

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

**Lecture – 32**

**Unconstrained Multivariable Optimization: Gradient Based Methods (Contd.)**

Welcome to lecture 32. This is week 7 and we are talking about Gradient Based Methods for Unconstrained Multivariable Optimization. In the previous week, we have talked about Cauchy's Steepest Descent Method. In this week, we will talk about another gradient based method name as Fletcher Reeves Conjugate Gradient Method, but before that we will make some comments about Cauchy's Steepest Descent Method.

(Refer Slide Time: 00:56)

**Cauchy's Steepest Descent Method**

The steepest descent method zigzags its way towards the optimum point.

Each step is orthogonal to previous step.

Update rule:  $x^{(k+1)} = x^{(k)} + \alpha^{(k)} s^{(k)}$  ✓

$\alpha^{(k)}$  is obtained by minimizing  $f(x^{(k)} + \alpha^{(k)} s^{(k)})$ . We set  $\frac{df}{d\alpha} = 0$

Upon differentiation:  $\nabla f(x^{(k)} + \alpha^{(k)} s^{(k)})^T s^{(k)} = 0$

$\Rightarrow \nabla f(x^{(k+1)})^T s^{(k)} = 0$

Since  $s^{(k+1)} = -\nabla f(x^{(k+1)})$ , we get  $(s^{(k+1)})^T s^{(k)} = 0$ .

This result is true only if the line search is performed exactly.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Cauchy's Steepest Descent method zigzags its way towards the optimum point. If you look at the figure, it will be clear that when you start from  $x_0$  and go to  $x_1$ , then go to  $x_2$  and then, go to  $x_3$ , you take zigzag path. This is because each step is orthogonal to previous step. In other words, each step is perpendicular to the previous step.

So, the entire sequence of paths will look like a zigzag way. This can be very easily shown that Steepest Descent method zigzag its way towards the optimum point. Remember, the update rule of Steepest Descent method which is  $x_{k+1} = x_k + \alpha_k s_k$ , where  $\alpha_k$  is the step length and  $s_k$  is the search direction that is gradient of the function at current estimate  $x_k$ .

Alpha  $k$  is obtained by minimizing the function  $f(x_k + 1)$  which is  $f(x_k + \alpha_k x_k)$  and you know that to determine this alpha will say  $\frac{df}{d\alpha} = 0$  and solve the resulting equation. Now, upon differentiation we get gradient of  $f$  at  $x_k + 1$  into  $s_k$  is equal to 0. Note that  $x_k + 1 + \alpha_k x_k$  is same as  $x_k + 1$ .

So, upon differentiation we get that gradient of  $f$  at  $x_k + 1$  into  $s_k$  equal to 0. That means the search direction which is a vector, so the product of search direction vector at  $k$ -th step and the gradient vector at  $k + 1$  step is equal to 0. That means they are perpendicular to each other.

Now, the search direction at  $k + 1$ th iteration is nothing, but the minus of the gradient vector at  $k + 1$ . This is what we know from Steepest Descent Method. So, we can write the search direction vector at  $k + 1$  into such direction at  $k$ -th step is equal to 0. That means two consecutive search directions are orthogonal to each other. So, that is why you see that the Steepest Descent method zigzags its way towards the optimum point.

Remember, this result is true only if the line search is performed exactly. That means the value of the alpha  $k$  is determined exactly.

(Refer Slide Time: 05:05)

The slide is titled "Gradient Based Methods: Role of Step Length". The text on the slide reads: "The update rule of a Gradient Based Method requires a step length controlling the amount of gradient updated to the current point at each iteration. A too large step length can lead to divergence. A too small step length takes longer time to converge. Backtracking line search and exact line search are two methods to find step length." Below the text, there is a handwritten equation in pink:  $x^{k+1} = x^k + \alpha \frac{\partial f}{\partial x}$ . The  $\alpha$  and the fraction are underlined with arrows pointing to them. In the bottom right corner, there is a small video inset showing a man speaking. The slide footer includes the IIT KHARAGPUR logo and the text "NPTEL ONLINE CERTIFICATION COURSES".

So, what is the role of step length in Gradient based methods. The update rule of a Gradient based method requires a step length controlling the amount of gradient updated

to the current point at each iteration because we have seen that  $x_{k+1}$  is  $x_k$  plus  $\alpha_k$  and  $s_k$ . So, this is the gradient and this is the step length.

So, the step length decides how much will move in the direction of search direction vector. A too large step length can lead to divergence; a too small step length will take much longer time to converge. That is why we need an optimum step length, so that we get convergence and we get convergence in minimum time. There are two methods to find the step length which are commonly used; the backtracking line search and the exact line search. The big backtracking line search is an in exact line search and the exact line search finds the  $\alpha_k$  exactly, whereas backtracking or in exact line search finds an approximate value of the step length.

