Distributed Optimization and Machine Learning

Prof. Mayank Baranwal

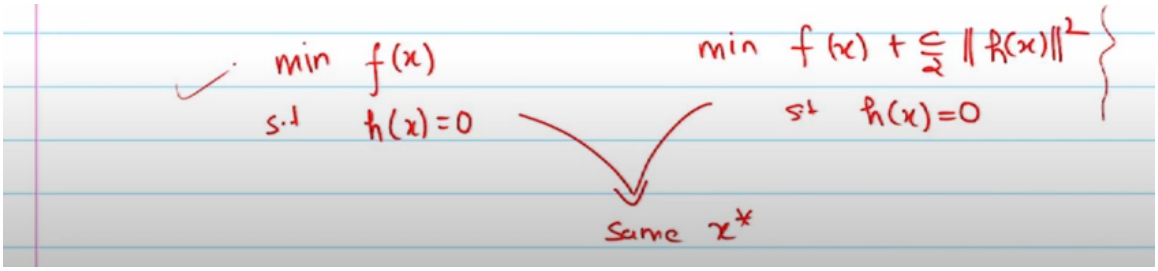
Computer Science & Engineering, Electrical Engineering, Mathematics

Indian Institute of Technology Bombay

Week-6

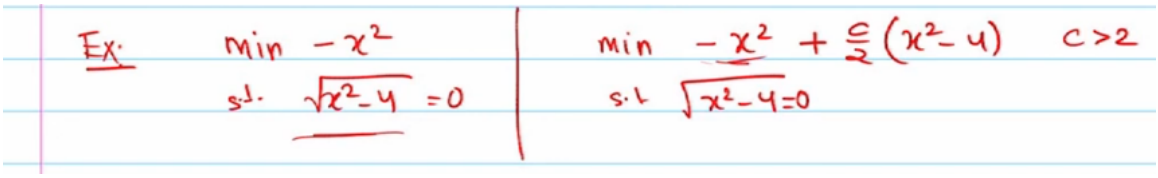
Lecture - 22: Augmented Lagrangian

So, tell me one thing let us say I have an optimization problem of this form minimize f of x subject to h of x equal to 0. versus another optimization problem of this form minimize f of x plus let us say c by 2 by the way h of x can be a vector valued function. So, it is. So, do they share the same optimal optimizer x star can also assume f is convex h is convex and so on right. So, but what would be what could be a motivation behind looking at this kind of problem over this kind of problem the original problem. In both cases you have to work with the constraint right.



So, let us take an example ok. Let us take an example. Suppose I want to minimize minus x square very simple example subject to x is scalar subject to square root of let us say x square minus 4 equal to 0. Is this a convex problem? Is minus x square a convex function? No right, so the objective function is not convex.

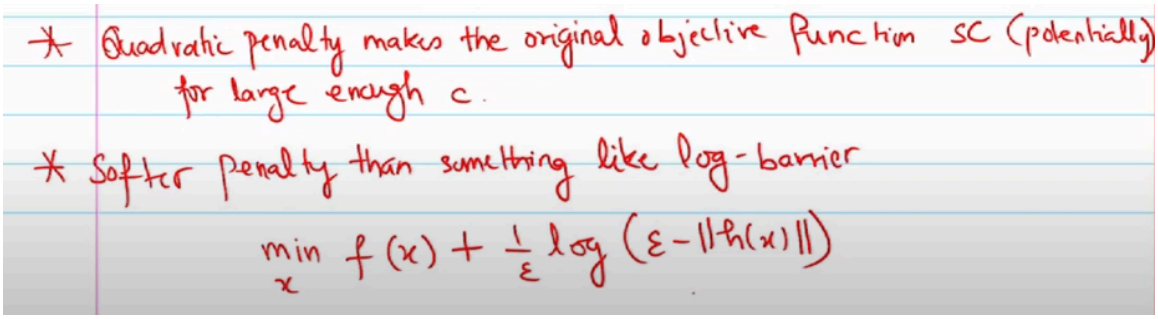
But if I consider an equivalent problem, by the way what is the optimal solution to this? there is just one point x square is equal to 4 is the only solution to it in fact. So, negative 4 is the optimal value of this objective function. But if I consider another optimization problem in this form, this is a simple example, but let us say c by 2 x square minus 4 subject to x square minus 4. And I choose c to be greater than 2. By the way here c is greater than 0.



