Essential Mathematics for Machine Learning

Prof. S. K. Gupta
Department of Mathematics
Indian Institute of Technology, Roorkee

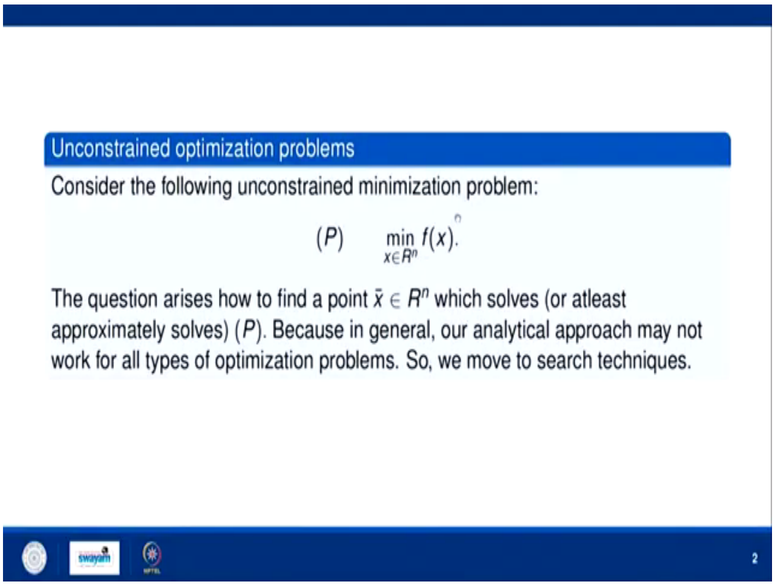
Lecture - 29
Steepest Descent Method

Hello friends. Welcome to lecture series on Essential Mathematics for Machine Learning. So however, we have seen analytical matrix to solve our non-linear problems. We have seen if a problem is a convex optimization problem, then the KKT conditions become sufficient. We can write the KKT condition and solve a problem a non-linear problem.

Now, always analytic methods may not work; because of the complexity of the problem, because the problems are not convex or because of some other reason. So, we have some such techniques numerical such techniques to handle such problems. So, this lecture is basically devoted to numerical optimization algorithms; that what numerical optimization algorithm which are useful in machine learning.

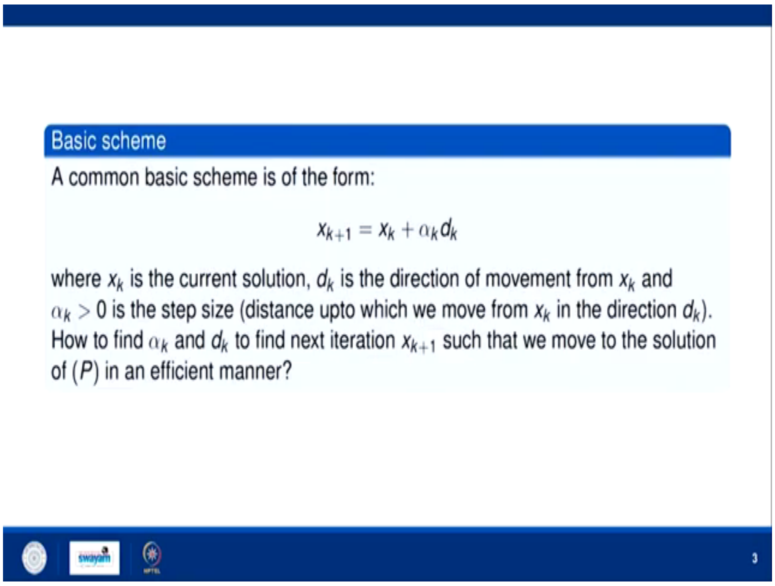
So, first let us understand; if you are having an unconstrained optimization problem. We have already discussed what unconstrained means. Unconstraint means; without any constraint without any restrictions.

(Refer Slide Time: 01:28)



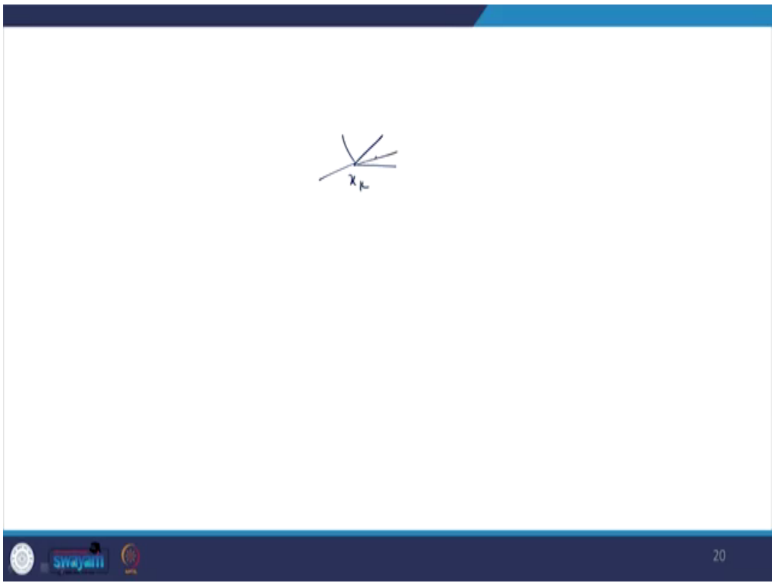
So; that means, we have to simply minimize f x subject to x belongs to R n. If we have this problem and we want to minimize it. So, so, the question arises; how can you find a point x bar in R n which solves the problem p or at least approximately solve the problem p. Because, in general our analytical approach may not work for all types of optimization problems. So, we move to such techniques or numerical optimization algorithms.

