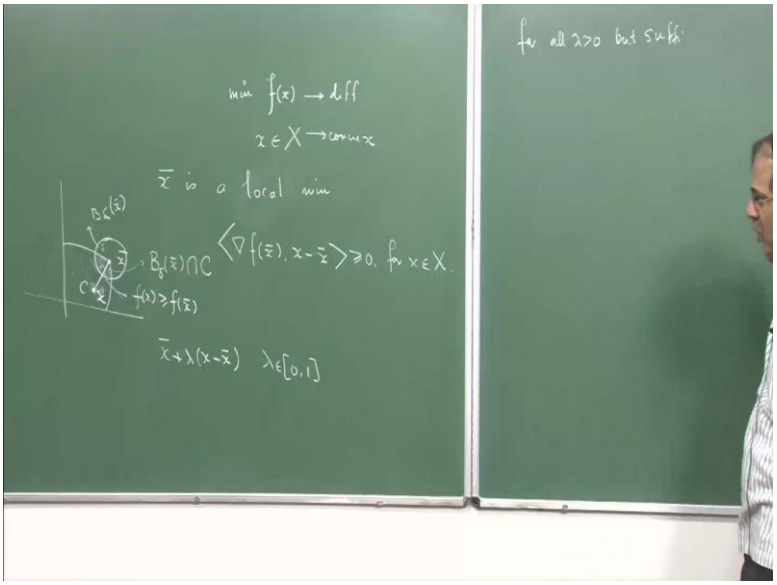
Foundation of Optimization
Prof. Dr. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 28

In the last class, we had spoken about why the projected gradient method works for the convex optimization case.

(Refer Slide Time: 00:37)



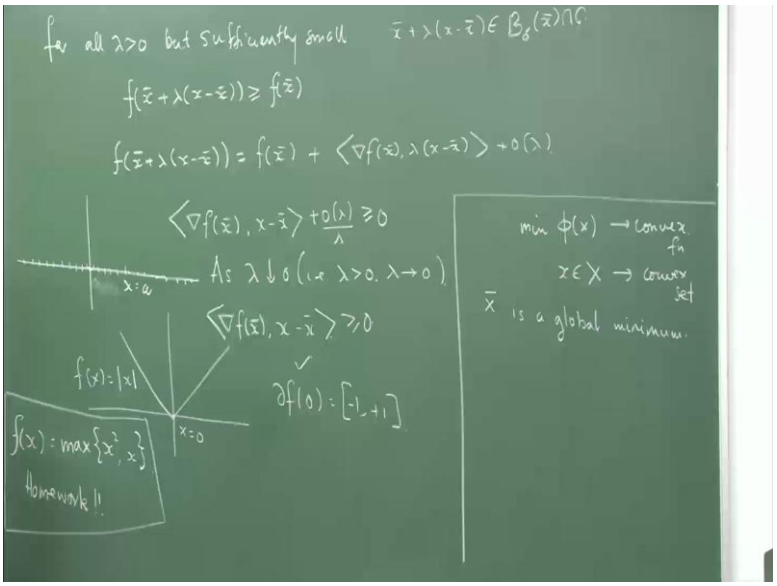
And, I asked you a question that two questions rather; one question was that suppose instead of a convex function, if I replace the objective function by a function, which is just differentiable, not convex and try to minimize over a set x, which is a convex; then, will the projected gradient method be a suitable method? That is a very, very important question is, whether the projected gradient method will be a suitable method? The issue is to observe that, in unconstrained optimization, we were actually not seeking a minimum; we were seeking something called a critical point. But while seeking the critical point, at every step, we were trying to decrease the value of the function. So, when we get a critical point, because every local minimum is a critical point, we were trying to convince ourselves that, under certain good conditions, this point – x that we get, be quite near a local minimum. Or we are just happy to get a critical point, which we can then test to be a local minimum or which we know from numerical experience is usually near the local

minimum, because we are always decreasing the value of the function at every step or… So, our algorithms are called monotone algorithms.

Now, for this problem, if x bar is a local minimum; no longer a global minimum, because I am not talking about convexity. So, let us take a local minimum; then, what is the condition for optimality? The condition for optimality here again, noting the fact that it is a convex at x, can be given in this form. It is not much difficult to prove it because of this following fact that… Now, suppose this is the C; and, this is the C and this x bar is the local minimum, which would mean that, there exists a wall of radius say delta around this x; that is, this is nothing but B delta x bar such that for any point, which lies in the intersection of B delta x bar and C, f x must be bigger than f x bar. For any x here; any x that is coming from this region, f of x must be bigger than f x bar. So, that is exactly the minimum idea of a local minimum.

