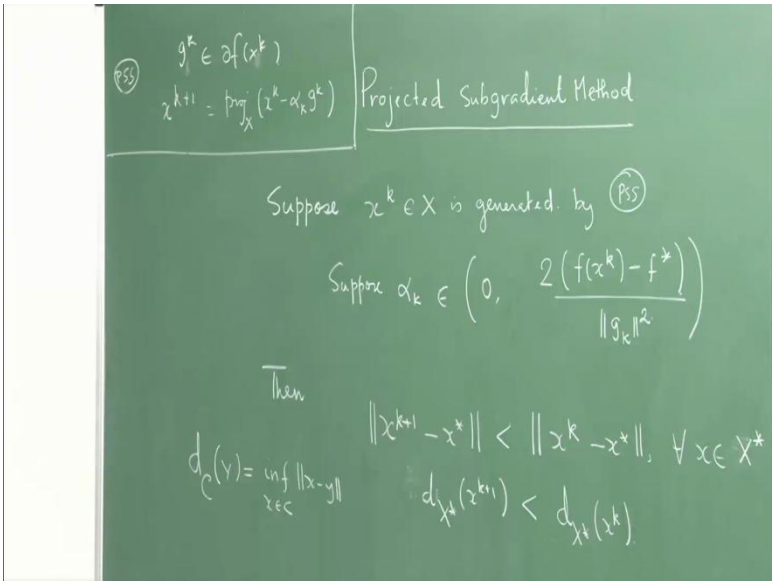
Foundation of Optimization
Prof. Dr. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 35

When you begin a course like this, you have ambitious plans to tell the viewers a lot of things; you keep on doing it, and planning more and more adding things that you actually like yourself, because most what a person can do at the most is to, tell in many ways a story of his love, so you would really talk about what he actually likes in the subject. But suddenly one realize is that we sort of video courses are marked by certain amount of fixed lecturers, and then suddenly one realize is only at the end of the course, today is the last part 1 lecturer. So, it is a good idea that we would do some advance topics as we were discussing yesterday.

(Refer Slide Time: 01:28)



And we will talk about the projected sub gradient method; I am not going to write down all the details what I had written down yesterday, because it was already done. Now, tomorrow possibly I have to wind up this course with giving some very major ideas, but lot of things in optimization which I am not been able to speak here. For example, what happens if the data is a corrupted by noise that is a random variable which has coming to the data chance based input.

Around which the designer who ever in the special in the engineering stream has to know control on. What we can probably have some idea, from the experience you can have an idea about what sort of noise has coming what sort of probability chosen such random variable agrees to. So, the presence of noise of stochastisity is a very, very fundamental thing in optimization, such a thing cannot be a part of an elementary course, because one needs to be trained in much more things, in order to handle stochastisity in optimization.

But, for engineering problems stochastisity is a very, very big requirement, it is a fact that comes in naturally, we are not, we are not discussing this stochastisity here. So, the question is that these issues remember, that whenever I have a algorithm and I always trust that you need to show that, your algorithm works by showing the convergence analysis, does not mean that that algorithm is superbly efficient that that, once you start with the starting point it will rush towards the solution.

For example, the Stephen's (( )) method that we discussed long back in the course, is a very, very good one and has a very good convergence for the quadratic, he is very good convergence for for example, a very good convergence for the convex cases. But, the problem is the following that it may not rush towards the solution it can be very slow, so and the so these sort of critics in that are convergence analysis, does not lead an algorithm to be very good enough for a given problem; and these criticisms actually are valid.

But, these convergence analysis also beings in to light, the quality of the algorithm itself, the nature of the behavior of the optima themselves, the nature of the behavior of the iterative iteration sequence; it tells you a lot about the nature of a particular algorithm, which has huge qualitative value. So, for example, I would like to show to you at least through this, so this is the original paper of written by Boris Polyak, called the minimization of minimization of un smooth functional.