(Refer Slide Time: 02:01)



So, what is the basic scheme? See, the basic scheme of any algorithm of any numerical algorithm is; we have a initial we have a current point x k a direction d k from x k there are. So, many directions say if you are having a point x k.

(Refer Slide Time: 02:20)

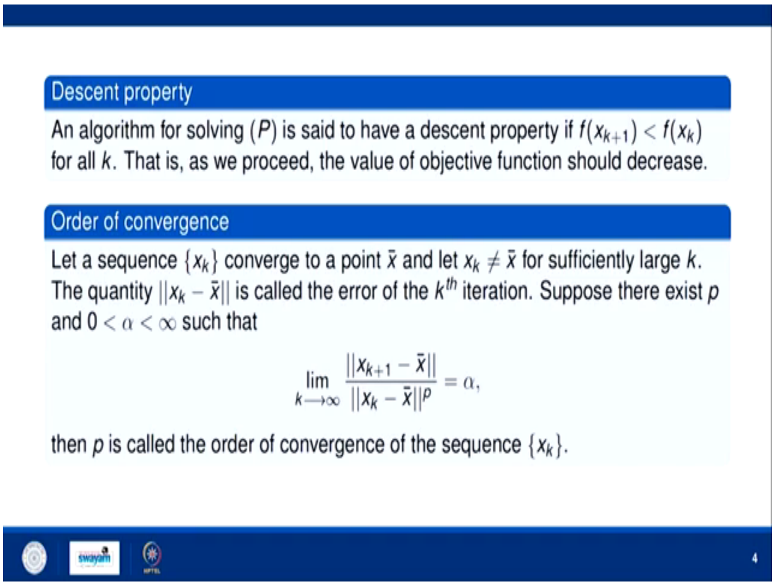


If you are having a point x k from this x k there are infinite directions in which direction we should move. So, that direction is basically d k the optimal direction. alpha k is a step size, optimal step size and using this relation x the next iteration x k plus 1 which is equal to x k plus alpha k d k the next iteration x k plus 1 can be found.

So, here x k is the current solution, d k is the direction of movement from x k and alpha k is a step size. The distance up to which we move from x k in the direction of d k to obtain x k plus 1 ok. Now how to find alpha k and dk? This alpha k and d k these are not known to us, only x k the current solution is known to us. How can you find alpha k and dk such that we get x k plus 1 from x k in an efficient way ok? So, this is the next question.

So, if we have having a minimization type problem as we have discussed here the minimization type problem.

(Refer Slide Time: 03:29)



Then we have descent property. What is a descent property? See, if we are moving from a current solution x k to x k plus 1 such that the value of f at x k plus 1 is less than value of f at x k; that means, we are going in a descent direction.

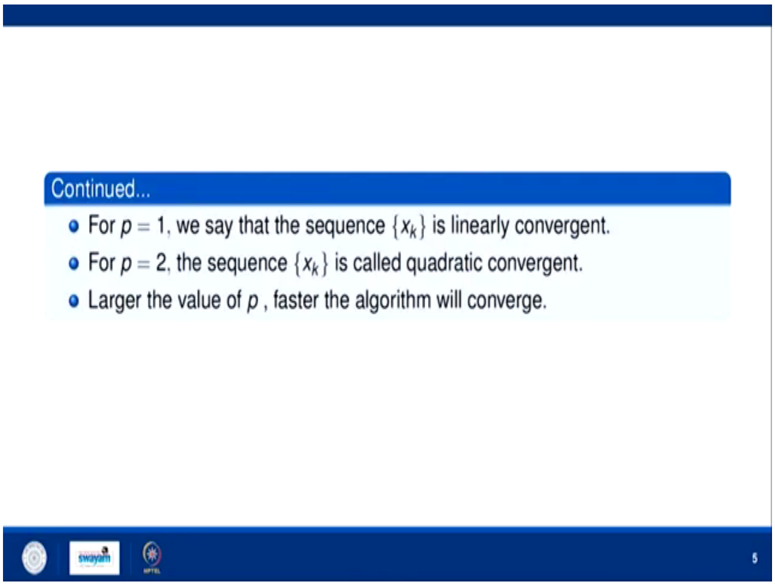
If it is a maximization type and f k plus f at x k plus 1 is more than f at x k, then we say that this is an ascent direction ok. So, an algorithm were solving a P is said to have a descent property if this inequality holds for every k and that is as we proceed the value of objective function decreases.

Now, if you are developing any algorithm any numerical algorithm. So, we have certain terms related to that algorithm. First is order of convergence. What is the order of convergence of that algorithm? So, how we define order of convergence? So, let a sequence x k converges to a point x bar ok.

And let x k is not equal to x bar for sufficiently large k. The quantity the norm of x k minus x bar is called the error of the kth iteration of course, because k x k converges to x bar and the norm of x k minus x bar because we have to anyhow approach to x k. So, this is nothing but the error term at the kth iteration.

Suppose there exists a p and alpha been between 0 and infinity such that norm of x k plus 1 minus x bar divided by norm of x k minus x bar whole raise to power p as limit k tends to infinity if it is alpha. Then p is called order of convergence of sequence x k; that means, this must be this value must be finite. If this value is finite for some p, then that p is called order of convergence of sequence x k ok.