Now, here I have said, this is true for every x you take in x. So, how can I show that? Now, if you take any x here, which is not in this intersection between the ball and the set, because this is the intersection and… So, local in the sense at that the x bar is a global minimum of f over this particular set; that is local over C. Now, if you take an x here and you join this with x bar by a line, which you can do, because this and this whole line would be in C because the set is convex. And then, you see that, any point here on this line can be written as x bar plus lambda of x minus x bar; where, lambda belongs to 0, 1. So, when lambda is 0, I get x bar; and lambda is 1, I get x. So, when I move lambda from 1 to 0, I am moving from x to x bar. I am moving along this line; from x to x bar. Now, if that is the case, then there will be a threshold value of lambda, for example, here; beyond which all these values x bar, these vectors would lie in this domain.

(Refer Slide Time: 05:48)



That is, for all lambda greater than 0 are sufficiently small. This is the standard way of writing what I have just told that… Or, you can write that, there exists a lambda naught such that for all lambda strictly bigger than 0, but strictly less than lambda naught; such a thing will happen. B del… This will be inside this. So, for all lambda greater than 0, but sufficiently small, you have that – have x bar plus lambda x minus x bar belonging to this. Once this is done, this means that, f of x bar plus lambda of x minus x bar is greater than equal to f of x bar. This is true by definition.

Now, what I will again do is to apply the Taylor's theorem or apply the definition of differentiability of the function to get… So, this is equal to f of x bar plus del f of x bar comma lambda of x minus x bar plus o of lambda. And, that simply means that del f of x bar comma x minus x bar plus o of lambda by lambda greater than or equal to 0. Now, a very definition of small o lambda, which we have discussed earlier when we were discussing the differentiability of a function from R n to R; then, this means as lambda tends to 0 plus… As lambda tends to… This means that is lambda is greater than 0 and lambda tends to 0. This will go to 0 in the limit. Because x is arbitrarily chosen, I can repeat this argument for every x proving what I had asked for here that, this should be greater than equal to 0 for every x.

The same argument about the projected gradient method can now be repeated for this case. But, in that case, what we will get at the end is no longer a solution, but a critical

point (( )). So, at the end, what we will get; the end product of the what is that called – of the projected gradient method has a sequence converges. It converges is not to a solution exactly, but to a point, which satisfies the fact a point which satisfies something is exactly this condition.

Now, it is also important to understand what would happen? I asked a question, what would happen, if f is not differentiable? Then, what sort of optimality condition you write and when, because when f is not differentiable, taking on a convexity would only complicate matters and which is beyond the scope of this syllabus. So, here we will take a convex function phi, which is convex to be minimized over x element of x; convex function, convex set. And, here we no longer ask phi to be differentiable. But we know that, at every point, even if derivative does not offer the subgradient is there; again, I recall for example, this function f x is equal to mod x; subgradient at 0, where it is not differentiable. That we have done already.

