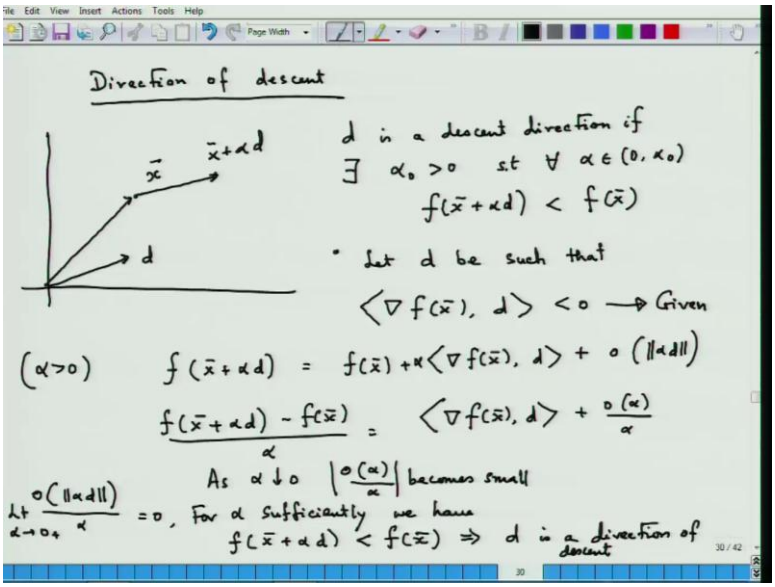
Foundation of Optimization
Prof. Dr. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

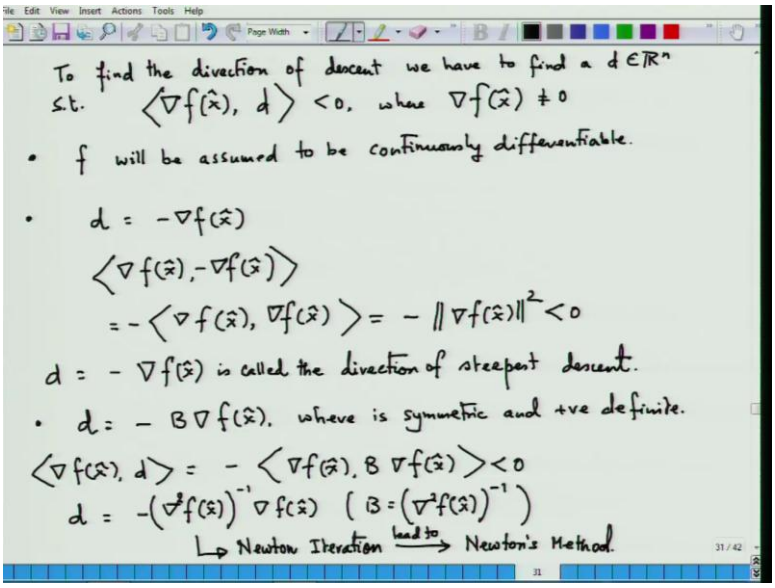
Lecture - 4

(Refer Slide Time: 00:24)



Welcome once again to this discussion on very fundamental issues about optimization. Now, we learnt about what is the descent direction in the last lecture, and we found that if this sort of condition this criteria, this grad f x bar into d is strictly less than 0 and this d becomes a descent direction.

(Refer Slide Time: 00:47)



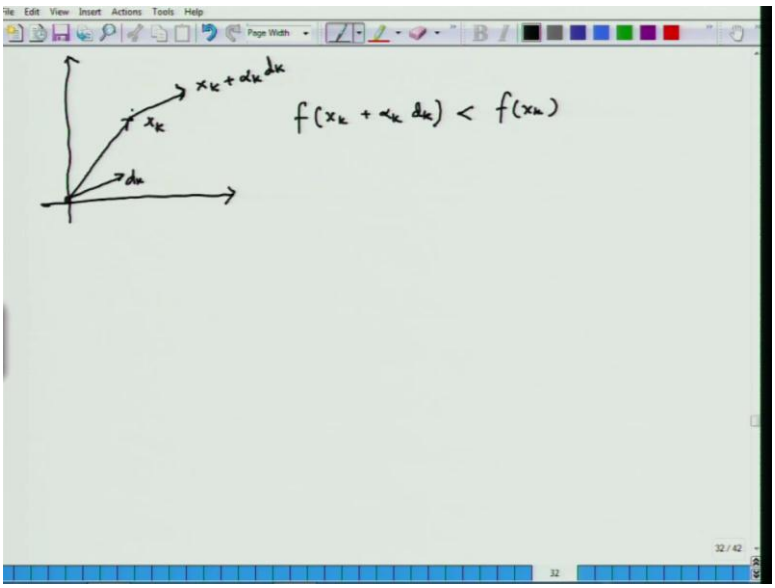
So, for us to find the direction of descent we have to find a d, small d such that at some x delta where grad f x bar delta is not equal to 0. We have to remember that in our setup grad f x bar is not equal to the function f is assumed to be differentiable not only differentiable, we can assume it to be continuously differentiable. So, for all purposes I again want to state that f will be assumed to be continuously differentiable; is not only the function is differentiable all the partial derivatives that you have of this function, they are also differentiable as functions of x, 1 x to dot dot dot x n. Now let us give some examples of directions of descent. Suppose I choose d is equal to minus grad f at x hat. So, then your grad f of x at is not 0 then putting this. So grad f x, what d is now this. So, this is equal to minus of grad f of x hat into grad f of x hat, not into in a product.

