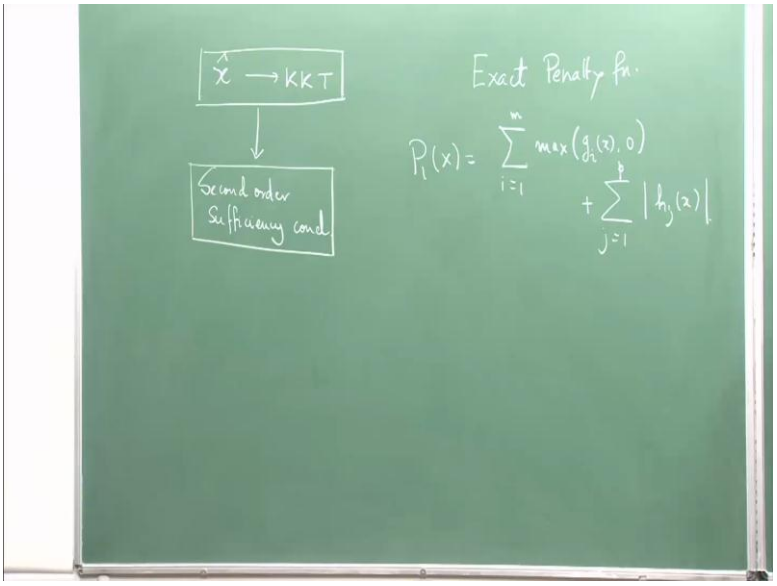
Foundation of Optimization
Prof. Dr. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 31

Today, we begin by telling, giving a brief idea, what is exact penalty function. And then, try to introduce to you a second order sufficient condition for optimality.

(Refer Slide Time 00:36)

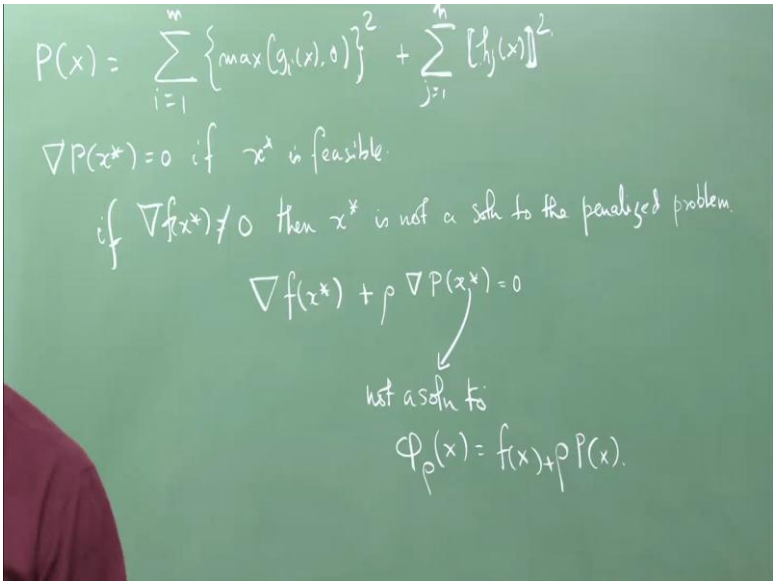


See what happens is that suppose you have x hat to be a KKT point of the standard mathematical programming problem P, which we had given in terms of equality and inequality constraints. But how do I know that this x hat is the local minimum of the problem? If all the functions are convex and for example, the equality constraints are all are fine; then it is guaranteed that such an x hat is not only a local minimum; it is a global minimum. The problem that in order to know whether x hat a KKT point is a local minimum, you need to find out something called second order sufficiency condition. So, this will guarantee you that yes, this x hat is the local minimum – a second order sufficiency condition.

