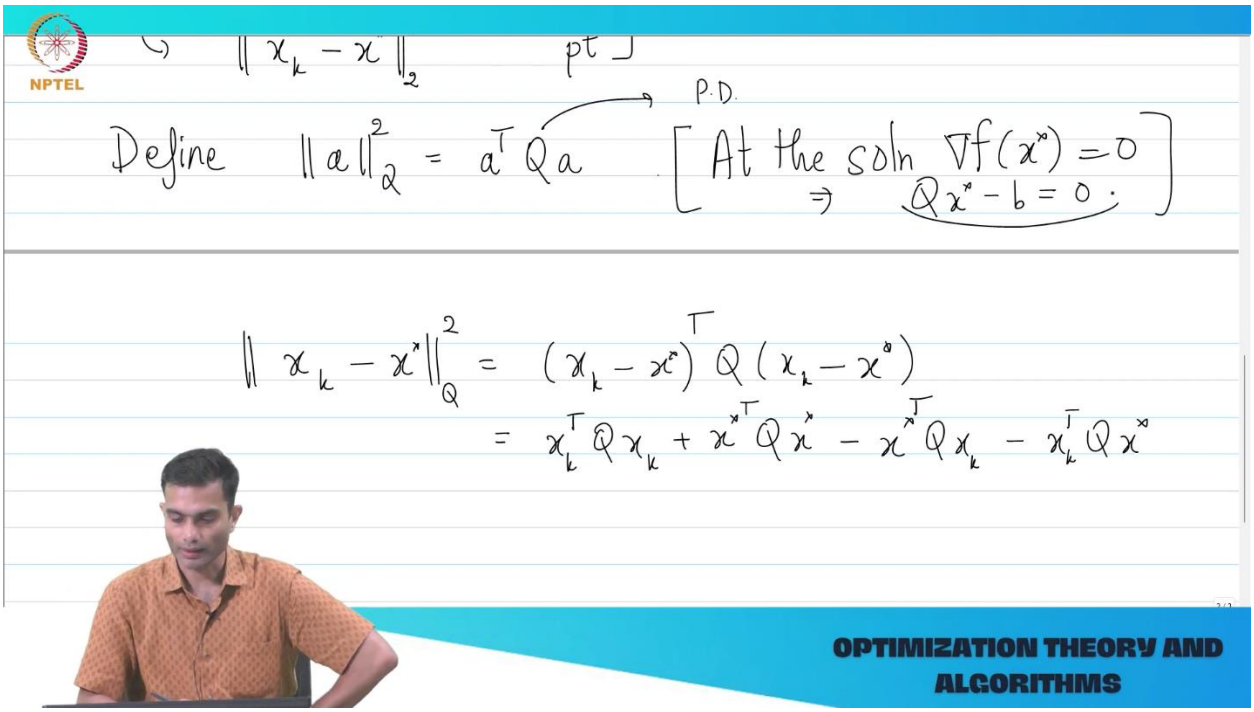
Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 05
Lecture - 31

Convergence analysis of a descent algorithm - 2

All right. So I'm going to use this norm as in the rest of the analysis. Now we should know where we are heading. We are heading towards a stationary point. If I'm heading towards a stationary point, what is one of the qualities of the stationary point? How will I identify it? $\nabla f(x^*)$ should be 0, right. We have studied this right in the beginning that the signature of a stationary point is its gradient is 0, right.

