PHILIPPE RIGOLLET: We're talking about generalized linear models. And in generalized linear models, we generalize linear models in two ways. The first one is to allow for a different distribution for the response variables. And the distributions that we wanted was the exponential family.

And this is a family that can be generalized over random variables that are defined on c or q in general, with parameters rk. But we're going to focus in a very specific case when y is a real valued response variable, which is the one you're used to when you're doing linear regression. And the parameter theta also lives in r.

And so we're going to talk about the canonical case. So that's the canonical exponential family, where you have a density, theta of x, which is of the form, exponential plus. And then, we have y, which interacts with theta only by taking a product. Then, there's a term that depends only on theta, some dispersion parameter phi. And then, we have some normalization factor. Let's call it c of y phi. So it really should not matter too much, so it's c of y phi, and that's really just the normal position factor. And here, we're going to assume that phi is known.

I have no idea what I write. I don't know if you guys can read. I don't know what chalk has been used today, but I just can't see it. That's not my fault. All right, so we're going to assume that phi is known. And so we saw that several distributions that we know well, including the Gaussian for example, belong to this family. And there's other ones, such as Poisson-- Poisson and Bernoulli. So if the PMF has this form, if you have a discrete random variable, this is also valid.

And the reason why we introduced this family is because there are going to be some properties that we know that this thing here, this function, b of theta, is essentially what completely characterizes your distribution. So if phi is fixed, we know that the interaction is the form. And this really just comes from the fact that we want the function to integrate to one. So this b here in the canonical form encodes everything we want to know. If I tell you what b of theta is-- and of course, I tell you what phi is, but let's say for a second that phi is equal to one. If I tell you this b of theta, you know exactly what distribution I'm talking about. So it should encode everything that's specific to this distribution, such as mean, variance, all the moments that you would want. And we'll see how we can compute from this thing the mean and the variance, for example.

So today, we're going to talk about likelihood, and we're going to start with the likelihood function or the log likelihood for one observation. From this, we're going to do some computations, and then, we'll move on to the actual log likelihood based on n independent observations. And here, as we will see, the observations are not going to be identically distributed, because we're going to want each of them, conditionally on x to be a different function of x, where theta is just a different function of x for each of the observation.

So remember, the log likelihood-- and this is for one observation-- is just the log of the density, right? And we have this identity that I mentioned at the end of the class on Tuesday. And this identity is just that the expectation of the derivative of this guy with respect to theta is equal to 0. So let's see why. So if I take the derivative with respect to theta, of log f, theta of x, what I get is the derivative with respect to theta of f, theta of x, divided by f theta of x.

Now, if I take the expectation of this guy, with respect to this theta as well, what I get is that this thing-- what is the expectation? Well, it's just the integral against f theta. Or if I'm in a discrete case, I just have the sum against f theta, if it's a pmf. Just the definition, the expectation of x, is either the integral-- well, let's say of h of x-- is integral of h of x. F theta of x-- if this is discrete or is just the sum of h of x, f theta of x. If x is discrete-- so if it's continuous, you put this soft sum. This guy is the same thing, right? So I'm just going to illustrate the case when it's continuous. So this is what? Well, this is the integral of partial derivative with respect to theta, of f theta of x, divided by f theta of x, all time f theta of x-- dx.

And now, this f theta is canceled, so I'm actually left with the integral of the derivative, which I'm going to write as the derivative of the integral. But f theta being density for any value of theta that I can take, this is the function. As a function of theta, this function is constantly equal to 1. For any theta that I take it, it takes value of 1. So this is constantly equal to 1. I put three bars to see that for any value of theta, this is 1, which actually tells me that the derivative is equal to 0. OK, yes?

AUDIENCE: What is the first [INAUDIBLE] that you wrote on the board?

PHILIPPE RIGOLLET: That's just the definition of the derivative of the log of a function?

AUDIENCE: OK.

PHILIPPE RIGOLLET: Log of f prime is f prime over f. That's a log, yeah. Just by elimination.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: I'm sorry.

AUDIENCE: When you write a squiggle that starts with an l, I assume it's lambda.

PHILIPPE RIGOLLET: And you do good, because that's probably how my mind processes. And so I'm like, yeah, l. Here is enough information. OK, everybody is good with this? So that was convenient. So it just said that the expectation of the derivative of the log likelihood is equal to 0. That's going to be our first identity. Let's move onto the second identity, using exactly the same trick, which is let's hope that at some point, we have the integral of this function that's constantly equal to 1 as a function of theta, and then use the fact that its derivative is equal to 0.

So if I start taking the second derivative of the log of f theta, so what is this? Well, it's the derivative of this guy here, so I'm going to go straight to it. So it's second derivative, f theta of x, times f theta of x, minus first derivative of f theta of x, times first derivative of f theta of x. Here is some super important stuff-- no, I'm kidding. So you can still see that guy over there? So it's just the square. And then, I divide by f theta of x squared.

So here I have the second derivative, times f itself. And here, I have the product of the first derivative with itself. So that's the square. So now, I'm going to integrate this guy. So if I take the expectation of this thing here, what I get is the integral. So here, the only thing that's going to happen when I'm going to take my integral is that one of those squares is going to cancel against f theta, right? So I'm going to get the second derivative minus the second derivative squared. And then, I'm divided by f theta. And I know that this thing is equal to 0.