I choose c to be greater than 2. So what happens if I choose c to be greater than 2? It becomes a convex function right. So potentially non-convex problem can also become a convex problem without like without you having to change the minimizer right. And not just its convex in fact it becomes strongly convex right. Let's say if I work with equality constraint like linear constraints here and if I take the norm of it you will get quadratic type of constraints and that would potentially make the original problem strongly convex for a large enough c even if you have the non-convex term sitting over here.

should be convex. So, anyway let us say we are working with I am saying that linear constraints I mean anyway do not have to do with I mean they are both they are just a fine constraint they can be both convex concave however you want to say it right. So, let us say you just have linear constraints Ax equal to b right and if I take the norm of Ax minus b whole square you will get something in terms of x square right and if A is let us say full row rank you would in fact get a strongly convex objective for a large enough value of c. So, the same optimization problem which was a non-convex problem to start with becomes a potentially convex problem right. And this is something really great and this is where this method of multiplier sort of comes into picture.

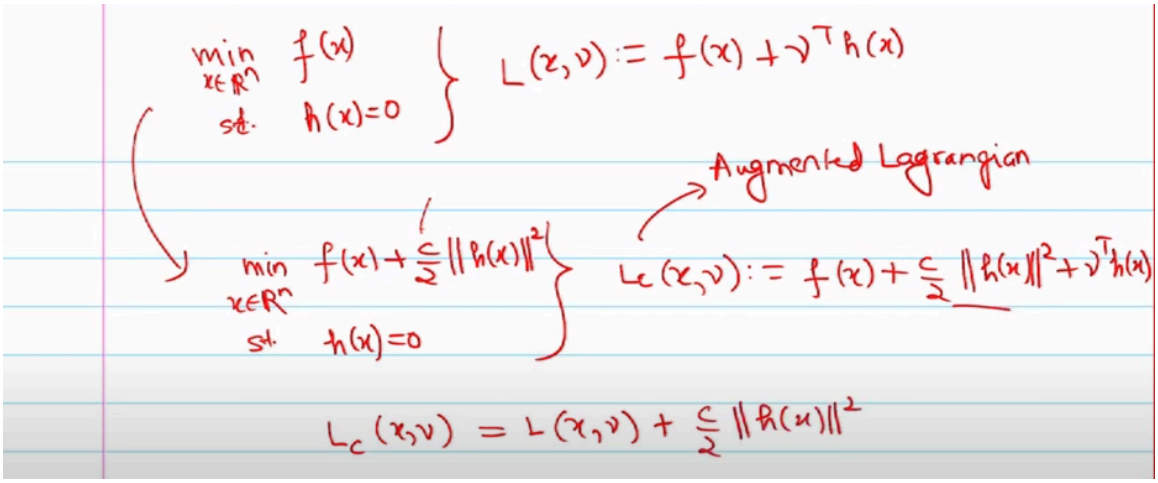
So, augmenting this quadratic penalty term or this quadratic cost to your original cost this has two benefits right. So, one is quadratic penalty makes the original objective function strongly convex or rather potentially strong, potentially right I mean that may not be true for all kinds of equality constraint. So, it makes the original objective function strongly convex for large enough c. The other thing is it basically has a softer penalty than something like log barrier. So what do we mean by that? So I could have written the same, I could have modified my objective as something like this.



log of something like this right, this would have been this would have still been fine, but

