EECS 126 Notes

Tyler Zhu

September 23, 2020

"A good stock of examples, as large as possible, is indispensable for a thorough understanding of any concept, and when I want to learn something new, I make it my first job to build one."

– Paul Halmos.

These are course notes for the Spring 2019 rendition of EECS 126, Probability Theory and Random Processes, taught by Professor Kannan Ramchandran. I began these notes five lectures in, but the missing material is mostly CS 70 level probability. Special thanks to Evan Chen for parts of his .sty file, and Zhang Qiaowei for pointing out many typos.

Contents

The missing days:

1 Tuesday, January 22nd

2 Thursday, January 24th

3 Tuesday, January 29th

4 Thursday, January 31st

5 Tuesday, February 5th

6 Thursday, February 7th

6.1 Announcements

Couple of announcements.

- HW 3 is due next Wednesday.
- Lab 2 is due on Friday.
- Self grades for both HW and Lab due on Monday.
- Readings are B&T, Ch. 3, 4.1-3 and 4.6

6.2 Mins and Maxes of Exponentials

We saw already that the exponential random variable has pdf $f_T(t) = \lambda e^{-\lambda t}$, and has $\mathbb{E}[T] = 1/\lambda$ and $var(T) = 1/\lambda^2$.

We also saw how $\min(T_1, T_2, \ldots, T_n) \sim \text{Expo}(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$ for independent exponential variables $T_1, T_2, \ldots, T_n$. The key idea used was the memoryless property.

One example is if we have a hundred lightbulbs that burn out in time proportional to Expo(1). Then what is the expected time it would take for the first lightbulb to burn out? Of course, this would be 1/100 since the min is modeled by Expo(100).

But what about the maximum? If we have n i.i.d. Expo(1) random variables T_i, what is $\mathbb{E}[\max(T_1, \ldots, T_n)]$? Continuing our previous analogy, we can think of this as asking what the expected time for the last lightbulb to burn out is.

There's a way to leverage what we've already calculated to do this calculation however! Since the exponential r.v. is memoryless, once the first bulb burns out, we have the same situation again, but with $n - 1$ bulbs instead. Hence,

$$\mathbb{E}[\text{time for } n \text{ bulbs to b.o.}] = \mathbb{E}[\text{1st bulb to b.o.}] + \mathbb{E}[n - 1 \text{ remaining bulbs to b.o.}]$$

If we let S_n be the r.v. counting the time it takes for n bulbs to burn out (and using the fact that $\mathbb{E}[S_1] = 1$), we can rewrite this as

$$\begin{aligned}
\mathbb{E}[S_n] &= \frac{1}{n} + \mathbb{E}[S_{n-1}] \\
&= \frac{1}{n} + \frac{1}{n-1} + \mathbb{E}[S_{n-1}] \\
&= \sum_{k=1}^{n} \frac{1}{k} = H_n \approx \ln n + \gamma
\end{aligned}$$

where γ is the Euler-Mascheroni constant.

Another quick remark: Geometric variables are just discretizations of exponential random variables. (See book).

Here's a teaser to cap off this section. Suppose you are in line at a post office, and ahead of you two people are waiting to be served with probability Expo(1) each. Once one of the two are served, you take their place waiting to be served. What is the probability you will be served before the other person? Answer is 1/2 thanks to the memoryless property.

6.3 Standard Normal Distributions

Definition 1. The PDF of the *standard normal* distribution $\mathcal{N}(0,1)$, i.e. with mean 0 and variance 1, is

$$f_X(x) \propto e^{-x^2/2} \quad \text{for } -\infty < x < \infty.$$

This is a probability distribution for an appropriate choice of c for which $c\int_{-\infty}^{\infty} e^{-x^2/2}dx = 1$.

You can probably convince yourself that this integral converges, but it's another question to figure out exactly *what* it converges to!

To do this, we'll introduce a variable α (effectively computing a more *general* integral) and consider the integral

$$I = \int_{-\infty}^{\infty} e^{-\alpha x^2/2}dx.$$

Then the trick is to consider I^2, since we get

$$I^2 = \int_{\mathbb{R}} e^{-\alpha x^2/2}dx \int_{\mathbb{R}} e^{-\alpha y^2/2}dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\alpha(x^2+y^2)/2}dxdy.$$

This should remind you of polar coordinates, where $x^2 + y^2 = r^2$ represents the radius. Hence, we will change our variables to achieve that, using

$$\begin{aligned} x^2 + y^2 &= r^2 \\ dx\,dy &= r\,dr\,d\theta \end{aligned}$$

(you can get these by computing the Jacobian (ew)). Substituting gives

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty} e^{-\alpha r^2/2} r dr d\theta = 2\pi \int_0^{\infty} re^{-\alpha r^2}dr \\ &= (2\pi)\left(-\frac{1}{\alpha}e^{-\alpha r^2/2}\right)\Big|_0^{\infty} = \frac{2\pi}{\alpha} \end{aligned}$$

Finally, to get our desired integral, we set $\alpha = 1$ so that $I^2 = 2\pi$ and hence, $I = \sqrt{2\pi}$. This means our PDF is actually

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

Let's verify that the mean and variance of this PDF are indeed 0 and 1. We know by symmetry about 0 that $\mathbb{E}[X] = 0$. For the variance, our calculations are simplified because of this, so

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-\alpha x^2/2}dx.$$

Now we'll employ differentiation under the integral, which states that if we have some function $f(x,\alpha)$, then

$$\int_a^b \frac{\partial}{\partial\alpha} f(x,\alpha)dx = \frac{d}{d\alpha}\int_a^b f(x,\alpha)dx.$$

We have an integral of the form on the left side, and our goal will be to get to the right side. This means we solve for f to get

$$\frac{\partial}{\partial \alpha} f(x, \alpha) = x^2 e^{-\alpha x^2/2} \implies f(x, \alpha) = -2e^{-\alpha x^2/2}$$

and plugging this in gives

$$\text{Var}(X) = \frac{1}{\sqrt{2\pi}} \frac{d}{d\alpha} \int_{\infty}^{\infty} -2e^{-\alpha x^2/2} dx.$$

But we already know this integral! It's the one we just calculated, I, which we know is $\sqrt{\frac{2\pi}{\alpha}}$. So we can just substitute this back in to get

$$\text{Var}(X) = -2 \frac{d}{d\alpha} \frac{1}{\sqrt{\alpha}} = \alpha^{-3/2}.$$

Our goal is when $\alpha = 1$, which gives us $\text{Var}(X) = 1$ as desired.

6.4 Applications of the Standard Normal

Sometimes we're interested in integrating the PDF only over certain intervals; this is the CDF, which is defined as

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Usually only $\Phi(y)$ for $y \geq 0$ are recorded due to the symmetry of the PDF.

We can also extend the standard normal distribution to distributions with arbitrary mean and variance. Let $Y = \mathcal{N}(0, 1)$. We write $X \sim \mathcal{N}(\mu, \sigma^2)$ for a normal distribution with mean $\mathbb{E}[X] = \mu$ and variance $\text{Var}(X) = \sigma$. Note that $X = \sigma Y + \mu$. The PDF of X is

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

There's some examples on calculating values from normal distributions, but honestly an AP Statistics book is probably a better source for this than anything I can write down.

6.5 Derived Distributions (Transformations of RVs)

Now let's look at deriving distributions from other distributions.

Theorem 2. Let Y and X be random variables, such that $Y = aX + b$, i.e. Y is linear in terms of X. Then if we know $f_X(x)$, we can derive $f_Y(y)$ as

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

for $a \neq 0$.

Here's an application.

Example 6.1. Suppose we have $X = \sigma Y + \mu$, where $Y \sim \mathcal{N}(0,1)$. We know that

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2},$$

and so

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

which matches with what we knew already.

Ramchandran's advice is to not memorize these formulas for derived distributions, but instead using the following general rule for derived distributions.

Theorem 3 (Finding Derived Distributions)**.** Suppose we have $Y = g(X)$. Then to find the density of Y:

1. Calculate $F_Y(y) = \int_{\{x|g(x)\leq y\}} f_X(x)dx$.
2. Then $f_Y(y) = \dfrac{dF_Y(y)}{dy}$.

Example 6.2. Let $Y = X^2$. We proceed in the above steps.

1. For $y \geq 0$,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(x^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq x \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

2. Differentiating gives

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) - \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}).$$

6.6 Multiple Random Variables

Literally page 13 of the book.

6.7 Bayes' Rule for Continuous Random Variables

We end with Bayes' Rule applied in the continuous case, which is pretty much exactly what you'd expect.

Theorem 4 (Bayes' Rule)**.** Suppose I have continuous density functions $f_X(x)$ and $f_Y(y)$ for random variables X, Y. Then,

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

Example 6.3. A lightbulb has an exponential lifetime $Y \sim \text{Expo}(\lambda)$, but λ itself is a random variable $\lambda \in U[1, 3/2]$. We test a lightbulb and record its lifetime. What can we say about λ?

Solution. Let Λ be the distribution of λ, i.e. $U[1, 3/2]$. Suppose we observe that the lightbulb has lifetime y. Then by Bayes',

$$f_{\Lambda|Y}(\lambda|y) = \frac{f_\Lambda(\lambda) f_{Y|\Lambda}(y|la)}{\int_1^{3/2} f_\Lambda(t) f_{Y|\Lambda}(y|t) dt} = \frac{2\lambda e^{-\lambda y}}{\int_1^{3/2} 2te^{-\lambda t} dt}.$$

□

7 Tuesday, February 12th

7.1 Announcements and Agenda

Announcements:

1. HW 3 due tomorrow.
2. Lab 3 due on Friday.
3. Midterm 1 next Tuesday.
4. Reading B&T 4.3-4.6, 5.1, W 13.7.

Agenda will be order statistics, then convolution, and finally transforms: MGFs.

7.2 Order Statistics

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. RVs with common density $f_X(x)$ and CDF $F_X(x)$. Let $X^{(k)}$ be defined as the kth smallest of $X_1, X_2, \ldots, X_n$; $X^{(1)}$ is the min while $X^{(n)}$ is the max. Order statistics comes from the fact that we're concerned with the order of these RVs.

Question. What is the pdf of $X^{(k)}$, equivalently $f_{X^{(k)}}(x)$?

Solution. By definition,

$$\mathbb{P}(X^{(k)} \in (x, x+dx)) \approx f_{X^{(k)}}(x)dx.$$

In order for the kth smallest point to lie between x and $x+dx$, we need three things to happen:

1. $k-1$ points must lie in the interval $(-\infty, x)$
2. One point must lie in $(x, x+dx)$
3. $n-k$ values to lie in the interval $(x+dx, \infty)$.

Now it's just counting. We have n choices for our one point, and $\binom{n-1}{k-1}$ choices to distribute the rest, so there are $n\binom{n-1}{k-1}$ ways to distribute our points. Combining these with the proper probabilities gives

$$f_{X^{(k)}}(x)dx = n\binom{n-1}{k-1} f_X(x)[F_X(x)]^{k-1}[1-F_X(x)]^{n-k}dx$$

so cancelling dx's gives us our desired density. [1] □

Order doesn't matter here by symmetry; for every ordering of the $k-1$ points we have the same number of orderings of the other $n-k+1$ points, etc.

For example, suppose we have a uniformly drawn RV $X \sim U[0,1]$, where $f_X(x) = 1$ and $F_X(x) = x$ for $0 \le x \le 1$. Then the kth order statistic for X is

$$f_X^{(k)}(x) = n\binom{n-1}{k-1} x^{k-1}(1-x)^{n-k}.$$

Now we can do statistics on the kth order statistic, which is kinda cool.

Question. What is the probability that the 9th smallest out of 10 drawings from $X \sim U[0,1]$ is greater than 0.8?

Solution. You kinda just do it. Answer is

$$f_{X^{(k)}}(x) = \frac{10!}{8!1!}x^8(1-x) = 10x^8 - 90x^9 \implies \mathbb{P}(X^{(9)} > 0.8) = \int_{0.8}^{1}(90x^8 - 90x^9)dx.$$

□

[1] It really should be $1 - F_X(x+dx)$, but in the limit it doesn't matter.

7.3 Convolutions

Suppose we have RV $Z = X + Y$ for independent CRVs X, Y, and that we are given $f_X(x)$ and $f_Y(y)$. We want to find $f_Z(z)$.

We can begin by looking at conditional CDFs, so

$$\begin{aligned} F_{Z|X}(z|x) &= \mathbb{P}(Z \leq z | X = x) \\ &= \mathbb{P}(Y \leq z - x | X = x) \\ &= \mathbb{P}(Y \leq z - x) \\ &= F_Y(z - x) \end{aligned}$$

since X and Y are independent. Now we can differentiate both sides of this equation with respect to z to get

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

We're in the home stretch now. All that's left is to get rid of the dependence of Z on Y, so we marginalize it out by integrating to get

$$f_Z(z) = \int_X f_{Z|X}(z|x) f_X(x) dx = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx = (f_X \star f_Y)(z)$$

which is just a convolution!

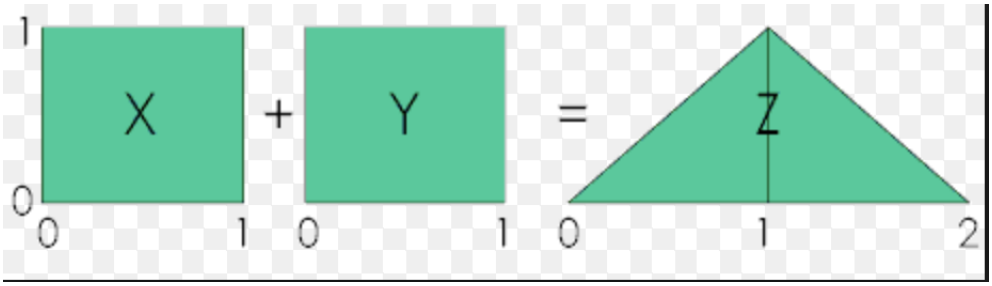

Figure 1: The convolution of two uniform distributions $U[0, 1]$.

This is actually why the distribution of dice rolls has a peak at 7 and decreases in either direction, because it is the convolution of two uniform distributions (namely $U[1, 6]$).

In the discrete case, this is just

$$\mathbb{P}(Z = k) = \sum_m \mathbb{P}(X = m)\mathbb{P}(Y = k - m).$$

7.4 Moment Generating Functions (Transforms)

Recall from calculus that the exponential function has the series expansion

$$e^{sX} = 1 + sX + \frac{s^2}{2!}X^2 + \frac{s^3}{3!}X^3 + \dots.$$

We can let X be a RV. Then we can take expectation of both sides and apply linearity of expectation to get

$$M_X(s) = \mathbb{E}[e^{sX}] = 1 + s\mathbb{E}[X] + \frac{s^2}{2!}\mathbb{E}[X^2] + \frac{s^3}{3!}\mathbb{E}[X^3] + \dots.$$

We call $M_X(s)$ the **moment generating function**, or **transform**, of X, for the following reason. All of the moments, i.e. RVs of the form X^k, are on the right hand side, and we can sift

out whichever moment we need with a cute trick. If we want $\mathbb{E}[X]$, then we can differentiate both sides with respect to s and set $s = 0$, killing all the terms but $\mathbb{E}[X]$. In symbols,

$$\frac{d}{ds}[M_X(s)] = \mathbb{E}[X] + s\mathbb{E}[X^2] + \frac{s^2}{2!}\mathbb{E}[X^3] + \dots$$

If we wanted $\mathbb{E}[X^2]$, we can just take another derivative to get

$$\frac{d^2}{ds^2}[M_X(s)] = \mathbb{E}[X^2] + s\mathbb{E}[X^3] + \dots$$

and set $s = 0$. In general, if we take n derivatives, we find

$$\frac{d^n}{ds^n}[M_X(s)] = \mathbb{E}[X^n] + s\mathbb{E}[X^{n+1}] + \dots$$

from which setting s to 0 gives us the nth moment.