Now, one of these guys here-- sorry, why do I have-- so I have this guy here. So this guy here is going to cancel. So this is what this is equal to-- the integral of the partial, so the second derivative of f theta of x, because those two guys cancel-- minus the integral of the second derivative. And this is telling me what? Yeah, I'm losing one, because I have some weird sequences. Thank you.

OK, this is still positive. I want to say that this thing is actually equal to 0. But then, it gives me some weird things, which are that I have an integral of a positive function, which is equal to 0.

Yeah, that's what I'm thinking of doing. But I'm going to get 0 for this entire integral, which means that I have the integral of a positive function, which is equal to 0, which means that this function is equal to 0, which sounds a little bad-- basically, tells me that this function, f theta, is linear. So I went a little too far, I believe, because I only want to prove that the expectation of the second derivative-- Yes, so I want to pull this out.

So let's see, if I keep rolling with this, I'm going to get-- well, no because the fact that it's divided by f theta, means that, indeed, the second derivative is equal to 0. So it cannot do this here.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: OK, but let's write it like this. You're right, so this is what? This is the expectation of the partial with respect to theta of f theta of x, divided by f theta of x squared. And this is exactly the derivative of the log, right? So indeed, this thing is equal to the expectation with respect to theta of the partial of l-- of log f theta, divided by partial theta. All right, so this is one of the guys that I want squared. This is one of the guys that I want. And this is actually equal, so this will be equal to the expectation--

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Oh, right, so this term should be equal to 0. This was not 0. You're absolutely right. So at some point, I got confused, because I thought putting this equal to 0 would mean that this is 0. But this thing is not equal to 0. So this thing, you're right. I take the same trick as before, and this is actually equal to 0, which means that now I have what's on the left-hand side, which is equal to what's on the right-hand side. And if I recap, I get that e theta of the second derivative of the log of f theta is equal to minus-- because I had a minus sign here-- to the expectation with respect to theta, of log of f theta, divided by theta squared. Thank you for being on watch when I'm falling apart.

All right, so this is exactly what you have here, except that both terms have been put on the same side. All right, so those things are going to be useful to us, so maybe, we should write them somewhere here. And then, we have that the expectation of the second derivative of the log is equal to minus the expectation of the square of the first derivative.

And this is, indeed, my Fisher information. This is just telling me what is the second derivative

of my log likelihood at theta, right? So everything is with respect to theta when I take these expectations. And so it tells me that the expectation of the second derivative-- at least first of all, what it's telling me is that it's concave, because the second derivative of this thing, which is the second derivative of kl divergence, is actually minus something which is must be non-negative. And so it's telling me that it's concave here at this [INAUDIBLE]. And in particular, it's also telling me that it has to be strictly positive, unless the derivative of f is equal to 0. Unless f is constant, then it's not going to change.

All right, do you have a question? So now, let's use this. So what does my log likelihood look like when I actually compute it for this canonical exponential family. We have this exponential function, so taking the log should make my life much easier, and indeed, it does. So if I look at the canonical, what I have is that the log of f theta of x, it's equal simply to y theta minus b of theta, divided by phi, plus this function that does not depend on theta.

So let's see what this tells me. Let's just plug-in those equalities in there. I can take the derivative of the right-hand side and just say that in expectation, it's equal to 0. So if I start looking at the derivative, this is equal to what? Well, here I'm going to pick up only y. Sorry, this is a function of y.

I was talking about likelihood, so I actually need to put the random variable here. So I get y minus the derivative of b of theta. Since it's only a function of theta, I'm just going to write b prime, is at OK-- rather than having the partial with respect to theta. Then, this is a constant. This does not depend on theta, so it goes away.

So if I start taking the expectation of this guy, I get the expectation of this guy, which is the expectation of y, minus-- well, this does not depend on y, so it's just itself-- b prime of theta. And the whole thing is divided by phi. But from my first equality over there, I know that this thing is actually equal to 0. We just proved that. So in particular, it means that since phi is non-zero, it means that this guy must be equal to this guy-- or phi is not infinity. And so that implies that the expectation with respect to theta of y is equal to b prime of theta.

I'm sorry, you're not registered in this class. I'm going to have to ask you to leave. I'm not kidding.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: You are? I've never seen you here. I saw you for the first lecture. OK.

All right, so e theta of y is equal to b prime of theta. Everybody agrees with that? So this is actually nice, because if I give you an exponential family, the only thing I really need to tell you is what b theta is. And if I give you b of theta, then computing a derivative is actually much easier than having to integrate y against the density itself. You could really have fun and try to compute this, which you would be able to do, right?

And then, there's the plus c of y phi, blah, blah, blah-- dy. And that's the way you would actually compute this thing. Sorry, this guy is here. That would be painful. I don't know what this normalization looks like, so it would have to also explicit that, so I can actually compute this thing. And you know, just the same way, if you want to compute the expectation of a Gaussian-- well, the expectation of a Gaussian is not the most difficult one. But even if you compute the expectation of a Poisson, you start to have to work in a little bit. There's a few things that you have to work through. Here, I'm just telling you, all you have to know is what b of theta is, and then, you can just take the derivative.

Let's see what the second equality is going to give us. OK, so what is the second equality? It's telling me that if I look at the second derivative, and then I take its expectation, I'm going to have something which is equal to negative this guy squared. Sorry, that was the log, right?

