Probability Theory

Muchang Bahng

Spring 2023

Contents

An overview of probability using measures. We will denote probability measures defined over σ-algebras with $\mathbb{P}$ and probability functions defined over some sample space Ω or $\mathbb{R}$ with P or p. When we say countable, we mean finite or countably infinite. I have used resources from:

1. Rick Durret's *Elementary Probability* and *Probability* textbooks.
2. Dr. Krishna's *Probability Foundation for Electrical Engineers* lectures at IIT.
3. Various quant interview books and websites for examples.

1 Measure Spaces

Let's do a little refresher on measure theory.

1.1 Sigma Algebras and Measures

Definition 1.1 (σ-algebra)

A **σ-algebra** on a set X is a collection of subsets of X, denoted $\mathcal{A} \subset 2^X$, satisfying
1. Contains Empty Set: $\emptyset \in \mathcal{A}$
2. Stability under Complementation: $A \in \mathcal{A} \implies A^c \in \mathcal{A}$, where $A^c = X - A$
3. Stability under Countable Union: If $\{A_i\}$ is a countable sequence of sets, then

$$\bigcup_i A_i \in \mathcal{A} \tag{1}$$

At first, we might wonder why we need σ-algebras in the first place. We want to identify sets that are measurable in the way that their size can be determined, but why not just use 2^X? This is because of the Banach-Tarski paradox, which gives you contradictions if you try to define a measure over 2^X.

Lemma 1.1 (Additional Property of σ-Algebras)

A commonly known property of any σ-algebra $\mathcal{A}$ is that it is stable under countable intersections, too.

$$A_1, A_2, \ldots, \in \mathcal{A} \implies \bigcap_{k=1}^{\infty} A_k \in \mathcal{A} \tag{2}$$

Proof.

We can utilize the fact that

$$\bigcap_{k=1}^{\infty} A_k = X \setminus \bigcup_{k=1}^{\infty} A_k^c \tag{3}$$

A σ-algebra is similar to the topology τ of topological space. Both $\mathcal{A}$ and τ require $\emptyset$ and X to be in it. The three differences are that (i) τ does not allow compelmentation, (ii) τ allows any (even uncountable) union of sets (condition is strengthened), and (iii) τ allows only finite intersection of sets (condition is weakened). Now in order to construct σ-algebras, the following theorems are useful since they allow us to construct σ-algebras from other σ-algebras. It turns out that the intersection of σ-algebras is a σ-algebra, but not for unions.

Theorem 1.1 ()

Let $\{\mathcal{A}_k\}$ be a family of σ-algebras of X. Then, $\cap \mathcal{A}_k$ is also a σ-algebra of X.

Proof.

Clearly, $\emptyset, X$ is in $\cap \mathcal{A}_k$. To prove complementation,

$$A \in \bigcap \mathcal{A}_k \implies A \in \mathcal{A}_k \; \forall k \implies A^c \in A_k \; \forall k \implies A^c \in \bigcap \mathcal{A}_k \tag{4}$$

To prove countable union, let $\{A_j\}_{j \in J}$ be some countable family of subsets in $\cap \mathcal{A}_k$. Then,

$$A_j \in \bigcap \mathcal{A}_k \; \forall j \in J \implies A_j \in \mathcal{A}_k \; \forall k \forall j \implies \bigcup A_j \in \mathcal{A}_k \; \forall k \implies \bigcup A_j \in \bigcap \mathcal{A}_k \tag{5}$$

This allows us to easily prove the following proposition, which just establishes the existence of σ-algebras.

Proposition 1.1 ()

Let $F \subset 2^X$ be a collection of subsets of X. Then there exists a unique smallest σ-algebra $\sigma(F)$ containing F. $\sigma(F)$ is called the σ-algebra **generated** by F.

Proof.

Let us denote $\mathcal{M}$ as the set of all possible σ-algebras $\mathcal{B}$ of X. $\mathcal{M}$ is nonempty since it contains 2^X. Then, the intersection

$$\bigcap_{\mathcal{B} \in \mathcal{M}} \mathcal{B} \tag{6}$$

is the unique smallest σ-algebra.

Now, how do we measure a size on subsets of X? We use measures.

Definition 1.2 (Measure)

Given a measurable space $(X, \mathcal{A})$, a **measure** is a function $\mu : \mathcal{A} \longrightarrow [0, +\infty]$ satisfying

1. Positive Definite: $\mu(A) \geq \mu(\emptyset) = 0$
2. Countable Additivity: For all countable collections $\{A_k\}_{k=1}^{\infty}$ of pairwise disjoint subsets $A_k \in \mathcal{A}$,

$$\mu\left(\bigsqcup_{k=1}^{\infty} A_k \right) = \sum_{k=1}^{\infty} \mu(A_k) \tag{7}$$

Remember that we are allowed to take countable unions inside our σ-algebra, so this makes sense. Disjointness is clearly important since if it wasn't, then $\mu(A) = \mu(A \cup A) = 2\mu(A)$, which is absurd.

The triplet $(X, \mathcal{A}, \mu)$ is called a **measure space**.

Theorem 1.2 (Properties of Measure)

Let μ be a measure on $(X, \mathcal{A})$.

1. Monotonicity: If $A \subset B$, then $\mu(A) \leq \mu(B)$.
2. Subadditivity: If $A \subset \cup_{i=1}^{\infty} A_i$, then $\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i)$
3. Continuity from Above: If $A_1 \subset A_2 \subset A_3 \subset \ldots$, then

$$\mu\left(\bigcup_{k=1}^{\infty} A_k \right) = \lim_{k \to \infty} \mu(A_k) \tag{8}$$

4. Continuity from Below: If $A_1 \supset A_2 \supset A_3 \supset \ldots$ and $\mu(A_1) < \infty$, then

$$\mu\left(\bigcap_{k=1}^{\infty} A_k \right) = \lim_{k \to \infty} \mu(A_k) \tag{9}$$

Proof.

Listed.

1. Let $B \setminus A \coloneqq B \cap A^c$. Then, since A and $B \setminus A$ are disjoint, we have

$$\mu(B) = \mu\big(A \cup (B \setminus A)\big) = \mu(A) + \mu(B \setminus A) \geq \mu(A) \tag{10}$$

2. We again try to divide this union into disjoint sets. Let $A_i' = A \cap A_i$, and let $B_1 = A_1'$ with

$$B_i = A_i \setminus \bigcup_{j=1}^{i-1} A_j' \tag{11}$$

Since B_i's are disjoint with $B_i \subset A_i$, we can use the first property to get

$$\mu(A) = \sum_{i=1}^{\infty} \mu(B_i) \leq \sum_{i=1}^{\infty} \mu(A_i) \tag{12}$$

3. This is the first time we introduce limits. With the fact that $\mu(A_k)$ must be nondecreasing, we can use real analysis and see that it is bounded by ∞, meaning that it must have a limit. But why does this limit equal to the left hand side? We can see that

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \mu(A_1) + \sum_{k=2}^{\infty} \mu(B_k) \tag{13}$$

$$= \mu(A_1) + \lim_{k\to\infty} \sum_{k=2}^{\infty} \mu(B_k) \tag{14}$$

$$= \lim_{k\to\infty} \mu(A_1 \cup B_2 \cup \dots B_k) = \lim_{k\to\infty} \mu(A_k) \tag{15}$$

where $B_k = A_k \setminus A_{k-1}$.

4. The $\mu(A_1) < \infty$ is a necessary condition, since if we take $A_k = [k, \infty)$ on the real number line, then we have $\cap_{k=1}^{\infty} A_k = \emptyset$, but the limit of the measure is ∞. Well we can define $B_k = A_k \setminus A_{k+1}$ and write $\cap_{k=1}^{\infty} A_k = A_1 \setminus \cup_{k=1}^{\infty} B_k$, which means that

$$\mu\left(\bigcap_{k=1}^{\infty} A_k\right) = \mu\left(A_1 \setminus \bigcup_{k=1}^{\infty} B_k\right) \tag{16}$$

$$= \mu(A_1) - \mu\left(\bigcup_{k=1}^{\infty} B_k\right) \tag{17}$$

$$= \mu(A_1) - \sum_{k=1}^{\infty} \mu(B_k) \tag{18}$$

$$= \mu(A_1) - \lim_{K\to\infty} \sum_{k=1}^{K} \mu(B_k) \tag{19}$$

$$= \lim_{K\to\infty} \left(\mu(A_1) - \sum_{k=1}^{K} \mu(B_k)\right) \tag{20}$$

$$= \lim_{K\to\infty} \mu\left(A_1 \setminus \bigcup_{k=1}^{K} B_k\right) = \lim_{K\to\infty} \mu(A_K) \tag{21}$$

Now the first line uses the fact that if $A \subset B$, then $\mu(B \setminus A) + \mu(A) = \mu(B)$, and with the further assumption that $\mu(A) < \infty$, we can subtract on both sides like we do with regular arithmetic.

Theorem 1.3 (Inclusion Exclusion Principle)

One familiar property commonly seen in probability and combinatorics is the inclusion exclusion

principle. If $A, B \in \mathcal{A}$,

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \tag{22}$$

and by induction, if $A_1, \dots, A_n \in \mathcal{F}$, then

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i) - \sum_{i<j} \mu(A_i \cap A_j) + \sum_{i<j<k} \mu(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n-1} \mu\left(\bigcap_{i=1}^{n} \mu(A_i)\right) \tag{23}$$

Finally, here is a definition which will be useful shortly when talking about how σ-algebras model knowledge.

Definition 1.3 (Sub-σ-Algebras)

Given a σ-algebra $\mathcal{F}$, a **sub-σ-algebra** of $\mathcal{F}$ is a σ-algebra $\mathcal{G}$ s.t. $\mathcal{G} \subset \mathcal{F}$.

This will allows us to compare σ-algebras by taking two σ-algebras $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$, which μ is guaranteed to be defined on since it is defined over $\mathcal{F}$.

1.1.1 Construction of Measure on $\mathbb{R}^n$

Let $\mathbb{R}^n$ be the continuum and $\mathcal{R}^n$ be the **Borel σ-algebra**, defined as the σ-algebra generated by the open sets of $\mathbb{R}^n$.

Example 1.1 (Stieltjes Measure Function)

Measures on $(\mathbb{R}, \mathcal{R})$ are defined by giving a **Stieltjes measure function** with the following properties:

1. F is nondecreasing
2. F is right continuous:

$$\lim_{y \downarrow x} F(y) = F(x) \tag{24}$$

Theorem 1.4 ()

Associated with each Stieltjes measure function F there is a unique measure μ on $(\mathbb{R}, \mathcal{R})$ with

$$\mu((a, b]) = F(b) - F(a) \tag{25}$$

When $F(x) = x$, then the resulting measure is called the **Lebesgue measure**.

This is quite a hard proof, but we outline the construction of this measure on $\mathbb{R}$. First, we would like to define a "nice" set of half-open half-closed intervals, which we show is a semialgebra $\mathcal{S}$. We can easily define a measure μ on this semialgebra. We can extend this semialgebra to an algebra $\overline{\mathcal{S}}$, along with a proper extension $\overline{\mu}$ that is a unique measure on $\overline{\mathcal{S}}$.

Definition 1.4 (Semialgebra, Algebra)

A collection $\mathcal{S}$ of sets is said to be a **semialgebra** if

1. it is closed under intersection
2. If $S \in \mathcal{S}$, then S^c is a finite disjoint union of sets in $\mathcal{S}$

A collection $\mathcal{A}$ of subsets is said to be an **algebra** if

1. it is closed under union
2. it is closed under complementation
3. the first two imply that it is closed under intersection

We can see that a set that is a σ-algebra $\implies$ it is an algebra.

Here is an example of a semialgebra, which we will utilize in building a measure on $\mathbb{R}^n$.

Example 1.2 ()

Let $\mathcal{S}_d$ be the empty set plus all sets of the form

$$(a_1, b_1] \times \ldots \times (a_d, b_d] \subset \mathbb{R}^d \tag{26}$$

where $-\infty \leq a_i < b_i \leq +\infty$. $\mathcal{S}_d$ is a semialgebra since

$$\left(\prod_i (a_i^1, b_i^1] \right) \cap \left(\prod_i (a_i^2, b_i^2] \right) = \prod_i (\max\{a_i^1, a_i^2\}, \min\{b_i^1, b_i^2\}] \tag{27}$$

and ...

Now, we show that we can extend this semialgebra to an algebra.

Lemma 1.2 ()

If $\mathcal{S}$ is a semialgebra, then $\overline{\mathcal{S}} = \{\text{finite disjoint unions of sets in } \mathcal{S}\}$ is an algebra, called the algebra generated by $\mathcal{S}$.

Proof.

Example 1.3 ()

Given $\mathbb{R}$ and its semialgebra $\mathcal{S}_1$, then $\overline{\mathcal{S}}_1$ consists of the empty set and all sets of the form

$$\bigcup_{i=1}^n (a_i, b_i] \text{ where } -\infty \leq a_i < b_i \leq +\infty \tag{28}$$

Now as for extending our measure function to $\overline{\mathcal{S}}$, we can simply use the properties. Note that since since an algebra is constructed from finite disjoint unions of a semialgebra, given that the finite collection $\{A_i\}_{i=1}^n$ all reside in $\mathcal{S}$ and are disjoint, then their disjoint union must be in $\overline{\mathcal{S}}$ and must be measurable, defined as

$$\overline{\mu}\left(\bigsqcup_{i=1}^n A_i \right) = \sum_{i=1}^n \mu(A_i) \tag{29}$$

Definition 1.5 (σ-finite measure)

Given a measure on an algebra $\mathcal{A}$, μ is said to be **σ-finite** if there is a sequence of sets $A_1, A_2, \ldots \in \mathcal{A}$ s.t. $\mu(A_i) < \infty$ and $\cup_i A_i = \Omega$.

Theorem 1.5 ()

Let $\mathcal{S}$ be a semialgebra and let μ defined on $\mathcal{S}$ have $\mu(\emptyset) = 0$. Suppose

1. if $S \in \mathcal{S}$ is a finite disjoint union of sets $\{S_i\}_{i=1}^n$, then

$$\mu(S) = \sum_{i=1}^n \mu(S_i) \tag{30}$$

2. f S is a countably infinite disjoint union of sets $\{S_j\}_{j=1}^{\infty}$, then

$$\mu(S) \leq \sum_{j=1}^{\infty} \mu(S_j) \tag{31}$$

Then, μ has a unique extension $\bar{\mu}$ that is a measure on $\overline{\mathcal{S}}$, the algebra generated by $\mathcal{S}$. If $\bar{\mu}$ is σ-finite, then there is a unique extension ν that is a measure on $\sigma(\mathcal{S})$ (the smallest σ-algebra containing $\mathcal{S}$).

1.2 Probability Spaces

Definition 1.6 (Probability Space)

A **probability space** is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}(\Omega) = 1$.
1. Ω is called the **sample space** and an element $\omega \in \Omega$ is called an outcome.
2. $\mathcal{F}$ is called the **event space** and an element $A \in \mathcal{F}$ is called an event.
3. The measure of an event $\mathbb{P}(A)$ is called the **probability** of that event.

We can think of the sample space Ω as the set of all conceivable futures and an event $F \in \mathcal{F}$ as some subset of conceivable futures. The probability $\mathbb{P}(F)$ represents our degree of certainty that our future will be contained in such an event. If some measure space X has a finite total measure, we can construct a probability space from it by normalizing the measure.

1.2.1 Sigma-Algebras as Models of Knowledge

Let us focus on the σ-algebra $\mathcal{F}$. We can see that the σ-algebra *models our knowledge of the experiment.* That is, given some outcome space Ω, let us have two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ such that $\mathcal{F} \subset \mathcal{G}$, i.e. $\mathcal{F}$ is a sub-σ-algebra of $\mathcal{G}$. What does this mean? Remember that the elements of the event space are the events that can be measured. If $\mathcal{G}$ is *finer* than $\mathcal{F}$, then every set F that is $\mathcal{F}$-measurable is also $\mathcal{G}$-measurable, and so someone who has knowledge of μ over $\mathcal{G}$ knows more than another who has knowledge of μ over $\mathcal{F}$.

For example, let us have a dice roll, with $\Omega = \{1, 2, 3, 4, 5, 6\}$.

1. Abby's knowledge is modeled by $\mathcal{F} = \{\emptyset, \{1, 2, 3\}, \{4, 5, 6\}, \Omega\}$, with

$$\mathbb{P}(F) = \begin{cases} 0 & \text{if } F = \emptyset \\ 1/2 & \text{if } F = \{1, 2, 3\}, \{4, 5, 6\} \\ 1 & \text{if } F = \Omega \end{cases} \tag{32}$$

2. Bob's knowledge is modeled by $\mathcal{G} = 2^{\Omega}$ with the following values

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = \frac{1}{6} \tag{33}$$

We can see that $\mathcal{F} \subset \mathcal{G}$ and that Bob has more information than Abby since from the values of $\mathbb{P}$ over his σ-algebra, he can deduce that $\mathbb{P}(\{1, 2, 3\}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = 1/2$ (and likewise for $4, 5, 6$). All Abby knows is that the probability that the roll is $1, 2, 3$ is $1/2$, but in her view, the individual probabilities may not be uniformly $1/6$ at all (it could be $\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = 0$ and $\mathbb{P}(\{3\}) = 1/2$, for example). More specifically, Bob has *complete* information of the experiment since $\mathcal{G} = 2^{\Omega}$, so he knows the probability of every possible event. But no matter how little information one has about the experiment, *everybody* will always know that the probability that *any* outcome will happen is 1 (hence $\mathbb{P}(\Omega) = 1$) and the probability that no outcome will happen is 0 ($\mathbb{P}(\emptyset) = 0$), which is consistent with the definition of σ-algebras requiring to have $\emptyset$ and Ω. Note that we can have two σ-algebras s.t. both model incomplete information and aren't strictly finer than one another, i.e. $\mathcal{F} \not\subset \mathcal{G}$ and $\mathcal{G} \not\subset \mathcal{F}$.

Note that given the same random experiment, we don't need to always have the same sample space or the same random variable. For example, let's have a coin toss. One could be interested in whether it lands heads or tails, which means $\Omega = \{0, 1\}$, but another could be interested in the number of times the coin flips midair, in which $\Omega = \mathbb{N}_0$. We could even be interested in the set of all trajectories of the coin, which would result in a huge space of all trajectories of the flip, or the velocity at which it lands on the table, which would lead to $\Omega = \mathbb{R}^+$.

Note that as you get more and more information, your σ-algebra can "grow" and get closer to something that models complete information. This means that given some σ-algebra $\mathcal{F}$ that models complete information, we can take a sequence of nondecreasing sub-σ-algebras of $\mathcal{F}$

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_i \subset \ldots \tag{34}$$

such that $\mathcal{F}_i \subset \mathcal{F}$, which models our increasing knowledge of the experiment.

Definition 1.7 (Filtration)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and I be an index set with total order $\geq$ (usually, $\mathbb{N}, \mathbb{R}$). For every $i \in I$, let $\mathcal{F}_i$ be a sub-σ-algebra of $\mathcal{F}$ satisfying

$$\mathcal{F}_i \subset \mathcal{F}_j \text{ if } i \geq j \tag{35}$$

Note that we do not write it as a sequence like before since I may be uncountable. Then, a **filtration** $\mathbb{F} = \{\mathcal{F}_i\}_{i \in I}$ is a family of σ-algebras that are ordered nondecreasingly. If $\mathcal{F}$ is a filtration, then $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ is called a **filtered probability space**.

Example 1.4 (Filtration of 3 Coin Tosses)

Let us describe a concrete example of a 3-coin toss filtration. The probability space is

$$\Omega = \{000, 001, 010, 011, 100, 101, 110, 111\} \tag{36}$$

which has 8 outcomes so a complete σ-algebra would consist of $2^8 = 256$ outcomes.

1. Before the experiment, we have no information at all, so

$$\mathcal{F}_0 = \{\emptyset, \Omega\} \tag{37}$$

   which has $2^{2^0} = 2$ elements.
2. After the first coin toss, we would have information on what the first flip landed on (whether it was of form $(0, *, *)$ or $(1, *, *)$), so we have a σ-algebra generated by these two events

$$\begin{aligned} \mathcal{F}_1 &= \sigma(\{(0, *, *)\}, \{(1, *, *)\}) \\ &= \sigma(\{000, 001, 010, 011\}, \{100, 101, 110, 111\}) \\ &= \{\emptyset, \{000, 001, 010, 011\}, \{100, 101, 110, 111\}, \Omega\} \end{aligned}$$

   which has $2^{2^1} = 4$ elements.
3. After the second coin toss, we would have information on what the first two flips landed on (whether it was of form $(0, 0, *)$, $(0, 1, *)$, $(1, 0, *)$, $(1, 1, *)$), so we have a σ-algebra generated

by these 4 events

$$\begin{aligned}\mathcal{F}_2 &= \sigma(\{(0,0,*)\},\{(0,1,*)\},\{(1,0,*)\},\{(1,1,*)\}) &(38)\\ &= \sigma(\{000,001\},\{010,011\},\{100,101\},\{110,111\}) &(39)\\ &= \{\emptyset,\{000,001\},\{010,011\},\{100,101\},\{110,111\}, &(40)\\ &\quad \{000,001,010,011\},\{000,001,100,101\},\{000,001,110,111\}, &(41)\\ &\quad \{010,011,100,101\},\{010,011,110,111\},\{100,101,110,111\}, &(42)\\ &\quad \{000,001,010,011,100,101\},\{000,001,010,011,110,111\}, &(43)\\ &\quad \{000,001,110,101,110,111\},\{010,011,110,101,110,111\},\Omega\} &(44)\end{aligned}$$

which has $2^{2^2} = 16$ elements.

4. After the third coin toss, we would have information on what the first three flips landed on (all 8 possibilities in Ω), so we have a σ-algebra generated by these 8 events

$$\mathcal{F}_3 = \sigma(\{000\},\{001\},\{010\},\{011\},\{100\},\{101\},\{110\},\{111\}) \tag{45}$$

This is too big to write down explicitly, but it has $2^{2^3} = 256$ elements.

1.2.2 Types of Probability Spaces

Definition 1.8 (Discrete Probability Space)

If Ω is a countable set, then we can take its σ-algebra $\mathcal{F}$ to be the power set of Ω and construct the measurable space $(\Omega, 2^\Omega, \mathbb{P})$. From the axioms, for any event $A \in \mathcal{F}$, we have

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) \text{ and } \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1 \tag{46}$$

The greatest σ-algebra $F = 2^\Omega$ describes the complete information. The cases $\mathbb{P}(\{\omega\}) = 0$ is permitted by the definition, but rarely used since such ω can safely be excluded from the sample space. Therefore, we can define the probability measure $\mathbb{P}$ by simply defining it for all singleton sets $\{\omega\}$.

This may be confusing, since for discrete spaces, it looks like we're assigning probabilities to each $\omega \in \Omega$, but we are actually assigning them to singleton *sets*. We should be writing $\mathbb{P}(\{\omega\})$, but sometimes we abuse notation and write $\mathbb{P}(\omega)$.

Example 1.5 ()

Consider the flip of a fair coin with outcomes either hands or tails. Then, $\Omega = \{H, T\}$. The σ-algebra $F = 2^\Omega$ contains $2^2 = 4$ events:

$$\begin{aligned}\{\} &= \text{Neither heads nor tails}\\ \{H\} &= \text{Heads}\\ \{T\} &= \text{Tails}\\ \{H,T\} &= \text{Either heads or tails}\end{aligned}$$

That is, $\mathcal{F} = \{\{\}, \{H\}, \{T\}, \{H, T\}\}$. Our probability measure $\mathbb{P}$ is defined

$$\mathbb{P}(f) = \begin{cases} 0 & f = \{\} \\ 0.5 & f = \{H\} \\ 0.5 & f = \{T\} \\ 1 & f = \{H, T\} \end{cases} \tag{47}$$

Being able to consider the event space as 2^X is very nice, since countability of X allows us to avoid the Banach-Tarski paradox. It doesn't matter whether $\mathcal{F} = 2^X$ itself is uncountable or not.

Example 1.6 (3 Coin Tosses)

A fair coin is tossed 3 times, creating 8 possible outcomes.

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \tag{48}$$

The complete information is described by the σ-algebra $\mathcal{F} = 2^\Omega = 2^8 = 256$ events, where each of the events is a subset of Ω.
Alice knows the outcome of the second toss only. Thus, her incomplete information is described by the partition

$$\Omega = A_1 \sqcup A_2 = \{HHH, HHT, THH, THT\} \sqcup \{HTH, HTT, TTH, TTT\} \tag{49}$$

and the corresponding σ-algebra is

$$\mathcal{F}_{Alice} = \{\emptyset, A_1, A_2, \Sigma\} \tag{50}$$

Bryan knows only the total number of tails, so his partition contains 4 parts:

$$\begin{aligned} \Omega &= B_0 \sqcup B_1 \sqcup B_2 \sqcup B_3 \\ &= \{HHH\} \sqcup \{HHT, HTH, TTH\} \sqcup \{TTH, THT, HTT\} \sqcup \{TTT\} \end{aligned}$$

When we calculate Bryan's event space, we have

$$\begin{aligned} \mathcal{F}_{Bryan} = \big\{&\emptyset, \{HHH\}, \{HHT\}, \{HTH\}, \{THH\}, \{HHT, HTH\}, \{HHT, THH\}, \\ &\{TTH, THT\}, \{TTH\}, \{THT\}, \{HTT\}, \{TTH, THT\}, \{TTH, HTT\}, \\ &\{THT, HTT\}, \{TTT\}, \Omega\big\} \end{aligned}$$

Note that the event space of Bryan (and Alice) is not merely just 2^Ω since we have some predetermined knowledge of the outcome space Ω. Therefore, we can partition it into 4 cases and construct the event space by putting only the events that are subsets of each partition. For example, it wouldn't make sense to have an event

$$\{HHH, TTT\} \tag{51}$$

since the events $\{HHH\}$ and $\{TTT\}$ are in completely different outcome spaces (given the number of tails). That is, if we knew that 3 tails were thrown, the event $\{HHH, TTT\}$ wouldn't make any sense. However, the event Ω or $\emptyset$ is viable since they describe the case of whether the coin was tossed at all or not. Furthermore, $\mathcal{F}_{Alice}$ and $\mathcal{F}_{Bryan}$ are incomparable. That is, $\mathcal{F}_{Alice} \not\subseteq \mathcal{F}_{Bryan}$ and $\mathcal{F}_{Bryan} \not\subseteq \mathcal{F}_{Alice}$, even though both are subalgebras of 2^Ω.

Example 1.7 (Geometric Measure on $\mathbb{N}$)

Let $\Omega = \mathbb{N}$ and $\mathcal{F} = 2^{\mathbb{N}}$. We can completely define the probability measure by assigning them to singletons $k \in \mathbb{N}$. One such assignment is

$$\mathbb{P}(\{k\}) = \frac{1}{2^k} \tag{52}$$

or more generally,

$$\mathbb{P}(\{k\}) = p(1-p)^{k-1} \tag{53}$$

Example 1.8 (Poisson Measure on $\mathbb{N}_0$)

Let $\Omega = \mathbb{N} \cup \{0\}$. Then, $\mathcal{F} = 2^{\Omega}$ and we can define $\mathbb{P}$ on the singleton sets as

$$\mathbb{P}(\{k\}) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{54}$$

for any $\lambda > 0$. We can then compute the probability of, say all primes, by taking

$$\mathbb{P}(\text{primes}) = \sum_{k \text{ prime}} \mathbb{P}(\{k\}) \tag{55}$$

which we know to be monotonically increasing and bounded above, so it must converge. Whether this has a closed form solution is another matter. Again, in reality we are assigning probability measures on all $\mathcal{F}$-measurable sets, but just doing it through assignment of measure through singleton sets.

Example 1.9 (Voters)

If 100 voters are to be drawn randomly from among all voters in California and asked whom they will vote for governor, then the set of all sequences of 100 Californian voters would be the sample space Ω. We assume that sampling without replacement is used: only sequences of 100 different voters are allowed. For simplicity an ordered sample is considered, that is a sequence $\{Alice, Bryan\}$ is different from $\{Bryan, Alice\}$. We also take for granted that each potential voter knows exactly his/her future choice, that is he/she doesn't choose randomly.
Alice knows only whether or not Arnold Schwarzenegger has received at least 60 votes. Her incomplete information is described by the σ-algebra $\mathcal{F}_{Alice}$ that contains:

1. the set of all sequences in Ω where at least 60 people vote for Schwarzenegger
2. the set of all sequences where fewer than 60 vote for Schwarzenegger
3. the whole sample space Ω
4. the empty set $\emptyset$

Bryan knows the exact number of voters who are going to vote for Schwarzenegger. His incomplete information is described by the corresponding partition $\Omega = B_0 \sqcup B_1 \dots B_{100}$ and the σ-algebra $\mathcal{F}_{Bryan}$ consists of 2^{101} events.
In this case Alice's σ-algebra is a subset of Bryan's: $\mathcal{F}_{Alice} \subset \mathcal{F}_{Bryan}$. Bryan's σ-algebra is in turn a subset of the much larger "complete information" σ-algebra 2^{Ω} consisting of $2^{n(n-1)\dots(n-99)}$ events, where n is the number of all potential voters in California.

Now if we move to uncountable outcome spaces, then things are not as nice, which is why we need to machinery of measure theory to study them. Let us try to model a probability measure on $\Omega = [0, 1]$. It is uncountable, and it turns out that 2^{Ω} has cardinality strictly greater than even the continuum. If we try to model a uniform probability measure $\mathbb{P}$, then for some subset $A \in 2^{\Omega}$, it should be the case that $\mathbb{P}(A) = \mathbb{P}(A \oplus k)$, where $A \oplus k$ is just some translated version of A still contained within $[0, 1]$. This applies to singleton sets, and it turns out that if we try to assign a nonzero probability measure to any singleton

$\{k\}$, then the probability measure of Ω blows up to infinity, which we can't have. So the only thing we can do is have every singleton have zero probability. Remember that a measure by definition has the *countable additivity* property, which says that

$$\mu\left(\bigsqcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k) \tag{56}$$

for all *countable* collections $\{A_k\}$. Summation is not defined for uncountable collections, and so having a probability 0 on every singleton does not imply that the probability of any uncountable set has is 0. That is, having $\mathbb{P}(\{k\}) = 0$ for all $k \in [0,1]$ does not tell you what $\mathbb{P}([0,1])$ is. So now rather than assigning probabilities to singletons, like we did with discrete sets, the approach is to assign probabilities directly to our event space $\mathcal{F}$. We can do this by directly assigning the Lebesgue measure to the Borel algebra of $[0,1]$, which has the properties

1. $\mathbb{P}((a,b)) = \mathbb{P}([a,b)) = \mathbb{P}((a,b]) = \mathbb{P}([a,b]) = b - a$
2. Translation invariance as stated above.

Over uncountable Ω, we cannot afford to work with 2^Ω, since there is an impossibility theorem that says that there is no measure defined on $2^{[0,1]}$ with the two properties above. Therefore, we must work with a smaller σ-algebra. Since the subsets of interest are usually intervals (or more generally, open sets), people usually take the Borel σ-algebra of open intervals on $[0,1]$. The Lebesgue measure on $\mathbb{R}$ is not a probability measure since it $\lambda(\mathbb{R}) = \infty$, but we can construct a uniform probability measure on any bounded set of $\mathbb{R}$. Usually, these continuous probability spaces are $\mathbb{R}^n$, and we define some measure μ directly on its σ-algebra.

Definition 1.9 (Atom)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be uncountable. If for some $\omega \in \Omega$, $\mathbb{P}(\{\omega\}) \neq 0$, then ω is called an **atom**.

Now, given a general (discrete or continuous, or a combination of both) distribution, the set of all the atoms are an at most countable (maybe empty) set whose probability is the sum of probabilities of all atoms (by countable additivity). That is, given $\omega_1, \omega_2, \dots$ atoms,

$$\mathbb{P}\left(\bigsqcup_{i=1}^{\infty} \{\omega_i\}\right) = \sum_{i=1}^{\infty} \mathbb{P}(\{\omega_i\}) \tag{57}$$

1. If this sum is equal to 1 then all other points can be safely excluded from the sample space Ω, returning us to the discrete case.
2. If this sum is 0 then we just have some continuous sample space. This means $\mathbb{P}(\{\omega\}) = 0$ for all $\omega \in \Omega$, and so Ω must be uncountable (since if it was countable, then we should be able to sum the $\mathbb{P}(\{\omega\})$'s to get 1, but it's 0). Remember that summation is only defined for at most countable elements.
3. If the sum of probabilities of all atoms is strictly between 0 and 1, then the probability space decomposes into a discrete, atomic part and a non-atomic, continuous part.

1.2.3 Conditioning on Events

Definition 1.10 (Conditional Probability w.r.t. Events)

Given a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, let B be an event such that $\mathbb{P}(B) > 0$. The **conditional probability** of A given B is defined

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{58}$$

Note that we can't condition on events that have probability 0, which is why we need the $\mathbb{P}(B) > 0$ condition. If this is the case, it doesn't even make sense to talk about a conditional probability $\mathbb{P}(A \mid B)$. For example, if we take the probability space $[0,1]$ with its Borel algebra and the Lebesgue measure, then we cannot

condition something on the rationals, e.g. $\mathbb{P}(\{\omega < 0.5\} \mid \omega \in \mathbb{Q})$ does not make sense. In fact, doing so can lead to contradictions, one being the **Borel-Kolmogorov paradox**.

An extremely useful theorem is that the conditional probability taken as a measure gives us a new viable measure on the same probability space Ω.

Theorem 1.6 ()

Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Then, $\mathbb{P}(\cdot \mid B) : \mathcal{F} \longrightarrow [0,1]$ is a probability measure on $(\Omega, \mathcal{F})$.

Proof.

We prove the properties of a probability measure.

1. The empty set has measure 0.

$$\mathbb{P}(\emptyset \mid B) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = \frac{0}{\mathbb{P}(B)} = 0 \tag{59}$$

2. The entire space has measure 1.

$$\mathbb{P}(\Omega \mid B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1 \tag{60}$$

3. Countable additivity of disjoint events. Let $A_i \in \mathcal{F}$ for $i = 1, 2, \dots$ which are disjoint. Then, their union is in $\mathcal{F}$ by definition of σ-algebra. Now,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \middle| B \right) &= \frac{1}{\mathbb{P}(B)} \mathbb{P}\left[\left(\bigcup_{i=1}^{\infty} A_i \right) \cap B \right] \\
&= \frac{1}{\mathbb{P}(B)} \mathbb{P}\left[\bigcup_{i=1}^{\infty} (A_i \cap B) \right] \\
&= \frac{1}{\mathbb{P}(B)} \sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B) \\
&= \sum_{i=1}^{\infty} \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i \mid B)
\end{aligned}$$

Lemma 1.3 (Law of Total Probability)

Suppose $A_1, A_2, ..., A_n$ is a partition of Ω. Then,

$$\{B \cap A_k\}_{k=1}^{n} \tag{61}$$

is a partition of B, and

$$\mathbb{P}(B) = \sum_{k=1}^{n} \mathbb{P}(B|A_k)\, \mathbb{P}(A_k) \tag{62}$$

This is also called the *Partition rule*.

Theorem 1.7 (Bayes Rule)

Let $A, B \in \mathcal{F}$. Then,

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\,\mathbb{P}(B)}{\mathbb{P}(A)} \tag{63}$$

Proof.

We know that

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ and } \mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(B)} \tag{64}$$

and so we can write

$$\mathbb{P}(A \mid B)\,\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A) \tag{65}$$

1.3 Distributions, Random Variables, Measurable Functions

Random variables are motivated by the following. When you have a random experiment, the experimenter may not be interested in the specific elementary outcomes. So if you have sample space Ω, you may not be concerned about what $\omega \in \Omega$ shows up, but more interested in some numerical function of the elementary outcome. For example, if you toss a coin 10 times, you're not interested in what sequence in $\{0,1\}^{10}$ shows up, but you may want to just know how many heads came up. In other words, your interest defines a numerical function $X : \Omega \to \mathbb{R}$. This is useful, since in many cases the sample space Ω can be extremely complicated (e.g. the sample space of all weather conditions) and the elementary outcomes also complicated, so you may want to know some simpler aspect (e.g. the temperature).

