PHILIPPE RIGOLLET: The chapter is a natural capstone chapter for this entire course. We'll see some of the things we've seen during maximum likelihood and some of the things we've seen during linear regression, some of the things we've seen in terms of the basic modeling that we've had before. We're not going to go back to much inference questions.

It's really going to be about modeling. And in a way, generalized linear models, as the word says, are just a generalization of linear models. And they're actually extremely useful. They're often forgotten about and people just jump onto machine learning and sophisticated techniques. But those things do the job quite well.

So let's see in what sense they are a generalization of the linear models. So remember, the linear model looked like this. We said that y was equal to x transpose beta plus epsilon, right? That was our linear regression model.

And it's-- another way to say this is that if-- and let's assume that those were, say, Gaussian with mean 0 and identity covariance matrix. Then another way to say this is that the conditional distribution of y given x is equal to-- sorry, I a Gaussian with mean x transpose beta and variance-- well, we had a sigma squared, which I will forget as usual-- x transpose beta and then sigma squared. OK, so here, we just assumed that-- so what is regression is just saying I'm trying to explain why as a function of x.

Given x, I'm assuming a distribution for the y. And this x is just going to be here to help me model what the mean of this Gaussian is, right? I mean, I could have something crazy. I could have something that looks like y given x is n0 x transpose beta. And then this could be some other thing which looks like, I don't know, some x transpose gamma squared times, I don't know, x, x transpose plus identity-- some crazy thing that depends on x here, right?

And we deliberately assumed that all the thing that depends on x shows up in the mean, OK? And so what I have here is that y given x is a Gaussian with a mean that depends on x and covariance matrix sigma square identity. Now the linear model assumed a very specific form for the mean. It said I want the mean to be equal to x transpose beta which, remember, was the sum from, say, j equals 1 to p of beta j xj, right?

It's where the xj's are the coordinates of x. But I could do something also more complicated,

right? I could have something that looks like instead , replace this by, I don't know, sum of beta j log of x to the j divided by x to the j squared or something like this, right?

I could do this as well. So there's two things that we have assumed. The first one is that when I look at the conditional distribution of y given x, x affects only the mean. I also assume that it was Gaussian and that it affects only the mean. And the mean is affected in a very specific way, which is linear in x, right?

So this is essentially the things we're going to try to relax. So the first thing that we assume, the fact that y was Gaussian and had only its mean [INAUDIBLE] dependant no x is what's called the random component. It just says that the response variables, you know, it sort of makes sense to assume that they're Gaussian.

And everything was essentially captured, right? So there's this property of Gaussians that if you tell me-- if the variance is known, all you need to tell me to understand exactly what the distribution of a Gaussian is, all you need to tell me is its expected value. All right, so that's this mu of x. And the second thing is that we have this link that says, well, I need to find a way to use my x's to explain this mu you and the link was exactly mu of x was equal to x transpose beta.

Now we are talking about generalized linear models. So this part here where mu of x is of the form-- the way I want my beta, my x, to show up is linear, this will never be a question. In principle, I could add a third point, which is just question this part, the fact that mu of x is x transpose beta.

I could have some more complicated, nonlinear function of x. And then we'll never do that because we're talking about generalized linear model. The only thing with generalize are the random component, the conditional distribution of y given x, and the link that just says, well, once you actually tell me that the only thing I need to figure out is the mean, I'm just going to slap it exactly these x transpose beta thing without any transformation of x transpose beta.

So those are the two things. It will become clear what I mean. This sounds like a tautology, but let's just see how we could extend that. So what we're going to do in generalized linear models-- right, so when I talk about GLNs, the first thing I'm going to do with my x is turn it into some x transpose beta. And that's just the l part, right?

I'm not going to be able to change. That's the way it works. I'm not going to do anything non-

linear.

But the two things I'm going to change is this random component, which is that y, which used to be some Gaussian with mean mu of x here in sigma squared-- so y given x, sorry-- this is going to become y given x follows some distribution. And I'm not going to allow any distribution. I want something that comes from the exponential family.

Who knows what the exponential family of distribution is? This is not the same thing as the exponential distribution. It's a family of distributions.

All right, so we'll see that. It's-- wow. What can that be? Oh yeah, that's actually [INAUDIBLE]. So-- I'm sorry?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: I'm in presentation mode. That should not happen. OK, so hopefully, this is muted.

So essentially, this is going to be a family of distributions. And what makes them exponential typically is that there's an exponential that shows up in the definition of the density, all right? We'll see that the Gaussian belongs to the exponential family. But they're slightly less expected ones because there's this crazy thing that a to the x is exponential x log a, which makes the potential show up without being there.

So if there's an exponential of some power, it's going to show up. But it's more than that. So we'll actually come to this particular family of distribution.

Why this particular family? Because in a way, everything we've done for the linear model with Gaussian is going to extend fairly naturally to this family. All right, and it actually also, because it encompasses pretty much everything, all the distributions we've discussed before.

All right, so the second thing that I want to question-- right, so before, we just said, well, mu of x was directly equal to this thing. Mu of x was directly x transpose beta. So I knew I was going to have an x transpose beta and I said, well, I could do something with this x transpose beta before I used it to explain the expected value.

But I'm actually taking it like that. Here, we're going to say, let's extend this to some function is equal to this thing. Now admittedly, this is not the most natural way to think about it. What you

would probably feel more comfortable doing is write something like mu of x is a function. Let's call it f of x transpose beta.

But here, I decide to call f g inverse. OK, let's just my g inverse. Yes.

AUDIENCE: Is this different then just [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah. I mean, what transformation you want to put on your x's?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Oh no, certainly not, right? I mean, if I give you-- if I force you to work with x1 plus x2, you cannot work with any function of x1 plus any function of x2, right? So this is different.

All right, so-- yeah. The transformation would be just the simple part of your linear regression problem where you would take your exes, transform them, and then just apply another linear regression. This is genuinely new.