And this second order sufficiency condition is of fundamental importance when we talk about exact penalty function. In exact penalty function – when you study except penalty function, what we are trying to do is that, we are trying to devise a penalty function, so

that when I solve the penalty problem for a large value of rho, I will get in most cases, a solution to the original problem. So, in this case, the penalty function is…

(Refer Slide Time 03:10)



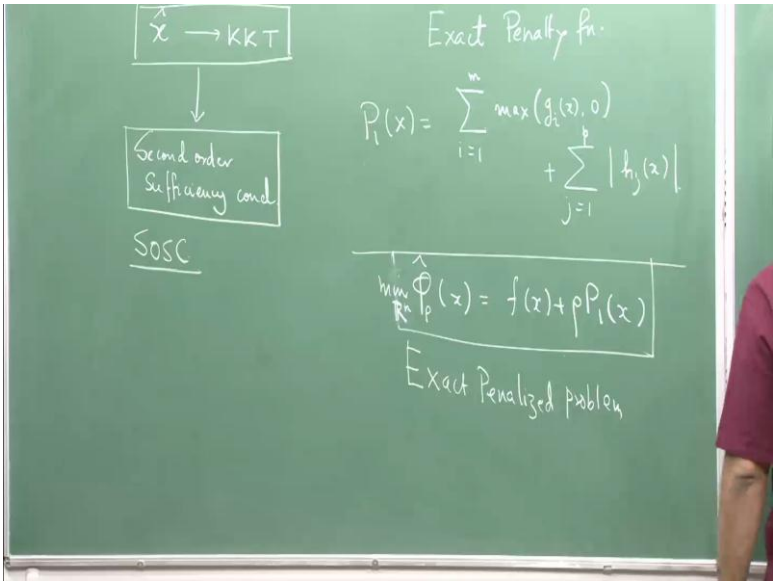
The reason why the other penalty function, where you have squares – here – fails to give you an exact solution – to give you a solution to exact or give you a solution to the original problem is essentially, because at the solution or at any feasible point, if you look at this one, when the original penalty function that we had studied in the last lectures; now, this function – if x star is feasible, then grad of… So, any such point does not satisfy you the Karush-Kuhn-Tucker condition unless you have grad of f x star equal to 0. If grad f x star is equal to 0, then such an x star can be thought of given as the Karush-Kuhn-Tucker point.

But the problem largely is that, if this is not equal to 0; which can be in most cases, because you are talking about constant optimization. So, this will never lead to… Such a problem would never give me a Karush-Kuhn-Tucker condition, will not satisfy the optimality condition. And, hence, we would be in trouble in the sense that… because that derivative is 0. So, basically, then if you want to solve it, because this is 0, because unless grad f x star is not equal to 0, we cannot say that x star is a solution to the panelized problem itself unless I have grad f x star…

No, grad f x star is not equal to 0; then, x star is not a solution to the panelized problem. So, if you have a feasible point; then at that feasible point, this is equal to 0. So, if x star

has to be a solution to the penalized problem with some rho, then this must be true. So, if this is 0; then, unless this is 0, you cannot say that this whole sum would be 0. So, if this is not 0 and this is 0, the whole sum is not 0; so, x star cannot be a solution to the panelized problem; that is, this x star is not a solution or local minimum to… which is not an unconstrained local minimum. So, that was the question I asked you; that means there cannot be a solution of the original problem, which is the solution to the panelized problem. So, how can we remedy this situation? That, if I have a solution to an original problem; then, under certain conditions, I will have a solution to a penalty problem. So, if I continue solving a penalty problem for large value of rho, I will come across a penalty problem, where whose solution would be actually the solution of the original problem.

(Refer Slide Time 07:01)

