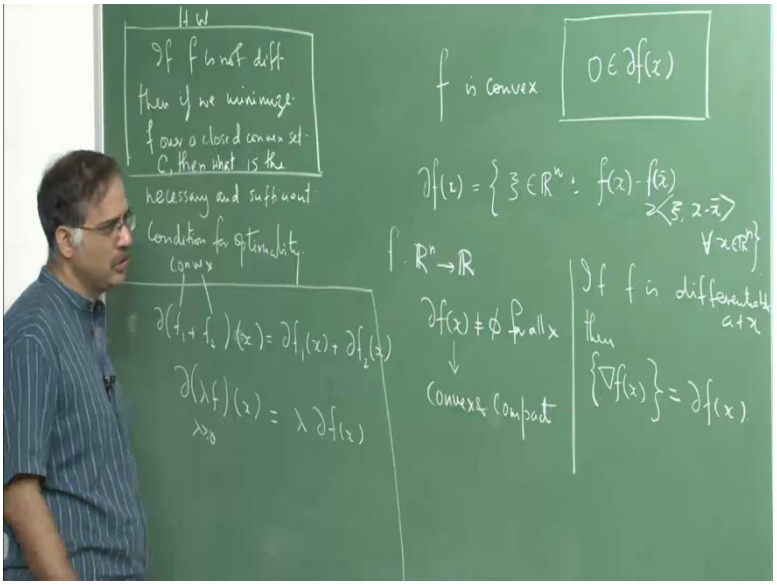
Foundation of Optimization
Prof. Dr. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 27

(Refer Slide Time: 00:36)



So, it is important to tell you that or rather recall that in the last lecture, we ended our discussion with the introduction of the sub differential of the convex function f. So, f is convex and we define something called the sub differential of f at x. Of course, I am assuming the function is from R n to R. So, this is f of so this set is now, always for a for an f is from R n to R n sorry R n to R then del of f at any x is not equal to phi, for all x and this is both convex and compact. So, it is convex close in bounded.

So, these are the fundamental properties of the set valued map or a or at the set, which mimics the derivative at every point x where the function is not differentiable, and also I give as a home work to just check out that if f is differentiable then differentiable at x, I would say then this single tone set actually equals the sub differential, as the sub differential has only one sub gradient and that is nothing but the gradient. So, these are few features of the sub gradient you should know, and obviously the most important feature. This is a necessary and sufficient condition for a point x to be minimum of the function f over the whole set R n.

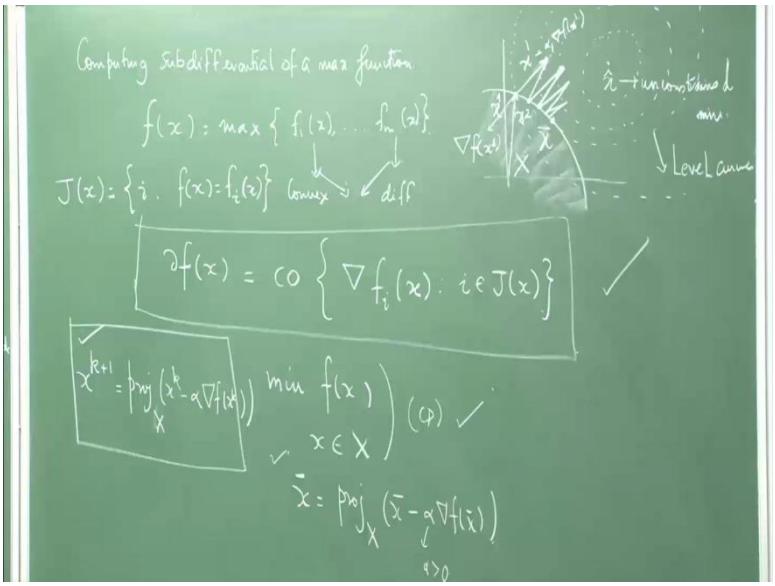
Of course, I now leave it you as a home work and I will put it up in the if a FAQ's which will be in the course website is that, if f is not differentiable in the non differential means, it does not have a derivative at every point then, if we minimize f over a closed convex set c, then what is then what is the necessary and sufficient condition that is your home work. So, those who can they can check out it from my lecturers on convex optimization in the same as NPTEL series, then you will be able to figure it out what is it then, what is the necessary and sufficient condition for optimality.

