EECS126 Notes

Lecturers: Kannan Ramchandran and Abhay Parekh
Scribe: Michael Whitmeyer

Jan 2020 - May 2020 (with some examples from Fall 2018)

Contents

1 Lecture 1: Overview and beginning of CS70 Review

1.1 Motivation

1. Uncertainty is all around us!
2. This course is about formalizing how to predict things.
3. Actually has origins in gambling
4. First need to develop **model** (requires understanding of the problem as an experiment), and then need to **solve** (using combinatorics, calculus, common sense, etc). As engineers, we need to do both, but often it is (perhaps unexpectedly) the modelling that is more difficult than the solving.
5. Last but certainly not least (for many of you), foundational for ML/AI.

1.2 Content

Definition 1.1. A **Sample Space** Ω of an experiment is the set of all outcomes of the experiment. The outcomes must be *mutually exclusive* (ME) and *collectively exhaustive* (CE)

Example 1.2. Toss two fair coins. Then we have $\Omega = \{HH, HT, TH, TT\}$. Can check that these outcomes are mutually exclusive and collectively exhaustive.

Definition 1.3. An **Event** is simply an allowable subset of Ω.

Example 1.4. In Ex 1.2 an event would be getting at least 1 Head

Definition 1.5. A **Probability Space** $(\Omega, \mathcal{F}, P)$ is a mathematical construct that allows us to model these "experiments". Here $\mathcal{F}$ denotes the set of all possible events, where each event is a set containing 0 or more base outcomes (for discrete Ω this is simply the power set of Ω). And $P : \mathcal{F} \mapsto [0, 1]$ is a function assigning probabilities to each event.

All of Probability Theory rests on just 3 (2.5?) axioms (Kolmogorov):

1. $\mathbf{Pr}(A) \geq 0$ for all $A \subseteq \Omega$
2. $\mathbf{Pr}(\Omega) = 1$
3. $\mathbf{Pr}(A_1 \cup A_2 \cup ...) = \mathbf{Pr}(A_1) + \mathbf{Pr}(A_2) + \ldots$ for all disjoint $A_1, A_2, \ldots$. This can be finite or we can take $n \to \infty$ and this becomes *countable additivity*.

We immediately have the following fundamental facts:

1. $\mathbf{Pr}(A^c) = 1 - \mathbf{Pr}(A)$
2. $\mathbf{Pr}(A \cup B) = \mathbf{Pr}(A) + \mathbf{Pr}(B) - \mathbf{Pr}(A \cap B)$
3. Union bound: $\mathbf{Pr}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbf{Pr}(A_i)$
4. Inclusion-Exclusion:

$$\mathbf{Pr}\left(\bigcup_{i=1}^{n} A_i\right) = \sum \mathbf{Pr}(A_i) - \sum_{i<j} \mathbf{Pr}(A_i \cap A_j) + \sum_{i<j<k} \mathbf{Pr}(A_i \cap A_j \cap A_k) - \ldots$$

In the discrete setting, just from the axioms, we have that $\mathbf{Pr}(A) = \sum_{\omega \in A} \mathbf{Pr}(\omega)$. If our sample space is *uniform*, then we have that $\mathbf{Pr}(A) = \frac{|A|}{|\Omega|}$

1.3 Conditional Probability

Definition 1.6. In general, we use the notation $\mathbf{Pr}(A|B)$ = the probability that event A has occurred *given* that we know that B has occurred.

Proposition 1.7 (Bayes Rule)**.**

$$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)}$$

"B is the new Ω"

Example 1.8. We roll two six sided die, and observe that the sum of the two die is 11. What is is the probability that the first die was a 6? Here let A = event of a 6 on the first die and B = event of sum being 11.

$$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)} = \frac{\mathbf{Pr}(\{6,5\})}{\mathbf{Pr}(\{6,5\}) + \mathbf{Pr}(\{5,6\})} = \frac{1}{2}$$

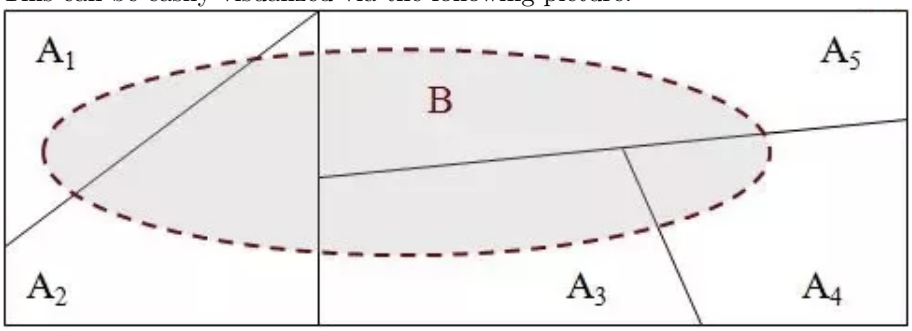
Bayes Rule directly extends to the **Product Rule**, which says that

$$\mathbf{Pr}(A_1 \cap A_2 \cap ... \cap A_n) = \mathbf{Pr}(A_1)\,\mathbf{Pr}(A_2|A_1)\,\mathbf{Pr}(A_3|A_1 \cap A_2) \cdots \mathbf{Pr}(A_n|A_1 \cap ... \cap A_{n-1})$$

We also can develop the **Law of Total Probability**, which says that for mutually exclusive and collectively exhaustive events $A_1, ..., A_n$, we have that

$$\mathbf{Pr}(B) = \mathbf{Pr}(A_1 \cap B) + ... + \mathbf{Pr}(A_n \cap B) = \sum_{i=1}^{n} \mathbf{Pr}(A_i)\,\mathbf{Pr}(B|A_i)$$

This can be easily visualized via the following picture:



In the picture, we partition the sample space indicated by the whole box into the mutually exclusive and collectively exhaustive events A_1, A_2, A_3, A_4, A_5. Conditioned on each of these, there is some probability that B occurs, and so we can find the total probability that B occurs by considering each case separately.

2 Lecture 2: Independence, Bayes Rule, Discrete Random Variables

We will begin with a cool example.

Example 2.1 (Birthday Paradox)**.** Want to estimate $\mathbf{Pr}$(at least two people in a group of size n share the same birthday). First we note that $|\Omega| = k^n = 365^n$. The problem is that this event is a bit complicated. So we will consider the complement: A^c = "no two people share a birthday" Then, since the distributions are uniform, we have

$$\begin{aligned}\mathbf{Pr}(A^c) = \frac{|A^c|}{|\Omega|} &= \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n} \\ &= 1(1-\frac{1}{k})(1-\frac{2}{k})\cdots(1-\frac{n-1}{k}) \\ &\approx e^{-1/k}e^{-2/k}\cdots e^{-(n-1)/k} \\ &= e^{-\frac{1}{k}(1+\cdots+n-1)} \\ &\approx e^{-n^2/k}.\end{aligned}$$

So then we have that

$$\mathbf{Pr}(A) = 1 - \mathbf{Pr}(A^c) \approx 1 - e^{-n^2/k}.$$

It turns out, for $k = 365$ and $n = 23$, we get a roughly 50% chance of two people having the same birthday!

2.1 Bayes Theorem

Bayes Theorem was motivated by disease testing.

Example 2.2 (False Positive Quiz)**.** We are testing for a rare disease, and our test has the following properties:

- If person has disease, we detect with 0.95 probability.
- If person doesn't have the disease, test is negative wp 0.95
- Random person has disease wp 0.001

Let A be the event that the person has the disease, and B be the event that the person tests positive. We would like to calculate $\mathbf{Pr}(A|B)$. We have via Bayes Theorem that

$$\begin{aligned}\mathbf{Pr}(A|B) &= \frac{\mathbf{Pr}(B|A)\,\mathbf{Pr}(A)}{\mathbf{Pr}(B|A)\,\mathbf{Pr}(A) + \mathbf{Pr}(B|A^c)\,\mathbf{Pr}(A^c)} \\ &= \frac{(0.95)(0.001)}{(0.95)(0.001)+(0.999)(0.05)} = 0.0187\end{aligned}$$

Most doctors, when asked, said this probability was 95%. The main contributing factor here is the fact that the prior $\mathbf{Pr}(A) = 0.001$ is so small. If we change the scenario and have $\mathbf{Pr}(A) = 0.01$, then our new probability $\mathbf{Pr}(A|B) = 0.16$, so we should be more worried. Note that the doctor would actually be correct if the disease were present in 1/2 of the population.

Definition 2.3. Two events are **Independent** if the occurrence of one provides no information about the occurrence of the other. i.e.

$$\mathbf{Pr}(A|B) = \mathbf{Pr}(A)$$

which is equivalent to saying

$$\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A)Pr(B)$$

Extending this, a collection of events S are independent if

$$\mathbf{Pr}(\bigcap_{i \in S} A_i) = \prod_{i \in S} \mathbf{Pr}(A_i)$$

Remark 2.4. Pairwise independence does not imply joint independence.

Remark 2.5. Being disjoint does not imply independence, nor does the implication hold in the other direction. If A, B are disjoint, then $\mathbf{Pr}(A \cap B) = 0$, and independence tells us that $\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A)\,\mathbf{Pr}(B)$, which would tell us for two events to be both disjoint and independent, at least one of the two events must have zero probability of occurring. Note that this tells us that base outcomes of our probability space, which are all disjoint by definition, and all have nonzero probability by definition, must not be independent.

Definition 2.6. Conditional Independence is when $\mathbf{Pr}(A \cap B|C) = \mathbf{Pr}(A|C)\,\mathbf{Pr}(B|C)$. Then we say that A and B are "conditionally independent given C".

Example 2.7. Say we have two coins, one with tails on both sides, one with heads on both sides. We pick one up at random, and we flip it twice. We also let H_i be the event that that ith flip is a heads. Note immediately that H_1 and H_2 are decidedly not independent. Now we denote A as the event of us picking the two-headed coin. Then we have that $\mathbf{Pr}(H_1 \cap H_2|A) = \mathbf{Pr}(H_1|A)\,\mathbf{Pr}(H_2|A \cap H_1) = \mathbf{Pr}(H_1|A)\,\mathbf{Pr}(H_2|A)$. So this is an example of events that are conditionally independent but not themselves independent.

Exercise 2.8. Construct an example of RVs that are independent, but not conditionally independent.

Example 2.9. I roll two fair die. What is the probability I see a 6 before I see a 7? Let's use independence to attack this problem.

Lets condition on the first roll. Let S be the event that the first roll of the two die is a 6, and T be the event that the first roll is a 7. Let E be the event we are looking for, that we see a 6 before we see a 7. Then we have

$$\begin{aligned}\mathbf{Pr}(E) &= \mathbf{Pr}(E|S)Pr(S) + \mathbf{Pr}(E|T)Pr(T) + \mathbf{Pr}(E|(S \cup T)^c)\,\mathbf{Pr}((S \cup T)^c) \\ &= 1 \times 5/36 + 0 \times 6/36 + \mathbf{Pr}(E) \times 25/36 \\ &\implies \mathbf{Pr}(E) = 5/11\end{aligned}$$

2.2 Discrete Random Variables

Definition 2.10. Random Variables associate a real number with each possible outcome. They are inherently a *function* $f : \Omega \to \mathbb{R}$.

Why is this useful? If we are stuck with only events, we have no numbers to work with, we can't calculate means and variances and we cannot do statistics. Heads and tails only gets us so far, but if we assign the value 0 or 1 now we can do *math*.

Example 2.11 (Some random variables). 1. The RV X has value i if the throw of a die is i.

2. X^2 is a perfectly valid random variable.

Consider rolling two four-sided dice. Then M_k is the *event* that the min is k, whereas we can say M is the *random variable* that is equal to the value of the minimum of the two die. By enumerating all the possible values of the two die roll (which are all equal probability), we can see that $M = 1$ wp $7/16$, $M = 2$ wp $5/16$, $M = 3$ wp $3/16$, and $M = 4$ wp $1/16$. This mapping from values of a RV to probabilities for discrete random variables is known as a **probability mass function** or **PMF**, and in a way it defines the random variable.

There are nigh on an uncountable number of notations you will see for PMFs, but I'll just briefly go over the one's you'll see in these notes. We let the PMF of a RV X be P_X, and we say $P_X(x) = \mathbf{Pr}(\{X = x\})$, often simply denoted $\mathbf{Pr}(X = x)$. We also have to have (in order for our PMF to be valid), that

$$\sum_x P_X(x) = 1, \qquad P(X \in S) = \sum_{x \in S} P_X(x)$$

Example 2.12 (chess). Imagine Vishy Anand is playing Kasparov in chess (when they are at the height of their power). They play 10 games, and for each individual game, the probability Anand wins is 0.3, the probability that Kasparov wins is 0.4, and then probability that they draw is 0.3. The first to win a game wins the match, and if there are ten consecutive draws then the match is drawn. **Question:** what is the PMF of the duration of the match L? We have that

$$P_L(l) = \begin{cases} 0.3^9 & l = 10 \\ 0.3^{l-1} \cdot 0.7 & 1 \leq l \leq 9 \end{cases}$$

Question: What is the probability that Anand wins the match?

$$\mathbf{Pr}(\text{A wins the match}) = \sum_{l=0}^{9} (0.3)^l (0.3)$$

which can be simplified using the formula for a geometric series.

3 Lecture 3: Expectation, Uniform, Geometric, Binomial and Poisson Distributions

Agenda:

1. Recap of Discrete RVs and Probability Mass Functions (PMF)
2. Expectation
3. Some popular Discrete RVs
4. Variance

As a reminder, **discrete random variables** (DRVs) associate a real number with each possible outcome, so they are really just functions from $\Omega \to \mathbb{R}$. The distribution or **PMF** is the collection of values $\{a, P_X(a) : a \in \mathcal{A}\}$ where $\mathcal{A}$ is the set of all possible values taken by the RV X

Remark 3.1 (Functions of RVs are still RVs)**.** Let $Y = g(X)$. Then we have

$$P_Y(y) = \sum_{\{x|g(x)=y\}} P_X(x)$$

An RV itself is a function, and the function of a function is still a function!

Example 3.2. Let $Y = |X|$, where X is uniformly distributed between -2 and 2. So then $P_X(x) = 1/5, \quad \forall x \in \{-2, -1, 0, 1, 2\}$. Then $P_Y(y) = 2/5$ for $y = 1, 2$ and $P_Y(y) = 1/5$ for $y = 0$.

3.1 Expectation

Definition 3.3. We have the **expectation** of a discrete RV X that takes on values in a set $\mathcal{X}$ is

$$\mathbf{E}[X] = \sum_{x \in \mathcal{X}} x P_X(x)$$

Alternatively, we also have $\mathbf{E}[X] = \sum_{\omega \in \Omega} X(\omega) P(\omega)$

Theorem 3.4 (Expectations of Functions of RVs)**.** *Let $Y = g(X)$. Then we have that*

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

Note that there are no restrictions on the function g, so it holds for any function.

Proof. We start by noting

$$\mathbf{Pr}(Y = y) = \sum_{x:g(x)=y} \mathbf{Pr}(X = x)$$

Then, we have that

$$\begin{aligned}\mathbf{E}[Y] = \sum_y y \,\mathbf{Pr}(Y = y) &= \sum_y y \sum_{x:g(x)=y} \mathbf{Pr}(X = x) \\ &= \sum_y \sum_{x:g(x)=y} g(x)\,\mathbf{Pr}(X = x) \\ &= \sum_x g(x)\,\mathbf{Pr}(X = x). \qquad \square\end{aligned}$$

We can then use the above theorem to prove linearity of expectations!

Theorem 3.5. *(Linearity of Expectation)*
We have

$$E[X + Y] = E[X] + E[Y]$$

for arbitrary X and Y that are defined on the same probability space. This of course generalizes (via induction) to more than just two RVs.

Proof. We let $g(X, Y) = X + Y$. Then, according to the above theorem, we have that

$$\begin{aligned}\mathbf{E}[X + Y] &= \sum_{x,y} (x + y)\,\mathbf{Pr}(X = x, Y = y) \\ &= \sum_{x,y} x\,\mathbf{Pr}(X = x, Y = y) + \sum_{x,y} y\,\mathbf{Pr}(X = x, Y = y) \\ &= \sum_x \sum_y x\,\mathbf{Pr}(X = x, Y = y) + \sum_y \sum_x y\,\mathbf{Pr}(X = x, Y = y) \\ &= \sum_x x \sum_y \mathbf{Pr}(X = x, Y = y) + \sum_y y \sum_x \mathbf{Pr}(X = x, Y = y) \\ &= \sum_x x\,\mathbf{Pr}(X = x) + \sum_y y\,\mathbf{Pr}(Y = y) \qquad \text{(law of total probability)} \\ &= \mathbf{E}[X] + \mathbf{E}[Y]. \qquad \square\end{aligned}$$

The extremely important property of linearity of expectations, which you may already be familiar with, is that this holds *even if X and Y are dependent on each other*. The fact that we don't need any assumptions about independence is what makes the linearity property of expectation so powerful and useful!

Example 3.6. The average of the sum of two rolls of the dice X_1, X_2 is

$$\mathbf{E}[X] = \mathbf{E}[X_1 + X_2] = \mathbf{E}[X_1] + \mathbf{E}[X_2] = 7$$

Example 3.7. Suppose Prof Ramchandran collects homeworks from n students, shuffles them randomly, and then hands them back (at random). What is the expected number of students who get their homework back? More formally, what is the expected number of fixed points in a random permutation of n points?

Solution: Let X_i be the indicator RV that equals 1 if student i gets their homework back, and equals 0 otherwise. Then we can note that the number of students who get their homework back X, is exactly equal to $X_1 + ... + X_n$. So then

$$\mathbf{E}[X] = \mathbf{E}[X_1 + ... + X_n] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum i = 1^n\,\mathbf{Pr}(X_i = 1) = n \cdot \frac{1}{n} = 1$$

Remarkably, we see that the expected number of fixed points is always 1, regardless of how large or small n is. This technique of defining indicator RVs and applying linearity of expectation is extremely powerful and will come up over and over again in this course. When in doubt, come up with some indicators!

Remark 3.8. The X_i's in the previous example are **not** independent (exercise: why?), yet we can still apply linearity of expectation!

Definition 3.9. We define the **Variance** of a RV X, sometimes denoted σ_X^2 is

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

And furthermore the **standard deviation** is

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

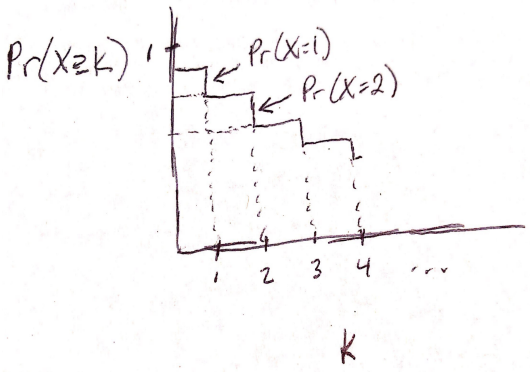
Exercise 3.10. From the definition of variance derive that

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

Theorem 3.11 (Discrete Tail Sum Formula)**.** *The (discrete) tail sum formula that we know and love for* ***positive valued*** *random variables is*

$$\mathbf{E}[X] = \sum_x x \, \mathbf{Pr}(X = x) = \sum_{k=1}^{\infty} \mathbf{Pr}(X \geq k)$$

There is a derivation that just does some tricks with algebra inside summations, but first I will give a hopefully more intuitive picture of what is going on here.



Consider the above picture. The regular formula for expectation, $\mathbf{E}[X] = \sum_{k=1}^{\infty} k \, \mathbf{Pr}(X = k)$, is equivalent to calculating the area of the above graph horizontally, while the tail sum formula $\sum_{k=1}^{\infty} \mathbf{Pr}(X \geq k)$, is equivalent to calculating the area of the above graph vertically.

Proof. Now we give the (less intuitive) algebraic proof:

$$\mathbf{E}[X] := \sum_{x=1}^{\infty} x \, \mathbf{Pr}(X = x)$$

notice here that $x \, \mathbf{Pr}(X = x) = \sum_{k=1}^{x} \mathbf{Pr}(X = x)$, so then we have:

$$\begin{aligned}\mathbf{E}[X] &= \sum_{x=1}^{\infty}\sum_{k=1}^{x}\mathbf{Pr}(X=x)\\ &= \sum_{k=1}^{\infty}\sum_{x=k}^{\infty}\mathbf{Pr}(X=x)\\ &= \sum_{k=1}^{\infty}\mathbf{Pr}(X\geq k)\end{aligned}$$

□

Exercise 3.12. Convince yourself that $\sum_{x=1}^{\infty}\sum_{k=1}^{x}\mathbf{Pr}(X=x) = \sum_{k=1}^{\infty}\sum_{x=k}^{\infty}\mathbf{Pr}(X=x)$. It may help to draw a graph of an arbitrary distribution, with $\mathbf{Pr}(X=x)$ as the y-axis and x as the x-axis.

3.2 Some Popular Discrete Random Variables

Definition 3.13 (Discrete Uniform RV)**.** The discrete uniform distribution over $[n] = \{1, ..., n\}$ has PMF:

$$P_X(k) = \frac{1}{n}, \forall k \in [n]$$

We can easily see that for uniform X, we have $\mathbf{E}[X] = \frac{n+1}{2}$.

Definition 3.14 (Bernoulli ("coin flip") RV)**.** The Bernoulli(p) RV takes on the value 1 with probability p, and 0 with probability $1-p$. Explicitly:

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

And we can easily calculate that $\mathbf{E}[X] = p$. We also have

$$\mathbf{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = p - p^2 = p(1-p)$$

Definition 3.15 (Indicator RV)**.** An **indicator RV** of an event A takes on the value of 1 if A happens/is true and 0 otherwise:

$$X = \{1\}_A = \mathbb{1}_A = \begin{cases} 1 & \text{A is true} \\ 0 & \text{else} \end{cases}$$

We can note then that

$$\mathbf{E}[\mathbb{1}_A] = \sum_x x\,\mathbf{Pr}(X=x) = \mathbf{Pr}(A)$$

Definition 3.16 (Binomial Random Variable)**.** If $X \sim Bin(n,p)$ then we define the PMF (probability mass function) as:

$$P_X(k) = \mathbf{Pr}(X=k) = \binom{n}{k}p^k(1-p)^{n-k}$$

The Binomial distribution is by definition also just the sum of n iid Bernoulli variables with parameter p. $X = \sum_{i=1}^{n} B_i$ where $B_i \sim Ber(p)$

It is not too difficult to calculate the expectation and variance of a binomial random variable, precisely because it can be represented as the sum of n i.i.d. Bernoullis. We have that we can use linearity of expectations to calculate the expectation of $X \sim Bin(n, p)$. We have

$$\mathbf{E}[X] = \mathbf{E}[\sum_{i=1}^{n} B_i] = \sum_{i=1}^{n} \mathbf{E}[B_i] = \sum_{i=1}^{n} p = np$$

Example 3.17. Let $Y = aX + b$. Then we have that

$$\begin{aligned}
\mathbf{Var}(Y) = \mathbf{Var}(aX + b) &= \mathbf{E}[(aX + b) - \mathbf{E}[aX + b]] \\
&= \mathbf{E}[((aX + b) - (a\,\mathbf{E}[X] + b))^2] \\
&= \mathbf{E}[(aX - a\,\mathbf{E}[X])^2] \\
&= a^2\,\mathbf{E}[(X - \mathbf{E}[X])^2] \\
&= a^2\,\mathbf{Var}(X)
\end{aligned}$$

Note that adding a constant does not affect the variance, nor should it intuitively, as we are simply shifting where the variable occurs and not affecting the spread of the variable at all. Multiplying by a constant, however, should and does affect the spread and therefore the variance of an RV.

Definition 3.18 (Geometric Random Variable)**.** A geometric random variable counts the time until the first success. We have that the PMF of a geometric random variable with parameter p (the probability of success is p) is as follows:

$$\mathbf{Pr}(X = k) = (1 - p)^{k-1} p$$

The above formula makes sense because we need the first $k-1$ events to be failures, which happen with probability $1 - p$, and then we need the k^{th} event to be a success, which happens with probability p. Also intuitively, if we want to calculate the probability $\mathbf{Pr}(X > k)$, then we need the first k events to all be failures, and it does not matter at all what happens after that. Therefore, $\mathbf{Pr}(X > k) = (1-p)^k$, which tells us that the CDF of a geometric RV is $\mathbf{Pr}(X \leq k) = 1 - \mathbf{Pr}(X > k) = 1 - (1 - p)^k$. As a sanity check, we can differentiate the CDF and find that it does indeed equal the PDF.

We can also calculate more easily the expectation of geometric random variable using the tail sum formula. We have that

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} \mathbf{Pr}(X \geq k) = \sum_{i=1}^{\infty} (1 - p)^{k-1} = \frac{1}{p}$$

Where the last step follows from the formula for infinite geometric series.

4 Lecture 4: (Co)variance, Correlation, Conditional / Iterated Expectation, Law of Total Variance

Agenda

1. Recap of expectation, their properties, and popular RVs
2. Memoryless property of Geometric(p) RVs
3. Conditional RV and Iterated Expectation
4. Covariance

4.1 Geometric RV and Properties, Poisson RV

Example 4.1 (Coupon Collector Problem). Imagine we have N balls of different colors, and we sample with replacement. What is the expected number of trials before we see all of the colors? To address this problem, we start by defining a few variables. Let C_r be the number of samplings required until we see at least r distinct colors. Then we know that $C_1 = 1$ and is in fact not at all random. We further define X_i as the number of samplings required to see i distinct colors given that we have already seen $i-1$ colors. We note here also that each X_i is a geometric random variable with parameter (probability of success) $p = \frac{N-i+1}{N}$ We also note that we have

$$C_N = \sum_{i=1}^{N} X_i$$

Then,

$$\mathbf{E}[C_N] = \sum_{i=1}^{N} \mathbf{E}[X_i] = 1 + \frac{N}{N-1} + \frac{N}{N-2} + ... + N \approx N \log N$$

Definition 4.2 (Poisson Random Variable). We define $X \sim Pois(\lambda)$ with the following PMF:

$$\mathbf{Pr}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

In general, the parameter λ describes a *rate*, i.e. the number of customers entering the store in a hour. We can calculate for $X \sim Pois(\lambda)$ the expectation:

$$\mathbf{E}[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

Exercise 4.3. Prove that for $X \sim Pois(\lambda)$:

$$\mathbf{Var}(X) = \lambda$$

Exercise 4.4 (Poisson Merging). Prove that for $X \sim Pois(\lambda)$ and $Y \sim Pois(\mu)$ we have $X + Y \sim Pois(\lambda + \mu)$ (this is done in discussion)

Exercise 4.5 (Poisson splitting)**.** Prove that if we "split" $X \sim Pois(\lambda)$ into two paths, by having an arrival take one path with probability p and the other with probability $1-p$. Prove that the number of arrivals to the first path Y is $Pois(p\lambda)$ and is moreover independent of the number of arrivals to the second path Z, which is distributed according to $Pois((1-p)\lambda)$. (this is a homework problem)

We now explore the relationship between the binomial distribution and poisson distribution. The poisson distribution actually turns out to be a limit of the binomial distribution, as we let n get large and p go to zero. Specifically, we must have that $\lim_{n\to\infty} np_n = \lambda$. Consider letting $p = \frac{\lambda}{n}$. Then we get the PMF for the binomial becomes:

$$\begin{aligned}\mathbf{Pr}(X=k) &= \binom{n}{k}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k} \\ &= \frac{n(n-1)...(n-k+1)}{n^k}\frac{\lambda^k}{k!}(1-\frac{\lambda}{n})^{n-k} \\ &= \frac{n}{n}\frac{n-1}{n}...\frac{n-k+1}{n}\frac{\lambda^k}{k!}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^k.\end{aligned}$$

Now, as we let $n \to \infty$, we can see that the first k left terms go to 1, as well as the rightmost term. We also know that the second to rightmost term approaches $e^{-\lambda}$. This leaves us with the pdf for the poisson distribution: $\mathbf{Pr}(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$.

Example 4.6 (St. Petersburg Paradox)**.** I keep tossing a fair coin until I get heads. If this takes n tosses, then I get 2^n dollars. How much should I pay to play this game? Well, if W is the amount I win, we can calculate:

$$\mathbf{E}[W] = \sum_{k=0}^{\infty} 2^k \frac{1}{2^k} = 2(\frac{1}{2} + 4\frac{1}{4} + 8\frac{1}{8} + \dots) = 1+1+1+.... = \infty$$

So I should pay an unbounded amount to play this game? Bernoulli said we should actually calculate $\log U$ where U is our utility/payout, in which case we would only pay \$4 to play this game.

Lemma 4.7. *If X and Y are independent,*

$$\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$$

Proof.

$$\begin{aligned}\mathbf{E}[XY] &= \sum_x \sum_y xyP_{XY}(x,y) \\ &= \sum_x \sum_y xyP_X(x)P_Y(y) = \sum_x xP_X(x)\sum_y yP_Y(y) = \mathbf{E}[X]\,\mathbf{E}[Y].\end{aligned}$$

□

Remark 4.8. The converse is generally **not** true:

$$\mathbf{E}[X]\,\mathbf{E}[Y] = \mathbf{E}[XY] \implies \text{X is independent of Y}$$

Lemma 4.9. *If X and Y are independent,*

$$\mathbf{Var}(X+Y) = \mathbf{Var}(X) + \mathbf{Var}(Y)$$

Proof. Without loss of generality (WLOG), we can say $\mathbf{E}[X] = \mathbf{E}[Y] = 0$, since variance is not affected by the shifting of a constant and therefore subtracting out the means does not alter the variance. Then if we let $Z = X + Y$, we have

$$\begin{aligned}\mathbf{Var}(Z) = \mathbf{Var}(X+Y) = \mathbf{E}[(X+Y)^2] &= \mathbf{E}[X^2 + Y^2 + 2XY] \\ &= \mathbf{E}[X^2] + \mathbf{E}[Y^2] + 2\,\mathbf{E}[X]\,\mathbf{E}[Y] = \mathbf{Var}(X) + \mathbf{Var}(Y).\end{aligned}$$ □

The above two lemmas of course generalize to more than just two independent variables via induction. We can also use the above lemma to calculate the variance of a binomial very easily, since a binomial $X \sim Bin(n,p)$ is equal to $B_1 + \cdots + B_n$, where each $B_i \sim Bern(p)$. Then, we have that (since the B_i are iid)

$$\mathbf{Var}(X) = \mathbf{Var}(\sum_{i=1}^{n} B_i) = \sum_{i=1}^{n} \mathbf{Var}(B_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

The above lemma raises the question, what if X_1 and X_2 are not independent, but we would like to calculate $\mathbf{Var}(X_1 + X_2)$?

Definition 4.10 (Covariance)**.** Consider $\mathbf{Var}(X+Y) = \mathbf{E}[(X+Y-\mathbf{E}[X]-\mathbf{E}[Y])^2]$. Now let $\hat{X} = X - \mathbf{E}[X]$ and $\hat{Y} = Y - \mathbf{E}[Y]$. Then

$$\mathbf{Var}(X+Y) = \mathbf{E}[(\hat{X}+\hat{Y})^2] = \mathbf{E}[\hat{X}^2] + \mathbf{E}[\hat{Y}^2] + 2\,\mathbf{E}[\hat{X}\hat{Y}].$$

This last term, $\mathbf{E}[\hat{X}\hat{Y}]$, is called the **covariance** of X and Y, and it tells us how they change with each other. We have that

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

Intuitively, the covariance between two variables is related to how they affect each other. If X_1 increasing causes X_2 to generally increase, then the covariance will be positive. If X_1 increasing causes X_2 to generally decrease, then the covariance will be negative.

Definition 4.11 (Correlation Coefficient)**.**

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

The above is known as the correlation coefficient of two variables, and is always between -1 and 1. This can be proved using the Cauchy-Schwarz Inequality (try it!)

