PROFESSOR: So I'm using a few things here, right? I'm using the fact that KL is non-negative. But KL is equal to 0 when I take twice the same argument. So I know that this function is always non-negative. So that's theta and that's KL P theta star P theta. And I know that at theta star, it's equal to 0. OK?

I could be in the case where I have this happening. I have two-- let's call it theta star prime. I have two minimizers. That could be the case, right? I'm not saying that-- so K of L-- KL is 0 at the minimum. That doesn't mean that I have a unit minimum, right? But it does, actually.

What do I need to use to make sure that I have only one minimum? So the definiteness is guaranteeing to me that there's a unique P theta star that minimizes it. But then I need to make sure that there's a unique-- from this unique P theta star, I need to make sure there's a unique theta star that defines this P theta star.

Exactly. All right, so I combine definiteness and identifiability to make sure that there is a unique minimizer, in this case cannot exist. OK, so basically, let me write what I just said. So definiteness, that implies that P theta star is the unique minimizer of P theta maps to KL P theta star P theta. So definiteness only guarantees that the probability distribution is uniquely identified.

And identifiability implies that theta star is the unique minimizer of theta maps to KL P theta star P theta, OK? So I'm basically doing the composition of two injective functions. The first one is the one that maps, say, theta to P theta. And the second one is the one that maps P theta to the set of minimizers, OK?

So at least morally, you should agree that theta star is the minimizer of this thing. Whether it's unique or not, you should agree that it's a good one. So maybe you can think a little longer on this.

So thinking about this being the minimizer, then it says, well, if I actually had a good estimate for this function, I would use the strategy that I described for the total variation, which is, well, I don't know what this function looks like. It depends on theta star. But maybe I can find an estimator of this function that fluctuates around this function, and such that when I minimize this estimator of the function, I'm actually not too far, OK?

And this is exactly what drives me to do this, because I can actually construct an estimator. I can actually construct an estimator such that this estimator is actually-- of the KL is actually close to the KL, all right? So I define KL hat. So all we did is just replacing expectation with respect to theta star by averages. That's what we did.

So if you're a little puzzled by this error, that's all it says. Replace this guy by this guy. It has no mathematical meaning. It just means just replace it by. And now that actually tells me how to get my estimator. It just says, well, my estimator, KL hat, is equal to some constant which I don't know. I mean, it certainly depends on theta star, but I won't care about it when I'm trying to minimize-- minus 1/n sum from i from 1 to n log f theta of x.

So here I'm reading it with the density. You have it with the PMF on the slides, and so you have the two versions in front of you, OK? Oh sorry, I forgot the xi.

Now clearly, this function I know how to compute. If you give me a theta, since I know the form of the density f theta, for each theta that you give me, I can actually compute this quantity, right? This I don't know, but I don't care. Because I'm just shifting the value of the function I'm trying to minimize. The set of minimizers is not going to change.

So now, this is my estimation strategy. Minimize in theta KL hat P theta star P theta, OK? So now let's just make sure that we all agree that-- so what we want is the argument of the minimum, right? arg min means the theta that minimizes this guy, rather than finding the value of the min.

OK, so I'm trying to find the arg min of this thing. Well, this is equivalent to finding the arg min of, say, a constant minus 1/n sum from i from 1 to n of log f theta of xi. So that's just-- I don't think it likes me. No. OK, so thus minimizing this average, right? I just plugged in the definition of KL hat.

Now, I claim that taking the arg min of a constant plus a function or the arg min of the function is the same thing. Is anybody not comfortable with this idea? OK, so this is the same. By the way, this I should probably switch to the next slide, because I'm writing the same thing, but better. And it's with PMF rather than as PF.

OK, now, arg min of the minimum is the same of arg max-- sorry, arg min of the negative thing is the same as arg max without the negative, right? arg max over theta of 1/n from i equal equal 1 to n log f theta of xi. Taking the arg min of the average or the arg min of the sum, again, it's not going to make much difference. Just adding constants OR multiplying by constants does not change the arg min or the arg max.

Now, I have the sum of logs, which is the log of the product. OK? It's the arg max of the log of f theta of x1 times f theta of x2, f theta of xn. But the log is a function that's increasing, so maximizing log of a function or maximizing the function itself is the same thing. The value is going to change, but the arg max is not going to change. Everybody agrees with this?

So this is equivalent to arg max over theta of pi from 1 to n of f theta xi. And that's because x maps to log x is increasing. So now I've gone from minimizing the KL to minimizing the estimate of the KL to maximizing this product.

Well, this chapter is called maximum likelihood estimation. The maximum comes from the fact that our original idea was to minimize the negative of a function. So that's why it's maximum likelihood. And this function here is called the likelihood.

This function is really just telling me-- they call it likelihood because it's some measure of how likely it is that theta was the parameter that generated the data. OK, so let's go to the-- well, we'll go to the formal definition in a second. But actually, let me just give you intuition as to why this is the distribution of the data. Why this is the likelihood-- sorry. Why is this making sense as a measure of likelihood?

Let's now think for simplicity of the following model. So I have-- I'm on the real line and I look at n, say, theta 1 for theta in the real-- do you see that? OK. Probably you don't. Not that you care. OK, so--

OK, let's look at a simple example. So here's the model. As I said, we're looking at observations on the real line. And they're distributed according to some n theta 1. So I don't care about the variance. I know it's 1. And it's indexed by theta in the real line.

OK, so this is-- the only thing I need to figure out is, what is the mean of those guys, OK? Now, I have this n observations. And if you actually remember from your probability class, are you familiar with the concept of joint density? You have multivariate observations. The joint density of independent random variables is just a product of their individual densities.

