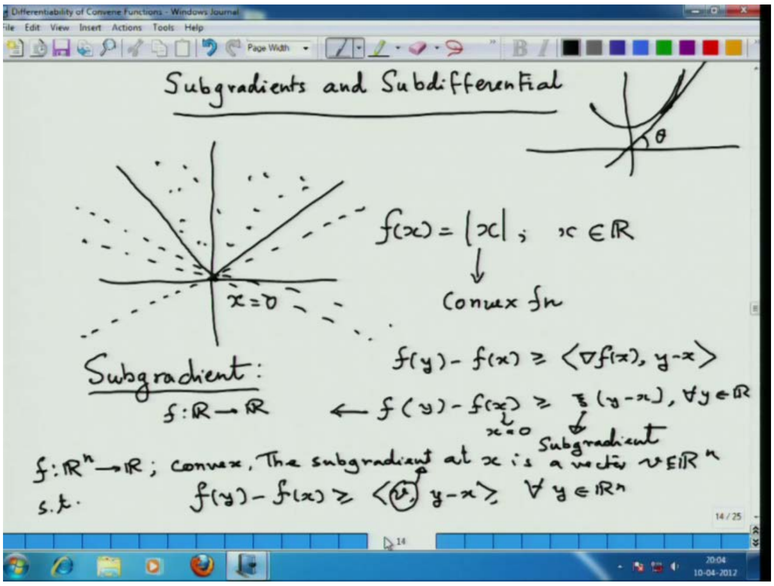
Convex Optimization

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

Lecture No. # 08

(Refer Slide Time: 00:32)



So, welcome once again to the course on convex optimization. Good evening to all of you those who are viewing this lecture. If you remember, in the last lecture I had asked you this simple question, not simple, but the question is simple, may the answer might not be that. If you consider the function f(x) equal to absolute value of x, of course, x is in real number. Then at x equal to 0 the function possess as a minimum, while you also know from very simple calculus that it does not possess a derivative where x is equal to 0, because the derivative when you approach from this side is minus 1 when you approach form the other side this 1 is plus 1.

So, left and right derivatives are different; left derivative been minus 1 and right derivative been plus 1. So, derivative does not exist. Then how… If you are faced such a function and observe that if you look at the epigraph of this function, this function naturally is a convex function. The question is how would you really talk about the derivative when the derivative is actually absent? Can you really talk about the derivative?

This issue was solved by introducing the notion of a subgradient. Let us intuitively motivate the notion of a subgradient. Now, what is the derivative? Derivative is nothing but the slope of the tangent drawn at a given point on the graph. So, if you are finding derivative at of the function at x, then at the point x f(x) if you draw the tangent, if you have a nice function like this. If you want to find the derivative of the function at this point you draw the tangent at the point x f(x), I should be it more nicer and then the slope of this tangent basically tan of theta is your derivative.

Here we cannot talk about such a single nice tangent. For example, at x equal to 0, (( )) ok, this line is a tangent this line is a tangent, equally correct, because they are not touching the body of the epigraph and they just touching the curve at this thing at only one point x equal to 0, x equal to 0 itself this line, y is equal to 0 itself is the tangent; why of course this is not a tangent, there could be. So, any line whose slope is varying from minus 1 to plus 1 is a tangent. You can say that they are tangents, because this is minus 45 to you this plus 45 to you.

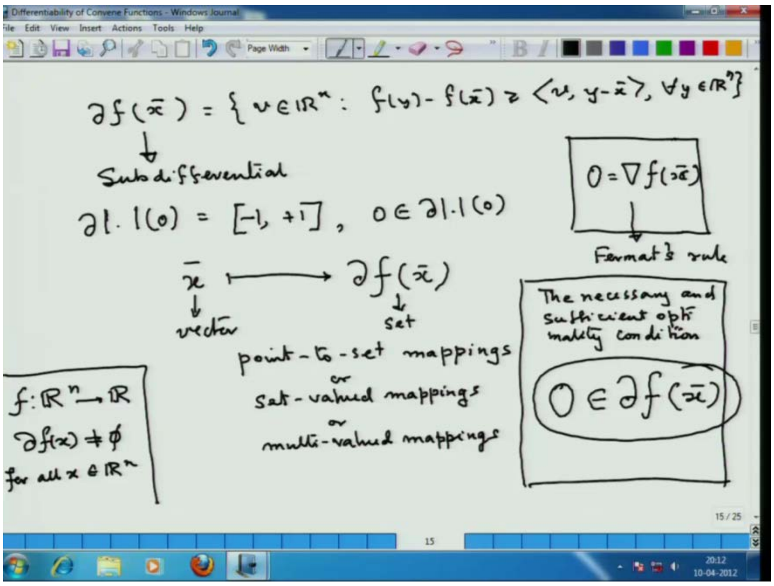
Now, what does gradient satisfy for a convex function? If a convex function is differentiable then we have this by now well known formula for you also. (No audio from 03:48 to 03:53) Now, here if I consider each of the lines is a tangent, each of these dotted lines are tangent. Then if you observe, that all of these tangents are lying below the graph. So, which means that if I considered for this case a slope for this particular function f, the slope of each of this graphs, each of this lines, then you will have this formula f(y) minus f(x). Some sort of slope which we are which, let us write it as some psi into y minus x. I am writing psi into y minus x, because we are in the real setting.