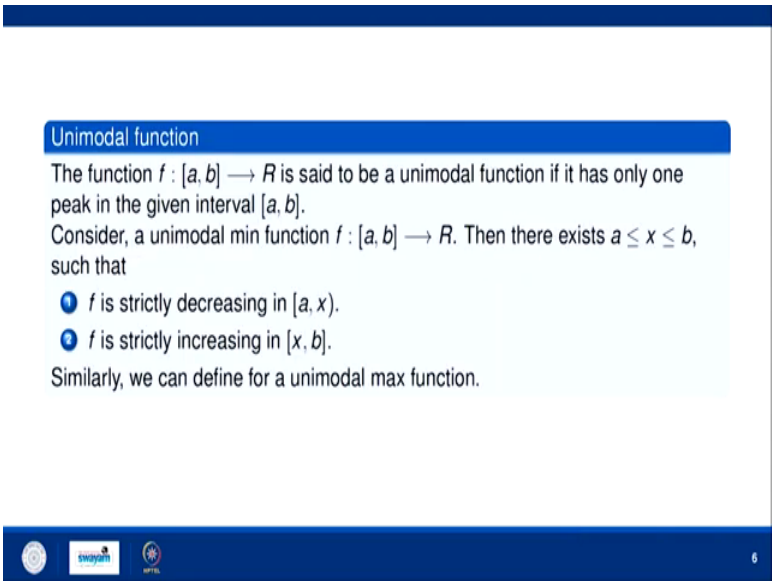
(Refer Slide Time: 05:40)



Now, if p equal to 1, if for some algorithm; the sequence x 1, x 2 x 3 up to x k and. So, on if that sequence converges for p equal to 1, then we say that the sequence is linearly convergent ok. If in this p comes out to be 1; that means, that sequences linearly convergent. If p equal to 2, then we say that sequence is quadratic convergent or convergent of order 2.

Of course a larger value of p, what indicates a larger value of p if you are having p 3 p 3.5 or something for some algorithms; that means, the algorithm is faster. The larger value of p indicate the faster the algorithm will converge ok.

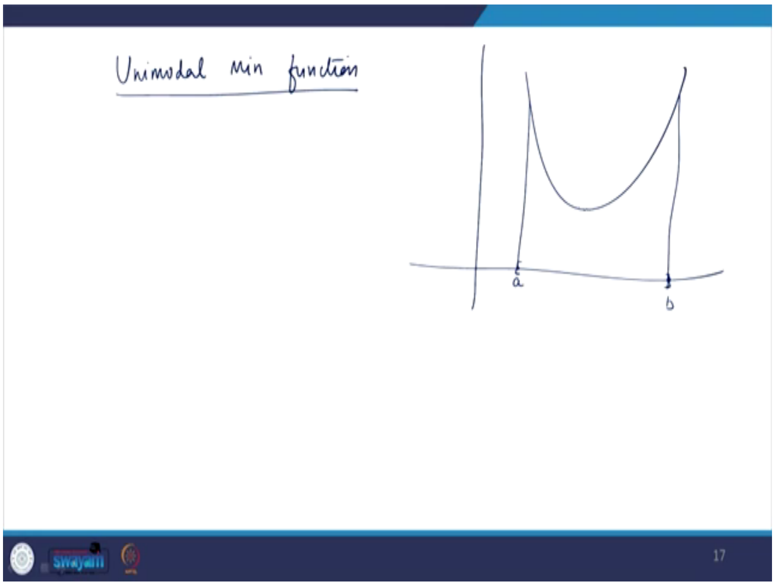
(Refer Slide Time: 06:29)



Now, we are having unimodal function. What do you mean by unimodal function? The function f from an interval a b. Now this initial interval is called interval of uncertainty. See, you are having an close interval a b and you are interested to find out point x which is the optimal solution of a given unconstrained optimization problem. So, this initial interval is called interval of uncertainty.

So, this problem the function from a b to r is said to be unimodal function if it has only one peak in the given interval a b; if it is only one peak either minima or maxima only one peak is called unimodal function. Let us suppose you are having a unimodal minimum function f from a b to r then, there exists x between a and b such that. See, if you are having a unimodal function, if you are having a unimodal minimum function.

(Refer Slide Time: 07:29)

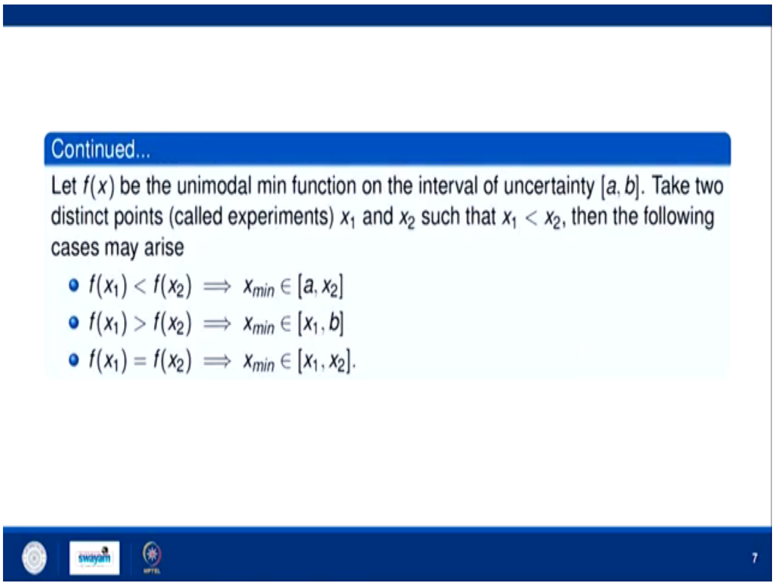


So, it may be of this type which is only one peak. So, it is unimodal minima and minimum also. So, it is unimodal minimum function.

