**PHILIPPE RIGOLLET:** So welcome back. We're going to finish this chapter on maximum likelihood estimation. And last time, I briefly mentioned something that was called Fisher information. So Fisher information, in general, is actually a matrix when you have a multivariate parameter theta. So if theta, for example, is of dimension d, then the Fisher information matrix is going to be a d by d matrix.

You can see that, because it's the outer product. So it's of the form gradient gradient transpose. So if it's gradient gradient transpose, the gradient is the d dimensional. And so gradient times gradient transpose is a d by d matrix. And so this matrix actually contains-- well, tells you it's called Fisher information matrix. So it's basically telling you how much information about the theta is in your model.

So for example, if your model is very well-parameterized, then you will have a lot of information. You will have a higher-- so let's think of it as being a scalar number, just one number now-- so you're going to have a larger information about your parameter in the same probability distribution. But if start having a weird way to parameterize your model, then the Fisher information is actually going to drop.

So as a concrete example think of, for example, a parameter of interest in a Gaussian model, where the mean is known to be zero. But what you're interested in is the variance, sigma squared. If I'm interested in sigma square, I could parameterize my model by sigma, sigma squared, sigma to the fourth, sigma to 24th. I could parameterize it by whatever I want, then I would have a simple transformation. Then you could say that some of them are actually more or less informative, and you're going to have different values for your Fisher information.

So let's just review a few well-known computations. So I will focus primarily on the one dimensional case as usual. And I claim that there's two definitions. So if theta is a real valued parameter, then there's basically two definitions that you can think of for your Fisher information. One involves the first derivative of your log likelihood. And the second one involves the second derivative.

So the log likelihood here, we're actually going to define it as l of theta. And what is it? Well, it's simply the likelihood function for one observation. So it's l-- and I'm going to write 1 just to make sure that we all know what we're talking about one observation-- of-- which is the order again, I think it's X and theta. So that's the log likelihood, remember?

So for example, if I have a density, what is it going to be? It's going to be log of f sub theta of X. So this guy is a random variable, because it's a function of a random variable. And that's what you see expectations of this thing. It's a random function of theta. If I view this as a function of theta, the function becomes random, because it depends on this random X.

And so I of theta is actually defined as the variance of l prime of theta-- so the variance of the derivative of this function. And I also claim that it's equal to negative the expectation of the second derivative of theta. And here, the expectation and the variance are computed, because this function, remember, is random. So I need to tell you what is the distribution of the X with respect to which I'm computing the expectation and the variance. And it's the theta itself.

So typically, the theta we're going to be interested in-- so there's a Fisher information for all values of the parameter, but the one typically we're interested in is the true parameter, theta star. But view this as a function of theta right now. So now, I need to prove to you-- and this is not a trivial statement-- the variance of the derivative is equal to negative the expectation of the second derivative. I mean, there's really quite a bit that comes into this right.

And it comes from the fact that this is a log not of anything. It's the log of a density. So let's just prove that without having to bother too much ourselves with some technical assumptions. And the technical assumptions are the assumptions that allow me to permute derivative and integral.

Because when I compute the variances and expectations, I'm actually integrating against the density. And what I want to do is to make sure that I can always do that. So my technical assumptions are I can always permute integral and derivatives.

So let's just prove this. So what I'm going to do is I'm going to assume that X has density f theta. And I'm actually just going to write-- well, let me write it f theta right now. Let me try to not be lazy about writing. And so the thing I'm going to use is the fact that the integral of this density is equal to what? 1. And this is where I'm going to start doing weird things.

That means that if I take the derivative of this guy, it's equal to 0. So that means that if I look at the derivative with respect to theta of integral f theta of X dX, this is equal to 0. And this is where I'm actually making this switch, is that I'm going to say that this is actually equal to the integral of the derivative. So that's going to be the first thing I'm going to use.

And of course, if it's true for the first derivative, it's going to be true for the second derivative. So I'm going to actually do it a second time. And the second thing I'm going to use is the fact the integral of the second derivative is equal to 0.

So let's start from here. And let me start from, say, the expectation of the second derivative of l prime theta. So what is l prime prime theta? Well, it's the second derivative of log of f theta of X. And we know that the derivative of the log-- sorry-- yeah, so the derivative of the log is 1 over-- well, it's the derivative of f divided by f itself.

Everybody's with me? Just log of f prime is f prime over f. Here, it's just that f, I view this as a function of theta and not as a function of X.

So now, I need to take another derivative of this thing. So that's going to be equal to-- well, so we all know the formula for the derivative of the ratio. So I pick up the second derivative times f theta minus the first derivative squared divided by f theta squared-- basic calculus.

And now, I need to check that negative the expectation of this guy is giving me back what I want. Well what is negative the expectation of l prime prime of theta? Well, what we need to do is to do negative integral of this guy against f theta. So it's minus the integral of--

That's just the definition of the expectation. I take an integral against f theta. But here, I have something nice. What's happening is that those guys are canceling. And now that those guys are canceling, those guys are canceling too.

So what I have is that the first term-- I'm going to break this difference here. So I'm going to say that integral of this difference is the difference of the integrals. So the first term is going to be the integral of d over d theta squared of f theta. And the second one, the negative signs are going to cancel, and I'm going to be left with this.

Everybody's following? Anybody found the mistake? How about the other mistake? I don't know if there's a mistake. I'm just trying to get you to check what I'm doing. With me so far?

So this guy here is the integral of the second the derivative of f of X dX. What is this?

**AUDIENCE:** It's 0.

**PHILIPPE RIGOLLET:** It's 0. And that's because of this guy, which I will call frowny face. So frowny face tells me this. And let's call this guy monkey that hides his eyes. No, let's just do something simpler. Let's call it star. And this guy, we will use later on.

So now, I have to prove that this guy, which I have proved is equal to this, is now equal to the variance of l prime theta. So now, let's go back to the other way. We're going to meet halfway. I'm going to have a series-- I want to prove that this guy is equal to this guy. And I'm going to have a series of equalities that I'm going to meet halfway.