What are the advantages of MGFs?

1. Much easier to find the *moments* of X.
2. Much easier to *multiply* than *convolve*
3. Great analytical tool for proving things (CLT).

Here's some properties.

Theorem 5. The moment generating function $M_X(s)$ of a RV X satisfies the following properties.

(1) $M_X(0) = 1$.

(2) For $Y = aX + b$, $M_Y(s) = e^{sb}M_X(as)$.

Proof. Part (a) is obvious. (Hint: use the very deep fact that $1 + 0 = 1$.)

For part (b), just do the math. You get

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

□

Let's get our hands dirty and find the MGFs of some common distributions.

Example 7.1 (Exponential MGF). Suppose we have a RV $X \sim \text{Expo}(\lambda)$ which has pdf $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Then

$$\mathbb{E}[e^{sX}] = \int_0^\infty e^{sx} f_X(x)dx = \lambda \int_0^\infty e^{-\lambda x}e^{sx}dx = \lambda \frac{e^{-(\lambda - s)x}}{-(\lambda - s)}\Big|_0^\infty = \frac{\lambda}{\lambda - s}$$

where $s < \lambda$ must hold in order for the integral to converge. Using this we can get $\mathbb{E}[X] = M_X'(0) = \frac{\lambda}{(\lambda - s)^2}\Big|_{s=0} = \frac{1}{\lambda}$, and $\mathbb{E}[X^2] = M_X''(0) = \frac{2}{\lambda}$.

Example 7.2 (Poisson MGF). Now lets do the same for the Poisson distribution. Let $X \sim \text{Poisson}(\lambda)$, so that $\mathbb{P}(X = k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$ for $k \geq 0$. Then

$$M_X(s) = \mathbb{E}[e^{sX}] = \sum_{k=0}^{\infty} e^{sk}\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^s)^k}{k!} = e^{-\lambda+\lambda e^s}.$$

From this we can calculate $M_X'(0) = \lambda$ and $M_X''(0) = \lambda^2 + \lambda$.

Example 7.3 (Normal MGF). Finally lets try the same for the normal distribution. Let $X \sim \mathcal{N}(0,1)$ so that $f_X(x) = \dfrac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Then

$$M_X(s) = \mathbb{E}[e^{sX}] = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}e^{sx}dx.$$

Now we're going to do a little bit of magic. The inside of this integral is $\exp(-x^2/2 + sx)$ which is a quadratic in x, so we're going to *complete the square*. Hence, we get

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{x^2-2xs+s^2}{2}\right)e^{s^2/2}dx = e^{s^2/2}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-(x-s)^2/2}dx\right].$$

But the bracketed term is precisely our pdf integrated over its domain, which we already know to be 1. Hence, $M_X(s) = e^{s^2/2}$. This is super important, so know it.

If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $M_Y(s) = e^{s\mu}M_X(\sigma s)$, so $M_Y(s) = e^{s\mu}e^{\sigma^2 s^2/2}$ is the MGF for general normal distributions.

It is left as an exercise to verifty that $M_X'(0) = 0$ and $M_X''(0) = 1$.

7.5 Inversions of transforms

Turns out $M_X(s)$ contains all of the info in $f_X(x)$ (under the mild condition that the moments are finite). This is known as the bilateral Laplace transform of $f_X(x)$. We can do inversions of these transforms using "pattern matching," which is really just educating guessing lol.

Example 7.4. Suppose we have an MGF of $M_X(s) = \dfrac{1}{2}e^{-3s} + \dfrac{1}{4}e^{200s} + \dfrac{1}{4}e^{s}$. Recall that in the discrete case, transforms are a sum of terms of the form e^{sx}, so by comparing with the general formula

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

we can recover our pdf as

$$P(X = k) = \begin{cases} 1/2 & \text{when } k = -3 \\ 1/4 & \text{when } k = 200 \\ 1/4 & \text{when } k = 1 \end{cases}$$

Here's the capstone on why we care so much about MGFs: we don't have to work with convolutions if we use them! Suppose we have $Z = X + Y$, where X, Y are independent RVs.

Then

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

avoiding any use of convolutions whatsoever! In summary, the MGF of the sum of two RVs is the product of their MGFs.

A quick application before we close out the day. Suppose $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \mu_Y^2)$. Then

$$M_Z(s) = M_X(s)M_Y(s) = \exp\left(\left(\frac{\sigma_X^2 + \sigma_Y^2}{2}\right)s^2 + (\mu_X + \mu_Y)s\right) = MGF(\mathcal{N}(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)).$$

8 Thursday, February 14th

Happy Valentine's day to all the non-CS majors out there. As for the rest of you...

8.1 Moment Generating Functions (Review)

The moment generating function of a random variable X is $M_X(s) = \mathbb{E}[e^{sx}]$, named as such because we can recover all of the moments $\mathbb{E}[X^n]$ by taking derivatives. A quick result of this is that

$$\mathbb{E}\left[\exp\left(s\sum_{i=1}^{n} X_i\right)\right] = \prod_{i=1}^{n} \mathbb{E}[e^{sX_i}].$$

8.2 Law of Total Variance

Theorem 6 (Total Variance)**.** For random variables X and Y,

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

Proof. We expand intelligently;

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X|Y] + \mathbb{E}[X|Y] - \mathbb{E}[X])^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X|Y])^2] + 2\mathbb{E}[(X - \mathbb{E}[X|Y])(\mathbb{E}[X|Y] - \mathbb{E}[X])] + \mathbb{E}[(\mathbb{E}[X|Y] - \mathbb{E}[X])^2]\end{aligned}$$

where the last step was performed by grouping the first two terms and the last two terms and expanding. "It is left as an exercise to show the middle term goes to 0," so we will focus on what we get from the remaining two terms.

We will make use of the following fact.

Theorem 7 (Iterated Expectation)**.** For random variables X and Y,

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

You can think of this as conditional expectation; we condition on Y and find the expectation on the inside, and then "sum out" over all values of Y by taking an expectation on the outside. Hence, we can rewrite

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X|Y])^2] &= \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y]] = \mathrm{Var}(\mathbb{E}[X|Y]) \\ \mathbb{E}[(\mathbb{E}[X|Y] - \mathbb{E}[X])^2] &= \mathbb{E}[(\mathbb{E}[X|Y] - \mathbb{E}[\mathbb{E}[X|Y]])^2] = \mathbb{E}[\mathrm{Var}(X|Y)]\end{aligned}$$

from which our result follows □

Here's an application as a quick reality check. [2]

[2] Thanks to Efe for the Piazza explanation.

Example 8.1. A chocolate store receives $B \sim \text{Bin}(n, p)$ types of chocolate. When you go to the store, for each type of chocolate in the store, you toss an independent coin which has a probability q of success. In summary, the amount of chocolate you buy is $C = \sum_{i=1}^{B} \mathbb{1}_i$ where $\mathbb{1}_i$ is a Bernoulli RV with probability q.

We can compute the variance of C using the law of total variance, which tells us that

$$\text{Var}(C) = \text{Var}(\mathbb{E}[C|B]) + \mathbb{E}[\text{Var}(C|B)].$$

First, notice that $C|B$ is a RV according to $\text{Bin}(B, q)$; there are B chocolates, and we have a probability q of buying each one. Then $\mathbb{E}[C|B]$ is Bq by linearity of expectation, so

$$\text{Var}(Bq) = q^2\text{Var}(B) = q^2np(1-p).$$

Also, $\text{Var}(C|B)$ is $Bq(1-q)$, so

$$\mathbb{E}[Bq(1-q)] = q(1-q)\mathbb{E}[B] = npq(1-q).$$

Putting it all together gives

$$\text{Var}(C) = npq^2(1-p) + npq(1-q) = npq(q - pq + 1 - q) = npq(1-pq),$$

which is precisely the variance of $\text{Bin}(B, pq)$! This shouldn't be surprising, since we expected that the total number of things compounds in this way.

8.3 Tail Bounds: Markov and Chebyshev

This consists of a lot of CS 70 material (Markov's, Chebyshev's), but also some new bounds (namely the Chernoff bound).

Theorem 8 (Markov's Inequality)**.** For a nonnegative random variable X,

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

for all $t > 0$.

This is important because tail bounds help us bound rare things that are away from the expectation. Think of a statistician who wants to bound the probability of errors in his/her data

Proof. We condition on values of X, so

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X|X \leq t]\mathbb{P}(X \leq t) + \mathbb{E}[X|X \geq t]\mathbb{P}(X \geq t) \\ &\geq 0 + t\mathbb{P}(X \geq t)\end{aligned}$$

where we can lower bound by 0 since X is nonnegative and we can lowerbound $\mathbb{E}[X|X \geq t]$ by t since we're conditioning on $X \geq t$. □

The proof in our book uses coupling, which is creating a new random variable which is 0 when $X \leq t$ and exactly t when $X \geq t$, which leads to a similar proof as shown here.

Let's try applying Markov's inequality to an exponential distribution $X \sim \text{Expo}(\lambda)$. We have $\mathbb{P}(X \geq t) = e^{-\lambda t}$ by the CDF, while Markov's inequality gives a bound of $\frac{1}{\lambda t}$, which is much looser. This might lead one to think that Markov's inequality is weak, but if all you know

about a distribution is its mean, Markov's inequality is actually *tight*! For a fixed t, there are distributions for which equality holds.

Of course, if you know more information you can obtain a better bound. Namely,

Theorem 9 (Chebyshev)**.** For a random variable X,

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

for $t > 0$.

Proof. Using the very deep fact that $|a| > b \implies a^2 > b^2$, we find that

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

where we can use Markov's inequality since $(X - \mathbb{E}[X])^2$ is nonnegative. □

Example 8.2. Let $X \sim \text{Bin}(n, p)$. By Markov,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|]}{t} \leq \frac{\mathbb{E}[|X| + |\mathbb{E}[X]|]}{t} = \frac{2\mathbb{E}[X]}{t} = \frac{2np}{t}$$

where we used the triangle inequality to deduce that $|X - \mathbb{E}[X]| \leq |X| + \mathbb{E}[X]$ (triangle inequality is $|a + b| < |a| + |b|$), and $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$ since expectation of a constant is a constant.

If we use Chebyshev, we get a bound of

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\text{Var}(x)}{t^2} = \frac{np(1-p)}{t^2}$$

which is a factor of t tighter than Markov.

8.4 Chernoff Bounds

Before we go into Chernoff bounds, let's try to motivate them and see why they are interesting. Let's try bounding the probability $\mathbb{P}(X \geq t)$ by the moments. By Markov, we have

$$\mathbb{P}(f(X) \geq f(t)) \leq \frac{\mathbb{E}[f(X)]}{f(t)}$$

for $f(X) \geq 0$ and for all $f(t) > 0$. Hence, we can see that

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[f(X)]}{f(t)}$$

provided that $\mathbb{P}(X \geq t) \leq \mathbb{P}(f(X) \geq f(t))$; this can be achieved if f is monotonically increasing, but this is merely sufficient, not necessary.

Using this derived bound, we can bound probabilities by moments of our random variable by taking $f(t) = t^n$ to get

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X^n]}{t^n}$$

for $X \geq 0$, which is pretty good. The Chernoff bound is similar, offering us a bound in terms of the moments.

Theorem 10 (Chernoff Bound)**.** For a random variable X,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}}$$

for all t and $\lambda > 0$.

Proof. This should be second nature at this point. . Since $\lambda > 0$,

$$\mathbb{P}(X \geq t) = \mathbb{P}(\lambda X \geq \lambda t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}}.$$

□

Important question of the day: why do we even need a λ at all?

Answer: Like the French revolution, it gives you freedom! (over how sharp of a bound you get). One thinks of it as a knob that you can adjust to give you a better (or worse) bound. Note that higher λ isn't always better.

Another note is that Chernoff bounds aren't always the best; if you have a Chernoff bound with a fixed λ, I can always come up with a better moment bound.

Example 8.3. Let's apply the Chernoff bound to normal random variables. We find that

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[e^{-\lambda X}]}{e^{\lambda t}} = \frac{e^{-\lambda^2\sigma^2/2}}{e^{\lambda t}}$$

for $X \sim \mathcal{N}(0, \sigma^2)$. We want to pick the λ that minimizes this, so we'll take logs and then differentiatie, because

$$\arg\min_{\lambda} \frac{e^{-\lambda^2\sigma^2/2}}{e^{\lambda t}} = \arg\min_{\lambda} \frac{\lambda^2\sigma^2}{2} - \lambda t.$$

Differentiating this gives a solution $\lambda = t/\sigma^2$, which gives a bound of $\mathbb{P}(X \geq t) \leq e^{-t^2/2\sigma^2}$, which is literally as tight as possible since it's our CDF. Magic.

9 Thursday, February 21st

9.1 Announcements

HW 5 due next Wednesday. Reading is B&T 5.2-5.6, Walrand 2.1-2.3. Apparently Chapter 2 is quite difficult to read, so proceed with caution.

9.2 Weak and Strong Law of Large Numbers (Modes of Convergence)

The idea of the weak law of large numbers is to look at the behavior of say coin flips in the long run. There's two questions we can start off our discussion with: how many heads (mean) will we get, and how variable are our results (variance)?

We can describe it formally as such. We perform an experiment n times independently and note

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (X_i's \text{ i.i.d. })$$

where X_i has mean μ and variance σ^2. Then

$$\mathbb{E}[M_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathbb{E}[X_1]\cdot n = \mathbb{E}[X_1] = \mu,$$

and

$$\text{Var}(M_n) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{\sigma^2}{n}$$

since the X_i are i.i.d. If we take $n \to \infty$ to look at the long term behavior, $\mathbb{E}[M_n] = \mu$ and $\text{Var}(M_n) = 0$.

This is cool and all, but wouldn't it be dope if we also knew the *rate* at which the variance decreased to 0? We can do this by using Chebyshev (Theorem 9) to do a tail bound:

$$\mathbb{P}(|M_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2},$$

and just like that we've derived the weak law of large numbers.

Theorem 11 (Weak LLN)**.** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. RVs with mean μ. Then for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$.

What does the Weak LLN really mean? In one way, it means that $\lim_{n\to\infty}\mathbb{P}(|M_n - \mu| \geq \epsilon) = 0$. Now recall delta-epsilon limits from calculus: for any $\epsilon, \delta > 0$, there exists $n(\epsilon, \delta)$ (meaning n is a function of ϵ, δ), large enough such that

$$\mathbb{P}(|M_n - \mu| \geq \epsilon) \leq \delta$$

for $n > n(\epsilon, \delta)$. We can think of these variables as representing the following:

$$\begin{aligned} \epsilon &: \text{ “accuracy level” or error} \\ \delta &: \text{ confidence level} \\ n(\epsilon, \delta) &: \text{ threshold for a given accuracy/confidence} \end{aligned}$$

Motivated by our above findings, we make an important definition:

Definition 12. We say a sequence of random variables $(M_n)_{n=1}^{\infty} = M_1, M_2, \dots$ **converges in probability** if for any $\epsilon > 0$, $\mathbb{P}(|M_n - \mu| \geq \epsilon) \to 0$ as $n \to \infty$, and denote it as $M_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu$.