So, it was published in US resort, and then later on it was you know translated to English by D Brown, but this paper and even now read the name, minimization of they did not write non smooth functional, but they have written un smooth functional; so that was a English translation. And it talks about it gives a unified method for both the smooth case and un smooth case where the Polyak step length that we had yesterday discussed, had actually been considered.

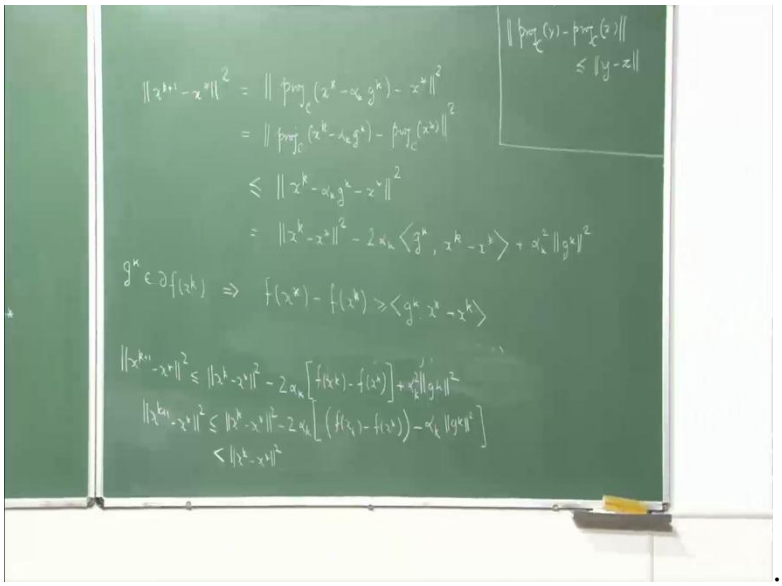
Now, let us suppose we have a sequence generated by the projected sub gradient method, the projected sub gradient method is, if you want to be more were you write, has to be written like this sorry, and x k plus 1 is projection on x. So, this is the projection projected sub gradient iteration scheme, and that is what we have going to use. Now, but suppose we had given that x k, which are in x is generated by the projected sub gradient scheme.

Generated, if I call this as PSS scheme, Projected Sub Gradient scheme generated by PSS, so suppose x k is generated by PSS. And your step length alpha k, suppose now your step length is chosen like this, so it is a number lying in the open interval, so at every k your step length will change depending on your choice of g k. Of course, here we are assuming that we somehow have come to know about the infimum value, but we do not know the x star at which the infimum value would be achieved, and the whole process is to find that x star. The whole process is to find that x star using this scheme of iteration.

If this happens, then the most fundamental fact about the projected sub gradient method, that it does not tell you that the functional value is reducing, only tells you that the k plus 1th iterate is much nearer to the solution x star. So, whatever solution x star you take, this fact is always true sorry, sorry capital X star; so, distance of x k plus 1 from the solution set is always little less than the distance of x k from the solution set that is the mining of this.

Because, now you can take infimum over x star on both sides to come to the conclusion, that distance a d of x k plus 1 distance function, this strictly less than the distance, so construct the infimum. The distance of the set see if I want to, to remind you, distance of a set over a point y outside a set c is given as the infimum of norm x minus y, where infimum is taken over all x in the set c that is in that, so that is what happens.

(Refer Slide Time: 10:11)



So, I will prove this fact to you, so how do I write this one, so here is little bit of art of doing algorithmic analysis, so this estimate is very important; this actually brings out the character of the projected sub gradient method. So, this whole square is x k plus 1 is nothing but, the projection, a very important fact about projection is, is it Lipchitz behavior. But, we have not spoken anything about Lipchitz functions in detail in this course.

