EECS 126 Notes

Kanyes Thaker

Spring 2019

0.1 Introduction

This document is an overview of EECS 126, Probability and Random Processes, at UC Berkeley. These notes are largely based off of Introduction to Probability by Dimitris P. Bertsekas and John N. Tsitsiklis, and lectures by Shyam Parekh. This is not an introductory class to probability, and these notes assume a basic understanding of probability from CS70, STAT134, or similar. These are not a replacement for lectures, labs, or discussions, but should solid enough for review!

Contents

1 Sample Spaces and Probability

Probability is an attempt to discuss an uncertain situation. It's not a concept that's uniformly or universally shared or understood, and we study it in order to create a more concrete understanding of uncertainty. Some people define probability as a **frequency of occurrence**, where we try to examine the number of successful occurrences in a large number of trials. We can also define it as an expression of **subjective belief** – what would you say the probability is that you'll do your laundry today? Here we aim to construct a more well-defined notion of probability.

Sets

A **set** is a collection of objects (the **elements** of the set), denoted as

$$\mathcal{S} = \{x_1, ..., x_n\}.$$

If there exists some mapping $f : \mathbb{N} \mapsto \mathcal{S}$, we call the set countably infinite. We can denote a set by the property it satisfies, i.e. $\mathcal{S} = \{x|P(x)\}$. If the set does not have elements which can be enumerated by the positive integers (i.e. a continuous range) we call the set **uncountable**. The universal set is denoted as Ω, and typically we are concerned with $\mathcal{S} \subseteq \Omega$.

The **complement** of $\mathcal{S}$ relative to Ω is the set $\{x \in \Omega | x \notin \mathcal{S}\}$. The **union** of two sets is the set of elements which belong to either set or both sets ($\mathcal{S} \cup \mathcal{T}$), and the **intersection** of two sets is the set of elements which belong to both sets ($\mathcal{S} \cap \mathcal{T}$). We have a general notation for the unions and intersections of multiple sets.

$$\bigcup_{n=1}^{\infty} \mathcal{S}_n = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \ldots = \{x | \exists \mathcal{S}_n \in \mathcal{S} : x \in \mathcal{S}_n\}$$

$$\bigcap_{n=1}^{\infty} \mathcal{S}_n = \mathcal{S}_1 \cap \mathcal{S}_2 \cap \ldots = \{x | \forall \mathcal{S}_n \in \mathcal{S} : x \in \mathcal{S}_n\}.$$

Two sets are **disjoint** if their intersection is empty. A collection of sets is a **partition** of $\mathcal{S}$ if $\bigcup \mathcal{S}_n = \mathcal{S}$ and $\bigcap \mathcal{S} = \emptyset$. Two of the most important properties of sets are given by **DeMorgan's Laws:**

DeMorgan's Laws:

$$\left(\bigcup_c \mathcal{S}_n\right)^c = \bigcap_n \mathcal{S}_n^c, \quad \left(\bigcap_n \mathcal{S}_n\right)^c = \bigcup_n \mathcal{S}_n^c.$$

1.1 Probabilistic Models

Probabilistic Models consist of a sample space Ω and a probability law that assigns a probability $\mathbf{P}(A)$ to each event A that encodes our knowledge about the "likelihood" of the elements in A. In the real world, we use **experiments** (3 coin flips, 2 dice rolls, etc.) in order to produce exactly one of the several possible outcomes. Elements in a sample space should be **mutually exclusive** and also **collectively exhaustive** – they shouldn't overlap and together they should encompass every possible outcome.

Probability Axioms:

1. **(Nonnegativity)** $\mathbf{P}(A) \geq 0$ for every event A
2. **(Additivity)** If A and B are two disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) \cup \mathbf{P}(B).$$

More generally, for a sequence of **disjoint** events, $\mathbf{P}(\bigcup A_i) = \sum \mathbf{P}A_i$.
3. (Normalization) The probability of the entire sample space is 1, that is $\mathbf{P}(\Omega) = 1$.

Properties of Probability Laws:
Consider a probability law and events A, B, and C.

1. If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.
2. $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$
3. $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$
4. $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^{\cup} \cap B^c \cap C)$

We can generalize property (c) in the following:

The Union Bound

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

1.2 Conditional Probability

Conditional probability gives us a way to reason about an experiment based on **partial information**. More exactly, given an experiment, sample space, and probability law, say that we know the outcome is part of some event B. We want to know the likelihood that the outcome also belongs to a different event A. In other words, we want to know the **conditional probability of A given B**, or $\mathbf{P}(A \mid B)$. Conditional probabilities can be thought of as a probability law on a **new universe** B, since the questions we're asking are entirely focused on B. We can logically reason that the probability of A given B would be the probability of being in the intersection of A and B divided by the total probability of B, or:

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \iff \mathbf{P}(A \cap B) = \mathbf{P}(B)\mathbf{P}(A|B).$$

Using this, we can visualize the probability of a large event A as the intersection of the probability of $A_1, ..., A_n$, where A_i is one of many events that need to happen for A to happen. This is expressed here:

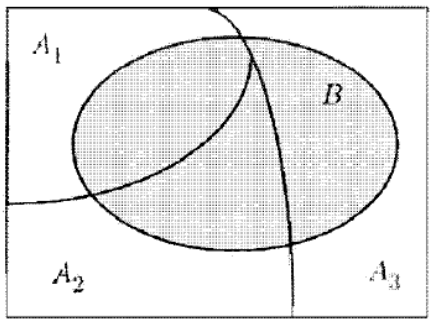
Multiplication Rule

$$\mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbf{P}(A_1) \prod_{i=2}^{n} \mathbf{P}\left(A_i | \bigcap_{j=1}^{i-1} A_j\right)$$

1.3 Total Probability Theorem and Bayes' Rule

Let $A_1, ..., A_n$ be a partition of Ω, i.e. every single point in Ω is in exactly one of $A_1, ..., A_n$. Then for any event B,

$$\mathbf{P}(B) = \mathbf{P}(A_1 \cap B) + \ldots + \mathbf{P}(A_n \cap B)$$



We've discussed the conditional probability of $\mathbf{P}(A|B)$, but we would like to relate this to a similar quantity, $\mathbf{P}(B|A)$. We have a nicely stated rule that helps us form this relationship.

Bayes' Rule

$$\mathbf{P}(A_i|B) = \frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\mathbf{P}(B)}.$$

Bayes' rule is often used in **inference**. We often have a large number of causes that may result in a certain effect – we examine the effect and try to determine the likelihood of each probable **cause**. $\mathbf{P}(B|A_i)$ is the probability that B occurs given the cause is A_i. Conversely, given that effect B has been observed, we want to find $\mathbf{P}(A_i|B)$ that the cause is in fact A_i. We refer to $\mathbf{P}(A_i|B)$ as the **posterior** of A_i, and we refer to $\mathbf{P}(A_i)$ as the **prior**.

1.4 Independence

Conditional probability ($\mathbf{P}(A|B)$) Allows us to capture the information that B provides about event A. We then have to consider the case when B provides *no* such information. Two events A and B are **independent** if

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

If $\mathbf{P}(B) > 0$, independence is equivalent to the condition

$$\mathbf{P}(A|B) = \mathbf{P}(A),$$

meaning that the fact that B occurred doesn't give us new information on whether or not A happened. If A and B are independent, then A and B^c are

also independent. Two events A and B are **conditionally independent** to another event C with $\mathbf{P}(C) > 0$ if

$$\mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C).$$

If $\mathbf{P}(B \cap C) > 0$, then we can further equate this to

$$\mathbf{P}(A|B \cap C) = \mathbf{P}(A|C).$$

Keep in mind that just because two events A and B are independent, they are not necessarily conditionally independent.

We say that *several* events $A_1, A_2, ..., A_n$ are **independent** if for every subset $\mathcal{S}$ of $\{1, 2, ..., n\}$,

$$\mathbf{P}\left(\bigcap_{i \in \mathcal{S}}\right) = \prod_{i \in \mathcal{S}} \mathbf{P}(A_i).$$

If any pair of events within a set of events are independent of each other, they are **pairwise independent**. Note that just because a set of events are pairwise independent, they aren't necessarily independent (and vice versa). Why? The intuition behind independence is that for a group of events, removing *any number* of those events doesn't effect the probability of the remaining events.

1.5 Counting

Here are some basic counting rules:

1. **Permutations** of n objects: $n!$
2. k**-Permutations** of n objects: $n!/(n-k)!$
3. **Combinations** of k of n objects: $\binom{n}{k} = n!/[k!(n-k)!]$
4. **Partitions** of n objects into r grouops, with the ith group having n_i objects: $\binom{n}{n_1, n_2, ..., n_r} = n!/(n_1! n_2! \dots n_r!)$

2 Discrete Random Variables

In some probabilistic models, the outcomes are always numerical – they're prices, instrument readings, or gathered data. In many others, however, the outcomes aren't numerical but they might be *related* to a number. For instance, we might be looking at how many students get a certain GPA. When dealing with data like this, we want to assign numbers to the notion that a certain event has a certain outcome. We introduce for this the concept of the **random variable**.Given an experiment and a set of outcomes, a **random variable** assigns each outcome to a number (the **value** of the random variable). A random variable could be "the sum of two rolls of a die" or "the time needed to make a trip." While these concepts themselves aren't numerical, they are tied to real numbers, and random variables assign numbers to those concepts. In more formal terms, a random variable is a **function** $f : \Omega \mapsto \mathbb{R}$ of the experimental outcome.

A random variable is **discrete** if the set of values it can take on is either finite or countably infinite. Random variables that can take on an infinite number of values (i.e. "a^2 where a is drawn from $[-1, 1]$") are not discrete.

2.1 Probability Mass Functions

We are most interested in knowing the *probabilities* of each of the possible values of the random variable. For a random variable X, we assign the **probability mass function** (PMF) of X, p_X. More specifically, if x is a value that X can take on, the **probability mass** of x is $p_X(x)$, or the probability of the event $\{X = x\}$:

$$p_X(x) = \mathbf{P}[X = x].$$

Note that we must follow the additivity and normalization axioms of probability laws, meaning that the events $X = x$ are disjoint and form a partition of Ω, and additionally

$$\sum_x p_X(x) = 1.$$

Finding the PMF is simple – for each value $x \in X$, collect all the possible values that could result in $X = x$, and sum their probabilities to get $p_X(x)$.

2.1.1 The Bernoulli Random Variable

Consider an event with only two possible outcomes, such as the flip of a coin. The **Bernoulli** random variable takes on the values 1 or 0, with probabilities

of p and $1-p$, respectively. Formally,

$$X = \begin{cases} 1 & \text{if a head} \\ 0 & \text{if a tail.} \end{cases}$$

Bernoulli PMF

$$p_X(x) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0. \end{cases}$$

While the Bernoulli random variable is extremely simple to understand, its real power comes from when several Bernoulli random variables are combined.

2.1.2 The Binomial Random Variable

Instead of flipping a single coin, now we flip n coins, each with probability p of heads, independent of each other. Let X be the number of heads after n tosses. X is a **binomial** random variable with **parameters** n and p. The PMF of X is as follows:

Binomial PMF

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, ..., n.$$

There are $\binom{n}{k}$ ways to select which of the n coins could be heads, and the probability that they will is $p^k(1-p)^{n-k}$. Note that here we follow the normalization property as well:

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

2.1.3 The Geometric Random Variable

Let's ask a different question now: Let's toss the same coin over, where $\mathbf{P}[\text{heads}] = p$. Then X =the number of tosses for a head to come up for the *first* time is a **geometric** random variable.