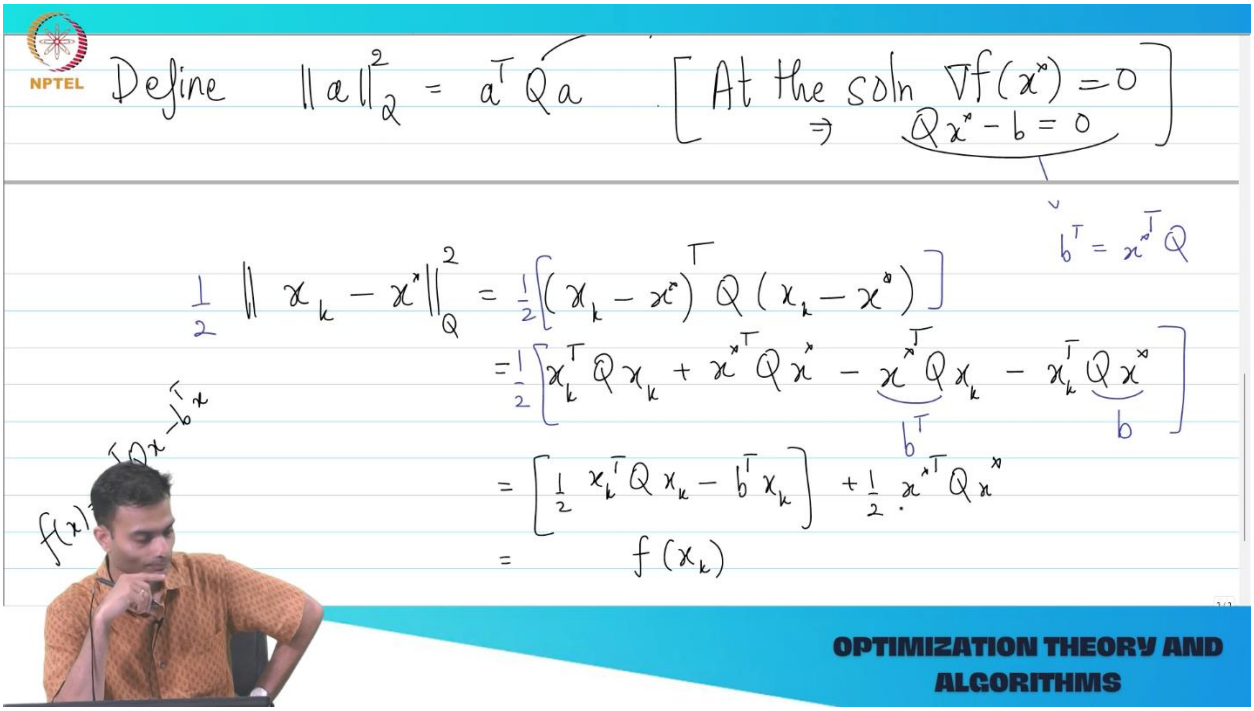
So let us keep a note, we should always know where we are heading otherwise it just looks like random math, right. So add the solution $\nabla f(x^*) = 0$. Do I have an explicit expression for ∇f? We do, right? It's simply $Qx - b$, right? So $Qx^* - b = 0$. Right? So this is just an aside that we will use.

Now let's get back to what we're interested. What is the distance of the iterate from the solution? But we are going to use the Q norm instead of the two norm. So what we need to compute is this. So, we will just plug in the definition of this norm. I am going to get $(x_k - x^*)^T Q(x_k - x^*)$, ok.



If I open this up, how many terms do I expect? 2, 4, 6, 8, how many terms? 4 terms, right. you can let us explicitly open it up. $x_k^T Q x_k$ is one term plus $x^{*T} Q x^*$ is another term and then I have the cross terms. Okay. Let's keep in mind this expression over here.

Can I simplify a few of the terms over here? No, I've just opened it differently. The first two terms are positive, the next two terms are negative, right? So do I see a Qx^* anywhere in this expression? Is there a Qx^* hanging out somewhere? There's a Qx^* hanging out over here. is there a transpose of that hanging out somewhere? There is a transpose of that also hanging out over here. So this term is going to be b and therefore this is b^T. If you take transpose of this expression what will you get? You will get $b^T = x^{*T}Q$.