The name "random variable" is very misleading. It's not random nor a variable. It is a deterministic function $X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow \mathbb{R}$ that assigns numbers to outcomes. The only source of randomness itself is which $\omega \in \Omega$ is chosen. But we can't just choose any function on Ω; they must satisfy the nice property of measurability. Now, to talk about random variables, recall that the definition of a measurable function $f : (X, \mathcal{A}) \longrightarrow \mathbb{R}$ is one where the preimage of every Borel set $B \in \mathcal{R}$ is in $\mathcal{A}$. With a potential measure μ, this allows us to define the Lebesgue integral of f. Note that this is also equivalent to the more easily provable fact that the preimage of every half-interval $(-\infty, t)$ is in $\mathcal{A}$. That is, $f^{-1}((-\infty, t]) \in \mathcal{A}$ for all $t \in \mathbb{R}$.

Definition 1.11 (Random Variable)

A **random variable** X on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is an $\mathcal{F}$-measurable function $X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow \mathbb{R}$. That is, for every subset $B \in \mathcal{R}$, its preimage

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F} \tag{66}$$

The reason we want X to be $\mathcal{F}$-measurable is because now we can define probabilities on Borel sets B of $\mathbb{R}$ by computing the probabilities of the preimage of B, which must be $\mathcal{F}$-measurable. In a way, a random variable "pushes forward" the probability measure $\mathbb{P}$, originally defined on $\mathcal{F}$, to $\mathcal{R}$.

Definition 1.12 (Probability Law of Random Variable X)

Let X be a random variable on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The **probability law of** X is a function $\mathbb{P}_X : \mathcal{R} \longrightarrow [0,1]$ defined, for each Borel set B of $\mathbb{R}$, as

$$\mathbb{P}_X(B) \coloneqq \mathbb{P}\big(X^{-1}(B)\big) = \mathbb{P}\big(\{\omega \in \Omega \mid X(\omega) \in B\}\big) \tag{67}$$

Note that $\mathbb{P}$ refers to the probability measure on $\mathcal{F}$, and $\mathbb{P}_X$ refers to the probability law on $\mathcal{R}$. In shorthand, we can write $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$. By abuse of notation, it is generally written

$$\mathbb{P}(X \in B) \tag{68}$$

It is important to get used to this notation. Whenever we write $\mathbb{P}(X \dots)$, we are always working in the probability law of X. Furthermore, whatever condition we put within the parentheses describes a measurable set. For example,

1. $\mathbb{P}(X = x)$ describes the probability law of X evaluated on the set $\{x\}$.
2. $\mathbb{P}(X \leq x)$ describes the probability law of X evaluated on the set $(-\infty, x]$.
3. $\mathbb{P}(Y \leq y)$ describes the probability law of Y evaluated on the set $(-\infty, y]$.
4. $\mathbb{P}(a \leq Y < b)$ describes the probability law of Y evaluated on the set $[a, b)$.
5. $\mathbb{P}(Z \in \mathbb{Q})$ describes the probability law of Z evaluated on the set $\mathbb{Q}$.

Theorem 1.8 (σ-Algebra Generated by Random Variable X)

Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X, $\mathbb{P}_X$ is a probability measure on $(\mathbb{R}, \mathcal{R})$. Now, it turns out that the collection of all preimages of Borel sets under X forms a σ-algebra on Ω. We call it

$$\sigma(X) \coloneqq \left\{ A \subset \Omega \mid A = X^{-1}(B) \text{ for some } B \in \mathcal{R} \right\} \tag{69}$$

which is a σ-algebra of Ω. Since X is a measurable function, every $X^{-1}(B)$ is $\mathcal{F}$-measurable, and so $\sigma(X)$ is a sub-σ-algebra of $\mathcal{F}$. It is never the case that $\sigma(X) \supset \mathcal{F}$, since that means that X itself is not $\mathcal{F}$-measurable.

Theorem 1.9 ()

Given $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, let us define a probability law $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$. Then,

$$(\mathbb{R}, \mathcal{R}, \mathbb{P}_X) \tag{70}$$

is a probability space.

This theorem is extremely useful, since in practical applications, one does not consider an abstract Ω and works immediately in $\mathbb{R}$. Once we have determined our numerical values of interest (heads or tails, number of heads, sum of dice rolls) with our random variable X, we can just throw away $(\Omega, \mathcal{F}, \mathbb{P})$ and work directly in probability space $(\mathbb{R}, \mathcal{R}, \mathbb{P}_X)$. Therefore, we don't actually control Ω by explicitly defining it as we said before.

We could just leave Ω to be some arbitrary large set, and construct an appropriate random variable X that will generate an appropriate σ-algebra $\sigma(X)$ that captures the information of the experiment. This allows us to "simplify" the σ-algebra $\mathcal{F}$ to the scope of the random variable. That is, let Ω be the sample space of all trajectories of a coin flip before it comes to rest. If we are just looking at whether it is heads or tails, we can define X to have image $\{0, 1\}$. Then, $\sigma(X)$ will be a sub-σ-algebra of $\mathcal{F}$ that looks at only the four subsets $\emptyset, \Omega$, the set of all trajectories landing heads, and the set of all trajectories landing tails. This simplifies $\mathcal{F}$ to a scope that we are interested in.

Let us review once more on the hierarchy of random variables. We usually classify random variables X by the smallest σ-algebra that they generate, which is $\sigma(X)$. That is, not only is X $\sigma(X)$-measurable, but for all σ-algebras $\mathcal{G}$ s.t. $\sigma(X) \subset \mathcal{G} \subset \mathcal{F}$, X is also $\mathcal{G}$-measurable. Remember, since this is the case, the only relevant measure on these random variables is how coarse/small $\sigma(X)$ is.

1. The finest random variable has $\sigma(X) = \mathcal{F}$.
2. The coarsest random variable is a constant random variable, which has $\sigma(X)$ to be the trivial σ-algebra $\mathcal{H} = \{\emptyset, \Omega\}$. Note that a constant random variable is still $\mathcal{F}$-measurable.
3. Every other random variable X has $\mathcal{H} \subset \sigma(X) \subset \mathcal{F}$.

1.3.1 Cumulative Distribution Function

Now, remember that the Borel algebra $\mathcal{R}$ is generated by the semi-infinite intervals of form $(-\infty, t]$ (for all $t \in \mathbb{R}$), which are considered "nice" Borel sets. So, $\mathbb{P}_X((-\infty, t])$ is well defined for all $t \in \mathbb{R}$. In fact, this has a name, and when we talk about the "distribution" of some random variable, we refer to the CDF.

Definition 1.13 (Cumulative Distribution Function)

Given $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$. Then, the **cumulative distribution function** of X is defined

$$F_X(x) = \mathbb{P}\big(\{\omega \in \Omega \mid X(\omega) \leq x\}\big) \tag{71}$$

We can also define this with the probability law $\mathbb{P}_X$ as

$$F_X(x) = \mathbb{P}_X\big((-\infty, x]\big) \tag{72}$$

By abuse of notation, we will write the CDF as $P(X \leq x)$. It satisfies the properties:

1. Limits:
$$\lim_{x \to -\infty} F_X(x) = 0 \text{ and } \lim_{x \to \infty} F_X(x) = 1 \tag{73}$$
2. Monotonicity:
$$x \leq y \implies F_X(x) \leq F_X(y) \tag{74}$$
3. Right-continuity: For all $x \in \mathbb{R}$
$$\lim_{\epsilon \to 0^+} F_X(x + \epsilon) = F_X(x) \tag{75}$$

So, if there are jumps, the hole can exist as the function approaches a value from the left. What is remarkable is that any function satisfying these three properties satisfies these 3 properties gives you a viable CDF (and as shown below, completely determines a unique random variable).

So if you give me the probability law for all Borel sets of $\mathbb{R}$, then I can easily define the CDF since $(-\infty, x]$ are also Borel sets. It turns out that if we know *just* the CDF, then since the semi-infinite intervals form a generating class of $\mathcal{R}$, it turns out that we can completely define $\mathbb{P}_X$. The proof of the theorem below is a bit more involved, using π-systems, but it is good to know.

Theorem 1.10 ()

The CDF $F_X(\cdot)$ uniquely specifies the probability law $\mathbb{P}_X$ for any random variable X.

To summarize, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable just pushes a measure onto the measure space $(\mathbb{R}, \mathcal{R})$. If we only care about the values of the random variable, then we can forget about Ω and only look at $(\mathbb{R}, \mathcal{R}, \mathbb{P}_X)$. The CDF on $(\mathbb{R}, \mathcal{R}, \mathbb{P}_X)$ will be well defined since semi-finite intervals are also Borel. If I am just given a CDF $F_X(\cdot)$, then this is enough for me to specify a unique probability measure $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{R})$. So although $\mathbb{P}_X$ contains the complete description of the random variable X, in practice we will use F_X since it also captures all the information of X and it's easier to work with.

1.3.2 Types of Random Variables

You classify random variables based on the nature of the measure $\mathbb{P}_X$ induced on the real line. Note that we can have a continuous probability space Ω with a discrete random variable X (e.g. coin flips). There are only three fundamental types of measures: discrete, continuous and singular random variables. In fact, a result in measure theory called *Lebesgue's Decomposition Theorem* says that every measure on $\mathbb{R}$ are either one of these 3 or mixtures thereof. We are used to the first two; the third one is very bizzare and has little to no practical applications.

Note that if we are working in a discrete probability space Ω, then we can simply take the σ-algebra to be

2^Ω, and so we can take any function on Ω as a random variable since its preimage will always be in 2^Ω.

Definition 1.14 (Discrete Random Variable)

Given $(\Omega, \mathcal{F}, \mathbb{P})$, let us have a random variable X that induces a probability law on $(\mathbb{R}, \mathcal{R})$. X is said to be **discrete** if there exists a countable set $E \subset \mathbb{R}$ s.t. $\mathbb{P}_X(E) = 1$ (i.e. E's preimage has probability measure 1). Since E is at most countable, we can enumerate it $E = \{e_1, e_2, \dots\}$, and by countable additivity of disjoint sets, we have

$$1 = \mathbb{P}_X(E) = P\left(\bigcup_{i=1}^{\infty} \{e_i\} \right) = \sum_{i=1}^{\infty} \mathbb{P}_X(\{e_i\}) = \sum_{i=1}^{\infty} P(X = e_i) \tag{76}$$

and for any $B \in \mathcal{R}$,

$$\mathbb{P}_X(B) = \sum_{x \in E \cap B} P(X = x) \tag{77}$$

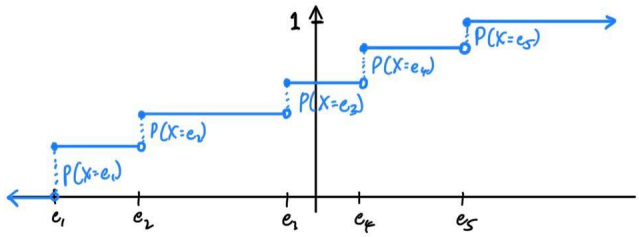
Therefore, the entire probability measure is determined by the probabilities of the singleton sets $P(X = e_i)$. Therefore, the function

$$p_X(x) \coloneqq P(X = x) \tag{78}$$

is called the **probability mass function** of X, and we can compute using the Lebesgue integral, which reduces to the summation:

$$\mathbb{P}_X(B) = \int_B p_X(x)\, d\mathbb{P}_X = \sum_{x \in E \cap B} p_X(x) \tag{79}$$

Sometimes, the definition of discrete X involves having a countable image in $\mathbb{R}$, but our definition allows us to have some $B \in \mathcal{R}$ where its preimage is not necessarily the sample space Ω, but a smaller subset of measure 1. What's nice about the discrete random variable is that the probability mass function p_X completely describes its probability law. The CDF of a discrete probability function will look like an increasing series of steps. If we have $E = \{e_1, e_2, e_3, e_4, e_5\}$, its CDF would look like:



If E was countable, then it would have countably infinite discontinuities. Now we'll give some examples of discrete random variables, and in here we'll completely ignore the sample space Ω, since once we have a random variable X, we can just work in $(\mathbb{R}, \mathcal{R}, \mathbb{P}_X)$. Remember that we will write $P(X = x)$ as shorthand for $\mathbb{P}_X(\{x\})$.

Definition 1.15 (Indicator/Bernoulli Random Variable)

Given $(\Omega, \mathcal{F}, \mathbb{P})$, let $A \in \mathcal{F}$ be an event. A useful random variable is the **indicator random variable** $1_A : \Omega \longrightarrow \mathbb{R}$ defined

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \tag{80}$$

This is a random variable since the preimages of $\emptyset, \{0\}.\{1\}, \{0, 1\}$ are $\emptyset, A^c, A, \Omega$, which are all $\mathcal{F}$-measurable. Since the probability measure of A is $\mathbb{P}(A) = p$, then $\mathbb{P}(A^c) = 1 - \mathbb{P}(A) = 1 - p$, and so

we get the PMF

$$p_{1_A}(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases} \tag{81}$$

The CDF of this function will look like a step function

$$F_{1_A}(x) = \begin{cases} 0 & \text{if } x < 0 \\ P(A^c) & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x \end{cases} \tag{82}$$

Example 1.10 (Uniform Random Variable)

Given a finite set $E = \{e_i\}_{i=1}^n \subset \mathbb{R}$, we define the PMF as

$$p_X(e_i) = \mathbb{P}(X = e_i) = \frac{1}{n} \; \forall i = 1, 2, \dots n \tag{83}$$

which induces the probability measure $\mathbb{P}_X(B) = \sum_{x \in E \cap B} p_X(x)$.

The Bernoulli RV leads to the geometric and binomial random variables.

Example 1.11 (Geometric Random Variable)

Given $E = \mathbb{N}$, we can define the PMF associated with random variable $X \sim \text{Geometric}(p)$ as

$$p_X(k) = \mathbb{P}(X = k) = (1-p)^{k-1}p \text{ for } k \in \mathbb{N}, \; p \in [0, 1] \tag{84}$$

which induces the probability measure $\mathbb{P}_X(B) = \sum_{x \in E \cap B} p_X(x)$. We can interpret this as the number of times you have to (independently) toss a p-coin (probability of heads is p) until you get a heads.

Example 1.12 (Binomial Random Variable)

We let $E = \mathbb{N}_0$ and define the PMF associated with random variable $X \sim \text{Binomial}(n, p)$ as

$$p_X(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k \in E, p \in [0, 1] \tag{85}$$

We can interpret this as the number of heads occurring in a sequence of n independent tosses of a p-coin.

Example 1.13 (Poisson Random Variable)

We let $E = \mathbb{N}_0$ and define the PMF of $X \sim \text{Poisson}(\lambda)$ as

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k \in E, \; \lambda > 0 \tag{86}$$

Definition 1.16 (Negative Binomial Distribution)

The negative binomial distribution, denoted $\text{NB}(r, p)$ is defined as

$$\mathbb{P}(X = x) \equiv \binom{k+r-1}{k} (1-p)^r p^k \tag{87}$$

It can be interpreted as the distribution that models the number of successes in a sequence of iid Bernoulli-p trials before a specified number r failures occurs.

A slight generalization of a discrete random variable is a simple random variable. Recall that the indicator random variable is a function $1_A : \Omega \to \mathbb{R}$ defined

$$1_A(\omega) \coloneqq \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if else} \end{cases} \tag{88}$$

As simple random variable generalizes this into multiple sets that form a partition of Ω. It is analogous to a simple function, introduced in measure theory.

Definition 1.17 (Simple Random Variable)

Let $\{A_i\}_i$ form a partition of probability space Ω. A **simple random variable** X is a random variable of the form

$$X(\omega) = \sum_i a_i 1_{A_i}(\omega) \tag{89}$$

that assigns value a_i if the input $\omega \in A_i$.

Now, let's move on to continuous random variables.

Definition 1.18 (Absolutely Continuous Measures)

Let μ, ν be measures defined on $(\Omega, \mathcal{F})$. We say that ν is **absolutely continuous** w.r.t. μ if for every $N \in \mathcal{F}$ s.t. $\mu(N) = 0$, we have $\nu(N) = 0$.

Definition 1.19 (Continuous Random Variable)

A random variable X is **continuous** if its induced measure $\mathbb{P}_X : (\mathbb{R}, \mathcal{R}) \to [0, 1]$ is absolutely continuous w.r.t. the Lebesgue measure $\lambda : (\mathbb{R}, \mathcal{R}) \to \mathbb{R}$, i.e. if for every Borel set N of Lebesgue measure 0, we have $\mathbb{P}_X(N) = 0$ also.

A common misconception is that a random variable X is continuous if the induced measure on every singleton set in $\mathcal{B}(R)$ is 0, i.e. $\mathbb{P}_X(\{x\}) = 0$ for all $x \in \mathbb{R}$. The definition above implies this since the Lebesgue measure of every singleton set is 0.

We introduce a theorem that is useful to know, but we won't prove it.

Theorem 1.11 (Radon-Nikodym Theorem (Special Case))

Let X be a continuous random variable. Then, there exists a nonnegative measurable function $f_X : \mathbb{R} \longrightarrow [0, \infty)$ s.t. for any $B \in \mathcal{R}$, we have

$$\mathbb{P}_X(B) = \int_B f_X \, d\lambda \tag{90}$$

where the above is the Lebesgue integral. Note that we must define using the Lebesgue integral because Riemann integral is not compatible with any Borel set. f_X is called the **probability den-**

sity function, aka **PDF**. Furthermore, we can get f_X from $\mathbb{P}_X$ by taking the **Radon-Nikodym derivative** (which we will not define now)

$$f_X = \frac{d\mathbb{P}_X}{d\lambda} \tag{91}$$

which basically says that if we have a set of very small Lebesgue measure $d\lambda$ tending to 0, then its probability measure $\mathbb{P}_X$ will also be very small, and the infinitesimal ratio of these two measures on an arbitrarily small set is f_X. Also, note that the integral does not change if the value of f changes on sets of Lebesgue measure 0, and so there is no unique PDF describing $\mathbb{P}_X$. It is unique up to sets of Lebesgue measure 0, so when we refer to such a PDF f_X, we are really talking about an equivalence class of functions.

This theorem guarantees the existence of some f_X that completely describes the probability law P_X! Take a special case of when $B = (-\infty, x]$), and we can define the CDF as

$$F_X(x) = P_X((-\infty, x]) = \int_{(-\infty, x]} f_X \, d\lambda \tag{92}$$

If the set of integration is an interval (and the function is continuous a.e.), then the Lebesgue integral and Riemann integral coincides, and we get the familiar formula

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt \tag{93}$$

and we can differentiate it to get back the PDF f_X (or more accurately, some function that agrees with f_X a.e.). We can show that the CDF of a continuous random variable X

1. is absolutely continuous, and
2. is differentiable almost everywhere, which means that its PDF will be defined almost everywhere (and we can fill in the undefined points however we want).

Note that the PDF f_X itself has no interpretation as a probability (indeed, we can change its value at a countable number of points to anything we want). It is only when we integrate it over some Borel set that gives us a probability.

Example 1.14 (Uniform Random Variable)

Let us define the uniform probability measure P_X on $(\mathbb{R}, \mathcal{R})$ with the CDF

$$F_X = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } 1 < x \end{cases} \tag{94}$$

It is differentiable almost everywhere except for at the two points $x = 0$ and $x = 1$. Therefore, the PDF f_X is defined for all real numbers except $x = 0$ and $x = 1$. But it doesn't matter: we can assign any value f_X we want on 0 and 1 since it won't affect the integral of it. In this example, we just set

$$f_X = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if else} \end{cases} \tag{95}$$

Example 1.15 (Exponential Random Variable)

The exponential random variable has the following CDF:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \text{ for } \lambda > 0 \tag{96}$$

which is differentiable everywhere except at $x = 0$. Differentiating it (and assigning a convenient value at $x = 0$ $f(0) = \lambda$) gives the PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if else} \end{cases} \tag{97}$$

Example 1.16 (Gaussian Random Variable)

The PDF is easier to specify for the Gaussian, so we define the Gaussian RV as having PDF

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \text{ for } \mu \in \mathbb{R}, \sigma > 0 \tag{98}$$

Note that this PDF decreases very quickly as we get further from μ. The CDF cannot be written in closed form, and we call the CDF of the standard Gaussian the **error function**:

$$\text{Erf}(x) = F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \tag{99}$$

Example 1.17 (Cauchy Random Variable (Standardized))

The Cauchy random variable gives the PDF

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2} \text{ for } x \in \mathbb{R} \tag{100}$$

Integrating this gives the inverse tangent, which after scaling it down by π satisfies the conditions of the CDF. Note that the Cauchy distribution falls off much more slowly around the mean (at a rate of $\frac{1}{1+x^2}$, like a power law) than the Gaussian (which is even *faster* than an exponential, it is at the rate of e^{-x^2}). If such a PDF falls off at a slow rate, like a power law, then this is called a *heavy-tailed random variable*.

Example 1.18 (Gamma Random Variable)

The PDF associated with random variable $X \sim \text{Gamma}(n, \lambda)$ is defined

$$f_X(x) = \frac{\lambda^n x^{n-1}}{\Gamma(n)} e^{-\lambda x} \text{ for } x \geq 0 \tag{101}$$

where Γ is the gamma function, which is an extension of the factorial function to the domain of complex numbers.

$$\Gamma(x) \coloneqq \int_0^\infty z^{x-1} e^{-z}\, dz, \qquad \text{Re}(x) > 0 \tag{102}$$

Example 1.19 (Beta Random Variable)

The PDF associated with random variable $X \sim \text{Beta}(\alpha, \beta)$, for positive reals α, β, is defined

$$f_X(x) \equiv \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \text{ where } B(\alpha,\beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{103}$$

and Γ is the Gamma function.

Example 1.20 (Uniform RV defined on Cantor Set)

The cantor set $C \subset [0,1]$ is defined by removing $(1/3, 2/3)$ from $[0,1]$ and then removing the middle third from each interval that remains. We define the distribution on this set by defining its CDF: We set

1. $F(x) = 0$ for $x \leq 0$ and $F(x) = 1$ for $x \geq 1$.
2. $F(x) = 1/2$ for $x \in [1/3, 2/3]$,
3. $F(x) = 1/4$ for $x \in [1/9, 2/9]$ and $F(x) = 3/4$ for $x \in [7/9, 8/9]$, ...

and extend F to all of $[0,1]$ using monotonicity.

Example 1.21 (Dense Discontinuities)

Let $q_1, q_2, \dots$ be an enumeration of the rationals. Let $\alpha_i > 0$ have $\sum_{i=1}^{\infty} \alpha_i = 1$, and let

$$F(x) = \sum_{i=1}^{\infty} \alpha_i 1_{[q_i,\infty)}(x) \tag{104}$$

where $1_{[q_i,\infty)}(x) = 1$ if $x \in [q_i, \infty)$ and 0 if otherwise.

To summarize, once we have a random variable $X : \Omega \to \mathbb{R}$, we can throw away the sample space and work in $(\mathbb{R}, \mathcal{R}, \mathbb{P}_X)$ with the induced measure $\mathbb{P}_X$, which is known as the **probability distribution** of X.

1. If X is discrete, then let there be some at most countable set $E = \{e_i\}$ where $P(E) = 1$. it turns out that $\mathbb{P}_X$ can be completely defined by a probability mass function $p_X : \mathbb{R} \to \mathbb{R}$ defined

$$p_X(x) = \mathbb{P}_X(\{x\}). \tag{105}$$

   Given that we have this PMF , we can define $\mathbb{P}_X$ as such: Given any Borel $B \in \mathcal{R}$,

$$\mathbb{P}_X(B) = \sum_{x \in E \cap B} p_X(x) \tag{106}$$

2. If X is continuous, then the Radon-Nikodym Theorem asserts the existence of a nonnegative probability density function f_X that completely describes the probability law $\mathbb{P}_X$. Given that we have this PDF, we can then define $\mathbb{P}_X$ as such: Given any Borel $B \in \mathcal{R}$,

$$\mathbb{P}_X(B) = \int_B f_X \, d\lambda \tag{107}$$

1.3.3 Space of Measurable Functions

Now it turns out that the space of $\mathcal{F}$-measurable functions $X : \Omega \to \mathbb{R}$ forms a function space, which means that the set of all random variables on Ω forms a vector space. We formally show it here.

Lemma 1.4 ()

The set of all $\mathcal{F}$-measurable functions $X : (\Omega, \mathcal{F}) \to \mathbb{R}$ forms a vector space, denoted $L_{\mathcal{F}}(\Omega; \mathbb{R})$, or $L_{\mathcal{F}}(\Omega)$ for short.

Proof.

Naturally, we can put the L^p-norm on this space, defined

$$||X||_p := \left(\int_\Omega |X|^p \, d\mathbb{P} \right)^{1/p} \tag{108}$$

Moreover, if $p = 2$, then we can put an inner product defined

$$\langle X, Y \rangle = \left(\int_\Omega XY \, d\mathbb{P} \right)^{1/2} \tag{109}$$

Definition 1.20 ()

The Banach space of $\mathcal{F}$-measurable functions is denoted $L^p_{\mathcal{F}}(\Omega)$, and the Hilbert space is denoted $L^2_{\mathcal{F}}(\Omega)$.

This means that if we have some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, then any $\mathcal{G}$-measurable function is also $\mathcal{F}$-measurable, since if the preimage of every $B \in \mathcal{R}$ is in $\mathcal{G}$, then it $B \in \mathcal{F}$. This immediately results in the following.

Theorem 1.12 ()

If $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$, then $L_{\mathcal{G}}(\Omega)$ is a subspace of $L_{\mathcal{F}}(\Omega)$.

This means that as we get coarser and coarser random variables, the space in which these random variables live in get smaller and smaller, until we get to the constant random variables, which form a 1-dimensional line in $L_{\mathcal{F}}(\Omega)$. The origin is simply the constant 0 random variable.

1.4 Independence

Definition 1.21 (Independence of 2 Events)

Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, events $A, B \in \mathcal{F}$ are said to be **independent under** $\mathbb{P}$ if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \, \mathbb{P}(B) \tag{110}$$

This leads to the immediate property that if $\mathbb{P}(B) > 0$, with A, B independent, then

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) \tag{111}$$

Note that A and B may be independent under one measure, but not under another measure. The property that $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ is *not* the definition of independence, since it has the more restricting property that $\mathbb{P}(B) > 0$, so only refer to the definition that $\mathbb{P}(A \cap B) = \mathbb{P}(A) \, \mathbb{P}(B)$. This is the true definition of independent events that we should rely on, not the one that says that A and B are independent if "one does not affect the other." This old definition is misleading and false. For example, take the probability space $[0, 1]$, with Borel σ-algebra, and Lebesgue measure $\mathbb{P} = \lambda$, and let $A = \mathbb{Q}$ and $B = \mathbb{R} \setminus \mathbb{Q}$. Then, contradictory to our

old definition, A and B are independent since $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) = 0$! By the definition, an event A is independent of itself if $\mathbb{P}(A) = 0$ or 1 (e.g. A is rationals, irrationals, cantor set, $\emptyset$, Ω, etc.).

Definition 1.22 (Independence of n Events)

Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$,

1. Let us have a finite collection of events $A_1, A_2, \dots, A_n \in \mathcal{F}$. They are **independent** if for all nonempty $I_0 \subset \{1, 2, \dots n\}$,
$$\mathbb{P}\left(\bigcap_{i \in I_0} A_i \right) = \prod_{i \in I_0} \mathbb{P}(A_i) \tag{112}$$
Note that it is not enough to just prove that
$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \prod_{i=1}^{n} \mathbb{P}(A_i) \tag{113}$$
We must verify this for all 2^n possible choices (to be precise, we don't need to prove for $I_0 = \emptyset$ and $I_0 = \{A_i\}$), so for $2^n - n - 1$ choices.
2. Let $\{A_i\}_{i \in I}$ be a collection of events indexed by a possibly uncountable I. They are independent if for all nonempty and finite $I_0 \subset I$, we have
$$\mathbb{P}\left(\bigcap_{i \in I_0} A_i \right) = \prod_{i \in I_0} \mathbb{P}(A_i) \tag{114}$$

Now when we are trying to compare two σ-algebras, the measure defined for one may not even be defined on the other. To ensure that a measure is defined on both, it makes sense to take its σ-algebra and construct two sub-σ-algebras, which μ is guaranteed to be defined on.

Definition 1.23 (Independence of σ-Algebras)

Let us have probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. Let $\mathcal{F}_1, \mathcal{F}_2$ be two sub-σ-algebras of $\mathcal{F}$. $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent if for any $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$, A_1 and A_2 are independent.
2. Let $\{\mathcal{F}_i\}_{i \in I}$ be an arbitrary collection of sub-σ-algebras of $\mathcal{F}$, indexed by possibly uncountable I. Then, they are independent if for any choices of $A_i \in \mathcal{F}_i$ for $i \in I$, $\{A_i\}_{i \in I}$ are independent events.

Definition 1.24 (Independent Random Variables)

Two random variables X, Y are **independent** if $\sigma(X)$ and $\sigma(Y)$ are independent σ-algebras. That is, for any Borel sets $B_1, B_2 \in \mathcal{R}$, the events $X^{-1}(B_1)$ and $Y^{-1}(B_2)$ are independent:
$$\mathbb{P}\big[X^{-1}(B_1) \cap Y^{-1}(B_2)\big] = \mathbb{P}(X^{-1}(B_1))\,\mathbb{P}(Y^{-1}(B_2)) \tag{115}$$
or by abusing notation,
$$\mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}(X \in B_1)\,\mathbb{P}(Y \in B_2) \tag{116}$$

If X, Y are independent, then we can say something about the CDFs
$$F_{X,Y}(x, y) = F_X(x)\,F_Y(y) \tag{117}$$

In fact, we can say something stronger.

Theorem 1.13 ()

X and Y are independent RVs if and only if

$$F_{X,Y}(x,y) = F_X(x)\,F_Y(y) \tag{118}$$

Moving onto multiple variables, we can define that $X_1, X_2, \dots, X_n$ are independent RVs if $\sigma(X_1), \dots, \sigma(X_n)$ are independent σ-algebras.

1.5 Functions of Random Variables

In many applications, it happens that we are interested not in the value of the random variable X, but a function of it. That is, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let us have a random variable $X : \Omega \to \mathbb{R}$. We can then define another function $f : \mathbb{R} \to \mathbb{R}$ and consider the potential random variable $f \circ X : \Omega \to \mathbb{R}$. We say potential because we don't know yet whether $f \circ X$ is measurable (i.e. the preimage of every Borel set in $\mathbb{R}$ is in $\mathcal{F}$). This condition suffices if f itself is a measurable function, i.e. for every Borel set $B \in \mathcal{R}$, its preimage $f^{-1}(B)$ is Borel in $\mathbb{R}$, and by measurablility of X, its preimage under X is $\mathcal{F}$-measurable, making $f \circ X$ a viable random variable. With this new random variable $f \circ X$, we would now like to answer the question: What is the probability law $\mathbb{P}_{f\circ X}$ of $\mathbb{R}$?

This also works for joint random variables, which we will learn later. Given a joint random variable $(X_1, X_2, \dots X_n) : \Omega \to \mathbb{R}^n$, we can define a measurable function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and define the scalar random variable $f \circ (X_1, \dots X_n)$ on Ω. But again, we want to find what the CDF of this composition.

1.5.1 Maximum/Minimum of Random Variables

Let $X_1, X_2, \dots, X_n$ be random variables of $(\Omega, \mathcal{F}, \mathbb{P})$ with joint CDF $F_{X_1 \dots X_n}(x_1, \dots, x_n)$. Let $Y_n = \min(X_1, \dots, X_n)$ and $Z_n = \max(X_1, \dots, X_n)$. Note that Y_n and Z_n are also functions of Ω to $\mathbb{R}$. To prove that they are random variables, we just have to prove that min and max are measurable functions from $\mathbb{R}^n$ to $\mathbb{R}$, which we can do by proving that the preimage of all semi-infinite interval $(-\infty, x]$ are Borel in $\mathbb{R}^n$.

1. The preimage of $(-\infty, x]$ under max is just the set of all n-vectors whose max is less than x, which is just the semi-infinite cuboid $(-\infty, x]^n \subset \mathbb{R}^n$, which is Borel in $\mathbb{R}^n$.
2. The preimage of $(-\infty, x]$ under min is the set of all n-vectors whose min is less than x, i.e. at least one element must be less than x. But this is just the complement of all vectors that have elements all greater than x, which is $\mathbb{R}^n \setminus (x, +\infty)^n \subset \mathbb{R}^n$, which is Borel in $\mathbb{R}^n$.

Now we must determine the CDF of Y_n and Z_n.

1. We have

$$\begin{aligned} F_{Z_n}(z) &= \mathbb{P}(\{\omega \mid Z_n(\omega) \le z\}) \\ &= \mathbb{P}(\{\omega \mid X_1(\omega) \le z, \dots, X_n(\omega) \le z\}) \\ &= F_{X_1 \dots X_n}(z, \dots, z) \end{aligned}$$

   where the last equality is describes simply the joint CDF of the joint distribution $(X_1, \dots, X_n)$. If we assume independence of X_i's, it simplifies out to

$$\prod_i F_{X_i}(z) \tag{119}$$

   and if iid, then we have $[F_X(z)]^n$, where X is the common distribution.

2. For Y_n, we work with complements again and have

$$\begin{aligned} F_{Y_n}(y) &= \mathbb{P}(\{\omega \mid Y_n(\omega) \leq y\}) \\ &= 1 - \mathbb{P}(\{\omega \mid Y_n(\omega) > y\}) \\ &= 1 - \mathbb{P}(\{\omega \mid X_1(\omega) > y, \dots X_n(\omega) > y\}) \end{aligned}$$

where $\mathbb{P}(\{\omega \mid X_1(\omega) > y, \dots X_n > y\})$ can be calculated from the joint distribution. If we assume independence of X_i, it simplifies out to

$$1 - \prod_i \mathbb{P}(\{\omega \mid X_i(\omega) > y\}) = 1 - \prod_i \big(1 - F_{X_i}(y)\big) \tag{120}$$

and if iid, then we have $1 - [1 - F_X(y)]^n$.

Example 1.22 (Uniforms)

Let X_1, X_2 be iid distributed as Uniform$[0,1]$, and let $Z = \max(X_1, X_2)$ with $Y = \min(X_1, X_2)$, i.e. Z is the greater of the two and Y is the lesser. We would expect the PDF of Z to have more mass towards 1 and the PDF of Y to have more mass towards 0. Our common CDF is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } 1 < x \end{cases} \tag{121}$$

Let's calculate the CDF of Z.

$$\begin{aligned} F_Z(z) &= \mathbb{P}(\{\omega \mid Z(\omega) \leq z\}) \\ &= \mathbb{P}(\{\omega \mid X_1(\omega) \leq z, X_2(\omega) \leq z\}) \\ &= F_{X_1,X_2}(z,z) \\ &= [F_X(z)]^2 = \begin{cases} 0 & \text{if } x < 0 \\ x^2 & \text{if } x \in [0,1] \\ 1 & \text{if } 1 < x \end{cases} \end{aligned}$$

This CDF is differentiable everywhere except the two points 0 and 1, so we can get the PDF to be $f_Z(z) = 2z$ for $z \in (0,1)$ and 0 otherwise. For the values of f_Z at 0 and 1, we can fill it in with anything we want (since the measure of these sets are 0), so we will just defined $f_Z(0) = 0$ and $f_Z(1) = 2$, getting

$$f_Z(z) = \begin{cases} 2z & \text{if } z \in [0,1] \\ 0 & \text{if else} \end{cases} \tag{122}$$

Let's calculate the CDF of Y.

$$\begin{aligned} F_Y(y) &= \mathbb{P}(\{\omega \mid Y(\omega) \leq y\}) \\ &= 1 - \mathbb{P}(\{\omega \mid Y(\omega) > y\}) \\ &= 1 - \mathbb{P}(\{\omega \mid X_1(\omega) > y, X_2(\omega) > y\}) \\ &= 1 - \mathbb{P}(\{\omega \mid X_1(\omega) > y\})\, \mathbb{P}(\{X_2(\omega) > y\}) \\ &= 1 - [1 - F_X(y)]^2 = \begin{cases} 0 & \text{if } y < 0 \\ 1 - (1-y)^2 & \text{if } y \in [0,1] \\ 1 & \text{if } y > 1 \end{cases} \end{aligned}$$

and differentiating it (with setting any values of the PDF at the nondifferentiable points 0 and 1)

gives

$$f_Y(y) = \begin{cases} 2-2y & \text{if } y \in [0,1] \\ 0 & \text{if else} \end{cases} \tag{123}$$

Example 1.23 (Exponentials)

Let $X_1, X_2, \dots, X_n$ be independent exponential random variables with parameters $\lambda_1, \dots, \lambda_n$, respectively (not identical!). Then, for each X_i, its CDF is

$$F_{X_i}(x) = 1 - e^{-\lambda_i x} \text{ for } x \geq 0 \tag{124}$$

and let $Y = \min(X_1, \dots, X_n)$. Then, we have

$$\begin{aligned} F_Y(y) &= 1 - \prod_{i=1}^{n} [1 - F_{X_i}(y)] \\ &= 1 - \prod_{i=1}^{n} e^{-\lambda_i x} \\ &= 1 - e^{-(\sum_{i=1}^{n} \lambda_i)x} \end{aligned}$$

which is the CDF of an exponential distribution. So,

$$Y \sim \text{Exponential}(\lambda_1 + \ldots + \lambda_n) \tag{125}$$

This is nice, since the minimum of a bunch of exponentials is an exponential. However, this is not the case for the maximum.

