Section Notes for POLSCI 598: Mathematics for Political Scientists

Patrick Y. Wu

Fall 2017

Note: I have written, but not necessarily read, these notes. Please let me know if you see any typos.

1 Tuesday, September 19

1.1 Welcome / Introduction

- Patrick Wu. E-mail: pywu@umich.edu. Office: Haven Hall 6564.
- This course will primarily rely on Iain's in-class notes. The optional textbooks are on the syllabus. You can use them for alternative explanations (sometimes it takes another explanation to fully understand something!).
- I will post section notes right after section. This should not be a reason why you do not attend section.
- If you need to miss section for a legitimate reason, e-mail me. It is important to note that Iain's policies described in the syllabus applies to both section and lectures.

1.2 Discrete versus Continuous Random Variables

- The material in this section is taken from `http://mathinsight.org/probability_density_function_idea`.
- From above, it seems like we can more generally classify sample spaces into two types according to the number of elements they contain. Sample spaces can be either *countable* or *uncountable.*
- If the elements of a sample space can be put into a 1-1 correspondence with a subset of integers, the sample space is countable; otherwise, it is uncountable.
- People usually get the difference in the countable versus uncountable case. But people get very confused when it comes to the difference between discrete random variables and continuous random variables.
- Remember, the probability mass function (PMF) of a **discrete** variable X gives the probability that X is equal to a particular value x. This is notated $f_X(x) = P(X = x)$. Easy enough to understand. Similarly, the cumulative density function (CDF) of a **discrete** variable X gives the probability that X is less than or equal to a particular value x. It is notated $F_X(x) = P(X \leq x)$.

- But now let's say that random variable X is a continuous random variable. If we try to solve $P(X = x)$, what does this equal? $P(X = x) = 0$. Thus, we can see that in the case of continuous random variables, $f_X(x) \neq P(X = x)$. This is why there is no such thing as a probability mass function when it comes to continuous random variables; rather, $f_X(x)$ is known as the probability density function (PDF).

- Let's say that I am thinking of a number (let's call it X) between 0 and 10 (inclusive). If I do not tell you what rule I am using to pick the number, you would probably just assume that the probability that $X = i$ is 1/11 for any integer i from 0 to 10. We formally write this as

$$Pr(X = i) = \frac{1}{11} \text{ for i = 0, 1, 2, ..., 10}$$

- Now, implicit in this assumption is that the probability that X is any other number is 0. That is, you're assuming that I won't pick 5.5. This implicit assumption is formalized as

$$Pr(X = x) = 0 \text{ if x is not one of \{0,1,2,...,10\}}$$

- Now, what if I am thinking of a number X between 0 and 1 (inclusive)? You may be assuming again that I just mean 0 and 1 so the probability of either option is 1/2. Or, you might guess that I had more than two options in mind. Maybe I was thinking of 1/2, 3/4, 7/8, etc. The possibilities are endless. I could even be thinking of some wacky irrational number, like $1/\sqrt{\pi}$. Thus, there are an infinite number of possibilities.

- So the probability is the same for every possible number between 0 and 1, inclusive. Let's call this constant probability c. We need all the probabilities across all possible numbers to add up to 1. But there are also an infinite number of possibilities. So what could the individual probability c be? If c were any finite number greater than 0, then adding across an infinite number of possibilities means that the probability also goes off to infinity. This means that c must be infinitesimally small, i.e., that $c = 0$. Thus, we formally define, when X is any possible number between 0 and 1 inclusive, the following probability rule:

$$Pr(X = x) = 0 \text{ for any real number x}$$

- But this also still doesn't make sense! How can every probability for each possibility be 0 but the total probability is 1?

- This is where the idea of *probability density* comes in. Rather than thinking of the probability that X is equal to an exact specific number, we should think about the probability that X is close to a single number.

- This is formally calculated using the probability density function. If the probability density around a point x is large, this means that the random variable X is likely to be close to x. If, on the other hand, $f_X(x) = 0$ in some interval, then X won't be in that interval.

- Thus, in order to determine the probability that X is in any subset A of the real numbers, we simply add up the values of $f_X(x)$ in the subset. By "add up" we are talking about integration. The probability that X is in A is precisely:

$$Pr(X \in A) = \int_A f_X(x)dx$$

- So if I is the interval $I = [a, b]$ with $a \leq b$, then the probability that $a \leq X \leq b$ is

$$Pr(x \in I) = \int_a^b f_X(x)dx$$

1.3 Bivariate Distribution Examples

1.3.1 Example 1: A Bivariate Density Function

This example is taken from John A. Rice's *Mathematical Statistics and Data Analysis* (2005).

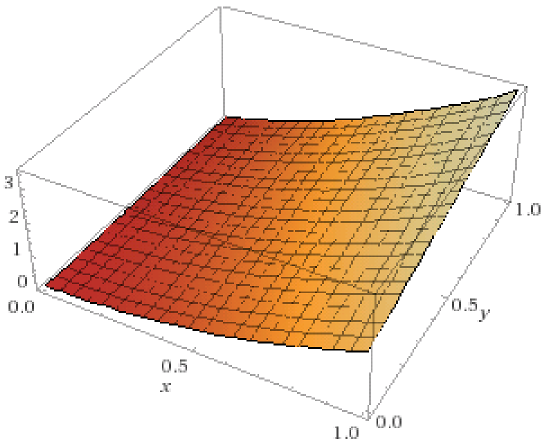
Let us consider the bivariate density function

$$f(x, y) = \frac{12}{7}(x^2 + xy)$$

with the conditions

$$0 \leq x \leq 1, 0 \leq y \leq 1$$

Graphically, it looks like this:



What is $P(X > Y)$? This can be found by integrating f over the set $\{(x, y)|0 \leq y \leq x \leq 1\}$:

$$\begin{aligned}
P(X > Y) &= \frac{12}{7}\int_0^1 \int_0^x (x^2 + xy)dydx \\
&= \frac{12}{7}\int_0^1 yx^2 + \frac{1}{2}y^2x\Big|_{y=0}^{x} dx \\
&= \frac{12}{7}\int_0^1 \frac{3}{2}x^3 dx \\
&= \frac{12}{7}\frac{3}{8} \\
&= \frac{9}{14}
\end{aligned}$$

What is the marginal density of X? Remember, in the bivariate case, $f_X(x) = F'_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$.

$$\begin{aligned} f_X(x) &= \frac{12}{7}\int_0^1 (x^2 + xy)dy \\ &= \frac{12}{7}\left[yx^2 + \frac{1}{2}xy^2\right]\Bigg|_{y=0}^{1} \\ &= \frac{12}{7}(x^2 + \frac{x}{2}) \end{aligned}$$

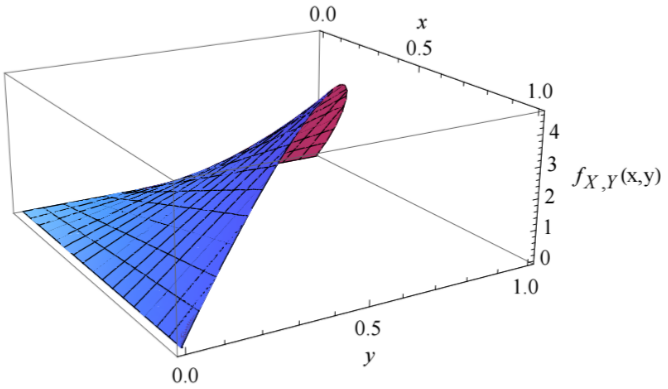
What is the marginal density of Y? Similarly,

$$\begin{aligned} f_Y(y) &= \frac{12}{7}\int_0^1 (x^2 + xy)dx \\ &= \frac{12}{7}\left[\frac{1}{3}x^3 + \frac{1}{2}x^2y\right]\Bigg|_{x=0}^{1} \\ &= \frac{12}{7}(\frac{1}{3} + \frac{y}{2}) \end{aligned}$$

1.3.2 Example 2: Another Bivariate Density Function

You are given the following joint probability density function for X and Y: $f_{X,Y}(x,y) = 12xy$. Both these random variables can take on any value between 0 and 1, but it must be the case that $x^2 + y \leq 1$. In other words, $x \leq \sqrt{1-y}$ (or that $y \leq 1 - x^2$).

The joint probability density function looks like this:



Let's do the following:

1. show that this is a valid probability density function
2. find the marginal density of X and the marginal density of Y and then find $E[X]$
3. determine whether X is independent of Y
4. find the conditional distribution $f_{X|Y}(x|y)$ and find $E[X|Y]$
5. confirm that the law of iterated expectation, $E[E[X|Y]] = E[X]$
6. find the covariance of X and Y

1. First, we confirm that this is a valid probability density function. We know that, from the above, that $0 \leq x \leq \sqrt{1-y} \leq 1$. We can also rephrase this as $0 \leq y \leq 1 - x^2 \leq 1$.

$$\begin{aligned}
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy &= \int_0^1 \int_0^{\sqrt{1-y}} 12xydxdy \\
&= \int_0^1 6x^2y \Big|_{x=0}^{\sqrt{1-y}} dy \\
&= 6\left(\frac{y^2}{2} - \frac{y^3}{3}\right)\Big|_{y=0}^{1} \\
&= 1
\end{aligned}$$

2. Now let's find the marginal density of X and the marginal density of Y. Remember, we know from the above that $0 \leq x \leq \sqrt{1-y} \leq 1$. We can also rephrase this as $0 \leq y \leq 1 - x^2 \leq 1$. We can then use the marginal distribution of X to determine E[X]. For the latter, it will be helpful for us to know that $\frac{1}{3} - \frac{2}{5} + \frac{1}{7} = \frac{8}{105}$.

$$\begin{aligned}
f_X(x) &= \int_0^{1-x^2} 12xydy \\
&= 6xy^2 \Big|_{x=0}^{1-x^2} \\
&= 6x(1 - 2x^2 + x^4) \\
&= 6(x - 2x^3 + x^5)
\end{aligned}$$

$$\begin{aligned}
f_Y(y) &= \int_0^{\sqrt{1-y}} 12xydx \\
&= 6x^2y \Big|_{x=0}^{\sqrt{1-y}} \\
&= 6(y - y^2)
\end{aligned}$$

So this means that we can use the above to find $E[X]$.

$$E[X] = \int_0^1 x6(x - 2x^3 + x^5)dx = 6(\frac{x^3}{3} - \frac{2x^5}{5} + \frac{x^7}{7} = 6(\frac{1}{3} - \frac{2}{5} + \frac{1}{7}) = \frac{48}{105} = \frac{16}{35}$$

3. Can we determine if X is independent of Y? Yes, we can. We find that X is not independent of Y. Why?

$$f_X(x)f_Y(y) = 6(x - 2x^3 + x^5)6(y - y^3) \neq 12xy = f_{X,Y}(x,y)$$

4. Now let's find the conditional distribution $f_{X|Y}(x|y)$. Note that conditional on some particular $Y = y$, x can only fall in the range $(0, \sqrt{1-y})$.