So, if my f is a function from R to R just like the one above this is what I am actually getting. Because I am considering each such line and here there is infinite such line where in between minus 1 and plus 1. So, any number that is lying between minus 1 and plus 1 is a slope to any of any one of these lines. So, given a number between minus 1 and plus 1 there is a tangent line to the graph at x is equal to 0 whose slope is that particular number.

So, such a psi is called a subgradient. So, in general, when you come for a function from R n to R, a function which is convex; so, here x what I am talking about is x equal to 0, because I am looking at the tangent at 0. The subgradient at x is a vector psi in R n, do not confuse with this psi, this psi was in R if you are confused I would just rub it off and maybe I will give it a new name. Instead of psi I call it any p element of R n such that f(y) minus f(x) is greater than equal to v times y minus x and this should be true for all y in R n. You observe that this is this fact what I am writing down has to be, because this is true for every pair (x,y). So, if I fix my x equal to 0 here, then at for every y this they can actually holds. So, this is actually true for all y in all.

So, this v is what is called a subgradient; this v is my subgradient. Now, the question would be which v would represent the gradient? There are many, many vs many, many competing these you see the many, many competing vs, I can claim, this is my favorite v and I want this to be the representation of the gradient when the gradient is actually absent. Somebody comes along and say no, no, that is not fair, you know x equal to 0 is the minimum then 0 must be equal to the gradient. So, this y is equal to 0 line should be the gradient. So, this can be resolved by taking a more holistic view of the thing and making a amazing jump in the imagination. This jump is one of the most fundamental advances in optimization in the last past few decades. Though people would not compare it with the interior point of evolution which took place in linear program and which we will talk in the later course. But this is a immense importance to the development of convex analysis and convex optimization theory in general. So, people said ok.

(Refer Slide Time: 08:11)



Let me put all the subgradients in one set and I call that set the sub differential of f at x. So, how the sub differential looks like? It consists of all possible v in R n which satisfies the property f(y) minus f(x) bar is greater than equal to v times y minus x bar for all y in R n. So, I collect all such possible vs. So, in our possible this scenario of f(x) equal to mod x function or maybe I should write it is like this del of mod at 0. This is nothing but the correction minus 1 to plus 1. So, this is the sub differential of the absolute value function. So, this is what it means if I go by this then this is what is the sub differential.

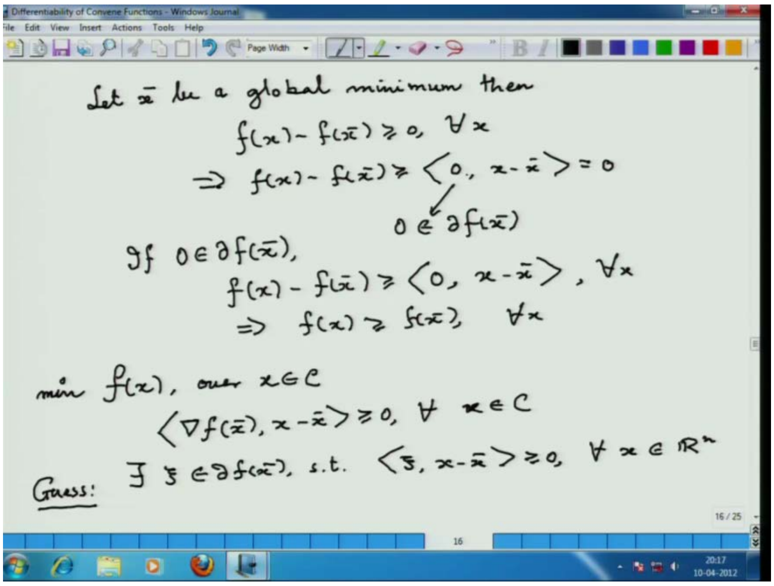
Now, the interesting part of this is that you have done this, but observe there is something very, very important. Now, I have made a very importantly my gradient is no longer of vector, my gradient is a set. So, given an x bar my gradient is not a vector, but a set. So, I am taking a vector in R n and by the sub differential or something which I am claiming to replace the gradient of a convex function where it is non differentiable by a set. So, such mappings are called point to set mappings or set valued mappings or multi valued mappings whichever you want to call them. Now, why such a choice is correct, why do you think that this sort of approach is a correct approach? See whatever we do even if we come to the regime of non differentiable function, we have to note that the Fermat's law for optimality that is when you have a differential function and f has a local minimizer at x bar. This is the fundamental rule that it should at least follow should satisfy this.

