Probability Bootcamp Summer 2024

Instructor: Ian McPherson

Contents

1 Lecture: Sources of Randomness and Observables

1.1 Probability Triple $(\Omega, \mathcal{F}, \mathbb{P})$

When we model the *randomness* that occurs from physical experiments - meant in the broadest possible interpretation of the word - we with to quantify these notion of randomness that makes explicit our implicit common rational understanding of the notion of *probabilities*. Today, we review how we can use the mathematical frameworks from *set-theory* and *measure-theory* to make explicit these intuitions about randomness. In words,

probabilistic events take on a physical and geometric intuition, allowing that can be manipulated under a well-studied framework.

Let's make this precise by defining each constituent of the *probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$, comprised of a sample space, σ-algebra, and probability measure respectively.

Definition 1.1: Sample Space: Ω

The set of all possible outcomes of an experiment is called the *sample space* and is denoted by Ω. We refer to specific outcomes as $\omega \in \Omega$.

Notice, this is a generic set without any structure. However, one is interested in being able to assign probabilities to events - that are combinations of outcomes in Ω. Morally, we need to allow operations such as

$$A \cup B, \quad A \cap B, \quad A^c,$$

where $A, B \subseteq \Omega$, as this captures the statements of A and B, A or B occurring, and not A occurring respectively. This motivates the following definition

Definition 1.2: σ-field: $\mathcal{F}$

A collection $\mathcal{F}$ of subsets of Ω is called a *σ-field* if it satisfies the following:

a) $\emptyset \in \mathcal{F}$;

b) If $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$;

c) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.

If $A \subseteq \Omega$ is such that $A \in \mathcal{F}$, we say that A is a (measurable) *event*.

Warning. One is initially tempted to simply allow any subset of Ω, however, by *measure-theoretic considerations* we would **not be able to assign a probability measure to such a collection**. Specifically, it becomes possible to construct a pathological $A \subseteq \Omega$ such that $\mathbb{P}(A) = 0$ and $\mathbb{P}(A) = 1$.

Lastly, we consider a function that spits out the probability of any event in $\mathcal{F}$:

Definition 1.3: Probability Measure: $\mathbb{P}$

A *probability measure* $\mathcal{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying

a) **Event of Everything and Nothing**: $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$;

b) **Sum of Constituent Subevents**: if $A_1, A_2, \dots \in \mathcal{F}$ are pairwise disjoint, then

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

This codifies the intuition that mathematical probability allows us to *geometrize* events.

1.2 Basic Properties of the Probability Measure

One can easily derive the following properties from basic set-theory:

Lemma 1.1: Basic Properties of $\mathbb{P}$

The following hold:

a) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;

b) If $B \supseteq A$, then
$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A);$$

c) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;

Remark. This can be made more general.

Proof. Exercise. □

In fact, for the proof of (b) there is a general idea used. We will briefly discuss the general technique in the proof sketch of the following *technical* property of $\mathcal{P}$:

Lemma 1.2: Continuity of $\mathbb{P}$

Let $(A_i)_{i=1}^{\infty} \uparrow A$, i.e.,
$$A = \bigcup_{i=1}^{\infty} A_i = \lim_{i \to \infty} A_i,$$
then $\mathbb{P}(A) = \lim_{i \to \infty} \mathbb{P}(A_i)$.

Analogously, if $(B_i)_{i=1}^{\infty} \downarrow B$, i.e.
$$B = \bigcap_{i=1}^{\infty} B_i = \lim_{i \to \infty} B_i,$$
then $\mathbb{P}(B) = \lim_{i \to \infty} \mathbb{P}(B_i)$.

Proof. (*Sketch*) Disjointification. More specifically, we can see
$$A = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \dots.$$
□

In these two lemmas, we are careful about overcounting/overlaps. However, it is often useful to have a coarse overestimate by not caring about overlaps, we record the following obvious but useful tool:

Lemma 1.3: Union Bound

Let $(A_i) \subseteq \mathcal{F}$ be a collection of events (finite or countably infinite). Then
$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i).$$

Moral Analysis Intuition. One generally considers taking approximations of some object of interest - in this setting this may probabilities, in other settings this may be functions. Analysis gives us many tools to bound this error, but one should be as cheap/coarse as the application allows. Some upper bounds can be extremely tight, but require a lot of assumptions to get working, when in reality one may not need to wrestle with such edge cases for a desired result to hold. In words, it is often better to be lazy and then tighten approximation bounds as needed, instead of the other way around.

1.3 Random Variables, Distribution Function, and Multivariate Extensions

Now that we have recalled how one models probabilities, we focus on the reason we even modeled these objects in the first place: *random variables.* Since we are not interested in the randomness of the outcomes themselves, but rather the consequence of those random outcomes,

random variables allow us to quantify randomness to these consequences of interest.

Morally, one can think of such random variables as *maps* from Ω to $\mathbb{R}$, which we denote $X : \Omega \to \mathbb{R}$, where we can assign probabilities numerical outcomes.

Naively, one would wish to assign a probability to every outcome $\omega \in \Omega$, that is one would like there to be an $f : \mathbb{R} \to [0, 1]$ such that

$$f(x) = \mathbb{P}(X = x).$$

However, this approach fails in general (try and remember why!). However, we can always attach with these outcomes a *distribution function* $F : \mathbb{R} \to \mathbb{R}$

$$F_X(x) = \mathbb{P}\{\omega \in \Omega : X(\omega) \leq x\}.$$

Since $\mathbb{P}$ is a function on $\mathcal{F}$, this motivates the following definition for a *random variable*

Definition 1.4: Random Variable

A *random variable* is a function $X : \Omega \to \mathbb{R}$ such that

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$. Such a function is said to be $\mathcal{F}$-measurable.

Remark. One can show more rigorously that such a definition allows us to assign probabilities to consequences of events as desired. Namely, we are able to make an association between $\mathcal{F}$ and $\mathcal{B}$, the Borel σ-algebra on $\mathbb{R}$.

We will return to the focus on random variables in the next lecture, before then we will focus on the motivating object of the definition for the random variable, the *distribution function*:

Definition 1.5: Distribution Functions

The *distribution function* of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ given by $F_X(x) = \mathbb{P}(X \leq x) := \mathbb{P}\{\omega \in \Omega : X(\omega) \leq x\}$.

These are incredibly important associated functions to random variables. As such, we compile some characterizations and consequences of the definition.

Lemma 1.4: Equivalent Characterization of Distribution Function

A function $F : \mathbb{R} \to [0, 1]$ is a *distribution function* if and only if

1. $\lim_{x\to-\infty} F(x) = 0$, and $\lim_{x\to+\infty} F(x) = 1$;
2. If $x < y$, then $F(x) \leq F(y)$;
3. F is right-continuous, that is $\lim_{h\downarrow 0} F(x + h) = F(x)$

Remark. This captures the usual visual intuition of a distribution function.

The following makes explicitly clear the remark from earlier, the distribution function really connects the outputs of random variable X with the associated probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Lemma 1.5: Consequences of Definition

Let F be a distribution function of X. Then

a) $\mathbb{P}(X > x) = 1 - F(x)$;

b) $\mathbb{P}(x < X \leq y) = F(y) - F(x)$;

c) $\mathbb{P}(X = x) = F(x) - \lim_{y \uparrow x} F(y)$.

From these consequences, we recover how the probabilities *tail events* of our random variable - that is the extreme values - can be quantified through the distribution function. In modern applications, we are quite interested in recovering quantitative bounds on different types of events. We will recall some basic non-asymptotic bounds on concentration in a future lecture, as well as hint at what is used at the forefront of modern research. We now highlight how we can these distribution functions also allow us to recover probabilities for the outcomes of collections of random variables - through *random vectors*.

Consider how a random variable X has an associated distribution function F_X defined by $F_X(x) = \mathbb{P}(X \leq x)$ for all $x \in \mathbb{R}$. Analogously, suppose we have a random vector $(X_1, X_2, \dots, X_n) : \Omega \to \mathbb{R}^n$, where each component is in it's own right is a random variable. Then, $(X_1, \dots, X_n)$ is a *random vector* with the corresponding *joint* distribution function

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n),$$

where $x_1, \dots, x_n \in \mathbb{R}$ are variables.

Remark. There is a nice geometric interpretation of the above distribution function if the components are *independent* of one another. More on this later.

Using boldface $\mathbf{X}$ and $\mathbf{x}$ to refer to the vector-valued versions of random variables and variables, and $\mathbf{x} \leq \mathbf{y}$ to denote $x_i \leq y_i$ for all $i \in [n]$, we have the following:

Definition 1.6: Joint Distribution Function

The *joint distribution function* of a random vector $\mathbf{X} = (X_1, \dots X_n)$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is the function $F_{\mathbf{X}} : \mathbb{R}^n \to [0,1]$ given by $F_{\mathbf{X}}(x) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) = \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) \leq \mathbf{x}\})$ for all $\mathbf{x} \in \mathbb{R}^n$.

One can easily define analogous consequences of Lemma 1.4 (guess!), as the construction did not depend on the dimensionality of the random vector to be 1.

1.3.1 From Distribution Functions to Mass and Densities

We finish with recalling the connections with the distribution function with the familiar concepts heavily emphasized in an undergraduate probability course: *probability mass functions* (discrete probabilities and discrete random variables) and *probability densities* (continuous probabilities and continuous random variables). Before recalling these notions, we make clear

the distinction between discrete vs. continuous is (mostly) negligible - when considering the framework of measure theory.

While there are strong distinctions between discrete and continuous, when we are considering the *framework* of probability theory, measure theory gives us a way to ***compress*** the main ideas and tools being used. The distinction manifests in the computations carried out - and thus specific outcomes and inferences from said computations - but morally the logic behind the tools used are unified by the notation of framework of measure theory.

Remark. In undergraduate probability, one goes through the exact same constructions twice, once for discrete probabilities/random variables and again for the continuous analog. It should not be surprising that since the logic is precisely the same that the differences are somewhat negligible once sufficiently abstracted.

To make this point clear, suppose we wish to quantify the probability of a *specific outcome*, or in other words an infinitesimal consequence.