This is nothing but norm of, norm you know with a positive quantity in this case because grad f x bar is not equal to 0. So, this is strictly less than 0. So, this direction d is called the direction of steepest descent. We will come very soon to study the problem of solving an optimization problem unconstant optimization problem by using the method of steepest descent, but for the moment just know that it is called the method of steepest descent and we will actually explain why it is called the method of steepest descent or the direction of steepest descent. Another one could be like this for example, if you take d to be minus B time where B is symmetric and positive definite where B is symmetric and positive definite. So, let us see what would happen with grad f x delta into d.

Now you see this vector d which is the direction of steepest descent d cannot be equal to 0 because if d is equal to 0 it cannot be direction of steepest descent and we do not want 0 vectors because we actually in some point where you know the optimized is not achieved or local minimum is not even achieved. Then you want to move away from that point; you want to move to a point where the function value decreases sufficiently decreases. So, you move certain distances away from the point. So, if you put d is equal to 0 then of course that has no meaning. So, in this particular case we will have nothing but. Now grad of f x delta is not equal to 0 that is the basic assumption but B in positive definite this would mean that this whole thing is strictly less than 0, because this part is strictly greater than 0 this part, and so this is strictly less than 0.

So, this is also a direction of descent; one of the important directions of descent is to have something like this at f x bar take the Hessian matrix, we have already known what Hessian matrix is. So in place of B, I have put this matrix. So, here this is a particular case with B equal to the inverse of the Hessian matrix. Now this is B is positive symmetric definite. So, we had this. But it might if say the Hessian matrix itself is positive semidefinite at x hat, then B itself is also the inverse is also positive semidefinite. So, when d is taken in this way this is called the Newton iteration. Newton iteration would lead to Newton's method which is a very, very important class of methods; we shall study in detail. So, we have this very basic studying information; now see what we can do.

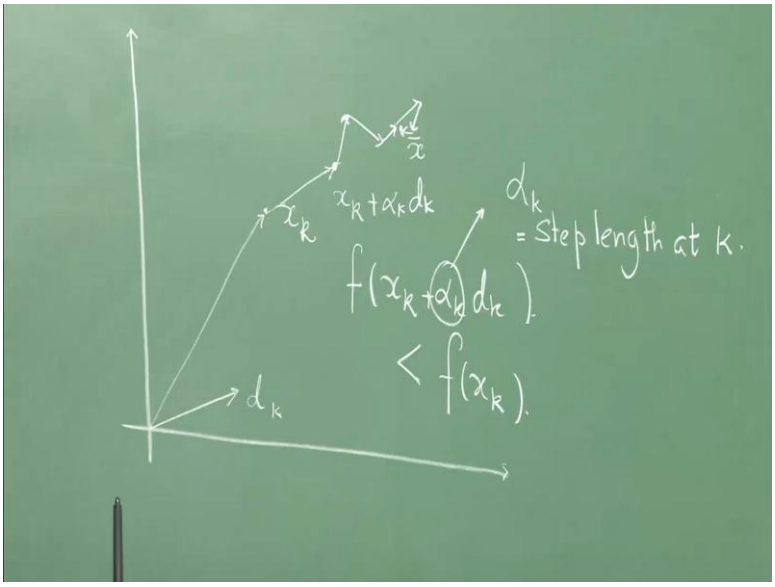
(Refer Slide Time: 07:19)



So, now what we know is that if this is my x k at the present moment. So, x k is the current iterate; so it is the k th iteration and you know the grad of f x k is not equal to 0 and suppose this is my direction of descent. So, I want to move from x k in this direction. So, I move a distance I come to a point x k plus some alpha k d k, and what I want to have is at that point f of x k plus alpha k or you can say this is d k that is a descent direction chosen for x k to move from x k; this one must be strictly less than.

So, now you once you know the descent direction its fine, but you cannot move very little from x k. Then your movement is not very good even if your function value decreases you have not moved away quite a good distance because if you do not move away quite a good distance, then you are not improving your algorithm; you are not going quite faster towards the minimum. The goal of optimization algorithms is to make you start from the starting point and reach the optimum solution as fast as possible but that has not been possible here.

(Refer Slide Time: 08:59)

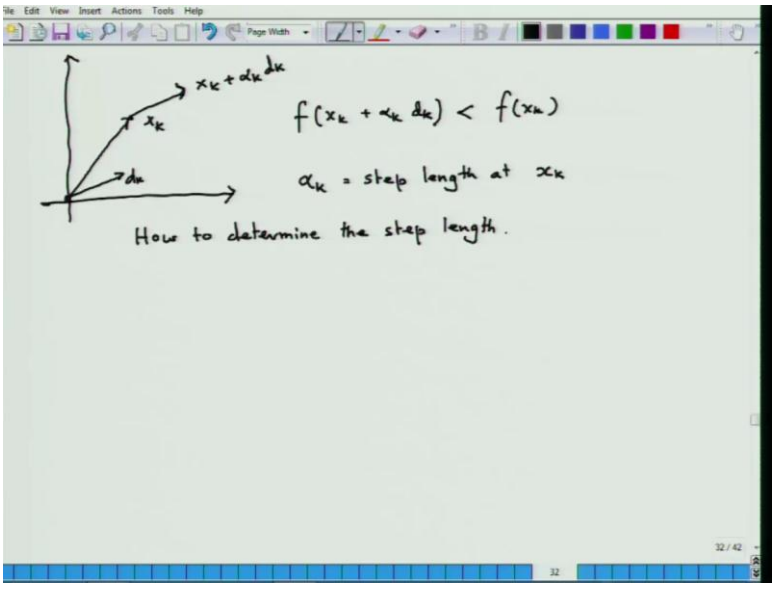


So, maybe I will use a chalk for a while; see what happens here is the following. Now here you have the x k and your actual optima are here your x bar. So, now here you have a descent direction. Now here you are moving; in the direction d k if you move then your function value decreases, but you cannot move long for a long while because you know that there is threshold alpha not beyond which this is all possible.

