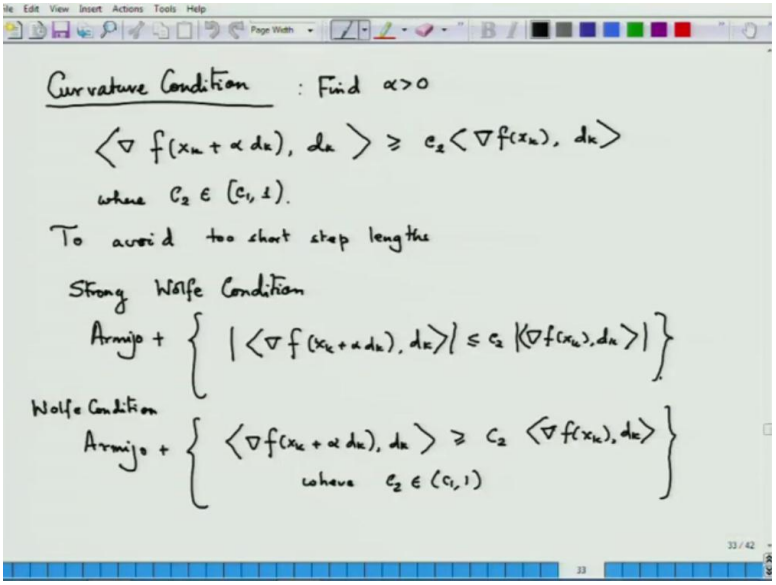
Foundation of Optimization
Prof. Dr. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

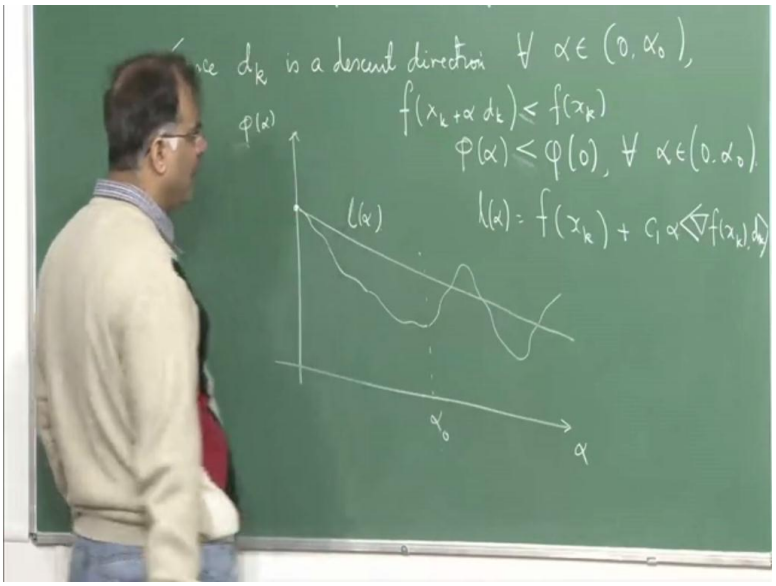
Lecture - 7

(Refer Slide Time: 00:22)



Yesterday we were speaking about the Armijo condition plus this additional curvature condition here which is called the, which combined with this Armijo is called Wolfe condition. Now, the question is that it is very important that we have a much better geometric understanding on this Wolfe condition. And how does it avoid this short step length. And once it is done, we will show that we will show that in fact, the Wolfe condition can the there could be an alpha. We can show the existence of an alpha, which will satisfies the Wolfe condition. Now, it is important to understand this function phi alpha once again.