We've already computed this first derivative of the likelihood. It's just the expectation of the square of this thing here. So expectation of the derivative, with respect to theta of log, f theta of x, divided by partial theta squared. This is equal to the expectation of the square of y, minus b theta, divided by phi squared-- b prime, theta squared.

OK, sorry, I'm actually going to move on with the-- so if I start computing, what is this thing? Well, we just agreed that this was what? The expectation of theta, right? So that's just the expectation of y. We just computed it here.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, that's b prime. There's a derivative here. So now, this is what? This is simply-- anyone? I'm sorry? Variance of y, but you're scaling by phi squared. OK, so this is negative of the right-hand side of our inequality. And now, I just have to take one more derivative to this guy. So now, if I look at the left-hand side now, I have that the second derivative of log, of f theta of y,

divided by partial of theta squared. So this thing is equal to-- well, no, I'm not left with much. The white part is going to go away, and I'm left only with the second derivative of theta, minus the second derivative theta, divided by phi.

So if I take expectation-- well, it just doesn't change. This is deterministic. So now, what I've established is that this guy is equal to negative this guy. So those two things, the signs are going to go away. And so this implies that the variance of y is equal to b prime prime theta. And then, I have a phi square in denominator that cancels only one of the phi squares, so it's time phi.

So now, I have that my second derivative-- since I know phi is completely determining the variance. So basically, that's why b is called the cumulant generating function. It's not generating moments, but cumulants. But cumulants, in this case, correspond, basically, to the moments, at least for the first two. If I start going farther, I'm going to have more combinations of the expectation of y3, y2, and y itself. But as we know, those are the ones that are usually the most useful, at least if we're interested in asymptotic performance. The central limit theorem tells us that all that matters are the first two moments, and then, the rest is just going to go and say well, it doesn't matter. It's all going to [INAUDIBLE] anyway.

So let's go to a Poisson for example. So if I had a Poisson distribution-- so this is a discrete distribution. And what I know is that f-- let me call mu the parameter of y.

So it's mu to the y, divided by y factorial, exponential minus mu. OK so mu is usually called lambda, and y is usually called x, that's why it takes me to a little bit of time. But it usually it's lambda to the x over factorial x, exponential minus lambda.

Since this is just the series expansion of the exponential when I sum those things from 0 to infinity, this thing sums to 1. But then, if I wanted to start understanding what the expectation of this thing is-- so if I want to understand the expectation with respect to mu of y, then, I would have to compute the sum from k equals 0 to infinity of k, times mu to the k, over factorial of k, exponential minus mu-- which means that I would, essentially, have to take the derivative of my series in the end. So I can do this. This is a standard exercise. You've probably done it when you took probability. But let's see if we can actually just read it off from the first derivative of b.

So to do that, we need to write this in the form of an exponential, where there is one

parameter that captures mu, that interacts with y, just doing this parameter times y, and then something that depends only on y, and then something that depends only on mu. That's the important one. That's going to be our B. And then, there's going to be something that depends only on y. So let's write this and check that this f mu, indeed, belongs to this canonical exponential family. So I definitely have an exponential that comes from this guy. So I have minus mu. And then, this thing is going to give me what? It's going to give me plus y log mu. And then, I'm going to have minus log of y factorial.

So clearly, I have a term that depends only on mu, terms that depend only on y, and I have a product of y and something that depends on mu. If I want to be canonical, I must have this to be exactly the parameter theta itself. So I'm going to call this guy theta. So theta is log mu, which means that mu is equal to e to the theta. And so wherever I see mu, I'm going to replace it by [INAUDIBLE] the theta, because my new parameter now, is theta. So this is what? This is equal to exponential y times theta. And then, I'm going to have minus e of theta. And then, who cares, something that depends only on mu. So this is my c of y, and phi is equal to 1 in this case. So that's all I care about. So let's use it.

So this is my canonical exponential family. Y interacts with theta exactly like this. And then, I have this function. So this function here must be b of theta. So from this function, exponential theta, I'm supposed to be able to read what the mean is.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Because since in this course I always know what the dispersion is, I can actually always absorb it into theta from one. But here, it's really of the form y times something divided by 1, right? If it was like log of mu divided by phi, that would be the question of whether I want to call phi my dispersion, or if I want to just have it in there. This makes no difference in practice. But the real thing is it's never going to happen that this thing, this version, is going to be an exact number. If it's an actual numerical number, this just means that this number should be absorbed in the definition of theta. But if it's something that is called sigma, say, and I will assume that sigma is known, then it's probably preferable to keep it in the dispersion, so you can see that there's this parameter here that you can, essentially, play with. It doesn't make any difference when you know phi.

So now, if I look at the expectation of some y-- so now, I'm going to have y, which follows my Poisson mu. I'm going to look at the expectation, and I know that the expectation is b prime of

theta. Agreed? That's what I just erased, I think. Agreed with this, the derivative?

So what is this? Well, it's the derivative of e to the theta, which is e to the theta, which is mu. So my Poisson is parametrized by its mean. I can also compute the variance, which is equal to minus the second derivative of-- no, it's equal to the second derivative of b. Dispersion is equal to 1. Again, if I took phi elsewhere, I would see it here as well. So if I just absorbed phi here, I would see it divided here, so it would not make any difference. And what is the second derivative of the exponential? Still the exponential-- so it's still equal to mu. So that certainly makes our life easier.