4.2 Conditioning of RVs

When we consider $X|Y$, the first thing to note is that this is just another random variable, with its own PMF $P_{X|Y}(x|y) = \mathbf{Pr}(X = x | Y = y)$. Therefore, we must still have that

$$\sum_{x} P_{X|Y}(x|y) = 1$$

Lemma 4.12 (Memorylessness of Geometric RVs). *We have that for geometric RV X*

$$\mathbf{Pr}(X = k + m | X > k) = \mathbf{Pr}(X = m)$$

Proof.

$$\begin{aligned}\mathbf{Pr}(X = k + m | X > k) &= \mathbf{Pr}(X = k + m \cap X > k) / \mathbf{Pr}(X > k) \\ &= \frac{\mathbf{Pr}(X = k + m)}{\mathbf{Pr}(X > k)} = \frac{(1-p)^{k+m-1}p}{(1-p)^k} \\ &= (1-p)^{m-1}p = \mathbf{Pr}(X = m).\end{aligned}$$

□

We can use a clever conditioning trick, along with the memorylessness property, to calculate the variance of a geometric random variable. First, we need $\mathbf{E}[X^2]$. We have by total probability:

$$\begin{aligned}\mathbf{E}[X^2] &= \mathbf{E}[X^2 | X = 1] \mathbf{Pr}(X = 1) + \mathbf{E}[X^2 | X > 1] \mathbf{Pr}(X > 1) \\ &= p + (1-p) \mathbf{E}[(1+X)^2],\end{aligned}$$

where $\mathbf{E}[X^2 | X > 1] = \mathbf{E}[(1+X)^2]$ follows from the memorylessness property (convince yourself this is true). Then,

$$\begin{aligned}\mathbf{E}[X^2] &= p + (1-p)(1 + \frac{2}{p} + \mathbf{E}[X^2]) \\ &\implies p \mathbf{E}[X^2] = 1 + \frac{2-2p}{p} = \frac{2-p}{p} \\ &\implies \mathbf{E}[X^2] = \frac{2-p}{p^2}.\end{aligned}$$

Then we have that

$$\mathbf{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

Example 4.13 (Romance is dead). $2m$ people form couples. 50 years from now, the probability that any person is alive is p. Now suppose that there are A people alive after 50 years. Let S be the number of couples for which both people are still alive. We would like to find $\mathbf{E}[S|A = a]$. In order to do this, we further define X_i as the indicator that the first person of couple i survives, and Y_i as the indicator that the second person of couple i survives. Then $S = \sum_i X_i Y_i$. Then, we have

$$\begin{aligned}\mathbf{E}[S|A = a] = \mathbf{E}[\sum_i X_i Y_i | A = a] &= \sum_i \mathbf{E}[X_i Y_i | A = a] \\ &= m \mathbf{E}[X_i Y_i | A = a] \\ &= m \mathbf{Pr}(X_i Y_i = 1 | A = a) \\ &= m \frac{a}{2m} \frac{a-1}{2m-1} \\ &= m \frac{\binom{2m-2}{a-2}}{\binom{2m}{a}}.\end{aligned}$$

Why have we included the last equality, rather than simplifying further? Because it lends itself to an alternate interpretation of the solution. Consider couple i. What is the probability that they survive, given that $A = a$? Well $\binom{2m-2}{a-2}$ is the number of ways for a people to survive including this specific

couple, and $\binom{2m}{a}$ is the number of ways for a people to survive in general. More formally, we have:

$$\mathbf{Pr}(X_iY_i = 1|A = a) = \frac{\mathbf{Pr}(A = a|X_iY_i = 1)\,\mathbf{Pr}(X_iY_i = 1)}{\mathbf{Pr}(A = a)}$$

Now, A is a $Bin(2m, p)$ and $A|X_iY_i = 1$ is a $Bin(2m-2, p)$. So we have:

$$\begin{aligned} &= \frac{\binom{2m-2}{a-2}p^{a-2}(1-p)^{2m-a}p^2}{\binom{2m}{a}p^a(1-p)^{2m-a}} \\ &= \frac{\binom{2m-2}{a-2}}{\binom{2m}{a}}. \end{aligned}$$

And all we have to do is use linearity of expectations (and multiply by m) to get the same answer as above.

5 Lecture 5: Iterated Expectation, Continuous Probability, Uniform, Exponential Distributions

Agenda

1. Law of Iterated Expectations
2. Continuous probability (CDF, Uniform, Exp)

5.1 Iterated Expectation

Recall how **conditional expectation** works:

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

We say that $\mathbf{E}[X|Y=y]$ is the "expectation of X w.r.t. the distributions of X *conditioned on* $Y=y$, and it is really just a number.

Definition 5.1. Let X and Y be RVs. Then $\mathbf{E}[X|Y]$ is also a RV, the conditional expectation of X given Y, which has the value $\mathbf{E}[X|Y=y]$ with probability $\mathbf{Pr}(Y=y)$. It is important but subtle to note that $\mathbf{E}[X|Y]$ is a RV itself.

Example 5.2. Suppose we roll a die N times. Let X be the sum of the die rolls. Then we have that

$$\mathbf{E}[X|N=1] = \frac{7}{2}$$
$$\mathbf{E}[X|N=2] = 7$$

and in general, $\mathbf{E}[X|N=n] = \frac{7n}{2}$, and in general:

$$\mathbf{E}[X|N] = \frac{7N}{2}$$

The difference here is subtle, but the last equality is actually a much stronger statement, as it equates random variables rather than just numbers.

Out of this comes a natural question: since $\mathbf{E}[X|Y]$ is a RV, what is its expectation?

Theorem 5.3 (Iterated Expectations/Tower Rule)**.**

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

Proof.

$$\begin{aligned}
\mathbf{E}[\mathbf{E}[X|Y]] &= \sum_y \mathbf{E}[X|Y=y] \, \mathbf{Pr}(Y=y) \\
&= \sum_y \sum_x x \, \mathbf{Pr}(X=x|Y=y) \, \mathbf{Pr}(Y=y) \\
&= \sum_x x \sum_y \mathbf{Pr}(X=x, Y=y) \\
&= \sum_x x \, \mathbf{Pr}(X=x) = \mathbf{E}[X].
\end{aligned}$$

□

Example 5.4. We roll a die N times where $N \sim Geom(p)$. As before, X represents the sum of the N die rolls. Then we have:

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|N]] = \mathbf{E}\left[\frac{7N}{2}\right] = \frac{7}{2}\mathbf{E}[X] = \frac{7}{2p}$$

Example 5.5 (Drunken walk on a line)**.** Suppose we take a random walk, starting at the origin, on a discretized line. Then if X_{n+1} is our location at time $n+1$, then we have the recurrence:

$$X_{n+1} = X_n + \mathbb{1}_+ - \mathbb{1}_-$$

where $\mathbb{1}_+$ is an indicator for drunk taking a $+1$ step, and likewise $\mathbb{1}_-$ is an indicator for drunk taking a -1 step. Then we have:

$$\mathbf{E}[X_{n+1}] = \mathbf{E}[X_n] + 1/2 - 1/2 = \mathbf{E}[X_n] = 0$$

But what about the variance of the walk? $\mathbf{E}[X_n^2] = ?$. We have

$$\mathbf{Pr}(X_{n+1}^2 = (k+1)^2 | X_n = k) = \mathbf{Pr}(X_{n+1}^2 = (k-1)^2 | X_n = k) = 1/2$$

So then we have:

$$\mathbf{E}[X_{n+1}^2 | X_n = k] = \frac{(k+1)^2 + (k-1)^2}{2} = k^2 + 1$$

$$\implies \mathbf{E}[X_{n+1}^2 | X_n] = X_n^2 + 1$$

Then we can calculate

$$\mathbf{E}[X_{n+1}^2] = \mathbf{E}[\mathbf{E}[X_{n+1}^2 | X_n]] = \mathbf{E}[X_n^2] + 1$$

$$= \mathbf{E}[X_{n-1}^2 + 1 + 1$$

and then, after noting $\mathbf{E}[X_0^2] = 0$, we can see that

$$\mathbf{Var}(X_n) = \mathbf{E}[X_n^2] = n$$

5.2 Continuous Probability

Continuous RVs is a concept you should be relatively familiar with from CS70, but we will go over it quickly again and there are some subtleties to make sure are clear.

In most settings, a **continuous sample space** is more natural than a discrete one (such as distance, time, temperature, etc). For a continuous RV, there is no such thing as $\mathbf{Pr}(X = x)$. Well there is, but it's just equal to zero and generally pretty meaningless. We need to instead define probability over sets that have "length" and quantify "allowable events". What "allowable events" refers to here gets more into measure theory, which we are not going to get into in this course, as in virtually all engineering applications the distinction is unimportant. So rather than talking about $\mathbf{Pr}(X = x)$, we instead talk about f_X, which is the **probability density function (PDF)** of a continuous random variable.

Definition 5.6. X is a **continuous RV** if

1. $\exists$ a non-negative function f_X s.t.

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

is well-defined.

2. It must hold that
$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

The function f is characterized by the random variable X, hence the subscript.

Remark 5.7. $\mathbf{Pr}(X = a) = 0$, which means that $\mathbf{Pr}(X < a) = \mathbf{Pr}(X \leq a)$ for continuous random variables, and henceforth I will be lazy and interchange $<$ and $\leq$ for continuous RVs at will.

This density has the property that if we wish to calculate the probability that our random variable falls in a small δ sized interval, we have

$$\mathbf{Pr}(X \in [x, x+\delta]) = \int_x^{x+\delta} f_X(t)dt \approx f_X(x)\delta$$

for small enough delta, of course. Then we have

$$f_X(x) \approx \frac{\mathbf{Pr}(X \in [x, x+\delta])}{\delta}$$

Hence the name "density function". Note that it is perfectly fine for the density to be greater than 1 at any particular point, as it is not a probability. We have only the requirement that the **integral of f_X over its domain must be equal to 1** (think about why this must be), and that the **density must be nonnegative**. Another useful interpretation may be to think of PDF values at certain points as relative likelihoods; that is, if $f_X(s) = 2f_X(t)$, then we are twice as likely to see values in a small δ neighborhood around s than values in a small δ neighborhood around t (if the density is continuous).

Example 5.8. let $f_X(x) = \frac{1}{2\sqrt{x}}$ for $0 < x < 1$ and take on the value 0 otherwise. Then we have that it is nonnegative, and that
$$\int_0^1 f_X(x)dx = 1$$
So this is a valid PDF.

Now we mention the **cumulative distribution function**, which completely analogously to the discrete case is simply $\mathbf{Pr}(X < x)$. Since the CDF would be

$$F(x) = \int_{-\infty}^{x} f_X(t)dt$$

The CDF has the following properties:

1. $F_X(\infty) = 1$
2. $F_X(-\infty) = 0$
3. if X is discrete, then
$$\mathbf{Pr}(X = k) = F_X(k) - F_X(k-1)$$
and in the continuous case:
$$f_X(x) = \frac{d}{dx}F_X(x)$$

Where the last fact follows from the fundamental theorem of calculus (if F is differentiable). This can be a very useful fact, as often the CDF is easier to calculate than the PDF.

Example 5.9. Imagine throwing darts at a unit circle. We model this by saying that the location of where the dart lands in the circle is completely random (i.e. uniform over the circle). We would like to find the CDF and PDF of Y, which is the distance from the origin of where the dart lands. We have that

$$\begin{aligned}\mathbf{Pr}(Y \leq y) &= \frac{\text{area of circle of radius y}}{\text{area of whole circle}} \\ &= \frac{\pi y^2}{\pi} = y^2\end{aligned}$$

. Then we have that simply

$$f_Y(y) = \frac{d}{dy} F_Y(y) = 2y$$

We can calculate that:

$$P(0.5 < Y < 0.6) = F_Y(0.6) - F_Y(0.5) = 0.36 - 0.25 = 0.11$$

.

We have some analogous definitions and lemmas that pretty much follow from the discrete case:

Definition 5.10. The **expectation** of a continuous RV X is

$$\int_{-\infty}^{\infty} x f_X(x) dx$$

Lemma 5.11.

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

Lemma 5.12. *if X, Y are independent, then*

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

Now we go over some popular continuous RVs.

Definition 5.13 (Uniform RV)**.** If $X \sim Unif[a, b]$, then it must have constant probability density between a and b and zero density everywhere else, which tells us that

$$f_X(x) = \frac{1}{b-a}$$

for $x \in [a, b]$.

We can calculate for $X \sim U[a, b]$ that:

$$\mathbf{E}[X] = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}$$

and also

$$\mathbf{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \frac{(b-a)^2}{12}$$

Exercise 5.14. verify the Variance of a uniform RV between a and b is actually what we claimed above.

Definition 5.15 (Exponential RV)**.** Let's say we wanted to find a continuous RV that had the same "memoryless" property as the discrete Geometric RV, and analogously measured "time to success" (or failure, however you want to look at it). But now this time, time is a continuous thing, say the amount of time before a lightbulb burns out. Specifically, for the memoryless property, we want $\mathbf{Pr}(X > t+s|X > s) = \mathbf{Pr}(X > t)$. That is, we want

$$\frac{\mathbf{Pr}(X > t+s \cap X > s)}{\mathbf{Pr}(X > s)} = \frac{\mathbf{Pr}(X > t+s)}{\mathbf{Pr}(X > s)} = \mathbf{Pr}(X > t)$$

The question then becomes, what function $g(t) = \mathbf{Pr}(X > t)$ satisfies $\frac{g(s+t)}{g(s)} = g(t)$? Well, eventually, we might notice that $g(x) = e^x$ works! The problem is, this increases $g(x)$ as x increases, which is not the behavior we want if we are to keep the analogy. Well, $g(x) = e^{-x}$ also works, and it is monotonically decreasing, so that is better! In fact, we can even throw in a constant $g(x) = e^{-\lambda x}$, for increased versatility, and it still is monotonically decreasing and memoryless. Then we have

$$F_X(x) = 1 - \mathbf{Pr}(X > x) = 1 - e^{-\lambda x}$$

$$\implies \frac{d}{dx}F_X(x) = f_X(x) = \lambda e^{-\lambda x}$$

for any $\lambda > 0$. We can further check that this integrates to 1 over its domain (since it is measuring time to success, this is a positive random variable):

$$\int_0^\infty f_X(x) = \lambda \int_0^\infty e^{-\lambda x} = 1$$

as desired. And with that I conclude the most long winded introduction to the exponential random variable that has ever been.

Exercise 5.16. Show that if $X \sim Exp(\lambda)$, then

$$\mathbf{E}[X] = \frac{1}{\lambda}$$

$$\mathbf{Var}(X) = \frac{1}{\lambda^2}$$

Definition 5.17 (Laplace Distribution)**.** Let $Z = X - Y$, where $X, Y \sim exp(\lambda)$, and X and Y are independent. Then how is Z distributed? Well, if $X > Y$, then by the memoryless property we have that Z is simply an exponential RV. This happens with probability $1/2$, so we have $f_Z(z) = \frac{1}{2}\lambda e^{-\lambda z}$. What if then $Y > X$? Then once again by the memoryless property, we get that Z is simply a negated exponential RV: $f_Z(z) = \frac{1}{2}\lambda e^{+\lambda z}$. So putting this together we have what is known as the **Laplace Distribution:**

$$f_Z(z) = \frac{1}{2}\lambda e^{-\lambda|z|}$$

6 Lecture 6: Normal Distribution, Continuous Analogs, Derived Distributions

Agenda: see above

6.1 Review

Exercise 6.1. Let R be the distance from the origin of a point randomly sampled on a unit ball (in $\mathbb{R}^3$).

1. what is the CDF of R?
2. PDF?
3. Expectation?

6.2 Normal Distribution

Definition 6.2 (Normal Distribution)**.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$, where μ is the mean of the distribution and σ is the standard deviation. Here is the PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

We call the pdf of $X \sim \mathcal{N}(0,1)$ is $F_X(x) = \Phi(x)$, which cannot be expressed in elementary functions

The PDF of the normal is clearly positive. We would like to also show that it integrates to 1:

Proof. We will show this when $\mu = 0$ and $\sigma^2 = 1$. The idea is to show that

$$\left(\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x\right)^2 = 1$$

We have that:

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x\right)^2 &= \left(\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} f_Y(y)\mathrm{d}y\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-(x^2+y^2)/2}\mathrm{d}x\mathrm{d}y \\
&= \int_{0}^{2\pi}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-(r^2)/2} r dr d\theta \qquad \text{(using polar integration, where } \mathrm{d}y\mathrm{d}x = r\mathrm{d}r\mathrm{d}\theta) \\
&= \int_{-\infty}^{\infty} e^{-(r^2)/2} r dr.
\end{aligned}$$

We can use u substitution to solve this integral, which will evaluate to 1 (think about the pdf of the exponential RV! Or just do it manually).

□

Some properties of Normal distributions:

1. if X, Y are independent normals, then $Z = X + Y$ is also normal $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

2. The sum of two dependent normals isn't always Normal. Consider $X \sim \mathcal{N}(0,1)$, and $Y = X$ w.p. $1/2$ and $-X$ wp $1/2$. Then both X and Y are normal but $X + Y$ is not normal.

3. We have that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then
$$\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

Example 6.3. Let $X \sim \mathcal{N}(2, 16)$. We wish to find $\mathbf{Pr}(-2 < X < 6)$. We have

$$\mathbf{Pr}(-2 < X < 6) = \mathbf{Pr}(-4 < X - 2 < 4) = \mathbf{Pr}(-1 < \frac{X-2}{4} < 1)$$

$$= \mathbf{Pr}(-1 < \mathcal{N}(0,1) < 1) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \approx 0.68$$

Where Φ is the CDF of the standard normal distribution.

Exercise 6.4. convince yourself that $\Phi(1) - \Phi(-1) = 2\Phi(1) - 1$ if you haven't already.

Example 6.5. Suppose male height is distributed as $\mathcal{N}(70, 5)$ and female height is $\mathcal{N}(64, 4)$. What's the probability that a random chosen male is taller than a randomly chosen female? Express your answer in terms of Φ.

Ans: Let X be the boys height and Y the girls height. We want to calculate $P(X - Y > 0) = P(Y - X < 0)$. Note that $Y - X \sim \mathcal{N}(-6, 9)$, and so

$$\frac{Y - X + 6}{\sqrt{9}} \sim \mathcal{N}(0,1)$$

So we have

$$P(Y - X < 0) = P(\frac{Y - X + 6}{3} < 2) = \Phi(2)$$

6.3 Continuous Analogs of Discrete RVs

For **joint distributions**, we can generalize from the discrete case:

$$P(A) = \sum_{(x,y)\in A} P_{X,Y}(x, y)$$

and analogously:

$$P(A) = \int_A f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y$$

The definitions of marginal probabilities, conditional probabilities, multiplication rule, and Bayes Rule all carry over naturally into the domain of continuous probability, all you need to do is replace summations with integrals and p_X's with f_X's.

Example 6.6. We can have discrete and continuous RVs defined jointly. For example. For example, let X be the outcome of a die roll, and $Y \sim Exp(X)$. Then we have

$$p_X(x) = \frac{1}{6}$$

and

$$f_{Y|X}(y|x) = xe^{-xy}$$

Example 6.7. Let $X \sim Bern(1/2)$ and $Y = 2X$. We have that the distribution of Y is

$$\mathbf{Pr}(Y = y) = \mathbf{Pr}(2X = y) = \mathbf{Pr}(X = \frac{y}{2})$$

more generally, if X is discrete RV, and $Y = f(X)$, then

$$\mathbf{Pr}(Y = y) = \mathbf{Pr}(f(X) = y) = \mathbf{Pr}(X \in f^{-1}(y))$$

Be careful! Is it then true that in the continuous case if $X \sim U[0, 1]$ and $Y = 2X$. Is it then true that

$$f_Y(y) = \mathbf{Pr}(Y = y) = \mathbf{Pr}(2X = y) = \mathbf{Pr}(X = y/2) = f_X\left(\frac{y}{2}\right)$$

NO. There are many things wrong here, first of all the quantity $\mathbf{Pr}(Y = y) = 0, \forall y$. Second, this does not integrate to 1:

$$\int_0^2 f_Y(y)\mathrm{d}y = \int_0^2 f_X(y/2)\mathrm{d}y = 2$$

Instead, we have to derive the CDF of Y properly using the CDF. It IS true that:

$$F_Y(y) = \mathbf{Pr}(Y \leq y) = \mathbf{Pr}(2X \leq y) = \mathbf{Pr}(X \leq \frac{y}{2}) = F_X(\frac{y}{2}$$

and then

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_X(y/2) = \frac{1}{2}f_X(\frac{y}{2})$$

Which we can check does integrate to 1.

Example 6.8 (More on the relationship between Exponential and Geometric RVs)**.** Toss a coin every δ seconds, and let the probability of heads $p = 1 - e^{-\lambda\delta}$, with $\delta << 1$. Let $N \sim Geom(p)$ and $X \sim exp(\lambda)$. Then we have that $F_N(n) = \mathbf{Pr}(N < n) = 1 - e^{-\lambda n\delta} = F_X(n\delta)$. If you graph $F_N(n)$ and $F_X(n\delta)$, then you can see how the exponential is the limit of the geometric as $\delta \to 0$.

It is useful to know that the Covariance is a multilinear function, meaning

$$\mathbf{Cov}(X + Y, W + Z) = \mathbf{Cov}(X, W) + \mathbf{Cov}(X, Z) + \mathbf{Cov}(Y, W) + \mathbf{Cov}(Y, Z)$$

And it is also useful to note that the variance $\mathbf{Var}(X) = \mathbf{Cov}(X, X)$. We also have that $\mathbf{Cov}(aX + b, Y) = a\,\mathbf{Cov}(X, Y)$. This yields the following useful identity:

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

The **Tower Rule** or **Iterated Expectation** or the **Law of Total Expectation** also holds in the continuous case. We have:

$$\begin{aligned}
\underset{Y}{\mathbf{E}}[\underset{X}{\mathbf{E}}[X|Y]] &= \int_Y f_Y(y) \int_X x f_{X|Y}(x|y)\mathrm{d}x\mathrm{d}y \\
&= \int_Y \int_X x f_{X|Y}(x|y) f_Y(y)\mathrm{d}x\mathrm{d}y = \int_X \int_Y x f_{X,Y}(x, y)\mathrm{d}y\mathrm{d}x \\
&= \int_X x f_X(x)\mathrm{d}x = \mathbf{E}[X].
\end{aligned}$$

The above result should make intuitive sense when you think about it, and the intuition is quite similar to the intuition behind discrete total probability. If we want to find $\mathbf{E}[X]$, and the instances of Y subdivide our probability space, it may be easier to calculate $\mathbf{E}[X|Y]$ for every Y. But then we have to weight each expectation the probability that particular instance of Y happens, hence the outside expectation over the Y variable.

Example 6.9. Consider trying to estimate X given some information Y with the estimate $\mathbf{E}[X|Y]$. Well we have that the error E is $E = X - \mathbf{E}[X|Y]$, and we further have that

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

and therefore $\mathbf{E}[X|Y]$ is called an *unbiased estimator*. We will learn more about this later in the semester though when we talk about MMSE.

7 Lecture 7: Order Statistics, Convolution, Moment Generating Functions

Agenda:

1. Law of Total Variance
2. Order Statistics
3. Convolution
4. Moment Generating Functions

7.1 Conditional Variance and Law of Total Variance

Definition 7.1 (Conditional Variance)**.** Let X, Y be RVs. We can define the conditional variance $\mathbf{Var}(X|Y=y)$ as the variance of the conditional distribution $P(X=x|Y=y)$.

Remark 7.2. $\mathbf{Var}(X|Y)$ is a RV that assumes the value $\mathbf{Var}(X|Y=y)$ with probability $\mathbf{Pr}(Y=y)$.

Lemma 7.3 (Total Variance)**.** *We have that*

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

I will now try to offer some sort of intuition before the formal proof. We want to answer the question: how much does X vary? Well, if we fix Y, we could take the expectation over all the $y \in Y$ of $\mathbf{Var}(X|Y)$. But even if we are fixing Y, there is still some variance in X, and therefore some variance in $\mathbf{E}[X|Y]$, which is where the second term comes into play. The first term is the expected variance from the mean of $X|Y$; the second is the variance of that mean.

Proof. We have that

$$\begin{aligned}
\mathbf{Var}(X) &= \mathbf{E}[X^2] + \mathbf{E}[X]^2 \\
&= \mathbf{E}[\mathbf{E}[X^2|Y]] - (\mathbf{E}[\mathbf{E}[X|Y]])^2 \\
&= \mathbf{E}[\mathbf{Var}(X|Y) + \mathbf{E}[X|Y]^2] - (\mathbf{E}[\mathbf{E}[X|Y]])^2 \\
&= \mathbf{E}[\mathbf{Var}(X|Y)] + \left(\mathbf{E}[\mathbf{E}[X|Y]^2] - (\mathbf{E}[\mathbf{E}[X|Y]])^2\right) \\
&= \mathbf{E}[\mathbf{Var}(X|Y)] + \mathbf{Var}(\mathbf{E}[X|Y]).
\end{aligned}$$

□

Example 7.4. We have a biased coin, we toss is n times, and we let X be the number of heads, and $Y \sim U[0,1]$ be the probability of heads (the bias of the coin). First, we have that

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]] = \mathbf{E}[nY] = n\,\mathbf{E}[Y] = \frac{n}{2}$$

Now, we can calculate the variance:

$$\begin{aligned}
\mathbf{Var}(X) &= \mathbf{Var}(\mathbf{E}[X|Y]) + \mathbf{E}[\mathbf{Var}(X|Y)] \\
&= \mathbf{Var}(nY) + \mathbf{E}[nY(1-Y)] \\
&= n^2\,\mathbf{Var}(Y) + n\,\mathbf{E}[Y] - n\,\mathbf{E}[Y^2] \\
&= \frac{n^2}{12} + \frac{n}{2} - \frac{n}{3} = \frac{n^2}{12} + \frac{n}{6}.
\end{aligned}$$

Compare this value to tossing a fair coin n times, which has variance $\frac{n}{4}$.

Example 7.5 (Random number of Random Variables)**.** Say we have $Y = X_1 + ... + X_N$, where the X_i are all independent and N is also random. What is $\mathbf{Var}(Y)$? First, we have:

$$\mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y|N]] = \mathbf{E}[N\,\mathbf{E}[X_i]] = \mathbf{E}[N]\,\mathbf{E}[X_i]$$

Where the second to last equality follows from linearity of expectations. Also, since N is given in the inner expectation, we can treat it as a constant (until the outer expectation). We then have:

$$\begin{aligned}\mathbf{Var}(Y) &= \mathbf{E}[\mathbf{Var}(Y|N)] + \mathbf{Var}(\mathbf{E}[Y|N]) \\ &= \mathbf{E}[N\,\mathbf{Var}(X_i)] + \mathbf{Var}(N\,\mathbf{E}[X_i]) \\ &= \mathbf{E}[N]\,\mathbf{Var}(X_i) + \mathbf{E}[X_i]^2\,\mathbf{Var}(N).\end{aligned}$$

7.2 Order Statistics

Let X be a continuous RV for which $x_1, x_2, ..., x_n$ are values of a random sample of size n. We can then reorder the x_i's from smallest to largest (we don't need to worry about ties, as we are in a continuous sample space here!).

Example 7.6. Suppose $X \sim U[0,1]$, and $n = 4$, and we observe $x_1 = 0.5, x_2 = 0.7, x_3 = 0.2, x_4 = 0.1$. Then we can order them as

$$x^{(1)} = 0.1, \quad x^{(2)} = 0.2, \quad x^{(3)} = 0.5, \quad x^{(4)} = 0.7$$

where $x^{(i)}$ i the i^{th} smallest observation

We call $X^{(i)} = \left(x^{(i)}\right)$ the i^{th} **order statistic**.

Theorem 7.7. *If X has pdf $f_X(x)$, the marginal pdf of the i^{th} order statistic is*

$$f_{X^{(i)}}(y) = \frac{n!}{(i-1)!(n-i)!}(F_X(y))^{i-1}(1 - F_X(y))^{n-i} f_X(y)$$

Proof. We present a sketch of the proof. We would like to calculate

$$\mathbf{Pr}(X^{(i)} \in \{y, y + dy\}) \approx f_{X^{(i)}}(y)dy$$

We need $i - 1$ of the samples to be less than y, which is the $(F_X(y))^{i-1}$ term. We also need exactly one to be right around y, which is approximately $f_X(y)dy$. Finally, we need $(n - i)$ of the samples to be greater than y, which is the $(1 - F_X(y))^{n-i}$ term. Lastly, we have to count how many ways we can pick with ones come first and which one is the the i^{th} largest (which exactly determines which ones come after y), which is $n * \binom{n}{i-1} = \frac{n!}{(i-1)!(n-i)!}$. Combining all of these together yields the exact expression we were looking for! □

Example 7.8 (Special case when X is uniform)**.** Suppose $X \sim U[0,1]$. Recall that $f_X(x) = 1$, and $F_X(x) = x$ (convince yourself if you've forgotten why this is true!). Then we can plug in and see that

$$f_{X^{(i)}}(y) = \frac{n!}{(i-1)!(n-i)!} y^{i-1}(1-y)^{n-i}$$

for $0 < y < 1$. This is a special case of a *Beta Distribution*.

Exercise 7.9. What is the probability that the 9th smallest of of ten draws from $X \sim U[0,1]$ is greater than 0.8?

7.3 Convolution

Let $Z = X + Y$, where X and Y are both continuous and independent. We would like to calculate the PDF of Z. We can relate Z to X using total probability:

$$f_Z(z) = \int_x f_{X,Z}(x,z)$$

Furthermore, we have

$$F_{Z|X} = \mathbf{Pr}(X + Y \leq z | X = x) = \mathbf{Pr}(Y \leq z - x | X = x) == \mathbf{Pr}(Y \leq z - x) = F_Y(z - x)$$

$$\implies f_{Z|X}(z|x) = f_Y(z - x)$$

Now, incorporating this into our original expression for $f_Z(z)$, we have

$$f_Z(z) = \int_x f_X(x) f_Y(z - x) dx = (f_X * f_Y)(z)$$

which is called a **convolution.** Intuitively, the expression should make sense, as we are just integrating over all possible combinations of X and Y that could sum to z. The discrete case is entirely analogous:

$$\mathbf{Pr}(Z = z) = \sum_k \mathbf{Pr}(X = k)\, \mathbf{Pr}(Y = n - k)$$

Example 7.10. Suppose $X, Y \sim U[0,1]$ are independent. What is $f_Z(z)$, where $Z = X + Y$? We could do an integral and get the right answer via the definition of the convolution, but we can also visually see that it becomes a triangle:

> **TODO 1.** tikz :(

As a general remark, convolution always creates more uncertainty than we started out with. In your homework you will show that if $X, Y \sim \mathcal{N}(0,1)$ are independent, then $Z = X + Y \sim \mathcal{N}(0,2)$.

7.4 Moment Generating Functions (MGFs)

Definition 7.11 (Moment Generating Functions)**.** We define the **Moment Generating Function** of an RV X as

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

Whats the point of MGFs? It seems like a fairly arbitrary definition. Well, first recall the Taylor series for e:

$$e^{sX} = 1 + sX + \frac{(sX)^2}{2!} + \frac{(sX)^3}{3!} + ...$$

$$\implies \mathbf{E}[e^{sX}] = 1 + s\,\mathbf{E}[X] + \frac{s^2}{2!}\mathbf{E}[X^2] + \frac{s^3}{3!}\mathbf{E}[X^3] + ...$$

Then, we can observe that

$$\frac{d}{ds}\mathbf{E}[e^{sX}]\bigg|_{s=0} = \mathbf{E}[X]$$

and

$$\frac{d^2}{ds^2}\,\mathbf{E}[e^{sX}]\bigg|_{s=0} = \mathbf{E}[X^2]$$

continuing in this manner, we can see that

$$\frac{d^n}{ds^n}M_X(s)\bigg|_{s=0} = \mathbf{E}[X^n]$$

Which is an extremely useful property of the MGF and can help with many computations. We also note that $M_X(0) = 1$ must be true.