Geometric PMF

$$p_X(k) = (1-p)^{k-1} p, \quad k = 0, 1, ...,$$

since we're asking for the probability that there are $k-1$ consecutive tails and then 1 heads on the kth flip. Again we can see that this follows the normalization axiom:

$$\sum_{k=1}^{\infty}(1-p)^{k-1}p = p\sum_{k=0}^{\infty}(1-p)^k = p\cdot\frac{1}{1-(1-p)} = 1.$$

Moving away from the coin flip for a second, think of the geometric random variable as doing something over and over until we hit a "success", where "success" is a loose definition (basically the first time something "meaningful" happens).

2.1.4 The Poisson Random Variable

Let's define the PMF of the **Poisson** random variable before we provide an intuitive explanation for what it means.

Poisson PMF

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, ...,$$

where $\lambda > 0$ characterizes the PMF.

Again, this satisfies normalization:

$$\sum_{k=0}^{\infty}e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\ldots\right) = e^{-\lambda}e^{\lambda} = 1.$$

Think of a binomial random variable with a very small p and very large n. Let X be the number of typos in a book with n words. X is binomial (either a word is a typo or it isn't), but p =the probability a word is misspelled is very small. Instead of bothering with the complicated system of combinations associated with the binomial PMF, we can approximate X using the Poisson PMF. In general, the Poisson PMF is with parameter $\lambda = np$ is a good approximation for the binomial PMF with parameters n and p. Here n is very large and p is very small, i.e.

$$e^{-\lambda}\frac{\lambda^k}{k!} \approx \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}, \quad k = 0, 1, ..., n.$$

2.2 Functions of Random Variables

Given a certain random variable X, we can generate *more* random variables by transforming X. For example, let X represent the temperature in degrees Celsius, and Y be the temperature in degrees Fahrenheit – then $Y = 1.8X + 32$. Here Y is a **linear** function in X, i.e.

$$Y = g(X) = aX + b.$$

We also consider nonlinear functions of the general form $Y = g(X)$. If X is a random variable, then Y is also a random variable, because it is still taking outcomes from a probability space and assigning values to them. This means that for every value $p_X(x)$, we can find an equivalent $p_Y(y)$ by summing all $x \in X$ such that $g(x) = y$.

$$p_Y(y) = \sum_{\{x|g(x)=y\}} p_X(x).$$

2.3 Expectation, Mean, and Variance

The PMF of X gives us several numbers, all of the probabilities of every possible value of X. However, this set of numbers is usually too descriptive to be useful. Instead, we try to express the PMF ins a single representative number. We do this through the **expectation**, which is a weighted (by probability) average of all possible values of X.

Expectation

The **expected value, expectation,** or **mean** of a random variable X with PMF $p_X(x)$ is

$$\mathbf{E}[X] = \sum_x x p_X(x).$$

We interpret the mean as a "representative" value of X, somewhere in the middle of the range. If we take the "mass" portion of "probability mass function" a bit more literally, the expectation is the **center of gravity** of the PMF. If you've taken a physics class, it might be interesting to think about the question of where to place the fulcrum under a beam such that the beam is balanced, and relate it back to this concept!

2.3.1 Variance, Moments, and the Expected Value Rule

Let's examine some more metrics we can gather about the PMF. The **2nd moment** of X is the expected value of X^2 – more generally, the n**th moment** is $\mathbf{E}[X^n]$. The second most important quantity associated with a random variable (aside from $\mathbf{E}[X]$) is the **variance**, $\mathbf{Var}(X)$.

Variance

$$\mathbf{Var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2.$$

If we examine this expression, we see that we're measuring how far (on average) X deviates from its mean. The **standard deviation** is easier to interpret (since it's in the same units as X):

$$\sigma_X = \sqrt{\mathbf{Var}(X)}.$$

Calculating $(X - \mathbf{E}[X])^2$ can be tricky sometimes, even though it's possible to find it using our function properties above and the definition of $\mathbf{E}[X]$. However, we have an easier method to calculate $\mathbf{Var}(X)$.

Let X be a random variable with PMF p_X, and let $g(X)$ be a function of X. Then

$$\mathbf{E}[g(X)] = \sum_x g(x)p_X(x).$$

This is the expected value rule for functions of random variables. This simplifies everything greatly – we have now that

$$\mathbf{Var}(X) = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] = \sum_x (x - \mathbf{E}[X])^2 p_X(x).$$

2.3.2 Properties of Mean and Variance

If $Y = aX + b$, then $\mathbf{E}[Y] = a\mathbf{E}[X] + b$ and $\mathbf{Var}(Y) = a^2\mathbf{Var}(X)$.

2.3.3 Mean and Variance of Some Common Random Variables

1. **Bernoulli(p):** $\mathbf{E}[X] = p$, $\mathbf{Var}(X) = p(1-p)$
2. **Uniform(k):** $\mathbf{E}[X] = \frac{a+b}{2}$, $\mathbf{Var}(X) = \frac{(b-a)(b-a+2)}{12}$. a and b are the bounds of the uniform distribution. These can be verified through induction.

3. **Pois(λ):** $\mathbf{E}[X] = \lambda$, $\mathbf{Var}(X) = \lambda$
4. **Geom(p, k):** $\mathbf{E}[x] = \frac{1}{p}$, $\mathbf{Var}(X) = \frac{1-p}{p^2}$.

2.4 Joint PMFs of Multiple Random Variables

In the real world, we often want to examine models involving multiple random variables. Consider two discrete random variables X and Y. The probabilities of the values that X and Y can take is the **joint PMF** of X and Y, written as $p_{X,Y}$. If (x, y) is a pair of possible values of X and Y, then

$$p_{X,Y} = \mathbf{P}(X = x, Y = y).$$

Note that $\mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x \cap Y = y)$. We can calculate the individual PMFs of X and Y, the **marginal PMFs**, from the joint PMF:

$$p_X(x) = \sum_y p_{X,Y}(x, y), \quad p_Y(y) = \sum_x p_{X,Y}(x, y).$$

This comes from the law of total probability.

2.4.1 Functions of Multiple Random Variables

Now that we have this concept of joint PMFs, we can extend it to functions of multiple variables. Consider a function $Z = g(X, Y)$. Its PMF can be calculated from the $p_{X,Y}$, just like we did in the single variable case:

$$p_Z(z) = \sum_{\{(x,y)|g(x,y)=z\}} p_{X,Y}(x, y).$$

Likewise, the expected value for rule extends here as well:

$$\mathbf{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y).$$

We can generalize this: Let $Y = g(X_1, X_2, ..., X_n)$. Then

$$p_Y(y) = \sum_{\{(x_1,x_2,...,x_n)|g(x_1,x_2,...,x_n)\}} p_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n),$$

$$\mathbf{E}[Y] = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} g(x_1, x_2, ..., x_n) p_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n).$$

Here we also explore the a crucial property of expectation.

Linearity of Expectation

$$\mathbf{E}[a_1X_2+a_2X_2+...+a_nX_n+b] = a_1\mathbf{E}[X_1]+a_2\mathbf{E}[X_2]+...+a_n\mathbf{E}[X_n]+b.$$

2.5 Conditioning

Recall the conditional probabilities of events we discussed in the first section. Here we introduce conditional PMFs, the occurrence of an event given the value of another random variable. This is nothing new, just a natural extension of the concepts introduced earlier.

2.5.1 Conditioning a Random Variable on an Event

The **conditional PMF** of a random variable X, conditioned on an event A with $\mathbf{P}(A) > 0$, is defined as

$$p_{X|A}(x) = \mathbf{P}(X = x|A)\ \frac{\mathbf{P}(X = x \cap A)}{\mathbf{P}(A)}.$$

Since the events that compose X are disjoint (by definition), this leads us to the interesting observation that

$$P(A) = \sum_x \mathbf{P}(X = x \cap A).$$

2.5.2 Conditioning one Random Variable on Another

If X and Y are variables in the same experiment, having knowledge of Y gives some color as to the value of X. We capture this in the **conditional PMF** of X given Y, which is determined by specializing A from the previous example to $Y = y$.

$$p_{X|Y}(x|y) = \mathbf{P}(X = x|Y = y),$$

$$p_{X|Y}(x|y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(Y = y)}.$$

2.6 Conditional Expectation

The conditional PMF can be thought of as the normal PMF over a new universe determined by the condition. An analogous case follows for the conditional expectation (and conditional variance). The conditional expectation of X given an event A, $\mathbf{P}(A) > 0$, is

$$\mathbf{E}[X|A] = \sum_x x p_{X|A}(x), \quad \mathbf{E}[g(X)|A] = \sum_x g(x) p_{X|A}(x).$$

The conditional expectation of X given a value y of Y is

$$\mathbf{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y).$$

From the total probability theorem, we get an equivalent **total expectation theorem**,

$$\mathbf{E}[X] = \sum_y p_Y(y)\mathbf{E}[X|Y = y].$$

2.7 Independence

Independence with random variables is analogous to independence of events. The independence of a random variable from an event is similar to the independence of two events. X is independent from the event A is

$$\mathbf{P}(X = x \text{ and } A) = \mathbf{P}(X = x)\mathbf{P}(A) = p_X(x)\mathbf{P}(A), \quad \forall x.$$

This is essentially saying that for all possible values of $X = x$, x and A are independent.
We can extend this concept to two random variables. Two random variables are independent if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \quad \forall x, y.$$

Likewise, X and Y are **conditionally independent** given an event A if

$$\mathbf{P}(X = x, Y = y|A) = \mathbf{P}(X = x|A)\mathbf{P}(Y = y|A).$$

If X and Y are independent random variables, then

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y] \quad \mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y).$$

3 General Random Variables

Lots of random variables don't necessarily take on a finite set of discrete values. For **continuous random variables**, we try to measure events that can occur on an infinite spectrum, such as velocities or amounts of liquid.

Continuous Random Variables and PDFs

Continuous Random Variables

A random variable X is called **continuous** if there is a nonnegative function f_X, the **probability density function** or **PDF** of X,

$$\mathbf{P}(X \in B) = \int_B f_X(x)dx$$

for every subset B of the real line.

Here the integral used is the **Riemann/Darboux integral** from most calculus classes, and is implicitly assumed to be well-defined. To be more specific, the probability that the value of X falls within a range $[a, b]$ is

$$\mathbf{P}(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

To qualify as a PDF, f_X must be nonnegative ($f_X(x) \geq 0 \ \forall x$), and must also satisfy the normalization property:

$$\int_{-\infty}^{\infty} f_X(x)dx = \mathbf{P}(-\infty < X < \infty) = 1.$$

The **expected value** or **mean** of a continuous random variable X is

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

This is identical to the discrete case except the PMF is replaced by the PDF and the summation is replaced by integration. Imagine a variable $Y = g(X)$. Y is a random variable, but it isn't necessarily continuous. It follows the **expected value rule**, i.e.

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

The n**th moment** of a continuous r.v. X is $\mathbf{E}[X^n]$, and the *variance* is defined identically to the discrete case, as $\mathbf{E}[(X - \mathbf{E}[X])^2]$.

3.1 Exponential Random Variables

An **exponential** random variable has PDF

Exponential PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

An exponential random variable is a good model for the amount of time until something important takes place (a message arriving, a lightbulb burning out, etc). It is closely related to the geometric random variable, which is an analog in discrete time.

$$\mathbf{E}[X] = \frac{1}{\lambda} \quad \mathbf{Var}(X) = \frac{1}{\lambda^2}$$

3.2 Cumulative Distribution Functions

Until right now we've been dealing with discrete and continuous random variables differently, with the PMF and PDF, respectively. We want to be able to describe all kinds of random variables with a single concept. Here we introduce the **cumulative distribution function** or **CDF**.