So let's start from the other end. We started from the negative l prime prime theta. Let's start with the variance part. Variance of l prime of theta, so that's the variance-- so that's the expectation of l prime of theta squared minus the square of the expectation of l prime of theta.

Now, what is the square of the expectation of l prime of theta? Well, l prime of theta is equal to the partial with respect to theta of log f theta of X, which we know from the first line over there-- that's what's in the bracket on the second line there-- is actually equal to the partial over theta of f theta X divided by f theta X. That's the derivative of the log.

So when I look at the expectation of this guy, I'm going to have the integral of this against f theta. And the f thetas are going to cancel again, just like I did here. So this thing is actually equal to the integral of partial over theta of f theta of X dX.

And what does this equal to? 0, by the monkey hiding is eyes. So that's star-- tells me that this is equal to 0. So basically, when I compute the variance, this term is not. Going to matter. I only have to complete the first one.

So what is the first one? Well, the first one is the expectation of l prime squared. And so that guy is the integral of-- well, what is l prime? Again, it's partial over partial theta f theta of X divided by f theta of X. Now, this time, this guy is squared against the density.

So one of the f thetas cancel. But now, I'm back to what I had before for this guy. So this guy is now equal to this guy. There's just the same formula. So they're the same thing.

And so I've moved both ways. Starting from the expression that involves the expectation of the second derivative I've come to this guy. And starting from the expression that tells me about the variance of the first derivative, I've come to the same guy. So that completes my proof. Are there any questions about the proof?

We also have on our way found an explicit formula for the Fisher information as well. So now that I have this thing, I could actually add that if X has a density, for example, this is also equal to the integral of-- well, the partial over theta of f theta of X squared divided by f theta of X, because I just proved that those two things were actually equal to the same thing, which was this guy.

Now in practice, this is really going to be the useful one. The other two are going to be useful depending on what case you're in. So if I ask you to compute the Fisher information, you have now three ways to pick from. And basically, practice will tell you which one to choose if you want to save five minutes when you're doing your computations.

Maybe you're the guy who likes to take derivatives. And then you're going to go with the second derivative one. Maybe you're the guy who likes to extend squares, so you're going to take the one that involves the square of the squared prime. And maybe you're just a normal person, and you want to use that guy.

Why do I care? This is the Fisher information. And I could have defined the [? Hilbert ?] information by taking the square root of this guy plus the sine of this thing and just be super happy and have my name in textbooks. But this thing has a very particular meaning.

When we're doing the maximum likelihood estimation-- so remember the maximum likelihood estimation is just an empirical version of trying to minimize the KL divergence. So what we're trying to do, maximum likelihood, is really trying to minimize the KL divergence. And we're trying to minimize this function, remember?

So now what we're going to do is we're going to plot this function. We said that, let's place ourselves in cases where this KL is convex, so that the inverse is concave. So it's going to look like this-- U-shaped, that's convex. So that's the truth thing I'm trying to minimize.

And what I said is that I'm going to actually try to estimate this guy. So in practice, I'm going to have something that looks like this, but it's not really this. And we're not going to do this, but you can show that you can control this uniformly over the entire space, that there is no space where it just becomes huge. In particular, this is not the space where it just becomes super huge, and the minimum of the dotted line becomes really far from this guy.

So if those two functions are close to each other, then this implies that the minimum here of the dotted line is close to the minimum of the solid line. So we know that this is theta star. And this is our MLE, estimator, theta hat ml. So that's basically the principle-- the more data we have, the closer the dotted line is to the solid line. And so the minimum is closer to the minimum.

But now, this is just one example, where I drew a picture for you. But there could be some really nasty examples. Think of this example, where I have a function, which is convex, but it looks more like this. That's convex, it's U-shaped. It's just a professional U.

Now, I'm going to put a dotted line around it that has pretty much the same fluctuations. The bend around it is of this size. So do we agree that the distance between the solid line and the dotted line is pretty much the same in those two pictures?

Now, here, depending on how I tilt this guy, basically, I can put the minimum theta star wherever I want. And let's say that here, I actually put it here. That's pretty much the minimum of this line.

And now, the minimum of the dotted line is this guy. So they're very far. The fact that I'm very flat at the bottom makes my requirements for being close to the U-shaped solid curve much more stringent, if I want to stay close.

And so this is the canonical case. This is the annoying case. And of course, you have the awesome case-- looks like this. And then whether you deviate, you can have something that moves pretty far. It doesn't matter, it's always going to stay close.

Now, what is the quantity that measures how curved I am at a given point-- how curved the function is at a given point? The secondary derivative. And so the Fisher information is negative the second derivative. Why the negative?

Well here-- Yeah, we're looking for a minimum, and this guy is really the-- you should view this as a reverted function. This is we're trying to maximize the likelihood, which is basically maximizing the negative KL. So the picture I'm showing you is trying to minimize the KL.

So the truth picture that you should see for this guy is the same, except that it's just flipped over. But the curvature is the same, whether I flip my sheet or not. So it's the same thing.

So apart from this negative sign, which is just coming from the fact that we're maximizing instead of minimizing, this is just telling me how curved my likelihood is around the maximum. And therefore, it's actually telling me how good, how robust my maximum likelihood estimator is. It's going to tell me how close, actually, my likelihood estimator is going to be-- maximum likelihood is going to be to the true parameter.

So I should be able to see that somewhere. There should be some statement that tells me that this Fisher information will play a role when assessing the precision of this estimator. And remember, how do we characterize a good estimator? Well, we look at its bias, or we look its variance. And we can combine the two and form the quadratic risk.

So essentially, what we're going to try to say is that one of those guys-- either the bias or the variance or the quadratic risk-- is going to be worse if my function is flatter, meaning that my Fisher information is smaller. And this is exactly the point of this last theorem.

So let's look at a couple of conditions. So this is your typical 1950s statistics paper that has like one page of assumptions. And this was like that in the early days, because people were trying to make theories that would be valid for as many models as possible. And now, people are sort of abusing this, and they're just making all this lists of assumptions so that their particular method works for their particular problem, because they just want to take shortcuts.