(Refer Slide Time: 07:06)

The slide is titled "Gradient Based Methods: Role of Step Length". It contains the following text: "The update rule of a Gradient Based Method requires a step length controlling the amount of gradient updated to the current point at each iteration. A too large step length can lead to divergence. A too small step length takes longer time to converge. Backtracking line search and exact line search are two methods to find step length." Below the text are three contour plots of a 2D function. The first plot, labeled "Too large step size", shows a path that oscillates and diverges from the minimum. The second plot, labeled "Too small step size", shows a path that zig-zags slowly towards the minimum. The third plot, labeled "Appropri", shows a path that converges smoothly to the minimum. The slide also features logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES, and a small video inset of a speaker.

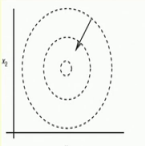
So, this is what happens when I use a step length that is too large. You see the search direction starts oscillating when you take too small step size. It takes too long time to hit the minimum. The minimum is shown by this term and we are showing a two dimensional functions contours. Similarly when you find an appropriate step size or step length, you will get convergence and you will get convergence in not so many iterations.

(Refer Slide Time: 08:05)

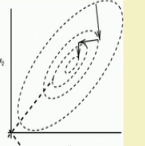
Page 3/10

## Cauchy's Steepest Descent Method: Scaling

Consider the two quadratic functions:






$f = x_1^2 + x_2^2$



$f = x_1^2 + ax_2^2$

Larger is the value of the parameter  $a$ , more elongated are the contours of the function  $f$ , and slower is the convergence rate.

Speed of convergence is related to the condition number of Hessian matrix. The condition number of a symmetric positive definite matrix is the ratio of largest to smallest eigenvalues. For a well conditioned Hessian matrix, the condition number is close to unity. Then the contours become more circular and the steepest descent method works best.

We want to make another comment on Steepest Descent method. This is about scaling. Let us consider two quadratic functions as shown in the figure. The first one is  $x_1^2 + x_2^2$  and you can see the contours are all perfectly circular. Now, if I consider the second function which is  $x_1^2 + ax_2^2$  then depending on the value of the parameter, the contours may be circular or may be elongated. If  $a$  equal to 1, this equation reduces to the first equation and the contours becomes circular, but if I take large value of  $a$ , the contours will be elongated as shown.

Now, you can note down that the first function  $x_1^2 + x_2^2$  if I start let us say from this point, you see that the search direction directly points to the minimum is a two-dimensional function and  $x_1^2 + x_2^2$ . So, the minimum is 0, but for the second function  $x_1^2 + ax_2^2$  you see that it takes longer time for Steepest Descent Method to go and hit the minimum at  $x_1$  equal to 0,  $x_2$  equal to 0. Hey sorry the function value equal to 0.

So, what we learn is that for a quadratic function whose contours are circular, the search direction in case of steepest method will directly point towards the minimum, but if the contours are elongated. Then, the convergence will be slow because in that case such direction does not directly point to the minimum starting from any arbitrary point. Speed of convergence is related to the condition number of the Hessian matrix, the condition number of a symmetric positive definite matrix is the ratio of largest to smallest eigenvalues.

The condition number of a symmetric positive definite matrix is the ratio of largest to smallest eigenvalues, condition number may generally be defined as the ratio of this singular values, but in case of symmetric positive definite matrix, the eigenvalues and the single value, singular values are same. So, the condition number of a symmetric positive definite matrix is the ratio of largest to smallest eigenvalues for a well conditioned Hessian matrix. The condition number is close to unity. If the condition number is close to unity, the contours will become more circular and the steepest descent method will work very well because in that case, the search direction will directly point towards the minimum.

(Refer Slide Time: 12:22)

**Cauchy's Steepest Descent Method: Scaling**

For the function  $f = x_1^2 + ax_2^2$  we have:

$$H = \nabla^2 f = \begin{bmatrix} 2 & 0 \\ 0 & 2a \end{bmatrix}$$

Eigenvalues are 2 and  $2a$ .

Thus, the condition number =  $a$

Convergence is fastest when  $a = 1$

The slide includes two contour plots. The left plot shows circular contours for  $f = x_1^2 + x_2^2$  with axes  $x_1$  and  $x_2$ . The right plot shows elliptical contours for  $f = x_1^2 + ax_2^2$  with axes  $x_1$  and  $x_2$ . A small video inset of a presenter is visible in the bottom right corner.