Now you have to move as far as possible from x k, because then you will have this new point x k plus alpha k d k. So, this new point you have to move in such a way so that my function value decreases as much as possible because I want to minimize. Because from here the new point, say, you need to go to some direction may be it is decreasing in this direction from there you need to go this direction and this direction and so on. Maybe it will come to this direction, this direction and this direction.

So, now what we are essentially doing; so you are moving along a line and then searching how much we have to move along that direction d k. So, this would be. So, we will decrease the function value. Our objective here is of course that you should have, but if I move alpha k very less if I come here, then my decrease is also very less. I am making a very less decrease which is not very, very good. So, the idea is that you cannot make such a small decrease, you cannot just come here. So, you have to make a sufficient decrease. So, how much is that sufficient decrease would be the question. Now what is this alpha k? This alpha k is called a step length at k and this alpha k the step length at k; this plays a fundamentally crucial role. So, this is something you need to take care.

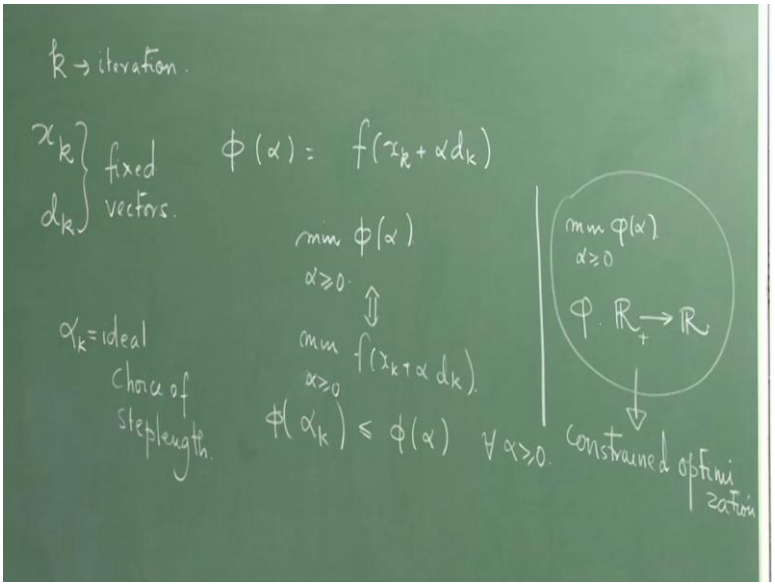
(Refer Slide Time: 11:37)



Here we can call alpha k. Now the question is how to determine the step length? So, whatever be your descent direction it does not matter; our point of concern would be how

to determine the step length and for that I would again go and do certain explanations in the board.

(Refer Slide Time: 12:28)



So, what you have here in this function phi alpha. So, once I know my x k. So, for a fixed given k, k th iteration. So, I have actually fixed. So, this x k and your d k these are now fixed vectors. So now, once this are fixed vectors let me construct this function phi alpha. Now you see this is the function value of f along this line, points on this line, alpha could be any number strictly bigger than 0.

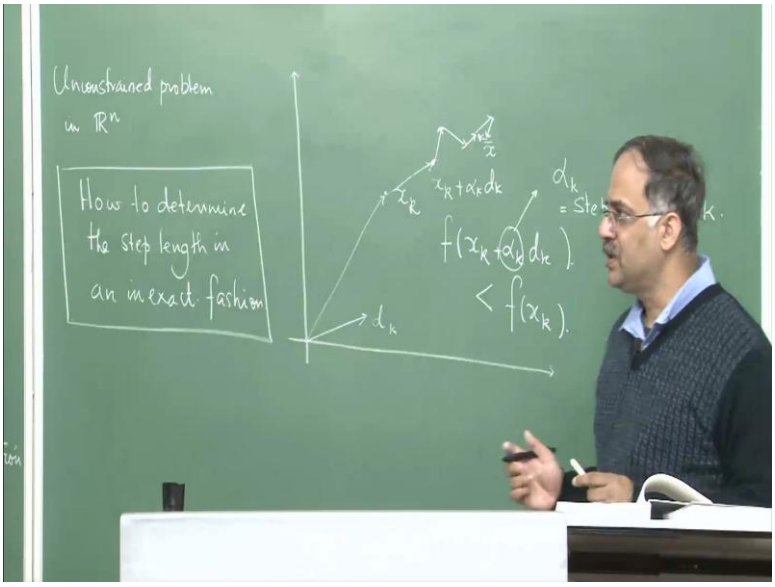
Now I have to know to what extent I can keep on decreasing the value of f because d is a descent direction. I know that there exist an alpha not beyond which for every alpha this function value is strictly less than f x k, but to what extent I will get the maximum decrease. The natural thing is to find an alpha for which along this line along with direction d k this function value is minimized.

Then I am basically trying to find or more specifically this is the same as writing. Now suppose I find a minimum of this problem that is alpha k, then the function value at alpha k that is. So, phi of alpha k is less than equal to phi alpha for all alpha greater than 0; step length is always positive. So, now what you see is that this alpha k should be ideally my choice of step length. So, alpha k is ideal choice of step length. The question is can we keep on doing this in practice. So, we have seen on the board as to what is what should be the ideal step length. The ideal step length would be to find the minimizer, but the

question is that can we really go on doing that; that is can we really find the ideal, can we really find the minimizer.