This has nice practical applications. For example, recall the memoryless property of the exponential, which nicely models radioactive decay. If we have n elements each decaying at some Exponential(λ_i) rate, then we can model the time at which the first alpha particle will emit amongst all n elements will also be an exponential. These processes where the inter-emission times are exponentials are called Poisson process, which we will discuss later.

Definition 1.25 (Order Statistic)

Let $X_1, X_2, ..., X_n$ be a finite collection of independent, identically distributed random variables. Suppose that they are continuously distributed with density f and CDF F. Define the random variable $X_{(k)}$ to be the kth ranked value, called the **kth order statistic**. This means that

$$X_{(1)} = \min\{X_1, X_2, ..., X_n\}, \quad X_{(n)} = \max\{X_1, X_2, ..., X_n\} \tag{126}$$

and in general, for any $k \in \{1, 2, ..., n\}$,

$$X_{(k)} = X_j \text{ if } \sum_{l=1}^{n} \mathbb{I}_{X_l < X_j} = k - 1 \tag{127}$$

which means that exactly $k-1$ of the values of X_l are less than X_j. Since F is continuous,

$$X_{(1)} < X_{(2)} < ... < X_{(n)} \tag{128}$$

holds with probability 1. This leads us to define the random variable $X_{(k)}$ representing the kth order statistic.

$$f_{(k)}(y) = \begin{cases} n \binom{n-1}{k-1} y^{k-1}(1-y)^{n-k} & y \in (0,1) \\ 0 & y \notin (0,1) \end{cases} \tag{129}$$

That is, $X_{(k)}$ has the Beta($k, n-k_1$) distribution.

1.5.2 Convolutions and Sums of Random Variables

Now given two random variables $X, Y : \Omega \to \mathbb{R}$ that each push their own probability laws $\mathbb{P}_X, \mathbb{P}_Y$ onto $\mathbb{R}$, their sum $Z = X + Y$ is also a random variable that pushes its own probability law $\mathbb{P}_Z$. We must actually prove that Z is a random variable, which we can do by proving that the preimage of every $(-\infty, x]$ is $\mathcal{F}$-measurable. Equivalently (by complementation), we must prove that the preimage of every $(x, +\infty)$ (that is, all sets of form $\{\omega \mid Z(\omega) > z\}$) is $\mathcal{F}$-measurable. Now we can write z as the sum of two numbers $z = q + (z - q)$, where $q \in \mathbb{R}$, and say that

$$\{\omega \mid Z(\omega) > z\} = \bigcup_{q \in \mathbb{R}} \{\omega \mid X(\omega) > q,\ Y(\omega) > z - q\} \tag{130}$$

But using the fact that $\mathbb{Q}$ is dense in $\mathbb{R}$, we can turn this from an uncountable union to a countable union and say

$$\{\omega \mid Z(\omega) > z\} = \bigcup_{q \in \mathbb{Q}} \{\omega \mid X(\omega) > q,\ Y(\omega) > z - q\} \tag{131}$$

$$= \bigcup_{q \in \mathbb{Q}} \big(\{\omega \mid X(\omega) > q\} \cap \{\omega \mid Y(\omega) > z - q\}\big) \tag{132}$$

and since I have a countable union of (an intersection of) these $\mathcal{F}$-measurable sets, $\{\omega \mid Z(\omega) > z\}$ is $\mathcal{F}$-measurable, and we are done. This equation above even gives us a hint of how to compute the CDF of Z.

Theorem 1.14 ()

Given random variables $X_1, X_2, \ldots, X_n$ of probability space $(\Omega, \mathcal{F}, \mathbb{P})$,
1. $X_1 + \ldots + X_n$ is a random variable.
2. $X_1 \cdot \ldots \cdot X_n$ is a random variable.

For simplicity, we will only consider jointly discrete or jointly continuous random variables. The probability law $\mathbb{P}_Z$ can be confusing to define, since given some Borel set $B \in \mathcal{R}$, we must now look at the preimage under the *sum* $X + Y$. A simpler way to approach this is to consider the joint distribution X, Y and look at its distribution, which we call the **convolution** of X and Y. This is especially simple to consider for discrete random variables.

Definition 1.26 (Sums of Discrete Random Variables)

Take two discrete random variables X, Y with their joint PMF $p_{X,Y}(x, y)$ and their sum $Z = X + Y$. We can see that the PMF of Z is

$$p_Z(z) = \sum_{(x,y)\,:\,x+y=z} p_{X,Y}(x, y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, z - x) \tag{133}$$

which by abuse of notation, we denote

$$\mathbb{P}(Z = z) = \sum_{x \in \mathcal{X}} \mathbb{P}(X = x, Y = z - x) \tag{134}$$

The CDF is very simple, since we just have to sum over all (x, y) such that their sum is less than z:

$$F_Z(z) = \sum_{(x,y)\,:\,x+y\leq z} p_{X,Y}(x, y) \tag{135}$$

which by abuse of notation, we write

$$\mathbb{P}(Z \leq z) = \sum_{(x,y)\,:\,x+y\leq z} \mathbb{P}(X = x, Y = y) \tag{136}$$

If X and Y are independent, then their joint distribution is the product of their singular distributions, and so we have

$$p_Z(z) = \sum_x p_X(x)\, p_Y(z-x) \coloneqq p_X * p_Y \tag{137}$$

where $p_Z = p_X * p_Y$ is called the convolution of p_X and p_Y. By abuse of notation,

$$\mathbb{P}(Z=z) = \sum_{x\in\mathcal{X}} \mathbb{P}(X=x)\,\mathbb{P}(Y=z-x) \tag{138}$$

Example 1.24 (Sums of Poisson RVs)

Let X_1 and X_2 be independent Poisson random variables with parameters $\lambda_1, \lambda_2 > 0$, and let $Z = X_1 + X_2$. The PMF of each X_i is

$$p_{X_i}(k) = \frac{e^{-\lambda_i}\lambda_i^k}{k!} \text{ for } k \in \mathbb{Z} \tag{139}$$

and taking the convolution gives the PMF of Z:

$$\begin{aligned} p_Z(z) &= (p_{X_1} * p_{X_2})(z) \\ &= \sum_{k=-\infty}^{+\infty} \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{z-k}}{(z-k)!} \\ &= \sum_{k=0}^{z} \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{z-k}}{(z-k)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{z!} \sum_{k=0}^{z} \binom{z}{k} \lambda_1^k \lambda_2^{z-k} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^z}{z!} \end{aligned}$$

for $z \in \mathbb{N}_0$, which is the PMF of another Poisson. So, $Z \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

This has a nice visualization, since the joint distribution of X and Y over $\mathbb{R}^2$ is being "summed up/integrated" over the diagonals of $\mathbb{R}^2$, i.e. the lines where $x + y = z$ for some z, sort of like marginalizing over these diagonals. This creates a new "diagonally marginal distribution" Z.

Definition 1.27 (Sums of Continuous Random Variables)

Take two continuous random variables X, Y with their joint PDF $f_{X,Y}(x,y)$ and their sum $Z = X+Y$. To calculate the CDF, we must basically integrate the joint PDF over the borel set $\{(x,y) \in \mathbb{R}^2 \mid x + y \leq z\}$.

$$\begin{aligned} \mathbb{P}(Z \leq z) = F_Z(z) &= \int_{(x,y)\,:\,x+y\leq z} f_{X,Y}(x,y)\, dy\, dx \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{z-x} f(x,y)\, dy\, dx \end{aligned}$$

We can see that the PDF of Z is

$$f_Z(z) = \int_{\mathbb{R}} f_{X,Y}(x, z-x)\, dx \tag{140}$$

If X and Y are independent, then

$$f_Z(z) = \int_{\mathbb{R}} f_X(x)\, f_Y(z-x)\, dx \coloneqq f_X * f_Y \tag{141}$$

where $f_Z = f_X * f_Y$ is the convolution of f_X and f_Y.

Definition 1.28 (Convolution)

Given two functions $f, g : \mathbb{R} \longrightarrow \mathbb{R}$, the **convolution** of f and g is a new function $f * g$ defined

$$(f * g)(t) \coloneqq \int_{\mathbb{R}} f(t)\, g(t-\tau)\, d\tau \tag{142}$$

Usually, when we take convolutions, it is not pretty and even for nice distributions like two Gaussians, convolving them is quite complicated. What we can do is transform them (using Laplace, Fourier, etc.) to make calculations easier and more elegant, which we will talk about later.

Example 1.25 ()

Let X_1 and X_2 be independent exponential with parameters λ_1, λ_2, with individual PDFs $f_{X_i}(x) = \lambda_i e^{-\lambda_i x}$ for $x \geq 0$. Let $Z = X_1 + X_2$. Then,

$$\begin{aligned} f_Z(z) = (f_{X_1} * f_{X_2})(z) &= \int_{-\infty}^{\infty} \lambda_1 e^{-\lambda_1 x}\, \lambda_2 e^{-\lambda_2 (z-x)}\, dx \\ &= \int_0^z \lambda_1 e^{-\lambda_1 x}\, \lambda_2 e^{-\lambda_2 (z-x)}\, dx \\ &= \lambda_1 \lambda_2 e^{-\lambda_2 z} \int_0^z e^{(\lambda_2 - \lambda_1)x}\, dx \\ &= \begin{cases} \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 z} - e^{-\lambda - 2z} \right) & \text{if } \lambda_1 \neq \lambda_2 \\ \lambda^2 z e^{-\lambda z} & \text{if } \lambda_1 = \lambda_2 = \lambda \end{cases} \end{aligned}$$

The distribution for when $\mu_1 = \mu_2$ is called the Erlang distribution, which has many applications, but the other case is an ugly form and not studied very well.

Theorem 1.15 (Sums of Discrete Variables)

Assume that X and Y are independent.

1. $X \sim \text{Binomial}(n, p)$, $Y \sim \text{Binomial}(m, p) \implies X + Y \sim \text{Binomial}(n + m, p)$.
2. $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\gamma) \implies X + Y \sim \text{Poisson}(\lambda + \gamma)$.
3. If $X_1, ..., X_n$ are Geometric(p), then $X_1 + ... + X_n$ is NB(n, p).

Theorem 1.16 (Sums of Densities)

Assume that X and Y are independent.

1. $X \sim \text{Normal}(\mu_1, \sigma_1^2)$, $Y \sim \text{Normal}(\mu_2, \sigma_2^2) \implies X + Y \sim \text{Normal}\ (\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
2. If $X_1, X_2, ..., X_n$ are Exponential(λ), then $X_1 + ... + X_n \sim \text{Gamma}(n, \lambda)$.
3. $X \sim \text{Gamma}(n, \lambda)$, $Y \sim \text{Gamma}(m, \lambda) \implies X + Y \sim \text{Gamma}(n + m, \lambda)$.
4. $X \sim \text{Gamma}\ (n, \lambda)$, $Y \sim \text{Exponential}\ (\lambda) \implies X + Y \sim \text{Gamma}(n + 1, \lambda)$.

1.5.3 Sum of Random Number of Random Variables

Now we consider a random variable where the number of terms we are summing is a random variable. Let $\{X_i\}_i$ be a countable sequence of independent random variables with CDF F_{X_i}. Let N be a positive integer-valued random variable with PMF $p_N(n) = \mathbb{P}(N = n)$. Assume that N is independent of X_i's. Now, consider the function

$$S_N \coloneqq \sum_{i=1}^{N} X_i \tag{143}$$

To interpret this, consider the sample space Ω. We have all X_i's and N defined on the same Ω. Once $\omega \in \Omega$ realizes, the $\{X_i\}$'s will realize as a sequence of numbers, and N will realize as a positive integer. We simply sum them up according to the rule S_N, and by this definition, S_N is a real-valued function on Ω. We first have to prove that S_N is a random variable (since we only know that a *fixed* sum of random variables is a random variable), and then we must find the CDF of S_N $\mathbb{P}(S_N \leq x)$.

First, note that the realization of N partitions the sample space as

$$\Omega = \bigsqcup_{n=1}^{\infty} \{\omega \mid N(\omega) = n\} \tag{144}$$

Once I have this partition, I can invoke the partition rule and write

$$\begin{aligned}
\mathbb{P}(S_N \leq x) &= \sum_{k=1}^{\infty} \mathbb{P}(S_N \leq x \mid N = k)\, \mathbb{P}(N = k) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(S_k \leq x \mid N = k)\, \mathbb{P}(N = k) && \text{(conditioned on } N = k) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(S_k \leq x)\, \mathbb{P}(N = k) && (N \text{ is indep. of } X_i\text{s})
\end{aligned}$$

where $\mathbb{P}(N = k)$ is known since we know the PMF of N, and the CDFs $\mathbb{P}(S_k \leq x)$ can be computed by computing the deterministic sums and computing their CDF.

Example 1.26 ()

Let X_i's be iid Exponential(λ), and $N \sim$ Geometric(p). We know that the deterministic sum of iid exponentials gives an Erlang. So, $S_N = \sum_{i=1}^{N} X_i$, and its CDF is

$$\mathbb{P}(S_N \leq x) = \sum_{k=1}^{\infty} \mathbb{P}(S_k \leq x)\, \mathbb{P}(N = k) \tag{145}$$

where $\mathbb{P}(N = k) = (1-p)^{k-1}p$. The PDF of the Erlang is

$$p_{S_k}(x) = \frac{\lambda^n x^{n-1}}{(n-1)!} e^{-\lambda x} \tag{146}$$

and doing the brute force calculations gives a clean $S_N \sim$ Exponential(λp).

1.5.4 General Transformations of Random Variables

Now we will look at more general transformations that are not just minimum, maximum, deterministic sums, or random sums. Let us have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable $X : \Omega \to \mathbb{R}$, and a measurable function $f : \mathbb{R} \to \mathbb{R}$. Now given that we know the CDF (and therefore distribution) of X, we want to find the CDF of random variable $Y = f(X) = f \circ X$ (which we have established as a random variable already due to measurability of f): $F_Y(y) = \mathbb{P}(Y \leq y)$, which is just $\mathbb{P}_Y((-\infty, y])$ (where $\mathbb{P}_Y$ is the probability law on

Y). But rather than trying to take the preimage of the entire composite random variable Y and calculating $\mathbb{P}\big(Y^{-1}((-\infty, y])\big)$ under the probability on $\mathcal{F}$, let's just take the preimage one step at a time. Note that $f^{-1}\big((-\infty, y]\big) = \{x \in \mathbb{R} \mid f(x) \leq y\}$. We can then write the CDF of Y in terms of the probability law of X:

$$\begin{aligned} F_Y(y) &= \mathbb{P}_X\big(f^{-1}((-\infty, y])\big) \\ &= \mathbb{P}_X\big(\{x \in \mathbb{R} \mid f(x) \leq y\}\big) \\ &= \mathbb{P}\big(X^{-1} \circ f^{-1}((-\infty, y])\big) \end{aligned}$$

Depending on how complicated f is, this may be easy or not, but conceptually, this is no problem. But theoretically, this is as far as we can go. Let's move onto some examples. We start with a practical way to generate a Gaussian distribution, which is how most modern software computes.

Example 1.27 (Box-Muller Transform)

Given that you have a uniform random number generator in $[0, 1]$, you can generate a normal $N(0, 1)$ by transforming it using the inverse CDF of the normal. This is usually computationally heavy since the inverse CDF of the Gaussian requires expensive operations. An easier way is to use the **Box-Muller transform**, where you take two uniforms u_1, u_2 and transform it as

$$\begin{aligned} x_1 &= \sqrt{-2\ln(u_1)}\cos(2\pi u_2) \\ x_2 &= \sqrt{-2\ln(u_1)}\sin(2\pi u_2) \end{aligned}$$

Once you have $x \sim N(0, 1)$, you can use $\mu + \sigma x \sim N(\mu, \sigma^2)$. You can extend this to a n-dimensional normal distribution $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$ and transform it to get $\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Example 1.28 (Chi-Squared Distribution)

Let $X \sim \mathcal{N}(0, 1)$ and $Y = f(X) = X^2$. Note that X takes values in $(-\infty, +\infty)$ and Y in $[0, +\infty)$. Then, we can write

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}_Y((-\infty, y]) \\ &= \mathbb{P}_Y([0, y]) && \text{(range of } Y\text{)} \\ &= \mathbb{P}_X(f^{-1}([0, y])) && \text{(work in prob. law of } X\text{)} \\ &= \mathbb{P}_X([-\sqrt{y}, \sqrt{y}]) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\, dx \end{aligned}$$

Rewriting this in our abuse of notation notation, we have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X^2 \leq y) \\ &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= 2\mathbb{P}(0 \leq X \leq \sqrt{y}) && \text{(Symmetry of Gaussian)} \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-x^2/2}\, dx \end{aligned}$$

and this is clearly differentiable, since it is written like an integral. Doing so gives the PDF

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \text{ for } y \geq 0 \tag{147}$$

This describes the PDF of a **Chi-Squared** distribution.

Example 1.29 (Log-Normal Distribution)

Let $X \sim \mathcal{N}(0,1)$ and $Y = f(X) = e^X$. Note that the range of f is $(0, +\infty)$. So,

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}_Y((-\infty, y]) \\ &= \mathbb{P}_Y((0, y]) \\ &= \mathbb{P}_X(f^{-1}((0, y])) \\ &= \mathbb{P}_X((-\infty, \ln y]) \\ &= \int_{-\infty}^{\ln y} f_X(x)\, dx \end{aligned}$$

Rewriting this in our abuse of notation notation, we have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(e^X \leq y) \\ &= \mathbb{P}(X \leq \ln(y)) \\ &= \int_{-\infty}^{\ln(y)} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \end{aligned}$$

We can differentiate this to get

$$f_Y(y) = \frac{1}{y\sqrt{2\pi}} e^{-\frac{(\ln y)^2}{2}} \text{ for } y \geq 0 \tag{148}$$

This describes the PDF of a **log-normal** distribution.

We now show a more specific formula under more specific assumptions about the transformation. Suppose X is a *continuous* random variable with density f_X and $g : \mathbb{R} \to \mathbb{R}$ a monotonic differentiable function. Then, the CDF of the random variable $Y = g(X)$ can be written in the probability law of X, which can then by written as an integral by invoking the Radon-Nikodym theorem:

$$\begin{aligned} \mathbb{P}(Y \leq y) &= \mathbb{P}_X(f^{-1}((-\infty, y]) \\ &= \int_{x\,:\,g(x) \leq y} f_X(x)\, dx \end{aligned}$$

Note that we can now talk about the actual inverse g^{-1} since differentaible and monotonic implies invertibility.

1. Assuming g is monotonically increasing, we can use the change of variables $x = g^{-1}(t)$ and $g(x) = t \implies g'(x)\, dx = dt$ to get the above integral as

$$\int_{-\infty}^{g^{-1}(y)} f_X(x)\, dx = \int_{-\infty}^{t} \frac{f_X\big(g^{-1}(t)\big)}{g'\big(g^{-1}(t)\big)}\, dt \tag{149}$$

   but since this is simply the CDF of Y, the PDF must equal

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(t))} \tag{150}$$

2. If g is monotonically decreasing, we get

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{-g'(g^{-1}(t))} \tag{151}$$

In general, we can consider both cases by putting an absolute value

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(t))|} \tag{152}$$

and $g'(g'(y))$ is the Jacobian, the same one that we use when we perform a change of variables in integration.

Example 1.30 (Log-Normal Revisited)

Given $X \sim \mathcal{N}(0,1)$ and $Y = e^X$ (which is monotonically increasing), we can simply plug in the formula to get the PDF:

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(t))|} = \frac{f_X(\ln y)}{|e^{\ln y}|} = \frac{1}{\sqrt{2\pi}y} e^{-(\ln y)^2/2} \tag{153}$$

for $y > 0$. This domain is important since $\ln y$ is only defined for $y > 0$.

Example 1.31 ()

Given $X \sim \mathcal{N}(0,1)$ and $Y = f(X) = X^2$, we cannot use the formula since f is not monotonic on the range of X, which is $(-\infty, +\infty)$.

Example 1.32 ()

Given $X \sim \text{Exponential}(\lambda)$ and $Y = f(X) = X^2$, it may seem like the formula is not applicable here, but f *is* monotonic on the range of X, which is $[0, +\infty)$.

However, there is much less chance of error by deriving using first principles, so I would recommend using it always rather than these formulas.

Let's do the n-dimensional version of this. Given random variables $X_1, X_2, \dots, X_n$ iid random variables with joint density $f_{X_1 \dots X_n}(x_1, \dots, x_n)$, we define the transformation $g : \mathbb{R}^n \to \mathbb{R}^n$ as

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} g_1(X_1) \\ \vdots \\ g_n(X_N) \end{bmatrix} \tag{154}$$

Then, the PDF of Y will be

$$\begin{aligned} f_{Y_1 \dots Y_n}(y_1, \dots, y_n) &= f_{X_1 \dots X_n}\big(\mathbf{g}^{-1}(\mathbf{y})\big) \cdot |\mathbf{J}(\mathbf{y})| \\ &= f_{X_1 \dots X_n}\big(g_1^{-1}(y_1), \dots, g_n^{-1}(y_n)\big) \cdot |\mathbf{J}(\mathbf{y})| \end{aligned}$$

where

$$\mathbf{J}(y) = \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial y_n} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix} \tag{155}$$

2 Integration

2.1 Construction and Properties

2.1.1 Simple Functions

Remember that Riemann integration is characterized by the approximation of step functions, which are the "building blocks" of Riemann integrable functions. To define the Lebesgue integral, we will consider a generalization of step functions called *simple functions*. A function will be Lebesgue integrable if it can be approximated by these simple functions in some appropriate way.

Definition 2.1 (Simple Functions)

For $A \subset X$ (any subset, not just in some σ-algebra), the **characteristic**, or **indicator function** of A is the function $1_A : X \longrightarrow \mathbb{R}$ defined

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if else} \end{cases} \tag{156}$$

A function $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ is called a **simple function** if it is a finite linear combination of characteristic functions.

$$\phi = \sum_{i=1}^{n} a_i 1_{A_i} \tag{157}$$

Lemma 2.1 (Measurability on Simple Functions)

Now, let $(X, \mathcal{A})$ be a measurable space. Then,

$$\phi = \sum_{i=1}^{n} a_i 1_{A_i} : (X, \mathcal{A}) \longrightarrow \mathbb{R} \tag{158}$$

is measurable if all A_i are measurable, i.e. $A_i \in \mathcal{A}$ for all i.

Proof.

Let T be an open set in $\mathbb{R}$. Then, for characteristic function 1_A,

$$1_A^{-1}(T) = \begin{cases} \emptyset & \text{if } 0, 1 \notin T \\ A & \text{if } 1 \in T, 0 \notin T \\ X \setminus A & \text{if } 0 \in T, 1 \notin T \\ X & \text{if } 0, 1 \in T \end{cases} \tag{159}$$

and so 1_A must be measurable if $A \in \mathcal{A}$ (which also by definition implies that $A^c = X \setminus A \in \mathcal{A}$). If 1_{A_i} is measurable, then the linear combination of measurable functions is also measurable.

Also observe that the coefficients need not be unique, since we can write

$$1 \cdot 1_{[0,1]} + 1 \cdot 1_{[0.5,1]} = 1 \cdot 1_{[0,0.5]} + 2 \cdot 1_{[0.5,1]} \tag{160}$$

If the E_i's are disjoint, then this decomposition is unique and is called the **standard representation** of ϕ.

Example 2.1 (Step Function as Simple Function)

For $a, b \in \mathbb{R}$, with $a < b$, let $f : [a, b] \longrightarrow \mathbb{R}$ be a step function. That is, there exists a partition $a = x_0 < x_1 < \ldots < x_n = b$ and constants $c_1, c_2, \ldots, c_n \in \mathbb{R}$ s.t. $f(x) = c_i$ for all $x \in (x_{i-1}, x_i)$ and each $i = 1, \ldots, n$. Then, f is equal to the following simple function, taken over all open intervals and the points x_j at the boundary of each interval.

$$f = \sum_{i=1}^{n} c_i 1_{(x_{i-1}, x_i)} + \sum_{j=0}^{n} f(x_j) 1_{\{x_j\}} \tag{161}$$

If we ignore the behavior of f on the partition points x_j's, then f agrees almost everywhere with the simple function

$$\sum_{i=1}^{n} c_i 1_{(x_{i-1}, x_i)} \tag{162}$$

If the A_i's above are just intervals in $\mathbb{R}$, then ϕ reduces to a step function. But the entire problem with intervals is that they are too coarse. We can't work with them, so we generalize them to all measurable sets in $(X, \mathcal{A})$. The Riemann integral is built on an approximation scheme of a function, which we usually want to be continuous to satisfy this approximation, and so, if we want to build an approximation scheme for Lebesgue integrals, we want a similar scheme, i.e. if we take a sequence of simple measurable functions, I can get arbitrarily close to any measurable function f. This is exactly what we show below.

Theorem 2.1 ()

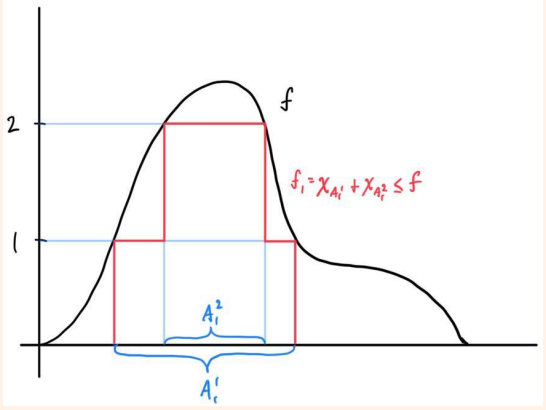
If $f : (X, \mathcal{A}) \longrightarrow [0, \infty]$ is measurable, there are simple measurable functions $f_k : (X, \mathcal{A}) \longrightarrow [0, \infty)$ s.t.

$$f_k \leq f_{k+1} \text{ and } f = \lim_{k \to \infty} f_k \tag{163}$$

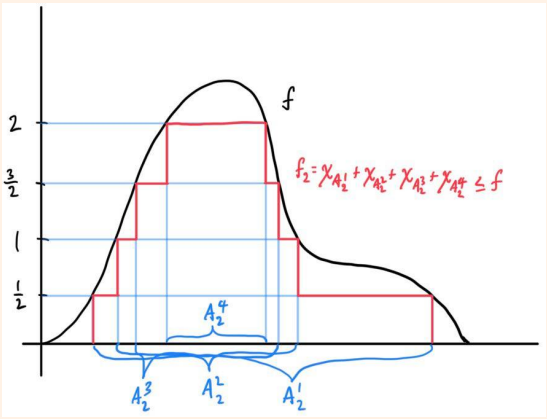
where the inequalities and limits are pointwise.

Proof.

We give a general picture of this proof for a function $f : \mathbb{R} \longrightarrow [0, \infty]$. We can first divide the codomain of the graph below into segments of $t = 1, 2, \ldots$, and take the preimage of all these units under f to get f_1. More specifically, $A_1^t = f^{-1}([t, \infty])$ for all t. By measurability of f, A_1^t is measurable, and we can assign $f_1 = 1_{A_1^1} + 1_{A_1^2} \leq f$.



Doing this again with finer subintervals of the codomain gives us, with $f_2 = 1_{A_2^1} + 1_{A_2^2} + 1_{A_2^3} + 1_{A_2^4} \leq f$.



and in general, we have $f_k = \sum_{j=1}^{\infty} \frac{1}{2^{k-1}} 1_{A_k^j}$. But we said a simple function is a *finite* sum, and if ∞ is in the range of f, then this becomes a problem. We can quickly fix this by just truncating the summation at a certain point in the codomain (f_1 only considers intervals up to 1, f_2 up to 2 and so on), ultimately giving us

$$f_k = \sum_{j=1}^{k2^{k-1}} \frac{1}{2^{k-1}} 1_{A_k^j} \tag{164}$$

2.1.2 Lebesgue Integral

Finally, we can learn how to integrate. We require the positiveness condition on f below because our previous theorem on approximating arbitrary functions with simple measurable functions f_k requires that it be positive, too.

Definition 2.2 (Lebesgue Integral of Positive Simple Functions)

If $f = \sum_{k=1}^{n} c_k 1_{A_k}$ is a positive simple Lebesgue measurable function on measure space $(X, \mathcal{A}, \mu)$, then the **Lebesgue integral** of f is

$$\int f \, d\mu = \sum_{k=1}^{n} c_k \mu(A_k) \tag{165}$$

This Lebesgue integral agrees with the Riemann integral for step functions. Let $c_1, \dots, c_n \in [0, \infty)$ and $a = x_0 < x_1 < \dots < x_n = b$ be a partition. Let $f : [a, b] \longrightarrow [0, \infty]$ be a step function taking the value c_i on the interval (x_{i-1}, x_i) for $i = 1, \dots, n$. Then the Riemann integral of f is simply

$$\int f(x) \, dx = \sum_{i=1}^{n} c_k |x_i - x_{i-1}| \tag{166}$$

The Lebesgue integral is

$$\begin{aligned} \int f \, d\mu &= \sum_{i=1}^{n} c_i \mu((x_{i-1}, x_i)) + \sum_{j=0}^{n} f(x_j) \mu(\{x_j\}) \\ &= \sum_{i=1}^{n} c_k |x_i - x_{i-1}| \end{aligned}$$

which agrees with the Riemann integral. In the Riemann integral, we write dx to indicate the variable that is being integrated over, but in the Lebesgue integral, we write $d\mu$, the measure which we are integrating over.

Therefore, there are many possible values that can come out of a Lebesgue integral of a certain function, while a Riemann integral outputs only one value if exists.

Example 2.2 ()

Consider the simple function (consisting of one characteristic function) $1_{\mathbb{Q}\cap[0,1]}$. $\mathbb{Q}\cap[0,1]$ is a Lebesgue measurable set of $\mathbb{R}$, and we have $1_{\mathbb{Q}\cap[0,1]} \geq 0$, so its Lebesgue integral is given by the above definition:

$$\int_{\mathbb{R}} 1_{\mathbb{Q}\cap[0,1]}\, d\lambda = 1 \cdot \lambda(\mathbb{Q}\cap[0,1]) = 0 \tag{167}$$

Definition 2.3 (Lebesgue Integral on Positive Measurable Functions)

If $f : (X, \mathcal{A}, \mu) \longrightarrow [0, \infty]$ is measurable, then

$$\int_X f\, d\mu = \sup\left\{ \int g\, d\mu \,\middle|\, g \text{ simple }, g \leq f \right\} \tag{168}$$

Unlike Riemann integration, which looks at both the supremum and infimum of integrals of simple functions, Lebesgue integration only looks at the supremum, given that f is nonnegative, so for all these f, the Lebesgue integral always exists. Defining Lebesgue integration for all real-valued functions, requires a simple extension.

Definition 2.4 (Lebesgue Integral)

Given a function $f : (X, \mathcal{A}, \mu) \longrightarrow \mathbb{R}$, we can split f into a positive and negative part:

$$f = f^+ - f^- \tag{169}$$

where $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$. Then, the Lebesgue integral of f is

$$\int f\, d\mu = \int f^+\, d\mu - \int f^-\, d\mu \tag{170}$$

given that at least one of these integrals is finite. If one is infinite and the other is finite, then we can call it infinite. If we have *both* infinite integrals, then the integral doesn't exist. It has the properties:

1. Monotonicity:
$$g \leq f \implies \int g\, d\mu \leq \int f\, d\mu \tag{171}$$
2. Scalar Multiplication:
$$\int cf\, d\mu = c \int f\, d\mu \tag{172}$$
3. Addition:
$$\int f + g\, d\mu = \int f\, d\mu + \int g\, d\mu \tag{173}$$

Since $|f| = f^+ + f^-$, f is also Lebesgue integrable if

$$\int |f|\, d\mu < \infty \tag{174}$$

since by triangle inequality, we have

$$\left| \int f\, d\mu \right| \leq \int |f|\, d\mu \tag{175}$$

Definition 2.5 ()

The set of all functions $f : (X, \mathcal{A}, \mu) \longrightarrow \mathbb{R}$ that are Lebesgue integrable is denoted $\mathcal{L}^1(X, \mathcal{A}, \mu; \mathbb{R})$, or for short $\mathcal{L}^1(X, \mathcal{A}, \mu)$.

Theorem 2.2 ()

$f : \mathbb{R} \longrightarrow \mathbb{R}$ is Riemann integrable iff it is continuous λ almost everywhere. If so, then f is Lebesgue measurable and

$$\int_{[a,b]} f \, d\lambda = \int_a^b f \, dx \tag{176}$$

for all $a < b$.

2.1.3 Integral Inequalities

We introduce 3 important inequalities on the integral.

Theorem 2.3 (Jensen's Inequality)

Suppose ϕ is convex, that is,

$$\lambda\phi(x) + (1-\lambda)\phi(y) \geq \phi(\lambda x + (1-\lambda)y) \tag{177}$$

for all $\lambda \in (0, 1)$ and $x, y \in \mathbb{R}$. If μ is a probability measure, and f and $\varphi(f)$ are integrable, then

$$\varphi\bigg(\int f \, d\mu \bigg) \leq \int \varphi(f) \, d\mu \tag{178}$$

Theorem 2.4 (Holder's Inequality)

If p, q are Holder conjugates, then

$$\int |fg| \, d\mu \leq ||f||_p ||g||_q \tag{179}$$

Corollary 2.1 (Cauchy-Schwarz Inequality)

Given that $p = q = 2$ above, then we have

$$\int |fg| \, d\mu \leq ||f||_2 ||g||_2 \tag{180}$$

which is similar to the familiar equation $\langle u, v \rangle \leq ||u|| ||v||$.

2.1.4 Convergence Theorems

Now, we want to give conditions that guarantee

$$\lim_{n\to\infty} \int f_n \, d\mu = \int \big(\lim_{n\to\infty} f_n \big) \, d\mu \tag{181}$$

Definition 2.6 (Convergence in Measure)

A sequence of functions $f_n \to f$ **in measure** if for any $\epsilon > 0$,

$$\mu\big(\{x : |f_n(x) - f(x)| > \epsilon\}\big) \to 0 \text{ as } n \to \infty \tag{182}$$

Theorem 2.5 (Bounded Convergence Theorem)

Let E be a set with $\mu(E) < \infty$. Suppose $f_n = 0$ on E^c, $|f_n(x)| \leq M$, and $f_n \to f$ in measure. Then,

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n d\mu \tag{183}$$

Lemma 2.2 (Fatou's Lemma)

If $f_n \geq 0$, then

$$\lim_{n\to\infty} \inf \int f_n \, d\mu \geq \int \Big(\lim_{n\to\infty} \inf f_n \Big) \, d\mu \tag{184}$$

Theorem 2.6 (Monotone Convergence Theorem)

Given a nondecreasing sequence of measurable nonnegative functions $\{f_n\}$, its limit $f_n \uparrow f$ always exists (since f_n is nondecreasing), is measurable, and

$$\int f_n \, d\mu \uparrow \int f \, d\mu \tag{185}$$

This allows us to integrate the limit of nice functions f_n by integrating these f_n first and then finding what the values converge to.

