}_{n-1}$ measurable. The last equality follows from the definition of A_n. □

For certain processes, Doob decomposition allows us to subtract the drift component and deal with martingales.

6 Poisson processes and Brownian motion

6.1 Poisson processes

Poisson processes model arrival processes. The graph of a Poisson process is the same as a step function and there is a natural correspondence between the process and when the arrivals happen. They are **continuous time processes**.

Definition (Number of arrivals)**.** *Define the* ***number of arrivals*** *in the time interval* $[0,t]$ *as* $N(t)$ *and let* T_k *be the time of the* $k-$*th arrival,i.e.* $T_k := \inf\{t \geq 0 : N(t) = k\}$.

Definition (Poisson process)**.** $\{N(t)\}_{t\geq 0}$ *is a homogeneous* ***Poisson process*** *with rate* λ *if the following conditions hold:*

1. $\{N(t)\}_{t\geq 0}$ *is a* ***counting processes****, i.e. it is increasing in* t*, it is integer-valued, and* $N(0) = 0$.
2. $\{N(t)\}_{t\geq 0}$ *has* ***independent increments****. That is,* $N_{t+s} - N_t$ *is independent of the natural filtration* $\mathfrak{F}_t$.
3. $\{N(t)\}_{t\geq 0}$ *has* ***stationary increments****:*

$$N_{t+s} - N_t =^d N_s.$$

4. $\{N(t)\}_{t\geq 0}$ *has* ***no simultaneous arrivals*** *that is*

$$\mathbb{P}\left(N(t+h) - N(t) = 1\right) = \lambda h + o(h)$$

and

$$\mathbb{P}\left(N(t+h) - N(t) \geq 2\right) = o(h)$$

almost surely as $h \to 0$.

Regarding property 4, if I look at a very short time interval, then the probability that there is an arrival is proportional to the length of the interval and the probability that there are two or more arrivals is negligibile.

Property (Properties of Poisson process)**.** *What are some characteristics of this process?*

1. ***Distribution of arrival times:*** *Consider interval* $[0,t]$*. Split into* $h-$ *length intervals. Then*

$$N_t \approx N_{\lfloor t/h\rfloor\cdot h} = \sum_{i=1}^{\lfloor t/h\rfloor}\left(N_{ih} - N_{(i-1)h}\right).$$

We can think of such intervals as being independent Bernoulli, hence their sum would be Binomial. From condition (4), $N_t \approx Bin\left(\lfloor t/h \rfloor, \lambda h\right)$ converges in distribution as $h \to 0$ to $Poi\left(\lambda t\right)$. Therefore $N\left(t\right) \sim Poi\left(\lambda t\right)$.

Note that we do not need to assume Poisson distributed arrivals. Suffices to assume properties 1-4, which is a set of weaker assumptions.

2. ***Distribution of inter-arrival times**:*

$$\left\{T_1 > t\right\} = \left\{N\left(t\right) = 0\right\} \implies \mathbb{P}\left(T_1 > t\right) = \mathbb{P}\left(N\left(t\right) = 0\right) = e^{-\lambda t} \implies T_1 \sim Exp\left(\lambda\right).$$

Similarly, $T_{i+1} - T_i \sim Exp\left(\lambda\right)$. Defining $T_0 = 0$ for simplicity, then $\left\{T_{i+1} - T_i\right\}_{i \geq 0}$ are iid $Exp\left(\lambda\right)$.

3. *Recall the **memorylessness property** of the exponential distribution*

$$\mathbb{P}\left(T_1 > t + s \mid T_1 > t\right) = \mathbb{P}\left(T_1 > s\right).$$

this follows from property 2 on independent increments.

4. *Can be shown that $\left\{N\left(t\right) - \lambda t\right\}_{t \geq 0}$ is a martingale.*
5. ***Superposition:** Suppose $M_t \sim PPP\left(\lambda\right)$ and $N_t \sim PPP\left(\mu\right)$. $R_t := M_t + N_t \sim PPP\left(\lambda + \mu\right)$. In words, the union of independent Poisson processes is $PPP\left(\lambda + \mu\right)$.*
6. ***Thinning:** Let $R_t \sim PPP\left(\lambda\right)$. Mark points as 0 or 1 i.i.d. with probability p. Let M_t be a collection of points with probability 0 and N_t be a collection of points with a mark. Then*

$$\left\{M_t\right\}_{t \geq 0} \sim PPP\left(\lambda\left(1 - p\right)\right), \quad \left\{N_t\right\}_{t\ geq0} \sim PPP\left(\lambda p\right)$$

and they are independent.

Lemma. *If $M_1, \ldots, M_k$ are independent Poisson processes with rate λ_i, $n_1 + \ldots + n_k = n$, $\lambda_1 + \ldots + \lambda_k = \lambda$, then*

$$\mathbb{P}\left(M_1 = n_1, \ldots, M_k = n_k \mid \sum_{i=1}^{k} M_i = n\right) = \frac{n!}{\prod_i n_i!} \cdot \prod_{i=1}^{k} \left(\frac{\lambda_i}{\lambda}\right)^{n_1}$$

$$= \mathbb{P}\left(multinomial\left(n, \left\{\frac{\lambda_i}{\lambda}\right\}_{i=1}^{k}\right) = \left\{n_i\right\}_{i=1}^{k}\right).$$

In words, if we partition an interval $\left(0, t\right)$ into subintervals and we know the number of arrivals N on the interval, then the arrivals for the subintervals have multinomial distribution.

