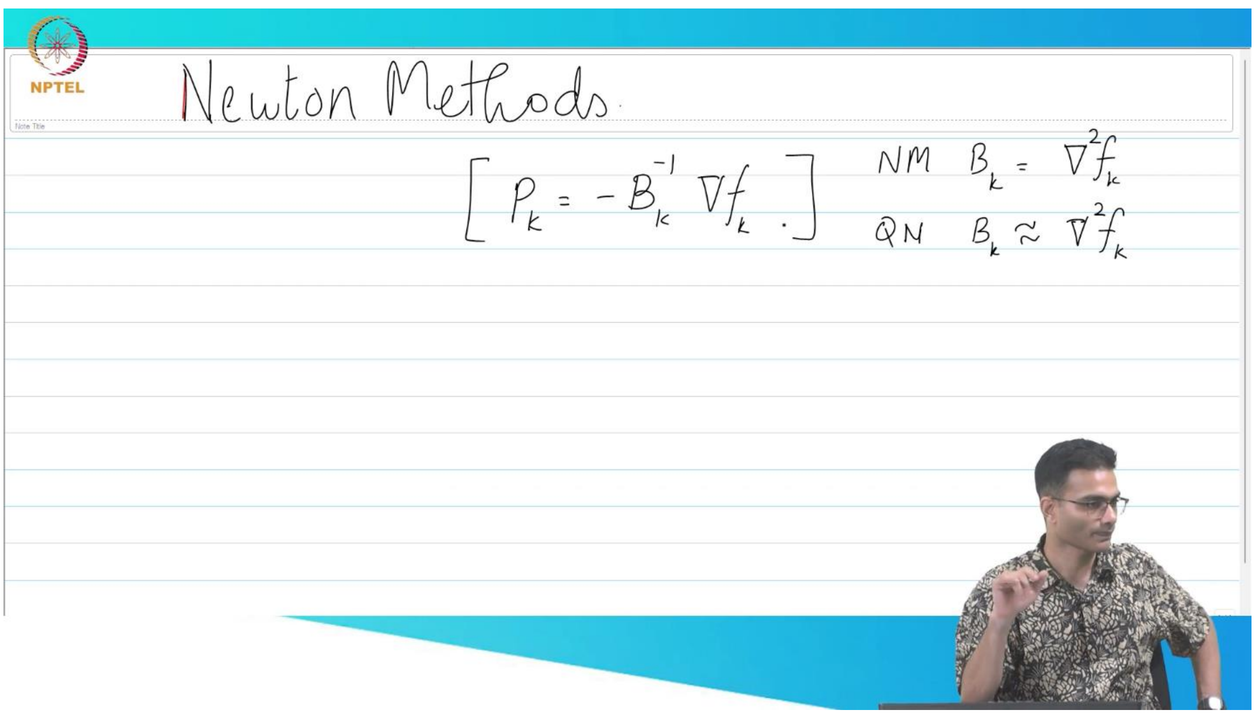
**Course Name: Optimization Theory and Algorithms**
**Professor Name: Dr. Uday K. Khankhoje**
**Department Name: Electrical Engineering**
**Institute Name: Indian Institute of Technology Madras**
**Week - 07**
**Lecture - 50**

# Newton methods and convergence

Yeah, without knowing the matrix how can vector matrix products be computed is the question. For example, if you want to know the curvature condition I need to know the Hessian. Now the point that I was trying to make earlier is we obviously need to know the Hessian. If I do not know something how can I calculate the product of that something with a vector? The point is how I store the matrix is where the innovation lies. I need not store the matrix in a dense way. That is what it means that matrix vector product is inexpensive if I mean can be evaluated efficiently as compared to storing $n^2$ entries.