So, if you are saying that this is I initial interval of uncertainty a b. If you are saying that these are this is a initial interval of uncertainty a b, then these are close interval a b. So, these points are also inclusive then; that means, that it first decreases up to some point and then increases, because it is a unimodal minimum function. So, that is what I have stated here that first of all there will be a x in between a and b and that access the minimum point where it first strictly decreases from a to x and then increases from x to b.

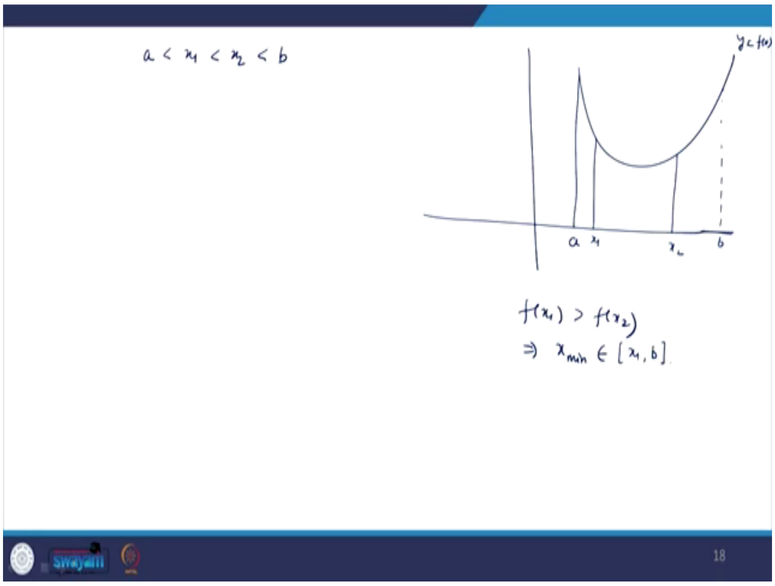
Now, similarly we can define unimodal minimum maximum function.

(Refer Slide Time: 08:38)



Now, in unimodal minimum function, the various cases may arise. So, let us discuss these cases now.

(Refer Slide Time: 08:44)

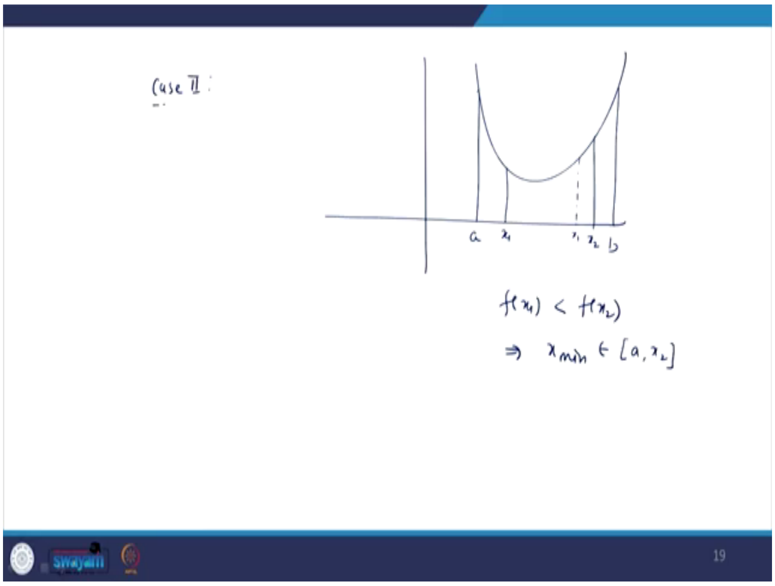


Suppose, you are having in between a and b, you are having two experiment x 1 and x 2. So, this is y equal to f x a function ok, this is a and this is some point b, this is the initial interval of uncertainty a b.

Now, you are having 2 experiments here. Now the two point x 1 x 2 may be any where in between a and b. And let us suppose x 1 is less than x 2 and between a and b. So, let us suppose x 1 is here and suppose x 2 is here. So, in this case in this case f x 1 is more than f x 2 because this height is more than this height.

So, where does maximum belongs to? It implies x minimum will belongs to which interval see this x minimum is somewhat here ok. Now this this x 2 may be here also this x 2 may be here also. So, we can say because this is f x 1 is less than fx 2. So, this point may be here also. So, we can say that this belongs to x 1 to b. Because x 1 cannot be here, if x 1 is here because x 1 is less than x 2 and if x 1 is here it is increasing. So, this when equality will not hold ok.

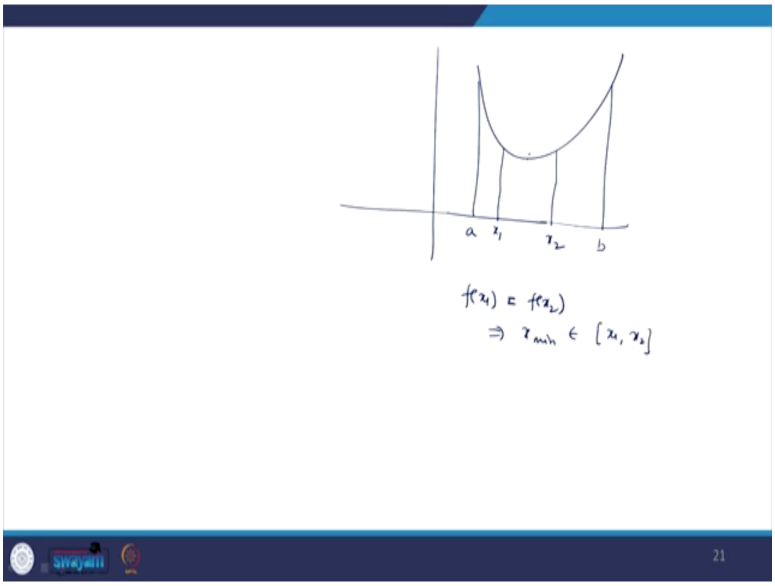
(Refer Slide Time: 10:26)