So, if I am trying to figure out the sub differential at a point of minimization like this particular case, then 0 somewhere must be lacking inside the sub differential. And here we are observing that 0 which means at this is really or this is a true extension of the gradient for the at the point of differentiability for a convex function. So, you now see the optimality condition changes. The necessary and sufficient condition - optimality condition; this is the following. It now changes form an equation to an inclusion to what would be called as the differential inclusion. So, x bar is a minimum of a convex function if and only if 0 belongs to this.

Now, any mathematical sensible person would ask me, you are talking about 0 belonging to a set and all this (( )) stuff but what if such a set is non empty empty. The important fact now comes to forth at one f is a convex function from R n to R, then this is not equal to zero, then this is not equal to the empty set not zero sorry this is not equal to the empty set for all x in R n. So, this non emptiness is a very, very important part. So, all this what we are telling make sense. So, whole of convex optimization the optimality criteria is actually nothing but trying to tell this in many, many ways. In fact, our recent book that I want to show you which has come out in convex optimization with myself being one of the authors is that - this book where I am also one of the coauthors with doctor Dhara Professor at IIT Gandhinagar. You see that this symbol 0 belonging to del f(x), the symbol has been in the cover.

So, in some way I am trying we are trying to say that whole of convex optimization. This book you know optimality conditions in convex optimization. This symbol by giving this symbol on the cover we are as authors where trying to say that optimality conditions in convex optimization essentially this story. So, here we are at the very hot of convex optimization theory.

(Refer Slide Time: 15:00)



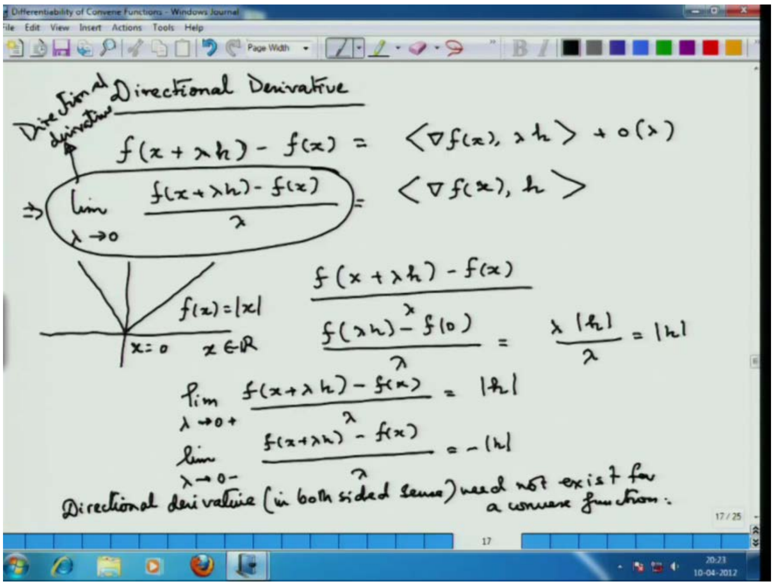
Now, once I know this necessary and sufficient optimality condition, let us try to prove this. Now, let x bar be a global minima, you know in convex optimization there is nothing like a local minima. Let x bar… (No audio from 15:02 to 15:17) Then x bar be a global minimum then f(x) minus f(x) bar is greater than equal to 0 for all x. This is simply implies at f(x) minus f(x) bar is bigger than the inner product of 0 vector with x minus x bar. And this simply says, because this is nothing but equal to 0 this simply says that 0 belongs to del f(x).

Now, I know that if 0 belongs to del f(x) bar by the very definition of a sub differential, I would have f(x) minus f(x) bar greater than equal to 0 or x minus x bar; so, 0 be in the sub gradient. So, this would immediately imply that f(x) is bigger than f x bar for all x; this is true for all x, for all x in R n of course. The next natural question would be… This is for unconstraint case, if I have to minimize a convex function f(x) over x element of c, how do you talk about necessary and sufficient optimality condition and I do not have differentiability of the function, say at the even at the minimum point.

Now, how to handle this thing? So, let us think it over such as intuitive if this was differentiable optimality condition is both necessary and sufficient; of course, f is a convex function, c is a convex set. In our study, we have to be very careful. Unless and until mentioned the functions are always convex and sets are always the functions are always convex and sets are always convex and usually closed. Unless specifically told to be open. So, this is… no sorry. So, if grad f(x) minus f(x) bar sorry you have sorry I think I have not done this thing. So, optimality condition that we know; that if f is differentiable then x bar is optimal if and only if. (No audio from 17:56 to 18:04) For all x in c. Now, the question is, if it is not differentiable how do I replace it with the subgradient? Let us make a guess. Guess is the following. There exists psi, some psi, I do not know what it is, in del f(x) bar, such that psi times x minus x bar is greater than equal to 0 for all x for all x element of R n.