Definition 1.7: Discrete Random Variables and Probability Mass Functions

The random variable X is called *discrete* if it takes on the values in some at most ***countable*** subset of values of $\mathbb{R}$, denoted $\{x_i\}_{i=1}^{\infty}$. Then, the discrete random variable X has a a *probability mass function* $f : \mathbb{R} \to [0,1]$ given by

$$f_X(x) = \mathbb{P}(X = x).$$

Definition 1.8: Continuous Random Variables and Probability Density Functions

The random variable X is called *continuous* if its distribution function can be expressed as

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u) du, \quad x \in \mathbb{R},$$

for some $f : \mathbb{R} \to [0, \infty)$, an integrable function called the *probability density function of X*.

Notice, that this interpretation also holds for the discrete random variables, where the integral becomes a summation: namely in the discrete setting

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i : x_i \leq x} \mathbb{P}(X = x_i).$$

In fact, one may have random variables that are a mixture of discrete and continuous random variables. We note that these definitions are **referring to properties of distribution functions, rather than the random variables (the functions) themselves**.

The extensions to random vectors is also immediate, we merely state the analogs to the probability mass functions and density functions, respectively:

$$f_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n),$$

and in the continuous case $f \in L^1_{\text{ac}}(\mathbb{R}^n)$ such that

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

In more generality, one can unify all of these ideas with the following notation:

$$\mathbb{P}(\mathbf{X} \leq \mathbf{x}) = \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) \leq \mathbf{x}\}) = \int_{\omega : X(\omega) \leq \mathbf{x}} d\mathbb{P}(\omega) = \int_{\Omega} 1_{\omega : X(\omega) \leq \mathbf{x}} d\mathbb{P}(\omega).$$

where the notation $d\mathbb{P}(\omega)$ refers to the probability mass/density that is assigned to the infinitessimal outcome denoted by ω. The subscript on the integral denotes what event we are considering, and thus we can interpret the integral as *adding up the probabilities of all of the outcomes* that fall under the event defined by the random variable/vector - or in words the consequence in question. We will return to these more general integrals in the next lecture.

2 Lecture : Independence and Expectation

2.1 Conditional Probabilities and Independence

2.1.1 Probabilistic Events

First, let's recall the notion of a *conditional probability* for a first pass - to codify how the probability of an event depends on another:

Definition 2.1: Conditional Probability

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple. Let $A, B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. Then, the *conditional probability* that A occurs given that B occurs is defined as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

The follow follows directly from definition:

Lemma 2.1: Bayes' Rule

Let $A, B \in \mathcal{F}$ such that $\mathbb{P}(A), \mathbb{P}(B) > 0$. Then, we have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

In a similar fashion to the *disjointification idea* of before, using simple ideas from *set-theory* we can consider a partition of Ω. Let $\{B_i\}_{i=1}^n$ be a partition of Ω such that for all $i \neq j$,

$$B_i \cap B_j = \emptyset \quad \text{and} \quad \bigcup_{i=1}^{n} B_i = \Omega.$$

This allows us to *bin* every $\omega \in \Omega$ into exactly one set of this partition. Thus,

Lemma 2.2: Law of Total Probability

For any events A and B such that $0 < \mathbb{P}(B) < 1$,

$$\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c).$$

More generally, let $B_1, \ldots, B_n$ be some partition of Ω such that $\mathbb{P}(B_i) > 0$ for all i. Then

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Proof. (*Scheme*) Use a disjoint union and definition of conditional probability. □

Directly from this Theorem comes the following important identity:

Lemma 2.3: Bayes Formula

Suppose that $\{B_i\}_{i=1}^n \subseteq \mathcal{F}$ forms a partition of our sample space Ω. Then, we have that if A is a fixed event such that $\mathbb{P}(A) > 0$, then for any $j \in [n]$:

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \cap B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$

Definition 2.2: Independent Events

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple. Then $A, B \in \mathcal{F}$ are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

More generally, a family $\{A_i : i \in I\}$ is called *independent* if

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$

for all finite subsets J of I.

Remark. This does not imply that $A \cap B = \emptyset$.

Morally, this captures the notion of the outcome of an event as not being dependent on the other event. Moreover, we may even weaken the notion of independent events to *pairwise independence*

Definition 2.3: Pairwise Independence

Let $\{A_i\}_{i \in I} \subseteq \mathcal{F}$ such that for all $i \neq j$ we have

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j).$$

Then, we say that this family is *pairwise independent.*

These notions become incredible useful when trying to understand complicated probabilistic events. Recall, last lecture we highlighted the general structure of probability spaces can be understood through the lens of *set theory* and *measure theory.* Some arguments required decomposition arguments, or in words breaking down an event into manageable and reasonable pieces. Conditional probabilities also play a role on understanding relationships between events, and independence highlights a scenario where it is very easy to manipulate the randomness. This will pop up when talking about some fundamental properties of expectation and variance, more on this later.

2.1.2 Extending to Random Variables

Recall that we were able to explicitly connect the outputs of our random variables with their sources of randomness through the *distribution function.* In light of the previous discussion, the following definition is clear

Definition 2.4: Independent Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple. Let $X, Y : \Omega \to \mathbb{R}$ be random variables with distribution function $F_X : \mathbb{R} \to [0,1]$ and $F_Y : \mathbb{R} \to [0,1]$. Consider the joint distribution function $F_{X,Y}$ on $(X,Y) : \mathbb{R} \times \mathbb{R} \to [0,1]$. We say X, Y are *independent* if and only if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \forall x, y \in \mathbb{R}.$$

Remark. We can read this as $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, where we define the events A and B accordingly.

Recall, that it suffices to consider the closed rays in $\mathbb{R}$ as this generates the Borel σ-algebra, and the distribution function is a *nice* function that allows us to transfer our understanding from $\mathcal{B}$ to $\mathcal{F}$. In words, we recover *independence* in the sense of the last discussion for our random variables - and we use the fact the distribution function carries the information of the randomness of our random variables for this to work.

For completeness, we list the other familiar notions of independence for random variables in the discrete and continuous case, stating that these are special cases of the above definition (why?):

- **Discrete Random Variables**: X and Y are independent if and only if events
$$\{X = x\} \quad \text{and} \quad \{Y = y\}$$
are independent for all x, y;
- **Continuous Random Variables**: X and Y are independent if and only if events
$$\{X \leq x\} \quad \text{and} \quad \{Y = y\}$$
are independent for all x, y.

We also note that one can recover the *factorization* of not just the joint distribution functions, but also the joint probability mass functions and joint probability densities.

Remark. This factorization hints at an underlying geometric interpretation, where independence is precisely the setting of being able to take the *tensor product* of some spaces. In general, the *product measure* marginalizes well, and it is precisely this product measure that reflects the notion of independence.

The nice outcomes of independence, and thus factorization, can truly be felt when considering outcomes in expectation. Beforehand, we make note of a notion that will be useful when discussing limiting phenomenon - such as the *Central Limit Theorem*

Definition 2.5: Independently Identically Distributed (IID) Random Variables

Let $\{X_i\}_{i \in I}$ be a collection of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, we say that this collection is *independently identically distributed* if they are *independent* of a collection of random variables, and have the *same distribution function.*

This definition is quite natural, if we assume we are making independent samples of some quantity governed by the same underlying distribution.

Remark. In modern regimes, we wish to relax this situation - for example in *online learning* - but we can still stay close to this case in some suitable sense of *close.*

2.2 Moments, Expectation and Variance

One often wishes to have a *single scalar* descriptions of the distribution of values taken by some random variable $X : \Omega \to \mathbb{R}$. To produce such scalar descriptions to describe notions such as the *center of mass*, *spread of mass* - given that we conceptualize the outputs of the random variable as mass - we use the *moments of a random variable*

Definition 2.6: Moments of a Random Variable

Let $X : \Omega \to \mathbb{R}$ be a random variable, and $n \in \mathbb{N}$. The *p-th moment* of X is defined as
$$\mathbb{E}[X^p] := \int_\Omega (X(\omega))^p d\mathbb{P}(\omega).$$
As hinted by the notation, the *expectation* of X is the 1st moment of X. Explicitly:
$$\mathbb{E}[X] := \int_\Omega X(\omega) d\mathbb{P}(\omega).$$

A few remarks. First, we use the notation introduced at the end of the previous section alluding to the use of the *Lebesgue integral.* Morally, one can still have the same intuitions of integration from Calculus, but this allows one to integrate more nasty functions - one can learn a more rigorous representation in the Probability Theory sequence and/or a Graduate Analysis course. Second, it is intuitive that the moments we have, the more specified our distribution is. This idea plays a role in estimating distributions from empirical samples. Moreover, it is important to state that in modern applied probability, the notion of *Gaussian Universality* essentially says that having the first two moments may suffice in practice. These ideas are still being fleshed out in the community, with many open problems!

Notice that the expressions for the expectation of discrete and continuous random variables are just special cases, respectively:

$$\mathbb{E}[X] = \sum_x x\mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}) \quad \text{and} \quad \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx,$$

where we explicitly use the *probability mass function* and *probability density function.*

2.2.1 Expectation Identities and Basic Properties

Before going through the standard properties, we mention some nice identities that can be used often:

- **Expectation of an Indicator.** Let $A \in \mathcal{F}$ be arbitrary. Then, if we define the *indcator function* I_A as

  $$I_A(\omega) := \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A, \end{cases}$$

  then we have $\mathbb{E}[I_A] = \mathbb{P}(A)$. Explicitly, notice

  $$\mathbb{E}[I_A] = \int_\Omega I_A(\omega)d\mathbb{P}(\omega) = \int_A d\mathbb{P}(\Omega) = \mathbb{P}(A).$$

  In other words, we are able to convert the probability of an event into an expectation - giving a dual perspective on how to manipulate these objects;

  Remark. In fact, one can arbitrarily approximate any random variable through the use of linear combinations of indicator functions in a dense fashion - take Probability Theory to see such an idea.

- **Tail Identity.** Let X be a *non-negative* random variable. Then,

  $$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)dx.$$

  To prove this, one uses *Fubini-Tonelli's Theorem*

  $$\int_0^\infty \mathbb{P}(X > x)dx = \int_0^\infty \int_\Omega I_{\{X>x\}}(\omega)d\mathbb{P}(\omega) = \int_\Omega \int_0^\infty I_{\{X>x\}}dxd\mathbb{P}(\omega) = \int_\Omega X(\omega)d\mathbb{P}(\omega) = \mathbb{E}[X].$$

  The only part that needs to be justified is switching the order of integration. Note, in the simple continuous or discrete case this reduces to the usual computations.