In particular,

$$\begin{aligned}\mathbb{P}\left(N_t = k \mid N_{t+s}\right) &= \mathbb{P}\left(N_t = k, N_{t-s} - N_t = l - k \mid N_{t+s} = l\right)\\ &= \binom{l}{k}\left(\frac{t}{t+s}\right)^k\left(\frac{s}{t+s}\right)^{l-k}.\end{aligned}$$

Conditional on $N_{t+s} = l$*, we have that* $N_t \sim Bin\left(l, \frac{t}{t+s}\right)$.

Lemma. *If* $Z \sim Poi(\lambda)$ *and given* Z, $X \sim Bin(Z, p)$ *and* $Y := Z - X$*, then* $X \sim Poi(\lambda p)$, $Y \sim Poi\left(\lambda(1-p)\right)$*, and* X, Y *independent.*

Proof. Let $\{N(t)\}_{t\geq 0}$ be a Poisson process with rate λ. $Z' := N_1, X' := N_p, Y' := N_1 - N_p$. $Z' \sim Poi(\lambda)$. By above calculation, conditionally, given Z', we have $X' \sim Bin(Z', p)$. By construction, $Y = Z' - X'$.
Then we have that $(X, Y, Z) =^d (X', Y', Z')$. We know that $X' \sim Poi(\lambda p), Y' \sim Poi\left(\lambda(1-p)\right)$ and X'/Y' are independent since the intervals $(0, p)$ and $(p, 1)$ do not overlap.
Same statement holds for X and Y. □

Example (Applications of superposition and thinning)**.** *A few simple examples. Proofs are left as an exercise.*

1. $X \sim Exp(\mu), Y \sim Exp(\lambda)$. *Claim that the minimum* $X \wedge Y \sim Exp(\mu + \lambda)$ *and* $\mathbb{P}(X \leq Y) = \frac{\mu}{\mu+\lambda}$.
2. *Waiting for a bus. Passengers* $PPP(\lambda)$ *and buses* $PPP(\mu)$*. Then the number of passengers getting on the bus is* $Geo\left(\frac{\mu}{\mu+\lambda}\right)$.

6.2 Continuous time Markov chains

Definition (Continuous time Markov chains)**.** *Let* S *be a state space.* $\{X_n\}_{n\geq 0}$ *is a discrete time Markov chain on* S *with transition matrix* P*. Take* $\{N_t\}_{t\geq 0} \sim PPP(\lambda)$ *independent of* $\{X_n\}_{n\geq 0}$*. Let* $Y_t := X_{N(t)}$ *move at the times of jumps in the Poisson process.*
Define transition matrix $P(t)$ *as*

$$\begin{aligned}\mathbb{P}\left(Y_t = b \mid Y_0 = a\right) &= \sum_{k=0}^{\infty}\mathbb{P}\left(N(t) = k, X_k = b \mid X_0 = a\right) = \sum_{k=0}^{\infty}\mathbb{P}\left(N(t) = k\right)\mathbb{P}\left(X_k = b \mid X_0 = a\right)\\ &= \sum_{k=0}^{\infty}\frac{(\lambda t)^k}{k!}e^{-\lambda t}\left(P^k\right)_{ab}\\ \implies P(t) &= e^{\lambda t(P-I)}\end{aligned}$$

where the above expression is defined through its power series. Let $Q := \lambda(P - I)$ be the ***transition rate matrix****. Row sums are one. $q_\alpha := \sum_{b:b\neq a} q_{ab}$.*
Moreover, note that $q_{aa} = -q_a$ and $\pi P(t) = \pi \iff \pi Q = 0$.

Note. *There are three possibilities for defining CTMC.*

1. *As above.*
2. *Every directed edge has a Poisson clock with parameter q_{ab}.*
3. *every state (vertex) has a Poisson clock with parameter q_a; when jump, jump accorting to P.*

Definition (Random counting measure). *$PPP(\lambda)$ is a random counting measure on $\mathbb{R}_+$. $N(A) \sim Poi(\lambda|A|)$. If $A_1, A_2, \ldots, A_k$ are disjoint, then $N(A_1), N(A_2), \ldots, N(A_k)$ are mutually independent.*
This concept generalizes. Take any measure μ on $\mathbb{R}_+$ (above $\mu = \lambda Leb$). $PPP(\mu)$ Poisson process on $\mathbb{R}_+$ with intensity measurr μ. A ***random counting measure*** *$N(\cdot)$ satisfies*

1. *$N(A) \sim Poi(\mu(A))$*

2. *If $A_1, .A_2, \ldots, A_k$ disjoint, then $N(A_1), N(A_2, \ldots, N(A_k))$ are independent.*

Example. *Suppose μ has a density $\{\lambda(t)\}_{t\geq 0}$ with respect to the Lebesgue measure. Known as a* ***nonhomogenous Poisson process*** *(think arrivals to a store).*
Can also take μ on any nice space like $\mathbb{R}^d$. Think a spatial Poisson process.