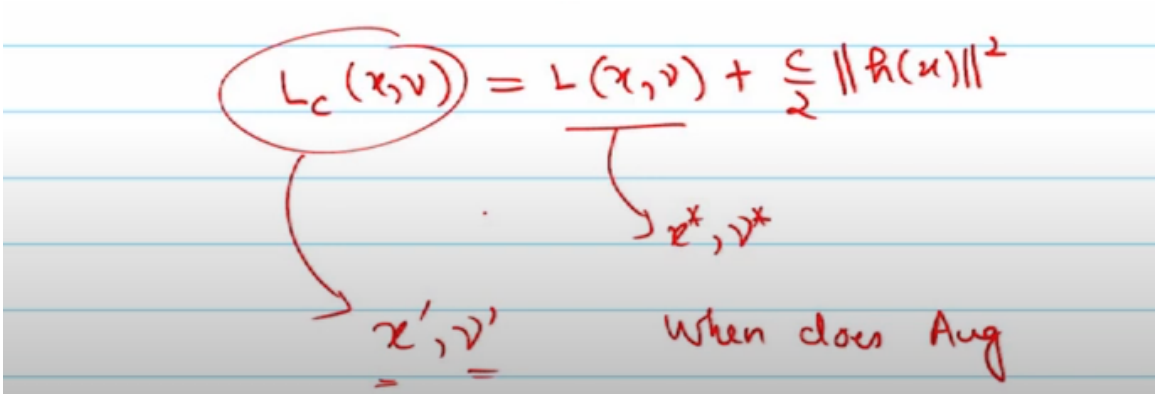
then because of this log barriers essentially you are saying that you do not let the h of x go like away from 0 right. So, because that is the equality constraint that you are trying to enforce. So, I mean you can look at log barrier methods which try to achieve something like this, but in that case what happens is that you are going to be working with in the interior of this thing right, because log is not defined for negative values. And that makes the iterates confined to a very small region and it basically adds to numerical instability of the algorithm that you are going to be designing right. So, iterates are confined in the strict in strictly confined values, let me write this So, iterates strictly confined in the interior.

Whereas if you add a quadratic penalty it has a much smoother or much nicer sort of landscape right? So, what is the original let us say if I get back to the problem statement. x in R n subject to h of x equal to 0. Still Lagrangian for this would be L x nu which is defined as f of x nu plus transpose h x. Now, if I add this quadratic penalty to this particular term or your objective function without changing the problem so essentially is I have this right what is the, so define a new Lagrangian here, let us call, let us denote this by L c, L sub c, where c is associated with this constant c here, right. So, L sub c x nu and this is defined as f of x plus c by 2 plus nu transpose h of x, So, from here you can see that by the way this is called augmented Lagrangian because you augment this quadratic penalty.



So, this is called augmented Lagrangian and Lc basically turns out to be Lx nu plus the So, if I try and run a saddle point problem on the original Lagrangian, let us say I get my x star and nu star right. Now, this is a modified problem. If I run a saddle point problem, I am going to get let us say x prime and nu prime as the saddle point for this. So, the question is I mean the point is that you should be able to first of all if I am trying to look at the unconstrained optimization problem, the solution to that like if I look at the

corresponding Lagrangian. they should return the same of like same optimal values of x and nu as the original problem right.



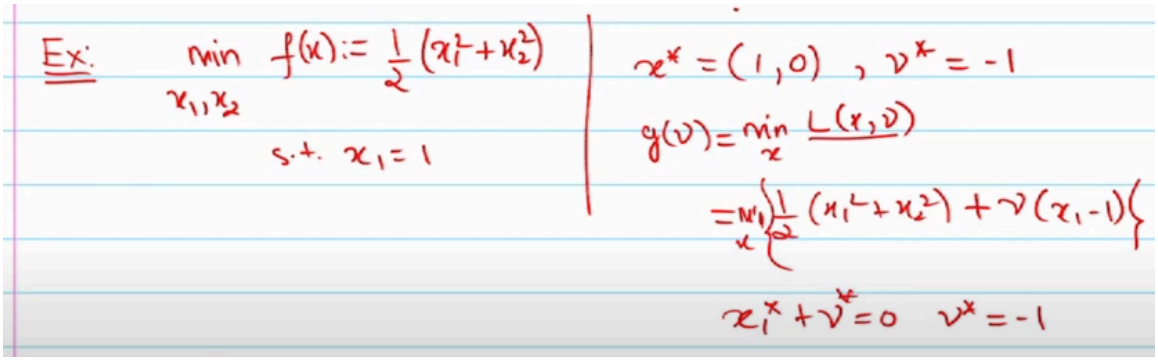
So, when does this work right? So, when does augmented Lagrangian work? As you can see that they would return different values of x star in here. Yeah, but then you have changed the function right. So, this is Lc is your new Lagrangian for the new problem. So, it would give you some like let us say you can take a dual of it and then you find nu star or the new prime corresponding to it and then you also find the primal of it x prime right that minimizes the that basically minimizes the Lagrangian and so on. So, you are going you are not guaranteed to get the same x star and u star that you would have gotten with the original problem.

