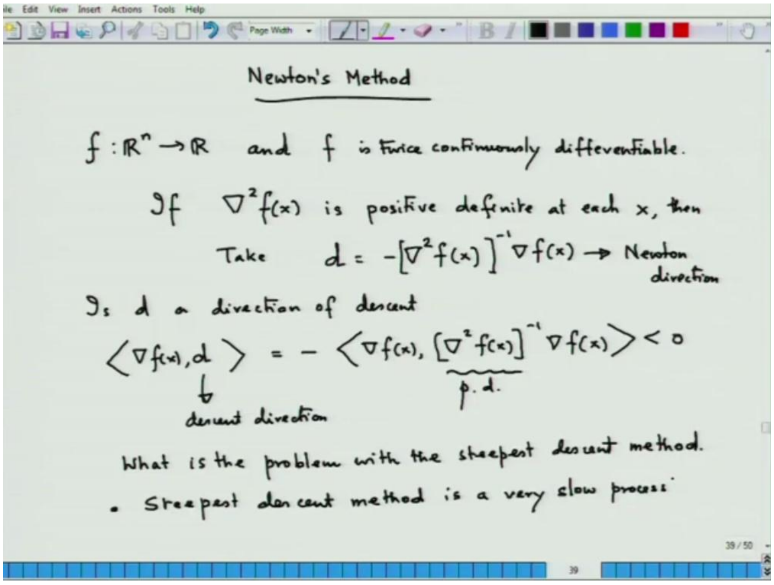
Foundation of Optimization
Prof. Dr. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 8

(Refer Slide Time: 00:28)



Newton's Method

$f : \mathbb{R}^n \to \mathbb{R}$ and f is twice continuously differentiable.

If $\nabla^2 f(x)$ is positive definite at each x, then

Take $d = -[\nabla^2 f(x)]^{-1} \nabla f(x) \to$ Newton direction

Is d a direction of descent

$$\underset{\substack{\downarrow \\ \text{descent direction}}}{\langle \nabla f(x), d \rangle} = - \langle \nabla f(x), \underbrace{[\nabla^2 f(x)]^{-1}}_{p.d.} \nabla f(x) \rangle < 0$$

What is the problem with the steepest descent method.

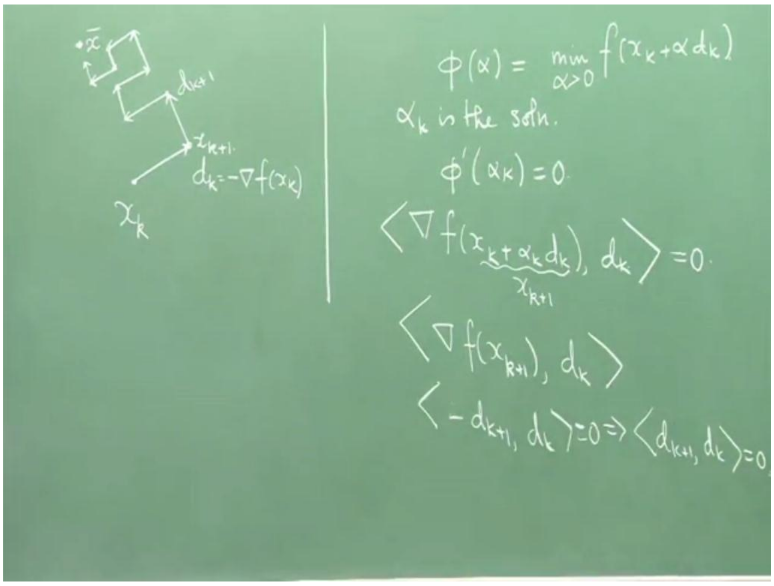
- Steepest descent method is a very slow process.

Today, we will start with discussing Newton's method. Newton's method is one of the most important methods in on for solving unconstraint optimization problems. So, we will consider a function R n to R and f would be assumed to be twice continuously differentiable. Now, if that is the case, the question is what is my descent direction? In this case, if the hessian matrix is positive definite, then d is equal to minus, because it is positive definite I can take the inverse of the hessian matrix this can act as a direction of descent. So, why because one of the question is d a direction of descent consider take is d a direction of descent.

So, let us see so what so suppose there as a point where grad of f x is not equal to 0, and let us see what does this give? This will give us now if a is positive definite. So is a inverse, so this being positive definite. So, this is also also p d matrix positive definite matrix p d for positive definite. So, this would become negative, and hence this is now this would give you a descent direction. So, this direction is often referred to in the literature as Newton direction, we will come why it is called the Newton direction, but we have few question to answer before this.