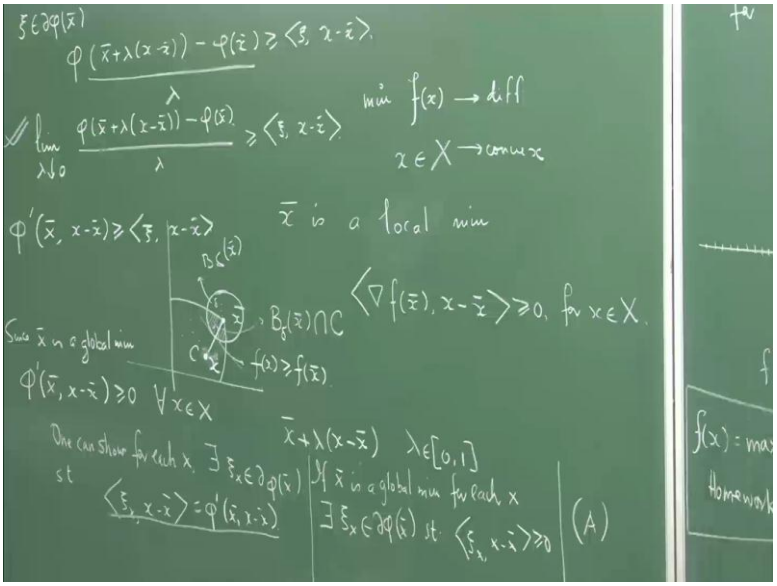
See the technique of finding a subgradient of the convex function is not so easy in general. But, for simple convex function from r 2 (( )) it is not very difficult either. Even for very slightly when a function in r 2 to r, it is not such a very difficult thing; the function is simple. For example, if I take the expression f x equal to max x square and x; we will try to draw this function and we will try to see how to compute the subgradient. The idea is very simple. Here if you observe all these points, the function is differentiable, except 0. At all these points, the function is differentiable. So, here along this line, on the negative side, I take a sequence x k, which is going to 0. Here the derivative is nothing but minus 1. So, it is minus 1, minus 1, minus 1. At every point of the sequence, the grad of x k is minus 1. Here it is plus 1 on this sequence. So, I have minus 1 plus 1; minus 1 plus 1; minus 1 and plus 1. These two values of grad of x k; if you take the limit of grad of x k, it will be minus 1 and plus 1. For this sequence minus 1 for this sequence and plus 1 for this sequence.

Now, I have to take the convex of all of that. If I take the convex of all of that, I will get exactly this one. Let me leave this as home work for you. So, what you do, the idea is the following; that if you have a point on the real line, x equal to 0 is the solution or x equal to a is the solution; does not matter. So, x equal to a is the solution on the minimum of the convex problem or (( )) or x equal to a is the point where you want to find the subgradient. Then, look at all the points. Construct a sequence x k, which is converging to

a from the left and all the points, which is converging to a from the right. Compute the gradients here and take the limit as x tends to a. Compute the gradients here; take the limit as x tends to a. And then, whatever you get there, whatever you get finally, just take the convex all of that; and then, basically, you get an interval, which is the subgradient. So, let us just try to draw this and try to do it as a homework.

Now, if x bar is a global minimum of this – the global minimum of this problem, how do I write down a necessary and sufficient optimality condition? Now, you understand subgradients would come, because there is no other choice, because I have not mentioned explicitly that, the function is differentiable. In this case, we have to use the subgradient. But how do we go ahead and use the subgradient? We will search an approach like this work, because I do not know that there is anything like a Taylor's theorem for subgradients. So, let us start thinking about it. Now, here you see I am no longer having an unconstrained case; I am having the case, where x is also there. There are many ways to view it, many ways to find the solution. But let us look at certain tools that are required to get this one.

(Refer Slide Time: 15:25)

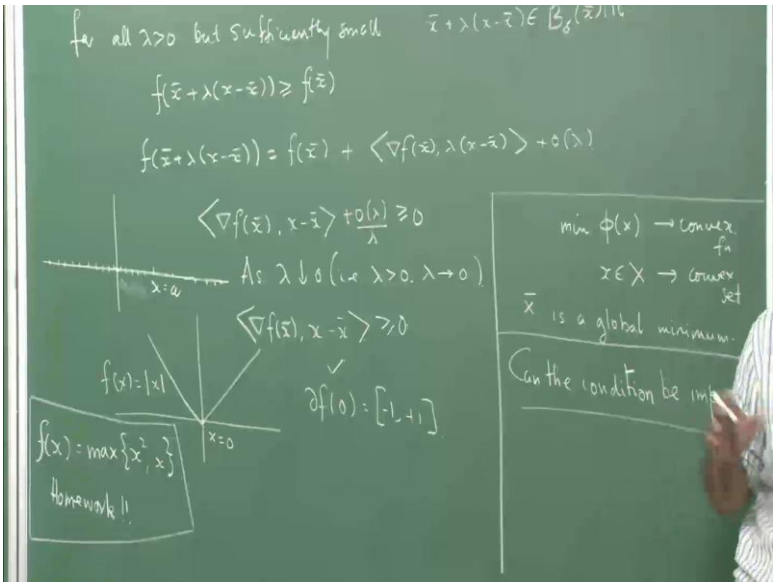


Now, let us just take a subgradient of the convex function phi. If xi is in set del of phi of x bar; and, let us see what happens if I take x bar and a point x forms x and construct a point, which still lies in x by the virtue that x is a convex set. And then, try to write down the definition of the subgradient. What I can do is of course, you will see here the lambda

would come extra lambda x minus x bar; which I can divide by lambda. Now, once you have this, what can you do? Shall you push the lambda to the limit? If I take limit; now, the question is, does the limit actually exists? Does such a limit exist? The answer surprisingly turns out to be yes, such a limit exists and it exists finitely. And, this limit is usually called the directional derivative of the function phi at the point x bar in the direction x minus x bar. So, this whole expression is now written as a directional derivative of phi at the point x bar in the direction x minus x bar.