So really, when I look at the product from i equal 1 to n of f theta of xi, this is really the joint density of the vector-- well, let me not use the word vector-- of x1 xn, OK? So if I take the product of density, is it still a density? And it's actually-- but this time on the r to the n.

And so now what this thing is telling me-- so think of it in r2, right? So this is the joint density of two Gaussians. So it's something that looks like some bell-shaped curve in two dimensions. And it's centered at the value theta theta. OK, they both have the mean theta.

So let's assume for one second-- it's going to be hard for me to make pictures in n dimensions. Actually, already in two dimensions, I can promise you that it's not very easy. So I'm actually just going to assume that n is equal to 1 for the sake of illustration.

OK, so now I have this data. And now I have one observation, OK? And I know that the f theta looks like this. And what I'm doing is I'm actually looking at the value of x theta as my observation. Let's call it x1.

Now, my principal tells me, just find the theta that makes this guy the most likely. What is the likelihood of my x1? Well, it's just the value of the function. That this value here. And if I wanted to find the most likely theta that had generated this x1, what I would need to do is to shift this thing and put it here.

And so my estimate, my maximum likelihood estimator here would be theta is equal to x1, OK? That would be just the observation. Because if I have only one observation, what else am I going to do?

OK, and so it sort of makes sense. And if you have more observations, you can think of it this way, as if you had more observations. So now I have, say, K observations, or n observations. And what I do is that I look at the value for each of these guys. So this value, this value, this value, this value. I take their product and I make this thing large.

OK, why do I take the product? Well, because I'm trying to maximize their value all together, and I need to just turn it into one number that I can maximize. And taking the product is the natural way of doing it, either by motivating it by the KL principle or motivating it by maximizing the joint density, rather than just maximizing anything.

OK, so that's why, visually, this is the maximum likelihood. It just says that if my observations are here, then this guy, this mean theta, is more likely than this guy. Because now if I look at the value of the function for this guy-- if I look at theta being this thing, then this is a very small value. Very small value, very small value, very small value. Everything gets a super small value, right? That's just the value that it gets in the tail here, which is very close to 0. But as soon as I start covering all my points with my bell-shaped curve, then all the values go up.

All right, so I just want to make a short break into statistics, and just make sure that the maximum likelihood principle involves maximizing a function. So I just want to make sure that we're all on par about how do we maximize functions. In most instances, it's going to be a one-dimensional function, because theta is going to be a one-dimensional parameter. Like here it's the real line. So it's going to be easy. In some cases, it may be a multivariate function and it might be more complicated. OK, so let's just make this interlude.

So the first thing I want you to notice is that if you open any book on what's called optimization, which basically is the science behind optimizing functions, you will talk mostly-- I mean, I'd say 99.9% of the cases will talk about minimizing functions. But it doesn't matter, because you can just flip the function and you just put a minus sign, and minimizing h is the same as maximizing minus h and the opposite, OK?

So for this class, since we're only going to talk about maximum likelihood estimation, we will talk about maximizing functions. But don't be lost if you decide suddenly to open a book on optimization and find only something about minimizing functions. OK, so maximizing an arbitrary function can actually be fairly difficult. If I give you a function that has this weird shape, right-- let's think of this polynomial for example-- and I wanted to find the maximum, how would we do it?

So what is the thing you've learned in calculus on how to maximize the function? Set the derivative equal to 0. Maybe you want to check the second derivative to make sure it's a maximum and not a minimum. But the thing is, this is only guaranteeing to you that you have a local one, right?

So if I do it for this function, for example, then this guy is going to satisfy this criterion, this guy is going to satisfy this criterion, this guy is going to satisfy this criterion, this guy here, and this guy satisfies the criterion, but not the second derivative one. So I have a lot of candidates. And if my function can be really anything, it's going to be difficult, whether it's analytically by taking derivatives and setting them to 0, or trying to find some algorithms to do this.

Because if my function is very jittery, then my algorithm basically has to check all candidates. And if there's a lot of them, it might take forever, OK? So this is-- I have only one, two, three, four, five candidates to check. But in practice, you might have a million of them to check. And that might take forever.

OK, so what's nice about statistical models, and one of the things that makes all these models particularly robust, and that we still talk about them 100 years after they've been introduced is that the functions that-- the likelihoods that they lead for us to maximize are actually very simple. And they all share a nice property, which is that of being concave. All right, so what is a concave function? Well, by definition, it's just a function for which-- let's think of it as being twice differentiable. You can define functions that are not differentiable as being concave, but let's think about it as having a second derivative.

And so if you look at the function that has a second derivative, concave are the functions that have their second derivative that's negative everywhere. Not just at the maximum, everywhere, OK? And so if it's strictly concave, this second derivative is actually strictly less than zero.

And particularly if I think of a linear function, y is equal to x, then this function has its second derivative which is equal to zero, OK? So it is concave. But it's not strictly concave, OK? If I look at the function which is negative x squared, what is its second derivative? Minus 2. So it's strictly negative everywhere, OK?

So actually, this is a pretty canonical example strictly concave function. If you want to think of a picture of a strictly concave function, think of negative x squared. So parabola pointing downwards.

OK, so we can talk about strictly convex functions. So convex is just happening when the negative of the function is concave. So that translates into having a second derivative which is either non-negative or positive, depending on whether you're talking about convexity or strict convexity. But again, those convex functions are convenient when you're trying to minimize something. And since we're trying to maximize the function, we're looking for concave.

So here are some examples. Let's just go through them quickly. OK, so the first one is-- so here I made my life a little uneasy by talking about the functions in theta, right? I'm talking about likelihoods, right? So I'm thinking of functions where the parameter is theta.

So I have h of theta. And so if I start with theta squared, negative theta squared, then as we said, h prime prime of theta, the second derivative is minus 2, which is strictly negative, so this function is strictly concave. OK, another function is h of theta, which is-- what did we pick-- square root of theta. What is the first derivative? 1/2 square root of theta.