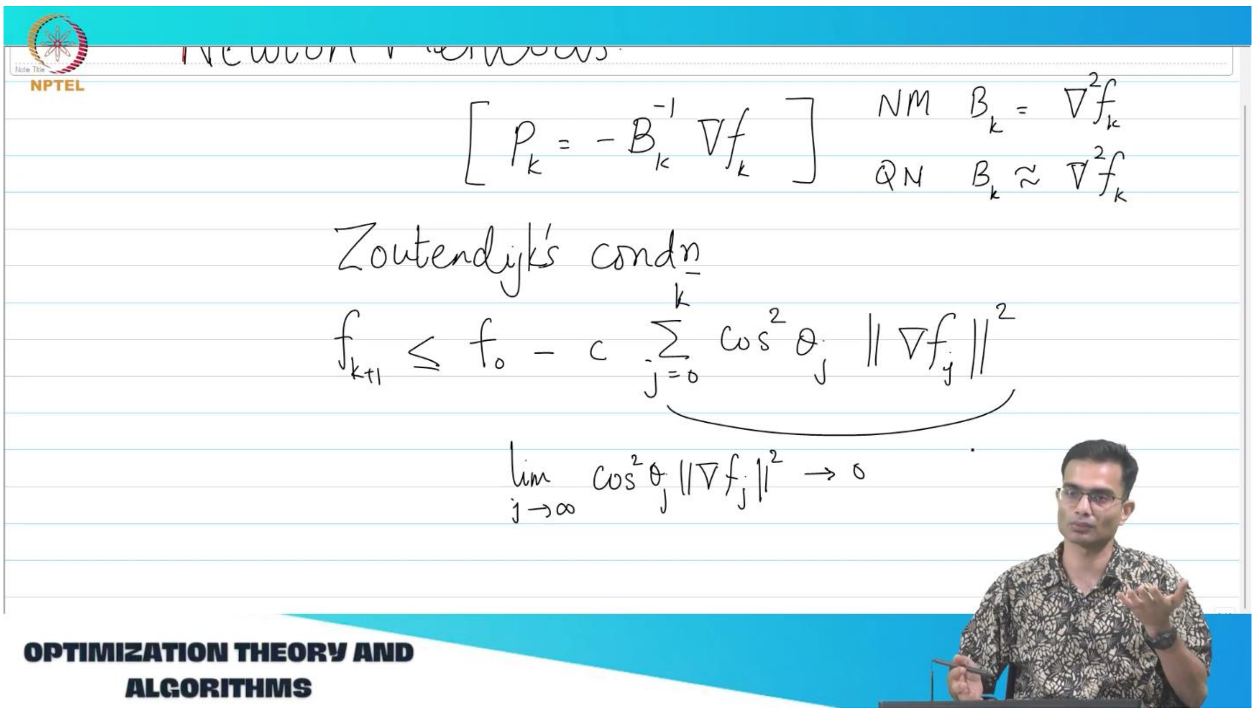
So, anyway this point we will not be covering in this course because it is a point of numerical linear algebra implementation. So, whoever has this doubt just come and talk to me after class. For non-linear CG method, is there a guarantee of convergence in $n$ steps like in linear CG? The point is that where did in the linear CG, where did $n$ come from? How did $n$ enter the problem itself? The size of $A$, it was $n \times n$, right. So, now there is no $A$ matrix in the non-linear CG method.




So, $n$ steps there is no guarantee. Even if your input vector is of length $n$, there is no guarantee of $n$ steps convergence. And is it only for convex methods? It is not only for convex methods, it is for general non-linear functions. Why cannot we do a linear slash quadratic approximation of the function followed by linear conjugate gradient method? Very good question. So, we are trying to do, we have a non-linear function, we find it difficult.

Why do not we linearize it or make it quadratic and then do linear CG? What is the problem with that? So, if I draw a complicated function like this, let us say ok, it would not be like this right. If I am at let us say at this point over here, if I do a linearized approximation I will get basically this line. So this line will help me in one small neighborhood, right. So if I am within this small neighborhood, it is approximately correct. The moment I go outside of it, again I have to update.

So if I do linearized approaches, they work and that is one legitimate way of doing it. You can be fancier, you can say let me approximate it here by a quadratic, I will get something like this, right. So if I am lucky enough to be close to the solution, it will work. But if I am sitting somewhere over here, now if I fit a quadratic, then supposing I fit something like this. Do I do? Basically the solution is go to either end and you will get minus infinity.

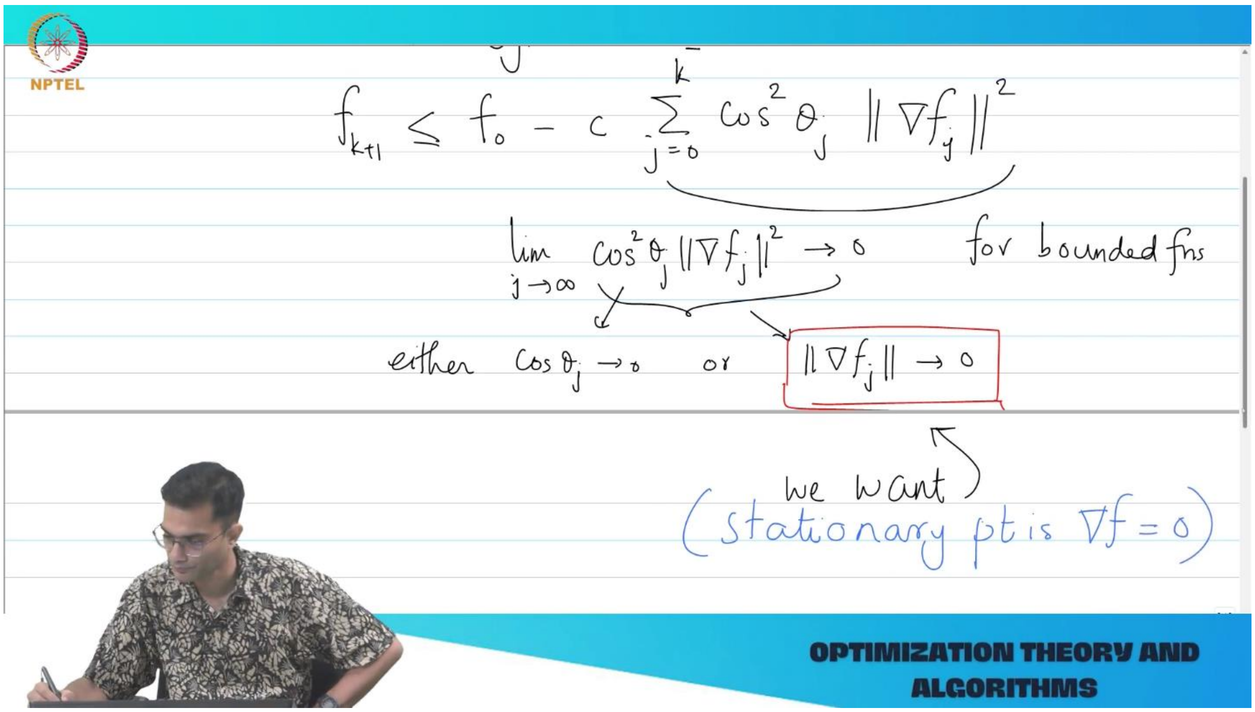



So, these tricks are used in fact what we will see later in quasi-Newton methods some kind of tricks like this are done. But what we have is the non-linear method which is a little bit more general than restricting yourself to first order or second order model. How are we going to approximate the Hessian matrix? We will come to that today. So, we are talking about Newton methods. So, the general recipe of the Newton method we saw last time that I am going to write the search direction as something times the gradient of the function.