How do I prove this? You might try to do it in the way we have done it for differentiable case, but remember there is nothing like a Taylor's theorem that you have seen in the non differentiable case. In order to do so, get an idea of this we shall refer back to what is and different approach to non differentiability of a convex function. But ultimately would be link with the sub definitional notion. That is a notion of a directional derivative.

(Refer Slide Time: 19:25)



(No audio from 19:25 to 19:38)

So, how do you define a gradient, what happens, if you write down the Taylor expansion. So, if I have if the function is differentiable then if you write down the Taylor expansion then you will have this lambda v minus… I have been using v for sub differential. So, let me put h, lambda is greater than 0 possibly for whatever, some lambda. This I am just writing down the definition of derivative. So, limit as lambda goes to 0. (No audio from 20:25 to 20:41) Because (( )) I will divide by lambda on both side, this will vanish and this is what I will have. Now, here lambda is supposed to go both from the right hand side to 0 and left hand side to 0. It could be an increment in any direction, it does not matter. But for a convex function, this thing may not always happen.

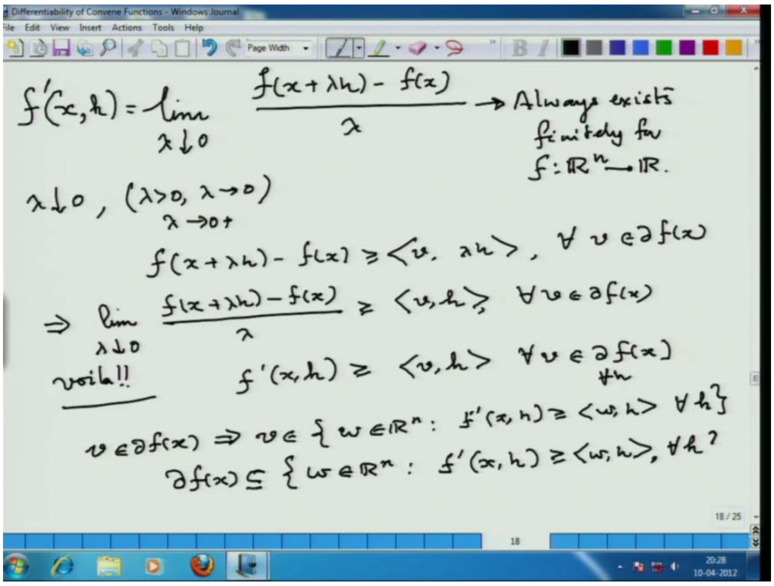
Again go back to the prototype example. (No audio from 21:06 to 21:13) So, if you look at the point 0 - x equal to 0; prototype example f(x) equal to absolute value of x, x is in R. So, let us look at this in this particular case x plus lambda h, x is 0 minus f(x) by lambda. First I take lambda from the right hand side - this side, lambda going to 0. So, this will be f of lambda h minus f(0) which is 0 by lambda. So, this will be lambda of mod h by lambda which is mod h. So, limit of… (No audio from 22:04 to 22:10) So, what I have done? I have made an increment along with given direction h – along a given vector h.

So, I have move from x, I move a little bit along vector h. So, this sort of derivative is usually called the directional derivative. Now, if the function is differentiable then the directional derivative is equal to as you can see from the Taylor expansion is equal to the inner product of the gradient and the direction. So, this sort of limit is usually called a directional derivative. (No audio from 22:42 to 22:53) So, lambda goes form the right side 0 plus. So, in this particular case, this is mod h. Now, when it will go from the left one.

(No audio from 23:07 to 23:22)

So, the both these two limits do not agree and so I cannot say anything about differentiability and which you know that is not differential. So, directional derivative in the both in both ways do not agree. So, directional derivative in this sense of this sense of the limit, limit been taken both ways need not exist for a convex function. (No audio from 23:42 to 23:57) So, directional derivative in both sided sense need not exist for a convex function. (No audio from 24:05 to 24:15) But an important thing happens.