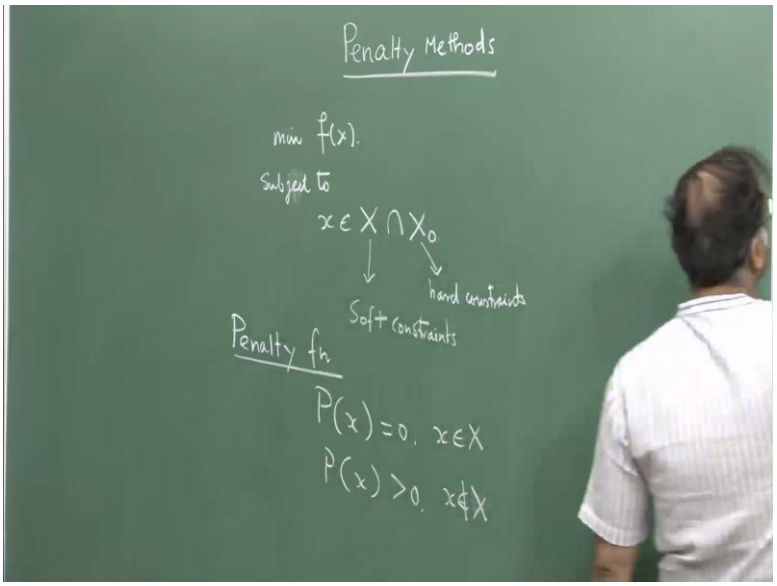
And, that now, because x bar is a global minimum, one thing is for sure is that, phi dash since x bar is a global minimum, but this is of course, true for all x. But what does it tell me about this part? It tells me nothing apparently. But there is a very deep result in convex analysis. I just do not want prove it, because I think more discussion is found in the convex optimization course. But a fact that we are not going to discuss too much about it. So, we are just stating this fact that… So, one can show, for each x, there exists xi x such that now, what does this show? This shows that, if x bar is a global minimum x bar is a global minimum. We can show that, for each x, there exists xi of x element of del phi of x bar… This xi is belonging to del phi of x bar such that because of this inequality, xi x, x minus x bar is greater than equal to 0. This looks to be a reasonable condition for optimality; that is, we are telling that, for each x, there exists one xi; for each x there exists one xi x belonging to this such that this will hold. If we change the x, this xi x is changing. My question to you is the following, which we will talk about tomorrow in the next class is that, can this condition be stringent? That is, can the condition if I write this as condition A; can the condition A be stringent?

(Refer Slide Time: 22:15)



If we can do this, it will be enough convex – non-differentiable convexity for us. Then, we will go on to the most standard things. See our aim is now to telling about some algorithms about unconstrained case. And, the condition be improved; means can I have only one xi, which will work for all x? That is what I am asking. Can there be only one xi, which will work for all x? That is our question, which we will try to answer tomorrow. Can I have only one xi, which will work for each and every x? But today, we will go and work on something called penalty function methods, which are very useful methods to solve constrained optimization problems. And that is exactly what we are going to do next.

(Refer Slide Time: 23:25)