  We note that this identity generalizes even further, without proof. If X is any random variable, and Φ is any increasing and differentiable function Φ, then we have

  $$\mathbb{E}[\Phi(|X|)] = \Phi(0) + \int_0^\infty \Phi'(t)\mathbb{P}[|X| > t]dt.$$

  This is useful for proving properties for *subgaussian random variables*, which are an important class of random variables in Statistics and Machine Learning.

Remark. The proof follows essentially from an application of Fundamental Theorem of Calculus.

We now recall familiar properties that follow directly from nice properties of the integral.

Lemma 2.4: Basic Properties of Expectation

a) **Linearity:** Let X, Y be random variables and $\alpha, \beta \in \mathbb{R}$, then we have

$$\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y];$$

b) **Deterministic Random Variables:** Suppose that $\mathbb{P}(X = b) = 1$ for some $b \in \mathbb{R}$, then one has

$$\mathbb{E}[X] = b.$$

c) **Bounded Random Variables:** Suppose that $\mathbb{P}(a \leq X \leq b) = 1$ for some $a, b \in \mathbb{R}$, then one has

$$a \leq \mathbb{E}[X] \leq b.$$

Proof. (*Scheme*) Follows from the definition of expectation. □

In addition, suppose we have some measurable function $g : \mathbb{R} \to \mathbb{R}$ - measurable in our context means some reasonable function - then we can easily compute the expectation of $g(X)$ when X is a random variable:

Lemma 2.5

Suppose that $g : \mathbb{R} \to \mathbb{R}$ is measurable, and X is a random variable. Then, we have the following realizations of the expectation: First, if X is discrete then

$$\mathbb{E}[g(X)] = \sum_x g(x)\mathbb{P}(X = x).$$

If continuous with probability density function f_X, then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

More generally, we can write this as

$$\mathbb{E}[g(X)] = \int_\Omega g(X(\omega)) d\mathbb{P}(\omega) = \int_{-\infty}^{\infty} g(x) dF_X(x),$$

where in the last expression we use the distribution function $F_X(x)$ of random variable X.

3 Lecture: Towards Multiple Random Variables

In the past two lectures, we build up the language for describing the randomness of a single random variable. Namely, we recalled the notion of the *probability triple*, and how the *distribution function* allows us to endow the outputs of a *random variable* with probabilities.

After initially finishing the discussion of variance, we will start considering how to understand the relationships of multiple random variables beyond independence, and then recall different *modes of convergence* that will allow one to make ***limiting statements***.

3.1 Variance, Covariance, and Correlation

We previously covered the *moments* of a random variable and highlighted why the *expectation* - the first moment - is a scalar of interest. Namely, this codifies the notion of *center of mass*. We now consider the *spread of mass*.

Definition 3.1: Variance

Let $X : \Omega \to \mathbb{R}$ be a random variable. Then, the *variance* of X is defined as

$$\mathrm{Var}[X] := \int_\Omega (X(\omega) - \mathbb{E}(X))^2 d\mathbb{P}(\omega).$$

This recovers the usual notions of variance for discrete and continuous random variables in an analogous fashion to before (check!).

At a coarse glance, one would guess that there is a connections with the second moment of X. One could then easily verify the following:

Lemma 3.1: Variance and Second Moments

Let X be a random variable with $\mathbb{E}[X], \mathbb{E}[X^2] < \infty$. Then,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Proof. This follows from the nice properties of expectation. Explicitly:

$$\mathrm{Var}(X) = \int_\Omega (X(\omega) - \mathbb{E}[X])^2 d\mathbb{P}(\omega) = \int_\Omega X(\omega)^2 d\mathbb{P}(\Omega) - 2\mathbb{E}[X]\int_\Omega X(\omega) d\mathbb{P}(\omega) + (\mathbb{E}[X])^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

where we expanded out the square and used Lemma 2.2 a and b in the second equality. □

Using the above proof as an example, understanding how to manipulate and control such quantities boils down to understanding how to work with these integrals.

With the *expectation* and *variance* we are able to get a coarse picture of the behavior of our random variable X. We now consider comparing two random variables **on the same probability triple** through the *covariance*:

Definition 3.2: Covariance of Two Random Variables

Let X and Y be two random variables. Then, we define the *covariance* between these random variables as

$$\mathrm{Cov}(X, Y) := \mathbb{E}\Big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\Big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Notice that the second equality follows from the exact same computation as in the proof of Lemma 3.1. To make it explicitly clear that these are on the same probability triple, we can more explicitly write:

$$\mathrm{Cov}(X, Y) = \int_{\Omega} (X(\omega) - \mathbb{E}[X])(Y - \mathbb{E}[Y]) d\mathbb{P}(\omega).$$

This definition *quantifies* how much two random variables are mutually affected by the randomness that generates both random variables. Moreover, we can also interpret the covariance as a *quantitative measure* of how **un-independent** X and Y are.

If X and Y were independent, then the covariance would be 0, as $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. In some contexts, we wish to consider the correlation, which is a *unitless* measure.

Definition 3.3: Correlation

Let X and Y be two random variables, such that $\mathrm{Var}(X), \mathrm{Var}(Y) > 0$. Then, we define the *correlation* of these two random variables as

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

We often like these unitless or dimensionless measures as this allows us to make *fair comparisons*. We note that for any X, Y random variables, we have that $|\rho(X, Y)| \leq 1$. We will come back to this point very soon.

3.1.1 Endowing Space of Random Variables with Hilbert Space Structure

Now the covariance, enjoys some nice properties:

Lemma 3.2: Elementary Properties of Covariance

Let X and Y be random variables. Then, we have that

a) **Symmetry:**

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X);$$

b) **Scaling:** Let $a \in \mathbb{R}$ be arbitrary. Then

$$\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y);$$

c) **Recovering Variance:**

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X);$$

Moreover, if we have some finite collections $\{X_i\}_{i=1}^n, \{Y_j\}_{j=1}^m$ of random variables, we also have

d) **Addition Preserved:**

$$\mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}(X_i, Y_j).$$

Hilbert Space Structure. In fact, we can see that these properties are enough to imbue the space of random variables with second finite moment with nice structure. Namely, we can recover a *Hilbert space structure* where the random variables $X, Y \in L^2(\mathbb{R})$. More explicitly:

since covariance is a bilinear, symmetric, and positive semi-definite real-valued function - it is an Inner Product on the space of finite-second moment random variables.

It should not be lost on us how convenient such a structure is, as it allows us to use tools from Functional Analysis and Linear Algebra to study random variables. In fact, this correspondence goes extremely deep - for example the theory of Markov Chains and more generally Markov Processes rely on ideas from these fields of mathematics to quantify the rates of *ergodicity*. Moreover, this Hilbert space structure will allow us to use ideas such as *projection* - this will be the underlying idea that allows us to study the theory of *linear regression* as regression functions can be thought of as projections into a function class given sampled data.

In short, one exploits this structure in many settings, for our purposes we will only use this structure in a simple fashion. Namely, $|\rho(X,Y)| \leq 1$ will follow immediately from *Cauchy-Schwartz* which we recall in *generality*:

Theorem 3.1: Cauchy-Schwartz Inequality

Let $\mathcal{H}$ be a Hilbert-space with real-valued inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. Then, *Cauchy-Schwartz* is

$$|\langle x, y\rangle| \leq \langle x, x\rangle^{\frac{1}{2}} \langle y, y\rangle^{\frac{1}{2}}, \quad \forall x, y \in \mathcal{H}.$$

In our setting, this becomes

$$|\text{Cov}(X,Y)| \leq \sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}.$$

There are many proofs of this fundamental result (my favorite is using the *discriminant* of real polynomials), but this should be covered in the Linear Algebra review, so we take it as fact.

Moreover, because we have this interpretation of covariance in connection with variances, we have the follow restatement of the *parallelogram law*.

Corollary 3.1: Parallelogram Law for Random Variables

Let $\{X_i\}_{i=1}^n$ be a collection of random variables, then

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, Y_j).$$

Notice that as a consequence, if the collection of random variables are *pairwise independent*, this implies the familiar rule:

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i).$$

Before continuing on to the next topic, we also mention we can capture the pairwise covariances for a finite group of random variables:

Definition 3.4: Covariance Matrix

Let $\{X_i\}_{i=1}^n$ be random variables. Then, the *covariance matrix* Σ is defined such that the $i, j \in [n]$ entries

$$\Sigma_{ij} = \text{Cov}(X_i, X_j).$$

Remark. This is a positive semi-definite matrix, as it is a Gram Matrix.

The covariance matrix is of great importance for statistical applications, and the study of the eigenvalues as well is very relevant. To get a more fine-grained understanding of a covariance matrix, and how we can get a good empirical estimate of these from data, one should take *Matrix Analysis*.

3.2 Marginalization, and Conditional Expectations

For simplicity sake, henceforth we will assume that our **random variables** are **continuous random variables**. Moreover, we will focus on when we have a pair of random variables X, Y - many interesting probabilistic statements concern pairs, and the ideas extend to any finite collection of random variables through induction.

Recall, for continuous random variables, we have the following *joint distribution function* described by integrating the *joint density function*

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v) du dv.$$

As previously discussed, for any sufficiently nice subset $B \in \mathbb{R}^2$ (namely a $B \in \mathcal{B}$), we can recover the probability of achieving such values by direct computation

$$\mathbb{P}((X, Y) \in B) = \int \int_{B} f(x, y) dx dy,$$

and this is derived from the *joint distribution function.*

A natural question is given *joint* information, can one recover the information of the distribution of one of the component random variables. This is done through *marginalization*

Definition 3.5: Marginal Distributions

Let X and Y be random variables. Then, the *marginal distribution functions* of X and Y are respectively denoted by

$$F_X(x) = \mathbb{P}(X \leq x) = F(x, \infty), \quad F_Y(y) = \mathbb{P}(Y \leq y) = F(\infty, y),$$

where we write $F(x, \infty) = \lim_{y \to \infty} F(x, y)$. Explicitly, this gives us

$$F_X(x) = \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} f(u, y) dy \right) dy,$$

which then gives us the *marginal density function* of X as

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Analogously for the random variable Y.

We may also ask ourselves of the distribution of one random variable given another random variable. This is especially of interest when trying to learn say a function $g : X \to Y$ given only access to the sample $\{X_i, Y_i\}_{i=1}^{n}$. First, we recall notions previously discussed in the setting of continuous random variables.