(Refer Slide Time: 24:25)



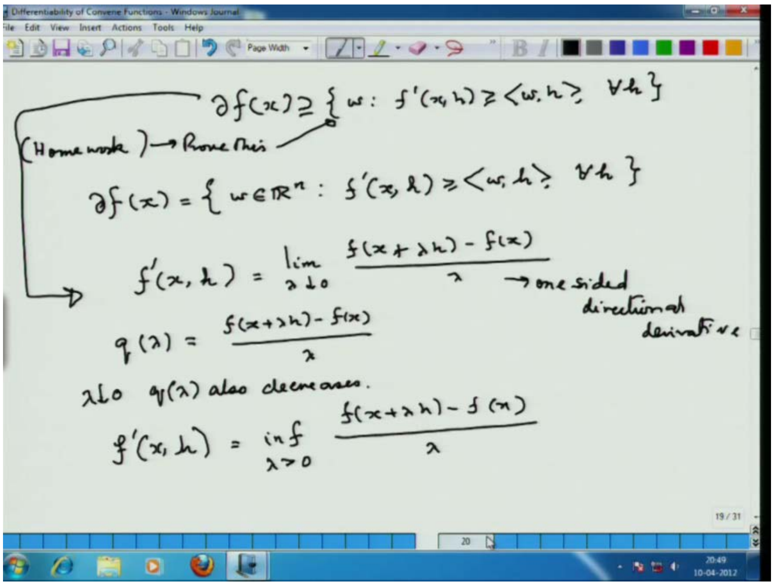
The directional derivative in a one sided sense always exist for a convex function that is if I take this limit where lambda is only coming from the right hand side. This symbol lambda down arrow 0 means lambda is greater than 0 and lambda is going to 0. Another another way of… These are shown another way of writing lambda is 0 plus. So, if f is a convex function then from R n to R, then this limit one sided limit always exist always exist finitely. (No audio from 25:01 to 25:13) Now, in this limit exist finitely, so what? How does it link with the subgradient? Can you give me a formula like this for the differentiable case; can the subgradient related to it like in this way? Let us figure it out.

By the definition of subgradient, I of course have following formula true. (No audio from 25:40 to 25:49) For every being, the sub differentiable of x, this is what will happen; just by the definition of the sub gradient. So, this would imply… (No audio from 26:00 to 26:20) So, voila has the French would say, thus French have a lot of contribution in optimization. This is the link between the differentiable case, between the differentiable case the link between the gradient and the directional derivative. Here we have a similar link between the subgradient and the one sided directional derivative. So, for a convex function, let us give it a particular symbol. We will call it the subgradient of f in at the point x in the direction h. And for a convex function give me a point then and give me any direction h this limit would always exist. Give me an h whatever point you take that limit would always exist; sub different the directional derivative always exist. Now, which means what I have?

(No audio from 27:21 to 27:35)

So, if… So, it means for every v in del f(x) this happens. So, v element of del f(x) implies v is element of the set, set of all say w such that f dash of (x,h) is bigger than w h for all h. So, if I fix this x and I keep on changing h, and for every h you take, and for any fix sub differential you take of for all, the basically this formula holds for all h also. So, you fix up of v then for all h this formula will hold. You fix up a h then for all with this formula will hold. So, basically this so, sub differential becomes this subset of this. (No audio from 28:44 to 28:50) It is all collection of all such w which satisfies this inequality for all h.

(Refer Slide Time: 29:01)



So, the next question is what is the convex. So, I am… See the whole idea is trying to parallel the differentiable case. So, I had this one. So, my question is when will I reverse it when will that be a subset. (No audio from 29:21 to 29:31) Now, I will leave this as homework, this is indeed true that is what you will have is that… So, here comes the beautiful relation between the sub differential and the directional derivative. A lot more thing to say that how do you prove this one. So, homework is through this. So… Because I have a lot more things to say about the directional derivative also; let us give you a hint, because it might not be so easy for those who are not in the subject to really prove this, because it is quite an advance thing. Because once you learn of all this things you are at the frontier of knowledge as far as convex of (( )) is concern.

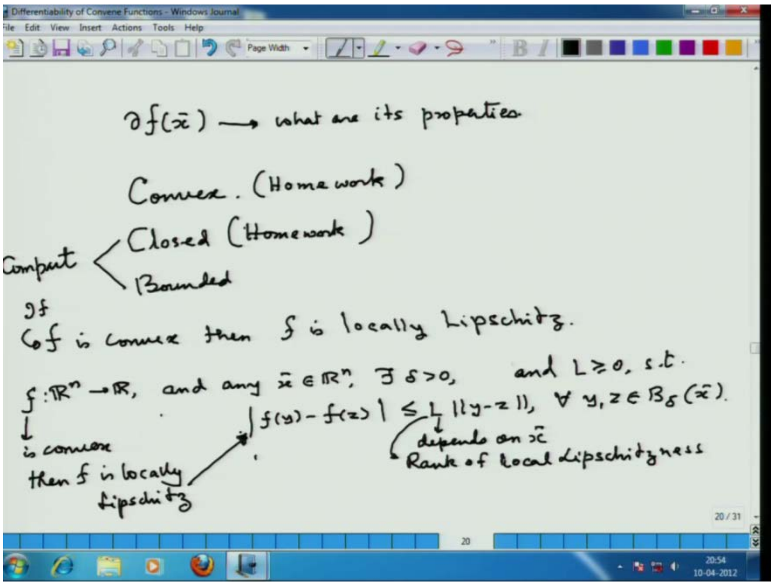
You see here you have to observe a very, very important fact that if you look at this expression. (No audio from 30:44 to 31:01) Let me just look at for a fixed lambda and fixed x n h, if I fix my x, fix n h then this is nothing but a function of lambda, sorry I will just look into this this differential quotient is a function of lambda. (No audio from 31:22 to 31:29) Now, as a function of lambda as lambda decreases, as this as lambda goes to 0 q lambda also decreases, and then the limit of q lambda should be the lower bound of q lambda as lambda goes to 0. So, in fact, one can also write that f dash (x,h) can be shown to be the infimum of lambda greater than 0 of the differential difference quotient. And this fact is fundamental to prove this fact, this inclusion. So, you take a careful note.

