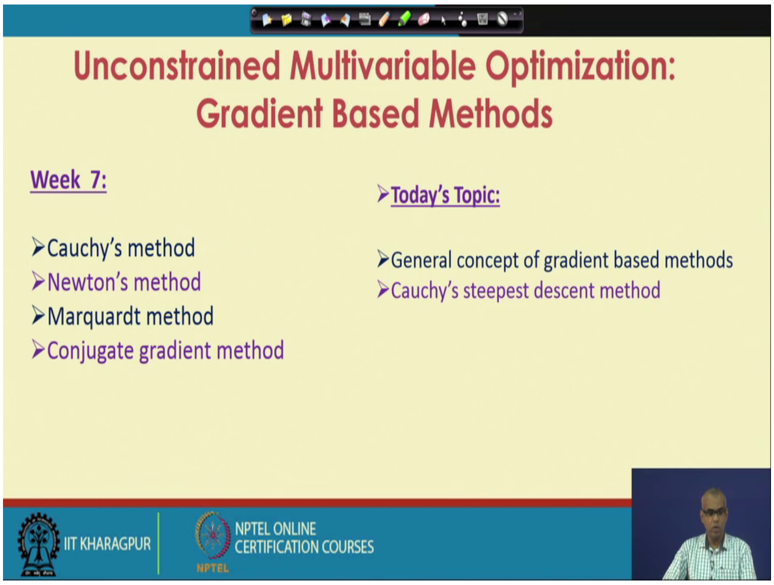
Optimization in Chemical Engineering
Prof. Debasis Sarkar
Department of Chemical Engineering
Indian Institute of Technology, Kharagpur

Lecture – 31
Unconstrained Multivariable Optimization: Gradient Based Methods

Welcome to lecture 31. This is the first lecture of week 7. In week 6 we have discussed direct search methods for Unconstrained Multivariable Optimization. In this week 7 we will talk about gradient based methods for Unconstrained Multivariable Optimization. In case of direct search methods you have seen that the optimization methods do not use information on gradient, they were only on the information of value of the objective functions.

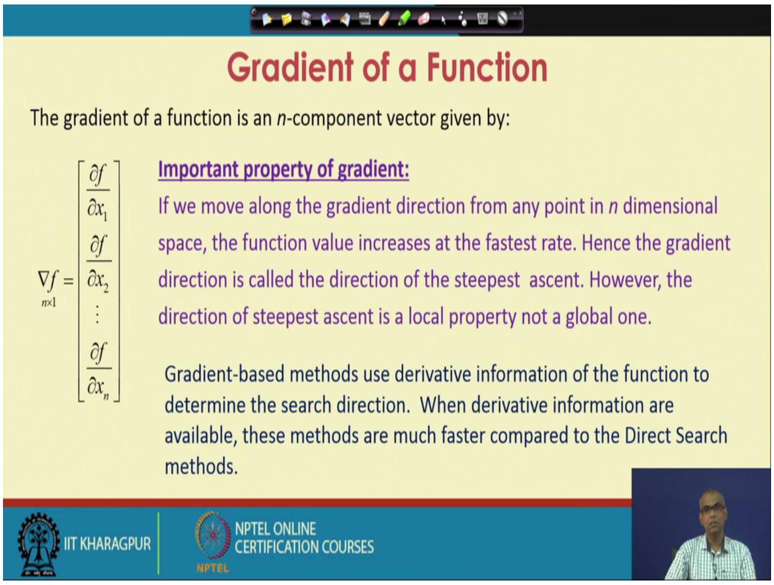
The gradient method based methods make use of informations on gradients, and when this gradient information is available the gradient mesh methods are much faster in terms of convergence compared to direct search methods. There are some methods which makes use of only gradient information, there are some methods which makes use of gradient information as well as hessian information that means, first partial derivative as well as second derivatives are required. Accordingly we call them first order methods or second order methods.

(Refer Slide Time: 01:16)



So, what we will plan to cover in this week is Cauchy's method, Newton's method, Marquardt method and Conjugate gradient method. So, today after giving a general concept of gradient based methods, we will talk about Cauchy's steepest descent method. So, Cauchy's method is also known as steepest descent method.