Now, the question is as follows why we at all need Newton's method? Because Newton method needs the function to be twice continuously differentiable means not only it is differentiable it is second derivative or the hessian matrix is also continuous for the function of x. But we were doing fine with the steepest descent method possibly, because we had just to bother about the gradient and need not bother about computing a matrix. So, we are increasing the computation cost by being in a hessian matrix. So, what, what is the problem with steepest descent method? That is the question. So, this fact, let us explain on the board what, what really is a problem with the steepest descent method?

(Refer Slide Time: 04:40)



$$\phi(\alpha) = \min_{\alpha > 0} f(x_k + \alpha d_k)$$

α_k is the soln.

$$\phi'(\alpha_k) = 0$$

$$\left\langle \nabla f(\underbrace{x_k + \alpha_k d_k}_{x_{k+1}}), d_k \right\rangle = 0$$

$$\left\langle \nabla f(x_{k+1}), d_k \right\rangle$$

$$\left\langle -d_{k+1}, d_k \right\rangle = 0 \Rightarrow \left\langle d_{k+1}, d_k \right\rangle = 0$$

So, in the steepest descent method, what you have is that your d is say the negative of grad of f x k where k is x k is the kth iteration point. Now, what we want to show is the following that if here I have x k, and suppose my descent direction from here is this; this is your d k which is negative of grad f x k. So, if this is my x k plus 1, then the descent direction from there that is negative of grad f x k plus 1 this would be perpendicular to d k. As a result the steepest descent method if it has to move if this is the actual optima my x bar, then it will move for in this way this sort of zigzag zagging pattern.

And this may just delay the progress towards the optimum, and that is the whole issue that is lies at the core of the steepest distance method been not used so much in practices. And one has to bring in such methods that we are going to discuss like the

Newton's method here. Now, let us see why this is so see one of the ways to find the step length that is how much you have to move along the direction alpha to obtain a sufficient decrease is obtained by theoretically by minimizing this function. Now, because I have put alpha strictly bigger than 0 and suppose there is a minima I have obtained a minima which is alpha k that is what I put as alpha k.

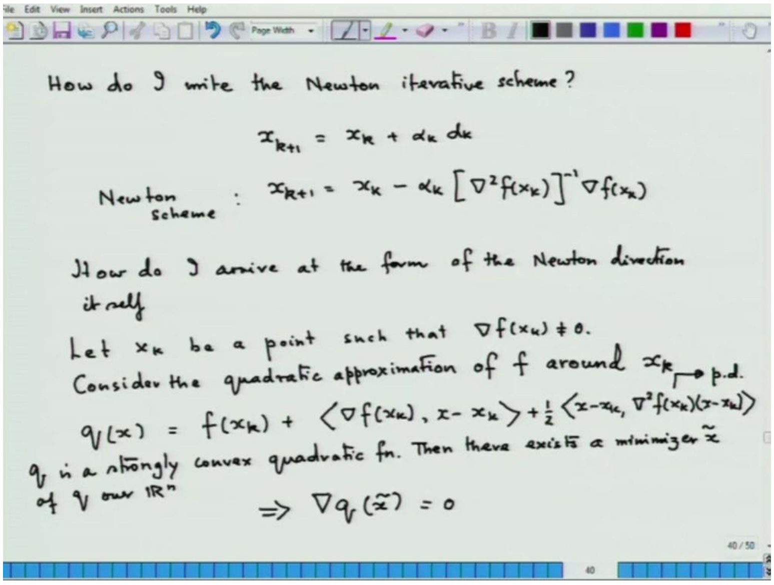
Now, suppose I have got the minima as alpha k here or alpha not whatever you want to say alpha k that is what we are what I say by the minimum. But here I have looked at only over alpha strictly bigger than 0 and this alpha strictly bigger than 0 is an open set. We have not yet discussed much about constraint and unconstraint optimization, because here what you have is constraint optimization. But I want to tell you something that if you optimize function or minimize a function of a open set the necessary optimality condition is nothing but the gradient of f of x is equal to 0. Because if you really talking about open sets then you would, we would all observe that as we will see later in the Karush Kuhn Tucker condition that all the constraint will become so called inactive constraints and the Langrange multipliers would be 0.

