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

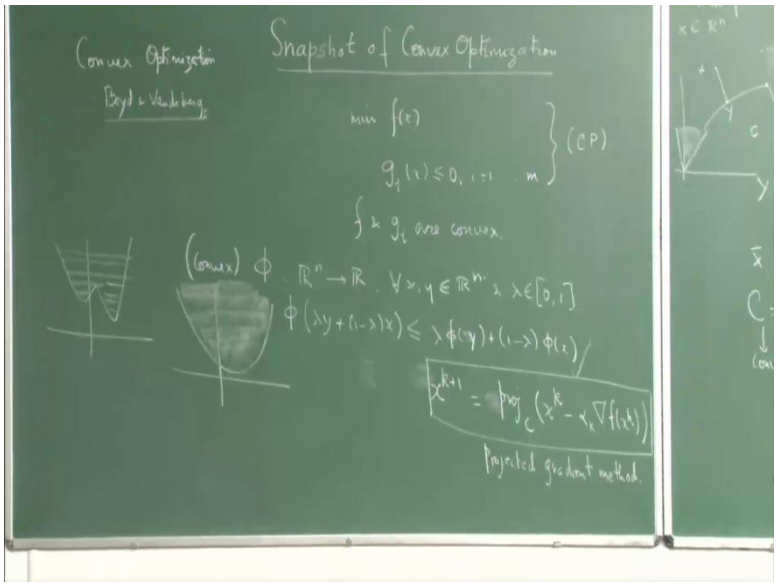
Lecture - 26

Welcome to this ongoing course on the foundations of optimization, in the last class we spoke about the SR1 method - the symmetric rank one method, but the it has easy to compute easy to update the matrices, which is approximating the inverse of the hessian in the Quasi-Newton octane method and it does not give you a possible definite matrices, but you know has some advantages.

So, and it has a good behavior, when you look at it means, it has a good behavior when you and when I apply it on, this on the, on solving the convex quadratic problem that is a quadratic problem with positive definite hessian. But there are lot of other issues which I have not spoken here for example, what is the rate of convergence of these sort of BFGS method or DFV method, it would, it would be stupidity just to mention you that the rate is super linear and so (( )), because you would not get a feeling about what does it mean, unless you really know the proof of that or you do the numerical experiments, any of the two has to be done which is not possible through this lecture. Then I largely want to keep it as a mixture of under grad and grad lecturers and graduate lecturers.

So, today instead of getting more into Quasi-Newton methods, which we can discuss if it will cause something; we can come back to quasi-Newton after some time, let us we bit more flexible rather rigid. So, you might get bored by doing Quasi-Newton, Quasi-Newton, quasi-Newton, so we just change over a bit for take a little bit of refreshment on the way.

(Refer Slide Time: 02:22)



And we would look at a snapshot of on convexity, convex optimization. Here we will take a snapshot, because I have a full course, on convex optimization, which is live and it is on you tube live on NPTEL, where a lot of details had been given about convex optimization. So, instead of really looking at convexity in too much detail, we are just give you a brief idea about what are the fundamental issues in convex optimization.

Convex optimization by the way is one of the most important areas of optimization because there are so much applications, which come as a form of convex optimization problem, and they need to be really handled. And the theory of convex optimization and the algorithms very are well suited and as Stephen Boyd of Stanford university, calls that convex optimization is almost a technology.

So, today we are going to introduce some aspects of convex optimization in fact, when we were talking about linear the Karush Kuhn Tucker condition for linear problems, for the fact that every (( )) on point has a normal (( )) on multiplier actually, I believe that the talk that a gave regarding the Karush Kuhn Tucker conditions for the linear programming problem the proof I thought I have been very clear, on a second thought I would feel like that.