(Refer Slide Time: 01:05)

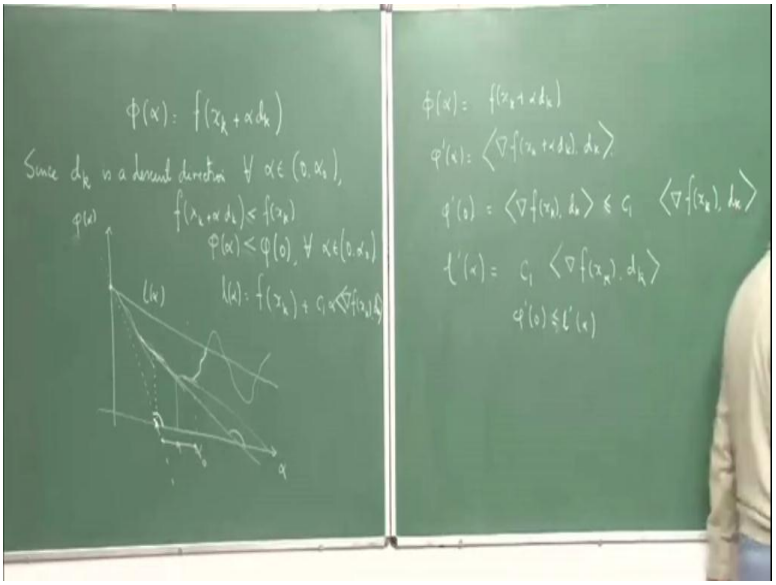


So, let us look at this function phi alpha which we introduced in the last lecture for a given k of course, this phi alpha you can say phi k alpha also does not matter. But let us keep phi alpha for the time being we know that we are analyzing things at the kth iteration and this is the descent direction. Now, what happens if I put alpha equal to 0 if I put alpha equal to 0, it will become f x k. Now, you know that f of x k is not the minimizer now which means in the neighborhood around the neighborhood of that point x k, there is a point where the function value is better means it it is lesser in the sense along the deduction d k the function value is actually decreasing

So, for certain alpha between 0 to alpha not which since d k is a descent direction, direction for all alpha element of some alpha naught I know that f of x k plus alpha d k is less strictly less than f of x k. So, this would allow us to draw the function better in the sense that you see that. Here so if this is my f of x k so this is alpha and this is my phi alpha. So, up to certain threshold alpha not here my function value is decreasing is less than this f of x k. Then of course, after that it can increase and then it can decrease and whatever so this is my phi alpha, but the threshold alpha naught till this threshold alpha naught it decreases. Now, my line l alpha so here what does it mean shows that phi of alpha is less than equal to phi of 0 for all, all other strictly less than phi of 0 for all alpha belonging to between 0 and alpha not could be equal to alpha naught, but.

So, phi alpha is a continuous function, because of this f is continuous and because of function f is differentiable. Now, this line l alpha goes like this and cuts up. So, what happens is you might question whether l alpha is below. This l alpha could be like this also it depends on the type of slope you take if you take a moderate slope here the line l alpha which you know is nothing but f of x k plus C 1 times alpha into grad f x k into d k in a product d. So, if you take a proper slope or proper modulation of C 1 then you know a for large chunk the l alpha line lies above this graph and that is a crucial fact. So, you can say this is my acceptable alpha naught this is my acceptable alphas. Now, what the curvature condition does is the following.

(Refer Slide Time: 05:13)



If you look at phi alpha, curvature condition is a second condition. So, curvature condition plus Armijo gives you the Wolfe condition. So, if you look at phi alpha here. So, this phi alpha is this function again f of x plus alpha d k. Now, let us do one thing I take a derivative of this, because this is a function; this derivative if you compute out very carefully is by applying chain rule, so first the gradient of this into the derivative of this, which is nothing but d k. So, it is so it is slope of the function phi dash at any alpha any alpha say this alpha; this is an alpha.

So, this is nothing but phi dash alpha the slope of this function, slope of this line is phi dash alpha. Now, what happens is that if you look at phi dash 0 if you look at phi dash 0, then pi dash 0 is nothing but grad of f x k d k now basically, then because c 1 is less than

1. This is bigger than c 1 times, say alpha times that is some alpha I would rather say this c 1 into. So, the slope of this now the slope of l dash what is the slope of, what is l dash alpha? What is what is l dash alpha? l dash alpha is nothing but C 1 into grad f x k d k it does not matter what is the alpha?