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{12xy}{6(y - y^2)} = \frac{2x}{1-y}$$

Thus, we can use this to solve $E[X|Y]$:

$$E[X|Y] = \int_0^{\sqrt{1-y}} x\frac{2x}{1-y}dx = \frac{2}{3}\frac{x^3}{1-y}\bigg|_{x=0}^{\sqrt{1-y}} = \frac{2}{3}\frac{(1-y)^{3/2}}{1-y} = \frac{2}{3}\sqrt{1-y}$$

5. Now we find $E[E[X|Y]]$ and we can show that it equals $E[X]$. Note that we have already found $E[X|Y]$ and the "outer expectation" in $E[E[X|Y]]$ is taken over potential values of Y. We already know the marginal density of Y. You will need to use integration by parts to solve this integral! Remember, integration by parts is as follows: $\int fg'dx = fg - \int f'gdx$.

$$\begin{aligned} E[E[X|Y]] &= \int_{-\infty}^{\infty} E[X|Y]f_Y(y)dy \\ &= \int_0^1 \frac{2}{3}\sqrt{1-y}6(y-y^2)dy \\ &= \int_0^1 4y(1-y)^{\frac{3}{2}}dy \end{aligned}$$

Now let $f = 4y$ and $g' = (1-y)^{\frac{3}{2}}$. This means that $g = -\frac{2}{5}(1-y)^{\frac{5}{2}}$. Thus,

$$\begin{aligned} &= -\frac{8}{5}y(1-y)^{\frac{5}{2}}\bigg|_0^1 + \int_0^1 \frac{8}{5}(1-y)^{\frac{5}{2}}dy \\ &= 0 + \left[-\frac{16}{35}(1-y)^{\frac{7}{2}}\right]\bigg|_0^1 \\ &= \frac{16}{35} \end{aligned}$$

6. A little fairy tells me that $E[Y] = \frac{1}{2}$. Let's find the covariance of X and Y. We know that $Cov(X,Y) = E[XY] - E[X]E[Y]$. So that means, finding $E[XY]$,

$$\begin{aligned} E[XY] &= \int_0^1 \int_0^{1-x^2} xy12xydydx \\ &= \int_0^1 4x^2y^3\bigg|_{y=0}^{1-x^2} dx \\ &= \int_0^1 4x^2(1-3x^2+3x^4-x^6)dx \\ &= 4(\frac{1}{3} - \frac{3}{5} + \frac{3}{7} - \frac{1}{9}) \end{aligned}$$

And by the magical properties of the math fairy...

$$= \frac{64}{315}$$

So that means the covariance is:

$$Cov(X,Y) = E[XY] - E[X]E[Y] = \frac{64}{315} - \frac{16}{35}\frac{1}{2} = -\frac{8}{315}$$

2 Tuesday, September 26

2.1 Some Consistent Mistakes in the Problem Set

- When you write out the `R` code, please also include the output. Having the code itself is not very useful.
- Be sure to note the difference between proving general statements (such as showing that the binomial distribution sums to unity with any parameters) and showing that specific statements, where the parameters are given, produce an expected result.
- When a problem has assumptions (such as $\lambda > t$), be sure to explicitly state the assumptions and also tell us why the assumption has to be made. This includes showing why when the assumption is not made, everything goes to hell. For instance, on the fourth problem of this problem set, we had to assume that $\lambda > t$. If we let $\lambda = t$, then the denominator would equal 0; if $\lambda < t$, the integral would be undefined. You should explicitly state why these assumptions have to be made.

2.2 Optional Problem on Problem Set 1

This is one of my favorite optional problems that Iain gives. It puts their understanding of distributions, conditional expectations, etc. to the test. Premise of the problem. Two actors 1 and 2 are bidding on an object which they both value at θ. Each gets a signal s_i of the value of the object before they bid such that $s_i = \theta + \eta_i$ where $\eta_i \sim Unif(-a, a)$. When we refer generically to one actor as i the other will be referred to as j. After receiving their signal, each actor privately makes a bid for the good, b_i. The highest bidder earns utility $u_i = \theta - b_i$ and the lowest bidder earns utility $u_j = 0$. Suppose that both actors decide to bid their signal, i.e. $b_i = s_i$. We will prove that their expected loss is $\frac{a}{6}$.

i.

Explain why the winner of the auction earns utility $u_i = -\eta_i$ with this bidding strategy. Therefore, $u_i = \theta - b_i = \theta - s_i = \theta - (\theta + \eta_i) = -\eta_i$.

ii.

Find the distribution function of η_j (i.e. find F_{η_j}) and then use this to find $Pr(\eta_j \geq \eta_i)$ in terms of η_i. The PDF of the uniform distribution (use Wikipedia!) is $\frac{1}{b-a}$ for a support of $[a, b]$. Since our support is $[-a, a]$, this means that our PDF is $\frac{1}{a-(-a)} = \frac{1}{2a}$. Thus, using this it is easy to find our PDF:

$$F_{\eta_j}(x) = \int_{-a}^{x} f_{\eta_j}(x)dx = \int_{-a}^{x} \frac{1}{2a}dx = \frac{x}{2a}\bigg|_{-a}^{x} = \frac{x}{2a} + \frac{1}{2}$$

So we can use this expression and find $Pr(\eta_j \geq \eta_i)$. Notice that we are looking at the distribution of η_j, NOT η_i. So we can treat η_i as a constant in this case. So this means that:

$$Pr(\eta_j \geq \eta_i) = 1 - F_{\eta_j}(\eta_i) = \frac{1}{2} - \frac{\eta_i}{2a}$$

iii.

Now let's find the expected earnings of i conditional on η_i when i uses the bidding strategy described above. That is, find $E[u_i|\eta_i]$. To get started on this, we will look at the utility payoffs when i wins the auction and when i loses the auction. In other words, we are interested in learning about the expected utility of i when he bids η_i. The utility u_i for the outcome that i wins the auction is $Pr(\eta_j < \eta_i)$ (that is, j bids lower than i). This happens with probability $1 - Pr(\eta_j \geq \eta_i) = 1 - \frac{1}{2} + \frac{\eta_i}{2a} = \frac{1}{2} + \frac{\eta_i}{2a}$ with utility $-\eta_i$. Now, the probability that i loses the auction is what we found in the previous part. This is just $P(\eta_j \geq \eta_i) = \frac{1}{2} - \frac{\eta_i}{2a}$ with utility 0 (since i does not get anything). Thus,

$$E[u_i|\eta_i] = -\eta_i(1 - Pr(\eta_j > \eta_i) + 0(Pr(\eta_j > \eta_i)) = -\eta_i\left(\frac{1}{2} + \frac{\eta_i}{2a}\right)$$

iv.

Now we can use the law of total expectations to find $E[E[u_i|\eta_i]] = E[u_i]$. We can think of a LOTUS integral where $E[u_i|n_i]$ is a function of the random variable η_i.

$$\begin{aligned}
E[u_i] &= E[E[u_i|\eta_i]] \\
&= -\int_{-a}^{a} \eta_i \left(\frac{1}{2} + \frac{\eta_i}{2a}\right) f_{\eta_i}(\eta_i) d\eta_i \\
&= -\int_{-a}^{a} \eta_i \left(\frac{1}{2} + \frac{\eta_i}{2a}\right) \frac{1}{2a} d\eta_i \\
&= -\left(\frac{\eta_i^2}{8a^2} + \frac{\eta_i^3}{12a^2}\right)\Bigg|_{-a}^{a} \\
&= -\frac{a}{6}
\end{aligned}$$

2.3 Directional Derivatives

Let's do an example of a directional derivative and expand directional derivatives a little bit.

2.3.1 Example of a Directional Derivative

Find $D_{\vec{v}} f(x, y, z)$, where $f(x, y, z) = x^2 z + y^3 z^2 - xyz$ and $\vec{v} =< -1, 0, 3 >$. Note that here we are concerned about the displacement vector $\vec{v}$; $\vec{v}$ is NOT a *point* (which is why $\vec{v}$ is described in the straight brackets). That is, we want to move in the direction given by this specific vector $\vec{v}$.

So we first solve for Df:

$$Df = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \end{bmatrix}$$

$$= \begin{bmatrix} 2xz - yz & 3y^2 z^2 - xz & x^2 + 2y^3 z - xy \end{bmatrix}$$

Now we solve for $\frac{v}{||v||} = \left\langle -\frac{1}{\sqrt{10}}, 0, \frac{3}{\sqrt{10}} \right\rangle$. Now solve for $Df \cdot \frac{v}{||v||}$. Remember it is a dot product!

$$Df \cdot \frac{v}{||v||} = -\frac{1}{\sqrt{10}}(2xz - yz) + 0(3y^2 z^2 - xz) + \frac{3}{\sqrt{10}}(x^2 + 2y^3 z - xy) = \frac{1}{\sqrt{10}}(3x^2 + 6y^3 z - 3xy - 2xz + yz)$$

2.3.2 Expanding Directional Derivatives

So we know that $D_{\vec{v}} f(x, y, z) = f_x(x, y, z)a + f_y(x, y, z)b + f_z x, y, zc =< f_x, f_y, f_z > \cdot < a, b, c >$ where f_x is the first partial derivative of f with respect to x (and analogously defined to the other letters) and $< a, b, c >$ is the direction of the derivative (assumed to be a unit vector—we can always standardized via norm if it's not!). Let's define the **gradient of** f as follows:

$$\nabla f =< f_x, f_y, f_z >$$

This simplifies our notation to the follow:

$$D_{\vec{v}}f = \nabla f \cdot \vec{v}$$

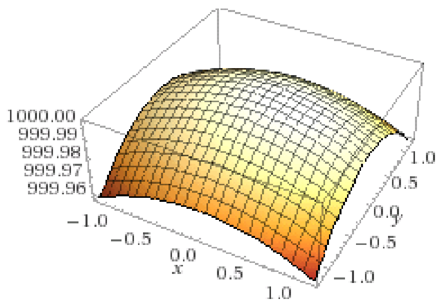
One very cool theorem that comes out of this is the following.

Theorem: The maximum value of $D_{\vec{v}}f$ (and hence then the maximum rate of change of the function $f(\vec{x})$) is given by $||\nabla f(\vec{x})||$ and will occur in the direction given by $\nabla f(\vec{x})$.

So let's look at a final, quick example.

2.3.3 Maximum Directional Derivative Example

Let's suppose that we model average income, in hundreds of dollars, in a neighborhood as follows: $z = 1000 - 0.01x^2 - 0.02y^2$, where x is the number of non-violent crimes that occur and y is the number of violent crimes that occur in a year within this neighborhood. Suppose that we currently have 60 incidents of non-violent crimes this year and 100 incidents of violent crimes this year. We can ask boring questions like how much non-violent crime should we decrease by or how much violent crime should we decrease by, but I want to know in what direction for both non-violent and violent crimes. In other words, in what direction should we move along our model to greatest increase the average income?