the moment you take it to the Lagrangian. The original the two optimization forms are the same right, but the moment you convert them into unconstrained optimization problem using Lagrangian. So, you may not get the same x star and u star with your augmented Lagrangian that you would have gotten with your x and u right because of this additional term. Now if I try to minimize it with respect to x I am going to get something else right. should still be satisfied, so that is fine.

I am saying that if I try and find a saddle point of this, so for a given value of c, you may get some certain, let's say if, for a given value, let's say if I fix c, the x that minimizes this Lagrangian, that gives me the dual, may not be the same x that minimizes the original Lagrangian, right? Because I have shifted everything, I may shift, end up shifting everything, right? Sorry. So, that that is. So, again let us let us I think in order to clear the confusion let us do that one one particular exercise. Let me just write the statement. So, let us take one example and I think it will be much clearer.

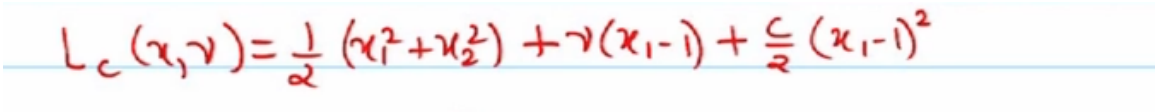
So, minimize with respect to x 1, x 2 these are a primal variables f of x which is defined to be half x 1 square plus x 2 square subject to x 1 equal to 1. What is the optimal solution? What is x star here? 1 comma 0 because x 1 is equal to 1 and this gets

minimized when x 2 is equal to 0. So, 1 comma 0 is the optimal x star. What is the optimal nu star here? So, let us quickly look at. So, g nu is right and first of all is the problem convex right, problem is convex and strong duality holds anyway there are no inequality constraints.

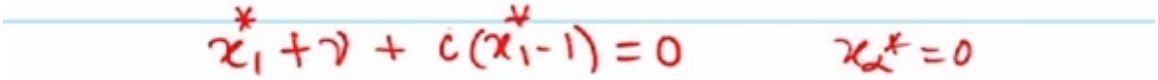


So, prime will explain, so essentially KKT conditions would be satisfied, they are both necessary and sufficient. So, let us, so you have half x 1 square plus x 2 square plus nu times x 1 minus 1 right. minimize with respect to x. And if I set the derivative with respect to x 2 to be 0, you get x 2 star to be 0. And if I set the derivative with respect to x 1, you get x 1 star plus nu is equal to 0 or nu star is equal to 0 or nu star is equal to negative 1.

So, nu star is equal to negative 1. Now, look at the augmented Lagrangian. So, this was the original Lagrangian. Now, let us So, as I said, we are going to add a constant to it. So, Lc is going to be the original function.