But really, the maximum likelihood estimator is basically as old as modern statistics. And so this was really necessary conditions. And we'll just parse that. The model is identified. Well, better be, because I'm trying to estimate theta and not P theta. So this one is good.

For all theta, the support of P theta does not depend on theta. So that's just something that we need to have. Otherwise, things become really messy. And in particular, I'm not going to be able to define likelihood-- Kullback-Leibler divergences.

Then why can I not do that? Well, because the Kullback-Leibler divergence has a log of the ratio of two densities. And if one of the support is changing with theta is it might be they have the log of something that's 0 or something that's not 0.

And the log of 0 is a slightly annoying quantity to play with. And so we're just removing that case. Nothing depends on theta-- think of it as being basically the entire real line as a support for Gaussian, for example.

Theta star is not on the boundary of theta. Can anybody tell me why this is important? We're talking about derivatives. So when I want to talk about derivatives, I'm talking about fluctuations around a certain point. And if I'm at the boundary, it's actually really annoying.

I might have the derivative-- remember, I give you this example-- where the maximum likelihood is just obtained at the boundary, because the function cannot grow anymore at the boundary. But it does not mean that the first order derivative is equal to 0. It does not mean anything. So all this picture here is valid only if I'm actually achieving the minimum inside.

Because if my theta space stops here and it's just this guy, I'm going to be here. And there's no questions about curvature or anything that comes into play. It's completely different. So here, it's inside. Again, think of theta as being the entire real line. Then everything is inside.

I is invertible. What does it mean for a positive number, a 1 by 1 matrix to be invertible? Yep.

**AUDIENCE:** It'd be equal to its [INAUDIBLE].

**PHILIPPE RIGOLLET:** A 1 by 1 matrix, that's a number, right? What is a characteristic-- if I give you a matrix with numbers and ask you if it's invertible, what are you going to do with it?

**AUDIENCE:** Check if the determinant is 0.

**PHILIPPE RIGOLLET:** Check if the determinant is 0. What is the determinant of the 1 by 1 matrix? It's just the number itself. So that's basically, you want to check if this number is 0 or not.

So we're going to think in the one dimensional case here. And in the one dimensional case, that just means that the curvature is not 0. Well, it better be not 0, because then I'm going to have no guarantees. If I'm totally flat, if I have no curvature, I'm basically totally flat at the bottom. And then I'm going to get nasty things.

Now, this is not true. I could have the curvature which grows like-- so here, it's basically-- the second derivative is telling me-- if I do the Taylor expansion, it's telling me how I grow as a function of, say, x squared. It's the quadratic term that I'm controlling. It could be that this guy is 0, and then the term of order, x to the fourth, is picking up. That could be the first one that's non-zero.

But that would mean that my rate of convergence would not be square root of n. When I'm actually playing central limit theorem, it would become n to the 1/4th. And if I have all a bunch of 0 until the 16th order, I would have n to the 1/16th, because that's really telling me how flat I am. So we're going to focus on the case where it's only quadratic terms, and the rates of the central limit theorems kick in.

And then a few other technical conditions-- we just used a couple of them. So I permuted limit and integral. And you can check that really what you want is that the integral of a derivative is equal to 0. Well, it just means that the values at the two ends are actually the same. So those are slightly different things.

So now, what we have is that the maximum likelihood estimator has the following two properties. The first one, if I were to say that in words, what would I say, that theta hat is-- Is what? Yeah, that's what I would say when I-- that's for mathematicians.

But if I'm a statistician, what am I going to say? It's consistent. It's a consistent estimator of theta star. It converges in probability to theta star.

And then we have this sort of central limit theorem statement. The central limit theorem statement tells me that if this was an average and I remove the expectation of the average-- let's say it's 0, for example-- then square root of n times the average blah goes through some normal distribution. This is telling me that this is actually true, even if theta hat has nothing to do with an average. That's remarkable. Theta hat might not even have a closed form, and I'm still having basically the same properties as an average that would be given to me by a central limit theorem.

And what is the asymptotic variance? So that's the variance in the n. So here, I'm thinking of having those guys being multivariate. And so I have the inverse of the covariance matrix that shows up as the variance-covariance matrix asymptotically. But if you think of just being a one dimensional parameter, it's one over this Fisher information, one over the curvature.

So the curvature is really flat, the variance becomes really big. If the function is really flat, curvature is low, variance is big. If the curvature is very high, the variance becomes very low. And so that illustrates everything that's happening with the pictures that we have.

And if you look, what's amazing here, there is no square root 2 pi, there's no fudge factors going on here. This is the asymptotic variance, nothing else. It's all in there, all in the curvature. Are there any questions about this?

So you can see here that theta star is the true parameter. And the information matrix is evaluated at theta star. That's the point that matters. When I drew this picture, the point that was at the very bottom was always theta star. It's the one that minimizes the KL divergence, am as long as I'm identified. Yes.

**AUDIENCE:** So the higher the curvature, the higher the inverse of the Fisher information?

**PHILIPPE RIGOLLET:** No, the higher the Fisher information itself. So the inverse is going to be smaller. So small variance is good.

So now what it means, actually, if I look at what is the quadratic risk of this guy, asymptotically-- what is asymptotic quadratic risk? Well, it's 0 actually. But if I assume that this thing is true, that this thing is pretty much Gaussian, if I look at the quadratic risk, well, it's the expectation of the square of this thing.

And so it's going to scale like the variance divided by n. The bias goes to 0, just by this. And then the quadratic risk is going to scale like one over Fisher information divided by n.

So here, the-- I'm not mentioning the constants. There must be constants, because everything is asymptotic. So for each finite n, I'm going to have some constants that show up. Everybody just got their mind blown by this amazing theorem?

So I mean, if you think about it, the MLE can be anything. I'm sorry to say to you, in many instances, the MLE is just going to be an average, which is just going to be slightly annoying. But there are some cases where it's not. And we have to resort to this theorem rather than actually resorting to the central limit theorem to prove this thing.