So, I would like to do that thing little in a different way, you see what is important to realize that given a linear programming problem, it does not mean that all the fridge on multipliers are normal. Given a linear programming problem monitoring always assort is

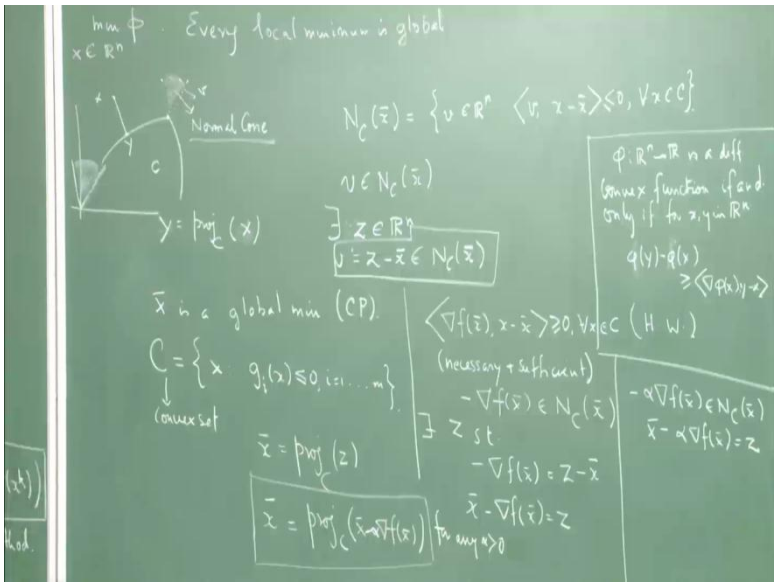
that there would exist a normal multiplier, an existence of a normal multiplier is equivalent to the satisfaction of the Karush Kuhn Tucker condition, existence of a normal fridge on multiplier, which I believed because I am trying to give you more broader view point.

Now, here we would be largely considered, concerned about convex optimization problems of this form where you minimize a convex function, subject to certain convex inequalities. So, f and g i are convex. So, those who have forgotten the definition of a convex function either you can see my earlier lecture on NPTEL, or I would just give you the definition if a function phi from R n to R n is convex, if you have that for all x y in R n.

So, this is convex this is a convex function for all x y in R n and lambda 0, 1 f of lambda y plus 1 minus lambda x is less than equal to lambda f x lambda phi x, for a phi y plus 1 minus lambda phi x, certain symmetry has to be observed here in spite of inequality is that if you have lambda y, here you have lambda phi y here this and this and this is actually a geometric fact, its convex functions looks like this usually.

So, the graph above the part of the plane for example, here above the graph of this function is called an epigraph, a function is convex if and only if its epigraph is convex and that is exactly that is that will you should assume that epigraph is convex, this is exactly what we will get. A function which is not convex, will not have its epigraph convex for example, this is not a convex set this is epigraph, what you can surely see the epigraph is not a convex set.

(Refer Slide Time: 07:09)



The importance of convex function arises from this very fundamental fact, as if you take a convex function phi and you want to minimize it over whole of R n, then every local minimized global the all these things are proved, in the last lecture last course. So, I am just giving a snap shot because I have a very different thing to tell you. What I want to re-assort here, and re collect here is the idea of a projection on a convex set which I have done and you already know about the notion of the normal cone possibly, that if you have a convex set c and if you take a point outside c, say x find the point on the boundary of c which is nearest to this, that is from x compute the distance of all the points in c and then you find, what is the point which minimizes that distance.

For example, if you come to here this is the nearest point come here. So, here you see there are lot of points whose nearest points is the same point. So, here they form a sort of cone and this cone is usually called a normal cone. So, if you have this point y, which is the nearest point and y is usually called the projection and x and x minus y is the perpendicular basically, projection on c x is called the projection of y on y is called the projection of x on c. So, you drop a perpendicular on c it will hit at y.

Now, what you do is a following the normal cone here, normal cone actually consists of these points at any points x bar in c consists a wall v in R n. So, I would request you to go back to the old course on convex optimization, and look in to the details and I think I have

given some details slight details earlier, but not to the extent that you are supposed to, but little bit of details.

So, if v is in N c x so, this v is in N c x this v so, v is in N c x for then it means, the following it means that basically, you are looking at the here the cone this cone contains 0 because 0 would satisfy it, you basically transfer the origin to this point and then give the geometry of this cone or you can look at it in this way that you draw parallels here. So, this is your cone translate them parallely to the origin so, any one of the ways you want to look at it. So, once you have this v element in c x bar means there exists some y or some z in R n because that z minus x bar, right z minus x bar is v is equal to z minus x bar is element of N c x bar.

