ConvexOptimizationI-Lecture11

Instructor (Stephen Boyd):Well, starting on our next application topic, which is statistical estimation. So last time we looked at estimation from a simple fitting point-of-view, approximation in fitting. Now we'll actually look at it from a statistical point of view. Actually it's interesting because you end up kind of with the same stuff. You just get a different viewpoint. So let me give the broad setting is actually just a – well, it's just pure statistics. So here is pure statistics. It's this. You have a probability density that depends on a parameter. So that's pure statistics and I guess our staff department, our contingent, are center front and will keep me honest or at least object when needed.

So in pure statistics there's just parameterized probability distributions and we have a parameter X and your job, you get one or more samples from one of these distributions and you're charge is to say something intelligent about which distribution, which is to say which parameter value, generated the sample. So that's statistics. So a standard technique is maximum likelihood estimations. In maximum likelihood estimation you do the following. You have an observation Y and you look at the density of the – you look at the density at Y or probability distribution if it's a distribution on like – if its got different points on atomic points.

You look at this and you actually choose the parameter value X that maximizes this quantity. That's the same as maximizing the log likelihood. We'll see actually why the log goes in there. The log goes in there for many reasons. One is if you have independent samples that splits into a sum. And the second is that many distributions are actually log concave this time in the parameter. Now, in fact, last time – so far we've looked at distributions that are concave in the argument. This is concave actually in the parameter and those are actually slightly different concepts. So we'll see this in examples. All right.

So the log of the – I'll just call it a density. The log of the density has a function now of the parameter here. Not the argument here. This is called the log likelihood function. In this problem it's often the case, for example, if some of the parameters involve things like variances or covariance matrixes or they're just parameters that are positive or something like that if they represent rates or intensities. They're clearly limited to some set. So you can add to this problem either in explicit constraint, that part that X has to be in subset C, or another way to just affect the same thing is simply to define P sub X of Y to be zero when X is outside the domain of P, at least the parameter argument.

That, of course, makes the log minus infinity and that makes it a very, very poor choice if you're maximizing. So that's the same as just making the constraint implicit. Okay. So these are very often convex optimization problems. By the way, they are sometimes not, if you do machine learning or statistics you'll kind of start to – you'll get a feel for when they're not. Typical examples would be missing – the most obvious and common example is missing data. But we'll see some examples and we'll focus on the cases where it is. So this is often a convex optimization problem. Remember, the variable here is X, which is a parameter in the distribution. Y here actually is a sample. It's an observed sample. Okay.

Let’s look at some examples. The first is just linear measurements with IID noise. So here is what you have. You have a bunch of measurements so you have YI is AI transpose X plus VI. Here X is the set of parameters that you want to estimate, but I guess if you’re a statistician you would say the following. You would say that the data YI are generated by a distribution. The distribution is parameterized by X. Okay. So that’s what it is and the way it’s parameterized is this way. In fact, X only affects, as you can see, the mean of the distribution. It doesn’t affect, for example, the variance, but that doesn’t have to be the case.

So, I mean, the way a normal person would say this is X is a set of parameter values that you want to estimate, but this is fine. The conceptual way to say this is you have a dist – Y is generated from a distribution parameterized by X. Okay. So we’ll take the VI’s to be IID measurement noises and we’ll let the density P of Z. So we’re just gonna make this scalar. It’s easier to work out what happens in the case when these are vector measurements. It’s very easy to do. Okay. So the density of this clearly is simply if you take YI minus AI transpose X and that, of course, has this density and they’re independent. So the probability density once if you fix X, this is the parameter it’s just simply the product of these because they’re independent samples.

So you get this and that’s the density. You take the log of that. That’s the log likelihood function, and you get this thing here. So this is maximized – you know, I think I’m missing a sign or something here. No, I’m not. Everything’s fine. Sorry. So your job you take the log of this. It splits out and you get the sum of the log of these probability densities and your job is to maximize that. So for this to be a convex problem in the general case in X. Remember YI are samples, they’re data. X are the variables here in an optimization problem. What you need to happen is you need P in this case, because this appears in P. That’s affine in X. You need P to be log concave.

So in this case, because the parameter appears afinely in P, it’s good enough for P to be log concave for this to be a convex problem. So let’s look at some examples. Here’s one. If the noise is Gaussian, so the density is just this. It’s nothing but that. Then the log likelihood function looks like this. It’s a constant here, has nothing to do with X, and then minus this term. But if you look at that it’s a positive constant times nothing but sort of an L2 residual, right? So nothing else. So, in this case, the maximum likelihood estimate is the least squares solution. So now you know if you’re doing least squares you’re actually doing maximum likelihood estimation with assuming that the noises are Gaussian.

In fact, it doesn’t matter what the variances – oh, sorry. All the noises have to have the same variance. If you do maximum likelihood estimation and the noises of the different measurements have different variances what does that translate to?

Student: Weighted.

Instructor (Stephen Boyd):Yes, there’s weighted least squares, of course. There’s diagonally weighted least squares. So that’s what you do. So you can turn it around and if you’re doing diagonally weighted least squares the truth is you’re doing diagonally weighed least squares if someone asked you. You could say because I trust that measurement more than I trust that one or something like that, but if a statistician friend stops you and say what are you doing? You say I’m doing maximum likelihood estimation with the noises here having different variances. So that’s what that means. Okay. That’s fine. So now you know what least squares is. Has a beautiful statistical interpretation.

Let’s try Laplassian. So in Laplassian noise it looks like this. It’s exponential and you actually can say a lot of things about this compared to a Gaussian. The most important by far is that the tails are huge compared to a Gaussian. So E to the minus X squared is a whole lot smaller than E to the minus X when X is big. So this has huge, huge tails. I mean, obviously distributions with huger tails, however, those distributions are not log concave, but nevertheless. So that’s the main difference here. You do this and work out what this is, and actually, all you’re doing is just this. You’re just maximizing the sum of the log of the probability density. You get a constant, that’s from this thing. And then over here you get a positive constant time the L1 norm. So that’s what you’re doing.

So if you do L1 estimation then what you’re really doing is you’re doing maximum likelihood estimation with Laplassian noise and this, sort of, makes sense actually. This makes sense that this is Y L1 estimation and you know what L1 estimation does. What L1 estimation does is it allows you to have some large out wires. Where L2 estimation would never allow you to do that. In return, it will actually drive a lot of the smaller residuals to zero. For example, to zero or small numbers. And now you can actually justify it statistically.

You can say, well, you’re really assuming that the noise, in this case, is more erratic. You’re saying that, in fact, V does not fall off like a Gaussian. If something falls off like a Gaussian then being six sigma out is actually a very, very unlikely event. And you will change X considerably to avoid a point where you’re actually making the estimate that one of these things is six sigma out. Okay? If this is Laplassian you are much more relaxed about that. It’s a rare event, but it is not essentially for all practical purposes impossible event. And you’ll actually allow a large out wire there.