What is that something? Not $\beta$, $\beta$ would be dangerous, $B$. Some matrix $B_k$ inverse and $B_k$ was what? So, now $B_k$ could be either depends on the context. In the Newton method $B_k$ was equal to Hessian and in quasi-Newton method approximation of Hessian. So, this common notation will cover both methods. When you see this do not immediately conclude that it is Newton or quasi-Newton, you have to ask further what it is, ok.

What I am going to do is I am going to show you a small property of convergence. We should have some proof that this method converges right. In the steepest descent method we had shown that the method converges. In the conjugate gradient method also we had shown that it converges in fact in at most $n$ steps. So, similarly there is going to be some proof of convergence.

The proof techniques are similar to what we did for steepest descent. So, there is nothing new really for you to see except just a reapplication of how. So, we look at that, then I will show you what the rate is, the rate is quadratic which is why Newton methods are so attractive, ok. And then we will come to these quasi-Newton methods which are also very nice, ok. So, that is roughly what we are going to do. So, does, do you remember this Zoutendijk condition? We had done this quite about a month ago right.



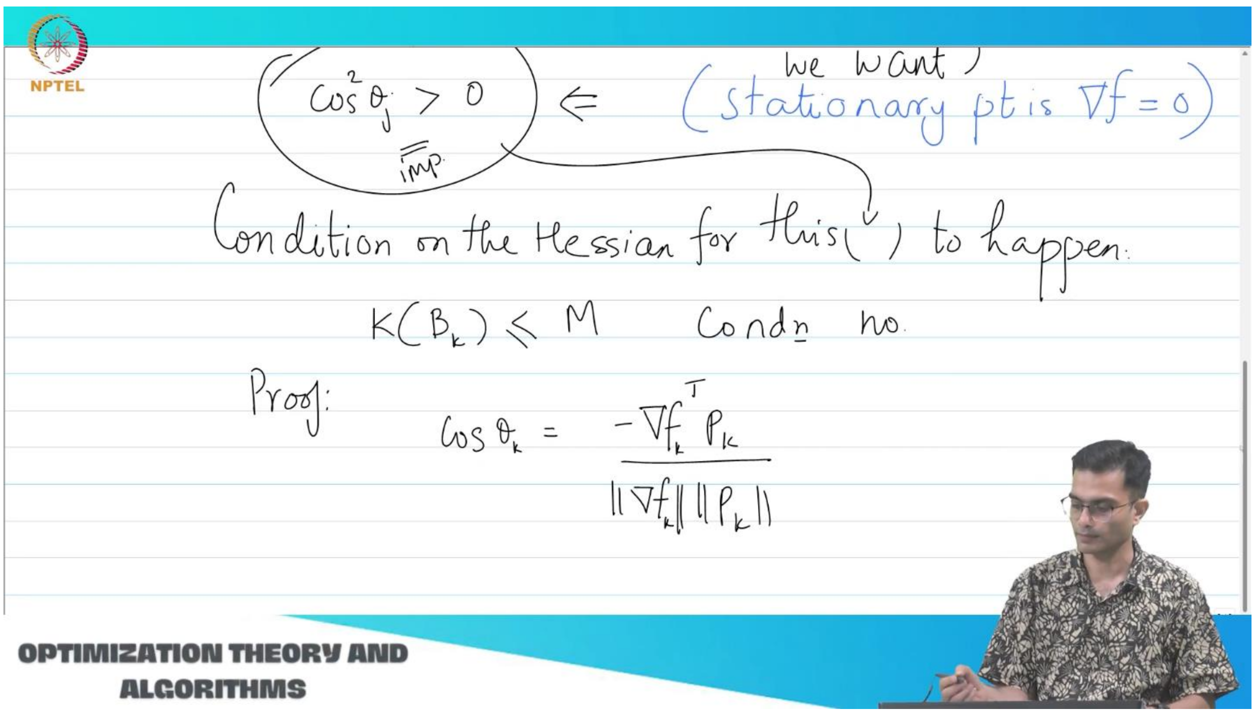

This was when we were trying to prove that the steepest descent or any line search method converges. So, the Newton method the way we wrote it is also a line search method. The only difference is the direction $p_k$ is different from steepest descent ok. So, this was also I think one of the last questions of your quiz. I had something like this.

So, this is this was telling us how $f_k$ was reducing in terms of this running summation, right. And we had said that if you look at the what is inside over here, this term will always be positive obviously, right. So, this summation $j$ is equal to 0. As $j$ tended to very large number of iteration steps, what did we want for the method to converge? This term $\cos(\theta_j) \times \nabla f$ should be tending to 0, right. Why do we want it to tend to 0 by the way? Can someone answer in plain English? What if it did not tend to 0? Then $f_k$ would become unbounded.