6.3 Brownian motion

Definition (Brownian motion). *A real-valued stochastic process $\{B(t)\}_{t\ geq0}$ is called a* ***(one-dimensional) Brownian motion*** *starting at $x \in \mathbb{R}$ if*

1. *$B(0) = x$;*
2. *The process has independent increments. That is for all $0 \leq t_1 \leq \ldots \leq t_n$*

$$B(t_2) - B(t_1), \ldots, B(t_n) - B(t_{n-1})$$

are independent random variables.
3. *Process has stationary increments that are Gaussian with mean 0 and variance equal to the increment length.*

$$\forall t \geq 0, \forall h > 0, B(t+h) - B(t) \sim \mathcal{N}(0,h)$$

4. *Function $t \mapsto B(t)$ is continuous almost surely.*

*If $x = 0$, then the process is called a **standard Brownian motion**. If $\left\{B\left(t\right)\right\}_{t\geq 0}$ is a standard BM, then $\left\{B\left(t\right)+x\right\}$ is a BM starting at x.*

Property (Finite-dimensional distributions of BM)**.** *Distribution of all finite dimensional RVs $\left(B\left(t_1\right),\ldots,B\left(t_n\right)\right)$ for all $0\leq t_1\leq\ldots\leq t_n$. Can determined these from properties (i)-(iii) above. Suppose $s<t$. Then use independent increments to find the covariance*

$$Cov\left(B\left(s\right),B\left(t\right)\right)=\mathbb{E}\left[B\left(s\right)B\left(t\right)\right]=\mathbb{E}\left[B\left(s\right)\left(B\left(t\right)-B\left(s\right)\right)\right]+\mathbb{E}\left(B\left(s^2\right)\right)=s.$$

By definition, $B\left(s\right)\sim\mathcal{N}\left(0,s\right)$. Putting these together, the joint distribution of s,t is

$$\left(B\left(s\right),B\left(t\right)\right)\sim\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}s&s\\s&t\end{pmatrix}\right)$$

Technique can be extended for additional increments.

Definition (Gaussian process)**.** *A stochastic process $\left\{X\left(t\right)\right\}_{t\in\mathcal{T}}$ is a **Gaussian process** if its finite dimensional distributions are Gaussian.*

Theorem (Wiener)**.** *Standard BM exists.*

Proof. General idea. Paul Levy's canonical construction uses a countable collection of i.i.d. standard Gaussians to construct Bm as a uniform limit of continuous functions, guaranteeing that the limit has continuous paths.
Alternatively, can define a BM has a limit of a random walk. □

Theorem (Donsher's invariance principle, functional CLT)**.** *Let $X_1,X_2,\ldots,X_n$ be i.i.d. with $\mathbb{E}\left(X_1\right)=0$ and $\mathbb{V}\left(X_1\right)=1$, and $S_n\sum_{i=1}^n X_i$. Let*

$$s\left(t\right)=s_{\lfloor t\rfloor}+\left(t-\lfloor t\rfloor\right)\left(s_{\lfloor t+1\rfloor}-s_{\lfloor t\rfloor}\right)$$

give a linear interpolation. Let

$$S_n^*\left(t\right)=\frac{S\left(nt\right)}{\sqrt{n}},\quad\forall t\in\left[0,1\right].$$

Note $\left\{S_n^*\left(t\right)\right\}_{t\in[0,1]}\in C\left[0,1\right]$. Consider $\left(C\left[0,1\right],\left\|\cdot\right\|_\infty\right)$. Then

$$S_n^*\xrightarrow{d}B$$

where $B = \left\{B(t)\right\}_{t\in[0,1]}$ is a standard BM on $[0,1]$. On the space $C[0,1]$ of continuous functions on the unit interval with the metric induced by the sup-norm, the sequence $\{S_n^*\}_{n\geq 1}$ converges in distribution to a standard BM $\left\{B(t)\right\}_{t\in[0,1]}$.
That is, for every bounded continuous function, $f : \left(C[0,1], \|\cdot\|_\infty\right) \to \mathbb{R}$, we have that

$$\mathbb{E}\left[f\left(S_n^*\right)\right] \xrightarrow{n\to\infty} \mathbb{E}\left[f(B)\right].$$

In words, this is a generalization of the central limit theorem. Properly scaled paths look like BM.

Proof. Rough sketch. Convergence of FDD follows from multivariate CLT. Note

$$\operatorname{Cov}\left(S_n^*(s), S_n^*(t)\right) = \frac{1}{n}(ns \wedge nt) = n \wedge t.$$

Takes some work to show that the limit is continuous almost surely. □

Property (Brownian motion)**.** *Let $\left\{B(t)\right\}_{t\geq 0}$ be a standard BM.*

1. ***Symmetry**: $\widetilde{B}(t) = -B(t), t \geq 0 \implies \left\{\widetilde{B}(t)\right\}_{t\geq 0}$ is also a standard BM.*
2. ***Time homogeneity**: Fix $s > 0$. $\widetilde{B}(t) = B(s+t) - B(s), t \geq 0$. Then $\left\{\widetilde{B}(t)\right\}_{t\geq 0}$ is also a standard BM.*
3. ***Time reversal**: Fix $T > 0$. $\widetilde{B}(t) := -B(T) + B(T-t)$. Then $\left\{\widetilde{B}(t)\right\}_{t\in[0,T]}$ is a standard BM.*
4. ***Scale invariance**. Fix $c > 0$. Let*

$$X(t) = \frac{1}{\sqrt{c}} B(ct).$$

Then $\left\{X(t)\right\}_{t\geq 0}$ is a BM. In words, sample paths of BM are random fractals.