Theorem 2.7 (Dominated Convergence Theorem)

If $f_n \to f$ a.e., $|f_n| \geq g$ for all n, and g is integrable, then

$$\int f_n \, d\mu \to \int f \, d\mu \tag{186}$$

2.1.5 Product Measures, Fubini's Theorem

Let $(X, \mathcal{A}, \mu_1)$ and $(Y, \mathcal{B}, \mu_2)$ be two measure spaces. Let

$$\Omega = X \times Y = \{(x, y) \mid x \in X, y \in Y\}$$
$$\mathcal{S} = \{A \times B \mid A \in \mathcal{A}, B \in \mathcal{B}\}$$

The sets in $\mathcal{S}$ are called **rectangles**. It is easy to see that $\mathcal{S}$ is a semi-algebra:

$$(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$$
$$(A \times B)^c = (A^c \times B) \cup (A \times B^c) \cup (A^c \times B^c)$$

Theorem 2.8 ()

There is a unique measure $\mu = \mu_1 \times \mu_2$ (or denoted $\mu_1 \otimes \mu_2$) on $\mathcal{F}$ with

$$\mu(A \times B) = \mu_1(A)\,\mu_2(B) \tag{187}$$

Theorem 2.9 (Fubini's Theorem)

Let $(X, \mathcal{A}, \mu_1)$ and $(Y, \mathcal{B}, \mu_2)$ be two measure spaces and $(X \times Y, \mathcal{F}, \mu = \mu_1 \times \mu_2)$ be their product space. Then, if $f \geq 0$ or $\int_{X\times Y} |f|\, d\mu < \infty$, then

$$\int_X \int_Y f(x,y)\,\mu_2\mu_1 = \int_{X\times Y} f\, d\mu = \int_Y \int_X f(x,y)\,\mu_1\mu_2 \tag{188}$$

2.2 Random Vectors

Now when we consider several random variables, they will all be defined on the same probability space. Given two random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$, they will each induce a probability law $\mathbb{P}_X$ and $\mathbb{P}_Y$ which completely characterizes them. Note that it is the same underlying randomness that is feeding these random variables, and so if I know some information about the value of X, then we know something about outcome ω, which can be used to find something about the value of Y. To capture this, we can imagine the map $(X, Y) : \Omega \longrightarrow \mathbb{R}^2$ defined $(X, Y)(\omega) \coloneqq (X(\omega), Y(\omega))$. And just like how X induces a measure P_X onto $\mathbb{R}$, we can imagine (X, Y) inducing a measure onto $\mathcal{B}(\mathbb{R}^2)$, which can be generated by all semi-infinite rectangles $(-\infty, x] \times (-\infty, y]$. Ideally, we would want to put a measure $\mathbb{P}_{X,Y}$ on $\mathbb{R}^2$ s.t.

$$\mathbb{P}_{X,Y}(B) \coloneqq \mathbb{P}((X, Y)^{-1}(B)) \tag{189}$$

where $(x, y)^{-1}(b) = \{\omega \in \omega \mid (x(\omega), y(\omega)) \in b\}$ denotes the preimage of (x, y). but is $(x, y)^{-1}(b)$ {-measurable? it turns out that it is.

Theorem 2.10 ()

Let $f : (X, \mathcal{A}, \mu) \longrightarrow \mathbb{R}^n$ have component functions $f_1, f_2, \ldots, f_n$. Then, f is measurable (i.e. $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}(\mathbb{R}^n)$) if and only if all of its component functions are measurable (i.e. $f_i^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}(\mathbb{R}^n)$).

From the theorem above, I have a probability law $\mathbb{P}_{X,Y}$ on all Borel sets of $\mathbb{R}^2$, making $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \mathbb{P}_{X,Y})$ a probability space. Now, since X and Y are both random variables dependent on the same $\omega \in \Omega$, we could expect certain "combinations" of X and Y to be more probable than other combinations.

Definition 2.7 (Joint Probability Law)

Given two random variables X, Y on $(\Omega, \mathcal{F}, \mathbb{P})$, the **joint random variable** $(X, Y) : \Omega \longrightarrow \mathbb{R}^2$ is a measurable function defined

$$(X, Y)(\omega) \coloneqq (X(\omega), Y(\omega)) \tag{190}$$

which induces a **joint probability law** $\mathbb{P}_{X,Y} : \mathcal{B}(\mathbb{R}^2) \longrightarrow [0, 1]$ defined

$$\mathbb{P}_{X,Y}(B) \coloneqq \mathbb{P}((X, Y)^{-1}(B))\ \forall B \in \mathcal{R} \tag{191}$$

of X, Y. This law captures everything there is about the interdependence of X and Y.

Given joint probability law $\mathbb{P}_{X,Y}$, we can get the probability laws of X and Y separately. For example, we can take a specific Borel set of $\mathbb{R}$ representing the outcomes of X and look at every single combination of it with every Y. But knowing $\mathbb{P}_X$ and $\mathbb{P}_Y$ is not enough to know the joint $\mathbb{P}_{X,Y}$.

Definition 2.8 (Marginal Probability Law)

Given a joint probability law $\mathbb{P}_{X,Y}$ of X, Y, we can get the **marginal probability law** of X by feeding in Borel sets of form $B \times \mathbb{R} \in \mathcal{B}(\mathbb{R}^2)$.

$$\mathbb{P}_X(B) = \mathbb{P}_{X,Y}(B \times \mathbb{R}) \tag{192}$$

and the marginal probability law of Y as

$$\mathbb{P}_Y(B) = \mathbb{P}_{X,Y}(\mathbb{R} \times B) \tag{193}$$

Definition 2.9 (Joint Cumulative Distribution Function)

Since sets of the form $(-\infty, x] \times (-\infty, y]$ are Borel in $\mathbb{R}^2$, the **joint cumulative distribution function**

$$\begin{aligned} F_{X,Y} &\coloneqq \mathbb{P}_{X,Y}\big((-\infty, x] \times (-\infty, y]\big) \\ &= \mathbb{P}\big(\{\omega \mid X(\omega) \leq x\} \cap \{\omega \mid Y(\omega) \leq y\}\big) \end{aligned}$$

is well-defined. By abuse of notation, we will write $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$. The marginal CDFs are defined

$$\begin{aligned} F_X(x) &\coloneqq \mathbb{P}_{X,Y}((-\infty, x) \times \mathbb{R}) \\ F_Y(y) &\coloneqq \mathbb{P}_{X,Y}(\mathbb{R} \times (-\infty, y)) \end{aligned}$$

Lemma 2.3 (Properties of Joint CDF)

Some common properties of the joint CDF are as follows:

1. Limits.
$$\lim_{(x,y)\to(+\infty,+\infty)} F_{X,Y}(x, y) = 1 \text{ and } \lim_{(x,y)\to(-\infty,-\infty)} F_{X,Y}(x, y) = 0 \tag{194}$$
2. Monotonicity.
$$x_1 \leq x_2,\ y_1 \leq y_2 \implies F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2) \tag{195}$$
3. Continuity from above.
$$\lim_{\epsilon\to 0^+} F_{X,Y}(x + \epsilon, y + \epsilon) = F_{X,Y}(x, y) \text{ for all } x, y \in \mathbb{R} \tag{196}$$
4. Maringal CDFs.
$$\lim_{y\to\infty} F_{X,Y}(x, y) = F_X(x), \quad \lim_{x\to\infty} F_{X,Y}(x, y) = F_Y(y) \tag{197}$$

2.2.1 Joint Discrete Random Variables

Definition 2.10 (Joint PMF)

Given discrete random variables X and Y, let their countable images be denoted $E_X, E_Y \subset \mathbb{R}$. Then, $E_X \times E_Y$ is also countable, and so the joint random variable (X, Y) is also discrete. This means that we can write for some Borel B of $\mathbb{R}^2$,

$$\mathbb{P}_{X,Y}(B) = \sum_{(x,y)\in(E_X\times E_Y)\cap B} \mathbb{P}_{X,Y}(\{(x, y)\}) \tag{198}$$

and we can define the PMF as $p_{X,Y}(x, y) \coloneqq \mathbb{P}_{X,Y}(\{(x, y)\})$. By abuse of notation, we write

$p_{X,Y}(x,y) = \mathbb{P}(X=x, Y=y)$ and write

$$\mathbb{P}_{X,Y}(B) = \sum_{(x,y)\in(E_X\times E_Y)\cap B} \mathbb{P}(X=x, Y=y) \tag{199}$$

If you give me a joint PMF $p_{X,Y}$, by the definition above this determines the entire probability law of $\mathbb{P}_{X,Y}$.

Definition 2.11 (Conditional PMF)

Let X, Y be discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The **conditional PMF** of X given $Y=y$ is defined

$$p_{X|Y}(x \mid y) \coloneqq \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{\mathbb{P}_{X,Y}(\{x,y\})}{\mathbb{P}_Y(\{y\})} \tag{200}$$

and again by abuse of notation, we can simply write

$$\mathbb{P}(X=x \mid Y=y) \coloneqq \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(Y=y)} \tag{201}$$

Theorem 2.11 (TFAE)

Let X and Y be discrete random variables. Then, the following are equivalent:

1. X and Y are independent.
2. For all $x, y \in \mathbb{R}$, the events $\{X=x\}$ (aka $X^{-1}(\{x\})$) and $\{Y=y\}$ (aka $Y^{-1}(\{y\})$) are independent. That is,

$$\mathbb{P}\big[X^{-1}(\{x\}) \cap Y^{-1}(\{y\})\big] = \mathbb{P}(X^{-1}(\{x\}))\,\mathbb{P}(Y^{-1}(\{y\})) \tag{202}$$

3. For all $x, y \in \mathbb{R}$, $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$.
4. For all $x, y \in \mathbb{R}$ such that $p_Y(y) > 0$, we have $p_{X|Y}(x \mid y) = p_X(x)$.

2.2.2 Joint Continuous Random Variables

Definition 2.12 ()

X and Y are jointly continuous if the joint law $\mathbb{P}_{X,Y}$ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^2$ (i.e. a Borel set of Lebesgue measure 0 must have $P_{X,Y} = 0$ also).

However, X and Y continuous does not always imply that (X,Y) are jointly continuous! If we have $X \sim \mathcal{N}(0,1)$ and $Y = 2X \sim \mathcal{N}(0,4)$. Jointly continuous allows us to define a PDF on it.

Theorem 2.12 (Radon-Nikodym Theorem)

If X and Y are jointly continuous RVs, then there exists a measurable $f_{X,Y} : \mathbb{R}^2 \longrightarrow [0,\infty)$ s.t. for any $B \in \mathcal{B}(\mathbb{R}^2)$, we have

$$\mathbb{P}_{X,Y}(B) = \int_B f_{X,Y}\, d\lambda \tag{203}$$

The Radon-Nikodym Theorem guarantees the existence of such $f_{X,Y}$. Taking $B = (-\infty, x] \times (-\infty, y]$, we can define the joint CDF as

$$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y) \coloneqq \mathbb{P}_{X,Y}\Big((-\infty,x] \times (-\infty,y]\Big) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)\, dt\, ds \tag{204}$$

2.3 Expectation

Definition 2.13 (Expectation)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \longrightarrow \mathbb{R}$, the **expectation** of X is defined

$$\mathbb{E}[X] \coloneqq \int_\Omega X \, d\mathbb{P} \tag{205}$$

Generally, if we are integrating over the entire probability space, then it is conventional to not write Ω in the integral at all: $\mathbb{E}[X] = \int X \, d\mathbb{P}$.

Definition 2.14 (Expectation of Discrete RV)

If X is a discrete random variable *that takes positive values*, then let $E = \{e_1, e_2, \dots\}$ denote the set where $\mathbb{P}_X(E) = 1$, and let $E_i = X^{-1}(\{e_i\}) \subset \Omega$. Then, we can see that since X is constantly e_i on E_i,

$$\int_{E_i} X \, d\mathbb{P} = e_i \cdot \mathbb{P}(E_i) = e_i \cdot \mathbb{P}_X(\{e_i\}) = e_i \cdot \mathbb{P}(X = e_i) \tag{206}$$

which implies

$$\mathbb{E}[X] = \int_\Omega X \, d\mathbb{P} = \sum_{i=1}^{\infty} \int_{E_i} X \, d\mathbb{P} = \sum_{i=1}^{\infty} e_i \cdot \mathbb{P}(X = e_i) \tag{207}$$

If X is discrete RV possibly taking negative values, then let $X = X^+ - X^-$, where $X^+ = \max(X, 0)$ and $X^- = -\min(X, 0)$. Then, we can compute

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] \tag{208}$$

which is well-defined as long as we don't have "$\infty - \infty$."

Note that the reason why expectations of the form $\infty - \infty$ are indeterminate is because of the Riemann rearrangement theorem.

Theorem 2.13 (Riemann's Rearragenement Theorem)

Given a series $\sum a_n$ that is conditionally convergent (i.e. converges but not absolutely convergent), the terms can be arranged so that the new series converges to an arbitrary real number, or diverges.

Lemma 2.4 (Properties of Expectation)

Let X and Y be random variables with finite expectations.

1. Monotonicity: If $X \leq Y$ (i.e. $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$), then
$$\mathbb{E}[X] \geq \mathbb{E}[Y] \tag{209}$$
2. Non-Negativity: This is implied from the above if we set the lower bound to the constant random variable 0. If $X \geq 0$, then
$$\mathbb{E}[X] \geq 0 \tag{210}$$
3. Linearity: For all $a, b, c \in \mathbb{R}$,
$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c \tag{211}$$

We now show a widely-used, but nontrivial, theorem.

Theorem 2.14 (Expectation of Independent Events)

Given independent RVs X and Y,

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{212}$$

Proof.

We show only for simple random variables which will give us a start in proving for all random variables in full generality. Let X and Y be simple random variables, i.e.

$$X = \sum_i a_i 1_{A_i} \text{ and } Y = \sum_j b_j 1_{B_j} \tag{213}$$

Since $\{A_i\}_i$ and $\{B_j\}_j$ are both partitions, $\{A_i \cap B_j\}_{i,j}$ is also a partition, and

$$XY = \sum_{i,j} a_i b_j \, 1_{A_i \cap B_j} \tag{214}$$

Its expectation can be expanded out by linearity, and since $\mathbb{E}[1_A] = \mathbb{P}(A)$, we have

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{i,j} a_i b_j \, \mathbb{P}(A_i \cap B_j) \\ &= \sum_{i,j} a_i b_j \, \mathbb{P}(A_i)\,\mathbb{P}(B_j) = \mathbb{E}[X]\,\mathbb{E}[Y] \end{aligned}$$

Now that we have proved for simple random variables, we can just approximate X from below using simple functions.

Theorem 2.15 (Tail Sum Formula)

If a discrete random variable X takes values in the non-negative integers $\{0, 1, 2, 3, ...\}$, then

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) \tag{215}$$

In any case (continuous or discrete), if X is a non-negative random variable, then

$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X > x)\, dx = \int_0^{\infty} 1 - F(x)\, dx \tag{216}$$

where F is the CDF of X.

Proof.

Suppose that X takes values in $\{0, 1, 2, 3, ...\}$. Then,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k\geq 1} k\,\mathbb{P}(X=k) \\
&= \sum_{k\geq 1}\sum_{j=1}^{k} \mathbb{P}(X=k) \\
&= \sum_{k\geq 1}\sum_{j=1}^{k} 1_{j\leq k}\,\mathbb{P}(X=k) \\
&= \sum_{j=1}^{\infty}\sum_{k\geq 1} 1_{j\leq k}\,\mathbb{P}(X=k) \\
&= \sum_{j=1}^{\infty}\sum_{k\geq j} \mathbb{P}(X=k) \\
&= \sum_{j=1}^{\infty} \mathbb{P}(X\geq j)
\end{aligned}$$

Corollary 2.2 ()

For any $m > 0$ and $\alpha > 0$,

$$\mathbb{P}\big(|X| > \alpha\big) \leq \frac{1}{\alpha^m}\mathbb{E}\big(|X|^m\big) \tag{217}$$

Example 2.3 (Geometric RV)

Recall that given $X \sim \text{Geometric}(p)$, we have $\mathbb{P}(X=i) = (1-p)^{i-1}p$ for $i \geq 1$. So,

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x\,\mathbb{P}(X=x) = \sum_{x=1}^{\infty} x\,(1-p)^{i-1}p = \frac{p}{(1-(1-p))^2} = \frac{1}{p} \tag{218}$$

Example 2.4 (Infinite Expectation)

Let us have discrete random variable s.t. $\mathbb{P}(X=k) = \frac{6}{\pi^2}\frac{1}{k^2}$ for $k \geq 1$. So,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k\,\mathbb{P}(X=k) = \frac{6}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{k} = +\infty \tag{219}$$

Example 2.5 (Undefined Expectation)

Let $\mathbb{P}(X=k) = \frac{3}{\pi^2}\frac{1}{k^2}$ for $k \in \mathbb{Z}\setminus\{0\}$. The expectation of this can be computed by getting the expectation of all the positive terms and the negative terms.

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] = \sum_{k=1}^{\infty} k\cdot\frac{3}{\pi^2}\frac{1}{k^2} + \sum_{k=1}^{\infty}(-k)\cdot\frac{3}{\pi^2}\frac{1}{k^2} = \infty - \infty \tag{220}$$

Note that by the Riemann rearrangement theorem, we can't just say that the expectation is 0 since the terms "cancel out." We could only do this if the series is absolutely convergent also, which works

if X takes positive values only.

Note that when we compute expectation, what we do it multiply the PMF/PDF by x and sum/integrate over it. The Cauchy distribution is a power function of form $\frac{1}{x^2}$, so if we multiply it by x, we have the new $\frac{1}{x}$ which is divergent.

2.3.1 Law of the Unconscious Statistician

Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random vector $X : \Omega \to \mathbb{R}^n$, this induces a probability law $\mathbb{P}_X$ acting as a measure on $\mathbb{R}^n$. Assume that this probability law $\mathbb{P}_X$ is known. Now introduce a function $g : \mathbb{R}^n \to \mathbb{R}$. We can create a new random variable $Y = g \circ X : \Omega \to \mathbb{R}$ with its own probability law $\mathbb{P}_Y$ on $\mathbb{R}$. Since we already know the probability distribution of X, so we can easily get the expected value of X as (in the discrete case)

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(X = x) \tag{221}$$

where $\mathcal{X}$ is the support of X. But what if we wanted to get the expected value of Y?

$$\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} y \cdot \mathbb{P}(Y = y) = ? \tag{222}$$

The problem is that we don't know the probability distribution of Y. But since we know that all the values of X are transformed by g, we are taught to compute it in terms of the probability distribution of X.

$$\mathbb{E}[Y] = \sum_{x \in \mathcal{X}} g(x) \cdot \mathbb{P}(X = x) \tag{223}$$

This "identity" that is often used must actually be treated as a rigorous theorem. This is like a change of basis formula that allows us to shift to a convenient space to compute integrals.

Theorem 2.16 (LOTUS)

Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable $X : \Omega \to \mathbb{R}^n$, and a function $g : \mathbb{R}^n \to \mathbb{R}$, the expectation of $g(X)$ is

$$\mathbb{E}[g(X)] = \int_\Omega g(X)\, d\mathbb{P} = \int_{\mathbb{R}^n} g\, d\mathbb{P}_X = \int_{\mathbb{R}} d\mathbb{P}_{g(X)} \tag{224}$$

It is usually the case that we don't know the distribution of $g(X)$ since g is too complicated (hard to compute the right integral) and we don't want to integrate over an abstract space Ω where we can't do calculus on (hard to compute the left integral). But we do know the distribution of X, so we can indeed compute the middle integral.

Note that if $g : \mathbb{R} \to \mathbb{R}$ is the identity function id, then we have

$$\mathbb{E}[X] = \int_\Omega X\, d\mathbb{P} = \int_{\mathbb{R}} \mathrm{id}\, d\mathbb{P}_X \tag{225}$$

1. For the discrete case, the above integral simplifies to

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g\, d\mathbb{P}_X = \sum_{x \in \mathcal{X} \subset \mathbb{R}^n} g(x) p_X(x) \tag{226}$$

2. For the continuous case, we have

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g\, d\mathbb{P}_X = \int_{\mathbb{R}^n} g(x)\, f_X(x)\, dx \tag{227}$$

2.3.2 Expectation w.r.t. Different Measures

Sometimes, we just write the expectation of a measurable function $f : (S, \mathcal{S}) \to \mathbb{R}$ as $\mathbb{E}[f]$. If we need to specify with respect to what measure we are integrating over, we write

$$\mathbb{E}_\mu[f] \coloneqq \int_S f \, d\mu \tag{228}$$

Usually, if f represents some transformation of a random variable $X : \Omega \to S$, then we assume that we are integrating w.r.t. the probability measure $\mathbb{P}$ defined on Ω or the probability law $\mathbb{P}_X$ induced by X.

$$\mathbb{E}[f] = \mathbb{E}[f(X)] = \int_\Omega f(X) \, d\mathbb{P} = \int_S f \, d\mathbb{P}_X \tag{229}$$

Example 2.6 (Expectation of Exponential RV)

The PDF of exponential random variable X is defined $f_X = ke^{-kx}$ for $x \geq 0$. So,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X \, d\lambda = \int_0^\infty xke^{-kx} \, dx = \frac{1}{k} \tag{230}$$

Similarly, if we want the expectation of X^2, then we can get

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f_X \, d\lambda = \int_0^\infty x^2 ke^{-kx} \, dx = \frac{2}{k^2} \tag{231}$$

Example 2.7 (Expectation of Gaussian RV)

The expectation of a Gaussian random variable X is

$$\mathbb{E}[X] = \int_{-\infty}^\infty x \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \mu \tag{232}$$

Example 2.8 (Expectation of One-Sided Cauchy)

If we have $f_X(x) = \frac{2}{\pi} \frac{1}{1+x^2}$ for $x \geq 0$, then

$$\mathbb{E}[X] = \int_0^\infty \frac{2}{\pi} \frac{x}{1+x^2} \, dx \tag{233}$$

and making the substitution $t = \frac{1+x^2}{,} \, dt = 2x$, we have

$$\int_1^\infty \frac{1}{\pi} \frac{1}{t} \, dt = \frac{\ln(t)}{\pi} \Big|_1^\infty = +\infty \tag{234}$$

Example 2.9 (Expectation of Two-Sided Cauchy)

The two-sided Cauchy is just another copy of the one sided into the negatives, so $f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}$ for $x \in \mathbb{R}$. The expectation of X should be split up into for positive and negative images, but computing it gives

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] = \int_0^\infty \frac{1}{\pi} \frac{x}{1+x^2} \, dx - \int_{-\infty}^0 \frac{1}{\pi} \frac{x}{1+x^2} \, dx = \infty - \infty \tag{235}$$

and so it is undefined.

With LOTUS, we can make sense of an extremely important inequality.

Theorem 2.17 (Jensen's Inequality)

If f is a convex function, then $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

Proof.

We will assume that f is differentiable for simplicity and let $\mathbb{E}[X] = \mu$. Define the linear function centered at μ to be $l(x) \coloneqq f(\mu) + f'(\mu)(x - \mu)$. Then, we know that $f(x) \geq l(x)$ for all x, so

$$\begin{aligned}\mathbb{E}[f(X)] &\geq \mathbb{E}[l(X)] \\ &= \mathbb{E}[f(\mu) + f'(\mu)\,(X - \mu)] \\ &= \mathbb{E}[f(\mu)] + f'(\mu)(\mathbb{E}[X] - \mu) \\ &= \mathbb{E}[f(\mu)] \\ &= f(\mathbb{E}[X])\end{aligned}$$

A nice way to visualize which side is greater (which I tend to always forget) is to think about a Bernoulli(p) distribution. $f(\mathbb{E}[X])$ is visualized to be lower than the region in which the $\mathbb{E}[f(X)]$ must lie.

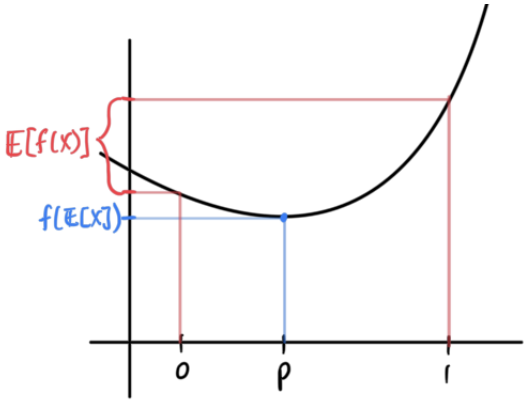


Figure 1: The function f essentially transforms the Bernoulli defined on $0, 1$ to the Bernoulli defined on $f(0), f(1)$. Therefore, $\mathbb{E}[f(X)] \in [f(0), f(1)]$, which lies completely over $f(\mathbb{E}[X])$.

2.4 Variance, Covariance, Correlation

Definition 2.15 (Variance)

Let X be a random variable and suppose $\mathbb{E}[X] < \infty$. The **variance** of X is defined

$$\text{Var}[X] = \sigma_X^2 \coloneqq \mathbb{E}[(X - \mathbb{E}[X])^2] \tag{236}$$

and $\sigma_X = \sqrt{\text{Var}[X]}$ is called the **standard deviation**. This is a measure of how much the probability distribution deviates from its mean. We can use linearity of expectation to write

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}\big[X^2 + \mathbb{E}[X]^2 - 2X\mathbb{E}[X]\big] \\ &= \mathbb{E}[X^2] + \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[X] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

which is often easier to compute, since it only requires us to compute the expectation of X and X^2.

Since variance is always nonnegative, we also know that $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$. The variance is always defined, whether it's finite or $+\infty$.

Likewise for expectation, the variance of a function f w.r.t. μ is

$$\text{Var}_\mu(f) = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 \tag{237}$$

Proposition 2.1 ()

The variance of a random variable X is 0 if and only if it constant almost everywhere on Ω.

Proof.

The if part is easy, so let's prove the only if part. Let $\mathbb{E}[(X - \mathbb{E}[X])^2] = 0$. Then, we can think of the function $x \mapsto (x - \mathbb{E}[X])^2$ and write the variance as

$$\text{Var}[X] = \int_\Omega (X - \mathbb{E}[X])^2 \, d\mathbb{P} = 0 \tag{238}$$

But by nonnegativity of the function, we know that $(X - \mathbb{E}[X])^2 = 0$ w/ probability 1, which implies that $X = \mathbb{E}[X]$ with prob. 1.

Lemma 2.5 (Properties of Variance)

Let X and Y be random variables with well-defined variances.

1. Translation Invariance: Given that $X + a$ is a new random variable defined $(X + a)(\omega) = X(\omega) + a$,
$$\text{Var}[X] = \text{Var}[X + a] \tag{239}$$
2. Quadratic Scaling: Given that aX is a new random variable defined $(aX)(\omega) = a\,X(\omega)$,
$$\text{Var}[aX] = a^2\text{Var}[X] \tag{240}$$

From the properties of expectation and variance, we can now **standardize** a random variable X. If X is a random variable with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \text{Var}(X)$, then the random variable

$$Y = \frac{X - \mu}{\sigma} \tag{241}$$

has mean $\mathbb{E}(Y) = 0$ and variance $\text{Var}(Y) = 1$.

Example 2.10 (Bernoulli)

Given $X \sim \text{Bernoulli}(p)$, we have

$$\mathbb{E}[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = p$$
$$\mathbb{E}[X^2] = 0^2 \cdot \mathbb{P}(X = 0) + 1^2 \cdot \mathbb{P}(X = 1) = p$$

and so $\text{Var}[X] = p - p^2 = p(1 - p)$.

Example 2.11 (Poisson)

Given $X \sim \text{Poisson}(X)$, then

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} \cdot \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \lambda \sum_{k=1}^{\infty} \cdot \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda$$
$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k^2 \cdot \frac{e^{-\lambda}\lambda^k}{k!} = \ldots = \lambda^2 + \lambda$$

So $\text{Var}[X] = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Example 2.12 (Uniform)

Let $X \sim \text{Uniform}[a, b]$. Then,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X \, d\lambda = \int_a^b x \cdot \frac{1}{b-a} \, dx = \frac{a+b}{2}$$
$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f_X \, d\lambda = \int_a^b \frac{x^2}{b-a} \, dx = \frac{a^2+ab+b^2}{3}$$

So $\text{Var}[X] = \ldots = \frac{1}{12}(b-a)^2$. This is consistent with the fact that if we spread out our measure over a wider interval, then the variance will be bigger.

Example 2.13 (Exponential)

Let $X \sim \text{Exp}(\lambda)$. Then, $\mathbb{E}[X] = \frac{1}{\lambda}$ and $\mathbb{E}[X^2] = \frac{2}{\lambda^2}$, so

$$\text{Var}[X] = \frac{1}{\lambda^2} \tag{242}$$

This is consistent with the fact that if λ is greater, then the PDF is much more concentrated at 0, making the variance small.

Just like how we explained that computing finiteness or infiniteness of expectation is similar to multiplying the PMF/PDF by x and determining if the series/integral converges or diverges, we can do the same for variance by multiplying the PMF/PDF by x^2. For a probability distribution of form $\frac{1}{x^2}$, it diverges if we multiply by x and also diverges if we multiply by x^2. But also, we could construct a distribution where the expectation may be finite, but the variance may be infinite. For example, if we have a distribution of form $\frac{1}{x^3}$, multiplying it by x leads to form $\frac{1}{x^2}$, which is finite (so finite expectation), but multiplying by x^2 leads to a harmonic, i.e. infinite variance.

Definition 2.16 (Moment)

The **nth (raw) moment** of a random variable X is $\mathbb{E}[X^n]$. Unlike the raw moment, which is calculated around the origin, the **nth central moment** of X is its moment centered around its mean $\mathbb{E}[(X - \mathbb{E}[X])^n]$.

1. the first moment is the mean $\mathbb{E}[X]$
2. the second central moment is the variance $\mathbb{E}[(X - \mathbb{E}[X])^2]$
3. the third central moment, divided by σ^3, is the skew $\frac{1}{\sigma^3}\mathbb{E}[(X - \mathbb{E}[X])^3]$

The variance is a measure for one random variable X, which measures how much it deviates from its mean. Now, the covariance is defined for two random variables and captures how they jointly vary.

Definition 2.17 (Covariance)

The **covariance** of random variables X and Y is defined as

$$\begin{aligned}\text{Cov}[X,Y] &= \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[X])\big] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

where the intermediate expectations are well-defined. X and Y are said to be **uncorrelated** if

$$\text{Cov}[X,Y] = 0 \tag{243}$$

The covariance is also easy to interpret. Given two random variables X and Y, if whenever X is greater than its expected value $\mathbb{E}[X]$, Y also tends to be greater than $\mathbb{E}[Y]$, then the covariance will be some positive number. If they tend to be on opposite sides of their expected values, then the covariance will be negative. And the degree with which these RVs lie on which side of the expected value determines the magnitude of the covariance.

Theorem 2.18 ()

If X and Y are independent random variables, then they are uncorrelated, meaning that independence is a stronger condition.

We show an example of why the converse is not true. Consider $X \sim \text{Uniform}[-1,1]$. We can show that x and $Y = X^2$ are dependent but uncorrelated. It is clearly dependent, but its covariance is

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \\ &= \mathbb{E}[X^3] - \mathbb{E}[X]\,\mathbb{E}[X^2] \\ &= \int_{-1}^{1} x^3 \cdot 1\,dx - 0 \cdot \mathbb{E}[X^2] = 0\end{aligned}$$

Theorem 2.19 (Variance of Sums of Random Variables)

If X and Y are two random variables, then

$$\text{Var}(X+Y) = \text{Var}[X] + \text{Var}(Y) + 2\text{Cov}(X,Y) \tag{244}$$

and by induction, we can show that

$$\text{Var}\bigg(\sum_i X_i\bigg) = \sum_i \text{Var}(X_i) + \sum_{i,j} \text{Cov}(X_i, X_j) \tag{245}$$

Proof.

Simple computation. The LHS expands to

$$\begin{aligned}\mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 &= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\ &= \big(\mathbb{E}[X^2] - \mathbb{E}[X]^2\big) + \big(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2\big) + 2\big(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\big) \\ &= \text{Var}[X] + \text{Var}(Y) + 2\text{Cov}(X,Y)\end{aligned}$$

Therefore, if we have n random variables $X_1, \dots, X_n$, then we can compute their pairwise covariance

$\text{Cov}(X_i, X_j)$ and compute their **covariance matrix** $\mathbf{\Sigma}$, which is an $n \times n$ symmetric matrix with entries

$$\mathbf{\Sigma}_{ij} = \text{Cov}(X_i, X_j) \text{ for } i, j = 1, \dots, n \tag{246}$$

Theorem 2.20 (Simple Bound on Covariance)

If X and Y are two random variables with finite variance, then the magnitude of their covariance is bounded by the following inequality.

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\ \text{Var}(Y)} = \text{std}(X)\ \text{std}(Y) \tag{247}$$

Finally we define the correlation.

Definition 2.18 (Correlation Coefficient)

The **correlation coefficient** of random variables X and Y is defined

$$\rho_{X,Y} = \text{Corr}(X, Y) \coloneqq \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]\,\text{Var}(Y)}} \tag{248}$$

By definition, this implies that $-1 \leq \text{Corr}(X, Y) \leq 1$. When $\text{Corr}(X, Y) > 0$ (which also means that $\text{Cov}(X, Y) > 0$), it is said that X and Y are *positively correlated*, and when $\text{Corr}(X, Y) < 0$ (which also means that $\text{Cov}(X, Y) < 0$), it is said that they are *negatively correlated*.

Theorem 2.21 ()

$\text{Corr}(X, Y) = \pm 1$ indicates a linear relationship between X and Y.

1. Let $\text{Corr}(X, Y) = 1$. Then, there exists a $m > 0$ and $b \in \mathbb{R}$ such that $Y = mX + b$.
2. Let $\text{Corr}(X, Y) = -1$. Then, there exists a $m < 0$ and $b \in \mathbb{R}$ such tat $Y = mX + b$.

This implies that $\text{Corr}(X, Y) = \pm 1$ indicates that the joint distribution of (X, Y) is concentrated on a line in $\mathbb{R}^2$.

2.4.1 Hilbert Space of Random Variables

In some sense the correlation is a scaled version of the covariance. It is scale-invariant, and it is always a number that lies between -1 and 1, making it a nice way to represent the correlation between two variables without having to worry about scale. We can prove this.

Theorem 2.22 (Cauchy-Schwartz)

For any two random variables X, Y, we have $|\text{Cov}(X, Y)| \leq \sigma_X \sigma_Y$, or in other words,

$$-1 \leq \rho_{X,Y} \leq 1 \tag{249}$$

Furthermore, whenever $\rho_{X,Y} = 1$ or -1, there exists a deterministic relationship between X and Y.

1. If $\rho_{X,Y} = 1$, there exists a $a > 0$ s.t.

$$Y - \mathbb{E}[Y] = a(X - \mathbb{E}[X]) \tag{250}$$

2. If $\rho_{X,Y} = -1$ there exists a $a < 0$ s.t.

$$Y - \mathbb{E}[Y] = a(X - \mathbb{E}[X]) \tag{251}$$

This implies that $\text{Corr}(X, Y) = \pm 1$ indicates that the joint distribution of (X, Y) is concentrated on a line in $\mathbb{R}^2$.

The fact that this is called the Cauchy-Schwartz inequality hints at the existence of inner products, norms, and vector spaces. That is, we can treat the random variables X, Y as vectors in the functional space of real-valued maps over Ω. In some sense, $\text{Cov}(X, Y)$ sort-of plays the role of an inner product.