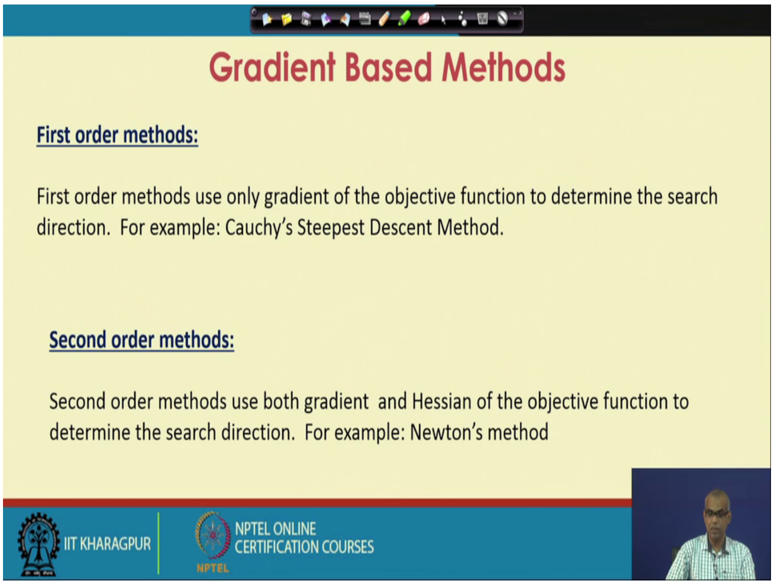
(Refer Slide Time: 02:16)



A quick review of gradient of a function, the gradient of a function is an n component vector given by del f del x 1, del f del x 2, del f del x 3 up to del f del x n. So, given an invariable function or objective function you have to take the partial derivative of the objective function with respect to each variable, and the vector thus obtained is the gradient of the function.

Now, there is an important property of gradient if we move along the gradient direction from any point in n dimensional space the function value increases at the fastest rate. So, gradient always points to the direction in which objective function value is changing at the fastest rate. Hence the gradient direction is called the direction of the steepest ascent however; the direction of steepest ascent is the local property and not a global property. Gradient based methods use derivative information of the function to determine the search direction, when derivative informations are available these methods are much faster compared to the direct search methods.

(Refer Slide Time: 04:02)

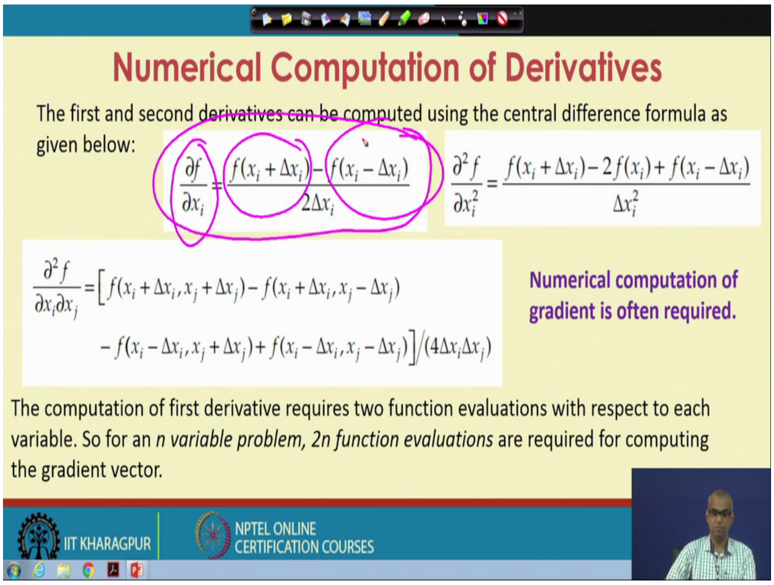


So, you can classify the gradient based methods as first order methods or second order methods. First order methods use only gradient of the objective function to determine the search direction for example, Cauchy's steepest descent method.

Second order methods use both gradient and hessian of the objective function to determine the search direction. For example, Newton's method is a second order method. So, the overall classification of the unconstrained multivariable optimization methods are direct search methods and gradient based methods. Sometimes you also call gradient based methods as indirect methods. So, all the techniques available for unconstrained multivariable optimization that we have discussed in this course are categorized into two. For example, direct search methods and indirect search methods or gradient based methods.

The gradient based methods are again categorized as first order methods and second order methods. First order methods use only gradient information, second order methods use gradient information as well as hessian information.