Now, let us summarize what we have learned till now. Convex function need not be differentiable at every point, the there a good chance and it is generating property that minimum, usually lies at the point of non differentiability. The non differentiability can be tackled by the introduction of subgradient and the notion of sub differential derivative which replaces the notion derivative. So, we are not really shown how it replaces. We are taking it as something in leave of the derivative. And for a convex function the directional derivative need not exist, but the one sided directional derivative exist. So, this is what we will call the one sided directional derivative.

(No audio from 33:00 to 33:08)

So, one sided directional derivative of f at x in the direction h and it is also linked with the sub differential in this fabulous fashion. So… And of course, we have the fundamental formula which is very, very important that 0 is element with del f x, but this holds true and so now, let us go back, let us not go back, let us just put go back to have an fresh look again at the sub differential.

(Refer Slide Time: 33:38)

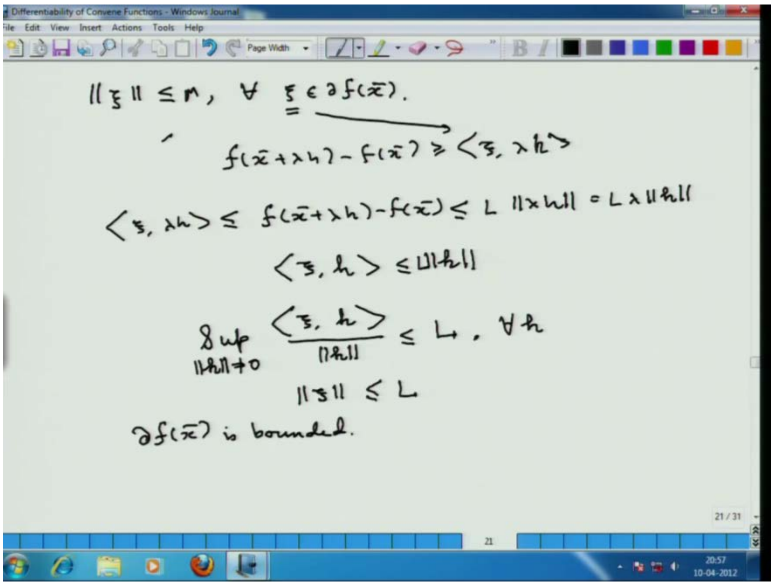


So, if I have the sub differential of a convex function from R n to R and a point x bar what are its properties. (No audio from 33:44 to 33:54) First property is that it is convex, this is homework. So, del f(x) is a convex set. The second property is that it is closed, this is homework for you. Those who are engineers might not be so convinced with (( )) how to prove closeness and all those things. Let it be not for the… But for the mathematics people all watching this show; they should definitely try to figure out how to prove it is closed. It is not very, very a big thing.

Another thing is that it is also bounded. So, this two makes it compact, because we are infinite definition and del f(x) is thus a convex and compact set. Now, how do you prove that it is bounded? That is a very, very important thing. But in order to prove that it is bounded we need an need to know some additional things about about a convex function. So, what we first should know that - a convex function f is convex, if f is convex or I should write if f is convex then f is locally Lipschitz. So, what do I mean we have why this statement are locally Lipschitz. So, for any function f from R n to R which is a convex function and any x bar in R n there exist delta greater than 0, and L greater than equal to 0 such that f(y) minus f(z) is less than or equal to L times norm of y minus z for all y and z belonging to this.

So, this L is highly dependent on your choice of x bar. Basically L depends on x bar. If you choose a different x bar then it the L changes - this L will change. This is called the rank of rank of the Lipschitz function - rank of rank of local Lipschitzness or local Lipschitzian rank whatever you want to call it. So, this constant will change if you change your x bar. But if this constant does not changes once you change your x bar which means then such a function is such a function is called globally Lipschitz. So, any convex function, if f is convex then f is locally Lipschitz. So, any convex function will satisfy this property.