1. It satisfies symmetricity:

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[YX] - \mathbb{E}[Y]\,\mathbb{E}[X] = \text{Cov}(Y, X) \tag{252}$$

2. It satisfies binlinearity. It suffices to show only for first argument, since we have symmetricity.

$$\begin{aligned}\text{Cov}(aX + bY, Z) &= \mathbb{E}[(aX + bY)Z] - \mathbb{E}[aX + bY]\,\mathbb{E}[Z] \\ &= a\mathbb{E}[XZ] + b\mathbb{E}[YZ] - a\mathbb{E}[X]\mathbb{E}[Z] - b\mathbb{E}[Y]\,\mathbb{E}[Z] \\ &= a\big(\mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z]\big) + a\big(\mathbb{E}[YZ] - \mathbb{E}[Y]\,\mathbb{E}[Z]\big) \\ &= a\,\text{Cov}(X, Z) + b\,\text{Cov}(Y, Z)\end{aligned}$$

3. We want the inner product of X with itself to always be greater than 0, with equality holding iff $X = 0$. Indeed, we have

$$\text{Cov}(X, X) = \text{Var}[X] \geq 0 \tag{253}$$

but it is not necessarily true that $\text{Var}[X] = 0 \implies X = 0$. We can say that X is equal to a constant almost everywhere at best. We can solve this problem by looking at the functional subspace of 0-mean random variables (which is a vector space due to linearity of expectation). So now all random variables X that are 0 almost everywhere have inner product 0, so we must add an equivalence class on this subspace that says two X, Y are equivalent if they agree almost everywhere.

The standard deviation σ_X and σ_Y act as norms on this quotient subspace of 0-mean random variables. So the correlation coefficient $\rho_{X,Y}$ can be interpreted as the cosine of the angle between X and Y. This now makes our desired space a Hilbert space, and our uncorrelated random variables are like orthogonal vectors.

Definition 2.19 ()

Let $L^2_{\mathcal{F}}(\Omega)$ be the function space consisting of equivalence classes of 0-mean random variables $X : (\Omega, \mathcal{F}) \to \mathbb{R}$ that are almost surely equal. Then,

1. we can define the inner product on this space as

$$\langle X, Y \rangle \coloneqq \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] = \int_\Omega XY\, d\mathbb{P} \tag{254}$$

2. which induces the L^2-norm on this space defined

$$||X||_2 \coloneqq \text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] = \int_\Omega X^2\, d\mathbb{P} \tag{255}$$

We set $L^2_{\mathcal{F}}(\Omega)$ to be a Banach space with bounded norm $\mathbb{E}[X^2] < \infty$.

3 Convergence

3.1 Borel-Cantelli Lemmas

There are many Borel-Cantelli lemmas, and we will introduce the two most famous ones. To understand what these lemmas say, given a sequence $A_1, A_2, \dots$ of events in σ-algebra $\mathcal{F}$, we must first understand what the daunting term

$$\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i \tag{256}$$

means. Now let's try to explain what the intersection of the unions mean. First, remember that σ-algebras are stable under both countable unions and countable intersections, this is also in $\mathcal{F}$. We can interpret

$$\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i = \{A_n \text{ i.o.}\} \tag{257}$$

as the *event that infinitely many A_n's occur*, where i.o. means "infinitely often." To parse this, let's start from the innermost term and call it

$$B_n = \bigcup_{i=n}^{\infty} A_i \implies \{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty} B_n \tag{258}$$

B_n is the event that at least one of the $A_n, A_{n+1}, A_{n+2}, \dots$ occurs, often referred to as the *nth tail event*. Now the intersection of all B_n's is the event that *all* B_n's occur. In other words, this is the event that for no matter how big of an $N \in \mathbb{N}$ I choose, there is always at least an event A_n with $n > N$ that occurs. This is shortly summarized as the event that infinitely many A_n's occur.

Lemma 3.1 (1st Borel-Cantelli Lemma)

Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, if $A_1, A_2, \dots$ is a sequence of events such that

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \tag{259}$$

the almost surely (with probability 1) only finitely many A_n's will occur.

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i \right) = 0 \tag{260}$$

Proof.

Setting B_n as above, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n \right) &= \lim_{n\to\infty} \mathbb{P}(B_n) && \text{(continuity of probability)} \\
&= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \right) && \text{(substitute } B_i\text{)} \\
&\leq \lim_{n\to\infty} \sum_{i=n}^{\infty} \mathbb{P}(A_i) = 0 && \text{(tail sum of convergent series is 0)}
\end{aligned}$$

The second Borel-Cantelli lemma is like a partial contrapositive to the first lemma, where it starts with the assumption that the sum of the $\mathbb{P}(A_n)$'s are infinite (along with the addition case that they are independent).

Lemma 3.2 (2nd Borel-Cantelli Lemma)

If $A_1, A_2, \ldots$ are independent events such that

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty, \tag{261}$$

then almost surely (with probability 1) infinitely many A_n's will occur. That is,

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i \right) = 1 \tag{262}$$

The intuition behind this lemma is challenging: We can let $\mathbb{P}(A_n) = P_n$ and interpret the sum as a series of P_n's. Since the series $P_1 + P_2 + \ldots$ is finite, this implies that

$$\lim_{n \to \infty} P_n = 0 \tag{263}$$

(but not the converse) and going to zero rather fast such that the series is finite. So, you are working with a sequence of events A_n that are becoming more and more unlikely rather fast. The lemma says that beyond a certain point n_0, none of the events A_n will occur almost surely. For the second lemma, we can go as far as we like in the sequence of A_n's, up to any A_{n_0}, but beyond that there is always an infinite number of A_n's that occur beyond A_{n_0}.

3.2 Transforms

3.2.1 Probability Generating Function (PGF)

The PGF is only defined for discrete random variable, and is analogous to the Z-transform in singal processing.

Definition 3.1 (Probability Generating Function)

Let X be a discrete random variable taking values in $\mathbb{N}_0$. Then, the **probability generating function** of X is defined

$$G_X(z) \coloneqq \mathbb{E}[z^X] = \sum_{i=0}^{\infty} z^i \, \mathbb{P}(X = i) \tag{264}$$

Now there is the problem of convergence, but we will not pay attention to this technicality for now and just consider the PGF as a tool.

Example 3.1 (PGF of Poisson)

The random variable $X \sim \text{Poisson}(\lambda)$ has pmf $\mathbb{P}(X = i) = \frac{e^{-\lambda} \lambda^i}{i!}$ for $i \in \{0, 1, \ldots\}$. Then,

$$G_X(z) = \mathbb{E}[z^X] = \sum_{i=0}^{\infty} z^i \, \frac{e^{-\lambda} \lambda^i}{i!} = \sum_{i=0}^{\infty} \frac{e^{-\lambda} (\lambda z)^i}{i!} = e^{\lambda(z-1)} \tag{265}$$

Example 3.2 (PGF of Geometric)

For $X \sim \text{Geometric}(p)$, its PGF is

$$G_X(z) = \sum_{i=1}^{\infty} z^i \, (1-p)^i p = \frac{pz}{1 - z(1-p)} \tag{266}$$

Lemma 3.3 (Properties of PGF)

Given random variable X and its PGF G_X, we have the following:

1. Evaluate at $z = 1$:
$$G_X(1) = \mathbb{E}[1^X] = \mathbb{E}[1] = 1 \tag{267}$$
2. Derivative at $z = 1$:
$$\left.\frac{dG_X(z)}{dz}\right|_{z=1} = \mathbb{E}[X] \tag{268}$$
3. kth derivative at $z = 1$:
$$\left.\frac{d^k G_X(z)}{dz^k}\right|_{z=1} = \mathbb{E}[X(X-1)(X-2)\dots(X-k+1)] \tag{269}$$
4. Transformation: Given the sum $Z = X + Y$ (where X, Y are independent), rather than computing its convolution, the PGF of Z is simply the product of the PGFs of X and Y:
$$G_Z(z) = G_X(z)\, G_Y(z) \tag{270}$$
For example, since a Poisson(λ) random variable has PGF of form $e^{\lambda(z-1)}$, if we have two Poissons X and Y with parameters λ, μ, then we can easily multiply their PGFs to get the PGF of $Z = X + Y$, which is $e^{(\lambda+\mu)(z-1)}$, which is the PGF of a Poisson($\lambda + \mu$) random variable.

3.2.2 Moment Generating Function (MGF)

Definition 3.2 (Moment Generating Function (MGF))

The **moment generating function** associated with a random variable X is a function $M_X : \mathbb{R} \longrightarrow [0, \infty]$ defined
$$M_X(s) \coloneqq \mathbb{E}[e^{sX}] \tag{271}$$
It is like an exponential moment. The region of convergence of M_X is the set $D_X = \{s \mid M_X(s) < \infty\}$. and we always have $M_X(0) = 1$, so $0 \in D_X$ always.

Lemma 3.4 (Properties of MGF)

Let X be a random variable with MGF $M_X(s)$.

1. $M_X(0) = 1$, so 0 is always in the region of convergence.
2. If $Y = aX + b$, then
$$M_Y(s) = e^{bs} M_X(as) \tag{272}$$
3. If X and Y are independent and $Z = X + Y$, then
$$M_Z(s) = M_X(s)\, M_Y(s) \tag{273}$$

Proof.

Listed.

1. $M_X(0) = \mathbb{E}[e^{0X}] = \mathbb{E}[1] = 1$.

2. We have

$$\begin{aligned} M_Y(s) &= \mathbb{E}[e^{s(aX+b)}] \\ &= \mathbb{E}[e^{asX} e^{bs}] \\ &= \mathbb{E}[e^{(as)X}]\,\mathbb{E}[e^{bs}] \\ &= e^{bs} M_X(as) \end{aligned}$$

where the penultimate step was due to independence of constant RV with any other RVs.

3. We can see

$$M_Z(s) = \mathbb{E}[e^{s(X+Y)}] = \mathbb{E}[e^{sX}\, e^{sY}] = \mathbb{E}[e^{sX}]\,\mathbb{E}[e^{sY}] = M_X(s)\, M_Y(s) \tag{274}$$

since X, Y independent means that any function of X and Y are independent.

Theorem 3.1 (Inversion Theorem)

Suppose $M_X(s)$ is finite for all $s \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$. Then, M_X uniquely determines the CDF of X. This implies that if X and Y are random variables such that $M_X(s) = M_Y(s)$ for all $s \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$, then X and Y have the same CDF.

This theorem is useful for comparing random variables with the MGFs, but a limitation is that it is not always clear that the MGF is defined beyond 0. Now, we explain why this is called a moment generating function.

Theorem 3.2 (Moment Generating Property)

Suppose $M_X(s) < \infty$ for $s \in [-\epsilon, \epsilon]$ with $\epsilon > 0$. Then, the derivatives at $s = 0$ generate the moments of X:

$$\frac{d^m M_X(s)}{ds^m}\bigg|_{s=0} = \mathbb{E}[X^m] \tag{275}$$

Proof.

A hand-wavy proof is that we can take the derivative and put it "in" the expectation.

$$\frac{d}{ds}\mathbb{E}[e^{sX}] = \mathbb{E}\Big[\frac{d}{ds} e^{sX}\Big] = \mathbb{E}[X e^{sX}] \tag{276}$$

which evaluates to $\mathbb{E}[X]$ when $s = 0$. Differentiating m times just gets $\mathbb{E}[X^m e^{sX}]$. However, this should be questioned, since the expectation is an integral and we are putting the derivative inside the integral.

Example 3.3 (Exponential RV)

The PDF of $X \sim \text{Exponential}(\mu)$ is $f_X(x) = \mu e^{-\mu x}$ for $x \geq 0$. The MGF is

$$M_X(s) := \int_0^\infty \mu e^{-\mu x} e^{sx}\, dx = \begin{cases} \frac{\mu}{\mu - s} & \text{for } s < \mu \\ \infty & \text{for } s \geq \mu \end{cases} \tag{277}$$

Example 3.4 (Gaussian RV)

The PDF of a standard Gaussian X is $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for $x \in \mathbb{R}$, and the MGF is

$$M_X(s) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{sx}\, dx = e^{s^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-s)^2}{2}}\, dx = e^{s^2/2} \tag{278}$$

which is valid for all $s \in \mathbb{R}$.

Example 3.5 (Cauchy RV)

If we have $f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}$ for $x \in \mathbb{R}$, the MGF is

$$M_X(s) = \int_{-\infty}^{\infty} \frac{e^{sx}}{\pi(1+x^2)}\, dx = \begin{cases} 1 & \text{if } s = 0 \\ \infty & \text{if } s > 0 \\ \infty & \text{if } s < 0 \end{cases} \tag{279}$$

So the region of convergence is just $\{0\}$. It is infinity everywhere else since the exponential function grows exponentially as $x \to \pm\infty$.

Example 3.6 ()

Given $X_1 \sim \text{Exponential}(\lambda_1)$ and $X_2 \sim \text{Exponential}(\lambda_2)$ are independent, the MGF of $Z = X_1 + X_2$ is

$$M_Z(s) = M_X(s)\, M_Y(s) = \frac{\lambda_1 \lambda_2}{(\lambda_1 - s)(\lambda_2 - s)} \text{ for } s < \min\{\lambda_1, \lambda_2\} \tag{280}$$

and we can perform our inverse transform on it.

3.2.3 Characteristic Function

We can see that the MGF has its limitations: for some random variables (like the Cauchy), its MGF was not defined at all beyond $\{0\}$. On the contrary, the characteristic function is always defined everywhere and is finite everywhere (in fact, is bounded by 1, shown below). Also, it is a bit easier to invert (similar to how the Fourier transform is a bit easier to invert than the Laplace).

Definition 3.3 (Characteristic Function)

Given a random variable $X : \Omega \longrightarrow \mathbb{R}$, the **characteristic function** is defined to be

$$\begin{aligned} \varphi_X(t) &= \mathbb{E}[e^{itX}] \\ &= \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)] \end{aligned}$$

If X admits a PDF, then the characteristic function is its Fourier transform with a small sign reversal.

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} f_X(x)\, dx \tag{281}$$

Theorem 3.3 (Properties of CF)

Let X be a random variable with CF $\varphi_X(t)$.

1. $\varphi_X(0) = 1$ and $|\varphi_X(t)| \leq 1$ for all $t \in \mathbb{R}$.

2. If $Y = aX + b$, then
$$\varphi_Y(t) = e^{ibt}\varphi_X(at) \tag{282}$$
3. If X and Y are independent random variables and $Z = X + Y$, then
$$\varphi_Z(t) = \varphi_X(t)\,\varphi_Y(t) \tag{283}$$
4. $\varphi_X(t)$ is uniformly continuous on $\mathbb{R}$, i.e. for all $t \in \mathbb{R}$, there exists a $\phi(h) \downarrow 0$ as $h \downarrow 0$ such that
$$|\varphi_X(t+h) - \varphi_X(t)| \leq \phi(h) \tag{284}$$
5. φ_X is a nonnegative-definite kernel, i.e. for any n reals $t_1, \dots, t_n$ and n complex numbers $z_1, \dots z_n$, we have
$$\sum_{i,j} z_i \varphi_X(t_i - t_j)\,\bar{z}_j \geq 0 \tag{285}$$

Proof.

Listed.

1. We just set $\varphi_X(0) = \mathbb{E}[e^{i0X}] = \mathbb{E}[1] = 1$, and for continuous random variables, we can bound
$$\begin{aligned} |\varphi_X(t)| &= \left| \int_{-\infty}^{\infty} e^{itx} f_X(x)\,dx \right| \\ &\leq \int_{-\infty}^{\infty} \left| e^{itx} f_X(x) \right| dx \\ &\leq \int_{-\infty}^{\infty} \left| e^{itx} \right| \cdot \left| f_X(x) \right| dx \\ &= \int_{-\infty}^{\infty} f_X(x)\,dx = 1 \end{aligned}$$
2. We have
$$\varphi_Y(t) = \mathbb{E}[e^{it(aX+b)}] = \mathbb{E}[e^{iatX}\,e^{ibt}] = \mathbb{E}[e^{i(at)X}]\,\mathbb{E}[e^{ibt}] = e^{ibt}\varphi_X(at) \tag{286}$$
3. We have
$$\varphi_Z(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}\,e^{itY}] = \mathbb{E}[e^{itX}]\,\mathbb{E}[e^{itY}] = \varphi_X(t)\,\varphi_Y(t) \tag{287}$$
4. We have
$$\begin{aligned} |\varphi_X(t+h) - \varphi_X(t)| &= |\mathbb{E}[e^{itX}(e^{ihX} - 1)]| \\ &\leq \mathbb{E}[|e^{itX}(e^{ihX} - 1)|] \\ &\leq \mathbb{E}[|e^{ihX} - 1|] \dots \end{aligned}$$

Now this next theorem states the uniqueness of each characteristic function. It is a highly nontrivial result.

Theorem 3.4 (Inversion Theorem)

If two random variables have the same characteristic function, then their CDFs are the same. Further, if X is a continuous random variable, then the PDF can be recovered from the characteristic function as follows:
$$f_X(x) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} e^{-itx} \varphi_X(t)\,dt \tag{288}$$
for every x where $f_X(x)$ is continuous.

Just like how we can recover moments from the MGF, we can always recover the moments from the characteristic function, with the added advantage that the CF will always exist.

Theorem 3.5 (Moment Generating Property)

Let X be a random variable and $\varphi_X(t)$ its CF.

1. If $\varphi_X^{(k)}(t)$ (the kth derivative) exists at $t = 0$, then

$$\mathbb{E}[|X^k|] < \infty \text{ for } k \text{ even}$$
$$\mathbb{E}[|X^{k-1}|] < \infty \text{ for } k \text{ odd}$$

2. If $\mathbb{E}[|X^k|] < \infty$, then
$$\varphi_X^{(k)}(0) = i^k \mathbb{E}[X^k] \tag{289}$$

3. Further, given that the moments are finite, we can expand the CF by moments of X as
$$\varphi_X(t) = \sum_{j=0}^{k} \frac{\mathbb{E}[X^j]}{j!}(it)^j + o(t^k) \tag{290}$$

Example 3.7 (Bernoulli)

Given $X \sim \text{Bernoulli}(p)$, we have

$$\begin{aligned}\mathbb{E}[e^{itX}] &= \sum_{x \in \{0,1\}} e^{itx} \cdot \mathbb{P}(X = x) \\ &= e^{it0}(1-p) + e^{it}p \\ &= 1 - p + pe^{it}\end{aligned}$$

Example 3.8 (Exponential)

Given $X \in \text{Exponential}(\lambda)$, we have

$$\begin{aligned}\mathbb{E}[e^{itX}] &= \int_0^\infty e^{itx} \lambda e^{-\lambda x}\, dx \\ &= \int_0^\infty \lambda\, e^{-(\lambda - it)x}\, dx \\ &= \frac{\lambda}{\lambda - it} \text{ for all } t \in \mathbb{R}\end{aligned}$$

where the complex integral requires some complex analysis.

3.3 Convergence of Random Variables

Unlike convergence of numbers, which is well-defined with respect to some metric or topology, there are many types of convergence of random variables. We must always specify which convergence when talking about them. Remember that a random variable X is just a function from Ω to $\mathbb{R}$, so we can talk about pointwise convergence. That is, given a sequence of random variables $\{X_n\}$ and some $\omega \in \Omega$, the sequence

$$X_1(\omega), X_2(\omega), X_3(\omega), \dots \tag{291}$$

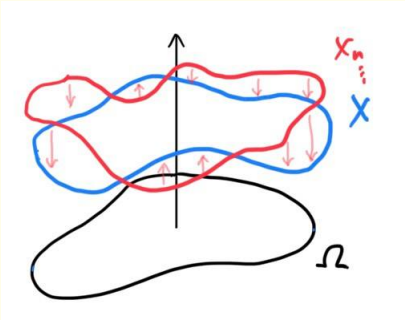
is simply a sequence of real numbers. If this sequence converges to the real number $X(\omega)$, then X_n converges to X at ω. If this occurs for all $\omega \in \Omega$, then we have sure convergence, and if this happens for an event (a $\mathcal{F}$-measurable subset of Ω) with probability 1, then we have almost sure convergence.

Definition 3.4 (Sure Convergence of RVs)

The sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to **converge pointwise** or **converge surely** to X if

$$X_n(\omega) \to X(\omega) \tag{292}$$

for every $\omega \in \Omega$. That is, we can choose *any* $\omega \in \Omega$, and the realized sequence $X_1(\omega), X_2(\omega), \dots$ will always converge to $X(\omega)$. We can visualize the function X with a surface defined over Ω and can imagine the X_n's as surfaces that converges to that of X.



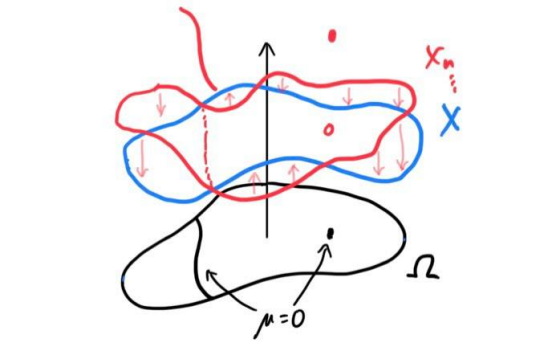
But this definition is too strong of a form of convergence, since in probability we don't care about values over sets of measure 0. That is, if we have two probability distributions that differ from each other on a set of measure 0, then they can be considered essentially the same probability distribution.

Definition 3.5 (Almost Sure Convergence of RVs)

The sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to **converge almost surely** or **converge with probability 1** to X if $X_n(\omega) \to X(\omega)$ on a subset of probability 1. That is,

$$\mathbb{P}\big(\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\big) = 1 \tag{293}$$

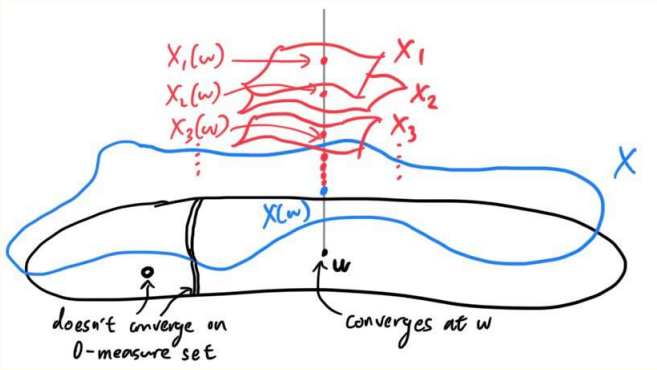
Considering small technicalities, it can be shown that this set of ω's can be considered an event in $\mathcal{F}$. This can be visualized similarly as sure convergence, but now the surfaces don't have to converge on sets of measure 0.



Crudely put, we just have to look at each $\omega \in \Omega$, see if $X_n(\omega)$ converges to $X(\omega)$ as $n \to \infty$, and determine if the set of all ω's that satisfy this have probability 1. In other words, let us have some experiment with outcome space Ω. With probability 1, some $\omega \in \Omega$ will be realized, which will realize the sequence of realized random variables

$$X_1(\omega), X_2(\omega), X_3(\omega), \dots \tag{294}$$

that will converge to $X(\omega)$. Visually, we can imagine selecting a random point in Ω, which will not hit the curve or point (with probability 1), and in these cases, the sequence of points will converge to $X(\omega)$.

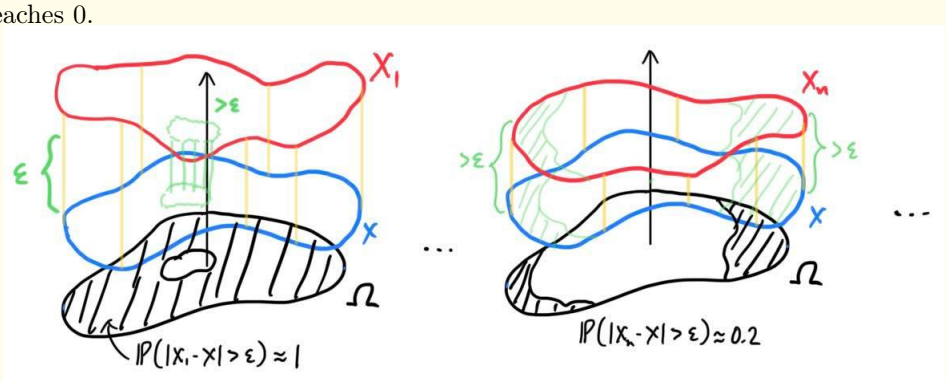


Definition 3.6 (Convergence in Probability)

The sequence of random variables $\{X_i\}_{i\in\mathbb{N}}$ is said to **converge to X in probability** if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| > \epsilon\big) = 0 \tag{295}$$

To understand what this means, fix an $\epsilon > 0$. Then, X_1 may be very far from X, meaning that the event $|X_1 - X| > \epsilon$, i.e. the set of all $\omega \in \Omega$ satisfying $|X_1(\omega) - X(\omega)| > \epsilon$ may be a larger portion of Ω. Now, as we increase n, this event will become smaller (in the way that it's probability decreases) until it reaches 0.

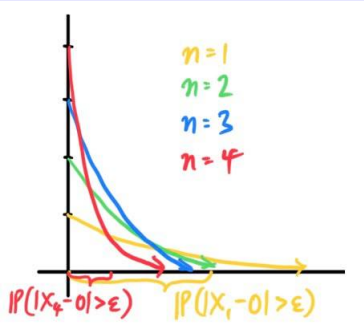


Example 3.9 ()

Given $X_n \sim \text{Exponential}(n)$ with $f_{X_n}(x) = ne^{-nx}$, we show that the sequence converges in probability to the 0 random variable. Given $\epsilon > 0$, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \epsilon) &= \lim_{n\to\infty} \mathbb{P}(X_n > \epsilon \cup X_n < -\epsilon) \\
&= \lim_{n\to\infty} \mathbb{P}(X_n > \epsilon) \\
&= \lim_{n\to\infty} \int_\epsilon^\infty ne^{-nx}\,dx \\
&= \lim_{n\to\infty} e^{-n\epsilon} = 0
\end{aligned}$$

We can imagine this since given any small $\epsilon > 0$, we can see that increasing n results in the distribution of X_n to decrease at a faster rate, and thus a bigger portion of the distribution would lie within ϵ of the 0 random variable.



Definition 3.7 (Convergence in rth Mean)

We say X_n **converges to** X **in the** r**th mean** if

$$\lim_{n\to\infty} \mathbb{E}\big[|X_n - X|^r\big] = 0 \tag{296}$$

For $r = 2$, X_n is said to converge to X in the **mean-squared sense**.

Definition 3.8 (Convergence in Distribution)

We say X_n **converges to** X **in distribution** if the CDF of X_n converges pointwise to the CDF of X, i.e.

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{297}$$

for all x where F_X is continuous.

So for practical purposes there are 5 notions of convergence that we will work with:

1. Sure convergence: $X_n \xrightarrow{p.w.} X$
2. Almost sure convergence: $X_n \xrightarrow{a.s.} X$
3. Convergence in probability: $X_n \xrightarrow{i.p.} X$
4. Convergence in rth mean: $X_n \xrightarrow{rth} X$ (Mean square: $X_n \xrightarrow{m.s.} X$)
5. Convergence in distribution: $X_n \xrightarrow{D} X$

Theorem 3.6 (Hierarchy of Convergence)

The following implications hold:
1. Pointwise c. $\implies$ almost sure c. $\implies$ c. in probability $\implies$ c. in distribution.
2. rth mean c. $\implies$ c. in probability $\implies$ c. in distribution.

Trying to understand these relationships can be very hard, so we will take some time to do that, with some examples. First, convergence in distribution is clearly the weakest, since convergence in distribution does not imply that the random variables need be close to each other. Take a look at the random variables $X \sim \text{Bernoulli}(1/2)$ and $Y = 1 - X$. X and Y are both Bernoulli(1/2) with the same distribution, but they are *not* the same random variable since $X - Y = 1$ always. Therefore, we can think of two random variables that have the same distribution but are not "close" to each other as functions over Ω that divide it into identical, but differently cut, distributions.

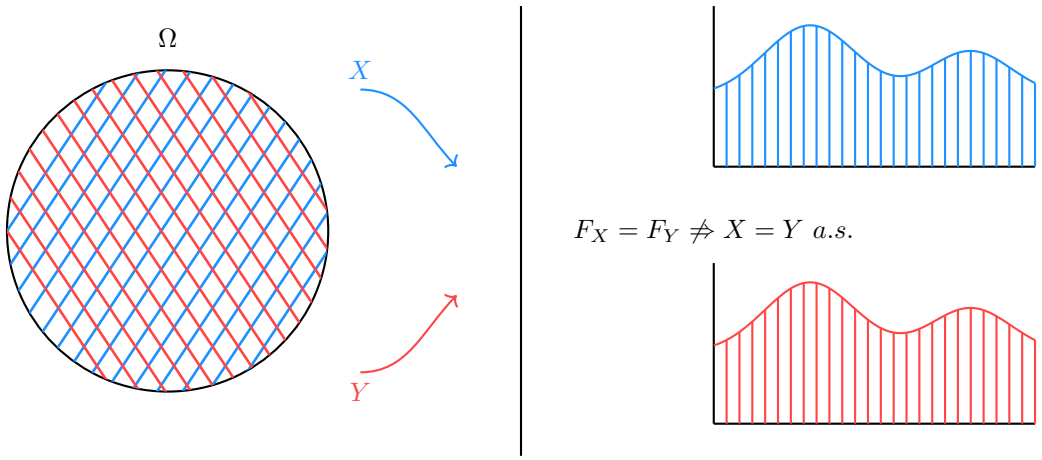


Figure 2: Two identical distributions may not model the same phenomenon.

Example 3.10 (C. in Distribution $\not\Rightarrow$ C. in Probability)

Let $X_1, X_2, \dots$ be such that $X_i = X$ for all i where $X \sim \text{Bernoulli}(1/2)$. This does not mean that the X_i's are iid Bernoulli; they are all copies of the same X, i.e. forms a constant sequence. Let $Y = 1 - X$. Clearly, $X_n \xrightarrow{D} Y$ since the CDF of every X_i is the same as that of Y, but $|X_n - Y| = 1$ for all n, so there is no convergence.

Example 3.11 (C. in Distribution $\not\Rightarrow$ C. in Probability)

Let $X_1, X_2, \dots \sim \mathcal{N}(0, 1)$ and $Y = -X$. Then, by symmetry of the standard Gaussian, both X and Y have the same CDF, but they are not the same random variable: their signs are opposite.

3.3.1 Convergence in Probability vs Almost Surely

Convergence almost surely and convergence with probability are very different. Almost sure convergence has the limit inside the probability, which indicates that we are talking about convergence of a sequence of random variables. On the other hand, convergence in probability has the limit on the outside, which talks about convergence of a sequence of probabilities. But a key point is that almost sure convergence implies convergence in probability. It happens so because there could exist a subset of small probability in Ω where the X_n's and X need not be close, but the probabilities of them deviating over whole Ω is small.

Example 3.12 (C. in Probability $\not\Rightarrow$ C. Almost Surely)

Consider the interval $\Omega = [0, 1]$ and the subsets $A_1 = [0, 0.1], A_2 = [0.1, 0.2], \dots$, such that at $A_{10} = [0.9, 1.0]$, the size with halve and will go to the left boundary, $A_{11} = [0, 0.05], \dots$. Then, the sequence of indicator random variables

$$X_n := 1_{A_n} \tag{298}$$

looks like it's converging to the 0 random variable. Indeed, $X_n \xrightarrow{i.p.} 0$ since the probability that X_n deviates from 0 by more than some small ϵ is simply the measure of A_n itself, which decreases to 0. That is, given some small $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \epsilon) = \lim_{n\to\infty} \mathbb{P}(1_{A_n} > \epsilon) = \lim_{n\to\infty} \mathbb{P}(A_n) = 0 \tag{299}$$

Now let's show that this doesn't converge almost surely. For *any* outcome $\omega \in \Omega$, the sequence of random variables $X_1(\omega), X_2(\omega), \dots$ will hit these intervals A_n infinitely many times and will not

converge to 0, since there will always be a 1 down the sequence. They will occur with decreasing frequency but they will always occur. Therefore, with probability 1, whatever realized sequence will not converge to the 0 random variable.

Here is another standard counterexample.

Example 3.13 (C. in Probability $\not\Rightarrow$ C. Almost Surely)

Let us take the sequence $X_1, X_2, \dots$ of independent random variables where $X_n \sim \text{Bernoulli}(1/n)$. That is,

$$\mathbb{P}(X_n = 1) = \frac{1}{n} \text{ and } \mathbb{P}(X_n = 0) = 1 - \frac{1}{n} \tag{300}$$

So, as n gets large we expect X_n to realize values of 0 more and more. Showing that $X_n \xrightarrow{i.p.} 0$ is easy, since we can compute for any $\epsilon > 0$

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \epsilon) &= \lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) \\ &= \lim_{n\to\infty} \mathbb{P}(X_n = 1) \\ &= \lim_{n\to\infty} \frac{1}{n} = 0 \end{aligned}$$

We want to show that this does not converge almost surely to 0, i.e. there is some set of nonzero measure such that for some ω in that set, the sequence $X_1(\omega), X_2(\omega), \dots$ does not converge to 0. This can be hard to see at first, but the fact that we have independence and the terms are $\frac{1}{n}$ hints at the Borel-Cantelli lemma. Let A_n be the event that $\{X_n = 1\}$ (i.e. the preimage of the singleton set under X_n: $X_n^{-1}(\{1\})$). Then, the A_n's are independent, and

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = +\infty \tag{301}$$

By the Borel-Cantelli lemma 2, this implies that almost surely infinitely many A_n's will occur. That is, we can choose as large of an n as we like, go down the sequence until we look at $X_n, X_{n+1}, \dots$, and we are guaranteed with probability 1 that at least one of the X_i's after n will realize a 1. This means that in every realization of $X_1, X_2, \dots$, we will get a sequence of 0s and 1s, but since BCL states that no matter how far down the road you will always get at least another 1, this sequence does not converge to 0.

The commonality between these two examples is that sequence of random variables satisfies convergence in probability as follows: As n increases, X_n is more and more likely to be near X (in the way that $|X_n - X| < \epsilon$ for some $\epsilon > 0$), ultimately satisfying this closeness property with probability 1 as $n \to \infty$. For example, we could have

$$\begin{aligned} \mathbb{P}(|X_1 - X| > \epsilon) &= 1 \\ \mathbb{P}(|X_2 - X| > \epsilon) &= 1/2 \\ \mathbb{P}(|X_3 - X| > \epsilon) &= 1/3 \\ \dots &= \dots \end{aligned}$$

This definitely satisfies convergence in probability, but this leaves open the possibility that $\mathbb{P}(|X_n - X| > \epsilon)$ an infinite number of times, although at infrequent intervals. Therefore, when looking at the sequence

$$X_1, X_2, X_3, \dots \tag{302}$$

each random variable *individually* may have less chance of being more than ϵ away from X, but since there is an infinite number of them in the sequence, the sequence *in totality* may contain an infinite number of

cases where $|X_n - X| > \epsilon$. Convergence almost surely tells us that we are *guaranteed* (with probability 1) that this sequence will converge to X. That is, we can specify an $N \in \mathbb{N}$ such that $|X_n - X| < \epsilon$ for all $n > N$.

Let us define some $\epsilon > 0$ and consider a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$. Given some outcome $\omega \in \Omega$, we will consider it a *success* if $|X_n(\omega) - X(\omega)| < \epsilon$ and *failure* if not. Then, convergence in probability tells us that the probability of failure goes to 0 as n goes to infinity and therefore we get better and better estimates of X. Convergence almost surely is a bit stronger and says that the total number of failures is *finite*. That is, after a certain point N, the random variable X_n will *always* estimate X within an error of ϵ (i.e. such that $|X_n - X| < \epsilon$). But since you don't know when you've exhausted all failures, there is not much of a difference from a practical point of view.