What is the second derivative? So that's theta to the negative 1/2. So I'm just picking up another negative 1/2, so I get negative 1/4. And then I get theta to the 3/4 downstairs, OK? Sorry, 3/2. And that's strictly negative for theta, say, larger than 0. And I really need to have this thing larger than 0 so that it's well-defined. But strictly larger than 0 is so that this thing does not blow up to infinity.

And it's true. If you think about this function, it looks like this. And already, the first derivative to infinity at 0. And it's a concave function, OK?

Another one is the log, of course. What is the derivative of the log? That's 1 over theta, where h prime of theta is 1 over theta. And the second derivative negative 1 over theta squared, which again, is negative if theta is strictly positive. Here I define it as-- I don't need to define it to be strictly positive here, but I need it for the log.

And sine. OK, so let's just do one more. So h of theta is sine of theta. But here I take it only on an interval, because you want to think of this function as pointing always downwards. And in particular, you don't want this function to have an inflection point. You don't want it to go down and then up and then down and then up, because this is not concave.

And so sine is certainly going up and down, right? So what we do is we restrict it to an interval where sine is actually-- so what does the sine function looks like at 0, 0? And it's going up. Where is the first maximum of the sine?

STUDENT: [INAUDIBLE]

PROFESSOR: I'm sorry.

STUDENT: Pi over 2.

PROFESSOR: Pi over 2, where it takes value 1. And then it goes down again. And then that's at pi. And then I go down again. And here you see I actually start changing my inflection.

So what we do is we stop it at pi. And we look at this function, it certainly looks like a parabola pointing downwards. And so if you look at the-- you can check that it actually works with the derivatives. So the derivative of sine is cosine. And the derivative of cosine is negative sine.

OK, and this thing between 0 and pi is actually positive. So this entire thing is going to be negative. OK? And you know, I can come up with a lot of examples, but let's just stick to those. There's a linear function, of course. And the find function is going to be concave, but it's actually going to be convex as well, which means that it's certainly not going to be strictly concave or convex, OK?

So here's your standard picture. And here, if you look at the dotted line, what it tells me is that a concave function, and the property we're going to be using is that if a strictly concave function has a maximum, which is not always the case, but if it has a maximum, then it actually must be-- sorry, a local maximum, it must be a global maximum. OK, so just the fact that it goes up and down and not again means that there's only global maximum that can exist.

Now if you looked, for example, at the square root function, look at the entire positive real line, then this thing is never going to attain a maximum. It's just going to infinity as x goes to infinity. So if I wanted to find the maximum, I would have to stop somewhere and say that the maximum is attained at the right-hand side.

OK, so that's the beauty about convex functions or concave functions, is that essentially, these functions are easy to maximize. And if I tell you a function is concave, you take the first derivative, set it equal to 0. If you find a point that satisfies this, then it must be a global maximum, OK?

STUDENT: What if your set theta was [INAUDIBLE] then couldn't you have a function that, by the definition, is concave, with two upside down parabolas at two disjoint intervals, but yet it has two global maximums?

PROFESSOR: So you won't get them-- so you want the function to be concave on what? On the convex cell of the intervals? Or you want it to be--

STUDENT: [INAUDIBLE] just said that any subset.

PROFESSOR: OK, OK. You're right. So maybe the definition-- so you're pointing to a weakness in the definition. Let's just say that theta is a convex set and then you're good, OK? So you're right. Since I actually just said that this is true only for theta, I can just take pieces of concave functions, right? I can do this, and then the next one I can do this, on the next one I can do this. And then I would have a bunch of them.

But what I want is think of it as a global function on some convex set. You're right. So think of theta as being convex for this guy, an interval, if it's a real line.

OK, so as I said, for more generally-- so we can actually define concave functions more generally in higher dimensions. And that will be useful if theta is not just one parameter but several parameters. And for that, you need to remind yourself of Calculus II, and you have generalization of the notion of derivative, which is called a gradient, which is basically a vector where each coordinate is just the partial derivative with respect to each coordinate of theta.

And the Hessian is the matrix, which is essentially a generalization of the second derivative. I denote it by nabla squared, but you can write it the way you want. And so this matrix here is taking as entry the second partial derivatives of h with respect to theta i and theta j. And so that's the ij-th entry.

Who has never seen that? OK. So now, being concave here is essentially generalizing, saying that a vector is equal to zero. Well, that's just setting the vector-- sorry. The first order condition to say that it's a maximum is going to be the same. Saying that a function has a gradient equal to zero is the same as saying that each of its coordinates are equal to zero. And that's actually going to be a condition for a global maximum here.

So to check convexity, we need to see that a matrix itself is negative. Sorry, to check concavity, we need to check that a matrix is negative. And there is a notion among matrices that compare matrix to zero, and that's exactly this notion. You pre- and post-multiply by the same x. So that works for symmetric matrices, which is the case here.

And so you pre-multiply by x, post-multiply by the same x. So you have your matrix, your Hessian here. It's a d by d matrix if you have a d-dimensional matrix. So let's call it-- OK. And then here I pre-multiply by x transpose. I post-multiply by x.

And this has to be non-positive if I want it to be concave, and strictly negative if I want it to be strictly concave. OK, that's just a real generalization. You can check for yourself that this is the same thing. If I were in dimension 1, this would be the same thing. Why?

Because in dimension 1, pre- and post-multiplying by x is the same as multiplying by x squared. Because in dimension 1, I can just move my x's around, right? And so that would just mean the first condition would mean in dimension 1 that the second derivative times x squared has to be less than or equal to zero. So here I need this for all x's that are not zero, because I can take x to be zero and make this equal to zero, right? So this is for x's that are not equal to zero, OK?