Now, the second case which may arise is, this is a this is b the second case which may arise is x 1 is somewhat here and x 2 is somewhat here; that means, fx 1 is less than fx 2. So, x minima belongs to in which interval? See, now x 1 is less than x 2. So, this x 1 may be here also; may be here also ok, but this x 2 cannot be here because x 1 is less than x 2.

So, we can say that minimum always belongs to a to x 2 in any case whether x 1 is here or here minima always belongs to a to x 2. So, these are second case 2.

(Refer Slide Time: 11:26)



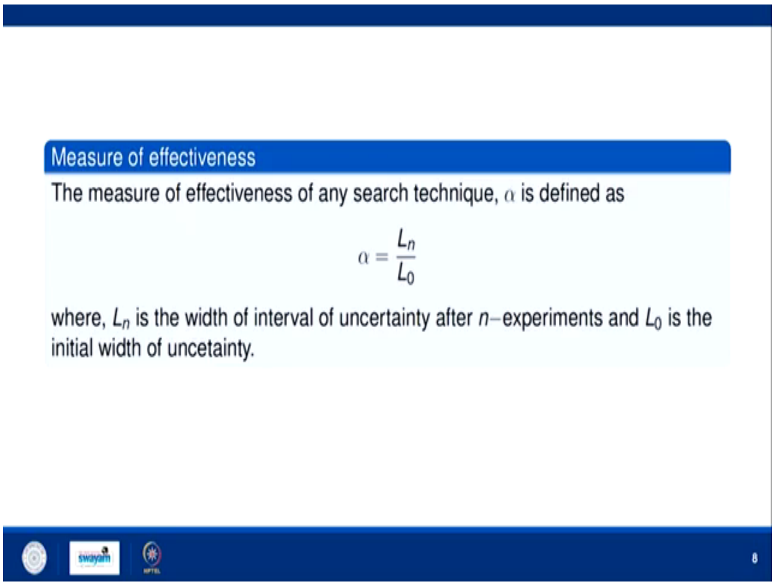
Now, what is the case 3? Case 3, it may be that you are having unimodal minimum function like this, this is a point a, this is a point b. And the 2 experiments which we have find which you have found x 1 and x 2, you are having 2 experiments such that f x 1 is equal to f x 2 ok.

So, you will be having only 3 cases; either f x 1 will be less than f x 2 or f x 1 is more than f x 2 or f x 1 is equal to f x 2 if it is a if it is equation then this implies x minimum will belongs to of course, in between x 1 and x 2, x minimum is here somewhat here. So, that is basically unimodal function. So, if we are dealing with a unimodal function and we are finding experiments by any method then either these three cases may arise.

So, if you have a unimodal function by any experiments you find experiments x 1 and x 2, then each time you are reducing the size of the experiment ok. See, if x 1 x 2 are like this that fx 1 is less than fx 2 then x minimum will belong to this then; that means, now the new interval of uncertainty is a to x 2 instead of a to b and then you will perform the same method on this interval.

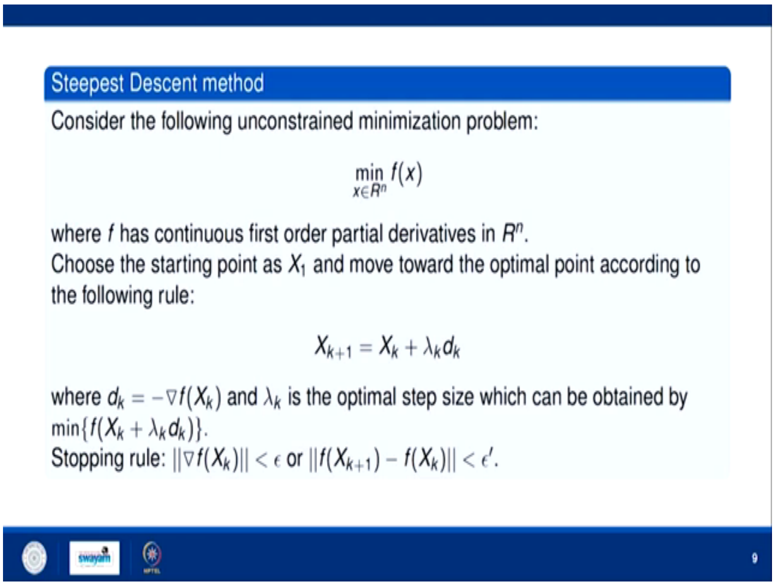
And if again it comes out to be this, then again the size of the interval will decrease and the process proceed; continue till you get the required accuracy.

(Refer Slide Time: 13:12)



Now, how we can find out measure of effectiveness of an algorithm? So, that will be a simply given by width of the nth interval of uncertainty divided by initial width of uncertainty. So, that is basically alpha and this alpha is always less than 1. So, this is how we can defined measure of effectiveness.