So, what, what is the very important to note here is that I can just obtain of alpha k is the solution, because there is x x k plus 1 x alpha k is a solution to this problem. And my necessary optimality condition is this though I have told you there is a paradigm shift between the between the constraint problem and unconstraint problem. But it is important to know that when you have open set the necessary optimality condition of a constraint problem and a unconstraint problem is the same. If you are minimizing the function over a open set please note that it is same as minimizing over the whole space.

Now, what would I get? From here I would get the following so alpha now there is a alpha k strictly bigger than 0 which solves. So, this happens is a necessary condition and this would give me what it will give me a grad of f of so my f dash this is nothing but x k alpha k d k into this. You see what I do is I apply the chain rule first I took the gradient with respect to this. And then took a derivative of this with respect to alpha which is nothing but the vector d k.

So, we have to write the inner product, because we cannot have multiplication, because we have functions for space R into R. Now, what is this, this part? This is nothing but x k plus 1. So, I can write this as grad of f of x k plus 1 into d k, but what is this by, by the definition of steepest descent method? This is nothing but the negative of the descent direction. And this would simply imply that this means that these two these two directions the conjugative descent directions are perpendicular to one another, and that is why it slows down the whole process. So, our conclusion here is steepest descent method is a very slow process. So, Newton method is one of the faster processes and so we need faster processes, because we just have slow processes here that is steepest descent is a slow process. Now, of course, it is it is not very difficult to write down the Newton scheme.

(Refer Slide Time: 11:35)



How do I write the Newton iterative scheme?

$$x_{k+1} = x_k + \alpha_k d_k$$

Newton scheme : $x_{k+1} = x_k - \alpha_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$

How do I arrive at the form of the Newton direction itself

Let x_k be a point such that $\nabla f(x_k) \neq 0$.

Consider the quadratic approximation of f around x_k → p.d.

$$q(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle x - x_k, \nabla^2 f(x_k)(x - x_k) \rangle$$

q is a strongly convex quadratic fn. Then there exists a minimizer $\tilde{x}$ of q over $\mathbb{R}^n$

$$\Rightarrow \nabla q(\tilde{x}) = 0$$

How do I write the Newton iterative scheme? So, this question is not difficult enough because know that you are looking at line search problems where your x of k plus 1 k plus 1 with iterative x k plus some scale factor alpha k into d k. But then I shown in the Newton scheme should possibly look like this. This is quite natural, because this is the minus of this is d k and this is alpha k. So, you are replacing d k with minus of this, but is that what we have all have to the Newton scheme, But the question is first now next natural question is how do I know that I have to choose d k in such a pattern? How do I know, how do I arrive at a form of the Newton direction itself? Now, here it is very crux issue which lies at the harder analysis, because it it it takes something from the tailors theorem.

(Refer Slide Time: 13:40)

See if you look at say a function like sign x. Now, if you look at all x which is very near 0, then sign x can probably be approximated by x, because there is very minute difference between x and sign x here. But as you move away from 0 there is a curvature here and the curve actually moves of from the straight line y equal to x. So, instead of having a straight line to approximate the curve which is the steepest descent method, we now use a quadratic function to approximate the curve. And if we do so if we use a quadratic function to approximate the curve, then we would lead to the Newton’s method.

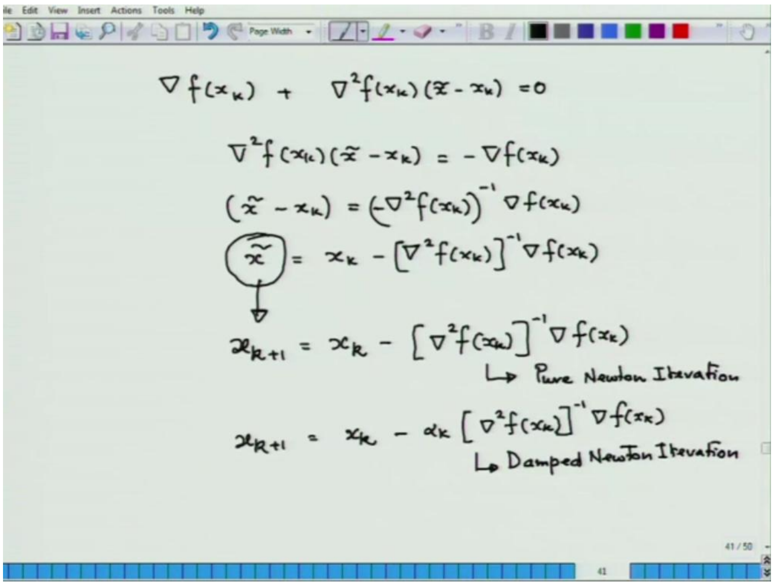
So, we are not using exact function or actual function, but we are using some sort of approximation of the function by instead of approximating it through a linear function. We are approximating it through quadratic function if you approximate it through linear function you get steepest descent method. If you approximated it through quadratic function you will get the Newton’s method. So, let us see what sort of thing that what sort of thing we intend to do here? So, let x k be a point such that so x k is not a minimum point. Now, what I require is the following that. So, consider the quadratic approximation of the function around x bar x k.

