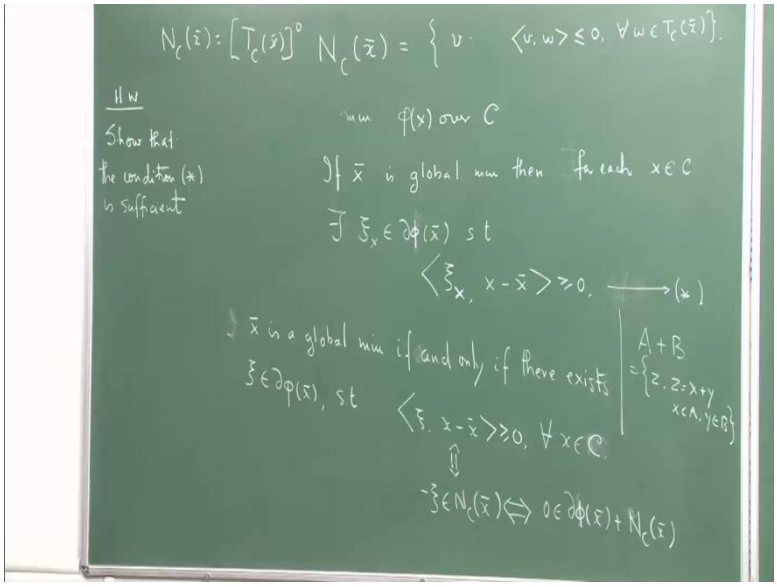
Foundation of Optimization
Prof. Dr. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 29

In our last class, we spoke about our projected gradient algorithm, we spoke about then optimality condition when a convex function is not differentiable and our optimality condition was something like this, that.

(Refer Slide Time: 00:37)



So, if you are trying to minimize a function say convex function phi over c, a convex set c then we said that if x bar is a local minima, the necessary condition sorry global minima, because this is x convex. So, we are writing down the necessary condition say x bar is a global min, then for each x in c, there exist psi x belonging to depending on the x you choose the sub differential of sorry y at x bar.

Such that, psi of x, x minus x bar is greater than equal to 0 means, for all x by I wanted to mean I wanted to say that, whenever you take an x, your psi x would change, this condition is also sufficient. So, as a home work, show that the, if I write down this as the condition star, the condition star agreed.

Now, this does not look very interesting, because at every x you need to change, but then how do you go ahead and prove something better, how do you improve this condition. By

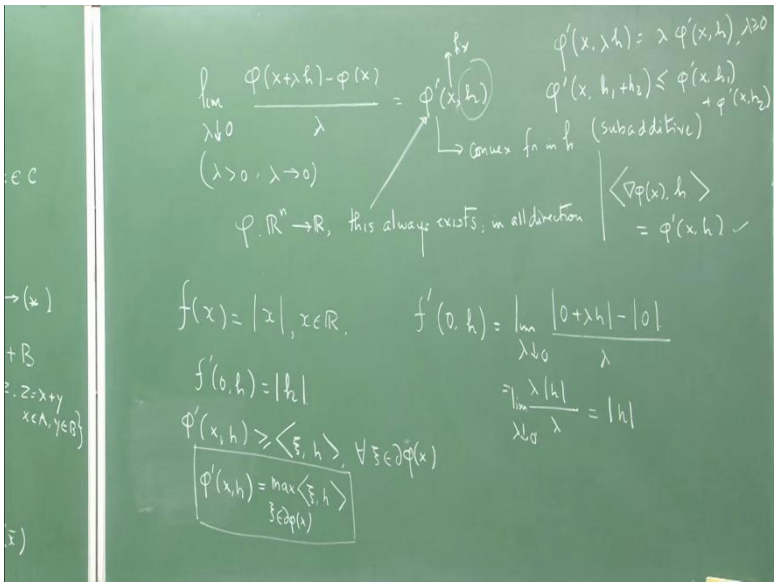
improvement, what we want to tell is that I want to get one psi, which will work for whatever x you take. Only one psi, which will work for whatever x you want to choose, right how do I get such an optimality condition.