Consider the function  $x_1^2 + ax_2^2$ . Compute the Hessian and you get the matrix  $\begin{bmatrix} 2 & 0 \\ 0 & 2a \end{bmatrix}$ . So, the eigenvalues are 2 and  $2a$ . So, the condition number is  $2a/2 = a$ . So, when the condition number is 1. That means  $a = 1$ , the convergence is fastest.

(Refer Slide Time: 12:59)


Page 5/7




## Cauchy's Steepest Descent Method: Scaling

For the function  $f = x_1^2 + ax_2^2$  we can define new variables  $y_1$  and  $y_2$  as follows:

$$y_1 = x_1, \quad y_2 = \sqrt{a}x_2$$

Then redefine  $f$  as:  $g = y_1^2 + y_2^2$  The condition number of the Hessian of this new matrix will be 1. Then the contours become circular and the steepest descent method works best.

$H = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ 


So now, again consider the function  $x_1^2 + ax_2^2$ . So, now if I perform appropriate scaling and if it is possible for me to redefine the function  $x_1^2 + ax_2^2$  as  $y_1^2 + y_2^2$  in terms of new variable  $y_1$  and  $y_2$ . Then the contours of this new function  $y_1^2 + y_2^2$  will be circular and Steepest Descent method will work very well.

(Refer Slide Time: 14:55)

Page 6/8

## Cauchy's Steepest Descent Method: Scaling

For the function  $f = x_1^2 + ax_2^2$  we can define new variables  $y_1$  and  $y_2$  as follows:


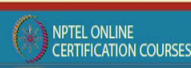
$$y_1 = x_1, \quad y_2 = \sqrt{a}x_2$$

Then redefine  $f$  as:  $g = y_1^2 + y_2^2$  The condition number of the Hessian of this new matrix will be 1. Then the contours become circular and the steepest descent method works best.

In general, if for a function  $f(\mathbf{x})$ , we scale the variables as:  $\mathbf{x} = \mathbf{T}\mathbf{y}$ , then the new function is

$$g(\mathbf{y}) = f(\mathbf{T}\mathbf{y}); \quad \text{Gradient: } \nabla g = \mathbf{T}^T \nabla f; \quad \text{Hessian: } \nabla^2 g = \mathbf{T}^T \nabla^2 f \mathbf{T}$$

Usually  $\mathbf{T}$  is chosen as a diagonal matrix such that Hessian of  $g$  has condition number close to unity.

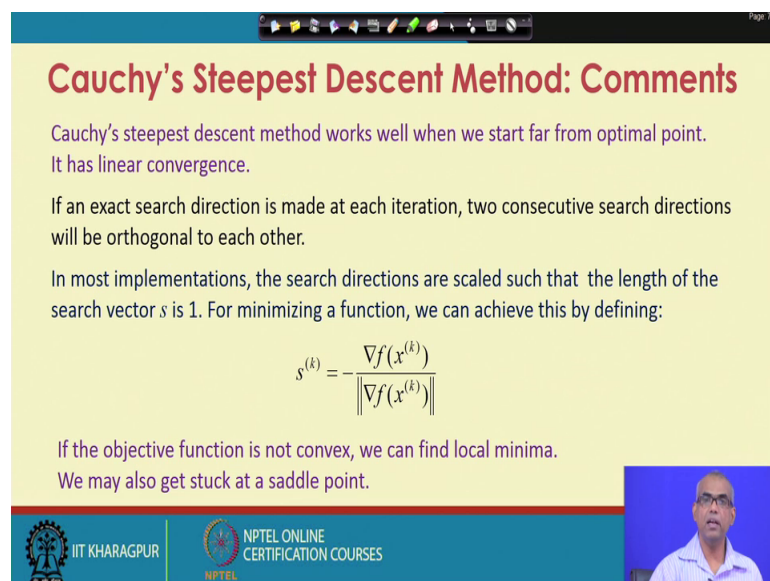
So, it is possible to do it here. Note that if I redefine  $y_1$  equal to  $x_1$  and  $y_2$  equal to  $\sqrt{a}x_2$ , then the function  $f = x_1^2 + ax_2^2$  can be written as  $g = y_1^2 + y_2^2$ . Note that, for this function, the condition number of the Hessian matrix will be 1, because for this function, the Hessian will be  $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ . So,

condition number is the positive symmetric definite matrix, symmetric positive definite matrix. So, the condition number will be 2 by 2 equal to 1 and the contours will become perfectly circular and the Steepest Descent method will work very well.

In general how do I perform scaling? So, to perform scaling we define the new variable  $y$  and take a diagonal matrix  $T$  and scale, the variable as  $x$  equal to  $T y$ . So, in terms of new variable  $y$ , I can scale the variables as  $x$  equal to  $T y$  where  $T$  i have to choose as a diagonal matrix. So, the new function will be now say  $g$  of  $y$  equal to function of  $T$  and  $y$ . So, the gradient of this new function gradient  $g$  can be computed as transpose of  $T$  into gradient of old function  $f$ .