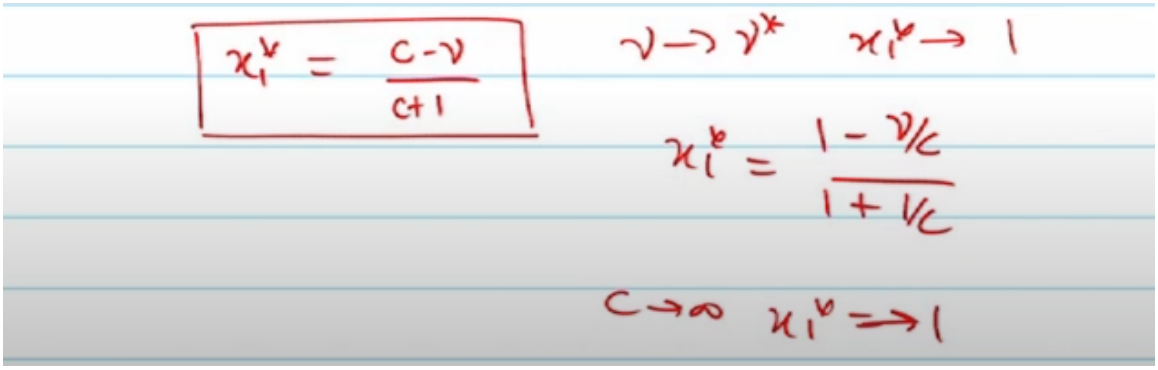


Yeah, because KKT conditions are satisfied, right. So, now let us let us try and minimize it with respect to x 1. So, we get x 1 star plus nu star or whether it is let us I am just trying to minimize with respect to x 1 and x 2. plus c times x 1 minus 1 that is equal to 0 right or x 1 star minus 1 that is equal to 0. So, which gives me x 2 star anyway is equal to 0.



So, x 2 star is equal to 0 is something that you directly get it from here. So, what is the value of x 1 star here then? It is basically c minus mu upon was c plus 1. So, this depends on c now, when only when nu is equal to nu star. So, that is the thing. So, when

nu is equal to nu star, when either when nu goes to nu star, you what do you get x1 star then goes to the optimal one that we wanted to we wanted it to get or there is another way out right.



I can write this as 1 minus nu over c and 1 plus 1 over c and as c goes to infinity again x1 star turns out to be 1 x1 star goes to 1. So, there are two ways through which you can approach the same objective. or same optimizer rather. So that is the whole point. Let's say I have a, so when do we ensure that the augmented Lagrangian and, so if I take new to be closer to new star, the original new star, right.

Again we don't know the original new star. If we somehow take this new, let me, I mean if this is not clear let me try to mean we do not know the original nu right, nu star. So I take this nu closer to nu star and I take c closer to infinity. Then we can show that x goes to x star right. So that is what we are trying to get here.

So how does and when does augmented, so the mechanism for this particular approach to work is you somehow try and take this nu closer to nu star. I mean you are not trying to solve the original problem first of all right. So you do not know where nu star is. Right.

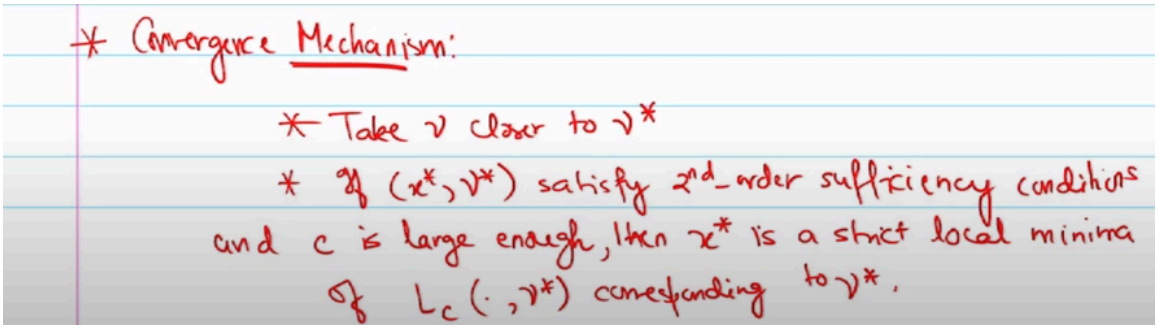
Also there will be a new star. There will be a new star, yeah. And that will match the old new star. Yeah, because the objectives are the same, yeah. So, that will like this, I mean that will match the new star. That match the nu, yeah that will match the nu star, yes.

So, the invariant x star from this Lc and the matrix. Right, x star would match, but then again you do not know what nu star is, right. But it will come as same as. Right, right. So if you are at new star you will get x star.