So, we would say that if, so for any projection mapping, so you take the projection on the set C, then the norm of projection of C, so you take any any y and z in r n, this is always less than equal to y and z can be in C also. So, I can write this thing as, and where x star is in c, the projection of x star on the set c is going to be x star itself. And that is what you have sorry (( )) must be inequality, but the important fact is not just that there is an inequality of the distance, that given take any x star this is always maintained, but when you take the infimum, you have the inequality here, less than equal to (( )).

Now, once you have this fact, then you can apply that result to show that this is nothing but, now I have to take the whole square, because I have squared both sides. Once I have done this, then I can open the norm by writing this as I will block this two together, and have, so this is how you open the norms. Now, once you have opened the norms like this, the next step is the following, now observe that g k is in del f x k, so g k is element of del

f x k that would imply f of x k minus f of x, is x star sorry, make a mistake f of x star minus x k is g k, x star minus x k.

So, what do you have is minus, so if you take the negative, so what you would have if you just take this to this side you will have g k x k minus x star, take it to this side you will have f x star minus f x k sorry, so you will have f of x k minus f of x star. Now, we are going to use the sub gradient inequality here, so I will just write down inequality again, so x k plus 1 minus x star whole square is now less than x k minus x star whole square, you can take club in the minus here. Once you club in the minus here, it will become plus here, and x star minus x k and that will immediately give you that it is less than f x k minus, it will give you f x star minus f x k.

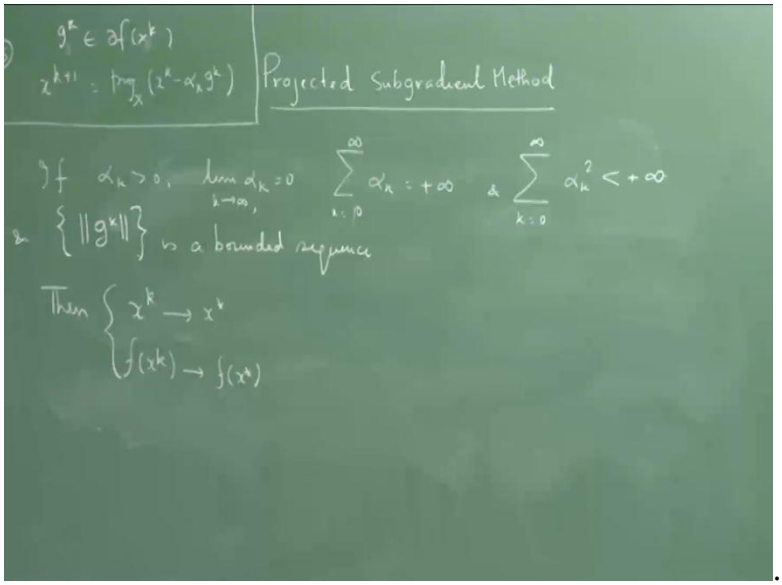
So, you (( )) just bother about the last this one, this is the same thing, so this means I will have minus 2 alpha k f x k minus f x star, which I can write as f star, (( )) plus alpha square norm alpha this. But, this one f x k minus, f x star is bigger, if x star is smaller, so this is positive, so and this is how this when you and this is positive, so this whole thing is negative, so this is and now alpha k, so this goes off; so this is less than this plus alpha k square. There is no doubt that alpha k is less than this quantity, so this is there is no doubt that alpha k is less than the square of this is positive.

And now here what we have to do, is now we have to use the fact that alpha k is less than this quantity, so you have x k plus 1 minus x star whole square is less than equal to, now you have to take in to account this one. So, you write x k minus x star whole square less than, now you have to come with come up with the alpha, so maybe you can take negative twice alpha k which is out; now you have f of x k minus f x star minus alpha k norm g k square. I would tell you to finish this calculations by putting in the values of alpha k, to show that this is norm less than norm x k minus x star square.