Just one quick from remark-- here's the function. I am giving you problem-- can the b function-- can it ever be equal to log of theta? Who says yes? Who says no? Why?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, so what I've learned from this-- it's sort of completely analytic, right? So we just took derivatives, and this thing just happened. This thing actually allowed us to relate the second derivative of b to the variance. And one thing that we know about a variance is that this is non-negative. And in particular, it's always positive. If they give you a canonical exponential family that has zero variance, trust me, you will see it.

That means that this thing is not going to look like something that's finite, and it's going to have a point mass. It's going to take value infinity at one point. So this will, basically, never happen. This thing is, actually, strictly positive, which means that this thing is always strictly concave. It means that the second derivative of this function, b, has to be strictly positive, and so that the function is convex.

So this is concave, so this is definitely not working. I need to have something that looks like this when I talk about my b. So f theta squared-- we'll see a bunch of exponential theta. And there's a bunch of them. But if you started writing something, and you find b-- try to think of the plot of b in your mind, and you find that b looks like it's going to become concave, you've made a sign mistake somewhere.

All right, so we've done a pretty big parenthesis to try to characterize what the distribution of y was going to be. We wanted to extend from, say, Gaussian to something else. But when we're doing regression, which means generalized linear models, we are not interested in the

distribution of y but really the conditional distribution of y given x. So I need now to couple those back together.

So what I know is that this same mu, in this case, which is the expectation-- what I want to say is that the conditional expectation of y given x-- this is some mu of x. When we did linear models, we said well, this thing was some x transpose beta for linear models.

And the whole premise of this chapter is to say well, this might make no sense, because x transpose beta can take the entire range of real values. Whereas, this mu can take only a partial range. So even if you actually focus on the Poisson, for example, we know that the expectation of a Poisson has to be a non-negative number-- actually, a positive number as soon as you have a little bit of variance. It's mu itself-- mu is a positive number. And so it's not going to make any sense to assume that mu of x is equal to x transpose beta, because you might find some x's for which this value ends up being negative.

And so we're going to need, what we call, the link function that relates, that transforms mu, maps onto the real line, so that you can now express it of the form x transpose beta. So we're going to take not this, but we're going to assume that g of mu of x is not equal to x transpose beta, and that's the generalized linear models.

So as I said, it's weird to transform x transpose beta-- a mu to make it take the real line. At least to me, it feels a bit more natural to take x transpose beta and make it fit to the particular distribution that I want. And so I'm going to want to talk about g and g inverse at the same time. So I'm going to actually take always g. So g is my link function, and I'm going to want g to be continuous differentiable. OK, let's say that it has a derivative, and its derivative is continuous. And I'm going to want g to be strictly increasing.

And that actually implies that g inverse exists. Actually, that's not true. What I'm also going to want is that g of mu spans-- how do I do this? So I want the g, as I arrange for all possible values of mu, whether they're all positive values, or whether they're values that are limited between the intervals 0, 1, I want those to span the entire real line, so that when I want to talk about g inverses define over the entire real line, I know where I started.

So this implies that gene inverse exists. What else does it imply about g inverse? So for a function to be invertible, I only need for it to be strictly monotone. I don't need it to be strictly increasing. So in particular, the fact that I picked increasing implies that this guy is actually increasing.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: That's the image. So this is my link function, and this slide is just telling me I want my function to be invertible, so I can talk about g inverse. I'm going to switch between the two. So what link functions am I going to get? So for linear models, we just said there's no link function, which is the same as saying that the link function is identity, which certainly satisfies all these conditions. It's invertible. It has all these nice properties, but might as well not talk about it.

For Poisson data, when we assume that the conditional distribution of y given x is Poisson, the mu, as I just said, is required to be positive. So I need a g that goes from the interval 0 infinity to the entire real line. I need a function that starts from one end and just takes-- not only the positive values are split between positive and negative values. And here, for example, I could take the log link. So the log is defined on this entire interval. And as I range from 0 to plus infinity, the log is ranging from negative infinity to plus infinity. You can probably think of other functions that do that, like 2 times log. That's another one. But there's many others you can think of.

But let's say the log is one of them that you might want to think about. It is unnatural in the sense that it's one of the first function we can think of. We will see, also, that it has another canonical property that makes it a natural choice. The other one is the other example, where we had an even stronger condition on what mu could be. Mu could only be a number between 0 and 1, that was the probability of success of a coin flip-- probability of success of a Bernoulli random variable.

And now, I need g to map 0, 1 to the entire real line. And so here are a bunch of things that you can come up with, because now you start to have maybe-- I will soon claim that this one, log of mu, divided by 1 minus mu is the most natural one. But maybe, if you had never thought of this, that might not be the first function you would come up with, right? You mentioned trigonometric functions, for example, so maybe, you can come up with something that comes from hyperbolic trigonometry or something.

So what does this function do? Well, we'll see a picture, but this function does map the interval 0, 1 to the entire real line. We also discuss the fact that if we think reciprocally-- what I want if I want to think about g inverse, I want a function that maps the entire real line into the unit interval. And as we said, if I'm not a very creative statistician or probabilist, I can just pick my favorite continuous, strictly increasing cumulative distribution function, which as we know, will

arise as soon as I have a density that has support on the entire real line. If I have support everywhere, then it means that my-- it is strictly positive everywhere, then, it means that my community distribution function has to be strictly increasing. And of course, it has to go from 0 to 1, because that's just the nature of those things.