And more importantly, even if this was an average, you don't have to even know how to compute the covariance matrix-- sorry, the variance of this thing to plug it into the central limit theorem. I'm telling you, it's actually given by the Fisher information matrix. So if it's an average, between you and me, you probably want to go the central limit theorem route if you want to prove this kind of stuff. But if it's not, then that's your best shot. But you have to check those conditions. I will give you for granted the 0.5.

Ready? Any questions? We're going to wrap up this chapter four. So if you have questions, that's the time. Yes.

**AUDIENCE:** What was the quadratic risk up there?

**PHILIPPE RIGOLLET:** You mean the definition?

**AUDIENCE:** No, the-- what is was for this.

**PHILIPPE RIGOLLET:** Well, you see the quadratic risk, if I think of it as being one dimensional, the quadratic risk is the expectation of the square of the difference between theta hat and theta star. So that means that if I think of this as having a normal 0, 1, that's basically computing the expectation of the square of this Gaussian divided by n. I just divided by square root of n on both sides.

So it's the expectation of the square of this Gaussian. The Gaussian is mean 0, so the expectation of the square is just a variance. And so I'm left with 1 over the Fisher information divided by n.

**AUDIENCE:** I see. OK.

**PHILIPPE RIGOLLET:** So let's move on to chapter four. And this is the method of moments. So the method of moments is actually maybe a bit older than maximum likelihood. And maximum likelihood is dated, say, early 20th century, I mean as a systematic thing, because as I said, many of those guys are going to be averages. So finding an average is probably a little older.

The method of moments, there's some really nice uses. There's a paper by Pearson in 1904, I believe, or maybe 1894. I don't know. And this paper, he was actually studying some species of crab in an island, and he was trying to make some measurements. That's how he came up with this model of mixtures of Gaussians, because there was actually two different populations in this populations of crab.

And the way he actually fitted the parameters was by doing the method of moments, except that since there were a lot of parameters, he actually had to basically solve six equations with six unknowns. And that was a complete nightmare. And the guy did it by hand. And we don't know how he did it actually. But that is pretty impressive.

So I want to start-- and this first part is a little brutal. But this is a Course 18 class, and I do not want to give you-- So let's all agree that this course might be slightly more challenging than AP statistics. And that means that it's going to be challenging just during class. I'm not going to ask you about the Weierstrass Approximation Theorem during the exams. But what I want is to give you mathematical motivations for what we're doing.

And I can promise you that maybe you will have a slightly higher body temperature during the lecture, but you will come out smarter of this class. And I'm trying to motivate to you for using mathematical tool and show you where interesting mathematical things that you might find dry elsewhere actually work very beautifully in the stats literature. And one that we saw was using Kullback-Leibler divergence out of motivation for maximum likelihood estimation, for example.

So the Weierstrass Approximation Theorem is something that comes from pure analysis. So maybe-- I mean, it took me a while before I saw that. And essentially, what it's telling you is that if you look at a function that is continuous on an interval a, b-- on a segment a, b-- then you can actually approximate it uniformly well by polynomials as long as you're willing to take the degree of this polynomials large enough.

So the formal statement is, for any epsilon, there exists the d that depends on epsilon in a1 to ad-- so if you insist on having an accuracy which is 1/10,000, maybe you're going to need a polynomial of degree 100,000, who knows. It doesn't tell you anything about this. But it's telling you that at least you have only a finite number of parameters to approximate those functions that typically require an infinite number of parameters to be described.

So that's actually quite nice. And that's the basis for many things and many polynomial methods typically. And so here, it's uniform, so there's this max over x that shows up that's actually nice as well. That's Weierstrass Approximation Theorem.

Why is that useful to us? Well, in statistics, I have a sample of X1 to Xn. I have, say, a unified statistical model. I'm not always going to remind you that it's identified-- not unified-- identified statistical model. And I'm going to assume that it has a density. You could think of it as having a PMF, but think of it as having a density for one second.

Now, what I want is to find the distribution. I want to find theta. And finding theta, since it's identified as equivalent to finding P theta, which is equivalent to finding f theta, and knowing a function is the same-- knowing a density is the same as knowing a density against any test function h. So that means that if I want to make sure I know a density-- if I want to check if two densities are the same, all I have to do is to compute their integral against all bounded continuous functions.

You already know that it would be true if I checked for all functions h. But since f is a density, I can actually look only at functions h that are bounded, say, between minus 1 and 1, and that are continuous. That's enough. Agreed? Well, just trust me on this. Yes, you have a question?

**AUDIENCE:** Why is this-- like, why shouldn't you just say that [INAUDIBLE]?

**PHILIPPE RIGOLLET:** Yeah, I can do that. I'm just finding a characterization that's going to be useful for me later on. I can find a bunch of them. But here, this one is going to be useful.

So all I need to say is that f theta star integrated against X, h of x-- so this implies that f-- if theta is equal to f theta star not everywhere, but almost everywhere. And that's only true if I guarantee to you that f theta and f theta stars are densities. This is not true for any function.

So now, that means that, well, if I wanted to estimate theta hat, all I would have to do is to compute the average, right-- so this guy here, the integral-- let me clean up a bit my board. So my goal is to find theta such that, if I look at f theta and now I integrate it against h of x, then this gives me the same thing as if I were to do it against f theta star. And I want this for any h, which is continuous and bounded.

So of course, I don't know what this quantity is. It depends on my unknown theta star. But I have theta from this. And I'm going to do the usual-- the good old statistical trick, which is, well, this I can write as the expectation with respect to P theta star of h theta of x. That's just the integral of a function against something.

And so what I can do is say, well, now I don't know this guy. But my good old trick from the book is replace expectations by averages. And what I get-- And that's approximately by the law of large numbers. So if I can actually find a function f theta such that when I integrate it against h it gives me pretty much the average of the evaluations of h over my data points for all h, then that should be a good candidate.