So it all, sort of, makes perfect sense statistically. As another example, you have uniform noise. So just suppose the noise is uniform. Which is interesting because it basically says this noise here cannot be bigger than in absolute value of A. Absolutely cannot be. On the other hand, between minus A and A you have absolutely no idea. You’re just completely uncommitted as to what its value is. It’s not even centered at zero. I mean, it’s centered at zero, sorry. But it’s not concentrated at zero. It’s not more likely to be small than it is to be large. It’s completely uniform in there. Sure enough the log likelihood function is this. It’s minus infinity unless all of these measurements are consistent, sort of, within A. They have to be.

But then it’s just a constant. So here the log likelihood function actually takes on only two values. It’s a constant and it’s minus infinity. So any point inside this polyhedron is a maximum likelihood estimate. Okay. So that's what happens here. And, of course, it's not unique and so on and so forth and if you were to add to the uniform noise even the slightest tilt towards the center you might get something else or something like that. So what this does is it allows you to translate or to talk statistically about your fitting, your penalty function. So, for example, lets do one.

Suppose I told you – suppose I'm doing fitting and I'm using a penalty function that looks like this. Okay. So it means I don't really – if your residual is within plus minus one I don't care at all. I just – it's fine. I'm not even interested. Once it's between an absolute value less than one it's fine. Then it grows linearly and so I want to know what's the statistical interpretation. This is maximum likelihood estimation. What noise density? What's the imputed noise density here? What would it be? You do. This is maximum likelihood estimation so what – this corresponds exactly to maximum likelihood estimation. What is it?

Student: It's uniformally [inaudible] and the [inaudible].

Instructor (Stephen Boyd):Exactly it. Yeah. So if you're doing dead zone linear penalty, which you could do and just defend it on very simple grounds and say, well, look if the residuals – look, I can't even measure it that accurately. So if it's between plus minus one I don't even care. That's fine. That's as good as my measurements are anyway. It makes no sense. And then you say, well, what about linear? Why not quadratic out here? Linear would say something like, well, sometimes there are some pretty large residuals or pretty large errors and things, so that's why I have this linear and not quadratic. Okay? So that would be it.

The statistical interpretation is exactly this. You're doing maximum likelihood estimation of Y equals AX plus V. The VI's are IID and they have a distribution that looks like this. That's supposed to be flat by the way. I don't – there. Okay. So it's a uniform distribution between plus minus one and then it falls off. It has exponential tails outside. So that's what you're doing statistically. Makes perfect sense. Here you'd adjust everything to have integral one, of course, but that makes no difference, in fact. It's just that that's the shape of what you have. So if you do uniform – by the way, this is, of course, log concave because the negative log of this is that. And, in fact, that's how I got – I mean, that's how this came by flipping this and then taking the X. Okay. So that's how that works.

So that's all you do. Okay. So there's more on that, but that's something you can do in homework and reading and stuff like that to sort of see how that works. And this is very simple case. You can actually get to the case where you're estimating things like covariance matrixes and things like that or other parameters and it's actually more interesting, but that's the basic idea. Okay. Let's look at another example of something different where the noises are not additive at all. They enter in a complex non-linear way. So we'll look at logistic regression as sort of a canonical example of other classes of maximum likelihood estimation problems that are convex. So let's look at that.

I have a random variable that’s in zero one and its distribution is the following. It has a logistic distribution so the probability that Y is one is E to the A transpose U plus V. It’s E to a number divided by one plus E to the number. So that looks like this, I guess, if you plot this function like that, so that’s zero. And what this says is – actually let’s even take a look at this. If this – for now just treat this as a number. A transpose U plus V. If this number is let’s say zero it means that sort of the equiprobable point. It says that the probability that Y is one is half. If A transpose U plus V is like two or three, then this is very likely, but not perfectly certain to be one.

If this number is say minus two or three this is also – it’s very likely to be zero, but not quite. So the transition zone is where this number is, let’s say, between plus minus one. I mean, you can chose that number some other way. But roughly speaking when this number is zero that’s the equiprobable point. When this number is between, I don’t know, plus minus a half, plus minus one, you can throw in your favorite number in there, the probability is some sort of number like 10 percent, 90, between something like that. You can work it all out. When this number is things like three and four, this is overwhelmingly probable to be one here and if it’s negative three or four it’s overwhelmingly probable to be zero and that’s, sort of, a picture like this. Okay.

Now, I guess to the statistician here A and B are parameters. So those are the parameters that generated the distribution and so your, our job is gonna be estimate A and B. Now, you actually – you can’t estimate A and B if I just give you one sample from this. So you’re gonna be given multiple samples. In other words, I’ll give you a bunch of samples and for each sample I’m gonna give you U. So U here is – these are often called explanatory variables. So they would be other things that you measure and associate with this sample.

So this would be, for example, I don’t know. If you were admitted to a hospital this would be things like blood pressure, weight, blood gas concentrations, all things like this. And this could maybe be if you died or something like that, right? So that would be that so – yeah, hopefully things will end up over here. All right. So that’s it. All right. So, okay. So that’s what this is. so these are the explanatory variables and what you want to estimate are the parameters A and B to parameterize this distribution. Okay. So what we’ll do is we’re given a bunch of samples and a sample looks like this. It’s a pair UI and YI here. These are people who checked in and this is what happened to them. I guess in this case, zero being the good outcome. Well, we have to decide that we’ll make zero the good outcome.

So you get a past data like from last year. You get this data, get a couple hundred of these or something like that and, actually, let’s even just talk about what you do with this. We’re actually gonna fit A and B. So we’re gonna do maximum likelihood estimation of the parameters A and B. By the way, once we have them we can do interesting things because someone can arrive tomorrow at the hospital, we can evaluate their U, evaluate this, and for example, if this turns out to be plus three that’s not good for them. If it’s minus three I guess we can reassure them or something like that. Okay. All right.

So this is the idea. Now, how do you do this? Do you want to work out the log likelihood as such is nothing, but the likelihood function is the product of these things because we're assuming these are independent samples from a logistic distribution. So you take the product of these. You could take, which is this, and what we've done is we've simply reordered the samples so that the first K1's were the ones, I guess, who died in this case and the last M minus K plus one or something like that, whatever this is, are the ones who didn't. So that's – you just reordered them. And, in that case, you get the – you just multiply all this out. You can see that in the denominator you always get this thing. That's for all of them.