So v is some z minus x bar or some lambda times z minus x bar. So, if you take lambda times z minus x bar does not make sense, you can take you pull it up you can take the z here also. So, you can always write like this now, when you have this. Now, suppose you have a global minima x bar, let me consider some facts about differentiable convex functions which I will just mention, I will not give any proof, but I will ask you to do a home is a pretty good exercise. Home work is this is on the differentiability property of convex functions. So, I do not remember whether I have actually spoken to you on this, but you might be wondering what this guy doing around here because he says he does not remember things.

When you teach science you teach at a flow, when this is your subject you possibly develop a private version of the subject everybody has a private version of the subject, if he is in love with the subject of course, then everybody gives is for a private version. The way I am teaching optimization does not mean that somebody else, some other researcher would teach optimization the same way the same spirit, I look at it more from a geometrical more end of view, and this is normal cone becomes a important vehicle of a expressing up to validity conditions and bring many other things.

So, for example a function phi from R n to R n this a differentiable convex function, if and only if for all x y in R n, f y minus f x sorry phi y minus phi x greater than equal to grad of phi x into y minus x for any pair x, y and it is an if and only if condition do it as homework, start with this definition apply or definition of differentiability and proceed.

Now, once you have that let us see what you can do now, if x bar is a solution so if x bar is a global minimum of this convex programming problem C P.

And if I have the fact that f is differentiable, how do I start analyzing it how do I characterize what conditions such a point would satisfy. So, that once I know the characterizing condition I can compute it, and because for convex case every critical point is a global minimum because here if you take any critical point here, then by this expression you will immediately know that if x is a critical point, then x is obviously global minimum this you can figure out or you listen to my other course, which will give you more details.

See here my idea is very different, here I am first going to consider the set C right. Now, instead of looking individually at these constraints first, I look just the set C as a convex set C that is all a closed convex set because we are assuming differentiability, may be or nothing or the set is always closed because every convex function is if from R n to r is a continuous function, that is something very important that one should remember, when trying to do something with convex function there is a advantage, the function if we define it from R n to r it is always a is continuous function. Examples of convex functions are abundant, and I would ask to do go and look at this book by Stephen Boyd convex optimization by a Boyd and Vandenberghe if you go to Boyd's website, you can find the whole book and download it.

Now, here so if x y is the global minimum of C P the following condition is a necessary and sufficient condition for optimality x in c means, all x which satisfies this so, this is necessary and sufficient. Once you look at the definition of the normal cone, you can immediately see that what I have here essentially tells me that the necessary, and sufficient condition has this geometrical expression, that the negative is always in the normal cone negative of the gradient.

Now, if that is so then there must exist a z such that so, this is minus grad f x bar is equal to z minus x bar right, if you have this then what does it show, what does it tell me minus grad f x bar is element of this. So, it tells me that x bar minus gradient f x bar is equal to z is z right, but what is x bar is the projection of z on projection of z on c. So, x bar is equal to projection of z on c, but hence that means. So, this is an if and only if condition now,

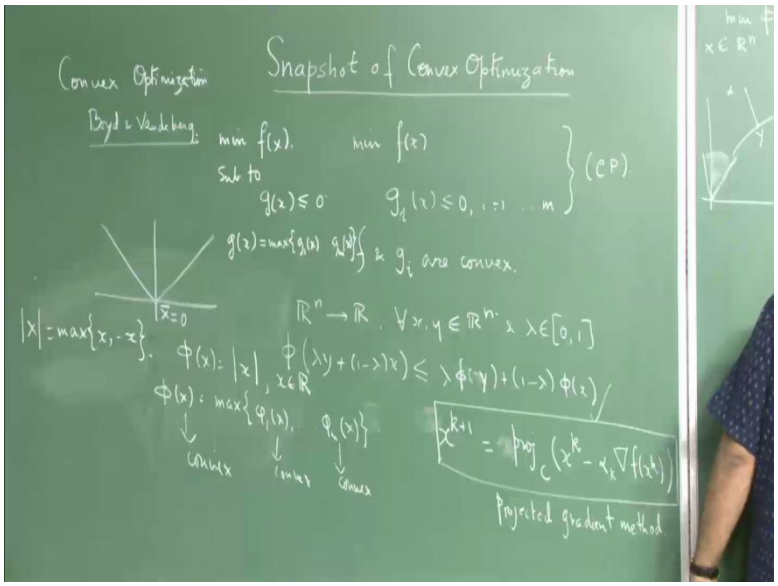
from this condition which looks very algebraic I write down a very simple condition it terms of projection.