8 Lecture 8: MGFs, Bounds/Concentration Inequalities (Markov, Chebyshev, Chernoff)

Agenda:

1. MGF's (examples and properties)
2. Limit theorems (Markov, Chebyshev, Chernoff)

8.1 Properties of MGFs

Recall that the Moment Generating Function (MGF) of an RV X is the transform:

$$M_X(s) = \mathbf{E}[e^{sX}] = \sum_{k=0}^{\infty} \frac{s^k \, \mathbf{E}[X^k]}{k!} = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx$$

and that

$$\frac{d^n}{ds^n} M_X(s)\Big|_{s=0} = \mathbf{E}[X^n]$$

Some utilities of the MGF:

1. Finding higher moments often becomes easier (derivatives are usually easier than integrals!)
2. Convolution becomes multiplication in the MGF domain, which is often much easier (again, avoiding integrals)
3. Great analytical tool to prove things (such as the CLT!)

And here are some properties to keep in mind:

1. $M_X(0) = 1$
2. if $X > 0$, then $M_X(-\infty) = 0$
3. if $X < 0$, then $M_X(\infty) = 0$
4. if $Y = aX + b$, we have

$$M_Y(s) = \mathbf{E}[e^{s(aX+b)}] = e^{sb} \, \mathbf{E}[e^{asX}] = e^{sb} M_X(as)$$

Example 8.1 (MGF of exponential RV)**.** Let $X \sim Exp(\lambda)$. Then we have that:

$$\begin{aligned} M_X(s) = \mathbf{E}[e^{sX}] &= \int_0^{\infty} e^{sX} \lambda e^{-\lambda x} \mathrm{d}x \\ &= \lambda \int_0^{\infty} e^{x(s-\lambda)} \mathrm{d}x \\ &= \lambda \frac{e^{x(s-\lambda)}}{s-\lambda}\Big|_0^{\infty} \\ &= \frac{\lambda}{\lambda - s}. \qquad \text{(assuming } \lambda > s\text{)} \end{aligned}$$

It is fine here that the MGF is not defined for all s, as we only need for it to be defined around $s = 0$ so that we can take derivatives evaluated at $s = 0$.

We can use the MGF of an exponential to easily calculate moments:

$$\mathbf{E}[X] = M'_X(0) = \frac{y}{(y-s)^2}\bigg|_{s=0} = \frac{1}{\lambda}$$

and we note that

$$\mathbf{E}[X^k] = \frac{d^k}{ds^k} M_X(s)\bigg|_{s=0} = \frac{\lambda k!}{(\lambda - s)^{k+1}}\bigg|_{s=0} = \frac{k!}{\lambda^k}$$

Example 8.2 (MGF of a Poisson)**.** We have that for $X \sim Pois(\lambda)$ that

$$\begin{aligned} M_X(s) &= \sum_{k=0}^{\infty} e^{sk} \mathbf{Pr}(X = k) \\ &= \sum_{k=0}^{\infty} e^{sk} \frac{e^{-\lambda}\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^s\lambda)k}{k!} \\ &= e^{-\lambda} e^{e^s\lambda} = e^{-\lambda+\lambda e^s}, \end{aligned}$$

which is valid for all values of s.

Example 8.3 (MGF of Normal RV)**.** Let $X \sim \mathcal{N}(0,1)$. Then we have

$$\begin{aligned} \mathbf{E}[e^{sX}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{-x^2/2} \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx - x^2/2} \mathrm{d}x \\ &= \frac{e^{s^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2/2 - sx + s^2/2)} && \text{(complete the square in the exponent)} \\ &= e^{s^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-s)^2/2} \\ &= e^{s^2/2}. \end{aligned}$$

In the last line we have used that fact that $\frac{1}{\sqrt{2\pi}} e^{-(x-s)^2/2}$ is the PDF of a standard normal that has been shifted by s, and so must integrate to 1.

Now, if $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = \sigma X + \mu$ and we have

$$\mathbf{E}[e^{sY}] = \mathbf{E}[e^{s(\sigma X + \mu)}] = e^{\mu s}\, \mathbf{E}[e^{\sigma Y s}] = e^{\mu s + \sigma^2 s^2/2}.$$

Remark 8.4. An interesting and useful fact (that we will not prove in this course) is that a given MGF corresponds to a unique CDF. This is related to the fact that $M_X(s)$ is just a Laplace transform of $f_X(x)$. Inversions are usually performed by just pattern matching.

Remark 8.5 (Convolving densities corresponds to multiplying their transforms)**.** Take $Z = X + Y$, and

assume X and Y are independent. Then we have that

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

Example 8.6 (MGF of binomial)**.** We can use the above remark very nicely in computing the MGF of a binomial RV, because we can use the fact that a binomial is simply the sum of bernoullis. We have $X \sim Bin(n,p) = Y_1 + ... + Y_n$, where $Y_i \sim Ber(p)$. We have then

$$M_{Y_i}(s) = \mathbf{E}[e^{Y_i s}] = (1-p)e^{s\cdot 0} + pe^s = 1 - p + pe^s$$

$$\implies M_X(s) = (1 - p + pe^s)^n$$

Example 8.7 (Summing of a random number of random variables)**.** Let $Y = X_1 + \cdots + X_N$, where $X_1, \ldots, X_N$ are i.i.d. and N is a RV. We then have that

$$\begin{aligned} M_Y(s) = \mathbf{E}[e^{Ys}] &= \mathbf{E}[\mathbf{E}[e^{Ys}|N]] \\ &= \mathbf{E}[\mathbf{E}[e^{s(X_1+\cdots+X_N)}|N]] = \mathbf{E}[M_X(s)^N] \\ &= \mathbf{E}[e^{N\ln(M_X(s))}] \\ &= M_N(\ln M_X(s)). \end{aligned}$$

Example 8.8 (Sum of Geometric number of exponential RVs)**.** We will begin with the fact that if $N \sim Geom(p)$, then

$$M_N(s) = \frac{pe^s}{1-(1-p)e^s}$$

Then, if $Y = X_1 + \cdots + X_N$, where each X_i is an iid exponential RV. Then, from the previous example, we have that $M_Y(s) = M_N(\ln(M_X(s)))$ We also have from before that $M_{X_i}(s) = \frac{\lambda}{\lambda - s}$. Then, we have

$$M_Y(s) = \frac{pM_{X_i}(s)}{1-(1-p)M_{X_i}(s)} = \frac{p\frac{\lambda}{\lambda-s}}{1-(1-p)\frac{\lambda}{\lambda-s}}.$$

8.2 Limiting Behavior of RV's

Suppose we observe a sequence $X_1, X_2, ..., X_n$ i.i.d. samples. We let

$$M_n = \frac{\sum X_i}{n}$$

be the **sample mean** (which makes sense, as it is just an average). We have:

1. $\mathbf{E}[M_n] = \frac{n\,\mathbf{E}[X_i]}{n} = \mu$
2. Assuming $\mathbf{Var}(X_i) < \infty$, we have

$$\mathbf{Var}(M_n) = \frac{1}{n^2}\sum \mathbf{Var}(X_i) = \frac{\mathbf{Var}(X_i)}{n} \implies 0$$

as $n \implies \infty$

A natural question is then: What happens to the "deviation" $|M_n - \mathbf{E}[M_n]| = |M_n - \mu|$?

Definition 8.9 (Markov Bound)**.** For a *non-negative random variable*, we have that

$$aP(X \geq a) \leq \mathbf{E}[X]$$

Proof. Define the indicator variable $Z = \begin{cases} 1 & X \geq a \\ 0 & \text{otherwise} \end{cases}$. Then we can see that $aZ \leq X$ by examining the two cases.

- If $X < a$ then $Z = 0$, so the condition is $0 \leq X$, which is true because we are considering X non-negative.
- If $X \geq a$ then $Z = 1$, so the condition is $a \leq X$, which is true by the condition with which we started this case.

Thus, we can take expectation on both sides:

$$\begin{aligned} \mathbf{E}[aZ] &\leq \mathbf{E}[X] \\ a\,\mathbf{E}[Z] &\leq \mathbf{E}[X] \\ a\mathbb{P}(X \geq a) &\leq \mathbf{E}[X]. \end{aligned}$$

□

Example 8.10. Let $X \sim U[0,1]$. Then we have

$$\mathbf{Pr}(X > 3/4) \leq \frac{1/2}{3/4} = 2/3$$

and

$$\mathbf{Pr}(X > 1) \leq \frac{1}{2}$$

Which seems pretty stupid/not very powerful. But it is this way because it makes very little assumptions on the RV. We don't hate on Markov too much because it is actually the building block for many other bounds, and is often very useful when we don't know much or anything about the higher moments of our RV.

Remark 8.11 (Markov Inequality intuition)**.** Say my distribution has a mean of μ, and I want to maximize the probability that $\mathbf{Pr}(X \geq k\mu)$. How would I do this? I would do this by letting X take on a value of $k\mu$ with probability $\frac{1}{k}$, and $X = 0$ otherwise. This achieves the correct expectation while still maximizing $\mathbf{Pr}(X \geq k\mu)$.

Definition 8.12 (Chebyshev's Inequality)**.** Chebyshev's Inequality states

$$\mathbf{Pr}(|X - \mu| \geq a) \leq \frac{\mathbf{Var}(X)}{a^2},$$

where $\mu = \mathbf{E}[X]$.

Proof. We know that $\mathbf{Var}(X) = \mathbf{E}[(X-\mu)^2]$. We then have that

$$\mathbf{Pr}((X-\mu)^2 \geq a) \leq \frac{\mathbf{Var}(X)}{a}$$

by Markov's inequality. This implies

$$\implies \mathbf{Pr}(|X-\mu| \geq \sqrt{a}) \leq \frac{\mathbf{Var}(X)}{a}$$

$$\implies \mathbf{Pr}(|X-\mu| \geq a) \leq \frac{\mathbf{Var}(X)}{a^2}.$$

□

Note that Chebyshev's inequality holds for any random variable, not just positive ones (in contrast to Markov's inequality. We also note the special case:

$$\mathbf{Pr}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

when X has mean μ and variance σ^2.

Remark 8.13 (Weak Law of Large Numbers)**.** We can use Chebyshev's inequality to show a result known as the **Weak Law of Large Numbers**. Suppose we have an average of a bunch of i.i.d. RVs $M_n = \frac{X_1+\cdots+X_n}{n}$. Then we have that $\mathbf{Var}(M_n) = \frac{n\,\mathbf{Var}(X_i)}{n^2} = \frac{\mathbf{Var}(X_i)}{n}$. This implies via Chebyshev's inequality that:

$$Pr(|M_n - \mathbf{E}[X_i]| \geq a) \leq \frac{\sigma}{na^2}$$

Definition 8.14 (Chernoff Bound)**.** Suppose we know the MGF of our random variable $M_X(s) = \mathbf{E}[e^{sX}]$. Note that this is a positive RV, so we can apply markov's inequality:

$$\mathbf{Pr}(e^{sX} \geq a) \leq \frac{\mathbf{E}[e^{sX}]}{a} = \frac{M_X(s)}{a}$$

$$\implies \mathbf{Pr}(e^{sX} \geq e^{as}) \leq \frac{\mathbf{E}[e^{sX}]}{e^{as}}$$

$$\implies \mathbf{Pr}(sX \geq as) \leq \frac{\mathbf{E}[e^{sX}]}{e^{as}}$$

where the last step follows since $f(x) = e^x$ is monotonic. Then, if $s > 0$ we have

$$\implies \mathbf{Pr}(X \geq a) \leq \frac{\mathbf{E}[e^{sX}]}{e^{as}}$$

alternatively, if $s < 0$ we have

$$\implies \mathbf{Pr}(X \leq a) \leq \frac{\mathbf{E}[e^{sX}]}{e^{as}}$$

Note that the Chernoff bound is a function of s. We often have to choose the optimal choice for s to get a good bound (take derivative and set to zero!). Also if we recall the Taylor series for e^x, the idea behind a Chernoff bound is that it can use all the moments of a RV to bound said RV. Compare this to Markov's, which only uses the first moment, and Chebyshev's, which only uses the second moment. This might lead one to think that Chernoff is *always* better than applying Markov/Chebyshev bounds, or even applying Markov's bound to higher moments of the random variable. This leads to the following remark:

Remark 8.15 (Is Chernoff always better than Markov/Chebyshev?)**.** In short, no. Consider using Markov's inequality to bound a higher moment of our RV X. This yields (provided the higher

moment is positive of course) $\mathbf{Pr}(X \geq a) \leq \frac{\mathbf{E}[X^k]}{a^k}$. Here I claim:

$$\inf_{k>0} \frac{\mathbf{E}[X^k]}{a^k} \leq \inf_{s>0} \frac{\mathbf{E}[e^{sX}]}{e^{as}}$$

Why is this true? Lets examine the RHS:

$$\begin{aligned}\frac{\mathbf{E}[e^{sX}]}{e^{as}} &= \frac{1}{e^{as}} \sum_k \frac{s^k \mathbf{E}[X^k]}{k!} \\ &= \sum_k \left(\frac{(as)^k e^{-as}}{k!} \right) \frac{\mathbf{E}[X^k]}{a^k}.\end{aligned}$$

Now, the above expression is simply averaging over the moment bounds where you let the moment be distributed as a Poisson random variable with parameter as, and in general, the minimum over the moment bounds will be smaller than the average (no matter how the averaging is done, and so minimizing over s doesn't change anything), and thus we get the result.

9 Lecture 9: Convergence, Weak and Strong Law of Large Numbers, Central Limit Theorem

Agenda:

1. Recap of Limit Theorems (Chernoff)
2. Laws of Large Numbers (WLLN, convergence in probability)
3. Central Limit Theorem

9.1 Recap of Bounds

Example 9.1. Let $X \sim \mathcal{N}(0,1)$. We can bound the tail probabilities of X using the chernoff bound:

$$\mathbf{Pr}(X \geq k) \leq \frac{\mathbf{E}[e^{sX}]}{e^{sk}}$$

Recalling the MGF of a normal distribution, we have:

$$= \frac{e^{s^2/2}}{e^{sk}} = e^{s^2/2 - sk}$$

Minimizing this expression over $s > 0$ corresponds to minimizing the exponent. Taking the derivative, we see

$$-k + s^* = 0 \Rightarrow s^* = k$$

Plugging this optimal value s^* in, we get:

$$\mathbf{Pr}(X \geq k) \leq e^{-k^2/2}$$

Which is actually exponential decreasing, which is much closer to the true behavior of the normal distribution.

Exercise 9.2. Extend the above exercise to show that for $X \sim \mathcal{N}(0,1)$, we have

$$\mathbf{Pr}(|X| \geq k) \leq 2e^{-k^2/2}$$

Definition 9.3 ((Weak) Law of Large Numbers (WLLN))**.** If we perform an experiment n times independently and

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- If X_i has mean μ and variance σ^2
- $\mathbf{E}[M_n] = \mathbf{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\mathbf{E}[X_i]n = \mu$
- $$\mathbf{Var}(M_n) = \frac{1}{n^2}\sum \mathbf{Var}(X_i) = \frac{\sigma^2}{n}$$

This tells us that if $X_1, ..., X_n$ are i.i.d. RV's with mean μ and finite variance, then for every $\epsilon > 0$, we have

$$\mathbf{Pr}(|M_n - \mu| \geq \epsilon) \to 0$$

as $n \to \infty$

Proof. The proof of the last claim is quite simple and only uses Chebyshev's inequality. It tells us at what "rate" this probability goes to zero as $n \to \infty$. We have

$$\mathbf{Pr}(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n} \to 0 \quad \text{as } n \to \infty$$

□

What does the WLLN tell us? It tells us that

$$\lim_{n\to\infty} \mathbf{Pr}(|M_n - \mu| \geq \epsilon) = 0$$

Remark 9.4. Similar to the definition of the limit. For any $\epsilon > 0$, $\delta > 0$, there exists some $n_0(\epsilon, \delta)$ such that

$$\mathbf{Pr}(|M_n - \mu| \geq \epsilon) \leq \delta)$$

for all $n > n_0(\epsilon, \delta)$. We say then that M_n **converges in probability** to μ.

Example 9.5. Let $Y_n = \min(X_1, ..., X_n)$ for $X_i \sim U[0, 1]$. We have that

$$\mathbf{Pr}(|Y_n - 0| \geq \epsilon) = \mathbf{Pr}(|X_1| > \epsilon, |X_2| \geq \epsilon, ..., |X_n > \epsilon) = (1 - \epsilon)^n$$

Which goes to zero for all $\epsilon > 0$ as $n \to \infty$. This tells us that Y_n converges in probability to 0.

Example 9.6. Suppose we have an arrival process where we divide the number line into exponentially increasing sized intervals:

$$I_k = \{2^k, 2^k + 1, ..., 2^{k+1} - 1\}$$

And suppose we have exactly one arrival in each interval. So we let $Y_n = 1$ if there is an arrival at time n, and $Y_n = 0$ if there is no arrivals. We then have that

$$\mathbf{Pr}(Y_1 = 1) = 1$$

$$\mathbf{Pr}(Y_2 = 1) = \mathbf{Pr}(Y_3 = 1) = 1/2$$

$$\mathbf{Pr}(Y_n = 1) = \frac{1}{2^k} \quad \text{if } n \in I_k$$

This implies that

$$\lim_{n\to\infty} \mathbf{Pr}(|Y_n - 0| \geq \epsilon) = \lim_{n\to\infty} \mathbf{Pr}(Y_n = 1) = \lim_{k\to\infty} \frac{1}{2^k} = 0$$

Which tells us that Y_n converges in probability to 0.

The above highlights the weakness of convergence of probability. We can see of course that for any finite n, there are certainly an infinite number of 1's (arrivals) after n, yet it still converges in probability. This is fixed by something known as **almost sure** convergence, which we will not get deep into in this course.

Question: What happens to $S_n = \sum_{i=1}^n X_i$. This is just a bunch of convolutions! In particular, if each $X_i \sim U[0, 1]$, we know that convolving two uniform pdfs looks like a triangle pdf. Convolving yet again gives us a quadratic polynomial. Each time we convolve the width gets higher (the variance blows up) and the order of the polynomial becomes larger. This general phenomenon happens for non-uniform iid RVs (amazingly) as well!

Our problem is that the mean and variance of S_n both blow up as $n \to \infty$. To fix this, we define

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

which we can verify has zero mean and unit variance.

Theorem 9.7 (Central Limit Theorem). *The CLT says that*

$$\lim \mathbf{Pr}(\widehat{S_n} \leq x) = \Phi(x)$$

where $\Phi(x)$ is the CDF of the standard normal distribution! This type of convergence is known as convergence in distribution.

Proof. We present a sketch of the proof. Note that $S_n \to \mathcal{N}(0,1)$ implies that $S_n \to \mathcal{N}(n\mu, n\sigma^2)$. □

Exercise 9.8. See **Sinho's notes** on modes of convergence (I will hopefully type up my own sometime soon, but these are very good)

Remark 9.9 (SLLN vs WLLN). The WLLN, as we already discussed, says that

$$\mathbf{Pr}(|M_n - \mathbf{E}[X_i]| \geq a) \leq \frac{\sigma}{na^2}$$

Which tells us that

$$\lim_{n\to\infty} \mathbf{Pr}(|M_n - \mathbf{E}[X_i]| \geq a) = 0$$

The Strong Law of Large Numbers, on the other hand, says something stronger. It says that:

$$\mathbf{Pr}(\lim_{n\to\infty} M_n = \mu) = 1$$

On the surface, these look similar. But the key difference is that for some $\epsilon > 0$, the SLLN says that $|M_n - \mu| > \epsilon$ will only happen a *finite* number of times (in other words, there exists some N such that $n > N \Rightarrow |M_n - \mu| < \epsilon$). On the other hand, the WLLN makes no such guarantee. More specifically, the WLLN says that M_n converges *in probability*, while the SLLN says M_n converges *almost surely* or *with probability one*. For more details on the difference between these two things, you should refer to Sinho's notes or the course notes.

10 Lecture 10: Information Theory

Agenda:

1. Recap of WLLN
2. Proof of CLT
3. Introduction to Information Theory (Entropy, Compression)

10.1 Proof of CLT

Recall the CLT:

Theorem 10.1 (Central Limit Theorem). *The CLT says that*

$$\lim \mathbf{Pr}(\widehat{S_n} \leq x) = \Phi(x)$$

where $\Phi(x)$ is the CDF of the standard normal distribution! This type of convergence is known as convergence in distribution.

where we had defined $\widehat{S_n}$ as

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

Proof. Let

$$Z_n = \frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}$$

where each X_i is iid and $\mathbf{E}[X_i] = 0$ and $\mathbf{Var}(X_i) = 0$. We also note that if $Y \sim \mathcal{N}(0,1)$, then $M_Y(s) = e^{s^2/2}$ and furthermore $\log M_Y(s) = s^2/2$. So it suffices to show that the log of the MGF of Z_n is $s^2/2$. We have

$$\begin{aligned} M_{Z_n}(s) &= \mathbf{E}[e^{sZ_n}] = \mathbf{E}[\exp(\frac{s}{\sqrt{n}}\sum_{i=1}^{n} X_i)] \\ &= \mathbf{E}[\exp(\frac{s}{\sqrt{n}}X_1)\cdots\exp(\frac{s}{\sqrt{n}}X_n)] \\ &= \mathbf{E}[\exp(\frac{s}{\sqrt{n}}X_1)]\cdots\mathbf{E}[\exp(\frac{s}{\sqrt{n}}X_n)] \\ &= \left[M_X(\frac{s}{\sqrt{n}})\right]^n \end{aligned}$$

Now, recall that $M_X(0) = 1$, and $M_X'(0) = 0$, and $M_X''(0) = 1$, by our assumptions and the properties of the MGF. Now we consider:

$$\lim_{n\to\infty} \log M_{Z_n}(s) = \lim_{n\to\infty}\left[n \log M_X(\frac{s}{\sqrt{n}})\right] = \lim_{n\to\infty}\left[\frac{\log M_X(\frac{s}{\sqrt{n}})}{\frac{1}{n}}\right]$$

Now, letting $y = \frac{1}{\sqrt{n}}$

$$= \lim_{y\to 0}\left[\frac{\log M_X(sy)}{y^2}\right]$$

Now notice that the limit of both the numerator and the denominator is zero, so we can use L'Hopital's rule!

$$= \lim_{y\to 0}\left[\frac{sM_X'(sy)}{2yM_X(sy)}\right]$$

The numerator and denominator once again both go to zero. L'Hopital again!

$$= \lim_{y \to 0} \left[\frac{s^2 M_X''(sy)}{2M_X(sy) + 2ysM_X'(sy)} \right] = \frac{s^2}{2}$$

□

Example 10.2 (Polling Example). Suppose we ask n randomly sampled voters if they support candidate X. So $X_i = 1$ if yes, and zero otherwise. Suppose we want a 95% confidence interval that $|M_n - p| < \epsilon$, where p is the true probability that each voter supports our candidate, and $M_n = \frac{1}{n}\sum X_i$ is the empirical mean. Well, Chebyshev tells us that

$$\mathbf{Pr}(|M_n - p| \geq a) \leq \frac{\mathbf{Var}(M_n)}{a^2}$$

But now we note that $\mathbf{Var}(X_i) = p(1-p) \leq 1/4$, which tells us that $\mathbf{Var}(M_n) = \frac{1}{n}\mathbf{Var}(X_i) \leq \frac{1}{4n}$. Now, suppose we want to know our p value to within 0.1 with probability at least 95%. Mathematically, we want:

$$\mathbf{Pr}(|M_n - p| \geq 0.1) \leq 0.05$$

and we know

$$\mathbf{Pr}(|M_n - p| \geq 0.1) \leq \frac{\mathbf{Var}(M_n)}{0.1^2} \leq \frac{1}{4n(0.01)}$$

Which implies that in order for us to obtain a 95% confidence interval, we need to set $n \geq 500$. If $a = 0.01$, then we would need $n \geq 50000$ for a 95% confidence interval!

Now, let's compare this with the CLT method. The CLT tells us that

$$\frac{M_n - \mathbf{E}[M_n]}{\sqrt{\mathbf{Var}(M_n)}} \to \mathcal{N}(0,1)$$

and we want

$$\mathbf{Pr}(|M_n - p| \geq 0.1) \leq 0.05$$
$$\iff \mathbf{Pr}(\frac{|M_n - p|}{\frac{1}{2\sqrt{n}}} \geq \frac{0.1}{1/(2\sqrt{n})}) \leq 0.05$$

But notice that the left hand side is roughly a standard normal. To get a 95% confidence interval for a normal distribution, we use the fact that we know 95% of the probability mass lies within 2 standard deviations, and in this case a standard deviation is 1. So we have

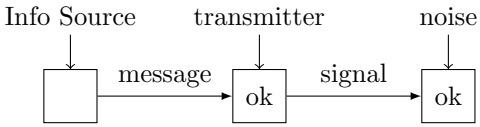
$$0.2\sqrt{n} \geq 2 \Rightarrow n \geq 100$$

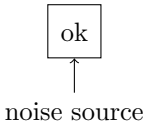
Which we can see is much better than the result Chebyshev gives us.

10.2 Intro to Info Theory

The field of information theory was pioneered by Claude Shannon in his seminal 1948 paper "A Mathematical Theory of Communication". There is a great textbook on the topic "Elements of Information" by Cover and Thomas, which is a highly recommended resource. Also, if you are interested in this topic further, you should take EECS 229A!

Shannon was concerned with the question, how much information can I reliably send over a noisy channel?





TODO 2. fix this diagram

There are two things we can concern ourselves with.

1. How much can we compress our information in the presence of no noise? This is known as the **Source Coding** problem.
2. How much information can we send in the presence of noise? This is known as the **Channel Coding Problem**

Shannon was able to answer both of these questions, and he was also even able to say that we can separately optimize for both of these criterion and arrive at a globally optimal solution!

11 Lecture 11: Info Theory, Binary Erasure Channel

Agenda:

1. Information theory overview (Entropy, AEP, Capacity of BEC)

Recall that there were two fundamental questions Shannon was exploring:

1. How much can we compress our information in the presence of no noise? This is known as the **Source Coding** problem.
2. How much information can we send in the presence of noise? This is known as the **Channel Coding** problem.

Theorem 11.1 (Source Coding Theorem). *Given N i.i.d. RV's $X_1, ..., X_n$, each having entropy $H(X)$, then these can be compressed with a source coding channel into no more than $N(H(X)+\epsilon)$ bits, $\forall \epsilon > 0$ as $N \to \infty$.*
Conversely, we also have that compression to fewer than $NH(X)$ bits is impossible without loss of information.

Definition 11.2 (Entropy). The entropy of a discrete RV X is defined as

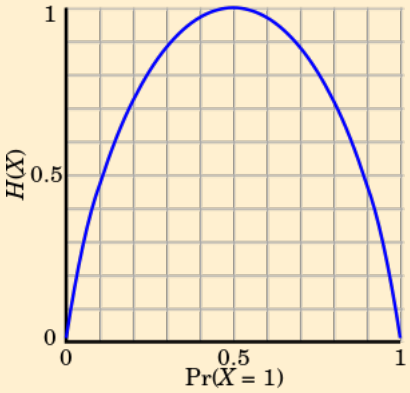
$$\begin{aligned} H(X) &:= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} \\ &= \mathbf{E}[\log \frac{1}{P_X(x)}]. \end{aligned}$$

We can interpret this definition very roughly by noting that the quantity $\log \frac{1}{P_X(x)}$ roughly corresponds to the "surprise" of seeing the outcome x. Then the entropy corresponds to the "average surprise" of our distribution. Another interpretation of entropy is that is correlated to the uncertainty of the random variable.

Example 11.3. When $X \sim Bern(p)$, then

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} =: H(p)$$

We can graph this quantity as p varies from 0 to 1. Note that at 0 and 1, the quantity is 0, and at $p = 1/2$, the quantity is 1. We also can calculate that $H(0.11) = 1/2$. This tells us that if we have a really long sequence of $Bern(0.11)$ RV's, then roughly half of the bits are "redundant", i.e. they can be compressed.



We further can naturally define

1.

$$H(X,Y) = \sum_{x,y} P_{X,Y}(x,y) \log \frac{1}{P_{X,Y}(x,y)}$$

Exercise 11.4. Show that

$$H(X,Y) = H(X) + H(Y|X)$$

where

$$\begin{aligned} H(Y|X) &:= \sum_{x\in\mathcal{X}} p(x)\, H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\ \log\, p(y|x) \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\ \log\, p(y|x) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log\, p(y|x) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)} \\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log \frac{p(x)}{p(x,y)}. \end{aligned}$$

Example 11.5. We now consider a motivating example for the AEP. Suppose we flip a coin n times independently. What is a "typical" sequence? Well, there are 2^n total sequences, but a "typical" sequence has np heads and $n(1-p)$ tails. The probability of a particular "typical sequence" S is:

$$\begin{aligned} P(S) &= p^{np}(1-p)^{n(1-p)} \\ &= 2^{np\log p} 2^{n(1-p)\log(1-p)} \\ &= 2^{n(p\log p + (1-p)\log(1-p)} = 2^{-nH(p)}. \end{aligned}$$

Our next question is then, how many such typical sequences are there? Well, there are exactly $\binom{n}{np}$, which it turns out is approximately $2^{nH(p)}$ for large n! How do we know this? Well it uses Stirling's approximation, and we won't go into detail here, but the first steps look something like this:

$$\binom{n}{np} = \frac{n!}{(np)!(n(1-p))!}$$

and we use the fact that $n! \approx \left(\frac{n}{e}\right)^n$.

What does this example tells us? Well, we have $2^{nH(p)}$ sequences, and all of these sequences occur with probability $2^{-nH(p)}$. This means virtually all of the probability must be used up by these "typical sequences"! This is known as the **Asymptotic Equipartition Property**, and is really quite a mind-boggling phenomenon, which is hopfully illustrated by this following example:

Example 11.6. Suppose our sequence of RVs are iid $Bern(0.11)$, and we are sending sequences of $n = 1000$ of these bits. We know that $H(p) = 0.5$. This tells us that our "typical set" is composed of the set of approximately 2^{500} sequences containing roughly $1000 \cdot 0.11$ 1's and $1000 \cdot 0.89$ 0's, and each of the sequences in this typical set have roughly equal probability. Then the source coding theorem

tells us we can transmit these sequences of 1000 bits with on average only around 500 bits!

How could we achieve this in practice? This is a very difficult question, and one that information theorists do not typically concern themselves with. We can, however, consider the following computationally infeasible scheme:

1. put each of the 2^{500} "typical" sequences into a lookup table with 2^{500} entries
2. if the input sequence is in the typical set, simply send a "0" following by the bit string that is the index of the typical sequence in the lookup table. The decoder can just look up the typical sequence in his copy of the lookup table when he receives the compressed message.
3. if the sequence is not "typical", just send a "1" followed by the whole sequence. This happens with probability that goes to zero as $n \to \infty$.

The bit at the beginning is simply to let the receiver know whether to look in the lookup table or to just look at the next 1000 bits. This scheme is entirely infeasible because we cannot store 2^{500} size lookup table in our computer, and much research in the last 50 years has been devoted to achieving the source coding theorem in practice.

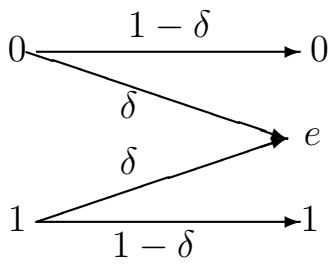
Theorem 11.7 (AEP). *We now formalize the Asymptotic Equipartition Property. If $X_1, ..., X_n$ are i.i.d. $\sim P_X(x)$, then*

$$-\frac{1}{n}\log P(X_1, ..., X_n) \longrightarrow H(X)$$

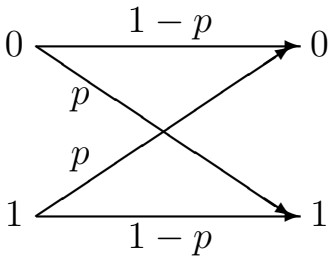
in probability as $n \to \infty$

11.1 Capacity of BEC

We have a **Binary Erasure Channel** looks like this:



Where this models the "noise" of a channel which takes a bit and erases it (maps it to e) with some probability δ. We also have a **Binary Symmetric Channel**, which looks like this:



As you can see, it flips each bit independently with probability p.

For the rest of this lecture, we will focus on the Binary Erasure Channel. Intuitively, we note that a binary erasure channel should have a higher capacity, which we will define shortly, than a BSC, because a BEC tells you exactly which bits have been corrupted.