A Summary of Main Results

Recap 1: Measure Theory

- **Sigma-algebra -** A collection of sets $\mathcal{E}$ is a sigma-algebra on a E if
  1. $\emptyset \in \mathcal{E}$
  2. it is closed under complements
  3. it is closed under countable union
- **Generated sigma-algebra -** The sigma-algebra generated by a collection of sets $\mathcal{C}$ is the smallest sigma-algebra containing $\mathcal{C}$ and it is denote as $\sigma(\mathcal{C})$. The sigma-algebra generated by a RV is

$$\sigma(X) = \left\{ B \subseteq \Omega : B = X^{-1}(A), A \in \mathcal{F} \right\}$$

- **Measurable space -** it is a non-empty set and a sigma-algebra $(E, \mathcal{E})$
- **Measure -** A set function $\mu : \mathcal{E} \to \overline{R}_+$ on a measurable space $(E, \mathcal{E})$ is a measure
  1. if $\mu(\emptyset) = 0$
  2. and it is countably additive
- **Measure Space -** it is a triplet $(E, \mathcal{E}, \mu)$.
- **Probability Space -** it is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathbb{P}(\Omega) = 1$.
- **Measurable function -** A function f from a measurable space $(E, \mathcal{E})$ to another measurable space $(F, \mathcal{F})$ is measurable if

$$f^{-1}(B) \in \mathcal{E}, \forall B \in \mathcal{F}$$

- **Checking on collections** $\mathcal{C}$ **-** if a sigma algebra $\mathcal{E}$ is generated by a collection of sets $\mathcal{C}$, then we can use just $\mathcal{C}$ to check:
  1. if two measures on $(E, \mathcal{E})$ are identical
  2. if two probability distributions are identical. If the codomain is endowed with $\mathscr{B}_{\mathbb{R}^n}$ then we can just use CDFs
  3. if a function f is measurable with respect to $\mathcal{F}$ and $\mathcal{E}$
- **Integration**
  - a function $f : E \to \overline{R}_+$ is a positive function
  - Every simple function is measurable (sum of indicators)
  - A positive function is measurable iff it is the limit of an increasing sequence of positive simple functions
  - if f is simple the integral is $\mu(f) = \sum_{i=1}^{n} a_i \mu(A_i)$
  - if f is measurable and positive the integral is $\mu(f) = \lim_{n\to\infty} \mu(d_n \circ f)$
  - if f is measurable the integral is $\mu(f) = \mu(f^+) - \mu(f^-)$

- the integral exists if at least one of $\mu\left(f^{+}\right)$ and $\mu\left(f^{-}\right)$ is finite, thus the integral always exists on $\overline{R}_{+}$ for positive functions
- a function is **integrable** if $\mu\left(|f|\right) < \infty$
- To **interchange limits and integrals** take f_n measurable $\forall\, n$:
  1. *Monotone Convergence Theorem* for increasing and positive $\left\{f_n\right\}_n$
  2. *Dominated Convergence Theorem* for dominated and positive $\left\{f_n\right\}_n$
  3. *Bounded Convergence Theorem* for bounded and positive $\left\{f_n\right\}_n$

$$\lim_{n\to\infty} \mu\left(f_n\right) = \mu\left(\lim_{n\to\infty} f_n\right) \qquad \text{(under 1,2,3)}$$

  4. *Fatou's Lemma* for positive $\left\{f_n\right\}_n$

$$\mu\left(\liminf_{n\to\infty} f_n\right) \leq \liminf_{n\to\infty} \mu\left(f_n\right)$$

- To **change order of integration**
  1. *Fubini*, if f is integrable, i.e.

$$\int_{E\times F} |f| d\left(\mu \times \nu\right) < \infty$$

  2. *Tonelli*, always if the integrand is positive
- **Expectations are integral**

Recap 2: Asymptotics and the Law of Large Numbers

- **General Markov Inequality -** Let X be a positive random variable and $f : \mathbb{R} \to \mathbb{R}_+$ be non-decreasing, then

$$\mathbb{P}\left(f(X) \geq \lambda\right) \leq \frac{\mathbb{E}\left[f(X)\right]}{f(\lambda)}, \quad \forall \lambda > 0$$

- **Chebyshev's Inequality -** Let X be a real-valued RV with $\mathbb{E}\left(X^2\right) < \infty$, then

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq \lambda\right) \leq \frac{\mathbb{V}(X)}{\lambda^2}, \quad \forall \lambda > 0$$

- **Chernoff's bound -** Let X be a random variable, then

$$\mathbb{P}(X \geq \lambda) \leq \inf_{t \geq 0} \frac{\mathbb{E}\left[e^{tX}\right]}{e^{t\lambda}}, \quad \forall \lambda > 0,$$

  It is Markov with $f(X) = e^{tX}\ \forall t \geq 0$
- **Convergence in probability -** $X_n \xrightarrow{p} X$ if

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}\left(|X_n - m| \geq \varepsilon\right) = \lim_{n\to\infty} \mathbb{P}\left(\left\{\omega \in \Omega : |X_n(\omega) - m| \geq \varepsilon\right\}\right) = 0$$

- **Almost sure convergence -** $X_n \xrightarrow{a.s.} X$ if

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = 0\right\}\right) = \mathbb{P}\left(\lim_{n\to\infty} X_n = 0\right) = 1.$$

- **Borel-Cantelli**
  I. $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \implies \mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 0$
  II. $\{A_n\}_n$ disjoint, then $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \implies \mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1$
- **WLLN** If $\{X_n\}_n$ are integrable and $\mathbb{C}\left(X_i, X_j\right) = \rho\left(|i-j|\right), \rho(x) \to 0$, then

$$\overline{X}_n \xrightarrow{p} \mathbb{E}[X_i]$$

- **SLLN** If $\{X_n\}_n$ are integrable and i.i.d., then

$$\overline{X}_n \xrightarrow{a.s.} \mathbb{E}[X_i].$$

- To prove convergence in probability use bounds
- To prove convergence almost surely rely on Borel-Cantelli lemmas or SLLN