(Refer Slide Time: 05:43)

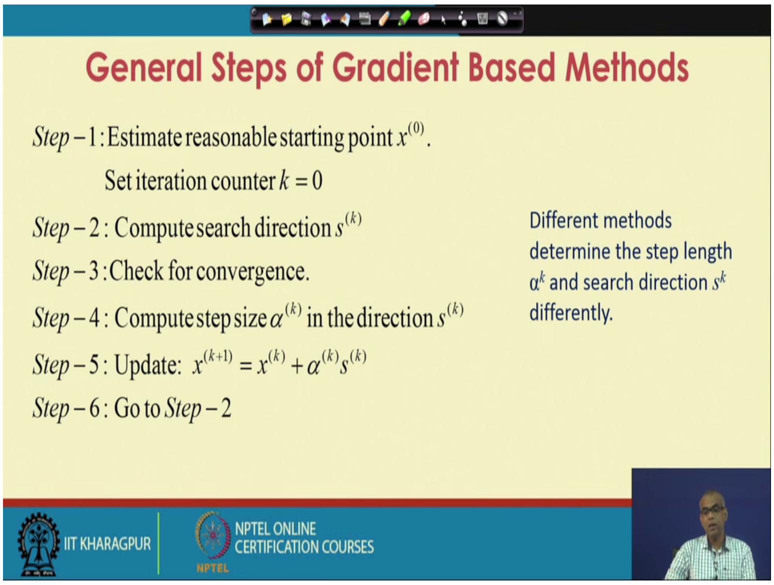


Now, it may not be always possible to have analytical derivatives of the objective function, and you often have to use numerical methods to compute gradient of the objective function. If the objective function is a simple mathematical expression most likely we will be able to compute the derivatives and gradients analytically, but suppose the value of the objective function is available to you after doing some simulations. The simulations may involve the solution of model equations it may be a large number of set of equations and after the simulations of all the equations you are able to compute the objective function value.

Analytical computation of the gradient it is not possible under such circumstances, and under such circumstances you have to make use of numerical computation of gradient. So, the slide shows you how you can compute first partial derivative as well as second partial derivatives of a function using central difference formula. There are several other formulas also to compute numerically, these first partial derivatives as well as second partial derivatives what is show here is central difference formula.

Note, del f del x i equal to f x i plus delta x i minus f x i minus delta x i by 2 delta x i. So, we can realize that for computation of first derivative it is necessary that we evaluate the function twice. So, the computation of first derivative requires two function evaluations with respect to each variable. So, with respect to each variable we need to evaluate the objective function twice. So, for n variable objective functions we will require 2 n function evaluations for computing the gradient vector.

(Refer Slide Time: 09:34)

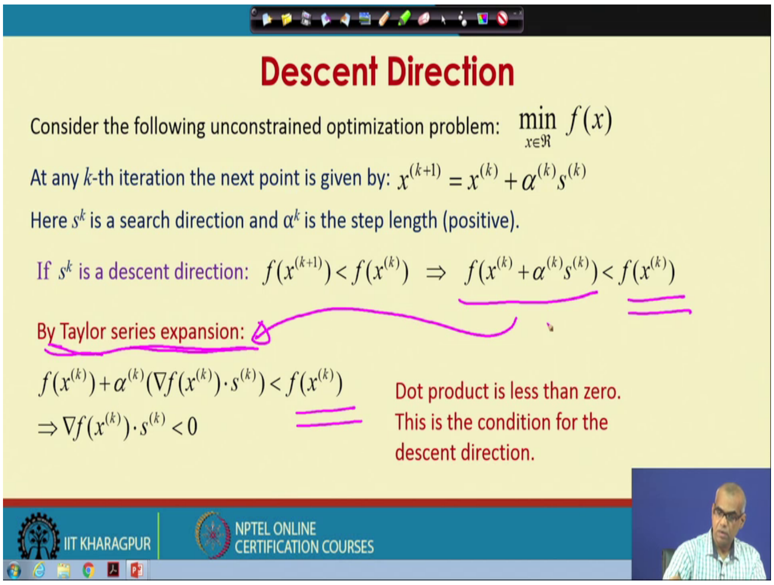


Now, here is the general step of any gradient based method. So, all the gradient based methods can be represented by a sequence of general steps as shown on the slides. So, in the first step we start with a reasonable estimate of optimal point, say I am minimizing a function. So, I start with a reasonable guess of the minimum point that is x 0 for me that is my starting point. So, we set the iteration counter k equal to 0.

In step 2, we compute search direction s. So, we need to find out the direction in which x will change. Step 3 we check for convergence if convergence is achieved we stop else we go to step 4. In step 4 we compute step size alpha in the direction s. Once I have the step size in the direction s I can update x as, x k plus 1 equal to x k plus alpha k s k that means, the improved guess for the minimum point is equal to the current estimate of the minimum point plus step size into the direction vector.