Definition 11.8 (Capacity)**.** We say that the **capacity** of a channel is the maximum rate of reliable communication for that channel. Mathematically,

$$Rate = R = \frac{L_n}{n}$$

where L_n is the length of your message, and n is the length of your encoding.

Say m is the message your encoder receives (so $m \in \{0,1\}^{L_n}$), and at the end your decoder outputs a guess $\hat{m}$. We would like to minimize the probability of error:

$$P_e^{(n)} = \max_m \mathbf{Pr}[m \neq \hat{m}]$$

We say that rate R is **achievable** for the channel if for every positive number n that is "long enough", there exists and encoder and decoder functions f_n and g_n respectively such that

$$P_e^{(n)} \to 0$$

as $n \to \infty$. The largest achievable rate R is called the **capacity** of our channel.

Theorem 11.9. *We have that the capacity of a BEC channel is*

$$C_{BEC(p)} = 1 - p$$

bits per channel use.

This is really a remarkable result (think about why!). We have to show two things:

1. The **converse**: We need to be able to show that it is not possible to achieve a rate of $1 - p + \epsilon$ for any $\epsilon > 0$.
2. **achievability**: We would like to show that there is actually a scheme (even if it is computationally infeasible) that achieves this $1 - p$ rate

The proof of the converse goes as follows: Suppose there is a genie which is actually helping you encode and decode your message by *telling you in advance* exactly which bits will be erased. We can show that even with this help, we cannot achieve a capacity better than $1 - p$ as $n \to \infty$.

12 Lecture 12: Wrapup of Info Theory

Agenda:

1. Info theory wrapup (Capacity of BEC, converse and achievability)
2. Quick Note on Huffman Codes
3. Markov Chain Intro

Recall the setup of the BEC, which erases each bit independently with probability p.

$$m \in \{0,1\}^{L_n} \xrightarrow{encoder} X^{(n)} \xrightarrow{channel} Y^{(n)} \xrightarrow{decoder} \hat{m}$$

Suppose we have an input of L_n bits into our channel. We encode our L_n bits into a sequence of n bits, where $n \geq L_n$ (to account for the noise of the channel). Then we have the rate of our channel is

$$Rate = R = \frac{L_n}{n} \quad \text{bits per channel use}$$

and the capacity C is simply the maximum rate which we can *reliably* communicate (i.e. not lose any information with high probability). We say that the probability of error is:

$$P_e^{(n)} = \max \mathbf{Pr}(\hat{m} \neq m)$$

We would like to more thoroughly prove the theorem from last time, namely that

$$C_{BEC} = 1 - p$$

As we mentioned last time, we need to show two things:

1. The **converse**: We need to be able to show that it is not possible to achieve a rate of $1 - p + \epsilon$ for any $\epsilon > 0$.
2. **achievability**: We would like to show that there is actually a scheme (even if it is computationally infeasible) that achieves any rate up to $1 - p$

Proof.

1. **Converse:** The idea behind the converse is to have a genie tell you exactly which bits will be erased *beforehand*. Even with this information, you cannot achieve a rate better than $1-p$ bits/channel use of reliable communication.

2. **Achievability:** We would like to show that we can achieve a rate $R = 1 - p - \epsilon$ for any $\epsilon > 0$. Shannon's insight was that we can leverage the SLLN to do this! By the SLLN, the probability that the channel erases exactly np of the n input symbols is exactly 1 as $n \to \infty$. Shannon's idea was then to create a *massive* lookup table. Each row of the table corresponded to an input, so there are 2^L rows. There were then n columns, so each input corresponded to an n bit string, which is represented by $X^{(n)}$ in our diagram. Then how was this table populated? Shannon's idea was to populate the table with iid $Bern(1/2)$ coin flips! This lookup table is called the **Codebook** $\mathcal{C}$. Each row c_i corresponds to a "codeword" corresponding to a specific input message m.

   WLOG, we can assume that the BEC channel erases the last np bits (we know it erases almost exactly np random bits, so we might as well assume they all come at the end). Then we can just consider $\mathcal{C}'$, which is a truncated codebook with all np bits shoved to the end of each codeword (so the entire right half of the original codebook is now just erasures).

   Now, how should the receiver decode? The decoder simply consults his own codebook (he has his own copy), to see which one of the 2^L codewords matches on the $n(1-p)$ bits that were sent. **When do we get an error**? We get an error when *more than one codeword is consistent with the* $n(1-p)$ *bits the decoder receives from the channel.*

Analysis: We can assume further WLOG that message 1 was sent. We have then that

$$\begin{aligned} P(\text{error}) &= \mathbf{Pr}(c_1' \text{ is not unique}) \\ &= \mathbf{Pr}\left(\bigcup_{i=2}^{2^L} \{c_i' = c_1'\}\right) \\ &\leq \sum_{i=2}^{2^L} 2^{-n(1-p)} \\ &\leq 2^L 2^{-n(1-p)} \\ &= 2^{nR-n(1-p)} = 2^{n(R-(1-p))}, \end{aligned}$$

which goes to zero if $R < 1 - p$. In particular, if $R = 1 - p - \epsilon$, then

$$P_e^{(n)} \leq 2^{-n\epsilon} \xrightarrow{n\to\infty} 0,$$

exponentially fast!

□

Example 12.1. If $n = 10000$, $p = 0.5$, $\epsilon = 0.01$. We know that

$$C_{BEC(1/2)} = \frac{1}{2}$$

which implies our capacity is 5000 bits. But we back off, we have $L = 10000(1 - 0.5 - 0.01) = 4900$. Then we have

$$P_e \leq 2^{-10000*0.01} = 2^{-100}$$

which is extremely small.

Here is a theorem we won't prove:

Theorem 12.2. *The capacity of a BSC channel is*

$$C_{BSC(p)} = 1 - H(p)$$

This should make sense, as we get a capacity of zero when $p = 1/2$, and a capacity of 1 when $p = 0$ or $p = 1$. In general, we have the following theorem from Shannon:

Theorem 12.3. *For a general Discrete Memoryless Channel (DMC) which has a conditional probability $P(Y|X)$, we have that:*

$$C = \max_{P(X)} I(X;Y) = \max_{P(X)} H(X) - H(X|Y)$$

Intuitively, the *mutual information* $I(X;Y)$ (which you explore a bit more in discussion and homework), tells us how much information you learn about X being given Y. Then of course we should maximize this in order to be able to transmit the most information through our channel!

12.1 Huffman Coding

Suppose we have a an alphabet $X \in \{A, B, C, D\}$ where $P_A = 0.4$, $P_B = 0.35$, $P_C = 0.2$, and $P_D = 0.05$. We can calculate that

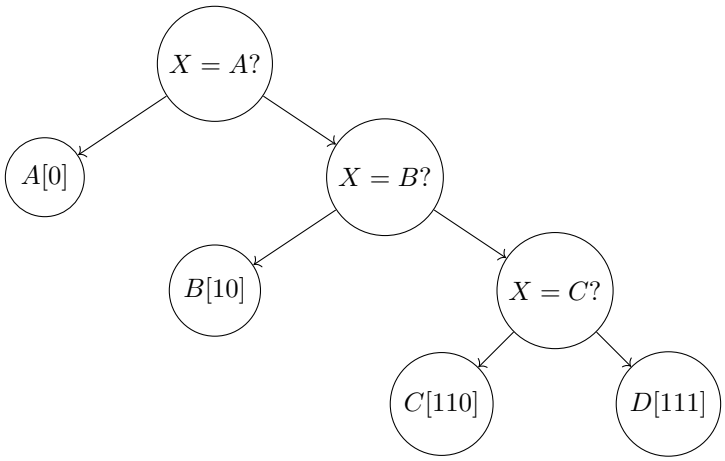
$$H(X) = 1.74$$

Then Huffman's algorithm to create an encoding for this alphabet is as follows:

1. Remove two members of our alphabet with the smallest probabilities, and assign them a the bits 0 and 1 respectively. Then, add their probabilities, concatenate the letters, and add the combined letters back into the alphabet.

2. keep doing this until there is only one giant combined member of our alphabet (which will have probability 1).

After running this algorithm, we can read off the encodings by running backward through the binary tree we created when running the algorithm.



The result of running this algorithm on our example (work this out yourself!) gives

$$A \to 0 \quad B \to 10 \quad C \to 110 \quad D \to 111$$

We can further calculate that the expected number of bits we have to use is 1.85. Notice that this is not quite optimal in that it does not achieve the entropy. This is because we have to work with integers, whereas the entropy does not. We can show that Huffman coding is actually optimal when each member of our alphabet has a probability of the form $p = \frac{1}{2^i}$.

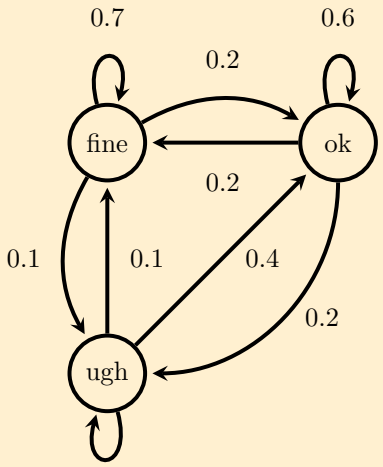
13 Lecture 13: Markov Chains

Definition 13.1 (Markov Chain)**.** A Markov chain is a sequence of random variables $X_0, X_1, X_2, ...$ satisfying the following 3 conditions:

1. The assumption that
$$Pr(X_{n+1} = c_{n+1} | X_0 = c_0, ..., X_n = c_n) = Pr(X_{n+1} = c_{n+1} | X_n = c_n)$$
2. $X_0, X_1, ...$ take on values from some set S
3. X_0 is an arbitrary pmf on S.

For a homogeneous discrete time Markov chain, we say $Pr(X_{n+1} = i | X_n = j) = P_{ji}$

Example 13.2. Here is an example of a markov chain, represented with a diagram. It represents the three fundamental states of any Berkeley student.



We can specify what our initial state X_0 is, and answer questions such as what is the $Pr(X_5 = ugh)$? Over 16 weeks, what fraction of time are you ok?

Example 13.3 (PageRank)**.** PageRank is google's algorithm for returning search results. It is now much more complicated, but at its core it uses Markov chains to determine how popular each website on the internet is. There are a few ways to formulate this notion:

1. Score each page i with π_i, such that
$$\pi_i = \sum_j \pi_j P_{ji}$$
$$\sum_i \pi_i = 1$$
The first equation is called a balance equation. More on that later.
2. bot randomly picks link on each page it visits. π_i is the equal to the probability that the bot is on page i at some point in time $t \gg 0$.

3. π_i = fraction of time bot spends on page i.

All three of these formulations are equivalent.

Let's start working our way more towards these balance equations. First, we are interested in what happens to finite states as $n \to \infty$. We define $r_{ij}(n)$ as the probability of going from state i to state j in n time steps. Well, $r_{ij}(1) = P_{ij}$, since there is only one way to get from i to j in one time step. $r_{ij}(n)$ is more complicated, but luckily we can actually write it in terms of $r_{ij}(n-1)$ as follows:

$$r_{ij}(n) = \sum_{k \in \mathcal{S}} r_{ik}(n-1) P_{kj}$$

The above are a form of the *Chapman Kolmogorov Equations*. They should intuitively make sense: the only way to get from i to j in n steps is if you first get to somewhere else in $n-1$ steps, and then make the last step to state j. We are just summing over all the possible places you could be at time step $n-1$. Lets examine $r_{ij}(2)$. We have that

$$r_{ij}(2) = \sum_{k \in \mathcal{S}} r_{ik}(1) P_{kj} = \sum_{k \in \mathcal{S}} P_{ik} P_{kj}$$

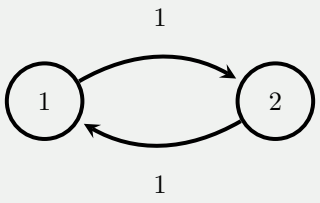
$$= \begin{bmatrix} P_{i1} & P_{i2} & \cdots & P_{im} \end{bmatrix} \begin{bmatrix} P_{1j} \\ P_{2j} \\ \vdots \\ P_{mj} \end{bmatrix}$$

Now further recall from CS70 our transition probability matrix:

$$\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ P_{m1} & P_{m2} & \cdots & P_{mm} \end{bmatrix}$$

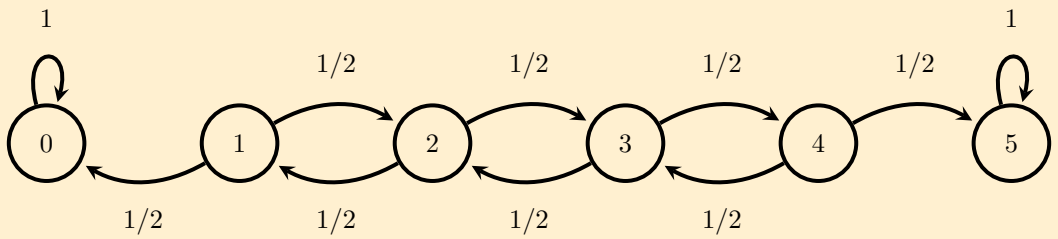
We can then see that $r_{ij}(2) = (P^2)_{ij}$. This is quite convenient! It is also very easy to then see that $r_{ij}(n) = (P^n)_{ij}$, or the $(i,j)^{th}$ entry of P^n. If the values of each column of P^n converge to the same value, then this tells us that no matter where we start out, you have an equal probability of ending up in a given state.

Remark 13.4. Consider the following Markov chain:



It is easy to see that we will "ping pong" infinitely back and forth, and it is entirely deterministic which state we are in at any given time (given that we know where we started).

Example 13.5 (2 spiders, 1 fly)**.** Consider the following Markov chain:



We will intuitively always end up either at state 0 or state 5, and it is much more likely that we get stuck at 5 if we start in state 4 than if we start in state 1. So once again, $Pr(X_n = i)$ is **not** always independent of where we start. In which situations is it? Stay tuned...

We first note that a more concise way to write (1) in the PageRank example would be $\pi P = \pi$ where $\sum_i \pi_i = 1$. Such a π satisfying these two equations is called the **stationary distribution** of a markov chain. We further note the following definitions:

Definition 13.6. State i is **recurrent** if, starting in state $X_0 = i$, the chain will revisit i at some point with probability one. Furthermore, we say that a state is **positive recurrent** if $\mathbf{E}[T_i] < \infty$, where T_i is the time to return to state i after leaving it. Otherwise if $\mathbf{E}[T_i] = \infty$ then the state is called **null recurrent** (provided the probability we return to state i is still 1). If a state is not recurrent then it is called **transient**.

Definition 13.7. The **class** of a state i is $\{j : j \text{ accessible from } i \text{ and } i \text{ accessible from } j\}$

Proposition 13.8. *The states in a class are either all recurrent or all transient.*

Proof. Let i and j be in the same class and suppose towards a contradiction that i is recurrent while j is transient. Since there is a path from i to j, there must be some $n \geq 1$ such that $P_{ij}^n > 0$. Now, given that we start in state i, I will revisit i infinitely often by recurrence. It takes $Geom(P_{ij}^n)$ visits to i before I will successfully land in state j n steps later. Hence, j is recurrent since geometric RVs are almost surely finite. □

Definition 13.9. Consider $s_i = \{n : r_{ii}(n) > 0\}$. Then we define the **periodicity** of a state as $GCD(s_i)$. In english, the periodicity of a state is the GCD of the all the possible times we could return to that state. In the "ping pong" example, both states have a periodicity of 2. If a state has a self loop, then its periodicity is trivially one.

Proposition 13.10. *All the states in a class have the same period.*

Proof. We start by denoting $d(s)$ as the period of state s. Once again, consider i and j in a communicating class together. We know i and j are accessible from each other, so WLOG consider a path of length n from i to j and a path of length m from j to i. Then there is a path of $n + m$ from i to i, so it follows from the definition of the period that $n + m$ is divisible by $d(i)$. Consider any path from j to j. Say it has length t. This creates yet another path from i to i of length $n + t + m$ (first go from i to j, then j to j, then back to i). B the same logic, we have that $n + m + t$ is divisible by $d(i)$. This implies that t is divisible by $d(i)$, for all t such that there is a path of length t from j back to j. Since this holds for all t, this means that $d(i)$ is

a factor of $\{n : r_{jj}(n) > 0\}$, and by definition it is less than or equal to the greatest common factor, $d(j)$[1]. Reversing the roles of i and j in the above argument implies that $d(j) \leq d(i)$, which implies that $d(j) = d(i)$, as desired. □

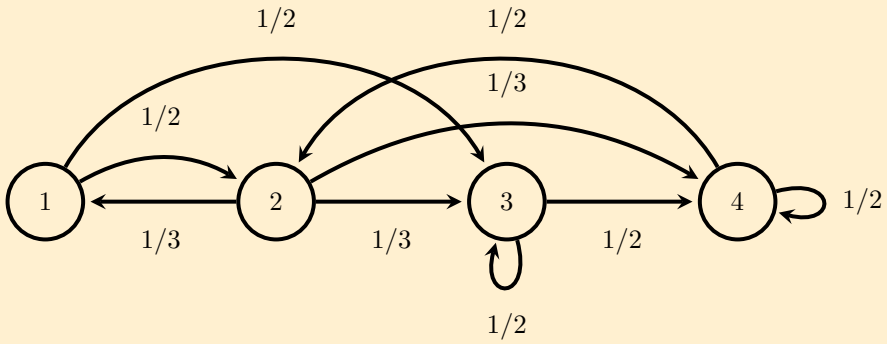
If the above proof was confusing, it is very helpful to draw it out!

In general, any MC with a single aperiodic recurrent class (and some transients) must converge in the following sense:

1. for each state j,
$$\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j, \forall i$$
2. The π_j are given by a system of equations:
$$\pi_j = \sum_{k=1}^{m} \pi_k P_{kj} \qquad \sum_i \pi_i = 1$$
3. $\pi_i = 0$ if state i is transient, and $\pi_i > 0$ if i is recurrent.

Now, what if we wanted to find the expected amount of time to get from one state to another, given that we are in stationarity? We will now develop a tool known as **first-step equations** to deal specifically with this omnipresent problem. It is actually easiest to see with an example.

Example 13.11. Consider the MC below.



We further define x_i as the expected amount of steps we must take to reach a certain special state, say 1 in this case. Then trivially we can observe that $x_1 = 0$. What about x_2? Well, with 1/3 probability, we are done, but we could also could go to states 3 and 4. By splitting up into cases, we can see

$$\begin{aligned} x_2 &= 1 + \mathbf{Pr}(\text{we go to 1})\,\mathbf{E}[\text{time to 1 from 1}] \\ &\quad + \mathbf{Pr}(\text{we go to 2})\,\mathbf{E}[\text{time to 1 from 2}] \\ &\quad + \mathbf{Pr}(\text{we go to 3})\,\mathbf{E}[\text{time to 1 from 3}] \\ &\quad + \mathbf{Pr}(\text{we go to 4})\,\mathbf{E}[\text{time to 1 from 4}] \\ &= 1 + 1/3 \cdot 0 + 0 \cdot x_2 + (1/3)x_3 + (1/3)x_4 \end{aligned}$$

Note that we include the 1+ since we need to take at least one step no matter what happens. Using similar logic, we can come up with the following system of equations for the other nodes:

$$x_3 = 1 + (1/2)x_3 + (1/2)x_4$$

$$x_4 = 1 + (1/2)x_2 + (1/2)x_4$$

Which leaves us with three equations and three unknowns, which means we can solve for each x_i.

[1] We can even claim that $d(j)$ is divisible by $d(i)$, but why overcomplicate things?

In general, we can use the same idea to define a **mean recurrence time** t_s^* = the average number of steps the MC takes to return to state s. Then we have

$$t_s^* = 1 + \sum_{i=1}^{m} P_{si} t_i$$

where t_i is of course the expected amount of time to get from state i to state s.

Example 13.12. We can consider the same Markov chain from the previous example, but a more general hitting time problem. Given sets A and B such that $A \cap B = \emptyset$, we want to find the probability that we reach a node in set A before we reach a node in set B. Using the MC from the previous example, we can let $A = \{1\}$ and $B = \{4\}$. Then now we can define x_i = probability that we reach A before B given we start in state i. Then trivially, $x_1 = 1$ and $x_4 = 0$. We also have by splitting into cases.

$$\begin{aligned} x_2 &= \mathbf{Pr}(\text{we go to 1}) \cdot \mathbf{Pr}(\text{we get to } A \text{ first given go to 1}) \\ &\quad + \mathbf{Pr}(\text{we go to 2}) \cdot \mathbf{Pr}(\text{we get to } A \text{ first given go to 2}) \\ &\quad + \mathbf{Pr}(\text{we go to 3}) \cdot \mathbf{Pr}(\text{we get to } A \text{ first given go to 3}) \\ &\quad + \mathbf{Pr}(\text{we go to 4}) \cdot \mathbf{Pr}(\text{we get to } A \text{ first given go to 4}) \\ &= 1/3 \cdot 1 + 0 \cdot x_2 + 1/3 \cdot x_3 + 1/3 \cdot 0 \end{aligned}$$

Similarly, we can formulate another equation for x_3, which would allow us to solve for our two unknowns.

Remark 13.13. There is an inherent connection to the material we have (or will) learned about in CS188. We can think about collecting a reward R_i every time to MC is in state i, and then we can further define r_i as the expected reward we get starting from state i until we reach some set A. Then we have

$$r_i = R_i \quad \forall i \in A$$

$$r_i = R_i + \sum_j P_{ij} r_j \quad \forall i \notin A$$

We can also think about adding in a "discount factor" so if $X(n) = i$ then we receive reward $\beta^n R_i$, where β is the discount factor. Then similarly we have:

$$r_i = R_i \quad \forall i \in A$$

$$r_i = R_i + \beta \sum_j P_{ij} r_j \quad \forall i \notin A$$

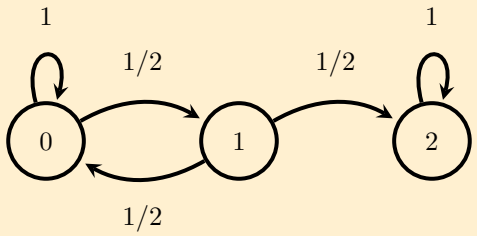
Exercise 13.14. Suppose Alice commutes between 2 houses every week. If the weather is great (this happens with probability p), she grabs her fishing rod and fishes on her way to the other house. There are N fishing rods. Assuming this has been going on for a very long time already, what is the probability that she has no rods when the weather is good? *Hint:* Set up a MC with $N + 1$ states.

14 Lecture 14: More Markov Chains

14.1 First-step equation modeling

Example 14.1. What is the expected number of tosses until you see two consecutive heads in a row (assuming the coin is fair)?

We can model this with a Markov chain with three states: the state of having seen zero heads,



one head, and two heads.

Then, we define

$$T_2 = \min\{n \geq 0 | X_n = 2\}$$

and then

$$\beta(i) = \mathbf{E}[T_2 | X_0 = i]$$

clearly, $\beta(2) = 0$. We also have that

$$\beta(0) = 1 + \frac{1}{2}\beta(0) + \frac{1}{2}\beta(1)$$

$$\beta(1) = 1 + \frac{1}{2}\beta(0) + \frac{1}{2}\beta(2)$$

We can solve these equations, since there are three variables and three unknowns (try it!)

In general, for **First Step Equations**, for a MC over state space $\mathcal{X} = \{1, 2, ..., N\}$ and $A \subset \mathcal{X}$ we let

$$T_A = \min\{n \geq 0 | X_n \in A\}$$

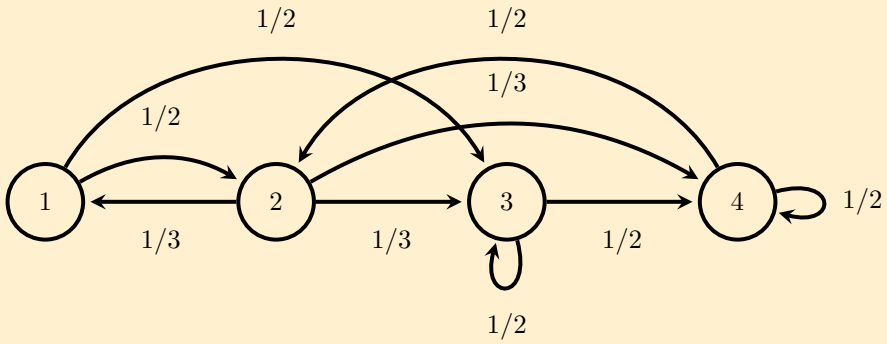
and then the expected hitting time from each node as

$$\beta_A(i) = \mathbf{E}[T_A | X_0 = i] \quad \forall i \in \mathcal{X}$$

and note we have that

$$\beta_A(i) = 0 \quad \forall i \in A$$

Example 14.2. Consider the following MC.



Considering the above MC chain, a question we could possibly ask is: what is the probability that I hit state 1 before I hit state 4? This is obviously dependent on what state we start in. In a very similar manner to the first step equations, we can define some clever variables and solve a system of equations. We let

$$\alpha(i) = \mathbf{Pr}[T_1 < T_4 | X_0 = i]$$

Which is simply the probability that we hit state 1 before we hit state 4 given that we start in state i. Then we can note immediately that

$$\alpha(1) = 1, \quad \alpha(4) = 0$$

and we can also set up the relations:

$$\begin{aligned}\alpha(2) &= \frac{1}{3}\alpha(1) + \frac{1}{3}\alpha(3) + \frac{1}{3}\alpha(4) \\ &= \frac{1}{3} + \frac{1}{3}\alpha(3) \\ \alpha(3) &= \frac{1}{2}\alpha(3) + \frac{1}{2}\alpha(4)\end{aligned}$$

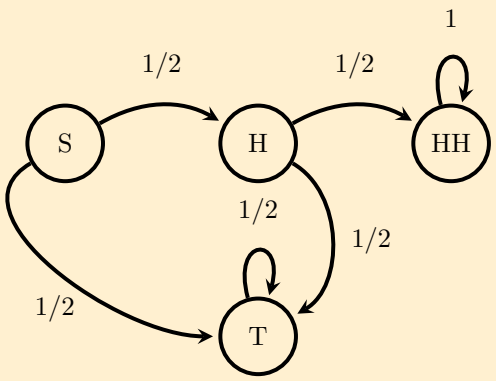
and so

$$\begin{aligned}\alpha(3) = \alpha(4) &= 0 \\ \alpha(2) = \frac{1}{3} + \frac{1}{3}\alpha(3) &= \frac{1}{3},\end{aligned}$$

which gives the final result $\alpha(i) = \begin{bmatrix}1 & 1/3 & 0 & 0\end{bmatrix}$.

Example 14.3. We once again flip a fair coin until we see two consecutive heads. Whats the expected number of tails I see?

We can model this via a Markov chain seen below, with a reward function g:



Then we have that $g(S) = g(H) = g(HH) = 0$ and $g(T) = 1$. And we can define $\gamma(s)$ as the expected total reward given that we start in state s. We have that

$$\gamma(S) = \frac{1}{2}\gamma(H) + \frac{1}{2}\gamma(T)$$

$$\gamma(H) = \frac{1}{2}\gamma(HH) + \frac{1}{2}\gamma(T)$$

$$\gamma(T) = 1 + \frac{1}{2}\gamma(T) + \frac{1}{2}\gamma(H)$$
$$\gamma(HH) = 0$$

we can solve this system of equations to get $\gamma(S) = 3$.

Theorem 14.4. *If MC is irreducible, aperiodic, and positive recurrent, then*

$$\lim_{n\to\infty} \mathbf{Pr}[X_n = j | X_0 = i] = \pi(j)$$

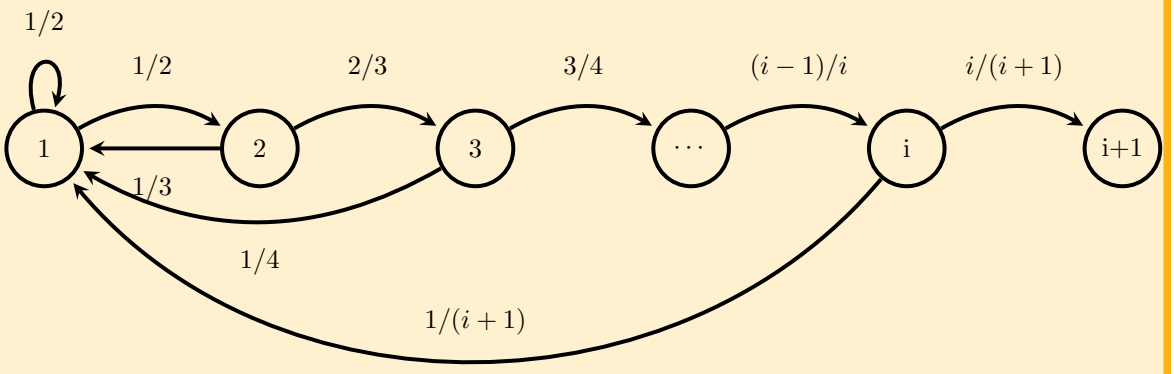
and the MC is said to be ***asymptotically stationary****.*

In general, in the finite setting, every irreducible MC has all its states being positive recurrent. On the other hand, in the infinite case, we can have all states being positive recurrent, null recurrent, or transient.

Theorem 14.5 (Big Thm for MCs)**.**

1. *A MC is either irreducible or reducible*
2. *if the MC is irreducible, then it is either transient (in which case no π exists), positive recurrent (a unique π exists), or null recurrent (no π exists).*
3. *furthermore, if the chain is positive recurrent, then it is either periodic or aperiodic. If it is aperiodic, then no matter where we start we always converge to π.*

Example 14.6 (Another null recurrent MC)**.** Consider



Is this MC transient or recurrent? Recall that being transient is the same as the probability of **not** returning to a state is greater than 0. We have

$$\mathbf{Pr}[\text{We do not return to state 1 at time n — start at state 1}] = \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{3}{4}\cdots\frac{n}{n+1} = \frac{1}{n+1}$$

This goes to zero as $n \to \infty$, so the chain is recurrent. But is it positive or null recurrent? We have that

$$\beta(1) = \mathbf{E}[T_1 | X_0 = 1]$$

14.2 Reversibility of MCs

Assume we have an irreducible and positive recurrent Markov Chain, initialized at its invariant distribution

π. The notion of *reversing* a chain is as follows: Suppose for every n, $(X_0, X_1, ..., X_n)$ has the same joint pmf as its "time reversed" version $(X_n, ..., X_0)$, then we call the chain **reversible**.

15 Lecture 15: Wrapup (reversible) Markov Chains, and beginning Poisson Processes

Agenda:

1. Reversible MCs
2. Poisson Processes

15.1 Reversible MCs

Consider an irreducible MC $\{X_n\}_{n=0}^{\infty}$ on the finite state space $\mathcal{X}$ with transition probability matrix P. The question is: When does a MC "look the same" whether it is run forward or backward? More formally, when does running the chain backwards give the same transition probabilities and invariant distribution? Here are some facts:

1. A MC run backward is always still a MC (needs proof)
2. If the MC is reversible, then the backward chain is the same MC as the forward chain.