And so for example, I can take the Gaussian, that's one such function. But I could also take the double exponential that looks like an exponential on one end, and then an exponential on the other end. And basically, if you take capital phi, which is the standard Gaussian cumulative distribution function, it does work for you, and you can take its inverse. And in this case, we don't talk about, so this guy is called logit or logit. And this guy is called probit. And you see it, usually, every time you have a package on generalized linear models. You are trying to implement. You have this choice. And for what's called logistic regression-- so it's funny that it's called logistic regression, but you can actually use the probit link, which in this case, is called probit regression. But those things are essentially equivalent, and it's really a matter of taste. Maybe of communities-- some communities might prefer one over the other.

We'll see that again, as I claimed before, the logistic, the logit one has a slightly more compelling argument for its reason to exist. I guess this one, the compelling argument is that it involved the standard Gaussian, which of course, is something that should show up everywhere. And then, you can think about crazy stuff. Even crazy gets name-- complimentary log, log, which is the log of minus, log 1 minus. Why not? So I guess you can iterate that thing. You can just put a log 1 minus in front of this thing, and it's still going to go. So that's not true. I have to put a minus and take-- no, that's not true.

So you can think of whatever you want. So I claimed that the logit link is the natural choice, so here's a picture. I should have actually plotted the other one, so we can actually compare it. To be fair, I don't even remember how it would actually fit into those two functions. So the blue one, which is this one, for those of you don't see the difference between blue and red-- sorry about that. So this the blue one is the logistic one. So this guy is the function that does e to the x, over 1 plus e to the x. As you can see, this is a function that's supposed to map the entire real line into the intervals, 0, 1. So that's supposed to be the inverse of your function, and I claimed that this is the inverse of the logistic of the logit function.

And the blue one, well, this is the Gaussian CDF, so you know it's clearly the inverse of the inverse of the Gaussian CDF. And that's the red one. That's the one that goes here. I would

guess that the complimentary log, log is something that's probably going above here, and for which the slope is, actually, even a little flatter as you cross 0. So of course, this is not our link functions. These are the inverse of our link function. So what do they look like when actually, basically, flip my thing like this? So this is what I see. And so I can see that in blue, this is my logistic link. So it crosses 0 with a slightly faster rate. Remember, if we could use the identity, that would be very nice to us. We would just want to take the identity. The problem is that if I start having the identity that goes here, it's going to start being a problem.

And this is the probit link, the phi verse that you see here. It's a little flatter. You can compute the derivative at zero of those guys. What is the derivative of the-- So I'm taking the derivative of log of x over 1 minus x. So it's 1 over x, minus 1 over 1 minus x. So if I look at 0.5-- sorry, this is the interval 0, 1. So I'm interested in the slope at 0.5. Yes, it's plus, thank you. So at 0.5, what I get is 2 plus 2. Yeah, so that's the slope that we get, and if you compute for the derivative-- what is the derivative of a phi inverse? Well, it's a little phi of x, divided by little phi of capital phi, inverse of x. So little phi at 1/2-- I don't know.

Yeah, I guess I can probably compute the derivative of the capital phi at 0, which is going to be just that. 1 over square root of 2 pi, and then just say well, the slope has to be 1 over that. Square root 2 pi. So that's just a comparison, but again, so far, we do not have any reason to prefer one to the other.

And so now, I'm going to start giving you some reasons to prefer one to the other. And one of those two-- and actually for each canonical family, there is something which is called the canonical link. And when you don't have any other reason to choose anything else, why not choose the canonical one? And the canonical link is the one that says OK, what I want is g to map mu onto the real line. But mu is not the parameter of my canonical family. Here for example, mu is e of theta, but the canonical parameter is theta. But the parameter of a canonical exponential family is something that lives in the entire real line. It was defined for all thetas. And so in particular, I can just take theta to be the one that's x transpose beta. And so in particular, I'm just going to try to find the link that just says OK, when I take g of mu, I'm going to map, so that's what's going to be. So I know that g of mu is going to be equal to x beta. And now, what I'm going to say is OK, let's just take the g that makes this guy equal to theta, so that this is theta that actually model, like x transpose beta.

Feels pretty canonical, right? What else? What other central, easy choice would you take? This was pretty easy. There is a natural parameter for this canonical family, and it takes value

on the entire real line. I have a function that maps mu onto the entire real line, so let's just map it to the actual parameter.

So now, OK, why do I have this? Well, we've already figured that out. The canonical link function is strictly increasing. Sorry, so I said that now I want this guy-- so I want g of mu to be equal to theta, which is equivalent to saying that I want mu to be equal to g inverse of theta. But we know that mu is what-- b prime of theta. So that means that b prime is the same function as g inverse. And I claimed that this is actually giving me, indeed, a function that has the properties that I want, because before I said, just pick any function that has these properties. And now, I'm giving you a very hard rule to pick this, though you need still to check that it satisfies those conditions and particular, that it's increasing and invertible.

And so for this to be increasing and invertible, strictly increasing and invertible, really what I need is that the inverse is strictly increasing and invertible, which is the case here, because b prime as we said-- well, b prime is the derivative of a strictly convex function. A strictly convex function has a second derivative that's strictly positive. We just figured that out using the fact that the variance was strictly positive. And if phi is strictly positive, then this thing has to be strictly positive.