So first, we find the gradient vector.

$$\nabla f(\vec{x}) =< f_x, f_y >=< -0.02x, -0.04y >$$

The maximum rate of change of average income levels will then occur in the direction of

$$\nabla f(60, 100) =< -1.2, -4 >$$

Thus, the maximum rate of change of average income levels at this point will be

$$||\nabla f(60, 100) = \sqrt{(-1.2)^2 + (4)^2} = \sqrt{17.44} = 4.176$$

And so we know we are climbing this hill, so to speak, because we are at point (60,100) and the direction is given by the vector $< -1.2, -4 >$. So we know we are increasing income (because we are climbing the hill). Thus, if we move in the direction of $< -1.2, -4 >$, our instantaneous rate of change in average income is 4.176, which is the maximum increase in average income that we can get given that we started at $(60, 100)$.

3 Tuesday, October 4

3.1 Example of Taylor Series Application: Brownian Motion

Why am I teaching you Brownian motion? Brownian motion is used by certain political scientists at the very forefront of what is known as social physics. They're using Brownian motion to model decisionmaking in the face of uncertainty, extended to a wide range of actors. You can go to Wikipedia to find an example of Brownian motion.

Let's just assume that we are given this:

$$f(x,t+\tau)=\int_{-\infty}^{\infty} f(x-\epsilon,t)\phi(\epsilon)d\epsilon$$

That is, we are interested in finding the probability that a particle is at position x after τ time. Our initial starting position is $x-\epsilon$. Let ϕ be some arbitrary probability density function symmetrical around 0 with mean 0. Let this be the PDF of random variable Z. How the hell do we solve this? We don't even know the functional form. Thankfully, we have Taylor approximations to save us. We know that, if we let $x_1 = t+\tau$ and $x_0 = t$,

$$f(x,t+\tau)\approx f(x,t)+(t+\tau-t)\frac{\partial f(x,t)}{\partial t}+...$$

And if we let $x_1 = x-\epsilon$ and $x_0 = x$

$$f(x-\epsilon,t)\approx f(x,t)+(x-\epsilon-x)\frac{\partial f(x,t)}{\partial x}+\frac{1}{2!}(x-\epsilon-x)^2\frac{\partial^2 f(x,t)}{\partial x^2}+...$$

So plugging them into our given equation, we get:

$$f(x,t+\tau)\approx f(x,t)+\tau\frac{\partial f(x,t)}{\partial t}=\int_{-\infty}^{\infty}\left[f(x,t)-\epsilon\frac{\partial f(x,t)}{\partial x}+\frac{\epsilon^2}{2}\frac{\partial^2 f(x,t)}{\partial x^2}+...\right]\phi(\epsilon)d\epsilon$$

Now distributing out the $\phi(\epsilon)$,

$$f(x,t)+\tau\frac{\partial f(x,t)}{\partial t}=\int_{-\infty}^{\infty} f(x,t)\phi(\epsilon)-\epsilon\frac{\partial f(x,t)}{\partial x}\phi(\epsilon)+\frac{\epsilon^2}{2}\frac{\partial^2 f(x,t)}{\partial x^2}\phi(\epsilon)+...d\epsilon$$

Now distributing out the integral,

$$f(x,t)+\tau\frac{\partial f(x,t)}{\partial t}=f(x,t)\int_{-\infty}^{\infty}\phi(\epsilon)d\epsilon-\frac{\partial f(x,t)}{\partial x}\int_{-\infty}^{\infty}\epsilon\phi(\epsilon)d\epsilon+\frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}\int_{-\infty}^{\infty}\epsilon^2\phi(\epsilon)d\epsilon$$

Now, notice that the first term's integral just integrates to 1. It is the integral of a PDF over its entire support. Notice that in the second integral, we have the definition of the expected value of Z. In our assumptions, we assumed that ϕ had a mean of 0 so that means this integral is just 0. Now, the very last term's integral is just the variance of Z—notice that this integral is just $E[(\epsilon-0)^2]$. And now look on the left hand side. We can now cancel out the $f(x,t)$. So we're left with:

$$\tau\frac{\partial f(x,t)}{\partial t}=\frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}Var(Z)$$

Which can be further simplified to

$$\frac{\partial f(x,t)}{\partial t}=\frac{1}{2\tau}\frac{\partial^2 f(x,t)}{\partial x^2}Var(Z)$$

Lastly, we can plug this back into

$$f(x,t+\tau) \approx f(x,t) + (t+\tau-t)\frac{\partial f(x,t)}{\partial t}$$

To get

$$f(x,t+\tau) \approx f(x,t) + \tau(\frac{1}{2\tau}\frac{\partial^2 f(x,t)}{\partial x^2}Var(Z)) = f(x,t) + \frac{1}{2}\frac{\partial^2 f(x,t)}{\partial x^2}Var(Z)$$

So the likelihood that the particle is at position x at time $t+\tau$ compared to other possible points is given by the likelihood that the point is at point x at our initial time t plus some behavioral changes of the particle given a change in the position x, multiplied by the variance of Z (which is where we draw our initial starting point from). So without even knowing the functional form of f, using some Taylor approximations we were able to do some really cool stuff.

3.2 Linear Algebra, Part I

I love linear algebra. For my own research, linear algebra is everything. In fact, we can think of ordinary least squares as an exercise in linear algebra. You can derive most properties of OLS without any statistical knowledge. I'm also a visual person. So I'll be teaching an alternative, more visually-focused version of linear algebra that'll complement what Iain is doing nicely. I'll also try to relate linear algebra to statistical concepts.

3.2.1 Why Vectors and Not Points?

	I	II
a	1	3
b	3	2
c	-2	-1
d	2	-2

Let's say we have a dataset. We will show these as vectors rather than points. Why? We can move them in ways that follow mathematical operations. We can shift them around, in essence, because vectors have an end point (the terminal point) that is tied to a relative position (the origin). Vectors allow us to shift the data points around in a space; point geometry does not.

3.2.2 Vector Addition

Shifting one vector along another until at the other's terminal point. If $a = \begin{bmatrix} a_1 & a_2 \end{bmatrix}$ and $b = \begin{bmatrix} b_1 & b_2 \end{bmatrix}$ then $a + b = \begin{bmatrix} a_1 + b_1 & a_2 + b_2 \end{bmatrix}$.

3.2.3 Vector Subtraction

Shifting a reflected vector, since $a - b = a + (-1b)$. Using the previous example, $a - b = \begin{bmatrix} a_1 - b_1 & a_2 - b_2 \end{bmatrix}$.

3.2.4 Scalar Multiplication

Changes vector by a factor of k. So the new vector is k times longer than the original vector. A negative constant points it the other direction (opposite quadrant). Scalar multiplication means, theoretically, we could reach any collinear line created by the vector. We just need to pick a proper value of k. This is called *generating a subspace* of that vector. Note that a subspace is NOT the subset.

3.2.5 Vector Subspace, Linear Independence, and Linear Dependence

Remember that we can generate a subspace of a vector because we can reach any point along the collinear line of a vector. When we add two vectors together, we get a two-dimensional subspace. This is a linear combination of two vectors. Linear independence is when the only solution for $\alpha_1 a + \alpha_2 b = 0$, where α_1 and α_2 are constants, are for both the constants to be equal to 0. If this is not true, then the vectors are linearly dependent. Notice that if we add a third vector in the two-dimensional space, then this third vector will be linear dependent—we can use some scalar of the two vectors and add them together to get the third vector. We'll explore this idea of linear independence and dependence more later on.

3.2.6 Scalar Product and Vector Length

Also known as a dot product. $a \cdot b$ is the sum of products of corresponding elements. If a has m-elements and b has m-elements, then $a \cdot b = \sum_{j=1}^{m} a_j b_j$. You can take dot products of a vector with itself: $X \cdot X = X^2 = \sum_{j=1}^{m} x_j x_j = \sum_{j=1}^{m} x_j^2$. Notice that this is just its length squared.

We calculate vector length using the norm vector, $||X||$. Now let's say we take a vector's mean, $\bar{X}$, and subtract it from each of the vector's elements. So we get $\tilde{X} = \begin{bmatrix} x_1 - \bar{x} & x_2 - \bar{x} & ... & x_n - \bar{x} \end{bmatrix}$. Now take the dot product of $\tilde{X}$ with itself: $\tilde{X} \cdot \tilde{X} = [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2] = \sum_{i=1}^{n} (x_i - \bar{x})^2 = ||\tilde{X}||^2$. If we multiply this by $\frac{1}{n-1}$, this is the variance of X! Thus, the length of X is proportional to the variance of X.

The scalar product of two vectors is $X \cdot Y = ||X||||Y||cos(\theta_{XY})$ where θ_{XY} is the angle between two vectors. Now let's say that X and Y are two data vectors (say, two variables) that have been centralized. That is, every data point of X has had its mean $\bar{X}$ subtracted from it and every day point of Y has had its mean $\bar{Y}$ subtracted from it. From the last part, we know that

$$||X||||Y|| = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

and

$$X \cdot Y = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

So using our equation $X \cdot Y = ||X||||Y||cos(\theta_{XY})$ we notice that

$$cos(\theta_{XY}) = \frac{X \cdot Y}{||X||||Y||} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

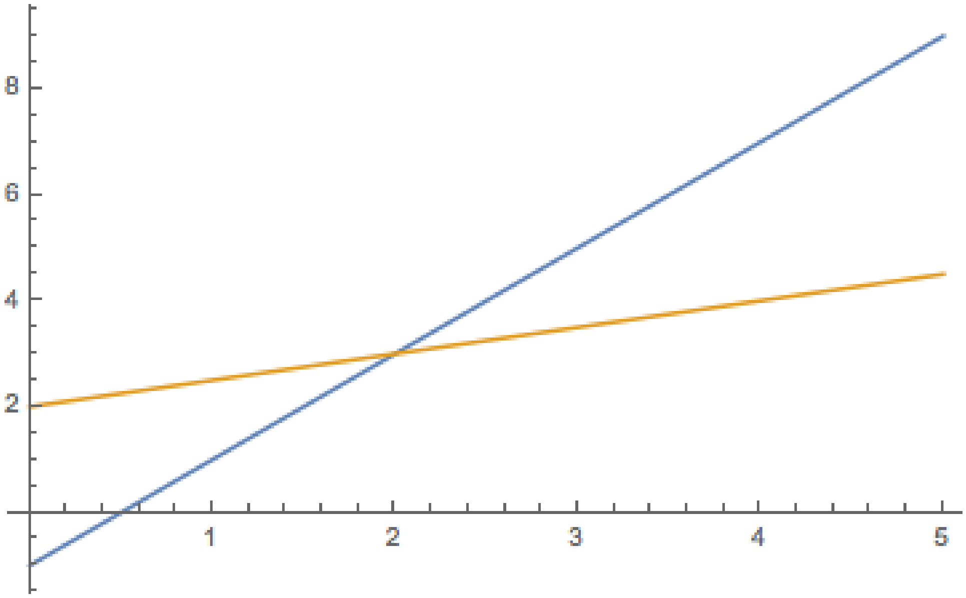
Notice that the numerator is the covariance of X and Y, and the denominator is the standard deviations of X and Y (the n's are not shown here, but they all cancel out). So that means if X and Y are two centralized data vectors, the cosine of the angle between the two vectors is the correlation! This makes sense—$cos(0) = 1$, $cos(90) = 0$, and $cos(180) = -1$. This is why orthogonal vectors are 90 degrees apart.