Any other question? All right, so this function g and the reason why I sort of have to, like, stick to this slightly less natural way of defining it is because that's g that gets a name, not g inverse that gets a name. And the name of g is the link function.

So if I want to give you a generalized linear model, I need to give you two ingredients. The first one is the random component, which is the distribution of y given x. And it can be anything in what's called the exponential family of distributions.

So for example, I could say, y given x is Gaussian with mean mu x sigma identity. But I can also tell you y given x is gamma with shared parameter equal to alpha of x, OK? I could do some weird things like this.

And the second thing is I need to give you a link function. And the link function is going to become very clear how you pick a link function. And the only reason that you actually pick a link function is because of compatibility.

This mu of x, I call it mu because mu of x is always the conditional expectation of y given x, always, which means that let's think of y as being a Bernoulli random variable. Where does mu of x live?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: 0, 1, right? That's the expectation of a Bernoulli. It's just the probability that my coin flip gives me 1.

So it's a number between 0 and 1. But this guy right here, if my x's are anything, right-- think of any body measurements plus [INAUDIBLE] linear combinations with arbitrarily large coefficients. This thing can be any real number.

So the link function, what it's effectively going to do is make those two things compatible. It's going to take my number which, for example, is constrained to be between 0 and 1 and map it into the entire real line. If I have mu which is forced to be positive, for example, in an exponential distribution, the mean is positive, right?

That's the, say, don't know, inter-arrival time for Poisson process. This thing is known to be positive for an exponential. I need to map something that's exponential to the entire real line. I need a function that takes something positive and [INAUDIBLE] everywhere.

So we'll see. By the end of this chapter, you will have 100 ways of doing this, but there are some more traditional ones [INAUDIBLE]. So before we go any further, I gave you the example of a Bernoulli random variable. Let's see a few examples that actually fit there. Yes.

AUDIENCE: Will it come up later [INAUDIBLE] already know why do we need the transformer [INAUDIBLE] why don't [INAUDIBLE]

PHILIPPE RIGOLLET: Well actually, this will not come up later. It should be very clear from here because if I actually have a model, I just want it to be plausible, right? I mean, what happens if I suddenly decide that my-- so this is what's going to happen.

You're going to have only data to fit this model. Let's say you actually forget about this thing here. You can always do this, right? You can always say I'm going to pretend my y's just happen to be the realizations of said Gaussians that happen to be 0 or 1 only. You can always, like, stuff that in some linear model, right?

You will have some least squares estimated for beta. And it's going to be fine. For all the points that you see, it will definitely put some number that's actually between 0 and 1.

So this is what your picture is going to look like. You're going to have a bunch of values for x.

This is your y. And for different-- so these are the values of x that you will get.

And for a y, you will see either a 0 or a 1, right? Right, that's what your Bernoulli dataset would look like with a one dimensional x. Now if you do least squares on this, you will find this.

And for this guy, this line certainly takes values between 0 and 1. But let's say now you get an x here. You're going to actually start pretending that the probability it spits out one conditionally in x is like 1.2, and that's going to be weird.

Any other questions? All right, so let's start with some examples. Right, I mean, you get so used to them through this course.

So the first one is-- so all these things are taken. So there's a few books on generalizing, your models, generalize [INAUDIBLE] models. And there's tons of applications that you can see.

Those are extremely versatile, and as soon as you want to do modeling to explain some y given x, you sort of need to do that if you want to go beyond linear models. So this was in the disease occurring rate. So you have a disease epidemic and you want to basically model the expected number of new cases given-- at a certain time, OK?

So you have time that progresses for each of your reservation. Each of your reservation is a time stamp-- say, I don't know, 20th day. And your response is the number of new cases.

And you're going to actually put your model directly on mu, right? When I looked at this, everything here was on mu itself, on the expected, right? Mu of x is always the expected-- the conditional expectation of y given x. right?

So all I need to model is this expected value. So this mu I'm going to actually say-- so I look at some parameters, and it says, well, it increases exponentially. So I want to say I have some sort of exponential trend. I can parametrize that in several ways.

And the two parameters I want to slap in is, like, some sort of gamma, which is just the coefficient. And then there's some rate delta that's in the exponential. So if I tell you it's exponential, that's a nice family of functions you might want to think about, OK?

So here, mu of x, if I want to keep the notation, x is gamma exponential delta x, right? Except that here, my x are t1, t2, t3, et cetera. And I want to find what the parameters gamma and delta are because I want to be able to maybe compare different epidemics and see if they

have the same parameter or maybe just do some prediction based on the data that I have without-- to extrapolate in the future.

So here, clearly mu of x is not of the form x transpose beta, right? That's not x transpose beta at all. And it's actually not even a function of x transpose data, right?

There's two parameters, gamma and delta, and it's not of the form. So here we have x, which is 1 and x, right? I have two parameters.

So what I do here is that I say, well, first, let me transform mu in such a way that I can hope to see something that's linear. So if I transform mu, I'm going to have log of mu, which is log of this thing, right? So log of mu of x is equal, well, to log of gamma plus log of exponential delta x, which is delta x.

And now this thing is actually linear in x. So I have that this guy is my first beta 1. And so that's beta 1 finds 1.

And this guy is beta 2-- times, sorry that said beta 0-- times 1, and this guy is beta 1 times x. OK, so that looks like a linear model. I just have to change my parameters-- my parameters beta 1 becomes the log of gamma and beta 2 becomes delta itself.

And the reason why we do this is because, well, the way we put those gamma and those delta was just so that we have some parametrization. It just so happens that if we want this to be linear, we need to just change the parametrization itself. This is going to have some effects.

We know that it's going to have some effect in the fissure information. It's going to have a bunch of effect to change those things. But that's what needs to be done to have a generalized linear model.

Now here, the function that I took to turn it into something that's linear is simple. It came directly from some natural thing I would do here, which is taking the log. And so the function g, the link that I take, is called the log link very creatively. And it's just the function that I apply to mu so that I see something that's linear and that looks like this. So now this only tells me how to deal with the link function. But I still have to deal with 0.1.