Now, once this is done we would like to go, and so this is also a motivation for the use of the Polyak step length, so if an alpha k is chosen of this form, and you are the norm of sure alpha k shown between these two numbers, then this is actually happening. In fact, this is strictly in fact this becomes strictly negative, so because alpha k is not less than or equal to alpha k is strictly less than this quantity, strictly less this, so this becomes strictly less. So, if an alpha k is taken to be equal to that this become less than or equal

to, so it make sense to use that particular choice of alpha k, and that is from where this Polyak step length is important.

(Refer Slide Time: 19:50)



Now, we are going to try to prove or give a sketch of the proof of what we intended to show yesterday or we mentioned yesterday that, if the assumption one means if the step length, if alpha k is strictly greater than 0 limit of alpha k is equal to 0 as k tends to infinity. Summation alpha k k is equal to 1 to infinity is plus infinity, and summation all of this has to be satisfied, the step length has to be chosen k equal to 0, this k 1 by k plus 1.

So, if this happens, if g k is bounded now, because I have not given you the mathematics of the fact that the sub differentials are, if you take finite valued function from one and two are you, the local what they something called, local boundedness of the sub differential; and that would lead to the fact that norm g k is bounded right. So, g k is element of g x k, and x k is, no we cannot say that sorry I made a mistake in the last class last class also, let me admit that mistake.

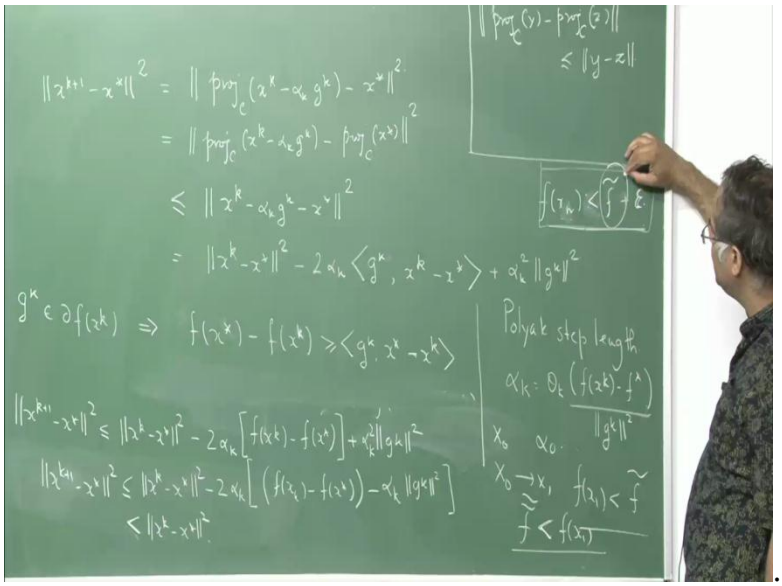
We have not yet proved that the x k is going to x star, if x k goes towards x star, then we can say that g k is bounded, so that is what we really have to prove that x k is going towards x star. So, g k has to, the norm of g k this is assumption, it is a bounded sequence then we are proving that, actually you have unless you have x k going x star you cannot use that, whatever I am calling as local boundedness property. So, please

now, what you about the local boundedness property at all, but remember when Polyak wrote those things these are essentially done for functions which are extended valued of course, x k are all in the domain.

Because, they are in a feasible set, when over which has find it, so must the domains intersection with the feasible set must be known empty, so you need not get too much poked away with it, but we have to take this thing, but if you take the Polyak sequence you do not have to bother about this boundedness of this norm g k. If this is assumed then x k goes to x star and obviously by continuity of a function f, if it is of course, if you do not have continuity have to show that, but if you have continuity of the function f then it is automatic this goes to this, so these two things happen.

But, our case this will happen will imply that this will imply this, so you forget about this word local boundedness that I have used in that lecture; so this is assumption I am making. So, you might be asking me why I am proving this result, and not the result with Polyak step length, see Polyak step length has this issue.