Similarly, the Hessian of the new function can be computed as Hessian of  $g$  equal to transpose of  $T$  into Hessian of old function  $f$  into matrix  $T$ . So, you have to choose the matrix  $T$  as a diagonal matrix, such that the Hessian of the new function  $g$  has condition number close to unity.

(Refer Slide Time: 16:46)



**Cauchy's Steepest Descent Method: Comments**

Cauchy's steepest descent method works well when we start far from optimal point. It has linear convergence.

If an exact search direction is made at each iteration, two consecutive search directions will be orthogonal to each other.

In most implementations, the search directions are scaled such that the length of the search vector  $s$  is 1. For minimizing a function, we can achieve this by defining:

$$s^{(k)} = -\frac{\nabla f(x^{(k)})}{\|\nabla f(x^{(k)})\|}$$

If the objective function is not convex, we can find local minima. We may also get stuck at a saddle point.

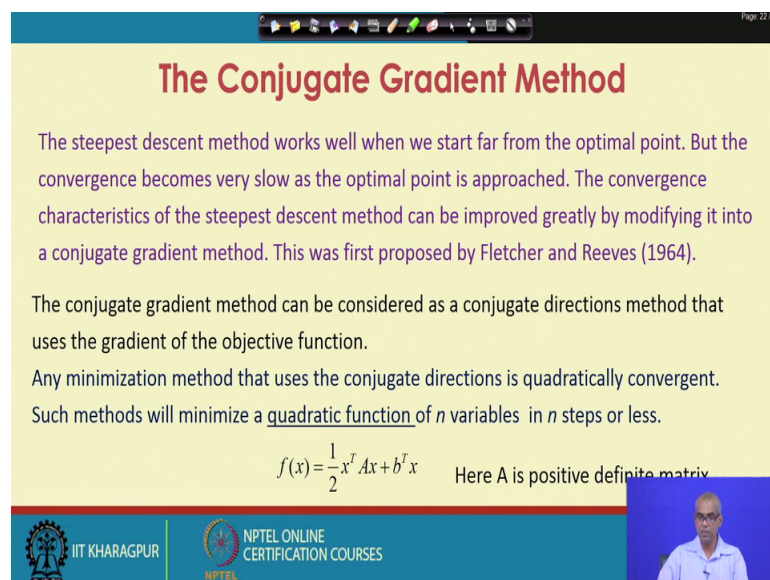
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Let us now make some more comments about Steepest Descent Method because Steepest Descent method works very well. When we start far from optimal point, it has linear convergence. If an exact search direction is made at each iteration, two consecutive search directions will be orthogonal to each other and the Steepest Descent method will zigzag its way towards the optimum.

In most implementations, the search directions are scaled such that the length of the search vector is 1. So, for minimizing a function we can scale the search direction by dividing the gradient by the norm of the gradient. So, instead of taking the search direction  $s$  at  $k$ -th iteration as minus gradient of  $f$  at  $x_k$ , we take  $s_k$  equal to minus gradient of  $f$  at  $x_k$  divided by norm of the gradient of  $f$  at  $x_k$ . If the objective function is not convex, we can find local minima while minimizing, but if the objective function is convex, you know that local minima is also a global minimum. We may also get stuck at a saddle point, but these are the characteristic feature of any Gradient based method.

Now, we will talk about Conjugate Gradient Method.

(Refer Slide Time: 18:41)



**The Conjugate Gradient Method**

The steepest descent method works well when we start far from the optimal point. But the convergence becomes very slow as the optimal point is approached. The convergence characteristics of the steepest descent method can be improved greatly by modifying it into a conjugate gradient method. This was first proposed by Fletcher and Reeves (1964).

The conjugate gradient method can be considered as a conjugate directions method that uses the gradient of the objective function.

Any minimization method that uses the conjugate directions is quadratically convergent. Such methods will minimize a quadratic function of  $n$  variables in  $n$  steps or less.

$$f(x) = \frac{1}{2} x^T A x + b^T x$$

Here A is positive definite matrix.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

The Steepest Descent method works well when you start far from the optimal point, but the convergence becomes very slow. As the optimal point is approached, the convergence characteristics of the Steepest Descent method can be improved greatly by modifying it into a conjugate gradient method. This was first proposed by Fletcher and Reeves in 1964.

The Conjugate Gradient method can be considered as a conjugate directions method that uses the gradient of the objective function. Any minimization method that uses the conjugate directions is quadratically convergent. Such methods will minimize a quadratic function of  $n$  variables in  $n$  steps or less.