(Refer Slide Time: 37:40)

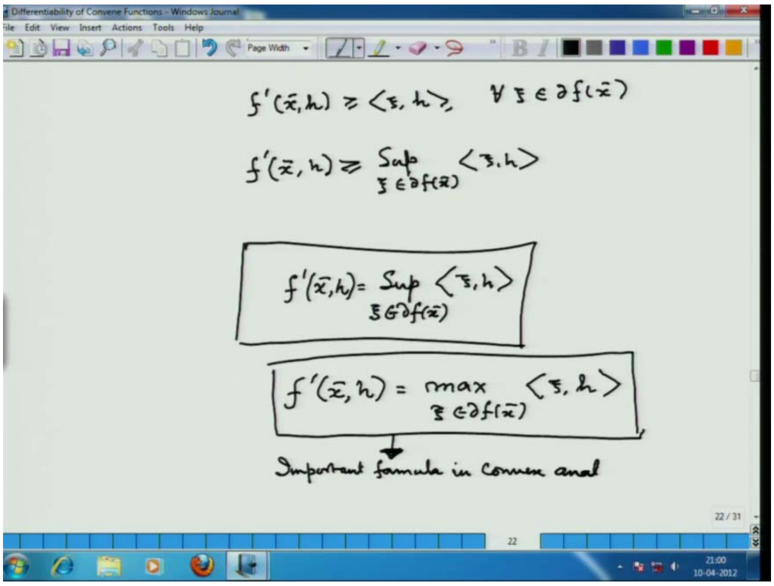


Now, this property would be used to show that sub differential is bounded set. To take any psi… What do you mean by bounded? I have to prove that norms are if is less than a fixed number say m for all psi in del of f x bar. Now, how do I do this? So, observe this fact. Psi… I will just be more clear in the very beginning. f of x bar plus lambda h minus f of x bar is greater than equal to psi times lambda v for any psi belonging to this, this is true, not psi times lambda v, psi times lambda h. Psi is a subgradient.

Now, what happens is that? Once I know this I can I will just write it in a slightly different way, just change the sites. Just change the way of writing. So, when lambda is sufficiently small, x bar plus lambda h is very near x bar. So, it is within some neighborhood which you want. And then applying local Lipschitzness you will get, this is nothing but lambda times norm of h. So, this is nothing but L times lambda norm of h. So, this cancels out this lambda cancels out to this psi of h. Now, here comes the more mathematically oriented thing.

Now, I can write this as psi of h of course, the h is 0 and h is 0 this is anyway satisfy. So, if an h is not equal to 0 I can divide. So, there is a L here. Again divide it by norm of h and like this. Now, because these two for L for all h; whatever h you take this is this is always be true, again take h repeat the argument, take another h repeat the argument. So, you can take supremum over… This is slightly mathematical just bear with me those who are not so involved in mathematics. So, this is by very, very standard technique, you can show that this is nothing, this is defined as norm of psi; this thing that sup of this is nothing but defined as norm of psi and that is less than equal to L. So, this is true for all psi. So, that shows that del f x bar is bounded. So, it has good properties.

(Refer Slide Time: 40:34)

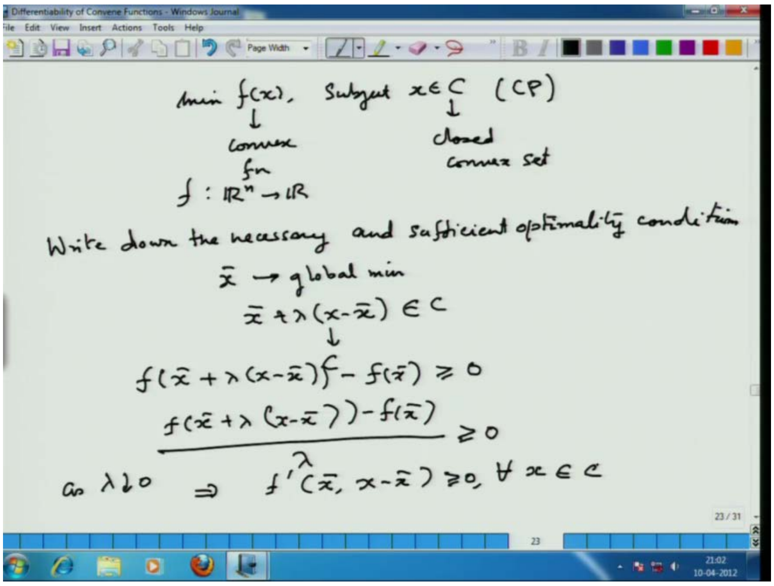


Now, what you have observed is that the directional derivative x in the direction h is always satisfying this property. See if I fix the x and fix the h this result is true for all psi in del f x bar. So, let me take x bar here. The question is, is this only a inequality or actually it is equality or strict inequality holds. The beauty and this is how the subgradient is finally link with the derivative is that the sub gradient is basically becomes a quite a good representative of the derivative, you see here this formula. So, I can now write this is greater than equal to sup of for any psi in del f x.

