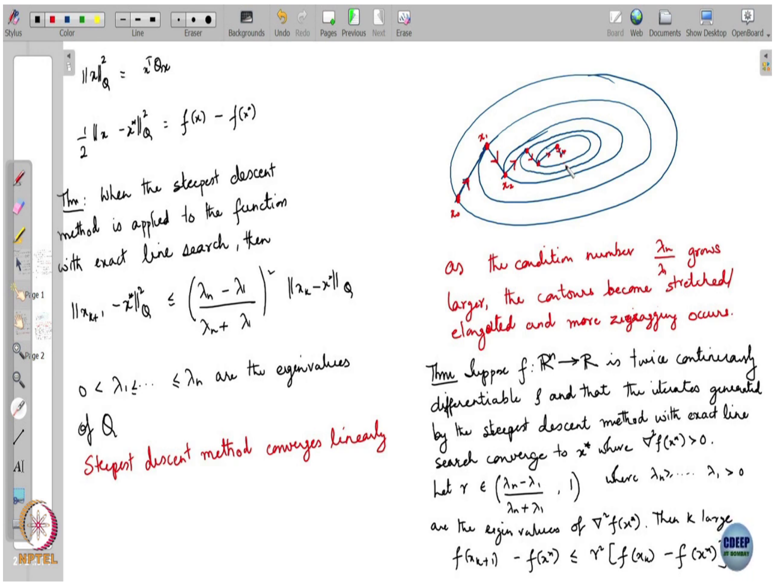
Optimization from Fundamentals
Prof. Ankur Kulkarni
Department of Systems and Control Engineering
Indian Institute of Technology, Bombay

Lecture - 20B
Newton’s method

So, as you can see because of this zigzagging nature of the steepest descent method, actually is the main reason for it to be slow, and it will, and it will be slow when the condition number is large.

(Refer Slide Time: 00:21)



$\|x\|_Q^2 = x^T Q x$

$\frac{1}{2}\|x - x^*\|_Q^2 = f(x) - f(x^*)$

Thm: When the steepest descent method is applied to the function with exact line search, then

$\|x_{k+1} - x^*\|_Q^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \|x_k - x^*\|_Q$

$0 < \lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of Q

Steepest descent method converges linearly

as the condition number $\frac{\lambda_n}{\lambda_1}$ grows larger, the contours become stretched/ elongated and more zigzagging occurs.

Thm Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice continuously differentiable & and that the iterates generated by the steepest descent method with exact line search converge to x^* where $\nabla^2 f(x^*) > 0$. Let $r \in \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1\right)$ where $\lambda_n \geq \cdots \lambda_1 > 0$ are the eigen values of $\nabla^2 f(x^*)$. Then k large

$f(x_{k+1}) - f(x^*) \leq r^2 [f(x_k) - f(x^*)]$

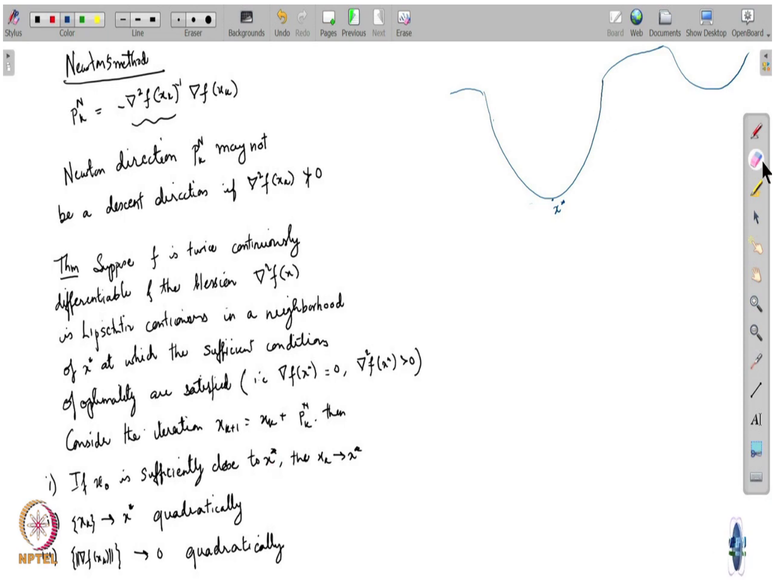
So, what we can now think of is, ok; I we need a method as I mentioned in the previous lecture, we need a method that takes into account that does not simply follow the steepest descent, but rather takes into account also how this direction itself is going to change, right.

So, it must take into account not only the gradient which tells you the direction of the decrease of the function, but also how that gradient itself is going to change which is. And now that the way the gradient itself is changing is captured by the curvature of the function or in other words the Hessian of the function, right.

