ConvexOptimizationI-Lecture08

Instructor (Stephen Boyd):We'll start with an announcement. It should be kind of obvious anyway. You should start reading Chapter 5. So I'll go fast, not that fast, on our next topic, but you should be reading if you want all the details and things. You should be reading along with or even maybe ahead of us, Chapter 5. Okay. So we're really gonna start this topic of duality. I think last time I did nothing but say a few things about it. They were kind of incoherent, but maybe I'll make them coherent today. One way to think about duality is it's a way to handle the constraints by incorporating them into the objective. That's the basic idea. So let's see how that works and it fits in exactly with what I was talking about last time, which had to do with this concept of irritation versus value for a function and the idea of a hard constraint where you go from completely neutral to infinite irritation as opposed to a soft, something like an objective, and an objective, it's just a linear irritation function or something like that that means the larger the function is the more irritated you get and the smaller it is, the happier you get. That's gonna tie into the idea of duality. All right. So first we start with a couple of definitions. We start with a problem like this so minimize the objectives, some inequality constraints and quality constraints. We are not going to assume this is convex and in fact, some of the most interesting applications of the material we're gonna talk about now, occur when this problem is not only not convex, but it's a problem known to be hard. So that's gonna be some of the most interesting applications. We form a lagrangian and the lagrangian is simply this, I mean, it's really kind of dumb. You take the objective and to the objective, you add a linear combination of the constraint functions and the equality functions.

So that's it. Here you say, for example, that lambda three is the lagrange multiplier or dual variable or price, I'll justify that term soon, associated with the third inequality, so FI of X less than zero, that's what lambda three is. And in fact, you can even sort of get a hint at what's going on here. If, for example, lambda three were six that would mean the following: up here this says that if F3 is less than or equal to zero, it's perfectly okay; if it's bigger than zero, it's infinitely bad. When you replace this constraint by this charge, lambda becomes the price or you can call it a penalty and it basically says FI can be positive, that's not a problem now, but you're gonna pay for it, and for every unit that FI is positive, you're gonna pay six here, whatever those units are in. Now, the flipside is actually quite interesting. You actually get a subsidiary for coming in under budget. Up here, as we said last time, there's absolutely no difference between F3 of X, for example, being minus .001 and minus 100, absolutely no difference, both are feasible and you have no right to say that you like one better than another because it's not an objective, it's a constraint. Now, when you convert that one constraint to this sort of thing, something interesting happens. When F3 is less than zero, you actually get a subsidy because it's 6 times whatever margin you have, so when you convert this constraint to a term in a lagrangian, you actually get a subsidy for coming in under budget and you do like F3 equals minus 100 more than you like it minus .0001. You get a subsidy. I'm just hinting at what you're gonna see soon. That's a lagrangian. Then we look at the so-called dual function or the lagrange dual function, it's got other names, but let's look at it. So the lagrange dual function is actually very, very simple. It just says minimize this lagrangian over all Xs. That's what it says. So you just minimize the lagrangian. Now, actually it's very interesting because what it’s really saying is this, and by the way, it’s a function now of these prices or dual variables. All sorts of ways to interpret this. You can interpret this as sort of the free market approach or something like that.

This is the constrained approach where you simply say by fiet FI of X has to be less than zero, HI of X has to be zero then you could say, no, you know what, we’re gonna let FI float above zero, no problem, we’ll charge you for it the lambda I’s are positive. That would be the meaning here. But if FI goes less than zero, we’ll actually pay you for it. You’ll be subsidized and this is sort of the optimal cost under these sort of market prices. All of this will become clearer as we move on with this. Now, here’s an interesting fact. If you look at this function as a function of lambda and nu, what kind of function is this as a function of lambda and nu for each X. It’s affine. Yeah. It’s linear plus that’s a constant. Okay. So it’s affine. And the [inaudible] of a family of affine functions is of course concave. So this function G, you can think of it as the optimal – we’ll get to it later – as a function of the prices, even if the original problem – even if these things are not convex, sorry, these are affine, that’s not convex. This dual function is absolutely concave so that’s – all right. Now, we get to something very simple, but it’s one of those things where you get a sequence of 12 simple things and you know the right sequence of 12 simple things will lead you to a very interesting thing. So trust me, that’s what we’re doing here. It looks very complicated, it’s quite profound. It says, basically, if you can evaluate G as a dual function, you’re gonna get a lower bound on the optimal value of the original problem. That’s what it says. So the argument is just embarrassingly simple. Look at this thing and imagine X is feasible. For any feasible X FI of X is less than or equal to zero, but if lambda I is bigger or equal this term is less than or equal to zero so therefore, this whole thing is less than zero. If you’re feasible, HI of X is zero so it doesn’t even matter what the side of new I, this is zero and, therefore, it says that the lagrangian is less than F0 of X or any feasible X.