This gives us a notion of convergence for random variables/distributions, much like convergence of functions from Calculus.[3]

Example 9.1. Suppose $X_1, X_2, \dots, X_n \sim U[-1, 1]$ are i.i.d., and $Y_n = \frac{X_n}{n}$. Then to find the density, we first find that

$$Y_n \leq y \implies X_n \leq ny,$$

so

$$F_{Y_n}(y) = F_X(ny) \implies f_{Y_n}(y) = nf_X(ny).$$

If we plot $f_{Y_n}(y)$, it would be a rectangle with endpoints at $y = -1/n$ and $y = 1/n$ with height $n/2$ since $f_{Y_n}(y) = n/2$ for all y in its domain. Then $\mathbb{P}(|Y_n| \geq \epsilon) = 0$ if $n > \frac{1}{\epsilon}$, which is what it means to converge in probability.

Example 9.2. Let $Y_n = \min(X_1, X_2, \dots, X_n)$ where X_i's are i.i.d. in $U[0, 1]$. Then

$$\mathbb{P}(|Y_n - 0| \geq \epsilon) = \mathbb{P}(X_1 \geq \epsilon, X_2 \geq \epsilon, \dots, X_n > \epsilon) = (1 - \epsilon)^n \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

In other words, the probability that the minimum is greater than any ϵ you pick in the long run is 0, which makes sense.

Example 9.3. Suppose time is in discrete units[a] $(1, 2, \dots)$ and $Y_n = 1$ if there is an arrival at time n, $Y_n = 0$ otherwise. Define $I_k = \{2^k, 2^k + 1, \dots, 2^{k+1} - 1\}$, so that every next interval is twice as large as the next one. Suppose there is exactly 1 arrival in each interval (equally likely). Then

$$\begin{aligned} \mathbb{P}(Y_1 = 1) &= 1 \\ \mathbb{P}(Y_2 = 1) = \mathbb{P}(Y_3 = 1) &= \frac{1}{2} \\ &\dots \\ \mathbb{P}(Y_n = 1) &= \frac{1}{2^k} \text{ if } n \in I_k. \end{aligned}$$

So $\lim_{n\to\infty} \mathbb{P}(Y_n = 1) = \lim_{k\to\infty} \frac{1}{2^k} = 0$, meaning that $\mathbb{P}(Y_n = 1)$ converges to 0 in probability.

This should be confusing however! Given any finite n, there are certain to be an infinite number of arrivals after n. Hence, we know for a *fact* that $\mathbb{P}(Y_n = 1)$ will be nonzero infinitely often.

This example demonstrates the weakness of the weak LLN, and tells us that perhaps there are stronger notions of convergence than just convergence in probability.

[a] "Bold move." - Phil

Before I state Strong LLN, I will first state what this stronger notion of convergence is.

[3] Remember $\delta - \epsilon$ limits? yea those disgusting things.

Definition 13. Let $(M_n)_{n\geq 1}$ be a sequence of random variables. Then we say that M_n converges **almost surely** to μ if

$$\mathbb{P}\left(\lim_{n\to\infty} M_n = \mu\right) = 1$$

and denote it by $M_n \xrightarrow{\text{a.s.}} \mu$.

Theorem 14 (Strong LLN)**.** Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. RV's with mean μ. Then,

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

as $n \to \infty$ with probability 1. In other words, if we let $M_n = \frac{1}{n}\sum X_i$, then $M_n \xrightarrow{\text{a.s.}} \mu$.

Proof. Walrand Chapter 2, but only for the brave. □

Let's illustrate the difference between the Weak and Strong LLN with an example of rolling 6-sided die. The Strong LLN states that *every* realization converges to μ. So if we were to draw a plot of all different realizations, the Strong LLN states that all of them tend towards the line $X = \mu$.

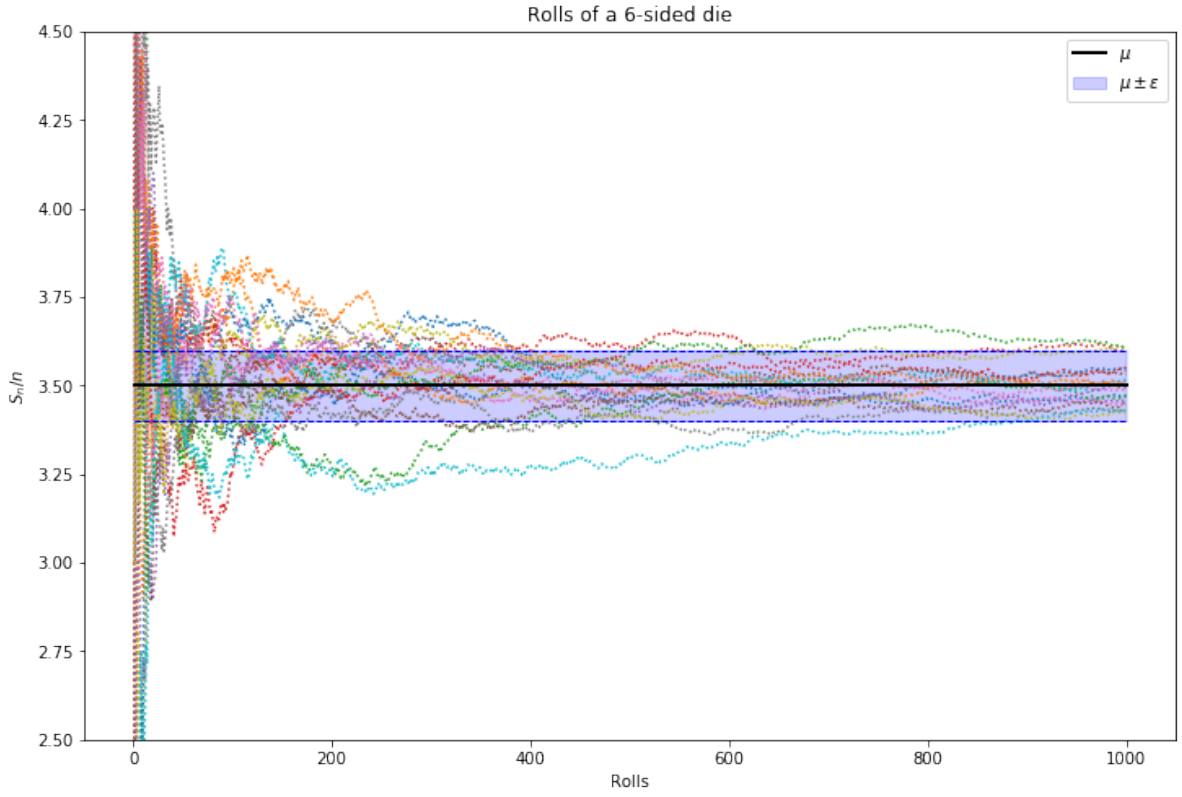

Figure 2: The weak LLN bounds the long-run values with the blue bounding box $X \in [\mu - \epsilon, \mu + \epsilon]$, while strong LLN asserts all values become the black line, $X = \mu$.

On the other hand, the Weak LLN only places a bounding box of width 2ϵ around our mean, and says that the probability of any realizations outside this box is 0 in the long run. It doesn't say anything about occurences within the box, which is why our wacky example from above technically converges in probability. Hence, all we are guaranteed is that the *fraction* of realizations outside $\mu \pm \epsilon$ for all $\epsilon, \delta > 0$ converges to 0.

9.3 The Central Limit Theorem

Question. What happens to $S_n = \sum_{i=1}^n X_i = X_1 + X_2 + \cdots + X_n$ as $n \to \infty$?

Answer. You always end up with a normal distribution. Try it out for $X_i \sim U[0,1]$ or $X_i \sim \text{Exp}(1)$ if you don't believe me. □

If we look at the variance and mean of $S_n \to \infty$ as $n \to \infty$, then we find that $\mathbb{E}[S_n] = n\mu$ and $\text{Var}(S_n) = n\sigma^2$. Hence we should normalize by defining

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

You can check that this now has 0 mean and variance 1. As before, we might expect that this new random variable is normal if the X_i are uniform or exponential, so that Z_n is just a standard normal distribution. Amazingly, this holds in general, which is what the CLT states.

Theorem 15 (Central Limit Theorem)**.** Suppose $S_n = \sum_{i=1}^n X_i$ where the X_i are i.i.d. RVs with mean μ and variance σ^2, and define Z_n as above. Then, (amazingly),

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le x) = \Phi(x)$$

where $\Phi(x)$ is the c.d.f of the standard normal distribution $\mathcal{N}(0,1)$.

For large enough n, $Z_n \sim \mathcal{N}(0,1)$ in distribution, i.e.

$$S_n \xrightarrow[n\to\infty]{d} \mathcal{N}(n\mu, n\sigma^2)$$
$$Z_n \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

where $\xrightarrow[n\to\infty]{d}$ is convergence in distribution, which we will not go into detail about. There are two implications.

1. The distribution of S_n and Z_n "wipe out" all the information except for μ and σ^2.
2. If there are a large number of small independent factors, the *aggregate* of these factors will be normally distributed, which is just noise.

We end the day by outlining the proof of CLT.

Proof of CLT. If $Y \sim \mathcal{N}(0,1)$, $M_Y(s) = \mathbb{E}[e^{sY}] = e^{s^2/2}$. Then suppose $X_1, X_2, \ldots, X_n$ are i.i.d. with mean 0 and variance 1 (WLOG). Let $M_X(s)$ be the MGF of each X_i, and let

$$Z = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \implies \mathbb{E}[Z] = 0, \text{Var}(Z) = 1.$$

Then

$$M_Z(s) = \mathbb{E}[e^{sZ}] = \mathbb{E}[\exp(\frac{s}{\sqrt{n}}(X_1 + X_2 + \cdots + X_n))].$$

We can finish the proof then by decomposing $M_Z(s)$ and using Taylor expansion. The rest of the proof will be discussed next week. □

10 Tuesday, February 26th

HW 5 is due tomorrow, and the readings are Walrand Ch 1, 2.4, 13.3, B&T 6.1-6.4.

10.1 Wrapping up CLT

From last lecture, suppose we have $S_n = \sum_{i=1}^n X_i$ where X_i are i.i.d. with mean μ and variance σ^2. Then if let

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

as $n \to \infty$, Z_n has mean 0 and variance 1, giving us information about the long-term behavior of S_n. The Central Limit Theorem tells us more precisely that $\mathbb{P}(Z_n \leq x) = \Phi(x)$ for every x. Let's prove CLT.

Proof of CLT. Let $X_1, X_2, \ldots, X_n$ be i.i.d. with mean 0 and variance 1 and let $M_X(s)$ be the MGF of each of the X_i's. Note that by definition,

$$Z = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \implies \mathbb{E}[Z] = 0, \mathrm{Var}(Z) = 1.$$

Our goal is to be able to find the MGF of Z as well. If we expand, we will get

$$\begin{aligned} M_Z(s) = \mathbb{E}[e^{sZ}] &= \mathbb{E}\left[e^{s\frac{1}{\sqrt{n}}(X_1+X_2+\cdots+X_n)}\right] \\ &= \mathbb{E}\left[e^{\frac{sX_1}{\sqrt{n}}}\right]\mathbb{E}\left[e^{\frac{sX_2}{\sqrt{n}}}\right]\ldots\mathbb{E}\left[e^{\frac{sX_n}{\sqrt{n}}}\right] \\ &= \mathbb{E}\left[e^{\frac{sX_i}{\sqrt{n}}}\right]^n = \left[M_X\left(\frac{s}{\sqrt{n}}\right)\right]^n \end{aligned}$$

Now recall Taylor's theorem: any infinitely differentiable function can be written as $f(x) = f(a) + f'(a)(x-a) + \cdots + f^{(n)}(a)(x-a)^n + \ldots$ (the *Taylor* Series). So,

$$\begin{aligned} M_X(s) &= M_X(0) + M_X'(0)s + M_X''(0)\frac{s^2}{2!} + M_X'''(0)\frac{s^3}{3!} + \ldots \\ &= 1 + \mathbb{E}[X]s + \mathbb{E}[X^2]\frac{s^2}{2} + \mathbb{E}[X^3]\frac{s^3}{6}\ldots \\ &= 1 + \frac{1}{2}s^2 + \frac{s^3}{6}\mathbb{E}[X^3] \end{aligned}$$

where we used our earlier facts that $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2 = 1$. Plugging this into our MGF for Z gives

$$\begin{aligned} M_Z(s) = \left[M_X\left(\frac{s}{\sqrt{n}}\right)\right]^n &= \left[1 + \frac{s^2}{2n} + \frac{s^3}{6n^{3/2}}\mathbb{E}[X^3] + \ldots\right]^n \\ \implies \lim_{n\to\infty} M_Z(s) &= \lim_{n\to\infty}\left[1 + \frac{s^2}{2n} + \frac{s^3}{6n^{3/2}}\mathbb{E}[X^3] + \ldots\right]^n. \end{aligned}$$

This looks really similar to our classic limit of the form $\lim_{n\to\infty}(1 + \frac{x}{n})^n = e^x$, but its got these higher order terms that we'd ideally like to ignore. Turns out that those terms actually don't matter; the only ones we care about are the first two. We write this as $o(\frac{1}{n})$ (little o notation) to show that in the limit these terms disappear, so

$$\lim_{n\to\infty} M_Z(s) = \lim_{n\to\infty}\left[1 + \frac{s^2/2}{n} + o\left(\frac{1}{n}\right)\right]^n = e^{s^2/2} = M_Y(s)$$

completing our proof. □

Here's an application of CLT to testing light bulbs.

Example 10.1. Light bulbs have i.i.d Expo(λ) lifetimes. We want to make sure that $\frac{1}{\lambda} > 1$. Say we measure the average lifetimes A_n of $n = 100$ bulbs and find $A_{100} = 1.2$. Then $A_n = \frac{1}{n}\sum_{i=1}^{n} S_i$, so

$$\mathbb{E}[A_n] = \frac{1}{\lambda}$$
$$\text{Var}(A_n) = \frac{1}{n^2} \cdot n \cdot \frac{1}{\lambda^2} = \frac{1}{n\lambda^2}.$$

Question. What is the confidence that we have $\frac{1}{\lambda} > 1$?

Let $Z_n = \dfrac{A_n - \frac{1}{\lambda}}{\frac{1}{\lambda\sqrt{n}}}$, so that Z_n has 0 mean and variance 1. Then by CLT, $Z_n \sim \mathcal{N}(0,1)$. Taking $n = 100$ gives

$$Z_{100} = \frac{A_{100} - \frac{1}{\lambda}}{\frac{1}{10\lambda}} = 10(1.2\lambda - 1) = 12\lambda - 10.$$

So to find our probability, we simply calculate

$$\mathbb{P}(\lambda < 1) = \mathbb{P}(12\lambda - 10 < 2) = \mathbb{P}(\mathcal{N}(0,1) < 2) = 97.5\%$$

Note this approximation is an asymptotic estimation and not a bound, but a damn good one at that.

10.2 Information Theory

This entire field was born with Claude Shannon's 1948 paper *A mathematical theory of communication*, which was actually rejected from the publishing journal Shannon sent it to for not being rigorous enough. The reviewer of the paper remarked 30 years later that, "One of my biggest regrets was rejecting that paper."

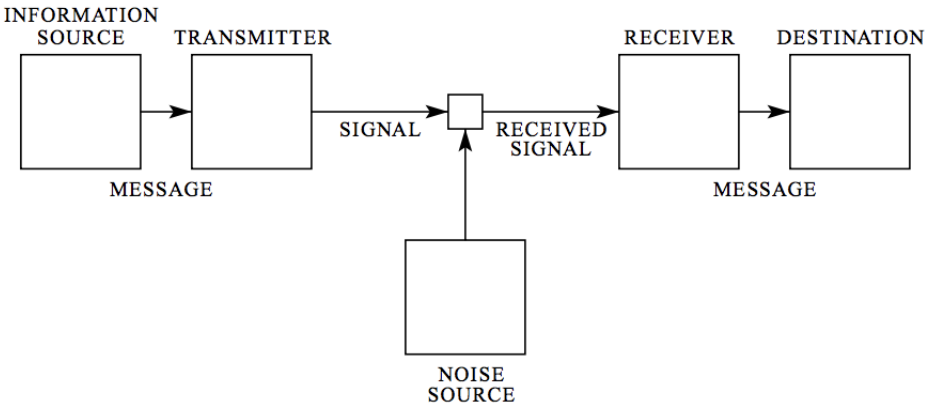

Fig. 1—Schematic diagram of a general communication system.