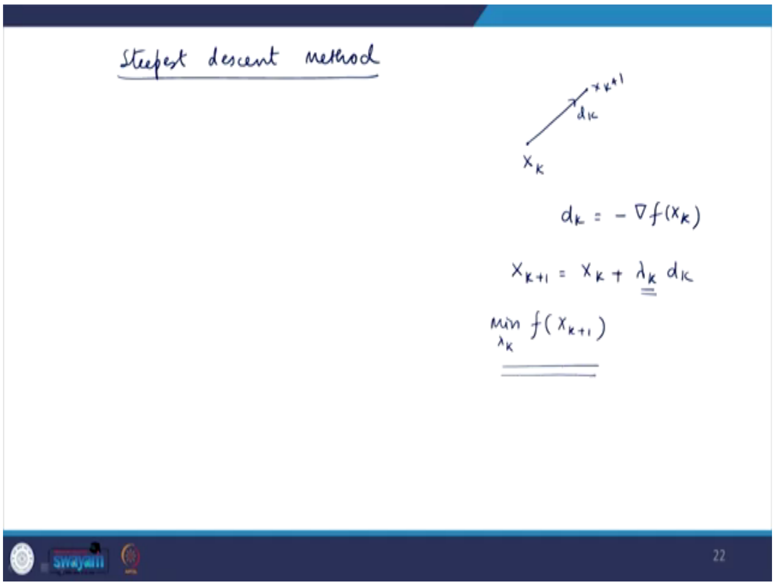
(Refer Slide Time: 13:35)



Now, let us come to the first method; which is steepest descent method. Now what is a steepest descent method? See suppose you are having an unconstrained optimization problem of this type minimization of fx x belongs to R n. Where f has continuous first order partial derivatives in R n ok. Now what is the numerical algorithm the algorithm is you first choose a starting point as x 1. We can take a point initial point as x 1 or initial guess as x 1 and move toward the optimal point according to the following rule.

So, rule is the same rule which we have discussed initially that; X k plus 1 is X k plus here alpha instead of alpha k you are having lambda k lambda k d k. Now d k is the direction and here is a minimization type problem. So, that direction must be must have descent property descent property means that f of X k plus 1 should be less than f at X k and lambda k is the optimal step size which can be obtained by this. So, how we will obtain this? Let us discuss.

(Refer Slide Time: 14:50)



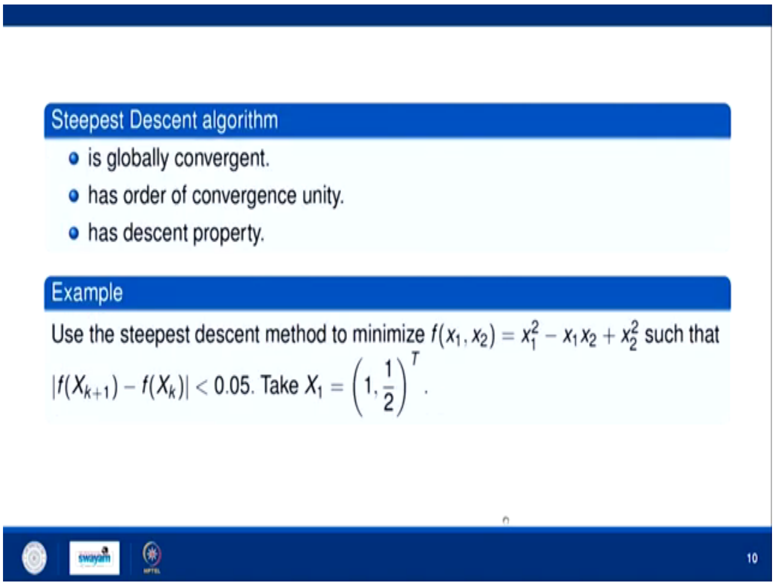
So, first of all first of all in a steepest descent method, this point is x k it may be x 1 x 2 or any one any point. From this x k as we already discussed there are in finite direction in which direction we should move that that is the first thing.

Now, it is a minimization type problem and we know that the rate of change of a function decreases most rapidly if we move along negative of gradient of f so; that means, if we move along negative of gradient of f at this point x k and take that as a descent direction. So, that will definitely the direction where the value of the f decreases.

Now, the question is how we can find out the optimal step size see we have find a direction d k which is nothing but negative of gradient of f at x k ok. And x k plus 1 is nothing but x k plus lambda k dk. Now this the alpha k is nothing, but this alpha k is see from here we have to reach a point x k plus 1 and this is the step size. So, how we can find out the optimal step size?

So, if we put this x k plus 1 this f of x k plus 1 and if you minimize this over lambda k then this will give the optimal step size ok. So, let us discuss it by an example first you will see few properties of this and what is the stopping rule up to how much we have to perform this iterations. We first start with x 1 then find x 2 x 3 and so on. So, the stopping rule is either norm of gradient of x f x k should be less than epsilon is the tolerance level which will be given to you or the difference of the 2 successive algorithm 2 successive iterations, the value of f and a 2 sub save iterations the norm of this should be less than epsilon dash that is the stopping rule.