Recap 3: Central Limit Theorem and Characteristic Functions

- **Weak Convergence -** X_n weakly converges to X if CDF of X_n converges pointwise to the CDF of X for all the continuity points of X. To prove it
  1. use definition above
  2. if $\mathbb{E}\left[g\left(X_n\right)\right] \Rightarrow \mathbb{E}\left[g\left(X\right)\right]$ for all continuous and bounded functions g
  3. convergence of characteristic functions (Levy's continuity)
  4. tightness (Prohorov's theorem)
- **CLT -** If $\left\{X_n\right\}_n$ are iid with finite second moments, mean-zero and variance 1, then

$$\frac{S_n - \mathbb{E}\left[S_n\right]}{\sqrt{\mathbb{V}\left(S_n\right)}} \Rightarrow N\left(0, 1\right)$$

- **Characteristic Functions -** $\phi_X\left(t\right) = \mathbb{E}\left[e^{itX}\right]$
  1. if X and Y are independent, then $\phi_{X+Y} = \phi_X \phi_Y$
  2. if $\phi_X = \phi_Y$, then $X \overset{d}{=} Y$
  3. Moments can be computed from characteristic functions by taking the $k-$th derivative and evaluating it at 0

$$\phi^{(k)}\left(0\right) = \mathbb{E}\left[\left(iX\right)^k\right] = i^k \mathbb{E}\left[X^k\right]$$

  4. Levy's Continuity theorem.
     - $F_n \Rightarrow F$ implies that $\phi_{X_n} \to \phi_X$ pointwise in t.
     - if $\phi_{X_n} \to \phi_X$ pointwise in t and ϕ_X is continuous at $t = 0$, then $F_n \Rightarrow F$
- **Uniform Tightness -** X_n is uniformly tight if

$$\forall\, \varepsilon > 0, \exists\, M \in \mathbb{R}_{++} : \mathbb{P}\left(|X_n| \geq M\right) \leq \varepsilon, \forall, n \in \mathbb{N}$$

- **Prohorov's Theorem**
  1. if $F_n \Rightarrow F$ then F_n is uniformly tight
  2. if F_n is uniformly tight then there exists a subsequence $\left\{F_{n_k}\right\}_k$ that weakly converges

Recap 4: Markov Chains

- **Markov Property -** Given a stochastic process $\{X_n\}_{n\geq 0}$ with state space S

$$\mathbb{P}\left(X_{n+1}=a_{n+1}|X_n=a_n,X_{n-1}=a_{n-1},\dots,X_0=a_0\right)=\mathbb{P}\left(X_{n+1}=a_{n+1}|X_n=a_n\right)$$

- Right multiply the transition matrix to get

$$\left(P^n f\right)(x)=\mathbb{E}\left[f\left(X_n\right)\mid X_0=x\right] \tag{forward}$$

- Left multiply the transition matrix with $\mu\left(y\right)=\mathbb{P}\left(X_0=y\right)$ to get

$$\left(\mu P^n\right)=\mathbb{P}\left(X_n=x\right) \tag{backward}$$

- **Classification of states/MC**
  - a state A is *closed* if $\mathbb{P}\left(A\to A^c\right)=0$
  - The closure of A is the smallest closed set containing A
  - a state A is irreducible (communicating class) if $A=\overline{A}$ (minimal closed) and all its parts are connected $\forall\, x,y\in A,\exists\, n_0\in\mathbb{N}:\left(P^{n_0}\right)_{xy}>0$
  - elements of communicating classes share the same properties, such as recurrence, transience, period
  - a state is *recurrent* if it is part of an irreducible component
  - a state is *transient* if it is not recurrent
  - the *period* of a state is the GCD of the length of the walks returning to it
  - a state is *aperiodic* if it has period 1
  - a state is *ergodic* if it is irreducible and aperiodic
- If S is finite then there always exists an irreducible component. If S is not finite this need not be true.
- **Stationary Distribution -** $\pi=\pi P$, therefore the stationary distribution is the left eigenvector of P with eigenvalue 1 with norm normalized to 1
  - **existence:** π exists if the chain is irreducible
  - **uniqueness:** π is unique iff MC is irreducible
  - **convergence:** if MC is ergodic then $X_n\Rightarrow X$, if it is irreducible with period k then

$$\frac{1}{k}\sum_{j=1}^{k}\mathcal{L}\left(X_{n+j-1}\right)\Rightarrow\pi$$

- **ergodic theorem (SLLN) -** if MC is irreducible, then $n^{-1}\sum_i f\left(X_i\right)\overset{a.s.}{\to}\mathbb{E}\left[f\left(X\right)\right]$
- **Polya's recurrence theorem -** The SSRW on $\mathbb{Z}^d$ is recurrent if $d\in\{1,2\}$ and transient if $d\geq 3$
- Let $T_x:=\inf\left\{n\in\mathbb{N}_+:X_n=x\right\},N_x=\sum_{n=1}^{\infty}\mathbb{1}\left(X_n=x\right)$
  1. **Transient:** if $\mathbb{P}\left(T_x<\infty\mid X_0=x\right)<1$ or $\mathbb{E}\left[N_x\right]<\infty$
  2. **Recurrent:** if $\mathbb{P}\left(T_x<\infty\mid X_0=x\right)=1$ or $\mathbb{E}\left[N_x\right]=\infty$
     - **Positive Recurrent**: if $\mathbb{E}_x\left[T_x\right]<\infty$
     - **Null Recurrent**: if $\mathbb{E}_x\left[T_x\right]=\infty$