The problem is that's a lot of functions to try. Even if we reduced that from all possible functions to bounded and continuous ones, that's still a pretty large infinite number of them. And so what we can do is to use our Weierstrass Approximation Theorem. And it says, well, maybe I don't need to test it against all h. Maybe the polynomials are enough for me.

So what I'm going to do is I'm going to look only at functions h that are of the form sum of ak-- so h of x is sum of ak X to the k-th for k equals 0 to d-- only polynomials of degree d. So when I look at the average of my h's, I'm going to get a term like the first one. So the first one here, this guy, becomes 1/n sum from i equal 1 to n sum from k equal 0 to d of ak Xi to the k-th. That's just the average of the values of h of Xi.

And now, what I need to do is to check that it's the same thing when I integrate h of this form as well. I want this to hold for all polynomials of degree d. That's still a lot of them. There's still an infinite number of polynomials, because there's an infinite number of numbers a0 to ad that describe those polynomials. But since those guys are polynomials, it's actually enough for me to look only at the terms of the form X to the k-th-- no linear combination, no nothing.

So actually, it's enough to look only at h of x, which is equal to X to the k-th for k equal 0 to d. And now, how many of those guys are there? Just d plus 1, 0 to d. So that's actually a much easier thing for me to solve.

Now, this quantity, which is the integral of f theta against X to the k-th-- so that the expectation of X to the k-th here-- it's called moment of order k, or k-th moment of P theta. That's a moment. A moment is just the expectation of the power. The mean is which moment? The first moment.

And variance is not exactly the second moment. It's the second moment minus the first moment squared. That's the variance. It's E of X squared minus E of X squared.

So those are things that you already know. And then you can go higher. You can go to E of X cube, E of X blah, blah. Here, I say go to E of X to the d-th.

Now, as you can see, this is not something you can really put in action right now, because the Weierstrass Approximation Theorem does not tell you what d should be. Actually, we totally lost track of the epsilon I was even looking for. I just said approximately equal, approximately equal.

And so all this thing is really just motivation. But it's essentially telling you that if you go to d large enough, technically you should be able to identify exactly your distribution up to epsilon. So I should be pretty good, if I go to d large enough. Now in practice, actually there should be much less than arbitrarily large d. Typically, we are going to need d which is 1 or 2.

So there are some limitations to the Weierstrass Approximation Theorem. And there's a few. The first one is that it only works for continuous functions, which is not so much of a problem. That can be fixed. Well, we need bounded continuous functions.

It works only on intervals. That's annoying, because we're going to have random variables that are defined beyond intervals. So we need something that just goes beyond the intervals. And you can imagine that if you let your functions be huge, it's going to be very hard for you to have a polynomial approximately [INAUDIBLE] well. Things are going to start going up and down at the boundary, and it's going to be very hard.

And again, as I said several times, it doesn't tell us what d should be. And as statisticians, we're looking for methods, not like principles of existence of a method that exists. So if E is discrete, I can actually get a handle on this d.

If E is discrete and actually finite-- I'm going to actually look at a finite E, meaning that I have a PMF on, say, r possible values, x1 and xr. My random variable, capital X, can take only r possible values. Let's think of them as being the integer numbers 1 to r. That's the number of success out of r trials that I get, for example. Binomial rp, that's exactly something like this.

So now, clearly this entire distribution is defined by the PMF, which gives me exactly r numbers. So it can completely describe this distribution with r numbers. The question is, do I have an enormous amount of redundancy if I try to describe this distribution using moments? It might be that I need-- say, r is equal to 10, maybe I have only 10 numbers to describe this thing, but I actually need to compute moments up to the order of 100 before I actually recover entirely the distribution.

Maybe I need to go infinite. Maybe the Weierstrass Theorem is the only thing that actually saves me here. And I just cannot recover it exactly. I can go to epsilon if I'm willing to go to higher and higher polynomials. Oh, by the way, in the Weierstrass Approximation Theorem, I can promise you that as epsilon goes to 0, d goes to infinity.

So now, really I don't even have actually r parameters. I have only r minus parameter, because the last one-- because they sum up to 1. So the last one I can always get by doing 1 minus the sum of the first r minus 1 first. Agreed? So each distribution r numbers is described by r minus 1 parameters. The question is, can I use only r minus moments to describe this guy?

This is something called Gaussian quadrature. The Gaussian quadrature tells you, yes, moments are actually a good way to reparametrize your distribution in the sense that if I give you the moments or if I give you the probability mass function, I'm basically giving you exactly the same information. You can recover all the probabilities from there.

So here, I'm going to denote by-- I'm going to drop the notation in theta. I don't have theta. Here, I'm talking about any generic distribution. And so I'm going to call mk the k-th moment.

And I have a PMF, this is really the sum for j equals 1 to r of xj to the k-th times p of xj. And this is the PMF. So that's my k-th moment. So the k-th moment is a linear combination of the numbers that I am interested in.

So that's one equation. And I have as many equations as I'm actually willing to look at moments. So if I'm looking at 25 moments, I have 25 equations. m1 equals blah with this to the power of 1, m2 equals blah with this to the power of 2, et cetera.

And then I also have the equation that 1 is equal to the sum of the p of xj. That's just the definition of PMF. So this is r's. They're ugly, but those are r's.

So now, this is a system of linear equations in p, and I can actually write it in its canonical form, which is of the form a matrix of those guys times my parameters of interest is equal to a right hand side. The right hand side is the moments. That means, if I did you the moments, can you come back and find what the PMF, because we know already from probability that the PMF is all I need to know to fully describe my distribution. Given the moments, that's unclear.

Now, here, I'm going to actually take exactly r minus 1 moment and this extra condition that the sum of those guys should be 1. So that gives me r equations based on r minus 1 moments. And how many unknowns do I have? Well, I have my r unknown parameters for the PMF, the r values of the PMF.

Now, of course, this is going to play a huge role in whether the are many p's that give me the same. The goal is to find if there are several p's that can give me the same moments. But if there's only one p that can give me a set of moment, that means that I have a one-to-one correspondence between PMF and moments. And so if you give me the moments, I can just go back to the PMF.