Figure 3: Shannon's proposed communication framework (Shannon 1948).

Theorem 16 (Separation)**.** There is no loss of optimality in separating source-coding (compression) from channel coding (reliable communication).

Example 10.2 (Source Coding)**.** Let $X \sim \{0,1\}$ with $\mathbb{P}(X = 1) = p$, and let $X^{(n)} = \{010\dots\}$ be an n-length Ber(p) source. Suppose we have a file of length 10,000. and it was Ber(p) i.i.d. What is the compression limit?

The entropy $H(X)$ is

$$H(X) = -\sum_{x\in\{0,1\}} p(x)\log p(x) = -p\log p - (1-p)\log(1-p) = h(p).$$

If we plot $h(p)$, we can see a parabola like shape with a maximum of 1 at $p = 0.5$, so we can't compress it at all. If $p = 0.11$, then $h(p) = 0.5$ between symbols, so a 10,000 length file has a 5,000 length compression limit.

If one is interested in more Information Theory, look into taking EE 229A.

10.3 Asymptotic Equipartition Property

If I have an n-length Ber(p) sequence, we will have np heads and $n(1-p)$ tails. Then AEP says that out of the 2^n total sequences, the number of sequences that I can expect to see, $\binom{n}{np}$, is approximately $2^{nh(p)}$ (using Stirling's apprx.), so each sequence appears with probability $2^{-nh(p)}$.

We also talked about Binary Erasure Channels and their capacities. See the next lecture for more info.

11 Thursday, February 28th

HW 6 is released, due next Wednesday. Reading is Walrand Chapter 1, 13.3 and B&T Chapters 7.1-7.4, plus the notes on BECs.

11.1 Capacity of the BEC

Recall that a **Binary Erasure Channel** is a model where a transmitter sends a bit, either 0 or 1, through a channel, and the receiver either receives the bit or is notified that the message was erased. Here is an illustration.

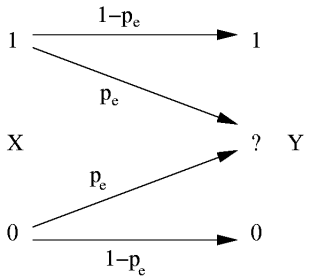

Figure 4: In a BEC, our transmissions are erased with probability p_e. Source: Wikipedia

The goal of today is to find the *capacity* of the BEC(p) channel. Capacity is the maximum rate of reliable communication, which we can define as

$$\text{Rate } = \frac{\#\text{ of bits reliably sent}}{\#\text{ of channel uses}} = \frac{L(n)}{n} \text{ bits/ch use.}$$

We also make the following definitions (which can be seen in the diagram):

$$f_n(\cdot) \text{ is the encoding function that maps } \{0,1\}^L \to \{0,1\}^n$$
$$g_n(\cdot) \text{ is the decoding function that maps } \{0,1,*\}^n \to \{0,1\}^L$$

We can let $P_e^{(n)} = \max_{m=\{0,1\}^L} \mathbb{P}(m \neq \hat{m})$ denote the probability of error. Let me explain why this makes sense. We take the max to get the "worst case" error, and we take $m = \{0,1\}^L$ as an approximation of sending many messages. Of course, the probability is precisely the event of an error, when our sent message is different from the decoded message.

Finally, let $R = L/n$ bits/channel use. We say that rate R is *achievable* for the channel if for every n, there exists encoding and decoding functions such that $P_e^{(n)} \xrightarrow[n\to\infty]{} 0$ (which exact mode of convergence is t.b.d.). The largest achievable rate R is called the **capacity** of the channel, denoted $C_{\text{BEC}(p)}$.

Now the stage is set for us to state Shannon's theorem.

Theorem 17 (Shannon 1948). $C_{\text{BEC}(p)} = 1 - p$ bits per channel use.

There are two statements hidden within this theorem. The capacity of the BEC is at most $1 - p$, and that this maximum is also attainable.

Proof. We first show that the capacity cannot exceed $1 - p$. Assume we have a friendly genie who relays instantaneously to the sender (TX) whenever the received symbol is a $*$. Then the best rate is to resend whichever symbols are erased. Hence, the time for a bit to get through

the channel is approximately Geo$(1-p)$, so the expected time it takes a bit to get through is $1/(1-p)$ channels per bit. Hence, $C \leq 1-p$ bits per channel use.

Now we do the forward direction to show this maximum is attainable. We'll show that $R = 1-p-\epsilon$ for all $\epsilon > 0$ is achievable. Shannon's insight was to leverage the Strong LLN. By Strong LLN, the probability of channel erases exactly np symbols is 1. In other words, as $n \to \infty$, $\mathbb{P}(np \text{ bits erased}) \to 1$.

Next we populate a lookup table of size 2^L by n with i.i.d. Ber$(1/2)$ entries. Call this table a *codebook* $\mathcal{C}$, and allow it to be shared between the sender and the receiver before hand.

	1	2	3	...	n
1	1	1	0	...	1
2	0	1	1	...	0
3	1	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2^L	1	0	0	...	1

Figure 5: An example of a codebook $\mathcal{C}$.

Suppose we transmit a message in the ith row of $\mathcal{C}$. On average, by SLLN, there are np bits that are erased. WLOG assume they are at the end of the message. The receiver will drop the last np columns of the codebook to obtain a truncated codebook $\mathcal{C}'$. Then he will follow these rules for decoding:

1. If c'_j is the *only* entry in $\mathcal{C}'$ matching $Y^{(n(1-p))}$ (Y^n without the last np bits), then decode $\hat{m} = j$.
2. Else, declare ERROR.

If this sounds like a really dumb idea you're not wrong. But it turns out this is just enough for us to attain the maximum. To see this, we need to calculate the probability of an error.

$$\begin{aligned}\mathbb{P}(\text{error}) = \mathbb{P}(c'_i \text{ is not unique}) &= \mathbb{P}\left(\bigcup_{i\neq j}\{c'_i = c'_j\}\right) \\ &= \sum_{i\neq j} 2^{-n(1-p)} < 2^L 2^{-n(1-p)}\end{aligned}$$

by the union bound. Hence, $P_e \leq 2^{n(R-(1-p))}$. In order for $P_e \xrightarrow[n\to\infty]{} 0$, we need $R-(1-p) < 0$, or $R < 1-p$. So we just make $R = 1-p-\epsilon$ so that $P_e \leq 2^{-n\epsilon} \xrightarrow[n\to\infty]{} 0$ for all $\epsilon > 0$. □

Here's a quick engineering example with some numbers.

Example 11.1. Suppose we take $n = 10,000$, $p = 1/2$ and $\epsilon = 0.01$. Then

$$C_{\text{BEC}(\frac{1}{2})} = \frac{1}{2} \text{ bits/ch. use} \implies C = 5000 \text{ bits.}$$

So we set our R to $1 - \frac{1}{2} - 0.01 = 0.49$ which means $L = 4900$ bits. Hence $P_e \leq 2^{-n\epsilon} = 2^{-100} \approx 0$. This means that we can can send 4900 with basically no errors. Neat.

11.2 Markov Chains

We will often want to study stochastic processes $X = \{X_t\}_{t \in T}$, which are a collection of RVs, where the index t often refers to a representation of time. X models the evolution of a sequence of RVs as a function of time. Some examples are stock prices, your wealth, customers, etc.

In general, to characterize the behavior of $X : (X_1, X_2, \ldots, X_n)^\infty$, we would need the joint pdf of $X_1, X_2, \ldots, X_n$. This is a *bad idea*, since it will very quickly grow too large for any reasonable computation. Hence we impose some structure on the process and get a markov chain.

Definition 18. Let $\mathcal{X}$ be a finite set (called the state space) with random variables X_i drawn from it. Then if

$$\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}),$$

we call X_n a **Markov Chain**.

The above property is called the "amnesia" or Markov property, since your state today only depends on your state yesterday and so on. Based on the Markov Chain, you can also come up with its transition matrix, which is just a matrix that encodes the Markov property, $\mathbb{P}(X_n | X_{n-1})$.

12 Tuesday, March 5th

I couldn't go this day, but all we convered was CS 70 level Markov Chain. Note 21 would be a good review here.

13 Thursday, March 7th

HW 7 is due next Wednesday, Lab 4 due tomorrow. Reading is Walrand Chapters 1, 2.4-2.6, 13.3 and B&T Sections 7.1-7.4.

These notes are still under construction; there's some pictures I need to add, edits I need to make but I have important deadlines these two weeks :(.

13.1 Discrete Time Markov Chains

Recall that we have two important notions regarding DTMCs:

- If a finite Markov Chain is *irreducible*, it has a unique invariant distribution π^*.
- If a Markov Chain is also *aperiodic*, $\lim_{n\to\infty} \pi_n^* = \pi^*$.

Example 13.1. Flip a fair coin repeatedly until you get 3 successive heads. What is the expected number of coin flips to 3 consecutive heads?

We can do this by defining a Markov Chain where the states are the number of successive heads, since the only information that matters for us is the number of heads in a row.

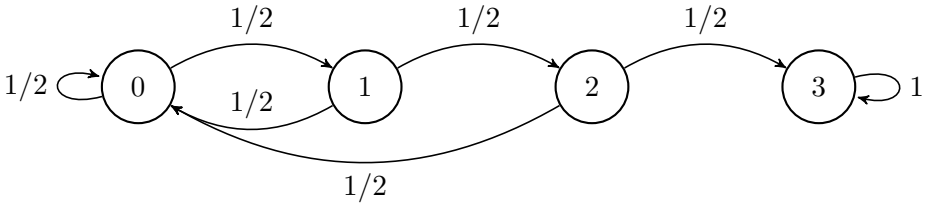

Figure 6: Casting the problem with a Markov Chain

We are looking for $T_3 = \min(n \geq 0, X_n = 3)$. Define $\beta(i) = \mathbb{E}[T_3 | X_0 = i]$ for $i = 0, 1, 2, 3$, so that $\beta(i)$ is the expected number of steps it takes to reach state 3 from state i. Our First-Step Equations (FSEs) are

$$\begin{aligned}
\beta(0) &= 1 + \frac{1}{2}\beta(1) + \frac{1}{2}\beta(0) \\
\beta(1) &= 1 + \frac{1}{2}\beta(2) + \frac{1}{2}\beta(0) \\
\beta(2) &= 1 + \frac{1}{2}\beta(3) + \frac{1}{2}\beta(1) \\
\beta(3) &= 0.
\end{aligned}$$

Solving gives us $\beta(3) = 0$, $\beta(2) = 8$, $\beta(1) = 12$, and $\beta(0) = 14$ which is our desired answer.

Example 13.2. A clumsy drunk monkey climbs a ladder. He goes up one step with probability $p = 0.8$, otherwise he slips to the ground. What is the average time taken to go to the nth rung (say $n = 10$)?

The states here should be the number of rungs climbed so far. Clearly we are interested in hitting time of the state n from the state 0. Here's the Markov Chain for this situation:

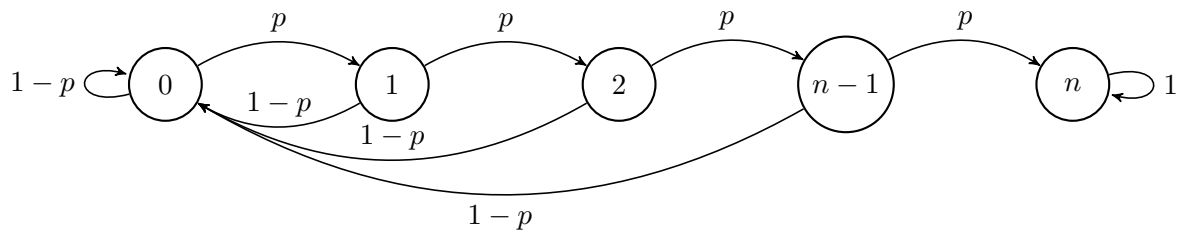

Figure 7: The drunk monkey Markov Chain

We can calculate our First Step Equations to be

$$\beta_m = 1 + p\beta_{m+1} + (1-p)\beta_0 \text{ for } m = 0, 1, \ldots, n-1$$
$$\beta_n = 0$$

It is *obvious* to see from here that the solution is $\beta_0 = \dfrac{1-p^n}{p^n - p^{n+1}}$. If $p = 0.8$ and $n = 10$ as we desire, then $\beta_0 = 41.5$, which is a very long time for the poor monkey.

13.2 General Hitting Times

Moving to generality, we can state hitting times of not just singular states, but sets of states, in Markov Chains. Suppose we have a Markov Chain $\mathcal{X}$ on a state space $\mathcal{S}$ with transition matrix P. Then for some subset $A \subseteq \mathcal{S}$, the **hitting time** T_A of that subset is defined as

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

We can define First Step Equations as usual in the following way. Let $\beta(i) = \mathbb{E}[T_A | X_0 = i]$. Then,

$$\beta(i) = \begin{cases} 1 + \sum_j P(i,j)\beta(j) & \text{if } i \notin A \\ 0 & \text{if } i \in A. \end{cases}$$

Here's another applciation of the same approach. I want to find the probability of hitting a set A before hitting a set B, where A and B are disjoint sets (why is this important?). Define $\alpha(i) = \mathbb{P}(T_A < T_B | X_0 = i)$ as the probability we hit A before B when we start at i, where

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

is the hitting time of set A. Then our FSEs in this case are

$$\alpha(i) = \begin{cases} 0 & \text{if } i \in B \\ 1 & \text{if } i \in A \\ \sum_j P(i,j)\alpha(j) & \text{if } i \notin A \cup B. \end{cases}$$

A classic application of this setup is the Gambler's Ruin, but we won't discuss it for the sake of time.

13.3 Classification of General DTMCs

Definition 19. A state i is said to be **transient** if given that we start at state i, there is a nonzero probability that we never return to state i. Otherwise the state is **recurrent**, i.e. we will return to state i with probability 1.

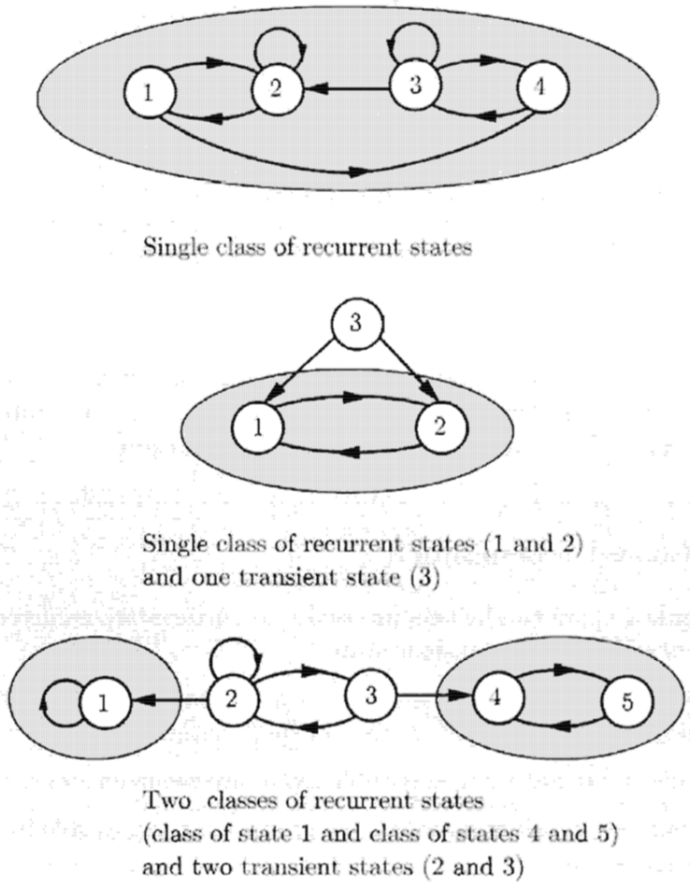

Figure 7.9: Examples of Markov chain decompositions into recurrent classes and transient states.