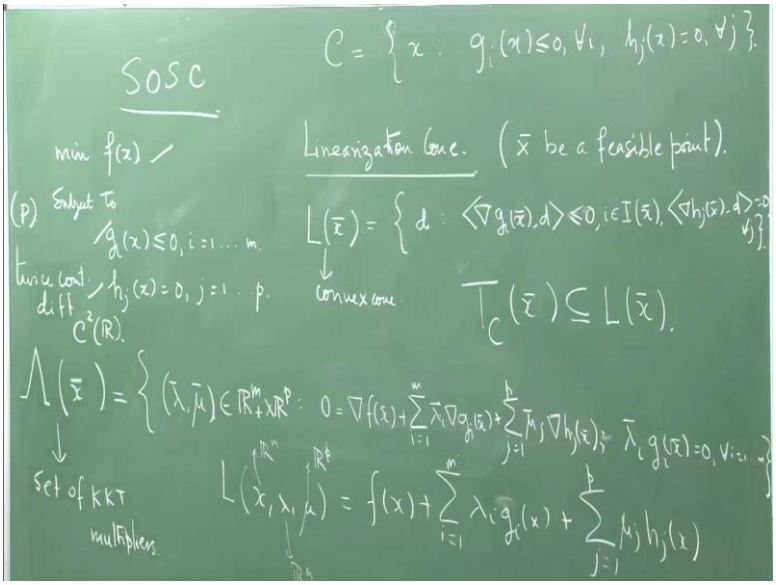


Now, that is done using what is called the exact penalty function, which we will now write as phi rho, because this will entertain this property; that if the functions are twice continuously differentiable, then if x hat is a local minimum of the original problem; then, x hat is also a solution or local minimum to this problem. So, this is very very important; that a solution of the original problem is also a solution to the penalized problem. So, for rho sufficiently large, if you solve this problem; then, the solution that we have got – if it is equal to the original problem, is actually a solution to the original problem. So, this thing is called the exact penalty function – exact penalized problem rather I would say is to take a mean of this over R n.

Now, in order to understand how such a thing happens that, how does the original solution, original local minimum is also a local minimum of this penalized rho for some rho's sufficiently large. For that, we need to understand what is called the second order sufficiency condition or SOSC – second order sufficiency condition. But, this will play a very vital role in establishing this fact. So, we are trying to remedy the problem that we have for the standard penalty function by using what is called the exact penalty function and what the exact panelized problem.

So, now, we have talked about KKT conditions; we have talked about… See in unconstrained case, we know when grad f x equal to 0; if x star is a point, which satisfies grad f x equal to 0; and, if the (( )) at x star is positive, definite we assure that x star is a local minimum of the problem – straight local minimum rather. Now, what we are trying to say is that, in case of the constrained problem, we had never given or made any discussion what could be a sufficient condition. Hardly, it is discussed in optimization courses. But, I believe that in some way, though it might not be so easily checkable, it is very important in understanding of the theory.

(Refer Slide Time 10:08)



For example, here in this exact penalized problem, such a condition plays a crucial role. So, we would now go and look into the second order sufficiency conditions. Given a KKT point, what extra condition the point should satisfy, so that we can guarantee that such a point is a local minimum? So, that is what we are trying to do now. Today, our

job would be to describe you what is the second order sufficiency condition. And then, tomorrow, we will discuss about the exact penalty function and move over at the end of tomorrow's lecture to what is called Lagrangian duality, which is also very very fundamental thing in optimization.

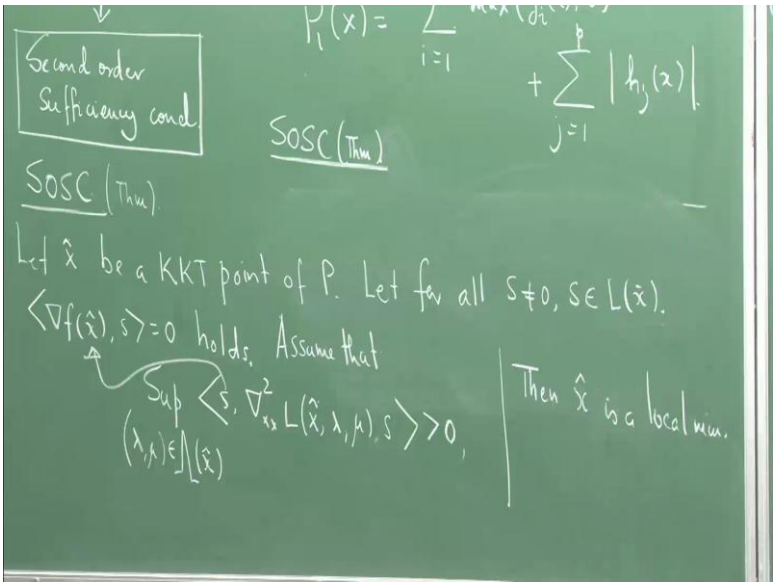
Suppose now you have this problem, the one which we were studying – problem P. Now, suppose x hat is a KKT point of this problem. But, you know corresponding to the KKT point, there could be many multipliers with respect to which x hat is a KKT point. So, given a point x bar, this is defined as a set of KKT multipliers. So, this consists of all elements lambda, mu – lambda bar mu bar in fact – element of R m plus cross R p such that 0 is element of grad of f x bar plus summation i is equal to 1 to m lambda i grad of g i x bar plus summation j is equal to 1 to p mu j grad of h x bar and then… And also, I put lambda bar. So, lambda bar i g i x bar is equal to 0 for all i. So, this is a set of KKT multipliers; that is, at x bar, find all these multipliers lambda bar and mu r, which would satisfy the KKT condition, which will make x bar a KKT point. So, we need this set.