Proof. 1. We have that

$$\begin{aligned}&\mathbf{Pr}[X_k = i_k | X_{k+1} = i_{k+1}, ..., X_{k+n} = i_{k+n}] \\ &= \frac{\mathbf{Pr}[X_k = i_k X_{k+1} = i_{k+1}, ..., X_{k+n} = i_{k+n}]}{\mathbf{Pr}[X_{k+1} = i_{k+1}, ..., X_{k+n} = i_{k+n}]}\end{aligned}$$

Now, we use the Markov Property, and I drop the i_j's for simplicity:

$$\begin{aligned}&\frac{\mathbf{Pr}[X_{k+n}|X_{n+k-1}]\,\mathbf{Pr}[X_{n+k-1}|X_{n+k-2}]\cdots\mathbf{Pr}[X_{k+1}|X_k]\,\mathbf{Pr}[X_k]}{\mathbf{Pr}[X_{k+n}|X_{n+k-1} = i_{n+k-1}]\,\mathbf{Pr}[X_{n+k-1}|X_{n+k-2}]\cdots\mathbf{Pr}[X_{k+2}|X_{k+1}]\,\mathbf{Pr}[X_k]} \\ &= \frac{\pi(i_k)P(i_k, i_{k+1})}{\pi(i_{k+1})}\end{aligned}$$

Notice that this is only a function of i_{k+1}, which shows that the backwards chain satisfies the Markov Property! Furthermore, if we denote $\tilde{P}$ as the transition probability matrix for the reversed chain, then we have that:

$$\tilde{P}_{i_{k+1},i_k} = \frac{\pi(i_k)P(i_k, i_{k+1})}{\pi(i_{k+1})}$$

So then, if our chain is reversible, we have the condition that

$$\begin{aligned}\tilde{P}_{i_{k+1},i_k} = P_{i_{k+1},i_k} &= \frac{\pi(i_k)P(i_k, i_{k+1})}{\pi(i_{k+1})} \\ \implies P_{i_{k+1},i_k} &= \frac{\pi(i_k)P(i_k, i_{k+1})}{\pi(i_{k+1})}.\end{aligned}$$

□

Theorem 15.1. *If a MC is reversible, it has an invariant distribution π*

Proof. We need to show that $\forall j, \quad \pi(j) = \sum_i \pi(i)P_{ij} \iff \pi = \pi P$. We have that:

$$\sum_i \pi(i)P_{ij} = \sum_i \pi(j)P_{ji} = \pi(j)\sum_i P_{ji} = \pi(j). \quad \square$$

Summary: Detailed balance equations being satisfied are *sufficient but not necessary* to have a stationary distribution.

Remark 15.2 (Sufficient condition for MC to be reversible)**. Fact:** Start with a graph associated with a MC, forget self loops and make all the arrows undirected. Then we have an undirected graph. If this resulting graph is a tree, then detailed balance equations hold. Note that the converse does not necessarily hold. Also note that this is not a necessary condition for detailed balanced equations to hold, only sufficient.

15.2 Poisson Processes

A Poisson Process is the continuous time analog of a "coin flip" or Bernoulli process. Some motivation:

1. Good model for arrivals of packets at a router, customers arriving at a cashier, photons at a detector, etc.

Definition 15.3 (Poisson Process)**.** We denote N_t as the total number of arrivals we have at any time t.

TODO 3. draw graph

We denote T_i as the time of the i^{th} arrival, and S_i as the *inter-arrival* times, so

$$S_i = T_i - T_{i-1}$$

and we define:

$$S_1, ..., S_n \sim_{iid} Expo(\lambda)$$

Formally, we define:

$$N_t = \begin{cases} \max_{n \geq 1}\{n | T_n \leq t\} & t \geq 0 \\ 0 & t < T_1 \end{cases}$$

Recall for an exponential distribution $\tau \sim Expo(\lambda)$, we have

1. $F_\tau(t) = 1 - e^{-\lambda t}$
2. $\mathbf{E}[\tau] = \frac{1}{\lambda}$
3. $\mathbf{Var}(\tau) = \frac{1}{\lambda^2}$
4. Memorylessness: $\mathbf{Pr}[\tau > t + s | \tau > s] = \mathbf{Pr}[\tau > t]$
5. $\mathbf{Pr}[\tau \leq t + \epsilon] | \tau > t] = \lambda\epsilon + o(\epsilon)$

Note that the "little-o" notation just refers to any function such that:

$$\lim_{\epsilon \to 0} \frac{o(\epsilon)}{\epsilon} = 0$$

Note that for example $\epsilon^2 \in o(\epsilon)$

proof of (5).

$$\mathbf{Pr}[\tau \geq t + \epsilon] | \tau > t] = \mathbf{Pr}[\tau > \epsilon] = e^{-\lambda\epsilon} = 1 - \lambda\epsilon + o(\epsilon) \quad \square$$

The above probability is the probability that there are no arrivals in a tiny ϵ amount of time. Then the probability of one arrival is approximately $\lambda\epsilon$, and the probability of more than one arrival in a tiny ϵ amount of time is $o(\epsilon)$, which we assume is essentially zero. So in a tiny amount of time, we only ever see one or zero arrivals. In Bertsekas, we start with this assumption and then derive that the interarrival times must be exponential, but we follow the Walrand book and assume that they are exponential and then derive this property.

Theorem 15.4. *Poisson Processes are memoryless.*

In pictures, if $N_t \sim PP(\lambda)$, then so is $(N_{t'+s} - N_t)$. The implication is that if we take increments of a poisson process that do not overlap in time, then these increments are **independent** and **stationary** (meaning if we shift in time, the statistics remain the same).

$\Rightarrow \forall 0 \leq t_1 < t_2 < ... < t_n \quad \{(N_{t_{n+1}} - N_{t_n})\}$are independent and distribution depends only on $(t_{n+1} - t_n)$

Proof. Straightforward from the memoryless property of the exponential. □

Theorem 15.5. *If $N = \{N_t | t \geq 0\}$ is a $PP(\lambda)$, then N_t = the number of arrivals in $(0, t)$ is distributed according to a poisson distribution with parameter λt:*

$$\mathbf{Pr}(N_t = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$

Proof. We start by finding the joint probability density of $T_1, ..., T_k, T_{k+1}$, where we know there are k arrivals in $(0, t)$. We have

$$\begin{aligned} f(t_1, ..., t_k)dt_1 \cdots dt_{k+1} &= \mathbf{Pr}[T_1 \in \{t_1, t_1 + dt_1\}, ..., T_k \in \{t_k, t_k + dt_k\}, T_{k+1} > t] \\ &= \mathbf{Pr}[S_1 \in \{t_1, t_1 + dt_1\}, S_2 \in \{t_2 - t_1, t_2 - t_1 + dt_2\}, ..., \\ &\qquad S_k \in \{t_k - t_{k-1}, t_k - t_{k-1} + dt_k\}, S_{k+1} > t - t_k] \\ &= \left(\lambda e^{-\lambda t_1} dt_1\right) \left(\lambda e^{-\lambda(t_2 - t_1)} dt_2\right) \cdots \left(\lambda e^{-\lambda(t_k - t_{k-1}} dt_k\right) e^{-\lambda(t - t_k)} \\ &= \lambda^k e^{-\lambda t} dt_1 dt_2 \cdots dt_k. \end{aligned}$$

Note that this joint distribution does not(!) depend on $t_1, ..., t_k$, which tells us that conditioned on the number of arrivals in an interval, those arrivals are *uniformly distributed* in the interval! Of course, we must have that they are in the correct order still. □

16 Lecture 16: Properties of Poisson Processes

Agenda:

1. Recap of Poisson Processes
2. Proof that number of arrivals in (0,T) is $\sim Pois(\lambda T)$.
3. examples
4. Merging and Splitting of PPs
5. Erlang Distribution
6. Random-Incidence-Paradox

Recall we have this theorem:

Theorem 16.1. *If $N = \{N_t | t \geq 0\}$ is a $PP(\lambda)$, then N_t = the number of arrivals in $(0,t)$ is distributed according to a Poisson distribution with parameter λt:*

$$\mathbf{Pr}(N_t = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$

and we had begun working through a proof:

Proof. We start by finding the joint probability density of $T_1, ..., T_k, T_{k+1}$, where we know there are k arrivals in $(0,t)$. We have

$$\begin{aligned}
f(t_1, \ldots, t_k)dt_1 \cdots dt_{k+1} &= \mathbf{Pr}[T_1 \in \{t_1, t_1 + dt_1\}, \ldots, T_k \in \{t_k, t_k + dt_k\}, T_{k+1} > t] \\
&= \mathbf{Pr}[S_1 \in \{t_1, t_1 + dt_1\}, S_2 \in \{t_2 - t_1, t_2 - t_1 + dt_2\}, \ldots, \\
&\quad S_k \in \{t_k - t_{k-1}, t_k - t_{k-1} + dt_k\}, S_{k+1} > t - t_k] \\
&= \left(\lambda e^{-\lambda t_1} dt_1\right) \left(\lambda e^{-\lambda(t_2 - t_1)} dt_2\right) \cdots \left(\lambda e^{-\lambda(t_k - t_{k-1}} dt_k\right) e^{-\lambda(t - t_k)} \\
&= \lambda^k e^{-\lambda t} dt_1 dt_2 \cdots dt_k \\
\implies f_{T_1,\ldots,T_k}(t1,...,t_k) &= \lambda^k e^{-\lambda t}
\end{aligned}$$

Note that this joint distribution does not depend on $t_1, ..., t_k$, which tells us that conditioned on the number of arrivals in an interval, those arrivals are *uniformly distributed* in the interval! Of course, we must have that they are in the correct order still. The way I like to think about this is simply dropping these arrivals randomly in the interval, and then assigning the arrivals to be in the correct order. Now, we have

$$\begin{aligned}
N_T(k) &= \int_{t_1} \int_{t_2} \cdots \int_{t_k} f_{T_1,\ldots,T_k}(t_1, ..., t_k)dt_1 \cdots dt_k \\
&= \lambda^k e^{-\lambda t} \int_0^t \cdots \int_0^t dt_1 \cdots dt_k.
\end{aligned}$$

The above expression is only correct if we have the condition that $t_1 < ... < t_k$. Let S be the support of the pdf. Then this equals

$$\lambda^k e^{-\lambda t} Vol(S)$$

What is $Vol(S)$? If we had no constraints as to order, then the volume would be t^k. But by symmetry, all of the possible orderings (of which there are $k!$) have the same volume, so we have to divide by $k!$. If this is not intuitive, then consider the case when $k = 2$.

TODO 4. draw square for k =2, and show how vol(S) = $t^2/2!$

Then we have

$$N_t(k) = \frac{\lambda^k e^{-\lambda t} t^k}{k!}$$

which is a Poisson distribution with parameter λt, as desired! □

Example 16.2 (Fishing). Bob catches fish according to a PP($\lambda = 0.6/hr$). If he catches at least one fish in the first two hours, he quits. Otherwise, he continues until he has caught his first fish.

1. What is the probability that bob fishes for more than two hours?
2. What is the probability that Bob catches at least two fish?
3. What is the expected number of fish Bob catches?
4. What is the expected fishing time, given that he has been fishing for 4 hours already?

Answers:

1. $Pr[N_2 = 0] = e^{-\lambda \cdot 2} = e^{-1.2}$
2. $1 - \mathbf{Pr}(N_2 = 1) - \mathbf{Pr}(N_2 = 0) = 1 - e^{-1.2} - 1.2e^{-1.2}$
3. Total number of fish caught = fish caught in [0,2] + fish caught in [2,∞]. Then we can use linearity of expectation and we have

$$\mathbf{E}[\text{total fish caught}] = \mathbf{E}[Pois(1.2)] + 1 \cdot \mathbf{Pr}[\text{still fishing in } (2, \infty)]$$

$$1.2 + e^{-1.2}$$

4. We of course use the memoryless property, and we get

$$4 + \mathbf{E}[Expo(0.6)] = 4 + \frac{1}{0.6} = 5.66hrs$$

16.1 Merging and Splitting

Merging:

Suppose we have N_1 is a $PP(\lambda_1)$ and $N_2 \sim PP(\lambda_2)$. Then we have the following fact:

$$N = N_1 + N_2 \sim PP(\lambda_1 + \lambda_2)$$

This fact follows easily from the fact that the sum of two independent poisson random variables is still poisson with their parameters added, which we have proven earlier in the semester (can be proven with MGFs or with a convolution).

Splitting:

Suppose we have $N \sim PP(\lambda)$, and then we take each arrival and send it to N_1 with probability p, and send it to N_2 with probability $1 - p$. We do this independently for each arrival. Then here are some facts:

1. $N_1 \sim PP(\lambda p)$
2. $N_2 \sim PP(\lambda(1 - p))$
3. $N_1(t)$ and $N_2(t)$ are independent RVs.

These again follow from the proof of poisson random variable splitting we did in homework earlier in the semester.

Example 16.3. Suppose we have two lightbulbs have independetly and exponentially distributed lifetimes T_a and T_b respectively, with parameters λ_a and λ_b respectively. What is the distribution of $Z = min(T_a, T_b)$?

Well, we can recall that this is simply an exponential RV with rate $\lambda_a + \lambda_b$ (or can easily derive this by examining the CDF). But a *way cooler* way is to notice that that T_a and T_b are the times of the first arrivals of two independent Poisson Processes with rates λ_a and λ_b. The $\min(T_a, T_b)$ is the first arrival of the *merged* PP, which we know has rate $\lambda_a + \lambda_b$, so then we know that

$$\min(T_a, T_b) \sim Expo(\lambda_a + \lambda_b)$$

16.2 Erlang Distribution

Recall that we defined $T_k = S_1 + ... + S_k$ to be the time of the k^{th} arrival of the poisson process. We can observe:

1. $E[T_k] = \sum \mathbf{E}[S_i] = \frac{k}{\lambda}$
2. $\mathbf{Var}(T_k) = k\,\mathbf{Var}(S_i) = \frac{k}{\lambda^2}$

We would like to know the pdf of T_k. The distinctly uncool way to calculate this would be with a big convolution. But we can do something easier:

$$f_{T_k}(t)dt = \mathbf{Pr}[\text{k-1 arrivals in (0,t)}]\,\mathbf{Pr}(\text{1 arrival in (t,t+dt)})$$

$$= \frac{e^{-\lambda t}(\lambda t)^{k-1}}{(k-1)!} \cdot \lambda dt$$

$$\implies f_{T_k}(t) = \frac{e^{-\lambda t}\lambda^k t^{k-1}}{(k-1)!}$$

This is known as the k^{th}**-order Erlang Distribution**.

17 Lecture 17: CTMCs

Agenda:

1. Quick Recap of PPs
2. Random Incidence Paradox
3. CTMC's: Introduction
4. Rate Matrix and Stationary Distributions

17.1 Random Incidence Paradox (RIP)

Consider a Poisson process with rate λ, and suppose it has been going on infinitely long. The question we would like to answer is: if I pick a random t^*, what is the expected length of the interval in which it falls?

It is very tempting to say that, since each interarrival is $Expo(\lambda)$, that then the expected length of the interval is just the expectation of that exponential random variable, which is $\frac{1}{\lambda}$. Let's call L the length of the interval in which t^* falls. We claim then that L is distributed according to an Erlang-2(λ) distribution. Why is this the case?

Lets say t^* falls between T_i and T_{i+1}. Lets call $U := T_i$ and $V := T_{i+1}$. Then we have that

$$L = (t^* - U) + (V - t^*)$$

We have by the memorylessness property that $V - t^*$ is distributed according to an exponential distribution! Now what about $t^* - U$? We have:

$$\begin{aligned}\mathbf{Pr}(t^* - U > x) &= \mathbf{Pr}(\text{more than } x \text{ sec have elapsed since last arrival}) \\ &= \mathbf{Pr}(\text{no arrivals in } [t^* - x, t^*]) \\ &= \mathbf{Pr}(N(x) = 0) = e^{-\lambda x}.\end{aligned}$$

Now we note that this looks like $1 - F_X(\lambda)$ where $X \sim Expo(\lambda)$ (and F is the CDF of X), therefore $t^* - U$ must be an exponential random variable as well. Hence L is the sum of two independent exponential random variables and therefore an Erlang-2 distribution.

Remark 17.1. The key takeaway from the section above is that a Poisson process run backwards is still a Poisson process with exponential interarrival times.

Remark 17.2. Here is some more intuition for RIP. Suppose the bus schedule is fixed deterministically, a bus comes after 5 mins, then after 55 minutes, then after 5 minutes, then after 55 minutes, etc. Then we have that the average interarrival time is of course 30 minutes. However, what if we just randomly show up to the bus stop? Then we the expected length of the interval we arrive in is actually:

$$(5/60) \cdot 5 + (55/60) \cdot 55 = 50.83$$

17.2 Continuous-Time Markov Chains (CTMCs)

Similar to a discrete time Markov Chain, we start with a countable set $\mathcal{X}$ of states. Then the process is $\{X_t : t \geq 0\}$, defined via the following:

1. One is given an initial prob distribution over $\mathcal{X}$
2. a **rate matrix** Q where
   (a) $Q(i,j) \geq 0 \quad \forall i \neq j$.
   (b) $\sum_j Q(i,j) = 0$.

Example 17.3. We could have

$$Q = \begin{bmatrix} -4 & 3 & 1 \\ 0 & -2 & 2 \\ 1 & 1 & -2 \end{bmatrix}$$

(Note: This follows the convention in Walrand's book, not Bertsekas).

Definition 17.4. A CTMC with initial distribution π and rate matrix Q is a process $\{X_t, t \geq 0\}$ such that $\mathbf{Pr}(X_0 = i) = \pi(i)$, and

$$\mathbf{Pr}(X_{t+\epsilon} = j | X_t = i, X_u, u < t) = \begin{cases} \epsilon Q(i,j) + o(\epsilon) & i \neq j \\ 1 + \epsilon Q(i,i) + o(\epsilon) & i = j \end{cases}$$

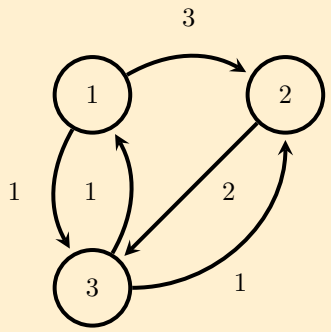
The above definition follows the traditional Markov property we are used to (when ϵ is small enough), and note that

$$\mathbf{Pr}(X_{t+\epsilon} \neq i | X_t = i) = \epsilon \sum_{i \neq j} Q(i,j)$$

$$\implies \mathbf{Pr}(X_{t+\epsilon} = i | X_t = i) = 1 - \epsilon \sum_{i \neq j} Q(i,j) = 1 + \epsilon Q(i,i)$$

Where in the above I have dropped the $o(\epsilon)$ terms.

Example 17.5. Consider the Markov chain below:



Where

$$Q = \begin{bmatrix} -4 & 3 & 1 \\ 0 & -2 & 2 \\ 1 & 1 & -2 \end{bmatrix}$$

And suppose our initial π is $\pi = (1/3, 1/3, 1/3)$. Then we have:

$$\mathbf{Pr}(X_{t+\epsilon} = 3 | X_t = 1) = \epsilon Q(1,3)$$

We can set up the following *jump chain*:

and we define $q_i = |Q_{ii}|$.

18 Lecture 18: More on CTMCs

Agenda:

1. Review of CTMC definition and properties
2. Rate matrix Q and stationary distribution π
3. Examples
4. First Step Equations for CTMC
5. Simulating a CTMC using a DTMC

Recall that a CTMC is defined by a **rate matrix** Q where

$$Q(i,j) \geq 0 \quad \forall i \neq j$$

$$\sum_j Q(i,j) = 0$$

And by definition we have

$$Q(i,j) \geq 0 \quad \forall i \neq j$$

$$Q(i,i) \leq 0$$

The **holding time** in state i, which is the amount of time we wait in state i before making a jump, is $\sim Expo(q(i))$ where $q(i) = -Q(i,i)$

Recall we also have the continuous analog of the Markov Property:

$$\mathbf{Pr}(X_{t+\epsilon} = j | X_t = i, X_u, u < t) = \begin{cases} \epsilon Q(i,j) + o(\epsilon) & i \neq j \\ 1 + \epsilon Q(i,i) + o(\epsilon) & i = j \end{cases}$$

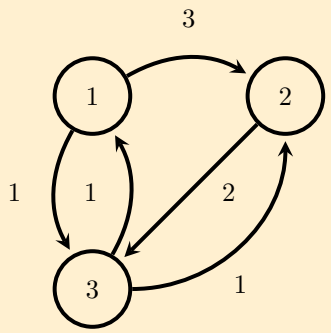
And in particular we have

$$\mathbf{Pr}(X_{t+\epsilon} = i | X_t = i) = 1 - \epsilon \sum_{i \neq j} Q(i,j) = 1 - \epsilon(-Q(i,i)) = 1 - \epsilon q(i)$$

If the current state is i, the time to "jump" is $Expo(q(i)$.

Lets go back to the example we were looking at then end of last lecture:

Example 18.1. Consider the Markov chain below.



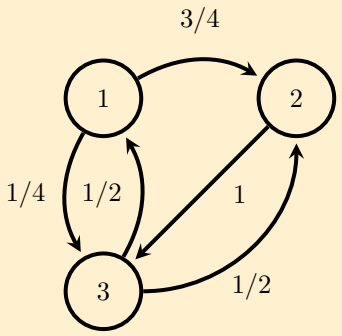
Where

$$Q = \begin{bmatrix} -4 & 3 & 1 \\ 0 & -2 & 2 \\ 1 & 1 & -2 \end{bmatrix}$$

And suppose our initial π is $\pi = (1/3, 1/3, 1/3)$. Then we have:

$$\mathbf{Pr}(X_{t+\epsilon} = 3 | X_t = 1) = \epsilon Q(1, 3)$$

We can setup the following *embedded DTMC*:



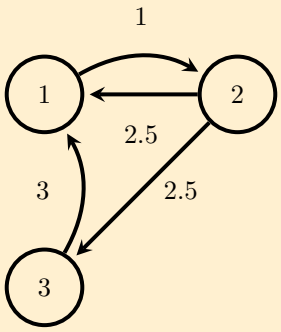
This DTMC models the jumps of the CTMC, but potentially does not model how long we are waiting in each state in the CTMC. We get these probabilities by remembering the properties of exponential splitting. Then we have

$$\mathbf{Pr}(X_{t+\epsilon} = 3 | X_t = 1) = (4\epsilon)(1/4)$$

How did we get this? Well, we need to jump in time $(0, \epsilon)$, which happens roughly with probability $q(i)\epsilon$. Then, when we jump, we need to jump to the correct state j, which happens with probability $\frac{Q(i,j)}{q(i)}$. Both of these things need to happen in order to transition to state j from state i, so we have

$$\mathbf{Pr}(X_{t+\epsilon} = j | X_t = i) = \epsilon q(i) \frac{Q(i, j)}{q(i)} = \epsilon Q(i, j)$$

Example 18.2 (B&T 7.14)**.** Say we have a Normal state (1) and a test state (2) and a repair state (3).



What is the stationary distribution of this Markov Chain? Recall that for discrete Markov Chains, we needed to solve for $\pi = \pi P$. However, for a continuous time markov chain, we need to solve for

$$\pi Q = 0 \quad \sum_i \pi_i = 1$$

This is known as the "Rate conservation principle" or "rate in = rate out" (can you see why?). Then

we have

$$Q = \begin{bmatrix} -1 & 1 & 0 \\ 2.5 & -5 & 2.5 \\ 3 & 0 & -3 \end{bmatrix}$$

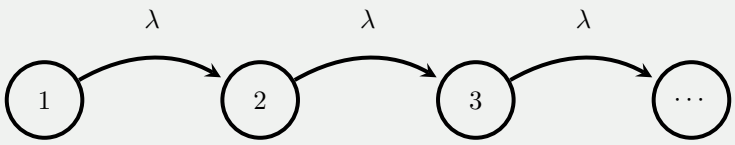
Writing out the equations, we have

$$\pi(1) \cdot 1 = \pi(2)\frac{5}{2} + \pi(3) \cdot 3$$

Notice that the left hand side is the flow coming out of state 1, while the right hand side is the flow going into state 1! For me, it is easier to remember this concept than to write out the equations by just remembering the matrix formula. We can set up the rest of the equations and solve to find

$$\pi = (30/41, 6/41, 5/41)$$

Remark 18.3 (Poisson Processes are CTMCs)**.** If we have a $PP(\lambda)$, we can model it as a CTMC as follows:



Example 18.4. We can consider the two state markov chain which transitions from state 0 to state one with rate λ, and from state 1 to state 0 with rate μ. Then by the flow equations we have

$$-\lambda\pi_0 = \mu\pi_1 - \mu\pi_1 \qquad = \lambda\pi_0\pi_0 + \pi_1 = 1$$

We can solve these equations to find that

$$\pi_0 = \frac{\mu}{\lambda + \mu} \quad \pi_0 = \frac{\lambda}{\lambda + \mu}$$

So when $\lambda = 1$ and $\mu = 2$, we are "parked" in state 0 twice as much as we are parked in state 1. However, the embedded DTMC has the stationary distribution $(1/2, 1/2)$, no matter what λ and μ are. Clearly, we need to do something different if we want our DTMC to have the same stationary distribution as the CTMC.

18.1 Hitting Times

Example 18.5. Consider 20 lightbulbs that have indep. lifetimes that are exponentially distributed with rate 1 (month). How long before all the bulbs die out?
We can model this with 21 states, corresponding to the number of lightbulbs still alive. We transition from state 20 to state 19 with rate 20 (min of 20 exponentials!), and from 19 to 18 with rate 19, and so on. How long does it take for us to get to zero? We have to come up with **first step equations**, which are quite analogous to the ones we saw for DTMCs. The biggest difference is that we are not transitioning after 1 time step now, we are transitioning on average after $1/q(i)$ amount of time (the mean of the exponential) if we are in state i. We can define x_i as the expected amount of time to hit

zero given you are in state m. We have then

$$x_i = \frac{1}{i} + x_{i-1}$$

This implies that

$$x_{20} = \frac{1}{20} + \frac{1}{19} + \cdots + 1 \approx 3.6$$

Example 18.6. Now assume the burnt out bulbs are replaced after an exponential amount of time with mean 0.1 month (meaning $\lambda = 10$). Now what is the expected amount of time until all our bulbs have burnt out? Now our FSE looks like:

$$\beta(20) = \frac{1}{20} + \beta(19)$$

and more generally for $1 \leq m \leq 19$:

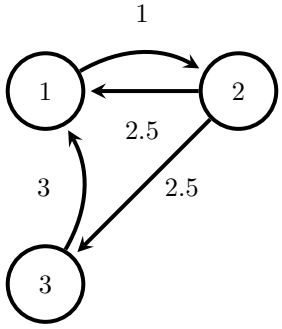
$$\beta(m) = \frac{1}{m+10} + \frac{m}{m+10}\beta(m-1) + \frac{10}{m+10}\beta(m+1)$$

and finally $\beta(0) = 0$. Solving for these equations recursively yields

$$\beta(20) \approx 2488$$

18.2 Simulating a CTMC with a DTMC

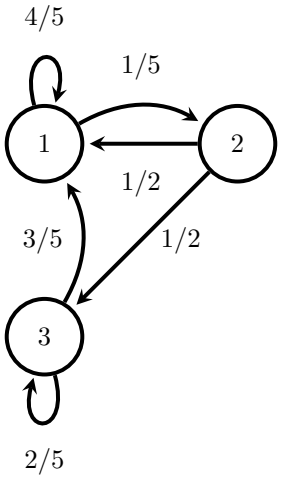
Let's consider the same example as before.



where

$$Q = \begin{bmatrix} -1 & 1 & 0 \\ 2.5 & -5 & 2.5 \\ 3 & 0 & -3 \end{bmatrix}$$

We then define $\Lambda = \max_i q(i) = 5$. This is intuitively our "clock delay". We can form the DTMC by dividing all the transition rates by Λ and adding self loops when necessary:



We can verify that this π is the same for this DTMC as for the CTMC.

19 Lecture 19: Random Graphs

Agenda:

1. Wrapup of CTMCs (simulating CTMC with DTMC)
2. Random Graphs
   (a) intro, definition
   (b) Erdös-Renyi $G(n,p)$ random graph model
   (c) Threshold conditions for graph connectivity

19.1 recap of CTMCs

Why would we want to simulate a CTMC with a DTMC? For one, it's easier to implement and do on a computer. Two key points about simulating CTMCs with DTMCs:

1. The **jump** or **embedded** chain has no self loops, and in general the stationary distribution of this chain and the corresponding CTMC is *not* the same. Intuitively, this is because if we have different values of $q(i)$ for each state i, then we are "waiting" longer at certain states, which is not reflected in the jump chain.
2. If we want the corresponding DTMC to have the same stationary distribution, we have to form a DTMC by dividing transition by $\Gamma = \max_i q(i)$, and adding self loops where necessary. This effectively adds in a waiting time to the corresponding DTMC. Here we can think of $\Gamma = \max_i q(i)$ as the **clock rate** of the markov chain.

The stationary distribution remains the same in the second case because our new matrix $P = I + Q/\Gamma$. We can verify that an eigenvalue of Q is also an eigenvalue of P, so therefore the stationary distributions must be the same.

19.2 Random Graphs

Some motivation:

1. Graphs, and random graphs, are everywhere. They have applications to the behavior of social networks, biological networks, recommendation systems (matrix completion), etc.
2. Modeling epidemics (very topical) involves random graphs.

Definition 19.1 (Erdös-Renyi (ER) Random Graphs)**.** Given a positive integer n and a probability value $p \in (0,1]$. Then $\mathcal{G}(n,p)$ is a **random graph** which is *undirected* graph on n vertices such that each of the $\binom{n}{2}$ edges are present independently and with probability p.

Intuitively, we are just drawing n vertices, and for each pair of vertices, we flip a biased coin (prob p of heads), and if it comes out heads, we draw an edge between these two vertices.

Erdös and Renyi stated a number of results that are based on "thresholds" of p needed for certain *structural properties* of the graph to emerge.

Example 19.2.
1. If $p = \frac{1}{n^2}$, then we see at least one edge with high probability.
2. if $p = \frac{1}{n^{3/2}}$, then the first "3-node trees" (incomplete triangles) start to emerge.
3. if $p = \frac{1}{n}$, then the first cycles begin emerging

4. if $p = \frac{1}{n}$, then the first "Giant Component" emerges. Specifically, if $p = \frac{1-\epsilon}{n}$, then the largest connected components are of size $O(\log n)$, but if $p = \frac{1+\epsilon}{n}$, then suddenly the size of the largest component is of size $O(n)$.

We will focus today on arguably the most important threshold for random graphs: the **threshold for connectivity**.