Figure 8: Page 350 of Bertsekas and Tsitsiklis.

There's a good illustration of recurrent and transient states on page 350 of B&T. They're pretty easy to understand lol.

Futhermore, if an irreducible Markov Chain is recurrent, then

$$\text{if } \begin{cases} \mathbb{E}[T_i | X_0 = i] < \infty & \text{positive recurrent} \\ \mathbb{E}[T_i | X_0 = i] = \infty & \text{null recurrent} \end{cases}$$

The idea is that there are two different classes of recurrent. Recurrence implies that you eventually return to your state; positive recurrence just means you return within some finite time, whereas null recurrence means you return to that state "in theory", but realistically you really don't.

There's a cool example of a markov chain called the *Gambler's Ruin* problem which is either transient, positive recurrent, or null-recurrent. I'll add it in once I create time to do so.

Now we have a Big theorem.

Theorem 20 (Walrand 13.2)**.** Suppose we have a finite Markov Chain. Then we can classify when a Markov Chain is transient, positive recurrent, or null-recurrent. To be added.

We end on a cute example of a null recurrent Markov Chain.

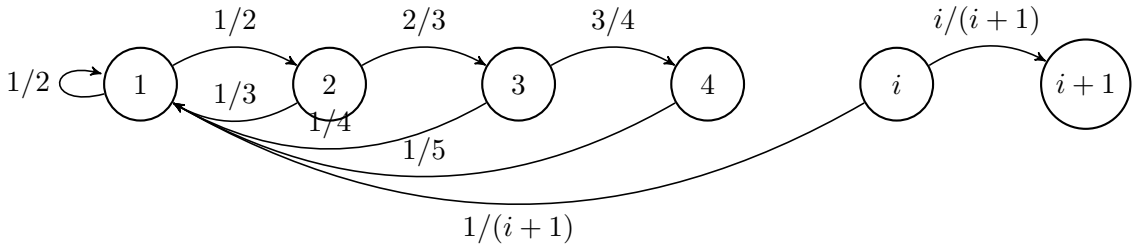

Figure 9: Null Recurrence

If we calculate the hitting time of T_1, we find that

$$\mathbb{E}[T_1|X_0 = 1] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} \cdot 2 + \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdot 3 + \dots$$

TBF.

14 Tuesday, March 12

HW 7 is due tomorrow, Lab 5 due Friday. Reading for DTMC's is Walrand Ch. 1, 2.4-2.6, and 13.3; B%T 7.1-7.4. Reading for Poisson Processes is Walrand 13.4 and B&T 6.1-6.3.

14.1 DTMC's: Recap of Recurrence, Transience

This was a recap of the gambler's ruin like example. Didn't catch it, but I'll edit it in once I get time (spring break?).

14.2 Reversibility Markov Chains

Assume we have an irreducible and positive recurrent Markov Chain started at its unique invariant distribution π. Suppose for every n, $(X_0, X_1, \ldots, X_n)$ has the same joint pmf as the time-reversed chain $(X_n, X_{n-1}, \ldots, X_0)$. Then we call the chain *reversible*.

Fact 21. Reversible or not, if we start a MC at its invariant distribution π, the time-reversed sequence is a Markov Chain.

Proof. We have that

$$\begin{aligned}\mathbb{P}(X_k = i | X_{k+1} = j, X_{k+2} = i_{k+2}, \ldots, X_n = i_n) &= \frac{\mathbb{P}(X_k = i, X_{k+1} = j, \ldots, X_n = i_n)}{\mathbb{P}(X_{k+1} = j, X_{k+2} = i_{k+2}, \ldots, X_n = i_n)} \\ &= \frac{\pi(i)P_{i,j}P_{j,i_{k+2}} \cdots P_{i_{n-1},i_n}}{\pi(j)P_{j,i_{k+2}} \cdots P_{i_{n-1},i_n}} \\ &= \frac{\pi(i)P_{i,j}}{\pi(j)}\end{aligned}$$

Hence, $\tilde{P}_{j,i} = \dfrac{\pi(i)P_{i,j}}{\pi(j)}$, which only depends on i, j, so the sequence is a Markov Chain. □

From this proof, we can see that for reversibility, we need

$$\tilde{P}_{ji}(\text{reverse chain}) = P_{ij}(\text{forward chain}).$$

This is satisfied when

$$\pi(j)\tilde{P}_{j,i} = \pi(i)P_{i,j}$$

for all $i, j \in \mathcal{X}$ in a reversible Markov Chain. We call these the **detailed balance equations**, but Ramchandran likes to call them Local Balance Equations.

Theorem 22. If a Markov Chain is reversible, then it has an invariant distribution π.

Recall that if a Markov Chain has an invariant distribution π, $\pi = \pi P$. We call this the *Global Balance Equation*, as opposed to the Local Balance Equations from above.

Proof. Suppose we have some distribution of states π. Then

$$\sum_i \pi(i)P_{i,j} = \sum_i \pi(j)\tilde{P}_{j,i} = \pi(j)\sum_i \tilde{P}_{j,i} = \pi(j).$$

Hence for every j, the local balance equations imply a global balance equation, so $\pi = \pi P$, and so our Markov Chain has an invariant distribution. □

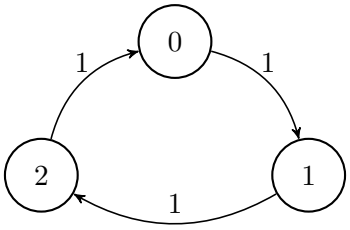

Figure 10: Is this chain reversible?

We can employ reversibility to solve for transience in the earlier example. Here's another quick application.

Is this chain reversible? If we list the possible states going forwards and backwards, they are

$$\text{Forwards: } 012012\ldots$$
$$\text{Backwards: } 210210\ldots$$

but we can never get from 2 to 1 in this chain. Hence, it is not reversible. Another way to see this is that by the local balance equations,

$$\pi_0(1) = \pi_1(0) = 0,$$

so $\pi_0 = 0$. Similarly, $\pi_1 = 0$ and $\pi_2 = 0$ which is a contradiction, since they must add up to 1.

There's some other cool ways to determine if a Markov Chain is reversible. One of them is Kolmogorov's criterion, which gives a necessary and sufficient condition based on only the transition probability. Out of scope though.

14.3 Introduction to Poisson Processes

Possion Processes are continuous versions of the "coin-flipping" or Bernoulli processes, where "arrival" times are continuous. Some motivation is that it's a good model for arrivals of packets at a router, photons hitting a photon detector, etc.

One way to define Poisson Processes, which is the way Walrand follows and so will we, is to have $S_1, S_2, \ldots, S_n$ which are i.i.d. Expo(λ) ($\lambda > 0$) RVs, and define T_i's as the CT arrival times. Then we can count arrivals as

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

15 Thursday, March 14

Happy π day everyone. HW 8 is out, due next Wednesday, Lab 5 is due on Friday (tomorrow), and the Reading is on Poisson Processes, Walrand 13.4 and B&T 6.1-6.3.

15.1 Poisson Processes

We can define Poisson processes by their arrival times, but also their interarrival times S_i, where $S_i \sim \text{Expo}(\lambda_i)$. Then $T_n = \sum_{i=1}^{n} S_i$ is our Poisson Process.

Let's do a quick review of Exponential RVs first. Recall that the density of the exponential distribution is $f_{S_i}(t) = \lambda e^{-\lambda t}$. Then the CDF is just

$$F_{S_i}(t) = \mathbb{P}(S_i \leq t) = \begin{cases} 1 - e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{else} \end{cases}$$

We also have that $\mathbb{E}[S_i] = 1/\lambda$ and $\text{Var}(S_i) = 1/\lambda^2$, and of course the memoryless property, which says that

$$\mathbb{P}(S_i > t + s | S_i > s) = \mathbb{P}(S_i > t),$$

i.e. the probability of an arrival is independent of how long we've waited.

Also,

$$\mathbb{P}(S_i \leq t + \epsilon | S_i > t) = \lambda\epsilon + o(\epsilon),$$

where $o(\epsilon)$ are the higher order terms, i.e. terms of the form ϵ^2, ϵ^3, etc. Formally, $\lim_{e \to 0} \frac{o(\epsilon)}{\epsilon} = 0$.

Why? Well by memoryless,

$$\mathbb{P}(S_i > t + \epsilon | S_i > t) = \mathbb{P}(S_i > \epsilon) = e^{-\lambda\epsilon} = 1 - \lambda\epsilon + o(\epsilon)$$

by a Taylor Series expansion. Then subtracting from 1 gives us our result above.

Now divide up an interval into ϵ long chunks. For each interval, let's ask the question of how many arrivals we will have in such an interval. Using our previous derivations, we find that

$$\begin{aligned} \mathbb{P}(\text{no arrivals in this interval}) &= 1 - \lambda\epsilon + o(\epsilon) \\ \mathbb{P}(1 \text{ arrival in an } \epsilon \text{ interval}) &= \lambda\epsilon + o(\epsilon) \\ \mathbb{P}(2 \text{ or more arrivals in the interval}) &= 1 - (1 - \lambda\epsilon + \lambda\epsilon + o(\epsilon)) = o(\epsilon) \end{aligned}$$

Hence as $\epsilon \to 0$, the probability of there being more than 1 arrival in any interval quickly goes to 0. So we can treat the Poisson Process also as arrivals within ϵ wide intervals, which is how the Bertsekas textbook defines PPs. From here, one can show that the interarrival times follow exponential distributions, but this is pretty clunky, so we defined them the other way around like Walrand does. We'll see later the motivation for doing so this way.

Here's a theorem that shouldn't be surprising.

Theorem 23. Poisson processes are memoryless.

One can think of Poisson processes as inheriting this memoryless property from exponential distributions. But what does it mean for a process to be memoryless? We know what it means for RVs to be memoryless; turns out the equivalent notion is similar. In pictures, if N_t is a PP(λ), then so is $N_{t+s} - N_t$.

Here's some implications. For starters, PP(λ) has *independent* and *stationary* increments. For any $0 \leq t_1 \leq t_2 \leq \ldots$, $\{N_{t+s} - N_t\}$ are independent and distribution depends only on $t_{n+1} - t_n$.

Proof with words. A PPλ has inter-arrival times that are independent Expo(λ) RVs. So, for $t > T_3$ in the picture, it is obvious that the inter-arrival times are Expo(λ) by constraint. The only possible issue is with the *first* arrival (i.e. $s < T_3 - t$). But, by the memoryless property of the Expo(λ) RV, the first inter-arrival distribution is also Expo(λ). □

You're probably also wondering by this point, why the hell are these called Poisson Processes? Literally everything about these processes has been exponential, so where does the Poisson come in? Here's why.

Theorem 24 (Walrand 13.7)**.** If $\mathcal{N} = \{N_t, t \geq 0\}$ is a PP(λ), then N_t, the # of arrivals in $(0, t)$, has a Poisson(λt) distribution, i.e.

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

for $k \geq 0$.

There's two proofs, one which uses a clever symmetry argument and another with differential equations. We're going to sketch the former thankfully. Bertsekas actually assumes this property of Poisson Processes and goes backwards to show inter-arrival times are exponential. Two equivalent ways of motivating Poisson Processes.

Proof. If our goal is to find the distribution of N_t, then we need a density for the number of arrivals in $(0, t)$. This means we will first find the joint density of $T_1, T_2, \ldots, T_k, T_{k+1}, \ldots$. This is equivalent to

$$\mathbb{P}(T_1 \in \{t_1, t_1 + dt_1\}, T_2 \in \{t_2, t_2 + dt_2\}, \ldots, T_k \in \{t_k, t_k + dt_k\}, T_{k+1} > t).$$

We'd like to convert these to S_i's however, because they're exponentials and easy to work with. So doing so gives

$$\mathbb{P}(S_1 \in \{0, t+dt_1\}, S_2 \in \{t_2-t_1, t_2-t_1+dt_2\}, \ldots, S_k \in \{t_k-t_{k-1}, t_k-t_{k-1}+dt_k\}, S_{k+1} > t-t_k).$$

We can approximate the probability of $S_i \in t_i + dt_i$ as a rectangular region whose area is $f_{S_i}(t_i) \cdot dt_i = \lambda e^{-\lambda t_i} dt_i$, so this probability reduces to

$$\begin{aligned}(\lambda e^{-\lambda t_1} dt_1)(\lambda e^{-\lambda(t_2-t_1)} dt_2)(\lambda e^{-\lambda(t_3-t_2)} dt_3) \ldots (\lambda e^{-\lambda(t_k-t_{k-1})} dt_k)(e^{-\lambda(t-t_k)}) \\ = \lambda^k e^{-\lambda t} dt_1 dt_2 \ldots dt_k\end{aligned}$$

by cancellation (!! sick !!). Hence,

$$\boxed{f_{T_1,T_2,\ldots,T_k}(t_1, t_2, \ldots, t_k) = \lambda^k e^{-\lambda t}}.$$

Now notice that f is *uniform* over the support of $(T_1, T_2, \ldots, T_k)$. Hence,

$$N_t(k) = \int_{t_1}\int_{t_2}\cdots\int_{t_k} f_{T_1,T_2,\ldots,T_k}(t_1, t_2, \ldots, t_k) dt_1 dt_2 \ldots dt_k.$$

The inside is independent of $t_1, \ldots, t_k$, so

$$N_t(k) = \lambda^k e^{-\lambda t}\int_0^t\int_0^t\cdots\int_0^t dt_1 dt_2 \ldots dt_k$$

subject to the constraint that the arrivals happen in order, i.e. $S : t_1 < t_2 < \cdots < t_k$.

Without any constraints, $Vol(S) = t^k$. But this accounts for *all* permutations of $(t_1, \ldots, t_k)$. By symmetry, all permutations have the *same* volume though! There are $k!$ different permutations, of which 1 is actually legal. Hence the volume we actually care about is $\frac{1}{k!}t^k$. Combining all together, we get that

$$N_t(k) = \lambda^k e^{-\lambda t}\left(\frac{1}{k!}t^k\right) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

as desired. □

Cleeeeean.

15.2 Splitting and Merging of Poisson Processes

If we have have two independent PPs $\text{PP}(\lambda_1)$ and $\text{PP}(\lambda_2)$, we can **merge** them to be a $\text{PP}(\lambda_1 + \lambda_2)$.

Likewise, if we have a $\text{PP}(\lambda)$, then it can split into two Poisson Processes $\text{PP}(\lambda p)$ and $\text{PP}(\lambda(1-p))$ where p is the probability of choosing an interval to split or not. That's basically all there is to splitting and merging, so let's do some examples.

Example 15.1 (Fishing). Bob catches fish according to a Poisson Process with rate $\lambda = 0.6$ fish/hr. If he catches at least one fish in the first 2 hours, he quits (after two hours). Else, he continues till he has caught the first fish.