Even as I am going to infinite terms, I am continuously adding something finite, adding something finite, that would mean that $f_k$ would become unbounded. On the other hand, we started the whole discussion by saying that it is common sense only to optimize functions which are bounded from below. They do not go to minus infinity. For example, if someone says optimize $\log(|x|)$, find a minima of $\log(|x|)$.

We know that as $x$ tends to 0, this thing goes to what? Minus infinity. So, it is nonsensical to optimize it. So, we start off by saying let us deal with functions that are bounded from below. So, that means this should tend to 0 for bounded functions, ok. Now, for this to tend to 0, what are the two possibilities? Either the first term is 0 or the second term is 0 obviously right.

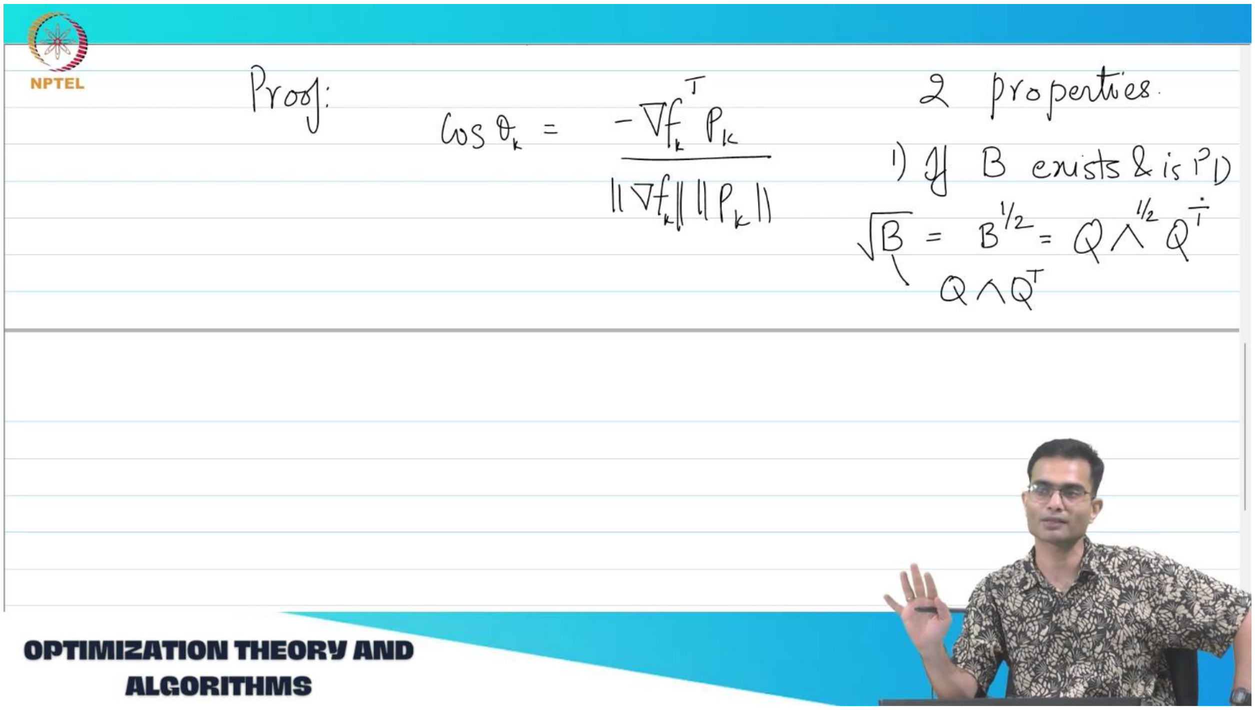
So, either $\cos\theta_j$ tends to 0 or the norm of this. Which one do we want? If you want the method to converge to a stationary point, what do we want? The second guy, right? The second guy had better be true, ok. Why? Because stationary point is $\nabla f = 0$, this is why we want. Straight away the implication therefore is that this $\cos\theta$ term therefore has to be strictly greater than 0, otherwise, my stationary point condition may not hold. So, $\cos^2\theta_j$ should be strictly greater than 0, ok, not greater than or equal to, but strictly ok.




So, in the case of the Newton method, what ends up happening it is quite easy to show is that there is a certain condition on the Hessian, ok. So, there is a condition on the Hessian condition on the Hessian for this, by this, I mean. And that condition is that the condition number of the Hessian should be, should have a bound. It should not be arbitrarily large, ok.

So, that condition is. So, $\kappa(B_k)$ is less than $M$. Now, we will show why this is the case, ok. This is the condition number. That kind of, before we get into some algebraic proof for this, can you think of an intuitive reason why this makes sense? Why should we have a bound in the condition number? What is condition number doing after all? Remember what was, it is amplifying the errors. So, obviously if the amplification goes very very large, every time you make a step in a certain direction you are going to have some errors over here and this is a proof of convergence, we want the method to converge.

So, intuitively you can expect that putting a control on the condition number is going to help us to arrive at the solution. This is a very very rough hand-wavy argument, but we can look at it a little bit more precisely, ok. So, what is, how does this lead to $cos\theta_j$ being greater than or equal to 0? So, let us look at the proof. Now, this $cos\theta$ was the angle between what and what? This is related to what we call the descent direction.