And the numerator take this and you take the log of this and this thing is the log of the product of the X. That comes out here as this afinely term and this one over here is minus log sum X. Now, of course, log – well, by the way, why did I call this log sum X? Well, I'm going fast now, but someone want to justify? How did I know just immediately that this thing was convex?

Student:A longer sum of the two.

Instructor (Stephen Boyd):Yeah. But there's a one there. That one is expo of zero. This is the log of E to the zero plus E to the A transpose UI plus V, like that. Okay. So it's log sub X. It's the two-term log sub X culled with zero comma A transpose UI plus V. Okay. So this is convex in this argument that's afine this whole thing is convex. With a minus sign it's concave. So this is concave. Okay. So that's the picture. By the way, if you want you can actually draw this function as a function of AI transpose U plus – yeah, that would be called the logistic. Well, if I put in a negative sign here, which I haven't done yet, it would be called the logistic loss function. And it would be something like this. I wonder if I can get it right. Let's see. Which do you want? If this is small, then that is about zero and it's that, so I think it looks like that at zero.

I may have gotten that right, but if it doesn't look like this then flip it over and if that doesn't work flip it this way and if that doesn't work flip it this way again. So in one of those orientations with some number of minus signs and so on the logist – I guess actually people do logistic loss they refer to it as this. It's a function that looks like that. That's where this gets linear in this case. Approximately linear. Anyway, so but this is something to be maximized. But is this covered in – I know a bunch have taken 229, ES229. Is this there? Come on. Is logistic regression covered in?

Student: It is, yeah. Yes.

Instructor (Stephen Boyd):It is. Okay. All right. So here's just an example. It's a super simple example. This is, as with most examples that you can show on a slide, they're examples of what – they're examples where you definitely did not need any advance theory to figure out how these things work. Okay. So if, in fact, there's only one explanatory variable, which is here, U, and then a bunch of outcomes here, then you definitely don't need anything in advance. Just look at the data with your eye. So that's not the point of this. I'll give you some examples of where logistic regression works. Where it's – we're not talking what you do with your eye.

So here's a whole bunch of data. It's 50 measurements. And each measurement consists of this. It's a value of U here and then it's a one or zero. So that's all it is. I mean, this could be anything. All right. And you can see roughly the following. That when U is larger you – there you get more ones. I mean, you can see that because there's sort of a concentration up here and fewer down here and you can see that if U is smaller there's fewer – you're more likely to be zero. Okay. But you can see there's lots of exceptions. Here's somebody with a very low U that turned out to be one and somebody with a high U that turned out t be zero. Okay.

Oh, I should mention one thing. What do you – well, okay. What this shows is A and B have been estimated. A and B do nothing but control what here in the logistic distribution? B controls the point where this is neutral and A controls the spread. It's the width of the transition zone. So A controls how stretched out this thing is and B controls, sort of, where it is and you can see that it is lined up about at what would be a very good decision point if you were forced to make a hard classifier here. It's lined up quite well with something that approximately separates, not perfectly, but approximately separates these cases from these.

And this spread here tells you roughly how much indecision there is. Okay. Let me ask you a couple of questions here. What do you imagine would happen if the data had looked like this? What if there were none there and there were none there? Okay. So basically there's this data and there's this data. What do you think? What do you think the logistic – well, first of all, let's just talk informally here about what you think the logistic estimate – the logistic maximum likelihood is gonna be? Just draw a picture. What do you think? What's it gonna look like? I mean, like this? No. It's gonna be much sharper. In fact, it's gonna be infinitely sharp. It's gonna basically look like this.

It's gonna be a perfect thing like that. Right? And the truth is, this thing can actually move left or right and it makes absolutely no difference. And the point is that actually the data here is actually linearly separable, right? There's a perfect classifier. That corresponds exactly to the logistic regression problem being unbounded. If you're doing maximization it's unbounded above. Okay. So unboundedness above of the log likelihood. It means you get an estimate. It basically means you can make the the log likelihood function as large as you want. Okay. And that corresponds to this perfect separation like this. Okay.

You have to check me on this. So this is the – okay. What would be uses for this? I am not quite sure, but I believe actually some of the best spam filters are done using logistic regression. You hear both things. Something called support vector machine, which you'll see very shortly, and logistic regression. So my guess, I think it's both. So that's how that works. There, of course, the dimensions are very different. They're not one, right? You've got thousands and thousands of features and things like that and your estimating over very, very large data sets. Okay. This makes sense?

So this is an example where it’s not linear remotely obviously. I mean, this here it enters in a very, very complicated way. The measure – if you like to think of this is a measurement or an outcome work this way. By the way, I should mention to those of you in areas like signal processing and things like that you might think, well, this doesn’t have anything to do with me. Actually, I beg to differ. This is a one-bit measurement. Okay. So this is exactly – well, with a particular distribution. You change the distribution it’s something else. So you will – if you do signal processing with low bit rate measurements, including, for example, single bit measurements, you will exactly get problems like this. Not quite this one. If it’s Gaussian and then you take a sign you’re actually gonna get – it’s called probit regression and you get a different function here, but it’s the same everything otherwise.

It’s kind of the same. Okay. So that was our whirlwind tour of a more complex parameter estimation – a maximum likelihood estimation problem. Okay. And let me mention, actually, with this one I can actually mention some of the cases where the canonical cases where you get non-convex problems and this you probably look at in machine learning and other applied classes and things like that. What would happen is you get a bunch of data samples like this, but for some of the samples some components of the UI’s would be missing. That’s the missing data problem. And there it’s not – the resulting problem is not convex. Okay.

There’s some very good uristics that you can imagine. Like assuming a value of U, carrying out the regression, then going back, plugging in the most likely value, you know, estimating U and alternating between these two and these are all schemes that are used and so on, but, okay. Now we’re gonna look at hypothesis testing. So how does hypothesis testing works? It works like this and for this we go back to the simplest possible case. You can get very complicated other cases like multiple hypotheses and not just one, but multiple ones. Sorry, two, but, sorry. The single hypothesis testing is actually quite simple, but anyway. So we’ll go to the simplest one.

This is where you have two and we’ll just take a random variable on the letters one through m. So just very, very simple random variable. And what’s gonna happen is either a sample, which is literally – it’s just an integer between one and m. It was either generated by the distribution of P or the distribution of Q. So these are non-negative numbers that add up to one. Good as vectors actually. That P and Q. The distribution because we have a finite set here. So that’s the question. I give you a sample and you – I give you a sample or I give you a couple samples of something like that and your job is to estimate was it this distribution or this one?

Now, of course, this could be very easy. For example, if X turns out to be one and this is kind of intuitive, and P1 is .9 and Q1 is .002, right? Then it’s just intuitively clear. It’s quite clear. It’s a very good guess. By the way, not a perfect guess, but a very good guess that, in fact, X came from this distribution of P and not Q. Okay. So that’s roughly – it’s not that unobvious what to do here as in these other things. It’s not that it’s intuitively unobvious. The question is how exactly do you estimate and how do you make a better estimator than an intuitive one. So we'll look at the idea of a randomized detector. So if a randomized detector looks like this.