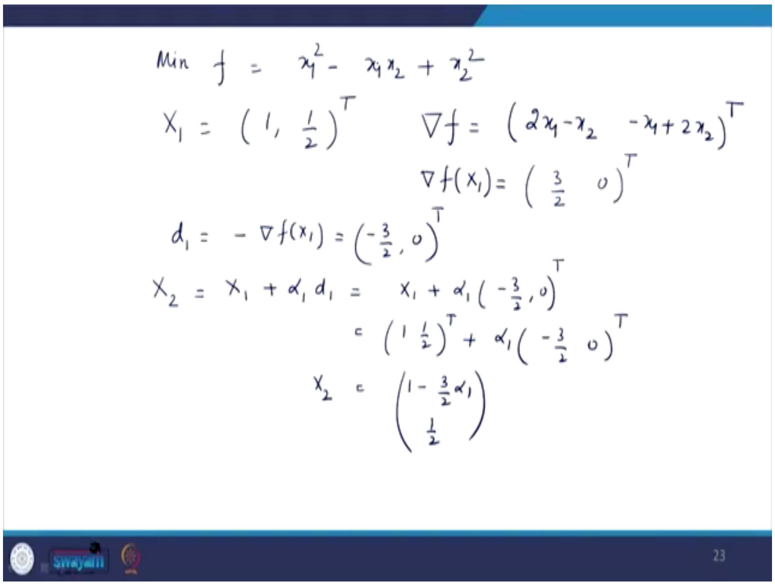
(Refer Slide Time: 17:14)



Now, what are properties of this method. The first property is, it is globally convergent ok. The second property is it has order of convergence unity; that is p equal to 1 for this case and it has a descent property. Of course, it a descent property it; that means, f of X k plus 1 is less than f X k for all k. So, that is the important property and it is very easy to understand, easy to apply that is why we are using this algorithm in machine learning.

So, let us discuss this method by a example. Suppose you want to minimize this function x 1 square minus x 1 x 2 plus x 2 square such that we want at the difference of two consecutive values should be less than 0.05 and it is given to us that initial guess as 1 and half.

(Refer Slide Time: 18:14)



So, what is the problem? Problem is minimization of f which is equal to x 1 is square minus x 1 x 2 plus x 2 square this is the problem ok.

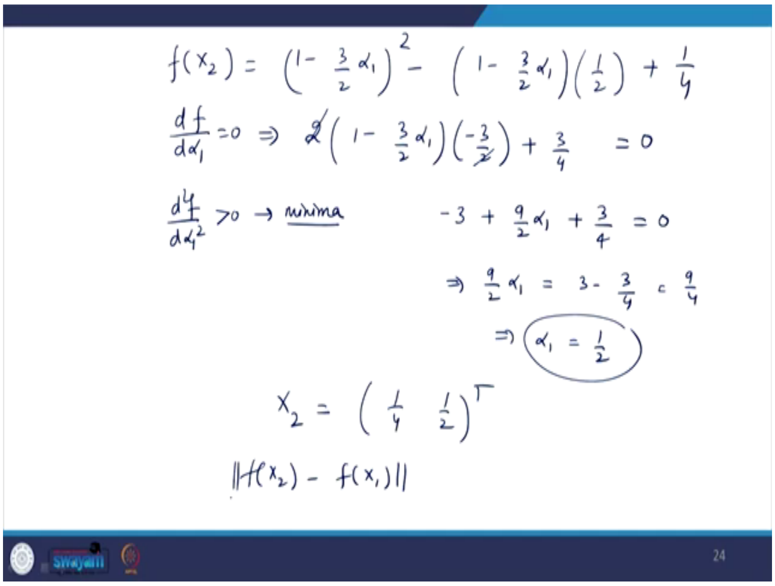
So, what is the initial guess? Initial guess is 1 comma 1 by 2 half. So, first let us find gradient of f. So, what is gradient of f? Grad of is 2 x 1 minus x 2 here it is minus x 1 plus 2 x 2 whole transpose. So, this will be the gradient of f.

Now, what is gradient of f at x 1 at initial guess as x 1. So, put x 1 equal to 1 and x 2 equal to half. So, 2 minus half is 3 by 2 and it is 1 by 2 so, that is 0. So, it is 0 transpose. So, this will be gradient of f. Now what will be d 1 d 1 is nothing, but negative of gradient of f x 1 and that will be nothing but minus 3 by 2 and 0 transpose ok.

So, the x 2 from the steepest descent method by the recursive algorithm that will be nothing but x 1 plus alpha 1 or lambda 1 d 1, that will be nothing but x 1 plus alpha 1 times minus 3 by 2 comma 0. So, that is this transpose. So, what is x 1? x 1 is 1 1 by 2 transpose plus alpha 1 minus 3 by 2 0 whole transpose. So, that will be 1 minus 3 by 2 alpha 1 and 1 by 2. So, that will be x 2.

Now, how to find x 2? You simply substitute this point in this function and now it will be a function of single variable alpha 1 and try to minimize that function. We have to find out alpha 1 for which that f, f is minimum. So, simply substitute this over here.

(Refer Slide Time: 20:29)



So, let us try to find out. So, what is fx 2? fx 2 will be fx 2 is you replace here you replace x 1 by 1 minus 3 by 2 alpha 1. So, you replace x 1 by 1 minus 3 by 2 alpha 1 whole square minus 1 minus 3 by 2 alpha 1 and x 2 is what? x 2 is half. So, it is half and plus 1 by 4.