Let us, first start with this thing, try to give you an answer and then try to pass on to our study of penalty methods, where we are suppose to study today an approach, which will be specifically done for equality and inequality constants. We will tell you how to write down the penalty function so what would happen, if you take a sequence of local minimum of the penalty problem. These are all helpful, when you actually study algorithms because you really know what the thing is going on suppose, if you apply penalty method.

Let us look into this problem for a while, you can look into my course in NPTEL, it is on convex optimization, you would get more detail on this. But, we are not going to going to so much detail but try to tell you a bit more. So, how do I go about this so what we intend to show here is the following, that that x bar is a global minimum, if and only if, there exist x bar is a global minima, if and only if, there exist psi element of such that so this psi no longer depends on this x e.

So, this actually means, it is equivalent to the following condition, which is equivalent to the condition. So, this is set addition that is, which there we are adding 2 sets, I want to recollect that, adding 2 sets a and b in R N means, is a collection over z. And z can be expressed as, x plus y with x belonging to a, and y belonging to b. So, you take arbitrary element of x and a, and any arbitrary element of b and make a vector addition of them and then new vector that is formed, that is in the set a plus b. So, this is what ultimately we need to prove, in order to prove such a thing we need to get into a slightly different fact.

(Refer Slide Time: 06:52)



The different fact is that, for a convex function suppose, you calculate this particular limit, fix up the h vector and then calculate the limit at any x over whole of R N. Now, this symbol actually means, lambda is strictly bigger than 0 and lambda goes to 0, this is called lambda down arrow 0. Now, this is if this limit at all exist then this is called the directional derivative of the function phi, at the point x and in the direction h, phi dash of x comma h.

Now, for a convex function phi from R N to R so if phi is from R N to R, this always exist and this basically means, this this are always exist in all directions. So, at a given point h in x in R N, whatever direction you choose and you try to compute the directional derivative, a directional derivative would exist, even if the original function does not have a derivative. For example, if you take this one, f x is absolute value of x and you now try to compute the directional derivative at h at 0.

Because, 0 is the point where, point of the derivative so it is 0 plus lambda h minus mod of 0 because phi of x is phi of 0 is 0, divided by lambda. So, basically it finally means, lambda lambda by mod h, the lambda is positive by lambda, which is mod a limit. So, finally because lambda lambda cancels because lambda goes to 0, lambda need not be 0 say, it cancels. So, as lambda goes to 0 so this is nothing but mod h because mod h is independent of lambda.

Now, so this is a function of h and this look there is a interesting thing, it looks like that this is the same, it has the same function structure. Now, so this once I fix the x, if I fix x once I fix the x this is the function of h so this is a very important point to remember, as a function of h, phi is positively homogeneous. That is, phi dash and also phi dash x of h 1 plus h 2, it is sub-additive, it is positively homogeneous and also sub-additive (No audio from 10:31 to 10:41).

So, these two properties are this is sub-additivity and the other is positive homogeneity, if a function has these two properties then this function is called sub linear. So, this is obviously, it is a convex function, you can show that this is obviously, phi dash is a convex function in h. So, it is very important is to know that, this is a convex function in H, and if you have the derivative, so if phi is actually differentiable then you will have this is exactly the value of the directional derivative at h.

And here you see, if it works for lambda negative, that is lambda comes from the negative side and goes to 0 then if the limit exist and is equal then of course the derivative is there. So, we have some few examples so what is this relation with the sub-differential so relation with the sub-differential is the following. It is very simple to show, which I am not going into detail at the present moment and you had need to go to a to my convex optimization course or may be in future, I should give a course on convex analysis proving all these things in detail.