And it is also important to know that if I say that this is imitating the derivative, then this must be able to give me some calculus rules. So, it really gives me some calculus rules, but there will be one rule which differentiates this convex non differentiable calculus from the standard differential calculus that you know. So, it says that del of if you have two convex functions see here everything is convex.

So, I am not writing convex, if you want I just write convex I am giving you the most simplest. I am giving you all these in a most more simpler descriptions actually in more advanced text you have, f from R n to R bar, R bar is an extended real line which includes minus infinity and plus infinity. So, this set sorry computed at any x in R, R n would now is same as the Minkoskey sum of these sets.

That is any element here is the sum of one element here plus 1 element here. So, another calculus rule which is simple to understand, you take any lambda for where lambda is greater than equal to 0, then compute this function lambda f at x from the integral sub differential this is same as doing this. So, there is a some rule like the derivative there is this rule of getting the constraint out, there is also composition rule what which is must more complicated. So, I will not talk about, but there is this rule about the max function, which may be of interest.

(Refer Slide Time: 07:22)



I just give you the formula rather than more detailed description for that, you really have to go to the convex analysis course, convex optimization course. So, suppose you have function f x, and you see that here I assume that all these functions are convex and differentiable, all of them are convex and differentiable. Now, f is convex, but not differentiable. So, the standard composition rule composition by some functional method max is not applicable here, you cannot use the derivative. So, you need to compute the sub differential of this.

The sub differential is actually computed in this way first you call for a given point x, you construct the set J x which consists of all the indexes, I chosen from 1 to m for which f of x this function equals to that f i x because here, I have a finite number of functions. So, once I put plug in n x at least one of them must be equal to this because f x is the max value of among all this.

So, once you this particular set this is nothing but compute that gradient of f i at x collect those, compute the gradient of gradient of f i at x, but those i's which belongs to the set J x. So, you do not compute for every function i. So, you reduce the computational effort once so given an x, a computer program can immediately recognize the set J x and so, for that you just have to consider for that x only those i's, which belongs to J x or your computational effort reduces, and then you have to take collect those elements and take

the convex hull. So, there are only finite number of elements so, it will be a polyhedral set and polyhedral compact set.

So, give me any I can definitely give me an x and if the function is not differentiable at that point, right that is when at there will be at least more than one i when J x is more not just singleton those are the points, where the function is usually not differentiable and then if you ask me to give me an element of the sub gradient. Sub gradient at that point as the element of the sub differential, I can use very simply compute by using this formula.

Of course, how do you really construct the convex hull to geometrically view it, that is a very difficult thing to do you can use random method many, many exciting things can be done out of here, but from a computational point of view, but the interesting part is that it tells you that if I just want an element. I can give you an element because this can be easily enumerated. So, computer program even if this is large and then immediately you can know which are the grad f i's you have to compute, and then you just compute an element of the convex hull, you choose any thing choose the lambda 1, lambda 2, lambda n in the way you want and so, make some random choice and you will get an element.

So, you see this is a beautiful formula, and that is what makes convex analysis so powerful and this convex non smooth calculus so, powerful. This has no analog in the smooth calculus or the standard differential calculus that you know. And one must also realize that the fact, that non smoothness or non differentiability does not just arise out of the blue, it comes by taking the maximization of certain functions or minimization of certain function, it typically arises in this form. So, then we really have a nice method a constructive method to compute the sub gradient, and that is why the subject is so beautiful and so powerful.

So, I do not get into too much of details with convex issues, but let me go back to this so called projected gradient method that I had wrote down in the last lecture, that if I am to solve minimize a convex function f x over a closed convex set x, then I know that x bar is a solution if and only if it is a projection on x of x bar minus alpha grad f x bar of course, I am assuming differentiability, where alpha is anything greater than 0 this is an if and only if condition.

So, from this we could we said that then we generate generated an iterative scheme. So, you really have to choose a step length alpha carefully for the shown x, y well I wrote x

here. So, this is called the projected gradient method, and we are going to today discuss why this works and I want to tell you what happens if the function is not differentiable, can you have a similar sort of formula for then, can you have a similar sort of necessary and sufficient optimality condition, which would lead to a projected gradient method.