It's a two by n matrix. So it looks like this. And what it means is this. Each column is actually a probability distribution on the two outcomes like hypothesis one and hypothesis two. Okay. And it says that if this is the outcome that occurs you should guess one with this probability or guess two with this probability. So that's the idea. Now, of course, often these things look like that. Okay. Something like that. What this is if this is what you observe in X you simply guess it came from one. If this is what actually is observed you guess it came from hypothesis two or Q. Okay. So if these are zero one entries it's a deterministic detector.

And the deterministic detector is silly. It just means basically you have partition X, the outcome space, into two groups where if the outcome is in one set you declare that you guess its hypothesis one. If it's in the compliment you say it's hypothesis two. I mean, that, sort of, makes sense. However, you can have something like this. You can have something weird like that. That means that if this is the outcome of observed you then go off and you toss a coin and with 80 percent probability you guess that it was hypothesis one and with 20 percent probability you guess it's hypothesis two. Now, by the way, this is sort of like randomized algorithms in computer science.

This just looks weird like why on earth if you're trying to make an intelligent statement or guess what happened it doesn't seem like gambling on your own is actually gonna help anything. I mean, why on earth would you say, well, tell me happened. You go, excuse me for a minute and get out a coin and flip it. It just doesn't look – there's something that doesn't make sense. Which is kind of like in a randomized algorithm, right? When someone says, here's my problem instance. I'm talking computer science and you say how do you solve it? And you go hang on just a minute and you get out the coin and you start flipping it and you say, look, what are you doing? And you say, no, I'm, gonna try to solve your problem. You say my problem has no probability in it.

I gave you a problem instance, please give me the answer. You go hey, back off. Come on. And you keep flipping the coin. It just looks – you know what I'm saying. After you get used to this intellectually it's okay, but at first it looks very odd. So the idea of having a deterministic detector – also, it's weird. Because you say, hey, output three happened and you go, yeah, sure that came from hypothesis one. Then the next day you say output three happened and you go, yep, that was hypothesis two. All right. So this looks very inconsistent and strange. Actually, we're gonna soon see what a randomized detector can do for you. I mean, it's still weird, I guess. It's like if you're in physics you have to – some day you have to either come to some – you have to sum an equilibrium, for example, with quantum mechanics and things like that. It doesn't make any sense and same with these. Or not. But then you just know it's an unresolved intellectual issue. Okay.

So this is what a detector matrix is and it describes a randomized detector and if you're uncomfortable with that you can make it all zero one matrix. Of course, there is a one in each column and this becomes an encoding of a partition of the outcome space. Okay. Now, if I simply multiply this matrix T by P and Q. So if I take T and then I multiply by PQ, like this, that's lined up like this, then I get four probabilities here and they're actually quite interesting. So I'll tell you what they are. And I guess we should start with – well, I don't know. We could start with the one one entry.

The one one entry is the probability that you guess hypothesis one is correct when, in fact, the sample is generated from hypothesis one. So this entry is a good entry – sorry, it's an entry that you want near one. Actually, if it were one it would mean that you would be absolutely infallible. You could never – this entry. Let's go down and look at this one. This entry is that's the probability that you guess hypothesis two when X is generated by distribution one. Okay.

So that's the – this is a false positive. We're taking, I guess, hypothesis two to be positive. Okay. So that's a false positive. Now, we'll get to that. This one is two two entry so these two add up to one. I mean, they're like a conditional distribution in fact. I mean, they're conditional probabilities. This entry is the probability that you were – that, in fact, the distribution was generated by distribution two and you guessed correctly distribution two. So that's this one. Again, that's something you want one. And this is – so these are the ones you want small are these off-diagonals that's probability of false positive, probability of false negative and you want those small. So, in fact, what you really have here is a bi-criterion problem because what you really like – well, of course, you'd really like this matrix to be the identity matrix.

That means you're absolutely infallible. That means whenever distribution – it comes from distribution one you correctly estimate the distribution one and so on. You make no mistakes of any kind. No false positives. No false negatives. But you really have – in general, you have a tradeoff between these two false alarm probabilities. Okay. And I guess there's lots of names for this tradeoff. I actually don't know what it's called in statistics. I know what it's called in signal processing. It's called the ROC, which is the receiver operating characteristic, but which goes back to like World War II or something like that. What's it called in statistics?

Student:[Inaudible]

Instructor (Stephen Boyd):Really. Maybe it came from you guys. Who knows. It doesn't sound like it to me, but – okay. All right. So this is – that's the bottom you want and let – we can actually talk about some of the extreme points. Let's see. How small could I make the probability of false positive be and how might I do that? How small can I make the false probability – the probability of a false positive?

Student: Zero.

Instructor (Stephen Boyd):I can make it zero. How?

Student: Always guess one.

Instructor (Stephen Boyd):Yeah. You just guess one no matter what happens. Then, in that case, you will never be accused of having guessed two when, in fact, one generated the sample. Okay. So that – by the way, that scheme also has another huge advantage, which is great simplicity. You don't have to even listen to what happened to guess where it came from. And, obviously, I could make false negative zero as well by simply always saying it's positive. All you do is you always guess it's – sorry, guess it's negative always. And I think we've got it wrong. This was – sorry, this was. To make a – no, sorry, that's right. You guess one here, you guess two always. Okay.

So will – I mean, this is actually correctly viewed as a bi-criterion problem. Okay. So it's a bi – oh, and by the way, what happens if there's multiple hypotheses like 12? What happens if there's 12? Let's talk about it. Suppose there's 12 because this is not exactly complicated stuff. We're just multiplying the matrix out. What happens in the case of 12 hypotheses? Wow, this would be a good homework problem or something like that. So what happens?

Student:[Inaudible]

Instructor (Stephen Boyd):You have a 12 by 12 matrix. And how many of these bad guys off the diagonal do you have?

Student:144 minus 12, so 132.

Instructor (Stephen Boyd):Yeah. But some of them aren't – oh, okay. That's fine. Yeah. That's fine. Okay. Yeah. So roughly 70 or something like that, right? Okay. So we've got a bunch of them, right? And so now you have a big tradeoff where you might have strong concerns about – you might have different concerns about guessing hypothesis three when, in fact, the truth came from whatever hypothesis two. That might be one you really care about making low or something and you can imagine after then it would get interesting, but we're just looking at it in the simple case. Okay.