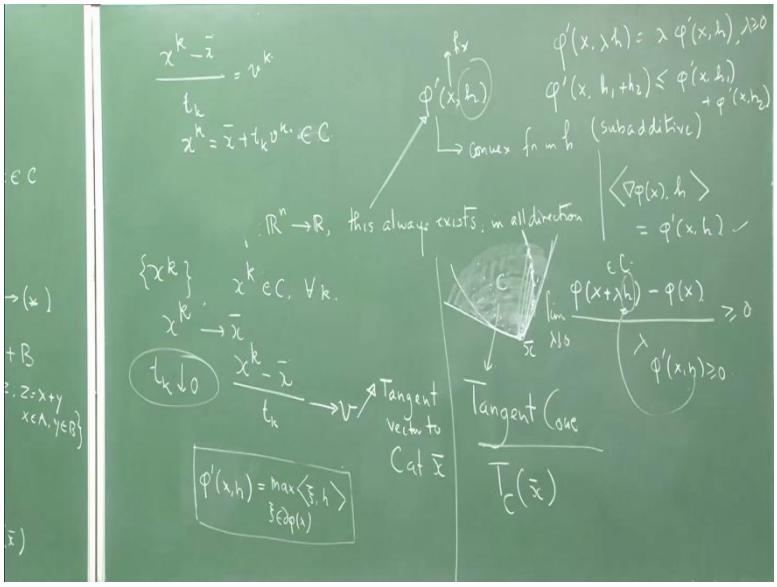
And so here the it is not very difficult to prove, which I will keep as home work so whatever psi you take, at the whatever psi you take from the sub differential of phi at x of the convex function, this result always holds. In fact, it can be shown that, this is a very important fact and it is truly a very fundamental fact in the sense, that it shows the power of it shows the power of separation theorem. But, we have not dealt with separation theorems here in so much rigorous space so we are just going to give you some very basic idea, what happens is that, you can always write.

So, once if I fix my h and x, this is becomes function in psi, a linear function in psi and your maximizing it over convex compact set. And in fact, because of the compactness and the continuity of the linear function, there exist a psi, for every x, there exist a psi for which, this will be true. Now, this duality that is, this is going back and forth between the

directional relative and the sub differential is a key thing in convex analysis, which is a key tool in understanding modern convex optimization or modern optimization as such.

So, here we tend to first look into so we will try looking to a very fundamental optimality condition, which relates, which makes use of this in the constraint sense, in this sense. So, minimizing it over c, we will not talk about minimizing over R N because minimizing over R N is not a very difficult thing of, x is the minimizer of phi over R N, if and only if phi dash x dash is greater than equal to 0, for every such direction each.

(Refer Slide Time: 15:26)



Now, I once I try to look into the set c, the geometry of the set c would play a significant role in understanding, what should be the necessary condition for optimality. So, let us have a convex set c and let me see, may be this is the point, which is optimum. So, what do you expect at this point, what why would you call it optimum because if you move from x bar along any line and the points to you stay in the visible set c, at every point the function value would always tend to be more than, it should be more than f of x bar.

So, function value at this point should be more than the function value at x bar. So, if I move along any such points, any such direction, basically a directional derivative should be always greater than equal to 0. Basically, if I have for say, if I take lambda between 0 a very small number, much less than half say 1 and when I am pushing it to 0 and suppose, for lambda sufficiently small, phi of x plus lambda h for a given direction H is in the set, this set c.

For all such lambdas then it is immediate and for such directions h, this should be greater than equal to 0 or phi dash x h would be greater than equal to 0, for such directions h for which, this is true that x plus lambda h is in c. So, for every such directions now, for example, if I take these direction obviously, it is not true because it is so much outside the set from the very beginning.

So, ultimately if you look at all such directions, you would finally come to a set of directions like this on which, this condition would be true. This set of directions here form a cone and that is called the tangent cone, to the set c at x bar. So, this takes us slightly into, what we would call non smooth geometry and we would write down a bit of definition about the tangent cone.

Let us again start looking at this cone and see, how we can mathematically describe this cone, what is happening here is that, if you look at this, we are basically looking at say, if you look at sequence or points coming along this boundary towards x bar, then you keep on drawing the rays or joining them the x bar by the line segment. If you keep on doing so you see ultimately it comes and convert this to the tangent line, the very simple notion over tangent line, which you already know from your high school.

So, basically if you look at sequences, if you take sequences, if you say why sequences why not a curl or why not a straight line the question is that, here this definition can be extended even the non convex sets which, could be even disconnected. And in the sense that, their sequence would be much more easy to describe things so if we take any sort of sequence of x case in c coming and converging to x bar and then you take those line segments and scale them up.

And then they see the direction to which they are converging then that sort of direction is usually called a tangent direction motivated by the fact of the definition of the tangent. Because, you are using the same procedure, as you would like to draw the tangent to a curve at a given point, the same procedure is done to generate the whole cone. So, this cone is usually known as the tangent cone or the bulligan tangent cone, it will be just for us let us just call it the tangent cone, rather than getting too much bothered with historical correctness.