And this, again, is just some modeling. Given some data, some random data, what distribution do you choose to explain the randomness? And this-- I mean, unless there's no choice, you know, it's just a matter of practice, right? I mean, why would it be Gaussian and not, you know,

doubly exponential?

This is-- there's matters of convenience that come into this, and there's just matter of experience that come into this. You know, I remember when you chat with engineers, they have a very good notion of what the distribution should be. They have y bold distributions. You know, they do optics and things like this.

So there's some distributions that just come up but sometimes just have to work. Now here what do we have? The thing we're trying to measure, y-- as we said, so mu is the expectation, the conditional expectation, of y given x.

But y is the number of new cases, right? Well it's a number of. And the first thing you should think of when you think about number of, if it were bounded above, you would think binomial, baby.

But here, it's just a number. So you think Poisson. That's how insurers think. I have a number of, you know, claims per year. This is a Poisson distribution.

And hopefully they can model the conditional distribution of the number of claims given everything that they actually ask you in the surveys that I hear you now fail in 15 minutes. All right, so now you have this Poisson distribution. And that's just the modeling assumption.

There's no particular reason why you should do this except that, you know, that might be a good idea. And the expected value of your Poisson has to be this mu i, OK? At time i.

Any question about this slide? OK, so let's switch to another example. Another example is the so-called pray capture rate.

So here, what you're interested in is the rate capture of preys yi for a given prey. And you have xy, which is your explanation. And this is just the density of pray. So you're trying to explain the rate of captures of preys given the density of the prey, OK?

And so you need to find some sort of relationship between the two. And here again, you talk to experts and what they tell you is that, well, it's going to be increasing, right? I mean, animals like predators are going to just eat more if there's more preys. But at some point, they're just going to level off because they're going to be [INAUDIBLE] full and they're going to stop capturing those prays.

And you're just going to have some phenomenon that looks like this. So here is a curve that sort of makes sense, right? As your capture rate goes from 0 to 1, you're increasing, and then you see you have this like [INAUDIBLE] function that says, you know, at some point it levels up. OK, so here, one way I could-- I mean, there's again many ways I could just model a function that looks like this.

But a simple one that has only two parameters is this one, where mu i is this a function of xi where I have some parameter alpha here and some parameter h here. OK, so there's clearly-- so this function, there's one that essentially tells you-- so this thing starts at 0 for sure. And essentially, alpha tells you how sharp this thing is, and h tells you at which points you end here.

Well, it's not exactly what those values are equal to, but that tells you this. OK, so, you know-- simple, and-- well, no, OK. Sorry, that's actually alpha, which is the maximum capture. The rate and h represent the pre-density at which the capture weight is.

So that's the half time. OK, so there's actual value [INAUDIBLE]. All right, so now I have this function.

It's certainly not a function. There's no-- I don't see it as a function of x. So I need to find something that looks like a function of x, OK?

So then here, there's no log. There's no-- well, I could actually take a log here. But I would have log of x and log of x plus h. So that would be weird.

So what we propose to do here is to look, rather than looking at mu i, we look 1 over mu i. Right, and so since your function was mu i, when you take 1 over mu i, you get h plus xi divided by alpha xi, which is h over alpha times one over xi plus 1 over alpha.

And now if I'm willing to make this transformation of variables and say, actually, I don't-- my x, whether it's the density of prey or the inverse density of prey, it really doesn't matter. I can always make this transformation when the data comes. Then I'm actually just going to think of this as being some linear function beta 0 plus beta 1, which is this guy, times 1 over xi.

And now my new variable becomes 1 over xi. And now it's linear. And the transformation I had to take was this 1 over x, which is called the reciprocal link, OK?

You can probably guess what the exponential link is going to be and things like this, all right?

So we'll talk about other links that have slightly less obvious names. Now again, modeling, right?

So this was the random component. This was the easy part. Now I need to just poor in some domain knowledge about how do I think this function, this y, which is which is the rate of capture of praise, I want to understand how this thing is actually changing what is the randomness of the thing around its mean. And you know, something that-- so that comes from this textbook. The standing deviation of capture rate might be approximately proportional to the mean rate. You need to find a distribution that actually has this property. And it turns out that this happens for gamma distributions, right? In gamma distributions, just like say, for Poisson distribution, the-- well, for Poisson, the variance and mean are of the same order.

Here is the standard deviation that's of the same order as the [INAUDIBLE] for gammas. And it's a positive distribution as well. So here is a candidate. Now since we're sort of constrained to work under the exponential family of distributions, then you can just go through your list and just decide which one works best for you.

All right, third example-- so here we have binary response. Here, essentially the binary response variable indicates the presence or absence of postoperative deforming for kyphosis on children. And here, rather than having one covariance which was before, in the first example, was time, in the second example was the density, here there's three ways that you measure on children.

The first one is age of the child and the second one is the number of vertebrae involved in the operation. And the third one is the start of the range, right-- so where it is on the spine. OK, so the response variable here is, you know, did it work or not, right? I mean, that's very simple.

And so here, it's nice because the random component is the easiest one. As I said, any random variable that takes only two outcomes must be a Bernoulli, right? So that's nice there's no modeling going on here.

So you know that y given x is going to be Bernoulli, but of course, all your efforts are going to try to understand what the conditional mean of your Bernoulli, what the conditional probability of being 1 is going to be, OK? And so in particular-- so I'm just-- here, I'm spelling it out before we close those examples. I cannot say that mu of x is x transpose data for exactly this picture that I drew for you here, right?

There's just no way here-- the goal of doing this is certainly to be able to extrapolate for yet unseen children whether this is something that we should be doing. And maybe the range of x is actually going to be slightly out. And so, OK I don't want to see that have a negative probability of outcome or a positive one-- sorry, or one that's lower than one.