Cumulative Distribution Function

The **CDF of** X is denoted F_X and provides the quantity $\mathbf{P}(X \leq x)$.

$$F_X(x) = \mathbf{P}(X \leq x) = \begin{cases} \sum_{k \leq x} p_X(x), & X \text{ is discrete,} \\ \int_{-\infty}^{x} f_X(t)dt & X \text{ is continuous.} \end{cases}$$

$F_X(x)$ "accumulates" the probability "up to" x. Some interesting properties:

1. F_X is monotonically nondecreasing since the PDF and PMF are strictly nonnegative.

2. $F_X(x)$ tends to 0 as $x \to -\infty$ and 1 as $x \to \infty$.

3. If X is discrete, then $F_(X)(x)$ is a piecewise constant function of x.

4. If X is continuous, then $F_X(x)$ is a continuous function of X.

It is interesting to note that the CDF of the geometric $(1-(1-p)^n)$ and exponential $(1-e^{-\lambda x})$ random variables are related – the CDF of the exponential is the limit of the CDF of the geometric.

3.3 Normal Random Variables

Normal Random Variable

A continuous random variable X is **normal** or **Gaussian** if its PDF is of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}.$$

Here $\mathbf{E}[X] = \mu$ and $\mathbf{Var}(X) = \sigma^2$.

A normal random variable Y with zero mean and unit variance is called **standard normal**. Its CDF, denoted by Φ, is

$$\Phi(y) = \mathbf{P}(Y \leq y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}\exp\left\{\frac{-t^2}{2}\right\}dt.$$

Because the normal distribution is **symmetric**, note that $\Phi(y) = 1-\Phi(-y)$, which is useful for negative values of y.

Let $X \sim \mathcal{N}(\mu, \sigma)$. Then we can "standardize" X by defining a variable Y such that

$$Y = \frac{X-\mu}{\sigma}.$$

Y now has mean 0 and variance 1.

A vital property to keep in mind: *the sum of a large number of independent and identically distributed (i.i.d.) (not necessarily normal) random variables has an approximately normal CDF, regardless of the CDF of the individual random variables.*

3.4 Joint PDFs of Multiple Random Variables

We can now extend the notion of PDFs to the case of multiple random variables. Just like in the discrete case, we introduce the idea of joint, marginal, and conditional PDFs.

Two continuous random variables are **jointly continuous** and can be described in terms of a **joint PDF** $f_{X,Y}$ if $f_{X,Y}$ satisfies

$$\mathbf{P}((X,Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x,y)dxdy,$$

For every subset $B \in \mathbb{R}^2$. In the case where B is a rectangle of the form $\{(x,y)|a \leq x \leq b,\ c \leq y \leq d\}$ we have

$$\mathbf{P}(a \leq X \leq b,\ c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy.$$

We can interpret the joint PDF as a "probability per unit area" in the vicinity of a certain point. To find the marginal density with respect to X, we simply take

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

Likewise the marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

3.5 Joint CDFs

If X and Y are two random variables associated with the same experiment, we define their joint CDF:

$$F_{X,Y}(x,y) = \mathbf{P}(X \leq x,\ Y \leq y).$$

Again, we use the CDF because it works for both discrete and continuous random variables. In particular, if X, Y are described by the joint PDF $f_{X,Y}$, then

$$F_{X,Y}(x,y) = \mathbf{P}(X \leq x,\ Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)dtds$$

$$\Updownarrow$$

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y).$$

Likewise, by the expected value rule,

$$\mathbf{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)dxdy$$

This can be naturally extended into a a greater number of random variables.

3.6 Conditioning

Similar to how we can condition discrete random variables on events (or other random variables), we can do the same for the continuous random variables.The **conditional PDF** of a continuous random variable X on an event A is defined as a nonnegative function $f_{X|A}$ that satisfies

$$\mathbf{P}(X \in B|A) = \int_B f_{X|A}(x)dx.$$

In the case where X is an element of the conditioning event, we can use a "Bayesian" approach:

$$\mathbf{P}(X \in B|X \in A) = \frac{\mathbf{P}(X \in B, X \in A)}{\mathbf{P}(X \in A)} = \frac{\int_{A\cap B} f_X(x)dx}{\mathbf{P}(X \in A)}.$$

Let $A_1, A_2, ..., A_n$ be a partition of Ω. Then

$$f_X(x) = \sum \mathbf{P}(A_i) f_{X|A_i}(x).$$

The **conditional PDF** of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

This definition is analogous to the discrete case. This can be generalized in a single expression which relates the joint, marginal, and conditional PDFs:

$$f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x|y),\ f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y)dy.$$

Additionally we can find the conditional probability as an integral of the joint PDF:

$$\mathbf{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx.$$

3.7 Conditional Expectation

The justifications behind the below are all in accordance with the discrete case:

$$\mathbf{E}[X|A] = \int_{-\infty}^{\infty} x f_{X|A}(x)dx$$

$$\mathbf{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx$$

For a function $g(X)$,

$$\mathbf{E}[g(X)|A] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x)dx$$

$$\mathbf{E}[g(X)|Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)dx$$

For a partition of the sample space,

$$\mathbf{E}[X] = \sum \mathbf{P}(A_i)\mathbf{E}[X|A_i]$$

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} \mathbf{E}[X|Y=y] f_Y(y)dy.$$

For functions of several random variables,

$$\mathbf{E}[g(X,Y)|Y=y] \int g(x,y) f_{X|Y}(x|y)dx,$$

$$\mathbf{E}[g(X,Y)] = \int \mathbf{E}[g(X,Y)|Y=y] f_Y(y)dy.$$

3.8 Independence

Two continuous random variables X and Y are **independent** if their joint PDF is the product of their marginal PDFs:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y),\ \forall x,\ y.$$

The remainder of the properties are the same as in the discrete case, i.e. $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$, $\mathbf{Var}(X+Y) = \mathbf{Var}(X) + \mathbf{Var}(Y)$, etc.

3.9 Continuous Bayes

Many times, we have some unobserved phenomenon X with PDF f_X, but an observed noisy measurement Y which we model with the conditional PDF $f_{X|Y}$. We can get information about X through **Bayes' Theorem**:

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)}$$

4 Further Topics on Random Variables

4.1 Derived Distributions

Say we are given $Y = g(X)$. Given the PDF of X, we should be able to calculate the PDF of Y. We do this in two steps:

1. Calculate the CDF F_Y of Y using the formula

$$F_Y(y) = \mathbf{P}(g(X) \leq y) = \int_{\{x|g(x)\leq y\}} f_X(x)dx$$

2. Differentiate to obtain the PDF of Y

$$f_Y(y) = \frac{dF_Y}{dy}(y).$$

4.2 Convolutions

Consider the function $Z = X+Y$ of two independent, integer-valued random variables with PMFs p_X and p_Y. Then for any integer z

$$\begin{aligned} p_Z(z) &= \mathbf{P}(X + Y = z) \\ &= \sum_{\{(x,y)|x+y=z\}} \mathbf{P}(X = x, Y = y) \\ &= \sum_x \mathbf{P}(X = x, Y = z - x) \\ &= \sum_x p_X(x)p_Y(z - x) \end{aligned}$$

p_Z is called the **convolution** of the PMFs of X and Y.

Continuous Convolution

For X and Y continuous, and $Z = X + Y$, we get

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x)dx.$$

Notice that this is equivalent to the definition of signal convolution from EE120.

4.3 Covariance and Correlation

So far we've been able to examine what random variables are and how we can create relationships (functions) using them. Here we explore how to find the direction and inherent relationship between two random variables.

Covariance

The **covariance** $\mathbf{cov}(X, Y)$ of two random variables X and Y is defined by

$$\mathbf{cov}(X, Y) = \mathbf{E}\big[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\big].$$

If $\mathbf{cov}(X, Y) = 0$, the two random variables are **uncorrelated**.

The **correlation coefficient** $\rho(X, Y)$ is defined as

$$\rho(X, Y) = \frac{\mathbf{cov}(X, Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}$$

We can use the covariance to find a general formula for the variance of the sum of several (not necessarily independent) random variables.

$$\mathbf{Var}\left(\sum X_i\right) = \sum \mathbf{Var}(X_i) + \sum_{i \neq j} \mathbf{cov}(X_i, X_j)$$

4.4 Conditional Expectation and Variance, Revisited

We can reformulate the law of total expectation into the **law of iterated expectations** and create a **law of total variance** that relates conditional and unconditional variance. As long as X is well defined and has a finite expectation $\mathbf{E}[X]$, the law of iterated expectation says that

$$\mathbf{E}[\mathbf{E}[X|Y]] = \mathbf{E}[X].$$

The law of total variance states:

$$\mathbf{Var}(X) = \mathbf{E}[\mathbf{Var}(X|Y)] + \mathbf{Var}(\mathbf{E}[X|Y]).$$

4.5 Transforms

Here we introduce a **transform** associated with X, also known as the **moment generating function** $M_X(s)$ of a scalar s is

$$M_X(s) = \mathbf{E}[e^{sX}].$$

If X is a discrete random variable, the transform is given by

$$M(x) = \sum_x e^{sx} p_X(x),$$

and if X is continuous it is

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx.$$

In order to get the **moment** from the **transform**, we can differentiate with respect to s. We take

$$\begin{aligned} \frac{d}{ds} M(s) &= \frac{d}{ds} \int_{-\infty}^{\infty} e^{sx} f_X(x) dx \\ &= \int_{-\infty}^{\infty} \frac{d}{ds} e^{sx} f_X(x) dx \\ &= \int_{-\infty}^{\infty} x e^{sx} f(X)(x) dx. \end{aligned}$$

Clearly we can see that evaluating this at $s = 0$ yields $\mathbf{E}[X]$. Likewise, evaluating the nth derivative at $s = 0$ will yield $\mathbf{E}[X^n]$ – try this for yourself if you're not convinced. The transform $M_X(s)$ is **invertible**, meaning we can determine the probability law (CDF, PDF, or PMF) of X. The proof is beyond the scope of this course. The formulas are difficult and cumbersome to use, so we usually invert them through pattern matching to known tables of distribution-transform pairs.

Transforms are particularly useful for the sum of random variables. The **addition of independent random variables corresponds to the multiplication of transforms**, which can provide for us a convenient alternative to the convolution formula. Let $Z = X + Y$. Then

$$M_Z(s) = \mathbf{E}[e^{sZ}] = \mathbf{E}[e^{s(X+Y)}] = \mathbf{E}[e^{sX} e^{sY}] = \mathbf{E}[e^{sX}]\mathbf{E}[e^{sY}] = M_X(s) M_Y(s).$$

More generally:

$$Z = \sum X_i \implies M_Z(s) = \prod M_{X_i}(s).$$

4.6 Order Statistics

Let $n \in \mathbb{Z}^+$. Let $X_1, ..., X_n$ be i.i.d. continuous random variables with PDF f and CDF F. Then for $i = 1, ..., n$ let $X^{(i)}$ be the ith smallest of the rando

mvariables. Then $X^{(i)}$ is known as the **ith order statistic.** The CDF of the order statistic is

$$\mathbf{P}(X^{(i)} \leq x) = \sum_{k=i}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}.$$

Differentiating the CDF gives us the PDF, namely

$$f_{X^{(i)}}(x) = n \binom{n-1}{i-1} f(x) F(x)^{i-1} (1 - F(x))^{n-i}.$$

5 Limit Theorems

Here we examine what happens to sequences of random variables as the size of the sequence becomes large. For a sequence of i.i.d random variables with mean μ and variacne σ^2, we define the sum of the first n variables as

$$S_n = \sum_{i=1}^{n} X_n.$$

Since they are independent, $\mathbf{Var}(S_n) = \sum \mathbf{Var}(X_i) = n\sigma^2$, meaning the variance of S_n becomes large linearly as n, so we can't get any meaningful information here. What we *can* get meaningful information from is the *sample mean*,

$$M_n = \frac{1}{n} S_n.$$

Then

$$\mathbf{E}[M_n] = \mu, \quad \mathbf{Var}(M_n) = \frac{1}{n}\sigma^2,$$

which follows what we expect from some of the laws of large numbers you may have seen in CS70. So now we can see that the variance of the sample mean approaches 0 asymptotically, while the variance of S_n approaches infinity asymptotically. Let's find a happy medium between the two. We can create the zero-mean random variable $S_n - n\mu$ and divide by $\sqrt{n\sigma^2}$ to get

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Then

$$\mathbf{E}[Z_n] = 0, \quad \mathbf{Var}(Z_n) = 1.$$

The **central limit theorem** asserts that as $n \to \infty$, the standardized sum of n random variables approaches standard normal.