Now, for infeasible Xs that’s false, but for feasible Xs it’s absolutely true that L of X, lambda nu is less than or equal to F0 of X because that’s zero and that’s less than or equal to zero. By the way, note the level of the mathematics being employed here. It’s quite deep. It relies on the very deep fact that the product of a non-positive number and a non-negative number is non-positive, and also that you can add such numbers up and you still have something less than or equal to zero. So I just want to point out nothing has been done here. It’s just embarrassingly simple. Okay. Now, if you then [inaudible] this over X well then obviously it’s less than F0 of X tilde. This is true for any feasible X tilde and there’s no conclusion possible other than this. Okay. So let’s look at some examples. Let’s do least norm solution of linear equations, so here’s – now, of course this is stupid. We know how to solve this problem analytically, you know how to, it doesn’t even matter what the solution is. It doesn’t matter. We know everything there is to know about this, but just as an example let’s see how this works. Well, the lagrangian function is this; 2X transpose X, that’s the objective, we add new transpose AX minus B so here, by the way, you have to write this as AX minus B = zero. You have to decide what H is. H is either AX minus B or something that’s B minus AX so all that would happen there is the sign on nu would flip or something, but it wouldn’t matter. I’ve written it this way, AX minus B = zero so. We’re gonna minimize this over X, that’s completely trivial because that's a convex quadratic, that's affine in X and you just take the gradient, you get 2X plus A transpose nu = zero and this is the optimal, that's the X that minimizes the lagrangian here. All right. Now, we take that X and we plug it back into here to get G so when you plug this in here that's the X that minimizes the lagrangian, you get this and you get some – well, first of all, let's take a look at it.

Evidently, this function here is concave because that is a positive semi-definite quadratic form. All we care about is a positive semi-definite quadratic form, but it happens to be positive definite. No, actually it doesn't matter in this case because I'm making no assumption about A whatsoever in this case; everything is true no matter what I assume about A. Okay. So this whole function is concave quadratic so there it is, which we knew had to happen because the dual function is always concave. Now, here's what's interesting. And we've already learned something. It's not a big deal because I know how to solve this problem, but look at this. What it says is the following; if you come up with any vector nu at all, which is I guess the size of B, the height of B, let's call that M, you come up with any vector nu at all and you simply evaluate this function then whatever number you get is a lower bound on the optimal value of this problem. In this case, it's totally useless. If that's a small problem – say, X has a thousand variables or something like that, this is goofy because we know how to solve this problem extremely efficiently, get the optimal solution, we don't need a lower bound on it. This is actually immediately useful. Let's look at a standard form LP. So I want to minimize C transpose X subject to AX = B X figured or = to zero. Just standard LP. Okay. The lagrangian is C transpose X plus, and I'm gonna add a lagrange multiplier for this, that's nu, transpose AX minus B and then here – to put this in standard form you really want to write it. You have to write it as minus X is less than zero so these are the FIs over here. Okay. So if I simply form this thing and that explains the minus sign here on the lambda. Okay. So that's your lagrangian. What kind of function is the lagrangian in X? Hey, look at that, that says linear doesn't it? And it seems to me that that is false, I believe. That's the name for something that's not true. Isn't that correct? You agree with me? The lagrangian is not linear in X. That's ridiculous. Is that affine? L is affine in X. It's affine because there's this constant term minus nu transpose B here. Okay. Now, this leads us to something. It's actually related to something on the homework from this week, which was – here's a very, very simple linear program. Ready for it? No constraints. How do you minimize an affine function? How do you minimize an affine function? More than a small number of people were stumped by this because it's such a stupid question, I think. Actually, it's not a stupid question, sorry, it's just that they wanted to make it more complicated. How do you minimize a linear function?

Student:

[Inaudible]

Instructor (Stephen Boyd):What's that?

Student:[Inaudible]

Instructor (Stephen Boyd):What’s the minimum of a linear function? Just in R2 I have a linear – what’s the minimum of X1 minus X3?

Student:[Inaudible]

Instructor (Stephen Boyd):X2. Sorry. Go ahead. What is it?

Student: It’s minus infinity.

Instructor (Stephen Boyd):Of course. It’s minus infinity. You have level curves which are hyper planes and you just go as far as you can in the direction minus C, it’s minus infinity. So there’s nothing, there’s no mystery here and what’s the minimum of an affine function. There’s no mystery here. Okay. By the way, notice that that’s a valid – so if G is minus infinity, it’s cool, it’s just minus infinity and note that it is indeed a lower bound, however it’s an informative lower bound because if somebody comes up to you and says, “Can you help me get a lower bound on this,” you can just automatically say minus infinity. Exactly one. What’s that? When the function is zero. And that’s it. That’s the only way. So in fact, if you minimize this over X, you will get minus infinity. One exception. If C plus A transpose nu minus lambda is zero then this whole thing goes away and you get that. So the dual function for the standard form LP is a very interesting function we’re going to look at it very closely. It’s this. It is a function which is linear on this weird affine set and it’s minus infinity otherwise. By the way, that function is concave. Right? You can just visualize it. I mean, it’s kind of weird. It would be an R2 like my drawing a line like this and here are the values, 0, 1, 2, -1 and then you have to visualize this so if you want to make a graph coming up. Slope this line up and now what I want to do is everywhere off that line the function guy is minus infinity so just make it fall off a cliff everywhere else. That function is concave. Well, it has to be concave because we know the G is always concave. Okay. But if you want to visualize it, that’s what it looks like so it’s a weird thing, it’s linear, but then off this thin affine set, it falls to minus infinity. Okay. So that’s what G is. Is there a question?

Student:[Inaudible] LP is affine [inaudible]?