So, which could be slow actually in practice might be, but mathematically you would lead to a method, which have a beautiful convergence, in the sense that under certain very simple conditions all the sequence. The sequence that you generate the any accumulation point of such a sequence is a solution of the convex programming problem, this one CP. So, it is important to have a look at, it is important to have a look at this sort of formula projection formula because now, this formulas are also used from much more different sort of problems, if there is special stocks as like SDPN and other conic optimization problem.

So, we have we want to discuss this one so, you might ask me why do you think that such a formula would work, how do you think that it will give us a solution the idea could be very simply like this that what happens in optimization is suppose this is your constraint boundary, feasible set boundary and this is your minima unconstraint minima of the convex function, and just for the time being assume that this level sets are level sets that is set of all x for which f x is some alpha that is f x values are fixed, objective values.

Now this, so the minima the constraint minima, because as the function values are increasing in this direction; so, a constraint minima is here so it may be x constraint minima is here this is x bar, but I may not be able to start with x bar immediately. So, this is your unconstrained minima x hat that is unconstrained minima, x bar is unconstrained minima. And you have to notice that this sets these lines, which I have also called level sets level lines or level curves. So, anything inside this is called level sets or lower level set.

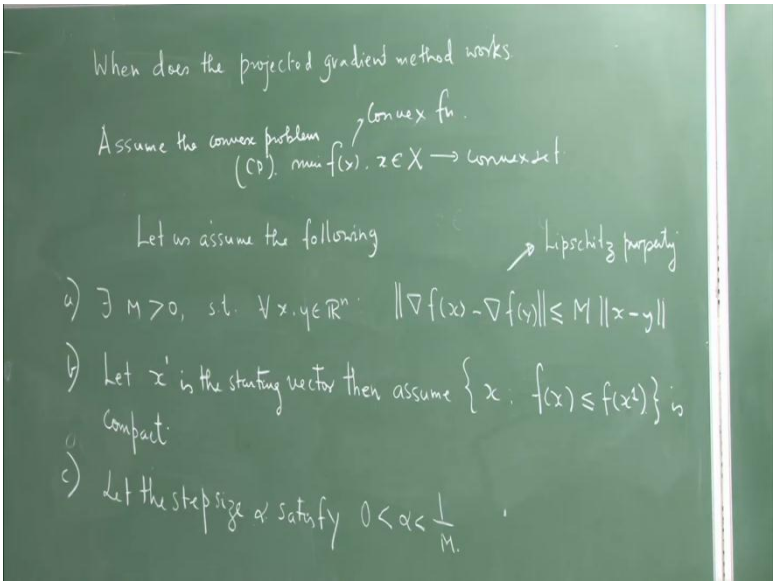
Let me do tell you something now what would happen, if I start from a point here a level curve is also passing through this point naturally. So, this x, x 1 see here the gradient of this function is always outward normal to this curve, to the level curve gradient at that point to f. So, here again if you want to find the gradient it is actually here so, if it is a gradient of f at x 1.

So, I am looking at the negative of the gradient right so, this is my x 1 suppose and from there I move in the direction of negative of the gradient. So, this is my up to certain distance alpha like so which is x 1 minus alpha grad f x 1, which you see have gone outside the set x right which is outside the set x is this point, and from there I can draw a projection or a perpendicular on this set that would give me x 2.

And with x 2 I can do for example, if the level curve is here its again the similar thing again, again from here which is x 2 minus some alpha this is alpha 1, I guess alpha 2 grad of x 2 you drop this you do this, and come like this towards the solution. So, this is a geometrical meaning of the projected gradient method. Now, once the geometrical meaning is settled it is very important to tell you mathematically, when and how this actually works that the solutions that has generated, would really give me the actual solution.

So, I will now use some of my notes here to figure out, why this method works and after that I will give you a question. Suppose, I just consider a differentiable function over a closed convex set, I want to minimize just a differential function can I use the projected gradient method projected gradient means because you are projecting the gradient, which is obviously projecting the negative of gradient in some sense, a gradient at a point negative of the gradient and projecting this.