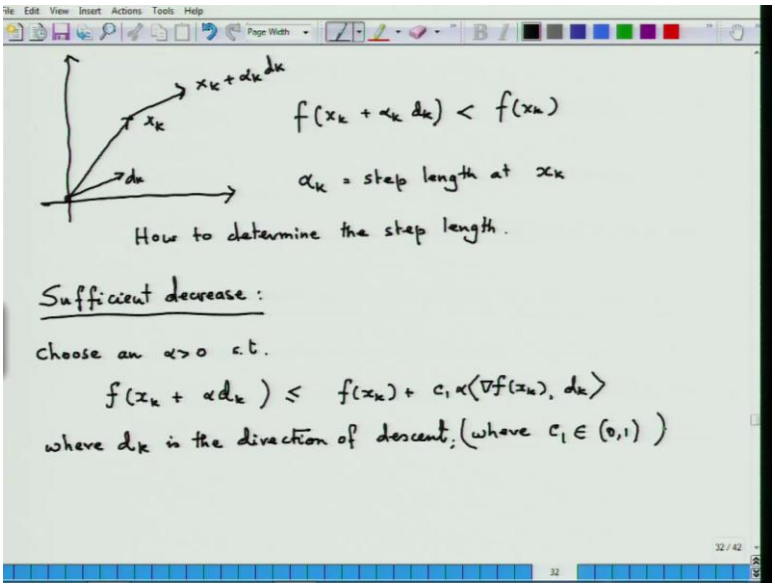
So, you see here the problem becomes difficult; difficult in the sense is that we want to minimize function phi alpha over alpha greater than equal to 0 where phi is a function from R or R plus to R; R plus is a set of all nonnegative real numbers. Now here this problem is a constraint optimization problem because you have an additional constraint that alpha is restricted to the nonnegative part of R. It is the constrained optimization problem. So, what we are having that. So, we have an unconstrained problem in a higher dimension. So, an unconstrained problem in R n needs the help of a constrained problem in R. So, this is a quite an important issue, but there are lot of methods of how to solve such a problem by bisection method, this method, that method. So, but we are not going to get into the details of this. So, the trick here is the following question as how to determine the step length.

(Refer Slide Time: 17:24)



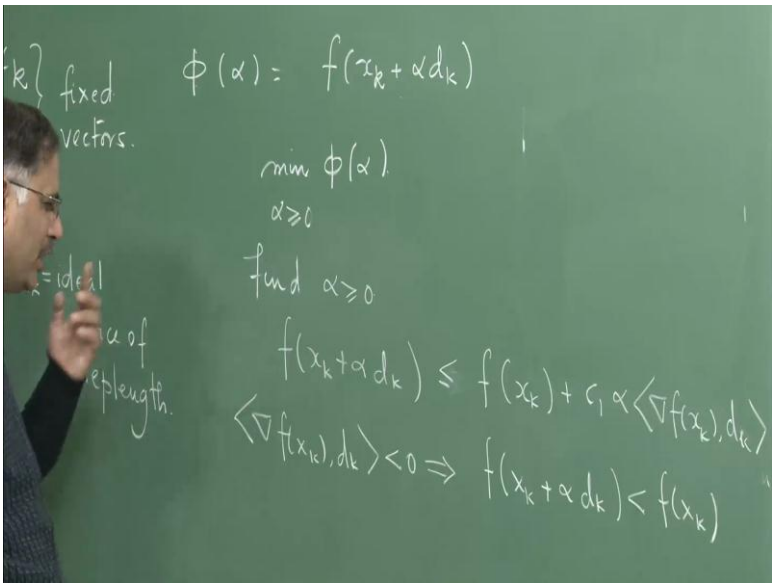
But the more important question would be how to determine the step length in an exact fashion; that is you need not find the exact alpha which minimizes this problem, but you can find an alpha which makes a sufficient decrease. So, what is this sufficient decrease called?

(Refer Slide Time: 18:04)



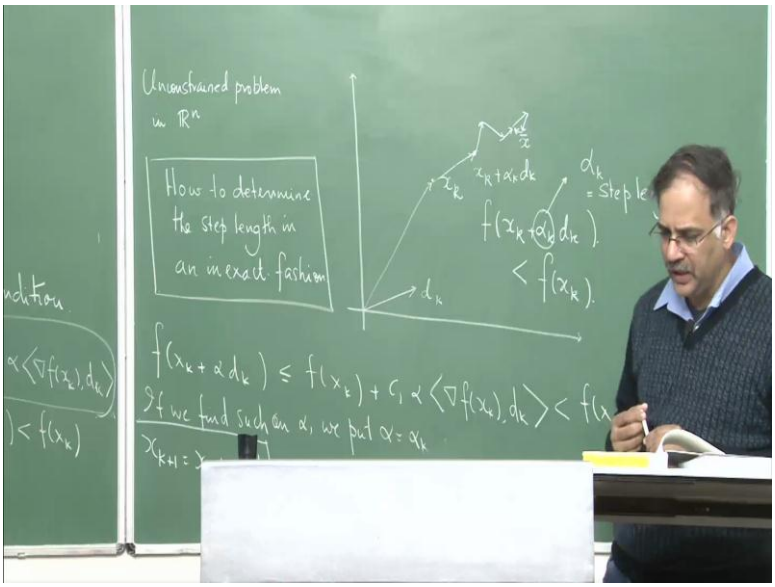
So, let us define what we would like to have meant by sufficient decrease. What we are doing now is essentially the fundamentals of line search method. So, we are going to write down something, but what we are doing here or thing is called line search method. There is something called trust region method we will come later on. Now what do you mean by sufficient decrease. So, what should be the alpha? So, choose an alpha greater than 0 such that f of x k plus alpha times d k is less than equal to f of x k plus c 1 times alpha times grad of f of x k into d k where d k is the direction of descent. So, let us remove this little part here because we have already started this in exact business. So, let us see what does this statement means, the sufficient decrease condition actually means. So, now of course you have to choose c 1 where is c 1 from; c 1 some number chosen between 0 and 1 excluding 0 and 1. So, what we have done is the following.