Recap 5: Martingales

- **Conditional Expectation -** Let $(\Omega, \mathcal{F}, \mathbb{P})$ and X be a real-valued RV. The conditional expectation of X given $\mathcal{G} \subseteq \mathcal{F}$ is a RV Y such that:
  1. Y is $\mathcal{G}-$measurable
  2. $\int_G Y d\mathbb{P} = \int_G X d\mathbb{P}, \forall\, G \in \mathcal{G}$
- **Filtration -** Given a totally ordered set $\mathcal{T}$, a filtration is a collection of sigma-algebras $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ such that $\mathcal{F}_s \subseteq \mathcal{F}_t$ for all $s < t$.
- **Natural Filtration -** Given a stochastic process $\{X_n\}_{n\geq 0}$ its **natural filtration** is given by

$$\mathcal{F}_n := \sigma\left(X_0, X_1, \ldots, X_n\right).$$

- **Adapted -** $\{X_t\}_{t\in\mathcal{T}}$ is **adapted** to filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ if X_t is $\mathcal{F}_t$-measurable for all $t \in \mathcal{T}$.
- **Martingale -** $\{X_t\}_{t\in\mathcal{T}}$ is a Martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ if
  1. $\{X_t\}_{t\in\mathcal{T}}$ is adapted to $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$.
  2. $\forall\, t \in \mathcal{T}$, X_t is integrable. That is $\mathbb{E}\left[|X_t|\right] < \infty$.
  3. $\forall\, s, t \in \mathcal{T}$ such that $s < t$,

$$\mathbb{E}\left[X_t \mid \mathcal{F}_s\right] = X_s. \qquad \text{(martingale property)}$$

- Using the definition and LIE we get $\mathbb{E}\left[X_t\right] = \mathbb{E}\left[X_0\right], \forall\, t \in \mathcal{T}$
- **Supermartingale -** $\mathbb{E}\left(X_{n+m} \mid \mathcal{F}_n\right) \leq X_n, \quad \forall\, m.$
- **Submartingale -** $\mathbb{E}\left(X_{n+m} \mid \mathcal{F}_n\right) \geq X_n, \quad \forall\, m.$
- **Stopping time -** T is a stopping time if $\forall\, n \in \mathbb{N}, \left\{\omega \in \Omega : T(\omega) = n\right\} \in \mathcal{F}_n$.
- **Stopped process -** If X_n is sub/super/martingale, then $X_n^T := X_{T\wedge n}$ is a stopped process
- If $\{X_n\}_{n\geq 0}$ is a sub/super/martingale, so is $\left\{X_n^T\right\}_{n\geq 0}$ hence

$$\forall\, n \in \mathbb{N}, \mathbb{E}\left[X_{T\wedge n}\right] = \mathbb{E}\left[X_{T\wedge 0}\right] = \mathbb{E}\left[X_0\right], \qquad \lim_{n\to\infty} \mathbb{E}\left[X_{T\wedge n}\right] = \mathbb{E}\left[X_0\right]$$

- **Optional stopping theorem -** If $\{X_n\}_{n\geq 0}$ is a martingale and T is an a.s. finite stopping time, then

$$\mathbb{E}\left[X_0\right] = \lim_{n\to\infty} \mathbb{E}\left[X_{T\wedge n}\right] = \mathbb{E}\left[\lim_{n\to\infty} X_{T\wedge n}\right] = \mathbb{E}\left[X_T\right]$$

  if at least one the following three conditions is satisfied :
  1. bounded stopping time
  2. bounded process until stopping time
  3. bounded increments and finite expected stopping time

- **Martingale convergence -** $\{X_n\}_{n\geq 0}$ is a sub/super/martingale with respect to $\{\mathcal{F}_n\}_{n\geq 0}$. Assume $\exists\, K < \infty$ such that $\mathbb{E}\left[|X_n|\right] \leq K$ for all $n \geq 0$, then

  $$X_n \xrightarrow{a.s.} X_\infty.$$

  Moreover $\mathbb{E}\left[|X_\infty|\right] \leq K$.
- If $\{X_n\}_{n\geq 0}$ is a nonnegative supermartingale, then $X_n \xrightarrow{a.s.} X_\infty$.
- If $\{X_n\}_{n\geq 0}$ is a submartingale such that $X_n \leq K$ for some finite K, then $X_n \xrightarrow{a.s.} X_\infty$.
- **Doob decomposition - Given an integrable stochastic process $\{X_n\}_{n\geq 1}$ adapted to the filtration $\{\mathcal{F}_n\}_{n\geq 0}$, there exists a decomposition**

  $$X_n = M_n + A_n$$

  where $\{M_n\}_{n\geq 0}$ is a martingale with respect to filtration $\{\mathcal{F}_n\}_{n\geq 0}$ and $\{A_n\}_{n\geq 0}$ is a $\{\mathcal{F}_n\}_{n\geq 0}$ predictable process, i.e. it is $\mathcal{F}_{n-1}-$measurable. Moreover, this decomposition is unique up to the value of M_0.

B Relevant Mathematical Facts

B.1 Series

The following series (Harmonic series) diverges[2]

$$\lim_{n\to\infty}\sum_{n=1}^{\infty}\frac{1}{n}=\infty$$

and it is a benchmark to verify convergence of other series. Indeed, if the summands decay slower than $1/n$ we now that their sum diverges, whereas if the summands decay faster their sum will converge. Two examples

$$\lim_{n\to\infty}\sum_{n=1}^{\infty}\frac{1}{\log n}=\infty,\qquad \lim_{n\to\infty}\sum_{n=1}^{\infty}\frac{1}{n^{\kappa}}<\infty,\ \ \kappa>1$$

If $\kappa=2$, this is known as the Basel problem and

$$\lim_{n\to\infty}\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi}{6}<2$$

The general setting is the Bertrand series

$$\sum_{n\geq 2}\frac{1}{n^{\alpha}\left(\ln n\right)^{\beta}}<\infty \iff \alpha>1,\ \text{or}\ \alpha=1,\beta>1$$

Note that the Bertrand series does not converge if $\alpha=\beta=1$, indeed

$$\int_{2}^{\infty}\frac{1}{x\ln\left(x\right)}\mathrm{d}x=\Big[\ln\left(\ln\left(x\right)\right)\Big]_2^{\infty}=\infty$$

by the integral test.