So if b prime, prime is strictly positive-- this is the derivative of a function called b prime. If your derivative is strictly positive, you are strictly increasing. And so we know that b prime is, indeed, strictly increasing. And what I need also to check-- well, I guess this is already checked on its own, because b prime is actually mapping all of our into the possible values. When theta ranges on the entire real line, then b prime ranges in the entire interval of the mean values that it can take.

And so now, I have this thing that's completely defined. B prime inverse is a valid link. And it's called a canonical link. OK, so again, if I give you an exponential family, which is another way of saying I'll give you a convex function, b, which gives you some exponential family, then if you just take b prime inverse, this gives you the associated canonical link for this canonical exponential family.

So clearly there's an advantage of doing this, which is I don't have to actually think about which one to pick if I don't want to think about it. But there's other advantages that come to it, and we'll see that in the representations. There's, basically, going to be some light cancellations that show up.

So before we go there, let's just compute the canonical link for the Bernoulli distribution. So remember, the Bernoulli distribution has a PMF, which is part of the canonical exponential family. So the PMF of the Bernoulli is f theta of x. Let me just write it like this. So it's p to the y, let's say-- one minus p to the 1 minus y, which I will write as exponential y log p, plus 1 minus y, log 1 minus p. OK, we've done that last time. Now, I'm going to group my terms in y to see how y interacts with this parameter p. And what I'm getting is y, which is times log p divided by 1 minus p. And then, the only term that remains is log, 1 minus p.

Now, I want this to be a canonical exponential family, which means that I just need to call this guy, so it is part of the exponential family. You can read that. If I want it to be canonical, this guy must be theta itself. So I have that theta is equal to log p, 1 minus p. If I invert this thing, it tells me that p is e to the theta, divided by 1, plus e to the theta. It's just inverting this function. In particular, it means that log, 1 minus p, is equal to log, 1 minus this thing. So the exponential thetas go away. So in the numerator, this is what I get. That's the log 1 minus this guy, which is equal to minus log 1, plus e to the theta.

So I'm going a bit too fast, but these are very elementary manipulations-- maybe, it requires one more line to convince yourself. But just do it in the comfort of your room. And then, what you have is the exponential of y times theta, and then, I have minus log, 1 plus e theta. So this is the representation of the p and f of a Bernoulli distribution, as part of a member of the canonical exponential family. And it tells me that b of theta is equal to log 1, plus e of theta. That's what I have there. From there, I can compute the expectation, which hopefully, I'm going to get p as the mean and p times 1, minus p as the variance. Otherwise, that would be weird. So let's just do this.

B prime of theta should give me the mean. And indeed, b prime of theta is e to the theta, divided by 1, plus e to the theta, which is exactly this p that I had there. OK just for fun-- well, I don't know. Maybe, that's not part of it. Yeah, let's not compute the second derivative. That's probably going to be on your homework at some point-- if not, on the final.

So b prime now-- oh, I erased it, of course. G, the canonical link is b prime inverse. And I claim that this is going to give me the logit function, log of mu, over 1 minus mu. So let's check that. So b prime is this thing, so now, I want to find the inverse. Well, I should really call my inverse a function of p. And I've done it before-- all I have to do is to solve this equation, which I've actually just done it, that's where I'm actually coming from. So it's actually telling me that the solution of this thing is equal to log of p over 1 minus p.

We just solve this thing both ways. And this is, indeed, logit of p by definition of logit. So b prime inverse, this function that seemed to come out of nowhere, is really just the inverse of b prime, which we know is the canonical link. And canonical is some sort of ad hoc choices that we've made by saying let's just take the link, such that d of mu is giving me the actual canonical parameter of theta. Yeah?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: You're right. Now, of course, I'm going through all this trouble, but you could see it immediately. I know this is going to be theta. We also have prior knowledge, hopefully, that the expectation of a Bernoulli is p itself. So right at this step, when I say that I'm going to take theta to be this guy, already knew that the canonical link was the logit-- because I just said oh, here's theta. And it's just this function of mu [INAUDIBLE].

OK, so you can do that for a bunch of examples, and this is what they're going to give you. So the Gaussian case, b of theta-- we've actually computed it, actually, once. This is theta squared over 2. So the derivative of this thing is really just theta, which means that g or g inverse is actually equal to the identity. And again, sanity check-- when I'm in the Gaussian case, there's nothing general about general linear models if you don't have a link.

The Poisson case-- you can actually check. Did we do this, actually? Yes we did. So that's when we had this e of theta. And so b is e of theta, which means that the natural link is the inverse, which is log, which is the inverse of exponential. And so that's logarithm link, which as I said, I used the word natural. You can also use the word canonical if you want to describe this function as being the right function to map the positive real line to the entire real line.

The Bernoulli-- we just did it. So b-- the cumulative enduring function is log of 1, plus e of theta, which is log of mu over 1 minus mu. And gamma function-- where you have the thing you're going to see is minus log of minus [INAUDIBLE]. You see the reciprocal link is the link that actually shows up, so minus 1 over mu. That maps.