So l dash 0 is also this l dash 0 is also nothing but but the same things. So, what I am expecting is that the slope of this line must be bigger so so phi dash 0 is bigger than l dash 0 l dash alpha that is that is this slope which is which is natural, because C 1 is so C 1 is between 0 and 1. So, this whole thing is nothing but a fraction of this. So, this one is bigger than this one.

Now, you can say these all are negative bigger than this one in the sense I made a mistake fraction of this one so because I made a mistake, this is negative this is negative. Now, you take a fraction, so fraction of the whole thing so its minus 1, you take minus half so minus half would become bigger than minus 1. So, this slope has to be smaller so this slope has to be smaller means what? I made a mistake here it had to be less than equal to those who have recorded greater than equal to please change it, this is not positive; this is negative; this d k is a descent direction. So, this slope this slope has to be smaller means what that the angle. So, this is this is less negative and this is more negative, more negative means what? So, this is an obtuse angle; this is the slope which is a negative slope and I should have some another slope I should have a line whose slope is like this. Let me consider a line whose slope is like this.

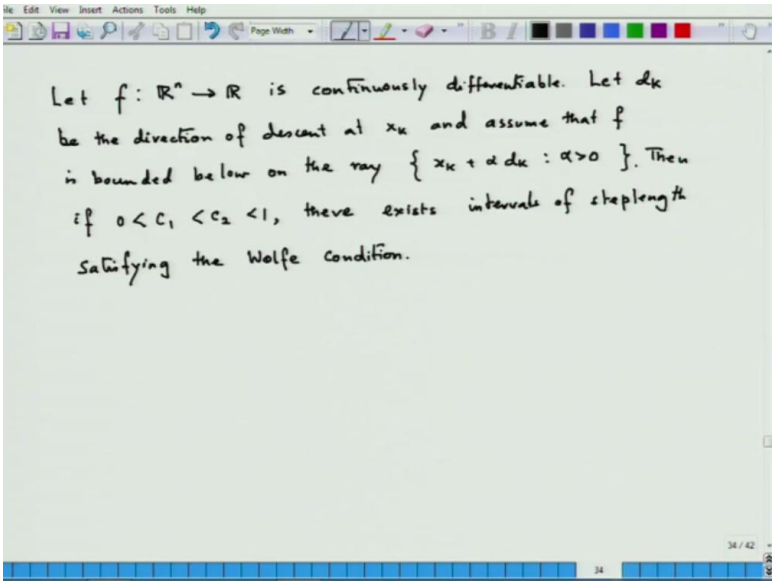
So, slope of the function at 0 so slope of the function at 0 is this which is less than C 1 times this. So, when the obtuse angle is more the tangent would be lesser and so here the tangent. So, here this slope this one that is for slope at phi dash 0 can be viewed as a slope of a line can we can basically draw a tangent to the curve of phi dash 0 and where ever this is cutting the line, cutting the x axis that point till that point I would not accept any alpha. So, beyond this for all alpha's this whatever slope you are taking, whatever slope you are taking all of the slopes would satisfy this condition see some slope should be for example, here up to alpha naught.

So, here you know the slopes are all coming down. So, this particular slope has to be for all this form starting point from here the slope this slope is always less than this slope which means that find an alpha. So, corresponding to this alpha on the curve,

corresponding to this alpha on the curve I start accepting my alpha. So, I am not very near to the starting point 0. So, this is the whole idea that you have another curve whose slope is another, you basically draw a tangent to this curve at the point alpha equal to 0 tangent to the curve phi alpha.

And then you see that where that curve or where that tangent cuts the x axis from there you consider from right to that till alpha naught you consider your acceptable alpha. So, this your acceptable alpha now this part is your acceptable alpha. So, this is basically the idea of the Wolfe condition. And now we will write down a very fundamental result which says that the Wolfe condition will actually gets satisfied. The Wolfe condition gets satisfied for very, very simple scenario.