Definition of e^x

$$e^x=\sum_{n=0}^{\infty}\frac{x^n}{n!}$$

B.2 Inequalities

$$\frac{1}{e}>\frac{1}{3},\quad e<3$$

[2] Most of this results hold even if the series starts at 0.

Some inequalities with the max operator

$$\max_{1 \le i \le n} |X_i| \le \sum_{i=1}^{n} |X_i|$$
$$\exp\left(\max_{1 \le i \le n} X_i\right) = \max_{1 \le i \le n} \exp(X_i) \le \sum_{i=1}^{n} \exp(X_i)$$

Bernoulli's inequality

$$(1-x)^n \ge 1 - nx$$

B.3 Integration by parts

If you don't recall the formula for integration by parts, just use the definition of derivative of a product and take the integral of both sides

$$\frac{\partial}{dx}\left[f(x)\,g(x)\right] = f'(x)\,g(x) + f(x)\,g'(x)$$
$$f(x)\,g(x) = \int f'(x)\,g(x)\,dx + \int f(x)\,g'(x)\,dx$$
$$\int f(x)\,g'(x)\,dx = f(x)\,g(x) - \int f'(x)\,g(x)\,dx$$

B.4 List of limits

$$\lim_{x\to+\infty}\left(\frac{x}{x+k}\right)^{x} = e^{-k}$$
$$\lim_{x\to 0}(1+x)^{\frac{1}{x}} = e$$
$$\lim_{x\to 0}(1+kx)^{\frac{m}{x}} = e^{mk}$$
$$\lim_{x\to+\infty}\left(1+\frac{k}{x}\right)^{mx} = e^{mk}, k, m \in \mathbb{R}$$
$$\lim_{x\to 0}\left(1+a\left(e^{-x}-1\right)\right)^{-\frac{1}{x}} = e^{a}$$
$$\lim_{x\to 0} xe^{-x} = 0$$
$$\lim_{x\to\infty} xe^{-x} = 0$$
$$\lim_{x\to 0}\left(\frac{a^{x}-1}{x}\right) = \ln a$$
$$\lim_{x\to 0}\left(\frac{e^{x}-1}{x}\right) = 1$$
$$\lim_{x\to 0}\left(\frac{e^{ax}-1}{x}\right) = a$$

B.5 Taylor Expansions (around 0)

$$\begin{aligned}
\ln(1-x) &= x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots - \frac{x^n}{n} + O\left(x^{n+1}\right)\\
\ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + \frac{(-1)^{n+1}x^n}{n} + O\left(x^{n+1}\right)\\
e^{x} &= 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + O\left(x^{n+1}\right)\\
\frac{1}{1+x} &= 1 - x + x^2 - x^3 + x^4 \cdots + (-1)^n x^n + O\left(x^{n+1}\right)\\
\frac{1}{1-x} &= 1 + x + x^2 + x^3 + x^4 \cdots + (-1)^n x^n + o\left(x^{n}\right)
\end{aligned}$$

B.6 Exact Taylor Expansion

$$f(y) = f(x) + \nabla f(x)'(y-x) + \frac{1}{2}(y-x)'\nabla^2 f(z)(y-x), \quad z \in [x, y]$$

B.7 List of Distributions

- **Bernoulli**, $X \sim Be\left(p\right)$, $p \in [0,1]$, $\mathcal{X} = \{0,1\}$

$$f_X\left(x;p\right) = p^x\left(1-p\right)^{1-x}, \quad \mathbb{E}\left[X\right] = p, \quad \mathbb{V}\left[X\right] = p\left(1-p\right)$$

- **Binomial**, $X \sim Bi\left(n,p\right)$, $\left(n,p\right) \in \mathbb{N} \times [0,1]$, $\mathcal{X} = 0,1,\ldots,n$

$$f_X\left(x;n,p\right) = \binom{n}{x} p^x\left(1-p\right)^{n-x}, \quad \mathbb{E}\left[X\right] = np, \quad \mathbb{V}\left[X\right] = np\left(1-p\right)$$

- **Geometric**, $X \sim Geo\left(p\right)$, $p \in [0,1]$, $\mathcal{X} = 0,1,\ldots$

$$f_X\left(x;p\right) = \left(1-p\right)^{x-1}p, \quad \mathbb{E}\left[X\right] = \frac{1-p}{p}, \quad \mathbb{V}\left[X\right] = \frac{1-p}{p^2}$$

- **Poisson**: $X \sim Poi\left(\lambda\right)$, $\lambda \in \mathbb{R}_{++}$, $\mathcal{X} = 0,1,\ldots$

$$f_X\left(x;\lambda\right) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \mathbb{E}\left[X\right] = \lambda, \quad \mathbb{V}\left[X\right] = \lambda$$

- **Exponential**: $X \sim Exp\left(\lambda\right)$, $\lambda \in \mathbb{R}_{++}$, $\mathcal{X} \in \mathbb{R}_{+}$

$$f_X\left(x;\lambda\right) = \lambda e^{-\lambda x}\mathbb{1}_{[0,\infty)}\left(x\right), \quad \mathbb{E}\left[X\right] = 1/\lambda, \quad \mathbb{V}\left[X\right] = 1/\lambda^2$$