Now, how do I go back? Well, by inverting this matrix. If I multiply this matrix by its inverse, I'm going to get the identity times the vector of p's equal the inverse of the matrix times the m's. So what we want to do is to say that p is equal to the inverse of this big matrix times the moments that you give me. And if I can actually talk about the inverse, then I have basically a one-to-one mapping between the m's, the moments, and the matrix.

So what I need to show is that this matrix is invertible. And we just decided that the way to check if a matrix is invertible is by computing its determinant. Who has computed a determinant before? Who was supposed to compute a determinant at least than just to say, no, you know how to do it.

So you know how to compute determinants. And if you've seen any determinant in class, there's one that shows up in the exercises that professors love. And it's called the Vandermonde determinant. And it's the determinant of a matrix that have a very specific form.

It looks like-- so there's basically only r parameters to this r by r matrix. The first row, or the first column-- sometimes, it's presented like that-- is this vector where each entry is to the power of 1. And the second one is each entry is to the power of 2, and to the power of 3, and to the power 4, et cetera.

So that's exactly what we have-- x1 to the first, x2 to the first, all the way to xr to the first, and then same thing to the power of 2, all the way to the last one. And I also need to add the row of all 1's, which you can think of those guys are to the power of 0, if you want. So I should really put it on top, if I wanted to have a nice ordering.

So that was the matrix that I had. And I'm not asking you to check it. You can prove that by induction actually, typically by doing the usual let's eliminate some rows and columns type of tricks that you do for matrices.

So you basically start from the whole matrix. And then you move onto a matrix that has only one 1's and then 0's here. And then you have Vandermonde that's just slightly smaller. And then you just iterate. Yeah.

**AUDIENCE:** I feel like there's a loss to either the supra index, or the sub index should have a k somewhere [INAUDIBLE]. [INAUDIBLE] the one I'm talking about?

**PHILIPPE RIGOLLET:** Yeah, I know, but I don't think the answer to your question is yes. So k is the general index, right? So there's no k. k does not exist. k just is here for me to tell me for k equals 1 to r.

So this is an r by r matrix. And so there is no k there. So if you wanted the generic term, if I wanted to put 1 in the middle on the j-th row and k-th column, that would be x-- so j-th row would be x sub k to the power of j. That would be the--

And so now, this is basically the sum-- well, that should not be strictly-- So that would be for j and k between 1 and r. So this is the formula that get when you try to expand this Vandermonde determinant. You have to do it only once when you're a sophomore typically. And then you can just go on Wikipedia to do it.

That's what I did. I actually made a mistake copying it. The first one should be 1 less than or equal to j. And the last one should be k less than or equal to r. And now what you have is the product of the differences of xj and xk.

And for this thing to be non-zero, you need all the terms to be non-zero. And for all the terms to be non-zero, you need to have no xi, xj, and no xj, xk that are identical. If all those are different numbers, then this product is going to be different from 0.

And those are different numbers, because those are r possible values that your random verbal takes. You're not going to say that it takes two with probability 1.5-- sorry, two with probability 0.5 and two with probability 0.25. You're going to say it takes two with probability 0.75 directly.

So those xj's are different. These are the different values that your random variable can take. Remember, xj, xk was just the different values x1 to xr-- sorry-- was the different values that your random variable can take. Nobody in their right mind would write twice the same value in this list.

So my Vandermonde is non-zero. So I can invert it. And I have a one-to-one correspondence between my entire PMF and the first r minus 1's moments to which I append the number 1, which is really the moment of order 0 again. It's E of X to the 0-th, which is 1.

So good news, I only need r minus 1 parameters to describe r minus 1 parameters. And I can choose either the values of my PMF. Or I can choose the r minus 1 first moments.

So the moments tell me something. Here, it tells me that if I have a discrete distribution with r possible values, I only need to compute r minus 1 moments. So this is better than Weierstrass Approximation Theorem. This tells me exactly how many moments I need to consider.

And this is for any distribution. This is not a distribution that's parametrized by one parameter, like the Poisson or the binomial or all this stuff. This is for any distribution under a finite number. So hopefully, if I reduce the family of PMFs that I'm looking at to a one-parameter family, I'm actually going to need to compute much less than r minus 1 values.

But this is actually hopeful. It tells you that the method of moments is going to work for any distribution. You just have to invert a Vandermonde matrix.

So just the conclusion-- the statistical conclusion-- is that moments contain important information about the PMF and the PDF. If we can estimate these moments accurately, we can solve for the parameters of the distribution and recover the distribution. And in a parametric setting, where knowing P theta amounts to knowing theta, which is identifiability-- this is not innocuous-- it is often the case that even less moments are needed.

After all, if theta is a one dimensional parameter, I have one parameter to estimate. Why would I go and get 25 moments to get this one parameter. Typically, there is actually-- we will see that the method of moments just says if you have a d dimensional parameter, just compute d moments, and that's it.

But this is only on a case-by-case basis. I mean, maybe your model will totally screw up its parameters and you actually need to get them. I mean, think about it, if the function is parameterized just by its 27th moment-- like, that's the only thing that matters in this distribution, I just describe the function, it's just a density, and the only thing that can change from one distribution to another is this 27th moment-- well, then you're going to have to go get the 27th moment. And that probably means that your modeling step was actually pretty bad. So the rule of thumb, if theta is in Rd, we need d moments.

So what is the method of moments? That's just a good old trick. Replace the expectation by averages. That's the beauty. The moments are expectations. So let's just replace the expectations by averages and then do it with the average version, as if it was the true one.

So for example, I'm going to talk about population moments, when I'm computing them with the true distribution, and I'm going to talk about them empirical moments, when I talk about averages. So those are the two quantities that I have. And now, what I hope is that there is. So this is basically-- everything is here. That's where all the money is.

I'm going to assume there's a function psi that maps my parameters-- let's say they're in Rd-- to the set of the first d moments. Well, what I want to do is to come from this guy back to theta. So it better be that this function is-- invertible. I want this function to be invertible.