After this you again go to step 2 where we further find the search direction for the next step, then we proceed until we reach convergence. Now, this is a general steps for all gradient based methods. So, we have different methods which makes use of gradient information, they differ in determining the step length alpha and the search direction s. So, different methods determine the step length, alpha and the direction s differently.

(Refer Slide Time: 13:07)

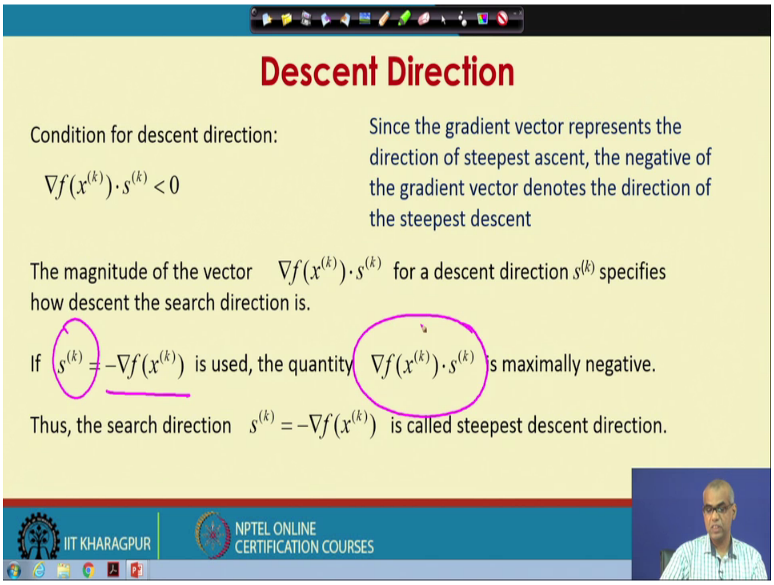


We have talked about descent direction before let us quickly review one more time what do you understand by descent direction. Consider an unconstrained optimization problem minimization of f x, x can take any real value. At any kth iteration the next point that means, at iteration number k plus 1 is obtained as x k plus 1 equal to x k plus alpha k into s k, where s k is the search direction and alpha k is the step length and we choose a positive step length.

If the direction s k is a descent direction the function value at k plus 1th iteration will be less than the function value at kth iteration. So, if s k is descent direction f of x k plus 1 must be less than f of x k. We know x k plus 1 is x k plus alpha k s k. So, you can substitute this here and we write as f of x k plus alpha k s k is less than f of s k, if s k is a descent direction. Now, I can perform a Taylor series expansion for f of x k plus alpha k s k and if I retain only the first order terms I write f of s k plus alpha k gradient of f at x k into s k that must be less then f of s k which is same as this.

So, this part comes from Taylor series expansion of this. Now, look at this expression you have on both sides the function value at x k. So, this inequality says you must have gradient f at x k into s k must be less than 0. Note that alpha the step length is positive. So, if s k is the descent direction the dot product of the function at x k with the s k must be less than 0.

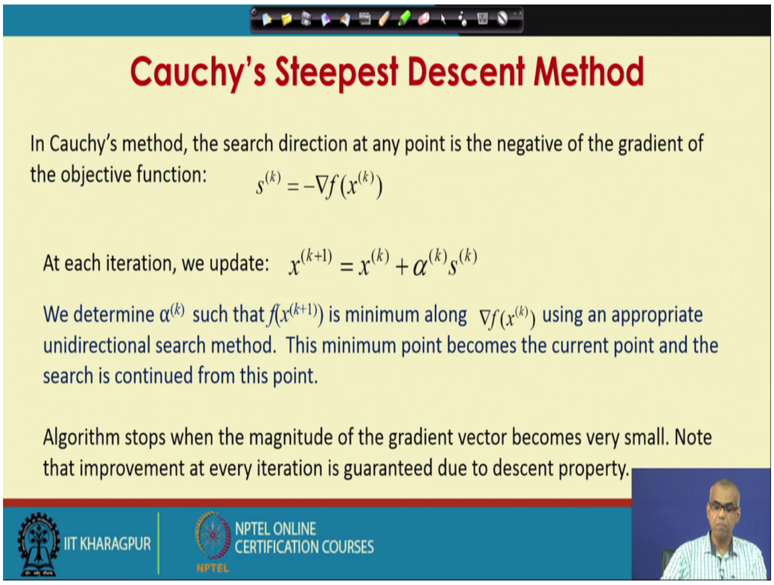
(Refer Slide Time: 16:42)



Now, since the gradient vector represents the direction of steepest ascent the negative of the gradient vector denotes the direction of steepest descent. So, we discussed few slides back that the gradient vector always points to the direction in which the function value is changing at the fastest rate. So, that is why it is known as direction of steepest ascent.