So you end up with this. If it's a bi-criterion problem. You want to minimize these two things subject to – this just says the column sums in T are one, obviously, and this says that they're non-negative and so this is obvious. By the way, notice that everything here is linear. So if this a bi-criterion LP. Actually, it's a trivial bi-criterion LP. So let's scalorize it, so we put in a weight lambda. If you put in a weight lambda you add this up and this is silly. That's a linear function of T of the entries in T. It's extremely straightforward. This says subject to that. It's not only a little – I mean, so this is one of those linear programs, I guess, that one of those trivial linear programs that you can just directly solve.

It's kind of obvious. In fact, you can optimize each column of T completely separately because that's linear and so it's separable in T. Gives you a sum of contributions from each column. The constraints on T are simply that each column is in the unit simplex and I think you just did a problem on that or you did one a while ago or something. Anyway, and its' extremely simple how to choose? You simply choose this. If P is bigger than lambda Q you choose the first hypothesis. The other way around, you choose this one. So you end up with a deterministic detector and this is called a likelihood ratio. Likelihood ratio test because what you really do is you look at this and you compare it to a threshold and then you guess which distribution it came from.

By the way, this extends to the idea of continuous distributions and things like that and it's very old. It's from easily the 20's or something like that. Roughly, yeah, 20's I'd say. Maybe even earlier, right? This is –

Student:I don't know. I think that's about right.

Instructor (Stephen Boyd):20's? Okay. We'll just say the 20's, right? It goes back to – so now you see you have a nice mathematical way to go back to the simple case. You just – you look. How do you guess which one it is? Oh, by the way, if lambda is one then you simply choose whichever one is the more likely. By the way, we'll see what the lambda equals one means. Lambda equals one – well, you can tell me. What does lambda mean here? What is lambda equals one? Because a maximum likelihood test would simply choose whichever had the maximum likelihood. That corresponds exactly to lambda equals one. What's the meaning in the bi-criterion problem for lambda equals one? Has a very specific meaning. What does it minimize? That minimizes something for false positives and false negative maximum likelihood. What is it?

Student:[Inaudible] of error.

Instructor (Stephen Boyd):The sum? Yeah. Which is otherwise known as the – if you put a lambda – if lambda's one here it means, actually, that you're minimizing the sum of this and this. But the sum of this and this has an interpretation. It's called being wrong. Right? So you're minimi – it's called an error. So you're minimizing the probability of error. That corresponds exactly to lambda equals one. Okay. By the way, this ties back to maximum. Like this is, sort of, the justification for maximum likelihood. So you can say maximum likelihood detector you would then argue, you'd argue and you'd be correct, minimizes the probability of being wrong and you're absolutely neutral. False positive has the same weight as false negative. You have a question?

Student: Yeah. For this do we have to assume that there's kind of an equal likelihood of being distribution E and distribution 2?

Instructor (Stephen Boyd):Nope.

Student: Because it's –

Instructor (Stephen Boyd):We're doing statistics.

Student: Okay. But, I mean –

Instructor (Stephen Boyd):You're not even allowed to say that in statistics.

Student: Okay.

Instructor (Stephen Boyd):I mean, if there are –

Student: If you're Basian.

Instructor (Stephen Boyd):Yeah. If you're – no, no. If you're Basian, no, no, no. We're not doing that right now. So that's – if you say that in front of statisticians be careful. It could be physically dangerous. If there are some Basians around in a room with you and they're large you're safe. Go ahead and say that, but watch out.

Student: Okay. So, I guess, it seems then that if we get a data form and there's a – and we look at P and Q and there's a very small probability that that data point came from distribution P. A very large probability that it came from Q, but if we know that 90 percent of the time –

Instructor (Stephen Boyd):By the way, if you're all ready using all ready, you've identified yourself as a Basian and you'd be all ready in big trouble.

Student: Okay. Well, I don't –

Instructor (Stephen Boyd):That's okay. I'll tell you in a minute. Okay. You're safe here.

Student: Okay.

Instructor (Stephen Boyd):Don't try this out in certain parts of Sequoia. Okay. Don't try it. But go on.

Student: Okay. But, anyway, if you then – if you know that 90 percent of the time it's distribution P then even though it's a small probability that your outcome came from distribution P, the fact that distribution P is much more likely should probably influence your choice.

Instructor (Stephen Boyd):Absolutely. So, actually, this is a very good point to say this. So we're doing statistics. In statistics there is no prior distribution. That you can't say even something like this is more likely or what's the likelihood of it being this or that? You're just neutral. Okay. I'm – by the way, I'm neutral on this. I'm not partisan in any way on this. We have to be careful because a lot of people are in machine learning in this room too and I don't know where their sympathies lie. A peer statistician would never – you're not even allowed to say something like what's – this is a more probable out – which one is more probable? The minute you say that now you're doing Basian estimation.

Now, I should say this. That if you do Basian estimation, then instead of the maximum likelihood you do something called MAP, which is Maximum A Posteriori Probability.

Okay. And it’s, of course, quite sensible if you have some ideas about what – which of these things. If I told you ahead of time – by the way, you could redo all of this and it’s very, very easy in that case. If you do things like saying, well, yeah, it could come from these two distributions, but, in fact, with 30 percent chance it comes from P and 70 from Q. And you could work out everything that would happen here and you’d want something called the max. Then you could talk about the probability of the actual probability of being wrong. Things like that.

Now, the good news is this. By the way, we could have done that earlier as well in maximum likelihood estimation, like, for example, here we could have done MAP. All that happens is very cool. You add an extra term, so what is a log likelihood function or a likelihood function – a log likelihood function turns into a log of a conditional density. Then you multiply by the log of a prior density and you get regularization of terms here. So that’s what would happen. But you need to know just socially speaking that watch out you’re treading on very dangerous – just saying things like that can get you in big trouble. What’s that?

Student: That wasn’t a good thing to start.

Instructor (Stephen Boyd):No, I’m neutral. I’m totally neutral, but just you can say that in the wrong place and be very sorry. So the good news is from the convex optimization point of view they all lead to – so the way that would happen is if I was being – if I had some Basians coming on one direction and I had Daphne Collar coming on one side and who would be the most extreme one in statistics? Who would – Who?

Student: No, I –

Instructor (Stephen Boyd):Whoever you’re not gonna name any names.

Student: People are pretty flexible now.

Instructor (Stephen Boyd):They’re pretty flexible, yeah. So that was the right answer. And I had a statistical fundamentalist coming in. A group of them coming in on the left. I would – for us it’s very simple. You just add a term here, which we would just say is this regularization? It’s regular. And they’d say that’s not regularization, that’s log of a prior. And I’d say, no, no, this is just regularized maximum likelihood or whatever of that. Thanks for bringing that up because I just wanted to – I mean, if actually I think if you read the book it’s neutral and, I believe, honest about what it says. It just says if you choose to believe that you’re getting a sample from a parameterized distribution and your job is to estimate the distribution that’s a statistician. That’s statistics. You can do that.