3.3.2 Complete Convergence

When proving almost sure convergence, we'd ideally just look at all the $\omega \in \Omega$ where $X_n(\omega) \to X(\omega)$, and if this set has probability measure 1, then we are done. But this is not very practical, so we use the following theorem, which gives a sufficient condition for $X_n \xrightarrow{a.s.} X$.

Theorem 3.7 ()

If for all $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \epsilon) < \infty \tag{303}$$

then $X_n \xrightarrow{a.s.} X$. This condition is a bit stronger, since not only are we saying that $\mathbb{P}(|X_n - X| > \epsilon)$ tends to 0 as $n \to \infty$, but that it goes down fast enough to keep the series convergent.

Proof.

Let the event that $|X_n - X| > \epsilon$ be denoted $A_n(\epsilon)$ (i.e. the preimage of (ϵ, ∞) under the map $|X_n - X|$, which is a $\mathcal{F}$-measurable set). Since the sum of their probabilities is finite, by the Borel-Cantelli lemma 1, finitely many $A_n(\epsilon)$'s will occur with probability 1. This means that for any $\epsilon > 0$, $|X_n - X| \leq \epsilon$ for all large enough n, meaning that it converges to 0.

3.4 Laws of Large Numbers

Theorem 3.8 (Weak Law of Large Numbers)

Let $X_1, X_2, ..., X_n$ be a sequence of iid random variables, with finite mean $\mathbb{E}[X]$. Then, the average of the random variables S_n/n converges in probability to $\mathbb{E}[X]$.

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{i.p} \mathbb{E}[X] \tag{304}$$

That is, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) - \mathbb{E}[X]\right| > \epsilon\right) = 0 \tag{305}$$

Proof.

We first do the proof assuming additionally that X has finite variance, so $\text{Var}[X] < \infty$. We will show that the random variable S_n/n converges in mean square to $\mathbb{E}[X]$, which will imply convergence in

probability. Note that $\mathbb{E}[S_n/n] = \mathbb{E}[X]$, and

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[\left|\frac{S_n}{n} - \mathbb{E}[X]\right|^2\right] &= \lim_{n\to\infty} \mathbb{E}\left[\left|\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right|^2\right] \\ &= \lim_{n\to\infty} \text{Var}\left(\frac{S_n}{n}\right) \\ &= \lim_{n\to\infty} \frac{\text{Var}(S_n)}{n^2} \\ &= \lim_{n\to\infty} \frac{\text{Var}[X]}{n} = 0\end{aligned}$$

Theorem 3.9 (Strong Law of Large Numbers)

Let $X_1, X_2, ..., X_n$ be a sequence of iid random variables, with finite mean $\mathbb{E}(X_k)$ and with finite variance. Then, the average of the random variables S_n/n converges almost surely to $\mathbb{E}[X]$.

$$\frac{S_n}{n} \xrightarrow{a.s.} \mathbb{E}[X] \tag{306}$$

That is,

$$\mathbb{P}\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^{n} X_i(\omega)\right) = \mathbb{E}[X]\right\}\right) = 1 \tag{307}$$

Now let's compare these two laws. They both deal with averages of random variables, i.e. we keep sampling from X and compute the averages $\overline{X}_n$. The weak law states that for a specified large n, the average $\overline{X}_n$ is likely to be near $\mathbb{E}[X]$. But it leaves open the possibility that $|\overline{X}_n - \mathbb{E}[X]| > \epsilon$ happens an infinite number of times (although less frequently). So no matter how big of an n we choose, there could always be an $\overline{X}_n$ in the future that fails to satisfy $|\overline{X}_n - \mathbb{E}[X]| > \epsilon$. However, the strong law shows that this almost surely will not occur. That is, with probability 1, we have for any $\epsilon > 0$ the inequality $|\overline{X}_n - \mathbb{E}[X]| < \epsilon$ for all large enough n greater than a certain N. Note that the weak law does not guarantee the existence of such an N.

This result is very useful because it justifies experiments that estimate some value by taking averages.

Example 3.14 (Estimating Speed of Light)

Say that we are conducting an experiment to justify the speed of light, which will have true value μ. The laws of large numbers say that in theory, after obtaining enough data, we can get arbitrarily close to the true speed of light. Choose $\epsilon > 0$ arbitrarily small. We can obtain n estimates $X_1, \dots, X_n$ of the speed of light and compute the average

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{308}$$

As we obtain more data, we can compute $\overline{X}_n$ for each $n = 1, 2, \dots$. The weak law says that $\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \to 0$ as $n \to \infty$, i.e. the probability of our estimate being off by more than ϵ goes to 0 (though it may happen with nonzero probability if we consider the infinite sequence). The strong law says that the number of times $|\overline{X}_n - \mu|$ is greater than ϵ is finite (with probability 1), and after a certain point our estimates will perfectly lie within the error ϵ. This gives us considerable confidence in the value $\overline{X}_n$ because it guarantees the existence of some $N \in \mathbb{N}$ s.t. $|\overline{X}_n - \mu| < \epsilon$ for all $n > N$, i.e. the average *never* fails for $n > N$.

3.5 Concentration Inequalities

Concentration inequalities give you probability bounds on random variables taking atypical values. For example, given a random variable with certain mean and variance, the probability of that random variable taking values outside a certain range around the mean is very small. It's called concentration because the probability concentrates around a certain range.

The basic question here is that we would like to model a random variable X over a probability space Ω and have some data $X_1, X_2, \dots, X_n$ iid according to X. Let us have a fixed function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ that transforms the joint random variable $(X_1, \dots, X_n)$ to create a new scalar RV

$$f(X_1, \dots, X_n) = f \circ (X_1, \dots, X_n) : \Omega \longrightarrow \mathbb{R} \tag{309}$$

$f(X_1, \dots, X_n)$ is a random variable so it has a mean, denote it $\mathbb{E}[f]$. Then concentration generally refers to the probability that the value of f is at least some distance further from its mean.

$$\mathbb{P}\big(|f(x) - \mathbb{E}[f]| \geq t\big) \leq \epsilon \tag{310}$$

for some small positive ϵ. Usually, we would like this ϵ to be an exponentially decaying function of t so that the bound goes down fast. This is what's so great about the Gaussian, which is why we'll introduce it here.

Theorem 3.10 (Gaussian Tail Inequality)

Given $X \sim \mathcal{N}(0, 1)$, the inequality says that the probability of X taking values past a certain t decays exponentially.

$$\mathbb{P}\big(|X| > t\big) \leq \frac{2e^{-t^2/2}}{t} \tag{311}$$

If we have $X_1, \dots, X_n \sim \mathcal{N}(0, 1)$, then

$$\mathbb{P}\big(|\overline{X}| > t\big) \leq \frac{2}{\sqrt{n}t} e^{-nt^2/2} \tag{312}$$

We can assume that the coefficient is less than 1 if n is large. The above tells us that this bound exponentially decays with t but also with the number of samples n.

Proof.

We can simply check

$$\phi(s) = \frac{1}{\sqrt{2\pi}} e^{-s^2/2} \implies \phi'(s) = s\,\phi(s) \tag{313}$$

and use this to evaluate

$$\begin{aligned}
\mathbb{P}(X > t) &= \int_t^\infty \phi(s)\, ds \\
&= \int_t^\infty \frac{s}{s}\phi(s)\, ds \\
&< \frac{1}{t}\int_t^\infty s\phi(s)\, ds \\
&= \frac{1}{t}\int_t^\infty \phi'(s)\, ds \\
&= \frac{\phi(t)}{t}
\end{aligned}$$

Due to the exponential nature of the probability bound, we are extremely confident in getting the majority

of our samples from a small interval. If we had taken some distribution like a Cauchy, with PDF of form

$$f(x) \propto \frac{1}{1+x^2} \tag{314}$$

Then we see that even though the shape looks like a Gaussian at first glance, the fat tails go down at the rate of $1/x^2$. It turns out that due to this, when we sample numerically, we occasionally get extreme values.

Theorem 3.11 (Markov's Inequality)

If X is a non-negative random variable of finite expectation and $\alpha > 0$, then

$$\mathbb{P}(X > \alpha) \leq \frac{\mathbb{E}[X]}{\alpha} \tag{315}$$

That is, the probability that X takes a value greater than α is at most the expectation of X divided by α. This is meaningful only when $\mathbb{E}[X] < \alpha$, since otherwise the RHS will be greater than 1.

Proof.

Given any $\alpha > 0$, we can set

$$X = X \cdot 1_{X \leq \alpha} + X \cdot 1_{X > \alpha} \tag{316}$$

and by linearity,

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X \cdot 1_{X \leq \alpha} + X \cdot 1_{X > \alpha}] \\ &\geq \mathbb{E}[X \cdot 1_{X > \alpha}] \\ &\geq \alpha \mathbb{E}[1_{X > \alpha}] \\ &= \alpha\, \mathbb{P}(X > x) \end{aligned}$$

In other words, the probability that $X > \alpha$ goes down at least as fast as $1/\alpha$. For example, setting $\alpha = 2\mathbb{E}[X]$, the probability that X takes value that is at least twice its expectation is at most $1/2$. Furthermore, as X gets very large, the probability that it will take a value beyond a large α goes down faster than $1/\alpha$. But this is a very conservative inequality, and usually the probability goes down much faster.

Markov's inequality is very conservative but very general, too. If we make further assumptions about the random variable X, we can often make stronger bounds. Chebyshev's inequality assumes a (possibly negative) random variable with finite variance and states that the probability will go down as $1/x^2$.

Theorem 3.12 (Chebyshev Inequality)

Given (possibly negative) random variable X, if $\mathbb{E}[X] = \mu < +\infty$ and $\text{Var}(X) = \sigma^2 < +\infty$, then for all $\alpha > 0$,

$$\mathbb{P}\big(|X - \mu| > k\sigma\big) \leq \frac{1}{k^2} \iff \mathbb{P}(|X - \mu| > \alpha) \leq \frac{\text{Var}[X]}{\alpha^2} \tag{317}$$

That is, the probability that X takes a value further than k standard deviations away from μ goes down by $1/k^2$. Therefore, if σ is small, then this bound will be small since there is more concentration in the mean.

Proof.

We apply Markov's inequality to the non-negative random variable $|X - \mu|$.

$$\mathbb{P}(|X - \mu| > \alpha) = \mathbb{P}(|X - \mu|^2 > \alpha^2) \leq \frac{\mathbb{E}(|X - \mu|^2)}{\alpha^2} = \frac{\text{Var}[X]}{\alpha^2} \tag{318}$$

since the numerator on the RHS is the definition of variance.

Chebyshev inequality is just Markov's inequality applied to X^2 (assuming 0 mean), and often yields a better bound. But even Chebyshev's inequality turns out to be quite loose, and even this $1/k^2$ is not a very nice bound. We could apply Markov's inequality to higher powers of X, e.g. given a random variable X, we can apply Markov's inequality to the kth power of nonnegative random variable $|X - \mathbb{E}[X]|$:

$$\mathbb{P}(|X - \mathbb{E}[X]| > \alpha) = \mathbb{P}\big(|X - \mathbb{E}[X]|^k > \alpha^k\big) \leq \frac{\mathbb{E}(|X - \mathbb{E}[X]|^k)}{\alpha^k} \tag{319}$$

The natural culmination of all this is to apply Markov's inequality to e^X (or, for a little flexibility, e^{tX}, where t is a constant to be optimized). This gives us an exponential bound on $\mathbb{P}(X > \alpha)$.

Example 3.15 (Gaussian)

For the normal distribution, recall the 67-95-99.7 rule. It is well known that the probability of a random variable taking values within 2 standard deviations from the mean is 95%, so the probability that it takes outside is 5%, or 1/20, which is less than the $1/2^2 = 1/4$ bound given by Chebyshev.

3.5.1 Chernoff Bound and MGFs

Theorem 3.13 (Chernoff Bound)

Given a (possibly negative) random variable X, assume that its moment generating function $M_X(s) = \mathbb{E}[e^{sX}]$ is finite for every $s \in [-\epsilon, \epsilon]$. Then, since $x \mapsto e^{sx}$ is monotonically increasing, we have the identity

$$\mathbb{P}(X > \alpha) = \mathbb{P}(e^{sX} > e^{s\alpha}) \text{ for } s > 0 \tag{320}$$

But since the new random variable e^{sX} is nonnegative, we can now go back to Markov inequality and write

$$\mathbb{P}(X > \alpha) = \mathbb{P}(e^{sX} > e^{s\alpha}) \geq \frac{\mathbb{E}[e^{sX}]}{e^{s\alpha}} = M_X(s)\, e^{-s\alpha} \tag{321}$$

for $s > 0$ (for identity above to hold) *and* $s \in D_X$ (and it is in domain of convergence). Now, we have an exponentially decaying bound in terms of α. We have the freedom to choose s, since our bound is in terms of α, so we must choose s that minimizes $M_X(s)\, e^{-s\alpha}$. Ultimately, our best bound is

$$\mathbb{P}(X > \alpha) \leq \inf_{s>0} M_X(s)\, e^{-s\alpha} \tag{322}$$

After we optimize over s what remains on the RHS is a function of α.

Now, we can calculate the MGF of X directly if we knew the distribution of X, but we can also get bounds on it given some coarse statistics of X.

Lemma 3.5 ()

Let X be a 0-mean random variable s.t. $a \leq X \leq b$ with probability 1. Then for all $t > 0$,

$$\mathbb{E}[e^{tX}] \leq e^{t^2(b-a)^2/8} \tag{323}$$

Proof.

We can write $x = \lambda a + (1 - \lambda b)$, $0 \leq \lambda \leq 1$, and convexity of the exponential tells us that

$$e^{tx} \leq \lambda e^{ta} + (1 - \lambda) e^{tb} \tag{324}$$

Plugging in $\lambda = (b-x)/(b-a)$ then gives

$$e^{tx} \leq \frac{b-x}{b-a}e^{tx} + \frac{x-a}{b-a}e^{tb} \tag{325}$$

Take expectations of both sides, and using linearity of expectation and the fact that $\mathbb{E}[X] = 0$.

$$\mathbb{E}[e^{tX}] \leq \frac{b-\mathbb{E}X}{b-a}e^{ta} + \frac{\mathbb{E}X-a}{b-a}e^{tb} = \frac{be^{ta}-ae^{tb}}{b-a} \leq e^{t^2(b-a)^2/8} \tag{326}$$

3.5.2 Hoeffding's Inequality

Hoeffding's inequality is one of the most important inequalities in concentration of measure. The proof of this inequality involves many useful tricks.

Theorem 3.14 (Hoeffding's Inequality)

Let $X_1, X_2, \dots, X_n$ be independent (not necessarily identical) random variables s.t. $a_i \leq X_i \leq b_i$ almost surely. Consider the random variable $\overline{X} = \frac{1}{n}(X_1 + \dots + X_n)$. Then, for all $t > 0$, we have the two inequalities

$$\mathbb{P}\big(\overline{X} - \mathbb{E}[\overline{X}] \geq t\big) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)$$
$$\mathbb{P}\big(\overline{X} - \mathbb{E}[\overline{X}] \leq -t\big) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)$$

which can be combined to produce

$$\mathbb{P}\big(\big|\overline{X} - \mathbb{E}[\overline{X}]\big| \geq t\big) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right) \tag{327}$$

We can create an equivalent bound on the sum $S_n = X_1 + \dots + X_n$:

$$\begin{aligned}\mathbb{P}\big(|S_n - \mathbb{E}[S_n]| \geq t\big) &= \mathbb{P}\big(n|\overline{X} - \mathbb{E}[\overline{X}]| \geq t\big) \\ &= \mathbb{P}\big(|\overline{X} - \mathbb{E}[X]| \geq \frac{t}{n}\big) \\ &\leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)\end{aligned}$$

Proof.

We will prove just with the case where $X_1, \dots X_n$ are all bounded by $[a, b]$, which gives

$$\mathbb{P}\big(|\overline{X} - \mathbb{E}[\overline{X}]| \geq t\big) \leq 2\exp\left(-\frac{2nt^2}{(b-a)}\right)$$

Now, we can write

$$\begin{aligned}
\mathbb{P}(\overline{X}_n > \epsilon) &= \mathbb{P}\Big(\sum_{i=1}^n X_i \geq n\epsilon\Big) \\
&= \mathbb{P}\big(e^{t\sum X_i} \geq e^{tn\epsilon}\big) && \text{(Variational Technique)} \\
&\leq e^{-tn\epsilon}\,\mathbb{E}[e^{t\sum X_i}] && \text{(Markov's Inequality)} \\
&= e^{-tn\epsilon}\left(\mathbb{E}[e^{tX_i}]\right)^n && \text{(Independence)} \\
&\leq e^{-tn\epsilon} e^{n\frac{t(b-a)^2}{2}} && \text{(prev. lemma)}
\end{aligned}$$

The step where we introduce an extra parameter t is called a variational technique, used for optimization, and we can adjust t to make it as small as possible. Taking the derivative of the final expression w.r.t. t and solving for 0 gives us $t = \frac{4\epsilon}{(b-a)^2}$, and substituting into the expression gives the bound as

$$\mathbb{P}(\overline{X}_n > \epsilon) \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \tag{328}$$

By further rearranging, we can write it as

$$\mathbb{P}\left(|\overline{X} - \mathbb{E}[\overline{X}]| \geq t\sqrt{\frac{\sum_{i=1}^n (b_i-a_i)^2}{n^2}}\right) \leq 2\exp(-2t^2) \tag{329}$$

which now looks like our Chebyshev inequality, but without a notion of standard deviation. But note the fact if $a_i \leq X_i \leq b_i$, then $\text{Var}(X_i) \leq (b_i - a_i)^2$ (since $\text{Var}(X_i) = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \leq \mathbb{E}[(b_i-a_i)^2]$). So, we have

$$\text{Var}(\overline{X}) \leq \frac{\sum_{i=1}^n (b_i-a_i)^2}{n^2} \implies \mathbb{P}\left(|\overline{X} - \mathbb{E}[\overline{X}]| \geq t\sqrt{\frac{\sum_{i=1}^n (b_i-a_i)^2}{n^2}} \geq \text{Var}(\overline{X})\right) \leq 2\exp(-2t^2) \tag{330}$$

which allows us to interpret Hoeffding's inequality in a more familiar way. It says that the probability that the sample average is more than t standard deviations from its expectation is at most $2e^{-2t^2}$.

Corollary 3.1 ()

If $X_1, X_2, \ldots, X_n$ are independent with $\mathbb{P}(a_i \leq X_i \leq b_i) = 1$ and common mean μ, then

$$\mathbb{P}\left[|\overline{X}_n - \mu| \leq \sqrt{\frac{\sum_{i=1}^n (b_i-a_i)^2}{2n^2}\log\left(\frac{2}{\delta}\right)}\right] \geq 1-\delta \tag{331}$$

Example 3.16 (Bernoulli)

Applying Hoeffding's inequality to a sequence of n p-coin tosses $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ gives

$$\mathbb{P}\big(|\overline{X}_n - p| > \epsilon\big) \leq 2\exp^{-2n\epsilon^2} \tag{332}$$

Example 3.17 (Mean)

Suppose we have $X_1, X_2, \ldots X_n \sim \text{Bernoulli}(p)$, all iid. Then, by Hoeffding's inequality, the average $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ is tightly concentrated around p.

$$\mathbb{P}\big(|\overline{X} - p| \geq t\big) \leq 2e^{-2nt^2} \tag{333}$$

Note that $b_i - a_i = 1 - 0 = 1$ for all i. There is an exponential decay in the probability of the sample

mean deviating from its expectation.

Example 3.18 (Hypercube)

Pick $X \in [-1,+1]^d$ uniformly at random, i.e. choose iid $X_1, \dots, X_d \sim \text{Uniform}[-1,+1]$. The expectation is

$$\mathbb{E}||X||^2 = \sum_{i=1}^d \mathbb{E}X_i^2 = \sum_{i=1}^d \int_{-1}^1 x^2 f_X(x)\,dx = \sum_{i=1}^d \int_{-1}^1 \frac{1}{2}x^2\,dx = \frac{d}{3} \tag{334}$$

Then, it can be shown that $||X|| =$ is tightly concentrated around $\sqrt{d/3}$. We show this again with Hoeffding's inequality by showing the concentration of $||X||^2$ around $d/3$.

$$\mathbb{P}\left(\left|||X||^2 - \frac{d}{3}\right| \geq t\right) \leq 2\exp\left(-\frac{dt^2}{2}\right) \tag{335}$$

This tells us that if we choose the uniform random vector $X \in [-1,+1]^d$, the vast majority of our samples will have $||X|| \approx \sqrt{d/3}$.

Hoeffding's inequality does not use any information about the random variables expect for the fact that they are bounded. If the variance of X_i is small, then we can get a sharper inequality from Bernstein's inequality.

Theorem 3.15 (Bernstein's Inequality)

If $\mathbb{P}(|X_i| \leq c) = 1$ and $\mathbb{E}[X_i] = 0$, set $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for any $t > 0$,

$$\mathbb{P}\big(\big|\overline{X}\big| > \epsilon\big) \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}\right) \tag{336}$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \text{Var}(X_i)$.

3.5.3 Concentration of Lipshitz Functions

Observing the Hoeffding bound, one might wonder whether such concentration applies only to averages or sums of random variables. After all, what's so special about averages? It turns out that the relevant feature of the average that yields tight concentration is that it is smooth in the way that if we change the value of one random variable the function does not change dramatically.

Theorem 3.16 (Bounded Difference Inequality)

Let has have independent random variables $X_1, X_2, \dots, X_n$ and a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ that satisfies the **bounded difference property** that

$$\big|f(x_1, \dots, x_k, \dots, x_n) - f(x_1, \dots, x_k', \dots, x_n)\big| \leq c_k \tag{337}$$

for every $x, x' \in \mathbb{R}^n$. That is, the function changes by at most c_k if its kth coordinate is changed. Then, for all $t \geq 0$, we have the concentration inequality:

$$\mathbb{P}\big(f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq t\big) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^n c_k^2}\right)$$
$$\mathbb{P}\big(f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \leq -t\big) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^n c_k^2}\right)$$

Combining the two gives

$$\mathbb{P}\big(\big|f(X_1,\dots,X_n)-\mathbb{E}[f(X_1,\dots,X_n)]\geq t\big|\big)\leq \exp\left(-\frac{2t^2}{\sum_{k=1}^n c_k^2}\right)$$

In fact, any smooth function of bounded independent random variables is tightly concentrated around its expectation, and the notion of smoothness is Lipshitz continuity.

Definition 3.9 ()

A function $f:\mathbb{R}^n\longrightarrow\mathbb{R}$ is L-Lipschitz w.r.t. the l_p-metric if for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$,

$$|f(\mathbf{x})-f(\mathbf{y})|\leq L||\mathbf{x}-\mathbf{y}||_p \tag{338}$$

Example 3.19 ()

For $x=(x_1,x_2,\dots,x_n)$, we define the average $a(x)=\frac{1}{n}(x_1+\ldots+x_n)$. Then, a is $(1/n)$-Lipschitz w.r.t. the l_1 metric, since for any $\mathbf{x},\mathbf{y}$,

$$\begin{aligned}|a(\mathbf{x})-a(\mathbf{y})|&=\left|\frac{1}{n}\big[(x_1-y_1)+\ldots+(x_n-y_n)\big]\right|\\&=\frac{1}{n}\big(|x_1-y_1|+\ldots+|x_n-y_n|\big)\\&=\frac{1}{n}||\mathbf{x}-\mathbf{y}||_1\end{aligned}$$

It turns out that Hoeffding's bound holds for all Lipschitz functions w.r.t. the l_1 metric.

Theorem 3.17 ()

Suppose $X_1,X_2,\dots,X_n$ are independent and bounded with $a_i\leq x_i\leq b_i$. Then, for any $f:\mathbb{R}^n\longrightarrow\mathbb{R}$ that is L-Lipschitz w.r.t. the l_1-metric, we have

$$\begin{aligned}\mathbb{P}[f\geq\mathbb{E}(f)+t]=\mathbb{P}[f-\mathbb{E}(f)\geq t]&\leq\exp\left(-\frac{2t^2}{L^2\sum_{i=1}^n(b_i-a_i)^2}\right)\\\mathbb{P}[f\leq\mathbb{E}(f)-t]=\mathbb{P}[f-\mathbb{E}(f)\leq -t]&\leq\exp\left(-\frac{2t^2}{L^2\sum_{i=1}^n(b_i-a_i)^2}\right)\end{aligned}$$

and combining these inequalities gives

$$\mathbb{P}[|f-\mathbb{E}(f)|\geq t]\leq\exp\left(-\frac{2t^2}{L^2\sum_{i=1}^n(b_i-a_i)^2}\right) \tag{339}$$

3.6 Central Limit Theorem

By the law of large numbers, the sample averages converge almost surely (and therefore converge in probability) to the expected value μ as $n\to\infty$. The CLT describes the size and the distributional form of the stochastic fluctuations around μ during this convergence. That is, it states that as n gets larger, the distribution of the difference $\overline{X}_n-\mu$ approximates a $\mathcal{N}(0,\sigma^2/n)$ distribution, where σ^2 is the variance of X.

Roughly speaking, the (weak) law of large numbers says that

$$\frac{S_n-n\mathbb{E}[X]}{n}\xrightarrow{i.p.}0 \tag{340}$$

That is, if we consider the sequence of functions $\{S_n - n\mathbb{E}[x]\}_{n\in\mathbb{N}}$, this sequence is sublinear (i.e. is $o(n)$). CLT does two things:

1. It specifically quantifies this fluctuation $\{S_n - n\mathbb{E}[X]$ by saying that it is approximately of order $\sqrt{n}$.
2. Furthermore, this fluctuation, when divided by $\sqrt{n}$ converges in distribution to a Gaussian.

$$\frac{S_n - n\mathbb{E}[X]}{\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, \sigma_X^2) \tag{341}$$

Theorem 3.18 (Central Limit Theorem)

Let $X_1, X_2, X_3, ...$ be a sequence of iid random variables, with mean $\mu = \mathbb{E}[X]$ and with variance $\text{Var}(X) = \sigma^2 < \infty$. Then, the sequence of random variables $\{\overline{X}_n\}_{n\in\mathbb{N}}$ converges in distribution to a Gaussian $\mathcal{N}(\mu, \sigma^2/n)$. That is,

$$\frac{\overline{X}_n - \mu}{\sigma\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1) \tag{342}$$

Proof.

Let $Z_i = \frac{X_i - \mu}{\sigma}$ and let $U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i$ (we can normalize the X_i's since they have finite mean and variance). Note that since we have finite second moments

$$\begin{aligned}
&\mathbb{E}[Z_i] = 0 < \infty \\
&\text{Var}[Z_i] = \mathbb{E}[(Z_i - \mathbb{E}[Z_i])^2] = \mathbb{E}[Z_i^2] = 1 < \infty
\end{aligned}$$

we can Taylor expand the characteristic function $\varphi_{Z_i}(t)$ up to at least the second order (from moment generating property theorem). So, we have

$$\begin{aligned}
\varphi_{Z_i}(t) &= 1 + \frac{\mathbb{E}[Z_i]}{1!}(it)^1 + \frac{\mathbb{E}[Z_i^2]}{2!}(it)^2 + o(t^2) \\
&= 1 + 0 + \frac{1}{2}(it)^2 + o(t^2) \\
&= 1 - \frac{1}{2}t^2 + o(t^2)
\end{aligned}$$

Now calculate the CF of U_n. Since the Z_i's are iid, we can get

$$\varphi_{U_n}(t) = \left(\varphi_{Z_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \tag{343}$$

The $o(t^2/n)$ term vanishes as $n \to \infty$, and using the limit $e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$, we have

$$\lim_{n\to\infty} \varphi_{U_n}(t) = \lim_{n\to\infty} \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n = e^{-t^2/2} \tag{344}$$

which is precisely the CF of a standard Gaussian random variable. Since CFs are unique, our result is proven. Essentially, we have proved convergence in distribution of random variables by showing convergence of their characteristic functions.

A big misconception is that this normalized sum has PDF that converges to a bell curve. It is the CDF (by definition of convergence in distribution) that converges to that of a Gaussian. That way, we can state this for discrete, continuous, mixtures: doesn't matter. They don't even need to have a density, since if we just took a bunch of Bernoulli's, the PMF of their sum would never be defined for an irrational number like π. But it would be defined for the CDF, and even though the CDF of a discrete random variable will

have jumps, these jumps would get smaller and smaller until it converges pointwise. Even if the X_i's had densities, the CLT does not say that their mean converges to the PDF of a normal. Just because the CDF converges, it doens't mean the PDF will look similiar.

It also turns out (?) that we can use CLT to prove the weak law of large numbers, since (roughly speaking) as n increases, the distribution of $\overline{X}_n$ concentrates more and more around μ, and therefore the probability of $|\overline{X}_n - \mu| < \epsilon$ tends to 1.

4 Conditional Expectation

Conditional expectation is extremely important, especially in the context of stochastic processes, which is talked about in more detail in another set of notes.

First, note that when we talk about the probability of event A happening, or equivalently, the probability of $\omega \in A$, we can write this as the expected value of the indicator function A.

$$\mathbb{P}(A) = \mathbb{E}[1_A] \tag{345}$$

This will come in handy later in connecting conditional probability and expectation.

Now conditional expectation is quite tricky to understand at first. We will start by defining it given a σ-algebra and then given a random variable.

Definition 4.1 (Conditional Expectation)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a sub-$\sigma$-algebra $\mathcal{G} \subset \mathcal{F}$, and an $\mathcal{F}$-measurable random variable X (with $\mathbb{E}[X] < \infty$), the **conditional expectation of X given $\mathcal{G}$** is defined to be the $\mathcal{G}$-measurable random variable $Y = \mathbb{E}[X \mid \mathcal{G}]$ satisfying

$$\int_A X \, d\mathbb{P} = \int_A Y \, d\mathbb{P} \tag{346}$$

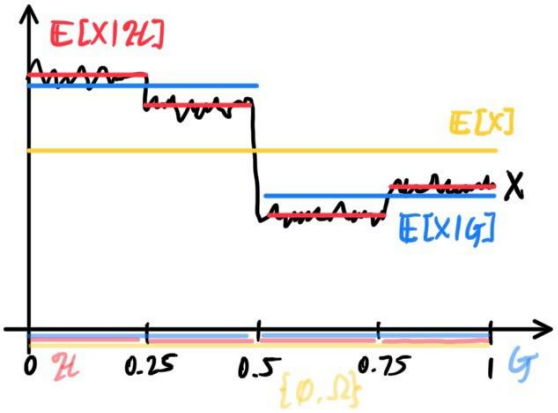
or equivalently,

$$\mathbb{E}[X \cdot 1_A] = \mathbb{E}[Y \cdot 1_A] \tag{347}$$

for all $A \in \mathcal{G}$. Any Y satisfying these two conditions is said to be a **version** of $\mathbb{E}[X \mid \mathcal{F}]$. The critical detail to note here is that the conditional expectation Y, has the same expected value as X does, not over just the whole $\mathcal{G}$, but *in every subset G of $\mathcal{G}$*.

We state without proof that $\mathbb{E}[X \mid \mathcal{G}]$ exists and is almost surely unique. For now, we can interpret this as the best approximation of the $\mathcal{F}$-measurable X with the $\mathcal{G}$-measurable Y. Here is a useful analogy. Say that we have some "fine" function X defined on the interval $[0, 1]$ with a fine Borel σ-algebra $\mathcal{F}$.

1. If we are given some sub-σ-algebra $\mathcal{G}$ composed of $\emptyset, [0, 0.5], (0.5, 1], [0, 1]$, then Y would be the step function defined constantly on these intervals.
2. If we are given a finer sub-σ-algebra $\mathcal{H}$ generated by $[0, 0.25), [0.25, 0.5), [0.5, 0.75), [0.75, 1]$, then this would give a $\mathcal{H}$-measurable function that is a better approximation of X.



Therefore, we can see that if $\mathbb{E}[X \mid \mathcal{G}]$ is $\mathcal{F}$-measurable, then

$$X = \mathbb{E}[X \mid \mathcal{G}] \tag{348}$$

since its value coincides with X for every event in $\mathcal{F}$. One way to think about it is that $\mathbb{E}[X \mid \mathcal{G}]$ is the conditional expectation of X (which is "detailed" up to resolution $\sigma(\mathcal{G})$) taken with a camera of resolution $\mathcal{G}$. The finer (bigger) the σ-algebra is, the higher the resolution.

4.1 Properties of Conditional Expectation

Theorem 4.1 (Tower Rule)

The expectation of X and its approximation always coincides.

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] \tag{349}$$

Lemma 4.1 ()

Let $\mathbb{E}[|X|], \mathbb{E}[|Y|] < \infty$. Then,

1. Conditional expectation is linear

$$\mathbb{E}[aX + bY \mid \mathcal{G}] = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}] \tag{350}$$

2. If $X \leq Y$, then

$$\mathbb{E}[X \mid \mathcal{G}] \leq \mathbb{E}[Y \mid \mathcal{G}] \tag{351}$$

Theorem 4.2 (Jensen's Inequality)

If φ is convex and $\mathbb{E}[|X|], \mathbb{E}[|\varphi(X)|] < \infty$, then

$$\varphi(\mathbb{E}[X \mid \mathcal{G}]) \leq \mathbb{E}[\varphi(X) \mid \mathcal{G}] \tag{352}$$

Theorem 4.3 ()

Conditional expectation is a contraction in L^p, $p \geq 1$.

Theorem 4.4 ()

If X is $\mathcal{F}$-measurable and $\mathbb{E}[|Y|], \mathbb{E}[|XY|] < \infty$, then

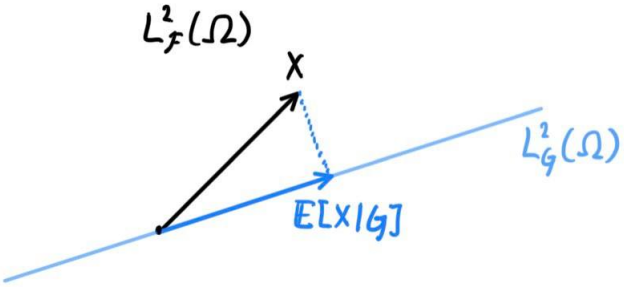
$$\mathbb{E}[XY \mid \mathcal{F}] = X\mathbb{E}[Y \mid \mathcal{F}] \tag{353}$$

Theorem 4.5 ()

Suppose $\mathbb{E}[X^2] < \infty$. Then, $\mathbb{E}[X \mid \mathcal{G}]$ is the $\mathcal{G}$-measurable function Y that minimizes the mean squared error

$$\mathbb{E}[(X - Y)^2] \tag{354}$$

This gives a nice geometric interpretation of $\mathbb{E}[X \mid \mathcal{G}]$. Given that X lives in the Hilbert space $L^2_{\mathcal{F}}(\Omega)$, $\mathbb{E}[X \mid \mathcal{G}]$ is the projection of X onto the subspace $L^2_{\mathcal{G}}(\Omega)$.



Therefore, we can change the way think about $\mathbb{E}[X]$. It is not just a value, but rather, we can think of it as our best prediction of X given no information. Specifically,

$$\mathbb{E}[X] = \mathbb{E}[X \mid \{\emptyset, \Omega\}] \tag{355}$$

That is, letting $\mathcal{G}$ be the trivial σ-algebra, we must find the best approximation of X that is $\mathcal{G}$-measurable. But any random variable that is $\mathcal{G}$-measurable must be constant, since if we take the preimage of any singleton set $\{x\} \in \mathcal{R}$, then it must be either $\emptyset$ (X does not map to it) or Ω (X maps all of Ω to it).

4.2 Perfect Information vs No Information

Now let us state some properties on how certain σ-algebras can change the conditional expectation of certain random variables.

Theorem 4.6 (Perfect Information)

If X is $\mathcal{G}$-measurable, then

$$\mathbb{E}[X \mid \mathcal{G}] = X \tag{356}$$

That is, the values of X are defined on $\sigma(X) \subset \mathcal{F}$ and so has a detail level of $\sigma(X)$. But if we condition it on an even finer $\mathcal{G} \supset \sigma(X)$, then we are taking a picture of X with something that has overly high resolution, and so our best approximation of X is X itself. Indeed, if X lives in $L_{\mathcal{G}}(\Omega)$, then its projection onto $L_{\mathcal{G}}(\Omega)$ is X itself.