So that is what we just saw right, for any value of c. If you are at new star you will get x star. Yeah, but here you are assuming that we do not know the mu, but mu can be calculated for this particular. I do not have the other problem where for a given value of c. So for a given value of c, let us say if c is below certain threshold.

even if nu is equal to nu star it is possible that c won't like I mean x won't be closer to x star. So, for this to work you want c to be like as long as c is sufficiently large this would work and in fact that is the result that we are going to be I mean not deriving today but stating today. So, there if c is sufficiently large and if nu is closer to nu star then x would be closer to x star. the corresponding x would be closer to x star. So, that is the idea behind augmented Lagrangian.

So, the convergence mechanism as we have seen through this example is, if you try and take nu somehow closer to nu star and c to infinity, then we can show that we can show that in fact x the optimal value of x basically goes to x star right. So, the idea is without having to solve for nu star if we can somehow if we are closer to that even let us say we are farther away from nu star, but if c is sufficiently large we can still converge to the optimal solution right. So, that is the idea. So, we do not have to solve for nu here. we are not really solving for nu, we are just trying to change like increase this c may be potentially like with every iteration we will be increasing the value of c and we will be performing certain updates.



So, that nu also tends to be closer to nu star. So, if those conditions are satisfied then for a sufficiently large c you are basically you get x star to be closer to the or the solution to the original lagrangian. So, that is the idea ok. So, the idea behind or the way this works is if x star nu star satisfy second order sufficiency conditions and c is sufficiently large or c is large enough. then x star is a So what this statement says, so let us say x star, nu star are the optimal solutions for the original Lagrangian right.

Now if you choose c to be sufficiently large and you also assume that, I will write down what second-order sufficiency conditions are, but let us say this pair x star, nu star satisfies the second-order sufficiency condition. Then as a, like if you fix here, if you fix a nu star and if you look at the augmented Lagrangian as a function of x. So this is this basically this function closer to x star looks like a strict local minima or it has a strict local minima around x. In fact x star turns out to be strict local minima. So when I say strict, so you only get x star as the only solution if you are closer to and its local because if that means you have to be closer to nu star.

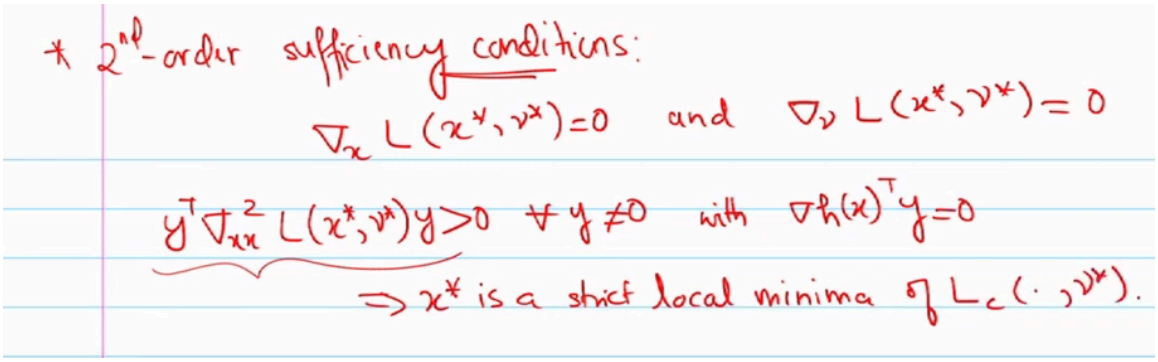
So around nu star you get a strict sort of bowl of optimality where you can live with x star being the optimal solution. but that happens only for sufficiently large C ok. But again the reason we want to work with the augmented Lagrangian is because it has a much nicer sort of convergence behavior than some of the problems well let us say when f of x is not when f of x is let us say convex, but not strongly convex. So, we have seen that for strongly convex we can accelerate optimization right. that may not be the case for simple convex functions and by make by adding this quadratic penalty you can potentially make the convergence much faster.