So I need to make this transformation. So what I need to do is to transform mu, which is, we know only a number. All we know is a number between 0 and 1. And we need to transform it in such a way that it maps the entire real line or reciprocally to say that-- or inversely, I should say-- that f of x transpose beta should be a number between 0 and 1.

I need to find a function that takes any real number and maps it into 0 and 1. And we'll see that again, but you have an army of functions that do that for you. What are those functions?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: I'm sorry?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Trait?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Oh.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, I want them to be invertible, right?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: I have an army of function. I'm not asking for one soldier in this army. I want the name of this army.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Well, they're not really invertible either, right? So they're actually in [INAUDIBLE] textbook. Because remember, statisticians don't know how to integrate functions, but they know how to turn a function into a Gaussian integral.

So we know it integrates to 1 and things like this. Same thing here-- we don't know how to build functions that are invertible and map the entire real line to 0, 1, but there's all the cumulative distribution functions that do that for us. So I can you any of those guys, and that's what I'm going to be doing, actually.

All right, so just to recap what I just said as we were speaking, so normal linear model is not appropriate for these examples if only because the response variable is not necessarily Gaussian and also because the linear model has to be-- the mean has to be transformed before I can actually apply a linear model for all these plausible nonlinear models that I actually came up with. OK, so the family we're going to go for is the exponential family of distributions.

And we're going to be able to show-- so one of the nice part of this is to actually compute maximum likelihood estimaters for those right? In the linear model, maximum-- like, in the Gauss linear model, maximum likelihood was as nice as it gets, right? This actually was the least squares estimator.

We had a close form. x transpose x inverse x transpose y, and that was it, OK? We had to just take one derivative.

Here, we're going to have a generally concave likelihood. We're not going to be able to actually solve this thing directly in close form unless it's Gaussian, but we will have-- we'll see actually how this is not just a black box optimization of a concave function. We have a lot of properties of this concave function, and we will be able to show some iterative algorithms.

We'll basically see how, when you opened the box of convex optimization, you will actually be able to see how things work and actually implement it using least squares. So each iteration of this iterative algorithm will essentially be a least squares, and that's actually quite [INAUDIBLE]. So, very demonstrative of statisticians being pretty ingenious so that they don't have to call in some statistical software but just can repeatedly call their least squares Oracle within a statistical software.

OK, so what is the exponential family, right? I promised to do the exponential family. Before we

go into this, let me just tell you something about exponential families, and what's the only thing to differentiate an exponential family from all possible distributions?

An exponential family has two parameters, right? And those are not really parameters, but there's this theta parameter of my distribution, OK? So it's going to be indexed by some parameter.

Here, I'm only talking about the distribution of, say, some random variable or some random vector, OK? So here in this slide, you see that the parameter theta that indexed those distribution is k dimensional and the space of the x's that I'm looking at-- so that should really be y, right? What I'm going to plug in here is the conditional distribution of y given x and theta is going to depend on x.

But this really is the y. That's their distribution of the response variable. And so this is on q, right? So I'm going to assume that y takes-- q dimensional-- is q dimensional. Clearly soon, q is going to be equal to 1, but I can define those things generally.

OK, so I have this. I have to tell you what this looks like. And let's assume that this is a probability density function. So this, right this notation, the fact that I just put my theta in subscript, is just for me to remember that this is the variable that indicates the random variable, and this is just the parameter. But I could just write it as a function of theta and x, right?

This is just going to be-- right, if you were in calc, in multivariable calc, you would have two parameter of theta and x and you would need to give me a function. Now think of all-- think of x and theta as being one dimensional at this point. Think of all the functions that can be depending on theta and x.

There's many of them. And in particular, there's many ways theta and x can interact. What the exponential family does for you is that it restricts the way these things can actually interact with each other. It's essentially saying the following.

It's saying this is going to be of the form exponential-- so this exponential is really not much because I could put a log next to it. But what I want is that the way theta and x interact has to be of the form theta times x in an exponential, OK?

So that's the simplest-- that's one of the ways you can think of them interacting is you just the product of the two. Now clearly, this is not a very rich family. So what I'm allowing myself is to

just slap on some terms that depend only on theta and depend only on x.

So let's just call this thing, I don't know, f of x, g of theta. OK, so here, I've restricted the way theta and x can interact. So I have something that depends only on x, something that depends only on theta. And here, I have this very specific interaction.

And that's all that exponential families are doing for you, OK? So if we go back to this slide, this is much more general, right? if I want to go from theta and x in r to theta and x theta in r-- to theta in r k and x in rq, I cannot take the product of theta and x.

I cannot even take the inner product between theta and x because they're not even of compatible dimensions. But what I can do is to first map my theta into something and map my x into something so that I actually end up having the same dimensions. And then I can take the inner product. That's the natural generalization of this simple product.

OK, so what I have is-- right, so if I want to go from theta to x, when I'm going to first do is I'm going to take theta, eta of theta-- so let's say eta1 of theta to eta k of theta. And then I'm going to actually take x becomes t1 of x all the way to tk of x.

And what I'm going to do is take the inner product-- so let's call this eta and let's call this t. And I'm going to take the inner product of eta and t, which is just the sum from j equal 1 to k of eta j of theta times tj of x. OK, so that's just a way to say I want this simple interaction but in higher dimension. The simplest way I can actually make those things happen is just by taking inner product.

OK, and so now what it's telling me is that the distribution-- so I want the exponential times something that depends only on theta and something that depends only on x. And so what it tells me is that when I'm going to take p of theta x, it's just going to be something which is exponential times the sum from j equal 1 to k of eta j theta tj of x. And then I'm going to have a function that depends only-- so let me read it for now like c of theta and then a function that depends only on x.

Let me call it h of x. And for convenience, there's no particular reason why I do that. I'm taking this function c of theta and I'm just actually pushing it in there.

So I can write c of theta as exponential minus log of 1 over c of theta, right? And now I have exponential times exponential. So I push it in, and this thing actually looks like exponential sum

from j equal 1 to k of eta j theta tj of x minus log 1 over c of theta times h of x.