Question. What is $\mathbb{P}(\text{Bob fishes for} > 2 \text{ hours})$?

Answer: Happens only if he catches 0 fish in $[0, 2]$ hours, which is $\mathbb{P}(N(2) = 0) = e^{-\lambda 2} = e^{-1.2}$ by using the CDF.

Question. What is $\mathbb{P}(\text{Bob fishes for time} t \in [2, 5] \text{ hours})$?

Answer: Two events must occur: he must catch no fish in $[0, 2]$, and also catch one fish in $[2, 5]$. The first event has probability $e^{-1.2}$, while the second event has probability $1 - e^{-1.8}$. By independence, our overall probability is $e^{-1.2}(1 - e^{-1.2})$.

Question. What is $\mathbb{P}(\text{Bob catch at least 2 fish})$?

Answer: Similar to the beginning of lecture. $1 - e^{-1.2} - 1.2e^{-1.2}$.

Question. What is $\mathbb{E}[\text{fish caught by Bob}]$?

Answer: By Linearity of Expectation,

$$\mathbb{E}[\text{fish caught}] = \mathbb{E}[\text{fish caught in}(0, 2)] + \mathbb{E}[\text{fish caught in}(2, \infty)]$$

$$= 1/(\lambda t) + 1 \cdot e^{-2\lambda} = 1.2 + e^{-1.2}.$$

16 Tuesday, March 19th

Readings on CTMC are Walrand 13.5 and B&T, 7.5.

16.1 Poisson Processes Recap

We did a lot of review of last week; the first new thing we did was the following example.

Example 16.1. Two light bulbs have burn out times distributed according to Expo(λ_a) and Expo(λ_b) respectively. What's the expected first burn out time of either light bulb?

We can treat the burn out times as a Poisson Process, where each of the lightbulbs have parameters λ_a and λ_b. Then the expected first burn out (a.k.a. the min of the two random variables) is equivalent to the first arrival of the merged processes, which would be distributed according to Expo($\lambda_a + \lambda_b$).

16.2 Erlang Distributions

16.3 Random Incidence Phenomenon

17 Thursday, March 21st

17.1 Continuous Time Markov Chains

17.2 Balance Equations

17.3 Hitting Times

18 Tuesday, April 2nd

HW 9 is due tomorrow. We have a midterm next Tuesday, and there are review sessions on Saturday, April 6th from 1-3pm. Reading on Random Graphs is to be posted.

18.1 Wrap up CTMC's

Here's an example of a continous time markov chain.

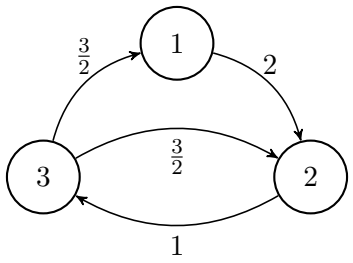

Figure 11: A Continuous Time Markov Chain

Our transition matrix is just

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

Then to find the stationary distribution, we just solve for the π such that $\pi Q = 0$ and $\sum \pi_i = 1$. We get

$$\pi = \frac{1}{19}[3 \quad 12 \quad 4].$$

But what if we wanted to calculate the amount of time we're spending to get from one state to another? It's clear what the transitions are since the rates are specified, but it's not so clear how we quantify time spent at each state. The out-going rates at state 1, 2, is slower than the out-going rate at state 3, 3, which would make it unfair to compare their times naively.

The solution then is to scale our times according to the "fastest" clock, which would be state 3's clock in this case. Suppose you want to find π based on equating it to a DTMC. Then we can construct the following discretized markov chain.

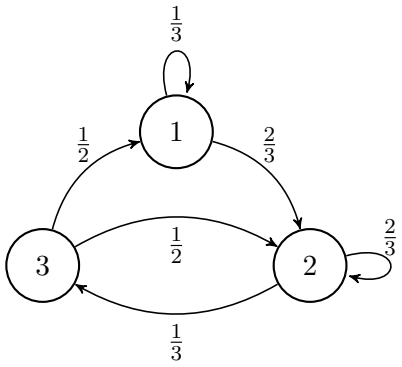

Figure 12: Discretizing our continuous markov chain

This alllows us to compute a stationary distribution according to the transition matrix

$$P = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

And now if you solve for $\pi P = \pi$, you'll get the same exact answer that we had before. Neat.

18.2 Random Graphs

We're stealing from MIT again, since their 6.207/14.15 notes are pretty good (who would've thought). It's actually a graduate course in economics, so it's interesting to see how probability theory still shows up in far away fields.

First off, there's lots of applications to the behavior of social networks, biological networks, matrix completion, etc. Some examples are understanding if a population will become extinct or if an epidemic will wipe out all of the species.

Definition 25. The models of random graphs we will use are called **Erdos-Renyi** Random Graphs. We denote them by $\mathcal{G}(n,p)$ where n is the number of vertices, and p is the probability of an edge being present or not. Typically $n \in \mathbb{Z}^+$, $0 \leq p \leq 1$.

Here's one way to view this. Flip a coin for every edge to determine if its present. This is independent for each edge in the graph.

We can ask some preliminary questions about random graphs, such as...

Question. What is $\mathbb{E}[\#\text{ of edges}]$?

Answer. Using linearity of expectation, $\mathbb{E}[\#\text{ of edges}] = \binom{n}{2}p = \frac{n(n-1)}{2}p$. □

Question. What is the distribution of D, the degree of a node? What is $\mathbb{E}[D]$?

Answer. There's $n-1$ potential other edges, each of which will have probability p of appearing. The density is $p_D(d) = \binom{n-1}{p}p^d(1-p)^{n-1-d}$. Hence $D \sim \text{Bin}(n-1,p)$, so $\mathbb{E}[D] = (n-1)p$. □

Question. If $n \to \infty, p \to 0$ while $(n-1)p = \lambda$ is a constant, how do you approximate $p_D(d)$?

Answer. We approximate D as Poisson(λ), so $p_D(d) \approx e^{-\lambda}\frac{la^d}{d!}$. □

Question. What is the probability q that a node is isolated?

Answer. All edges must not be present, so $q = (1-p)^{n-1}$. □

Erdos 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 18.1 (Structural Properties)**.** Here's some examples of what we mean by structural properties. Note that all of these properties occur in the limit $n \to \infty$.

- $p = \frac{1}{n^2}$ is a threshold for when the first edge appears w.h.p.
- $p = \frac{1}{n}$ is a threshold for a "Giant Component" to emerge. That is, if $p = \frac{1-\epsilon}{n}$, we will have lots of small components (much like raindrops) of size $O(\log n)$. On the other hand, if $p = \frac{1+\epsilon}{n}$, then we will still have a few small components, but there will also be a "Giant Component" of size $O(n)$.
- We will focus on perhaps the most important property: a threshold for **connectedness**, which is $p = \frac{\log n}{n}$.

In all of our random graph ventures, we pick an n, then pick a p based on what n is. Hence all of our p will implicitly be functions of n. We assume you know this, so for convenience sake we drop the notation. Now let's prove this last property.

Theorem 26 (Erdos-Renyi 1961)**.** Let $p(n) = \lambda\frac{\log n}{n}$. Then,

(a) If $\lambda < 1$, $\mathbb{P}(\mathcal{G}(n,p) \text{ is connected}) \to 0$ as $n \to \infty$.

(b) If $\lambda > 1$, $\mathbb{P}(\mathcal{G}(n,p) \text{ is connected}) \to 1$ as $n \to \infty$.

Quick remark: this should remind you of Shannon's capacity theorem. If the capacity is anything less than the maximum, you will eventually transmit all of your message with positive probability. Otherwise, you will never be able to. Same on-off principle.

Proof. We will prove part (a) first, opting to prove something stronger. It suffices to show that $\mathbb{P}(\text{no isolated nodes}) \to 0$ as $n \to \infty$.

Let X be the number of isolated nodes in $\mathcal{G}(n,p)$. Let's find $\mathbb{E}[X]$. We use our old friend indicator variables, letting I_i be an indicator that i is isolated. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[I_i] = \sum_{i=1}^{n} \mathbb{P}(\text{node } i \text{ is isolated}) = nq$$

where q is from our preliminary excursion above. Thus,

$$\mathbb{E}[X] = nq = n(1-p)^{n-1} \approx ne^{-p(n-1)} \approx ne^{-np} = ne^{-\lambda \log n} = n^{1-\lambda}.$$

The first approximation was just a Taylor Series approximation, while the second one assumes that $e^p \approx 1$, which is true as $p \to 0$. Hence, for any constant λ, our expectation blows up as $n \to \infty$, which is really good, since we want $\mathbb{P}(X = 0) = 0$.

But is this enough? Is knowing that our expectation goes to infinity enough to state that there is no mass at 0?

Turns out its not, and one can create bad eamples where this happens. For example, we can have a variable R for which $\mathbb{P}(R = 0) = \frac{1}{n}$ and $\mathbb{P}(R = n^2) = \frac{n-1}{n}$. Expectation goes to infinity, but we still have remaining mass at 0.

To fix this, we need to control the variance as well. This is kind of annoying, but there's a nice easy lemma we can make use of.

Lemma 27. If X is a nonnegative integer valued RV, then $\mathbb{P}(X = 0) \leq \dfrac{\text{Var}(X)}{[\mathbb{E}[X]]^2}$.

Proof. We have

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{P}(X=0)\mathbb{E}[X]^2 + \mathbb{P}(X=1)\mathbb{E}[(X-1)^2] + \mathbb{P}(X=2)\mathbb{E}[(X-2)^2] + \dots \\
&\geq \mathbb{P}(X=0)\mathbb{E}[X]^2.
\end{aligned}$$

from which the result follows. □

Hence if we can show that the variance goes to 0, then so does $\mathbb{P}(X = 0)$. Expanding results in

$$\begin{aligned}
\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} I_i\right) &= \sum_{i=1}^{n} \text{Var}(I_i) + \sum_{i \neq j} \text{Cov}(I_i, I_j) \\
&= n\text{Var}(I_i) + n(n-1)\boxed{\text{Cov}(I_1, I_2)}.
\end{aligned}$$

We can treat I_i as a Bernoulli random variable with probability q of being 1, from which $\text{Var}(I_i) = q(1-q)$.

All we have left to calculate is the boxed portion, the covariance, which is

$$\text{Cov}(I_1, I_2) = \mathbb{E}[I_1 I_2] - \mathbb{E}[I_1]\mathbb{E}[I_2].$$

The first quantity asks for the probability that two nodes are both isolated. This requires all $2(n-1) - 1 = 2n - 3$ edges out of both nodes to be absent, giving a probability of $(1-p)^{2n-3}$.

The second quantity was already calculated earlier. If $q = (1-p)^{n-1}$ is the probability that a node is isolated, the product is just $q \cdot q = q^2$. Putting this all together,

$$\text{Cov}(I_1, I_2) = \frac{q^2}{1-p} - q^2 = \frac{pq^2}{1-p}.$$

So our variance is

$$\text{Var}(X) = nq(1-q) + \frac{n(n-1)pq^2}{1-p}.$$

Finally, we can use our lemma (along with $\mathbb{E}[X] = nq$) to show that

$$\mathbb{P}(X = 0) \leq \frac{\text{Var}(X)}{\mathbb{E}[X]^2} = \frac{1-q}{nq} + \frac{(n-1)p}{n(1-p)} \to 0$$

as $n \to \infty$, completing our proof. □

19 Thursday, April 4th

Thanks to Megan Kawakami for her notes on this day.

19.1 More on Random Graphs

Recall from last lecture that we were discussing random graphs, and how they could be used to prove many *structural properties* of graphs. One of these is that of connectedness in graphs which is described by Theorem 26. We will reinterpret the theorem as follows.

Suppose we have a random graph $\mathcal{G}(n,p)$ where $p(n) = \frac{\lambda \log n}{n}$. Then our two threshold scenarios can be described as follows:

(a) If $\lambda < 1$, then $\mathbb{P}(\text{no isolated nodes}) \to 0$ as $n \to \infty$.

(b) If $\lambda > 1$, then $\mathbb{P}(\text{not connected}) \to 0$ as $n \to \infty$.

We proved part (a) last lecture previously, so now we will prove part (b). Before we do so, recall from the proof of part (a) that if X is the number of isolated nodes in $\mathcal{G}(n,p)$, then $\mathbb{E}[X] = nq \approx n^{1-\lambda} \to \infty$ as $n \to \infty$.

Proof of part (b). We will prove the converse; if $\lambda > 1$, then $\mathbb{P}(\text{connected}) \to 1$.

The key idea is that a graph being disconnected is equivalent to the existence of a set of size k (where $1 \le k \le \frac{n}{2}$) such that there's no edge between this set and its complement. Using this idea, we can find the probability of the latter event and use the union bound to upper bound our probability:

$$\begin{aligned}
\mathbb{P}(\mathcal{G}(n,p) \text{ is not connected}) &= \mathbb{P}\left(\bigcup_{k=1}^{\frac{n}{2}} \{\exists \text{ a set of size } k \text{ that is disconnected from its comp set}\}\right) \\
&\le \sum_{k=1}^{\frac{n}{2}} \mathbb{P}(\exists \text{ a set of size } k \text{ that's disconnected}) \\
&\le \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} \mathbb{P}(\text{a specific set of } k \text{ nodes is disconnected from comp}) \\
&= \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} (1-p)^{k(n-k)}
\end{aligned}$$

where the last step follows since the $k(n-k)$ cross edges between components need to all be missing, which happens individually with probability $1-p$. Through some tedious calculations with non-obvious inequailty bounds, we can show that this last summation goes to 0, as desired.

As a specific example, take $k = 1$. Then we have

$$\binom{n}{1}(1-p)^{n-1} = \mathbb{E}[X] \approx ne^{-p(n-1)} \approx n^{1-\lambda} \xrightarrow{n\to\infty} 0.$$

□

19.2 Inference: Detection and Bayes' Rule

There are two main approaches to performing inference. One is the *classical* or *frequentist* approach, where the unknowns are fixed and to be estimated. The other is the *Bayesian*

approach, where the unknowns are RVs whose distributions have to be estimated. Each of these will lead to a different inference method as we will see later.

The basic premise is that we have n possible exclusive causes $C_1, C_2, \ldots, C_n$ of a particular symptom. Each cause has a prior probability p_i and it has a probability q_i of causing the symptom. We call p_i our *priors* and q_i our *posteriors*, and can illustrate the setup with the following diagram.

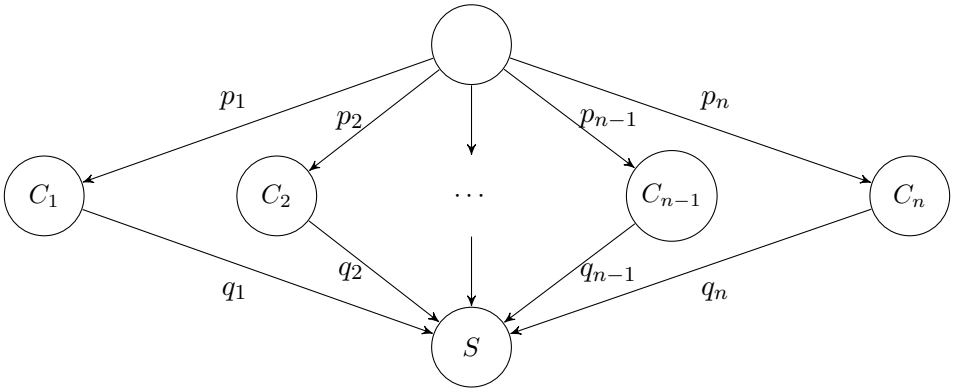

Figure 13: The basic inference setup, where the C_i's are possible causes of a symptom S.