(Refer Slide Time: 23:15)



So, what is for the Polyak step length, there is a fundamental assumption when the Polyak step length is something like this which, so f star the infimum value sorry, it becomes (( )) step it is changing, because of this. And we know there are particular choices of theta has to be done that we have written yesterday, now the problem is the

fact that I know f star, the in real computations you cannot have f star, you do not know f star of course, because you are supposed to find f star.

Because, you do not know f star, you really it cannot be inverse problem that I give you the f, I give you the f star then tell you to find x star, then you use the Polyak step length straight away. But, if you do not have such a thing and what you can do that you can take an arbitrary estimate of f star, and then try to say from x 0 go to x 1. If f x 1 is strictly less than f star then of course, your function values that f star is not a good estimate. So, what you have to come in to a situation, so you what you do, so you reduce the value of f star, you reduce it below f x 1.

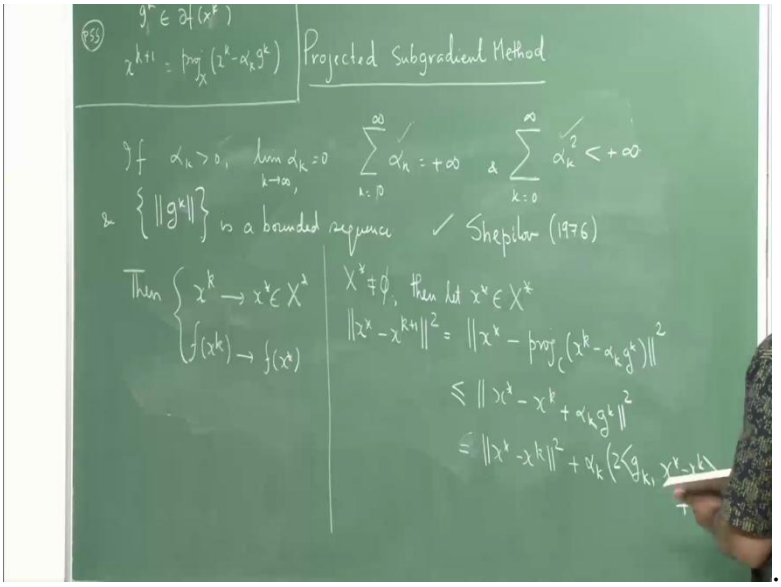
And then you again try, so you try for some times and see that you have got f x value less than equal to less than f star, so you can try with an x you take that x 1 and then try to the new estimate, you can try from going from x 1 to x 2 with the new estimate of f star, and the step length. So, from here I start with the point x 0, and then I calculate alpha 0 and then I go from x 0 to x 1 by this standard projection. So, now I check whether if f of x 1 is strictly less than my estimate of f star, then possibly this not a good estimate of f star is, f star cannot be strictly less than f x 1.

So, then I have to choose something which is say f which is strictly less than f x 1, and then come from x 0 to, and then take that x 1 as a starting point, and then I try to get to f x n or may get to f x 2. So, what I essentially would require at the end is that ok, I must have my f x n's of this form, it would be enough to have this situation in computation for some given epsilon greater than 0. This is this is the sorry yeah, if f is at infimum then if I do this and x n at least should be of this form, so for some n, x n you should immediate ask given fix epsilon you can have it in this form, so f star is, f at is the infimum.

I made a mistake f of x n is sorry, I made a mistake it is infimum, it should be less than or equal to rather strictly less than f hat plus epsilon. So, if this happens then we can take this f hat, this is the definition of epsilon, if definition of infimum, if f tilde in some sort of an estimate of the infimum suppose this is the infimum, then for some epsilon greater than 0 I can find an n, so that f of x n must be this. So, now if I when f hat and some f star I can I have an x n which is this, and that and if I lock my f x n value is decreasing, then this f hat would work.