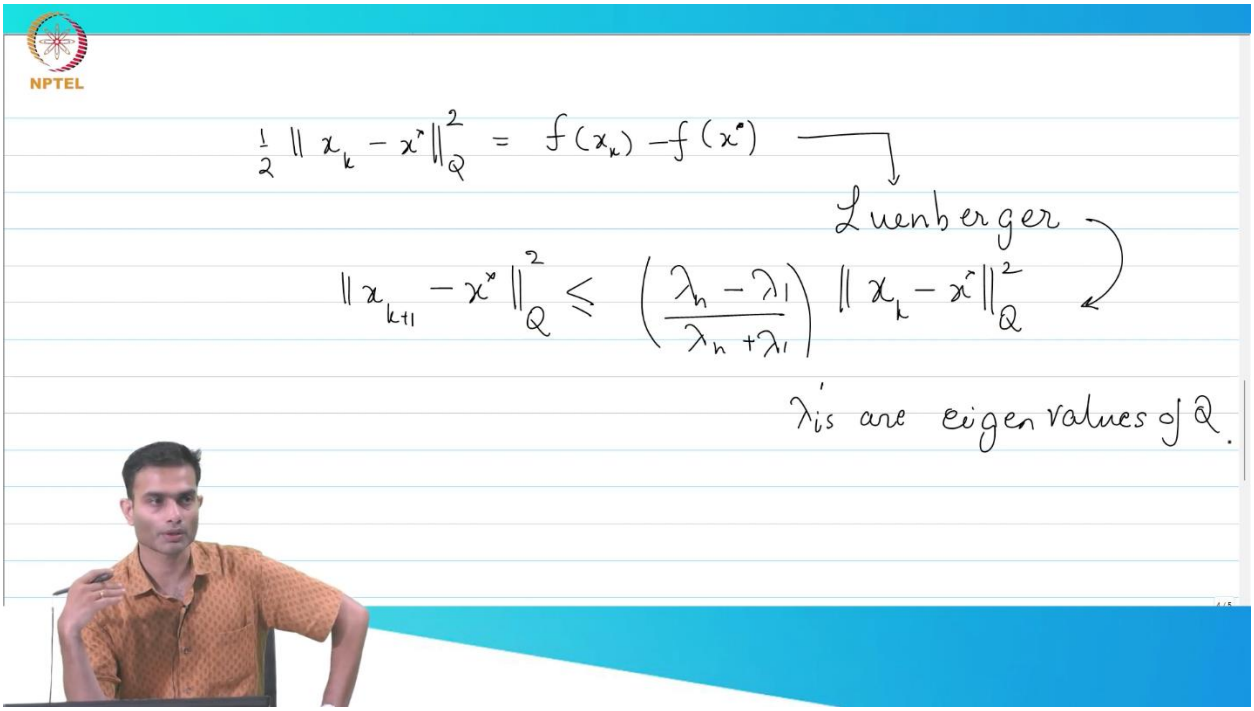
Q is symmetric so $Q^T = Q$. Anything else that I can pull out? Can I group some things common? I'm gonna put a half over here. Can I identify, can I simplify this entire expression of four terms in terms of f? Can you see if you can identify an f over here? What was my definition of f? $\frac{1}{2}x^TQx - bx$. Do I, am I seeing those terms anywhere? Do I see a $f(x_k)$ anywhere? What was $f(x)$? Let us write that once again. $\frac{1}{2}x^TQx - b^Tx$.

Okay, now do I see $f(x_k)$ anywhere? First term matches b^Tx_k. Is that there anywhere? First and which term? Third term. First and third term if I combine this is $f(x_k)$. If I combine term number 2 and 4, x_k^T, sorry yeah, So, is have I made a mistake anywhere or it is ok? x^Tx_k yeah ok. Term 3 and.

So, this is b^Tx_k and ok. So, have we made a mistake in algebra somewhere? 1, 3 and. Correct, half is not for the b's coefficient. Correct, correct, correct. So, let us get this out.