Geometrically, the dot product is fundamentally a *projection*. We'll talk about projections later, but geometrically, it just means taking the terminal point and drawing down at a 90 degree angle down to the space you want to project in. As shown in the figure, we can see how a dot product works geometrically. This is why we take the dot product in directional derivatives: we project the derivative onto the displacement vector, which is pointing in the direction we are interested in!

3.2.7 A Geometrical View of Gauss-Jordan Elimination

Gauss-Jordan elimination can be summarized as swap, scale, and pivot. But what exactly, intuitively, is happening when you do the row operations? Swap and scale do not change the graphs. Swapping does not change the graph because it is just reordering the equations. Scaling does not change the graph because if you have $2x - y = 2$, this is the same as $4x - 2y = 4$. But pivoting does change the graph. Let's take a visual example of what exactly Gauss-Jordan elimination is doing.

Let's say we have equations $2x - y = 1$ and $x - 2y = -4$. In matrix and graphical form, it looks like the following.

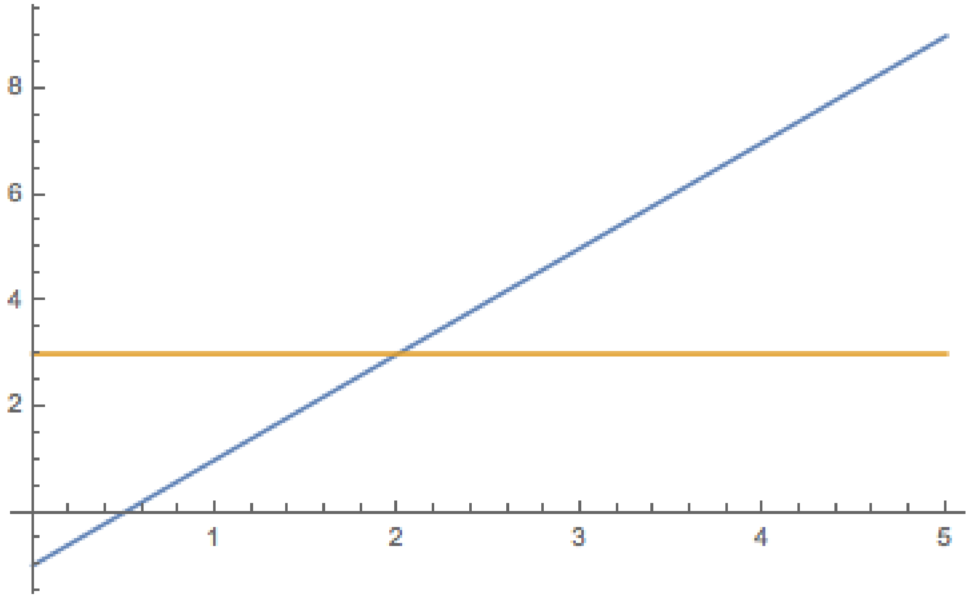


$$\left[\begin{array}{cc|c} 2 & -1 & 1 \\ 1 & -2 & -4 \end{array}\right]$$

Now let's begin applying Gauss-Jordan elimination steps to solve for x and y. Note that I won't draw out visually what happens when I swap or scale; those do not change the shape of the graph.

$$\left[\begin{array}{cc|c} 2 & -1 & 1 \\ 1 & -2 & -4 \end{array}\right] \xrightarrow{0.5r_1} \left[\begin{array}{cc|c} 1 & -0.5 & 0.5 \\ 1 & -2 & -4 \end{array}\right] \xrightarrow{r_2 - r_1} \left[\begin{array}{cc|c} 1 & -0.5 & 0.5 \\ 0 & -1.5 & -4.5 \end{array}\right]$$

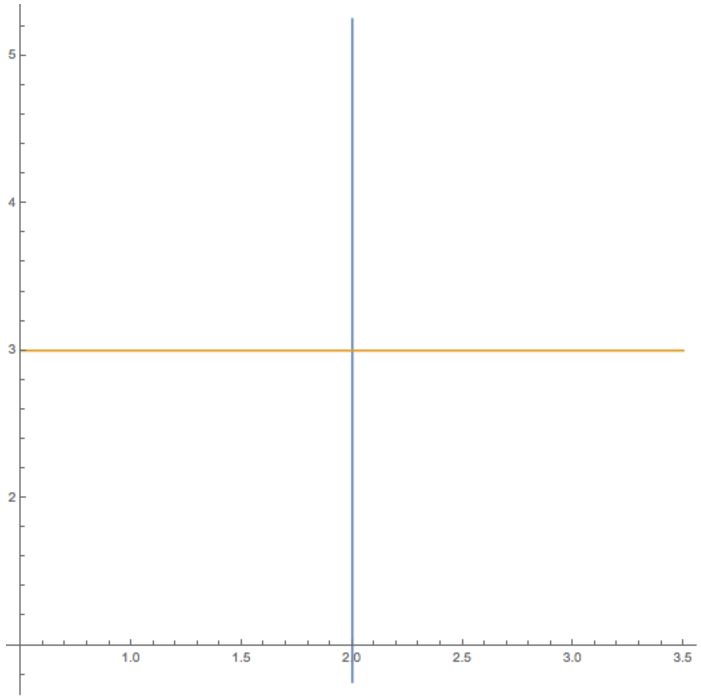
What does this do? It pivots the graph of $x - 2y = -4$ such that it is parallel to the x-axis.



Now, we continue with the above:

$$\left[\begin{array}{cc|c} 1 & -0.5 & 0.5 \\ 0 & -1.5 & -3 \end{array}\right] \xrightarrow{r_2/(-1.5)} \left[\begin{array}{cc|c} 1 & -0.5 & 0.5 \\ 0 & 1 & 3 \end{array}\right] \xrightarrow{0.5r_2+r_1} \left[\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & 3 \end{array}\right]$$

Notice that we now have the augmented matrix in reduced row-echelon form. The graph looks as follows:



So, geometrically, what Gauss-Jordan elimination is doing is a series of swaps, scales, and pivots. Although the pivots do change how the graph looks, the solution always remains the same. And notice that the reduced-row echelon form has 2 lines that are parallel to the two-axis. This generalizes to any dimension. If we were to apply Gauss-Jordan elimination to 3 equations of 3 unknowns, then the reduced row-echelon form would be 3 planes parallel to the x, y, and z axes.

4 Tuesday, October 10

4.1 Deriving OLS of One and Two Explanatory Variables Using the Concepts We've Learned So Far

We can derive the properties of OLS using all the concepts we've learned so far, along with the idea of the projection. Let's look at the case with a single variable. Let X be the data of the explanatory variable (as a vector) and Y be the dependent variable (again, as a vector). Now, assume like last week that X and Y have been centered—that is, all entries of X and Y have the means of their vector subtracted from each entry. Now, let's draw what this would look like. Now, let's draw the perpendicular projection from Y onto X (dotted line pointing up). Let's call this point $\hat{y}$. Notice that $e = Y - \hat{Y}$. Notice also that this is the shortest we can make the e vector. Moving it around—that is, making the projection non-perpendicular—makes the e vector longer. Thus, the projection minimizes the length of the vector e. So $\hat{y}$ lies in the subspace of X. We can get $\hat{y}$ by multiplying X by a constant. Call this constant b. So, $\hat{y} = bX$. Since e is perpendicular to X, this means that $e^T X = 0$. You may have heard of the key assumption of linear regression being that the errors and the X's are independent—that is, the explanatory variables do not explain the errors. This is where that assumption comes from. Thus, from this, we can find:

$$\begin{aligned}
e^T X = (Y - \hat{Y})^T X &= 0 \\
\Rightarrow (Y - bX)^T X &= 0 \\
\Rightarrow (Y^T - bX^T)X &= 0 \\
\Rightarrow Y^T X - bX^T X &= 0 \\
\Rightarrow Y^T X &= bX^T X \\
\Rightarrow b &= \frac{Y^T X}{X^T X} \\
\Rightarrow b &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$

Notice that this is the OLS regression coefficient for simple linear regression (one explanatory variable). We just derived the coefficient that minimizes the square distances between the data and the regression line without any calculus!

This logic extends to regressions with multiple variables. With two explanatory variables, first notice that the two explanatory variable creates a column space. This column space is created from the fact that we can reach all the points inbetween X_1 and X_2 using vector addition and scalar multiplication. This is why, in linear regression, one of the assumption is the absence of extreme collinearity—that is, we need the column space of X_1 and X_2 to be "sufficiently" big. That doesn't mean they can't be correlated! Next, notice that we simply choose a linear combination of the two explanatory variables, using constants b_1 and b_2, to get the $\hat{Y}$ line that is created by the projection of Y into the column space of X. So the same logic as the simple linear regression case applies. We won't get the nice values of the coefficients that we got with the simple regression case, but that's something we'll cover in 699.

4.2 The Geometry of Matrix Determinants

One more thing I wanted to go over is the determinant. Lots of people don't really understand determinants intuitively. They just sort of do them without really thinking about it. I wanted to give some motivation to what is so cool about determinants.

First, some properties of determinants:

- $det(A)$ gives the area or volume magnification factor for the linear transformation $x \to Ax$
- A is invertible iff $det(A) \neq 0$
- $det(AB) = det(A)det(B)$
- the most efficient way of calculating $det(A)$ is by row (or column) reducing A to triangular form

Intuitively, the determinant of a matrix reflects how the linear transformation associated with the matrix can scale or reflect objects. Let's see how this works in practice. Let's say that we consider the unit square $OIJK$ such that the coordinates representing the square are $O = (0,0); I = (1,0); J = (1,1); K = (0,1)$. Now suppose we were to multiply these coordinates by the matrix

$$\mathbf{T} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}$$

This is a transformation matrix that will stretch and rotate this unit box in a specific way. So let's see how it stretches it:

$$\mathbf{U} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 3 & 1 \\ 0 & 3 & 7 & 4 \end{bmatrix}$$

So the key aspect of this transformation has to do with the ratio of the area of the quadrilateral to the area of the original unit square. *The ratio of the two areas equals the determinant of the transformation matrix* $\mathbf{T}$. That is, $|\mathbf{T}| = 8 - 3 = 5$. This concept generalizes to matrices of order 3x3 and higher. In the 3x3 case, the determinant measures the ratio of volume between the original and transformed figures. In the 4x4 case and higher-order cases, the determinant measures the ratio of hypervolumes between original and transformed figures. Notice also why singular matrices have 0 determinant. Let's say we have the transformation matrix

$$\mathbf{T} = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}$$

Now, let's do the same steps as before to the unit square:

$$\mathbf{U} = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 3 & 1 \\ 0 & 2 & 3 & 1 \end{bmatrix}$$

Notice that the parallelogram collapses into a line, so it has 0 area. The ratio between the line (area of 0) and the unit square (area of 1) is 0.