Lemma 19.3. *If $p > \frac{\log n}{n}$, then our graph is connected w.h.p. Otherwise, if $p < \frac{\log n}{n}$ then our graph is not connected with high probability*

Remark 19.4. We have that the probability of a particular fixed graph G_0 with m edges appearing is

$$\mathbf{Pr}(\mathcal{G}(n,p) = G_0) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}$$

Question 1: What is $\mathbf{E}[\#$ edges in $G]$? We have that there are $\binom{n}{2}$ possible edges, each appearing with probability p, so by linearity of expectations (after defining appropriate indicators), we have

$$\mathbf{E}[\#\text{ edges in } G] = p\binom{n}{2}$$

Question 2: If we pick an arbitrary vertex and let D be its degree, then what is the distribution of D? What is the expected degree?
Answer: We have that $D \sim Binom(n-1, p)$, and therefore we have

$$\mathbf{Pr}(D = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d} \quad \forall d \in \{0, 1, ..., n-1\}$$

and $\mathbf{E}[D] = p(n-1)$.
Question 3: Suppose now that $p_n = \frac{\mu}{n}$ for a constant $\mu > 0$. What is the approximate distribution of D when $n \to \infty$, $p_n \to 0$, and $p_n n \to \mu$?
Answer:

$$D \sim Poisson(\mu)$$

And we have

$$\mathbf{Pr}(D = d) \approx \frac{e^{-\mu}\mu^d}{d!}$$

Question 4: What is the probability q that a node is isolated?
Answer: $q = (1-p)^{n-1}$

Theorem 19.5 (E-R '61). *Let $p_n = \lambda \frac{\log n}{n}$. Then*

1. *If $\lambda < 1$, then*
$$\mathbf{Pr}(\mathcal{G}(n,p) \text{ is connected}) \xrightarrow{n\to\infty} 0$$
2. *If $\lambda > 1$, then*
$$\mathbf{Pr}(\mathcal{G}(n,p) \text{ is connected}) \xrightarrow{n\to\infty} 1$$

Remark 19.6. if $p_n = \frac{\ln n + c}{n}$ for a constant $c \in \mathbb{R}$, then it can be shown that

$$\mathbf{Pr}(\mathcal{G}(n,p) \text{ is connected}) \xrightarrow{n\to\infty} e^{-e^{-c}}$$

Before we begin the proof of the the theorem, we need this quick lemma:

Lemma 19.7. *If X is a non-negative integer valued RV, then*

$$\mathbf{Pr}(X = 0) \leq \frac{\mathbf{Var}(X)}{\mathbf{E}[X]^2}$$

Proof.

$$\begin{aligned}\mathbf{Var}(X) &= \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{Pr}(X = 0)\,\mathbf{E}[X]^2 + \mathbf{Pr}(X = 1)[\mathbf{E}[X] - 1]^2 + \mathbf{Pr}(X = 2)[\mathbf{E}[X] - 2]^2 + \dots \\ &\geq \mathbf{Pr}(X = 0)\,\mathbf{E}[X]^2 \\ &\implies \mathbf{Pr}(X = 0) \leq \frac{\mathbf{Var}(X)}{\mathbf{E}[X]^2}.\end{aligned}$$

□

Now we can do the proof of the theorem:

Proof. For (1), it is sufficient to show that there will be isolated nodes with high probability. This is actually something stronger than we need, but we will prove it anyway. We would like to show

$$\mathbf{Pr}(\text{no isolated nodes}) \xrightarrow{n\to\infty} 0$$

Let X be the number of isolated nodes in our graph. First, lets find $\mathbf{E}[X]$. We define I_i as the indicator RV of the event that node i is isolated. Then we have that

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[I_i] = \sum_{i=1}^{n} \mathbf{Pr}(\text{node } i \text{ is isolated}) = nq = n(1-p)^{n-1}$$

Then we have that

$$\ln \mathbf{E}[X] = \ln n + (n-1)\ln(1-p) \approx \ln n + (n-1)(-p) = \ln n + (n-1)\frac{-\lambda \ln n}{n}$$

Where in the second equality we have used the fact that via Taylor series expansion, for small x, we have $\ln(1-x) \approx -x$. Then we have finally

$$\ln \mathbf{E}[X] \approx \ln n - \frac{n-1}{n}\lambda \ln n \approx \ln n(1-\lambda) \xrightarrow{n\to\infty} \infty$$

And we have that

$$\mathbf{E}[X] \approx e^{(\ln n)(1-\lambda)} = n^{1-\lambda}$$

Note that we *are not done yet*. Just because the expectation of a RV goes to infinity does not mean that the probability that it is not zero goes to zero. Consider the RV

$$W = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{n} \\ n^2 & \text{w.p. } \frac{1}{n} \end{cases}$$

Then we can see that $\mathbf{E}[W] \xrightarrow{n\to\infty} \infty$, but also the probability that $W = 0$ goes to one as n goes to infinity. To finish the proof, we need to get a handle on the variance of X, as well as use the lemma we proved above. We have

$$\mathbf{Var}(X) = \mathbf{Var}(\sum_i I_i)$$

We note here that the our indicators are not independent, so we have to do some more work.

$$\mathbf{Var}(X) = \sum_{i=1}^{n} \mathbf{Var}(I_i) + \sum_{j=1}^{n}\sum_{k\neq j} \mathbf{Cov}(I_j, I_k) \qquad = n\,\mathrm{var}(I_1) + n(n-1)\,\mathbf{Cov}(I_1, I_2)$$

Now, in order to continue, we need to figure out what the covariance term is. We have

$$\mathbf{Cov}(I_1, I_2) = \mathbf{E}[I_1 I_2] - \mathbf{E}[I_1]\,\mathbf{E}[I_2]$$

Here since I_i is an indicator, $\mathbf{E}[I_i]$ is just equal to the probability that node i is isolated, and $\mathbf{E}[I_1 I_2]$ is the probability that both nodes one and two are isolated, so we have

$$\mathbf{Cov}(I_1, I_2) = (1-p)^{n-1}(1-p)^{n-2} - (1-p)^{n-1}(1-p)^{n-1} = \frac{q^2}{1-p} - q^2$$

Now we can plug this back into our original expression, and we find that

$$\mathbf{Var}(X) = nq(1-q) + n(n-1)\left[\frac{q^2}{1-p} - q^2\right] = nq(1-q) + n(n-1)\frac{pq^2}{1-p}$$

Now, we use our lemma to upper bound the probability that $X = 0$:

$$\begin{aligned}\mathbf{Pr}(X=0) \leq \frac{\mathbf{Var}(X)}{\mathbf{E}[X]^2} &= \frac{nq(1-q) + n(n-1)\frac{pq^2}{1-p}}{n^2q^2} \\ &\leq \frac{1-q}{nq} + \frac{n-1}{n}\frac{p}{1-p}.\end{aligned}$$

Now we are done, because $\frac{1-q}{nq} \xrightarrow{n\to\infty} 0$ and $\frac{p}{1-p} \xrightarrow{n\to\infty} 0$. So we have proven the first part of the theorem.

How do we prove the second part? Namely, we want to show that if $\lambda > 1$, we want to show that

$$\mathbf{Pr}(\text{G not connected}) \xrightarrow{n\to\infty} 0$$

We will give only a proof sketch here. The idea is that the "Graph is disconnected" $\equiv$ "there exists a set of size k (where $1 \leq k \leq n/2$) such that there is no edge between this set and its complement". Next, we will apply the union bound twice to get the result.

$$\begin{aligned}\mathbf{Pr}(\text{graph is not connected}) &= \mathbf{Pr}(\bigcup_{k=1}^{n/2}(\exists\text{set of size } k \text{ disconnected from everything})) \\ &\leq \sum_{k=1}^{n/2}\binom{n}{k}\mathbf{Pr}(\text{a specific set of size } k \text{ is disconnected}) \\ &= \sum_{k=1}^{n/2}\binom{n}{k}(1-p)^{k(n-k)}.\end{aligned}$$

We can show that this summation goes to zero as n goes to infinity, but we omit the details here because it gets a bit messy. The details can be found in the appendix of the class notes.

□

20 Lecture 20: Wrapup of Random Graphs and Starting Statistical Inference

Agenda:

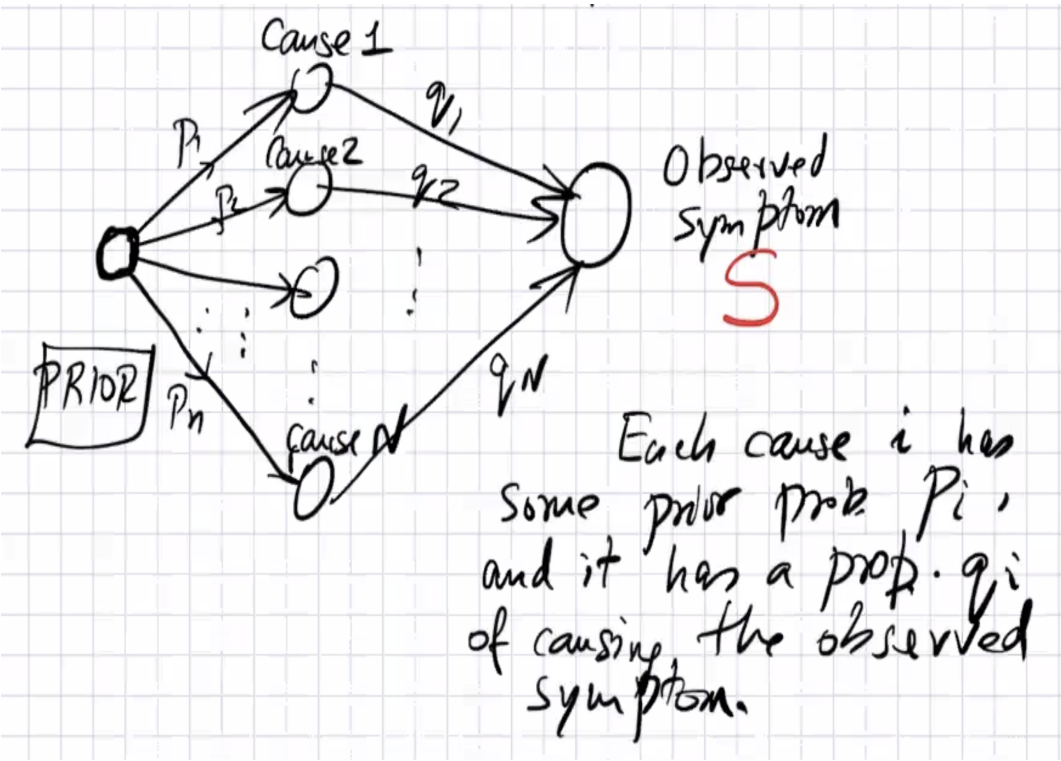
1. Random Graphs (finish proof of graph connectivity cutoff)
2. Statistical Inference (MAP and MLE and examples)

20.1 Statistical Inference

There are two schools of thought when it comes to statistical inference:

1. **Bayesian**: Treats unknowns as RVs with known distributions and priors, which effectively moves statistics to the realm of probability analysis.
2. **Frequentist:** Treats unknown as deterministic parameters to be estimated (for example, the mass of an electron). Bayesian priors reflect our knowledge.

The **basic premise**: There are some number N possible exclusive *causes* of a particular *symptom*. Exactly 1 of the N possibilities is the correct cause.



In the above figure, the p_i's are the *priors*, which tells us how likely each of our causes are *a priori*. The q_i's tell us the conditional probability of the symptom given a specific cause i. However, we are looking for the **posterior probability** π_i, which is the conditional probability of cause i given the symptoms we are seeing. We will use the all powerful Bayes Rule to achieve figure this out. We have:

$$\pi_i = \mathbf{Pr}(c_i|S) = \frac{\mathbf{Pr}(c_i \cap S)}{\mathbf{Pr}(S)} = \frac{\mathbf{Pr}(S|c_i)\,\mathbf{Pr}(c_i)}{\sum_i \mathbf{Pr}(S|c_i)\,\mathbf{Pr}(c_i)}$$

$$= \frac{p_i q_i}{\sum_i p_i q_i}$$

This is actually our first theorem:

Theorem 20.1 (Bayes' Theorem)**.**

$$\pi_i = \frac{p_i q_i}{\sum_i p_i q_i}$$

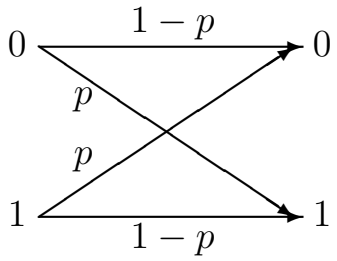
Then the **Maximum A Posteriori** (MAP) estimate of the cause given the symptom is $\arg\max_i \pi_i = \arg\max_i p_i q_i$.
The **Maximum Likelihood Estimate** (MLE) is the same thing, except we assume that all our priors are just uniformly distributed (so no cause is more likely than another). So then the MLE is $\arg\max_i q_i$.

More generally, we have that

$$MAP(X|Y=y) = \arg\max_x \mathbf{Pr}(X=x|Y=y)$$

$$MLE(X|Y=y) = \arg\max_x \mathbf{Pr}(Y=y|X=x)$$

Let's do some analysis on the MAP/MLE of a BSC channel. Recall the BSC channel looks like this:



Theorem 20.2. *For a BSC(p) with $p < 1/2$, we have*

$$MAP[X|Y=0] = \mathbb{1}_{p>1-\alpha}$$
$$MAP[X|Y=1] = \mathbb{1}_{p<\alpha}$$

and

$$MLE[X|Y] = Y$$

where $\alpha = \mathbf{Pr}(X=1)$.

Proof. We have $MAP[X|Y=1] = \arg\max_{i\in\{0,1\}} p_i q_i$. If $p_0 q_0 > p_1 q_1$, then we guess that $\hat{x} = 0$, and otherwise if $p_0 q_0 < p_1 q_1$, then our guess is $\hat{x} = 1$. We have

$$p_0 q_0 = (1-\alpha)p \qquad p_1 q_1 = \alpha(1-p)$$

Then we have that $p_0 q_0 < p_1 q_1 \Leftrightarrow p < \alpha$, which means that

$$\hat{X}_{MAP}(Y=1) = \mathbb{1}_{p<\alpha}$$

which is exactly what we wanted to show. We can do something very similar for the case when we observe $Y = 0$. Finally, we note that $MLE[X|Y]$ is equal to our MAP estimate, when we let $\alpha = 1/2$. Then we get the result claimed.

□

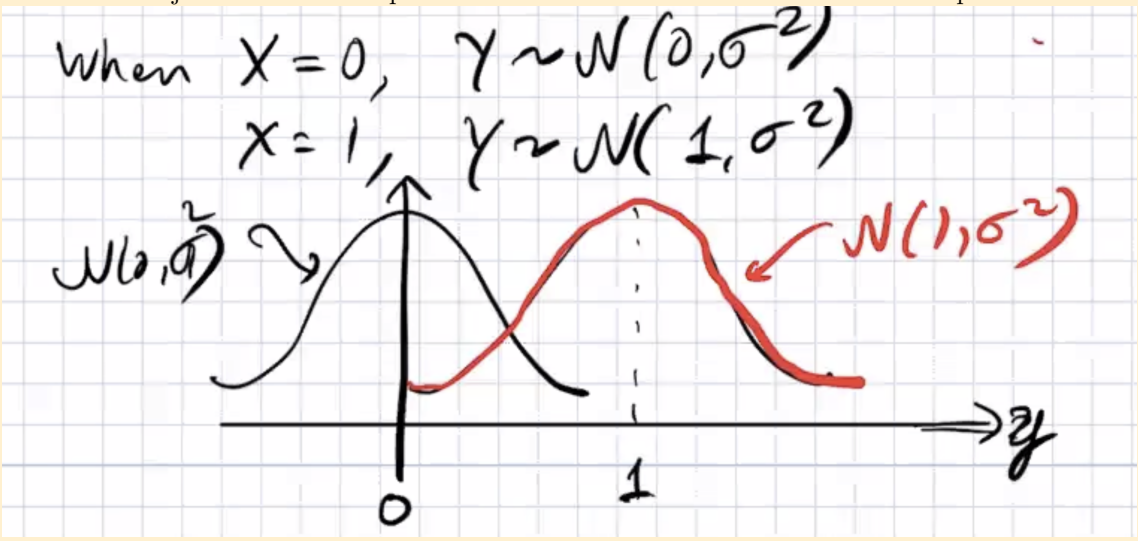
21 Lecture 21: Wrapup of MLE/MAP and Hypothesis Testing/Neyman-Pearson

Agenda:

1. Wrapup of MLE/MAP
2. Hypothesis Testing and Neyman-Pearson

21.1 Wrapup of MLE/MAP

Example 21.1 (AWGN channel)**.** AWGN (Additive White Gaussian Noise) channel is just a channel that takes in X and outputs $Y = X + Z$ where $Z \sim \mathcal{N}(0, \sigma^2)$. Assume X is Bernoulli, then when $X = 0$, we have $Y \sim \mathcal{N}(0, \sigma^2)$, and when $X = 1$ we have $Y \sim \mathcal{N}(1, \sigma^2)$. If we plot this, we can see that these are just two Gaussian pdf's with different means that have some overlap:



Our *symptom* is $Y = y$, while our causes are $X = 0$ or $X = 1$. Then we note that

$$\mathbf{Pr}(Y = y|X = 0) \approx f_0(y)\epsilon \qquad \mathbf{Pr}(Y = y|X = 1) \approx f_1(y)\epsilon$$

where f_0 and f_1 are the densities of the normal distribution centered at 0 and 1 respectively. Then our MAP rule says we should declare $\hat{X} = 0$ if

$$p_0 f_0(y)\epsilon \geq p_1 f_1(y)\epsilon$$

and otherwise we should declare $\hat{X} = 1$. Equivalently, we guess $\hat{X} = 0$ if

$$\frac{f_1(y)}{f_0(y)} \leq \frac{p_0}{p_1}$$

where this $\frac{f_1(y)}{f_0(y)}$ is known in the business as a **likelihood ratio** $L(y)$. In particular, we have

$$L(y) = \frac{\frac{1}{\sqrt{2\pi}} \exp -\frac{(y-1)^2}{2\sigma^2}}{\frac{1}{\sqrt{2\pi}} \exp \frac{-y^2}{2\sigma^2}}$$

$$= \exp\left(\frac{2y-1}{2\sigma^2}\right)$$

We can take the log of the likelihood $LL(y) := \log L(y)$

$$LL(y) = \frac{2y-1}{2\sigma^2}$$

Then we guess $\hat{X} = 0$ iff

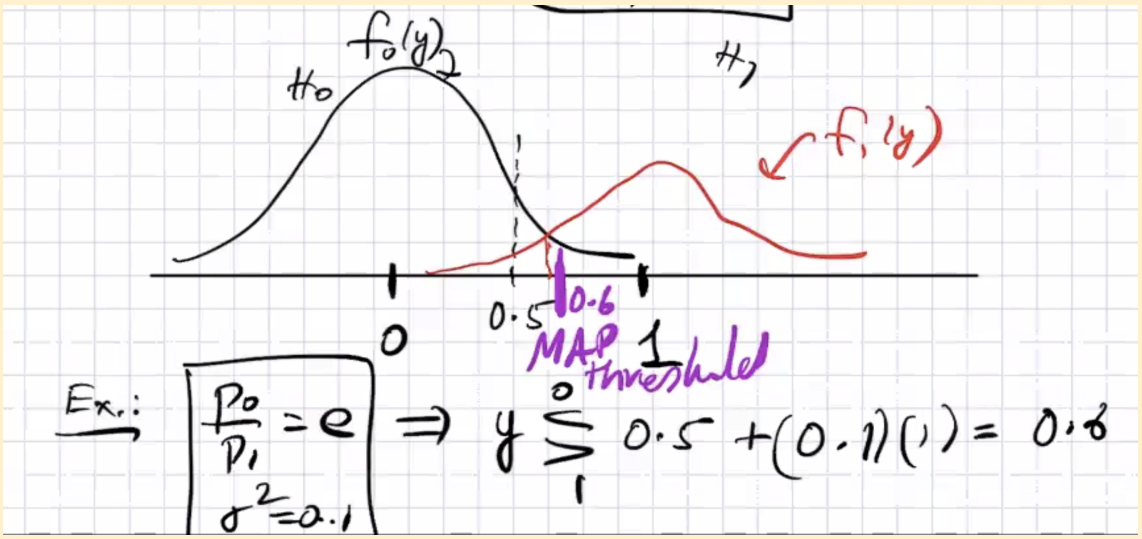
$$LL(y) \leq \log\left(\frac{p_0}{p_1}\right)$$

or in other words we guess zero if

$$y < \frac{1}{2} + \sigma^2 \log\left(\frac{p_0}{p_1}\right)$$

and this is our MAP rule. what about our MLE? Well we just set $p_0 = p_1$ in that case and our rule becomes to guess $\hat{X} = 0$ iff

$$y < \frac{1}{2}$$



In the above example, we can see how the MAP rule affects our decisions. If $\frac{p_0}{p_1} = e$ and $\sigma^2 = 0.1$, then our new cutoff is 0.6, indicating that we are inclined to believe that the cause was $X = 0$ even if the density is higher for $X = 1$.

Example 21.2 (German Tank Problem)**.** This problem was motivated by WWII, where the allies wanted to estimate how many tanks the Germans had just by observing the serial numbers of the captured/destroyed tanks. Suppose we have a bucket with N balls, each labeled 1 through N, where N is an unknown integer. What is

$$MLE[N|Y = m]$$

Let's say we observe $Y = 7$. What is the MLE of N given your observation? We have

$$MLE[N|Y = 7] = \arg\max_n \mathbf{Pr}[Y = 7|N = n] = \arg\max_n \begin{cases} 0 & \text{if } n < 7 \\ 1/n & \text{if } n \geq 7 \end{cases}$$

Which implies that $MLE[N|Y = 7] = \hat{N} = 7$, and in general

$$MLE[N|Y = m] = m$$

What if we observe k balls, $y_1, ..., y_k$? There are $\binom{n}{k}$ sets of k distinct numbers that are subsets

of $[n] = \{1, ..., n\}$. Each subset is equally likely, so

$$\mathbf{Pr}(y_1, ..., y_k | n) = \begin{cases} \frac{1}{\binom{n}{k}} & \text{if } y_1, ..., y_k \in [n] \\ 0 & \text{otherwise} \end{cases}$$

This is once again maximized when $n = \max\{y_1, y_2, ..., y_k\} = m_k$. This is once again a weird solution, and arises because the MLE is a biased estimator. It can be fixed, but we won't get into how in this course.

21.2 Hypothesis Testing

Motivation: In many settings, the notion of a prior doesn't make much sense. What's the prior of your house being on fire? MLE and MAP are "point-estimates", which are not meaningful when priors are not sensible to assign (e.g. alarm systems, spam filters, medical tests, etc.)
We are going to use the **Formulation of Neyman and Pearson:**
Consider $X \in \{0, 1\}$ is the inference RV of interest, and it induces a continuum of tradeoffs based on the observation. Our goal is to maximimize the *probability of correct detection*:

$$\max PCD = \max \mathbf{Pr}(\hat{X} = 1 | X = 1)$$

such that the *probability of false alarm* is less than some value β:

$$PFA = \mathbf{Pr}(\hat{X} = 1 | X = 0) \leq \beta$$

Then the Neyman-Pearson method is to:

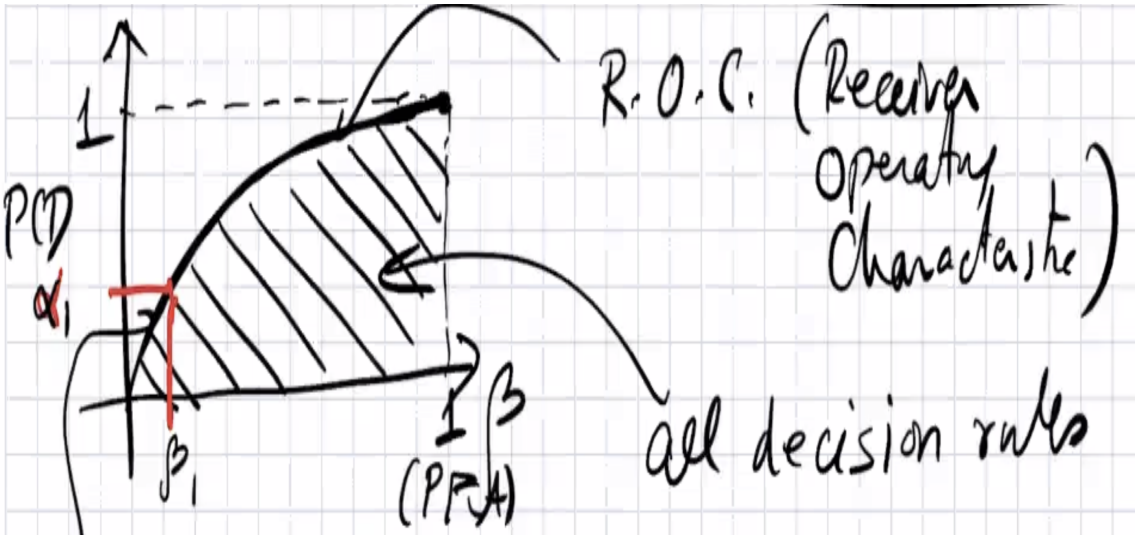
1. Observe Y
2. Two hypotheses:
   - $H_0 : Y \sim f(y|X = 0)$. This is called the *null hypothesis*
   - $H_1 : Y \sim f(y|X = 1)$. This is called the *alternative hypothesis*
3. formalize some decision rule
$$r : \mathbb{R} \to \{0, 1\}$$
where $\hat{X} = r(Y)$.

Alternatively, but equivalently, we could try to minimize $\mathbf{Pr}(r(Y) = 0 | X = 1]$ which is called a *false negative* subject to $\mathbf{Pr}(r(Y) = 1 | X = 0) \leq \beta$ which is called a *false positive*.

Remark 21.3. 1. There are no priors
2. there are only two hypotheses
3. False negatives are far more critical than false positives.

Remark 21.4. *False Positives* are also known as **type-I error**, while false negatives are also known as **type-II errors**. However, we will be following the notation found in Walrand's textbook.

Recall our goal of maximizes PCD while minimizing PFA. This can be drawn as what is known as an **ROC curve** (the reason for the name is historical, it's just jargon). If $\beta = 1$, then we can just always guess $\hat{X} = 1$, but we will have a lot of false alarms.



The above figure illustrates that all valid decision rules must fall under the ROC curve. The optimal decision rule, however, always falls on the ROC curve.

Theorem 21.5 (Neyman-Pearson). *Recall the Likelihood ratio* $L(y) = \frac{f(y|1)}{f(y|0)}$ *Then the optimal decision rule is*

$$r^*(Y) = \begin{cases} 1 & \text{if } L(Y) > \lambda \\ 0 & \text{if } L(y) < \lambda \\ 1 & \text{w.p. } \gamma \text{ if } L(y) = \lambda \end{cases}$$

where $\lambda > 0$ *and* $\gamma \in [0, 1]$ *are chosen to make sure that*

$$\mathbf{Pr}(\hat{X} = 1 | X = 0) = \beta$$

Intuitively, since we are allowed to err with probability β, we want to use all of this slack in order to maximize our probability of correct detection. If $L(y)$ is "large" then we are declaring $\hat{X} = 1$, while otherwise we guess zero. λ controls the "sensitivity" of your detector. We want to choose λ sensitive enough to *just* meet the PFA constraint β.

Example 21.6 (Bias of a Coin). Let H_0 correspond to the hypothesis that our coin is fair, while H_1 corresponds to the hypothesis is biased to land heads 60% of the time. We observe Y_i for the i^{th} coin toss, $i = 1, ..., n$. We would like the PFA $= \mathbf{Pr}(\hat{X} = 1 | X = 0) \leq 0.05$. We have

$$\mathbf{Pr}(Y_1 = y_1, ..., Y_n = y_n | X = 0) = (0.5)^n \, \mathbf{Pr}(Y_1 = y_1, ..., Y_n = y_n | X = 1) = (0.6)^H (0.4)^{n-H}$$

Then we have that

$$L(Y_1, ..., Y_n) = \frac{(0.6)^H (0.4)^{n-H}}{(0.5)^n} = \left(\frac{0.4}{0.5}\right)^n \left(\frac{0.6}{0.4}\right)^H$$

where H = number of heads observed. H is known in the business as a *sufficient statistic* for the detection problem, because n is given and beyond that L only depends on H. The Neyman-Pearson theorem says that

$$\hat{X} = \begin{cases} 1 & H \geq n_0 \\ 0 & H < n_0 \end{cases}$$

where $\mathbf{Pr}(H \geq n_0 | X = 0) = 0.05$. We technically need to find a threshold on $L(y)$, but since $L(y)$ depends monotonically on H, we can just equivalently find a threshold on H, which is intuitive and

also a lot easier in this particular problem. $H \geq n_0$ if and only if $L(y) \geq \lambda$. Now all we have to do is figure out what n_0 is for a PFA of 5%. If $X = 0$, then $H \sim Binom(n, \frac{1}{2})$, so $\mathbf{E}[H] = \frac{n}{2}$ and $\mathbf{Var}[H] = \frac{n}{4}$. We are trying to bound the probability $\mathbf{Pr}(H \geq n_0 | X = 0)$, as this is exactly the PFA. Here, we can use the CLT to argue that

$$\mathbf{Pr}(H \geq n_0 | X = 0) = \mathbf{Pr}\left(\frac{H - \mathbf{E}[H]}{\sqrt{\mathbf{Var}(H)}} \geq \frac{n_0 - \mathbf{E}[H]}{\sqrt{\mathbf{Var}(H)}}\right)$$

$$\approx \mathbf{Pr}\left(\mathcal{N}(0,1) \geq \frac{n_0 - \frac{n}{2}}{\sqrt{n/4}}\right)$$

$$= 1 - \mathbf{Pr}\left(\mathcal{N}(0,1) < \frac{n_0 - \frac{n}{2}}{\sqrt{n/4}}\right) = 0.05$$

$$\implies \frac{n_0 - n/2}{\sqrt{n}/2} = \Phi^{-1}(0.95) = 1.65$$

$$\implies n_0 = 0.825\sqrt{n} + n/2$$

if $n = 100$, then we find that $n_0 = 58.25$

Remark 21.7. In the above example, if the bias of the coin in our alternative hypothesis, then nothing in our calculations at the end of the problem changes, and our cutoff n_0 remains 58.25 for a 5% PFA. However, the PCD will be different (namely, it will be significantly higher) for this different alternative hypothesis. This phenomenon that a single rule works for a whole array of alternative hypotheses is sometimes call the "UMP rule" (uniformly most powerful).

22 Lecture 22: Wrapup of Hypothesis Testing and Beginning of LLSE

Agenda

1. Hypothesis Testing Wrapup (more examples and proof)
2. Estimation (Linear Least Squares Estimation (LLSE))

22.1 Proof of Neyman-Pearson

First, we show an example of why we might need a randomized rule (the case when $L(y) = \lambda$ and we have to "ring the alarm" (meaning guess $\hat{X} = 1$) with probability γ).
As slightly more motivation, Consider two hypotheses H_0 and H_1, but we have no observation Y. Then our decision is not a function but either $r = 0$ always or r is a random variable. If $r = 0$ is a constant then

$$\mathbf{Pr}(\hat{X} = 1|X = 0) = 0 \leq \beta$$

so our PFA is golden, but or PCD is terrible:

$$\mathbf{Pr}(\hat{X} = 1|X = 1) = 0$$

The idea is we can choose r randomly, in particular we can set

$$r = \begin{cases} 1 & \text{w.p. } \beta \\ 0 & \text{w.p. } 1 - \beta \end{cases}$$

Then we have that the PFA $= \mathbf{Pr}(r = 1|X = 0) = \mathbf{Pr}(r = 1) = \beta$, and the PCD $= \mathbf{Pr}(r = 1|X = 1) = \beta$ which is better than the other rule.
Now, we prove the Neyman-Pearson Theorem:

Proof. Consider a binary hypothesis testing problem. The idea of the proof is to show that any other decision rule (other than Neyman-Pearson rule) having the same PFA spec β will not result in a better PCD. Let's denote $\tilde{X}$ as our alternate decision rule, and $\hat{X}$ as our regular N-P decision rule. Mathematically, we would like to show that if $\mathbf{Pr}(\tilde{X} = 1|X = 0) \leq \beta$, then $\mathbf{Pr}(\tilde{X} = 1|X = 1) \leq \mathbf{Pr}(\hat{X} = 1|X = 1)$. To do so we need the following lemma:

Lemma 22.1.

$$(\hat{X} - \tilde{X})(L(y) - \lambda) \geq 0$$

Proof. if $L(y) > \lambda$, then $\hat{X} = 1$ which implies that $\hat{X} - \tilde{X} \geq 0$. Otherwise, if $L(y) < \lambda$, then $\hat{X} = 0$ which implies $\hat{X} - \tilde{X} \leq 0$ and the inequality still holds. Finally, if $L(y) = \lambda$, then we get $0 \geq 0$. □

If we expand the equation in the lemma, we get the equivalent:

$$\hat{X}L(y) - \tilde{X}L(y) \geq \lambda\hat{X} - \lambda\tilde{X}$$

Now, we take $\mathbf{E}[\cdot|X = 0]$ on both sides of the above equation to get:

$$\mathbf{E}[\hat{X}L(y)|X = 0] - \mathbf{E}[\tilde{X}L(Y)]|X = 0] \geq \lambda \left(\mathbf{E}[\hat{X}|X = 0] - \mathbf{E}[\tilde{X}|X = 0]\right)$$

$$= \lambda \left(\mathbf{Pr}(\hat{X} = 1|X = 0) - \mathbf{Pr}(\tilde{X} = 1|X = 0)\right) \geq 0$$

Here, we used the fact that $\lambda > 0$ and $\mathbf{Pr}(\hat{X} = 1|X = 0) = \beta$ while $\mathbf{Pr}(\tilde{X} = 1|X = 0) \leq \beta$. So now we have that

$$\mathbf{E}[\hat{X}L(y)|X = 0] \geq \mathbf{E}[\tilde{X}L(y)|X = 0]$$

Now, we have the LHS of the above equation is

$$LHS = \int g(Y)L(y)f_{Y|X}(y|0)dy$$
$$= \int g(y)\frac{f_{Y|X}(y|1)}{f_{Y|X}(y|0)}f_{Y|X}(y|0)dy$$
$$= \int g(y)f_{Y|X}(y|1)dy = \mathbf{E}[\hat{X}|X=1] = \mathbf{Pr}(\hat{X}=1|X=1) = PCD$$

Using very similar logic, we can show that the RHS is the PCD for the alternate decision rule. Then we find that the N-P outperforms any other decision rule that satisfies the same specs!