Definition 3.6: Conditional Distribution and Density Function

The *conditional distribution function* of Y given $X = x$ is the function $F_{Y|X}(\cdot \mid x)$ given by

$$F_{Y|X}(y \mid x) = \int_{-\infty}^{y} \frac{f(x,v)}{f_X(x)} dv$$

for any x such that $f_X(x) > 0$. Sometimes this is denoted $\mathbb{P}(Y \leq y \mid X = x)$.

The *conditional density function* of $F_{Y|X}$ is thus defined by

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)} dv$$

for any x such that $f_X(x) > 0$.

Note, this can be easily remembered by the usual notion of conditional probability: $f_{Y|X} = \frac{f_{X,Y}}{f_X}$.

3.2.1 Conditional Expectation and Variance

From the previous definitions, the following notion is immediate:

Definition 3.7: Conditional Expectation

The *conditional expectation* $\mathbb{E}[Y \mid X]$ is defined by

$$\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy.$$

Since this is merely an expectation with respect to a probability distribution, all of the nice properties for expectations also hold within this setting - for examples linearity of (conditional) expectation.

We can also define the notion of conditional variance

Definition 3.8: Conditional Variance

The *conditional variance* of Y, given that $X = x$, we have that

$$\mathrm{Var}(Y \mid X = x) = \mathbb{E}\Big[(Y - \mathbb{E}[Y \mid X = x])^2 \mid X = x\Big].$$

Again, analogously to the unconditional case, we can decompose the variance into simpler terms:

Lemma 3.3: Conditional Variance Formula

Let $\mathrm{Var}(Y \mid X)$ be a function of X defined by the conditional variance. Then, we have the *conditional variance formula*:

$$\mathrm{Var(Y)} = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X]),$$

where the expectation and variance are taken with respect to the distribution of X.

We don't give proofs, but these notions are covered in most Probability textbooks.

Lastly, with the view of decomposing and reconstructing complex objects from simple ones, we recall the following nice law

Theorem 3.2: Tower Property - Law of Total Probability

Let $\mathbb{E}[Y \mid X]$ denote the function of Y where the value at $Y = y$ is given at $\mathbb{E}[Y \mid X = x]$. Then,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]].$$

Explicitly, for the special cases of discrete and continuous random variables, we respectively have

$$\mathbb{E}[Y] = \sum_x \mathbb{E}[Y \mid X = x]\mathbb{P}\{X = x\},$$

and

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \mathbb{E}[Y \mid X = x] f_X(x) dx.$$

Remark. This is a very general property, and holds for all pairs of random variables, regardless of type.

Proof. We give the proof for the continuous case is analogous: Let $\psi(X) := \mathbb{E}[Y \mid X]$. Then, we have

$$\begin{aligned}\mathbb{E}(\psi(X)) = \int_{\mathbb{R}} \psi(x) f_X(x) dx = \int_{\mathbb{R}}\int_{\mathbb{R}} y f_{Y|X}(y \mid x) dy f_X(x) dx = \int_{\mathbb{R}}\int_{\mathbb{R}} y f_{X,Y}(x,y) dy dx \\ = \int_{\mathbb{R}} y \int_{\mathbb{R}} f_{X,Y}(x,y) dx dy = \int_{\mathbb{R}} y f_Y(y) dy = \mathbb{E}[Y].\end{aligned}$$

The third equality follows from definition of conditional probability and one can rigorously justify the the integration switch through Fubini's. □

We finish with an example of where the tower property can be used to analyze random variables that is neither discrete nor continuous.

Example of Generality of Tower Property

Let $\Omega = \{T\} \cup \{(H, x) : 0 \leq x < 2\pi\}$. This codifies the event of a coin toss with probability of heads p (this is a Bernoulli random variable with parameter p). If a head is tossed, then we fling a rod on the ground and measure it's angle (this is a uniform probability on $[0, 2\pi)$). This probability space will allow us to mix discrete and continuous notions.

Let $X : \Omega \to \mathbb{R}$ be given by

$$X(T) = -1, \quad X((H, x)) = x.$$

This random variable takes values in $\{-1\} \cup [0, 2\pi)$, this is a continuous random variable except for the point mass at -1. ***Compute*** $\mathbb{E}[X]$**.**

To do so, let A be the event that a tail turns up. Then, use the tower property:

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X \mid I_A]] \\ &= \mathbb{E}[(X \mid I_A = 1]\mathbb{P}(I_A = 1) + \mathbb{E}[X \mid I_A = 0]\mathbb{P}(I_A = 0) \\ &= \mathbb{E}[X \mid \text{tail}]\mathbb{P}(\text{tail}) + \mathbb{E}[X \mid \text{head}]\mathbb{P}(\text{head}) \\ &= -1 \cdot q + \pi \cdot p = \pi p,\end{aligned}$$

where at the end we used the properties of the Bernoulli and Uniform distribution.

4 Lecture: Limit Theorems Part 1 - Estimators for Moments

We slightly shift gears to discussing how one can get a handle on expressions of the flavor of *in the long run* and *on the average*. These statements reflect the faith of how repeated *iid* experiments show less and less random fluctuations as they settle down to some limit in some sense.

To make sense of the *precise* statements we can say about these kinds of expressions, we need to first recall some of the *different notions of convergence*. Then, one can start trying to get *rates of convergence* through some non-asymptotic statements, which will allow us to review Markov's and Chebyshev's Inequalities (both are secretly Markov's). Lastly, we will recall a few of the tools that are used to prove the crowning jewel of an undergraduate Probability Theory course - that is the tools to prove *Central Limit Theorem*.

4.1 Modes of Convergence

We recall the three *modes of convergence* of random variables taught in undergraduate courses, and remark on how there is a chain of implications that in general only go in one direction. Before doing so, we make it clear

this is just the tip of the iceberg, there are many notions of convergence of interest to probabilists and applied probabilists. When one makes statements about convergence, make sure to understand in what sense!

We begin with the strongest type of convergence, and progressively weaken the assumption on the mode of convergence. In words, the implications will go in a downwards direction.

Definition 4.1: Almost Sure Convergence

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. We say that X_n *converges almost surely* to X, $X_n \xrightarrow{a.s.} X$, if there exists a set $E \in \mathcal{F}$ such that $\mathbb{P}(E) = 1$ and for all $\omega \in E$, we have that

$$X_n(\omega) \to X(\omega).$$

Explicitly, this statement says that for every $\epsilon > 0$, there exists an $N := N(\omega) \in \mathbb{N}$ such that for all $n > N(\omega)$ we have that

$$|X_n(\omega) - X(\omega)| < \epsilon$$

We can also read off this statement as saying

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

Now, we can loosen the requirement of stating convergence in realizations - that is where we fix the ω a priori - to convergence in the overall probability of the random variables.

Definition 4.2: Convergence in Probability

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. Then, we say that X_n *converges in probability to* X, $X_n \xrightarrow{\mathbb{P}} X$, if for all $\epsilon > 0$ we have that

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

The key difference here is that we are not fixing the set of samples where the limits agree, that is we can consider a shifting sequence of events, as long as the limit holds.

Lastly, we can further weaken the notion of convergence from the probability of *deviation* being zero, to the statement being *morally* of the form "*eventually, the values of X_n will have a* ***distribution that resembles the distribution*** *of X.*"

Definition 4.3: Convergence in Distribution

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. Then, we say that X_n *converges in distribution to* X, $X_n \Rightarrow X$, if the *distribution functions converge.* That is, for all $x \in \mathbb{R}$ we have

$$\lim_{n\to\infty} F_{X_n}(x) = \lim F_X(x).$$

Explicitly, this says $\lim_n \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$ for all x.

4.1.1 Examples Showing Modes of Convergence are Distinct

Beyond the intuitive reasons why these are different notions of convergence, we give an explicit examples of how these notions are different from one another.

In order to highlight some examples of how these definitions are distinct from one another, we need to recall a tool that is used to make statements about events that occur *infinitely often.*

Lemma 4.1: Borel-Cantelli Lemma

Let $\{A_n\}_{n=1}^{\infty}$ be a collection of events. The event that A_n happens infinitely often is given by

$$A_n \text{ i.o.} = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k\geq n} A_k.$$

Then, we have that

a) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$;

b) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and $\{A_n\}_{n=1}^{\infty}$ are independent, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

To understand the notion of A_n i.o., it suffices to consider what happens when $\omega \notin \bigcap_{n\geq 1} \bigcup_{k\geq n} A_k$. This simply means that ω is only in finitely many A_n, and thus not infinitely often.

Proof. (*Scheme*) Statement a follows from rewriting $\mathbb{P}(A_n \text{ i.o.}) = 0$ as the equivalent $\mathbb{P}(N < \infty) = 1$, where

$$N = \sum_{n=1}^{\infty} I_{A_n}$$

counts the number of occurrences of the events. Then, translating the assumption into saying $\mathbb{E}[N] < \infty$, using Fubini-Tonelli.

Statement b follows by looking at

$$\mathbb{P}(A_n \text{ i.o.}) = \lim_{k\to\infty} \mathbb{P}\left(\bigcup_{n=k}^{\infty} A_n\right) = 1 - \mathbb{P}\left(\bigcap_{n=k}^{\infty} A_n^c\right),$$

then using the independence assumption and the numerical inequality $1 - x \leq e^{-x}$ to get $\mathbb{P}(\cap_{n=k}^{\infty} A_n^c) = 0$. □

This Lemma is useful because it gives us a common way of proving almost sure convergence. Namely, if we define

$$A_n := \{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \epsilon\},$$

then $\mathbb{P}(A_n \text{ i.o.}) = 0$ corresponds with the event that $X_n \xrightarrow{a.s.} X$ does **not** happen. With this in mind, we can highlight the sought-after examples of why these definitions are strictly different.

Remark. This is also a common tool to prove different $0-1$ laws and to make statements about tail probabilities.

Example Showing Differences in Modes of Convergence

Example 1. ***Convergence in Probability is Weaker Than Almost Surely.***

Let $X_n \sim \text{Bernoulli}(\frac{1}{n})$, $n \geq 1$, be independent random variables, and let $X \sim 0$. Then $X_n \xrightarrow{\text{i.p}} X$ but not almost surely.