And so some examples. Just look at this function. So now I have functions that depend on two parameters, theta1 and theta2. So the first one is-- so if I take theta to be equal to-- now I need two parameters, r squared. And I look at the function, which is h of theta. Can somebody tell me what h of theta is?

STUDENT: [INAUDIBLE]

PROFESSOR: Minus 2 theta2 squared? OK, so let's compute the gradient of h of theta. So it's going to be something that has two coordinates. To get the first coordinate, what do I do? Well, I take the derivative with respect to theta1, thinking of theta2 as being a constant. So this thing is going to go away. And so I get negative 2 theta1.

And when I take the derivative with respect to the second part, thinking of this part as being constant, I get minus 4 theta2. That clear for everyone? That's just the definition of partial derivatives.

And then if I want to do the Hessian, so now I'm going to get a 2 by 2 matrix. The first guy here, I take the first-- so this guy I get by taking the derivative of this guy with respect to theta1. So that's easy. So that's just minus 2.

This guy I get by taking derivative of this guy with respect to theta2. So I get what? Zero. I treat this guy as being a constant. This guy is also going to be zero, because I take the derivative of this guy with respect to theta1. And then I take the derivative of this guy with respect to theta2, so I get minus 4.

OK, so now I want to check that this matrix satisfies x transpose-- this matrix x is negative. So what I do is-- so what is x transpose x? So if I do x transpose delta squared h theta x, what I get is minus 2 x1 squared minus 4 x2 squared. Because this matrix is diagonal, so all it does is just weights the square of the x's.

So this guy is definitely negative. This guy is negative. And actually, if one of the two is non-zero, which means that x is non-zero, then this thing is actually strictly negative. So this function is actually strictly concave. And it looks like a parabola that's slightly distorted in one direction.

So well, I know this might have been some time ago. Maybe for some of you might have been since high school. So just remind yourself of doing second derivatives and Hessians and things like this. Here's another one as an exercise.

h is minus theta1 minus theta2 squared. So this one is going to actually not be diagonal. The Hessian is not going to be diagonal. Who would like to do this now in class? OK, thank you. This is not a calculus class. So you can just do it as a calculus exercise. And you can do it for log as well.

Now, there is a nice recipe for concavity that works for the second one and the third one. And the thing is, if you look at those particular functions, what I'm doing is taking, first of all, a linear combination of my arguments. And then I take a concave function of this guy.

And this is always going to work. This is always going to give me a complete function. So the computations that I just made, I actually never made them when I prepared those slides because I don't have to. I know that if I take a linear combination of those things and then I take a concave function of this guy, I'm always going to get a concave function.

OK, so that's an easy way to check this, or at least as a sanity check. All right, and so as I said, finding maximizers of concave or strictly concave function is the same as it was in the one-dimensional case. What I do-- sorry, in the one-dimensional case, we just agreed that we just take the derivative and set it to zero. In the high dimensional case, we take the gradient and set it equal to zero. Again, that's calculus, all right?

So it turns out that so this is going to give me equations, right? The first one is an equation in theta. The second one is an equation in theta1, theta2, theta3, all the way to theta d. And it doesn't mean that because I can write this equation that I can actually solve it. This equation might be super nasty. It might be like some polynomial and exponentials and logs equal zero, or some crazy thing.

And so there's actually, for a concave function, since we know there's a unique maximizer, there's this theory of convex optimization, which really, since those books are talking about minimizing, you had to find some sort of direction. But you can think of it as the theory of concave maximization. And they allow you to find algorithms to solve this numerically and fairly efficiently.

OK, that means fast. Even if d is of size 10,000, you're going to wait for one second and it's going to tell you what the maximum is. And that's what machine learning is about. If you've taken any class on machine learning, there's a lot of optimization, because they have really, really big problems to solve.

Often in this class, since this is more introductory statistics, we will have a close form. For the maximum likelihood estimator will be saying theta hat equals, and say x bar, and that will be the maximum likelihood estimator. So just why-- so has anybody seen convex optimization before?

So let me just give you an intuition why those functions are easy to maximize or to minimize. In one dimension, it's actually very easy for you to see that. And the reason is this. If I want to maximize the concave function, what I need to do is to be able to query a point and get as an answer the derivative of this function, OK?

So now I said this is the function I want to optimize, and I've been running my algorithm for 5/10 of a second. And it's at this point here. OK, that's the candidate. Now, what I can ask is, what is the derivative of my function here?

Well, it's going to give me a value. And this value is going to be either negative, positive, or zero. Well, if it's zero, that's great. That means I'm here and I can just go home. I've solved my problem. I know there's a unique maximum, and that's what I wanted to find.

If it's positive, it actually tells me that I'm on the left of the optimizer. And on the left of the optimal value. And if it's negative, it means that I'm at the right of the value I'm looking for.

And so most of the convex optimization methods basically tell you, well, if you query the derivative and it's actually positive, move to the right. And if it's negative, move to the left. Now, by how much you move is basically, well, why people write books. And in higher dimension, it's a little more complicated, because in higher dimension, thinks about two dimensions, then I'm only being able to get in a vector.

And the vector is only telling me, well, here is half of the space in which you can move. Now here, if you tell me move to the right, I know exactly which direction I'm going to have to move. But in two dimension, you're going to basically tell me, well, move in this global direction.

And so, of course, I know there's a line on the floor I cannot move behind. But even if you tell me, draw a line on the floor and move only to that side of the line, then there's many directions in that line that I can go to. And that's also why there's lots of things you can do in optimization.

OK, but still, putting this line on the floor is telling me, do not go backwards. And that's very important. It's just telling you which direction I should be going always, OK? All right, so that's what's behind this notion of gradient descent algorithm, steepest descent. Or steepest descent, actually, if we're trying to maximize.