5.1 The Markov and Chebyshev Inequalities

The inequalities mentioned here use the mean and variance of random variables to get an idea about the probabilities of certain events. They're especially useful when we can't determine the distribution of X but we *can* calculate its mean and variance easily.

The **Markov inequality** asserts that if a *nonnegative* random variable has

a small mean, the probability that it takes on a large value is small.

Markov's Inequality

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}, \quad \forall a > 0.$$

Note that this is not a tight bound, but it is a maximal upper bound.

The **Chebyshev inequality** makes a similar assertion. It states that if the variance of a random variable is small, the probability that it takes on a value far from the mean is also small. i.e. if X has mean μ and variance σ^2, then the following holds.

Chebyshev's Inequality

$$\mathbf{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}, \quad \forall c > 0.$$

Since the Chebyshev inequality uses both the mean and variance, it allows us to form a slightly tighter bound than the Markov inequality.

5.2 The Chernoff Bound

The Markov and Chebyshev bounds are extremely loose bounds, and as such fare poorly when discussing random variables that fall off exponentially the further you get from the mean. Here we introduce the **Chernoff bound**, which lets us calculate tail probabilities for **independent** random variables that fall off exponentially from the mean.

Let $X_1, ..., X_n$ be independent Poisson random variables with $\mathbf{P}[X_i = 1] = p_i$, and let $X = \sum X_i$ as standard. Then if $\mu = \mathbf{E}[X]$, the general Chernoff bound for $\delta \in (0, 1]$ is

$$\mathbf{P}[X < (1 - \delta)\mu] < \left(\frac{e^{-\delta}}{(1 - \delta)^{(1-\delta)}} \right)^{\mu}.$$

This bound is a bit clunky to work with, so we can relax it a bit to the following:

Chernoff Bound

$$\mathbf{P}[X < (1-\delta)\mu] < \exp\left\{-\mu\delta^2/2\right\}.$$

5.3 The Weak Law of Large Numbers

The **weak law of large numbers** asserts that the sample mean of a large number of i.i.d. random variables is very close to the true mean with high probability. We can use the Chebyshev inequality to elaborate.

The Weak Law of Large Numbers

For a set of i.i.d. random variables with mean μ, $\forall\, \epsilon > 0$ we have

$$\mathbf{P}(|M_n - \mu| \geq \epsilon) = \mathbf{P}\left(\left|\frac{1}{n}\sum X_i - \mu\right| \geq \epsilon\right) \to 0, \quad n \to \infty.$$

5.4 Convergence in Probability

The weak law of large number essentially states that M_n converges to μ. However, we need to define convergence a bit more tightly – after all, M_n is a *variable*, not a constant value. We begin with the traditional definition of convergence used in analysis: for a sequence of real numbers $a_1, a_2, ...$ and another real number a we say that a_n converges to a if:

$$\forall\, \epsilon > 0,\ \exists N \in \mathbb{N} : n > N \implies |a_n - a| \leq \epsilon.$$

Intuitively, this means that for a big enough n, a_n will be within ϵ of a for any ϵ. In probability, we say something similar: for a sequence $Y_1, Y_2, ...$ of random variables, and a real number a, Y_n converges to a if:

$$\forall\, \epsilon > 0,\ \lim_{n\to\infty} \mathbf{P}(|Y_n - a| \geq \epsilon) = 0.$$

We can also rephrase this more generally to define the probability to be **accurate** within some **confidence level** (this is more similar to the $\epsilon - \delta$ definition used in real analysis):

$$\forall\, \epsilon > 0,\ \forall \delta > 0,\ \exists N \in \mathbb{N} : \mathbf{P}(|Y_n - a| \geq \epsilon) \leq \delta.$$

5.5 The Central Limit Theorem

The Central Limit Theorem

Let $X_1, X_2, ...$ be a sequence of i.i.d. random variables with common mean μ and variance σ^2, and define

$$Z_n = \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}.$$

We are taking the sum of random variables, removing the mean, and keeping the variance fixed. Then the CDF of Z_n converges to the standard normal

$$\mathbf{\Phi}(z) = \frac{1}{\sqrt{2}} \int_{-\infty}^{z} \exp\left\{-\frac{x^2}{2}\right\} dx,$$

in the sense that

$$\lim_{n \to \infty} \mathbf{P}(Z_n \leq z) = \mathbf{\Phi}(z), \quad \forall z.$$

This theorem is extremely general – aside from requiring the random variables to be independent and with finite mean and variance, the distribution of the random variables does not matter. The sum of a large number of random variables is approximately normal. We can use this to treat a large sum $S_n = X_1 + \ldots + X_n$ as if it were normal by normalizing it and approximating with the standard normal CDF.

5.5.1 The De Moivre-Laplace Approximation of the Binomial

Remember the way we defined the binomial distribution earlier – as the sum of n independent Bernoulli random variables. We can use the central limit theorem now to approximate the probability of an event $\{k \leq S_n \leq l\}$, where k and l are given integers: the derivation is omitted.

The De Moivre-Laplace Approximation

If S_n is a binomial random variable with parameters n and p, with n large, and $k, l \in \mathbb{Z}_{\geq 0}$, then

$$\mathbf{P}(k \leq S_n \leq l) \approx \mathbf{\Phi}\left(\frac{l + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \mathbf{\Phi}\left(\frac{l - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

5.6 The Strong Law of Large Numbers

Here we also show that the sample mean converges to the tru mean, with a *different* type of convergence. In general, we can write the following statement.

The Strong Law of Large Numbers

Let $X_1, X_2, ...$ be a sequence of i.i.d random variables with mean μ. Then the sequence of sample means $M_n = \frac{1}{n}(X_1, X_2, ...)$ converges to μ **with probability 1** in the sense that

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} = \mu\right) = 1.$$

What then, is the difference betweenn the weak and strong laws? The weak law simply states that the probability that the sample mean deviates from the true mean is low as $n \to \infty$; however, that probability (that it *does* deviate significatly) still exists, but we don't know how how many deviations there will be. The strong law gives us this assurance, by saying that M_n converges to μ with probability 1, so for any ϵ, the probability that $|M_n - \mu|$ will exceed ϵ an infinite number of times is 0.

Note the difference in the convergence emphasized by the strong law versus the weak law. We here define this new "almost sure" definition of convergence in more detail.

Almost Sure Convergence

Let $Y_1, Y_2, ...$ be a sequence of random variables. Let $c \in \mathbb{R}$. Y_n converges to c **with probability 1** or **almost surely** if

$$\mathbf{P}(\lim_{n\to\infty} Y_n = c) = 1.$$

Almost sure convergence implies convergence in probability, but the converse is not generally true!

5.6.1 The Borel-Cantelli Lemma

Suppose that $A_1, A_2, ...$ is a series of events in some probability space Ω. Then the event that A *occurs infinitely many times* (denoted $A(i.o.)$) is:

$$A(i.o.) = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_k.$$

The Borel-Cantelli Lemma

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty \implies \mathbf{P}(A(i.o.)) = 0.$$

This is to say that if the sum of individual probabilities of each event A_i is finite, then the probability that infinitely many of the A_i occur is 0.

The converse of the lemma, sometimes called the **second Borel-Cantelli lemma**, is true if each A_i is independent. Then

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty \implies \mathbf{P}(A(i.o.)) = 1.$$

5.7 Binary Erasure Channels

Suppose we want to send a message over a noisy channel. We do this in three essential steps – we compress the message, add redundancy to deal with noise, and then send the message through the channel. In the 1940's, Claude Shannon proved that we can design our *source* and *channel* coding seperately without impacting the optimal rate.

A **binary erasure channel (BEC)** erases the input to the channel with probability $p \in (0, 1)$. How many bits can the transmitter send over the channel without error? Say we're encoding a length L message with a length $n > L$ message (to account for possible erasures). The ratio L/n is the **rate** of the message. Assume that the receiver can't contact us in the middle of the transmission to tell us which bits were erased. For an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, we also send an **encoding function** $f_n : \mathcal{X}^L \mapsto \mathcal{X}^n$ and a **decoding function** $g_n : \mathcal{Y}^n \mapsto \mathcal{X}^L$. In a BEC, the input alphabet is $\{0, 1\}$ and the output alphabet is $\{0, 1, \mathbf{e}\}$. Now we account for the noise in the channel.

Maximum Probability of Error

Let $X^{(n)}$ and $Y^{(n)}$ be n-length bit strings corresponding to the input and output, respectively. Then

$$P_e(n) := \max_{x \in \mathcal{X}^L} \mathbf{P}\{g_n(Y^{(n)}) \neq x | X^{(n)} = f_n(x)\}.$$

We say a rate R is **achievable** if for each positive integer n, there exists a set of encoding and decoding functions that encode L to n such that $P_e(n) \to 0$ as $n \to \infty$. The **capacity** of a channel is the largest achievable rate.

The capacity of a BEC with erasure probability p is $1 - p$. It is important that we use an intelligent encoding (i.e. Huffman coding) to make sure that the receiver has to ask as few questions as possible to decode the information, i.e. go over a small codebook.

The general result is stated in terms of the **mutual information** of random variables, defined as $I(X;Y) := H(X) + H(Y) - H(X,Y)$. Let X be the channel input and let Y be the channel output, and let $\mathcal{P}$ be the set of probability distributions on $\mathcal{X}$. The **channel capacity** is

$$C := \max_{p \in \mathcal{P}, X \sim p} I(X;Y).$$

Channel Coding Theorem

Any rate below C is achievable. Conversely, any sequence of codes with $P_e(n) \to 0$, $n \to \infty$ has rate $R \leq C$.

6 Discrete Time Markov Chains

Here we consider processes that depend on and (to an extent) can be predicted using what has happened in the past. We can summarize the effect of the past on the future as a **state**, which changes over time according to various probabilities. For this course, we only consider models with a finite number of state values, whose probabilities are time-invariant.

6.1 Discrete-Time Markov Chains

We first start with **Discrete-Time Markov Chains (DTMCs)**, where the state changes at discrete time intervals n. The state of the chain at time n is a random variable X_n, belonging to a finite set $\mathcal{X}$ of possible states (the **state space**). The chain is described by its **transition probabilities** $P(x_n, x_{n+1})$, describing the probability of going from state x_n to state x_{n+1}.

$$P(x_n, x_{n+1}) = \mathbf{P}(X_{n+1} = x_{n+1} | X_n = x_n), \quad x_n, x_{n+1} \in \mathcal{X}.$$

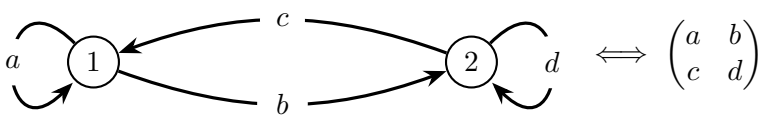
Note something interesting in the above – the probability of going between two states does **not** depend on the time at which we arrived at state x_n, and does not depend on how we got to state x_n in the first place. This is an important property of Markov chains.