Instructor (Stephen Boyd):Okay. So actually the question is this function, which I write this way. You just have to blur your eyes because I’m asking you what kind of function of X is it? Okay. So let’s look. That’s is a constant. That is a constant vector, but well, if I include the transpose here, it’s a constant roe vector, therefore, this is linear in X, I add a constant so it’s affine. Does that make sense? Okay. Okay. Now, this is really interesting. We can actually say what the lower bound property is. So the lower bound is this. This function is the dual function. It is always the lower bound on that LP. Now, of course, if you randomly pick lambda and nu, you’re gonna get minus infinity. Just minus infinity. Okay. In which case it’s still a valid lower bound, but it’s just a completely uninformative lower bound. It’s the lower bound that works for all problems. It’s the universal lower bound. So let’s see what we’ve come up with. It says the following. If you have this linear program here and someone says what is the optimal value of it, well, it depends on the context. If a person has a specific A, B, and C, and actually just wants to know solve my problem, you can run some code and solve the problem if it’s not too big and if that’s what they’re interested in. Okay. However, you can make a very general statement. You can say the following. If you find any vector nu by any means, it doesn’t matter how you find it, it’s no one’s business how you found such a nu, if you found a vector nu such as A transpose nu plus C is a non-negative vector, then you evaluate minus B transpose nu, that’s the lower bound on this LP. That strings together three or four totally obvious things, but I think you come up with some – that’s not obvious. Okay. Let’s do a quality constrain norm minimization. So here you minimize the norm of X subject to AX = B.

By the way, we’ve seen that – in fact, I guess we just did that problem two examples ago – not quite. We did the case where this was norm squared and where this was the two norm, now, it’s completely general. Well, the lagrangian is the [inaudible] of norm X minus and then some horrible thing here, which is affine and here we have to be able to minimize norm X minus nu transpose AX, this is a constant so it’s totally irrelevant and you have to be able to do that. Now, this goes back to the idea of a dual norm and let me go to that so let’s look at that and if I want to minimize this thing – the question is what is this thing, what do you get here, right? And the answer is actually pretty [inaudible] straight from dual norms. In fact, we can do it for the two norm first just as a warm up. So for the two norm you’d say, well, look, if norm Y is bigger than one in two norm then I can align X with it in that direction and then this thing sort of over powers, it has enough gain to overpower this one and I can make it go to minus infinity. Okay. Now, on the other hand, if norm Y is less than or equal to one [inaudible] tells me that this thing is less than that and, therefore, this whole thing is bigger or equal to zero. So I could never ever make this thing negative. On the other hand, by choosing X equals zero, I can make it zero, so that’s clearly the optimal.

So this is equal to and this generalizes now to a general norm. It’s either equal to minus infinity if the dual norm of Y is bigger than one or at zero otherwise. Okay. And in fact, that’s the dual norm. It’s from the definition of the dual norm. So this is what you get. All right. So applying this up here gives us exactly this. This is our Y and here you have it so once again, this dual function is not totally obvious. I mean, it’s not an obvious thing. It’s something, it’s linear, it’s a linear function, but it’s got a weird domain and in this case, it’s the set of points nu where the dual norm of A transpose nu is less than or equal to one. Okay. That’s that. Okay. And then you can go through the argument here and I won’t go through it, but actually now you’ve got something totally non-obvious. It basically says if you can come up with a vector nu, for which A transposed nu is less than or equal to one in dual norm, then B transpose nu is a lower bound on the optimal value of this problem. Here’s an example: ready for a dual feasible point? Nu equals zero. Well, let’s check. That’s zero. Zero is definitely less than one and now we have a drum roll to find out what lower bound we’ve come up with. The lower bound is zero and that’s actually not particularly interesting here because the objective is zero. Okay. So that’s what it says here. Okay. Okay. This gives you a parameterized lower bound. It’s parameterized by nu. Okay. Now, we’re gonna look at a problem and we’ll see it a whole bunch of times. It’s actually just a simple example, it’s a perfectly good working example of a hard problem. It’s two-way partitioning. It’s embarrassingly simple. It goes like this. I want to minimize a quadratic form subject to XI squared equals one and this means XI is plus or minus one. Okay. And let me first just say a little bit about this problem. We're gonna see it a lot and just so you get a rough idea of what it means.

So XI is plus minus one so we can really think of this as the following is you have a set of points, like, M points and what you want to do is you're gonna partition them into two groups. Okay. That's one group and then the other group will be this, okay. And we encode that by saying here's where XI is plus one and here's where XI equals minus one. So we're gonna use the variable here, which is XI, which is the plus minus one vector, to basically encode a partition. It's a partition. It's just that it's a numeric data structure to encode a partition, okay. All right. Let's look at what the objective is. The objective is sum XI XJ WIJ and let's just see what it is. So you sum over all pairs. If XI and XJ are in the same partition, what happens, what is XI XJ?

Student: One.