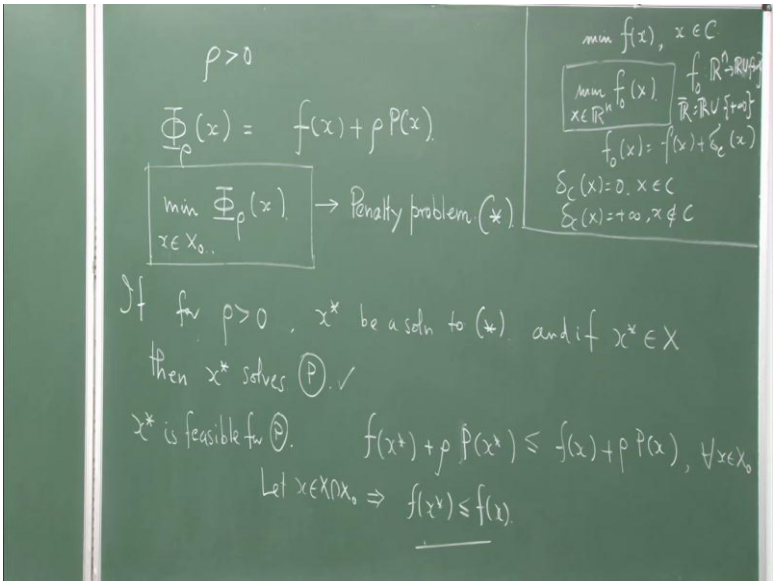


Let us now talk about the penalty method. So, what does it mean or what does it mean by a penalty? We will first do it in a very general form and then go to some specific cases. Let me ask you to minimize the function f x subject to x belonging to capital X intersection X naught. So, here I separate the 2's. I write the feasible set as the intersection of two sets. This could be for example, (( )) just inequality in inequality constant. And this could be a constant that x belongs capital X naught. What it means is the following that certain constraints, which we want to call soft constraints… And this is called hard constraints. The reason is this, the idea is that, we always want to convert a problem, which is a constrained problem into an unconstrained problem, because that is easier to solve. Penalty function what we do is we pick some constraints from the set of constraints and add it with the objective in some manner. And we will show what is that manner.

Now, every constraint cannot be added, because there are certain constraints like that x i's (( )) is between minus 1 and plus 1. These are variable bounds for example, in engineering problems. They have to be strictly adhered to. You cannot slightly even violate them. That would lead to safety concerns; for example, in design, issues would come that. But for these constraints issue, we will call a soft constraint; slight violation is not a very big thing to be bothered about. Once x is lying between in x naught, slight violation of these constraints here is not a cause of great concern. So, what we can do is to somehow plug in this to this and so make the problem much more simpler. That is, we

create a function called a penalty function. The penalty functions – penalty function… We construct something called a penalty function. Penalty function is something like this. We define a function P x, which will be equal to 0 if x belongs to x. Of course, in most of our operations, we should have x also in X naught. P x is strictly bigger than 0 if x is not element of X.

(Refer Slide Time: 27:11)



Now, take a number rho; take up some rho (( )) positive (( )) and construct the function. How do I think of symbolizing? Maybe I will choose the symbol, which is (( )) as chosen as… So, the essence of x belonging to x – soft constraints is encoded in this function; that is, if the soft constraints are satisfied, then P x is 0; function value remains, objective remains as f x. If it is not, then P x is positive; that is, you add up a penalty to the objective for violating the constraints. This is very very fundamental. And now, we can talk about minimizing the penalty of function over x element of X naught.

Now, if x naught is nothing but R n, you see this becomes an unconstrained optimization problem. But the question would remain – how does the solution of such an unconstrained or this simpler constrained problem relate to the solution of this original problem? And that is what we are going to investigate at this moment. So, this is what is called a penalty problem. What we do is that, we can keep on changing this rho and keep on successively solving this problem.

We shall now relate the solution of the penalty problem or succession of such problems to the solution of this problem. Before I go and do that relationship, let me tell you that, this has a theoretical underpinning, which is one of the frameworks of modern optimization. It is that, if you ask me to solve; if you ask me to look at this problem of minimizing a function x over set c, then what I can do, I can write this problem as an unconstrained problem; when I minimize a function f naught x over x element of R n; where, f naught x is written as f x plus del c x; where, del c x is usually called the indicator function a set; where, del c x is 0 if f x is in C and del c x is equal to plus infinity; x is not in C; means theoretically, I can attach infinite penalty when x is in the constraint is violated. If I attach infinite penalty in minimization when infinite values are not required; basically then, this is equivalent way of solving this problem – a minimizing over C, because only when x is element of C, this function value is finite. If x is not in C, this function value is not finite.