Student: Are also statisticians.

Instructor (Stephen Boyd):Now you know who I hang out with and what they are. Okay, sorry.

Student: It’s Frequentists is the –

Instructor (Stephen Boyd):Frequentists, Okay. So I’m not totally up on the details of these very schisms, but I do know that there’s been a lot of wars about it and a lot of bloodshed actually over these issues, but for us they’re all just convex problems and they’re just innocent little extra terms that end up in here. So that was a very long answer to your question. But it’s a good – just as a matter of safety to know. Asking that innocent question in the wrong place could get you in deep trouble. So not really, but – actually, maybe it could. Well, we won’t go in. All right. So let’s look at some examples now. Oh, we have our interpretation of maximum likelihood.

Maximum likelihood says you scan – if outcome three has occurred you scan the two probability distributions, you go to outcome three, and you see which has the higher likelihood. That’s a maximum likelihood detector that corresponds exactly to lambda equals one. Lambda equals one corresponds precisely to saying that you treat false positives and false negatives equally. Okay. That’s maximum likelihood. That will minimize the sum of the two, which is something like the probability of it being wrong. Okay. Now, you could also do a mini max detector.

You could say, well, I want to minimize the maximum of the two error probabilities. Now, this one actually – of course, it translates to an LP immediately, but this one actually, generally speaking, in fact, in the finite case essentially always has a non-deterministic solution and so let’s see what that is. So here’s an example where these two distributions and it’s kind of look, anybody can figure out what happened here. If it’s outcome one you should probably guess it came from P and if it’s outcome three you should probably guess it came from Q. Everyone’s gonna agree on that. Okay.

And there’s actually not much else here to do, right? Because two other outcomes and obviously if it’s outcome two you should kind of lean towards saying it was two, but you should equivocate or something like that. Something like that. Okay. All right. So here’s the ROC, or Receiver Operating Characteristic, although it’s generally drawn this way and I don’t know why, but anyway, it often is. So here are the two. There’s a probability of false positive and false negative. We’ve all ready talked about these things. This one is the detector that simply says – sorry, yeah. You have zero false negative that means you just simply always say it’s positive. You just announce it’s positive and you can have zero probability of false negative.

Here’s a tradeoff curve here and actually it’s on, of course, it’s piecewise linear. I mean that’s obvious because you’re minimizing a piecewise linear function over some linear inequality constraints. So this curve it’s – well, this region is polyhedral and the vertices here correspond to different things. In fact, the vertices correspond to the three thresholds in lambda. Lambda is the slope here and so as you vary lambda to get a tradeoff curve it’s kind of boring because if lambda is smaller than this you get that point. As you increase lambda you will click over to three, this point, and it’s just some point. These are the deterministic detectors, one, two, three and four. As you increase lambda like this it’s always the same.

Right at this point you switch over to a different threshold. Then you keep going like this and then eventually you get so high that the safest thing to do is simply to guess that nothing ever happened and that way you will have zero false positive always. Okay. So there's not many points on the tradeoff curve, but the mini max one that's very simple. It's the maximum of the two and the level curves of the maximum – this dash line here is equal, is the line of equal false probability and false negatives and you can see very clearly that it's not a determinant. It's right in between these two and it's not a deterministic detector.

So if you want to have a detector here which minimizes the maximum probability of making either a false positive or a false negative then you're gonna have to use a non-deterministic – a randomized detector. Okay. Now, I mean, this is all kind of stupid looking in cases like this. That's fine. But if you want to make this non-trivial, it's very, very simple. You just imagine something with 12 outcomes and vectors, which are – and probability solutions on a thousand points. Okay. Very, very simple. And now you want to decide on a detector and now you start throwing in insane things like how much you care about guessing it's No. 7 when, in fact, the truth was No. 3 and you start throwing in how much you care about all these things and all of a sudden it's not obvious exactly how to do this.

Then when you solve an LP you get something that – or an LP or whatever. Actually, it's always going to be an LP. When you solve an LP you'll actually get a detector that will beat, soundly beat, according to that measure whatever your criterion is. Okay. So that's the picture. Okay. So that's that. There's a lot more you can do with mini max detector – with detectors here. And there's a couple more topics in the book, which you will read about. Okay. Now, we're gonna look at our last topic in this whirlwind tours experiment design. This is quite useful, very useful, and I think not that well – in some fields it's not that well diffused, knowledge of experiment design.

Even in areas where people actually often do experiments and get – construct models from data and things like that. Okay. So experiment design goes like this. And we'll talk about how it works. We'll do the simplest case, which is linear measurements like this and we'll just say that the noises are IID and zero one. You can change all of this, but let's just do that. Well, the maximum likelihood of least squares estimate is just this. And the error is zero mean. That's X hap minus X is zero mean and has a covariance matrix, which is this thing. So that's the covariance.

By the way, here the noises all have noise power one. So the norm of A is something like a signal to noise ratio and now I'm talking if you're in signal processing or communications. So that's because I just made the noises have power one. So if A is large – if the norm of A is large that's a very good measurement. It's a high quality measurement. It's a high signal noise ratio. If the norm of A is small – the norm of A is zero, that's an utterly useless measurement because it's just a sample from noise and it has nothing to do with X. Okay. So that's if you want to get a rough idea. A larger A is larger signal to noise. That's what norm of A is. It's literally the signal to noise, something like that ratio because the noise power is one. Okay.

All right. And I'm sort of assuming there norm of X is on the order of one, but that – with that assumption that just scales the signal to noise. It's still the case that if you double A it's, sort of, twice as good from a noise point-of-view, standard deviation anyway as if you hadn't doubled A. Okay. Now, this is the error covariance and it's a sum of diags inverse. That's very interesting. Okay. So we can say a lot of interesting things about this, but first is if that matrix were singular that would not be good. Okay. And what that means is the matrix is singular it means you basically haven't taken a set of linearly independent – you haven't taken enough measurements basically. Sorry, measurements that span X. Okay. So that would be the case there.

So if you've taken enough measurements so that the measurement vectors span our end then that means this thing is not singular and then you take the inverse and, of course, that's the error covariance and now it's clear now I'm gonna be very rough here. To make this matrix – you want this matrix small and we'll talk about what small can mean. It can mean lots of things. You want the matrix small. Roughly speaking the first thing you want is you want what's inside it since that's an inverse to be big. Good. That corresponds perfectly because if A is big – if the A's are big then that means you have high signal to noise ratio and that makes this error covariance small.