In general, the determinant of n vectors of length n will give the volume (or parallelepiped) determined by those vectors in n-th dimensional Euclidean space.

4.3 The Geometry of Cramer's Rule

This section is largely taken from Wikipedia's page on Cramer's Rule.

Cramer's Rule is really weird, but it is useful because sometimes you can solve a system of equations very quickly. It states that given a system of n linear equations and n unknowns, $Ax = b$, where the $n \times n$ matrix A has a nonzero determinant and the vector $x = (x_1, ..., x_n)^T$ is the column vector of variables. Then the system of equations has a unique solution, whose individual values for the unknowns are given by

$$x_i = \frac{det(A_i)}{det(A)}$$

where A_i is the matrix formed by replacing the i-th column of A by the column vector b.

Why is this the case? Let's look at the geometry of Cramer's Rule. Let's say we have a system of equations:

$$a_{11}x_1 + a_{12}x_2 = b_1$$

$$a_{21}x_1 + a_{22}x_2 = b_2$$

We can rewrite this as

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Now, the area of the parallelogram determined by $\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}$ and $\begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}$ is

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

Now, the area of $x_1 \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}$ and $\begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}$ has to be x_1 times the area of the first parallelogram above. Now, this parallelogram, by Cavalieri's Principle, has the same area as the parallelogram determined by $\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = x_1 \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}$ and $\begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}$. But this means that

$$\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix} = \begin{vmatrix} a_{11}x_1 & a_{12} \\ a_{21}x_1 & a_{22} \end{vmatrix} = x_1 \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

5 Tuesday, October 24

5.1 Eigenvalues, Eigenvectors, and Definiteness

Definition: Let A be an $n \times n$ matrix. The number λ is an eigenvalue of A if there exists a non-zero vector v such that $Av = \lambda v$. In this case, v is called an eigenvector. We notice that $Av - \lambda v = (A - I\lambda)v = 0$. In order to prevent getting the trivial solution $v = 0$ (that is, if $A - I\lambda$ is non-singular, the only solution is $v = 0$), we solve $det(A - I\lambda) = 0$ in order to get the eigenvalues and their corresponding eigenvectors.

Eigenvalues and eigenvectors are used everywhere. They were used in some of the original face recognition software (Eigenfaces): classification was achieved by comparing the low-dimensional face pictures to a basis set. It is also used in principal components analysis, which Iain briefly discussed in class. Music visualizers also use eigenvalues and eigenvectors from the music in order to show the principal components of the music (it is a principal components analysis on the music).

Eigenvalues are also super useful for just about everything. Properties of eigenvalues include

- det(A) $= \prod_{i=1}^{n} \lambda_i$
- tr(A) $= \sum_{i=1}^{n} \lambda_i$
- The eigenvalues of A^{-1} are λ_i^{-1} (this is really useful for finding the determinant of an inverse matrix!)
- The eigenvalues of A^n are λ_i^n
- The eigenvectors of A^{-1} are the same as the eigenvectors of A
- If eigenvalues of A are distinct (not repeated), A can be eigendecomposed

Example: Let

$$A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$$

Then that means that

$$(A - \lambda I)v = \left(\begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right) v$$

So solving the determinant,

$$det[A - \lambda I] = 0 \Rightarrow det \begin{bmatrix} 2 - \lambda & -4 \\ -1 & -1 - \lambda \end{bmatrix} = (2-\lambda)(-1-\lambda) - (-4)(-1) = \lambda^2 - \lambda - 6 = (\lambda - 3)(\lambda + 2) = 0$$

So this implies that

$$\lambda_1 = 3, \lambda_2 = -2$$

So notice that because the eigenvalues are both negative and positive, this means that A is indefinite. Let's now find the corresponding eigenvector to $\lambda_1 = 3$.

$$\begin{bmatrix} 2 - 3 & -4 \\ -1 & -1 - 3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So from this, we obtain the duplicate equations:

$$-v_1 - 4v_2 = 0$$

$$-v_1 - 4v_2 = 0$$

If we let $v_2 = t$, notice that $v_1 = -4t$. So that means all eigenvectors of $\lambda_1 = 3$ are multiples of $\begin{bmatrix} -4 \\ 1 \end{bmatrix}$. Thus, the eigenspace corresponding to $\lambda_1 = 3$ is given by the span of $\begin{bmatrix} -4 \\ 1 \end{bmatrix}$. That is,

$$\left\{ \begin{bmatrix} -4 \\ 1 \end{bmatrix} \right\}$$

is a basis of the eigenspace corresponding to $\lambda_1 = 3$.

Now we do the same thing for the second eigenvalue, $\lambda_2 = -2$. Repeating the process above, we find that

$$4v_1 - 4v_2 = 0$$

$$-v_1 + v_2 = 0$$

Notice that if $v_1 = t$ then $v_2 = t$ as well. Notice that we can't find explicit values of v_1 and v_2 because they are two equations that are simply linear combinations of one another. Thus, the eigenspace corresponding to $\lambda_2 = -2$ is the subspace created by $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Notice that the eigenspace is of dimension 1.

There's a few things to notice. First, notice that $Av = \lambda v$, which implies that what matrix A does the eigenvector is that it stretches the eigenvector, but it doesn't change its direction. Second, eigenvectors are linearly independent. This is a good way to quickly check to see if your eigenvectors make sense. This is an easy proof for the case of two eigenvectors.

Proof that Two Eigenvectors are Independent: Let $Av_1 = \lambda_1 v_1$. Let $Av_2 = \lambda_2 v_2$. Assume that $\lambda_1 \neq \lambda_2$. Now assume that the two eigenvectors are not zero vectors and that they are linear dependent. That is, $cv_1 = v_2$. Then $Av_2 = \lambda_2 v_2 = c\lambda_2 v_1$ and $Av_2 = Ac_1 v_1 = cAv_1 = c\lambda_1 v_1$. Hence this means that $c(\lambda_2 - \lambda_1) = 0$ which implies that $c = 0$ which implies that $v_2 = 0$ which is a contradiction. Thus, they must be linearly independent. ■

Example: Let A be defined as:

$$A = \begin{bmatrix} 3 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix}$$

Then, finding its eigenvalues, we have:

$$\begin{aligned} \begin{vmatrix} 3-\lambda & 0 & 1 \\ 1 & 1-\lambda & 0 \\ 1 & 0 & 3-\lambda \end{vmatrix} &= (3-\lambda)(-1)^{1+1} \begin{vmatrix} 1-\lambda & 0 \\ 0 & 3-\lambda \end{vmatrix} + 1(-1)^{1+3} \begin{vmatrix} 1 & 1-\lambda \\ 1 & 0 \end{vmatrix} \\ &= (3-\lambda)(1-\lambda)(3-\lambda) - (1-\lambda) \\ &= ((3-\lambda)^2 - 1)(1-\lambda) \\ &= (8 - 6\lambda + \lambda^2)(1-\lambda) \\ &= (8 - 6\lambda + \lambda^2)(1-\lambda) \\ &= (\lambda - 4)(\lambda - 2)(1-\lambda) = 0 \end{aligned}$$

This means that

$$\lambda_1 = 4, \lambda_2 = 2, \lambda_3 = 1$$

From these eigenvalues, we can tell that this is positive definite. Let's use this find the eigenvectors. Let's start with $\lambda_1 = 4$.

$$(A - 4I)v_1 = 0 \Rightarrow \begin{bmatrix} -1 & 0 & 1 & 0 \\ 1 & -3 & 0 & 0 \\ 1 & 0 & -1 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} -1 & 0 & 1 & 0 \\ 1 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

This means that

$$-x_1 + x_3 = 0$$

$$x_1 - 3x_2 = 0$$

So that means if $x_1 = t$, then $x_2 = t/3$ and $x_3 = t$. One possible eigenvector is to let $t = 3$, so that means $v_1 = \begin{bmatrix} 3 \\ 1 \\ 3 \end{bmatrix}$.

Now let's look at $\lambda_2 = 2$. This means that

$$(A - 2I)v_2 = 0 \Rightarrow \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

This means that

$$x_1 + x_3 = 0$$

$$x_1 - x_2 = 0$$

So that means $x_1 = x_2$ and $x_1 = -x_3$. If we let $x_1 = 1$, then a possible eigenvector is $v_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$.

Lastly, let's look at $\lambda_3 = 3$. This means that

$$(A - I)v_3 = 0 \Rightarrow \begin{bmatrix} 2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 \end{bmatrix}$$

So this means that $x_1 = 0$, $x_3 = 0$, and $x_1 + 2x_3 = 0$. But notice that means that x_2 can take on any value. So a possible eigenvector is $v_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$.

Lastly, we can test for definiteness with leading principal minors. The third order principal minor is:

$$\begin{vmatrix} 3 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 3 \end{vmatrix} = 8$$

$$\begin{vmatrix} 3 & 0 \\ 1 & 1 \end{vmatrix} = 3$$

$$\begin{vmatrix} 3 \end{vmatrix} = 3$$

Therefore, the leading principal minors are all strictly positive meaning that this is positive definite, as expected.

6 Tuesday, October 31 (spooky!)

6.1 Some Clarifications...

A common mistake on the midterm: arguing that independence implies orthogonality. This is not true! Example: $\begin{bmatrix}1 & 1 & 3\end{bmatrix}$ and $\begin{bmatrix}2 & 1 & 5\end{bmatrix}$ are linearly independent vectors. I can't express one in terms of the other by multiplying one of the vectors by some constant. So they are linearly independent. But, $\begin{bmatrix}1 & 1 & 3\end{bmatrix} \cdot \begin{bmatrix}2 & 1 & 5\end{bmatrix} = 2 + 1 + 15 = 18$ so they are not orthogonal.

6.2 Elements of Real Analysis I

Real analysis is really cool. In my opinion, it is what gives mathematics and statistics its beauty. Real analysis gives us the language to talk about the really weird things in math in really precise ways. For instance, this notion that some infinities are bigger than other infinities. In the next few sections, we'll be talking about real analysis.