So, this certainly a complicated thing, and so here Polyak step length is largely a tool which would give me theoretical results, but the problem here again is that these step lengths. And immediately you can find it a step length satisfying this, but the problem again would be this, who check the boundedness of the norm of g k, how do you know that they are actually lying between something, so these are questions that coming. So, everything has an advantage everything has a disadvantage, because you know once you come from to sub gradient issues, but this can also be used in the gradient thing.

(Refer Slide Time: 28:43)



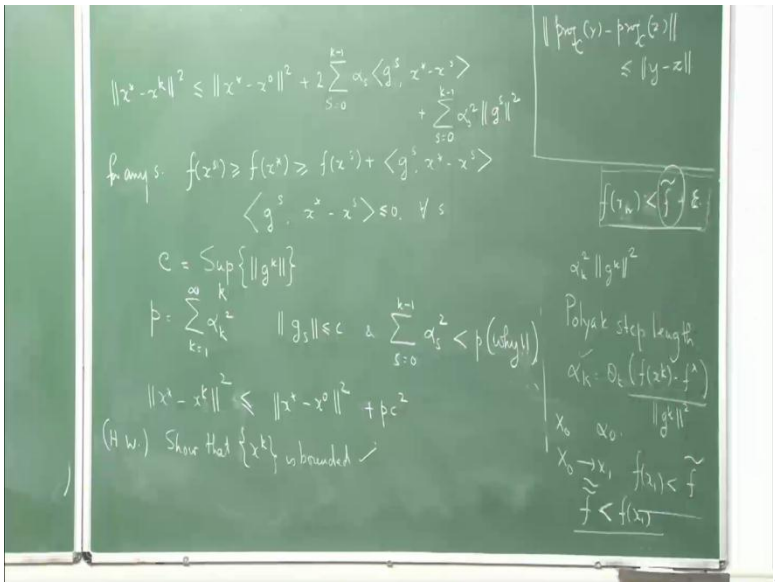
So, but only advantage in the Polyak case is that you do not have to assume the norm g k is bounded that that is the plus point, any way let us this was the proof given in 1965. And let me just tell you who was the author, this proof was by Shepilov 65, the proof that we are going to give is by the Russian mathematician Shepilov, Shepilov gave this proof in 1976. So, Polyak result is in 1969, so Shepilov improved it in the sense that is easy to find it, but this is again now draw back. But, luckily if this is happening, so this also happens, so practical purpose is this is pretty, the Shepilov result is pretty pretty interesting in that sense.

So, we try to do some proof of this, some outline of the proof I would not say the full proof, but I will just go and do some outline, give some idea us to how the proof can be achieved. But, that would constitute what would one call as a art of convergence analysis, so again you start with this fact that you assume that x star is a solution, so x

star, so this goes to, so we have assumed that x star is non empty. So, we have assumed that x star is not equal to phi, then let x star is element of x star; now compute this distance, from x star compute the distance of x k plus 1 square of the distance rather that is again the same as what we have done earlier.

So, this again by that same property written there, at that end of the board this property, would lead to the fact that minus minus plus this minus and this minus will be plus, and so this is what you have. And then if you write down open the square, this is exactly equal to, so alpha k in to, I am writing this I am not doing too much of stepping jumping a step. So, basically I have 2 alpha k g k in a product x star minus x k and of course, alpha k square plus norm g k square, so we are not combining those two terms and writing it immediately to get.

(Refer Slide Time: 32:33)



If I use this thing repeatedly what I have is the following, I have that x star minus x k whole square, so I I I repeat I start from x k go to x k minus 1, so here it will be x k then x k minus 1, so and so forth. So, and then keep on repeating to get the starting point, so obviously will be up to k minus 1, this things repeat at applications of the cone. So, s is a (( )) my variable instead of not g s here, what you have, now f of x s for any s for any s guaranteed.