So, if we have a; so the a better method would be one that takes in that knows at this point itself that although this is the direction of steepest descent, I should not actually be going here because this is not really a sustainable, will not give me a sustainable decrease I would have to again change my direction and go in the, in a in another direction etcetera.

So, what we would need is ideally a method that kind of that is when while you are sitting here itself identifies looking at the curvature and all this other information identifies a better direction to move in, ok. And that is what is a, that is that method is basically Newton’s method.

(Refer Slide Time: 02:02)



Newton's method

$p_k^N = -\underbrace{\nabla^2 f(x_k)^{-1}}\nabla f(x_k)$

Newton direction p_k^N may not be a descent direction if $\nabla^2 f(x_k) \nsucc 0$

Thm Suppose f is twice continuously differentiable & the Hessian $\nabla^2 f(x)$ is Lipschitz continuous in a neighborhood of x^* at which the sufficient conditions of optimality are satisfied (i.e $\nabla f(x^*) = 0$, $\nabla^2 f(x^*) \succ 0$) Consider the iteration $x_{k+1} = x_k + p_k^N$. then

i) If x_0 is sufficiently close to x^*, the $x_k \to x^*$

ii) $\{x_k\} \to x^*$ quadratically

iii) $\{\|\nabla f(x_k)\|\} \to 0$ quadratically

So, that is Newton's method, ok. So, thus, it is again a, it is again method like before, but now we are going to take the, what we are going to do is take a Newton step. So, P Newton at time k, at iteration k is defined in this way. It is defined as the Hessian of at x k is defined as the Hessian of x of the function at x k inverse gradient at x k.

Now, here is one important thing to note. See the Hessian itself may not be a it may not be positive definite, ok. So, Hessian need not be positive definite. And in that case the Newton step or the Newton direction may not actually give you descent, ok. So, the Newton direction may not be a descent direction, may not be a descent direction if this is not positive definite, right. So, as a consequence a lot of what we discussed so far does not actually directly hold for the for Newton's method.

Meaning that, we are not necessarily decreasing at every step with the Newton method. We are, we may not even be getting descent, we may in fact be getting ascent. We may be increasing the objective value, right. So, the so, when applying Newton method we have to be careful that we are actually, we have although the method is intelligent in the sense that it makes takes into account a curvature of the function, we have to make, we have to make sure that if that the that we are in fact, getting descent, right. So, ok.

So, let us. So, let us now discuss the rate of convergence of this, of the Newton method. Now, one other thing I want you to note here is in the Newton method is that the Newton's step is itself has baked in it already the step size. The step size has been has already found, has already been found for you by taking into account a curvature of the function through the Hessian. So, it one does not usually need in addition to this another step size.

Because the Newton step is a complete step meaning that it is not just a direction. It is a complete, it is a complete step to a, to the new to the next iteration alright, ok. So, here is the theorem. Suppose, f is twice continuously differentiable and the Hessian the at x is Lipschitz continuous in a neighbourhood of x star at which the sufficient conditions of optimality are satisfied. Now, what does this mean sufficient conditions of optimality are satisfied? It means that it is a point at which gradient is equal to 0.

And the Hessian is positive definite, ok. So, that means, the sufficient conditions of optimality are satisfied. And consider the iteration; so, right, show you in the bracket. So, the gradient at x star is equal to 0 and the Hessian at x star is positive definite, right. Consider the iteration x k plus 1 equals x k plus P N k, where P N k is as defined above. Then, if the starting point x 0 is sufficiently close to x star, then x k converges to x star and the iterates converge to x star.

Moreover, x k converges to x start quadratically. That means, if you look at the error between x k and x star that error decreases to 0 quadratically. Third, if you look at the norm of the gradients this norm of the gradients this also converges to 0 quadratically, ok.

So, here, so let us take a note of a few things. So, I mentioned to you; look at let us see what the theorem is saying that the if f is twice differential continuously differentiable and Hessian is Lipschitz continuously in the neighbourhood of x star at which the sufficient conditions of optimality are satisfied. That means, it is a local; that means it is a local minimum of your function f. And you look at this iteration which is x k plus 1 equals x k plus P N k, where you are taking a Newton step.

Now, if x; now here is the here is the main thing. If x 0 is sufficiently close to x star then you are guaranteed that this will converge to x star. That means, if you are sufficiently close, what does this mean? You are in a, in the neighbourhood of x star, means around the, you are in a part around of the space around x star where the function f looks convex. See, if the functions the function may very well do other things elsewhere, but around x you are close enough to x star.

So, at x star the function has a gradient equal to 0 and Hessian positive definite. So, in that in a neighbourhood around x star the function is actually convex and what this is saying is that you are starting your iteration in that sort of neighbourhood you are starting your iteration in the region where the function is convex.