So, consider the quadratic approximation of x around x k, which means I will define a function like this q of x is f of x k, so x k is fixed, so x is a variable. So, basically I am taking only the, I am taking the tailor series part if I want to express the function value of f at x in terms of the tailor series. This q x is nothing but the tailor series part without the remainder term without the error term, this is what I have. Now, suppose I have a point I now try to minimize this function. So, instead of minimizing the original function I try to minimize its quadratic approximation around x k. So, how what sort of a function is this in terms of x? In terms of x this is a strongly convex function, we have done a bit possibly about strong convexity, let us go back and have a look whether we have spoken about strong convexity at all. So, we have not spoken about strong convexity, we have not spoken about strong convexity.

So, let us speak about strong convexity, what sort of a function is this? Here, here we have spoken about strong convexity. So, you see we have spoken about strong convexity earlier. So, any quadratic function with a positive definite hessian is strongly convex and it has a unique minima. So, this is we have assumed that grad square of f hessian matrix is positive definite. So, this q is a strongly convex quadratic function then this will always have an minimum. So, you might ask me why it will have a minima? Unfortunately, the total detail explanation of those things are possibly beyond the scope of this course, because this is the foundation of this course, because that would need a lot of more mathematical analysis and details which many of the part people in the audience. For example, those who are in the engineering stream may not appreciate and may not would like may not like to go through it.

So, but it is instructive to know that strongly convex functions will always have a unique minimum and it always would pose as a minimizer then there exists. So, this is always p d that is assumed there exists a minimizer x tilde of q over R n; this would imply that q of x tilde gradient of q of x. This is 0 the standard necessary condition. So, what would this necessary condition give me that is the next question.

(Refer Slide Time: 20:05)



$$\nabla f(x_k) + \nabla^2 f(x_k)(\tilde{x} - x_k) = 0$$

$$\nabla^2 f(x_k)(\tilde{x} - x_k) = -\nabla f(x_k)$$

$$(\tilde{x} - x_k) = (-\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

$$\boxed{\tilde{x}} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

$$\downarrow$$

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

↳ Pure Newton Iteration

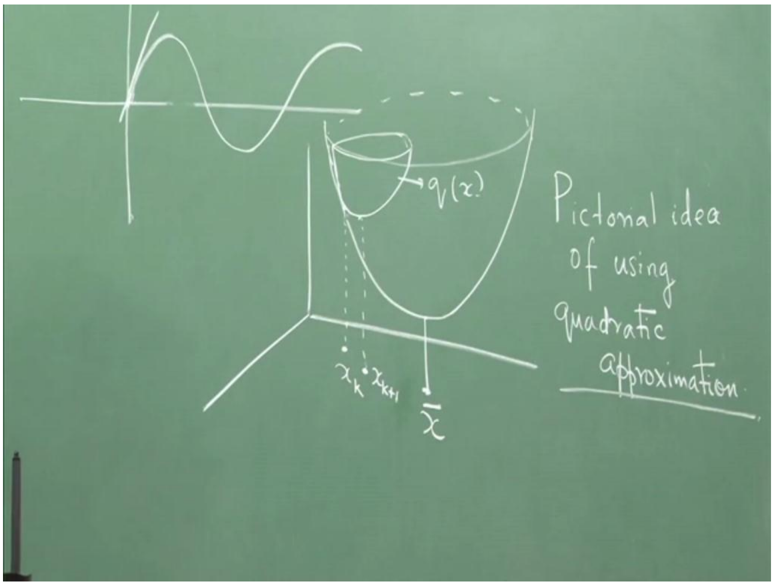
$$x_{k+1} = x_k - \alpha_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

↳ Damped Newton Iteration

See this would give me plus grad square f of x k x tilde minus x k is minus of grad f x k that would lead to the following, this would lead to because of the invertibility, so x tilde, so here I need to put the negative sign, negative sign. So, this can be written as x tilde is equal to x k minus, you will get this one and then observe the following that this x tilde can now be written as x k plus 1.