(Refer Slide Time: 20:21)



So, my question is, when does the projected gradient method works? So, let us make certain assumptions, assume the convex problem C P. So, take this problem so, where f is a convex function and x is the convex set fact, I think is should repeat because this is not really a convex optimization curve. So, this is convex function and this is convex set. Now, I will assume that the function is of type c 1, 1 may technically called c 1, 1 in the sense that the gradient function, which is a vector function the function from R n to R n a wave is a Lipschitz function, is a Lipschitz function. Of course, this projected gradient method works for a the case where you can have nice projections. So, I will give you tomorrow some cases, where you can have a very simple projection formula, and for those cases, which are also important in applications this sort of method will actually work.

Now, here so we will assume that the gradient of f has a nice behavior, let us make a list of assumptions exists m greater than 0 such that for all x y in R n this is a nice behavior, I can give you an example their quadratic function you will have this, take a convex quadratic function half x transpose a x then this is immediately satisfy and I urge you to find others.

This result is due to a Rosin's key it appears in this book, I think is should mention it appears in this book non-linear optimization published by Princeton University Press around 2007, this m times norm of x minus y. So, this is something called the Lipschitz property that the distance between the functional values, and the distance between the original points has this relation.

So, the distance between the functional values is less than some constraint of the original point, right if m is 1 then we call that map as a non expensive map B now, when we start an algorithm we start with the point x 1 and suppose, that x 1 is not the solution of the problem and then we assume then it is clear that the solution must lie in the following set, let x 1 in the is the starting point let x 1 is the starting vector, then assume this is a crucial assumption, it is the picture that is actually happening figure out why you see x 1 here if it is a starting point if it is not the if is a solution fine, if it is not the solution and the solution must lie in this set that is you are minimizing the level set of x 1 f x 1.

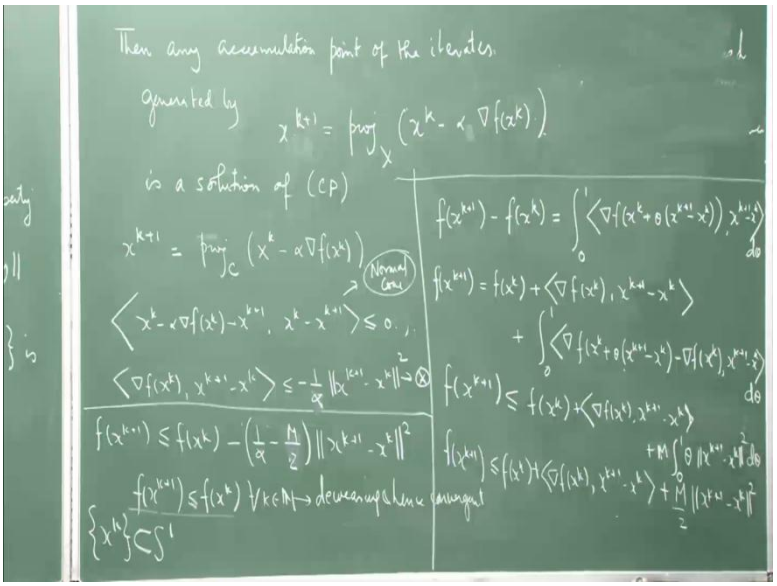
Now, if the solution is lying in this set and if this set is compact then I basically I have to minimize the function over this set, and then if functions is this is the functions

continuous, I will be able to pinpoint the minimize because minimizes are would exists because this function is this set is compact. This set is compact is guarantying that the minimizer is actually existing right. And what we will do we will generate iterates in this set that any x 2 must have the x 2 should in this set x 3, should be in this set x 4 should be in this set and the accumulation point of that sequence, should also be in this set because this set is closed and that would give us some insight. So, this is one major result a major thing.

Now, let the step size alpha so, alpha here is called the step size just like the line search method, line search method this will be x k minus alpha this is something like a Stephens gradient thing, but here you have the projection business. So, here phi naught there this alpha is I can always write as alpha k if you want, I can this is called the line it is called the step length as in the line search case.

Now, let the let the step size alpha satisfy so many assumptions are there and now, will prove. Now, if I have take if I have taken these assumptions then let us see what happens this is my convex problem CP. So, if these three assumptions hold these are not very unnatural quite natural for example, they would be very useful in solving quadratic optimization problems. This can be very useful in solving a class of problems called the linear complement directive problems where m is positive, whether the matrix is positive. So, may definite we will not get into that we will just give an, give those examples afterwards, but here let me do something here let me try to come to the conclusion.