We have in probability convergence as $\mathbb{P}(|X_n - X| \geq \epsilon) = \frac{1}{n}$ tends to 0 as $n \to \infty$. Now, to see that X_n does not converge almost surely, we apply Borel-Cantelli Lemma 1 with the A_n from before in mind.

Example 2. ***Convergence in Distribution is Weaker Than In Probability.***

Consider the uniform measure on the unit interval $\Omega = [0,1]$. Now, let $X_n \sim \text{Uniform}([0,1])$, specifically allowing $X_{2k}(\omega) = \omega$ and $X_{2k+1}(\omega) = 1 - \omega$, for all $k \in \mathbb{N}$. Then, allowing $X(\omega) = \omega$, we see that $X_n \Rightarrow X$, but the probability of deviation $\mathbb{P}(|X_n - X| \geq \epsilon)$ continues to oscillate between zero and a nonzero value.

4.2 Markov's Inequality and Law of Large Numbers

We now remind ourselves of the simple but powerful inequality, that allows us to make our first limiting statement of interest. Moreover, this will provide us with the first non-asymptotic inequalities one often sees in Probability - *Markov's* and *Chebyshev's* inequalities.

Theorem 4.1: Markov's Inequality

If X is a nonnegative random variable, then for any $a > 0$ we have:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Let $X \geq 0$. For simplicity, assume that X is a continuous random variable (analogous proof holds in general case). Then,

$$\mathbb{E}[X] = \int_\Omega X(\omega) d\mathbb{P}(\omega) = \int_0^\infty x f_X(x) dx \geq \int_a^\infty x f_X(x) dx = a \int_a^\infty f_X(x) dx = a\mathbb{P}(X \geq a).$$

□

From this, we can prove the *generalized Chebyshev inequality*, which reduces to the commonly known Chebyshev's as a special case.

Theorem 4.2: Generalized Chebyshev

Let X be a random variable with finite pth moment, given that $p \geq 2$. Then, it follows for all $a > 0$ that we have

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^p]}{a^p}.$$

Notice, when $p = 2$, we reduces to the familiar inequality:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{a^2} = \frac{\text{Var}(X)}{a^2}$$

Proof. We just apply Markov's inequality:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) = \mathbb{P}(|X - \mathbb{E}[X]|^p \geq a^p) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^p]}{a^p}.$$

□

These inequalities allow us to have a *quantifiable* control over randomness with only access to the mean and variance of our distribution of the random variable. Moreover, notice that if we have more moments, we can have tighter control on our deviation from mean. These are statements about the *tails* of our distribution, namely that they must exhibit specific rates of decay. This makes sense when thinking about when a function is integrable. Morally, one should somewhat equate *tail decay* with *concentration about the mean*, this idea is explored more when considering non-asymptotic statistics, and high-dimensional phenomena.

Remark. If one has exponential tail decay, such as when considering *Gaussian* and *Sub-Gaussian* random variables, one gets really nice guarantees.

As a corollary of Chebyshev's inequality, we obtain the *Weak Law of Large Numbers*:

Corollary 4.1: Weak Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of *iid* random variables, each having finite mean and variance, denoted $\mathbb{E}[X_i] = \mu$ and $\operatorname{Var}(X_i) = \sigma^2$ for all i, respectively. Then, for any $\epsilon > 0$:

$$\mathbb{P}\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0, \quad \text{as } n \to +\infty.$$

For notation sake, define the *sample mean* as

$$\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, this statement says that $\hat{\mu}_n \xrightarrow{\mathbb{P}} \mu$ as $n \to \infty$. In words, our *sample mean converges to the population mean in some sense* - that being in probability. The proof follows from Chebyshev's inequality, and thus also gives us a *non-asymptotic* handle on how this quantity concentrates around the mean.

Proof. Since X_i are independent, we have by the linearity and scaling properties of *expectation* and *variance* the following:

$$\mathbb{E}[\hat{\mu}_n] = \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu \quad \text{and} \quad \operatorname{Var}(\hat{\mu}_n) = \operatorname{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}.$$

Therefore, by Chebyshev's inequality, we have that

$$\mathbb{P}\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\} = \mathbb{P}\{|\hat{\mu}_n - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

□

Remark. This is a ***weak bound***, namely that of a $O(\frac{1}{n})$ term. However, the reason for the bound being so weak is that this statement is made under ***very general assumptions***. We only assume our sample process is *iid* on a distribution with finite mean and variance. More assumptions result in much stronger bounds.

Beyond rates of convergence one can obtain a stronger version of convergence of our sample mean to the true mean, namely almost surely, with an additional technical assumption:

Lemma 4.2: Strong Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of *iid* random variables. Consider the sample mean $\hat{\mu}_n$ as defined above. Suppose that for all i, X_i has finite mean and variance, μ and σ^2, as well as a assume ***bounded fourth moment***, $\mathbb{E}[X_i^4] < \infty$. Then, we have that $\hat{\mu}_n \xrightarrow{\text{a.s.}} \mu$ as $n \to \infty$.

These Law of Large Numbers answer a central question in *statistical inference*:

Does the sample mean concentrate around the true mean? Can I recover this true mean from my sample?

Moreover, the proofs give a non-asymptotic rate of convergence. This gives another avenue of application, where one can use *injected randomness* in order to do - otherwise numerically beastly - computations.

4.2.1 Application: Monte Carlo Integration

We highlight a simple application of these limiting theorems. Consider wanting to evaluate the integral of a function $f : \mathbb{R}^d \to \mathbb{R}$ on some set S. Now, suppose we have some probabilistic process that allows us to sample from S, that is the *evaluation of the integral* takes on the following form:

$$\int_S f(x) d\mathbb{P}(x),$$

where we are considering some probability measure $\mathbb{P}$ such that $\text{supp}(\mathbb{P}) = S$.

Now, instead of going about integration through the use of meshes and numerical approximations in the scheme of what would be done in a Numerical Analysis course - consider evaluating the integral probabilistically. Why?

When d is large, then most deterministic integration methods have an error that depends exponentially on dimension.

We will show that integrating with *injected* randomness will allow us to circumvent this exponential dependence on dimension.

Consider a random point X that takes values in S according to law $\mathbb{P}$. Then, we may interpret the integral of f as the expectation:

$$\int_{\S} f d\mathbb{P} = \mathbb{E}_{X\sim\mathbb{P}} f(X).$$

Now, if we were to take *iid* samples from S according to μ, then by the Law of Large Numbers we have that

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathbb{E}f(X)$$

almost surely as $n \to \infty$. Thus, intuitively, we can use the following Monte Carlo approximation for the integral:

$$\int_S f d\mathbb{P} \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Now, if we have a sufficiently nice function and domain, for instance, say $f : [0,1] \to \mathbb{R}$ is continuous, then the error rate can be shown to be $O(\frac{1}{\sqrt{n}})$ from an argument regarding the variance of the sample mean. This is merely the decay rate of the *standard deviation* as $n \to \infty$. notice that this rate has ***no dependence*** on the *dimension*, which is the main appeal of these approximations - *of course one has to consider if the sampling truly reveals what we want to learn.*

4.3 Concentration Phenomena - Basic Concentration Inequalities

When talking about the *generalized Chebyshev's inequality*, we noticed that with the pth moment, we recover a tail decay rate on the order of $O(\frac{1}{a^p})$. Explicitly:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^p]}{a^p} = O\left(\frac{1}{a^p}\right).$$

These are polynomial rates of decay, but if all moments are finite we should expect an exponential tail decay - something faster than polynomial for all degrees. This leads to modern *non-asymptotic bounds* of the flavor of *concentration inequalities.*

We state a few that one will see some that may be familiar to some of you:

Theorem 4.3: Hoeffding's Inequality

Let $X_1, \dots, X_n$ are *iid* random variables, and consider the *sample mean* $\hat{\mu}_n$. Now, suppose that for all $i \in [n]$ we have that $X_i \in [a, b]$ almost surely. Then, for all $\epsilon > 0$ we have that

$$\mathbb{P}(|\hat{\mu}_n - \mu| \geq \epsilon) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

Holding ϵ and the size of the interval *constant* in our minds, notice that this bound has an *exponential decay* with respect to the sample size.

However, in this setting we are replace the variance $\sigma^2 = n\mathrm{Var}(\hat{\mu}_n)$ with a possibly much larger quantity $\frac{(b-a)^2}{4}$. In words, there are two regimes of interest - this can be quantified and accounted for at a small cost in the following:

Theorem 4.4: Bernstein's Inequality

Let $X_1, \dots, X_n$ be independent random variables. Now, assume that $|X_i - \mathbb{E}X_i| \leq B$ for every $i \in [n]$. Then, *Bernstein's Inequality* states that for all $\epsilon > 0$

$$\mathbb{P}(|\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n]| \geq \epsilon) \leq 2\exp\left(-\frac{n\epsilon^2/2}{\sigma^2 + Bt/3}\right).$$

Both of these inequalities are proven through a *Chernoff bound technique*, roughly using a Laplace transform and a Fenchel conjugate. More roughly, one can prove these using the *Moment Generating Functions* we will introduce next lecture.

This is ***much beyond*** the scope of this review, however we make note that these ideas extend much beyond the study of bounded random variables. More generally, we are able to get a handle of properly scaled sums of *sub Gaussian* and *sub Exponential* random variables. Morally, these are random variables who's tails decay at least at the rate of a Gaussian or an Exponential, that is of the rate $O(e^{-x^2})$ or $O(e^{-x})$ respectively. These are covered in any course in Machine Learning or High Dimensional Probability.

4.4 Convolutions and Sums of Random Variables

In anticipation of next lecture, we recall the distribution of the sum of two *independent* random variables is given by the following operation:

Definition 4.4: Convolution

Let X and Y be independent continuous random variables. Let $f_X(x)$ and $f_Y(y)$ be the respective densities, and $F_X(x)$ and $F_Y(y)$ be the respective distribution functions. Then, *convolution* recovers the densities and distribution function of $X+Y$. Explicitly, the operation of *convolution of distribution functions* is defined as

$$F_{X+Y}(a) = (F_X * F_Y)(a) := \int_{-\infty}^{\infty} F_X(a-y)dF(y)$$

The operation of *convolution of density functions* is defined as

$$f_{X+Y}(a) = (f_X * f_Y)(a) := \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy.$$

Remark. We focus on continuous random variables for ease of notation, but the same idea holds for discrete random variables - and more general random variables as well.