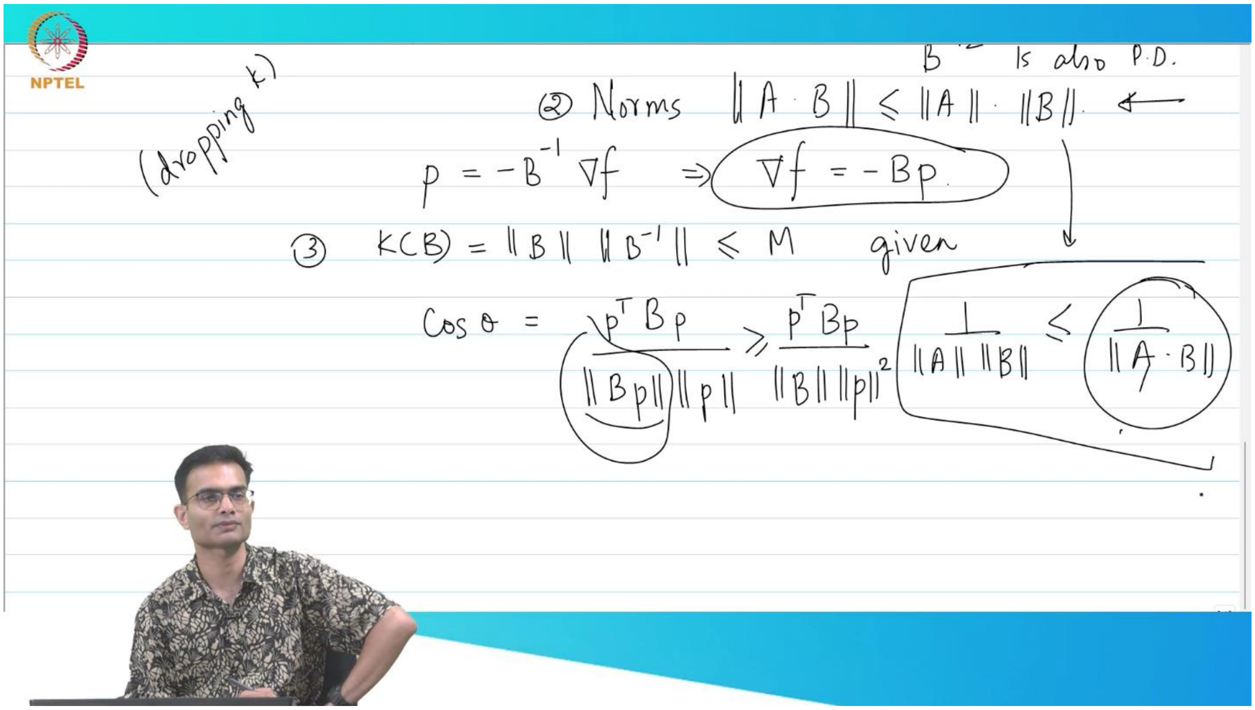



So, $p_k$ and $\nabla f$, right. So, $p_k$ and $-\nabla f$, right, the gradient, the negative gradient. So, $cos\theta$ therefore is the angle here $-\nabla f^\top p_k$ and normalize it. Yes. So, here I am using the everywhere I have a subscript. Now I am going to use two properties of norms to come to this result, two fairly straightforward properties of norms.

So, that these are, if this exists and is positive definite, supposing the matrix $B$ or $B_k$ exists and it is positive definite, can we say anything about $\sqrt{B}$? Will it also be positive definite? First of all, what do I mean by square root of a matrix? Does it make sense? Right. So, has anyone seen this ever written before for a matrix? It kind of looks funny, right? But what is it effectively? It is $B^{1/2}$. Now, if I tell you $A \times A$, you would say, oh, there is no problem multiplying a matrix with a matrix. So, $B^{1/2}$, what would it end up being? So, it is going to be that matrix which when multiplied with itself gives me back $B$. That is one way of looking at it, right.

And there is a very simple way of writing it. So, if I had the eigenvalue decomposition of $B$, what would it be? So, if $B$ was $Q\Lambda Q^\top$, then $B^{1/2}$ would be what? So, the square root simply transfers over here. And you can convince yourself that if I multiply two of these with each other what will I get? $Q^\top Q$ inside will become identity, $\Lambda^{1/2}$ multiplied by $\Lambda^{1/2}$ will become $\Lambda$, I get back $B$. So, this is legitimate candidate for $\sqrt{B}$, I am going to work with that. Now, is this guy positive definite? Clearly it is positive definite.

So therefore, this is also PD (positive definite) all positive definite, ok. Similarly, I can also talk about $B^{-1/2}$, it makes as much sense as $B^{1/2}$. All that I have is negative $1/\sqrt{\lambda}$ will appear in the diagonal term also is also positive definite. The second is the, is a property of norms which simply says that if I have the norm of a product, this is less than or equal to norm A times norm B. This is again follows from linear algebra, we will not prove it over here. Now, our search direction I am writing, I am dropping the $k$ for norm because we are sitting at the $k$'th iteration also.