(Refer Slide Time: 12:44)



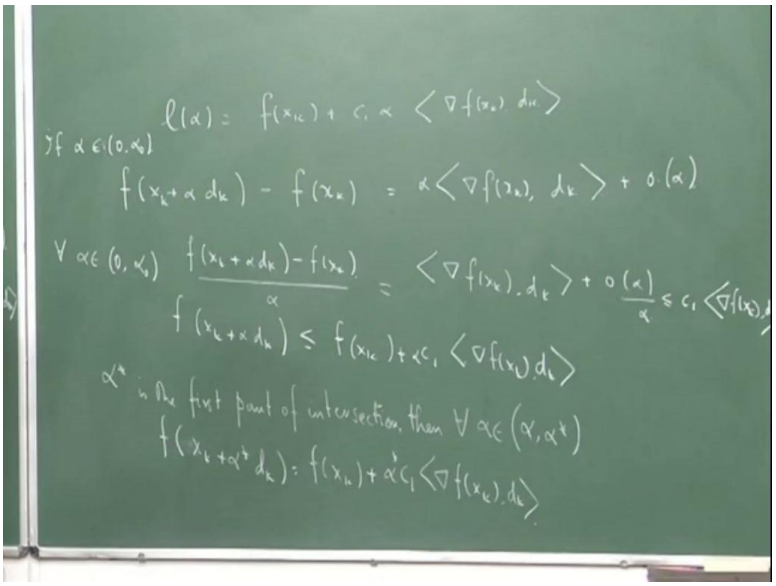
So, let us write down the result and then we will prove it on the black board. So, let f let be a function from R n to R is continuously differentiable which means smooth. So, it is continuously differentiable, let d k be the direction of descent at x k x k. And assume that f is bounded below on the ray it becomes a one dimensional minimization problem you see f is bounded below on the ray that it has to be this. Then if we, has this constants fix some constant there exists and if this happens there exists intervals of step length satisfying the Wolfe condition.

So, this is something you need to really remember it is not that you need to be very, very bothered about, but it is good to have an idea of the proof so that you get into the habit of

knowing that even in this issues of numerical optimization. We have to be really careful of taking care of the mathematics whatever statement we make you, we need to prove it even if you are making this sort of approximations with Wolfe conditions etcetera, and Armijo condition trying to find alpha approximately is imperative on us that our alpha that whatever we want such an alpha exists.

So, when we run the algorithm we are sure that if you are doing such an operation, this operation would actually give me something. So, now we will now go for the proof. Now, you see what happens is that just after alpha equal to 0 that is x equal to x k the function value decreases quite shortly, because that is what would happen till an interval alpha naught. And then it starts increasing a bit and goes up and you know this line l alpha this is unbounded below.

(Refer Slide Time: 16:02)



Because if you take the line l alpha this f x k plus C 1 times alpha into grad f x d k. Now, grad f x d k is negative C 1 is positive alpha is positive. So, I can now if this is negative I can make this alpha going towards infinity and make the whole thing negative and larger in the negative sense. So, this value will keep on dropping, keep on dropping keep on dropping, but but it will not have a bound now f is assumed to have a bound. So, basically, now this line l alpha at the very beginning, because what would happen is I can always have an alpha. So, for till alpha naught so here so till alpha naught the function values are decreasing. So, f of x k plus alpha d k is strictly minus f x naught if x k is strictly less than

0. So, this is a fixed for, for all alpha for a given a alpha. So, if, if alpha is between 0, and alpha naught by looking at this diagram. Then this difference is strictly less than 0. So, it is some number some k some.

Now, I can multiply this, this grad effects by this alpha part and multiply by a chosen C 1 between 0 and 1, so that I can make this. So this number this alpha k can be made to be less than C 1 into you can choose C 1 like this right we can always choose c 1 like this, because this is approximately this value, because by a by, by tailors theorem, what could happen is by tailors theorem? You can write this as nothing but or, or by the very basic definition of differentiability you can write this as small o of alpha. So, now this thing to which means that this thing is strictly less than 0; this whole thing; this thing is strictly less than 0. Now, I know that this is also strictly less than 0. Now, what I can do for a alpha very, very small I can now if I divide by alpha I will have f of x k plus alpha d k minus f of x k divided by alpha is equal to grad of f x k d k plus o alpha by alpha.