Generally, the operation of convolution can be thought of as a *weighted average* of one function against the other. In a sense, we are blending and smoothing out a function. In this setting, if we look at the convolution of the density functions, we are essentially convolving with respect to the density of the Y function. In signal processing, this is a natural operation, as convolutions become multiplications when moving the Fourier domain. This is the central insight that will motivate the use of *characteristic functions* in the proof of Central Limit Theorem - which will be introduced very soon.

5 Lecture: Distribution Limit Theorems, and Basic Stochastic Processes Terms

Recall, in the last lecture, we discussed the question of concentrating the sample mean around the true mean. We now ask ourselves:

Can we recover the actual distribution from a finite sample?

The celebrated ***Central Limit Theorem*** will give the general answer to this question. However, before we can discuss these theorem, we need to be able to make sense of the *distribtion* of the sum of random variables, and how to analytically manipulate these objects.

5.1 Generating Functions and Characteristic Functions, Studying Sums

We got a result about a coarse representation - namely the *expected value* - from the sum of random variables, properly scaled. We want to be able to get a *distributional result*, that is recover more than just moment information but rather a full characterization of a *limiting distribution.* Again, we will focus on independent random variables, as this gives us the framework to be able to actually analytically study these sums without too many tools.

First, we recall the distribution of the sum of two *independent* random variables is given by the following operation:

Definition 5.1: Convolution

Let X and Y be independent continuous random variables. Let $f_X(x)$ and $f_Y(y)$ be the respective densities, and $F_X(x)$ and $F_Y(y)$ be the respective distribution functions. Then, *convolution* recovers the densities and distribution function of $X+Y$. Explicitly, the operation of *convolution of distribution functions* is defined as

$$F_{X+Y}(a) = (F_X * F_Y)(a) := \int_{-\infty}^{\infty} F_X(a-y)dF(y)$$

The operation of *convolution of density functions* is defined as

$$f_{X+Y}(a) = (f_X * f_Y)(a) := \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy.$$

Remark. We focus on continuous random variables for ease of notation, but the same idea holds for discrete random variables - and more general random variables as well.

Generally, the operation of convolution can be thought of as a *weighted average* of one function against the other. In a sense, we are blending and smoothing out a function. In this setting, if we look at the convolution of the density functions, we are essentially convolving with respect to the density of the Y function. In signal processing, this is a natural operation, as convolutions become multiplications when moving the Fourier domain. This is the central insight that will motivate the use of *characteristic functions* in the proof of Central Limit Theorem - which will be introduced very soon.

Defining Moment Generating Functions and Characteristic Functions. The exponential function will be of much use when studying random variables. First, recall the Taylor expansion of the exponential:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{j=0}^{\infty} \frac{x^j}{j!}.$$

This motivates the following definition:

Definition 5.2: Moment Generating Function of X

Let X be a random variable. Then the *moment generating function* $M_X(t)$ of X is defined for all $t \in \mathbb{R}$ by

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_\Omega e^{tX(\omega)} d\mathbb{P}(\omega) = \int_\mathbb{R} e^{tx} dF_X(x).$$

Remark. Moment generating functions are in one-to-one correspondence with the random variable distributions, in some cases.

As hinted above, one can explicitly recover moments from this function, where the proof follows from a Taylor expansion and the linearity of expectation:

Lemma 5.1: Moments and Moment Generating Functions

Let $M_X(t)$ be the moment generating function of X. Then, for any $p \in \mathbb{N}$:

$$\frac{d^p}{dt^p} M_X(t)\bigg|_{t=0} = \mathbb{E}[X^p].$$

Recall, we have previously discussed how moment matching allows us to get a handle on some generating distribution. Since moment generating functions are in one-to-one correspondence with the random variable distributions, we are justified in that intuition!

Well-Defined? Moment Matching?

Moment generating functions are very useful, especially for discrete valued random variables. However, this occurs only if convergence happens, which is not always guaranteed.

Moreover, if one considers the *log-normal* distribution, one can construct an example of the moments matching for all moments, but the distributions do not coincide.

We have another object that uses the exponential where we augments the argument by i, the imaginary number. Because $|e^{itx}| \leq 1$, we can side-step these convergence guarantee issues:

Definition 5.3: Characteristic Function

Let X be a random variable. Then, the *characteristic function* of X is defined as

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \int_\Omega e^{itX(\omega)} d\mathbb{P}(\omega) = \int_\mathbb{R} e^{itx} dF_X(x).$$

Remark. There is a one-to-one correspondence between function distributions and characteristic functions.

Lemma 5.2: Independence Play Nice With Transforms

If X and Y are independent random variables, then we have the following identities:

- **Moment Generating Function Factorizes:** For all $t \in \mathbb{R}$

$$M_{X+Y}(t) = M_X(t) M_Y(t);$$

- **Characteristic Function:** For all $t \in \mathbb{R}$

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

Before continuing, notice that the Moment Generating Function and the Characteristic Function are precisely a special case of the *Laplace Transform* and *Fourier Transform* seen in Analysis. If you have seen this, this will make the following subsection immediate. Namely, in the language of signal processing, we transform our objects into *frequency space*, work with manipulations there, then *pull back into the spatial domain.*

5.1.1 Inversion and Continuity Theorems

The key reason one works with characteristic functions is because of how they are able to transform the nasty expression of *convolution* into simple multiplication, as well. To make this clear, we recall the *inversion theorem* which makes the remark after the definition clear.

First, we highlight the special case

Lemma 5.3: Special Case of Inversion Theorem

Let X be continuous with density function f and characteristic function ϕ. Then, at every x where f is differentiable, we have

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt.$$

This is saying with the right scaling we can reconstruct our density by considering all of the contributions at each frequency. Recall, how we discussed that we can put a Hilbert space structure on our random variables, this merely says that using some basis - *the Fourier eigenbasis* - we can decompose and recompose our density with respect to this eigenbasis.

More generally, we have the following tehcnical theorem that is *ommitted from Lecture*:

Theorem 5.1: Inversion Theorem

Let X have distribution function F and characteristic function φ. Define $\overline{F} : \mathbb{R} \to [0,1]$ by

$$\overline{F}(x) = \frac{1}{2}\left\{F(x) + \lim_{y\uparrow x} F(y)\right\}.$$

Then

$$\overline{F}(b) - \overline{F}(a) = \lim_{N\to\infty}\int_{-N}^{N} \frac{e^{-iat} - e^{-ibt}}{2\pi it}\varphi(t)dt.$$

As a ***direct corollary*** we recover the one-to-one correspondence between characteristic functions and distribution functions:

Corollary 5.1: One-to-One Correspondence, Characteristic and Distribution Function

Random variables X and Y have the same characteristic function if and only if they have the same distribution function.

Now, recall the notion of convergence in distribution, that is if F_{X_n} and F_X are the corresponding distribution functions to the respective random variables we say that $X_n \Rightarrow X$ if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Thus, we see that the point of this Corollary is that we can translate the convergence of the distributions of a sequence of random variables to the convergence of the characteristic functions. This motivates the next Theorem

Theorem 5.2: Continuity Theorem

Suppose that $F_1, F_2, \ldots$ is a sequence of distribution functions with corresponding characteristic functions $\varphi_1, \varphi_2, \ldots$.

a) If $F_n \to F$ for some distribution function F with characteristic function φ, then $\varphi_n(t) \to \varphi(t)$ for all t;

b) If $\phi(t) = \lim_{n\to\infty} \phi_n(t)$ exists and is continuous at $t = 0$, then φ is the characteristic function of some distribution function F, and $F_n \to F$.

This just establishes rigorously the idea that is written above. These are in fact the tools that make the Central Limit Theorem follow immediately.

5.2 Central Limit Theorem

We state the *Central Limit Theorem*. This follows naturally from the statements above.

Theorem 5.3: Central Limit Theorem - Assymptotic Normality

Let $X_1, X_2, ...$ be a sequence of *iid* random variables, each having mean μ and variance σ^2, both finite. Let $S_n := \sum_{i=1}^n X_i$. Then, the distribution of

$$\frac{S_n - n\mu}{\sqrt{\sigma^2 n}}$$

tends to the standard normal as $n \to \infty$. That is, for $-\infty < a < \infty$, we have that

$$\mathbb{P}\left\{\frac{S_n - n\mu}{\sqrt{\sigma^2 n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

as $n \to \infty$. That is our scaled and centered sum converges in distribution to $\mathcal{N}(0, 1)$.

There is an ***extraordinary*** fact that one can immediately notice - and we mean ***extraordinary*** in the full extent of the word:

Lack of Assumption on Distribution.

Notice that we only require finite variance - and thus mean - of the distribution of the random variable. It cannot be understated how general of a statement this is.

Proof. (*Idea*) Work with the characteristic function of $\frac{S_n - n\mu}{\sqrt{\sigma^2 n}}$, and get those characteristic functions to converge to the characteristic function of $N(0, 1)$, which is explicitly of the form $e^{-\frac{1}{2}t^2}$. Then, use the continuity theorem. □

We note that there are many generalizations of the central limit theorem - as well as the law of large numbers we have seen earlier. For example, one can deal with *dependent variables* and *differently distributed variables*, respectively.

In addition, we note that the Central Limit Theorem can also be interpreted as ***applying the correct scaling limit***. Scaling limits have a rich history in *Statistical Physics* where may consider the microscopic interactions are scaled in a suitable fashion to recover some macroscopic observable. For example, the random motions of molecules give rise to the observable temperature. Now, with that in mind, we can notice that the $\frac{1}{\sqrt{n}}$ is a type of scaling limit. The limit is strong enough to avoid diverging to some infinity, while it is weak enough to **not** eliminate any interesting phenomena. Interesting stuff!

More practically, the Central Limit Theorem gives rise to the confidence intervals that we are familiar with from an introductory Statistics course. In general, the Central Limit Theorem is a powerful tool used across science to give rigorous guarantees on experimental data - as we can more or less assume the data comes as a sequence of *iid* random variables.

5.3 General Stochastic Process Families

We completely switch gears from the first portion of these notes. We begin by briefly mentioning what a *stochastic process* is and highlighting some key families of stochastic processes. Then, at the end of these notes, we will focus purely on *Markov Chains*, a very nice discrete-time stochastic process that can serve as a toy-example for more complicated processes. We will focus in on a key theorem about *ergodic Markov Chains* often discussed at the end of an undergraduate course on Probability Theory, and if there is time highlight some connections with dynamics and PDE.