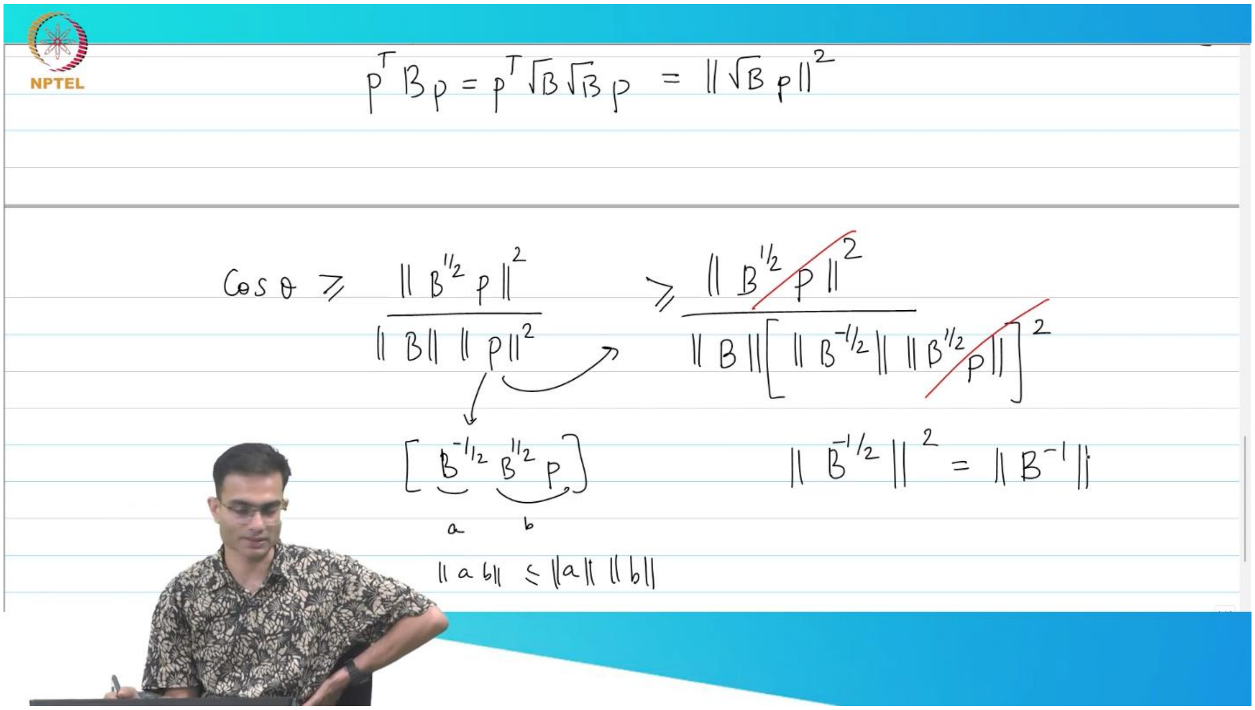



So, let us just write over here dropping $k$. $P$ was simply $-B^{-1}\nabla f$, right. This is for the Newton or quasi-Newton method, which in other words I can write this as $\nabla f = -B \times P$, ok. Now, what are we, the thing to keep in mind is this is the guy and what do we have to prove? What is the end goal? I have this expression for $\cos\theta_k$, what do I have to prove? It is positive, right? And I have to somehow relate it to this condition number business, ok. Now, you can see this property number 2 which I have used is going to help me in which part of this expression? Norm of A into B is less than equal to norm A into norm B, where can I, I can possibly use this somewhere, right? In the numerator or denominator, there are both places depending on how I take it. The third thing, what was the definition of the condition number? Just like the textbook definition.

No, that is a more applied definition, the ratio of singular values. That is one correct definition, but there is another way in terms of matrix norms, right. So, this was the product of norm of $B$ multiplied by norm of $B^{-1}$, ok. And we are given that this is less than or equal to $M$.

This is given to us in the problem, ok. So, now I am going to take my cosine expression and notice your cosine expression is the matrix $B$ inside it anywhere? It is only in terms of $\nabla f$ and $p$. So far $B$ is not there. If $B$ is not there, $M$ cannot enter the condition number, right. So the first trick would be to now somehow get $B$ into the picture. Is there a nice way by which I can get $B$? Yes, right, because $\nabla f = -B \times p$.

So that is our first clue, right. That let me at least get $B$ into the picture, then I can put some inequalities on it. So this guy over here, I am going to substitute and get rid of which guy? $\nabla f$. So, $\cos\theta$ would be equal to $-\nabla f$ would become, the minus and minus would become a plus. I have a $p^\top B$, $B^\top$ can I write it equal to $B$? Hessian is always symmetric because the mixed partial derivatives are sitting there, right. So, then I have a $p$, this is the numerator. And denominator has $\nabla$, so there is a sorry $B \times p$ and then what else do I have? I have a $p$.