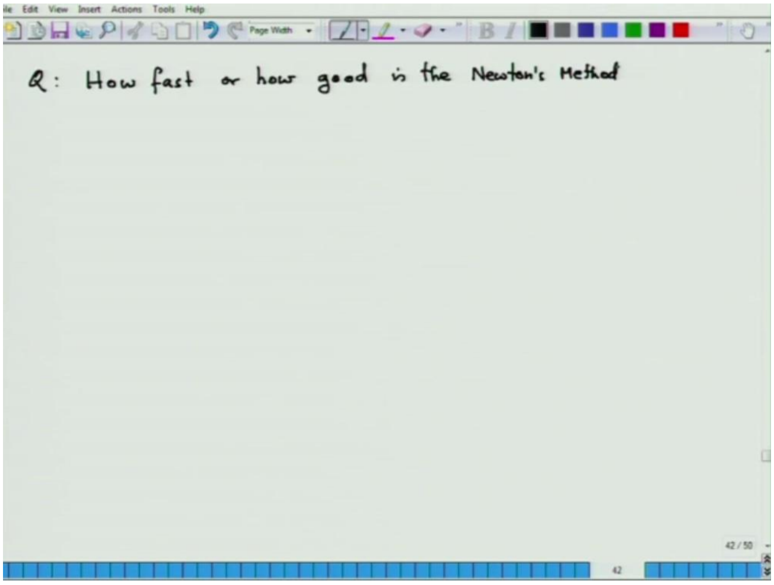
So, my Newton iteration so you might question me what about that alpha the line search parameter you just have the d, you have written x k plus d here the lines such parameter alpha k is equal to 1, we take it alpha k to be 1 a constant 1. So, this is called a pure Newton iteration, and if you take this one that we have written earlier which comes out naturally, finding some alpha k this is called a damped Newton iteration. What is the geometrical idea behind this? The geometrical idea is possibly of this form.

(Refer Slide Time: 23:17)



So, suppose you are trying to minimize some sort of parabolite or some nice looking function of this form I am just trying to give a some sort of 3 dimension view point. So, the minimizer is here actual minimizer and you are now at some x k here. So, you are trying to have around x k, you are trying to get some sort of, of course, f k is here. So, this is your f x k. So, you are trying to get certain some sort of another quadratic approximation to this problem to the function. So, this is your q x and then you minimize this which is so you have come possibly little bit near to x bar. So, this is the pictorical pictorial pictorical idea of the Newton method, idea of using quadratic approximation, because if you are approximating my strongly quadratic problem you are actually getting a nice result what, what we just have pictorial idea of using quadratic approximation. Now the important question is how fast or how good is the Newton method?

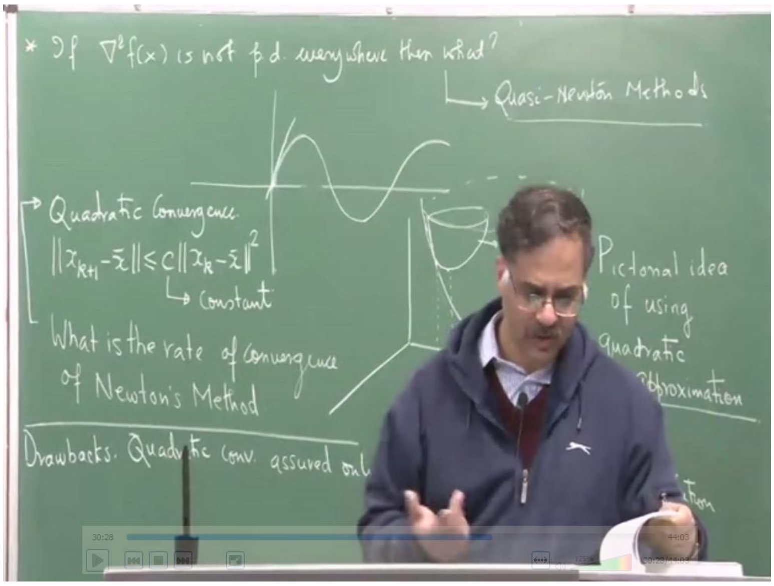
(Refer Slide Time: 25:43)



Q : How fast or how good is the Newton's Method

My question; how fast or how good is the Newton method? So, in other words I am asking this particular question.