Instructor (Stephen Boyd):One. Okay. And then you add this to this thing. Now, this is something we want to minimize. Okay. Now, if XI and XJ are in opposite partitions, this is negative so I think that means that WIJ is a measure of how much I hates J. Did I do that right? I believe so because if W is very high, it means that if XI and XJ have the same side, you're gonna be assessed a big charge in the cost. I mean, if XI is high and they're in opposite things, you're gonna decrement the cost a lot and happiness is gonna go up. Okay. So WIJ is basically how much I annoys J, but it's symmetric so it's the average of how much I annoys J and J annoys I. Okay. Now, if WI is small, it means they don't care much. So in fact, this now makes perfect sense as you have a group of people, you have social network or something like that and you want to partition it. There would be obvious ones. If the sign pattern in W were such that like everybody liked everybody except one then it would be very simple; you'd just isolate that one nod. But in general, actually finding a solution to this problem is basically extremely hard. You can't do it. So if this was a hundred or a couple of hundred, you can't do it. It just cannot be done. Okay. So that's the partitioning problem and for us, it's gonna be a canonical example of a hard problem. By the way, there's instances of it which are easy, I just mentioned when there's some obvious solution, but I'm talking about the general case here. Okay. By the way, it comes up in tons and tons of other – my interpretation was sort of a joke, but the point is it comes up in tons and tons of real applications, it comes up in partitioning, it comes up in statistics, I mean, just everywhere. So my interpretation was a joke but it's a very real problem with real applications. Okay. Now, the dual function is this. We simply take X transpose WX, we add as the lagrangian tells us to a linear combination of these functions. I write them as XI squared minus one so I get this and I have to calculate – this is the lagrangian and the lagrangian is quadratic. Okay. It's a quadratic function. We're gonna have a very short discussion about how do you minimize a quadratic function. The first thing you do is let's talk about how do you minimize a quadratic form? So what is the minimum of a quadratic form so what is the minimum of a quadratic form?

Student:[Inaudible]

Instructor (Stephen Boyd):What is it?

Student:[Inaudible]

Instructor (Stephen Boyd):It can be negative infinity. I agree. When would it not be?

Student: Within [inaudible]

Instructor (Stephen Boyd):If the quadratic form is positive semi-definite then the only values it takes on are non-negative so it couldn't be minus infinity then so that's exactly the condition. The minimum of a quadratic form is minus infinity if that matrix is not positive semi-definite so if it has one negative eigenvalue, the minimum is minus infinity. Okay. Otherwise, if it's positive semi-definite the minimum is zero because it can't be any lower than zero and it can be zero by plugging in zero. Okay. Let me tell you what the lagrange dual function is for you right now. It is a lower bound on an optimization problem, gives a lower bound. It of course can give you – it's parameterized by lambda and nu by these dual variables. Now, in some cases if you plug in some lambda nu you're gonna get the following lower bound minus infinity. But it's always a lower bound. In some cases, you plug in lambda nu and you're gonna get actually an interesting lower bound that's not obvious. So right now, you should just think of it as a lower bound. Lower bounds can be good, they can be tight, they can be crappy, we're gonna get to all that later. Okay. Okay. I just want to tie together the idea of the [inaudible] function and lagrange duality. So if you have a function with just linear inequality and equality constraints, a problem, and you work out what the dual function is, it's a minimum of F0 plus – I collect this together and multiply by X and then that's, of course, a constant. And what this means is the following. If I focus on this and then go and look up what the conjugate function is, which was the conjugate over X of Y transposed X minus F of X, that's the dual, if you plug in also all the right minuses, you get this. It's equal to that. Now, what that means is the lagrange dual function of this thing is exactly equal to this.

It's equal to that. Now, recall the conjugate functions often have domains that are not everything. It was actually the probability simplex was the domain of it. So that'll automatically impose some inequality constraints in here when that happens, but here's an example. The maximum [inaudible] problem is maximize sum minus XI log XI subject to some inequalities and equalities. And by the way, that's already a really interesting problem because it says lots of things. It says find me the maximum entropy distribution that has these expected values – these are just known expected values. These can be moments, it could be probabilities, it could be anything and these are inequalities on expected values. So it's really quite a sophisticated problem to ask. What's the maximum entropy distribution, for example, on these points that, for example, has the following variance and has the probability in the left tail less than that. You could go on and on and make it a very sophisticated thing. That's the maximum entropy problem. That's this thing. And if you work out what that is when FI of X is the negative entropy here, that's minimized negative entropy, you will actually get the sum of exponentials. So the dual function for a maximum entropy problem is gonna involve a sum of exponentials. Now, if you're in statistics – and I said statistics not probability, this will be very familiar to you because it's a connection between exponential families and maximum entropy and we'll see more of this later. Just a hint. Okay. Now, we get to the dual problem and to write down the dual problem – I mean, it’s the dumbest thing ever. If someone walks up to you and says, “I have a lower bound on my problem, but it’s parameterized by this vector lambda and this vector nu,” and then you say the only interesting thing about lower bound is, “Well, that it’s a lower bound,” and if someone has multiple lower bounds, obviously the higher the lower bound, the more interesting it is.