6.2.1 Cardinality of Sets

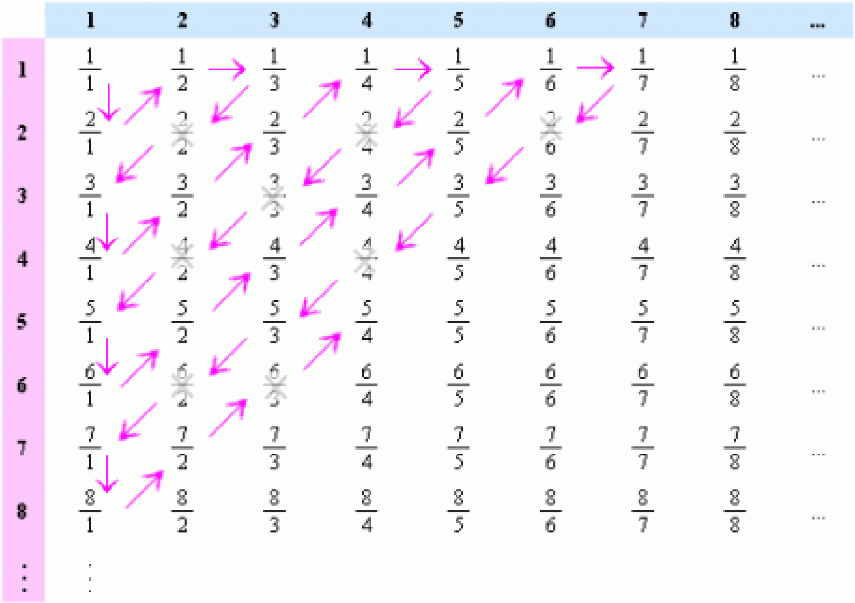
Cardinality is a measure of infinity. Georg Cantor was the guy who invented all of this. Cantor is arguably one of the most important modern mathematicians. He invented set theory, established the importance of one-to-one correspondence between the members of two sets, defined infinite and well-ordered sets, and proved that real numbers are more numerous than the natural numbers. In short, he found out that there is such thing as larger infinities. This has led to a lot of philosophical interests as well. For those who like political theory, Alain Badiou uses set theory in his philosophy. Real analysis has a lot of ties with analytic philosophy because it opens up a whole world of formal logic.

Countability is defined as the ability to map the natural numbers to the element of the set. For instance, the set $\{Apple, Oranges\}$ is countable because we can map "Apple" to 1 and "Oranges" to 2. This means that the cardinality of $\{1, 2, 3, ...\}$ is the same as the cardinality of $\{4, 5, 6\}$—that is, they're both countable. Uncountability means we cannot construct such a mapping.

Question: Are rational numbers countable? Are real numbers countable?

Answer: The rational numbers are countable. Can you prove that the rational numbers are countable? Let's construct it together:

We first write out all the fractions with 1 as the numerator. Then, we write out all the fractions with 2 as the numerator. Keep doing this, so we get:



	1	2	3	4	5	6	7	8	...
1	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	$\frac{1}{8}$	...
2	$\frac{2}{1}$	$\cancel{\frac{2}{2}}$	$\frac{2}{3}$	$\cancel{\frac{2}{4}}$	$\frac{2}{5}$	$\cancel{\frac{2}{6}}$	$\frac{2}{7}$	$\frac{2}{8}$	...
3	$\frac{3}{1}$	$\frac{3}{2}$	$\cancel{\frac{3}{3}}$	$\frac{3}{4}$	$\frac{3}{5}$	$\frac{3}{6}$	$\frac{3}{7}$	$\frac{3}{8}$	...
4	$\frac{4}{1}$	$\cancel{\frac{4}{2}}$	$\frac{4}{3}$	$\cancel{\frac{4}{4}}$	$\frac{4}{5}$	$\frac{4}{6}$	$\frac{4}{7}$	$\frac{4}{8}$	...
5	$\frac{5}{1}$	$\frac{5}{2}$	$\frac{5}{3}$	$\frac{5}{4}$	$\frac{5}{5}$	$\frac{5}{6}$	$\frac{5}{7}$	$\frac{5}{8}$	...
6	$\frac{6}{1}$	$\cancel{\frac{6}{2}}$	$\cancel{\frac{6}{3}}$	$\frac{6}{4}$	$\frac{6}{5}$	$\frac{6}{6}$	$\frac{6}{7}$	$\frac{6}{8}$	...
7	$\frac{7}{1}$	$\frac{7}{2}$	$\frac{7}{3}$	$\frac{7}{4}$	$\frac{7}{5}$	$\frac{7}{6}$	$\frac{7}{7}$	$\frac{7}{8}$	...
8	$\frac{8}{1}$	$\frac{8}{2}$	$\frac{8}{3}$	$\frac{8}{4}$	$\frac{8}{5}$	$\frac{8}{6}$	$\frac{8}{7}$	$\frac{8}{8}$	...
⋮	⋮								

Now snake along the path, and we can see that we will hit every number using this snaking pattern. Thus, the rationals are countably infinite. We can also do this with the negative rationals, and find that they're also countably infinite. The union of two countably infinite sets are countably infinite. But notice that if I give you any two rational numbers, I can always name a rational number in between! So this has a curious property. Notice that we have a *bijective* mapping between rational numbers and natural numbers: it is surjective because every element of the rational numbers map onto some natural number, and it is injective because every rational number only gets one natural number. So this is a great example of these concepts in action!

Now, are the real numbers countable? No, they are not. Real numbers have a different cardinality from rational numbers—that is, the infinity of real numbers is larger than the infinity of rational numbers. Let's prove this! Let's do this by contradiction. Suppose that the real interval $[0, 1]$ were a countable set. Therefore, we could write a small section of this hypothetical list as follows:

$$\begin{array}{ll} 1: & 0.\mathbf{a_1}\ a_2\ a_3\ a_4\ a_5\ldots \\ 2: & 0.b_1\ \mathbf{b_2}\ b_3\ b_4\ b_5\ldots \\ 3: & 0.c_1\ c_2\ \mathbf{c_3}\ c_4\ c_5\ldots \\ 4: & 0.d_1\ d_2\ d_3\ \mathbf{d_4}\ d_5\ldots \\ \vdots & \end{array}$$

Now here the numbers stretch off to infinity on the right as real numbers have infinite decimal expansions. The letters a_i, b_i, c_i, d_i represent numbers between 0 and 9, so $0.a_1a_2a_3...$ may be the number 0.12321.... It is unimportant what the specific numbers are, only that we can write them in such a list. Now, along the diagonal of this list the numbers that we have emboldened form a number: it begins $0.a_1b_2c_3d_4....$ With this we define a new number, called D, by adding 1 to each digit with the convention that if a digit is 9 then adding 1 goes to 0. In this way we have the first digit of D is $a_1 + 1$. So if the first number really did start 0.12321 then D would start 0.2, for example. Now this number D is surely a real number, and moreover it certainly lies somewhere between 0 and 1. So that means if the list really is complete, as we have claimed, then it should contain D. But this is where we hit a problem. The first digit of D is different from the first number (indeed, $a_1 + 1 \neq a_1$); the second digit of D is different from the second number; the third digit of D is different from the third number; and so on. You give me any number from this list, and D will have a digit that is off from that number that you give me by construction. In this way, we can see that D is different from every number on the list, so it can't be on the list at all. This contradiction shows that the real numbers are uncountably infinite. The infinity for the real numbers is, in a sense, larger than the infinity for rational numbers. We can also say that the real numbers are more dense than the rational numbers.

You might be asking—why is this useful?! It is important to know if you're dealing with countable or uncountable sets when you measure data. The above is also why discrete functions and continuous distributions are so different and have such different properties.

6.2.2 Limits

As Michael Spivak, the great author of the book *Calculus*, states, "The concept of the limit is surely the most important, and probably the most difficult one in all of calculus." When I learned theoretical mathematics in my undergraduate class, we spent one entire quarter on limits. Limits are what makes calculus possible. This is an incredibly shortened introduction to the logic of delta-epsilon proofs. Do let me know if you would like to discuss this more! These are used a lot in game theory.

An example: let's consider the function $f(x) = 3x$ with $a = 5$. Presumably, f should approach the limit 15 near 5. That is, we ought to be able to get $f(x)$ as close to 15 as we would like if we require that x be sufficiently close to 5. To be specific, suppose we want to make sure that $3x$ is within $\frac{1}{10}$ of 15. This means that we want to have:

$$15 - \frac{1}{10} < 3x < 15 + \frac{1}{10}$$

Which we can rewrite as

$$-\frac{1}{10} < 3x - 15 < \frac{1}{10}$$

To do this, we require that

$$-\frac{1}{30} < x - 5 < \frac{1}{30}$$

or simply

$$|x - 5| < \frac{1}{30}$$

There is nothing special about the number $\frac{1}{10}$. It is just easy to guarantee that $|3x - 15| < \frac{1}{100}$; simply require that $|x-5| < \frac{1}{300}$. In fact, if we take any positive number ϵ we can make $|3x-15| < \epsilon$ simply by requiring that $|x - 5| < \epsilon/3$. There is also nothing special about our choice $a = 5$. It is just as easy to see that f approaches the limit $3a$ at a for any a. That is, if we want to ensure that $|3x - 3a| < \epsilon$, then we have to require that $|x - a| < \epsilon/3$.

Now, for the famous definition of the limit:

Definition: The function f approaches the limit l near a means: for every $\epsilon > 0$ there is some $\delta > 0$ such that, for all x, if $0 < |x - a| < \delta$ then $|f(x) - l| < \epsilon$.

Example 1: $\lim_{x\to 5} 3x = 15$. Prove this using the definition above.

We know that in this case, $a = 5$ and $l = 15$. We want to find a suitable δ value such that the definition above holds true. Now, we know that $|3x - 15| < \epsilon$ for every epsilon. This implies that $|x - 5| < \epsilon/3$. Since $0 < |x - 5| < \delta$, we can simply let $\delta = \epsilon/3$. What does this mean? It means that for every value of ϵ you throw at me I can just divide it by 3 to ensure that if $|x - 5| < \epsilon/3$ then $|3x - 15| < \epsilon$.

Example 2: $\lim_{x\to 2} x^2 + 5x - 2 = 12$. Prove this using the definition above.

We know that in this case, $a = 2$ and $l = 12$. We want to find a suitable δ value such that the definition above holds true. Now, we know that $|x^2+5x-2-12| < \epsilon$ for every epsilon. This implies that $|x + 7||x - 2| < \epsilon$, so that means $|x - 2| < \epsilon/|x + 7|$. Since we are talking about values of x close to 2, this means that we can first restrict x such that it is at most 1 away from 2; meaning,

$$|x - 2| < 1 \Rightarrow 1 < x < 3 \Rightarrow 8 < x + 7 < 10$$

This means that the min value of $\epsilon/|x + 7|$ using our restrictions above is $\epsilon/10$. This means that

$$|x - 2| < \epsilon/|x + 7| < \epsilon/10$$

So we have two restrictions: $|x - 2| < 1$ and $|x - 2| < \epsilon/10$. So that means we can simply let $\delta = min\{1, \epsilon/10\}$. Technically, to complete this proof, we need to use our choice of δ to show that the definition holds.

7 Tuesday, November 7

7.1 Midterm Review

7.2 Visual Demonstration of Limits

https://www.youtube.com/watch?v=kfF40MiS7zA

7.3 2016 Presidential Election: A Discussion

The 2016 election has a lot of implications. The election was won by Trump, as the Washington Post analyzed, by 107,000 voters voting for him across MI, PA, and WI. See: https://www.washingtonpost.com/graphics/politics/2016-election/swing-state-margins/?tid=sm_fb This is the same number of people that can fit inside the Big House.

However, one of the key implications it has is for the field of political methodology. What did political methodology miss in this case? What does it say about the predictive power of big data? What does it say about doing research using Facebook or Twitter, which are slowly becoming common methods of political science research, especially with respect to machine learning.