(Refer Slide Time: 26:17)



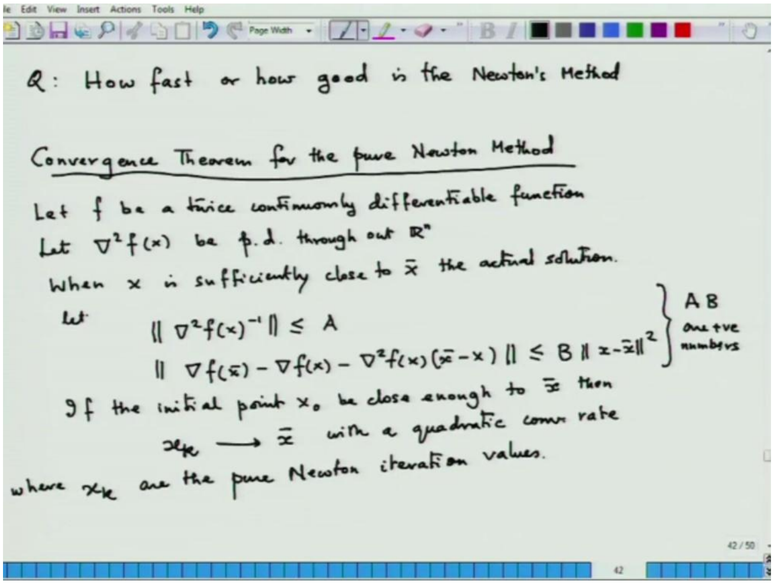
* If $\nabla^2 f(x)$ is not p.d. everywhere then what?
↳ Quasi-Newton Methods
Quadratic Convergence
$\|x_{k+1} - \bar{x}\| \leq c\|x_k - \bar{x}\|^2$
↳ Constant
What is the rate of convergence of Newton's Method
Drawbacks. Quadratic conv. assured onl
Pictorial idea of using quadratic approximation

I am asking that, what is the rate of convergence of the Newton’s method? So, this is an important question, but there is a crux the Newton method is very fast. In fact the rate of convergence the rate of convergence is quadratic that is the that is the difference between x k plus 1 with x bar is less than some constant times x k minus x bar whole square the c is a constant of course. So, this c is a constant, but there is a major crux here you know you cannot say I will start from anywhere any where I will my x 0 should be at any point in R n and I will very rapidly come to the solution.

The Newton method would be the greatest thing of all at least for strongly convex functions may be which are not quadratic you have these sort of these method will just work like magic. But it is not so always it is the problem in Newton method is that quadratic convergence is assured if I start very near the actual solution. But how do I know what is the actual solution? So, the drawbacks are as follows; quadratic convergence assured only if we start very near the solution very near the actual solution. Now, this is the drawback another major drawback is that what would happen if my problem is such that at every point I do not have a positive definiteness of the hessian, which means that if I do not have positive definiteness of the hessian I have to stop it and go back to the slow steepest descent method is in there are something which can help me.

So, question is if grad square f x is not positive that is the problem is not strongly convex is not positive definite everywhere, then what? So, these drawbacks has lead to what will study later as to Quasi Newton method, Quasi Newton methods which is one of the major methods for solving unconstraint optimization problems where there is a very clever way to immediate the Newton method. But handle situations where this is a true. Now, let us write down a important result associated with the Newton scheme with its convergence.

(Refer Slide Time: 30:43)



Q: How fast or how good is the Newton's Method

Convergence Theorem for the pure Newton Method

Let f be a twice continuously differentiable function

Let $\nabla^2 f(x)$ be p.d. through out $\mathbb{R}^n$

When x is sufficiently close to $\bar{x}$ the actual solution.

let

$$\left.\begin{aligned} &\| \nabla^2 f(x)^{-1} \| \leq A \\ &\| \nabla f(\bar{x}) - \nabla f(x) - \nabla^2 f(x)(\bar{x} - x) \| \leq B \| x - \bar{x} \|^2 \end{aligned}\right\} \text{A B are +ve numbers}$$