Now, this is true for any alpha if I make the alpha very, very small I can make this thing very, very small, and this this negative negative part will over run this even if it is positive. So, this will basically become negative. Now, what I can do this is of course, I can make the whole thing to be less than some C 1 between 0 and one this is of course, true. So, this whole thing is negative this is less than a fraction of this. So, fraction of this is bigger than I can take a C 1 in such a way that this whole thing is less than a fraction of this because this will become very very small. So, this will dominate and hence this can be made to be less than a fraction of this would, would which would be bigger in the negative sense which will be more near 0.

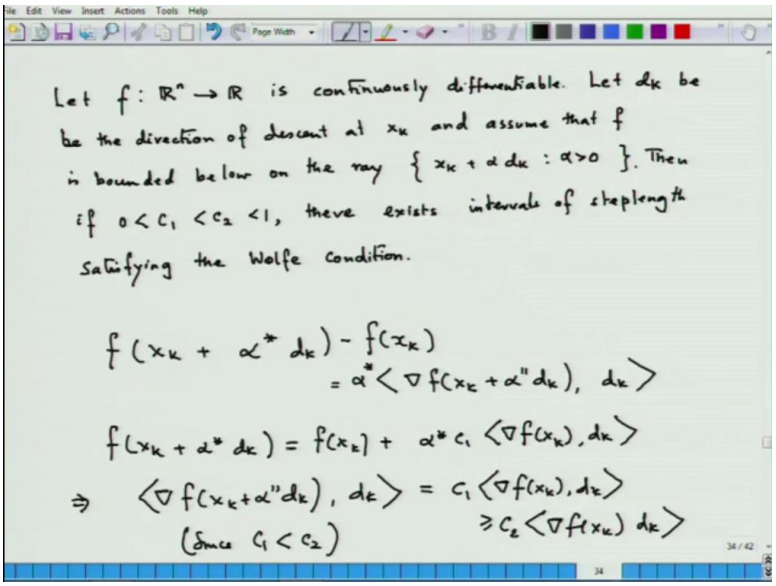
So, then what happens is that so for for alpha for all alpha it is between 0 and alpha not at least I can have f of x k plus alpha d k less than equal to f of x k plus C 1 into alpha times grad f x k d k. So, I meet the Armijo condition, now you see I have said that up to alpha naught till alpha naught, this point my Armijo conditions are met so maybe I take some alpha here where you got some alpha just slightly less than alpha naught. Now, in this interval this is a continuous function in this and in this interval it will have a minimum possibly the minimum is here and then it is going up like this. Now, what happens, because this is unbounded this line there will be a first point where it first intersects this curve now I can find a band of alpha. So, here it will intersect this curve, so if you observe

from the geometrical picture that, all the curve now is lying below this line. So, from this alpha star where it intersects I will have the sufficient decrease condition holding.

So that is so if alpha dash or alpha star if alpha star is the first point of the, is the first point of intersection, then for all alpha for all alpha between alpha and alpha star basically. So, for all these function values are less then f of x k here was my f of x k value here what would have happen is that in this particular picture starting from alpha dash, all the function values are basically less than f x k f x k is here all the function values are less than f x k. So, you that what happens is this thing comes down after f x k in a neighborhood. And then it starts raising up if the function value comes down which takes the minimum value in this certain interval and then it starts rising up where it comes and hits this curve that is it becomes equal.

So, then at that point what we would have is that f of x k plus alpha double dash alpha dash alpha star d k is f of x k plus alpha C 1 and alpha star star f of x k into d k. So, this is what will happen, so you see we so we, we have found the Armijo condition. So, this alpha star that means it intersects for all alpha star which is less for all alpha element of alpha star for alpha star we would have this. So, for all alpha element less then alpha star we would basically have this, that is what you can see from the diagram. Now, the next point is that we shall use the mean value theorem. So, let us do that here in this thing is that now you have got this alpha star.