And then, we need what is called the linearization cone associated with this problem. So, what is a linearization cone at a given point x bar? If x bar is a feasible point, consider x bar to be feasible. So, the linearization cone is given as follows. I will… Or, at L at x bar, the linearization cone is a set of all d such that… I am of course, assuming these are all… Now, let me assume not only continuous, but twice continuously differentiable. So, assumption would be problem data – this data, this function, this function and this function. They are twice continuously differentiable; that is, all of them are in C 2 R. So, this would consist of the vectors d, which satisfy this one for every i in the index – an active index set; and, grad of h x bar d is equal to 0 for all j. So, it is a collection of all the d's, which would satisfy this system of equality and inequality.

Now, you have to understand that, this linearization cone is a cone. The linearization cone – we are telling it as cone. But, the set L x bar is a cone and it is a convex cone. Now, let us denote the feasible set of the problem P as C; that is, C feasible set of P; that is, a set of all x such that g i x is less than equal to 0 for all i and h x is equal to 0 for all j. Now, if you calculate the tangent cone as we have done in the last classes, the tangent cone to C at the point x bar is contained in L x bar. So, these information are now read… With these information, we are now ready to write down the second order sufficiency conditions. But, another important construction that we have already done earlier; but,

now, we will do it again is the Lagrangian function associated with this problem. So, the Lagrangian function l x lambda mu is the function f x plus summation lambda i g i x plus summation – here I transform i from 1 to m and here I transform 1 to p. Now, here x is in R n; this is in R m plus; and, this is in R p. So, here there is a restriction on lambda. Lambda – all the components have to be greater than equal to 0. This is called the Lagrangian function.

(Refer Slide Time 17:51)



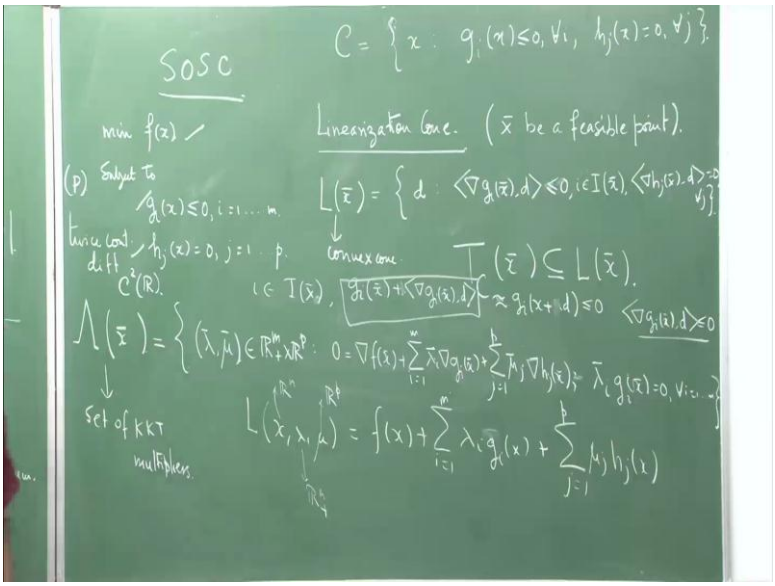
Second order sufficiency conditions – what will be important here that you should be able to compute the easier matrix of this function. And, that would lead us to develop what is called the… So, we will now state the main result. Now, let x hat theorem I would say – this is a very important rather. Our idea would be to sketch a proof of this fact. Let x hat be a KKT point of P. Now, here we are going to make certain crucial requirement; we will take some crucial requirement.