OK, so let's move on. So this course is not about optimization, all right? So as I said, the likelihood was this guy. The product of f of the xi's. And one way you can do this is just basically the joint distribution of my data at the point theta.

So now the likelihood, formerly-- so here I am giving myself the model e theta. And here I'm going to assume that e is discrete so that I can talk about PMFs. But everything you're doing, just redo for the sake of yourself by replacing PMFs by PDFs, and everything's going to be fine. We'll do it in a second.

All right, so the likelihood of the model. So here I'm not looking at the likelihood of a parameter. I'm looking at the likelihood of a model. So it's actually a function of the parameter. And actually, I'm going to make it even a function of the points x1 to xn.

All right, so I have a function. And what it takes as input is all the points x1 to xn and a candidate parameter theta. Not the true one. A candidate.

And what I'm going to do is I'm going to look at the probability that my random variables under this distribution, p theta, take these exact values, px1, px2, pxn. Now remember, if my data was independent, then I could actually just say that the probability of this intersection is just a product of the probabilities. And it would look something like this.

But I can define likelihood even if I don't have independent random variables. But think of them as being independent, because that's all we're going to encounter in this class, OK? I just want you to be aware that if I had dependent variables, I could still define the likelihood. I would have to understand how to compute these probabilities there to be able to compute it.

OK, so think of Bernoullis, for example. So here is my example of a Bernoulli. So my parameter is-- so my model is 0,1 Bernoulli p. p is in the interval 0,1. The probability, just as a side remark, I'm just going to use the fact that I can actually write the PMF of a Bernoulli in a very concise form, right?

If I ask you what the PMF of a Bernoulli is, you could tell me, well, the probability that x-- so under p, the probability that x is equal to 0 is 1 minus p. The probability under p that x is equal to 1 is equal to p. But I can be a bit smart and say that for any X that's either 0 or 1, the probability under p that X is equal to little x, I can write it in a compact form as p to the X, 1 minus p to the 1 minus x.

And you can check that this is the right form because, well, you have to check it only for two values of X, 0 and 1. And if you plug in 1, you only keep the p. If you plug in 0, you only keep the 1 minus p. And that's just a trick, OK? I could have gone with many other ways. Agreed?

I could have said, actually, something like-- another one would be-- which we are not going to use, but we could say, well, it's xp plus and minus x 1 minus p, right? That's another one. But this one is going to be convenient. So forget about this guy for a second.

So now, I said that the likelihood is just this function that's computing the probability that X1 is equal to little x1. So likelihood is L of X1, Xn. So let me try to make those calligraphic so you know that I'm talking about smaller values, right? Small x's. x1, xn, and then of course p.

Sometimes we even put-- I didn't do it, but sometimes you can actually put a semicolon here, semicolon so you know that those two things are treated differently. And so now you have this thing is equal to what? Well, it's just the probability under p that X1 is little x1 all the way to Xn is little xn. OK, that's just the definition.

All right, so now let's start working. So we write the definition, and then we want to make it look like something we would potentially be able to maximize if I were-- like if I take the derivative of this with respect to p, it's not very helpful because I just don't know. Just want the algebraic function of p. So this thing is going to be equal to what?

Well, what is the first thing I want to use? I have a probability of an intersection of events, so it's just the product of the probabilities. So this is the product from i equal 1 to n of P, small p, Xi is equal to little xi. That's independence.

OK, now, I'm starting to mean business, because for each P, we have a closed form, right? I wrote this as this supposedly convenient form. I still have to reveal to you why it's convenient. So this thing is equal to-- well, we said that that was p xi for a little xi. 1 minus p to the 1 minus xi, OK? So that was just what I wrote over there as the probability that Xi is equal to little xi. And since they all have the same parameter p, just have this p that shows up here.

And so now I'm just taking the products of something to the xi, so it's this thing to the sum of the xi's. Everybody agrees with this? So this is equal to p sum of the xi, 1 minus p to the n minus sum of the xi. If you don't feel comfortable with writing it directly, you can observe that this thing here is actually equal to p over 1 minus p to the xi times 1 minus p, OK?

So now when I take the product, I'm getting the products of those guys. So it's just this guy to the power of sum and this guy to the power n. And then I can rewrite it like this if I want to

And so now-- well, that's what we have here. And now I am in business because I can still hope to maximize this function. And how to maximize this function? All I have to do is to take the derivative. Do you want to do it? Let's just take the derivative, OK?

Sorry, I didn't tell you that, well, the maximum likelihood principle is to just maxim-- the idea is to maximize this thing, OK? But I'm not going to get there right now. OK, so let's do it maybe for the Poisson model for a second.

So if you want to do it for the Poisson model, let's write the likelihood. So right now I'm not doing anything. I'm not maximizing. I'm just computing the likelihood function. OK, so the likelihood function for Poisson.

So now I know-- what is my sample space for Poisson?

STUDENT: Positives.

PROFESSOR: Positive integers. And well, let me write it like this. Poisson lambda, and I'm going to take lambda to be positive. And so that means that the probability under lambda that X is equal to little x in the sample space is lambda to the X over factorial x e to the minus lambda. So that's basically the same as the compact form that I wrote over there. It's just now a different one.

And so when I want to write my likelihood, again, we said little x's. This is equal to what? Well, it's equal to the probability under lambda that X1 is little x1, Xn is little xn, which is equal to the product. OK? Just by independence.

And now I can write those guys as being-- each of them being i equal 1 to n. So this guy is just this thing where a plug in Xi. So I get lambda to the Xi divided by factorial xi times e to the minus lambda, OK? And now, I mean, this guy is going to be nice. This guy is not going to be too nice. But let's write it.