I can refine it further, I can refine it further in the sense that I will do this little jugglery, I will multiply this by alpha so, because it is a cone this cone would absorb the alpha because for a cone if you take any element x, alpha times x is also an element of the cone so it will absorb the alpha. So, what I will have is minus alpha times grad of f x bar is element of N c x bar. So, I can then what you will have is x bar minus alpha grad f x bar is z and so, here I can just tune it that for let say x bar is an optimal to the convex programming problem C P, if and only if this happens for any alpha greater than 0. Now, I am sure you are getting some feel of a line search type writing because here you have an x, k here you have the alpha here you have the Stephen distance gradient minus grad f.

Now, this is equal to this suppose the projection of this is not equal to this then the x bar is not a solution, but this projection procedure gives you an idea of to write down an algorithm, it tells you it is called the projected gradient algorithm and it tells you what to do it tells you that you, take this and write down iterative scheme of the following form x of k plus 1 is projection of c, x k minus alpha k times gradient of f x k. So, my intention over the next 2, 3 days is to study the projected gradient method, or I will study the projected gradient method in a slightly more general sense. This is the iterative scheme for the projected gradient method, and that is one of our aims because we have to learn some algorithms about for the constraint case.

And the convex optimization problem is one of the constraint optimization problems, the problem C P where some sort noise algorithms can be developed, compared to if this things are simply non convex. What we are going to now, take into account is the fact that a convex function can be non differentiable and a large number of convex functions, can be non differentiable. Simple example is the function which you have learnt at high school.

(Refer Slide Time: 23:53)



So, if you have an example like phi x then this has a minimum, this is a convex function because its upper half make a convex set, the shaded part would be convex at x equal to 0 at x bar equal to 0, this problem has a solution, but it does not have the derivative. So, you cannot apply your standard techniques to do anything, it is a unconstraint problem you cannot apply a Newton method anything else.

Now, here is a difficult issue so even in fact at the point of minimum the precisely, where the non differentiability sets and that is the standard case, and that that is the generic case which has been recently shown. Now, what is important is to know how to deal with this differentiability, non differentiability and one has to appreciate that non differentiability just does just does not appear arbitrarily, it usually appears when you try to define a new function out of maximizing certain given functions. So, if you have a sequence, if you have say m convex functions and at every x, you define these are the m convex functions and you define if phi of x as the maximum at every x, you take the maximum all these and define phi x.

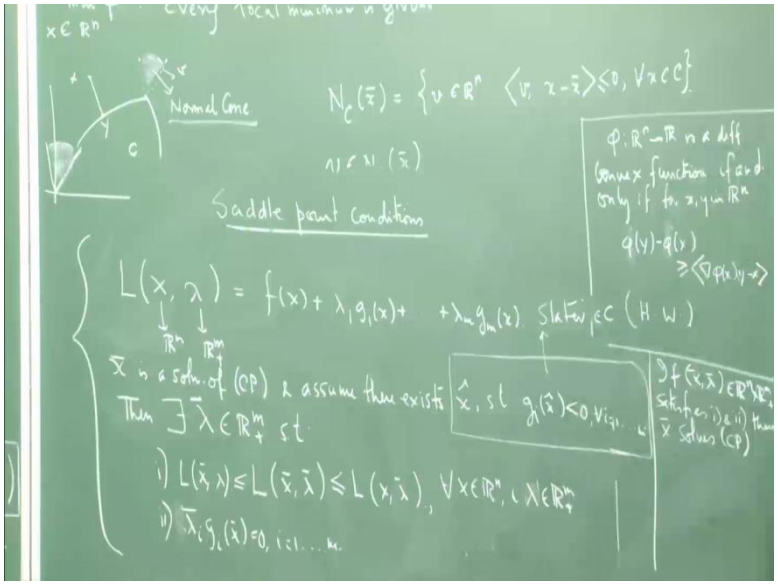
So, these are individually convex in most applications involving convex function, non differentiability arises precisely in this way. You see this max function is a very fundamental thing and it appears in optimization research pretty often, I can express this function with m constraints as a function with 1 constraint. So, I can write this problem c p

equivalently as minimize the function f x subject to, subject to g x less than equal to 0 where g x is the max of g 1 x.