The Markov Property

For all times n, for all states $x_n, x_{n+1} \in \mathcal{S}$, for all possible sequences of states $x_0, ..., x_{n+1}$,

$$\begin{aligned}\mathbf{P}(X_{n+1} = x_{n+1} | X_n = x_n, ..., X_0 = x_0) &= \mathbf{P}(X_{n+1} = x_{n+1} | X_n = x_n) \\ &= P(x_n, x_{n+1}).\end{aligned}$$

All elements of a Markov chain can be encoded in a **transition probability matrix** (which is **row stochastic**, i.e. has row sums of 1) and **graph**, drawn from a a 2-dimensional array with the element in row i, column j corresponding to p_{ij}.



$\iff \begin{pmatrix} a & b \\ c & d \end{pmatrix}$

6.1.1 The Probability of a Path

Given a chain, we can compute the probability of a sequence of future states, similar to the multiplication rule in probability tree models. In particular we have

$$\mathbf{P}(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = \pi_0 \prod_{k=1}^{n} p_{i_{k-1} i_k}.$$

Think of this as following the path from node to node in the transition graph and multiplying the probabilities at each step.

6.1.2 k-Step Transition Probabilities

We often want to find the probability law of the state in the *future*.

$$P_k(x, y) := \mathbf{P}(X_k = y | X_0 = x) = P^k(x, y).$$

Put another way, this is the probability that after k steps we will reach state y starting from state x, which is the (x, y) entry of the kth power of P We calculate this (the **k-step transition probabilities**) using the following recursion:

Chapman-Kolmogorov Equations

The k-step transition probabilities can be generated from

$$\begin{aligned} P_k(x, y)) &= \mathbf{P}(X_k = y | X_0 = x) \\ &= \sum_{x_1, ... x_{k-1} \in \mathcal{X}} \mathbf{P}(X_k = y, X_{k-1} = x_{k-1}, ..., X_1 = x_1 | X_0 = x) \\ &= \sum_{x_1, ... x_{k-1} \in \mathcal{X}} P(x, x_1) P(x_1, x_2) \dots P(x_{k-2}, x_{k-1}) P(x_{k-1}, y) \\ &= P^k(x, y) \end{aligned}$$

6.2 Classification of States

The probabilities associated with each state give us different information about each state. For instance, some states, once visited, are certain to be visited again, while others may never be visited again. Here we draw a relationship between the states of a Markov chain and the long-term frequency with which they are visited.

For each $x \in \mathcal{X}$, let us define a random variable $T_x^+ := \min n \in \mathbb{N} : X_n = x$, the **hitting time** of state x. Additionally define $\mathbf{P}_x$ and $\mathbf{E}_x$ to indicate that the chain begins at state x, i.e.

$$\mathbf{P}_x(\cdot) := \mathbf{P}(\cdot | X_0 = x)$$

$$\mathbf{E}_x(\cdot) := \mathbf{E}[\cdot | X_0 = x].$$

Then for $x, y \in \mathcal{X}$, let $\rho_{x,y} = \mathbf{P}_x(T_y^+ < \infty)$, the probability we are guaranteed to see state state y starting from state x. Let $\rho_x = \rho_{x,x}$. Clearly, if $\rho_x = 1$ (we are guaranteed to see state x from state x), we will visit x infinitely; this state is **recurrent**. Likewise, if $\rho_x < 1$, we are guaranteed to stop seeing x from x after a countably infinite number of steps; this state is **transient.** Classifying a state as transient or recurrent only depends on whether or not arrows *exist* in the transition probability graph; they do not depend on what the actual probabilities are.

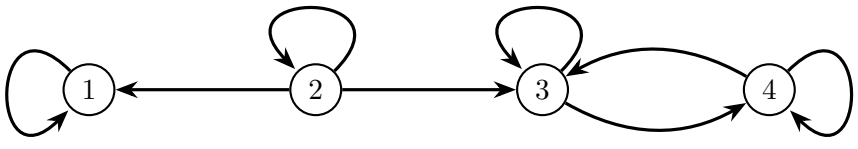
$\mathbf{P}_x$-a.s.

Let N_x denote the total number of visits to state x, that is, $N_x := \sum_{n \in \mathbb{N}} \mathbf{1}\{X_n = x\}$. If x is recurrent, then $N_x = \infty \mathbf{P}_x - a.s.$, so in particular $\mathbf{E}_x[N_x] = \infty$. If x is transient, then $\mathbf{E}_x[N_x] < \infty$; in fact,

$$\mathbf{E}_x[N_x] = \frac{\rho_x}{1 - \rho_x} < \infty.$$

In particular, $N_x < \infty \mathbf{P}_x - a.s.$

Let $A(x)$ be the set of all states accessible (through a series of state transitions) from x. If x is recurrent, $A(x)$ forms a **recurrent class** or **communicating class**, meaning all states in $A(x)$ are accessible from each other and no state outside $A(x)$ is accessible from them. In graph theory, we call this a **strongly-connected component**. If a Markov chain is **irreducible** if it has only a single communicating class.



Recurrent Transient Recurrent Recurrent

At least one recurrent state must be accessible from any transient state. We can decompose any Markov chain into one or more recurrent classes plus possibly a few transient states. Decompositions let us visualize the evolutions of states. Once the state enters in a class of recurrent states, it stays within that class, and since all states in the class are accessible from one another, all states in the class will be visited an infinite number of times. If the initial state is transient, then the state trajectory will contain an initial portion of transient states and a final portion consisting of recurrent states within the same class.

Periodicity

Consider a communicating class $\mathcal{R}$. This class is **periodic** if its states can be grouped in $d > 1$ disjoint subsets $\mathcal{S}_1, ..., \mathcal{S}_d$ so that all transitions from $\mathcal{S}_k$ lead to $\mathcal{S}_{k+1}$.

The class is **aperiodic** if and only if there exists a time k such that $P_k(x, y) > 0, \ \forall \ x, y \ \in \mathcal{R}$.

6.2.1 Positive Recurrence and Null Recurrence

A sequence of random variables is **stationary** if for all $k, n \in \mathbb{Z}_+$, and all events $A_i, ..., A_n$, then

$$\mathbf{P}(X_1 \in A_1, ..., X_n \in A_n) = \mathbf{P}(X_{k+1} \in A_1, ..., X_{k+n} \in A_n).$$

What does this mean? All we're saying here is that the distribution of $X_1, ..., X_n$ stays the same as the joint distribution if we shift the time index by k to $X_{k+1}, ..., X_{k+n}$. Many stochastic processes **converge** to this notion of stationarity.

Stationarity

Suppose that a Markov chain is irreducible with a stationary distribution π. Then, for teach $x \in \mathcal{X}$,

$$\pi(x) = \frac{1}{\mathbf{E}_x[T_x^+]}.$$

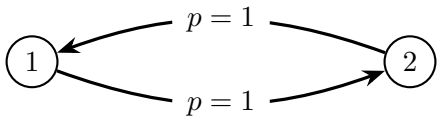
A Markov chain is, essentially, a way to introduce dependency into the idea of i.i.d. random variables. As such, our best shot at understanding them is by analyzing the i.i.d. structure hidden within the chain. If we're at state x for the 1st time, there is no difference in the distribution than if we were at state x for the kth time; let $T_x(k)$ be the kth time we hit state x. Then for all k, we can split the Markov chain into $X_{T_x(k)}, ..., X_{T_x(k+1}-1$ and treat each of these as i.i.d random variables, giving rise to the above expression.

Positive and Null Recurrence

State x is is **positive recurrent** if x is recurrent and $\mathbf{E}_x[T_x^+] < \infty$ (the expected time to return to x is finite).
State x is **null recurrent** if x is recurrent and $\mathbf{E}_x[T_x^+] = \infty$ (the expected time to return to x is infinite).

6.3 Steady-State Behavior

We'll examine both the long and short-term behavior of Markov chains. We begin by looking at what the k-step transition probabilities $P_k(x, y)$ are when k is very large. If a Markov chain has multiple classes of recurrent states, the limiting value of $P_k(x, y)$ must depend on our initial state. We focus selectively on chains that have only one recurrent class plus transient states. Even when a chain only has a single recurrent class, $P_k(x, y)$ may *not* converge! Examine the following Markov chain:



$$\iff P_k(x,y) = \begin{cases} 1, & k \text{ even} \\ 0, & k \text{ odd} \end{cases}$$

This is what happens when the recurrent class is periodic. Ignoring the two cases mentioned previously (multiple classes and periodicity), we can create the following theorem:

Steady-State Convergence Theorem

Consider a Markov chain with a single aperiodic recurrent class. Then each state x has an associated **steady-state probability** $\pi(x)$:

1. $\forall\, x$,
$$\lim_{n\to\infty} P_k(y,x) = \pi(x), \quad \forall i$$
2. $\pi(x)$ are the unique solutions of the following:
$$\pi(x) = \sum_{y\in\mathcal{X}} \pi(y)P(y,x) \quad j = 1,...,m,$$
$$1 = \sum_{x\in\mathcal{X}} \pi(x)$$

The steady-states sum to 1 and form a distribution, called the **stationary distribution**. This is because for all n, $\mathbf{P}(X_n = x) = \pi(x)$. This means that if some initial state is chosen according to the stationary distribution, the state at *any* future time will have the same distribution. The equations

$$\sum_{y\in\mathcal{X}} \pi(y)P(y,x)$$

are the **balance equations**. Together with the **normalization equation**

$$1 = \sum_{x\in\mathcal{X}} \pi(x).$$

An irreducible positive reccurent Markov chain has a unique stationary distribution; likewise, an irreducible Markov chain is positive recurrent if and only if a stationary distribution exists. What does it mean if the stationary distribution *doesn't* exist? If μ is the solution to the balance equations, then a stationary distribution does not exist if it is impossible to normalize μ, i.e. $\sum \mu(x) = 0$ or $\sum \mu(x) = \infty$. Otherwise, if $\sum \mu(x) = c$, then $\pi := c^{-1}\mu$. The fraction of time we spend in state x converges almost surely to $\pi(x)$ as the number of steps we take increases infinitely.

Here we also have the **first-step equations**:

First-Step Equations

For each state x:

$$\pi(x) = 1 + \sum_{y \in \mathcal{X}} P(x,y)\pi(y)$$

The expected time to return to a state x:

$$\mathbf{E}_x[T_x^+] = 1 + \sum_{y \in \mathcal{X}} P(x,y)\mathbf{E}_y[T_x^+]$$

6.4 Reversibility of Markov Chains

When does a Markov chain look the same whether we run it forwards in time or backwards in time? Fix a positive integer N and define $Y_n = X_{N-n}$; this is the **reversed chain**. If the original chain is irreducible and the chain is started from stationary distribution π, then the reversed chain is also irreducible with transition probabilities $\hat{P}(x,y) = \pi(y)P(x,y)/\pi(x)$. The stationary distribution for the reversed chain is also π. The reversed chain looks the same as the original chain if

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

The above is known as the **detailed balance equations**. The condition for stationarity is

$$\pi(y) = \sum_{x \in \mathcal{X}} \pi(x)P(x,y)$$

This is a stronger condition than the global balance equations; the global equations say the probability mass entering a state equals the probability mass exiting the state. The detailed equations express a local condition where the mass along each edge is balanced.