But what are some moral implications of these polls? Did they affect voter turnout? What did you think of Slate's "Votecastr" application, which constantly updated results throughout the day? Was this just a polling error that happened across the critical states?

Here are some major polls across the United States. Even Andrew Gelman didn't get it right. Let's try to break these down a little bit:

- http://projects.fivethirtyeight.com/2016-election-forecast/?ex_cid=rrpromo
- http://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html
- http://election.princeton.edu/2016/11/06/is-99-a-reasonable-probability/
- http://www.slate.com/articles/news_and_politics/politics/2016/09/trump_vs_clinton_who_s_winning_today_s_forecasts_of_who_will_win_the_election.html
- http://andrewgelman.com/2016/11/08/updating-forecast-election-night-r/

7.3.1 Correlated Errors

Correlated errors are bad. They will screw up the confidence intervals when you do regressions. They will mess with predictive powers. However, they don't mess with the point estimates. Many people are blaming correlated errors for the major polling errors that occurred. Similarly, correlated errors are what caused the financial crash of 2007 to happen—you should check out the movie or book *The Big Short* if you're interested in this. Basically in 2007, the bankers all assumed that the the chances of default rates were independent. That is, me defaulting on my house loan didn't affect the probability of you defaulting on your house loan. But that's not true. They're actually correlated. Here's a video explaining this: https://www.youtube.com/watch?v=3hG4X5iTK8M

So where did this all go wrong? We could argue that there were emotional voters. Answering questions to a pollster is different than voting in a voting booth. We can imagine someone that weakly answers Clinton in the polls but in the voting booth, in an emotional rage, cast their vote for Trump. So if a small but consistent proportion of people vote this way, then this is a source of correlated errors because no poll would be able to successfully predict this group of people. The errors across the polls are correlated, and nobody sees it.

Depressed voters might also be a major problem. What this means is that there are voters that are less than enthusiastic to vote. Even if they were intending to vote for Clinton, they might not have made it out to the polls because they felt they had better things to do and that the polls already had Clinton down as a certainty. Again, if a specific group of individuals do this this generates correlated polling errors.

Let's see how this works mathematically. Let's use the example from Homework 2. Let the total vote share be defined as

$$v(f, s) = k(f)^{\alpha}(s)^{1-\alpha}$$

where $0 < k, \alpha < 1$. Now, we know that f and s might be measured with some measurement error. Let's call these errors Δf and Δs. Then, we want to determine the absolute uncertainty of v, the vote share. Let's assume that the errors are independent: that is, the error vectors are orthogonal to each other. Then, this means that we can calculate the errors simply as follows, using the Pythagoras theorem:

$$[\Delta v]^2 = \left[\frac{\partial v}{\partial f}\Delta f\right]^2 + \left[\frac{\partial v}{\partial s}\Delta s\right]^2$$

Why is this the case? Let's say we have a function $f(x)$ and the measurement x, and there is a measurement error Δx. Then, this forms the delta-bounds around x, with its corresponding ε-bands in the function $f(x)$. Notice that the amount that Δx changes relative to $\Delta f(x)$ is simply

$$\frac{\Delta f(x)}{\Delta x} = \frac{df(x)}{dx}$$

Now, notice that we made the assumption that the errors are not correlated. If they were correlated, that means the above measurement greatly increases. Imagine a 3-4-5 right triangle; then, redo it with the angle between 3 and 4 being 120 degrees. This means that the new side is now $c^2 = a^2 + b^2 - 2acCos(\theta)$, which means it is around a length of 6.08—an increase of more than 20% in the correlated errors!

8 Tuesday, November 14

8.1 Proofs Techniques

One book that really goes in-depth on this issue is *How to Prove It: A Structured Approach* by Daniel J. Velleman. We'll go through some basic terminology and techniques in proofs. Professor Michael Hutchings, a mathematician at UC Berkeley, has also written a short guide on mathematical proofs: `https://math.berkeley.edu/~hutching/teach/proofs.pdf`.

8.1.1 Logical Operators

- **Not**
- **And**
- **Or**
- **If...then**
- **If and only if**

8.1.2 Proof by Cases

This is pretty straightforward—discuss the various cases that exist and show that it is true for all cases. Usually useful for a small number of cases.

Example: For every integer x, the integer $x(x+1)$ is even.
Proof: Let x be any integer. Then x is either even or odd. This gives us two different cases.

- Case 1: suppose x is even. Choose an integer k such that $x = 2k$. Then $x(x+1) = 2k(2k+1)$. Let $y = k(2k+1)$; then this means that y is an integer and $x(x+1) = 2y$ which means that $x(x+1)$ is even.
- Case 2: suppose x is odd. Choose an integer k such that $x = 2k+1$. Then $x(x+1) = (2k+1)(2k+2)$. Let $y = (2k+1)(k+1)$ so that means that $x(x+1) = 2y$, so that means $x(x+1)$ is even.

8.1.3 Proof by Contradiction

Suppose that we want to prove that the statement P is true. We begin by assuming that P is false. We then try to deduce a contradiction, i.e. some statement Q which we know is false. If we succeed, then our assumption that P is false must be wrong. So that means P must be true, and we complete the proof. This technique is really useful when you're trying to show something is uncountable, irrational, etc.

Example: Prove that $\sqrt{2}$ is irrational.
Proof: Suppose that $\sqrt{2}$ is rational. That means we can write $\sqrt{2}$ as:

$$\sqrt{2} = \frac{a}{b}$$

where p and q are integers. Assume that p and q have no common factors, because any common factors can be canceled out. Squaring both sides, we get:

$$2 = \frac{p^2}{q^2}$$

which implies that

$$2q^2 = p^2$$

Thus, this means that p^2 is even. The only way that this is true is if p itself is even. But that means that p^2 is actually divisible by 4, which means that q^2 and therefore q must be even. So that means p and q are both even, but that means they have a common factor of 2; therefore, this is a contradiction of our assumption that they had no common factors. Therefore, $\sqrt{2}$ must be irrational.

8.1.4 Proof by Induction

The logic is sort of similar to a row of falling dominoes. You first prove that the base case is true; then, you show that if the n case is true then it must also be true for the $n+1$ case. This is similar to dominoes because if it is true for the first case, and then you assume that it is true for the n case and then you show that it is true for the $n+1$ case, that means it must be true for the $n+2$ case because you can just say that $n+2 = m$ which means it must be true for the $m+1 = n+3$ case.

Example: Show that $0+1+2+...+n = \frac{n(n+1)}{2}$. *Proof*: Base Case: $n = 0$, so that means $0(0+1)/2 = 0$ which means it is true as expected.
Inductive Case: Assume that the base case $0+1+2+...+n = \frac{n(n+1)}{2}$ is true. Now let's check $n+1$, which means that we want to show that $0+1+2+...+n+(n+1) = \frac{(n+1)(n+2)}{2}$. But we know that $0+1+...+n = \frac{n(n+1)}{2}$ so that means that $\frac{n(n+1)}{2}+(n+1) = \frac{n(n+1)}{2}+\frac{2n+2}{2} = \frac{n^2+3n+2}{2} = \frac{(n+1)(n+2)}{2}$. So that means that we have proved that $0+1+2+...+n = \frac{n(n+1)}{2}$ is true.

8.1.5 Without Loss of Generality

Often abbreviated WLOG. For instance, if a proof involves cases where $b > 0$ and $b < 0$ but both have the same logic, you can just say assume without loss of generality that $b > 0$. Often used to simplify computations. For instance, in many OLS proofs, the proof might say "assume, WLOG, that $\bar{x} = 0$." The mean of x might make computations more difficult, but you end up with the same results.

8.2 Proof by Induction

The logic is sort of similar to a row of falling dominoes. You first prove that the base case is true; then, you show that if the n case is true then it must also be true for the $n+1$ case. This is similar to dominoes because if it is true for the first case, and then you assume that it is true for the n case and then you show that it is true for the $n+1$ case, that means it must be true for the $n+2$ case because you can just say that $n+2 = m$ which means it must be true for the $m+1 = n+3$ case.

Example: Show that $0+1+2+...+n = \frac{n(n+1)}{2}$. *Proof*: Base Case: $n = 0$, so that means $0(0+1)/2 = 0$ which means it is true as expected.
Inductive Case: Assume that the base case $0+1+2+...+n = \frac{n(n+1)}{2}$ is true. Now let's check $n+1$,

which means that we want to show that $0+1+2+...+n+(n+1) = \frac{(n+1)(n+2)}{2}$. But we know that $0+1+...+n = \frac{n(n+1)}{2}$ so that means that $\frac{n(n+1)}{2}+(n+1) = \frac{n(n+1)}{2}+\frac{2n+2}{2} = \frac{n^2+3n+2}{2} = \frac{(n+1)(n+2)}{2}$. So that means that we have proved that $0+1+2+...+n = \frac{n(n+1)}{2}$ is true.

8.3 Supremum and Infimum

Theorem: $\mathbb{N}$ is not bounded above.
Proof: Suppose that $\mathbb{N}$ were bounded above. Since $\mathbb{N} \neq \emptyset$, this means that there would be a least upper bound α for $\mathbb{N}$. Then,

$$\alpha \geq n \text{ for all } n \text{ in } \mathbb{N}$$

Consequently, this means that

$$\alpha \geq n+1 \text{ for all } n \text{ in } \mathbb{N}$$

since $n+1$ is in $\mathbb{N}$ if n is in $\mathbb{N}$. But this means that

$$\alpha - 1 \geq n \text{ for all } n \text{ in } \mathbb{N}$$

and this means that $\alpha - 1$ is also an upper bound for $\mathbb{N}$, contradicting the fact that α is the least upper bound.

9 Tuesday, November 21

9.1 Example of a Lagrangian

We want to find the following

$$\max f(x,y) = xy \text{ s.t. } x^2+y^2=1$$

First we rewrite the constraint $x^2+y^2=1$ in the form $g(x,y)=0$, that is

$$x^2+y^2-1=0$$

Our Lagrangian function is thus:

$$\mathcal{L}(x,y,\lambda) = xy + \lambda(x^2+y^2-1)$$

Let's find the critical points of the function $\mathcal{L}$ for three variables.

$$\mathcal{L}_x = y + 2\lambda x$$

$$\mathcal{L}_y = x + 2\lambda y$$

$$\mathcal{L}_\lambda = x^2+y^2-1$$

Now, setting $\mathcal{L}_x = \mathcal{L}_y = 0$, we get:

$$y = -2\lambda x$$

$$x = -2\lambda y$$

And if $x \neq 0$ and $y \neq 0$,

$$\frac{y}{-2x} = \lambda$$

$$\frac{x}{-2y} = \lambda$$

$$\frac{y}{x} = \frac{x}{y}$$

$$x^2 = y^2$$

which means that

$$x = \pm y$$

which means that

$$x^2+y^2=1 \Rightarrow x^2+x^2=1 \Rightarrow x = \pm\frac{1}{\sqrt{2}}$$

by symmetry,

$$y = \pm\frac{1}{\sqrt{2}}$$

Notice that we assumed that $x \neq 0$ and $y \neq 0$ in order to divide x and y. But we actually have to take those into account. Therefore, we also have to check $(0,1)$ and $(1,0)$ as possible points of maxima. But notice that $f(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}) = \frac{1}{2}$ and $f(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}) = -\frac{1}{2}$ while $f(1,0)=0$ and $f(0,1)=0$. This means we know that $(0,1)$ and $(1,0)$ are not points of minima or maxima (but you have to check them!!).