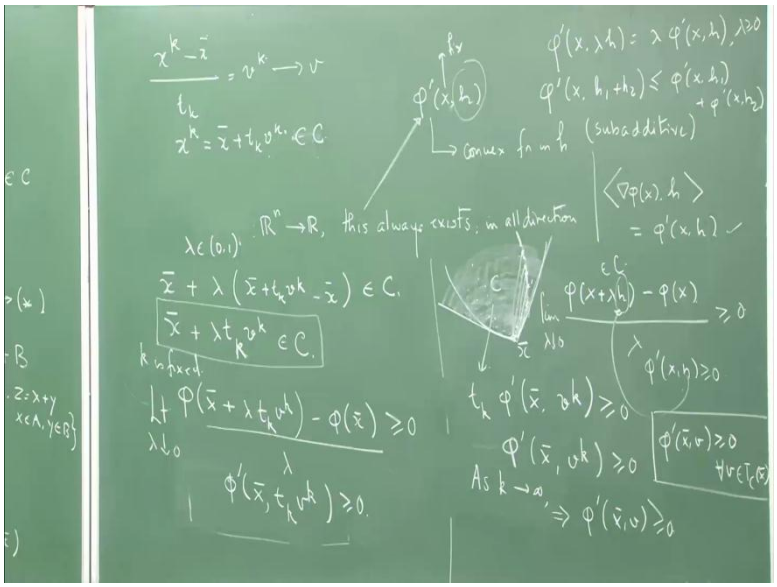
We should call it bulligan tangent cone, it came around 1938 so what it says is the following that so what you do, you consider a sequence x k with each x k element of c

and have such a sequence, which is converging to x bar. Now, scale it by some number so that, it just does not become the zero vector, to stop it from becoming zero vector. So that, you really observe, in which direction this is converging so divide it by some t k where, t k goes down to 0.

So, here as t k is going down to 0, the scale factor is increasing and stopping this norm of x k minus x bar to countdown to 0. So, this goes to certain vector v and this vector v is called the tangent vector to c at x bar so v is called the tangent vector. Now, once this is known, you can now the set of all such vectors v is bought up in a cone and this is called the tangent cone to c at x bar. So, v is called the tangent vector to c at x bar, if there exist an x k going to x bar such that, and of sequence t k going to 0 such that, this ratio the limit of this ratio goes to that vector v.

So, that is the meaning of the tangent cone now, what does the tangent cone, tell me, can it give me give me some information regarding, the information it gives is the following information. Now, what does it tell you so it tells you that, I can write, if I write this as v k then I have x k equal to x bar plus t k v k and this is an element of c, you can write like that also.

(Refer Slide Time: 23:35)



Now, since c is the convex set so I can write x bar plus some lambda times, lambda is between 0 and 1, this vector x bar plus t k v k minus x bar, this is also in c. So, once you know this then that will give you x bar plus lambda t k v k is element of c right. So, for all

lambda between 0 and 1, this would happen, this is this is the outcome of convexity. Now, if x bar is an optimal point then you write, you take the direction t k v k for a fixed k and then because you know that.

So, whatever k you take this means, that for any k you take, this is always in c so this minus phi x bar must be greater than or equal to 0. Now, if I divide this by lambda and take the limit as lambda tends to 0, down arrow 0, this will give me that phi. So, I fix that k, for a fixed k, k is fixed, phi of x bar t k v k is bigger than or equal to 0. Now, the next step is very simple, so because my k is fixed now, what I will do, once I know this and I know the there is a positive homogeneity of this directional derivative so this is the directional derivative.