(Refer Slide Time: 19:37)

**The Conjugate Gradient Method**

Powell's conjugate direction method requires  $n$  single-variable search per iteration. We then find new conjugate direction at the end of each iteration.

Therefore, to find the minimum of a quadratic function, in general, Powell's method requires  $n^2$  single-variable search.

But if we use the gradients of the objective function, we can set up a new conjugate direction after every single-variable search, and hence we can achieve faster convergence.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 11

Powell's Conjugate Direction method requires  $n$  single variable search per iteration. We then find new conjugate direction at the end of each iteration. Therefore, to find the minimum of a quadratic function in general, Powell's method requires  $n$  square single variable search. But if, we use gradients of the objective function, we can set up a new conjugate direction after every single variable search and hence, we can achieve faster convergence.

(Refer Slide Time: 20:13)

**The Conjugate Gradient Method**

The conjugate gradient method combines current information about the gradient vector with that of gradient vectors from previous iterations to obtain the new search direction.

The proposed search direction for this method is a linear combination of the current gradient and the previous search direction.

$$S^{(k+1)} = -\nabla f_{k+1} + \frac{\|\nabla f_{k+1}\|^2}{\|\nabla f_k\|^2} S^{(k)} = -\nabla f_{k+1} + \frac{\nabla^T f_{k+1} \nabla f_{k+1}}{\nabla^T f_k \nabla f_k} S^{(k)}$$

Note that conjugate gradient method is a first-order method. This method represents a major improvement over steepest descent method with only a marginal increase in computational effort.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

The Conjugate Gradient method combines current information about the gradient vector with that of gradient vectors from previous iterations to obtain the new search direction. The proposed search direction for this method is a linear combination of the current gradient and the previous search direction. So, the search direction becomes a linear combination of the current gradient and the previous search direction.

So, the search direction  $S$  at  $k$  plus 1th iteration is minus gradient  $f$  at  $k$  plus 1 plus norm square of gradient at  $k$  plus 1 divided by norm square of gradient  $f$  at  $k$ -th iteration into search direction at  $k$ -th iteration. So, basically the search direction at the current step is a combination of the current gradient and the previous search direction. So, this can also be written as minus gradient  $f$  plus gradient transpose into gradient  $f$  divided by gradient  $f$  transpose into gradient  $f$ . The numerator 1 is evaluated at  $k$  plus 1th iteration that is  $x$   $k$  plus 1 and the denominator is evaluated at  $x$   $k$  and the search direction is evaluated at  $x$   $k$ , that means search direction from the previous step.

So, in case of Conjugate Gradient method, the current search direction is always a linear combination of the current gradient and the previous search direction. So, the difference with the Steepest Descent method is this that in case of Steepest Descent method, the search direction is the current gradient direction negative of the current gradient direction, but we add this information from the previous step in case of Conjugate Gradient method.

Note that the Conjugate Gradient method is a first order method. We are still using first order informations only. First partial derivatives are involved. This method represents a major improvement over Steepest Descent method with only a marginal increase in the computational effort? Note that only this information above the previous search direction is included in the current gradient information. So, there is a marginal increase in the computational effort, but there is a major improvement over steepest descent method.

(Refer Slide Time: 24:02)

Page 25/38

## The Conjugate Gradient Method: Algorithm

Step-1: Start with an arbitrary initial point  $\mathbf{X}^0$ .



Step-2: Set the first search direction  $\mathbf{S}^0 = -\nabla f(\mathbf{X}^0) = -\nabla f_0$

Step-3: Find the point  $\mathbf{X}^1$  according to the relation  $\mathbf{X}^1 = \mathbf{X}^0 + \lambda_0^* \mathbf{S}^0$   
 where  $\lambda_0^*$  is the optimal step length in the direction  $\mathbf{S}^0$ . Set  $k=2$  and go to next step.

Step-4: Find  $\nabla f_k = \nabla f(\mathbf{X}^k)$ , and set  $\mathbf{S}^{(k+1)} = -\nabla f_{k+1} + \frac{\|\nabla f_{k+1}\|^2}{\|\nabla f_k\|^2} \mathbf{S}^{(k)}$

Step-5: Compute the optimum step length  $\lambda_k^*$  in the direction  $\mathbf{S}^k$  and find the new point  
 $\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} + \lambda_k^* \mathbf{S}^{(k)}$

Step-6: If convergence (small value of gradient) is achieved, stop.  
 Otherwise set the value of  $k=k+1$  and go to step 4.