So, now, if I take the negative of the gradient vector that must point to the direction in which the function value is decreasing at the fastest rate. So, the direction of negative gradient is the direction of steepest descent now. The magnitude of the vector gradient of f dot s for a descent direction s specifies how descent the search direction is if the magnitude is high is more decent. Now, if I said the search direction s as negative of the gradient of the objective function the dot product of the gradient with the search direction s will be maximally negative thus the search direction s equal to minus of gradient is called steepest descent direction.

(Refer Slide Time: 18:47)



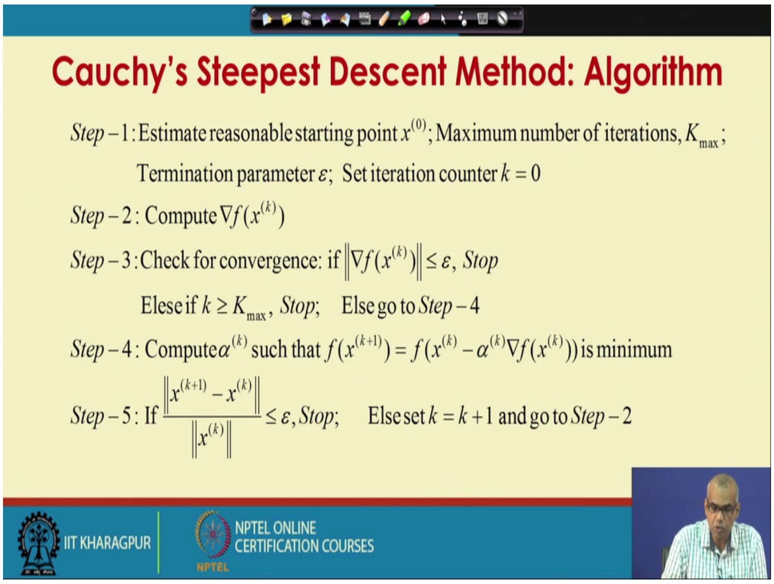
Now, let us talk about Cauchy's steepest descent method. So, this is the first gradient based method that we will discuss. It is one of the simplest method, and this method is very efficient when I start my search away from the minimum point. But as we get closer and closer to the minimum point the rate of convergence decreases. So, as we discussed that all the gradient based methods differ in determination of the search direction and in the determination of step length, otherwise the general structure is more or less same for all the gradient based methods. So, in case of Cauchy's steepest descent method we take the negative of the gradient as the search direction.

So, the search direction at any point is the negative of the gradient of the objective function. So, search direction at any point is the negative of the gradient of the objective function. At each iteration we will update the guess values for minimum using usual expression x k plus 1 equal to x k plus alpha k s k, where s k is set as negative of the gradient of the objective function at that point. Now, how to determine the step size alpha? We determine alpha such that f of x k plus 1 is minimum along gradient direction using an appropriate unidirectional search method.

This minimum point becomes the current point and the search is continued from this point. So, alpha k is determined such that f of x k plus 1 is minimum in the search direction and we can find this out using an appropriate unidirectional search method. So, once this minimum point along the search direction is obtained the minimum point becomes the current point and the search is continued from this point. So, what you do is at this point again you find out the gradient and update using the usual equation x k plus 1 equal to x k plus alpha k s k.

The algorithm will stop when the magnitude of the gradient vector becomes very small or there is no improvement in the objective function value. Note that improvement at every iteration is guaranteed this is because we are talking about a decent method. So, when the search direction is a descent direction the objective function value will keep on decreasing with each iteration, but when I say that we can stop the algorithm when there is no appreciable change in the objective function value. As long as we are using decent direction there will be changes in the objective function value the objective function value will keep on decreasing. But the amount by which it will decrease will be extremely small when the magnitude of the gradient becomes very very small and that is what happens when we approach the minimum point.