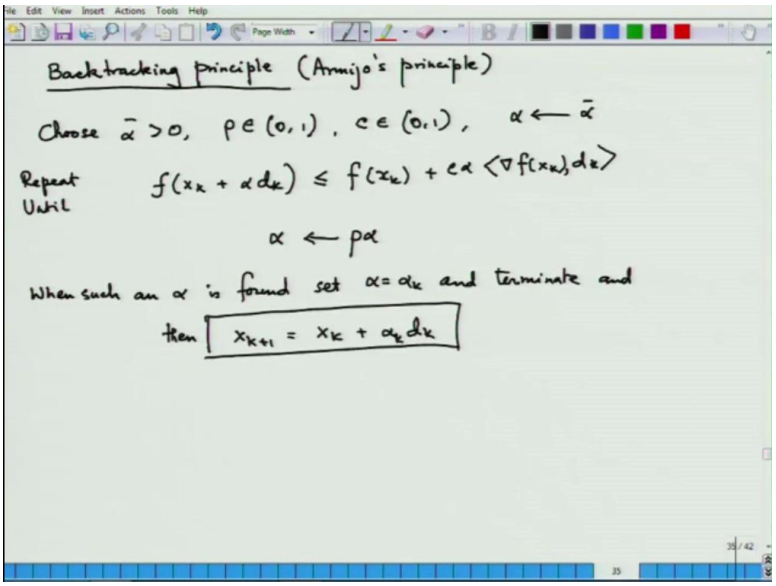
(Refer Slide Time: 25:00)



So, let us use f of the mean value theorem alpha star times d k minus f of x k this is equal to alpha star times grad of f of x k plus some alpha double dash d k, where this thing is lying between this; this point. And this point is a point inside lying in the interior of the line segment joining these 2 points. Now, what we have? We also have this expression that was from there where at this point alpha star it comes and intersects the line comes and intersects the curve. So, here what you can do is this now look at this, this difference. So, what, what I am having is a following I am having. So, this would imply that grad of f of x k plus alpha star d k alpha double, double dash d k d k is equal to this 1 C 1 times. Now C 2 is this is a negative quantity a C 2 is lesser than C 1.

Now C 2 is lying between bigger than C 1 C 2 is lying between C 1 and 1. So, naturally this would be this would be if C 1 is lesser then this would be less negative than this one, so C 1 is a positive quantity which is lesser than C 2 that is since C 1 is lesser than C 2 this quantity would be less negative than this quantity. So, my alpha dash double dash is the required. So, alpha double dash is actually lying between alpha star and 0 so alpha double dash is actually my required alpha dash by in for which the Wolfe condition is satisfied. So, here we have a proof. Now, we are going to discuss about how do I actually apply this sufficient decreasing decrease?

(Refer Slide Time: 28:14)



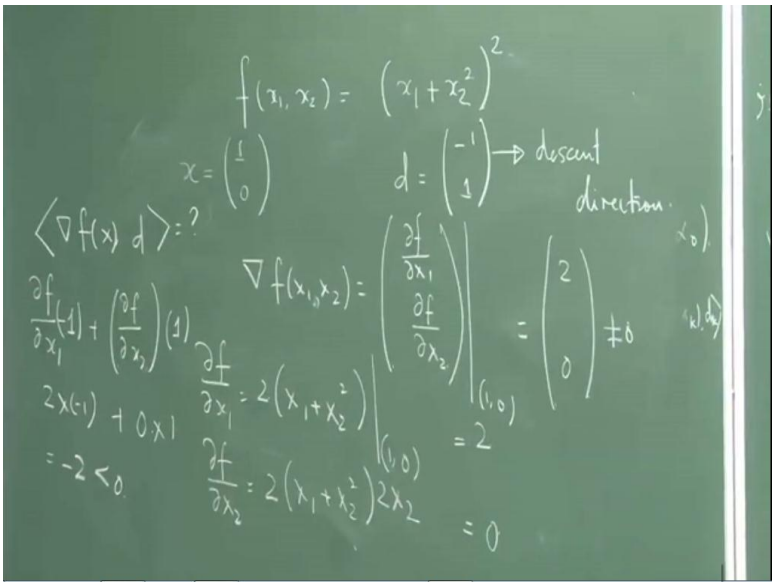
So, this sufficient decrease is done by a method called backtracking, how do I choose my backtracking principle? So, how do I choose my step length by using backtracking

principle? Now, this is also called Armijo principle due to mathematician Armijo. So, this is quite often applied in algorithms and you know to get a result quite fast. So, you choose alpha bar some alpha bar greater than 0 and choose a row between 0 and 1 and C between 0 and 1 that is the c 1 actually. Now, initialize alpha with alpha bar some alpha bar, you have chosen now repeat until repeat until you have f of x k plus alpha d k so until, so d k is a descent direction.