So, that is the idea behind working with augmented Lagrangian and also you can convert non-convex problem to convex problem. So, when you when you have a non-convex objective you would like in this case for instance you can in fact see that show that like for instance here when nu star is equal to 1 or nu star is equal to minus 1, no matter what your c was as long as c is positive you were getting the same like x1 star to be 1. For non-convex case you want c to be at least certain value, there will be some threshold on c. So, I mean in practice we do not know what c is. So, with every iteration we just maybe the new c would be 1.

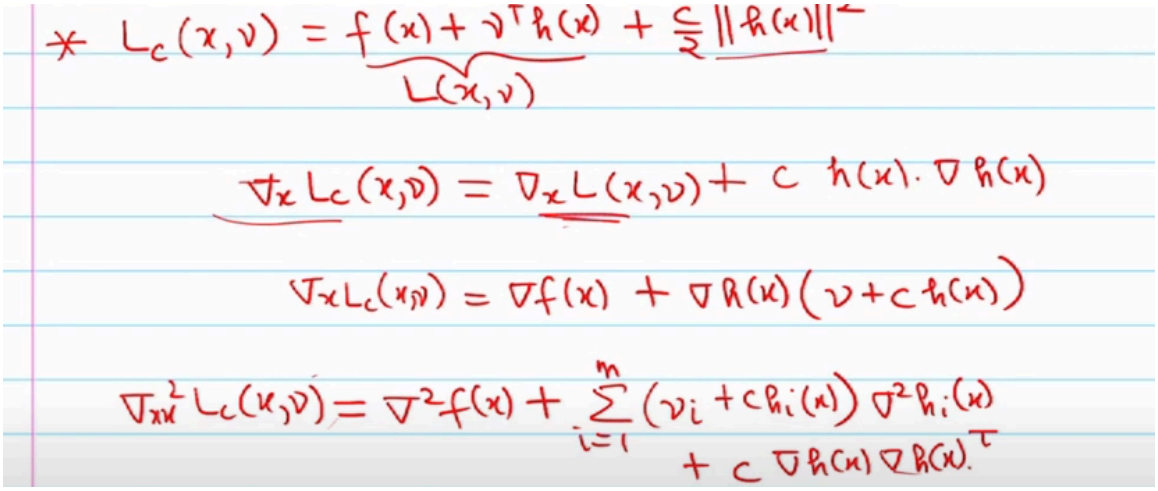
2 times the previous c and so on we keep on increasing. So, that once it hits a threshold then you are like you can you can basically solve for the optimal x star and u star. In general you cannot like I mean it really depends on the function that you can maybe for certain analytical functions you can quantify that, but in general it is not it is very difficult to know it a priori. So, you start with some value of c and you just keep increasing it. Not divergence, but it like there are some numerical instabilities if c becomes really large right. So, I mean you should not be starting at a very large value of c, so that the updates that you are going to be making.

are going to be dependent on c right. So, I mean you do not make very large updates on your x's in some sense. So, that kind of issues can be there, but I mean usually you start with the small value of c and you gradually increase it. So, that eventually you hit that threshold and that is how it usually works. So, second order sufficiency condition. So we say x star nu, the pair x star nu star satisfy second order sufficiency condition.

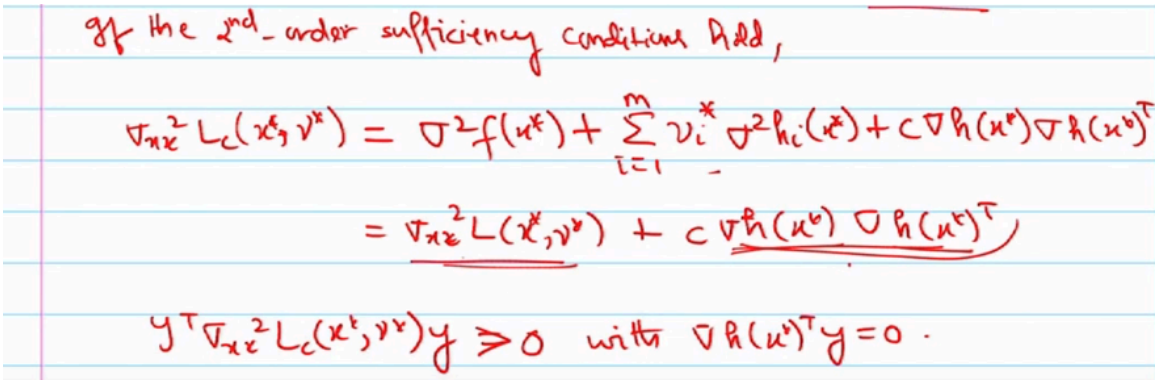
So these sufficiency conditions are, so this is with respect to the original Lagrangian, x star nu star is equal to 0. Basically you have stationarity in both x and nu. and you have y transpose. So, for every y which is orthogonal to gradient of H of x, you kind of have this constraint. So, as long as these two are satisfied, then you can show that x star is a strict local minimum.