We first highlight the three many things to keep in mind when considering a *stochastic process*, which is roughly a family of random variables *indexed* by some set T:

Definition 5.4: Stochastic Process

A *stochastic process* is a family of random variables $\{X_t\}_{t\in T}$ where for each $t \in T$, $X_t : \Omega \to S$ for some set S. One makes explicit the following things:

- ***State Space*** S**:** The possible values of X_t;
- ***Index Parameter*** T**:** How one indexes the random variables;
- ***Dependence Among*** X_t**:** How the random variables are related to one another.

We note morally we can regard T as a *time* index, either seen as *discrete-time* - such as when $T = \mathbb{Z}$ or $\mathbb{N}$ - or *continuous-time* - such as when $T = \mathbb{R}$ or $[0, \infty)$.

There is a vast literature introducing and studying broad families of *stochastic processes.* Broad families include *markovian processes*, *martingales*, *stationary processes*, *renewal processes*, *queues*, and *diffusions.* One can learn more about these in a *stochastic process* course, and one can devote an entire career to studying a subset of these processes. Moreover, some stochastic processes of interest are combinations of these general families. Thus, we briefly recall some definitions and defining traits, and then focus in on the process of Markov Chains next lecture.

First, a *Markov Process* is defined by probabilities of any particular future behavior ***only depend on exact knowledge of our current state.*** That is, additional past knowledge does not change the probabilistic behavior. Formally,

Definition 5.5: Markov Process

Let $\{X_t\}_{t\in T}$ be a stochastic process. Then this process is a *Markovian Process* one has

$$\mathbb{P}(a < X_t \leq b \mid X_{t_1} = x_1, \dots, X_{t_n} = x_n) = \mathbb{P}(a < X_t \leq b \mid X_{t_n} = x_n)$$

whenever $t_1 < \cdots < t_n < t$, for any a, b.

Remark. We are implicitly assuming $S \subseteq \mathbb{R}$.

We note that with the assumption that $S \subseteq \mathbb{R}$, one can consider any interval A on $\mathbb{R}$ - note this can be generalized. Then, the function

$$P(x, s; t, A) := \mathbb{P}(X_t \in A \mid X_s = s), \quad t > s,$$

is a *transition probability function* which codifies the probability of transitioning to within A given that we start at state x at time s. When there are a finite or denumerable number of states, these processes are called *Markov chains* and we can the transition probability function can be replaced with the familiar *transition matrix.*

Second, we recall the notion of a *martingale* which is somewhat related to the idea of a *Markov process*, albeit it is possible to have Markovian processes taht are *not* martingales. The intuition is that *previous history and information* does not impact the *expected current outcome of a martingale*:

Definition 5.6: Martingale

Let $\{X_n\}_{n=0}^{\infty}$ be a stochastic process. Then $\{S_n\}_{n=0}^{\infty}$ is a *margtingale* associated with this stochastic process, if for all $n \geq 0$

a) ***Absolutely Integrable:***

$$\mathbb{E}[|S_n|] < \infty$$

b) ***Expect No Change***: Let S_n be some outcome connected with X_n. Then, we have

$$\mathbb{E}[S_{n+1} \mid X_1, \dots, X_n] = S_n.$$

Remark. We assume discrete-time here for simplicity. In the continuous time case, one needs to consider the notion of a *filtration*, that is an *increasing sequence of σ-fields.*

We mention that the other three stochastic processes are not often seen during an undergraduate course, so we record them here with some of their intuitions. If time, we will mention them in lecture.

Stationary processes are characterized by the property that their *finite-dimensional distributions* are ***invariant under time shifts***. In words, they are stationary:

Definition 5.7: Stationary Processes

The process $\{X_t\}_{t\geq 0}$, taking values in $\mathbb{R}$ is called (*strongly*) *stationary* if the families

$$\{X_{t_1}, \dots, X_{t_n}\} \quad \text{and} \quad \{X_{t_1+h}, \dots, X_{t_n+h}\}$$

have the same joint distribution for all $t_1, \dots, t_n$ and $h > 0$.

This also implies that $X(t)$ has the same distribution for all t. We note that one can relax this assumption to *weakly* or *covariance stationary* which asks for the process to have constant means, and an *autocovariance function*, defined as

$$c(t, t+h) := \text{Cov}(X_t, X_{t+h})$$

to satisfy

$$c(t, t+h) = c(0, h), \quad \forall t, h \geq 0.$$

That is, it is only a function of the gap in time. More explicitly,

Definition 5.8: Weakly Stationary Process

The process $\{X_t\}_{t\geq 0}$ is called *weakly stationary* if for all t_1, t_2, and $h > 0$,

$$\mathbb{E}(X_{t_1}) = \mathbb{E}(X_{t_2}) \quad \text{and} \quad \text{Cov}(X_{t_1}, X_{t_2}) = \text{Cov}(X_{t_1+h}, X_{t_2+h}).$$

In some cases, we are interested in the successive occurrences of events. A common special case one sees is a *Poisson process*, but more generally such processes are called *renewal* or *counting processes*:

Definition 5.9: Renewal Process

A *renewal process* $\{N_t\}_{t \geq 0}$ is a process for which

$$N_t = \max\{n : T_n \leq t\}$$

where

$$T_0 = 0, \quad T_n = X_1 + \ldots X_n, \quad \text{for } n \geq 1,$$

where the X_m are *iid* non-negative random variables.

This describes N in terms of some underlying sequence $\{X_n\}$, but of course, one can describe this in absence of the sequence, defining the underlying sequence implicitly:

$$T_n = \inf\{t : N_t = n\}, \quad X_n = T_n - T_{n-1}.$$

The last process we won't explicitly define, but these are the *Wiener processes* or otherwise referred to as *Brownian motion.* These are continuous time processes that exhibit two basic properties: ***time-homogeneity*** and ***independent increments***. These are often seen at the end of an introductory *stochastic processes course.*

6 Lecture: Basic Markov Chains

Recall, *Markov chains are a special case of Markovian processes.* These are stochastic processes with the property that, conditional on their present value, the future is independent of the past. We will focus on *discrete-time* Markov chains, with *finite state space.* Markov chains will be the focus because

Markov chains provide a good toy example for stochastic processes, and allow us to think more about distributional convergence.

Before going into definitions, we first recall the notion of a directed graph, and how we can model Markov chains in such a manner:

Definition 6.1: Directed Graph

A *weighted directed graph* is defined by three objects $G = (V, E, W)$:

- $V = \{v_i\}_{i=1}^n$: This is the *vertex set* which comprises of members that have "relationships" with one another;
- $E = [n] \times [n]$: This is the *edge set* which comprises of all **ordered** pairs of nodes;
- $W : E \to \mathbb{R}_{\geq 0}$: This is the *edge weight function* that assigns a weight to $e \in E$. These individual weights are denoted as w_e. If $w_e > 0$, we draw a directed arrow from the first node to the second node.

Remark. The weight function in this case can be written as a matrix.

With this in mind, a Markov chain can be modeled as a *completely connected* directed graph, where the ***states*** play the role of the *nodes/vertices* and the ***probabilities of transition*** from one state to another is captured in the *edge weights.* In fact, we can even codify the weights W as a matrix where the (i, j) entry is given by the weight on the edge connected node i to j, this has a special meaning for Markov chains.

6.1 Markov Chains

Before continuing from before we will make the following assumption about the Markov chain:

Definition 6.2: Homogeneous Markov Chain

A Markov chain is called *homogeneous* if

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(X_1 = j \mid X_0 = i)$$

for all n, i, j.

In words, this says that the transition probabilities do not change over time. We will henceforth assume *Markov chains are homogeneous*, this is a common assumption unless specified.

Now, with directed graph in mind, we will first focus on the *weight matrix* of the directed graph, as this will dictate the *dynamics* of a Markov chain. With the assumption of homogeneity in mind, this *weight matrix will be fixed for all time.* For notational consistency, we will denote the weight matrix by P, as it defines the transition probabilities between states:

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix}.$$

Since we want to keep intuition of probabilities, we see this matrix has the following properties:

$$p_{ij} \geq 0, \quad \text{for all } i, j \in [n],$$
$$\sum_{j=1}^{n} p_{ij} = 1.$$

This property is referred to *row-stochastic.* Henceforth, the matrix will be referred to as the *transition matrix.*

Now, since random variables at a fixed time have a distribution for their outputs, we can associate with every vertex a probability. Namely, we can have a distribution amongst the vertices, and in this setting we will codify this distribution as a row vector whose entries sum to one

$$\mu_t = \begin{bmatrix} \mu_t(v_1) & \dots & \mu_t(v_n) \end{bmatrix}.$$

Here, μ_{t_i} expresses the probability that X_{t_i} has the value v_{t_i}.

With the distribution at time t given as a *row-vector* and the transition probabilities as *matrix*, it is natural assume that the distribution after one-time step can be computed as a row-matrix multiplication:

$$\mu_t P = \mu_{t+1}.$$

We ccan see this explicitly by carrying out the computation:

Row-Matrix Multiplication Evolves Distribution of Markov Chains by a Time Step

Fix our attention on some state v_j at time step $t+1$. Notice that if we consider the column vectors of P

$$P = \begin{bmatrix} & | & \\ \dots & p_{\cdot,j} & \dots \\ & | & \end{bmatrix},$$

the $p_{\cdot,j}$ column corresponds with all of the probabilities of ending up at state v_j. Then, notice that

$$\pi_{t+1}(v_j) = \pi_t p_{\cdot,j} = \sum_{i=1}^{n} \pi_t(v_i) p_{i,j}.$$

Notice that the sum on the right is merely summing the probabilities of ending up at v_j across all states, scaled by the probability of being at those states at the previous time step. Since this intuition holds for all v_j, it is clear that row-matrix multiplication corresponds with taking a time step into the future.

Thus, we can characterize the dynamics of the (homogenous) *Markov Chain* through the right-action of the matrix on the distribution row vector.