- **Gaussian**: $X \sim N\left(\mu,\sigma^2\right)$, $\left(\mu,\sigma^2\right) \in \mathbb{R} \times \mathbb{R}_{++}$, $\mathcal{X} \in \mathbb{R}$

$$f_X\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{\left(x-\mu\right)^2}{2\sigma}}, \quad \mathbb{E}\left[X\right] = \mu, \quad \mathbb{V}\left[X\right] = \sigma^2$$

C Exercises on Almost Sure Convergence

1) Let $X_1, X_2, \ldots$ be mutually independent random variables with the following distributions: $\mathbb{P}(X_n = 0) = 1 - 1/n^2$ and $\mathbb{P}\left(X_n = n^2\right) = 1/n^2$ for every $n \geq 1$. Show that $\mathbb{E}[X_n] = 1$ for every $n \geq 1$, and show also that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to 0 \quad a.s$$

Proof. Define the event $A_n := \{X_n \neq 0\}$, then $\mathbb{P}(A_n) = 1/n^2$. Note that

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

Thus we can apply Borel-Cantelli I and conclude that only finitely many of the $\{A_n\}_{n\geq 1}$ occur, thus $X_n \overset{a.s}{\to} 0$ which also implies the claim above. □

2) Suppose that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - m)\right| > \frac{1}{r}\right) \leq \frac{Cr^4}{n^2}, \quad \forall\, r \in \mathbb{N}$$

Conclude using the Borel-Cantelli lemma that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to m \quad \text{a.s.}$$

Proof. Define the event $A_{n,r} := \left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - m)\right| > \frac{1}{r}\right\}$. Fix $r \in \mathbb{N}$, then note that $\mathbb{P}(A_n) \leq \frac{Cr^4}{n^2}$ by assumption, thus

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) \leq \sum_{n=1}^{\infty} \frac{Cr^4}{n^2} = Cr^4 \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

By Borel-Cantelli we have that

$$\mathbb{P}\left(\limsup A_{n,r}\right) = 0, \forall\, r \in \mathbb{N}$$

which implies by the union bound that

$$0 \leq \mathbb{P}\left(\cup_{r=1}^{\infty} \limsup A_{n,r}\right) \leq \sum_{r=1}^{\infty} \mathbb{P}(A_{n,r}) = 0$$

Therefore, consider the complement of this event, i.e.

$$\left(\cup_{r=1}^{\infty} \cap_{N=1}^{\infty} \cup_{n=N}^{\infty} A_{n_r}\right)^c = \cap_{r=1}^{\infty} \cup_{N=1}^{\infty} \cap_{n=N}^{\infty} A_{n,r}^c$$

which is the exact definition of limit, indeed it reads

$$\forall\, r \in \mathbb{N}, \exists\, N \in \mathbb{N} : \forall\, n \geq N \left| \frac{1}{n} \sum_{i=1}^{n} (X_i - m) \right| \leq \frac{1}{r}$$

and this event has probability 1, thus it holds almost surely. Since it holds for all $r \in \mathbb{N}$, it must hold also in the limit, so we conclude that

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

□

3) Suppose X_n is a standard normal and you have the following bound

$$\frac{1}{\sqrt{2\pi}} \left(\frac{1}{x} - \frac{1}{x^3} \right) e^{-x^2/2} \leq \mathbb{P}\left(X_1 \geq x\right) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2}$$

Show that

$$\limsup_{n \to \infty} \frac{X_n}{\sqrt{2 \log n}} = 1, \quad \text{a.s.}$$

Proof. Note that we have only bounds on the probability of □

4) Let $\{U_i\}_{i=1}^n$ be a sequence of standard uniform distributions. Let $M_n := \max_{1 \leq i \leq n} U_i$. Show that $M_n \overset{p}{\to} 0$.

Proof. We want to show $\lim \mathbb{P}\left(|M_n| > \varepsilon\right) = 0, \forall\, \varepsilon \in (0, 1)$. First note that $|M_n| = M_n$, then

$$\mathbb{P}\left(M_n > \varepsilon\right) = \mathbb{P}\left(X_1 > \varepsilon\right)^n = (1 - \varepsilon)^n$$

which gives us that

$$\lim_{n \to \infty} \mathbb{P}\left(|M_n| > \varepsilon\right) = \lim_{n \to \infty} \mathbb{P}\left(M_n > \varepsilon\right) = \lim_{n \to \infty} (1 - \varepsilon)^n = 0$$

as $(1 - \varepsilon) \in (0, 1]$. □

Show that $M_n \overset{a.s.}{\to} 0$.

Proof. Consider the event $A_n := \{M_n \geq \varepsilon\}$, then

$$\mathbb{P}\left(A_n\right) = (1-\varepsilon)^n \implies \sum_{n=1}^{\infty} \mathbb{P}\left(A_n\right) = \frac{1}{\varepsilon} < \infty$$

Thus we can apply Borel-Cantelli I and have that

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 0$$

which gives us that

$$\limsup_{n\to\infty} M_n \geq \varepsilon, \forall\, \varepsilon > 0, \quad \text{a.s.}$$

thus this holds also in the limit for ε, thus

$$\limsup_{n\to\infty} M_n \geq 0 \quad \text{a.s.}$$

Since $M_n \geq 0$, we can also conclude that $\liminf M_n = \limsup A_n$ which finally implies that

$$\mathbb{P}\left(\lim_{n\to\infty} M_n = 0\right) = 1$$

□

Show that $\liminf_{n\to\infty} n^{\alpha} M_n \geq 1$ almost surely for $\alpha > 2$.