So are there any questions about the canonical links, canonical families? I use the word canonical a lot. But is everything fitting together right now? So we have this function. We have canonical exponential family, by assumption. It has a function, b, which contains every information we want. At the beginning of the lecture, we established that it has information

about the mean in the first derivative, about the variance in the second derivative. And it's also giving us a canonical link. So just cherish this b once you've found it, because it's everything you need. Yeah?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: I don't know, a political preference? I don't know, honestly. If I were a serious practitioner, I probably would have a better answer for you. At this point, I just don't. I think it's a matter of practice and actual preferences. You can also try both. We didn't mention it, but there's this idea of cross-validation-- well, we mentioned it without going too much into detail. But you could try both and see which one performs best on a yet unseen data set. In terms of prediction, just say I prefer this one of the two, because this actually comes as part of your modeling assumption, right?

Not only did you decide to model the image of mu through the link function as a linear model, but really what you're saying-- your model is saying well, you have two pieces of [INAUDIBLE], the distribution of y. But you also have the fact that mu is modeled as g inverse of x transpose beta. And for different g's, this is just different modeling assumptions, right? So why should this be linear-- I don't know. My authority as a person who has not examined the [INAUDIBLE] data sets for both things would be that the changes are fairly minor.

OK, so this was all for one observation. We just, basically, did probability. We described some density, some properties of the densities, how to compute expectations. That was really just probability. There was no data involved at any point. We did a bit of modeling, but it was all for one observation. What we're going to try to do now is given the reverse engineering to probability that is statistics, given data, what can I infer about my model?

Now remember, there's three parameters that are floating around in this model. There is one that was theta. There is one that was mu, and there is one that is beta. OK, so those are the three parameters that are floating around. What we said is that the expectation of y, given x, is mu of x. So if I estimate mu, I know the conditional expectation of y, given x, which definitely gives me theta of x. How do I go from mu of x to theta of x? The inverse of what-- of the arrow? Yeah, sure, but how do I go from this guy to this guy? So theta as a function of mu is?

AUDIENCE: [INAUDIBLE]

PHILIPPE Yeah, so we just computed that mu was b prime of theta. So it means that theta is just b prime

RIGOLLET: inverse of mu. So those two things are the same as far as we're concerned, because we know that b prime is strictly increasing it's invertible. So it's just a matter of re-parametrization, and we just can switch from one to the other whenever we want.

But why we go through mu, because so far for the entire semester I told you there was one parameter that's theta. It does not have to be the mean, and that's the parameter that we care about. It's the one on which we want to do interference. That's the one for which we're going to compute the Fisher information. This was the parameter that was our object of worship. And now, I'm saying oh, I'm going to have mu that's coming around. And why we have mu, because this is the mu that we use to go to beta. So I can go freely from theta to mu using b prime or b prime inverse. And now, I can go from mu to beta, because I have that g of mu of x is beta transpose x.

So in the end, now, this is going to be my object of worship. This is going to be the parameter that matters. Because once I set beta, I set everything else through this chain. So the question is if I start stacking up this pile of parameters-- so I start with my beta, which in turns give me a mu, which in turn, gives me a theta-- can I just have a long, streamlined-- what is the outcome when I actually start writing my likelihood, not as a function of theta, not as a function of mu, but as a function of beta, which is the one at the end of the chain? And hopefully, things are going to happen nicely, and they might no. Yeah?

AUDIENCE: [INAUDIBLE]

PHILIPPE Is G-- that's my link. G of mu of x-- now, its mu is a function of x, because its conditional on x.
RIGOLLET: So this is really theta of x, mu of x, but b is not a function of x, because it's just something to tells me what the function of x is.

AUDIENCE: [INAUDIBLE]

PHILIPPE Mu is the conditional expectation of y, given x. It has, actually, a fancy name in the statistics
RIGOLLET: literature. It's called-- anybody knows the name of the function, mu of x, which is a conditional expectation of y, given x?

AUDIENCE: [INAUDIBLE]

PHILIPPE That's the regression function. That's actual definition. If I tell you what is the definition of the

RIGOLLET: regression function, that's just the conditional expectation of why, given x. And I could look at any property of the conditional distribution of y given x. I could look at the conditional 95th percentile. I can look at the conditional median. I can look at the conditional [INAUDIBLE] range. I can look at the conditional variance. But I decide to look at the conditional expectation, which is called the regression function. Yes?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Oh, there's no transpose here. Actually, only Victor-Emmanuel used this prime for transpose, and I found it confusing with the derivatives. So primes here is only a derivative.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Oh, yeah, sorry, beta transpose x. So you said what? I said that g of mu of x is beta transpose x?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Isn't that the same thing? X is a vector here, right?

AUDIENCE: Yeah.

PHILIPPE RIGOLLET: So x transpose beta, and beta transpose x are of the same thing.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: So beta looks like this. X looks like this. It's just a simple number. Yeah, you're right. I'm going to start to look at matrices. I'm going to have to be slightly more careful when I do this.

OK so let's do the reverse engineering. I'm giving you data. From this data, hopefully, you should be able to get what the conditional-- if I give you an infinite amount of data, you would know exactly, of pairs xy, what the conditional distribution of y given x is. And in particular, you would know what the conditional expectation of y given x is, which means that you would know mu, which means that you would know theta, which means that you would know beta.

Now, when I have a finite number of observations, I'm going to try to estimate mu of x. But really, I'm going to go the other way around. Because the fact that I assume, specifically, that mu of x is of the form g of mu of x is x transpose beta, then that means that I only have to estimate beta, which is a much simpler object than the entire regression function. So that's what I'm going to go for. I'm going to try to represent the likelihood, the log likelihood, of my data as a function, not of theta, not of mu, but of beta-- and then, maximize that guy.