So you can just say okay, what’s the best lower bound, on the original problem, that you could establish by lagrange duality? What is the best lower bound? We don’t know if it’s sharp. We’re just saying what’s the best one and it kind of wouldn’t make any sense to really examine any other anyway. All right. That leads you just to this problem right here. Now, I want to point something out. This is always a convex optimization problem no matter what the primal problem was. Oh, by the way, this is called lagrange dual and sometimes it’s just shortened to the dual problem here. In fact, people say “the dual problem,” the same way we say “the optimal point,” even in situations where we don’t know that there’s an optimal point. We’re actually gonna see this multiple dual the way the word is used on the street. There’s lots of dual, but we’ll get there. For now, it actually really is “the lagrange dual problem.” And it says simply maximize the dual function subject to this. The subject to the lambda’s being positive, that’s all. Okay. Now, often what happens is G is minus infinity for some values of lambda and nu. We’ve already seen that a couple of times. That is a not interesting lower bound and it’s sure not gonna help you maximize something. To find a point where it’s minus infinity, you know, this thing could actually be minus infinity, that can happen, but the point is it’s not an interesting value. So in fact, often what happens is you pull the implicit constraints in G out and make them explicit. Okay. Now, here for example, let’s go back and look at this. The dual function for this LP is this weird thing that looks like this. I drew it somewhere. It was this sick thing here where this thing is kind of going up on a line, but off that line, the thing falls off to minus infinity and we’re just simply going to maximize that subject to lambda positive; however, it’s easier to simply take the implicit constraint out and you end up with something that looks like this.

Okay. So here’s the so called standard form LP and then this is what it looks like when you have actually pulled out this implicit constraint. Technically speaking, this is not the lagrange dual of that. However, people would call this the lagrange dual so you’re given a little bit of license to form the lagrange dual and do a little bit of trivial rearrangement and people would still call it the dual or something like that. This is equivalent under a very, very simple equivalence of this. The lagrange dual of this thing is that where G is the sick function. I just want to point this out. Okay. And by the way, let’s see what happens here. That is also an LP and let’s see what it says. Here I can say something about this problem. If you have a feasible nu here then minus B transpose nu is a lower bound on the optimal value of this problem. This thing says, “Okay, you have a family of lower bounds, please get for me the best lower bound.” That’s what the meaning of this problem is. This also has beautiful interpretations in a lot of cases. So for example in engineering design, it’ll make a lot of sense. X will be a sub optimal design, for example, sorry, any X here, if it’s feasible, it satisfies the constraints, but something that’s feasible here would be a sub optimal design. Okay.

The nu will have a beautiful interpretation. A dual feasible nu in that case is a certificate on a limit of performance. That's what it is. That's the meaning of nu here. Okay. Actually, we'll see that when you look at a real problem, it will have physical significance. We'll get lots of examples. If this is a question of how bad could something be bad, if, let's say you're at a bank and they want to know, okay, what's the worst thing that could possibly happen, then a lower bound actually gets interesting. Yeah.

Student: Could you please say why that was not, technically, a lagrange dual on the right?

Instructor (Stephen Boyd):This?

Student:

Yes.

Instructor (Stephen Boyd):Sure. That's the lagrange dual, that's why.

Student:I mean, relative to the LP.

Instructor (Stephen Boyd):No, they're not the same thing. They're not the same thing. This is minimizing a function, which is a weird function. It's equal to minus B transpose nu provided that A transpose nu plus C minus lambda equals zero, okay, something like that and it's minus infinity otherwise. Okay? So that's what this is. And by the way, if you were very careful, it would make a different. Let me explain that. For example: suppose I throw in a nu for which A transpose nu plus C is not a non-negative vector, okay? Then in this problem when I say how about this nu, how do you like that, what is sent back to me is actually the infeasible token is sent back to me saying your nu is infeasible. Okay. Over here, it's actually more interesting. Over here, if I throw such a nu in or whatever, what comes back to me is the object function sends me an OOD token, Out of Domain. Now, that's a concave function and that means it's minus infinity. You get two slightly exceptions are thrown in this thing. But I want to point out that these are just – you can call this just silly semantics and all that if you like, but it's very important to understand these are not the same problem. By the way, don't focus on these minor things. That's something you can think about, you can read this, think about it on your own. Don't let silly little technical things get in the way of what the picture is. The big picture is you have an optimization problem and you form another one called the lagrange dual. That lagrange dual problem, essentially, is saying what is the best lower bound on the optimal value of the first one I can get using the lagrange dual function. That is what's important. Okay. So now we get to the idea of weak and strong duality. Now, weak duality says that "D" star is less than "P" star. Now, here, let me see how this works. Okay. So in this context, the original problem is called the primal problem and the lagrange dual is then called the dual problem. Okay. So that's the primal and the dual and we'll call – well, we've already assigned the symbol P star to mean the optimal value here. We're gonna let D star be the optimal value of the dual problem. Okay. So optimal value of this is gonna be denoted D star. You always have D star as less than P star.

Why? Any dual feasible point is a lower bound P star so the best one is also a lower bound. This is called weak duality. It's called weak duality because let me review the deep mathematics required in establishing this, right, it hinged on properties such as the product of two positive numbers is positive in the sum of positive numbers.