(Refer Slide Time: 23:48)

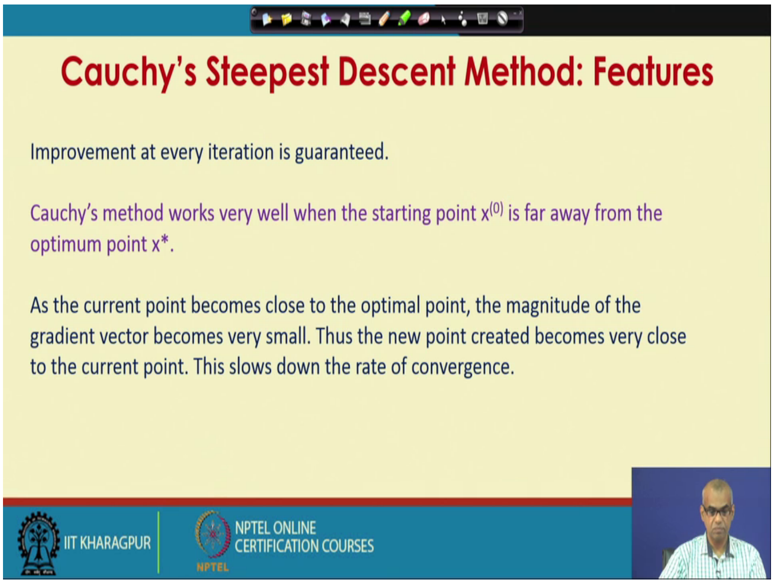


So, this is the algorithm for steepest descent method. In the step 1 we start with a reasonable guess of the minimum point we call it x 0, set the maximum number of iterations k max, set the termination parameter epsilon a small value, set the iteration counter k equal to 0.

Compute the gradient of the objective function at the given point let us say I am at k th iteration. So, this x k we check for convergence by checking if the magnitude of the gradient vector is less than the tolerance then the small termination parameter epsilon or not. So, the gradient vector is very small we stop, otherwise else you check whether the number of iteration has exceeded the maximum allowable number of iterations k max or not. So, if the number of iterations exceeds the maximum allowable number of iterations we stop, otherwise we go to next step which is step 4.

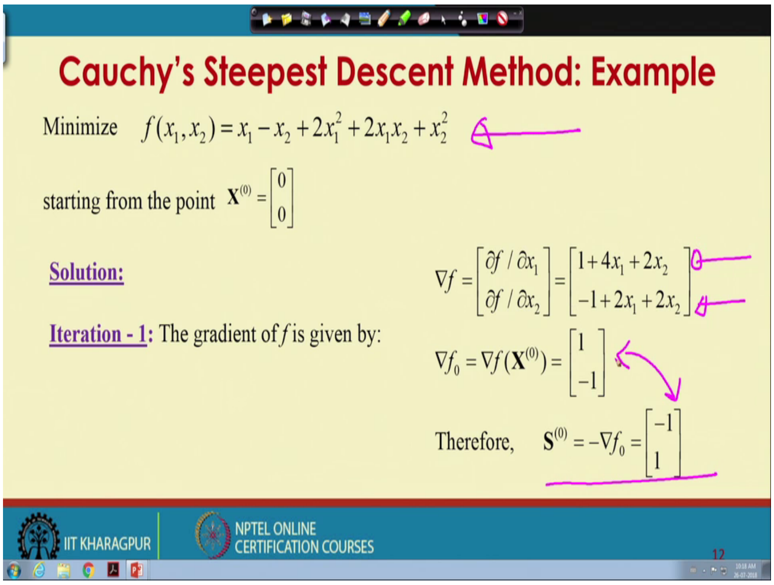
In step 4, we compute the search direction alpha k such that f of x k plus 1 which is f of x k minus alpha k and gradient of f at x k is minimum and this you can do using any unidirectional search technique. Now, again we can check if the guess value is not changing much we can stop, otherwise we set k equal to k plus 1 and go to step 2 where for the new or for the current estimate of the minimum point we again compute the gradient and continue with the procedure.

(Refer Slide Time: 26:29)



These are some interesting features of Cauchy's steepest descent method, improvement at every iteration is guaranteed because we are talking about a decent method. Cauchy's method works well when the starting point x 0 is far away from the optimum point. As the current point becomes close to the optimal point the magnitude of the gradient vector becomes very small. Thus, the new point created becomes very close to the current point. This slows down the rate of convergence.