So, half, okay let us write this down properly. So, I am going to get $\frac{1}{2}x_k^TQx_k$ and b^Tx_k and this is going to give me $-b^Tx_k$, right. This is what I have got. So, this is and what is left? Plus $\frac{1}{2}x^{*T}Qx^*$, right. So, what is left? Now, can I make somehow, can I introduce, can I write this in terms of $f(x^*)$? by some algebra, by some clever trick.

Is it possible? Yeah. So, this is clearly $f(x_k)$, right. What about this term? Minus f of this term is $-f(x^*)$, ok. Why is that x^*, let us see that. $x^* = Q^{-1}b$ Therefore? Okay.



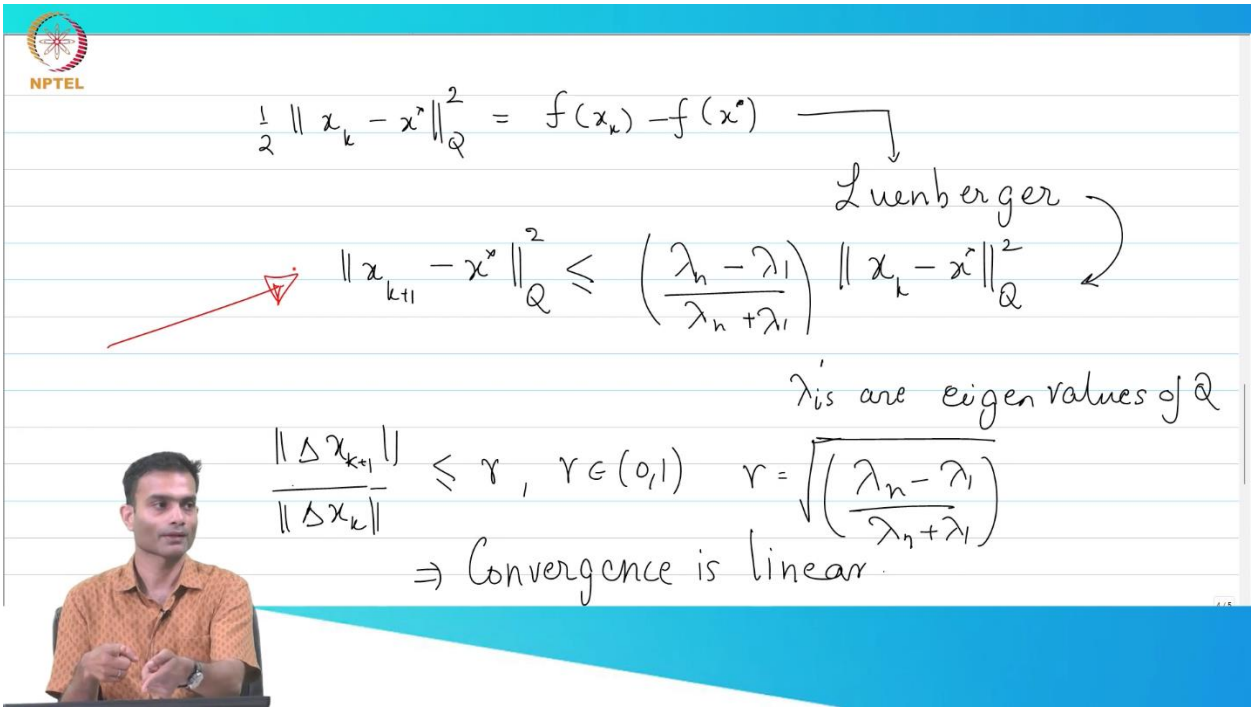
I mean it is correct but does everyone follow what has happened over here? How this became $-f(x^*)$? So let us try to write it out a little bit more explicitly. We have, what is the solution over here? $Qx^* = b$. $Qx^* = b$. From here what do we do next? So, $x^* = Q^{-1}b$, then? Substituted in the expression of f. Substituted in the expression of f, ok.