Theorem 4.7 (Irrelevant Information)

If X is independent of $\mathcal{G}$, i.e. $\sigma(X)$ and $\mathcal{G}$ are independent σ-algebras, then

$$\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X] \tag{357}$$

That is, our best approximation of X given information $\mathcal{G}$ is $\mathbb{E}[X]$ itself, i.e. if you don't know anything about X, then the best guess is the mean $\mathbb{E}[X]$. To see why, note that independence means that for all $A \in \mathcal{G}$ and $B \in \mathcal{R}$,

$$\mathbb{P}(X^{-1}(B) \cap A) = \mathbb{P}(X^{-1}(B)) \cdot \mathbb{P}(A) \tag{358}$$

Theorem 4.8 (Trivial Information)

If $\mathcal{G} = \{\emptyset, \Omega\}$, then

$$\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X] \tag{359}$$

This makes sense since we're trying to measure $\sigma(X)$-measurable X with the trivial $\mathcal{G}$, and the only function that is measurable w.r.t. the trivial σ-algebra is a constant function (since the preimage of every Borel set in $\mathcal{R}$ must be either Ω or $\emptyset$). This is the same as projecting X to the line of constant functions in $L_{\mathcal{F}}(\Omega)$.

Theorem 4.9 ()

If $\mathcal{F}_1 \subset \mathcal{F}_2$, then
1. $\mathbb{E}[\mathbb{E}[X \mid \mathcal{F}_1] \mid \mathcal{F}_2] = \mathbb{E}[X \mid \mathcal{F}_1]$.
2. $\mathbb{E}[\mathbb{E}[X \mid \mathcal{F}_2] \mid \mathcal{F}_1] = \mathbb{E}[X \mid \mathcal{F}_1]$.

In other words, the smaller σ-algebra always wins.

We can see this visually since in both cases, we are projecting X onto $L^2_{\mathcal{F}_1}(\Omega)$ and onto $L^2_{\mathcal{F}_2}(\Omega)$, but either way, we end up in $L^2_{\mathcal{F}_1}(\Omega)$. Additionally, this is also consistent with our camera analogy, where $\mathbb{E}[X \mid \mathcal{G}]$ is like taking a picture of random variable X with a camera of resolution $\mathcal{G}$. Conditional expectation is essentially an averaging/blurring operator. So, $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}]$ is like taking a picture of X with resolution $\mathcal{H}$ and then with $\mathcal{G}$. The lower resolution would always win.

4.3 Computation of Conditional Expectation

Definition 4.2 ()

Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the conditional expectation of Y given X is the random variable

$$\mathbb{E}[X \mid Y] \coloneqq \mathbb{E}[X \mid \sigma(Y)] \tag{360}$$

Note that since both X and Y are random variables, they are both $\mathcal{F}$-measurable. However, this doesn't mean that they may be $\mathcal{G}$-measurable for some sub-σ-algebra $\mathcal{G}$. So, as long as $\sigma(Y) \subset \sigma(X)$ (neither of which may be $\mathcal{F}$), we have some nontrivial approximation.

Now let's introduce a new way to think about expectation and conditional expectation in general.

1. The first step is to think of $\mathbb{E}[X]$ not as a value μ but as the best estimate for the value of a random variable X in the absence of any information. To minimize the squared error

$$\mathbb{E}[(X-e)^2] = \mathbb{E}[X^2 - 2eX + e^2] = \mathbb{E}[X^2] - 2e\mathbb{E}[X] + e^2 \tag{361}$$

we differentiate with respect to e to obtain $2e - 2\mathbb{E}[X] = 0 \implies e = \mathbb{E}[X]$. For example, if I throw a fair die and you have to estimate its value X, according to the analysis above, your best bet is to guess $\mathbb{E}[X] = 3.5$ since $\Omega = \{1,2,3,4,5,6\}$. On specific rolls of the die, this will be an over-estimate of an underestimate, but on the long run it minimizes the mean square error.

2. If we *do* have additional information, then we use conditional expectation. Suppose that I tell you that X is an even number. Then, I would guess that the possible values of X are $\{2,4,6\}$, and so the our conditional expectation is 4. Similarly, if I told you that X is odd, then the conditional expectation is 3. This additional information can be put into a random variable $Y : \Omega \to \mathbb{R}$ defined

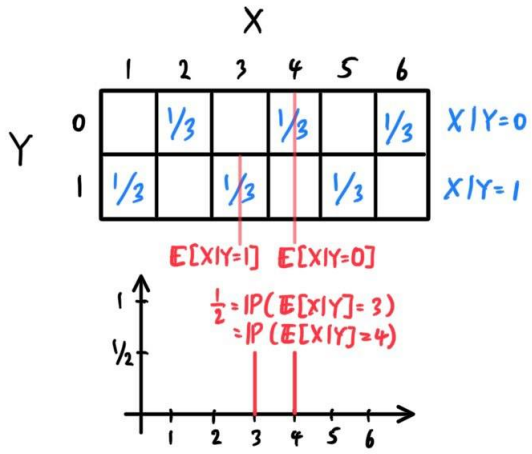
$$Y(\omega) = \begin{cases} 0 & \text{if } \omega = 2,4,6 \\ 1 & \text{if } \omega = 1,3,5 \end{cases} \tag{362}$$

Then, we can say that $\mathbb{E}[X \mid Y = 0] = 4$ and $\mathbb{E}[X \mid Y = 1] = 3$. We can interpret this as the conditional expectation given the σ-algebra generated by these two sets $\{2,4,6\}$ and $\{1,3,5\}$.

3. Now, imagine that I roll the die and I tell you the parity of X. You should see that a single numerical response cannot cover both cases. You would respond 3 if I tell you X is odd ($Y = 1$) and 4 if I tell you X is even ($Y = 0$). A single numerical response is not enough because the particular piece of information I give you is *itself random*. In fact, your response is necessarily a function of this particular piece of information, represented in our notation as

$$g(Y) = \mathbb{E}[X \mid Y] = \begin{cases} 3 & \text{if } Y = 1 \\ 4 & \text{if } Y = 0 \end{cases} \tag{363}$$

This is a function of Y, and it is consistent with our understanding of $\mathbb{E}[X \mid Y]$ as our "best estimate" of X with random variable Y.



From the visual above, we can see that we take the joint distribution $X \times Y$, and for each value $Y = y$, we can estimate X as $\mathbb{E}[X \mid Y = y]$. But now there is the additional uncertainty of what value Y will take, which turns this value estimate $\mathbb{E}[X \mid Y = y]$ in a distribution $\mathbb{E}[X \mid Y]$. So, for the discrete case,

$$\mathbb{P}\big(\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid Y = y]\big) = \mathbb{P}(Y = y) \tag{364}$$

Now we can talk about how to compute conditional expectation. In essence, the conditional expectation $\mathbb{E}[X \mid Y]$ is simply a function of a random variable Y that is this best approximation. Given a joint random variable $(X, Y) : \Omega \to \mathbb{R}^2$, we can fix a value of $Y = y$. Therefore, we are given the information that event $Y^{-1}(\{y\})$ happened, and so we can construct our conditional distribution $X \mid Y = y$, which defines a new probability measure. Taking the expectation of that gives us a number.

Definition 4.3 (Discrete Conditional Expectation Given $Y = y$)

Let X, Y be discrete random variables, with joint random variable $(X, Y) : \Omega \to \mathbb{R}^2$ and its joint PMF $p_{X,Y}(x, y)$. Recall that the conditional PMF is

$$p_{X|Y}(x \mid y) \coloneqq \frac{p_{X,Y}(x, y)}{p_Y(y)} \tag{365}$$

The **conditional expectation** of X given $Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x \, p_{X|Y}(x \mid y) \tag{366}$$

Definition 4.4 (Continuous Conditional Expectation Given $Y = y$)

Let X, Y be jointly continuous with joint PDF $f_{X,Y}(x, y)$. Recall that the conditional PDF is

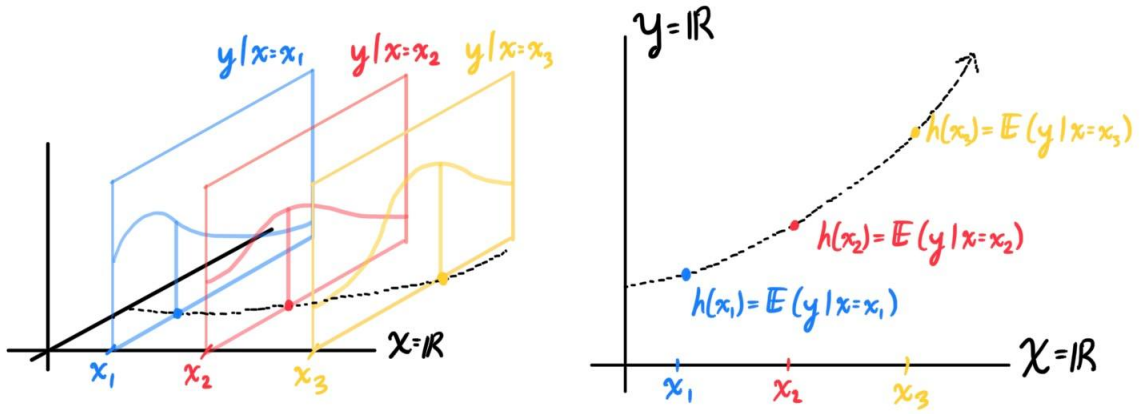
$$f_{X|Y}(x \mid y) \coloneqq \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{367}$$

The **conditional expectation** of X given $Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \int_{x \in \mathbb{R}} x \, f_{X|Y}(x \mid y) \, dx \tag{368}$$

Again, we can set $\psi(y) \coloneqq \mathbb{E}[X \mid Y = y]$, which is a function of y and therefore a random variable.

As a visual, we can take a "slice" of the joint distribution of some value of Y, look at the distribution of X on this slice, and compute its expectation. That is, for every value of $Y = y$, there exists some (conditional) distribution of X with PMF of $p_{X|Y}(x \mid y)$ or PDF of $f_{X|Y}(x \mid y)$.



So given a value of $Y = y$, we generally know something about X (e.g. if I know humidity, I know something about the temperature) and want to find the best estimate of X. This is precisely the conditional expectation $\mathbb{E}[X \mid Y = y]$, and we can interpret this as a regression function $\psi(y) \coloneqq \mathbb{E}[X \mid Y = y]$, which predicts the expected value of X given $Y = y$.

But since we don't know what exactly Y is, this process is random itself, and it is only after this Y is realized that we can provide the expected value of X. Thus, by replacing the little y's with the big Y's, we can construct a random variable that will estimate X for us given Y, denoted $\mathbb{E}[X \mid Y] = \psi(Y)$. This turns out to be a $\sigma(Y)$-measurable random variable itself.

Example 4.1 ()

Let $f_{X,Y}(x, y) = \frac{1}{x}$ for $0 < y \leq x \leq 1$. Find $\mathbb{E}[Y \mid X]$. We first calculate the marginal density of X, which will allow us to calculate the conditional density of Y:

$$f_X(x) = \int_0^x \frac{1}{x}\, dy = \frac{y}{x}\Big|_0^x = 1 \implies f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1/x}{1} = \frac{1}{x} \text{ for } 0 < y \leq x \leq 1 \tag{369}$$

Since the conditional density of Y is not dependent on x, Y is uniform from 0 to x. Now calculate the expectation:

$$\mathbb{E}[Y \mid X = x] = \int_0^x y \cdot \frac{1}{x}\, dy = \frac{x}{2} \tag{370}$$

and so the conditional expectation is

$$\mathbb{E}[Y \mid X] = \frac{1}{2}X \tag{371}$$

A fundamental result in statistical learning theory is that if we have two random variables X and Y, the best predictor of Y as a function of X is the conditional expectation $\mathbb{E}[Y \mid X]$.

Theorem 4.10 ()

Let us have two random variables X and Y, with $g(X) = \mathbb{E}[Y \mid X]$. Then, the function g minimizes the cost function $\mathbb{E}[(Y - h(X))^2]$. That is,

$$\inf_{h \text{ meas.}} \mathbb{E}[(Y - h(X))^2] = \mathbb{E}[(Y - g(X))^2] \tag{372}$$

We restate the tower rule again.

Theorem 4.11 (Tower Rule)

We know that $\mathbb{E}[Y \mid X]$ is the random variable $\psi(X)$ that is a transformed version of X. Then, we have

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y] \tag{373}$$

This is confusing notation due to the iterated expectations, but note that the term on the inside is a transformed random variable of X, while the expectation on the outside computes the expectation of this transformed random variable. So, letting $\psi(X) = \mathbb{E}[Y \mid X]$, we can equivalently write

$$\mathbb{E}[\psi(X)] = \mathbb{E}[Y] \tag{374}$$

Intuitively, this makes sense, since $\mathbb{E}[Y \mid X]$ is the random variable that tries to "model" Y given (random) information from X, so its expectation must be the expectation of Y itself.

Proof.

We can just expand this out. We will do it for the discrete case.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \sum_x p_X(x) \underbrace{\mathbb{E}[Y \mid X = x]}_{\psi(x)} \\
&= \sum_x \left(p_X(x) \cdot \sum_y y \cdot p_{Y|X}(y \mid x) \right) \\
&= \sum_x \left(p_X(x) \cdot \sum_y y \cdot \frac{p_{X,Y}(x,y)}{p_X(x)} \right) \\
&= \sum_{x,y} y \cdot p_{X,Y}(x,y) \\
&= \sum_y \left(y \cdot \sum_x p_{X,Y}(x,y) \right) \\
&= \sum_y y \cdot p_Y(y) \\
&= \mathbb{E}[Y]
\end{aligned}$$

Example 4.2 ()

Consider the random sum of random variables $S_N = \sum_{i=1}^N X_i$, where X_i are iid and N is independent of X_i's. Then, we can use the tower rule to write $\mathbb{E}[S_N] = \mathbb{E}[\mathbb{E}[S_N \mid N]]$. $\mathbb{E}[S_N \mid N]$ is a transformed random varibale of N, and to compute its closed form we should just compute $\mathbb{E}[S_N \mid N = n]$ and replace n with N.

$$\mathbb{E}[S_N \mid N = n] = \mathbb{E}[S_n] = \mathbb{E}\left(\sum_{i=1}^n X_i \right) = n\mathbb{E}[X] \tag{375}$$

remember that $\mathbb{E}[X]$ is just a number, so replacing n with N gives $\mathbb{E}[S_N \mid N] = N\mathbb{E}[X]$, i.e. the random variable N multiplied by $\mathbb{E}[X]$. Therefore,

$$\mathbb{E}[\mathbb{E}[S_N \mid N]] = \mathbb{E}[N\mathbb{E}[X]] = \mathbb{E}[N]\,\mathbb{E}[X] \tag{376}$$

This makes sense intuitively, since we want to approximate this value by taking the expected value of X and multiplying it by the expected number of summands.

Now, we can simply use the property to talk about what $\mathbb{P}(X \mid Y)$ means. Formally, using the fact that the probability that X will realize in A, i.e. $\mathbb{P}(X \in A) = \mathbb{E}_X[1_A]$, we can define the conditional probability as

$$\mathbb{P}(X \in A \mid Y) = \mathbb{E}[1_X \mid Y] \tag{377}$$

We can interpret this in multiple ways, in increasing level of rigor. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, with random variables X, Y with their probability laws $\mathbb{P}_X, \mathbb{P}_Y$.

1. When Y realizes, we can use this definition to have a better educated guess of where X will land. But since Y is random, so is our guess.
2. Let us have the joint distribution (X, Y). Given that $Y = y$, we can take the conditional distribution $X \mid Y = y$ and compute the event that this random variable lands in A by replacing the little y's with the big Y's.
3. Let A be an event in $\mathcal{R}$. Then, the probability that X will land in A is

$$\mathbb{P}(X \in A) = \mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A)) \tag{378}$$

where the first is abuse of notation, the second is the probability law of $\mathbb{R}$, and the third is the probability law of Ω. Then, 1_A generates a σ-algebra $\sigma(1_A)$ on Ω, consisting of the sets $\{\emptyset, X^{-1}(A), X^{-1}(A)^c, \Omega\}$.

4.4 Conditional Expectation given Multiple Random Variables

Now the σ-algebra generated by multiple random variables should intuitively be bigger than the σ-algebra generated by one random variable. We can't simply take the union of the individual σ-algebras.

Definition 4.5 (σ-algebra generated by Multiple Random Variables)

Given random variables $\{X_i\}_{i \in I}$ indexed by some set (possibly uncountable), the σ-algebra generated by this collection is defined

$$\sigma(X_1, \dots, X_n) \coloneqq \sigma\left(\bigcup_{i \in I} \sigma(X_i) \right) \tag{379}$$

Let us look at $\mathbb{E}[X \mid Y, Z]$ and compare it to $\mathbb{E}[X \mid Y]$. From this definition, we know that the information about X contained in $\sigma(Y, Z)$ is at least as great as the corresponding information in $\sigma(Y)$. Therefore, we can simply define conditional expectation as such:

Definition 4.6 (Conditional Expectation given Multiple Random Variables)

Given random variables $\{X_i\}_{i \in I}$, the conditional expectation is defined

$$\mathbb{E}[Y \mid \{X_i\}_{i \in I}] = \mathbb{E}[Y \mid \sigma(\{X_i\}_{i \in I}] \tag{380}$$

which is the random variable where

$$\int_A Y \, d\mathbb{P} = \int_A \mathbb{E}[Y \mid \{X_i\}_{i \in I}] \, d\mathbb{P} \tag{381}$$

for all $A \in \sigma(\{X_i\}_{i \in I})$.

4.5 Conditional Variance

Similar to conditional expectation, we can define the conditional variance $\text{Var}(Y \mid X = x)$ as a function $h(x)$ that outputs the variance of Y given $X = x$. We have

$$\text{Var}(Y \mid X = x) = \mathbb{E}\big[(Y - \mathbb{E}[Y \mid X = x])^2 \mid X = x\big] \tag{382}$$

Definition 4.7 (Conditional Variance)

The **conditional variance** of Y given X is defined as

$$\text{Var}(Y \mid X) = \mathbb{E}\big[\big(Y - \mathbb{E}[Y \mid X]\big)^2 \mid X\big] \tag{383}$$

which tells us how much variance is left if we use $\mathbb{E}[Y \mid X]$ to predict Y.

5 Order Statistics

Let $X_1, X_2, ..., X_n$ be a finite collection of independent, identically distributed random variables. Suppose that they are continuously distributed with density f and CDF F.

Definition 5.1 (Order Statistic)

Define the random variable $X_{(k)}$ to be the kth ranked value, called the *kth order statistic*. This means that

$$X_{(1)} = \min\{X_1, X_2, ..., X_n\}, \quad X_{(n)} = \max\{X_1, X_2, ..., X_n\} \tag{384}$$

and in general, for any $k \in \{1, 2, ..., n\}$,

$$X_{(k)} = X_j \text{ if } \sum_{l=1}^{n} \mathbb{I}_{X_l < X_j} = k - 1 \tag{385}$$

which means that exactly $k-1$ of the values of X_l are less than X_j. Since F is continuous,

$$X_{(1)} < X_{(2)} < ... < X_{(n)} \tag{386}$$

holds with probability 1. This leads us to define the random variable $X_{(k)}$ representing the kth order statistic.

$$f_{(k)}(y) = \begin{cases} n \binom{n-1}{k-1} y^{k-1}(1-y)^{n-k} & y \in (0,1) \\ 0 & y \notin (0,1) \end{cases} \tag{387}$$

That is, $X_{(k)}$ has the Beta$(k, n-k_1)$ distribution.

5.1 Poisson Arrival Process

A **Poisson Arrival Process** with rate $\lambda > 0$ on the interval $[0, \infty)$ is a model for the occurence of some events which may have at any time. We can interpret the process as a collection of random points in $[0, \infty)$ which are the times at which the arrivals occur.

1. *Interpretation 1.* Set $T_0 = 0$. The arrival times are random variables $0 < T_1 < T_2 < T_3 < ...$ such that the inter-arrival waiting times
$$W_k = T_k - T_{k-1}, \quad k \geq 0 \tag{388}$$
have the property that $\{W_k\}_{k=1}^{\infty}$ are independent Exp(λ) random variables.
2. *Interpretation 2.* For any interval $I \subset [0, \infty)$, let
$$N_I \equiv \text{ number of arrivals that occur in interval } I \tag{389}$$
Then, $N_I \sim$ Poisson$(\lambda|I|)$, and for any collection of disjoint intervals $I_1, I_2, ..., I_n$, the random variables
$$\{N_{I_k}\}_{k=1}^{n} \tag{390}$$
are independent.

Theorem 5.1 ()

These two interpretations of the arrival process are equivalent.

Proof.

In the 2nd interpretation, the statement $N_I \sim \text{Poisson}(\lambda|I|)$ means that

$$\mathbb{P}(N_I = m) = e^{-\lambda|I|}\frac{(\lambda|I|)^m}{m!}, \quad m = 0, 1, 2, 3, ... \tag{391}$$

where $|I|$ is the length of interval I. From the first perspective, notice that

$$T_k = W_1 + W_2 + ... + W_k \tag{392}$$

so that the kth arrival time T_k is a sum of k independent $\text{Exp}(\lambda)$ random variables. Thus,

$$T_k \sim \text{Gamma}(k, \lambda) \tag{393}$$

and therefore has density

$$\lambda e^{-\lambda t}\frac{(\lambda t)^{k-1}}{(k-1)!}, \quad t > 0 \tag{394}$$

Note that the arrival times T_i are not independent of each other, but the wait times W_i are indeed independent.

We can slightly modify this to create a Poisson arrival process over some finite time horizon $[0, L]$. Again, you can do this two ways:

1. Starting with independent $\text{Exp}(\lambda)$ random variables $W_1, W_2, ...$, we define

$$T_k = \sum_{i=1}^{k} W_i \tag{395}$$

   Once you have $T_k > L$, stop.

2. We let $N \sim \text{Poisson}(\lambda L)$, since we are only working in finite interval L. Given $N = n$, let $U_1, U_2, ..., U_n \sim \text{Uniform}([0, L])$. These define the arrival times, and let us order them to get

$$T_k = U_{(k)}, \quad k = 1, 2, ..., N \tag{396}$$

   where $U_{(k)}$ is the kth ordered point, with $T_1 = \min(U_1, ..., U_N)$.

Lemma 5.1 (Memoryless Property)

The $\text{Exp}(\lambda)$ distribution has the property that for all $t, s \geq 0$,

$$\mathbb{P}(W > t + s \mid W > t) = \mathbb{P}(W > s) \tag{397}$$

which is called the *memoryless property*. We can interpret this in the following way. Let W be the time you have to wait for the first arrival. Given that you already waited t units of time, the probability that you have the wait s additional units of time is just the probability that you wait at least s from the beginning. That is, knowing that t units of time have elapsed does not affect the distribution of the remaining waiting time.

Theorem 5.2 ()

Let W be a continuously distributed random variable. Then $W \sim \text{Exp}(\lambda)$ for some $\lambda > 0$ if and only if W satisfies the memoryless property.

6 Markov Chains

I have an entire set of notes dedicated to stochastic processes, but we talk about it on a basic level here.

6.1 Discrete Time Chains

Definition 6.1 (Markov Chain)

A **Markov chain** is a sequence of random variables $\{X_n\}_{n=0}^{\infty}$, which take values in some set $\mathcal{S}$, called the **state space** satisfying the **Markov property**. Since we are working with discrete time chains, we will assume that $\mathbb{S}$ is a countable (and in most cases, finite). Thus, the X_n will all be discrete random variables. We can also think of X_n as a discrete "time" index; that is, X_n is the state of the system at time n. Therefore, the sequence of random variables models a system evolving in a random way.

Definition 6.2 (Markov Property)

A sequence of random variables $\{X_i\}$ satisfies the **Markov property** if

$$\mathbb{P}(X_{n+1} = y \mid X_n = x_n, X_{n-1} = x_{n-1}, ..., X_0 = x_0\} = \mathbb{P}(X_{n+1} = y \mid X_n = x_n\} \tag{398}$$

holds for any choice oc states $y, x_n, x_{n-1}, ..., x_0 \in \mathcal{S}$ and for any $n \geq 1$.

Colloquially, given that one is at state $X_n = x_n$, knowing all the previous states does not help in predicting X_{n+1}. Knowing only the current state is relevant in predicting the next one. We can model this entire system using a matrix.

Definition 6.3 (Transition Matrix)

Assuming that the chain is *time-homogeneous*, the *transition probability matrix* P has elements P_{xy} defined

$$P_{xy} = P(x, y) = \mathbb{P}(X_1 = y \,|\, X_0 = x) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x) \tag{399}$$

which is the probability of moving from state x to state y in one step. The time homogeneous condition refers to the last equality; that is, the one-step transition probabilities don't change with the time index n. Note that if $\mathcal{S}$ is finite, then P is a $|S| \times |S|$ matrix, and if $\mathcal{S}$ is countably infinite, then P is an infinite-dimensional matrix. The axioms of probability imply that A^T is an entry-wise nonnegative stochastic matrix.

Example 6.1 (Random Walks)

A *random walk* on the integers $\mathcal{S} = \mathbb{Z}$ where a point has equal probability of moving right or left can be modeled with the probability function.

$$P(x, y) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x) = \begin{cases} \frac{1}{2} & y = x + 1 \\ \frac{1}{2} & y = x - 1 \\ 0 & otherwise \end{cases} \tag{400}$$

This can be generalized to multiple dimensional random walks on graphs with probability function

$$P(x, y) = \frac{1}{\deg(x)} \tag{401}$$

where $\deg(x)$ is the number of adjacent nodes to node x. In this way, the point hops randomly from node to node, and if the graph is connected, then the walker can visit any vertex in the graph.

Example 6.2 (Discrete Moran Model)

Consider a population of size N. Each individual is one of two types (say, red or blue). At each time step, the system evolves in the following way: First, one of the individuals is chosen uniformly at random to be eliminated from the population; and another individual is chosen uniformly at random to produce one offspring identical to itself. These two choices are made independently. So, if a red individual is chosen to reproduce, and a blue one is chosen for elimination, then the total number of red particles increases by one and the number of blue particles decreases by one. If a red is chosen for reproduction and a red is chosen for elimination, then there is no net change in the number of reds and blues. Let X_n be the number of red individuals at time n. The transition matrix for this chain is

$$P_{ij} = \begin{cases} \frac{i}{N}\left(\frac{N-i}{N}\right) & j = i-1, i \neq 0 \\ \left(\frac{N-i}{N}\right)\frac{i}{N} & j = i+1, i \neq N \\ 1 - 2\left(\frac{N-i}{N}\right)\frac{i}{N} & j = i \\ 0 & \text{otherwise} \end{cases} \tag{402}$$

Note that the states $X_n = 0$ and $X_n = N$ are absorbing states, which represents a phenomenon called *fixation*.

Definition 6.4 (Absorbing State)

A certain state F in the state space $\mathcal{S}$ of a Markov chain is called an **absorbing state** if

$$\mathbb{P}(X_{n+1} = F \mid X_n = F) = 1 \iff \mathbb{P}(X_{n+1} \neq F \mid X_n = F) = 0 \tag{403}$$

Theorem 6.1 ()

Let there exist a time homogeneous Markov chain with transition probability matrix P. Given a probability distribution ν_n (a row vector) representing the a state of a system at time $t = n$, the probability distribution of which state the system will be at when $t = n + 1$ can be calculated by

$$\nu_{n+1} = \nu_n P \tag{404}$$

The probability distribution of the state of the system at $t = n + k$ can be calculated by summing up all of the possible probabilities that lead to each state at $t = n + k$. It is calculated equivalently as matrix multiplication:

$$\nu_{n+k} = \nu_n P^k \tag{405}$$

Definition 6.5 (Initial Distribution)

The distribution ν of a Markov chain at time $t = 0$ is called the *initial distribution* for the chain. That is, ν is the initial distribution if

$$\mathbb{P}(X_0 = x) = \nu(x) \tag{406}$$

Definition 6.6 (Stationary Distribution)

An *invariant distribution*, or *stationary distribution*, is a probability distribution π such that

$$\pi P = \pi \tag{407}$$

This means that

$$\pi P^k = \pi \tag{408}$$

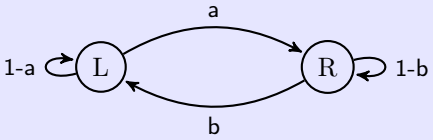
for all $k \in \mathbb{N}$. We can equivalently call π the left eigenvector of matrix P with eigenvalue 1. If π is an invariant distribution for the chain, and $X_0 \sim \pi$, then the distribution of X_n does not change with n; it is invariant. Note that this does not mean that X_n is constant; rather, it means that the distribution of X_n is not changing.

Example 6.3 ()

Let us have a two node system with nodes labeled L and R. That is, $\mathcal{S} = \{L, R\}$. Consider a chain on this state space with transition probability matrix.

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \tag{409}$$

which can be visualized in the following diagram below.



Then, the stationary distribution is

$$\pi = \left(\frac{b}{a+b}, \frac{a}{a+b} \right) \tag{410}$$

Notice that if $a = b = 0$, then this definition is ill-defined, and any probability distribution is invariant since $P = I_2$, the identity matrix.

Definition 6.7 (Recurrent)

A state $x \in \mathcal{S}$ is *recurrent* if

$$\mathbb{P}(X_n = n \text{ for some } n \geq 1 \,|\, X_0 = x\} = 1 \tag{411}$$

That is, if the initial state is x, the chain has probability 1 of returning to x at some later time. If a state is not recurrent, then the state is said to be *transient*. That is, if x is transient, there is some positive probability that the chain will never return to x.

Definition 6.8 (Communication)

Two states $x, y \in \mathcal{S}$ are said to *communicate*, denoted $x \leftrightarrow y$, if there are positive integers n and m such that

$$P^{(n)}(x, y) > 0 \text{ and } P^{(m)}(y, x) > 0 \tag{412}$$

That is, there is some positive probability that the chain can go from x to y and from y to x in some number of steps.

Definition 6.9 (Irreducible Chains)

If all pairs $x, y \in \mathcal{S}$ communicate, then the chain is said to be *irreducible*. If there exists a pair of states that do not communicate, then the chain is said to be *reducible*.

Note that the notion of communication is an equivalence relation between states. That is, it satisfies the properties.

1. $x \leftrightarrow x$.
2. $x \leftrightarrow y \implies y \leftrightarrow x$.
3. $x \leftrightarrow y, y \leftrightarrow z \implies x \leftrightarrow z$.

This relation partitions the state space $\mathcal{S}$ uniquely into transient states and irreducible sub-chains

$$\mathcal{S} = T \cup C_1 \cup C_2 \cup ... \tag{413}$$

More specifically, T is the set of all transient states, and the sets C_k are *closed communication classes*, meaning that

1. For all $x, y \in C_k$, $x \leftrightarrow y$.
2. $P(x, z) = 0$ whenever $x \in C_k$ but $z \notin C_k$.

Note that for all $x, y \notin T$, x and y communicate if and only if x and y are in the same class C_k. Moreover, once the chain reaches one of the sets C_k, it cannot leave C_k.

Definition 6.10 (Period)

For any state $x \in \mathcal{S}$, the *period* of x is defined to be

$$d(x) \equiv \gcd\{n \geq 1 \mid P^{(n)}(x, x) > 0\} \tag{414}$$

Theorem 6.2 ()

It follows that if two states x and y communicate, then they must have the same period: $d(x) = d(y)$. It naturally follows that if the chain is irreducible, then all states must have the same period, and we can define the period of the chain to be $d(x)$ for any x we choose.

Definition 6.11 ()

If an irreducible chain has period 1, the chain is said to be *aperiodic*. Otherwise, the chain is *periodic* with period $d > 1$.

Theorem 6.3 ()

Suppose $|\mathcal{S}| < \infty$. If the chain is irreducible, then there always exists a unique stationary distribution π. If the chain is also aperiodic, then for any initial distribution ν,

$$\lim_{k \to \infty} \nu P^k = \pi \tag{415}$$

Hence

$$\lim_{k \to \infty} P^{(k)}(x, y) = \pi(y) \tag{416}$$

for all $x, y \in \mathcal{S}$. Furthermore, for any function $F : \mathcal{S} \longrightarrow \mathbb{R}$, the limit

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} F(X_n) = \sum_{x \in \mathcal{S}} F(x)\, \pi(x) = \mathbb{E}\big(F(x)\big) \tag{417}$$

holds with probability 1. In particular, the limit does not depend on the initial distribution.

Proof.

The Frobenius Extension to Perron's theorem (Linear Algebra, Theorem 7.31) combined with its applications to stochastic matrices (Linear Algebra, Theorem 7.30) proves this statement.

Definition 6.12 (First Visit)

For each $x \in \mathcal{S}$, define the *first visit* to x by

$$T_x \equiv \min\{n \geq 1 \mid X_n = x\} \tag{418}$$

This T_x is an integer-valued random variable. We say $T_x = +\infty$ if X_n never reaches x. Then, we define the *mean return time* to x by

$$\mu_x \equiv \mathbb{E}\big(T_x \,|\, X_0 = x\big) \tag{419}$$

If x is transient, then $\mu_x = +\infty$, since there is positive probability that $T_x = +\infty$.

Definition 6.13 ()

It is possible that x is recurrent while $\mu_x = +\infty$. If this is the case, then x is said to be *null-recurrent*. If x is recurrent and $\mu_x < \infty$, then x is said to be *positive recurrent*.

Theorem 6.4 ()

An irreducible chain has a stationary probability distribution π if and only if all states are positive recurrent. If a chain is irreducible and all states are positive recurrent, then

$$\pi(x) = \frac{1}{\mu_x} \tag{420}$$

for all $x \in \mathcal{S}$. π is also unique.

6.1.1 Exit Probabilities

Suppose a chain is finite and irreducible. Let $a, b \in \mathcal{S}$ be given states, and let us define $h(x)$ to be the probability of hitting b before a, given that we start from x.

$$h(x) \equiv \mathbb{P}(X_n \text{ reaches } b \text{ before } a \,|\, X_0 = x) \tag{421}$$

Clearly, $h(b) = 1$ and $h(a) = 0$. By conditioning on the first jump out of x, we also have

$$\begin{aligned} h(x) &= \mathbb{P}(X_n \text{ reaches } b \text{ before } a \,|\, X_0 = x) \\ &= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \,|\, X_1 = y, X_0 = x)\, \mathbb{P}(X_1 = y \,|\, X_0 = x) \\ &= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \,|\, X_1 = y, X_0 = x)\, P(x, y) \\ &= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \,|\, X_1 = y)\, P(x, y) \\ &= \sum_y h(y)\, P(x, y) \end{aligned}$$

The sum is over all $y \in \mathcal{S}$ for which $P(x,y) \neq 0$. This gives us a linear system of equations to solve for h

$$\begin{aligned} h(x) &= \sum_y P(x,y)h(y) \ \forall x \in \mathcal{S} \setminus \{a,b\}, \\ h(b) &= 1, \\ h(a) &= 0 \end{aligned}$$

6.1.2 Exit Prize

Let $B \subset \mathcal{S}$ be some subset of the state space, and let $g : B \longrightarrow \mathbb{R}$ be some function. Consider the function

$$h(x) = \mathbb{E}\big(g(X_\tau) \,|\, X_0 = x\big) \tag{422}$$

where $\tau = \min\{n \geq 0 \,|\, X_n \in B\}$ is the first time that the chain reaches some state in the set B (this time is random). We can interpret $g(y)$ as a "prize" that is awarded if the chain first reaches B at state y, which means that $h(x)$ is the expected prize, given that $X_0 = x$. If $x \in B$, then $\tau = 0 \implies h(x) = g(x)$. But if $x \notin B$, then by the same argument as shown in exit probabilities, it is true that h satisfies the linear system of equations

$$\begin{aligned} h(x) &= \sum_g P(x,y)\, h(y), \ \ \forall x \in \mathcal{S} \setminus B, \\ h(x) &= g(x), \ \ x \in B \end{aligned}$$

Note that Exit probability system is a special case of the Exit prize system. In the former, we have defined $B = \{a,b\}$ and g defined by $g(a) = 0, g(b) = 1$.