(Refer Slide Time: 27:11)

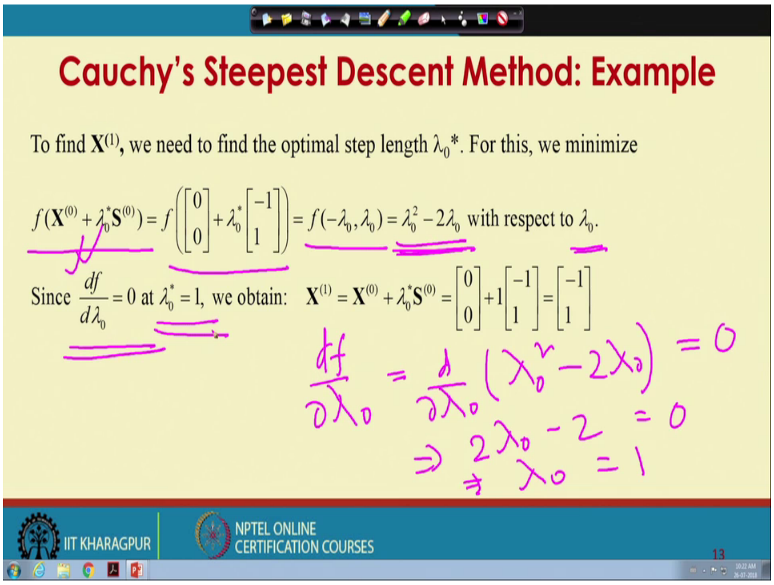


Now, let us take an example to understand how Cauchy's steepest descent method works. So, we have taken a two variable function x 1 minus x 2 plus 2 x 1 square plus 2 x 1, x 2 plus x 2 square we want to minimize it starting from the point 0 0. So, x 1 equal to 0, x 2 equal to 0.

So, let us start iteration number 1. At this step you can define the termination parameter epsilon which is a very small quantity may be 10 to the power minus 4, the starting point is given you can also define the maximum number of iterations that we wish to perform. So, next we calculate the gradient of the objective function at the given point 0 0. So, this is the objective function.

So, the gradient vector will be del f del x 1 del f del x 2. So, there computed as 1 plus 4 x 1 plus 2 x 2 which is del f del x 1 and minus 1 plus 2 x 1 plus 2 x 2 which is del f del x 2. So, find out the value of the objective function at 0 0 so set x 1 equal to 0 x 2 equal to 0 and you get 1 minus 1 as the gradient vector. So, negative of this vector will be considered as the search direction. So, s 0 is minus del f 0 is minus 1 1; compare this two. So, I have the search direction.

(Refer Slide Time: 29:36)

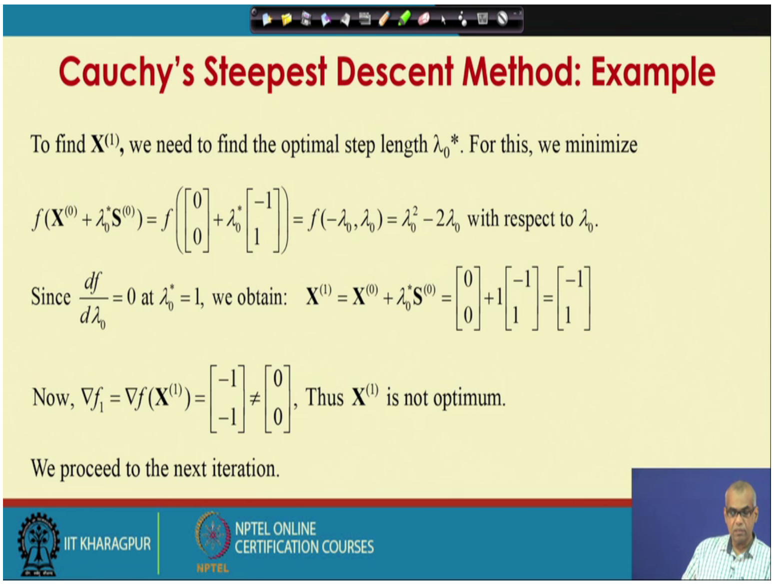


So, once I have the search direction I can find out the estimate of the minimum in the next iteration that is iteration number 1. So, I can find out x 1. So, how do I found out X 1? X 1 will be found out as X 0 plus step length into S 0. So, this is our scheme to update the estimate for the minimum point.