(Refer Slide Time: 20:25)



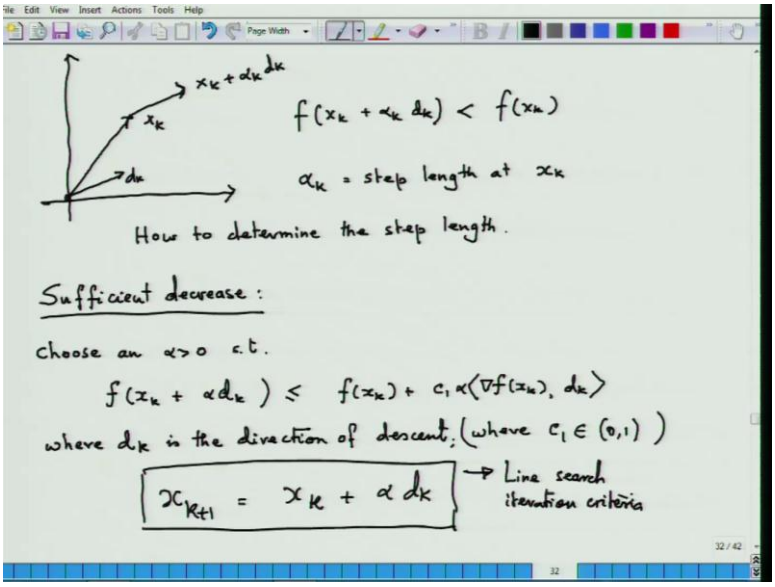
So, we have to find an alpha such that f of x k. Now because d k is the direction of descent we have this to be strictly less than 0, but alpha is strictly bigger than 0, c 1 is strictly bigger than 0. So, this would imply that for such an alpha, it would imply that for such an alpha, this is a negative quantity. So, f of x k plus a negative quantity is obviously strictly less than f of x k; this is a strictly negative quantity. So, f of x k plus alpha because this part become completely negative; you can see the interesting part is this but we are not only expecting this to be true, but we are expecting that this is lesser than some quantity which is quite less than this quantity.

(Refer Slide Time: 21:56)



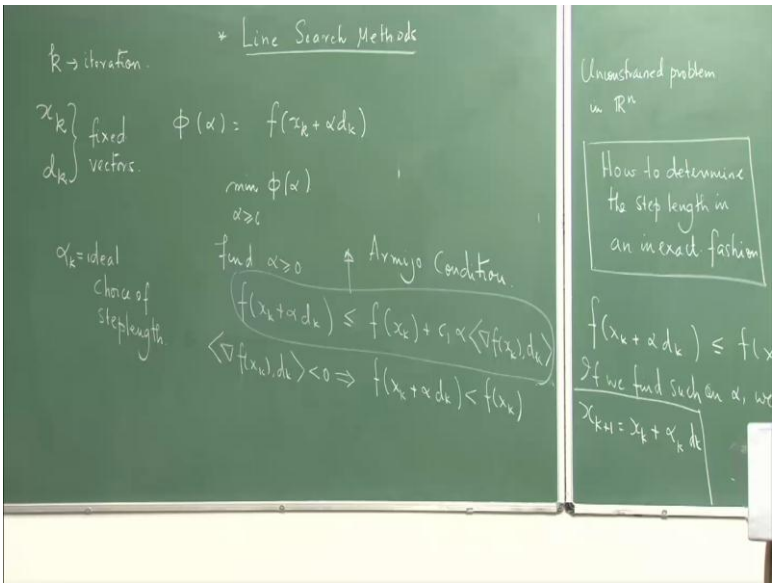
Essentially what we have done is that we want to have x k and alpha which has this property. So this if you can find such an alpha, if we find such an alpha we put alpha is equal to alpha k. We can now put x k plus 1 is equal to x k that is the k plus 1 iteration value that is this is my k plus 1 th chosen point; it could be a solution it could not be a solution alpha k d k .

(Refer Slide Time: 23:06)