In the Vandermonde case, this function with just a linear function-- multiply a matrix by theta. Then inverting a linear function is inverting the matrix. Then this is the same thing. So now what I'm going to assume-- and that's key for this method to work-- is that this theta-- so this function psi is one to one. There's only one theta that gets only one set of moments.

And so if it's one to one, I can talk about its inverse. And so now, I'm going to be able to define theta as the inverse of the moments-- the reciprocal of the moments. And so now, what I get is that the moment estimator is just the thing where rather than taking the true guys in there, I'm actually going to take the empirical moments in there.

Before we go any further, I'd like to just go back and tell you that this is not completely free. How well-behaved your function psi is going to play a huge role. Can somebody tell me what the typical distance-- if I have a sample of size n, what is the typical distance between an average and the expectation? What is the typical distance? What is the order of magnitude as a function of n between xn bar and its expectation.

**AUDIENCE:** 1 over square root of n.

**PHILIPPE RIGOLLET:** 1 over square root n. That's what the central limit theorem tells us, right? The central limit theorem tells us that those things are basically a Gaussian, which is of order of 1 divided by its square of n.

And so basically, I start with something which is 1 over square root of n away from the true thing. Now, if my function psi inverse is super steep like this-- that's psi inverse-- then just small fluctuations, even if they're of order 1 square root of n, can translate into giant fluctuations in the y-axis. And that's going to be controlled by how steep psi inverse is, which is the same as saying how flat psi is-- how flat is psi.

So if you go back to this Vandermonde inverse, what it's telling you is that if this inverse matrix blows up this guy a lot-- so if I start from a small fluctuation of this thing and then they're blowing up by applying the inverse of this matrix, things are not going to go well. Anybody knows what is the number that I should be looking for? So that's from, say, numerical linear algebra numerical methods. When I have a system of linear equations, what is the actual number I should be looking at to know how much I'm blowing up the fluctuations? Yeah.

**AUDIENCE:** Condition number?

**PHILIPPE RIGOLLET:** The condition number, right. So what's important here is the condition number of this matrix. If the condition number of this matrix is small, then it's good. It's not going to blow up much. But if the condition number is very large, it's just going to blow up a lot.

And the condition number is the ratio of the largest and the smallest eigenvalues. So you'll have to know what it is. But this is how all these things get together. So the numerical stability translates into statistical stability here.

And numerical means just if I had errors in measuring the right hand side, how much would they translate into errors on the left hand side. So the error here is intrinsic to statistical questions. So that's my estimator, provided that it exists. And I said it's a one to one, so it should exist, if I assume that psi is invertible.

So how good is this guy? That's going to be definitely our question-- how good is this thing. And as I said, there's chances that if psi is really steep, then it should be not very good-- if psi inverse is very steep, it should not be very good, which means that it's-- well, let's just leave it to that.

So that means that I should probably see the derivative of psi showing up somewhere. If the derivative of psi inverse, say, is very large, then I should actually have a larger variance in my estimator. So hopefully, just like we had a theorem that told us that the Fisher information was key in the variance of the maximum likelihood estimator, we should have a theorem that tells us that the derivative of psi inverse is going to have a key role in the method of moments. So let's do it.

So I'm going to talk to you about matrices. So now, I have-- So since I have to manipulate d numbers at any given time, I'm just going to concatenate them into a vector.

So I'm going to call capital M theta-- so that's basically the population moment. And I have M hat, which is just m hat 1 to m hat d. And that's my empirical moment.

And what's going to play a role is what is the variance-covariance of the random vector. So I have this vector 1-- do I have 1? No, I don't have 1. So that's a d dimensional vector.

And here, I take the successive powers. Remember, that looks very much like a column of my Vandermonde matrix. So now, I have this random vector. It's just the successive powers of some random variable X.

And the variance-covariance matrix is the expectation-- so sigma-- of theta. The theta just means I'm going to take expectations with respect to theta. That's the expectation with respect to theta of this guy times this guy transpose minus the same thing but with the expectation inside. Why do I do X, X1. I have X, X2, X3. X, X2, Xd times the expectation of X, X2, Xd.

Everybody sees what this is? So this is a matrix where if I look at the ij-th term of this matrix-- or let's say, jk-th term, so on row j and column k, I have sigma jk of theta. And it's simply the expectation of X to the j plus k-- well, Xj Xk minus expectation of Xj expectation of Xk. So I can write this as m j plus k of theta minus mj of theta times mk of theta.

So that's my covariance matrix of this particular vector that I define. And now, I'm going to assume that psi inverse-- well, if I want to talk about the slope in an analytic fashion, I have to assume that psi is differentiable. And I will talk about the gradient of psi, which is, if it's one dimensional, it's just the derivative.

And here, that's where notation becomes annoying. And I'm going to actually just assume that so now I have a vector. But it's a vector of functions and I want to compute those functions at a particular value. And the value I'm actually interested in is at the m of theta parameter.

So psi inverse goes from the set of moments to the set of parameters. So when I look at the gradient of this guy, it should be a function that takes as inputs moments. And where do I want this function to be evaluated at? At the true moment-- at the population moment vector. Just like when I computed my Fisher information, I was computing it at the true parameter.

So now, once they compute this guy-- so now, why is this a d by d gradient matrix? So I have a gradient vector when I have a function from rd to r. This is the partial derivatives.

But now, I have a function from rd to rd. So I have to take the derivative with respect to the arrival coordinate and the departure coordinate. And so that's the gradient matrix.

And now, I have the following properties. The first one is that the law of large numbers tells me that theta hat is a weakly or strongly consistent estimator. So either I use the strong law of large numbers or the weak law of large numbers, and I get strong or weak consistency.

So what does that mean? Why is that true? Well, because now so I really have the function-- so what is my estimator? Theta hat this psi inverse of m hat 1 to m hat k.

Now, by the law of large numbers, let's look only at the weak one. Law of large numbers tells me that each of the mj hat is going to converge in probability as n to infinity to the-- so the empirical moments converge to the population moments. That's what the good old trick is using, the fact that the empirical moments are close to the true moments as n becomes larger. And that's because, well, just because the m hat j's are averages, and the law of large numbers works for averages.