7 The Bernoulli and Poisson Processes

What is a **process**? A stochastic process is a model of a probabilistic experiment over time. Each value in the process is a random variable, so really a process is just a sequence of random variables. In processes, we focus on **dependencies** in the sequence of values, long-term **averages** involving the values, and the likelihood or frequencies of **boundary events**. Here we discuss *arrival processes*, where we care about the times between arrivals. If the arrivals occur in discrete time, we model this with the *Bernoulli process*, and if the arrivals occur in continuous time we model them with the *Poisson process*.

7.1 The Bernoulli Process

The Bernoulli process can be visualized as a sequence of independent coin tosses with probability p. Take a sequence of random variables, where success at the ith variable occurs with probability p. Recall the properties of the binomial and geometric distributions from earlier in this course. Additionally recall the **memoryless** property of the geometric distribution; the future of the process does not depend on how much time has already passed.

7.1.1 Interarrival Times

One of the reasons we study processes is to determine the time of the kth success, or **arrival**, Y_k. Because of this, we can express the events of the process as a sequence of geometric random variables $T_1, ...$ with parameter p, standing for the time in between arrivals. Then the sequence is $T_1, T_1 + T_2, T_1 + T_2 + T_3...$ We then have the following set of observations:

Properties of the kth Arrival Time

The kth arrival time is the sum of the first k interarrival times:

$$Y_k = T_1 + ... + T_k,$$

where T_i are independent geometric random variables with common parameter p.

$$\mathbf{E}[Y_k] \sum \mathbf{E}[T_i] = \frac{k}{p},$$

$$\mathbf{Var}(Y_k) = \sum \mathbf{Var}(T_1) = \frac{k(1-p)}{p^2}.$$

The PMF of Y_k is then

$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \quad t = k, k+1, ...,$$

the **Pascal PMF of order k.**

7.1.2 Splitting and Merging of Bernoulli Processes

Consider **splitting** a process: When there is an arrival, we keep it with probability q or discard it with probability $1-q$. Then the probability of there being an arrival we keep is pq and the probability of there being an arrival we discard is $p(1-q)$. We are essentially sorting our standard arrival process into two processes.

We can also **merge** two processes, recording an arrival in the merged process if there is an arrival in one of the two processes, with probability $p+q-pq$ $1-(1-p)(1-q)$.

7.2 The Poisson Process

The continuous-time variant of the Bernoulli process is the **Poisson process**. An arrival process is a Poisson process with rate λ if if the the probability of k arrivals in τ time, $P(k, \tau)$ is the same for all intervals of length τ. The number of arrivals during an interval is independent of the number of arrivals outside that interval. Additionally, it satisfies the following

small-interval probabilities:

$$P(0,\tau) = 1 - \lambda\tau + o(\tau),$$
$$P(1,\tau) = \lambda\tau + o_1(\tau),$$
$$P(k,\tau) = o_k(\tau),$$

Here

$$\lim_{\tau \to 0} \frac{o(\tau)}{\tau} = 0, \quad \lim_{\tau \to 0} \frac{o_k(\tau)}{\tau} = 0.$$

Think of $o(\tau)$ as the $O(\tau^2)$ terms in the Taylor expansion of $P(k,\tau)$. The probability of a single arrival is roughly $\lambda\tau$ with a negligible term. The probability of 0 arrivals is roughly $1 - \lambda\tau$.

Random Variables Associated with the Poisson Process

The number of arrivals N_τ in a Poisson process with parameter λ over an interval τ:

$$p_{N_\tau}(k) = P(k,\tau) = e^{-\lambda\tau}\frac{\lambda^k\tau^k}{k!}$$
$$\mathbf{E}[N_\tau] = \lambda\tau \quad \mathbf{Var}(N_\tau) = \lambda\tau.$$

The time until the first arrival T:

$$f_T(t) = \lambda e^{-\lambda t}, \quad \mathbf{E}[T] = \frac{1}{\lambda}, \quad \mathbf{Var}(T) = \frac{1}{\lambda^2}.$$

The Poisson process is also independent; i.e. two non-overlapping time sets can be considered independent processes, and the distribution of interarrival times is memoryless. Just as the first arrival after time t for the Bernoulli process is distributed geometrically, the first arrival after time t for the Poisson process is distributed exponentially.

The kth Arrival Time

As with the Bernoulli process, we can model the Poisson process as a sequence of independent exponential random variables with parameter λ, and record arrivals at times T_1, $T_1 + T_2$, ... Then the kth arrival time is the sum of the first k interarrival times:

$$Y_k = \sum_{i=1}^{k} T_i,$$

where T_i are independent exponential random variables with common parameter λ.

The mean and variance of Y_k are

$$\mathbf{E}[Y_k] = \sum_{i=1}^{k} \mathbf{E}[T_i] = \frac{k}{\lambda},$$

$$\mathbf{Var}(Y_k) = \sum_{i=1}^{k} \mathbf{Var}(T_i) = \frac{k}{\lambda^2}.$$

The PDF of Y_k is

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0,$$

the **Erlang PDF of order** k.

7.2.1 Poisson Splitting and Merging

Just as with the Bernoulli case, we can split a Poisson process into two Poisson processes (with rate λp) if we split with probability p (keep with probability p and discard with probability $1 - p$). Alternatively, we can merge two Poisson processes into a single process with merged rate $\lambda_1 + \lambda_2$; the probability that an arrival occurs in the first process is $\frac{\lambda_1}{\lambda_1+\lambda_2}$, and the probability that an arrival occurs in the other is $\frac{\lambda_2}{\lambda_1+\lambda_2}$. In fact, the sum of n Poisson processes is a Poisson process with parameter $\sum \lambda_i$.

7.2.2 Sums of Random Variables

Sums of Random Numbers of Random Variables

Let N and $X_1, X_2, ...$ be random variables, and let $Y = \sum_{i=1}^{N} X_i$.

1. If X_i is Bernoulli with parameter p and N is binomial with parameters m, q, then Y is binomial with parameters m and pq.
2. If X_i is Bernoulli with parameter p, and N is Poisson with parameter λ, then Y is Poisson with parameter λp.
3. If X_i is geometric with parameter p, and N is geometric with parameter q, then Y is geometric with parameter pq.
4. If X_i is exponential with parameter λ, and N is geometric with parameter q, then Y is exponential with parameter λq.

7.2.3 The Random Incidence Paradox

Here we examine an interesting property of the Poisson process. Assume we take some random time t^* during the process. We call this time a "random incidence," but be aware that this isn't a random variable, just an arbitrary time. Assume that t^* happens after the process has been running for a long time, so there has been an arrival sometime in the past. Here we consider the length L interval that contains t^*. We could argue that L is distributed exponentially, like a typical interarrival period. However this is not true – L is distributed according to an Erlang distribution with parameter 2. This is because the time from t^* to the next arrival is a Poisson process with parameter λ; likewise, the time from t^* to the previous arrival is *also* a Poisson process with parameter λ. Then L is distributed according the Erlang PDF of order 2 by our previous assertions.

8 Erdös-Rényi Random Graphs

Given $n \in \mathbb{Z}_+$, and some probability value $p \in [0,1]$, the graph $\mathcal{G}(n,p)$ is defined as an undirected graph on n vertices such that each of the $\binom{n}{2}$ edges of the graph is present (independently) with probability p. This means if $p = 0$, $\mathcal{G}$ is an empty graph and if $p = 1$ then it is a fully connected graph. We typically define p as a function of n, $p(n)$, and are especially interested in what happens as $n \to \infty$. Clearly, then, $\mathcal{G}(n,p)$ is a *distribution* over the set of graphs on n vertices! The "PDF" of $G \sim \mathcal{G}(n,p)$ can be easily calculated:

$$\mathbf{P}(G = G_0) = p^m (1-p)^{\binom{n}{2}-m}.$$

8.1 Sharp Threshold for Connectivity

Sharp Threshold

Let

$$p(n) = \lambda \frac{\ln n}{n}$$

for $\lambda > 0$. Then if $\lambda < 1$, then $\mathbf{P}(\mathcal{G}(n, p(n) \text{ is connected}) \to 0$; if $\lambda > 1$, then $\mathbf{P}(\mathcal{G}(n, p(n) \text{ is connected}) \to 1$. This is called a **sharp threshold** since a slight deviation in λ around 1 can drastically change the behavior of the limit.

9 Bayesian Statistical Inference

Here we deviate from our pure mathematical subject of **probability** (based on the principles from the first chapter) and move towards **statistics**. In probability theory, we assume some underlying model and attempt to predict future events using this information. In statistics, we observe some event, and attempt to figure out the process which could have led to those observations. If we are drawing red and blue balls from a bag, *probability* gives us information about what ball we'll draw next given that we know how many of each are in the bag. Statistics lets us *infer* how many of each are in the bag based on the balls we draw.

There are two schools of thought for looking at statistics – the **Bayesian** and the **classical** (or **frequentist**). In the former, unknown quantities are treated as random variables with known distributions; in the latter, unknown values are treated as deterministic quantities that just happen to have unknown values.

There are two forms of inference. In **model inference**, we take our observations and try to construct a model to explain the process behind those observations. In **variable inference**, we estimate the value of some unknown variable through some noisy observations.

We can roughly classify statistical inference into two buckets. In **parameter estimation**, we have a model with some unknown parameter θ, which we try to estimate using either a Bayesian or classical approach. In m**-ary hypothesis testing**, we take m possible hypotheses and use our data to determine which is true.

9.1 Bayesian Inference

The Bayesian method operates by defining a random variable Θ that represents our model, and a probability distribution $p_\Theta(\theta)$ (the **prior**). We can then make a set of observations $\mathbf{x}$ and derive a **posterior** distribution $p_{\theta|X}(\theta|x)$. There are three key principal Bayesian inference methods we explore here: **Maximum a posteriori (MAP)**, where we select the hypothesis/parameter with the highest posterior probability; **Least mean squares (LMS)**, where we select the hypothesis/parameter with the minimum mean squared error between the parameter and its estimate; and **Linear least mean squares (LLMS)**, where we select an estimator as a linear function

of the data that minimizes the mean squared error between the parameter and its estimate.

We assume a **prior** p_Θ or f_Θ for thee unknown Θ, assume a model $p_{X|\Theta}$ or $f_{X|\Theta}$ for the vector of **data** or observations X, and then use Bayes' rule (different versions depending on whether X and Θ are discrete or continuous) to determine the posterior $p_{\Theta|X}$ or $f_{\Theta|X}$.

4 Versions of Bayes' Rule

Let

$$\phi = \begin{cases} p_X(x) & X \text{ discrete,} \\ f_X(x) & X \text{ continuous.} \end{cases}$$

Then, Bayes rule is as follows:

$$\phi_{\Theta|X}(\theta|x) = \begin{cases} \dfrac{\phi_\Theta(\theta)\phi_{X|\Theta}(x|\theta)}{\displaystyle\sum_{\theta'} \phi_\Theta(\theta')\phi_{X|\Theta}(x|\theta')} & \Theta \text{ discrete,} \\ \dfrac{\phi_\Theta(\theta)\phi_{X|\Theta}(x|\theta)}{\displaystyle\int \phi_\Theta(\theta')\phi_{X|\Theta}(x|\theta')d\theta'} & \Theta \text{ continuous.} \end{cases}$$

9.2 Point Estimation, Hypothesis Testing, and the MAP Rule

Now that we've introduced the framework for Bayesian inference, we can introduce a method that we can then apply to estimation and hypothesis testing problems. We're given a value x of an observation, and we select a value of θ (normally denoted $\hat{\theta}$) that maximizes the posterior $p_{\Theta|X}(\theta|x)$ (or $f_{\Theta|X}(\theta|x)$ for continuous Θ). This is known as the **Maximum a Posteriori** rule (or **MAP**).