So it's weak because you can explain it to somebody in junior high school. I mean, they might not have taken those 14 steps, but the point is it has nothing in them that's hard so it's called weak. Okay. All right. That's weak duality. All right. It always holds. Convex, non-convex. It's absolutely universal. It could be stupid. You could, indeed, have D star equals minus infinity in which case your best lower bound is of no interest what so ever. Okay. That can happen, but this is always true. Okay. Now, if we go to the partitioning problem and we ask what is the best lower bound on the two-way partitioning problem you can get from the lagrange dual you will form this problem. That is a semi-definite program. And now, things are interesting because, although this is something that was not known 15 years ago, and absolutely inconceivable 20 years ago, I can tell you this, this SDP, you can solve it, people can solve it. You can solve it like that for a thousand variables. No problem here. And if you knew what you were doing you could go, easily, to problems with 10,000 and 100,000. The point is, you can solve this SDP and you will get a lower bound on the two-way partitioning problem. That is fantastically useful if you couple that with a uristic for partitioning. So you do some crazy uristic, there's lots of uristic; some of them work really well by the way. Now, you don't expect it to work all the time because you are solving, after all, an NP hard problem in general, so you don't expect it to work well all the time, but what happens is you'll do a partition and you'll say, "Here's my partition and here's the number I got," whatever it is. It's the X transpose WX and you want to know could there be a better one. You can solve this SDP and in fact, you'll see in a lot of times the numbers are pretty close. Okay. At least it's a good thing to know. You would know I have a partition, but there's no partition that's more than – I'm at most such and such sub optimal.

And you might just say, okay, that's good enough. All right. Okay. Strong duality, this will not rely on junior high math, okay. Strong duality is going to be that that lower bound is tight. That says, there's a lower bound that goes all the way up to the optimal value. That's strong duality and we'll see what its equivalent to, but that is not trivial. And by the way, it often doesn't happen. Okay. So in two-way partitioning problems, by the way, if it were true there, you'd have P = NP because this problem we can solve in polynomial time and so in fact, if P star were equal D star – and in fact, there's even approximation. If you know about complexity and you have something that's not even approximable or something like that then that tells you that you can't even get something where you can bound the gap or something like that but I won't go into that. Now, here's an interesting part. When a problem is convex, you usually have strong duality. Okay. So that's actually amazing. That's gonna actually have a lot of implications. It's gonna be equivalent to, by the way, it's gonna involve the separating hyper plane something. We'll see what it connects to. There are multiple books, multiple courses, not here, but at some other schools; you can take entire courses, read books, thousands of papers that elaborates on this one word, usually. Okay. Now, basically these are called constraint qualifications. So a constraint qualification theorem goes like this. It says if the primal problem is convex and then you insert your constraint qualification here, okay, then P star equals D star. That's a constraint qualification. You could devote your life to this. On occasion, these issues actually do come up, but maybe less frequently in applications than the people who devote their lives to it would like to think. I'm saying that of course because their grad students will watch this and then alert them to it.

So I'm just making trouble. Now, by the way, if you're in this industry, sub industry of constraint qualifications, then this like the big, the sledge hammer, the most unsophisticated one there could be possibly be, this is the basic one that everybody knows. Okay, this is the least squares or something like that of the constraint qualification worlds, its Slater's Constraint Qualification, although, actually, the correct name here would probably be Russian, but we won't get into that. So let's call it Slater's Constraint Qualification and it says this, if you have a convex problem like this, it says if there is a strictly feasible point, if there exists one, then P star equals D star. Strictly feasible means not just that you met the inequality constraints, but you do so with positive margin for each one. That's the condition. Okay. Now, I should add that basically, it's completely clear, that for most problems that covers everything in engineering, pretty much, I mean, as much as people would make fun of Slater's Constraint Qualification and give you reasons and they could make examples up why it's not sophisticated enough and sure enough, there are problems where you don't have a strictly feasible point, but for most problems that come up in engineering, anything in machine learning, pretty much anything, this makes perfect sense, right.

For example, if the third inequality was a limit on power, it doesn't make any sense to say – just think about it, right? If Slater's condition failed to hold, it means their existing circuit dissipates 100 milli-watts, but there's no circuit that dissipates 99.999999 because if there were, Slater's condition would hold. Everybody see what I'm saying here? If solving that problem relied on these most fantastically subtle facts as to whether strict inequalities held or weakened equalities and one, but not the other held, then I got news for you, you're not doing engineering, you're not doing statistics, you're not doing economics, you're doing something like peer analysis. Okay. So that's my little story on it. Again, there are actually cases where these come up in practice, but they're pretty rare. And mostly, I'm saying this to irritate at other universities, my colleagues, who will be alerted to this, watch this tape and be very angry. But I thought I'd mention this. Okay. All right. So let's go to the inequality form linear program. Here you want to minimize C transpose X subject to X less than B. G of lambda is C transpose X plus lambda transpose AX minus B because I put the B on the left-hand side to make this F less than zero. I do this and I infamize this, but we know how to infamize a affine function. You get minus infinity unless the linear part vanishes so I get this and so this is the dual problem. Notice this is actually not the dual problem. So if there's lawyers present, you would say, "This is a problem that is trivially equivalent to the dual problem," okay, but after a while if there are no lawyers present you'd just say that's the dual problem or something like that. So that's it. Okay. Now, Slater's condition says that if the feasible set – of course the feasible set is a polyhedron and by the way, one possibility is the feasible set could be empty, which in fact, is a polyhedron. What Slater's condition says geometrically is very simple. It says if that polyhedron has non-empty interior, that's what this means, it means, basically, that there’s an interior point, if it has a non-empty interior then you have strong duality so you have P star equals D star. Okay. So that’s the picture.