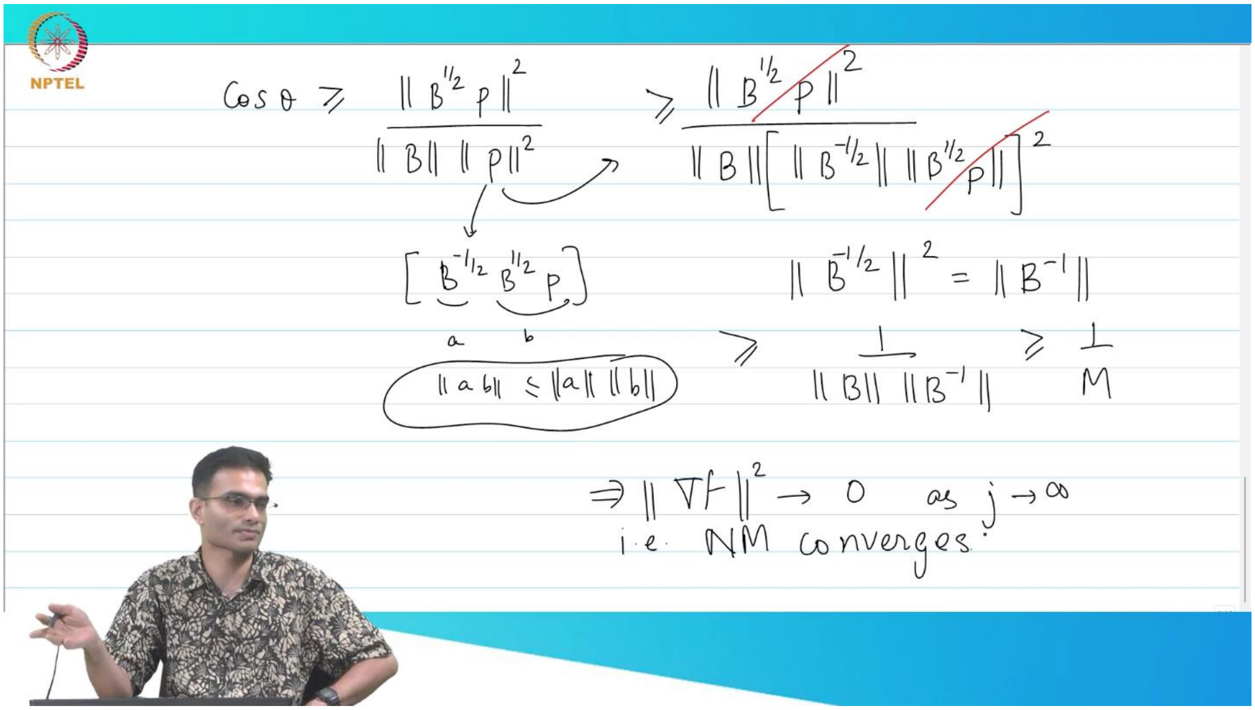

So, what we can do is we can take our this inequality from 2, ok. So, inequality from 2 is saying that $\| AB \|$ I can split as $\| A \| \| B \|$, right. If I take it in the reciprocal, what will this be? $\frac{1}{\|A\|\|B\|}$ would be all positive quantities I can do this. So, if I apply this over here, I am trying to open up this guy over here, right. So, this expression over here would be I can write it as greater than or equal to something, right.

This expression can be written as greater than or equal to what will I have? Numerator stays as it is. Greater than or equal to I am, this is the guy that this is think of this oops think of this as this ok. I should have written it that way. So, this is greater than or equal to I am going to get a $\| B \|$ and then $B^2$. Now, can we simplify this numerator in some nice way? Can I write it as the norm of something, norm squared of something for example? The hint is the square root $B$ guide that I had introduced, $p^\top B p$, can I write it as $p^\top \sqrt{B}\sqrt{B} p$? I can write it like that.

And will $\sqrt{B}$ also be symmetric? It will also be symmetric. So, this guy can therefore be written as $\| \sqrt{B} p \|^2$. Can I write it like that? So, I am trying to now simplify this numerator expression over here. So, what do I get? Is greater than or equal to I have a $\| \sqrt{B} p \|^2$, ok. And denominator I have $\| B \|$ and I have a $\| p \|^2$, ok.

Now, let us, we are basically what we are trying to do is somehow get rid of some of these extra $\sqrt{B}$'s just have one $B$ over there, then I can replace this condition number over there. This $p$

itself I can write as $B^{-1/2}, B^{1/2}$, supposing I write this $p$ like this, is it correct? $B^{-1/2}$ multiplied by $B^{1/2}$ and $p$, right? Now you can see what I am trying to do with this. I am going to try to split it in this way so that $B^{1/2}$ times $p$ is there in the numerator and denominator, that is going to be the strategy. So, let us get rid of this over here. So, this can further I am going to write this as $B^{1/2} \times p$ squared.