Maximum a Posteriori (MAP)

Given an observation value x, we select the value $\hat{\theta}$ that maximizes θ over the posterior $p_{\Theta|X}(\theta|x)$ or $f_{\Theta|X}(\theta|x)$ (depending on whether Θ is discrete or continuous).

$$\hat{\theta} = \arg\max_{\theta} p_{\Theta|X}(\theta|x) \iff \hat{\theta} = \arg\max_{\theta} f_{\Theta|X}(\theta|x).$$

By Bayes' rule, the form of the posterior distribution is a fraction, where the numerator is dependent on both x and θ. However, it is important to note that the denominator is the same for all values of θ (it is dependent on θ', which takes on all values in Θ). Therefore, we can exclude the denominator entirely, and are left with the following (using the definition of ϕ from the previous box):

$$\hat{\theta} = \arg\max_{\theta} \phi_{\Theta}(\theta)\phi_{X|\Theta}(x|\theta).$$

If Θ takes on a finite number of values, then MAP essentially minimizes the probability of selecting the wrong hypothesis.

9.2.1 Point Estimation

Point Estimates

An **estimator** is a random variable $\hat{\Theta}$ that is a function of our observations, i.e.

$$\hat{\Theta} = g(X).$$

An **estimate** is a value $\hat{\theta}$ of an estimator, which is determined by an actual observation x.

The **MAP estimator** selects $\hat{\theta}$ to be the θ that maximizes the posterior distribution over all θ given x.

The **conditional expectation estimator** (called the **LMS estimator**) selects the $\hat{\theta}$ that is $\mathbf{E}[\Theta|X = x]$. This is called the LMS estimator because it has the property that it minimizes the **mean squared error** over all estimators.

9.2.2 Hypothesis Testing

In a hypothesis testing problem, θ takes on one of a small set of values. The ith hypothesis, H_i, is the event $\Theta = \theta_i$. Once we've observed x of X, we can calculate the probability of each hypothesis given the outcome. We can then select the hypothesis that maximizes this posterior probability (the MAP rule). The probability of a correct decision is then

$$\mathbf{P}(\Theta = g_{MAP}(x)|X = x).$$

9.3 Bayesian Least Mean Squares

Here we discuss the second estimator we mentioned above, the **Least Mean Squared** or **LMS** estimator. Consider a simpler problem, where we don't have any observations x. Then the **mean squared error** $\mathbf{E}[(\Theta - \hat{\theta})^2]$ is minimized when $\hat{\theta} = \mathbf{E}[\Theta]$. We use the MSE because if we were to just use $\Theta - \hat{\theta}$, we have a random variable that cannot be minimized in θ. In this case:

$$\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta])^2\right] \leq \mathbf{E}[(\Theta - \hat{\theta})^2].$$

Now we consider our observation x of X; in this case, $\mathbf{E}[(\Theta - \hat{\theta})^2|X = x]$ is minimized at $\hat{\theta} = \mathbf{E}[\Theta|X = x]$:

$$\mathbf{E}[(\Theta - \mathbf{E}[\Theta]|X = x])^2|X = x] \leq \mathbf{E}[(\Theta - \hat{\theta})^2|X = x].$$

Finally, for all esimators $g(X)$, the MSE is minimized when $g(X) = \mathbf{E}[\Theta|X]$:

$$\mathbf{E}[(\Theta - \mathbf{E}[\Theta|X])^2] \leq \mathbf{E}[(\Theta - g(X))^2].$$

Properties of the Estimation Error

If we have the LMS estimator $\hat{\Theta} = \mathbf{E}[\Theta|X]$, then the **estimation error** is $\tilde{\Theta} = \hat{\Theta} - \theta$.

1. $\tilde{\Theta}$ is unbiased, i.e. it has 0 mean both conditionally and unconditionally ($\mathbf{E}[\tilde{\Theta}] = \mathbf{E}_\theta[\hat{\Theta}] - \theta = 0$, $\mathbf{E}[\tilde{\Theta}|X] = 0$). The value $\mathbf{E}_\theta[\tilde{\Theta}]$ is the **bias**, denoted $b_\theta(\hat{\Theta})$.
2. $\tilde{\Theta}$ is uncorrelated with $\hat{\Theta}$, i.e. $\mathbf{cov}(\hat{\Theta}, \tilde{\Theta}) = 0$.
3. $\mathbf{Var}(\Theta) = \mathbf{Var}(\hat{\Theta}) + \mathbf{Var}(\tilde{\Theta})$.

10 Classical Parameter Estimation

In a classical setting (as opposed to the Bayesian one we previously introduced), we view the unknown parameter θ as *deterministic* rather than random. X, the observation, is then random, and instead of dealing with a single probabilistic model we deal with several possible **candidate models**.

10.1 Classical Parameter Estimation

Recall the definitions of an **estimator** and **estimate** from the previous section. The definitions of **estimation error** and **bias** are the same. It is interesting to note that the **mean squared error** can also be written as

$$\mathbf{E}[\tilde{\Theta}^2] = b_\theta^2(\hat{\Theta}) + \mathbf{Var}(\hat{\Theta}).$$

This leads to an interesting problem in machine learning known as the **bias-variance tradeoff**, where reducing the bias term increases the variance term and vice versa.

10.1.1 Maximum Likelihood Estimation (MLE)

In classical parameter estimation, we can describe a **random vector** of observations X by a join PMF $p_X(x;\theta)$, dependent on a deterministic parameter θ. Then a **maximum likelihood estimate** is a value of $\hat{\theta}$ that maximizes $p_X(x = (x_1, ...x_n); \theta)$ over all θ. Thus,

$$\hat{\theta} = \arg\max_\theta p_X(x_1, ..., x_n; \theta).$$

Note that for a continuous X, we replace the PDF $p_X(x;\theta)$ with the PMF $f_X(x;\theta)$; we call this the **likelihood function**. If each observation is independent (which we usually assume), we can simplify this to

$$\hat{\theta} = \arg\max_\theta \prod p_{X_i}(x_i;\theta).$$

Since the logarithm is increasing on its domain, we can equivalently maximize the **log likelihood**, i.e.

$$\hat{\theta} = \arg\max_\theta \sum \log p_{X_i}(x_i;\theta).$$

Note the difference between the term *likelihood* and *probability*. We are not looking for the probability of θ, since we know that θ is a deterministic (but

unknown) quantity. Instead, we're asking "what would θ need to be to maximize the chance that we saw the observations we did?"

Note our maximization of $p_X(x;\theta)$ here, and how in our previous MAP formulation we maximized $p_\Theta(\theta)p_{X|\Theta}(x|\theta)$. In this way, we can consider MLE the same as MAP with a **uniform prior**, meaning that the prior for θ is the same for all θ.

10.2 Hypothesis Testing

Recall the issue of hypothesis testing we brought up in the previous section. However, now we assume that we have no prior information. We have two hypothesis we are deciding between – H_0, the **null hypothesis**, and H_1, the **alternative hypothesis**. More precisely, we say we have Θ_1 and Θ_2, and based on x, we want to know which is more likely; $\theta \in \Theta_0$ or $\theta \in \Theta_2$ (note the similarity to the parameter estimation problem above!). If $\theta \in \Theta_0$, then H_0 is correct; otherwise if $\theta \in \Theta_1$ then H_1 is correct. **For this class, we are mainly focused on frequentist hypothesis testing (i.e. H_0 being true is unknown but not random). We introduced the Bayesian idea above briefly, but it is not a focus of this class.**

Let's be more precise about what accepting a hypothesis means. We can split our total set of possible observations X into two categories: A and A^c. We call A the **acceptance region** of observations, i.e. we accept $H_0 \iff x \in A$.

Acceptance Region Examples

Some examples of acceptance regions:

- Reject H_0 if $x > t$
- Reject H_0 if $x = t$
- Reject H_0 with probability γ if $x > t$

As always however, we have to take into account **error**. If we reject the null hypothesis when the null hypothesis is actually correct, we have a **Type-I Error** or **probability of false alarm (PFA)**. Formally, let $\mathbf{P}_{H_i}(B)$ denote the probability of an event B occurring when H_i is true. Then the probability of type-I error (the **significance level**) is

$$\alpha(A) = \mathbf{P}_{H_0}(H_1) = \mathbf{P}_{H_0}(x \notin A).$$

Analogously, we have **type-II error**, i.e. the probability that we accept the null hypothesis when in reality the alternative is true. Formally:

$$\beta(A) := \mathbf{P}_{H_1}(H_0) = \mathbf{P}_{H_1}(x \in A).$$

Neyman-Pearson Hypothesis Testing

Define the **likelihood** to be:

$$L(x) = \frac{\mathbf{P}_{H_1}(x)}{\mathbf{P}_{H_0}(x)}, \quad L(x) = \frac{f_{H_1}(x)}{f_{H_0}(x)}.$$

Then the **Neyman-Pearson Test** states that, for some threshhold c:

1. Accept H_0 if $L(x) < c$
2. Reject H_0 with probability γ if $L(x) = c$
3. Reject H_0 if $L(x) > c$

Note that this is optimal when the following are true:

$$\mathbf{P}_{H_0}(L(x) > c) + \gamma \mathbf{P}_{H_0}(L(x) = c) = \alpha_0$$

$$\mathbf{P}_{H_1}(L(x) < c) + (1-\gamma)\mathbf{P}_{H_1}(L(x) = c) = \beta_0.$$

If you've taken EE127 or an equivalent optimization course, you may notice that this is a simple optimization problem:

$$\max_c 1 - \mathbf{P}_{H_1}(L(x) < c) - (1-\gamma)\mathbf{P}_{H_1}(L(x) = c)$$
$$s.t. \mathbf{P}_{H_0}(L(x) > c) + \gamma \mathbf{P}_{H_0}(L(x) = c) = z$$

This is all very complicated and can be difficult to compute. In this class, there is often a relationship $L(x) \iff B$ for some B which is much simpler to understand; for example, we'll have $L(x)$ be monotonic in x, meaning $L(x) > c \iff x > t$ or $x < t$, which makes all of the above much easier.

11 Hilbert Spaces, Estimation, and Kalman Filtering

As we conclude this course with estimation, we draw a parallel between spaces of random variables and vector spaces. Note that this portion is fairly heavy in linear algebra on the level of Math110. Before we begin, let's introduce once particular set (with X random):

$$\mathcal{H}\{X : X \in \mathbb{R} s.t.\ \mathbf{E}[X^2] < \infty\}.$$

11.1 A Brief Review of Linear Algebra

Recall that a **vector space** $\mathcal{V}$ is a collection of objects (**vectors**) including the **zero vector** on which the operations of **vector addition** and **scalar multiplication** are defined. Vector addition is commutative, associative, and satisfies the identity property with the zero vector. Scalar multiplication is distributive, commutative, and associative, and satisfies the identity property with the **1** vector.

For a set $\mathcal{S} \subseteq \mathcal{V}$, $\mathbf{span}(\mathcal{S})$ is the set of vectors achievable by only using scalar multiplication and vector addition (using vectors in $\mathcal{S}$), i.e.

$$\mathbf{span}(\mathcal{S}) = \left\{ \sum_{i=1}^{m} c_i \mathbf{v}_i, m \in \mathbb{N}, \mathbf{v}_i \in \mathcal{S}, c_i \in \mathbb{R} \right\}.$$

In other words, every element in the span of $\mathcal{S}$ is a **linear combination** of the vectors in $\mathcal{S}$. Additionally recognize that $\mathbf{span}(\mathcal{S})$ must be a vector space. Whenever we have $\mathcal{U} \subseteq \mathcal{V}$ and $\mathcal{U}$ is a vector space, we call $\mathcal{U}$ a **subspace**.