So, let us try and derive this. So, you have your augmented Lagrangian x nu which is basically f of x plus nu transpose h of x. So, this is nothing but your L of x nu. So, So what is this quantity? In the pointwise multiplication right of individual terms, this is nothing but gradient of f of x plus So, if you now take the double derivative of this because we want to show if it has to be a strict local minima then we have to show that even for the Lc here this is going to be greater than 0 right that is when it is a strict local minima. So, even for Lc this has to be greater than 0. if you take the double derivative of this, this turns out to be this term plus let us try to write it this way which is much easier.

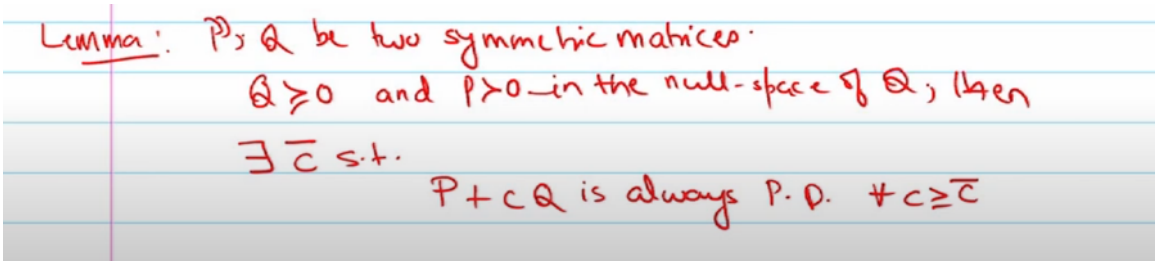


So, if the second order sufficiency conditions hold, okay and I consider y which are orthogonal to gradient of h of x. So, the moment I multiply this with y, sorry h of x transpose, the moment I sort of multiply this with y, this particular term is equal to 0 because of this constraint, right. So, let me first write down this particular term here in terms of x and then it will be clear.



transpose. So, this is nothing but. does this additional term because this is anyway going to be subsumed in the original Lagrangian. So, this is, now to this if I take y transpose y, this is going to be y transpose this particular term which is greater than 0. plus c times this particular term, which is which because of this particular thing. So, what what do we get? This is greater than 0 right for with ok. So, what is this kind of like? So, if I look at this matrix, this matrix is positive semi-definite right.

So, there is a result let me quickly write down this result. So, this matrix acts like a positive definite matrix because y transpose this thing is greater than 0, this is like a positive definite matrix. So, there is this result in linear algebra like if P and Q are the two symmetric matrices with Q being positive semi-definite and P to be positive definite in the null space of Q. again so this is positive definite only in the null space of Q right. So, in the null space of q then there exists a constant c bar greater than 0 such that P plus c Q is always positive definite for every c greater than equal to c bar. So, there exists a sufficiently large c such that this is always positive definite and therefore, you can say that this is strictly I mean this is a strict local minima and therefore, you have strict inequality and that is this c bar is what we are going to be looking at in when we look at few more examples.



So, unfortunately today I do not have like further examples to provide on this particular work, but then in the next lecture we are going to be looking at potentially non-convex functions. and then there you would see that unlike this particular case that we looked at

where this worked for all c's when nu was equal to nu star you would see that for those class of functions not all c would not work. Thank you.