(Refer Slide Time: 29:15)



So, if this happens, then any accumulation point or limit point of the, I directs generated by the projected gradient method. So, you choose your alpha step size should not exceed this quantity sorry x, k plus 1 that is it is generated by the projection, projected gradient method. Now, if you do not know again alpha has to be chosen by you right so, it could be alpha k. So, generated by this projected gradient method, then any accumulation point of the iterates generated by this is by this is a solution of C P.

So, this will generate and iterate sequence x 1, x 2, x k. So, this sequence x k actually will show lies here, because it is a compact set, and this it will be bounded sequence on every sequence has a in a compact set, every sequence has a convergence sub sequence. So, there will be a accumulation point, which is a or a limit point just forget about this for limit point limit may not exists whatever.

The limit exists, then it was limit points if not accumulation point, equated to that accumulation point. So, x k is bounded, so there will be a convergence sub sequence which we are not renaming and it will go to some x star. So, that x star is a solution of C P that x star is a solution of C P. So, our class today will end by the by justification of this fact. So, our class again, I say that class today will end by justification of what we have said which means, giving the proof.

So, let us see so we will start with some sort of integral so called integral form of the mean value theorem, simple nothing to this is a standard formula. We will rub this part

we will do work it on this part also is a convex combination of these two points, that will usually appear in the mean value thing under the d theta here. So, what do you essentially do is you add a term grad f x k inner product x k plus 1 minus f x k, if I add and subtract.

So, I can write now that f x k plus 1 is f of x k right plus grad of f x k x k plus 1 minus x k if I also add I also to subtract and then, I can put the whole thing down into the integral and write, I can write the I can subtract that part and now pull the whole thing into the integral because I can write any constraint into 1. So, 1 can be written as 0 to 1 integral d theta.

So, I can write this as see the idea to write like this, would be soon cleared that I can then use my Lipchitz property. So, basically when I add this term I can pull it into the integral just by writing 1 equal to 0 to 1 integral d theta, this is inner product d theta. Now, once you write down the, I can write this as less than equal to norm of this norm of this, which are Quasi's words.

So, finally I can use the Lipchitz property which is up there to write that this is less than f x k plus m times 0 to 1 theta norm x k plus 1 minus norm x k, this first norm would come x k plus 1 minus norm x k would come from the Quasi's words here, and then applying the this rule Lipchitz property here m will come out here, and theta is of course, the difference between this and this. So, there will be theta x k plus 1 minus x k and that theta would come out and that is positive. So, it will just come out of the norm so, that that is what you get here.

So, this will finally give me of course, with this you have to add of course, there is a another term added to this whole thing maybe I should write, I have to I should not forget this term, maybe I should write it in a better way x k plus 0 to f i plus m times 0 to 1 theta norm x k plus 1 minus x k whole square d theta. So, this would give me that f of x k plus 1 is less than f of x k plus grad of f of x k x k plus 1 minus x k plus m times or m by 2 into because theta square by 2 for 0 to 1 integral that is what you will get. So, let us keep this formula there, and try to work out the things here a part of which I will also encourage you to work out yourself, rather than just following this proof. You should know that here we are working with this sort of sorry projection. So, this is this has what happened.

Now, what does this mean that this is the projection of this, but just by the definition of projection means so, this vector minus this vector in the element of the normal cone. So, which means x k minus alpha grad f x k minus x k plus 1 into x k minus x k plus 1 this must be negative, this is just from the definition of the normal cone think about normal cone.

Now, this can be rearranged to write as so, you please do this check out this calculation and I am sure this calculation is correct. So, once you get this now apply this star apply this to estimate this thing here so, that would give you sorry alpha now because of this condition sorry m by 2 what I am writing, m by 2 m by 2 is here so because of this condition this is positive. So, here you have actually this whole quantity is a negative quantity which you are subtracting.

So, for each k for whatever be your k so, this is there. So, if I call this set as s 1 this set then this sequence this is so, what you have generated that for every k x k is a sub set of s 1 is in s 1, and this sequence is a monotonically decreasing sequence starting from this is true for all k's in the set of natural numbers. So, k is equal to 1, f x is bigger than f x 2, f x 2 bigger than f x 3 and so on.