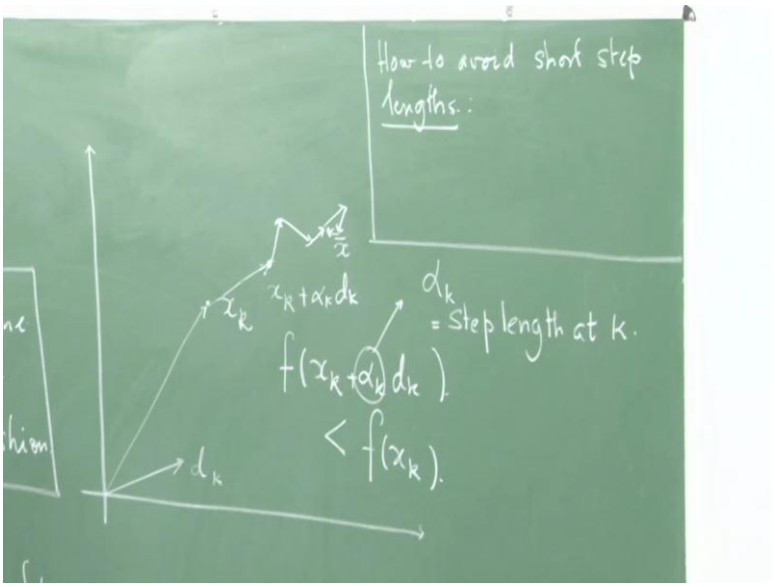
What we have done if you can find such an alpha here then we put x of k plus 1 is x k plus alpha times d k. So, this is what is called a line search iteration criteria or this is called line search. Direction of steepest descent it is sometimes called the Armijo condition; this is called the Armijo. This condition this steepest this sort of condition is just called the Armijo condition.

(Refer Slide Time: 23:53)



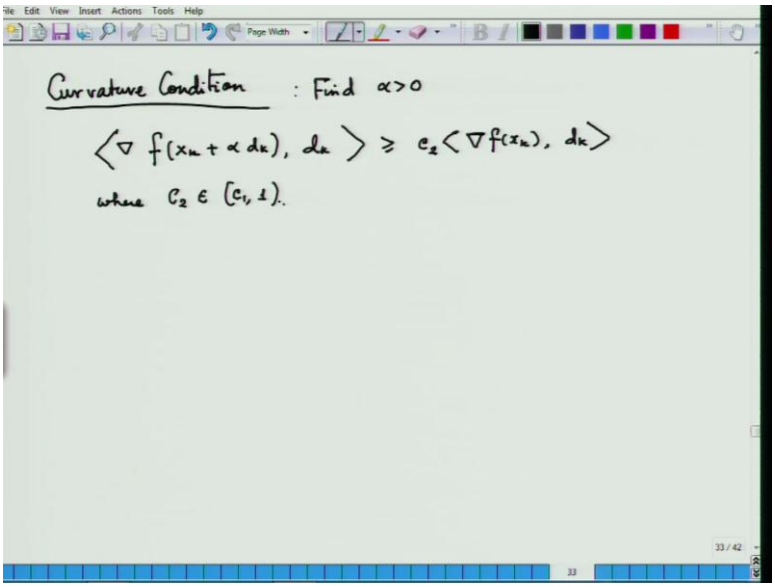
Now there has been an issue which comes up here is that you know even if you choose c 1 which is the fraction of this thing even to get this alpha, this alpha could be very, very small. This alpha will have to be very, very small, but if this alpha is becoming very, very small then you have not made much movement from x k; if alpha is very, very small then you are here. Then actually possibly you should be here for which you have the actual alpha which minimizes the function phi alpha. So, the whole thing is that then how do you stop such short step lengths. If you have a very short step length then you do not make much progress from x k and that is not desired when you are running an optimization algorithm.

(Refer Slide Time: 25:02)



So, the question is how to avoid short step lengths? To avoid short step lengths you have to consider another criteria called the curvature condition.

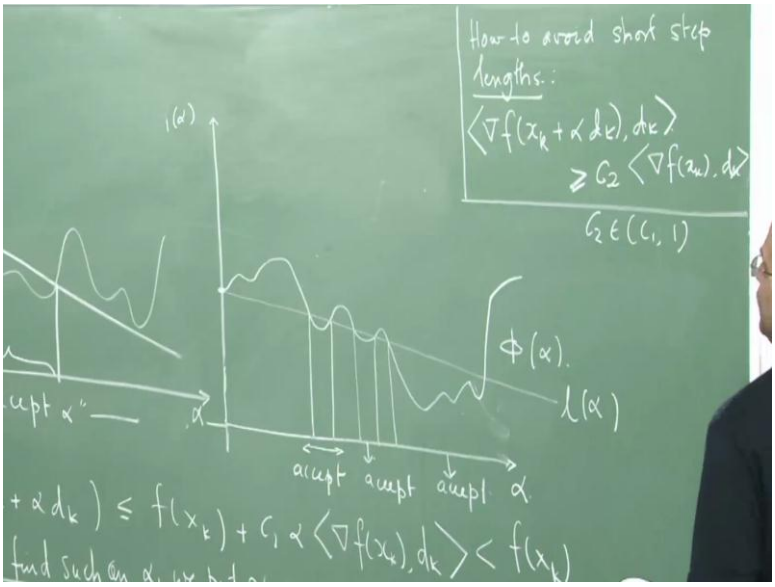
(Refer Slide Time: 25:31)



So, we have to now consider what is called the curvature condition. So, what does the curvature conditions tells us? Curvature condition tells us find an alpha such that you would have this. So, such that the gradient at f of x k plus 1 basically. See what happens is the following. Let us try to understand the sufficient decrease condition Armijo