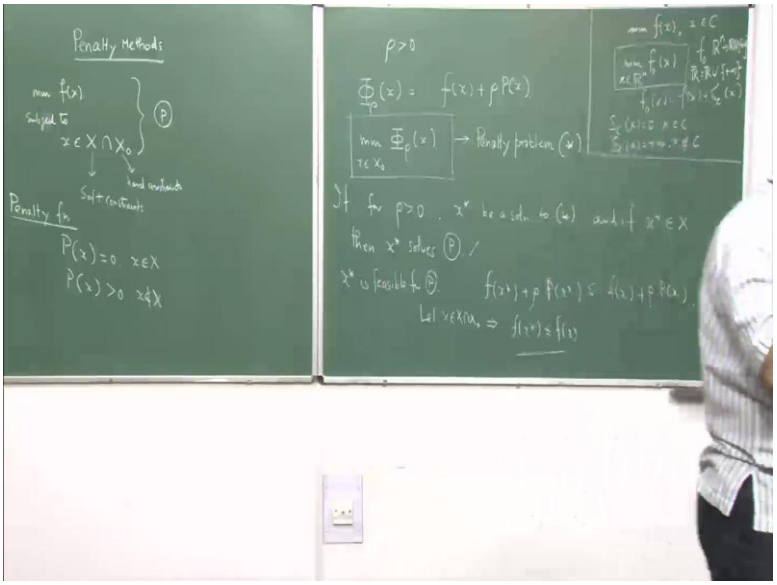
So, instead of solving this problem, I can now look at this problem; or the function f naught is no longer a function from R n to R, but is a function from R n to R union or I can just write R union plus infinity. That is also you can write as R bar as R union plus infinity; that is, I am considering extending the real line by adding the point plus infinity. So, we are now basically talking about extended real valued functions. So, real valued function over a set C can be viewed; minimization over function f over set C can be viewed as a minimization of an extended valued function about the whole R n. So, this idea, this framework actually come from this penalty function scheme.

Now, what we expect? We expect is the following; that, as I increase the rho; as I go on increasing the rho, there will be a rho large enough for which minimization of this problem would give me a solution of that problem; that is, I am expecting for rho large enough; there would be an x, which the solution would not only be inside this, which has to be inside this, but will also be inside x. So, let us see what could be the first such result. If for rho greater than 0, x star be a solution to this problem – to the penalty problem; maybe I can call this problem as a star problem; penalty problem to star.

And, if x star is also an element of x, then x star solves this problem – the original problem say P; x star solves P. This is not very difficult to prove, because whenever x star is a solution of the penalty problem, x star is in x naught. And, if you additionally say that, x star belongs to x; then, x star is a feasible point of P. So, we already have that x star. Now, you know that, x star is a solution to this. Basically, you know that, for that

rho, f of x star rho P of x star is less than equal to f of x plus rho P of x for any x in x naught. That is what is given to me. Now, take any x. Let x be a feasible element of this problem; that is, x is in this. Then, for such an x, P x is 0. And, because of this fact, P x star is 0. So, for this, it will immediately imply that f x star is less than equal to f x – proving what I had wanted that x star is a solution of the original problem.

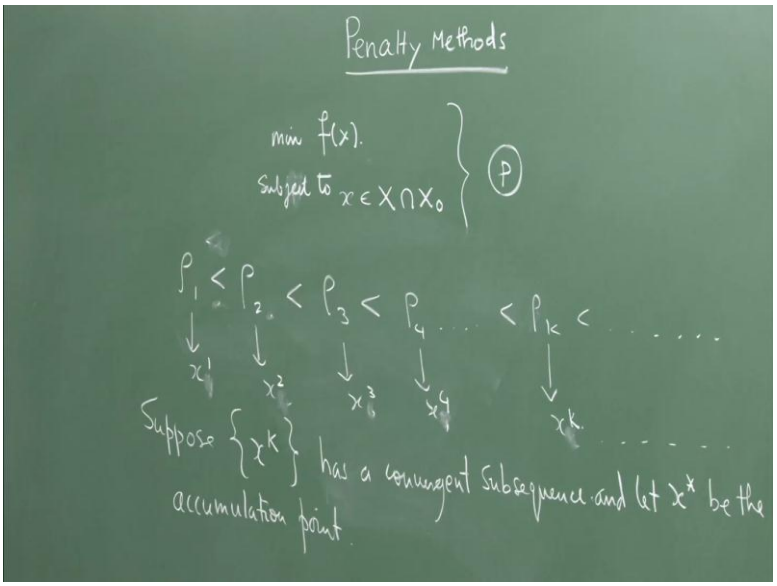
(Refer Slide Time: 36:23)



Now, we will try to show you that, if I actually keep on increasing rho; as I take a sequence of rho – rho 1, rho 2, rho 3, rho 4; then, as rho k goes to infinity. So, I start generating the solutions of this penalty problem for each k. So, for rho 1, it is x 1; for rho 2, it is x 2; for rho 3, it is x 3 and so on. So, I will, suppose this sequence has a boundary or at least has a convergence of sequence. Then, the limit – the accumulation point of that subsequence – the limit point of that subsequence is a solution to the original problem. And that makes so which means that when I actually try to implement it, I solve this problem for very large values of rho.