Let for all S not equal to 0 and s element of L x bar grad of f of x hat S is equal to 0. So, all elements of the critical cone – they satisfy this particular condition in terms of the objective functions gradient – gradient of the objective function. So, here you remember the objective function is also appearing in the description in the very beginning. So, x hat is a KKT point. If this is a local minimum and suppose constant qualification satisfy, it will be a KKT point.

Let for all S not equal to 0 with S element of S x, this holds. Now, assume that supremum of lambda mu – element of… Now, assume that for every S that you have here, you have this condition as a easier matrix at x hat. This matrix – this is strictly bigger than 0, where the supremum runs over all the possible lambda mu's in this. So, you are taking those lambda mu's, constructing those lambdas, those Lagrangian functions; and then, checking that for every S of this form, this holds for all such S – for all S, which satisfies this. Then, if this happens, my conclusion is that, then, x hat is a local minimum.

Now, how would you go about doing the proof of this? So, let us go about the theorem once again. It says that, if x hat is a KKT point, then of course, the KKT point would be a set of KKT multipliers, which could be single turn, which could be more than 1, whatever depending on the constant qualification. KKT point then… Let for all S, which is nonzero and S which belongs to the linearization cone, this cone is called the linearization cone. Why? Because it looks at only the linear part of the function.

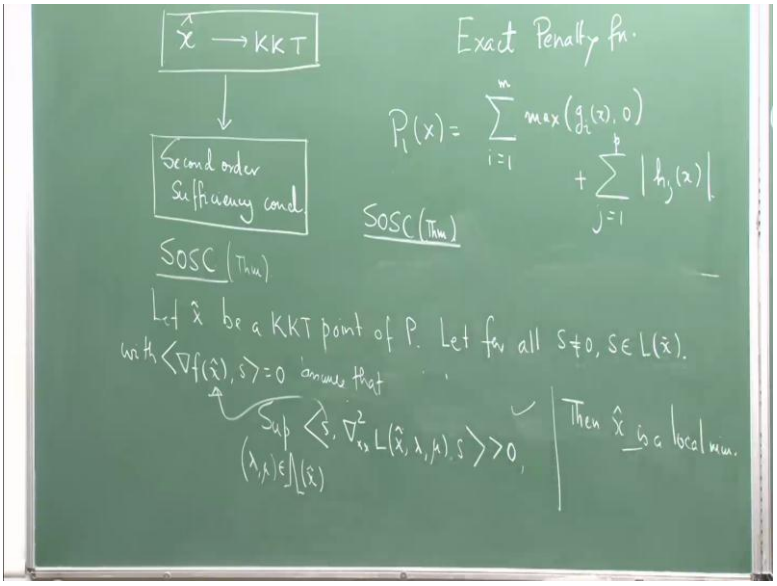
(Refer Slide Time 22:20)



Basically, if you express the functions g i with i element of I x naught; if you make a linear approximation of g i's, it will be g i's at… Around x bar, the linear approximation of g i is like this. So, this will be some g i of x plus lambda d. And, assume that x plus lambda d is a… When I am… Of course, I am not bothering about… So, lambda is here. But, if it is feasible, this is less than equal to 0 and this is equal to 0.

So, basically, then you will have grad g i x bar d is less than equal to 0. So, this is… So, basically, replacing the original problem, which is linearized version. So, you are replacing the constants with its linearized version, with its linear approximation. Around x bar, I can replace the original prob function by this function. Maybe you can forget about the lambda. And, replace the original function by this function. Since i is in the i x bar, this will be 0. So, this will be less than equal to 0. And, that is exactly what you want. So, you want the constant function around x bar to behave in this fashion in terms of linear function in d. So, that is why it is called linearized cone. So, what it says that, if you take any nonzero element from the linearized cone, 0 is obviously an element there.