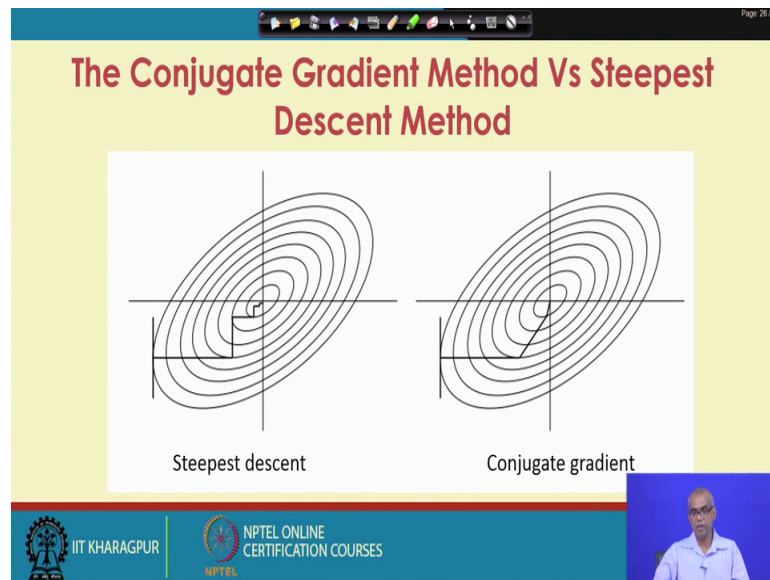



So, where is the algorithm for Conjugate Gradient method? In the step 1, we start with an arbitrary initial point  $\mathbf{X}^0$ . Step 2 we said the first search direction as  $\mathbf{S}^0$  equal to minus gradient of  $F$  at  $\mathbf{X}^0$  which is same as Steepest Descent method. In the step 3, we find the point  $\mathbf{X}^1$  according to the usual relation  $\mathbf{X}^1 = \mathbf{X}^0 + \lambda_0^* \mathbf{S}^0$ . So, at  $\lambda_0^*$  is the optimal step length in the direction  $\mathbf{S}^0$ . That means we need to perform the line search.

We set  $k$  equal to 2 and go to the next step. In the step 4, we find the gradient of  $F$ . Let us consider this is  $k$ -th step. So, you find the gradient of  $F$  at  $\mathbf{X}^k$  and set the search direction for  $k+1$  step as  $\mathbf{S}^{k+1} = -\nabla f_{k+1} + \frac{\|\nabla f_{k+1}\|^2}{\|\nabla f_k\|^2} \mathbf{S}^k$ . So, such direction at  $\mathbf{X}^{k+1}$  is the linear combination of the gradient at  $\mathbf{X}^{k+1}$  and the search direction at  $\mathbf{X}^k$ .

Now, compute the optimal step length  $\lambda_k^*$  in the direction  $\mathbf{S}^k$  by doing line search and find the new point as  $\mathbf{X}^{k+1} = \mathbf{X}^k + \lambda_k^* \mathbf{S}^k$ . If the convergence is achieved, we will stop. To understand the convergence is achieved or not, we have to check the magnitude of the gradient. So, if the norm of the gradient is very small, we will assume that the convergence is achieved will stop otherwise we will set the value of  $k$  equal to  $k+1$  and go to again step 4, where we find the gradient and then, the search direction and proceed.

(Refer Slide Time: 26:36)



So, this figure schematically shows the functioning of Steepest Descent method and Conjugate Gradient method. Note that, the Steepest method takes zigzag path to the optimal point and takes many more iterations compared to Conjugate Gradient method.

(Refer Slide Time: 26:57)

The slide contains the following text and equations:

**The Conjugate Gradient Method: Example**

Minimize  $f(x_1, x_2) = x_1 - x_2 + 2x_1^2 + 2x_1x_2 + x_2^2$

starting from the point  $\mathbf{X}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

**Solution:**

**Iteration - 1:** The gradient of  $f$  is given by:

$$\nabla f = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \end{bmatrix} = \begin{bmatrix} 1 + 4x_1 + 2x_2 \\ -1 + 2x_1 + 2x_2 \end{bmatrix}$$
$$\nabla f_0 = \nabla f(\mathbf{X}^{(0)}) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Therefore,  $\mathbf{S}^{(0)} = -\nabla f_0 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

The slide includes logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES, and a small video inset of a speaker.

Now, let us take an example. We are considering a quadratic function two variables and we will solve, that means we will minimize this function using Conjugate Gradient method starting from the point  $\mathbf{X}^{(0)}$  which is  $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ . So, in the iteration 1, the gradient of the objective function is evaluated first. So, you take the  $\partial f / \partial x_1$  and  $\partial f / \partial x_2$ . They

are evaluated as 1 plus 4 x one plus two x two 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, put the value of x 1 equal to 0 and x 2 equal to 0 and we get the value of the gradient at X 0 as 1 minus 1. So, initially the search direction is minus of gradient at X 0. So, the search direction is minus 1 1. Note that the gradient is 1 minus 1. So, the search direction is minus of the gradient, so minus 1 1.

