EECS 126 Notes

Tyler Zhu

March 1, 2019

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, whose .sty file never fails to impress, which is why I've stolen bits of it for these notes.

Contents

1 Thursday, February 7th

1.1 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

1.2 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.

1.3 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

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

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

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

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

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

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

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

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

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

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

1.4 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.

1.5 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 1.1

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 1.2

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

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

1.6 Multiple Random Variables

Literally page 13 of the book.

1.7 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 1.3

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',

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

□

2 Tuesday, February 12th

2.1 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.

2.2 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.

2.3 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!

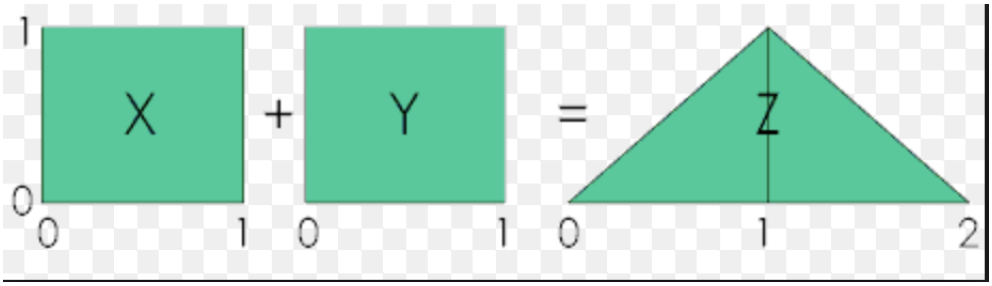

Figure 1: The convolution of two uniform distributions $U[0, 1]$.

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

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

2.4 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 2.1 (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 2.2 (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 2.3 (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 $M_X''(0) = 1$.

2.5 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 2.4

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

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

3 Thursday, February 14th

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

3.1 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

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

3.2 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

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

from which our result follows □

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

[2] Thanks to Efe for the Piazza explanation.

Example 3.1

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.

3.3 Tail Bounds

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 3.2

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.

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 3.3
Let's apply the Chernoff bound to normal random variables. We find that

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

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

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

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.

4 Thursday, February 21st

4.1 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.

4.2 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 $\epsilon > 0$, $\mathbb{P}(|M_n - \mu| \geq \epsilon) \to \infty$ 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 4.1

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 4.2

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 4.3

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

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

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

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 $X = \mu$.

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.

4.3 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. 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),

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

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

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

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. □

5 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.

5.1 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 5.1

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 $Z_n = \dfrac{A_n - \frac{1}{\lambda}}{\frac{1}{\sqrt{n}}\lambda}$, 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.

5.2 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."

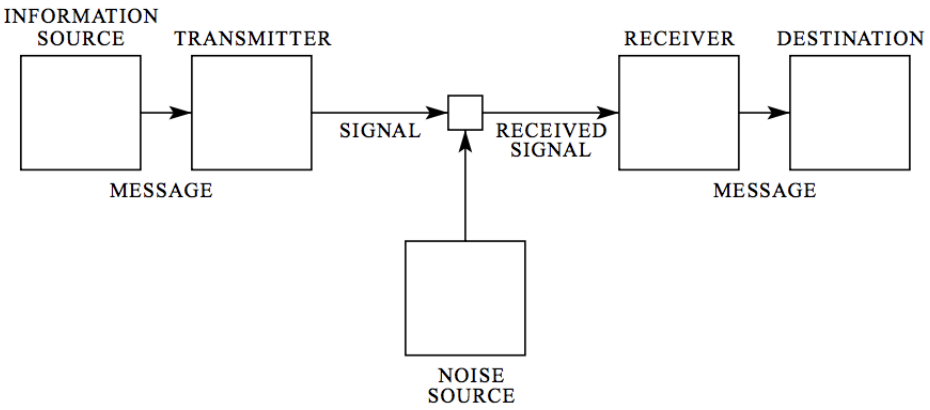

Fig. 1—Schematic diagram of a general communication system.

Figure 2: 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 5.2 (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.

5.3 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. Sorry I got really lazy and didn't bother taking notes about those. I'll add them back in once I get the time to. :)

6 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.

6.1 Capacity of the BEC

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 $1/(1 - p)$. 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.

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 $\hat{m} = j$.

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

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 6.1

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.

6.2 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, $\mathbb{P}(X_n|X_{n-1})$.

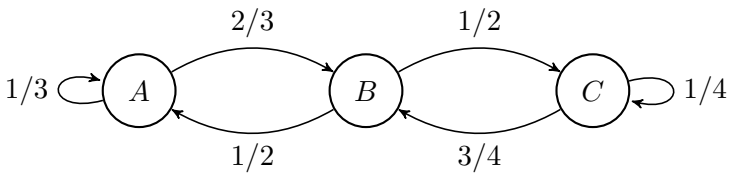

Figure 4: An example of a Markov Chain