Now we need to check all possible combinations of $x = \pm\frac{1}{\sqrt{2}}$ and $y = \pm\frac{1}{\sqrt{2}}$. Notice that λ depends on which combination we do. Let's form the bordered Hessian.

$$B = \begin{bmatrix} 0 & 2x & 2y \\ 2x & 2\lambda & 1 \\ 2y & 1 & 2\lambda \end{bmatrix}$$

Case 1: $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. In this case, $\lambda = -\frac{1}{2}$.

$$B = \begin{bmatrix} 0 & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -1 & 1 \\ \sqrt{2} & 1 & -1 \end{bmatrix}$$

We need to check $n - k = 1$ leading principal minors, such that $r \in \{2k+1, ..., k+n\} = \{3\}$. So $|B_3| = 8$; and since $(-1)^{3-1}(8) > 0$ this means we are at a maximum! Notice that $(-1)^1(8) < 0$ which means it is not a minimum.

Case 2: $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ In this case, $\lambda = -\frac{1}{2}$.

$$B = \begin{bmatrix} 0 & -\sqrt{2} & -\sqrt{2} \\ -\sqrt{2} & -1 & 1 \\ -\sqrt{2} & 1 & -1 \end{bmatrix}$$

Again, we need to check $n - k = 1$ leading principal minors. Again, determinant is 8. So we are a max.

Case 3: $(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. In this case, $\lambda = \frac{1}{2}$.

$$B = \begin{bmatrix} 0 & \sqrt{2} & -\sqrt{2} \\ \sqrt{2} & 1 & 1 \\ -\sqrt{2} & 1 & 1 \end{bmatrix}$$

So in this case determinant is -8. Notice that $(-1)^1(-8) > 0$ which means it is a min.

Case 4: $(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$. In this case, $\lambda = \frac{1}{2}$.

$$B = \begin{bmatrix} 0 & -\sqrt{2} & \sqrt{2} \\ -\sqrt{2} & 1 & 1 \\ \sqrt{2} & 1 & 1 \end{bmatrix}$$

Again determinant is -8 so by the same logic as above it is a min.

So now we found our max and min of this constrained optimization problem!

10 Tuesday, November 28

10.1 Kuhn-Tucker Conditions: First Two Examples from Iain's Notes

These two examples are from Iain's notes. We won't go into the details of K-T conditions—they're pretty complicated and the math is pretty wild!

10.2 Snake-Fighting Example

Credits to Jason Davis for this example. Here is the link to the backstory if you're interested: `https://www.mcsweeneys.net/articles/faq-the-snake-fight-portion-of-your-thesis-defense`. Let's say that you have 16 hours in your day. You can spend time on your thesis, which is x, and you can spend your time on learning snake-fighting, which is t. Let's say that as part of your graduation requirements you have to practice fighting snakes at least one hour a day. You can choose not to spend all 16 hours on these two activities. Let U be your utility function:

$$U(x,t) = 12\sqrt{x} + 3t$$

So now let's maximize our utility function given the constraints $x + t \leq 16$ and $t > 1$. This means that our Lagrangian will be as follows:

$$\mathcal{L}(x,t) = 12\sqrt{x} + 3t + \lambda_1(16 - x - t) + \lambda_2(t - 1)$$

Then we find our first-order conditions:

$$\frac{\partial \mathcal{L}}{\partial x} = 6x^{-1/2} - \lambda_1 = 0$$
$$\frac{\partial \mathcal{L}}{\partial y} = 3 - \lambda_1 + \lambda_2 = 0$$

So there are 6 Kuhn-Tucker conditions:

$$16 - x - t \geq 0, \lambda_1 \geq 0, \lambda_1(16 - x - t) = 0$$

$$t - 1 \geq 0, \lambda_2 \geq 0, \lambda_2(t - 1) = 0$$

So now let's solve the four equations for four unknowns. Let's assume first that $t = 1$. Then that means $\lambda_1(15 - x) = 0$. This implies that x could be either $x = 15$ or $x = 0$. If $x = 0$, then $\lambda_1 = 0$ and that means $\lambda_2 = -3$, which violates one of our K-T conditions. If $x = 15$, then $\lambda_1 \approx 1.5$, which means that $\lambda_2 \approx -1.5$ which is a contradiction to our K-T conditions. So that means that $t \neq 1$.

This means that we can assume that $t > 1$ and $\lambda_2 = 0$. That means $\lambda_1 = 3$ which means that $x = 4$ and $t = 12$. So $t = 12$ and $x = 4$ is our max. We can see that our K-T conditions are satsified. Now, let's look at the SOC. Remember, we only have to include the constraints that were binding! Notice that because $\lambda_2 = 0$ this meant that the $t > 1$ constraint was not binding.

$$B = \begin{bmatrix} 0 & -1 & -1 \\ -1 & -3x^{-3/2} & 0 \\ -1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & -1 \\ -1 & -0.375 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

There are $n = 2$ variables and 1 *binding* constraint, which means that we need to check the final leading principal minor of the bordered Hessian; $r = 3$. The determinant of the last leading principal minor is 0.375. Since $(-1)^{3-1}(0.375) > 0$ this means we are at a maximum. Hooray!

11 Tuesday, December 5

11.1 Implicit Functions and Differentiation of Implicit Functions

Implicit differentiation is nothing more than a special case of the chain rule. The usual differentiation problem involves functions y written explicitly as functions of x. So let's do an example.

11.1.1 Implicit Functions: Example 1

Assume that y is a function of x. That is, we can imagine y being $y(x)$. Find $y' = dy/dx = x^3 + y^3 = 4$. Differentiating both sides, we get

$$D(x^3) + D(y^3) = D(4)$$

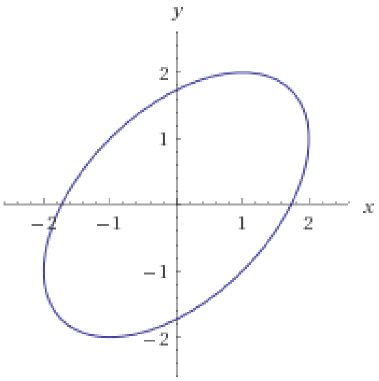
$$3x^2\frac{dx}{dx} + 3y^2\frac{dy}{dx} = 0$$

Solving for dy/dx we get (notice that we have a dx/dx in the first time but that equals 1).

$$\frac{dy}{dx} = -\frac{x^2}{y^2}$$

11.1.2 Implicit Functions: Example 2

The graph of $x^2 - xy + y^2 = 3$ is a "tilted" ellipse, which looks as follows below. Among all the points (x, y) on this graph, find the largest and smallest values of y. Among all the points (x, y) on the graph, find the largest and smallest values of x.



It is easy to see where the extreme values of x and y are located. Again, we treat y ias a function of x (so we can think of y being $y(x)$). The extreme values of x are located at the edges (where the derivative of y does not exist) and the extreme values of y are also located at the edges (where the derivate of y is 0). Begin by taking the derivatives of both sides:

$$D(x^2 - xy + y^2) = D(3)$$

$$2x - (x\frac{dy}{dx} + y) + 2y\frac{dy}{x} = 0$$

Solving for $\frac{dy}{dx}$.

$$-x\frac{dy}{dx} + 2y\frac{dy}{x} = -2x + y$$
$$\frac{dy}{dx} = \frac{y-2x}{2y-x}$$

So to find the max values of y, we just take what we found above and set it equal to 0. Notice that this is the case when $y - 2x = 0$ which implies that $y = 2x$. Substituting this into the original equation, we get:

$$x^2 - x(2x) + (2x)^2 = 3 \Rightarrow x^2 - 2x^2 + 4x^2 = 3x^2 = 3 \Rightarrow x = \pm 1$$

So this mean that the max point occurs at $(1, 2)$ while the minimum point occurs at $(-1, -2)$.

Now we find the max values of x. When does this occur? This occurs when the derivative of y, dy/dx does not exist. So that means $2y - x = 0$ which means that$x = 2y$. Plugging this into our original equation, we get:

$$(2y)^2 - 2y^2 + y^2 = 3 \Rightarrow 3y^2 = 3 \Rightarrow y^2 = 1 \Rightarrow y = \pm 1$$

So the max value of x occurs at $(2, 1)$ and the min value of x occurs at $(-2, -1)$.

11.1.3 Implicit Functions: Example 3

Assume that the supply and demand functions for a single commodity are:

$$Q = S(P, T)$$
$$Q = D(P, Y)$$

where Y is income, T is tax of the commodity, and P is the price. We do not assume any functional forms; however, we do assume that supply increases with price and decreases with tax, while the demand decreases with price and increases with income. So formally, assume that

$$S_P = \frac{\partial S}{\partial P} > 0$$
$$S_T = \frac{\partial S}{\partial T} < 0$$
$$D_P = \frac{\partial D}{\partial P} < 0$$
$$D_Y = \frac{\partial D}{\partial Y} > 0$$

At the equilibrium stage, we know that supply and demand are equal to each other. So that means $0 = S(P, T) - D(P, Y)$. The derivative of the right-hand side with respect to price is:

$$S_P - D_P > 0$$

Because $D_P < 0$ and $S_P > 0$. We can see that the above equation determines the price P as a function of income Y and T. Then, taking total derivatives with respect to Y, we get:

$$0 = S_P\frac{\partial P}{\partial Y} + D_P\frac{\partial P}{\partial Y} - D_Y$$

And if we take the total derivative with respect to T, we get:

$$0 = S_P \frac{\partial P}{\partial T} + S_T - D_P \frac{\partial P}{\partial T}$$

So that means, rearranging things around,

$$(S_P - D_P) \frac{\partial P}{\partial Y} = D_Y$$

And also:

$$(S_P - D_P) \frac{\partial P}{\partial T} = -S_T$$

So that means

$$\frac{\partial P}{\partial Y} = \frac{D_Y}{S_P - D_P} > 0$$

And

$$\frac{\partial P}{\partial T} = \frac{-S_T}{S_P - D_P} > 0$$

Thus, the price increases with both an increase in income and tax.

12 Tuesday, December 12

This is the last section! Woohoo!

Today, we'll be going over Sperner's Lemma, which is the combinatorial analog for Brouwer's fixed-point theorem. It's essentially a geometric look at Brouwer's fixed-point theorem, and shows us the cool things we can do with the fixed-point theorem, which includes solving some real world problems.

Theorem: *Sperner's Lemma.* Every Sperner coloring or Sperner labeling of a triangulation of an n-dimensional simplex contains a cell colored with a complete set of colors.