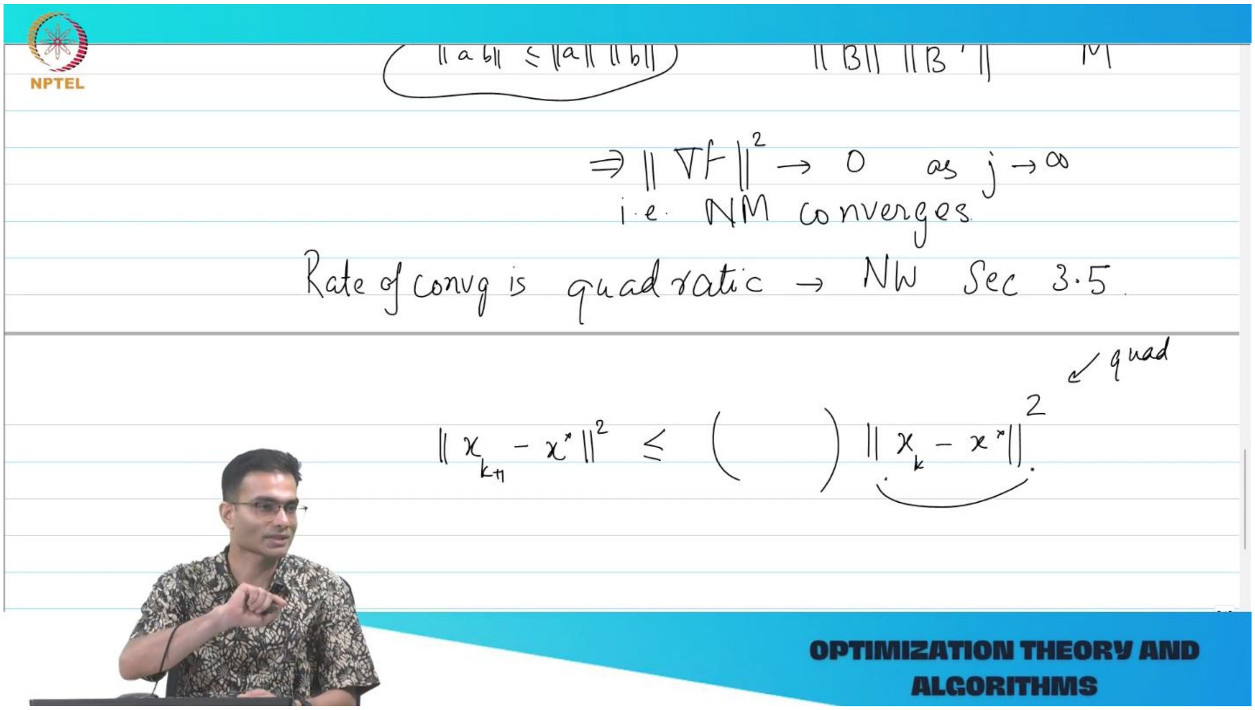



This $B$ remains over here, ok. And now then here is where I split it open. So, I have $B^{-1/2}$, then $B^{1/2} \times p$. So, what I did. Is the inequality still valid? I used that $\| A \times B \|$ is less than or equal to $\| A \| \times \| B \|$. So, I have used that trick a second time, right. I wrote it like this and then I club this as this is my $A$ term, this is my $B$ term and then $\| A \times B \|$ is less than or equal to $\| A \| \| B \|$.

So, when it goes into the denominator, it reverses. Now, can you see what needs to, what happens? Does the numerator cancel off? I have $B^{1/2} \times p \|^2$ is there. Is it there also in the denominator? It is also there in the denominator, right. So, this guy and this guy cancels off. There is another simple enough property of linear algebra that this is equal to, you can take the square inside basically, right. So, what do I get? This becomes $\frac{1}{\|B\|}$ and then this becomes $\| B^{-1} \|$, right.

And this expression is what? This is related to our condition number, right. So it's a little bit of, it's not difficult. It's just, if you know that you want to somehow get in the condition number there, I need to get $B$ and $B^{-1}$. These two have to appear together.

And so I repeatedly use this norm property over here to somehow keep splitting it until I got rid of the numerator and I was left purely with $B$. If I was stuck with a $p$ somewhere then I am in trouble because $p$ could be anything depending on the iteration, right. Here I have managed to get it regardless of the direction $p$, I only got it in terms of $B$. Now regardless of the iteration if you tell me that the condition number is at max $m$, this condition will always be valid. That means cosine $\theta$ is always greater than $1/m$, do whatever you want.




If that is the case then if you look back over here, if this is greater than 0 that means the only way this Zoutendijk condition is going to work is if $\nabla f = 0$. And if $\nabla f = 0$ that means I have arrived at a stationary point. So, these proofs of convergence typically will use these kind of inequalities, maybe Cauchy-Schwarz and this norm inequality and then a whole bunch of clever algebra. The first time you see it, it looks like you know rabbit being pulled out of the hat or you see it a few times, these are, you know, few standard tricks, nothing much to it. So, this we get this condition, therefore $\nabla f$ tends to 0 as $j$ tends to infinity, Newton method convergence.

Again practitioners almost they do not care about this because it is, you know, they care more about the rate at which convergence happens, ok. So, I am going to state without proof that this the rate of convergence is quadratic, ok. So, the rate of convergence is this is in Nocedal and Wright I think this is theorem or rather look at section 3.5 where they have a proof for this, ok. So, this is as I have been saying many times this is the reason why Newton methods are very popular because you get quadratic, right.

So, quadratic just simply means that if I look at this distance $x_{k+1} - x^*$, and this was less than or equal to some constant over here and then I have $x_k - x^*$. What is the difference between linear and quadratic? You are going to get a square here ok. So, this is from quadratic. If you had linear this would be to the power 1. So, this is showing you that as I get closer and closer to the solution this error in some sense is getting is going to 0 at a much faster rate, right.

Imagine that at step $k$ this guy $\| x_k - x^* \|$ was 0.1. In a linear method, the next iterate had to be less than constant times 0.1, but in a quadratic method this has to be some constant times 0.01. So, it is going to towards 0 at a much faster rate that is why quadratic is a huge deal, right. It is not just a scaling factor or a linear factor it is an exponential, I mean power 2 by which it is going.

Okay, so this is the whole reason for the charm of Newton method. Any questions on this? The proof just has, you know, some getting used to by looking at it a few times and then you master it.