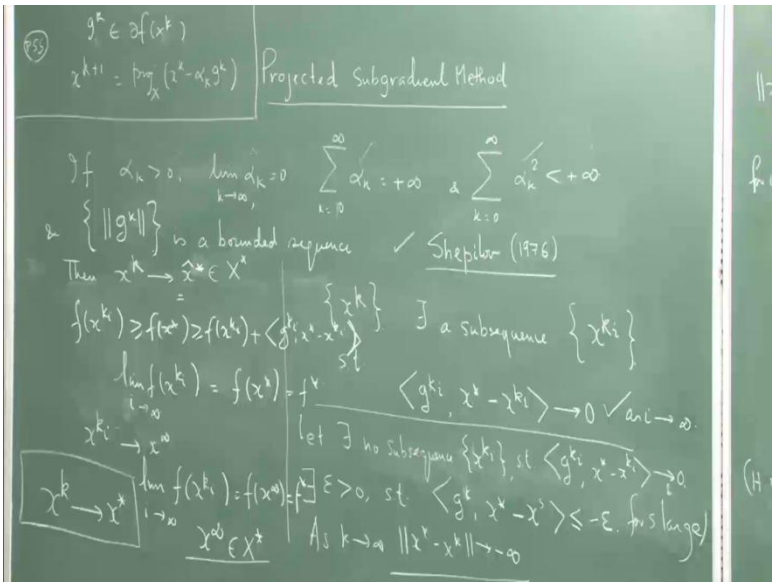
Now, define c, because you know that this is bounded this norm g k is bounded then define supremum of norm g k over k I define this, now say p I shift that the infinite sum,

because this is strictly less than infinity, this is p. And this is obviously less than c, and I would have sorry, why this happens I would this is a very basic question in analysis, so I would just ask you to figure out this, I would not repeat why such a things happens, this is the very basic question in very basic basic calculus basic analysis. So, because we had the advance level at the end of the course, certain advance level, so this is you can figure out why.

Now, then going back to what I have here, I can write that x star minus x k whole square is less than or equal to right, this part is less than equal to 0, so this negative this is goes I have this part, and this part. So, it will give me sorry, this is x k, this one is this plus, now this can be written as strictly less than, because this is strictly less than this this, so there will be a norm is less than equal to c. So, product is strictly less than unless of course, there is a 0 fact (( )), so this will simply show you that from here home work show that x k is bounded.

So, this is a very very important step in convergence analysis to show that the sequence that you are generating the is bounded. Because, then you are guaranteed there is a convergence sub sequence, so the idea always is that you show that there is a convergence sub sequence which will have a limit point, so that convergence sub sequence has a limit point. And then show actually the whole sequence foes to that limit, so and the whole sequence is actually going to that limit, and hence and show that the limit is having some properties like this that is element of the solution set. So, this can be showed as a home work, this is too easy for us to unnecessary detail on, because you have to understand, because x not is known, x star minus x not is actually a fixed quantity, and this is a fixed quantity; the crucial part of the proof lies here, which we will now analyze.

(Refer Slide Time: 39:14)



Crucial part of the proof says the following, but Shepilov shows is the following that in this sequence x k, there exists a sequence, there exists a sub sequence x k i such that that now what you are going to prove that there exists, in x k there is a sub sequence x k I, there exists some sub sequence or sub sequence such that g k i this goes to 0, the proof this though it is a very, very crucial step in the proof of Shepilov theorem, Shepilov analysis; this is a very, very interesting thing in the sense that it is proof is not very difficult, because we will use one of the most important tool of the mathematician tool works proof by contradiction, we will say that let us now assume that, there is no such sub sequence in x k for which this will happen. Let there exists no sub sequence such that, sub sequence x k I, but you know that this result is true, so though is less than equal to 0, the sub sequence value cannot tend to 0, for whatever be the sub sequence.

So, given any epsilon greater than 0, so the not given any epsilon there must exist some epsilon, so there must be a negative number which, which will bound this value, so this value cannot go above that negative number and move towards you. So, which means there exists epsilon greater than 0 such that, g k x star minus x s must be strictly less than minus epsilon, this is this is what you have. At least for sufficiently large values of x this must be true, that is for some finite number of values it may go beyond that, but after that for s for s large for s large sufficiently large it means m this has to be the case otherwise, it can go to 0.