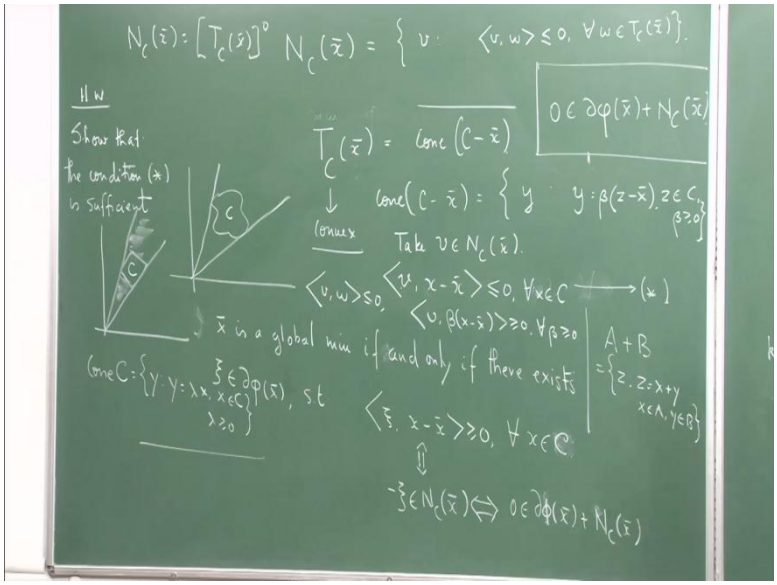
So, the positive homogeneity of the directional derivative would tell me that, I can take the t k out and as t k is strictly bigger than 0, I can divide by T k. So, I can write t k times phi dash x bar of course, v k, I am taking as non zero, zero is in the rangen cone, I do not want to bother about zero vector. Zero is obvious because if you put h equal to 0 here, this will become 0 by definition so this is greater than equal to 0 and phi dash x bar v k is greater than equal to 0.

Now, because for a fixed x bar, this is a convex function in h over the whole of R N, which means that, over the whole of R N, this is a continuous function because every convex function is continuous over the whole space. So, I can now take as limit as k tends to infinity, I can push the limit in here just by basic definition of spontaneity, that will imply that phi dash x bar v. Because, v k goes to v right, this v k obviously, goes to b v by the very definition of a tangent cone, this is greater than equal to 0.

But, this v was just one of the members of the tangent cone so if I take any other member, I can repeat the same argument. So, give me the following fundamental necessary and sufficient optimality condition, this is necessary as well as sufficient. So, why it is sufficient is, still there is a little bit of story, which I shall not tell you right now, may be you need not even bother about it. This is what you have, a tangent cone has a some sort of a strange relationship with the normal cone, not strange I will say rather nice geometrically. So, the tangent cone the normal cone to see at x bar can be written as, set of all v such that, v of w is less than equal to 0, for all w belonging to the tangent cone to

c at x bar. Symbolically, this means, this is called the polar, this side is called the polar of the tangent cone given by this notation.

(Refer Slide Time: 29:53)



Now, what is important here is to know that, how that is true, can I take an element of the normal cone and write it like that. Because, if c is a convex set because if c is a convex set then if the tangent cone has a more simplifying because of the convexity of c, tangent cone has a more simplifying expression. What you essentially have to do, you should take the cone generated by c minus x bar and then take the closure of that, that is exactly the tangent cone.

So, see the set c minus x bar and the cone generated by this, is nothing but the set of all y where, y is written as beta times, z minus x bar, z is element of c and beta is greater than equal to 0. So, you take c minus x bar basically, now, basically you have you have translated the cone to cones vertex at 0. Basically, this is now x y now, put at the origin and then you are at drawing a cone, you are generating a cone so you know how you will generate a cone.

If you just recall I guess, say if you have any set c so you can just, this is a way you generate a cone, of course, if it is a convex set c, it will all always generate a cone, which itself is convex. For example, this is this is but non convex sets can also generate convex cone, they can also generate non convex cone so this is this is the cone. So, the cone

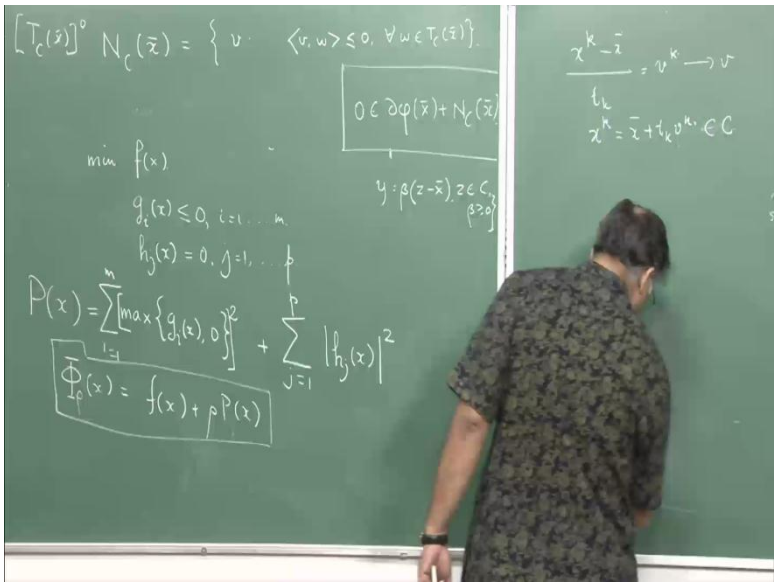
generated by a set c, is a set of all y where, y is lambda times x where, x is in c and lambda is greater than or equal to 0.