Note that if $\mathbf{v} \in \mathcal{S}$ and $\mathbf{v}$ is a linear combination of other vectors in $\mathcal{S}$, then $\mathbf{span}(\mathcal{S}) = \mathbf{span}(\mathcal{S} \backslash \{\mathbf{v}\})$. We call $\mathbf{v}$ a **redundant** vector, and if $\mathcal{S}$ contains no redundant vectors, we call $\mathcal{S}$ **linearly independent**. If $\mathcal{S}$ is linearly independent and $\mathbf{span}\mathcal{S} = \mathcal{V}$ (i.e. $\mathcal{S}$ is a **generating set** for $\mathcal{V}$) then we call $\mathcal{S}$ a **basis** for $\mathcal{V}$. Every vector space has a basis, and all bases for the same vector space have the same cardinality or size, known as the **dimension** of $\mathcal{V}$ and denoted $\mathbf{dim}\mathcal{V}$.

11.2 Inner Product Spaces and Hilbert Spaces

An inner product $\langle \cdot, \cdot \rangle$ maps $\mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}^{\geq}$. The inner product is **symmetric** ($\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$), **linear** ($\langle \mathbf{u} + c\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + c\langle \mathbf{v}, \mathbf{w} \rangle$), and **positive**

definite ($\langle \mathbf{u}, \mathbf{v} \rangle > 0,\ \mathbf{u} \neq 0$). Then we call $\mathcal{V}$ along with the map $\langle \cdot, \cdot \rangle$ a **real inner product space**. The inner product gives us two concepts: a **norm** and **angle**:

$$\| \cdot \| : \mathcal{V} \mapsto \mathbb{R}^{\geq}, \ \ \|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}, \quad \langle \mathbf{u}, \mathbf{v} \rangle = \|\mathbf{u}\| \|\mathbf{v}\| \cos\theta.$$

We are most interested in when $\theta = 0$, i.e. $\langle \mathbf{u}, \mathbf{v} \rangle = 0$; we then call $\mathbf{u}$ and $\mathbf{v}$ **orthogonal**.

$\mathcal{H}$ is a **Hilbert space** if it is a **real inner product space** that is **complete**. We don't need to worry about what completeness means for this course, but I'm putting the full definition here if you're interested. Additionally, we behave as if $\mathcal{H}$ is finite-dimensional (even though it usually isn't).

Formal introduction (not in scope): A **Hilbert space** is a **vector space** that is **complete** with respect to an **inner product** defined on that space (a metric space M is complete if every Cauchy sequence in M converges in M). It is an instance of a **Banach space** (defined by any convex set) wherein the **norm** is defined via the **inner product** (ellipse).

11.3 Projection

We want to estimate a random variable $Y \in \mathcal{H}$. However, we only know some X that is correlated with Y. We want to find a "best guess" function of X that estimates Y. This problem because much easier if we restrict ourselves to affine functions of the form $a + bX$. Then our "best guess" is the point $\mathbf{x} \in \mathcal{U}$, $\mathcal{U} = \mathbf{span}\{1, X\} \subseteq \mathcal{H}$ that is *closest* to Y. We find this by finding the shortest **orthogonal distance** to Y. Formally, the shortest orthogonal distance is

$$\mathcal{P} : \mathcal{V} \mapsto \mathcal{U} \quad \mathcal{P}\mathbf{y} := \arg\min_{\mathbf{x} \in \mathcal{U}} \|\mathbf{y} - \mathbf{x}\|^2.$$

Note that $\mathcal{P}\mathbf{y}$ is in $\mathcal{U}$ and $\mathbf{y} - \mathbf{P}\mathbf{y}$ is **orthogonal** to $\mathcal{U}$ (equivalently, $\mathbf{y} - \mathbf{P}\mathbf{y} \in \mathcal{U}^{\perp}$).

11.4 Gram-Schmidt Orthonormalization

Since $\mathcal{P}\mathbf{y} \in \mathcal{U}$ and $\mathbf{y} - \mathcal{P}\mathbf{y} \in \mathcal{U}^{\perp}$, $\mathbf{y} - \mathcal{P}\mathbf{y}$ must be orthogonal to every basis vector for $\mathcal{U}$. Since $\mathcal{P}\mathbf{y}$ is linear, $\mathcal{P}\mathbf{y} = \sum c_i \mathbf{v}_i$ for basis vectors $\mathbf{v}_i$. We can then compute $\mathcal{P}\mathbf{y}$ by solving:

$$\left\langle \mathbf{y} - \sum_{i=1}^{n} c_i \mathbf{v}_i, \mathbf{v}_j \right\rangle = 0.$$

What is the basis vectors were **orthornormal**? This makes our job much easier, since we would then be able to get each component of $\mathcal{P}$ by computing $\langle \mathbf{y}, \mathbf{v}_i \rangle$. How do we orthonormalize our basis vectors? We can use the following:

Gram-Schmidt Process

```
u[1] = v[1]/||v[1]||
for j in 1,...,n-1:
    let w[j+1] = v[j+1]
        - sum from i-1 to j of (<v[j+1], u[i]>u[i])
    set u[j+1] = w[j+1]/||w[j+1]||
```

11.5 Linear Least Squares Estimate (LLSE)

Recall our linear formulation $\hat{Y} = a + bX$ above as our estimate for Y. Then, given $X, Y \in \mathcal{H}$, we seek to minimize (over $b, a \in \mathbb{R}$):

$$\mathbf{E}[(Y - \hat{Y})^2] = \mathbf{E}[(Y - a - bX)^2].$$

The solution to this problem is the **linear least squares estimator** or **LLSE**. Notice this similarity to the classical parameter estimation problem from earlier; we're trying to solve

$$Y^* = \arg\min_{\hat{Y} \in \mathcal{U}} \|Y - \hat{Y}\|^2, \ \mathcal{U} = \mathbf{span}\{1, X\}.$$

Some change of basis and algebra magic gives us the following:

LLSE

For $X, Y \in \mathcal{H}$, where X is variable, the LLSE of Y given X is

$$\begin{aligned} L[Y|X] &= \mathbf{E}[Y] + \mathbf{E}\left[Y\left(\frac{X - \mathbf{E}[X]}{\sqrt{\mathbf{Var}X}}\right)\right]\frac{X - \mathbf{E}[X]}{\sqrt{\mathbf{Var}X}} \\ &= \mathbf{E}[Y] + \frac{\mathbf{cov}(X,Y)}{\mathbf{Var}X}(X - \mathbf{E}[X]). \end{aligned}$$

The squared error of the LLSE is then

$$\mathbf{E}[(Y - L[Y|X])^2] = \mathbf{Var}Y - \frac{\mathbf{cov}(X,Y)^2}{\mathbf{Var}X}.$$

11.6 Minimum Mean Square Estimation (MMSE)

Say instead of restricting X to linear functions, we let $\hat{Y}$ be any function of X, say $\phi(X)$. Then finding the best function ϕ to minimize

$$\mathbf{E}[(Y - \phi(X))^2]$$

is known as the **minimum mean squared error estimator (MMSE)**. One important condition is placed on this estimator; $Y - \phi(X)$ must be orthogonal to all other functions of X. The solution to this problem is the conditional expectation of Y given X, or $\mathbf{E}[Y|X]$, such that for all bounded continuous functions ϕ:

$$\mathbf{E}[(Y - \mathbf{E}[Y|X])\phi(X)] = 0.$$

The difference between MMSE and LLSE is that MMSE deals with *all* functions on X, even nonlinear ones (which is really hard to visualize). The optimum ϕ is guaranteed to exist (since the objective function is convex – see more in EE127), but finding it is often hard. For this course, just knowing these basic facts about MMSE will suffice.

11.6.1 Jointly Gaussian Random Variables

X and Y are **jointly Gaussian** if their joint PDFs are **multivariate Gaussian**, or equivalently if each linear combination of X and Y is Gaussian. Generally, $L[Y|X] \neq \mathbf{E}[Y|X]$. However, when X and Y happen to be jointly Gaussian, the MMSE is equivalent to the LLSE, i.e.

MMSE for Jointly Gaussian Random Variables

If X, Y are jointly Gaussian, then

$$\mathbf{E}[Y|X] = L[Y|X] = \mathbf{E}[Y] + \frac{\mathbf{cov}(X,Y)}{\mathbf{Var}X}(X - \mathbf{E}[X]).$$

Other important properties that can be used to calculate the LLSE/MMSE are shown below:

Key Properties

Both the LLSE and MMSE are **unbiased**, meaning

$$\mathbf{E}[X - \mathbf{E}[X]] = \mathbf{E}[\mathbf{E}[Y|X] - \mathbf{E}[Y]] = 0.$$

Additionally, $Y - L[Y|X]$ and X are uncorrelated and also independent (since $Y - L[Y|X]$ and X are jointly Gaussian).

11.7 Kalman Filtering

The Kalman filter is an optimal **state estimation** algorithm (akin to the Linear-Quadratic Regulator from EECS127). Consider a system with a **state** $\mathbf{X}(n)$ and an **output** (n) at time n. We can describe the system as follows:

$$\begin{aligned}\mathbf{X}(n+1) &= \mathbf{A}\mathbf{X}(n) + \mathbf{V}(n)\\ \mathbf{Y}(n) &= \mathbf{C}\mathbf{X}(n) + \mathbf{W}(n)\end{aligned}$$

Here, $\mathbf{X}(0)$, $\mathbf{V}$, and $\mathbf{W}$ are random and zero-mean, where $\mathbf{cov}\mathbf{V} = \Sigma_{\mathbf{V}}$ and $\mathbf{cov}\mathbf{W} = \Sigma_{\mathbf{W}}$. We want to recursively be able to recover $\mathbf{X}$ to the best of our ability based on our noisy observations $\mathbf{Y}$, i.e.

$$\hat{\mathbf{X}}(n) = L[\mathbf{X}(n)|(\mathbf{Y}(i))_{i=0}^{n}].$$

The Kalman Filter

The filter is defined recursively as follows:

$$\begin{aligned}\hat{\mathbf{X}}(n) &= \mathbf{A}\hat{\mathbf{X}}(n-1) + \mathbf{K}_n[\mathbf{Y}(n) - \mathbf{C}\mathbf{A}\hat{\mathbf{X}}(n-1)]\\ \mathbf{K}_n &= \mathbf{S}_n\mathbf{C}^\top[\mathbf{C}\mathbf{S}_n\mathbf{C}^\top + \Sigma_{\mathbf{W}}]^{-1}\\ \mathbf{S}_n &= \mathbf{A}\Sigma_{n-1}\mathbf{A}^\top + \Sigma_{\mathbf{V}} = \mathbf{cov}(\mathbf{X}(n) - \mathbf{A}\hat{\mathbf{X}}(n-1))\\ \Sigma_n &= (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{S}_n = \mathbf{cov}(\mathbf{X}(n) - \hat{\mathbf{X}}(n))\end{aligned}$$

Some key observations: $\mathbf{K}_n$ and Σ_n can be precomputed at time 0 since they do not depend on the observations (note: even though $\hat{\mathbf{X}}$ depends on the observations, the residual $\mathbf{X}(n) - \hat{\mathbf{X}}(n)$ does not). Additionally if $\mathbf{X}(0)$ and the noise variables $\mathbf{V}$ and $\mathbf{W}$ are Gaussian, then the Kalman filter simply computes the MMSE. Additionally, even though this entire box is a computational mess, it can be programmed very easily and it is incredibly simple to solve computationally by a computer.