If the initial point x_0 be close enough to $\bar{x}$ then

$x_k \longrightarrow \bar{x}$ with a quadratic conv rate

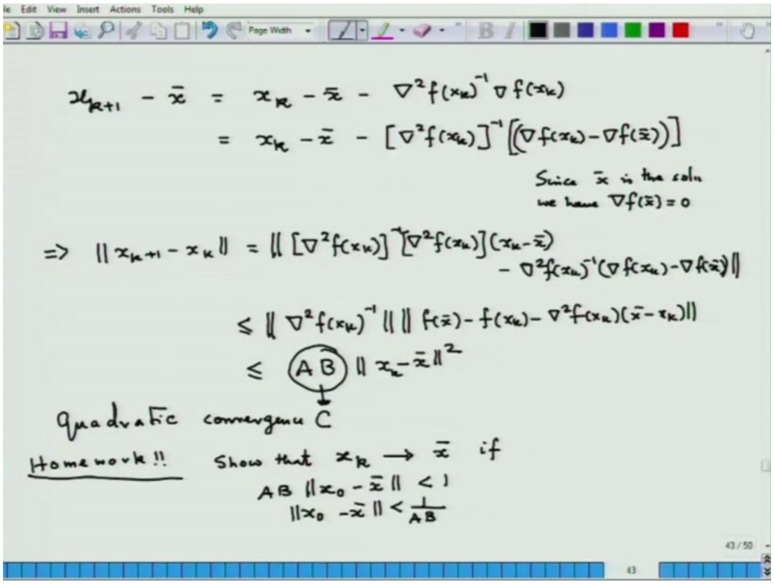
where x_k are the pure Newton iteration values.

So, this is a convergence theorem for new for the pure Newton method. Now, let us write down the result the theorem. So, let f be a twice continuously differentiable function. Now, we have to put assumptions on the hessian etcetera to get the convergence, but there are various authors giving various results we are writing one given by the in the following book mathematical methods of optimization by Lars christen pwares, please have a lot of this mathematical that is the Indian; this is a Indian addition. So, you can also use it but there are many many other approaches. Let grad square f x be p d positive definite throughout R n, then when x is sufficiently close to a sufficiently close to x bar close to x bar the actual solution when x is sufficiently close to x bar the actual solution let the matrix norm.

So, how how do you take the norm of a matrix? So, these are all symmetric matrices you have you have learnt how to take the norm of a matrix, but here possibly there are meaning the operator norm the norm of a matrix. So, those who are little uncomfortable please do not bother much about this terminologies at this movement. And so you are making some two strong assumptions. So, if the initial x 1 if the initial point x naught x 1 initial point x naught be close enough to x bar, then then x k tends to x bar with a quadratic convergence rate. Now, x k where x k are the pure Newton iterations are the pure Newton iteration values there is naught with there is no the alpha k is 1. So, this is a very very now you might ask me what, what this what about this too too much of conditions?

So, in the next class I will try to give a simpler of a certainly a different condition, but let us go and try to do the proof these A and B are constants. So, this A and B here A and B are constants of course, they are positive constraint, because a non negative constant, because here this, these are known straight. And the hessian matrix is out of 0 matrix or something like that, because 0 matrices cannot be invertible tight, because they are not p d matrices because all their Eigen values would become 0. Now, how do you go about proving this fact?

(Refer Slide Time: 36:21)



$$x_{k+1} - \bar{x} = x_k - \bar{x} - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

$$= x_k - \bar{x} - [\nabla^2 f(x_k)]^{-1} [(\nabla f(x_k) - \nabla f(\bar{x}))]$$

Since $\bar{x}$ is the soln we have $\nabla f(\bar{x}) = 0$

$$\Rightarrow \| x_{k+1} - x_k \| = \| [\nabla^2 f(x_k)]^{-1} [\nabla^2 f(x_k)](x_k - \bar{x}) - \nabla^2 f(x_k)^{-1} (\nabla f(x_k) - \nabla f(\bar{x})) \|$$

$$\leq \| \nabla^2 f(x_k)^{-1} \| \| f(\bar{x}) - f(x_k) - \nabla^2 f(x_k)(\bar{x} - x_k) \|$$