6.1.3 Occupation Times, Absorbing States

Suppose that a chain on a finite $\mathcal{S}$ is irreducible. Let $B \subset \mathcal{S}$ be some subset of states and let $A = \mathcal{S} \setminus B$ be the other states. Then for $x \in A$, we wish to know how many steps the chain will take before reaching a state in the set B. We define

$$\tau_B = \min\{n \geq 0 \,|\, X_n \in B\} \tag{423}$$

which represents the first time that X is in B, an integer valued random variable. We wish to compute

$$h(x) = \mathbb{E}(\tau_B \,|\, X_0 = x) \tag{424}$$

Clearly, $h(y) = 0$ for all $y \in B$. For $x \in A$, it takes at least one step to reach $B \implies h(x) \geq 1$ for $x \in A$. We condition on the first step from x. This leads to the system

$$h(x) = 1 + \sum_{y \in \mathcal{S}} P(x,y)\, \mathbb{E}(\tau_B \,|\, X_1 = y), \forall x \in A = \mathcal{S} \setminus B$$

Since the chain is time-homogeneous, this means that

$$h(x) = 1 + \sum_{y \in \mathcal{S}} P(x,y)\, h(y), \forall x \in A$$

Since $h(y) = 0$ for all $y \in B$, we now have

$$h(x) = 1 + \sum_{y \in A} P(x,y)\, h(y), \forall x \in A$$

To solve this system, let us define M as the $|A| \times |A|$ submatrix of P obtained by keeping only the entries $P(x, y)$ with $x, y \in A$. So, the system can be written as

$$h(x) = 1 + \sum_{y \in A} M(x, y)\, h(y), \forall x \in A$$

We can solve this system of equations through the equivalent matrix equation

$$(I - M)h = 1 \tag{425}$$

where $1 = (1, 1, ..., 1)^T$ is the column vector consisting of all 1's. The solution vector is therefore

$$h = (I - M)^{-1}1 \tag{426}$$

So, for a particular $x \in A$,

$$h(x) = \sum_{y \in A} (I - M)^{-1}(x, y) \tag{427}$$

Alternatively, we can slightly modify the chain to chain $\tilde{X}_n$ by replacing the transition probability matrix P with another one defined as

$$\tilde{P}(x, y) = \begin{cases} P(x, y) & x \in A, y \in \mathcal{S} \\ 1 & x = y \in B \\ 0 & \text{else} \end{cases} \tag{428}$$

This modification means that all transitions from state in A to any other state are preserved and the only transitions from a state $x \in B$ are self loops. In particular, all transitions from states $x \in B$ to states $y \in A$ are removed. Therefore, under this modified transition matrix, the states in B are absorbing states. The tail sum formula implies that

$$\mathbb{E}(\tau_B \,|\, X_0 = x) = \sum_{k=0}^{\infty} \mathbb{P}(\tau_B > k \,|\, X_0 = x) \tag{429}$$

Notice that since the chain X_n and $\tilde{X}_n$ have the same transition rules before hitting a state B, we have

$$P^{(k)}(x, y) = \tilde{P}^{(k)} = M^{(k)}(x, y) \tag{430}$$

where M is the $|A| \times |A|$ submatrix defined previously. Therefore, putting this all together, we have

$$\begin{aligned} \mathbb{E}(\tau_B \,|\, X_0 = x) &= \sum_{k=0}^{\infty} \mathbb{P}(\tau_B > k \,|\, X_0 = x) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(\tilde{X}_k \in A \,|\, X_0 = x) \\ &= \sum_{k=0}^{\infty} \sum_{y \in A} \tilde{P}^{(k)}(x, y) \\ &= \sum_{k=0}^{\infty} \sum_{y \in A} M^{(k)}(x, y) \\ &= \sum_{y \in A} \left(\sum_{k=0}^{\infty} M^{(k)} \right)(x, y) \end{aligned}$$

Using a theorem from linear algebra, we can show that if all the eigenvalues of a $d \times d$ matrix M have modulus strictly less than 1, then $I - M$ is invertible and

$$\sum_{k=0}^{\infty} M^{(k)} = (I - M)^{-1} \tag{431}$$

where I is the $d \times d$ identity matrix. If M is the $|A| \times |A|$ submatrix described above, one can show that M has his property and that $I - M$ is invertible. Hence,

$$\mathbb{E}(\tau_B \mid X_0 = x) = \sum_{y \in A} \left(\sum_{k=0}^{\infty} M^{(k)} \right)(x, y) = \sum_{y \in A} (I - M)^{-1}(x, y) \tag{432}$$

which refers to the (x, y) entry of the matrix $(I-M)^{-1}$. This is indeed consistent with our previous derivation of the formula for $h(x)$, the expected number of steps before the state reaches B.

6.2 Markov Chain Monte Carlo Algorithms

In statistics, Markov chain Monte Carlo (MCMC) methods comprise of a class of algorithms for sampling from a probability distribution by constructing a Markov chain that has the desired distribution as its equilibrium distribution. That way, by recording samples from the chain, one may get better approximations of the actual distribution.

Let there exist a state space $\mathcal{S}$ with some probability distribution $\pi(x)$ for every $x \in \mathcal{S}$. Clearly,

$$\sum_{x \in \mathcal{S}} \pi(x) = 1 \tag{433}$$

but the problem is that we do not know that π is. We do know, however, another function f that is directly proportional to π.

$$\pi(x) = \frac{f(x)}{c}, \text{ where } c = \sum_{x \in \mathcal{S}} f(x) \tag{434}$$

is the normalizing constant. It is often the case that c is unknown and the state space $\mathcal{S}$ is so large that computing c directly is expensive. Therefore, we construct Markov chains that can provide approximations to π.

6.2.1 Metropolis-Hastings Algorithm

This algorithm is useful because it does not require knowledge of the normalizing constant c. The algorithm only requires evaluations of

$$\frac{\pi(x)}{\pi(y)} = \frac{f(x)}{f(y)} \tag{435}$$

We first have the state space $\mathcal{S}$ consisting of all the possible states. We now construct (any) probability transition matrix q for a Markov chain on $\mathcal{S}$. Note that q is a $|\mathcal{S}| \times |\mathcal{S}|$ matrix and q^T is a stochastic matrix. This matrix is constructed by the user and is completely well-defined and known. We start off with any initial state $x_0 \in \mathcal{S}$ and iterate the following 2-steps to construct a Markov chain.

1. Given a state $X_n = x$, we generate a new state X_{n+1} by first proposing a new state $y \in \mathcal{S}$ with probability $q(x, y)$ (determined from the matrix q).
2. With this chosen state y, we decide whether to accept to reject the proposal. With probability

$$\min\left(1, \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}\right) \tag{436}$$

   we accept the proposal and set $X_{n+1} = y$. Otherwise, the proposal is rejected and the new state is the same $X_{n+1} = x$.

Note that there are two levels of randomness here: which state the new state y will be and whether to accept this state to be the next one or not. If step two did not exist (i.e. the probability of accepting the proposal is always 1), then this would just be a regular Markov chain represented by the matrix q. But the addition of step 2 means that while q is used in constructing the discrete chain X_n, it is *not* the transition probability matrix of X_n.

There is also a lot of flexibility on choosing q, although the performance of the algorithm (speed of convergence of the distribution of X_n to the stationary distribution) will depend on the choice.

Proposition 6.1 ()

For the chain defined by the Metropolis-Hastings algorithm, the distribution π is stationary.

Proof.

Let us write in shorthand

$$\alpha(x,y) = \frac{\pi(y)\,q(y,x)}{\pi(x)\,q(x,y)} \tag{437}$$

First, observe that if $x \neq y$, the transition probability for the chain defined by the algorithm is just

$$P(x,y) = q(x,y)\,\min\{1, \alpha(x,y)\} \tag{438}$$

Next, we claim that for all $x, y \in \mathcal{S}$,

$$\pi(x)P(x,y) = \pi(y)\,P(y,x) \tag{439}$$

This condition is called *detailed balance*. Assuming that $\alpha(x,y) \leq 1$, it is true that

$$\pi(x)P(x,y) = \pi(x)q(x,y)\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} = \pi(y)q(y,x) \tag{440}$$

In this case, we also have $\alpha(y,x) = 1/\alpha(x,y) \leq 1$. So,

$$\pi(y)P(y,x) = \pi(y)q(y,x) \tag{441}$$

and we have proved what we had claimed. Now, summing over x,

$$\sum_x \pi(x)P(x,y) = \sum_x \pi(y)P(y,x) = \pi(y)\sum_x P(y,x) = \pi(y) \tag{442}$$

since P^T is stochastic.

6.2.2 Gibbs Sampling

Let $\mathcal{A} = \{a_1, ..., a_k\}$ be some finite set. Suppose that the state space

$$\mathcal{S} = \mathcal{A} \times ... \times \mathcal{A} = \mathcal{A}^M \tag{443}$$

for some $M \in \mathbb{N}$. The following algorithm generates a Markov chain on $\mathcal{S}$ with stationary distribution

$$\pi(x) = \frac{f(x_1, x_2, ..., x_M)}{c}, \quad x = (x_1, x_2, ..., x_M) \in \mathcal{S} \tag{444}$$

where $c > 0$ is a normalizing constant. Note that $|\mathcal{S}| = k^M$, so computing c may be expensive when M is large. The current state of the chain is denoted

$$X_n = (X_n^1, X_n^2, ..., X_n^M) \tag{445}$$

We think of X_n as having M components, each component taking values in $\mathcal{A}$. We start off with any initial state $X_0 = (X_0^1, X_0^2, ..., X_0^M)$ and construct a Markov chain by iterating the following two steps.

1. Given $X_n = (X_n^1, X_n^2, ..., X_n^M)$, we generate the next state X_{n+1} by picking a component index $i \in \{1, ..., M\}$ uniformly at random.

2. With this chosen, well-defined i, we choose a random $Y^i \in \mathcal{A}$ according to the distribution

$$\mathbb{P}(Y^i = a) = \frac{f\big(X_n^1, ..., X_n^{i-1}, a, X_n^{i+1}, ..., X_n^M\big)}{\sum_{j=1}^k f\big(X_n^1, ..., X_n^{i-1}, a_j, X_n^{i+1}, ..., X_n^M\big)}, \quad a \in \{a_1, ..., a_k\} \tag{446}$$

3. Then, set $X_{n+1} = \big(X_n^1, ..., X_n^{i-1}, Y^i, X_n^{i+1}, ..., X_n^M\big)$.

Note that at each step, only one component of X_n is updated. Observe that the distribution above is also equal to

$$\mathbb{P}(Y^i = a) = \frac{\pi\big(X_n^1, ..., X_n^{i-1}, a, X_n^{i+1}, ..., X_n^M\big)}{\sum_{j=1}^k \pi\big(X_n^1, ..., X_n^{i-1}, a_j, X_n^{i+1}, ..., X_n^M\big)} \tag{447}$$

which is the marginal distribution of the ith component, given the values of the other components.

Proposition 6.2 ()

For the chain defined by this algorithm, the distribution π is stationary.

Proof.

We verify that the detailed balance condition holds. It is also helpful to note that $P(x, y) \neq 0$ if and only if x and y differ in one coordinate.

6.3 Continuous Time Markov Chains

As the name suggests, in a continuous time Markov chain X_t, the time parameter is continuous ($t \geq 0$). As before, the system jumps randomly between states in $\mathcal{S}$, but now the jumps may occur at any time and they occur randomly. This implies that there are *two* sources of randomness:

1. *where* the system jumps and
2. *when* the system jumps

Definition 6.14 (Continuous Time Markov Chain)

The Markov property in the continuous time case says that for any $s, t \geq 0$ and $y \in \mathcal{S}$,

$$\mathbb{P}(X_{t+s} = y \,|\, X_t) = \mathbb{P}(X_{t+s} = y \,|\, X_r \; \forall 0 \leq r \leq t) \tag{448}$$

Colloquially, the conditional distribution of X_{t+s} given the history up to time t is the same as the conditional distribution of X_{t+s} given only X_t. Thus, if we know the current state at t, knowing information about the past doesn't help us better predict the future state X_{t+s}.
In order for the Markov property to hold, the times between jumps must be exponentially distributed random variables because it is the only density that has the memoryless property. This fact has already been stated in a theorem when covering Poisson arrival processes. This is what makes $\text{Exp}(\lambda)$ so important for continuous time Markov chains.

Lemma 6.1 ()

Let $T_1, T_2, ..., T_n$ be independent exponential random variables with rates $\lambda_1, \lambda_2, ..., \lambda_n$, respectively. Then the random variable $T \equiv \min\{T_1, T_2, ..., T_n\}$ is

$$T \sim \text{Exp}\Big(\sum_{i=1}^n T_i\Big) \tag{449}$$

Moreover,

$$\mathbb{P}(T_k = \min\{T_1, ..., T_n\}) = \frac{\lambda_k}{\lambda_1 + ... + \lambda_n} \tag{450}$$

We can interpret the lemma above by imagining that we have n alarm clocks all set simultaneously, which will ring independently at random times. Suppose that clock k will ring after T_k units of time have expired, where T_k is a random variable distributed as $\text{Exp}(\lambda_k)$. Then, $T = \min\{T_1, ..., T_n\}$ is the time at which the first ring occurs.

Example 6.4 ()

The simplest and the most important continuous time Markov chains is the Poisson arrival process. The process really has a single parameter $\lambda > 0$ (the rate of process) by definition and is integer valued. At each jump time, the process increases by 1, and the time between jumps are independent, distributed as $\text{Exp}(\lambda)$.
Notice that when λ is large, the arrivals occur more frequently than when λ is small, because the expected time between arrivals is $1/\lambda$. The second way we can interpret it is to choose an interval of time t and let X_t be the number of jumps that have occurred up to time t. It is a fact that X_t is a integer-valued, $\text{Poisson}(\lambda t)$ distribution. That is,

$$\mathbb{P}(X_t = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!},\ k = 0, 1, 2, ... \tag{451}$$

In particular, $\mathbb{E}(X_t) = \lambda t$ and $\text{Var}(X_t) = \lambda t$.

6.4 Branching Processes

Definition 6.15 (Branching Process)

A *branching process* is a type of Markov chain modeling a population in which each individual produces a random number of children (possibly 0) and dies. The state space is $\mathcal{S} = \{0, 1, 2, 3, ...\}$. Furthermore, there is a discrete-time version and a continuous time version of the chain. In the discrete case, the state is Z_n, the size of the population at time $n = 0, 1, 2, ...$, and in the continuous case, the state is Z_t for $t \geq 0$.

6.4.1 Discrete-time Branching Process

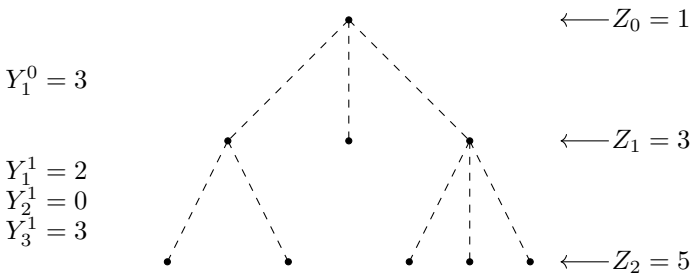
In the discrete case, all of the Z_n individuals in the current generation branch at the same time and immediately die. The branching is independent and distributed according to the *offspring distribution* $\{p_k\}_{k=0}^{\infty}$. Specifically, if $Z_n = m$, then

$$Z_{n+1} = Y_1^n + Y_2^n + ... + Y_m^n \tag{452}$$

where Y_i^n represents the number of offspring the ith individual in the nth generation has. All of them are distributed as

$$\mathbb{P}(Y_i^n = k) = p_k,\ k = 0, 1, 2, 3, ... \tag{453}$$

where p_k is the probability that a parent has k children. Note that if $p_0 \neq 0$, then there is positive probability that $Y_i^n = 0$ for all i, meaning that the population can go extinct. A sample branching process up to the second generation is shown below.



Suppose that the mean number of offspring of a single parent is finite.

$$\mu = \mathbb{E}(Y) = \sum_{k=0}^{\infty} k\,\mathbb{P}(Y = k) = \sum_{k=0}^{\infty} k\,p_k < \infty \tag{454}$$

If Y_1 and Y_2 are two independent, discrete random variables, we can define their convolution and use the fact that $\mathbb{P}(Y_i = k) = p_k$ to get

$$\begin{aligned}\mathbb{P}(Y_1 + Y_2 = k) &= \sum_j \mathbb{P}(Y_1 = k - j)\,\mathbb{P}(Y_2 = j) \\ &= \sum_{j=0}^{\infty} p_{k-j} p_j, \quad k = 0, 1, 2, ...\end{aligned}$$

This is a two-fold convolution of the sequence $\{p_k\}$ with itself, denoted

$$p_k^{*2} = \sum_{j=0}^{\infty} p_{k-j}\,p_j \tag{455}$$

Extending this, we can find the m-fold convolution of the sequence $\{p_j\}$ with itself, represented by the sequence $\{p_j^{*m}\}$, where p_k^{*m} is the kth term in this sequence. This gives us

$$p_k^{*n+1} = \sum_{j=0}^{\infty} p_{k-j}\,p_j^{*n} \tag{456}$$

for all $n \in \mathbb{N}$. Using this, we can write down the transition probabilities for the Markov chain Z_n.

$$\mathbb{P}(Z_{n+1} = k \,|\, Z_n = m) = \begin{cases} 0 & \text{if } m = 0 \\ p_k^{*m} & \text{if } m \geq 1, k \geq 0 \end{cases} \tag{457}$$

where $\mathbb{P}(Z_{n+1} = k \,|\, Z_n = m)$ represents the probability of the nth generation consisting of m individuals producing a total of k offspring for the $(n+1)$th generation. Thus, the branching process is completely determined by the distribution of Z_0 and the offspring distribution $\{p_k\}_{k=0}^{\infty}$.

Lemma 6.2 ()

Given this discrete-time branching process, let μ be the mean of the offspring distribution. Then,

$$\mathbb{E}(Z_n \,|\, Z_0 = 1) = \mu^n \tag{458}$$

If $\mu > 1$, the mean of Z_n grows exponentially, and if μ_1, the mean of Z_n decreases exponentially.

6.4.2 Continuous-time Branching Process

A continuous time branching process Z_t has very similar structure to the discrete time branching process, except that the times between branch events (for each individual) are independent exponentially distributed

random variables $\text{Exp}(\lambda)$, where the parameter $\lambda > 0$ is the branching rate. It is as though each individual has an independent alarm clock which rings as a time that is $\text{Exp}(\lambda)$, independently of all other clocks. So, if there are currently N individuals, then the next alarm will ring at rate λN; that is, the time until the next ring is distributed as $\text{Exp}(\lambda N)$, since it is the minimum of N independent $\text{Exp}(\lambda)$ random variables. When an individual branches (clock rings), that individual produces a random number of offspring, according to the offspring distribution $\{p_k\}$, as before. So, a continuous time branching process has the same geneological structure as the discrete time process, but the times between branch events is randomized. Consequently, whether or not the process eventually goes extinct, depends only on the offspring distribution, not on the branching rate λ.

Let $m_1(t) = \mathbb{E}(Z_t)$ denote the expected population size at time t. Then, it is a fact that $m_1(t)$ satisfies the ordinary differential equation

$$\frac{d}{dt} m_1(t) = \lambda(\mu - 1) m_1(t) \tag{459}$$

where

$$\mu = \sum_{k=1}^{\infty} k p_k \tag{460}$$

is the mean of the offspring distribution. Solving this equation reveals that

$$m_1(t) = e^{\lambda(\mu-1)t} m_1(0) \tag{461}$$

If $\mu > 1$, the mean population size grows exponentially, and if $\mu < 1$, the mean population size decreases exponentially.

6.4.3 Extinction Probability, Generating Functions

The expression for the transition probabilities of Z_n (disrete case) is quite difficult to work with. Alternatively, it can be convenient to work with generating functions.

Definition 6.16 (Generating Function)

The *generating function* for the offspring distribution is the function

$$G(s) \equiv \sum_{k=0}^{\infty} p_k \, s^k = \mathbb{E}(s^Y) \tag{462}$$

where $Y \sim \{p_k\}$ is a random variable representing the number of children produced by a given individual. Note that G is a power series that simply encodes information about the offspring distribution (also a sequence) $\{p_k\}_{k=0}^{\infty}$.

Theorem 6.5 (Properties)

Properties of the generating function.

1. The radius of convergence of $G(s)$ is at least 1. $G(s)$ defines a continuous function on $|s| \leq 1$.
2. On the interval $[0, 1]$, $G(s)$ is increasing and convex. If $p_0 + p_1 < 1$, then $G(s)$ is strictly convex for $s \in [0, 1]$.
3. $G(0) = p_0$.
4. $G(1) = 1$.
5. $G'(1^-) = \mu$ is the expected number of offspring of a single individual.

Proof.

We use the fact that

$$\sum_{k=0}^{\infty} p_k = 1 \text{ and } 0 \geq p_k \geq \ \forall k = 0, 1, 2, ...$$

Theorem 6.6 ()

Suppose that $Z_0 = 1$ and that $p_0 + p_1 < 1$. Then

$$\lim_{n\to\infty} \mathbb{P}(Z_n = 0) = \mathbb{P}(\text{eventual extinction}) = t \tag{463}$$

where $t \in [0, 1]$ is the smallest non-negative root of the equation $t = G(t)$. If $\mu \leq 1$, then $t = 1$ (clearly, since the population will exponentially decrease on average). If $\mu > 1$, there is a positive probability that the population never goes extinct.

Proof.

Let t be the probability that an individual's descendent family tree goes extinct. That is, $t = \mathbb{P}(Z_n = 0$ for some $n \geq 1 \mid Z_0 = 1)$. To derive the equation $t = G(t)$, let us condition on the first generation, with Y_1 denoting the number of offspring of the single parent.

$$\begin{aligned} t &= \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k)\, \mathbb{P}(Y_1 = k \mid Z_0 = 1) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k)\, p_k \end{aligned}$$

That is, given that there are k children of the first individual, the probability that this first individual's descendent family tree will go extinct is equal to the probability that each of the k children's trees go extinct. These k extinction events are independent. Therefore,

$$\mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k) = t^k \tag{464}$$

which implies that

$$t = \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k)\, p_k = \sum_{k=0}^{\infty} t^k\, p_k = G(t) \tag{465}$$

Additionally, under the hypothesis that $p_0 + p_1 < 1$, then $G(s)$ is strictly convex on $[0, 1]$. Hence if $G'(1) = \mu \leq 1$, the smallest non-negative root of $t = G(t)$ must be $t = 1 \implies$ extinction occurs with probability 1. On the other hand, if $G'(1) = \mu > 1$, then the smallest root of $t = G(t)$ occurs in the interval $[0, 1)$.

Note that this result applies to both the discrete time case and the continuous time case. In continuous-time chains, whether or not the population goes extinct does not depend on λ, the rate at which individuals give birth. The λ affects the time at which extinction occurs (if it occurs), but it does not affect the probability that it occurs. However, the extinction probability certainly does depend on the offspring distribution.

Definition 6.17 (Counting Variable)

A random variable X is a *counting variable* if it takes values in $\{0, 1, 2, ...\}$.

Note that generating functions is a mapping from X, the set of counting variables (all assumed to be pairwise independent) to the algebra of power series over variable s.

$$G : X \longrightarrow F[[s]] \tag{466}$$

Lemma 6.3 ()

Let X and Y be two independent random counting variables, with generating functions $G_X(s) = \mathbb{E}(s^X)$ and $G_Y(s) = \mathbb{E}(s^Y)$. Then, the generating function for the random variable $Z = X + Y$ is $G_Z(s) = G_X(s)G_Y(s)$. That is, the generating function mapping G is a homomorphism that maps addition to multiplication. In particular, if X and Y are iid, then $G_Z(s) = G_X(s)^2$.

Proof.

Since X and Y are independent,

$$G_Z(s) = \mathbb{E}(s^Z) = \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X s^Y) = \mathbb{E}(s^X)\mathbb{E}(s^Y) = G_X(s)G_Y(s) \tag{467}$$

Applying this argument iteratively, we get the following lemma.

Lemma 6.4 ()

Let $N \geq 1$ be a fixed positive integer. Let $Y_1, Y_2, ..., Y_N$ be independent, identically distributed random counting variables with generating function $G_Y(s) = \mathbb{E}(s^Y)$. Then, the generating function for the sum $Z = Y_1 + ... + Y_n$ is

$$G_Z(s) = G_Y(s)^N \tag{468}$$

Now, suppose that N is not fixed, but another random variable. We wish to describe the distribution of the sum of a random number of random variables.

Lemma 6.5 ()

Let $Y_1, Y_2, Y_3, ...$ be a collection of independent, identically distributed random variables with generating function $G_Y(s) = \mathbb{E}(s^Y)$. Let N be a random counting variable, independent of the Y_i. Let N have generating function $G_N(s)$. Then the generating function for $Z = Y_1 + Y_2 + ... + Y_N$ is

$$G_Z(s) = G_N\big(G_Y(s)\big) \tag{469}$$

Proof.

Just condition on $N = k$

$$\begin{aligned} G_Z(s) = \mathbb{E}(s^Z) &= \sum_{k=0}^{\infty} \mathbb{E}\big(s^Z \mid N = k\big)\, \mathbb{P}(N = k) \\ &= \sum_{k=0}^{\infty} \mathbb{E}(s^{Y_1 + \ldots + Y_k} \mid N = k)\, \mathbb{P}(N = k) \\ &= \sum_{k=0}^{\infty} G_Y(s)^k\, \mathbb{P}(N = k) \\ &= \mathbb{E}\big(G_Y(s)^N\big) = G_N\big(G_Y(s)\big) \end{aligned}$$

Theorem 6.7 ()

Let $G(s)$ be the generating function for the offspring distribution $G(s) = \sum_{k=0}^{\infty} p_k s^k$. Suppose that $Z_0 = 1$ and let $G_n(s) = \mathbb{E}(s^{Z_n})$ be the generating function for the random variable Z_n. Then,

$$G_{n+m}(s) = G_n\big(G_m(s)\big) = G_m\big(G_n(s)\big) \tag{470}$$

Hence,

$$G_n(s) = G(G(G(...(G(s))...))) \quad \text{n-fold composition} \tag{471}$$

Example 6.5 ()

Suppose the offspring distribution is

$$p_k = qp^k, \;\; k \geq 0 \tag{472}$$

for some $p \in (0, 1)$, where $q = 1 - p$. Thus, the number of children from a given parent is $Y = X - 1$, where $X \sim \text{Geom}(q)$. Then, $\mathbb{E}(Y) = \frac{1}{q} - 1 = \frac{p}{q}$. With some computation, this means that

$$G(s) = \frac{q}{1 - ps} \tag{473}$$

and $t = \min\{1, \frac{q}{p}\}$.

7 Common Distributions

7.1 Multivariate Gaussians

Recall $X \sim \mathcal{N}(\mu, \sigma^2)$ implies that its PDF is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Now we will consider a Gaussian random *vector*, which can be considered a vector of random variables

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

mapping from Ω to $\mathbb{R}^n$. It is not merely a vector where every X_i is Gaussian, as we will show later. That is, a joint distribution that has all n marginal distributions Gaussians does not make a multivariate Gaussian.

This measurable function $\mathbf{X} : \Omega \to \mathbb{R}^n$ induces a probability law $\mathbb{P}_X$ on $\mathcal{B}(\mathbb{R}^n)$, and the Radon-Nikodym theorem states the existence of a PDF f_X such that $\mathbb{P}_X(B) = \int_B f_X \, d\lambda$.

7.1.1 Bivariate Gaussians

Let us first begin with two-variable Gaussians.

Definition 7.1 (Standard Bivariate Gaussian RV)

A random variable (X, Y) is said to be a **standard bivariate Gaussian** if its PDF is of form

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right) \text{ for } \rho \in (-1, 1)$$

Proposition 7.1 ()

Given a standard bivariate Gaussian (X, Y),

1. X and Y are marginally distributed as $\mathcal{N}(0, 1)$. That is, if we integrate a variable (say, x) out, we will get a univariate standard Gaussian PDF of the other (y):

$$\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right) dx = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

2. $\rho_{X,Y}$, the correlation coefficient of X and Y, is equal to ρ.
3. The conditional distribution of X given $Y = y$ is $X \mid Y = y \sim \mathcal{N}(\rho y, 1 - \rho^2)$. That is,

$$f_{X|Y}(x \mid y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(x - (\rho y))^2}{2(1-\rho^2)} \right)$$

4. From (3), we can see that the conditional expectation $\mathbb{E}[X \mid Y = y] = \rho y$ since $X \mid Y = y$ has mean at ρy. Therefore, the conditional expectation of X given Y (which is a random variable) is

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

i.e. $\mathbb{E}[X \mid Y]$ is a linear function of Y.

The formula of the general bivariate Gaussian $\mathbf{X} = (X_1, X_2)$ PDF is messy, but we will put it here.

$$f_{X_1,X_2}(x,y) = \frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{\exp}\left[-\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2} \right) \right]$$

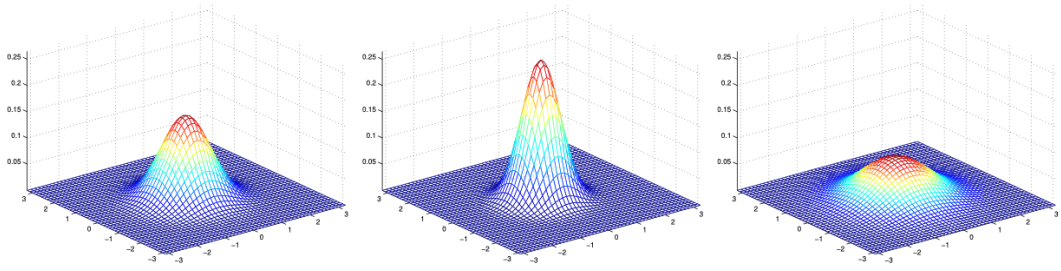
It is cleaner to put it into matrix form.

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sqrt{\det(\mathbf{\Sigma})}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2} \right)$$

where

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ mu_2 \end{pmatrix}, \quad \mathbf{\Sigma} = \mathbb{E}\big[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\big] = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1,X_2) \\ \text{Cov}(X_1,X_2) & \text{Var}(X_2) \end{pmatrix}$$

Note that visually, $\mathbf{\Sigma}$ will determine how much the Gaussian distribution is "stretched" on one way or another. Obviously, the "peak" of the distribution will be $\boldsymbol{\mu}$. If $\mathbf{\Sigma} = I$, then we could visualize the Gaussian distribution as being perfectly symmetric. However, if we scale the distribution up to a certain constant (below shown $\mathbf{\Sigma} = I$, $\mathbf{\Sigma} = 0.61I$, $\mathbf{\Sigma} = 2I$), we get



Now we've made a remark before that given a multivariate distribution $\mathbf{X} = (X_1, \dots, X_n)$, all of its marginal distributions being Gaussian does not mean that $\mathbf{X}$ is a multivariate Gaussian. We give a counterexample.

Example 7.1 ()

Let Y_1, Y_2 be iid random variables distributed according to the PDF

$$f_Y(y) = \sqrt{\frac{2}{\pi}}e^{-y^2/2} \text{ for } y > 0$$

which we can interpret as a one-sided Gaussian. Let $W \sim \text{Bernoulli}(\frac{1}{2})$ be independent of Y_1 and Y_2. Now, define the random variables

$$X_1 = W\,Y_1 \text{ and } X_2 = W\,Y_2$$

This is clearly not a multivariate Gaussian, even though the marginals are $X_1, X_2 \sim \mathcal{N}(0,1)$. We could make the degenerate case that $X_1 = X_2$, which would make the image of (X_1, X_2) just the line at $x_1 = x_2$, but we can think of this as a degenerate Gaussian with a singular $\mathbf{\Sigma}$.

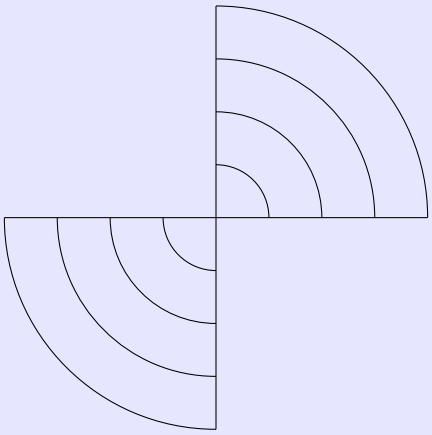


Figure 3: Now note that Y_1 and Y_2 are both positive, and since X_1 and X_2 are both dependent on the same value of W, it is either X_1 and X_2 are both positive or both negative. So, the joint distribution of X_1, X_2 will be on only the 1st and 3rd quadrant with no mass on the 2nd and 4th.

7.1.2 Multivariate Gaussians

There are three equivalent definitions of multivariate Gaussians of n-variables.

Definition 7.2 (Multivariate Gaussian)

Let us have a vector-valued random variable $\mathbf{X} = (X_1 \dots X_n)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. $\mathbf{X}$ is a **multivariate Gaussian distribution** with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and symmetric, positive-definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ if its probability density function is

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(x - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (x - \boldsymbol{\mu}) \right)$$

   The covariance matrix $\boldsymbol{\Sigma}$ is the $n \times n$ matrix whose (i,j)th entry is $\text{Cov}(X_i, X_j)$. That is, for any random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$, its covariance matrix

$$\boldsymbol{\Sigma} = \mathbb{E}\big[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\big] = \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

   is positive definite and symmetric, which implies by the spectral theorem we can break it down into n orthogonal eigenspaces of positive eigenvalues.
2. X is a multivariate Gaussian distribution if it can be expressed as

$$\mathbf{X} = \mathbf{D}\mathbf{w} + \boldsymbol{\mu}$$

   where $\mathbf{w}$ is a vector of independent $\mathcal{N}(0,1)$ Gaussians, $\boldsymbol{\mu} \in \mathbb{R}^n$, and $\mathbf{D} \in \mathbb{R}^{n \times n}$. The mean of $\mathbf{X}$ is $\boldsymbol{\mu}$ and its covariance is $\boldsymbol{\Sigma} = \mathbf{D}\mathbf{D}^T$; $\mathbf{D}$ is called the **standard deviation matrix**. When modeling high-dimensional Gaussians, this way is most computationally feasible.
3. X is a multivariate Gaussian distribution if for every $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{a}^T\mathbf{x}$ is a Gaussian RV. This means that if we take $\mathbf{a} = \mathbf{0}$, then the entire $\mathbf{X}$ is constantly 0, which we will take to be the degenerate Gaussian with mean, variance 0.

The n semi-axes of the $(n-1)$-dimensional isocontour ellipsoid formed by an n-dimensional Gaussian distribution are precisely the normalized eigenvectors of $\boldsymbol{\Sigma}$ multiplied by their eigenvalues.

If we let $\boldsymbol{\Sigma} = \mathbf{I}$, then this means that all the X_i's are pairwise uncorrelated since $\Sigma_{ij} = \text{Cov}(X_i, X_j) = 0$. In general, this does not mean that the X_i's are independent, but for joint Gaussians, this also implies independence!

Theorem 7.1 ()

Given multivariate Gaussian $\mathbf{X} = (X_1 \dots X_n)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the X_i's are pairwise independent if and only if they are uncorrelated.

Proof.

We can expand the PDF of $\mathbf{X}$ as

$$\begin{aligned} f_X(x) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(x-\mu)^T(x-\mu) \right) \\ &= \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(\sum_{i=1}^n -\frac{1}{2}(x_i - \mu_i)^2 \right) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu_i)^2 \right) \end{aligned}$$

which is the product of n single-variable Gaussians X_i. Therefore this means that independence and uncorrelation are equivalent!

Therefore, if the nondiagonal entries of the covariance matrix are all 0, then we know that the variables are all uncorrelated and therefore independent.