So, this is the standard definition, which we usually learned in the very basics so I am packing in a lot of convex analysis in this lecture. Now, once you know this so take any v, take v element of N c x bar, once you take the v element of N c x bar, your immediate formula that you know is this, x element of c. Now, once you know this, what would happen is the following now, I can take I can multiply it by beta.

So, I can do that for any given x and then I of course, can take the limit of such quantities so you simply have v of w. So, any element here would lie down here, how we know that any element here would be actually there, it is obvious, because x minus x bar, if x is in c, x minus x bar is naturally element of T c x bar and hence, this will be true. So, any element here is automatically here and by this again by this simple procedure, any element here is automatically here.

So, you know this fact so then keeping in this fact in mind and using the separation theorem, you from there you can conclude the following fact that, 0 belongs to. So, this is a very important fact and knowing some of this facts are important because we are going to study duality in after one or one class I guess. So, now we will go back to what we were trying to study about penalty functions so we will now, write down standard equality and inequality constant problem. And we will try to see, how do you construct the function p x right, the penalty function, rub this part of the board, I am sure that you have observed but anyway this is a visual you can go back and check.

(Refer Slide Time: 34:47)



So, what is very important here is to note the following is that is the following that, let me consider this problem. Now, for this problem, how do I construct the penalty function p x, is usually called the bentromy penalty function, see this penalty function is usually given like this. So, you basically take the max of between g i x or 0 and take the whole square, and you do it individually for all the i's and sum them up.

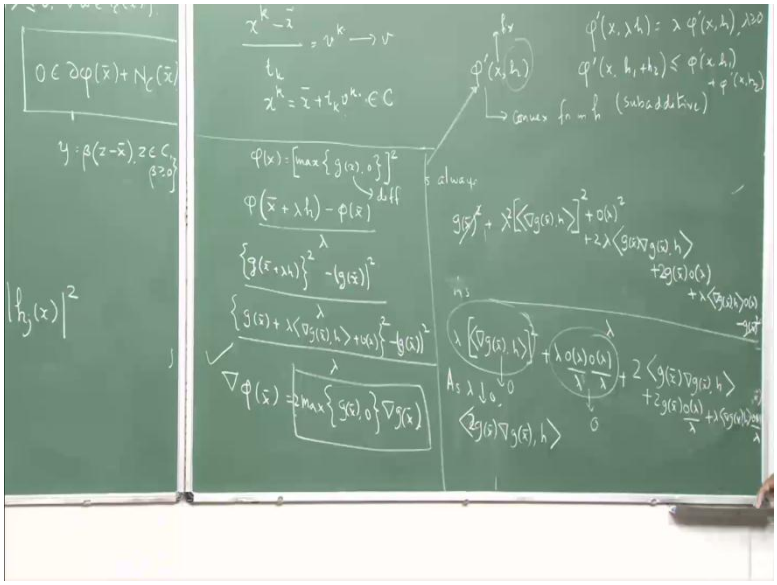
The interesting part here is that, if you do not take the square then max of g i x 0 is not a differentiable function, it is a non differentiable function in general but if you put the square, it becomes a differentiable function. It is very important at the outset to know that, why such a problem is (( )), how to compute the derivative of at least this one because you see we will basically consider our self studying this function where, p x in this case would be this.

When we would actually write some results about this, we can just forget this part, just concentrate on the inequality constraint for the time being and the other part can go as home work. Now, the interesting part is how do I compute the derivative of this, I have hardly seen, I have not seen any book, which actually does a little bit of gives a space to do it because everybody might not be so comfortable in doing computing the derivative of this.