When I'm going to take the product of those guys here, I'm going to pick up lambda to the sum of the xi's. Here I'm going to pick up exponential minus n times lambda. And here I'm going to pick up just the product of the factorials. So x1 factorial all the way to xn factorial.

Then I get lambda, the sum of the xi. Those are little xi's. e to the minus xn lambda. OK? So that might be freaky at this point, but remember, this is a function we will be maximizing. And the denominator here does not depend on lambda. So we knew that maximizing this function with this denominator, or any other denominator, including 1, will give me the same arg max. So it won't be a problem for me. As long as it does not depend on lambda, this thing is going to go away.

OK, so in the continuous case, the likelihood I cannot-- right? So if I would write the likelihood like this in the continuous case, this one would be equal to what? Zero, right? So it's not very helpful. And so what we do is we define the likelihood as the product of the f of theta xi.

Now that would be a jump if I told you, well, just define it like that and go home and don't discuss it. But we know that this is exactly what's coming from the-- well, actually, I think I erased it. It was just behind. So this was exactly what was coming from the KL divergence estimated, right?

The thing that I showed you, if we want to follow this strategy, which consists in estimating the KL divergence and minimizing it, is exactly doing this. So in the Gaussian case-- well, let's write it. So in the Gaussian case, let's see what the likelihood looks like.

OK, so if I have a Gaussian experiment here-- did I actually write it? OK, so I'm going to take mu and sigma as being two parameters. So that means that my sample space is going to be what? Well, my sample space is still R. Those are just my observations.

But then I'm going to have a N mu sigma squared. And the parameters of interest are mu and R. And sigma squared and say 0 infinity. OK, so that's my Gaussian model. Yes.

STUDENT: [INAUDIBLE]

PROFESSOR: No, there's no-- I mean, there's no difference.

STUDENT: [INAUDIBLE]

PROFESSOR: Yeah. I think the all the slides I put the curly bracket, then I'm just being lazy. I just like those concave parenthesis. All right, so let's write it. So the definition, L xi, xn. And now I have two parameters, mu and sigma squared.

We said, by definition, is the product from i equal 1 to n of f theta of little xi. Now, think about it. Here we always had an extra line, right? The line was to say that the definition was the probability that they were all equal to each other. That was the joint probability.

And here it could actually have a line that says it's the joint probability distribution of the xi's. And if it's not independent, it's not going to be the product. But again, since we're only dealing with independent observations in the scope of this class, this is the only definition we're going to be using. OK, and actually, from here on, I will literally skip this step when I talk about discrete ones as well, because they are also independent. Agreed?

So we start with this, which we agreed was the definition for this particular case. And so now all of you know by heart what the density of a-- sorry, that's not theta. I should write it mu sigma squared. And so you need to understand what this density. And it's product of 1 over sigma square root 2 pi times exponential minus xi minus mu squared divided by 2 sigma squared.

OK, that's the Gaussian density with parameters mu and sigma squared. I just plugged in this thing which I don't give you, so you just have to trust me. It's all over any book. Certainly, I mean, you can find it. I will give it to you. And again, you're not expected to know it by heart. Though, if you do your homework every week without wanting to, you will definitely use some of your brain to remember that thing.

OK, and so now, well, I have this constant in front. 1 over sigma square root 2 pi that I can pull out. So I get 1 over sigma square root 2 pi to the power n. And then I have the product of exponentials, which we know is the exponential of the sum. So this is equal to exponential minus. And here I'm going to put the 1 over 2 sigma squared outside the sum.

And so that's how this guy shows up. Just the product of the density is evaluated at, respectively, x1 to xn. OK, any questions about computing those likelihoods? Yes.

STUDENT: Why [INAUDIBLE]

PROFESSOR: Oh, that's a typo. Thank you. Because I just took it from probably the previous thing. So those are actually-- should be-- OK, thank you for noting that one. So this line should say for any x1 to xn in R to the n. Thank you, good catch. All right, so that's really e to the n, right? My sample space always.

OK, so what is maximum likelihood estimation? Well again, if you go back to the estimate that we got, the estimation strategy, which consisted in replacing expectation with respect to theta star by average of the data in the KL divergence, we would try to maximize not this guy, but this guy. The thing that we actually plugged in were not any small xi's. Were actually-- the random variable is capital Xi.

So the maximum likelihood estimator is actually taking the likelihood, which is a function of little x's, and now the values at which it estimates, if you look at it, is actually-- the capital X is my data. So it looks at the function, at the data, and at the parameter theta. That's what the-- so that's the first thing. And then the maximum likelihood estimator is maximizing this, OK?

So in a way, what it does is it's a function that couples together the data, capital X1 to capital Xn, with the parameter theta and just now tries to maximize it. So if this is just a little hard for you to get, the likelihood is formally defined as a function of x, right? Like when I write f of x. f of little x, I define it like that. But really, the only x arguments we're going to evaluate this function at are always the random variable, which is the data.

So if you want, you can think of it as those guys being not parameters of this function, but really, random variables themselves directly. Is there any question?

STUDENT: [INAUDIBLE] those random variables [INAUDIBLE]?

PROFESSOR: So those are going to be known once you have-- so it's always the same thing in stats. You first design your estimator as a function of random variables. And then once you get data, you just plug it in. But we want to think of them as being random variables because we want to understand what the fluctuations are. So we're going to keep them as random variables for as long as we can.

We're going to spit out the estimator as a function of the random variables. And then when we want to compute it from data, we're just going to plug it in. So keep the random variables for as long as you can. Unless I give you numbers, actual numbers, just those are random variables.

OK, so there might be some confusion if you've seen any stats class, sometimes there's a notation which says, oh, the realization of the random variables are lower case versions of the original random variables. So lowercase x should be thought as the realization of the upper case X. This is not the case here.