$$\leq \underset{\substack{\downarrow \\ C}}{\boxed{AB}} \| x_k - \bar{x} \|^2$$

quadratic convergence

Homework !! Show that $x_k \rightarrow \bar{x}$ if

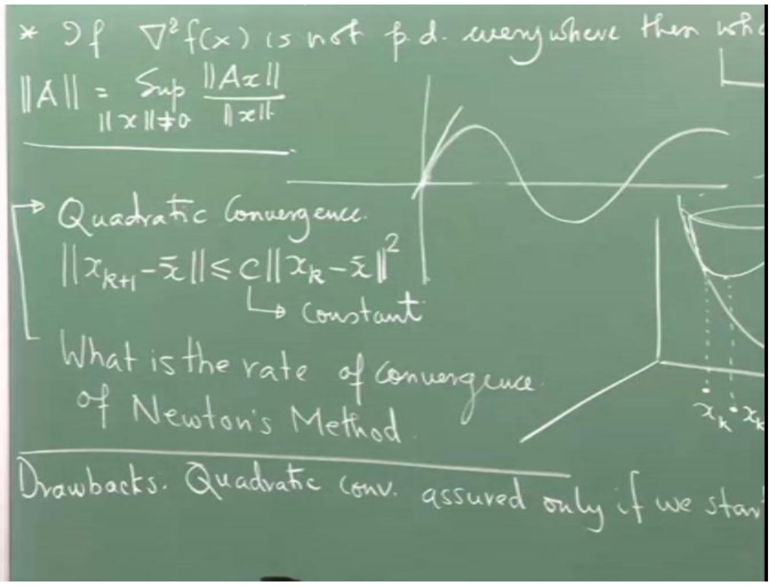
$$AB \| x_0 - \bar{x} \| < 1$$

$$\| x_0 - \bar{x} \| < \frac{1}{AB}$$

Now, from the Newton scheme will use now the Newton scheme so x k plus 1 minus x bar. So, we have to find the quadratic convergence is x k minus. So, I am writing x k plus 1 as x k minus the hessian matrix inverse into grad of f x k, now remember grad of f x bar is equal to 0, because that is the solution. So, I can write this as So, this would imply norm of x k plus 1 minus x k I can write this thing as norm of grad square f x k inverse into grad square f x k this whole thing into x k minus x bar minus grad square f x k inverse into grad f x k minus grad f x bar, we are taking the norm of this whole. Now, this would turn out to be like this that I can take all this, these things out common. So, this would be by Cauchy Schwarz inequality, basically the standard, the norm of a vector is norm of some a or norm of a x is less than equal to norm a into norm x that is the definition of the operator norm.

So, those who do not know about the operator norm please do not get into a fix just just I would like to go in, and tell you that the norm of a matrix the one which is been used here the operator norm.

(Refer Slide Time: 39:14)



* If $\nabla^2 f(x)$ is not p.d. everywhere then whi

$$\|A\| = \sup_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|}$$

Quadratic Convergence

$$\|x_{k+1} - \bar{x}\| \leq C\|x_k - \bar{x}\|^2$$

↳ constant

What is the rate of convergence of Newton's Method

Drawbacks. Quadratic conv. assured only if we star

A norm of a matrix a symmetric matrix here the norm of a matrix a is given as supremum of norm of the vector a x n cosine matrix by norm of x by norm of x is not equal to 0 so that is called the operator norm of the matrix. So, here we are using the operator norm and from here by definition you will see that this, this is a Cauchy Schwarz. This is a supremum of this so each of this is less than equal to norm of a so norm of A x is norm of a into norm of x. So, using that same thing we can write now f x inverse norm into norm of f x bar minus f x k minus grad square f x k x bar minus x k. And this is less than A so a and this whole thing by A because x is now for case sufficiently large.

So, this can be written as by, by the definition here x minus x bar or x k minus x bar whole square so A B is my constant C now this is what is called the quadratic rate of convergence, because C is the here is my this is my constant C. So, I have got the form of quadratic convergence, what is remaining which I want you give you as home work; show that x k actually goes and hits x bar. From this expression, find under what condition x k would actually go and hit x bar? So, I tell you the condition is that if I have A B into norm x naught which is the starting point x naught minus x bar if this is strictly less than 1, then I have a convergence.

So, if so basically, then what I require is that so if x naught is within such a distance from actual solution then x k thus things that we have the pure Newton iteration points will go and hit x bar. And so this gives us a nice way to sum of the quadratic convergence as well and well as well as the convergence of the iteration points. So, this is a brief study of the Newton method and then we will go into a modified Newton method in the next class, maybe telling a bit slightly more and giving you some more examples. And then we will study a very, very important class of method called the conjugate direction method which is, is very important and applied in many many, many places. So, we will do that and we will study conjugate direction method within in quite a big detail. And then give you a brief idea about trust region method before moving to study the theory of constraint optimization.

Thank you very much for today.