Now, where whether there is this can be a strict inequality or always equality holds. One of the far reaching results of convex analysis is to show that. (No audio from 41:46 to 41:55) Now, but because this is this is a compact set for a given x bar and this been a linear function. One can also write this as… So, this is how truly this is through this dual nature this is a this is how truly the directional derivative is linked with the linked with the sub differential. So, this is one of the most important for reaching formulas in convex analysis and as many applications.

(No audio from 42:37 to 42:48)

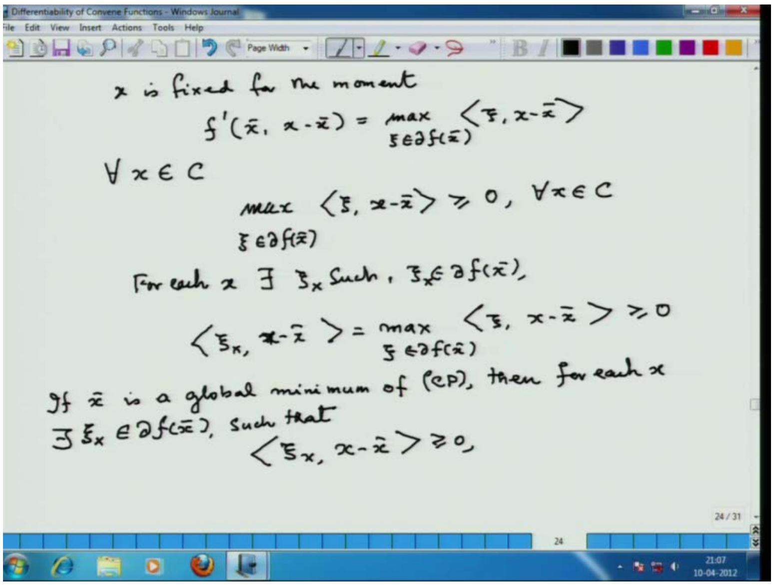
(Refer Slide Time: 42:53)



So, now let us go back to our question by which we are (( )) to let us talk. Minimize f(x) right subject to x element of c, f convex and c closed and convex convex function from R n to R and see closed convex set. What is the necessary and sufficient optimality condition? Write down the optimality condition that is what we wanted. Write down the necessary and sufficient optimality condition.

Now, if x bar is a global minima. So, this is my problem CP which I had already written convex programming. Say x bar is the global mean, consider this thing x bar plus lambda x minus x bar where x belongs to c, then by convexity of c this belongs to c, and by the global optimality of x bar this is greater than equal to 0. Now, I can divide by lambda now what I can do is I can divide by lambda. (No audio from 44:35 to 44:46) Now, I take the limit as lambda tends to 0 is that imply f dash x bar x minus x bar is greater than equal to 0 for all x you know. Basically this is the formula, because this is true for any arbitrary x.

(Refer Slide Time: 45:16)



Now, what happens? Suppose I know the fixed x, suppose my x is fixed for the moment x is a fixed for the moment. (No audio from 45:18 to 45:27) So, this is the directional derivative right and this is your whatever. This is your direction x minus x bar. So, this... So, I can write that f dash of… Now, I will use that formula which I am which I wrote as an important formula. So, this I can now write as f dash x bar x minus x bar is equal to max of… (No audio from 46:03 to 46:12) Right, this is what you can have.

Now, once you have this, but this is for every x this is greater than equal to 0. So, for all x, I will have now max of… (No audio from 46:34 to 46:50) So, this is my optimality condition. Now, for every x, for each x there exist psi, such that psi belongs to del f(x) bar, of course, psi depends on x. So, such that psi x of x minus x bar is equal to max of…

(No audio from 47:37 to 47:49)

So, what I have proved that if x bar is an optimal is a is is a global minimum of CP. (No audio from 48:02 to 48:15) Then for each x there exist psi x such that psi x, x minus x bar. Now, if I go back a little bit earlier, my formula say that there exist a psi element of this my am I guess was that this will happen for all psi for all the x this will hold true. But I seem to I have got something slightly different. What is the difference that for every x my x changes. There is no harm I have perfectly deduced it. So, mathematically this is fine. So, from here is this any way to go to the one I have made a case, can that is that case right. For that we have to study a bit more convex analysis and show that. That case is right and both these formula from here to there can be obtained and from there to here can be obtained. So, for that we have to little bit more study about convexity and convex analysis and that would be taken up in the next lecture. Thank you and good night.