And then essentially, for very large values of rho, for certain large value of rho I stop; and, say that this could be possibly the desired solution. But that solution may not be feasible. That will be in x naught, but may not be completely in x. So, we will expect a slight violation in x. But that is actually considered in many many practical cases, we will take that as a solution, because people do not mind a very slight violation in X. So, our result is of the following.

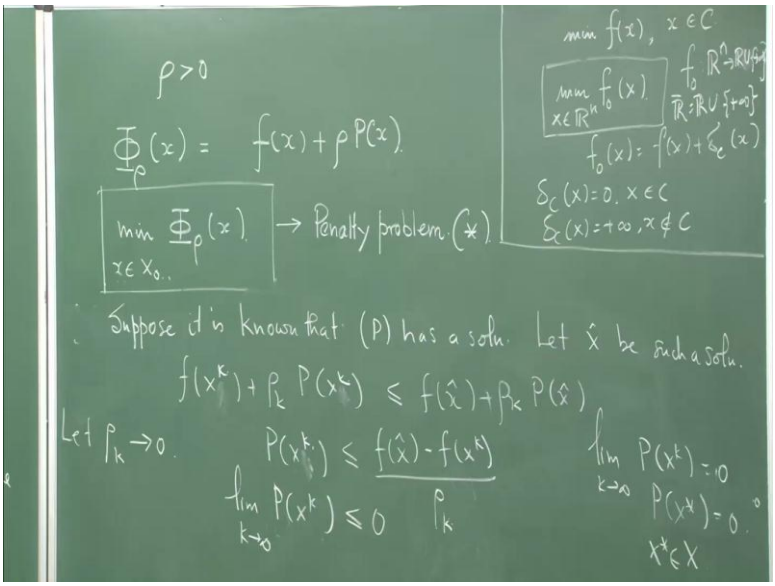
(Refer Slide Time: 37:11)



Now, you see… Let us I will just keep this as… Now, what we are trying to say is that, we are solving this problem for rho 1, rho 2, rho k. So, I am just increasing the… This is increasing; rho 1 bigger than rho 2. So, this is of this having this. (( )) increasing the values and I am solving them. And, for the each of these, I am solving the penalty problem. Suppose the penalty function problem has this solution. Of course, they do not have such ordering because these are all vectors will go on. Now, suppose the sequence x k maybe I should write it as 1, 2 (( )) in a sequence. So when I have a vector, I will write the sequence index on the top; and when I have a scalar, I will write it at the bottom, so x k, suppose x k has a convergent subsequence.

And let x star be the accumulation point. Now, suppose somebody now has given you the information that, x star that the original problem has a solution. Here you are trying to construct a solution. But somebody has given you a prior information that, the original problem has a solution. For example, the original problem could be a linear programming problem or quadratic programming problem and there are certain ways to know that they have a solution. Suppose the original problem has a solution. Suppose it is given that P has a solution.

(Refer Slide Time: 40:00)



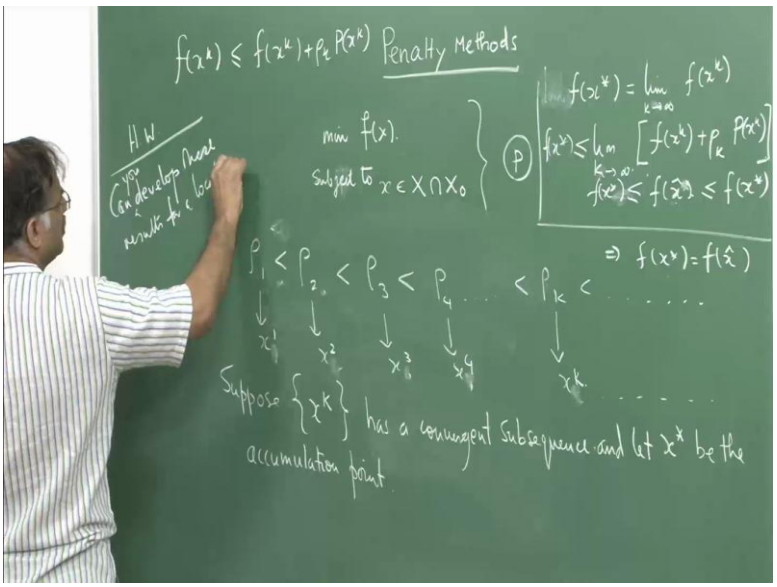
It is only known that P has a solution. Nobody knows what is a solution. So, I am now Penalty method is trying to construct that solution at least approximately. Suppose it is known that, original problem P has a solution. If it is known that, that has a solution; let that solution be x hat. Then, what you can write? Because x k is a solution of P k, you can write f of x k plus rho k P of x k; x k you do not know whether it is feasible to the original problem. So, this is not 0, is less than equal to f x hat plus rho k P x hat.