□

22.2 Estimation

There are numerous applications involving needing to estimate quantities of interest.

1. Radar and Lidar
2. Multi-antenna wireless systems (MIMO systems: multiple input multiple output)
3. GPS
4. sensor networks
5. IoT (internet of things, every device has a sensor on it)
6. Machine Learning

Canonical Estimation Problem: You are given Y, and use your estimator produce an estimate $\hat{X}$, which we then compare to the actual value X to get some error in estimation Δ. Our goal is naturally to estimate X for Y as accurately as possible, i.e. we want $\mathbf{E}[\Delta^2]$ to be "small".

Definition 22.2. The **LLSE** (linear least-squares estimate) is an estimate where we constrain $\hat{X}$ to be a linear function of Y:

$$\hat{X} = a + bY$$

In the LLSE setting, we would like to minimize

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

We can also consider adding a quadratic term, i.e. allowing $\hat{X} = a + bY + cY^2$. This would be the QLSE (quadratic least-squares estimator). One can easily imagine adding more and more terms, and more generally one can consider what is known as the **MMSE**, which is just the best function $\hat{X}(Y)$, and not constrained to be linear or even a polynomial, but we are getting ahead of ourselves.

For the moment, let's assume we know the joint statistics of X and Y, and derive the LLSE. Recall our goal is to minimize $f(a,b) = \mathbf{E}[(X-\hat{X})^2] = \mathbf{E}[(X-a-bY)^2]$

$$\mathbf{E}[(X-a-bY)^2] = \mathbf{E}[X^2 + a^2 + b^2Y^2 - 2aX + 2abY - 2bXY]$$

Then we have

$$\frac{\partial f}{\partial a} = 2a - 2\,\mathbf{E}[X] + 2b\,\mathbf{E}[Y] = 0$$
$$\frac{\partial f}{\partial b} = 2b\,\mathbf{E}[Y^2] + 2a\,\mathbf{E}[Y] - 2\,\mathbf{E}[XY]$$

Solving these equations for a, b gives us our best linear estimate:

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

23 Lecture 23: Geometry of RV's: LLSE and MMSE

Agenda:

1. Recap of LLSE
2. Hilbert Space of Random Variables
3. MMSE

Recall that The goal of LLSE is to estimate X as the most accurate linear function of Y:

$$L[X|Y] = \hat{X} = a + bY$$

Last lecture we solved for a and b with calculus to find the LLSE formula:

$$L[X|Y] = a + bY = \mathbf{E}[X] + \frac{\mathbf{Cov}(X,Y)}{\mathbf{Var}(Y)}\left[Y - \mathbf{E}[Y]\right]$$

23.1 Hilbert Spaces

Without getting too into the details, a Hilbert space is roughly a "complete inner product vector space". A vector space V has $0 \in V$, is closed under addition, closed under scalar multiplication. For example $\mathbb{R}^n$ is a Hilbert space. More details on the formulation can be found in the notes on the course website. An *inner product space* is also equipped with an inner product $\langle \cdot, \cdot \rangle : V \times V \to (0, \infty)$ which satisfies symmetry, linearity, and positivity (if any of these words are not clear, just google the definition of an inner product or read the notes). An inner product also always induces an *norm* $\|\cdot\| : V \to (0, \infty)$ where $\|v\| = \sqrt{\langle v, v \rangle}$. Now, the most important concept for our purposes will be that of an **orthogonal projection onto a subspace**. Suppose we have some subspace U of our vector space. Then the orthogonal projection onto U is the map

$$P : V \to U \quad \text{s.t.} \quad P_U(y) := \arg\min_{x \in U} \|y - x\|$$

An orthogonal projection satisfies the properties that 1) $P_U(y) \in U$ and $y - P_U y \in U^\perp$. Here we have

$$\|y - x\|^2 = \|y - P_U(y) + P_U(y) - x\|^2 = \|y - P_U(y)\|^2 + 2\langle y - P_U(y), P_U(y) - x \rangle + \|P_U(y) - x\|^2$$

Now we note that $y - P_U(y) \in U^\perp$ and $P_U(y) - x \in U$ so therefore $\langle y - P_U(y), P_U(y) - x \rangle = 0$. So then we have

$$\|y - x\|^2 = \|y - P_U(y)\|^2 + \|P_U(y) - x\|^2 \geq \|y - P_U(y)\|^2$$

with equality if and only if $x = P_U(y)$!

Recall the **Gram-Schmidt process** which is used to "orthogonalize" an arbitrary basis of vectors. Suppose we have two vectors v_1 and v_2 and we would like to make them orthonormal while still spanning the same space. Then we can do the following:

1. Set $u_1 = \frac{v_1}{\|v_1\|}$
2. set $w_2 = v_2 - \langle v_2, u_1 \rangle u_1$
3. set $u_2 = \frac{w_2}{\|w_2\|}$

This process can easily be generalized to orthonormalize a general set of vectors.

23.2 Properties of LLSE

For convenience here is the LLSE again:

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

And here are some properties of it:

1. $\mathbf{E}[\hat{X}] = \mathbf{E}[X]$
   you should check this! Also this means that the LLSE is *unbiased*.
2. **Projection Property:** $\mathbf{Cov}(X - \hat{X}, Y) = \mathbf{Cov}(\Delta, Y) = 0$

   > **Exercise 23.1.** verify this!

   This means that the estimation error is *uncorrelated* with the observation.

Let's rederive $L[X|Y]$ using geometry. To do so, we need to define the vector space perspective of random variables:
Assume X, Y are zero-mean RVs with finite second moments. Then we have the following association of RVs as a vector space:

RV X	**Geometry**
random variable X	a vector
RVs X and Y	two vectors with an angle θ between them
$\mathbf{E}[XY]$	$\langle X, Y\rangle = \|X\|\|Y\|\cos(\theta)$
$\mathbf{E}[XY] = 0$	$\theta = \frac{\pi}{2}$
$\mathbf{E}[X^2]$	$\langle X, X\rangle = \|x\|^2$ which is the norm of X
$\rho = \frac{\mathbf{E}[XY]}{\sqrt{\mathbf{E}[X^2]}\sqrt{\mathbf{E}[Y^2]}}$	$\frac{\langle X,Y\rangle}{\|X\|\|Y\|} = \cos(\theta)$

Table 1: Correspondences between random variables and their geometry

Now we can derive $L[X|Y]$ geometrically. Our basis for the observation space Y is $\{1, Y\}$, where Y may have nonzero mean. This means that the vector representation of 1 may not be orthogonal to Y. We would like to project X onto this basis, but first it must be orthogonal! This is where Gram-Schmidt comes into play. 1 as a vector is already a unit vector. Our orthonormal Y becomes

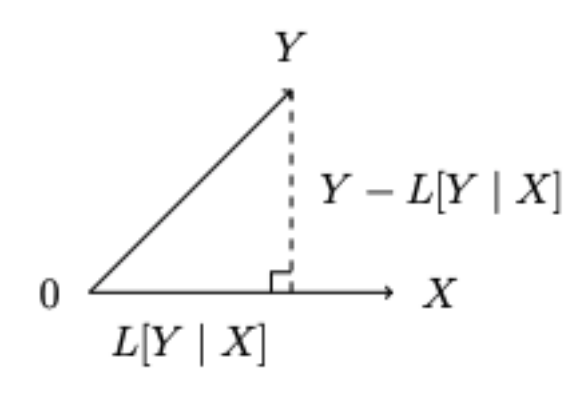
$$\bar{Y} = \frac{Y - \langle Y, 1\rangle}{\|Y - \langle Y, 1\rangle\|}$$

So $\{\bar{Y}, 1\}$ is an orthonormal basis, so then

$$\begin{aligned} P_{\{1,\bar{Y}\}}(X) &= P_{\{\bar{Y}\}}(X) + P_{\{1\}}(X) \\ &= \langle X, 1\rangle \cdot 1 + \langle X, \bar{Y}\rangle \cdot \bar{Y} \\ &= \mathbf{E}[X] + \mathbf{E}\left[\frac{X(Y - \mathbf{E}[Y])}{\sqrt{\mathbf{Var}(Y)}}\right]\frac{(Y - \mathbf{E}[Y])}{\sqrt{\mathbf{Var}(Y)}} \\ \implies L[X|Y] &= \mathbf{E}[X] + \frac{\mathbf{Cov}(X,Y)}{\mathbf{Var}(Y)}(Y - \mathbf{E}[Y]) \end{aligned}$$

Which is the answer we saw before!

Example 23.2. Let's look at when X and Y are both zero mean. Consider the picture:



In the picture, we have flipped Y and X from our normal convention, but this builds character. We have

$$L[Y|X] = bX = Proj_Y X = \frac{\langle X, Y\rangle}{\|X\|^2} X = \frac{\mathbf{E}[XY]}{\mathbf{E}[X^2]} X = \frac{\mathbf{Cov}(X,Y)}{\mathbf{Var}(X)} X$$

Which is exactly what we expected

Example 23.3. Suppose $Y = \alpha X + Z$ where X, Z are zero-mean. What is $L[X|Y]$? We have

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

And we have

$$\mathbf{E}[XY] = \mathbf{E}[X(\alpha X + Z)] = \alpha\,\mathbf{E}[X^2] + \mathbf{E}[XZ] = \alpha\,\mathbf{E}[X^2]$$

and

$$\mathbf{E}[Y^2] = \mathbf{E}[(\alpha X + Z)^2] = \alpha\,\mathbf{E}[X^2] + \mathbf{E}[Z^2]$$

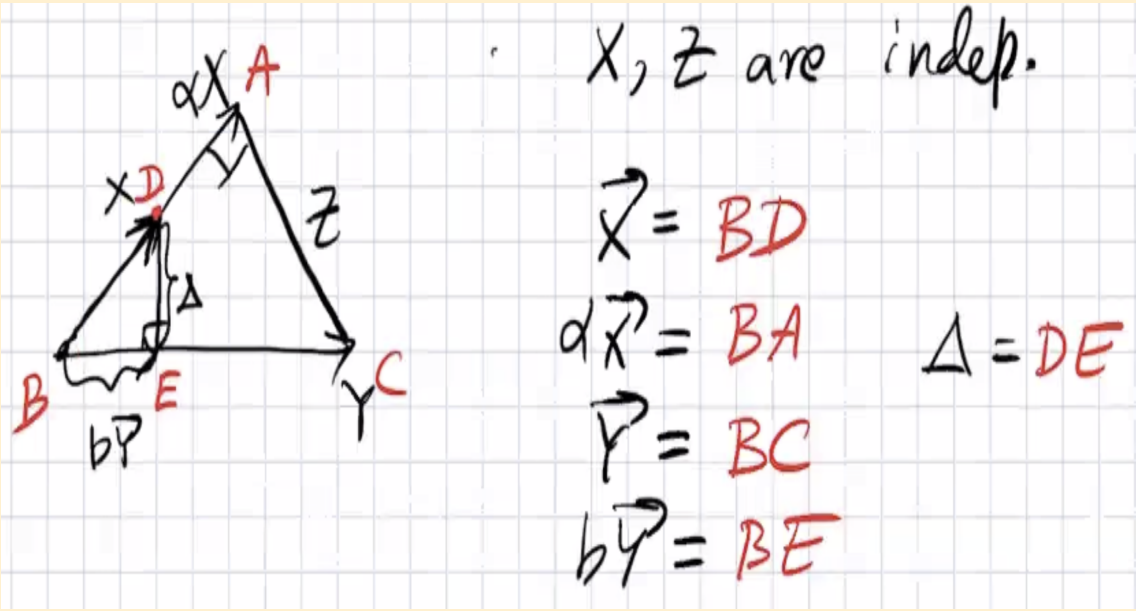
So finally, we have

$$L[X|Y] = \frac{\alpha\,\mathbf{E}[X^2]}{\alpha^2\,\mathbf{E}[X^2] + \mathbf{E}[Z^2]} \cdot Y = \frac{\alpha^{-1} Y}{1 + \frac{1}{SNR}}$$

where SNR stands for the *signal to noise ratio* and it is $\frac{\text{signal power}}{\text{noise power}} = \frac{\alpha^2\,\mathbf{E}[X^2]}{\mathbf{E}[Z^2]}$ Remarks:

1. If SNR $\gg 1$, then $L[X|Y] \approx \frac{1}{\alpha}$
2. If SNR $\ll 1$, then $L[X|Y] \approx 0$

Both of these should make intuitive sense. We can also think about this problem geometrically:



Now, we would like to find BE. But we can notive that the triangles BDE and BAC are similar triangles! This implies that

$$\frac{BE}{BD} = \frac{BA}{BC}$$

$$\implies BE = \frac{BA \cdot BD}{BC} \implies b\|Y\| = \frac{\alpha\|X\|\|X\|}{\|Y\|}$$

$$\implies b = \frac{\alpha\|X\|^2}{\|Y\|^2} = \frac{\alpha\,\mathbf{E}[X^2]}{\alpha^2\,\mathbf{E}[X^2] + \mathbf{E}[Z^2]}$$

Which is exactly the same as before, except this time we only used geometry!

A final remark before we end today: in the general LLSE problem with zero mean X and Y, we have

$$Error = \mathbf{E}[(X - \hat{X})^2] = \mathbf{E}[\Delta^2] = \mathbf{E}[X^2]\sin^2(\theta)$$

$$= \mathbf{E}[X^2](1 - \cos^2(\theta)) = \mathbf{E}[X^2](1 - \rho^2)$$

$$= \mathbf{E}[X^2]\left(1 - \frac{\mathbf{Cov}^2(X,Y)}{\mathbf{E}[X^2]\,\mathbf{E}[Y^2]}\right)$$

$$= \mathbf{E}[X^2] - \frac{\mathbf{Cov}^2(X,Y)}{\mathbf{E}[Y^2]} = \sigma_x^2 - \frac{\mathbf{Cov}^2(X,Y)}{\sigma_Y^2}$$

And we can notice that the above expression goes to σ_X^2 as $\mathbf{Cov}(X,Y)$ goes to zero. This makes intuitive sense, as the less related X is to Y, the less information we can gain about X by observing Y, so our error will remain nearly as large as our uncertainty about X.

24 Lecture 24: MMSE and Jointly Gaussian Random Variables

Agenda

1. Wrapup of LLSE/Geometry
2. Connection with Linear Regression
3. MMSE and its geometry
4. Jointly Gaussian Random Variables

24.1 Recap of LLSE

Recall that we derived the LLSE formula geometrically last time as the orthogonal projection. of X onto the the subspace of linear functions of the form $a + bY$. We also saw that:

1. The LLSE is *unbiased*, meaning $\mathbf{E}[\hat{X}] = \mathbf{E}[X]$
2. The estimation error Δ is *uncorrelated* with Y, meaning $\mathbf{Cov}(\Delta, Y) = \mathbf{E}[\Delta Y] = 0$

We have the proof that the LLSE is optimal using our standard trick:

Proof. let $g(Y) = a + bY$. Then we have

$$\|X - g(Y)\|^2 = \|X - L[X|Y] + L[X|Y] - g(Y)\|^2$$

$$= \|X - L[X|Y]\|^2 + 2\langle X - L[X|y], L[X|Y] - g(Y)\rangle + \|L[X|Y] - g(Y)\|^2$$

now note that $X - L[X|y] = \Delta$, and furthermore $L[X|Y] - g(Y)$ is some linear function of Y, so by our second fact above, we have $\langle X - L[X|y], L[X|Y] - g(Y)\rangle = 0$. Then,

$$\|X - g(Y)\|^2 = \|X - L[X|Y]\|^2 + \|L[X|Y] - g(Y)\|^2 \geq \|X - L[X|Y]\|^2$$

which tells us that the LLSE is our best linear estimator

□

We have derived all of this in the scalar case, but it is important to note that all of this holds essentially in the case when X and Y are vectors as well. In particular, we have:

$$L[X|Y] = \mathbf{E}[X] + \mathbf{Cov}(X, Y)\Sigma_Y^{-1}(Y - \mathbf{E}[Y])$$

Where $\Sigma_Y = \mathbf{E}[(Y - \mathbf{E}[Y])(Y - \mathbf{E}[Y])^T]$ and $\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])^T]$.

The details can be found in Walrand. But it looks almost exactly the same as the scalar case!

Remark 24.1. So far, we have assumed a *Bayesian* framework, assuming complete knowledge of our joint distribution of X and Y. If we take a non-probabilistic "data-driven" perspective, then this just becomes linear regression.

- We assume we have access to samples $\{(x_1, y_1), ..., (x_k, y_k)\}$
- Goal: construct $g(Y) = a + bY$ such that

$$\mathcal{E}(a, b) = \frac{1}{k}\sum_{i=1}^{k} |x_i - a - by_i|^2$$

To solve this, we take the partial derivatives of $\mathcal{E}$ wrt a and b and set them equal to zero. This is a good exercise, and you will find that we get the the same exact formula as we have previously derived! The only difference is that the means, variances, and covariance will be replaced with the

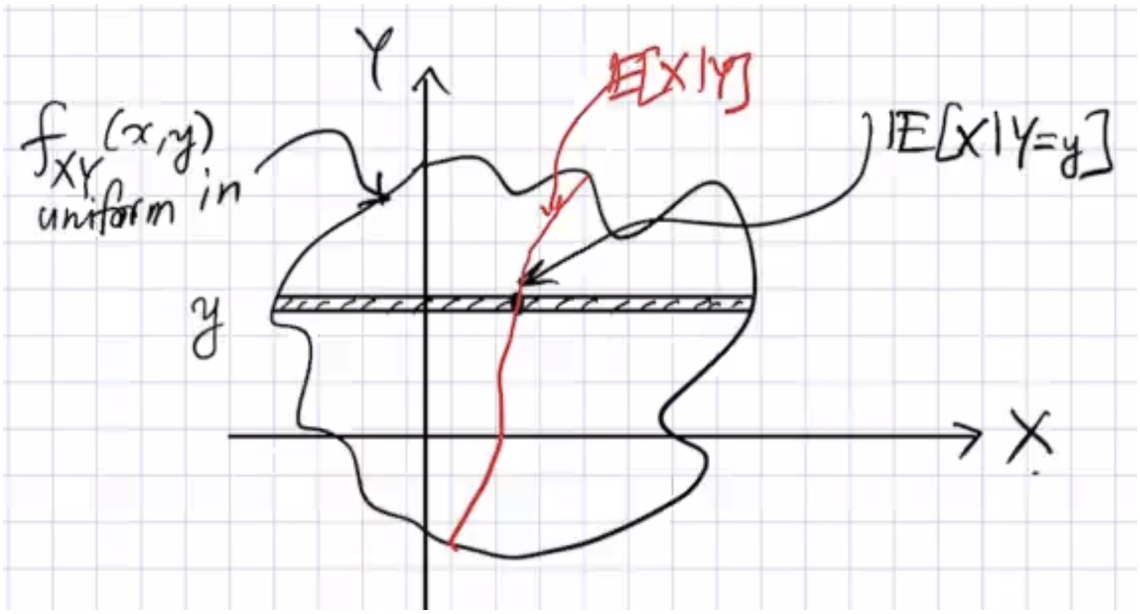
empirical means, variance, and covariance. This means that by the strong law of large numbers, that linear regression converges to the LLSE.

24.2 MMSE Estimation

Once again the goal is to minimize $\mathbf{E}[(X - \hat{X}(Y))^2]$, but now $\hat{X}(Y)$ can be *any* function of Y.
Intuition: Suppose we don't even observe Y. What is the MMSE estimate of X given nothing? Well it should be $\mathbf{E}[X]$! If this is not convincing enough, you can show that

$$\arg\min_c \mathbf{E}[(X-c)^2] = \mathbf{E}[X]$$

What if Y is given? The natural extension would be to guess that $MMSE[X|Y] = \hat{X}(Y) = \mathbf{E}[X|Y]$



In the above picture, the red line represents a kind of "center of mass" of the blob given some observation y. If we are given y in the above picture, then x could be any of the values inside the striped bar, but we choose the one that intersects with the center of mass line.

Before, we performed an orthogonal projection onto the space spanned by Y and 1 to get the LLSE of X given Y. Now, intuitively, we are projecting onto the space of all possible functions of Y, rather than just linear ones. We would like to show that $\mathbf{E}[X|Y]$ is indeed this orthogonal projection.

Theorem 24.2. *The MMSE of X given Y is given by*

$$g(Y) = \mathbf{E}[X|Y]$$

Lemma 24.3. *1. for all functions $\varphi(\cdot)$,*

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

This means that our projection is indeed orthogonal!

2. *If there exists a function $g(Y)$ such that*

$$\mathbf{E}[(X - g(Y))\varphi(Y)] = 0$$

for all $\varphi(\cdot)$, then $g(Y) = \mathbf{E}[X|Y]$
Together, these two claims mean that the orthogonal projection is both optimal and unique.

Proof. 1. We would like to show that $\mathbf{E}[\Delta\varphi(Y)] = 0$ for all φ

$$\Leftrightarrow \mathbf{E}[X\varphi(Y)] = \mathbf{E}[\mathbf{E}[X|Y]\varphi(Y)]$$

But this is immediate as we have

$$\mathbf{E}[\mathbf{E}[X|Y]\varphi(Y)] = \mathbf{E}[\mathbf{E}[\varphi(Y)X|Y]] = \mathbf{E}[\varphi(Y)X]$$

Where we have used the fact that $\varphi(Y)$ can be treated as a constant when Y is given, and then we have used iterated expectation.

2. For the proof of this see Walrand, but it is just algebra

□

Now we can prove the theorem:

Proof. The idea of the proof is the same as in the LLSE case. We have:

$$\mathbf{E}[|X - h(Y)|^2] = \mathbf{E}[|X - \mathbf{E}[X|Y] + \mathbf{E}[X|Y] - h(Y)|^2]$$

$$= \mathbf{E}[|X - \mathbf{E}[X|Y]|^2] + \mathbf{E}[|\,\mathbf{E}[X|Y] - h(Y)|^2] + 2\,\mathbf{E}[(X - \mathbf{E}[X|Y])(\mathbf{E}[X|Y] - h(Y))]$$

Again, we note that $X - \mathbf{E}[X|Y] = \Delta$ is our error term, and that $\mathbf{E}[X|Y] - h(Y)$ is just some function of Y. By the lemma, Δ is orthogonal to any function of Y so the term $2\,\mathbf{E}[(X - \mathbf{E}[X|Y])(\mathbf{E}[X|Y] - h(Y))] = 0$ and we have

$$\mathbf{E}[|X - h(Y)|^2] = \mathbf{E}[|X - \mathbf{E}[X|Y]|^2] + \mathbf{E}[|\,\mathbf{E}[X|Y] - h(Y)|^2] \geq \mathbf{E}[|X - \mathbf{E}[X|Y]|^2]$$

with equality iff $h(Y) = \mathbf{E}[X|Y]$. □

Consider the following picture (which can be found in Walrand) for intuition:

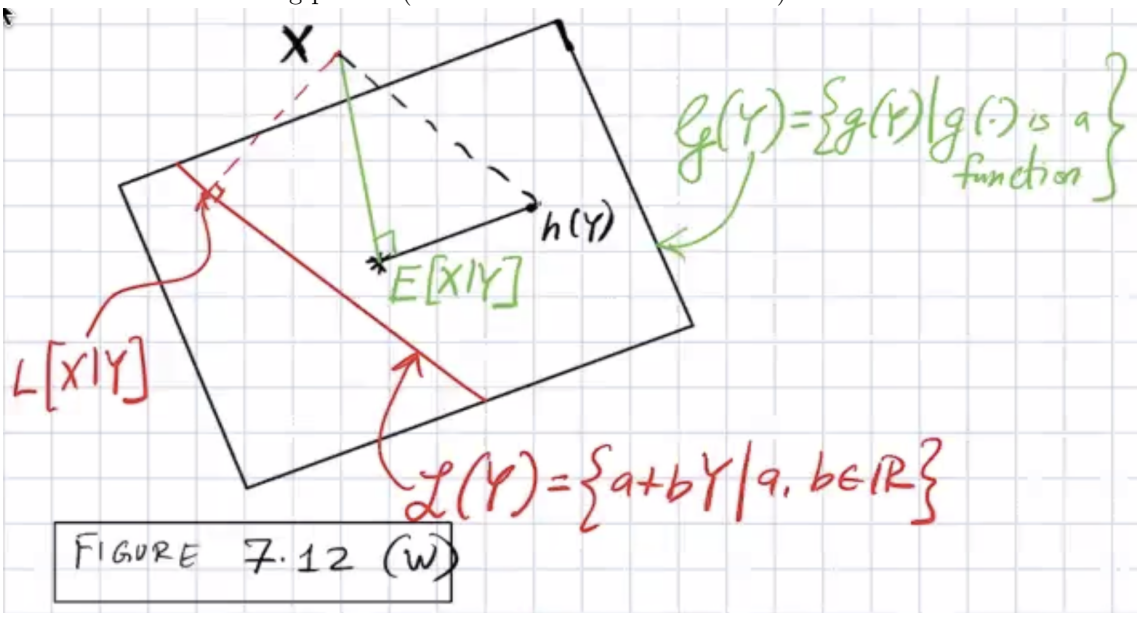


FIGURE 7.12 (W)

The above picture is highlighting that the set of linear estimators is just a small subspace of all possible estimators. In general, it does not happen that $MMSE[X|Y]$ equals the $L[X|Y]$. However, it does sometimes happen, and when it does happen this means that the best possible estimator of X given Y is a linear function.

Example 24.4. Suppose we know that Y is uniformly distributed between -1 and 1, and we are told that $X = Y^2$. Then is it clear that $\mathbf{E}[X|Y] = \mathbf{E}[Y^2|Y] = Y^2$ and this is the MMSE. However,

$$L[X|Y] = \mathbf{E}[X] + \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}Y$$

But $\mathbf{E}[XY] = \mathbf{E}[Y^3] = 0$ so $L[X|Y] = \mathbf{E}[X] = \frac{1}{3}$. In this case it is clear that the LLSE and the MMSE do not coincide.

Example 24.5. Suppose X, Y are i.i.d. RVs and you observe $X+Y$ and we want to estimate X given $X+Y$. We have by symmetry (since X and Y are indistinguishable RVs

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

but we also know that $\mathbf{E}[X+Y|X+Y] = X+Y$, so then it is clear by linearity of expectations that

$$\mathbf{E}[X|X+Y] = \frac{X+Y}{2}$$

Note that this is a special case when the LLSE and MMSE collide, i.e. that the MMSE is linear.

24.3 Jointly Gaussian RVs

We will start with a very important theorem:

Theorem 24.6. *If X and Y are jointly gaussian random variables, then*

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

First, we have to define what jointly gaussian means

Definition 24.7 (Jointly Gaussian RVs)**.** Let $f(X_1, X_2)$ be the joint pdf of X_1 and X_2. If $f(X_1, X_2)$ is such that

$$\alpha_1 X_1 + \alpha_2 X_2 \sim \text{Normal pdf}$$

for all $\alpha_1, \alpha_2 \in \mathbb{R}$. Alternatively, X_1 and X_2 are said to be JG RVs if every linear combination $\alpha_1 X_1 + \alpha_2 X_2$ is a normal pdf.

As a fun bonus, we also prove the law of total variance (from way back in the beginning of the course) geometrically:

Remark 24.8 (Geometric interpretation of the Law of Total Variance)**.** First, we perform some manipulation that will be useful later:

$$\begin{aligned}\mathbf{E}[\mathbf{Var}(X|Y)] &= \mathbf{E}[\mathbf{E}[(X - \mathbf{E}[X|Y])^2|Y]] \\ &= \mathbf{E}[(X - \mathbf{E}[X|Y])^2] \\ &= \mathbf{Var}(X - \mathbf{E}[X|Y])\end{aligned}$$

Now we will see that the law of total variance is simply an expression of the pythagorean theorem!

Consider:

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

In the geometric representation of RVs, $\mathbf{E}[X|Y]$ is as we know a projection of X onto the space of functions of Y. Then, we have that $\mathbf{Var}(X)$ is the square of the length of X (which is the standard deviation), and $\mathbf{Var}(X - \mathbf{E}[X|Y]) + \mathbf{Var}(\mathbf{E}[X|Y])$ (which are orthogonal by the definition of the MMSE) is the sum of the squares of the lengths of the two vectors that add to form X.

25 Lecture 25: Jointly Gaussian Random Variables and Scalar Kalman Filter

Agenda:

1. Jointly Gaussian RVs
2. Orthogonal Updates and Kalman Filters

25.1 Jointly Gaussian RVs

Recall that we see at the end of last lecture that:

Theorem 25.1. *If X and Y are jointly gaussian random variables, then*

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

First, we have to define what jointly gaussian means

Definition 25.2 (Jointly Gaussian RVs)**.** Let $f(X_1, X_2)$ be the joint pdf of X_1 and X_2. If $f(X_1, X_2)$ is such that

$$\alpha_1 X_1 + \alpha_2 X_2 \sim \text{Normal pdf}$$

for all $\alpha_1, \alpha_2 \in \mathbb{R}$. Alternatively, X_1 and X_2 are said to be JG RVs if every linear combination $\alpha_1 X_1 + \alpha_2 X_2$ has a normal pdf.

There is a completely equivalent definition that goes as follows. $Y_1, ..., Y_n$ are considered **Jointly Gaussian** if

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

has a multivariate normal pdf, which we will define below.

Definition 25.3. A random vector $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$ is JG with mean μ_y and covariance matrix Σ_Y if

$$Y = AX + \mu_Y$$

Where $\Sigma_Y = AA^T$ and $X \sim \mathcal{N}(0, I)$ so X is a just a vector of independent standard gaussians.

Here the covariance matrix Σ_Y describes the pairwise covariance between every element of the vector Y. It is easy to see then that Σ_Y is a symmetric matrix, since $(\Sigma_Y)_{i,j} = \mathbf{Cov}(Y_i, Y_j) = \mathbf{Cov}(Y_j, Y_i) = (\Sigma_Y)_{j,i}$. Here we can compute the covariance matrix $\mathbf{Var}(Y) = \Sigma_Y$ as follows:

$$\mathbf{Var}(Y) = \mathbf{E}[(Y - \mu_Y)(Y - \mu_Y)^T] = \mathbf{E}[AX(AX)^T] = A\,\mathbf{E}[XX^T]A^T = AA^T$$

Since $\mathbf{E}[XX^\top] = I$ since X is a standard multivariate normal distribution.

Example 25.4. for $n = 2$, suppose

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}}_{A} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

Then we can compute

$$\Sigma_Y = AA^\top = \begin{bmatrix} 5 & -1 \\ -1 & 2 \end{bmatrix}$$

Theorem 25.5. *If $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$ then its pdf is given by*

$$f_Y(y) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma_Y)}} \exp\left(-\frac{1}{2}(Y-\mu_Y)^T\Sigma_Y^{-1}(Y-\mu_Y)\right)$$

Note: *the level curves of this joint pdf are ellipses, and we will see some pictures of this later. The level curves are the set of y's that give $f_Y(y) = c$ for some constant c.*

Example 25.6. Again, we suppose $n = 2$, and furthermore suppose Y_1 and Y_2 are uncorrelated. Then we have:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}}_{A} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

Then we also have

$$\Sigma_Y = AA^\top = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \text{and} \quad \Sigma_Y^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix}$$

Now plugging into our pdf for the multivariate gaussian we have

$$f_Y(y_1, y_2) = \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2} \exp\Big(\underbrace{-\frac{1}{2}Y^\top\Sigma_Y^{-1}Y}_{-\frac{1}{2}\left[y_1^2/\sigma_1^2 + y_2^2/\sigma_2^2\right]} \Big)$$

$$= \Big(\frac{1}{\sqrt{2\pi}\sigma_1}\exp\Big(\frac{-y_1^2}{2\sigma_1^2}\Big)\Big)\Big(\frac{1}{\sqrt{2\pi}\sigma_2}\exp\Big(\frac{-y_2^2}{2\sigma_2^2}\Big)\Big) = f_{Y_1}(y_1)f_{Y_2}(y_2)$$

The above calculations show that Y_1 and Y_2 are actually independent! This is quite remarkable, because we only assumed that they were uncorrelated in the beginning of this example, and recall that in general two random variables being uncorrelated does not necessarily mean that they are independent. But in the case of jointly gaussian random variables, it does!