Suppose we want the posterior probability π_i of cause i given the symptom S. In other words, the probability that cause i caused the symptoms we observed. Then we can use Bayes' Rule to get

$$\pi_i = \mathbb{P}(C_i|S) = \frac{\mathbb{P}(S|C_i)\mathbb{P}(C_i)}{\sum_{j=1}^n \mathbb{P}(S|C_j)\mathbb{P}(C_j)} = \frac{p_i q_i}{\sum_{j=1}^n p_j q_j}.$$

This is important enough to be stated as a theorem.

Theorem 28 (Posterior Probability)**.** The posterior probability π_i of a cause i is given by

$$\pi_i = \frac{p_i q_i}{\sum_j p_j q_j}.$$

19.3 Inference: MAP and MLE, and the MAP Rule

There are two main inference methods: MAP and MLE.

Definition 29 (MAP)**.** The **maximum a posteriori**, or MAP, is defined to be

$$\text{MAP} = \arg\max_i \pi_i = \arg\max_i p_i q_i.$$

In other words, it is the best estimate of the cause given a symptom.

Definition 30 (MLE)**.** The **maximum likelihood estimate**, or MLE, is defined to be

$$\text{MAP} = \arg\max_i q_i,$$

which is just the MAP estimate under a uniform prior, i.e. $p_1 = \cdots = p_n$.

More generally,

$$\text{MAP}[X|Y=y] = \arg\max_x \mathbb{P}(X=x|Y=y),$$

which can be interpreted as finding "which cause best explains the observed symptom," and

$$\text{MLE}[X|Y=y] = \arg\max_x \mathbb{P}(Y=y|X=x),$$

which can be interpreted as finding "which cause best *generates* the observed symptom."

Let's do an inference example in digital communications. Suppose we're trying to send a message X across a channel. Unfortunately, the channel is noisy, so the message received is actually Y. Given this, our goal is to find an estimate for X, $\hat{X}$, given our observation Y. In other words, we want to find the MLE and MAP estimate of X given Y.

20 Thursday, April 11th

Lab 6 is due tomorrow, projects due 4/19, and HW 10 due next Wed. Readings for MLE/MAP are W 5.1-5.4, B&T 8.1-8.2, 9.1. For Hypothesis Testing, readings are W 5.5,5.6,6.5 and B&T 9.3-9.4.

20.1 Wrap up MLE/MAP

Recall that the basic premise is that we have n causes and an observed symptoms, and we are trying to perform inference on the causes. Refer to Figure 13 for more details. Then with this convention, MAP $= \arg\max_i p_i q_i$ and MLE $= \arg\max_i q_i$. where p_i are our priors and q_i are the posteriors. Let's see how we could model this in digital communications.

Suppose we have a Binary Symmetric Channel (BSC). Then using the MAP rule, we can derive a formulation for the likelihood...

20.2 Gaussian Channel

Now suppose we have a setup with a Gaussian channel. Then,

$$f_0(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2}.$$
$$f_1(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-1)^2/2\sigma^2}.$$

Then by MAP, we find

$$p_0 q_0 \underset{1}{\overset{0}{\gtrless}} p_1 q_1$$
$$p_0 f_0(y) \underset{1}{\overset{0}{\gtrless}} p_1 f_1(y).$$

We can interpret this as

We would call the thing on the left the *likelihood* $L(y)$. One thing we can do is take logs to make this easier to evaluate (hence the term *log-likelihood*). This gives

$$\ln f_0(y) - \ln f_1(y) \underset{1}{\overset{0}{\gtrless}} \ln(\frac{p_1}{p_0})$$
$$-\frac{y^2}{2\sigma^2} + \frac{(y-1)^2}{2\sigma^2} \underset{1}{\overset{0}{\gtrless}} \ln(\frac{p_1}{p_0}).$$

Solving for y gives us a MAP of

$$\text{MAP: } \boxed{y \underset{1}{\overset{0}{\lessgtr}} \frac{1}{2} + \sigma^2 \ln(\frac{p_0}{p_1})}.$$

Of course, we can also solve for the MLE, since we just set all of our priors equal, i.e. $p_0 = p_1$.

$$\text{MLE: } \boxed{y \underset{1}{\overset{0}{\lessgtr}} \frac{1}{2}}.$$

Example 20.1. Suppose $\frac{p_0}{p_1} = e$, and $\sigma^2 = 0.1$. Then our MAP decision would be $y \underset{1}{\overset{0}{\lessgtr}} 0.6$.

Read Chapter 5 of Walrand for details of applications to digital communications.

20.3 German Tank Problem

Suppose we have n balls in a bucket, and we're trying to figure out what n is by sampling. Each ball has a serial number from 1 to n.

Suppose we sample once and get a ball labeled 17. Then what is the MLE of the maximum serial number in the bucket, n?

It would just be 17, because the probability of getting 17 if $n < 17$ is 0, and the probability of getting 17 if $n \geq 17$ is $\frac{1}{n}$ which is strictly decreasing.

In other words, $MLE[n|Y_1 = 17] = 17$, since $ML = \arg\max_i q_i$. In fact, for any $m = 1, 2, \ldots, n$, $\mathbb{P}(Y = m|X = n) = \frac{1}{n}$, so $MLE[n|Y = m] = m$.

This is completely against what our intuition says. We should be able to know something more about the total number of balls that's not obvious.

The problem is that there is a sampling bias that we aren't accounting for in the problem. Suppose we make k observations. There are $\binom{n}{k}$ sets of k distinct #'s that are subsets of $[n] = \{1, 2, \ldots, n\}$. Each set is equally likely, so

$$\mathbb{P}_Y(y, n) = \begin{cases} \frac{1}{\binom{n}{k}} & \text{if } n \geq m_k = \max(Y_1, \ldots, Y_k) \\ 0 & \text{else.} \end{cases}$$

Then the estimator we are using, the maximum likelihood estimation $m_k = \max\{Y_1, Y_2, \ldots, Y_k\}$, is actually biased! One can show that $\mathbb{E}[m_k] \neq n$; in fact, $\mathbb{E}[m_k] = \frac{k(n+1)}{k+1}$. This is contrary to what we want, since an unbiased estimator should have an expectation equal to the true value.

The bias of MLE is $b(M_k) = \mathbb{E}[M_k] - n = \frac{k-n}{k+1}$.

If we want an unbiased estimator, then we could instead use $\hat{N}_k = \frac{k+1}{k} m_k - 1$. Note now that $\mathbb{E}[\hat{N}_k] = n$ as desired, so this is indeed an unbiased estimator.

Example 20.2. Suppose $n = 155$, and we draw $k = 4$ samples. Our samples are $Y = \{17, 62, 102, 120\}$. Then $m_k = 120$, so $\hat{N}_k = \frac{5}{4} \cdot 120j - 1 = 149$.

20.4 Hypothesis Testing: Neyman-Pearson Test

The motivation for hypothesis testing is that there are many things that you can't put priors on, like having cancer, or how likely my house is to be on fire. So thinking about them in terms of MLE and MAP isn't the right approach. Alternatively, we have the mindset of giving likelihoods of these events based on our risk-aversion, a.k.a. how tolerant we are to false alarms.

In general, the setup is as follows:

- Observe a random variable Y.
- Under hypothesis H_0 (i.e. given $X = 0$), $Y \sim f(y|0)$ (no cancer).
- Under hypothesis H_1 (i.e. given $X = 1$), $Y \sim f(y|1)$ (cancer).

Then our goal is to come up with a decision rule $r : \mathbb{R} \to \{0, 1\}$.

Perhaps the most important formulation is the Neyman-Pearson (N-P) formulation. In it, we have no priors, only two hypotheses, and false negatives are way more important than false positives.

Figure 14: Table of outcomes based on our decision and the true hypothesis.

Our goal is to the maximize PCD, the probability of correct detection, which is $\mathbb{P}(\hat{x}|X=1)$ such that the probability of a false alarm, PFA, is $\leq \beta = \mathbb{P}(\hat{x}|X=0)$ ($\hat{x}$ is Walrand's way of denoting $r(Y)$).

The Neyman-Pearson theorem gives us the optimal decision rule for maximizing PCD such that $PFA \leq \beta$.

Theorem 31 (Neyman-Pearson)**.** The decision $\hat{X}$ that maximized PCD under the constraint that PFA $\leq \beta$ is

$$\hat{X} = \begin{cases} 1 & \text{if } L(Y) > \lambda \\ 0 & \text{if } L(Y) < \lambda \\ 1 & \text{w.p. } \gamma \text{ if } L(Y) = \lambda \end{cases}$$

where $L(y) = \frac{f_{Y|X}(y|1)}{f_{Y|X}(y|0)}$ is called the *Likelihood Ratio* and $\lambda > 0$ and $\gamma \in [0,1]$ are chosen to ensure that PFA $= \beta$.

This rule is sometimes called the *threshold rule*. The Neyman-Pearson theorem claims that this rule is **the most powerful** rule at some PFA β

21 Tuesday, April 16th

21.1 Hypothesis Testing

21.2 Estimation: LLSE

22 Thursday, April 18th

22.1 Hilbert Space of Random Variables

22.2 Gram Schmidt Process

23 Tuesday, April 23rd

HW 11 is due tomorrow, Projects are due on Friday. Reading is chapters 6-8 of Walrand and 8.3-8.5 of B&T.

23.1 LLSE: Recap

Recall that a Linear Least Squares Estimator (LLSE) is trying to predict a random variable X given some observations Y, denoted as $L[X|Y]$. Since the estimator is linear, our predictions are always of the form $\hat{X} = a + bY$; the error of our prediction is $\Delta = X - \hat{X}$.

We also have that

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

Our LLSE has two properties:

1. Our estimator is unbiased, i.e. $\mathbb{E}[\hat{X}] = \mathbb{E}[X]$.
2. Our error Δ is uncorrelated with the observation, i.e. $\text{Cov}(\Delta, Y) = 0$.

Makes sense, right?

Last lecture we also went over interpretations of LLSE in terms of Hilbert space geometry (for zero-mean RVs); this allows us to do geometric things with random variables. Too bad it was erased before I could get it down rip. There's notes on the website however.

Using this interpretation, the LLSE is just the projection of X onto Y, so

$$L[X|Y] = \text{proj}_Y X = \frac{\langle X, Y\rangle}{||Y||^2}Y = \frac{\mathbb{E}[XY]}{\text{Var}(Y)}Y.$$

Example 23.1. Let $Y = X + Z$, where X and Z are independent and zero-mean. What is $L[X|Y]$?

Solution. Knowing that they're zero-mean allows us to use the geometric interpretation of LLSE, which is just the projection.

Let this projection be bY (since it lies on the vector Y). Then by similar triangles (we have right angles since X, Y have covariance 0),

$$\frac{||b\vec{Y}||}{||X||} = \frac{||X||}{||Y||} \implies ||bY|| = \frac{||X||^2}{||Y||} =$$

□

We can also ask what the error would be. By using similar triangles again,

$$\frac{||\Delta||^2}{||X||^2} = \frac{||Z||^2}{||Y||^2}.$$

23.2 Linear Regression

Let's take a look at Linear Regression from a Non-Bayesian view. So far, we have assumed a Bayesian framework (i.e. joint distribution of X, Y are known), but instead let's take a "data-driven" perspective.

Suppose we have access to $(X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k)$. Our goal is to construct $g(Y) = a + bY$ so that

$$\frac{1}{k}\sum_{i=1}^{k} |X_i - (a + bY_i)|^2$$

is minimized. This makes sense since our observations are Y, so given them we want to predict the original values, namely X.

It's also important to note that this is just a special case of the Bayesian approach, where each of our k data points is equally likely, and hence has a prior of probability $\frac{1}{k}$. Hence $(X, Y) \sim \text{Uniform}\{(x_i, y_i)_{i=1}^k\}$. So minimizing our error is equivalent to solving

$$\xi(a, b) = \frac{1}{k}\sum_{i=1}^{k} |X_i - (a + bY_i)|^2 \implies \frac{\partial \xi}{\partial a} = 0, \frac{\partial \xi}{\partial b} = 0.$$

Our estimator then is just

$$a + bY = \mathbb{E}_k[X] + \frac{\text{Cov}_k(X, Y)}{\text{Var}_k(Y)}(Y - \mathbb{E}_k(Y)).$$

But $\mathbb{E}_k[X] = \frac{1}{k}\sum_{i=1}^k X_i$ and $\mathbb{E}_k[Y] = \frac{1}{k}\sum_{i=1}^k Y_i$. Also

$$\text{Var}_k(Y) = \frac{1}{k}\sum_{i=1}^{k} Y_i^2 - \mathbb{E}_k^2[Y]$$

and

$$\text{Cov}_k(X, Y) = \frac{1}{k}\sum_{i=1}^{k} X_i Y_i - \mathbb{E}_k[X] \cdot \mathbb{E}_k[Y].$$

By SLLN, as $k \to \infty$, Linear Regression approaches the LLSE estimator.

This Linear Regression is also the same as the one we did in 16A and all your fancy ML classes. The linear algebra perspective is that we want to solve for x in $Ax = b$, where in this case,

$$\begin{bmatrix} 1 & Y_1 \\ 1 & Y_2 \\ \vdots & \vdots \\ 1 & Y_k \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$$

Least squares says that the solution is $x = (A^T A)^{-1} A^T b$. You can actually convince yourself that $\hat{X}_{LS} = \begin{bmatrix} a \\ b \end{bmatrix}$, where a, b are the same covariance/expectation values as from the Bayesian framework. So the approaches are equivalent, except our LLSE Bayesian approach gives us more generality.

23.3 MMSE

We can move up to higher degree estimators, such as $\hat{X} = a + bY + cY^2$, and you can imagine that doing so will give you better results. But you run the risk of overfitting to your data.

Instead of constraining ourselves as before to linear estimate models, we now simply want the best mean squared estimator. Assume you know the joint pdf of X, Y. Our goal is to find $g(Y)$ such that $\mathbb{E}[(X - g(Y))^2]$ is minimum. This is what we call the **Minimum Mean Square Error** estimate, or MMSE estimate. Note that this is over ALL functions $g(Y)$, so this is a pretty hard problem.

Suppose we knew nothing about the data. Then the best value we could pick is just the mean of the data (you can convince yourself why this minimizes the mean squared error). Now suppose we conditioned on $Y = y$; by the same logic, we should just pick the mean of the data where $Y = y$, which is precisely $\mathbb{E}[X|Y = y]$. It turns out that this is actually the optimal MMSE, which we will prove.

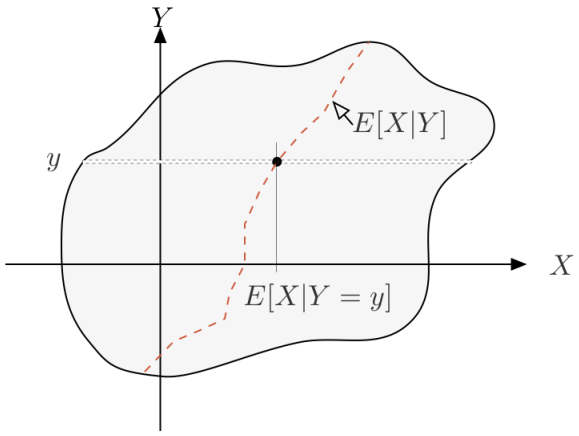

Figure 15: For a given y, $\mathbb{E}[X|Y = y]$ is just the mean of the mass where $Y = y$. Source: Walrand.

Theorem 32 (Walrand 7.4)**.** The minimum mean squares error estimate is

$$\text{MMSE}[X|Y] = \mathbb{E}[X|Y],$$

where

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

Before we start, we will need some lemmas that we state without proof.