So, this is same as x k plus 1 equal to x k plus alpha k s k. So, that alpha which is the step length has to be found out and we need to find out the optimal step length. Let us call we call the step length lambda, the optimal step length let us call lambda 0 star. So, what we do is we have to minimize the function f X 0 plus lambda 0 star S 0. So, this function becomes this put X 0 equal to 0 0, lambda 0 star and s 0 is minus 1 1. So, this becomes minus lambda 0 lambda 0.

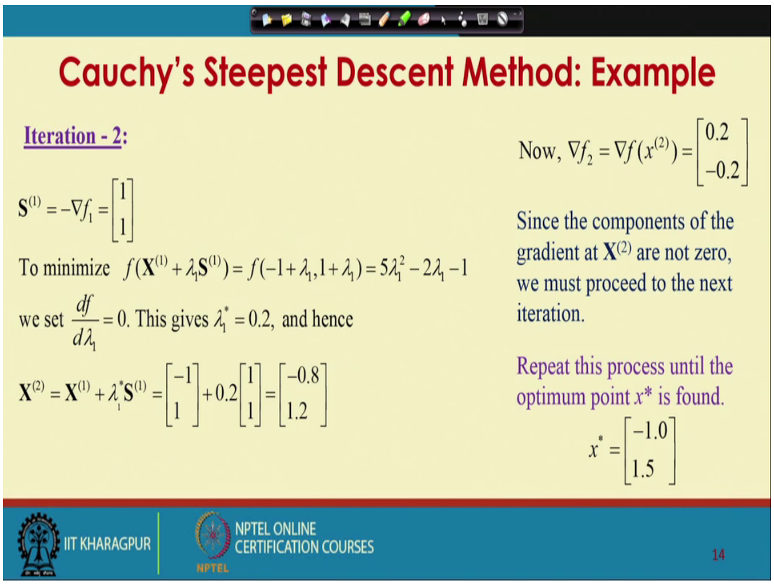
So, after putting X 1 equal to minus lambda 0 X 2 equal to lambda 0 in the given objective function we get f equal to or the function equal to lambda 0 square minus 2 lambda 0. So, this has to be made minimum with respect to lambda 0. So, by for that what we have to do is I have to take the derivative of the function with respect to lambda 0 and set that equal to 0. So, d f d lambda 0 is d lambda 0 and f at lambda is minus lambda 0 lambda 0 is nothing, but lambda 0 square minus 2 lambda 0. So, this will be equal to 0. So, this gives you 2 lambda 0 minus 2 equal to 0 which gives you lambda 0 equal to 1. So, the optimal value of lambda 0 which is being represented at lambda 0 star is equal to 1. So, once you have this we can find out X 1 easily, X 1 equal to X 0 plus lambda 0 star is 0 so 0 0, plus lambda 0 star is obtained as 1, S 0 was minus 1 1. So, X 1 is obtained as minus 1 1.

(Refer Slide Time: 33:38)



So, now, let us take the gradient of the function at this value X 1 minus 1 1, we see that the gradient evaluated at X 1 is minus 1 minus 1. So, it is not 0 0. So, thus X 1 is not optimum and we must proceed to the next iteration.

(Refer Slide Time: 34:14)



So, in the iteration 2 we will set the search direction as negative of the gradient at X 1. So, note the gradient was the gradient at X 1 is minus 1 minus 1. So, the search direction will be minus of this gradient so which will be 1 1.

Again I have to minimize f of X 1 plus lambda 1, S of 1 which becomes equivalent to minimizing 5 lambda 1 square minus 2 lambda 1 minus 1 with respect to lambda 1. So, again take the derivative with respect to lambda 1 set that equal to 0 and you get lambda 1 star as 0.2 the optimal value of the step length. Once I have that I can update X as X 2 equal to X 1 plus lambda 1 star into S 1. So, put X 1 which is minus 1 1 lambda 1 star is obtained as 0.2 and S 1 is 1 1. So, X 2 is obtained as minus 0.8, 1.2.

Now, again evaluate the gradient at this point X 2 and we check that the gradient is 0.2 minus 0.2. Note that the gradient is not 0. So, you must proceed to the next iteration, you remember at stationary point, the gradient vector will be 0 the first order necessary condition for optimality. So, gradient is not 0 neither very very small. So, you repeat this process go to iteration number 3 and check if the gradient is 0 or very close to 0. So, you need to repeat this process until the optimal point minus 1, 1.5 is obtained. But this two iterations shows you how to compute in each steps. So, each step is repetition of these steps.

(Refer Slide Time: 37:02)

With this we stop lecture 31 here.