Let us see first how do you compute the derivative of this, and finish our today's journey here and continue from tomorrow about the property of this class of and this class of

penalty function. What does it give me, what does it tell me, if I solve and get a local minima of each of these and what is where does the accumulation point of such a local minima lie, what is its property. So, here is a more here is a practical issue so in practice, are we really seeking optima are we really seeking an optimal point or we are happy with some point, which satisfies the Karush Kuhn Tucker conditions.

(Refer Slide Time: 38:30)



So, now, I am suppose to find the derivative so I will take a function like this, phi x is max of say, a single function g x 0 and I am taking a whole square of that. So, what would I do so here observe certain simple facts, simple facts are like this that, I can compute the gradient of this but how do I do it. So, basically I start computing a directional derivative and if I can express the directional derivative as some function v into h then v basically, I can take it for lambda in in any direction.

Then, I will and get the same answer then I have the derivative, that is for lambda minus 1 and plus 1 and lambda in the direction of minus one, and lambda in the direction of plus one. So, now, let me look at it so if I look at the directional derivative of this problem basically, then I am trying to compute the limit say, at x bar. I have I have to first compute this fact now, you see here, my g x bar could be 0, g x bar could be positive and for the moment, I assume this could be 0.

And this could be the problem is that, if this is positive, this has to be positive, if this is 0, it could be 0 or positive, it does not matter. If this is positive then this has to be positive

because of, the continuity of the function phi. Now, this function overall is a continuous function because of the continuity for lambda, sufficiently small whatever, direction h you choose, phi x bar lambda h is positive, by the continuity of the function.

So, suppose now, phi x bar is positive, if it is 0 then we can hope for the similar thing can be done then what can I do is.... So, what I have now is, g of x bar plus lambda h whole square minus g x bar whole square and that, divided by lambda. But now, here I open because I know that g is differentiable so g is given to be differentiable of course, here we are only studying differentiable problems, please note that we were studying only differentiable problems, it all differentiable. So, then what would happen is the following then I write down by the definition of derivative, if you look at this, a plus b plus c whole square that, then you will get basically a square plus b square plus c square plus 2 into a b. I just pulled in the g x bar into this, a b a c b c 2 of course, plus this into this.

So, now, I will separate g x bar from this thing and the whole thing actually is, now divided by lambda and you see this thing cancels. And now, what I have here first is, from if I divide by lambda, I have lambda. So, I have this o lambda square, I can write, I can multiply top and bottom by lambda and I have lambda, o lambda, o lambda. Because, multiply top and bottom by lambda, it will it will become lambda square in the bottom, it will become lambda lambda plus, now if I divide it will become 2 g x bar. Now, here it will become 2 g x bar o lambda by lambda plus lambda times, grad g x bar h o lambda by lambda. Now, you know small (()), when we define the derivative we have given in the in the definition, this you know this small o quantity, o lambda by lambda, that is actually going to 0.

Now, when lambda goes to 0, this will go to 0, this will obviously go to 0 because this is going to 0, this is going to 0, this is going to 0. Now, here it is free of lambda so this will remain but this will go to 0 and this will go to 0 and lambda, this part is going to 0, this part is going to 0. So, which means, now, what is left over as lambda goes to so as lambda goes to 0 plus in fact 0 plus 0 minus does not matter, whatever, basically as lambda goes to 0.

What I am getting here is 2 times g x bar grad g x bar h now, if g x bar was less this strictly less than 0 then it or equal to 0, this would become 0. So, this is my derivative actually, so but basically then this is either g x bar or 0 so I can now write, the gradient of

phi at x bar is, either it is 0 because this becomes 0 or it is this. So, basically it is max of twice into max of g x bar 0, this whole thing into grad of g x bar, this is my actually required derivative of this.

So, this calculation is quiet important, you can try it out on your own and once you learn some non smooth calculus, those who are really deeply interested in optimization, they will learn. So, then for them it will be simpler because that will be just an application of a non smooth calculus rule. So, with this, I end my talk here and tomorrow we will start to see the nature of this thing and we will talk about something called exact penalization. And then go over to study the use of Lagrangian duality N, may be a sort of idea of how to use Newton's method in the constant case.

Thank you very much.