Let’s look at a quadratic program. Let’s minimize X transpose PX subject to AX less than P. That’s minimizing quadratic form over a polyhedron, the dual function is this X transpose PX and we’re gonna assume P is positive definite. Actually, that’s so that I can avoid the horrible way to write down – it’s not that big of a deal, but the horrible to infamize a general quadratic function with a linear term because I don’t feel like doing it so this will work out. So here the dual function is you infamize over XX transpose PX plus lambda transpose AX minus B here like that and now I minimize over X. Now, the nice part is P is positive definite so I know how to minimize this. It’s P inverse times whatever something. I’m not even gonna do it because it’s easy to minimize a strictly convex quadratic function so I minimize it. I plug that X back in here and I get this thing, okay, which is I get minus one quarter lambda transpose A, P inverse, A transpose lambda something or other and my dual problem then looks like this. By the way, this really is the dual problem because in this problem, up here, notice that the dual function, the domain is all of our – let’s call it RM, it’s all of RM. Okay. So in this case, the dual function is domain is everything, which is to say, you get a lower bound for any – if you plug in random numbers lambda and you’re not gonna get a trivial lower bound. Okay. You might get a rather stupid one. For example, you might get the lower bound minus seven. Let’s talk about the lower bound minus seven here. Why is the lower bound minus seven valid for this problem? Because the objective is always non-negative, but the point is, you get a lower bound and you get this. So that’s the dual problem. And by the way, what we’re saying here is not obvious at all. What we’re doing is we’re saying, you want to solve this quadratic program – we haven’t yet told you how to do it or how it’s done or anything like that, but we’ll tell you this, if you come up with any vector lambda that’s non-negative and you evaluate this concave quadratic function, you get a lower bound on the optimal value of this thing. This has lots of uses. For example, suppose someone says I know how to solve this problem and you say, “How did you do it,” and they go, are you joking, – that’s, like, “If I told you, I’d have to kill you.” I’m patenting it right now in [inaudible]. Okay. I can’t tell you how I did it.

And you say, “Well, why should I believe that that’s the optimal X, how do you prove it? You say, “Well, watch this.” You say, “Check out this lambda, notice that it’s bigger than or equal to zero,” and you go, “Yeah,” then you evaluate that number and that number is equal to the value up here of the point. That, by the way, ends the discussion. That X is feasible and by the way, you would call that lambda a certificate proving it. Everybody got this? And notice that you didn’t have to say how you did it. Everyone got this? And then you’d say, “Hey, how’d you get the lambda,” and you go, “Like I’m gonna tell you that either.” Now, Slater’s condition says the following: If this polyhedron has non-empty interior, then these are absolutely equal then their always exists a certificate proving optimality of the optimal X. Always. So okay. By the way, a very small number of non-convex problems have strong duality. I’m not gonna go into it because it’s complicated and so on. This is actually covered in an appendix of the book and I would encourage you to read it. This one is not obvious. And actually, there’s a whole string of these. There’s, like, 15 of them or something like that and they’re just weird things that have to do with specific problems that are non-convex and just happen for deeper reasons to have zero duality gap.

The quadratic ones are the ones we collect at the end of the book in one of the appendices. There are others, you will see them, they’re kind of weird and some of them are quite strange. One I’ve seen recently where it involves complex polynomials of degree four. Right? And then something that should have zero duality gap and it comes down to something in algebraic geometry, but that’s always the way these are. These are not simple. This is just to say there are non-convex problems with zero duality gap. A few. Okay. Let’s look at the geometric interpretation. All right. So let’s see if we can do this right. So we’re gonna do a problem with just one constraint so what we’re gonna do is we’re gonna minimize – I’m gonna write the graph of the problem. What I’m gonna do is for each feasible X or each X in the domain, I’ll evaluate this pair. So although the problem may be happening in a hundred dimensions, for every X, I’m gonna plot a point which is in this plane; and one, basically, this tells you the objective value and this tells you the constraint function. So, basically, everything over here corresponds to feasible. Okay. And then the height corresponds to the objective value, so quite obviously, that’s the optimal value. Any point that ends up being colored there is optimal. Okay. So that’s the optimal value, P star. Everybody see that. So that’s the idea. So that point really has a very nice objective value, but it’s infeasible because it’s constraint function is positive. Okay. So that’s P star. Now let’s see what the dual is. How do you get lagrange duality in this picture? Well, lagrange duality works like this. You minimize F0 plus lambda F1. Now, on this plane, that corresponds to taking something here like this an it’s got a slop of – is it one over lambda or something like this, let’s see, it’s slope minus lambda so I take something like that.

So for example, if you fix lambda and then ask me to evaluate the dual, what you do is this. You fix a slope here and you march down this way until you just barely leave this set, and that would be right there. Okay. And then when you work out what G of lambda is, it’s this intersection here. Okay. So this is G of lambda and now the dual problem says, “Optimize over all lambda,” so if lambda is zero, you get this. You go down there and G of zero is this number right here, which is indeed a lower bound on P star, it has to be. Okay. Now, I crank up the slope and as I crank up the slope G is rising and it keeps rising until you just hit here, this point, at which point here its right there. Okay. Now as I keep increasing lambda what happens is the optimal point is actually here and this thing is rotating around – it’s not a fixed point, it’s rolling the context, but because it’s got sharp curves, it’s just rolling just slightly. It’s rolling along here and as I increase lambda, G gets worse and worse. In fact, if lambda is huge, it looks like this and G is very negative. It’s still a lower bound, just a crappy one. Everybody see this. So D star is that point. Questions?