Now, P x hat is 0, because x hat is a solution to the original problem. So, x hat is in x. So basically, what you now get is the following, you get P x k is less than equal to f x hat minus f x k by P k. It should be x k. Now, when we define penalty function, we will always consider this penalty function P to be a continuous function. That would be a binding on us that, and that would be helpful in algorithms that we take this P to be a continuous function. So, what we do here is the following.

Now, observe that there is a convergence subsequence x k, which is going to x star; and, this rho k is going to infinity. Let rho k goes to infinity. Now, for that particular subsequence, this is also true, because this is true for all k in the sequence. Hence, for that particular subsequence, you take the limit; just taking the limit for the particular subsequence; we are not relabeling this. Now, f x k if f is continuous of course, f x k is going see in our discussions, we largely talk about differential function. So, continuity of function is, always taken. So, please do not say – if it is a discontinuous function, then

things would not be the same as what we are trying to say. So, here f of x k… You see here f of x k would be like this; that f of x k would go to f of x star. So, it will be, the top part goes to finite value; while this part so the top part is bounded while this part is – 1 by rho k is going to 0. So, basically which means this whole this is going to 0. So, this is less than 0. But P of x k is greater than equal to 0. So, limit k tends to infinity P of x k greater than equal to 0. So, limit P of x k would also be greater than equal to 0. But here, we have less than equal to 0. So, finally, we will get this. But by the continuity of P, I can push this limit inside to get P of x star is equal to 0. So, by definition, P of x star is equal to 0 means x star must be in X. That is the definition. Definition is always if and only if; which means x star is in X.

(Refer Slide Time: 44:37)



Now, it is only left to show that x star is a solution of the original problem. So, let us see how can… what will this problem be. So, f of x star can be viewed as limit of k tends to infinity f of x k; of course, k is only running over the subsequence. We are not bothering about relabeling the sequence and the subsequence, you can without loss of generality you can take x k to be going to x star, no problem. But those are just extra writings, nothing else. Now, you see again always write f of x k… So I can always write f of x k is always less than f of x k plus rho k P x k, because P x k is a non-negative quantity and rho k is a positive quantity. So, if I add a positive quantity to a number, it will increase; which means this f of x star is less than equal to limit k tends to infinity f of x k plus rho k P of x k.

Now, this little thing here is obviously less than f of x hat. But f of x hat is independent of k; which means this is this thing; this limit is nothing but less than equal to f of x hat; which means f of x star is less than equal to f of x hat. But f of x star… Is this also a solution to the original problem? Maybe. When a solution means, we are expecting a global minimum of course. So, let f of x hat be a solution to the original problem. f of x hat is the solution to the original problem – global minimum. So, f of x star must be equal to f of x hat. But f of x hat is a solution to the original problem; which means that x star is feasible to this thing in the original problem. So, this has to be true; which would imply that, f of x star is equal to f of x hat. And hence, x star is also a solution.

Now, can you tell me if this idea can be extended to local minimum? Can you handle this? So, home work is – can you develop these results for local minimum or these are just for the global minimum, which we should not be; we should be able to find a local minimum? Can you develop these results for a local minimum? Thank you very much. So, we will continue our discussion of penalty functions there tomorrow in the next class, where we will take x as certain functions in the sense that x certain inequalities; and x naught is a fixed set. And then, we will try to discuss about it. And then, we will discuss a very important concept called exact penalty functions.

Thank you very much.