The above example essentially proves (a special case of) the following very important lemma:

Lemma 25.7. *If Y_1 and Y_2 are uncorrelated and jointly gaussian, then they are also independent.*

Remark 25.8. This is not super important, but we discuss here intuitively why the level sets of gaussians are ellipses. To do so, we just look at the 2-D case. Recall if Y_1 and Y_2 are uncorrelated (and therefore also independent), then we have

$$f_Y(y_1, y_2) = \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2} \exp\Big(-\frac{1}{2}\Big[\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2}\Big]\Big)$$

Then clearly level sets for this function correspond to solutions to the equation:

$$\frac{y_1^2}{2\sigma_1^2} + \frac{y_2^2}{2\sigma_2^2} = c$$

For the case when $\sigma_1 = \sigma_2$, the above equation just corresponds to a circle in the y_1, y_2 plane. If on the other hand $\sigma_1^2 = 2\sigma_2^2$, then the equation corresponds to an ellipse with its longer axis along the y_1 axis. An analogous phenonmenon occurs when $\sigma_2^2 = 2\sigma_1^2$.
The more complicated case is when we don't actually assume that Y_1 and Y_2 are uncorrelated. For example, if we assume as in the example above that

$$\Sigma_Y = \begin{bmatrix} 5 & -1 \\ -1 & 2 \end{bmatrix}$$

Then we can compute (with some work not shown) the pdf of Y as

$$f_Y(y) = \frac{1}{2\pi \cdot 3} \exp\Big(\underbrace{\frac{-1}{2} Y^\top \Sigma_Y^{-1} Y}_{-\left[\frac{2y_1^2 - 2y_1 y_2 + 5y_2^2}{18}\right]} \Big)$$

Now, the equation in the exponent (if we complete the square a couple times) is

$$\left(\frac{y_1 - a_1}{b_1}\right)^2 + \left(\frac{y_2 - a_2}{b_2}\right)^2 = c$$

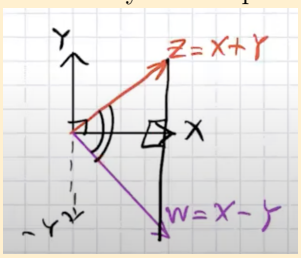
Which if we set equal to some constant c, is just the equation for a rotated ellipse in the y_1, y_2 plane. Moreover, the axes for this rotated ellipse are given by the eigenvectors of the inverse covariance matrix Σ_Y^{-1}, but this is getting way out of scope.

Recall the first definition of JG random variables. $(X_1, ..., X_n)$ are JG if and only if $\alpha_1 X_1 + ... + \alpha_n X_n$ is a univariate normal for everal value of $\alpha_1, ..., \alpha_n \in \mathbb{R}$.

Example 25.9. Suppose $X, Y \sim \mathcal{N}(0, 1)$ i.i.d. RV's. Let $Z = X + Y$ and $W = X - Y$. Are Z and W independent or not?
We have

$$\mathbf{Cov}(Z, W) = \mathbf{E}[ZW] - \mathbf{E}[Z]\,\mathbf{E}[W] = \mathbf{E}[(X+Y)(X-Y)] = \mathbf{E}[X^2] - \mathbf{E}[Y^2] = 0$$

Which means that Z and W are uncorrelated. However, they are also jointly gaussian, so by the lemma they are independent! This can be nicely visualized/proven geometrically:



In the above figure, we have plotted X and Y geometrically in the hilbert space of random variables we introduced earlier. Note that they are orthogonal by definition. Furthermore, not that the triangles $(0, X, Z)$ and $(0, X, W)$ are similar iscoceles triangles. These two facts tell us that the angle between Z and W must be 90 degrees, which means Z and W must also be orthogonal.

Remark 25.10. Linear combinations of JG RVs are still JG. This follows directly from the definition of jointly gaussian.

Example 25.11. Let

$$W = \begin{cases} 1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$$

And we let $X \sim \mathcal{N}(0,1)$ and X, W are independent. We set $Y = WX$
Question 1: are X and Y uncorrelated?
We have $\mathbf{E}[XY] = \mathbf{E}[X^2W] = \mathbf{E}[X^2]\,\mathbf{E}[W] = 0$, so yes.
Question 2: What is the distribution of Y?
By symmetry, $Y \sim \mathcal{N}(0,1)$
Question 3: Are X and Y independent?
Clearly $Y = WX$ is not indepedendent of x.
So what is going on here? Doesn't this contradict our lemma?
Question 4: Are X and Y jointly gaussian?
Consider

$$Z = X + Y = \begin{cases} 2X & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases}$$

Which is *not* gaussian so we cannot apply the lemma, so there is no contradiction above.

Now, we probably the most important theorem from today's lecture:

Theorem 25.12. *If X and Y are jointly gaussian, then $\mathbf{E}[X|Y] = L[X|Y]$.*

Recall:

1. If (X, Y) are JG, then all linear combinations of X and Y are jointly gaussian.
2. If (X, Y) are JG and uncorrelated, then X, Y are independent

Proof.
1. First, we know that $X - L[X|Y] \perp Y$ by the projection property of LLSE.
2. If (X, Y) are JG so are linear combinations of X, Y, namely $(X - L[X|Y])$ and Y.
3. Now, we can conclude that $X - L[X|Y]$ and Y must be independent, by the lemma.
4. This implies that $X - L[X|Y]$ and $\varphi(Y)$ are independent for all functions $\varphi(\cdot)$.
5. This implies that $X - L[X|Y]$ is orthogonal to $\varphi(Y)$ (since independence implies uncorrelated.)
6. This means that $L[X|Y]$ must be our MMSE estimate $\mathbf{E}[X|Y]$, since $X - \mathbf{E}[X|Y]$ is unique and must be orthogonal to every function of Y.

□

25.2 Kalman Filter

In general, there are many different forms of **inference** that pop up in the literature/real life settings:

1. **Filtering:** Filtering is done in real time, for example tracking position in real time. We are trying to guess $\hat{X}$ given a sequence of observations $Y_1, ..., Y_n$
2. **Prediction:** We are given T observations $Y_1, ..., Y_T$ and we are trying to predict what we will observe in the future, i.e. $Y_{T_i}, i = 1, 2, ...$ This is a lot of machine learning and can be used for radar tracking, stock market predictions, predictive coding (speech, image, video, etc.).
3. **Smoothing:** Infer $\hat{X}_t$ for $t \leq T$ given observations $Y_1, ..., Y_T$. This is offline (which is the biggest difference between smoothing and filtering), i.e. inferring the cause of a car crash by post processing a video.

4. **Max Likelihood State Estimation (MLSE)** Given $Y_0, ..., Y_T$, we want to output the most likely possible *sequence* of states $\hat{X}_0, ..., \hat{X}_T$. The big difference between this and smoothing is we are trying to get the entire sequence, not just one particular timestep. Examples of this include the convolutional coding (the Viterbi algorithm), auto-correct, and speech recognition.

The **Kalman Filter** is of course a filtering algorithm to update the estimate of the *state* $X(n)$ or X_n of a system. The system has a **state** $X(n)$ and an **observation** (or output) $Y(n)$ at time $n = 0, 1, ...$ according to the **State Space Equations:**

$$X(n+1) = AX(n) + V(n)$$
$$Y(N) = CX(n) + W(n)$$

In the above equations, there is sometimes an optional control input $BU(n)$ added to the equation for $X(n+1)$, but we won't consider that case in this class. Here, $W(n)$ and $V(n)$ are some zero mean and orthogonal noise terms. We denote the $\mathbf{Cov}(V_n) = \Sigma_V$ and $\mathbf{Cov}(W_n) = \Sigma_W$.
The objective of the kalman filter is to estimate $\hat{X} = L[X(n)|Y(0), ..., Y(n)]$

Theorem 25.13 (Kalman Filter)**.**

$$\hat{X}_{n|n} = A\hat{X}_{n-1|n-1} + k_n\Big(Y_n - CA\hat{X}_{n-1|n-1}\Big) \tag{1}$$
$$k_n = \Sigma_{n|n-1}C^\top\Big[C\Sigma_{n|n-1}C^\top + \Sigma_W\Big]^{-1} \tag{2}$$
$$\Sigma_{n|n-1} = A\Sigma_{n-1|n-1} + \Sigma_V \tag{3}$$
$$\Sigma_{n|n} = (I - k_nC)\Sigma_{n|n-1} \tag{4}$$

where

- $\Sigma_W = \mathbf{Cov}(W_n)$
- $\Sigma_V = \mathbf{Cov}(V_n)$
- $\Sigma_{n|n-1} = \mathbf{Cov}(\underbrace{X_n - A\hat{X}_{n-1|n-1}}_{\Delta_{n|n-1}})$
- $\Sigma_{n|n} = \mathbf{Cov}(\underbrace{X_n - \hat{X}_{n|n}}_{\Delta_{n|n}})$

Don't worry if the above notations and equations don't make much sense yet. We haven't explained a lot of it, we are just stating it above for the general vector case. Next lecture, we will derive the scalar case geometrically, which looks entirely analogous to the vector case and should hopefully provide some much needed intuition about what is going on with the Kalman filter.

26 Lecture 26: Kalman Filter

Agenda:

1. Recap of Kalman Filter setup
2. Orthogonal Updates
3. Derivation of the scalar Kalman filter using geometry.

In the scalar case, our state-space equations are as follows:

$$X_n = aX_{n-1} + V_{n-1}$$

$$Y_n = cX_n + W_n$$

However, we note here that without loss of generality, we can divide the second equation by c which yields the equation $Y_n/c = X_n + W_n/c$. This can be though of as just a different observation with a different noise on the X_n variable, so from now on WLOG we just set $c = 1$ and our equations are:

$$X_n = aX_{n-1} + V_{n-1}$$

$$Y_n = X_n + W_n$$

Recall that the objective of a kalman filter is to estimate our state at time n, X_n in an online fashion using our observations $Y_1, ..., Y_n$. Our estimate, since we are in Gaussian land, will be $\hat{X}_n = L[X_n|Y_n, ..., Y_1]$. Lets recall some notation:

- $\hat{X}_{n|n} = L[X_n|Y_1, ..., Y_n]$ is our estimate of X_n at time n
- Similarly, $\hat{X}_{n|n-1} = L[X_n|Y_1, ..., Y_{n-1}]$ is the LLSE estimate of X_n at time $n-1$.
- $\Delta_{n|n} = X_n - \hat{X}_{n|n}$
- $\Delta_{n|n-1} = X_n - \hat{X}_{n|n-1}$
- $\mathbf{E}[\Delta^2_{n|n}] = \sigma^2_{n|n}$
- $\mathbf{E}[\Delta^2_{n|n-1}] = \sigma^2_{n|n-1}$
- $\mathbf{E}[V_n^2] = \sigma_V^2$
- $\mathbf{E}[W_n^2] = \sigma_W^2$

26.1 Orthogonal Updates

Recall when we were looking at the LLSE the following lemma:

Lemma 26.1. *If X, Y, Z are zero-mean and $\mathbf{E}[YZ] = 0$ (so Y is orthogonal to Z), then*

$$L[X|Y, Z] = L[X|Y] + L[X|Z]$$

However, this only works if Y, Z are orthogonal. For the general case, we have the following theorem:

Theorem 26.2. *Again suppose X, Y, Z are zero-mean, but this time, Y and Z could be correlated. We have*

$$L[X|Y, Z] = L[X|Y] + L[X|\tilde{Z}]$$

Where $\tilde{Z} = Z - \text{Proj}_Y Z = Z - L[Z|Y]$

Proof. The key here is to just note that $\tilde{Z}$ is orthogonal to Y, so we can just apply the previous lemma. $\square$

Now, we are going to give an outline of the kalman filter. The kalman filter is a recursive way of updating estimates based on prediction and update (after observing the new sample point), *one sample at a time.*

Recursively estimate X_n given $Y^n = Y_1, ..., Y_n$ by first finding $L[X_n|Y_1, ..., Y_{n-1}]$ at time $n-1$, and then updating the estimate based on the new observation Y_n at time n. Specifically, we have

$$L[\underbrace{X_n}_{X} | \underbrace{Y_1, ..., Y_{n-1}}_{Y}, \underbrace{Y_n}_{Z}] = L[\underbrace{X_n}_{X} | \underbrace{Y_1, ..., Y_{n-1}}_{Y}] + L[\underbrace{X_n}_{X} | \underbrace{\tilde{Y}_n}_{\tilde{Z}}]$$

Where we have drawn the an analogy to the lemma above. Here $\tilde{Y_n} = Y_n - L[Y_n|Y_1, ..., Y_{n-1}]$. Here, $\tilde{Y_n}$ represents the **innovative** part of our latest observation. It is the component of the latest observation that is orthogonal to all of our previous observations, so it represents new information. Below, we highlight the "predict then update" logic of the Kalman Filter:

$$\begin{aligned} L[X_1|Y_1] &\xrightarrow{\text{observe } Y_2} L[X_2|Y_1, Y_2] = \overbrace{L[X_2|Y_1]}^{\text{predict}} + \overbrace{L[X_2|\tilde{Y}_2]}^{\text{update}} \\ &\xrightarrow{\text{observe } Y_3} L[X_3|Y_1, Y_2, Y_3] = \overbrace{L[X_3|Y_1, Y_2]}^{\text{predict}} + \overbrace{L[X_3|\tilde{Y}_3]}^{\text{update}} \\ &\xrightarrow{\text{observe } Y_4} \dots \end{aligned}$$

With these tools, we are now ready to geometrically derive the scalar Kalman Filter

26.2 Scalar Derivation of Kalman filter

In the scalar case, our state space equations look like:

$$X_n = aX_{n-1} + V_n$$

$$Y_n = X_n + W_n$$

and we have the Kalman update equations:

$$\hat{X}_{n|n} = \hat{X}_{n|n-1} + k_n(Y_n - \hat{X}_{n|n-1}) \tag{5}$$

$$k_n = \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_W} \tag{6}$$

$$\sigma^2_{n|n-1} = a^2\sigma^2_{n-1|n-1} + \sigma^2_V \tag{7}$$

$$\sigma^2_{n|n} = (1 - k_n)\sigma^2_{n|n-1} \tag{8}$$

Let's derive these equations in the scalar case. Note that after time $n-1$, we know $\hat{X}_{n-1|n-1}$ and $\sigma^2_{n-1|n-1}$. Then, at time n, we get Y_n, and we do some updates to get $\hat{X}_{n|n}$ and $\sigma^2_{n|n}$. We can derive the first equation algebraically:

$$\hat{X}_{n|n} = L[X_n|Y_1, ..., Y_n] = L[X_n|Y_1, ..., Y_{n-1}] + L[X_n|Y_n - L[Y_n|Y_1, ..., Y_{n-1}]]$$

Which, letting $Y^{(n)} = Y_1, ..., Y$ can be concisely written as

$$\hat{X}_{n|n} = L[X_n|Y^{(n)}] = L[X_n|Y^{(n-1)}] + L[X_n|Y_n - L[Y_n|Y^{(n-1)}]]$$

$$= \hat{X}_{n|n-1} + L[X_n|\tilde{Y}_n] = \hat{X}_{n|n-1} + k_n\tilde{Y}_n$$

We can also very easily derive $\hat{X}_{n|n-1}$ algebraically, so we will do that now:

$$\begin{aligned}\hat{X}_{n|n-1} &= L[X_n|Y_1,...,Y_{n-1}] \\ &= L[aX_{n-1}+V_n|Y_1,...,Y_{n-1}] \\ &= aL[X_{n-1}|Y_1,...,Y_{n-1}]+L[V_n|Y_1,...,Y_{n-1}] \\ &= a\hat{X}_{n-1|n-1}\end{aligned}$$

Where in the last step we have used the fact that V_n is orthogonal zero-mean noise. Lastly, we note that

$$\begin{aligned}\tilde{Y}_n &= Y_n - L[Y_n|Y^{(n-1)}] \\ &= Y_n - L[X_n+W_n|Y^{(n-1)}] \\ &= Y_n - L[X_n|Y^{(n-1)}] - L[W_n|Y^{(n-1)}] \\ &= Y_n - \hat{X}_{n|n-1}\end{aligned}$$

Now, we are ready to examine some geometry to derive the remaining equations. To start doing this, consider the diagram below:

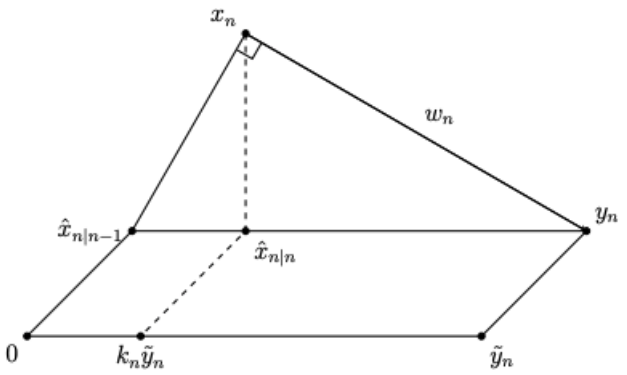


Figure 1: Geometry of the Kalman filter.

Try to think of the above diagram as kind of like a laptop which is at 90 degrees. The triangle $(\hat{x}_{n|n-1}, x_n, y_n)$ forms a plan which is orthogonal to the rectangle containing the origin at the bottom of the figure. There are a plethora of other things to notice about the above diagram.

- Note first that $\hat{X}_{n|n-1}$ is orthogonal to $(X_n - \hat{X}_{n|n-1})$. This is because $\hat{X}_{n|n-1}$ is the orthogonal projection of X_n onto the subspace spanned by $Y_1,...,Y_{n-1}$.
- $\tilde{Y}_n$ is orthogonal to $\hat{X}_{n|n-1}$. This is becauses $\tilde{Y}_n$ must be orthogonal to the subspace spanned by $Y_1,...,Y_{n-1}$, and as we mentioned already, $\hat{X}_{n|n-1}$ lives in that subspace.
- $k_n\tilde{Y}_n$ is the orthogonal projection of X_n onto $\tilde{Y}_n$. This is a little harder to see given the way the diagram is drawn (there are a lot of right angles in this diagram and it is hard to draw them all so that they actually look like right angles)
- W_n is orthogonal to everything, but in particular it is orthogonal to X_n, as denoted in the diagram.

We would like to figure out what k_n actually is based on this picture. To do this, we note that the triangles $(\hat{x}_{n|n-1}, \hat{x}_{n|n}, x_n)$ and $(x_n, \hat{x}_{n|n}, y_n)$ and $(\hat{x}_{n|n-1}, x_n, y_n)$ are all **similar triangles**. This means in particular that

$$\frac{\|k_n\tilde{Y}_n\|}{\|\Delta_{n|n-1}\|} = \frac{\|\Delta_{n|n-1}\|}{\|\tilde{Y}_n\|}$$

Where $\Delta_{n|n-1} = X_n - \hat{X}_{n|n-1}$ is the line segment connecting x_n and $\hat{X}_{n|n-1}$ in the above diagram. Rearranging, we can see that

$$k_n = \frac{\left\|\Delta_{n|n-1}\right\|^2}{\left\|\tilde{Y}_n\right\|^2} = \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_w}$$

Where in the last line we have used the pythagorean theorem to argue that $\left\|\tilde{Y}_n\right\|^2 = \left\|\Delta_{n|n-1}\right\|^2 + \|W_n\|^2$. Finally, we can verify the last equation by once again using similar triangles to note that:

$$\frac{\left\|\Delta^2_{n|n}\right\|}{\left\|\Delta^2_{n|n-1}\right\|} = \frac{\|W_n\|}{\left\|\tilde{Y}_n\right\|} = \frac{\left\|(1-k_n)\tilde{Y}_n\right\|}{\|W_n\|}$$

$$\implies \frac{\left\|\Delta^2_{n|n}\right\|^2}{\left\|\Delta^2_{n|n-1}\right\|^2} = \frac{\|W_n\|(1-k_n)\left\|\tilde{Y}_n\right\|}{\left\|\tilde{Y}_n\right\|\|W_n\|}$$

$$\implies \sigma^2_{n|n} = (1-k_n)\sigma^2_{n|n-1}$$

This is exactly what we wanted! For the last of the equations, we examine the following diagram:

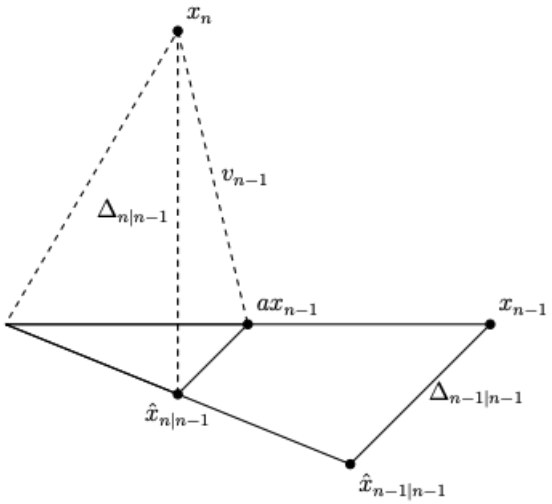


Figure 2: Geometry of the Kalman filter.

First, we note that $(x_n, ax_{n-1}, \hat{x}_{n|n-1})$ is a right triangle. We also note that $(0, \hat{x}_{n|n-1}, ax_{n-1})$ and $(0, \hat{x}_{n-1|n-1}, x_{n-1})$ are similar triangles. This implies that

$$\left\|ax_{n-1} - \hat{x}_{n|n-1}\right\| = a\left\|\Delta_{n-1|n-1}\right\|$$

We further have since V_{n-1} is orthogonal to $ax_{n-1} - \hat{x}_{n|n-1}$ that by the pythagorean theorem:

$$\left\|\Delta_{n|n-1}\right\|^2 = \left\|a\Delta_{n-1|n-1}\right\|^2 + \|V_n\|^2$$

$$\implies \sigma^2_{n|n-1} = a^2\sigma^2_{n-1|n-1} + \sigma^2_V$$

Which is exactly what we wanted! This was the last equation we have to verify, and so we are done. We have proved the scalar Kalman Filter entirely geometrically! The vector versions are just direct generalizations of the scalar equations, but they do not have such a nice geometric interpretation and must be derived algebraically.

Some final remarks:

1. At iteration n, the algorithm has inputs $\hat{X}_{n|n-1}, \sigma^2_{n-1|n-1}$ and new observation Y_n, and outputs $\hat{X}_{n|n}, \sigma^2_{n|n}$ for the next time step to use.
2. the kalman gain k_n and the errors $\sigma^2_{n|n-1}$ and $\sigma^2_{n|n}$ can be pre-computed, because they do not depend on the X and Y! Only $\hat{X}_{n|n}$ needs to be computed in real time.
3. This algorithm is easy to implement (just a few lines of code are required).
4. If V_n and W_n are Gaussians, then the Kalman filter is also giving us the MMSE estimate!

In the example below we highlight how we can precompute these variances without even seeing any data, and how often these variances converge to a "steady state"

Example 26.3. Suppose

$$X_n = \overbrace{\frac{1}{\sqrt{2}}}^{a} X_{n-1} + \overbrace{V_n}^{\mathcal{N}(0,1)}$$
$$Y_n = \underbrace{2}_{c} X_n + \underbrace{W_n}_{\mathcal{N}(0,1)}$$

Then we have from our Kalman filter equations:

$$k_n = \frac{2\sigma^2_{n|n-1}}{4\sigma^2_{n|n-1} + 1} \tag{9}$$
$$\sigma^2_{n|n-1} = \frac{1}{2}\sigma^2_{n-1|n-1} + 1 \tag{10}$$
$$\sigma^2_{n|n} = \sigma^2_{n|n-1}(1 - 2k_n) = \frac{\sigma^2_{n|n-1}}{4\sigma^2_{n|n-1} + 1} \tag{11}$$

1. When $n = 0$, we have $\hat{X}_{0|0} = 0$ and $\sigma^2_{0|0} = \mathbf{E}[X_0^2] = 2$, since $\sigma^2_X = \frac{1}{2}\sigma^2_X + \sigma^2_V \implies \sigma^2_X = 2\sigma^2_V = 2$.
2. When $n = 1$, we can compute
$$\sigma^2_{1|0} = \frac{1}{2}\sigma^2_{0|0} + 1 = 2$$
$$k_1 = \frac{2\sigma^2_{1|0}}{4\sigma^2_{1|0} + 1} = 4/9 = 0.444$$
$$\sigma^2_{1|1} = \sigma^2_{1|0}(1 - 2k_1) = 2/9 = 0.222$$
3. repeating this for $n = 2$, we find that $\sigma^2_{2|2} = 0.204$
4. Similarly for $n = 3$, we find $\sigma^2_{2|2} = 0.204$. We seem to have converged!

Indeed, if we take an analytical approach to the steady-state solution, using the equations we can derive

$$\sigma^2_{n|n} = \frac{\frac{1}{2}\sigma^2_{n-1|n-1} + 1}{2\sigma^2_{n-1|n-1} + 5}$$

and then using the fact that in steady state, $\sigma^2_{n|n} = \sigma^2_{n-1|n-1}$ to solve for $\lim_{n\to\infty} \sigma^2_{n|n}$ and we find that it equals 0.2037. Which had basically already achieved by time step $n = 3$!

27 Extra Content: Hidden Markov Models

In Kalman Filtering, we are given a sequence of observations coming at us in order $Y_1, ..., Y_n$ and we would like to come up with the most accurate prediction for our current actual state, $\hat{X_n}$, in real time. However, In many applications, such as autocorrect or speech recognition, we may want to actually guess the best estimate for our state at *all* time steps $\hat{X_1}, ..., \hat{X_n}$. This is, as we mentioned awhile back, known as **MLSE**, or Maximum Likelihood State Estimation. In this note we explore **Hidden Markov Models**, which are a way to model the underlying relationship between the true states and our observations. Formally, we have the following definition:

Definition 27.1. A **Hidden Markov Model** is a random sequence $\{(X(n), Y(n)); n \geq 0\}$ such that $X(n) \in \mathcal{X} = \{1, ..., N\}$ and $Y(n) \in \mathcal{Y} = \{1, ..., M\}$.
Here $\mathcal{X}$ and $\mathcal{Y}$ are simply our state and observation spaces, respectively. Here we also assume that our states $X(i)$ are characterized by a state transition matrix P with some initial distribution π_0. The *state observation model* Q tells us that when we are in state $x \in \mathcal{X}$, we observe $y \in \mathcal{Y}$ with probability $Q(x, y)$.

Below we have a nice visualization of an HMM. It represents an HMM that has been running for T time steps. The probabilities of the next states we visit is characterized by P, while the probabilities of the observations we see given our current state are characterized by our state observation model Q.

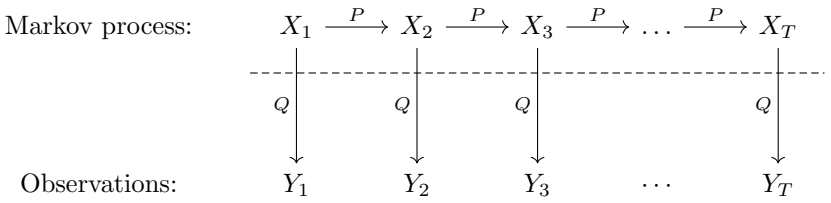


Figure 1: An example hidden markov chain that has run for T time steps.

In the context of speech recognition, the X_n may be segments of sentences or words, while the Y_n are sounds or something similar. Note that in this model, we are assuming that each successive state is only dependent on the previous state, and that each observation is only dependent on the current state. The structure of the language itself determines the relationship between the states X_i, while the relationship between X_i and Y_i can be speaker dependent. Our problem is as follows: suppose we have observed $\mathbf{Y}^n := (Y_0, ..., Y_n) = \mathbf{y}^n = (y_0, ..., y_n)$. What is the most likely sequence $\mathbf{X}^n$ that explains these observations? Formally, we want to find

$$MAP[\mathbf{X}^n | \mathbf{Y}^n = \mathbf{y}^n]$$

which is the same as finding the sequence $\mathbf{x}^n \in \mathcal{X}^{n+1}$ that maximizes

$$\mathbf{Pr}[\mathbf{X}^n = \mathbf{x}^n | \mathbf{Y}^n = \mathbf{y}^n] = \frac{\mathbf{Pr}[\mathbf{X}^n = \mathbf{x}^n \cap \mathbf{Y}^n = \mathbf{y}^n]}{\mathbf{Pr}[\mathbf{Y}^n = \mathbf{y}^n]}$$

Maximizing this expression is equivalent to maximizing the numerator. Now, from the definition of a the hidden markov model (which encodes our assumptions about the system), we have by applying Bayes rule multiple times along with the properties of Markov chains for $n = 1$:

$$\mathbf{Pr}[x_0, x_1, y_0, y_1]$$

$$= \mathbf{Pr}[y_1 | x_0, x_1, y_0] \, \mathbf{Pr}[x_1 | x_0, y_0] \, \mathbf{Pr}[y_0 | x_0] \, \mathbf{Pr}[x_0]$$

$$= \mathbf{Pr}[x_0] \, \mathbf{Pr}[y_0 | x_0] \, \mathbf{Pr}[x_1 | x_0] \, \mathbf{Pr}[y_1 | x_1] = \pi_0(x_0) Q(x_0, y_0) P(x_0, x_1) Q(y_1, x_1)$$

where in the last step we have substituted P and Q for our probabilities using the model we have set up. Extending this, it is not too hard to see then that

$$\arg \max_{\mathbf{x}^n \in \mathcal{X}^{n+1}} \mathbf{Pr}[\mathbf{X}^n = \mathbf{x}^n | \mathbf{Y}^n = \mathbf{y}^n]$$

$$= \arg \max_{\mathbf{x}^n \in \mathcal{X}^{n+1}} \Big[\pi_0(x_0)Q(x_0, y_0)P(x_0, x_1)Q(y_1, x_1) \cdots P(x_{n-1}, x_n)Q(y_n, x_n)\Big]$$

Now, since logarithms are monotonic, we can equivalently minimize the negative log of the above expression. This may seem mysterious at first, but actually has a very nice interpretation.

$$\arg \max_{\mathbf{x}^n \in \mathcal{X}^{n+1}} \mathbf{Pr}[\mathbf{X}^n = \mathbf{x}^n | \mathbf{Y}^n = \mathbf{y}^n]$$

$$= \arg \max_{\mathbf{x}^n \in \mathcal{X}^{n+1}} \Big[\underbrace{\log(\pi_0(x_0)Q(x_0, y_0))}_{-d_0(x_0)} + \sum_{m=1}^{n} \underbrace{\log(P(x_{m-1}, x_m)Q(x_m, y_m))}_{-d_m(x_{m-1}, x_m)}\Big]$$

$$= \arg \min_{\mathbf{x}^n \in \mathcal{X}^{n+1}} \Big[d_0(x_0) + \sum_{m=1}^{n} d_m(x_{m-1}, x_m)\Big]$$

Above we have define these terms as $-d_i$ since they are logs of probabilities, so in this way the d_i will be positive. This in turn lends itself very nicely to the interpretation of the above expression as a graph, known in this setting as a **trellis diagram**.

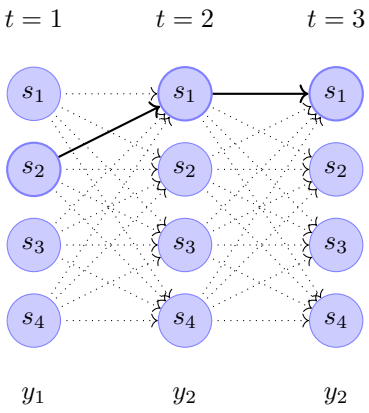


Figure 2: Trellis of the observation sequence y_1,y_2,y_2 for the above HMM. The thick arrows indicate the most probable transitions. As an example, the transition between state s_1 at time t=2 and state s_4 at time t=3 has probability $\alpha_2(1)a_{14}b_4(y_2)$, where $\alpha_t(i)$ is the probability to be in state s_i at time t.

28 Extra Content: Expectation Maximization