And this thing here, log 1 over c of theta, I call actually b of theta Because c, I called it c. But I can actually directly call this guy b, and I don't actually care about c itself. Now why don't I put back also h of x in there? Because h of x is really here to just-- how to put it-- OK, h of x and b of theta don't play the same role.

B of theta in many ways is a normalizing constant, right? I want this density to integrate to 1. If I did not have this guy, I'm not guaranteed that this thing integrates to 1. But by tweaking this function b of theta or c of theta-- they're equivalent-- I can actually ensure that this thing integrates to 1. So b of theta is just a normalizing constant.

H of x is something that's going to be funny for us. It's going to be something that allows us to be able to treat both discrete and continuous variables within the framework of exponential families. So for those that are familiar with this, this is essentially saying that that h of x is really just a change of measure. When I actually look at the density of p of theta-- this is with respect to some measure-- the fact that I just multiplied by a function of x just means that I'm not looking-- that this guy here without h of theta is not the density with respect to the original measure, but it's the density with respect to the distribution that has h as a density.

That's all I'm saying, right? So I can first transform my x's and then take the density with respect to that. If you don't want to think about densities or measures, you don't have to. This is just the way-- this is just the definition.

Is there any question about this definition? All right, so it looks complicated, but it's actually essentially the simplest way you could think about it. You want to be able to have x and theta interact and you just say, I want the interaction to be of the form exponential x times theta.

And if they're higher dimensions, I'm going to take the exponential of the function of x inner product with a function of theta. All right, so I claimed since the beginning that the Gaussian was such an example. So let's just do it.

So is the Gaussian of the-- is the interaction between theta and x in a Gaussian of the form in the product? And the answer is yes. Actually, whether I know or not what the variance is, OK? So let's start for the case where I actually do not know what the variance is.

So here, I have x is n mu sigma squared. This is all one dimensional. And here, I'm going to assume that my parameter is both mu and sigma square.

OK, so what I need to do is to have some function of mu, some function of stigma square, and take an inner product of some function of x and some other function of x. So I want to show that-- so p theta of x is what? Well, it's one over square root sigma 2 pi exponential minus x minus mu squared over 2 sigma squared, right? So that's just my Gaussian density.

And I want to say that this thing here-- so clearly, the exponential shows up already. I want to show that this is something that looks like, you know, eta 1 of-- sorry, so that was-- yeah, eta 1 of, say, mu sigma squared. So I have only two of those guys, so I'm going to need only two etas, right?

So I want it to be eta 1 of mu and sigma times t1 of x plus eta 2 mu 1 mu sigma squared times t2 of x, right? So I want to have something like that that shows up, and the only things that are left, I want them to depend either only on theta or only on x. So to find that out, we just need to expand.

OK, so I'm going to first put everything into my exponential and expand this guy. So the first term here is going to be minus x squared over 2 sigma square. The second term is going to be minus mu squared over two sigma squared. And then the cross term is going to be plus x mu divided by sigma squared.

And then I'm going to put this guy here. So I have a minus log sigma over 2 pi, OK? OK, is this-- so this term here contains an interaction between X and the parameters.

This term here contains an interaction between X and the parameters. So let me try to write them in a way that I want. This guy only depends on the parameters, this guy only depends on the parameter.

So I'm going to rearrange things. And so I claim that this is of the form x squared. Well, let's say-- do-- who's getting the minus? Eta, OK.

So it's x squared times minus 1 over 2 sigma squared plus x times mu over sigma squared, right? So that's this term here. That's this term here.

Now I need to get this guy here, and that's minus. So I'm going to write it like this-- minus, and now I have mu squared over 2 sigma squared plus log sigma square root 2 pi. And now this thing is definitely of the form t of x times-- did I call them the right way or not?

Of course not. OK, so that's going to be t2 of x times eta 2 of x eta 2 of theta. This guy is going to be t1 of x times eta 1 of theta. All right, so just a function of theta times a function of x-- just a function of theta times a function of x.

And the way combined is just by sending them. And this is going to be my d of theta. What is h of x?

AUDIENCE: 1.

PHILIPPE RIGOLLET: 1. There's one thing I can actually play with, and this is something you're going to have some three choices, right? This is not actually completely determined here is that-- for example, so when I write the log sigma square root 2 pi, this is just log of sigma plus log square root 2 pi.

So I have two choices here. Either my b becomes this guy, or-- so either I have b of theta, which is mu squared over 2 sigma squared plus log sigma square root 2 pi and h of x is equal to 1, or I have that b of theta is mu square over 2 sigma squared plus log sigma. And h of x is equal to what?

Well, I can just push this guy out, right? I can push it out of the exponential. And so it's just square root of 2 pi, which is a function of x, technically. I mean, it's a constant function of x, but it's a function.

So you can see that it's not completely clear how you're going to do the trade off, right? So the constant terms can go either in b or in h. But you know, why bother with tracking down b and h when you can actually stuff everything into one and just call h one and call it a day?

Right, so you can just forget about h. You know it's one and think about the right. H won't matter actually for estimation purposes or anything like this.

All right, so that's basically everything that's written. When stigma square is known, what's happening is that this guy here is no longer a function of theta, right? Agreed?

This is no longer a parameter. When sigma square is known, then theta is equal to mu only. There's no sigma square going on. So this-- everything depends on sigma square can be thought of as a constant.

Think one. So in particular, this term here does not belong in the interaction between x and theta. It belongs to h, right? So if sigma is known, then this guy is only a function of h-- of x.

So h of x becomes exponential x squared minus x squared over 2 sigma squared, right? That's just a function of x. Is that clear?

So if you complete this computation, what you're going to get is that your new one parameter thing is that p theta x is not equal to exponential x times mu over sigma squared minus-- well, it's still the same thing. And then you have your h of x that comes out-- x squared over 2 sigma squared. OK, so that's my h of x.

That's still my b of theta. And this is my t1 of x. And this is my eta one of theta. And remember, theta is just equal to mu in this case.