When I write this, it's the same way as I write f of x is equal to x squared, right? It's just an argument of a function that I want to define. So those are just generic x. So if you correct the typo that I have, this should say that this should be for any x and xn. I'm just describing a function. And now the only place at which I'm interested in evaluating that function, at least for those first n arguments, is at the capital N observations random variables that I have.

So there's actually texts, there's actually people doing research on when does the maximum likelihood estimator exist? And that happens when you have infinite sets, thetas. And this thing can diverge. There is no global maximum. There's crazy things that might happen.

And so we're actually always going to be in a case where this maximum likelihood estimator exists. And if it doesn't, then it means that you actually need to restrict your parameter space, capital Theta, to something smaller. Otherwise it won't exist.

OK, so another thing is the log likelihood estimator. So it is still the likelihood estimator. We solved before that maximizing a function or maximizing log of this function is the same thing, because the log function is increasing. So the same thing is maximizing a function or maximizing, I don't know, exponential of this function.

Every time I take an increasing function, it's actually the same thing. Maximizing a function or maximizing 10 times this function is the same thing. So the function x maps to 10 times x is increasing. And so why do we talk about log likelihood rather than likelihood?

So the log of likelihood is really just-- I mean the log likelihood is the log of the likelihood. And the reason is exactly for this kind of reasons. Remember, that was my likelihood, right? And I want to maximize it. And it turns out that in stats, there's a lot of distributions that look like exponential of something.

So I might as well just remove the exponential by taking the log. So once I have this guy, I can take the log. This is something to a power of something. If I take the log, it's going to look better for me.

I have this thing-- well, I have another one somewhere, I think, where I had the Poisson. Where was the Poisson? The Poisson's gone. So the Poisson was the same thing. If I took the log, because it had a power, that would make my life easier.

So the log doesn't have any particular intrinsic notion, except that it's just more convenient. Now, that being said, if you think about maximizing the KL, the original formulation, we actually remove the log. If we come back to the KL thing-- where is my KL? Sorry.

That was maximizing the sum of the logs of the pi's. And so then we worked at it by saying that the sum of the logs was-- maximizing the sum of the logs was the same as maximizing the product. But here, we're basically-- log likelihood is just going backwards in this chain of equivalences. And that's just because the original formulation was already convenient. So we went to find the likelihood and then coming back to our original estimation strategy. So look at the Poisson. I want to take log here to make my sum of xi's go down.

OK, so this is my estimator. So the log of L-- so one thing that you want to notice is that the log of L of x1, xn theta, as we said, is equal to the sum from i equal 1 to n of the log of either p theta of xi, or-- so that's in the discrete case. And in the continuous case is the sum of the log of f theta of xi.

The beauty of this is that you don't have to really understand the difference between probability mass function and probability distribution function to implement this. Whatever you get, that's what you plug in. Any questions so far?

All right, so shall we do some computations and check that, actually, we've introduced all this stuff-- complicate functions, maximizing, KL divergence, lot of things-- so that we can spit out, again, averages? All right? That's great. We're going to able to sleep at night and know that there's a really powerful mechanism called maximum likelihood estimator that was actually driving our intuition without us knowing.

OK, so let's do this so. Bernoulli trials. I still have it over there. OK, so actually, I don't know what-- well, let me write it like that. So it's P over 1 minus P xi-- sorry, sum of the xi's times 1 minus P is to the n.

So now I want to maximize this as a function of P. Well, the first thing we would want to do is to check that this function is concave. And I'm just going to ask you to trust me on this. So I don't want-- sorry, sum of the xi's. I only want to take the derivative and just go home.

So let's just take the derivative of this with respect to P. Actually, no. This one was more convenient. I'm sorry. This one was slightly more convenient, OK?

So now we have-- so now let me take the log. So if I take the log, what I get is sum of the xi's times log p plus n minus some of the xi's times log 1 minus p. Now I take the derivative with respect to p and set it equal to zero.

So what does that give me? It tells me that sum of the xi's divided by p minus n sum of the xi's divided by 1 minus p is equal to 0. So now I need to solve for p. So let's just do it.

So what we get is that 1 minus p sum of the xi's is equal to p n minus sum of the xi's. So that's p times n minus sum of the xi's plus sum of the xi's. So let me put it on the right. So that's p times n is equal to sum of the xi's.

And that's equivalent to p-- actually, I should start by putting p hat from here on, because I'm already solving an equation, right? And so p hat is equal to syn of the xi's divided by n, which is my xn bar.

Poisson model, as I said, Poisson is gone. So let me rewrite it quickly. So Poisson, the likelihood in X1, Xn, and lambda was equal to lambda to the sum of the xi's e to the minus n lambda divided by X1 factorial, all the way to Xn factorial.

So let me take the log likelihood. That's going to be equal to what? It's going to tell me. It's going to be-- well, let me get rid of this guy first. Minus log of X1 factorial all the way to Xn factorial. That's a constant with respect to lambda. So when I'm going to take the derivative, it's going to go.

Then I'm going to have plus sum of the xi's times log lambda. And then I'm going to have minus n lambda. So now then, you take the derivative and set it equal to zero. So log L-- well, partial with respect to lambda of log L, say lambda, equals zero. This is equivalent to, so this guy goes. This guy gives me sum of the xi's divided by lambda hat equals n.

And so that's equivalent to lambda hat is equal to sum of the xi's divided by n, which is Xn bar. Take derivative, set it equal to zero, and just solve. It's a very satisfying exercise, especially when you get the average in the end. You don't have to think about it forever.

OK, the Gaussian model I'm going to leave to you as an exercise. Take the log to get rid of the pesky exponential, and then take the derivative and you should be fine. It's a bit more-- it might be one more line than those guys.

OK, so-- well actually, you need to take the gradient in this case. Don't check the second derivative right now. You don't have to really think about it. What did I want to add? I think there was something I wanted to say. Yes.