(Refer Slide Time: 28:26)

**The Conjugate Gradient Method: Example**

To find  $\mathbf{X}^{(1)}$ , we need to find the optimal step length  $\lambda_0^*$ . For this, we minimize  $f(\mathbf{X}^{(0)} + \lambda_0^* \mathbf{S}^{(0)})$  with respect to  $\lambda_0$ .

Now,  $f(\mathbf{X}^{(0)} + \lambda_0^* \mathbf{S}^{(0)}) = f\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \lambda_0^* \begin{bmatrix} -1 \\ 1 \end{bmatrix}\right) = f(-\lambda_0, \lambda_0)$   $\lambda_1 = -\lambda_0$

Evaluate  $f(-\lambda_0, \lambda_0) = -\lambda_0 - \lambda_0 + 2\lambda_0^2 - 2\lambda_0^2 + \lambda_0^2 = \lambda_0^2 - 2\lambda_0$   $\lambda_2 = \lambda_0$

Set  $\frac{df}{d\lambda_0} = 0$  and we get  $\lambda_0^* = 1$ . Thus,  $\mathbf{X}^{(1)} = \mathbf{X}^{(0)} + \lambda_0^* \mathbf{S}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Now,  $\nabla f_1 = \nabla f(\mathbf{X}^{(1)}) = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$ .

We proceed to the next iteration.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

To find X 1, we now need to find the optimal step length lambda 0. For this we minimize f of X 0 plus lambda 0 into S 0 with respect to lambda 0.

So, let us now evaluate f of X 0 plus lambda 0 S 0. You know X 0 equal to 0 0 and S 0 is minus 1 1. So, this becomes f of minus lambda 0 lambda 0. So, we evaluate f of minus lambda 0 lambda 0 by putting X 1 equal to minus lambda 0 and X 2 equal to lambda 0 in the original expression for f and we get that as lambda 0 square minus 2 lambda 0.

So, f of minus lambda 0 lambda 0 equal to lambda 0 square minus 2 lambda 0.

(Refer Slide Time: 29:50)

Page 23 / 42

## The Conjugate Gradient Method: Example

To find  $\mathbf{X}^{(1)}$ , we need to find the optimal step length  $\lambda_0^*$ . For this, we minimize  $f(\mathbf{X}^{(0)} + \lambda_0 \mathbf{S}^{(0)})$  with respect to  $\lambda_0$ .

Now,  $f(\mathbf{X}^{(0)} + \lambda_0 \mathbf{S}^{(0)}) = f\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \lambda_0 \begin{bmatrix} -1 \\ 1 \end{bmatrix}\right) = f(-\lambda_0, \lambda_0)$

Evaluate  $f(-\lambda_0, \lambda_0) = -\lambda_0 - \lambda_0 + 2\lambda_0^2 - 2\lambda_0^2 + \lambda_0^2 = \lambda_0^2 - 2\lambda_0$



Set  $\frac{df}{d\lambda_0} = 0$  and we get  $\lambda_0^* = 1$ . Thus,  $\mathbf{X}^{(1)} = \mathbf{X}^{(0)} + \lambda_0^* \mathbf{S}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Now,  $\nabla f_1 = \nabla f(\mathbf{X}^{(1)}) = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$

We proceed to the next iteration.

*Handwritten notes:*  
 $\frac{d}{d\lambda_0} (\lambda_0^2 - 2\lambda_0) = 0$   
 $2\lambda_0 - 2 = 0$   
 $\Rightarrow \lambda_0^* = \frac{2}{2} = 1$

*Arrows pointing up from the handwritten notes to the equations above.*

So, to find out the optimal lambda 0, we will set d lambda 0 of f which is lambda 0 squared minus 2 lambda 0 is equal to 0. So, you get as 2 lambda 0 minus 2 equal to 0 and then, we get lambda 0 optimal is 2 by 2 equal to 1.

So, we get lambda 0 star as 1. So, once we get that, we can find X 1 as X 0 plus lambda 0 into S 0. So, X 0 is 0 0 lambda 0 have obtained as 1 s 0 is minus 1 1. So, X 1 is obtained as minus 1 1. So, find out the gradient of f at X 1 and we can find the gradient s minus 1 minus 1. So, now we proceed to the next iterations. So, you obtain gradient of f at X 1 as minus 1 minus 1.