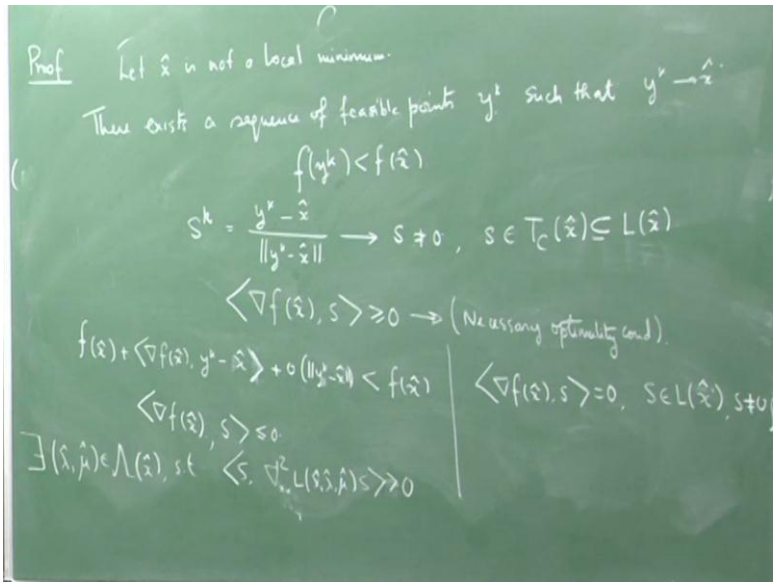
(Refer Slide Time 24:24)



Take a nonzero element is there; take a nonzero element from the linearized cone. And, for all such nonzero element with this equal to 0… I will just refine the writing. Let for all S not equal to 0, S element of this with this. So, take all those S, which is not equal to 0 and it is belonging to this and also satisfies this. Assume that, that is, let for all S not equal to 0 with S element of this with this been equal to 0. So, you cannot take only those nonzero S from this for which this is true. And, for those S's, assume that this is true. And then, x hat is the local minimum of the original problem. And, we will give a sketch of the proof of such a thing. The proof would be quite long. So, we will do it step by step.

Now, let me just remove this part; in the sense, I am rubbing it off, so that you do not get a feeling that the main result is not written; that means the main result is written. But I would just request you to remember this little linearization cone business and what I have just described and the tangent cone. But the tangent cone may not be exactly equal to the linearization cone. If it is equal, then we say some sort of an (( )) type qualification condition holds. So, we will not get into all that details, but just we will try to prove this.

(Refer Slide Time 26:13)



We will start with a contradiction that, let x hat is not a local minimum; which means that whatever neighborhood you take around x hat, there is one y hat for which f of x hat is strictly bigger than f of y hat; which means that, what I can do, let x hat is not a local minimum. Then, without loss of generality, I can say that, there is a sequence of feasible points y k, which is converging to x. But for each of those y k's, f of y k is strictly less than f x hat. So, there exists a sequence of feasible points y k such that y k tends to x hat and f of y k is strictly less than f of x hat. Now, once that is known, we would now construct some important things.

Now, let me construct a sequence S k. Now, this sequence is bounded, because each is of norm 1. And, this sequence is going to some S k; there is a unbounded… there is a convergence of sequence of S k, which goes to some S. Now, if you observe carefully, if I put this as T k; without loss of generality, let me assume that, S k goes to S and S is not equal to 0; of course, norm S is 1. Now, observe that, if I put this as T k, this T k is going

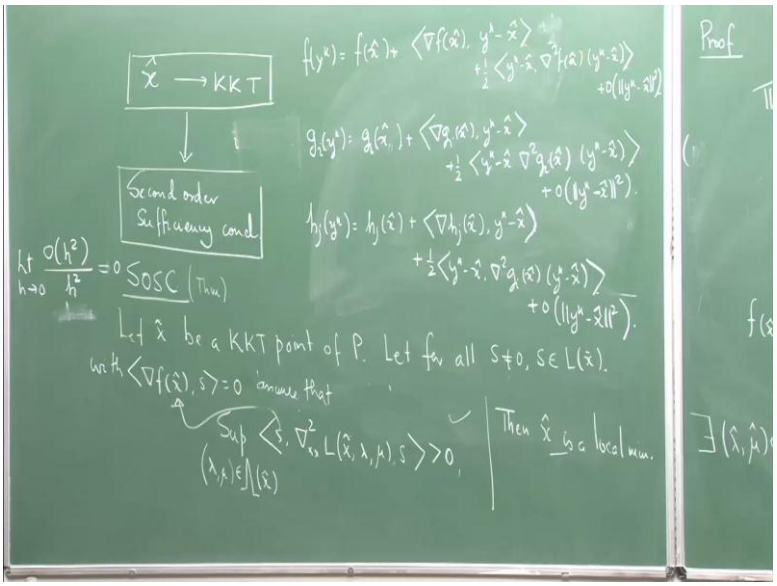
to 0 and y k is a element of the feasible set; y k is going to x hat; and, this whole thing is going to S, which simply says that S is an element of the tangent cone to C hat x hat; and, which is a subset of L of x hat. See immediately, certain things are known to us. And, that will help us to do some progress.