Now, the interesting thing is it's not scalar, it's actually matrix inverse and that's very interesting because it means what the error covariance is is not simply – it doesn't depend just on the individual A's. It's actually how they all go together. Okay. Oh, there's one exception. If the AI's are mutually orthogonal then this inverse you can just do it, change coordinates, and do it and they're independent and then they don't interact in any way whatsoever. But, in general, what this says is the following is you take a bunch of measurements, those are characterized by A1, A2, A3, and so on, and then you calculate the error covariance and the error covariance kind of depends on actually the geometry of the whole set of these things. I mean, this is kind of obvious, but that's the idea.

And, for example, if you want to make an error code – a confidence ellipsoid it would depend on this covariance thing. Okay. So experiment design is this. It says I give you a pallet of possible experiments you can carry out. So we'll call those V1 through VP and these are just experiments you can carry out and now your job is to choose some number of experiments that will make this error covariance matrix as small as possible. So another way of saying that is I give you a pallet of possible experiments you can choose and the simplest thing would be I give you 50 possible measurements and I say you may choose 1,000 measurements from these 50. Okay. And you choose a thousand and you want that choice of a thousand from those 50 to be mutually maximally informative because what's gonna happen is you're gonna take all those measurements together. You're gonna blend them, do least squares, and do all the good blending that least squares is gonna do.

Together they're gonna give you an error covariance like that. Let's talk about some choices. If someone said, well, since your one has the highest signal to noise ratio, so I'm gonna choose all 1,000 measurements to have the signal to noise ratio ten because all these others have signal to noise ratio one. Any comments on that? Is it a good choice? If

V1 has a signal noise ratio of ten and all the others have signal noise ratio one then you can choose any of them. So you can – why would you not always choose the first measurement. It's ten times cleaner than any of the other measurements. What's the problem?

Student: This is not an [inaudible].

Instructor (Stephen Boyd):Yeah. You do unbelievably badly here because if A – if all the AI's are equal to V1 this one is ranked one. And it is by a long shot not invertible. Everybody see that? So the point is you're gonna be forced. You can't do a greedy thing and just choose high quality measurements. It's actually something where you have to choose all of them together. Does this make sense? And actually if you're confused it's probably because this is actually quite trivial and I'm just making it sound complicated, but, anyway, okay. That's – but I do want to point this out. Okay.

So you can write this this way. It's a vector optimization problem over the positive semi-definite cone and here's what it says. So obviously all the – it doesn't matter what order you choose the measurements in that's clear because all you're gonna do is sum these dyads here. You just sum the dyads. So it doesn't matter which measurement I do first. What matters if I have a thousand measurements to take I have a pallet of 50 choices. What matters is how many measurements of type one do I take? How many of type two and so on. And we're gonna call those MK. So MK is the number of measurements of choice K I make. And so I get this problem here. Now the MP's are integers and they have to be non-negative and, yeah, they add up to my budget. By the way, this is just the simplest version. I can also put a cost on each measurement.

I cannot only put a pallet of measurements in front of you. I can also put a price tag on every measurement and you can have a budget in money or time and you'd get a – then this budget would be something different. I can add other constraints to this if I want. I can add a time, money, all sorts of other constraints. This is just a simplest case where all the measurements are equal. You have a – you're allowed to choose [inaudible]. Okay. Now, this problem here – if you just say experiment design this is what experiment design is. It's this problem right here. Okay. And this problem in general is hard to solve, but there are some regimes where it's relatively easy to solve.

Actually, the feasible set is essentially the set of partitions of M. By the way, if M is really small, like three, then this is pretty easy to solve, or four or something like that. Then this is easy to solve. The other extreme is when M is very large because what you do is you rewrite this this way and you let lambda K be MK over M and this is now the fraction of the experiments of the M total budget you're gonna use of which you will take experiment type – measurement type K. So this fraction. Okay. Now, I haven't changed anything here. Actually the truth would be something like this. The real problem would have M lambda K is in Z, like that. Okay.

So this real problem – this is the problem that gets you back to that one. It says that the lambda's I chose have to be multiples of M is a thousand. I think multiples of .001, right?

So I comment this out because can’t handle it basically. You comment this out and you get – now you get something called the relaxed experiment design because I’ve relaxed this constraint. You all ready know that because – wait, is that on the current homework? I don’t know because we’re working on homework’s like one and two ahead. Are you doing something with a relaxation and Boolean variables now?

Student: Yeah.

Instructor (Stephen Boyd):Okay. Good. Then you know what relaxation is. Okay. So relaxation is commenting out the constraints that you can’t handle. That’s really what it is. There was a question? Yeah?

Student: Yeah. I was wondering if you can’t handle [inaudible] problem guessing stuff convex? It’s like your M is not convex.

Instructor (Stephen Boyd):The only problem – the only thing that’s not convex is that.

Student: Yeah. But you can’t handle it.

Instructor (Stephen Boyd):You could, but not in right not in what we’re doing in the class. No, in general, you can’t. So these problems can be very difficult.

Student:[Inaudible]

Instructor (Stephen Boyd):I believe it is, but don’t –

Student: Because it looks like [inaudible].

Instructor (Stephen Boyd):Yeah. I think I could make – I’m pretty sure I could make this NPR, but there’d be no – I’d go to Google and I’d type experiment design, NP hard and there’d probably be five papers showing aversions of experiment design for NP hard. I’m guessing, but I don’t know. Just to make sure.

Student: Could you also get rid of that constraint if we could use like a steptastic experiment design?

Instructor (Stephen Boyd):Yes. Okay. So that actually is an excellent question and so let me explain how this works. When you do a relaxation – I mention this to you because you’re doing one now. Okay. So the way you deal with relaxation is take very good points – good time to talk about it. I’ll say more about it later in the class, but when you’re doing your relaxation. Now, first let’s talk about – let’s say what the truth is. The truth is you have a constraint you can’t handle and you just comment it out. It’s kind of like the truth of why do you least squares? You do least squares because you can and you know how to. That’s why you do it, okay. Then you can construct all sorts of stories, which are – some of which are sort of true, not true. And someone says why are you doing this? You go oh, everyone, well, maximum likelihood asymptotically blah blah blah estimator Gaussian blah blah blah.

And they say is the noise really Gaussian? And you go, yeah, yeah, sure it is. Yeah. Anyway, you do – also, by the way, if you repeat those lame excuses often enough you'll actually start to believe them, right? So you'll actually – you start it off as just you couldn't handle it, but what happens is after you've been successful doing least squares for like 15 or 20 years you actually start believing it that things are Ga- anyway. All right. So this is a great example. You're doing experiment designs. You can't handle this, so you simply comment it out and you're gonna solve this problem here. Now, by the way, in this problem you could get very close because you can bound how – how far can you be off if M is a thousand? Right?