So now, rather than thinking of just one observation, I'm going to have a bunch of observations. So this might actually look a little confusing, but let's just make sure that we understand each other before we go any further. So I'm going to have observations, x1, y1, all the way to xn, yn, just like in a natural regression problem, except that here my y's might be 0 one valued. They might be positive valued. They might be exponential. They might be anything in the canonical exponential family.

OK so I have this thing, and now, what I have is that my observations are x1, y1, xn, yn. And what I want is that I'm going to assume that the conditional expectation of yi, given-- the conditional distribution of yi, given xi, is something that has density. Did I put an i on y-- yeah.

I'm not going to deal with the phi and the c now. And why do I have theta i and not theta is because theta i is really a function of xi. So it's really theta i of xi. But what do I know about theta i of xi, it's actually equal to b-- I did this error twice-- b prime inverse of mu of xi.

And I'm going to assume that this is of the form beta transpose xi. And this is why I have theta i-- is because this theta i is a function of xi, and I'm going to assume a very simple form for this thing. Sorry, sorry, sorry, sorry-- I should not write it like this. This is only when I have the canonical link. So this is actually equal to b prime inverse of g, of xi transpose beta. Sorry, g inverse-- those two things are actually canceling each other.

So as before, I'm going to stack everything into some-- well, actually, I'm not going to stack anything for the moment. I'm just going to give you a peek at what's happening next week, rather than just manipulating the data. So here is how we're going to proceed at this point. Well now, I want to write my likelihood function, not as a function of theta, but as a function of beta, because that's the parameter I'm actually trying to maximize. So if I have a link-- so this thing that matters here, I'm going to call h. By definition, this is going to be h of xi transpose beta. Helena, you have a question?

AUDIENCE: Uh, no [INAUDIBLE]

PHILIPPE RIGOLLET: So this is just all the things that we know. Theta is just the, by definition of the fact that mu is b prime of theta, the mean is b prime of theta-- it means that theta is b prime inverse of mu. And then, mu is modeled from the systematic component. G of mu is xi transpose beta, so this is g inverse of xi transpose beta. So I want to have b prime inverse of g inverse. This function is a bit annoying to say, so I'm just going to call it h. And when I do the composition of two inverses, the inverse of the composition of those two things in the reverse order-- so h is really the inverse of g, composed with b prime, g of b prime inverse.

And now, if I have the canonical link, since I know that g is b prime inverse, this is really just the identity. As you can imagine, this entire thing, which is actually quite complicated-- would just say oh, this thing, actually, does not show up when I have the canonical link. I really just have that theta can be replaced by xi of beta. So think about going back to this guy here.

Now, theta becomes only xi transpose beta. That's going to be much more simple to optimize, because remember, when I'm going to log likelihood, this thing is going to go away. I'm going to sum those guys. And so what I'm going to have is something which is essentially linear in beta. And then, I'm going to have this minus b, which is just minus the sum of convex functions of beta. And so I'm going to have to bring in the tools of convex optimization. Now, it's not just going to be take the gradient, set it to 0. It's going to be more complicated to do that. I'm going to have to do that in an iterative fashion.

And so that's what I'm telling you, when you look at your log likelihood for all those functions. You sum, the exponential goes away because you had the log, and then, you have all these things here. I kept the b. I kept the h. But if h is the identity, this is the linear function, the linear part, yi times xi transpose beta, minus b of my theta, which is now only xi transpose beta. And that's the function I want to maximize in beta.

It's a convex function. When I know what b is, I have an explicit formula for this, and I want to just bring in some optimization. And that's what we're going to do, and we're going to see three different methods, which are really, basically, the same method. It's just an adaptation or specialization of the so-called Newton-Raphson method, which is essentially telling you do iterative local quadratic approximations through your function-- so second order [INAUDIBLE] expansion, minimize this guy, and then do it again from where you were. And we'll see that this can be, actually, implemented using what's called iteratively re-weighted least squares, which means that every step-- since it's just a quadratic, it's going to be just squares in there--

can actually be solved by using a weighted least squares version of the problem.

So I'm going to stop here for today. So we'll continue and probably not finish this chapter, but finish next week. And then, I think there's only one lecture. Actually, for the last lecture, what do you guys want to do? Do you want to have doughnuts and cider? Do you want to just have some more outlooking lecture on what's happening post 1975 in statistics? Do you want to have a review for the final exam-- pragmatic people.

AUDIENCE: [INAUDIBLE] interesting, advanced topics.

PHILIPPE RIGOLLET: You want to do interesting, advanced-- for the last lecture?

AUDIENCE: Something that we haven't thought of yet.

PHILIPPE RIGOLLET: Yeah, that's, basically, what I'm asking, right-- interesting advanced topics, versus ask me any question you want. Those questions can be about interesting, advanced topics, though. Like, what are interesting, advanced topics? I'm sorry?

AUDIENCE: Interesting with doughnuts-- is that OK?

PHILIPPE RIGOLLET: Yeah, we can always do the doughnuts.

[LAUGHTER]

AUDIENCE: As long as there are doughnuts.

PHILIPPE RIGOLLET: All right, so we'll do that. So you guys have a good weekend.