So, this is a convergent sequence right this is a convergent sequence, this is a convergent sequence. So, this is decreasing and hence convergent obviously, this is all bounded below because of there is an infimum here, and the infimum is achieved because the infimum lies in this set and it is achieved. And that f x star value must be is the lowest among all of them, and you are decreasing you show there is a decreasing sequence which has a lower bound so it is convergent. So, the bounded sequence, which is convergent we applying very basically analysis.

Now, once this is known what do you have then, then it is your job to prove from this equation that x k plus 1 minus x k this norm is going to 0 show is a very simple thing, if you know some bit of manipulation and it can do a little bit of analysis, you can easily show from here that this result actually holds. So, if that result holds that would simply tell you that limit k tends to infinity projection c x k minus alpha grad f x k minus x k is equal to 0.

Now, let x star be the accumulation point so, let us take x star so x k is in s 1 which is bounded. So, x star is accumulation point for accumulation point for x k that is it is so,

there is a convergence sub sequence in x k which goes to x star. Projection mapping by the way, let me just tell you because for a convex set every part of given point of projection point is unique. So, the projection mapping is a continuous function it is a function which actually satisfies, this sort of property Lipschitz with m equal to 1.

So, the projection mapping is a continuous function then by using continuity what I will have is projection J x star minus alpha grad because f is convex function, it is a continuous grad f is also continuous because of this Lipschitzian property. So, any way it is continuous a convex function which is continuous on whole of R n, is convex function which is differentiable on the whole of R n is also continuously differentiable, but we also have this property here. So, this so ultimately we get this equal to 0. What does this equation tells us?

It tells us that x star must be same as the projection of J x star minus alpha grad f x star where alpha is the. Now, sorry I should be also doing a little bit of more analysis, here I should be putting alpha x k because this alpha k would change with, if I fix the alpha then I basically I have got a solution because this is a if and only if condition, but if I want to now put alpha as alpha k, then what I will have here is alpha k.

Now, here observe that alpha k has to be always within this limit, for whatever be your k alpha k has to be within this limit. So, alpha k is a bounded sequence and hence, it will have a convergent sequence which goes to some alpha star, and that is the proper analysis and here you have the alpha star, and so this so this would imply that I have found an alpha star which is greater than 0. Of course, alpha star is bounded and should have had the k, but for our simplicity our discussion alpha is fixed. So, if it is fixed then I do not have to bother.

Now, actually I should put a question now, because I do not know what would happen because there is strictly bigger than 0 here. Now, I put a question, if I want to vary the alpha in this scheme here alpha is fixed alpha greater than 0 is fixed, and alpha is between one by m and 0. Now, I want to say that how can I change alpha at every step. So, basically if I want to make a change here then I have to make a change in this so, instead of having this greater than just 0, I should have some number say beta which is strictly greater than 0 and it is here and then it will be a bounded sequence, bounded between two positive quantities. So, there limits it its limit it will be a bounded sequence.

So, its limit cannot go beyond this two positive quantities the limit has to be in the closed interval beta and 1 by m and then you can push that and put an alpha star here. So, when I change the step size to alpha k so, we had worked with the step size constant fixed. Now, if I change the step size change step size at each step, can you repeat this, can you repeat this argument. So, the home work would be to repeat this argument.

So, let me just make a short summary, we start with this projection method where my alpha is fixed if alpha is not fixed, then I have to do this at every k alpha k is has to be chosen. Now, let the step size alpha k whatever alpha you choose, step size at each at each k step size alpha k at each k satisfy this, then you can actually run down this argument. You can I have an alpha k here and that alpha k will jump zoom to some alpha star lying between beta and 1 by m and then you will again have the same thing, here with the alpha star and that will again tell you that x star is a solution of C P.

So, here I will so again I want to assert in our first starting, in the first write up my alpha was fixed then I had then I am changing my alpha at every step, but keeping it holding it between these two quantities, and then if that that can be done the whole argument can be repeated, and you can reach the same conclusion. So, with this I end and in the next class, we will talk about penalty methods for solving non-linear programming problem.

Thank you very much.