(Refer Slide Time: 31:03)

Page 23 / 43

## The Conjugate Gradient Method: Example

### Iteration - 2:

$\mathbf{S}^{(1)} = -\nabla f_1 + \frac{\|\nabla f_1\|^2}{\|\nabla f_0\|^2} \mathbf{S}^{(0)}$ ; Here,  $\|\nabla f_0\|^2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}^T \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 2$  and  $\|\nabla f_1\|^2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}^T \begin{bmatrix} -1 \\ -1 \end{bmatrix} = 2$



Therefore,  $\mathbf{S}^{(1)} = -\nabla f_1 + \frac{\|\nabla f_1\|^2}{\|\nabla f_0\|^2} \mathbf{S}^{(0)} = -\begin{bmatrix} -1 \\ -1 \end{bmatrix} + \left(\frac{2}{2}\right) \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$

To minimize  $f(\mathbf{X}^{(1)} + \lambda_1 \mathbf{S}^{(1)}) = f(-1, 1 + 2\lambda_1) = 4\lambda_1^2 - 2\lambda_1 - 1$

we set  $\frac{df}{d\lambda_1} = 0$ . This gives  $\lambda_1^* = 0.25$ , and hence

$\mathbf{X}^{(2)} = \mathbf{X}^{(1)} + \lambda_1^* \mathbf{S}^{(1)} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 0.25 \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1.5 \end{bmatrix}$

*Handwritten notes:*  
 $\frac{d}{d\lambda_1} (4\lambda_1^2 - 2\lambda_1 - 1) = 0$   
 $\Rightarrow \lambda_1^* = 0.25$

So, in the iteration 2, we first now find the search direction S 1 which can be obtained as a linear combination of the gradient at X 1 and the search direction at X 0. So, let us first find out the norm square at of f. The norm square of gradient of f at X 0 and norm square of gradient f at X 1 which are obtained as 2 and 2. So, you can now find such direction at X 1 which is obtained as 0 2. So, again minimize f of X 1 plus lambda 1 s 1 because you have to find out the optimal lambda 1. Now, this is obtained as 4 lambda 1 square minus 2 lambda 1 minus 1.

So, we have to now take dd lambda 1 of 4 lambda 1 square minus 2 lambda 1 minus 1 equal to 0 and if we do that, we will get lambda 1 star is equal to point 2 5 or 1 by 4.

(Refer Slide Time: 32:49)

**The Conjugate Gradient Method: Example**

**Iteration - 2:**

$$S^{(1)} = -\nabla f_1 + \frac{\|\nabla f_1\|^2}{\|\nabla f_0\|^2} S^{(0)}; \quad \text{Here, } \|\nabla f_0\|^2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}^T \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 2 \quad \text{and} \quad \|\nabla f_1\|^2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}^T \begin{bmatrix} -1 \\ -1 \end{bmatrix} = 2$$

Therefore,  $S^{(1)} = -\nabla f_1 + \frac{\|\nabla f_1\|^2}{\|\nabla f_0\|^2} S^{(0)} = -\begin{bmatrix} -1 \\ -1 \end{bmatrix} + \left(\frac{2}{2}\right) \begin{bmatrix} -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$

To minimize  $f(X^{(1)} + \lambda_1 S^{(1)}) = f(-1, 1 + 2\lambda_1) = 4\lambda_1^2 - 2\lambda_1 - 1$

we set  $\frac{df}{d\lambda_1} = 0$ . This gives  $\lambda_1^* = 0.25$ , and hence

$$X^{(2)} = X^{(1)} + \lambda_1^* S^{(1)} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 0.25 \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1.5 \end{bmatrix} \quad \text{This is the optimum.}$$

Note,  $\nabla f_2 = \nabla f(x^{(2)}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ ;

$\|\nabla f_1\|^2 = 2$ ;  $\|\nabla f_2\|^2 = 0$

Thus,  $S^{(2)} = 0$

No search direction to reduce  $f$  further.

$$x^* = \begin{bmatrix} -1.0 \\ 1.5 \end{bmatrix}$$

So, X 2 is obtained as X 1 plus lambda 1 into S 1 and it is obtained as minus 1 1.5. In fact, X 2 equal to minus 1 1.5 is the optimal solution. To check that this is the optimal solution let us find out the gradient at this value X 2 and we find that the gradient is 0 0.

So, gradient 0 is the first order optimality criteria if you remember. Also find out the norm of the gradient of f at X 2 which is 0. So, S 2 will be 0. So, there is no search direction to reduce the function value further. So, it shows that x to minus 1 1.5, that means X 1 equal to minus 1 and X 2 equal to 1.5 is the optimal solution. So, this is how the Conjugate Gradient method works and you see that we had a two variable quadratic function and we could get the optimal point in exactly two iterations.

With this we stop our discussion on Conjugate Gradient Method here.