So, how do you find this alpha bar? This is the following way let us see. So, if you find an alpha take an alpha say alpha bar and you see that this condition is not met for a given that for chosen c, this condition is not met what you do? So, if this condition is not met, you reduce alpha by a certain amount you take a fraction of alpha and say if it is alpha bar in the starting.

So, you take a fraction of alpha bar and put it to the new alpha and then check with the new alpha weather this condition is satisfied with this particular C, if it is not and then go on basically doing. Then when such an alpha is found set alpha equal to alpha k and terminate, and then x k plus 1 a new point is x k plus alpha k d k. So, this is what is done in actual practice. Now, let me give you a 1 or 2 examples. For example, how would you find, let me take an ex example from the exercise of Nocedal. So, let us try to summarize what we have learnt in this section on descent directions, and finding a control step length by which you control the movement from x k.

(Refer Slide Time: 31:25)

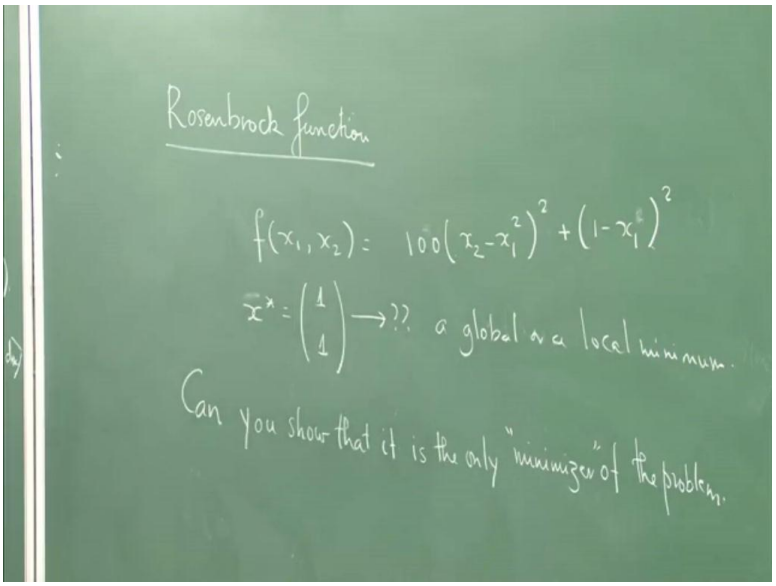


So, you take a function of 2 variables and the function of 2 variables here is x 1 plus x 2 square whole square. And they are taking a x the point x as 1 0 and they giving us a deduction d minus 1 1. Now, a question is, is this d a direction of descent from this point 1 0. So, how do I first know whether I need to have a descent to have to have a descent I first have to find the gradient of this function So, the gradient of this function; this function if you look at it I have to compute this value and this value at the points 1 0. So, what is my del f del x 1 here? del f del x 1 is 2 times x 1 plus x 2 square and into x 1 is 1. So, by chain rule so del of del x 2 is 2 times x 1 plus x 2 square into 2 x 2, now if I put here I compute the value at 1 0 that is I put x 1 equal to 1 and x 2 equal to 0, then what I get here is 2. And if I put x 1 equal to 1 and x 2 equal to 0 then what I get here is 0. So, this value at 1 0 is nothing but 2 0 and this is not equal to 0 and hence this point is not a point or local minimizer.

So, this is not a critical point so then I have to move to a better point from 1 0. So, what I would do to check that weather this is actually a gradient or not. To do this I have to check this one. I have to see weather grad of f x into d, what is, what is this value? Which means I take, take the inner product del of del x 1 into 1 plus into minus 1 plus del of del x 2, both are obviously computed at the point 1 0, point 1 0 into 1. So, this computed at the point 1 0 is 2 2 into minus 1; this computed at the point 1 0 is 0 so 0 into 1. So, what I get is minus 2 which is strictly less than 0 showing that this is indeed a descent direction.