So if I ask you prove that this distribution belongs to an exponential family, you just have to work it out. Typically, it's expanding what's in the exponential and see what's-- and just write it in this term and identify all the components, right? So here, notice those guys don't even get an index anymore because there's just one of them. So I wrote eta 1 and t1, but it's really just eta and t.

Oh sorry, this guy also goes. This is also a constant, right? So it can actually just put sigma divided by sigma square root 2 pi. So h of x is what, actually? Is it the density of--

AUDIENCE: Standard [INAUDIBLE].

PHILIPPE RIGOLLET: It's not standard. It's centered. It has mean 0.

But it variance sigma squared, right? But it's the density of a Gaussian. And this is what I meant when I said h of x is really just telling you with respect to which distribution, which measure you're taking the density. And so this thing here is really telling you the density of my Gaussian with mean mu is equal to-- is this with respect to a centered Gaussian is this guy, right?

That's what it means. If this thing ends up being a density, it just means that now you just have a new measure, which is this density. So it's just saying that the density of the Gaussian with mean mu with respect to the Gaussian with mean 0 is just this [INAUDIBLE] here. All right, so let's move on. So here, as I said, you could actually do all these computations and forget about the fact that x is continuous. You can actually do it with PMFs and do it for x is discrete. This actually also tells you if you can actually get the same form for your density, which is of the form exponential times the product of the the interaction between theta and x is just taking

this product, then a function only of theta and of function only of x, for the PMF, it also works.

OK, so I claim that the Bernoulli belongs to this family. So the PMF of a Bernoulli-- we say parameter p is p to the x 1 minus p to the 1 minus x, right? Because we know so that's only for x equals 0 or 1.

And the reason is because when x is equal to 0, this is 1 minus p. When x is equal to 1, this is minus 0. OK, we've seen that when we're looking at likelihoods for Bernoullis.

OK, this is not clear this is going to look like this at all. But let's do it. OK, so what does this thing look like?

Well, the first thing I want to do is to make an exponential show up. So what I'm going to write is I'm going to write p to the x as exponential x log p, right? And so I'm going to do that for the other one.

So this thing here-- so I'm going to get exponential x log p plus 1 minus x log 1 minus p. So what I need to do is to collect my terms in x and my terms in whatever parameters I have, see here if theta is equal to p. So if I do this, what I end up having is equal to exponential-- so determine x is log p minus log 1 minus p. So that's x times log p over 1 minus p.

And then the term that rest is just-- that stays is just 1 times log 1 minus p. But I want to see this as a minus something, right? It was minus b of theta.

So I'm going to write it as minus-- well, I can just keep the plus, and I'm going to do-- and that's all [INAUDIBLE]. A-ha!

Well, this is of the form exponential-- something that depends only on x times something that depends only on theta-- minus a function that depends only on theta. And then h of x is equal to 1 again. OK, so let's see.

So I have t1 of x is equal to x. That's this guy. Eta 1 of theta is equal to log p1 minus p.

And b of theta is equal to log 1 over 1 minus p, OK? And h of x is equal to 1, all right? You guys want to do Poisson, or do you want to have any homework?

It's a dilemma because that's an easy homework versus no homework at all but maybe something more difficult. OK, who wants to do it now? Who does not want to raise their hand

now? Who wants to raise their hand now?

All right, so let's move on. I'll just do-- do you want to do the gammas instead in the homework? That's going to be fun.

I'm not even going to propose to do the gammas. And so this is the gamma distribution. It's brilliantly called gamma because it has the gamma function just like the beta distribution had the beta function in there.

They look very similar. One is defined over r plus, the positive real line. And remember, the beta was defined over the interval 0, 1.

And it's of the form x to some power times exponential of minus x to some-- times something, right? So there's a function of polynomial [INAUDIBLE] x where the exponent depends on the parameter. And then there's the exponential minus x times something depends on the parameters.

So this is going to also look like some function of x-- sorry, like some exponential distribution. Can somebody guess what is going to be t2 of x? Oh, those are the functions of x that show up in this product, right?

Remember when we have this-- we just need to take some transformations of x so it looks linear in those things and not in x itself. Remember, we had x squared and x, for example, in the Gaussian case. I don't know if it's still there.

Yeah, it's still there, right? t2 was x squared. What do you think x is going-- t2 of x here. So here's a hint. t1 is going to be x.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, [INAUDIBLE], what is going to be t1? Yeah, you can-- this one is taken. This one is taken.

What? Log x, right? Because this x to the a minus 1, I'm going to write that as exponential a minus 1 log x.

So basically, eta 1 is going to be a minus 1. Eta 2 is going to be minus 1 over b-- well, actually the opposite. And then you're going to have-- but this is actually not too complicated.

All right, then those parameters get names. a is the shape parameter, b is the scale parameter. It doesn't really matter. You have other things that are called the inverse gamma distribution, which has this form.

The difference is that the parameter alpha shows negatively there and then the inverse Gaussian distribution. You know, just densities you can come up with and they just happened to fall in this family. And there's other ones that you can actually put in there that we've seen before. The chi-square is actually part of this family.

The beta distribution is part of this family. The binomial distribution is part of this family. Well, that's easy because the Bernoulli was.

The negative binomial, which is some stopping time-- the first time you hit a certain number of successes when you flip some Bernoulli coins. So you can check for all of those, and you will see that you can actually write them as part of the exponential family. So the main goal of this slide is to convince you that this is actually a pretty broad range of distributions because it basically includes everything we've seen but not anything there-- sorry, plus more, OK? Yeah.

AUDIENCE: Is there any example of a distribution that comes up pretty often that's not in the exponential family?

PHILIPPE RIGOLLET: Yeah, like uniform.

AUDIENCE: Oh, OK, so maybe a bit more complicated than [INAUDIBLE].

Anything Anything that has a support that depends on the parameter is not going to fall-- is not going to fit in there. Right, and you can actually convince yourself why anything that has the support that does not-- that depends on the parameter is not going to be part of this guy. It's kind of a hard thing to-- in fact, you proved that it's not and you prove this rule.