But we are telling that there is no such sub sequence for which it goes, so for s sufficiently large or this s that you are generating or the, so whatever sub sequence you take. So, for sufficiently large index of x s in the in the sequence, because we are choosing x s from the same sequence x k, this has to be true, now if this has to be true what is the consequence. So, remember this is a very, very crucial thing, so we will just write here that will just write here that this statement of a of the result final result sorry very bad mistake.

Now, if this is true I just want to remind you the sequence x star that I that x k I generate need not go to that x star that I have chosen here, so I will put here that this as some x x star x hat star, x goes to x hat star, so it goes to some x star, x hat star. So, I started with some x star in x star, because x star is non empty, but that does not mean that the sequence I generate goes to this one. Now, what would be the consequence if this is happening, so from here the consequence is this is less than minus epsilon, so why I say that this is a very crucial step, because now you will see the use of this results, the use of this we have used this results, but we have not used this fact, now because you have this, we have essentially used this and this, but we have not used this result, this is not used.

So, this has been used, this has been used of course, alpha k is any where greater than 0, this has not been used, we will not show the use of this we will just keep it, we will not bother about at this moment, this has not been used. So, if I use this, if I say that this is less than equal to some minus epsilon, then this as k becomes large these goes to infinity, so this whole thing goes towards minus infinity as k tends to k tends to infinity. So, as k tends to infinity this would imply that x star minus x k goes to minus infinity, which is obviously not true.

Because I have said that we have already showed that this is this is a bounded thing, so I cannot so this will contradict the fact that this is going, so what I have said is this fact that there exists no sequence is not true, and there exists a sequence where this takes place. Now, once this is done, when you know this fact, then if you know that there is a sub sequence, now you come here you put x k i here, x k i here, and this x k i and x k i here, and you take the limit, so from here what I will now do, I will take this inequality and in this inequality.

Let me just have a use this fact here, so I am taking now this inequality, so I am taking this inequality and what I am proving, I am proving that I am proving this following fact that f of x k i is greater than f of x star is greater than f of x k i plus g s x star minus x k i. Now, as x k i sorry, g k i g k i as k i goes to 0, as k i as i goes to infinity. So, as i goes to infinity which means what, that i goes to infinity this is going to 0, which means the limit of this by using the sandwich theorem, limit of f x k i.

Now, which is which I can call as f star, now x k i is a bounded sequence, the bounded sequence f x k I, there is a convergence sub sequence, we assume without loss of generality it is at x k i goes to some x infinity. So, then I can immediately write by the continuity continuity of the functions f that, limit of x k i, i tending to infinity is f of x infinity is equal to f star. And x infinity, is obviously an element of x star, because this value is equal to f star the solution, the optimal value; so x infinity is now in x star.

And my job the problem of whole proof would end, once I can show that x k goes to x star sorry x k, the whole sequence x k will go to x infinity. So, this is x infinity is my x x hat star, so basically now I have the proof will end, once I prove that x k goes to x star, and this part of the proof I will not prove, because we are running out of time. But, this is a whole glimpse, what I have done here is to tell you that o k this is the way things are done, so there is a crucial point where you start applying this idea. And this idea, this idea would be used, when you are trying to prove this, you see that we are using everything, and that that is very, very important.

So, I have just told you the crucial thing, so we have shown that, we have we have generated a solution, we have generated a sub sequence of x k from which I have some behavior from, which goes to some solution. And actually we will show that, no not x star sorry I made a mistake, the whole thing actually x k goes to x infinity, please take care of this. So, that is what we have to prove, at the end which have not proved and with this we end our last today's lecturer, which is the last one, basically for the course and tomorrow we will sum up what we have studied.