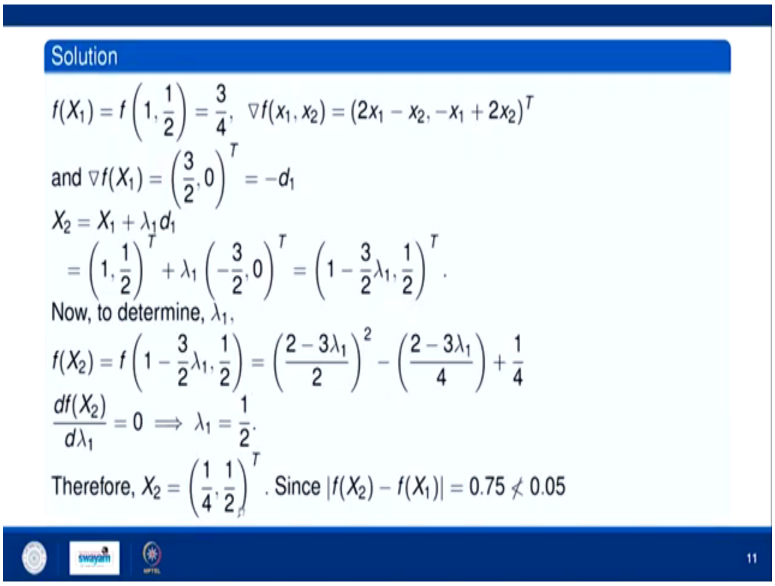
So, now it is a function of single variable alpha 1. So, your differentiate with respect to alpha 1 and put it equal to 0 for maxima and minima. So, it will be 2 times 1 minus 3 by 2 alpha 1 minus 3 by 2 minus minus plus 3 by 4 and put it equal to 0. Second derivative is of course, second derivative here is positive; that means, minima. Second derivative with respect to alpha 1 is positive; that means, minima ok.

So, let us compute alpha 1 from here. So, 2 2 cancels out. So, it is nothing but minus 3 and it is plus 9 by 2 alpha 1 plus 3 by 4 equal to 0. So, that will be this implies 9 by 2 alpha 1 should be equal to 3 minus 3 by 4. So, it is 12 minus 3 that is 9. So, which is 9 by 4 alpha 1. So, this 9 by 4. So, this is simply 9 by 4 and this imply alpha 1 is 1 by 2.

So, in this way we obtain alpha 1. So, what will be x 1 x 2 now? So, x 2 is nothing but, now I substitute alpha 1 here alpha 1 is what? 1 by 4. So, 1 by 4; that means, it is 1 by 4 and half. So, it is 1 by 4 and half whole transpose. So, this is x 2. Now we will compute f x 2 minus f x 1 and it is norm or modulus.

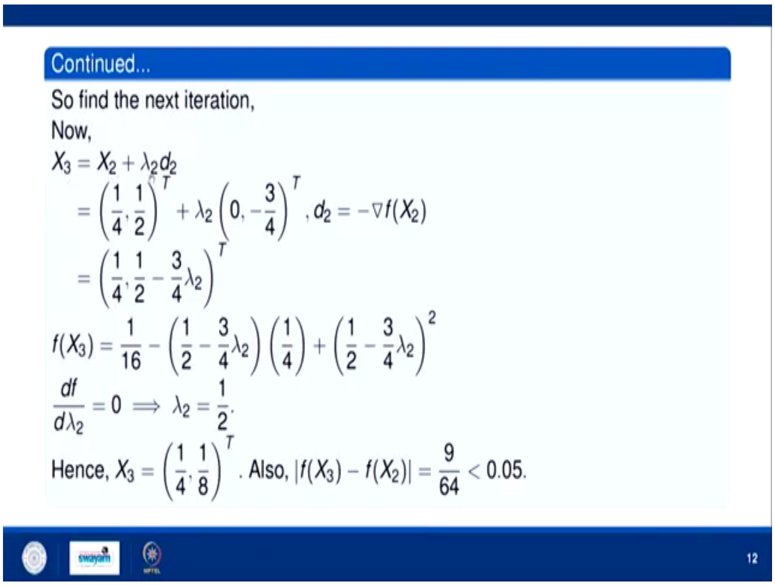
We will see whether this value is less than 0.05 or not, because that is a stopping rule here, that is a stopping rule here. If this is come out to be less than 0.05; that means, we will take x 2 as we will take x 2 as the approximate solution of this given problem. So, what is what is this value?

(Refer Slide Time: 23:06)



So, now we will see here. So, we have we have find I have shown for X 2 and the you can you can verify that f X 2 minus f X 1 is 0.75 which is not less than 0.05. So, hence this means we have to proceed further.

(Refer Slide Time: 23:25)



So, in the same way we will find X 3 now. Now X 3 will be given as X 2 plus alpha 2 d 2 alpha 2 or lambda 2, both are both are ok. So, lambda 2 d 2, X 2 is what? X 2 from here is 1 by 4 1 by 2. So, you can substitute X 2 as 1 by 4 1 by 2 lambda 2 is lambda 2 and d 2; how you will find d 2? d 2 is nothing but negative of gradient of f at X 2.

So, gradient of f we have already computed, gradient of f is this. So, you have to find now gradient of f at X 2. X 2 is 1 by 2 1 by 1 by 4 1 by 2. So, you substitute X 1 as 1 by 4 and X 2 is 1 by 2 and find out gradient of f at X 2 at this point. So, we will obtain this ok. Now again we substitute this X 3 in the function and try to minimize try to find out that lambda 2 for which, for which f is minimum and after solving again you will get lambda 2 as half.

So, when you substitute lambda 2 equal to half here we will get X 3 as 1 by 4 1 by 8 ok. And now if you find out the mod of f X 3 minus f X 4, it comes out to be 9 by 64, which is less than 0.05. Now we will stop here and we will say that X 3 is the approximate optimal solution of the given problem.

So, in this way we have seen that; if we are having a non-linear unconstrained optimization problem, then using a steepest descent method we can find out at least approximate optimal solution of a given problem. In the next lecture we will see some more numerical such techniques for solving constrained optimization problems.

Thank you.