As before, we are interested in the *long-term behavior* of the Markov Chain. In fact, we have access to description of both the *short term* behavior, described by the transition matrix P, and longer-term behaviors which are described in the following fashion:

Definition 6.3: n-Step Transition Matrix

The *n-step transition matrix* $P(m, m+n) = (p_{ij}(m, m+n))$ is the matrix of n-step transition probabilities

$$p_{ij}(m, m+n) = \mathbb{P}(X_{m+n} = j \mid X_m = i).$$

Now, since we assumed that the Markov Chain is *homogenous* we have that $P(m, m+1) = P$. Moreover, we have the following important fact:

Theorem 6.1: Chapman-Kolmogorov Equations

$$p_{ij}(m, m+n+r) = \sum_k p_{ik}(m, m+n)p_{kj}(m+n, m+n+r).$$

Therefore, it follows that

$$P(m, m+n+r) = P(m, m+n)P(m+n, m+n+r),$$

and

$$P(m, m+n) = P^n,$$

that is the nth power of the transition matrix P.

Remark. More generally, since we are working with a homogenous Markov chain, this says that the family of transition matrices forms a *Markov semigroup*. In short, these give rise to a whole rise of tools from dynamics to study these systems as well.

Proof. (*Scheme*) Use the fact that $p_{ij}(m, m+n+r) = \mathbb{P}(X_{m+n+r} = j \mid X_m = i)$, and $\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid B \cap C)\mathbb{P}(B \mid C)$. □

More succinctly, this theorem allows us to relate *long-term* with *short-term* in a succinct fashion:

Corollary 6.1

$$\mu^{(m+n)} = \mu^{(m)}P^n \quad \text{and} \quad \mu^{(n)} = \mu^{(0)}P^n.$$

Thus, we see that the

random evolution of the Markov chain is determined by the transition matrix P and the initial mass function $\mu^{(0)}$.

We note, from last time we recalled there are many *models* for *random dynamics*, but Markov chains give a particularly *simple* model to analyze. We can envision this as a walker randomly moving from one state to another according to the transition probabilities, with a memoryless assumption - the *Markov assumption*. Moreover, under the assumption of *homogeneity* it follows that these transition probabilities don't change over time.

Because of the simplicity of this model, it follows that *limiting phenomena* of Markov Chains can often be reduced to the *algebraic properties of these transition matrices*. The goal is to recover distributions of the following form, as well as rates of convergence to such distributions if convergence occurs:

Definition 6.4: Stationary Distribution

We say π is a *stationary distribution*, if it satisfies the property that

$$\pi P = \pi.$$

Thus, we are interested in the existence and uniquness of the such a limiting distribution given some initial distribution μ_0:

$$\lim_{n\to\infty} \mu_0 P^n = \pi.$$

Let's consider an easy to compute example of a Markov Chain.

Markov Chain Example. Consider a system that consists of three states. Let the *transition matrix* have the following form:

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

One can check that the transition matrix is *row-stochastic*.

Then, if we set $\mu_0 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$, then, we have that

$$\mu_0 P = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} = \mu_1.$$

More generally, at time t, by Corollary 6.1, we have that $\mu_t = \mu_0 P^t$, and for example one can compute that

$$\mu_3 \approx \begin{pmatrix} 0.2778 & 0.3611 & 0.3611 \end{pmatrix}, \quad \mu_4 \approx \begin{pmatrix} 0.2593 & 0.3704 & 0.3704 \end{pmatrix}, \quad \mu_{100} \approx \begin{pmatrix} 0.25 & 0.375 & 0.375 \end{pmatrix}.$$

It appears that as $t \to \infty$, we are converging to some *stationary distribution*, in this case

$$\mu_\infty = \begin{pmatrix} \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{pmatrix}.$$

In fact, this is true under mild assumptions ***regardless of the initial distribution***. Moreover, the rate of convergence can be specified. Moreover, one can compute that this really is a left eigenvector corresponding with eigenvalue 1, as hinted at before.

6.2 Markov Chain State Classifications, Ergodic Markov Chains

We now recall the conditions that ensure the basic limit theorem on Markov Chains. This boils down to classifying the states of the Markov Chain, we want to make sense of the following definition

Definition 6.5: Ergodic State

A state is called *ergodic* if it is persistent, non-null, and aperiodic.

Each component has an intuitive meaning behind them. First, persistence morally says that we always return to the state with probability 1

Definition 6.6: Persistent State

State i is called *persistent* if

$$\mathbb{P}(X_n = i \quad \text{for some } n \geq 1 \mid X_0 = i) = 1.$$

If this is strictly less than 1, then the state is called *transient*.

Now, while we can ensure that we return to a state with probability 1, we want to make sure that on average it does not take forever. This is codified by the following notion. First, define

$$T_j = \min\{n \geq 1 : X_n = j\}$$

as the first time we visit j. If $T_j = \infty$, we say that we never visit. Now, one defines the following:

Definition 6.7: Mean Recurrence Time

The *mean recurrence time* μ_i of a state is defined as

$$\mu_i := \mathbb{E}[T_i \mid X_0 = i].$$

This may be infinite even if i is persistent. Thus, to rule out this possibility, we define

Definition 6.8: Non-Null State

For a persistent state i,

$$i \text{ is called } \begin{cases} \textbf{null}, & \text{if } \mu_i = \infty, \\ \textbf{non-null}, & \text{if } \mu_i < \infty. \end{cases}$$

These technical conditions simplify in the finite Markov chain case. Lastly, we make the following *technical* assumption that forces uniqueness of the stationary distribution:

Definition 6.9: Aperiodic States

The *period* $d(i)$ of a state is defined by $d(i) = \gcd(n : p_{ii}(n) > 0\}$. We call i *periodic* if $d(i) > 1$ and *aperiodic* if $d(i) = 1$.

Thus, we have the notion of an *ergodic state.*

In order to describe the entire system, we just have to make mild assumptions on *how these states interact.* This is captured in the notion of *communication.* This merely means that if you're at one state, it is possible to get to another state, and vice versa:

Definition 6.10: Communicates and Intercommunicates

We say state i *communicates with* state j, $i \to j$, if $p_{ij}(m) > 0$ for some $m \geq 0$. Moreover, we say i and j *intercommunicate* if $i \to j$ and $j \to i$, in which case we write $i \leftrightarrow j$.

These were all statements about individual states of the Markov Chain, we have the natural definitions about the entire system:

Definition 6.11: Ergodic Markov Chains and Irreducible Markov Chains

We say that the Markov Chain is

- *Irreducible* if for all $i, j \in S$ we have $i \leftrightarrow j$;
- *Ergodic* if for all $i \in S$ we have that i is *ergodic.*

6.2.1 A Fundamental Limit Theorem for Markov Chains

We have the following Theorem, which we refer to as a *Fundamental Theorem of Ergodic Markov Chains*

Theorem 6.2: Fundamental Theorem

Let P be a transition matrix that describes the dynamics of an *ergodic, irreducible* Markov chain. Then, regardless of the initial distribution μ_0, there exists a unique *stationary distribution* π, such that

$$\lim_{n \to \infty} \mu_0 P^n = \pi.$$

Moreover, we can even characterize the stationary distribution in terms of the *mean recurrence time*

Lemma 6.1: Characterizing Probability

For an irreducible aperiodic chain, we have that

$$p_{ij}(n) \to \frac{1}{\mu_j}$$

as $n \to \infty$ for all $i \to j$.

Recalling that $p_{ij}(n) := \mathbb{P}(X_n = j \mid X_0 = i)$, this says that regardless of i, we end up at the same stationary distribution. In other words, the chain forgets the origin, and we have that the convergence to the stationary distribution regardless of how we initialize the initial distribution.

6.3 Quantifying Convergence of Distributions

As applied mathematicians, we are perhaps unsatisfied with the presentation of Markov chains given so far. Namely, these statements are asymptotic, and we often want to ask ourselves if we can give robust guarantees on how long one has to run a system before being at stationary. In fact, this is a current field of *research*, where one may see this from a *Variational Bayesian Perspective* or an *Optimization Perspective*. As a common mantra from Analysis:

If we wish to give rates of convergence, we need to specify in what sense!

Thus, we close with highlighting a few different notions of metric or distance that are often used to describe the distance between distributions.

Definition 6.12: Total Variation Distance

Consider some measurable space $(\Omega, \mathcal{F})$ with two probability measures/distributions that can be defined on this space, namely P and Q. Then the *total variation distance* can be defined as

$$d_{TV} := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|.$$

Intuitively, this gives a worse-case difference between the two distributions. It returns the largest possible difference in how the distributions assign probabilities to measurable events.

Definition 6.13: Kullback-Leibler (KL) Distance

Consider some measurable space $(\Omega, \mathcal{F})$ with two probability measures/distributions that can be defined on this space, namely P and Q. Then the KL-Distance is defined as

$$D_{KL}(P \| Q) = \int_\Omega \log\left(\frac{P(d\omega)}{Q(d\omega)}\right) dP(\omega).$$

In a more special case setting, namely a discrete probability space, we have that

$$D_{KL}(P \| Q) = \sum_{\omega \in \Omega} P(\omega) \log\left(\frac{P(\omega)}{Q(\omega)}\right).$$

In other words, this is the expectation of the *logarithmic difference* between the probability distributions P and Q with respect to the distribution P.

This distance, otherwise called relative entropy, can be interpreted as the expected excess surprise from using Q as a model when the actual distribution is P. While not a metric, this finds many uses in an information-theoretic context.

Remark. This is not a metric, it is merely a distance.

Definition 6.14: Wasserstein Distances

Suppose that we have a measurable space $(\Omega, \mathcal{F})$ with two probability measures P and Q. Then the p-Wasserstein distance on Ω is defined as

$$W_p(P,Q) := \inf_{\Pi \in \mathcal{U}(P,Q)} \left(\int_{\Omega \times \Omega} d(x,y)^p d\Pi(x,y) \right)^{\frac{1}{p}},$$

where d is the distance on Ω, and $\mathcal{U}$ denotes all possible *couplings*. A coupling dictates a *transportation plan*, or in other words a way of moving one distribution to another.

This notion may take many times for it to sink in what this truly means but the intuition lies on how one can transport a sand pile from one orientation to another. This is far beyond the scope of what is expected, but this is a common distance used in many *modern* applications. Feel free to ask me more about it after lecture.

With these notions of distances, one can then begin to make statements about what kind of assumptions ensure specific rates of convergence, which we do not state here.

In closing, overall there is a lot of incredibly interesting mathematical arguments used to get a handle on really important applications - however at the heart of it all are fundamental fields such as Probability Theory, Linear Algebra, and Statistics. It is always worth going back to your basics and sharpening your fundamentals!