So, when you take the max of few finite number of convex functions, if these are all convex then this is also convex. So, this problem remains the convex optimization problem it is an equivalent problem it is the same problem, but instead of many constraints I have one constraint, but the price that you have to pay for going moving from one many constraint to one constraint is that, you lose differentiability even if all these functions are smooth this need not be smooth.

For example, if you look at this function phi x the mod x, mod x can be written as the max of so, if the non differentiability here also arises because this can be written as the max function and in convex case most cases this happens. Now, how do you handle non differentiability, we you would like to device some sort of projected gradient type method, when I do not really have differentiability. So, the question would first come how do you handle non differentiability in your studies when suppose, you have this problem, how would you handle non differentiability.

(Refer Slide Time: 28:09)



So, first let us write down a very general type of optimality condition, which is called the saddle point condition, which I had not spoken, when we was talking about Karush Kuhn Tucker condition and all those things. Now, let us look into this thing pretty nicely now, how do I try to characterize the optimality of a point x bar which is a global minimum, to

do that we would need a something called saddle point conditions. We have to we will get something called saddle point condition, but before that saddle point conditions we come to something called the lagrangian function.

This lagrangian function would be very important would be part of a study for the next few days, when we come to also discuss the projected gradient type method. So, for the last next few classes 3, 4 classes we would be discussing this part of convex optimization and then end of the course would be a talking bit about the direct search methods, and other miscellaneous things.

Now, let me talk about the lagrangian associated with a function with this equal to 0 at every if I write this as lambda 1, lambda 2 dot lambda m then all of them are greater than or equal to 0. So, this is a function f x which is written like basically you pull in all the constraints. Now, suppose x bar suppose x bar is a solution is a global minimum, solution of C P which means, it is a global minimum and assume additionally that the Slater condition holds means, assume that this set feasible set c has an interior in the sense.

And assume there exists x at such that g i x at strictly less than 0, for all I then it this happens if this is given to me this is given and this is also given that that also happens, then there exists this is the sign of there exists lambda element of R m plus sorry lambda bar element of R m plus such that, L of x bar lambda bar is less than equal to L of x lambda bar less than equal to L of x bar lambda, this is true for all x in R m and all and this means for all x in R n and for all lambda in R m plus basically, this is almost unconstrained minimization in x, this is this condition is hold, this condition holds lambda bar i g i x bar is 0.

Now, this is just not a necessary condition, this is a sufficient condition also that if there is a x bar and a lambda bar a pair which in R n cross R n plus or R n cross R m plus, and it satisfy these two condition then x bar is a solution of the original problem. So, here you have an if and only if condition. So, we also have this fact which we now write down on the side is that once you know this, if so these two these condition if this particular condition is called the saddle point condition, and this is called the complementarities, complementary slap knees condition as you had seen in the fridge on condition and all which we have studied earlier, if x bar lambda then x bar solves C P.

I am not going into the proof because those who are interested in the proof, should see my earlier course, but this information is pretty important and to what we are trying to do this after this. So, here you see there is no assumption of any differentiability it just a condition this saddle point conditions are free of differentiability, but these saddle point conditions are important because they would lead to something called duality theory, which we will have a idea later on.

Some we will take have idea about them, but then the question would be if you do not have differentiability is that is there something, which we when replace the derivative is there something via, which through which we can work because here it is giving me these sort of information and it is also easily computable, you cannot compute a x bar by doing this through this inequalities.

So, there must be some sort of a equations solving or some sort of thing which give me more information of course, when you really solve a problem you do not use a really solve you do not try to solve Karush Kuhn Tucker conditions, or whatever because it is an if and only if in the case of convex functions. Now, I ask you a interesting question now suppose, this condition was not true what would have been the shape of this?

So, this a homework to you if I do not have this condition, we just call the Slater condition, Slater. So, if I do not have the Slater condition then what would be the shape of the saddle point conditions, that is what I want to ask you that would be a homework. Now, once that is done you can realize the link between this and the fridge on conditions and all those things, this is very important to do this exercise yourself.