Student:[Inaudible]

Instructor (Stephen Boyd):We’re gonna talk about that, but it depends very much – so for example, in a non-convex primal in two way general partitioning problems, NP is hard, but the dual is a SDP. That’s easy. In that case, it can be infinitely far away. Now, in the case of a convex problem, now it gets interesting. So in a convex problem, you will see later that they both solve the problem and a lot of people get all excited and they go, "Oh, how cool, I can solve my problem by the dual." It turns out that if you really know what you're doing, the complexity of the primal and dual are equal if you really know what you're doing. You will in about four weeks. Three. Whatever it is. Yes?

Student: How did you rule out the bottom point for P star? You can't just say it was that.

Instructor (Stephen Boyd):How did I do it?

Student: Yes.

Instructor (Stephen Boyd):Well, the first thing I asked is I asked – this shows you the objective and the constraint function for every possible point in the domain, okay, now, that points not good, for one thing, it's got a high objective, but it's also infeasible. Anybody who landed on the right here is infeasible. So in fact, these are very interesting, but they're not relevant as far as the optimization problem is concerned so we simply look at these. Now, every point that got shaded in here is feasible. Okay. The height tells you the objective value and so you want the lowest point among these. That's clearly right there and you go across here and that's P star.

Student: Why would the right-hand be infeasible?

Instructor (Stephen Boyd):Because your first coordinate here is your constraint function and F1 has to be less than or equal to zero. That's what it means to be feasible. Okay. So that's the picture. So here you have a gap. By the way, this thing strongly suggests something very interesting and you can see why convexity of the problem is gonna come in. When F1 and F0 are convex, this weird set G – now, what I'm about to say is actually not true, but it's close to true – that weird set G is convex, okay, when something is convex, you have a gap here because this blob is non-convex so if this thing had to be convex you can't have a gap. Everybody see this? That's what is gonna happen. Now, I'll tell you the truth. G is actually not convex, but its lower left corner, which is what we care about, is. Now I've corrected it and said the truth. By the way, you can also see how Slater's condition works so if you take not G, but A, that's the set of points that you can meter beat in a bi-criterion problem, so basically, if you take A then color in all these points here and now you can see A will actually be convex if that's convex and that's convex so A will have to look like this. Slater condition says that somewhere A goes a positive amount or it goes into the left side. These are the kinds of things you would study in a one of these whole courses on this topic. So that's the idea. So you can even get how Slater's condition connects to all of this. Okay. I'm gonna mention one more thing. We'll get to one more topic. It's a complimentary slackness. So let's assume that strong duality holds and actually, I don't care if the problem is primal or feasible. Okay. Convex. What I said made no sense whatsoever so let's start over. What I meant to say was I don't care if the primal problem is convex. That's what I meant to say, but it just came out a weird permutation. Okay. So I don't care if the primal problem is convex, of course the dual problem is always convex. So let's assume strong duality holds and let's suppose X star's primal optimal and lambda star and nu star are dual optimal. That says this. By the way, this is basically what it comes down to, it says X star is an optimal point, lambda star and nu star, you can think of then as a certificate establishing optimality of X star. Okay. By the way, these ideas, we're gonna use them from now on. They're gonna come up computationally. All algorithms are gonna work this way. All modern methods – you haven't done it yet, but whenever you solve a problem, it doesn't just say here's X and you have to trust the software or whatever.

It doesn't work that way, although you haven't seen it return you yet. They also return, no exceptions, a certificate proving that it's the solution so you don't have to trust the implementation. Everybody see what I'm saying? These ideas are gonna diffuse through everything we do. So basically you think of that as an optimal point, optimal design, whatever you want to call it, this is a certificate proving that's optimal because that's what it is. That's a lower bound on P star, that's a point that's feasible and satisfies and has objective value equal to this lower bound of P star, therefore, it is P star. Now, by definition this thing is the infinium over all X of G with these optimal lagrange multipliers. Okay. But if it's the infinium over all X, it's certainly less than or equal to this when I plug in a particular X and I'm gonna choose to plug in X star. Okay. So I plug in X star and I have the following. Very interesting. This says F0 of X star is less than or equal to F0 of X star plus something where every term in that is less than or equal to zero. Okay. And every term in that is zero. So this one is not relevant. Okay. We'll get to that. Okay. Yes, everything here is zero. And now you say, wait a minute here, if this thing is less than or equal to that thing and that's the same as that, then they're all three equal and we have no choice but to conclude that the sum of lambda I star times FI of X star is zero. Okay. But there's more than that. Wait a minute. This is a sum of numbers, all of which is less than or equal to zero. If you have a sum of numbers, which are less than or equal to zero and it's equal to zero, there's only one conclusion; every single one of those numbers has to be zero. And that says the following: lambda I star times FI of X star is actually equal to zero for all of these. Okay. And that's known as complimentary slackness and what this means is the following: it says if you have any primal optimal point and any dual optimal point, the following must hold; if the optimal lagrange multiplier is positive or zero then that thing has to be tight. If a constraint is loose at the optimal point, these lagrange multipliers have to be zero. Okay. So this is gonna have lots of implications and when we give other interpretations of what all this means, it's all gonna tie in, like, with these things being prices for example. But we'll quit here for today.

Student:[Inaudible]

Instructor (Stephen Boyd):Exactly.

[End of Audio]

Duration: 77 minutes