So now, plus if I look at my continuous mapping theorem, then I have that psi inverse is continuously differentiable. So it's definitely continuous. And so what I have is that psi inverse of m hat 1 m hat d converges to psi inverse m1 to md, which is equal to of theta star. So that's theta star. By definition, we assumed that that was the unique one that was actually doing this.

Again, this is a very strong assumption. I mean, it's basically saying, if the method of moment works, it works. So the fact that psi inverse one to one is really the key to making this guy work.

And then I also have a central limit theorem. And the central limit theorem is basically telling me that M hat is converging to M even in the multivariate sense. So if I look at the vector of M hat and the true vector of M, then I actually make them go-- I look at the difference for scale by square root of n.

It goes to some Gaussian. And usually, we would see-- if it was one dimensional, we would see the variance. Then we see the variance-covariance matrix.

Who has never seen the-- well, nobody answers this question. Who has already seen the multivariate central limit theorem? Who was never seen the multivariate central limit theorem?

So the multivariate central limit theorem is basically just the slight extension of the univariate one. It just says that if I want to think-- so the univariate one would tell me something like this-- and 0. And then I would have basically the variance of X to the j-th. So that's what the central limit theorem tells me.

This is an average. So this is just for averages. The central limit theorem tells me this. Just think of X to the j-th as being y. And that would be true. Everybody agrees with me?

So now, this is actually telling me what's happening for all these guys individually. But what happens when those guys start to correlate together? I'd like to know if they actually correlate the same way asymptotically.

And so if I actually looked at the covariance matrix of this vector-- so now, I need to look at a matrix which is d by d-- then would those univariate central limit theorems tell me-- so let me right like this, double bar. So that's just the covariance matrix. This notation, V double bar is the variance-covariance matrix.

So what this thing tells me-- so I know this thing is a matrix, d by d. Those univariate central limit theorems only give me information about the diagonal terms. But here, I have no idea where the covariance matrix is. This guy is telling me, for example, that this thing is like variance of X to the j-th.

But what if I want to find off-diagonal elements of this matrix? Well, I need to use a multivariate central limit theorem. And really what it's telling me is that you can actually replace this guy here-- so that goes in distribution to some normal mean 0, again. And now, what I have is just sigma of theta, which is just the covariance matrix of this vector X, X2, X3, X4, all the way to Xd.

And that's it. So that's a multivariate Gaussian. Who has never seen a multivariate Gaussian? Please, just go on Wikipedia or something. There's not much to know about it.

But I don't have time to redo probability here. So we're going to have to live with it. Now, to be fair, if your goal is not to become a statistical savant, we will stick to univariate questions in the scope of homework and exams.

So now, what was the delta method telling me? It was telling me that if I had a central limit theorem that told me that theta hat was going to theta, or square root of n theta hat minus theta was going to some Gaussian, then I could look at square root of Mg of theta hat minus g of theta. And this thing was also going to a Gaussian. But what it had to be is the square of the derivative of g in the variance.

So the delta method, it was just a way to go from square root of n theta hat n minus theta goes to some N, say 0, sigma squared, to-- so delta method was telling me that this was square root Ng of theta hat N minus g of theta was going in distribution to N0 sigma squared g prime squared of theta. That was the delta method.

Now, here, we have a function of those guys. The central limit theorem, even the multivariate one, is only guaranteeing something for me regarding the moments. But now, I need to map the moments back into some theta, so I have a function of the moments.

And there is something called the multivariate delta method, where derivatives are replaced by gradients. Like, they always are in multivariate calculus. And rather than multiplying, since things do not compute, rather than choosing which side I want to put the square, I'm actually just going to take half of the square on one side and the other half of the square on the other side.

So the way you should view this, you should think of sigma squared times g prime squared as being g prime of theta times sigma squared times g prime of theta. And now, this is completely symmetric. And the multivariate delta method is basically telling you that you get the gradient here.

So you start from something that's like that over there, a sigma-- so that's my sigma squared, think of sigma squared. And then I premultiply by the gradient and postmultiply by the gradient. The first one is transposed. The second one is not.

But that's very straightforward extension. You don't even have to understand it. Just think of what would be the natural generalization.

Here, by the way, I wrote explicitly what the gradient of a multivariate function is. So that's a function that goes from Rd to Rk. So now, the gradient is a d by k matrix.

And so now, for this guy, we can do it for the method or moments. And we can see that basically we're going to have this scaling that depends on the gradient of the reciprocal of psi, which is normal. Because if psi is super steep, if psi inverse is super steep, then the gradient is going to be huge, which is going to translate into having a huge variance for the method of moments.

So this is actually the end. I would like to encourage you-- and we'll probably do it on Thursday just to start. But I encourage you do it in one dimension, so that you know how to use the method of moments, you know how to do a bunch of things. Do it in one dimension and see how you can check those things.

So just as a quick comparison, in terms of the quadratic risk, the maximum likelihood estimator is typically more accurate than the method of moments. What is pretty good to do is, when you have a non-concave likelihood function, what people like to do is to start with the method of moments as an initialization and then run some algorithm that optimizes locally the likelihood starting from this point, because it's actually likely to be closer. And then the MLE is going to improve it a little bit by pushing the likelihood a little better.

So of course, the maximum likelihood is sometimes intractable. Whereas, computing moments is fairly doable. If the likelihood is concave, as I said, we can use optimization algorithms, such as interior-point methods or gradient descent, I guess, to maximize it.

And if the likelihood is non-concave, we only have local heuristics. Risk And that's what I meant-- you have only local maxima. And one trick you can do-- so your likelihood looks like this, and it might be the case that if you have a lot of those peaks, you basically have to start your algorithm in each of those peaks. But the method of moments can actually start you in the right peak, and then you just move up by doing some local algorithm for maximum likelihood.

So that's not key. But that's just if you want to think about algorithmically how I would end up doing this and how can I combine the two. So I'll see you on Thursday. Thank you.