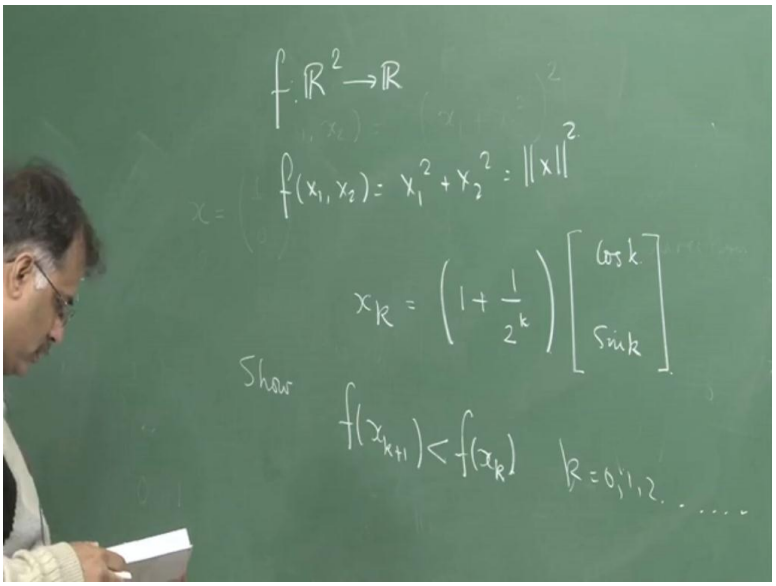
Now, let me give you a home work from the book of Nocedal and Wright and I expect that you really go ahead and look into this home work, because that will give you a little practice, because still a beginning we have not done any practice here. So, this course is essentially a course at the advance graduate level and advanced under grad level and beginning graduate level. So, we really need to pull up our socks and look into examples. So, now I give you a example of very important test function which is used in optimization to do the job do various demonstrations of algorithms.

(Refer Slide Time: 35:57)



So, here this is called the Rosen Brock function. Let me again show you the reference of the book that I am using the book called numerical optimization. And it is written by Nocedal and Wright it is published by Springer. And it is now available in Indian edition I have already mentioned it only minimizer of the problem. Can you show that so hence can you show that it is the only minimizer of the problem? So, I am giving a apostrophe here. So, I am asking you also is this a global or a local minimum tomorrow we will start discussing the steepest descent method. And after we discuss steepest descent method I will give you a solution of this problem. But I expect all of you to really have a look at this problem at home, and try to do this problem this is a very, very important function and has a many many gives a lot of demonstrations when you study numerical algorithms (( )). Now, what I want to ask you is the following; function f x equal to norm of x is square x is in R 2.

(Refer Slide Time: 38:03)



So, f is a function from R 2 to R and my f of x 1 x 2 is equal to x 1 square plus x 2 square which is non x square. Now, let me write down an iterative algorithm to minimize it that is I am writing down some sort of iteration that is how I go from x k to x k plus 1. Now, let me take a kth iteration is of the form into Cos k and Sin k. What I am asking is to show that, this also we can discuss tomorrow after we discuss steepest descent method.

So, this just show this so these would be two problems for you which you have to show, but here you see the iterates the function value is decreasing think about the geometry of this. But please keep a note that you have to do this. Thank you very much for today. And tomorrow, we will get into this exciting thing of steepest descent method and we will use steepest method for the particular class of quadratic optimization problem where the hessian matrix is positive definite. And we will show how we can understand rates of convergence and other issues related to an optimization algorithm.

So, the optimization algorithm in the unconstraint sense is not just solving grad f x equal to 0 that equation, but it also has to entail that every x k plus one every k plus 1 iteration, my function value has to be better than f of x k. So that is that is the important thing that one has to realize; one does when one uses the optimization algorithm.

So, thank you very much for today, see you tomorrow.