Lemma 33 (Walrand 7.6)**.** For any function $\phi(\cdot)$,

(a) $\mathbb{E}[(X - \mathbb{E}[X|Y]) \cdot \phi(Y)] = 0$; in other words, $\Delta = X - \mathbb{E}[X|Y]$ is independent of $\phi(Y)$ for all $\phi(\cdot)$.

(b) If $\exists g(Y)$ such that $\mathbb{E}[(X - g(Y)) \cdot \phi(Y)] = 0$ for all $\phi(\cdot)$, then $g(Y) = \mathbb{E}[X|Y]$.

Proof of Theorem 7.4. Let $G(Y) = \{g(Y)|g(\cdot) \text{ is a function}\}$ be the space of all functions on Y.

As with the LLSE, $X - \mathbb{E}[X|Y]$ is independent to the projection of X on to a general member of the functional space, say $h(Y)$. Now if we could show that

$$\mathbb{E}[|X - h(Y)|^2] \geq \mathbb{E}[|X - \mathbb{E}[X|Y]|^2|],$$

then we're done, since this shows that no matter what functional you pick, $\mathbb{E}[X|Y]$ is always king.

So let's take the left hand side and expand it using the world's most *annoying* trick, adding and subtracting some quantity.

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

Here's the slick part though; since $\mathbb{E}[X|Y]$ and $h(Y)$ both belong to $G(Y)$, their difference is still a function of Y, so $\phi(Y) = \mathbb{E}[X|Y] - h(Y) \in G(Y)$. So by part (a) of our lemma, the product is just 0.

Also, our second term is always nonnegative since it's the expected value of something squared, so

$$\mathbb{E}[|X - h(Y)|^2] = \mathbb{E}[|X - \mathbb{E}[X|Y]|^2] + \mathbb{E}[|\mathbb{E}[X|Y] - h(Y)|^2] \geq \mathbb{E}[|X - \mathbb{E}[X|Y]|^2]$$

and so we're done. □

23.4 Jointly Gaussian

In general, our MMSE won't be the same as our LLSE, since the projection of X onto the general space $G(Y)$ won't be the same as our projection onto $a + bY$, which is a subspace of that general space. In other words, $L[X|Y] \neq \mathbb{E}[X|Y]$.

So when are they equal then? One case is obvious; when $g(Y)$ is linear. But if you think about it more generally, this will always hold when X and Y are *jointly Gaussian*, or JG.

This is really important, since the main problem with the MMSE is that finding it numerically is extremeley tough. But if X, Y are jointly Gaussian, we can use our formula for the LLSE from before. This is big enough to be stated as a theorem.

Theorem 34 (Walrand 7.8)**.** If (X, Y) are jointly Gaussian, then

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

What does being Jointly Gaussian mean? It means that $X = (X_1, X_2)$ has a multivariate normal pdf.

24 Thursday, April 25th

24.1 Jointly Gaussian Random Variables

24.2 Kalman Filtering

25 Tuesday, April 30th

25.1 Kalman Filter

26 Thursday, May 2nd

Last lecture woo

HW 12 and Lab 7 are due tomorrow, and an optional lab is due next Friday. Stay tuned for review sections.

26.1 Hidden Markov Models

Suppose we had a Markov Model or Chain, but some of the states are hidden to us. In other words, there are states we can't observe, and those we can. Then we call this a **Hidden Markov Model** since some states are "hidden"[4] from us, and depict it as such. The X_i are state variables and belong to a state space $\mathcal{X}$ (discrete or continuous), while the $y_i \in \mathcal{Y}$, which also may or may not be discrete.

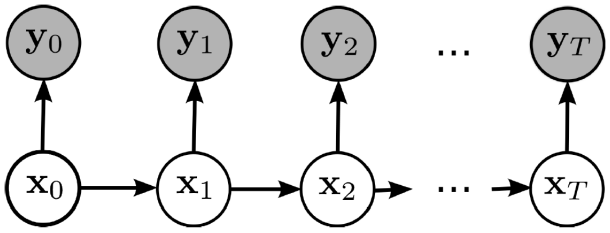

Figure 16: A Hidden Markov Model. The grey states are observed, while the white ones are hidden.

Example 26.1. If we had $T = 2$, then the joint probability $\mathbb{P}(x_0, y_0, x_1, y_1)$ would be

$$\mathbb{P}(x_0, y_0, x_1, y_1) = \mathbb{P}(x_0)\mathbb{P}(y_0|x_0)\mathbb{P}(x_1|x_0, y_0)\mathbb{P}(y_1|x_1, x_0, y_0) = \mathbb{P}(x_0)\mathbb{P}(y_0|x_0)\mathbb{P}(x_1|x_0)\mathbb{P}(y_1|x_1).$$

Notice that due to the structure of the HMM, we can simplify our probabilities a lot. This will be very useful! (This is akin to how the power of a Markov Chain is that the joint distribution is modeled by just a transition matrix P).

In general,

$$\mathbb{P}(x_0, x_1, \ldots, x_n, y_0, y_1, \ldots, y_n) = \pi_0(x_0)Q(x_0, y_0)P(x_0, x_1)Q(x_1, y_1)\ldots P(x_{n-1}, x_n)Q(x_n, y_n)$$

where π_0 is our initial state, Q models our transition probabilities between hidden states and observations, and P models transitions between hidden states.

Our goal is to find the MLSE, or maximum likelihood sequence estimate, based on our observations $y_0, \ldots, y_n$. In other words, we want

$$MAP[X^n|Y^n = y^n] \implies \begin{cases} \text{infer the best sequence of (hidden) states} \\ \text{that best explain the observed sequence.} \end{cases}$$

What's one application of Hidden Markov Models and finding the MLSE? Speech Recognition. You receive sounds as observations, and your goal is to find the most likely sequence of words corresponding to the sounds.

HMMs are super flexible, so there's a couple of ways we can use them depending on what you are trying to find.

[4] Sometimes we refer to them as *latent* as well.

- **Filtering**: We feed in $Y_0, Y_1, \ldots, Y_T$ to our filter and expect out $\hat{X}_T$, the last hidden state. Think Kalman filtering.

$$Y_0, Y_1, \ldots, Y_T \to \boxed{\text{Filter}} \to \hat{X}_T$$

  Some examples are tracking positions in real time or monitoring current health of patient given symptoms $\{Y_0\}_{i=0}^T$.

- **Prediction**: We feed in $Y_0, Y_1, \ldots, Y_T$ and want to predict $\hat{Y}_{T+1}$.

$$Y_0, Y_1, \ldots, Y_T \to \boxed{\text{Predict}} \to \hat{Y}_{T+1}$$

  Some examples are radar tracking, stock price predictions, or predictive coding (?).

- **Smoothing**: We feed in $Y_0, Y_1, \ldots, Y_T$ and want to find $\hat{X}_t$ for a choice of $t \leq T$. In other words, if I gave you a value t, what is the most likely value of $\hat{X}_t$ given our observations?

$$Y_0, Y_1, \ldots, Y_T \to \boxed{\text{Smooth}} \to \hat{X}_t \quad t \leq T$$

  Some examples are inferring the cause of a car-crash or "post-mortem" analysis.

- **MLSE**: We feed in $Y_0, Y_1, \ldots, Y_T$ and want the most likely *sequence* $\hat{X}_0, \hat{X}_1, \ldots, \hat{X}_T$ that explains our observations. This differs from smoothing, where we only care about maximizing over a single hidden state.

$$Y_0, Y_1, \ldots, Y_T \to \boxed{\text{MLSE}} \to \{\hat{X}_0, \hat{X}_1, \ldots, \hat{X}_T\}$$

  Some examples are speech recognition, auto-correction, and convolutional coding (Viterbi algorithm).

We're focusing on just the last one today (MLSE), both to preserve one's sanity but also because time isn't a construct and very realistically limits us.

26.2 The Viterbi Algorithm

Let's now try to write out what the MLSE would be algebraically. We're looking for

$$\begin{aligned} x^{n^*} &= \arg\max_{x^n \in \mathcal{X}^n} \mathbb{P}[X^n = x^n | Y^n = y^n] \\ &= \arg\max_{x^n} \left[\pi_0 Q(x_0, y_0) P(x_0, x_1) Q(x_1, y_1) \ldots P(x_{n-1}, x_n) Q(x_n, y_n)\right]. \end{aligned}$$

But I don't like taking arg maxes of products, so I'll make an easy fix by taking logs on both sides.

$$x^{n^*} = \arg_{x^n \in \mathcal{X}^n} \max \left[\log \pi_0(x_0) Q(x_0, y_0) + \sum_{m=1}^{n} \log\left[P(x_{m-1}, x_m) Q(x_m, y_m)\right]\right].$$

So we've reduced the MLSE problem to optimizing over the right hand side. To make it more compact, let's define

$$\begin{aligned} d_0(x_0) &= -\log \pi_0(x_0) Q(x_0, y_0) \\ d_m(x_{m-1}, x_m) &= -\log\left[P(x_{m-1}, x_m) Q(x_m, y_m)\right] \end{aligned}$$

so that all the $d_i's$ are positive (why?).
Then we have the following.

Definition 35 (MLSE)**.** Our maximum likelihood sequence estimate reduces to the optimization problem:

$$x^{n^*} = \arg\min_{x^n} \left[d_0(x_0) + \sum_{m=1}^{n} d_m(x_{m-1}, x_m) \right]$$

where x^{n^*} is the optimal sequence explaining our observations.

Since things are getting very abstract, we're going to turn to an example to ground our intuition. Ah, good 'ol engineering maths.

Say you're at a "nearly honest casino" which uses a fair die most of the time, but switches to a loaded die occasionally. We can model their transitions (P) between using the fair die and the loaded die as the following Markov Chain (so assume we know these parameters).

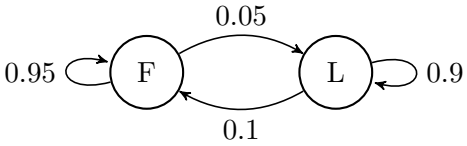

Figure 17: Transitions for the casino.

Additionally, we know that for the fair die, the probability of each outcome is equally likely (as expected), so $\mathbb{P}(F = i) = \frac{1}{6}$ for $i = 1, \ldots, 6$. For the loaded die, we know that $\mathbb{P}(L = 6) = \frac{1}{2}$ and $\mathbb{P}(L = 1) = \cdots = \mathbb{P}(L = 5) = \frac{1}{10}$ so that the die is biased towards rolling a 6.

Then given an observed sequence of die rolls (say $6, 6, 1, 6, 2, \ldots$), we want to infer the most likely sequence of "hidden" states (say $F, F, L, F, L, \ldots$).

We can use a technique called a "TRELLIS" diagram, which looks like this:

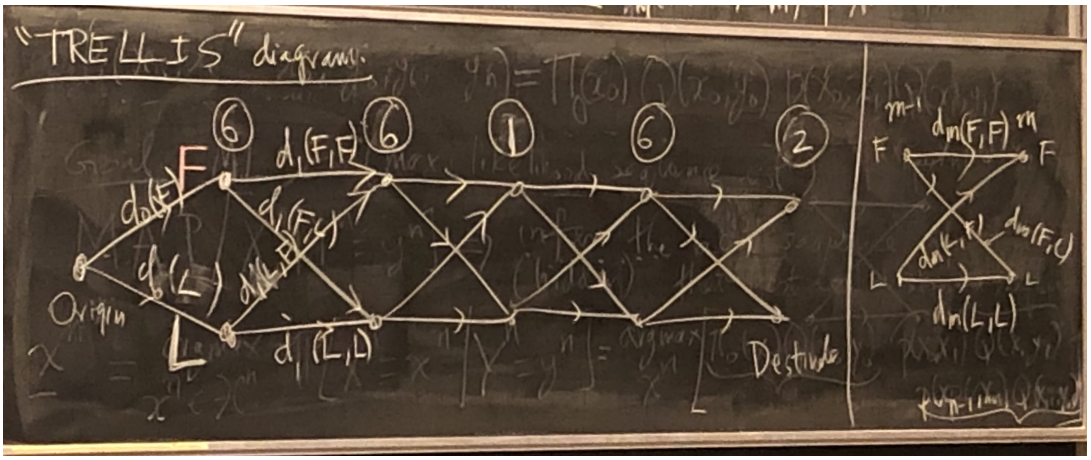

Figure 18: A TRELLIS diagram. Looking quite DAG-like.

To find the minimum length path from stage 0 to stage n, we just need a good shortest path problem. Bellman-Ford seems like a good choice, especially since its dynamic programming nature lends itself very well to such a calculation (recall that shortest paths on DAGs are best solved by DP). This technique of filling out the diagram using a dynamic programming method is something that Viterbi discovered first, so we call this the **Viterbi Algorithm**.

First thing's first. We have to calculate our edge weights, a.k.a. all the d_m's. The computation would look something like this:

$$d_m(F,F) = -\log\left[P(F,F)Q(F,Y_m)\right]$$
$$d_m(F,L) = -\log\left[P(F,L)Q(L,Y_m)\right]$$
$$d_m(L,F) = -\log\left[P(L,F)Q(F,Y_m)\right]$$
$$d_m(L,L) = -\log\left[P(L,L)Q(L,Y_m)\right]$$

If you plug in those numbers for every possible m based on our observations, your TRELLIS diagram will look something like this:

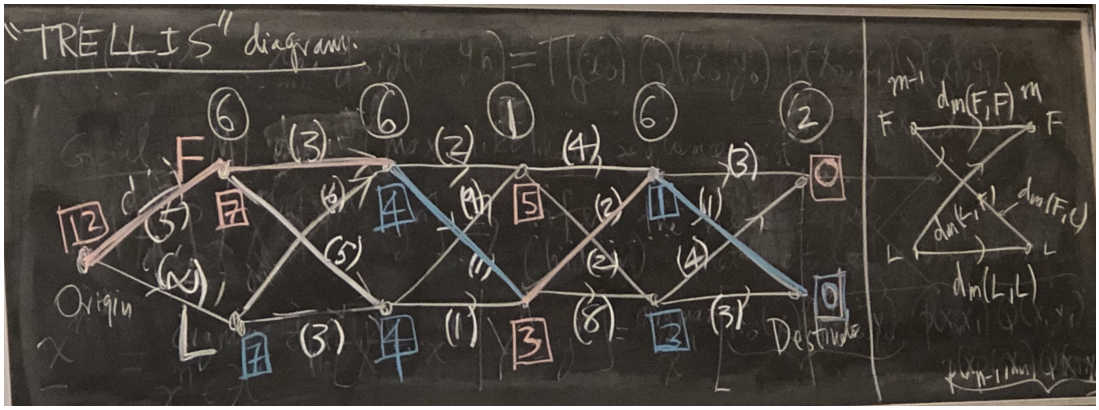

Figure 19: The TRELLIS diagram filled in with edge weights and the shortest path.

The circled numbers are the observations, and the numbers along every edge are the weights. The boxed numbers are the shortest path values from that node to the final stage; blue numbers represent transitions to a loaded die, while red numbers represent transitions to a fair die. Our initial transition from the origin to L is infinity because we're told that the casino will start with a fair die.

If we work our way back through this diagram, we can find that the MLSE estimate is (F, F, L, F, L).

Finally, we can do a quick analysis of how much time each of these methods take. The cost of populating trellis is $O(N^2 n)$ where N is the number of states and n is the number of stages. If we have a populated trellis, it only takes $O(Nn)$ to find the shortest path (since we only have one node at each stage to consider). Note that the naive completion is just $O(N^n)$, so we're doing really well. We turned a computational infeasible problem into a pretty efficient one.