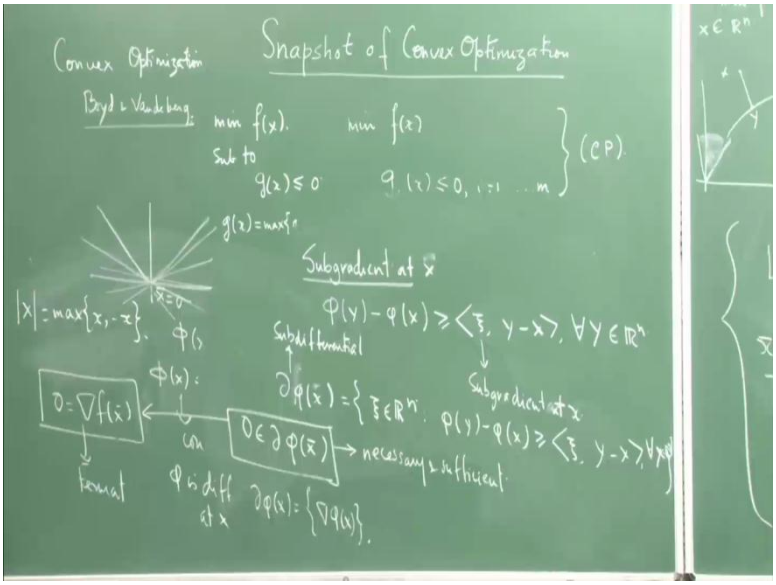
Now, another interesting question I want to ask you, what would happen if a my set in all these constraints that you see g i x less than equal to 0, they are all linear or affine would you need this condition to get this, that is that is a interesting question and I would be happy if you tried these things out. So, if these g i's or not at not non-linear convex, but they are affine they are convex, but affine functions linearly you can take linear also then y n f is just the convex function, what would be the shape of this saddle point conditions.

Now, if I want to talk about how to replace the derivative at the point of non differentiability, the best thing to realize is the derivative is nothing but for a function from r to r is the slope a of the tangent at that given point. Now, at the point x bar here to this curve I can not only draw one tangent, I can draw many tangent in fact there are infinite

tangents, whose slope varies from here whose which are the slope of one to here which are the slope minus 1.

So, because all of this tangents can be represented if tangents at this point, we collect them into 1 set and call that this set is now, replacing the derivative at this point of non point of non differentiability, and that gives rise to the notion of sub gradient of a convex function. And this would be an important thing that we will use in our projected gradient method. So, we are not bothering much anything about differentiability of the convex function, but we are talking about sub gradient.

(Refer Slide Time: 38:00)



If you observe for all this slopes of this tangent, which is varying from minus 1 to plus 1 they will all follow this fact, take any one of them and take the point 0. So, I am defining the sub gradient at a point x. So, I call a point psi a sub gradient at x sub gradient at x, sub gradient at x if this relation holds for all y in R n, you can try to verify at least with 1 or 0 all these things.

Now, here you see there are not only one sub gradient, there are infinite sub gradients. So, I collect them in one particular set, called the sub differential of the convex function phi at x bar, at a point x, and this is the set, which replaces the derivative at the point of non differentiation, you ask me why? And this set is actually non empty at every point in R n, it is very interesting that now, I am talking telling that the derivative is the set, it is not a number function on a function.

So, the sub differential this is very useful because you see, at if the function was differentiable and if x bar was the minimum, local minimum and a global minimum for a convex function, then grad of f x bar phi x bar would be 0. But here you observe if x bar is a local minimum here, this line x axis is also part of is one of the tangent whose slope is 0. So 0, so for a general convex function phi, this becomes a necessary and sufficient condition for optimality in the unconstrained case.

And of course, if the function is differentiable at a point x, then this is nothing but the singleton derive. So, if phi is differentiable and this set consists of nothing but the gradient of phi at x. So, you see differentiable at x, so you see this now this has come from, so instead of the equality, which we had called the (( )) in the beginning. So, we have this which is called the non smooth (( )).

So, the equality becomes a inclusion here because here, I have set it is not just a vector with that you can equate it to 0 it is a collection of vectors, and this is what is very important and just like the ordinary derivative has a calculus, this also has a calculus, but it has one interesting thing that you can compute. The sub differential of the max function which you have to do because that is how a non differential functions come, but a max function a function, which expresses as the maximum of a finite number of functions is never differentiable. it is usually non differentiable.

So, for a max function there is nothing no calculus would be the usual derivative because that is actually a composition of two functions. The rule has to be with the sub differential and with this we will stop. We will just give a very brief idea of the sub differential, and then we will come to the projected gradient method in terms of the sub differential, and we will try to analyze it in a next class.