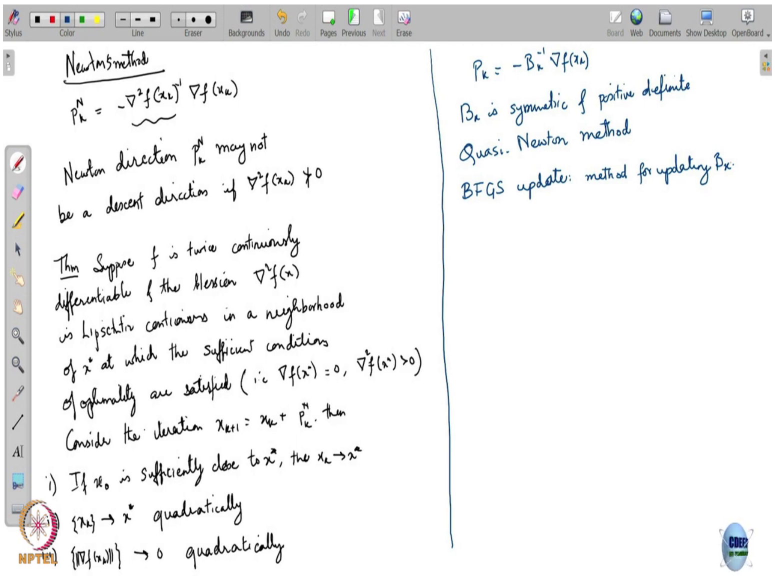
If the function is convex, then in that case if you look at the Newton direction, the Newton direction in that case that if the Hessian, then if the function is convex then the Hessian actually is positive definite, right. If the Hessian is positive definite this is actually then a descent direction.

And then, so if you start your iterations in this sort of region, then the Newton direction is a descent gives you a descent direction, ok. So, and so what does this theorem say? That if you start your iteration somewhere close enough then the iterates converge to x star. Moreover they converge quadratically and even your gradient vanishes quadratically. The norm of the gradients also vanish quadratically.

So, if your, so if you start your iteration somewhere in this sort of region then the Newton method not only converges it converges quadratically, it converges faster than the steepest descent method. So, the main thing here is that because Newton direction is not guaranteed to be a direction of descent, we have to include this rider that you are starting in this kind of basin of attraction where the function is actually in a neighbourhood of the 2 minimum where your function is convex.

Now, if the function actually is convex everywhere then things become easier and then this sufficiently close really is has no bearing on the final, on the final result. But the main thing to note here is how we have got now quadratic convergence as opposed to linear convergence which is what, which is what steepest descent was giving us. Now, so we can also do other variants of a of the Newton method. So, we one can do for example.

(Refer Slide Time: 13:17)



Newtons method

$p_k^N = -\underbrace{\nabla^2 f(x_k)^{-1}} \nabla f(x_k)$

Newton direction p_k^N may not be a descent direction if $\nabla^2 f(x_k) \nsucc 0$

Thm Suppose f is twice continuously differentiable & the Hessian $\nabla^2 f(x)$ is Lipschitz continuous in a neighborhood of x^* at which the sufficient conditions of optimality are satisfied (i.e $\nabla f(x^*) = 0$, $\nabla^2 f(x^*) \succ 0$) Consider the iteration $x_{k+1} = x_k + p_k^N$. then

i) If x_0 is sufficiently close to x^*, the $x_k \to x^*$

ii) $\{x_k\} \to x^*$ quadratically

iii) $\{\|\nabla f(x_k)\|\} \to 0$ quadratically

$p_k = -B_k^{-1} \nabla f(x_k)$

B_k is symmetric & positive definite

Quasi. Newton method

BFGS update: method for updating B_k.

One can do instead of taking; so, in situations where the Hessian is not positive definite, one can do, one can come up with variants the of a new of the Newton method that tend to make the behaviour of the Newton method. So, for example, you could do simply P k equals some minus B k inverse gradient of f at x k, where B k is symmetric and positive definite, right. So, B k is symmetric and positive definite. So, this kind of iterate is what is called a Quasi-Newton method, Quasi-Newton method.

Now, what this tends to do is it brings you a little bit of information about the curvature alongside and also gives you the properties of guaranteed descent. Now, the way the B k is obtained is through the past derivatives and past information that you have obtained about the function. And so, there are, many different ways of updating this B k, one of the, one of the, one of the sort of most popular updates is what is called the BFGS update.

So, we do not have the time to go into all details around all this. The BFGS gives you a update. It gives you a way of updating method for, it is a method for updating B k. Now, what sort of, what sort of result can we get for this kind of a; alright. So, now; so, now, with this, I think there is this gives us a wide gamut of methods for solving unconstrained optimization problems. So, from here now we will move on to constraint optimizations.