Now, if u have remembered the optimality condition that we had spoken off in the last class, is that, if u have the tangent cone and if you have a differentiable function minimizing a… Not last class, last to last class I guess. If you have a differentiable function and you want to minimize it over a close set C, then the optimality condition is that, the gradient of f x in a product with all the tangent vectors must be greater than equal to 0. So, in this case, we should always have… This must be greater than equal to 0, because S belongs to T c x as a part. So, this comes from necessary optimality condition.

Now, if we look at this condition – this one and apply the Taylor's theorem here; that is, if you write this as f of y k as f of x hat plus y k minus x hat plus o of norm y k minus x hat is Taylor's expansion. And this is less than f of x hat. Now, if I divide both sides, this will cancel of course giving this is strictly less than 0. See if I divide this by this divide this – on both the sides by norm y k minus x hat and then I make k tends to infinity; and, what I will finally have from is grad of f of x hat, because this is going to S. What I will have finally, is this. So, this combined with these two gives me this fact that, grad of f of x hat s is equal to 0. And of course, S is element of L x hat and also S is not equal to 0. So, all these things are combined in this story.

Now, once you have combined the things, now, you have an S, which satisfies all these. So, that S must satisfy this. So, the supremum of this is strictly bigger than 0 for all lambda mu; which means there must be at least one lambda hat mu hat in this for which that this in a product is strictly greater than 0. So, it would imply there exists – just by the definition of supremum, lambda hat mu hat element of lambda x hat such that S times grad square (( )) of this Lagrangian into the required S is strictly bigger than 0. This is just by the definition of supremum. What we will do, our job would be to contradict this fact. Therefore, now, we would use the second order expansion of a function – Taylor expansion. So, what is that second order Taylor expansion? Here in the second order Taylor expansion, because we will now write this second order term…

(Refer Slide Time 33:57)

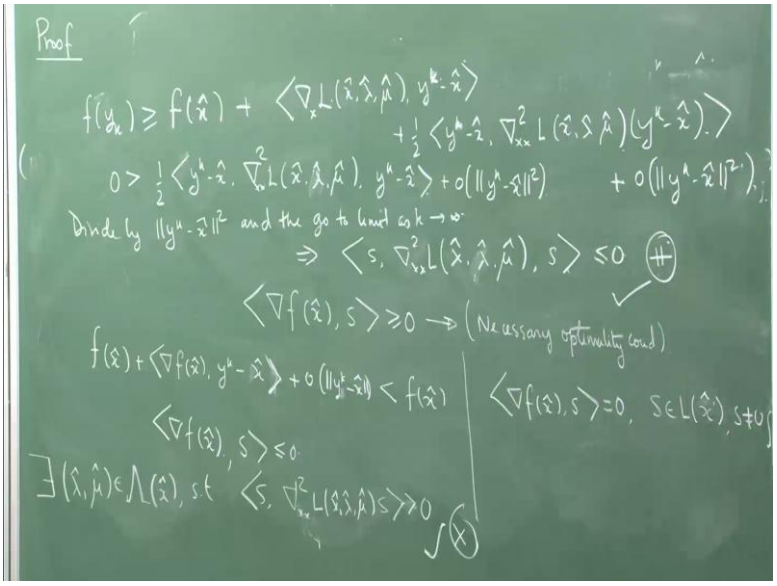


After the second order term, there will be an error term, which would be of the form this – o of h square. So, o of h square by h square – that goes to 0; h is a real variable – as h tends to 0. So, limit… This is going to 0 – is equal to 0. So, we would have some quantity like this. So, what we would now have is f of y k is f of x hat plus grad f of x hat y k minus x hat plus half of y k minus x hat grad square f of x hat y k minus x hat plus o of norm y k minus x hat whole square. The same argument will go through for each of g i's – maybe just for the active indices; does not matter.