When I have a function that's concave and I'm on, like, some infinite interval, then it's true that taking the derivative and setting it equal to zero will give me the maximum. But again, I might have a function that looks like this. Now, if I'm on some finite interval-- let me go elsewhere.

So if I'm on some finite interval and my function looks like this as a function of theta-- let's say this is my log likelihood as a function of theta-- then, OK, there's no place in this interval-- let's say this is between 0 and 1-- there's no place in this interval where the derivative is equal to 0. And if you actually try to solve this, you won't find a solution which is not in the interval 0, 1.

And that's actually how you know that you probably should not take the derivative equal to zero. So don't panic if you get something that says, well, the solution is at infinity, right? If this function keeps going, you will find that the solution-- you won't be able to find a solution apart from infinity. You are going to see something like 1 over theta hat is equal to 0, or something like this.

So you know that when you've found this kind of solution, you've probably made a mistake at some point. And the reason is because the functions that are like this, you don't find the maximum by setting the derivative equal to zero. You actually just find the maximum by saying, well, it's an increasing function on the interval 0, 1, so the maximum must be attained at 1.

So here in this case, that would mean that my maximum would be 1. My estimator would be 1, which would be weird. So typically here, you have a function of the xi's. So one example that you will see many times is when this guy is the maximum of the xi's. And in which case, the maximum is attained here, which is the maximum of this.

OK, so just keep in mind-- what I would recommend is every time you're trying to take the maximum of a function, just try to plot the function in your head. It's not too complicated. Those things are usually squares, or square roots, or logs. You know what those functions look like. Just plug them in your mind and make sure that you will find a maximum which really goes up and then down again.

If you don't, then that means your maximum is achieved at the boundary and you have to think differently to get it. So the machinery that consists in setting the derivative equal to zero works 80% of the time. But o you have to be careful. And from the context, it will be clear that you had to be careful, because you will find some crazy stuff, such as solve 1 over theta hat is equal to zero.

All right, so before we conclude, I just wanted to give you some intuition about how does the maximum likelihood perform? So there's something called the Fisher information that essentially controls how this thing performs. And the Fisher information is, essentially, a second derivative or a Hessian.

So if I'm in a one-dimensional parameter case, it's a number, it's a second derivative. If I'm in a multidimensional case, it's actually a Hessian, it's a matrix. So I'm going to actually take in notation little curly L of theta to be the log likelihood, OK? And that's the log likelihood for one observation. So let's call it x generically, but think of it as being x1, for example.

And I don't care of, like, summing, because I'm actually going to take expectation of this thing. So it's not going to be a data driven quantity I'm going to play with. So now I'm going to assume that it is twice differentiable, almost surely, because it's a random function. And so now I'm going to just sweep under the rug some technical conditions when these things hold.

So typically, when can I permute integral and derivatives and this kind of stuff that you don't want to think about? OK, the rule of thumb is it always works until it doesn't, in which case, that probably means you're actually solving some sort of calculus problem. Because in practice, it just doesn't happen.

So the Fisher information is the expectation of the-- that's called the outer product. So that's the product of this gradient and the gradient transpose. So that forms a matrix, right? That's a matrix minus the outer product of the expectations.

So that's really what's called the covariance matrix of this vector, nabla of L theta, which is a random vector. So I'm forming the covariance matrix of this thing. And the technical conditions tells me that, actually, this guy, which depends only on the Hessian, is actually equal to negative expectation of the-- sorry. It depends on the gradient. Is actually negative expectation of the Hessian. So I can actually get a quantity that depends on the second derivatives only using first derivatives.

But the expectation is going to play a role here. And the fact that it's a log. And lots of things actually show up here. And so in this case, what I get is that-- so in the one-dimensional case, then this is just the covariance matrix of a one-dimensional thing, which is just a variance of itself. So the variance of the derivative is actually equal to negative the expectation of the second derivative.

OK, so we'll see that next time. But what I wanted to emphasize with this is that why do we care about this quantity? That's called the Fisher information. Fisher is the founding father of modern statistics. Why do we give this quantity his name?

Well, it's because this quantity is actually very critical. What does the second derivative of a function tell me at the maximum? Well, it's telling me how curved it is, right? If I have a zero second derivative, I'm basically flat. And if I have a very high second derivative, I'm very curvy.

And when I'm very curvy, what it means is that I'm very robust to the estimation error. Remember our estimation strategy, which consisted in replacing expectation by averages? If I'm extremely curvy, I can move a little bit. This thing, the maximum, is not going to move much.

And this formula here-- so forget about the matrix version for a second-- is actually telling me exactly-- it's telling me the curvature is basically the variance of the first derivative. And so the more the first derivative fluctuates, the more your maximum is actually-- your org max is going to move all over the place. So this is really controlling how flat your likelihood, your log likelihood, is at its maximum.

The flatter it is, the more sensitive to fluctuation the arg max is going to be. The curvier it is, the less sensitive it is. And so what we're hoping-- a good model is going to be one that has a large or small value for the Fisher information. I want this to be-- small? I want it to be large.

Because this is the curvature, right? This number is negative, it's concave. So if I take a negative sign, it's going to be something that's positive. And the larger this thing, the more curvy it is. Oh, yeah, because it's the variance. Again, sorry. This is what-- OK.

Yeah, maybe I should not go into those details because I'm actually out of time. But just spoiler alert, the asymptotic variance of your-- the variance, basically, as n goes to infinity of the maximum likelihood estimator is going to be 1 over this guy. So we want it to be large, because the asymptotic variance is going to be very small.

All right, so we're out of time. We'll see that next week. And I have your homework with me. And I will actually turn it in. I will give it to you outside so we can let the other room come in. OK, I'll just leave you the--