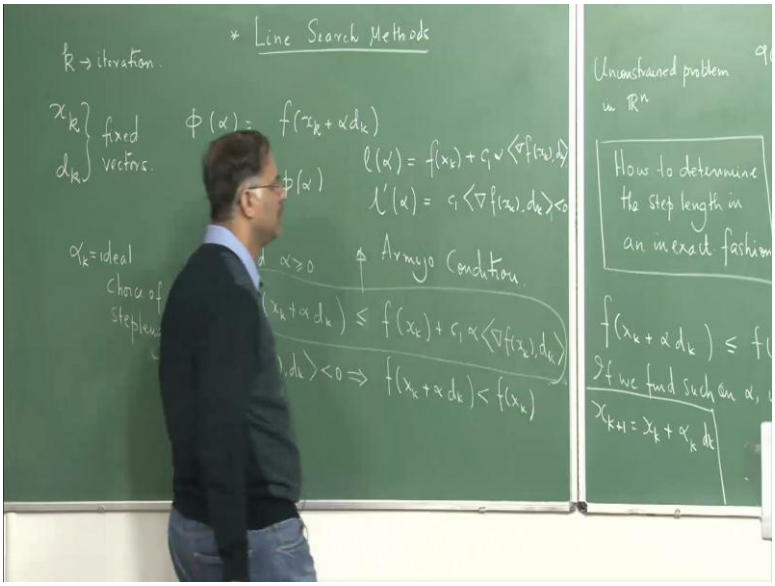
condition in a more geometric fashion. So, let us use this diagram here; we will just make certain changes. Let us try to understand this in a better way.

(Refer Slide Time: 26:57)



So, let us consider this part as alpha and this is my phi alpha. When alpha is 0 then it is f of x k basically. So, phi alpha is f of x k some value here. Now as alpha is changing this is say phi alpha, this is say the graph of phi alpha. This is the graph of phi alpha.

(Refer Slide Time: 27:46)



Now what happens? Look at this expression here; let me call this as l alpha. Let we write l alpha is f of x k plus c 1 times alpha into grad f x k d k. Now at alpha equal to 0 this is f

of x k; this is your f of x k. Now let me take the derivative of l alpha. So, the slope of the line l alpha; so l alpha given an alpha this is y equals m x plus c form. So, l alpha is straight line, but how is it slope in which direction it is moving. So, its slope l alpha at direction f dash x at alpha is nothing but c 1 into grad f x k d k.

So, it is strictly less than 0. So, now l alpha is some line like this. So, this is my l alpha. I am going to accept that alpha for which this value or phi of alpha. So what I want to, I want to accept alpha. So, my idea is simple; accept alpha if phi of alpha is less than l of alpha. So, you see there is a negative slope. The angle at which it will hit the x axis is obtuse.

So, let us see what happens here is that you see here phi of alpha is bigger than l alpha. So, I do not accept that alpha; here it is less. So, this is the acceptable zone, accept alpha. So, these alphas are acceptable, but these alphas are not, these alphas are not, this alpha is again acceptable while this is not, again this alpha is acceptable; right. So, this is accept. You need not move away so much, right; may be it is usually something like this, like this; I am just changing the drawing a bit because you cannot have all alphas acceptable.

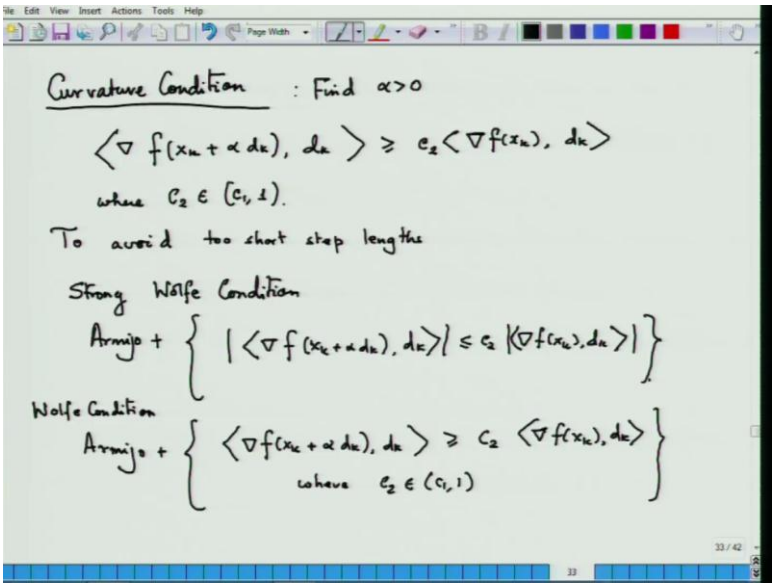
So, you see that. So, these are the regions of acceptance. So, these are the alphas you would accept as your choice, but once you have this curvature, but now here you see that there is no problem, but it could be slightly problematic if we had this. Now let me change the diagram a little bit and let me look into this situation. If I look into this situation where you have say phi alpha in this fashion.

So, this is my phi alpha and this alpha. So, this is my l alpha. So, you see this is my acceptable zone, this is accept zone, accept alpha zone. So, this is my accept alpha. Now here I could choose an alpha which is very, very small. So, I have not moved sufficiently off from x k. So, in order to avoid that I need to have the curvature condition which tells me that too much short step lengths cannot be taken which condition is nothing but as we have written in the other format. We need to find an alpha which now you are asking something to be bigger.

Now this quantity is negative where c 2 is of course taken between c 1 and 1. This quantity is negative, but this quantity could be positive also or this quantity could be 0 also or this quantity is negative also. Of course, positive would not give me any

information then d k would not be the descent direction at; of course, it will be positive because there is no guarantee that d k is the descent direction at x k at this point. So, this is called the curvature condition which we have written down.