That's kind of a little difficult, but the way you can convince yourself is that remember, the only interaction between x and theta that I allowed was taking the product of those guys and then the exponential, right? If you have something that depends on some parameter-- let's say you're going to see something that looks like this. Right, for uniform, it looks like this.

Well, this is not of the form exponential x times theta. There's an interaction between x and theta here, but it's actually certainly not of the form x exponential x times theta. So this is

definitely not going to be part of the exponential family.

And every time you start doing things like that, it's just not going to happen. Actually, to be fair, I'm not even sure that all these guys, when you allow them to have all their parameters free, are actually going to be part of this. For example-- the beta probably is, but I'm not actually entirely convinced. There's books on experiential families.

All right, so let's go back. So here, we've put a lot of effort understanding how big, how much wider than the Gaussian distribution can we think of for the conditional distribution of our response y given x. So let's go back to the generalized linear models, right?

So [INAUDIBLE] said, OK, the random component? y has to be part of some exponential family distribution-- check. We know what this means. So now I have to understand two things.

I have to understand what is the expectation, right? Because that's actually what I model, right? I take the expectation, the conditional expectation, of y given x.

So I need to understand given this guy, it would be nice if you had some simple rules that would tell me exactly what the expectation is rather than having to do it over and over again, right? If I told you, here's a Gaussian, compute the expectation, every time you had to use that would be slightly painful. So hopefully, this thing being simple enough-- we've actually selected a class that's simple enough so that we can have rules.

Whereas as soon as they give you those parameters t1, t2, eta 1, eta 2, b and h, you can actually have some simple rules to compute the mean and variance and all those things. And so in particular, I'm interested in the mean, and I'm going to have to actually say, well, you know, this mean has to be mapped into the whole real line. So I can actually talk about modeling this function of the mean as x transpose beta.

And we saw that for the [INAUDIBLE] dataset or whatever other data sets. You actually can-- you can actually do this using the log of the reciprocal or for the-- oh, actually, we didn't do it for the Bernoulli. We'll come to this. This is the most important one, and that's called a logit it or a logistic link.

But before we go there, this was actually a very broad family, right? When I wrote this thing on the bottom board-- it's gone now, but when I wrote it in the first place, the only thing that I wrote is I wanted x times theta. Wouldn't it be nice if you have some distribution that was just x times theta, not some function of x times some function of theta?

The functions seem to be here so that they actually make things a little-- so the functions were here so that I can actually put a lot of functions there. But first of all, if I actually decide to re-parametrize my problem, I can always assume-- if I'm one dimensional, I can always assume that eta 1 of theta becomes my new theta, right? So this thing-- here for example, I could say, well, this is actually the parameter of my Bernoulli.

Let me call this guy theta, right? I could do that. Then I could say, well, here I have x that shows up here. And here since I'm talking about the response, I cannot really make any transformations.

So here, I'm going to actually talk about a specific family for which this guy is not x square or square root of x or log of x or anything I want. I'm just going to actually look at distributions for which this is x. This exponential families are called a canonical exponential family.

So in the canonical exponential family, what I have is that I have my x times theta. I'm going to allow myself some normalization factor phi, and we'll see, for example, that it's very convenient when I talk about the Gaussian, right? Because even if I know-- yeah, even if I know this guy, which I actually pull into my-- oh, that's over here, right?

Right, I know sigma squared. But I don't want to change my parameter to be mu over sigma squared. It's kind of painful. So I just take mu, and I'm going to keep this guy as being this phi over there.

And it's called the dispersion parameter from a clear analogy with the Gaussian, right? That's the variance and that's measuring dispersion. OK, so here, what I want is I'm going to think throughout this class-- so phi may be known or not.

And depending-- when it's not known, this actually might turn into some exponential family or it might not. And the main reason is because this b of theta over phi is not necessarily a function of theta over phi, right? If I actually have phi unknown, then y theta over phi has to be-- this guy has to be my new parameter. And b might not be a function of this new parameter.

OK, so in a way, it may or may not, but this is not really a concern that we're going to have because throughout this class, we're going to assume that phi is known, OK? Phi is going to be known all the time, which means that this is always an exponential family. And it's just the simplest one you could think of-- one dimensional parameter, one dimensional response, and I

just have-- the product is just y times or, we used to call it x.

Now I've switched to y, but y times theta divided by phi, OK? Should I write this or this is clear to everyone what this is? Let me write it somewhere so we actually keep track of it toward the [INAUDIBLE].

OK, so this is-- remember, we had all the distributions. And then here we had the exponential family. And now we have the canonical exponential family.

It's actually much, much smaller. Well, actually, it's probably sort of a good picture. And what I have is that my density or my PMF is just exponential y times theta minus b of theta divided by phi.

And I have plus phi of-- oh, yeah, plus phi of y phi, which means that this is really-- if phi is known, h of y is just exponential c of y phi, agreed? Actually, this is the reason why it's not necessarily a canonical family.

It might not be that this depends only on y. It could depend on y and phi in some annoying way and I may not be able to break it. OK, but if phi is known, this is just a function that depends on y, agreed? In particular, I think you need-- I hope you can convince yourself that this is just a subcase of everything we've seen before.

So for example, the Gaussian when the variance is known is indeed of this form, right? So we still have it on the board. So here is my y, right? So then let me write this as f theta of y. So every x is replaceable with y, blah, blah, blah.

This is this guy. And now what I have is that this is going to be my phi. This is my parameter of theta. So I'm definitely of the form y times theta divided by phi.

And then here I have a function b that depends only on theta over phi again. So b of theta is mu squared divided by 2. OK, then it's divided by 6 sigma square. And then I have this extra stuff.

But I really don't care what it is for now. It's just something that depends only on y and known stuff. So it was just a function of y just like my h. I stuff everything in there.