So the point is if you solve this problem with lambda being just a real number these are numbers between zero and one. It's a probability distribution. Each such number is at most .0005 away from an integer multiple .001. So you could actually bound that, right? In this case. So you could actually – you don't have to fall back on some totally lame excuse and something like that, but the alternative is actually better because it works universally. It was your suggestion and it goes like this. When someone says what are you doing? Well you say I'm solving this problem here and you go yeah, but this thing has to be an integer multiple of M. And you go oh, no, that's very unsophisticated. I'm actually doing a – what I'm really doing is starcastic experiment. It's like randomized – actually its exactly randomized detector.

And you go this is very sophisticated. This is how – this is like a randomized algorithm, randomized detector. This is much more sophisticated than just committing to 179 of type one, 28 of type two, and so on that add up to a thousand. You go no, no, no, this is much more sophisticated. I'm coming up with a randomized experiment design and I'm gonna come up with a probability distribution on the possible measurements and then what I will do is I will ask for a measure – each time someone asks I can carry out a measurement. I'll flip a coin and do a randomized experiment and I'll use those probabilities.

Now, of course, this is totally lame, this argument, but it often works. So I do recommend trying it and when you relax something and see if you get away with it because it just makes things easier. Did I answer your question?

Student: One other quick one. Can you do better like that?

Instructor (Stephen Boyd):What? With this?

Student: Using starcastic design can you do better than if you deterministics?

Instructor (Stephen Boyd):Can you do better? Oh, yes, because you've removed a constraint. So you always do better in some sense, right? I mean, the problem, the only problem with solving this problem and not the original one, is when someone comes along and says yeah, but I have a thousand measurements. You've gotta come up with some numbers that add up to a thousand and you go, yeah, well, 187.22. And they go what does that mean? I can't do 187.22 experiments of type one. I can do 187 or 188, which is it? So that's the problem. And you go no, but the 22 is more sophisticated because if you had to do a hundred thousand you'd be doing 18722 or something like that. So anyway. I mean, don't – in this case, if M is large you can bound how far off you can be and it's not a big deal.

In the Boolean case you're doing right now, by the way, in those it's often not the case that you can bound how far you're off. Okay. All right. So this is the problem. By the way, these things work really, really well. I should also add the same thing you're doing in your homework now or are doing or will be doing shortly. The same methods work for experiment design. They work unbelievably well. So if you were to – if someone actually said choose a thousand experiments to carry out from this palate of 20 you'd solve this problem, you'd get some lambda's, you'd just round them to numbers. Like, you'd say, okay, 187 of the first one. By the way, that would be a valid choice of experiment design, right?

And it would have a certain value of E, which is a covariance matrix. Okay. The relaxation gives you a lower bound and then you'd say someone could say is that the optimal? And then you could honestly say don't know, but I can't be more than .01 percent off. So it's good enough. So the same thing. You only know what I'm talking about if you've started on the homework, which is maybe not very many people, but anyway, we'll just move on.

Okay. So the other question is how do you scalorize the fact that it's a vector problem? Well, it is a vector problem. It's covariance matrices and, by the way, its experiment design. You get interesting stuff and it's just what you think. I mean, basically what happens is this. I'll just draw a confidence ellipsoid. You make one experiment design and you might get this confidence ellipsoid. Okay. You choose another blend of experiments – well, if you choose another blend of experiments and you get this it's very simple. The second was a better choice than the first. Okay. This is the clear unambiguous section, but, in fact, the way it really works is something like this.

You choose one set of experiments and you get that as your confidence ellipsoid and you can choose another set of experiments and you get this. Okay. And now you can say which suite of – which experiment design is better? The answer is it means that it's total nonsense. It's a multi-criterion problem. Now, by the way, if you're estimating X1, which one is better? Two. Obviously the second one is better, right? If for some reason you wanted to estimate something along this axis obviously – oh, sorry. That axis, then one would be better. Okay. So you have to scalorize this and there's lots of ways to scalorize it and each way to scalorize it, by the way, ends up with a name of experiment design.

The most common one by far multiple – many books written on it is D-Optimal Experiment Design. I'm guessing D comes from determinant, but honestly I don't know. So the D-Optimal Experiment Design you minimize the log depth of the inverse or you maximize the determinant of the covariance matrix. Okay. So that's what you do. As a beautiful interpretation of confidence ellipsoids you are minimizing the volume of the confidence ellipsoid. Okay. So that's the way – that would be the geometric interpretation. There's others. If you put a trace – if you minimize the trace of this which, by the way, has another beautiful interpretation. The trace, by the way, is the expected this. It's the expected value of X hat minus X true norm squared. So if you minimize the trace and that's called A-Optimal design is the other one and there's others.

There's E-Optimal design and they go on and on. Okay. So this is clearly – I mean, it's a convex problem because we're doing relaxed experiment design. It's a convex problem. You know what we should do? We should just do a homework problem on that. Now that they know what relaxations are. Just super simple one. Sorry, we're way behind on that. You should do that where you get the bound and then you do the design. Okay. We'll do that. So here I – actually I'll think I'll quit here, but just we'll look at an example just to see how this works and then I'll talk about this last business.

So here's some possible experiments that you can carry out. So these are the A's. So basically what it says is you're gonna measure – you're actually gonna make a linear measurement and I noticed that these things are farther out. They have a larger magnitude than these. That means that this is a set of sensors here that have higher signal to noise ratio than these measurements. Okay. Now, it should be kind of obvious what you really want to do, if possible, is to have high signal to noise ratio measurements which are orthogonal because if they're orthogonal it turns out it's gonna transfer – when you do the inverse the condition number is small and big will translate to small when you invert the matrix and so on. And so, sure enough, I think we add something like 20 possible experiments – the palate of experiments offered was 20, two were selected.

And the two that were selected make tons of sense. Didn't select any of these because these measurements kind of do the same thing, but with a better signal to noise ratio. So these were opted for No. 1. No. 2 it shows these ones that, sort of, were farthest apart from each other. The measurements that were farther apart from each other. By the way, if you do things like GPS and stuff like that these things will make perfect sense to you, right? These are the – it says you want to take measurements. In fact, you'd even call this like geometric gain or something like that is what you'd call this. So this is what happened.

So obviously in this case you did not need to solve a convex problem to be told that you should take high signal noise ratio measurements instead of low ones and you should, if possible, take ones that are nearly orthogonal. You don't need that. Trust me. You have a problem where you're estimating 50 variables, or ten even, and you have a hundred possible measurements. There is absolutely no way you can intuit what the correct mixture of what the right experiment design is and the result can be very, very good. But anyway, I thought we all ready decided. You'll find out. You'll do one. Okay. We'll quit here.

[End of Audio]

Duration: 77 minutes