(Refer Slide Time: 33:27)

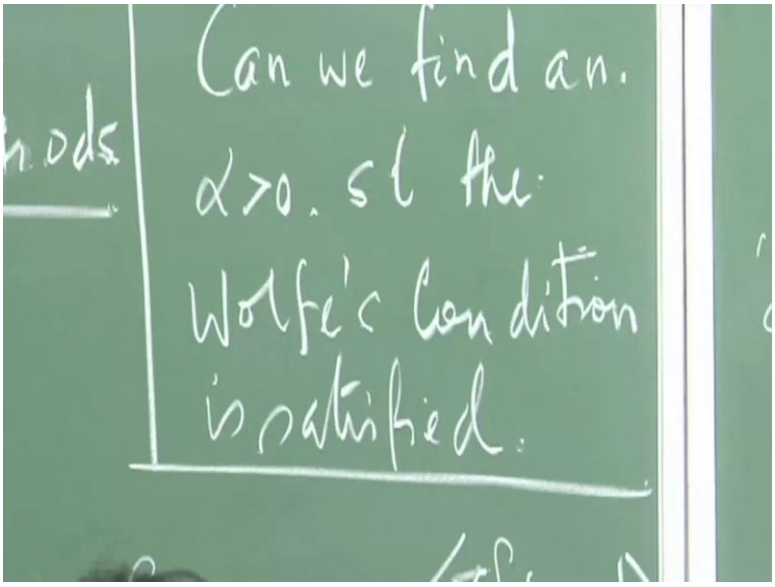


So, this is to avoid too short lengths. So, sometimes in algorithms people use Armijo in jointly with the curvature condition. So, what I want to say is the following that you can combine these two conditions, but there has been a result by the famous optimization theorist Wolfe. This is sometimes when you combine these two things Armijo conditions with the curvature conditions it called the weak Wolfe condition. So, you could also define the strong Wolfe condition. The strong Wolfe condition is Armijo plus this fact rife on to have the alpha satisfying the Armijo plus also having the alpha satisfying not just this. So, alpha has to satisfy both of the things.

So, Armijo this is called the strong Wolfe condition, another is called Wolfe condition. This is Armijo plus this condition that. Now we will show that if we consider the Wolfe condition then there exists interval lengths that is you can find intervals of alphas which satisfies both this conditions. Now, under a very mild condition that you will always find alpha; the very important thing is that if you are telling that I would take this condition and that condition I want to find an alpha; before you really do the numerical algorithms you need to really confirm that, yes I can actually go ahead and find an alpha. So, how do we find that alpha that, yeah, we can guarantee that such an alpha which satisfies the

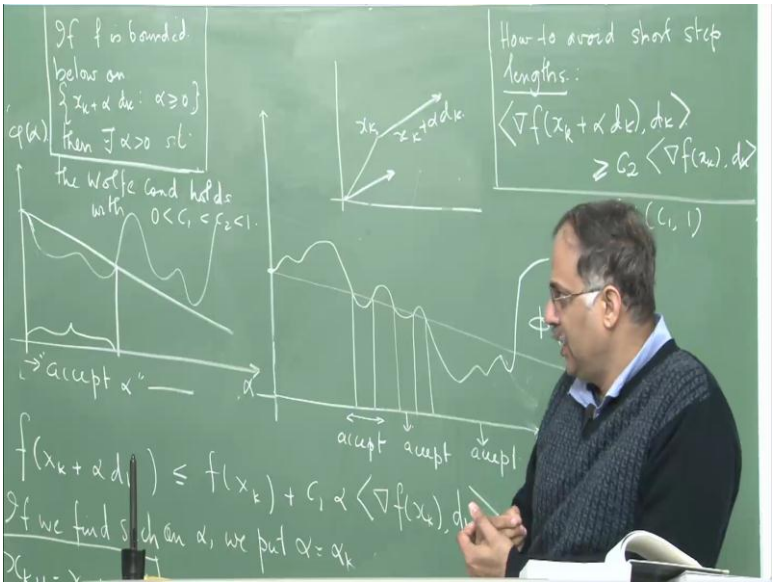
Wolfe condition or may be the strong Wolfe condition we will not bother about the strong Wolfe condition at all here.

(Refer Slide Time: 36:39)



So, my next question would be following. Can we find an alpha greater than 0 such that the Wolfe's condition is satisfied. So, tomorrow's lecture would begin with the study of this fact that, okay, yes, I will show that under a very, very mild condition which is nothing but which is assuming that the f is actually bounded below on this line.

(Refer Slide Time: 37:20)



That is if f is bounded below on this line at x k and here is again your line. This is the ray x k plus alpha d k where d k is the direction of descent, then d k is the direction of descent if along this line or along this ray where alpha is strictly greater than 0 greater than or equal to. If the function f is bounded below on this, then you are always sure to find an alpha which will satisfy the Wolfe condition or the strong Wolfe condition. So, this is a very, very important result. So, it says that that is what we are going to prove tomorrow. If f is bounded below on d k then there exist alpha greater than 0 such that the Wolfe condition holds. I have not written directly; that is Wolfe condition holds with, that is what happens.

So, this is what we will discuss tomorrow first we will prove this fact. Once we have proven this fact we would like to show how to use the Armijo condition; that is we do not use the Wolfe condition. In real practice we can actually use the Armijo condition and that is called the backtracking line search. So, we will see how we can us the backtracking line search and go ahead and do the algorithm and then we will really talk about a particular type of method called the steepest descent method. We will then talk about the convergence and the rate of convergence and all this fundamental issues related to algorithms.

So, when we run an algorithm we have generate those iterates but how do I assure that this iterates this sequence of iterates would converge in the mathematical sense of convergence of a sequence to a point which is at least a critical point of the function; that is at least a point where grad over x k is 0; more happily it would be if it goes to a point which is minimum. But we cannot guarantee that it will go to a minimum for just any function but for certain types of function it does. So, that is what we will do. So, we will stop here today.

Thank you very much.