The b, though, this thing here, this is actually what's important because in the canonical family, if you think about it, when you know phi-- sorry-- right, this is just y times theta scaled by a

known constant-- sorry, y times theta scaled by a known constant is the first term. The second term is b of theta scaled by some known constant.

But b of theta is what's going to make the difference between the Gaussian and Bernoullis and gammas and betas-- this is all in this b of theta. b of theta contains everything that's idiosyncratic to this particular distribution. And so this is going to be important. And we will see that b of theta is going to capture information about the mean, about the variance, about likelihood, about everything.

Should I go through this computation? I mean, it's the same. We've just done it, right? So maybe it's probably better if you can redo it on your own.

All right, so the canonical exponential family also has other distributions, right? So there's the Gaussian and there's the Poisson and there's the Bernoulli. But the other ones may not be part of this, right?

In particular, think about the gamma distribution. We had this-- log x was one of the things that showed up. I mean, I cannot get rid of this log x.

I mean, that's part of it except if a is equal to 1 and I know it for sure, right? So if a is equal to 1, then I'm going to have a minus 1, which is equal to 0. So I'm going to have a minus 1 times log x, which is going to be just 0. So log x is going to vanish from here.

But if a is equal to 1, then this distribution is actually much nicer, and it actually does not even deserve the name gamma. What is it if a is equal to 1? It's an exponential, right?

Gamma 1 is equal to 1. x to the a minus 1 is equal to 1. b-- so I have exponential x over b divided by b. So 1 over b-- call it lambda. And this is just an exponential distribution.

And so every time you're going to see something-- so all these guys that don't make it to this table, they could be part of those guys, but they're just more-- they're just to-- they just have another name in this thing. All right, so you could compute the value of theta for different values, right? So again, you still have some continuous or discrete ones.

This is my b of theta. And I said this is actually really what captures my theta. This b is actually called cumulant generating function, OK?

I don't have time. I could write five slides to explain to you, but it would just only tell you why it's

called cumulant generating function. It's also known as the log of the moment generating function. And the way it's called cumulant generating function is because if I start taking successive derivatives and evaluating them at 0, I get the successive cumulance of this distribution, which are some transformation of the moments.

AUDIENCE: What are you talking about again?

PHILIPPE RIGOLLET: The function b.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: So this is just normalization. So this is just to tell you I can compute this, but I really don't care. And obviously I don't care about stuff that's complicated.

This is actually cute, and this is what completes everything. And the rest is just like some general description. You only need to tell you that the range of y is 0 to infinity, right?

And that is essentially telling me this is going to give me some hints as to which link function I should be using, right? Because the range of y tells me what the range of expectation of y is going to be. All right, so here, it tells me that the range of y is between 0 and 1. OK, so what I want to show you is that this captures a variety of different ranges that you can have.

OK, so I'm going to want to go into the likelihood. And the likelihood I'm actually going to use to compute the expectations. But since I actually don't have time to do this now, let's just go quickly through this and give you spoiler alert to make sure that you all wake up on Thursday and really, really want to think about coming here immediately.

All right, so the thing I'm going to want to do, as I said, is it would be nice if, at least for this canonical family, when I give you b, you would be able to say, oh, here is a simple computation of b that would actually give me the mean and the variance. The mean and the variance are also known as moments. b is called cumulant generating function.

So it sounds like moments being related to cumulance, I might have a path to finding those, right? And it might involve taking derivatives of b, as we'll see. The way we're going to prove this by using this thing that we've used several times.

So this property we use when we're computing, remember, the fisher information, right? We

had two formulas for the fisher information. One was the expectation of the second derivative of the log likelihood, and one was negative expectation of the square-- sorry, expectation of the square, and the other one was negative the expectation of the second derivative, right?

The log likelihood is concave, so this number is negative, this number is positive. And the way we did this is by just permuting some derivative and integral here. And there was just-- we used the fact that something that looked like this, right? The log likelihood is log of f theta.

And when I take the derivative of this guy with respect to theta, then I have something that looks like the derivative divided by f theta. And if I start taking the integral against f theta of this thing, so the expectation of this thing, those things would cancel. And then I had just the integral of a derivative, which I would make a leap of faith and say that it's actually the derivative of the integral.

But this was equal to 1. So this derivative was actually equal to 0. And so that's how you got that the expectation of the derivative of the log likelihood is equal to 0.

And you do it once again and you get this guy. It's just some nice things that happen with the [INAUDIBLE] taking derivative of the log. We've done that, we'll do that again.

But once you do this, you can actually apply it. And-- missing a parenthesis over there. So when you write the log likelihood, it's just log of an exponential. Huh, that's actually pretty nice.

Just like the least squares came naturally, the least squares [INAUDIBLE] came naturally when we took the log likelihood of the Gaussians, we're going to have the same thing that happens when I take the log of the density. The exponential is going to go away, and then I'm going to use this formula. But this formula is going to actually give me an equation directly-- oh, that's where it was.

So that's the one that's missing up there. And so the expectation minus this thing is going to be equal to 0, which tells me that the expectation is just the derivative. Right, so it's still a function of theta, but it's just a derivative of b.

And the variance is just going to be the second derivative of b. But remember, this was some sort of a scaling, right? It's called the dispersion parameter.

So if I had a Gaussian and the variance of the Gaussian did not depend on the sigma squared which I stuffed in this phi, that would be certainly weird. And it cannot depend only on mu, and

so this will-- for the Gaussian, this is definitely going to be equal to 1. And this is just going to be equal to my variance.

So this is just by taking the second derivative. So basically, the take-home message is that this function b captures-- by taking one derivative of the expectation and by taking two derivatives captures the variance. Another thing that's actually cool and we'll come back to this and I want to think about is if this second derivative is the variance, what can I say about this thing? What do I know about a variance?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, that's positive. So I know that this is positive. So what does that tell me?

Positive? That's convex, right? A function that has positive second derivative is convex.

So we're going to use that as well, all right? So yeah, I'll see you on Thursday. I have your homework.