Now, here I have g i x hat; so g i x hat would be 0. I am just taking i element of I x naught. That is exactly what is required – i x hat in the Karush-Kuhn-Tucker thing, because lambda i g i x hat is 0. For all inactive cases, lambda is 0. So, I do not have to bother about those things. So, let me just take this part. And then, we will again write down everything in terms of the scalar thing. Let me take this out. I will have the same grad g i x hat y k minus x hat plus half of y k minus x hat grad square g i x hat. In fact, you do not need to even take this, but I am taking this just to make it easier, simpler as that, because that is what you will get at the end. If you feel uncomfortable, some people might try to confuse to take it for every i; take this for every i – this expression. For everyone, you will have this order quantity at the end. So, this is… Similarly, for h j's, you will have the same expression.

Now, what I do is multiply all of these with lambda i and this with u j's and then add up the whole thing. So, if I add up this, add up this; now, I multiply with lambda, this with mu j's. So, if I add up this, this and this; this multiplied with lambda, this with lambda mu j; then, what I would have is this part, would give me the first condition of the Karush-Kuhn-Tucker condition. This part would give me 0, because lambda i g i x hat is 0; h j x hat is anyway 0. So, this part will go. This part is anyway g i y hat; g i y hat is less than equal to 0 and this is equal to 0. So, what I will have is finally, if you add these things completely; after addition, what I want to do is this – multiply all of these by lambda i's – these ones; multiply all of these by mu j's.

(Refer Slide Time 40:08)



Then, what you will finally get is f y k greater than equal to f of x hat plus grad of L x hat and lambda i hat of course; same lambda hat, mu hat – the same ones; y k minus x hat plus half of y k minus x hat… So, you multiply this with this, this with this. And then, added all the three. So, you will get some resulting equalities from here. And then, add all the three to get this one. Now, if you add and… this order quantity, this order quantity, this order quantity will again be that sort of order quantity of same type. Now, what you do, you basically know what has happened; you know that f y k minus f x hat – that is strictly less than 0. That was our basic assumption at f x hat was bigger. So, I will have… But this zero (( )) because x hat is the Karush-Kuhn-Tucker condition. So, this is 0. So, I will have 0 – here this one. So, this is strictly less than 0. This will be strictly less than half of y k minus x hat grad square lambda hat mu hat y k minus x hat plus…

Now, divide both the sides – all the sides by y k minus x hat whole square and then take the limit as k tends to infinity. So, divide by norm and then go to limit as k tends to infinity. So, now, noting this fact, you will simply see that, when dividing, y minus x hat square will come here – y minus x hat and here y minus y k minus x hat. And then, that will go to S. So, you will finally have S of grad square. So, that would imply… because this will go to 0, this will become… So, less than equal to 0. This finally becomes… So, what I have done, when I take the division by this, I separate the y k minus x hat whole square into norm y k minus x hat into norm y k minus x hat; one I push under this; one I push under this.

And, here I push the whole norm of y k minus x hat whole square. So, when I take k is going to infinity, y k is going to x hat; norm y k minus x hat whole square is also going to 0; norm y k minus x hat is going to 0. So, this will go to 0. And, this will go to S – this part and this will go to S. And, this strict inequality will go to equality. And, that is exactly what we get. But, this is a contradiction to what we have. So, this one contradicts this one.

And hence our assumption that x hat is not a local minimum (( )) x hat is a local minimum. This is a very very important result, though it is not easy to check. But it is mathematically very beautiful and it shows up to what level of complexity one can have when one actually handles constraint optimization problems.

Thank you.

And, tomorrow we would again start our discussion on exact penalty functions.