What will I get? Do I get? and then I will do the same in the first term ok. So, this is $Q^{-1}b$, this is you want to leave it as it is b^TQ^{-1T} then. So, the QQ^{-1} goes to identity right, then what does this become $\frac{1}{2}b^TQ^{-1T}b$ and then and that transpose? Q is symmetric so I can get rid of this right and then I have a $b^TQ^{-1}b$ ok. Then and then this half and half becomes equal to $-\frac{1}{2}b^TQ^{-1}b$ ok. I am still not there yet I want this expression.

What do I do next? substitute, I can substitute b as Qx^*, right. So, now if I substitute $b = Qx^*$, I am going to get exactly this expression, right. So, little bit of algebra, right, but we get this expression. This is, let us write this as an aside. So what did I, let us summarize it over here, $\frac{1}{2}\|x_k - x^*\|_Q^2 = f(x_k) - f(x^*)$, ok.

Yeah, question. The bx_k, that is here, it had a, there were two of them then they got multiplied by half, so the coefficient became 1. The term number 3 and term number 4, these guys. A^TB is equal to, I mean I can swap it, B^TA, that is how I got this. So 3 and 4 combine into this and the second term is what needed a little bit of algebra to see it is actually $f(x^*)$. Anyone having trouble in this step? So if you write along as you are watching this, it will make more sense.

Otherwise, it just looks like a bad movie. So this is what you get. Now, unfortunately, at this point, the proof is equal to pulling a rabbit out of the hat. There is a lot of hard work done by another scientist. So we are going to write his name over here.



It is Luenberger. who takes it from this point. So I have related the distance between two function between the iterate and the convergent point to the difference of function values. So that is a great simplification. And Luenberger takes this further and says that this norm over here is actually less than equal to

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

this proof is not going to be done in class, but we will take it on faith. So after relating this norm of the iterate distance to the function values, this is what you get.

Now what are, anyone wants to guess what the lambdas are? Eigenvalues of Q. These are the λ_i's are eigenvalues of Q, ok. There is a similar question that you will also find in the tutorial about this, ok. So now this is a very very useful result because if I wanted to look at, if you look at the definition of linear convergence what was the expression that you had? What was the ratio that we had in when we spoke about linear, linear convergence? The ratio of what to what should be less than something, what is it? It was x_k, so let us call it $\| \Delta x_{k+1} \|$ divided by $\| \Delta x_k \|$, right. Δx is $x_{k+1} - x^*$.

Was what? Less than equal to r, where r was 0 to 1. Does it look like that? Does this expression that we have got, does it look like that? It does look like that. What is my r? Square root of this expression, right? Because I have got the squares on both sides. So,

$$r = \sqrt{\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}}$$

Again, just a quick linear algebra refresher.

I have called these the eigenvalues of Q. If I call these the singular values of Q, would I be correct or incorrect? Correct, because for a not just a square matrix right, for a positive definite matrix, so obviously which is also symmetric, the eigenvalues and the singular values are identical right. So, I can talk in either terms ok. So, this proves that convergence, we proved earlier that convergence happens and now we are saying that convergence is what? Linear. Now this expression over here with the red arrow contains in it actually a lot of intuition, a lot of geometric intuition is contained inside it if you pay attention to it.

So remember when we had drawn the zigzag contour of the steepest descent trajectory, remember the zigzag contour, that zigzag contour had happened when I had an elliptical elliptical cup so to speak, right? And we had said that the curvature of the cup is related to the properties of the Hessian, second order derivatives gives me curvature information. Where is the Hessian over here going to, I mean in this quadratic cost function which I have, the Hessian is going to be what? Q is going to be the Hessian. And what have I got over here? The r term, does it have the properties of the Hessian? It directly has the properties of the Hessian, the eigenvalues. Now if I had a square bowl or a square cup in n dimensions, what would happen to the eigenvalues? They would actually all be the same. No matter which way you approach from, the curvature is the same.

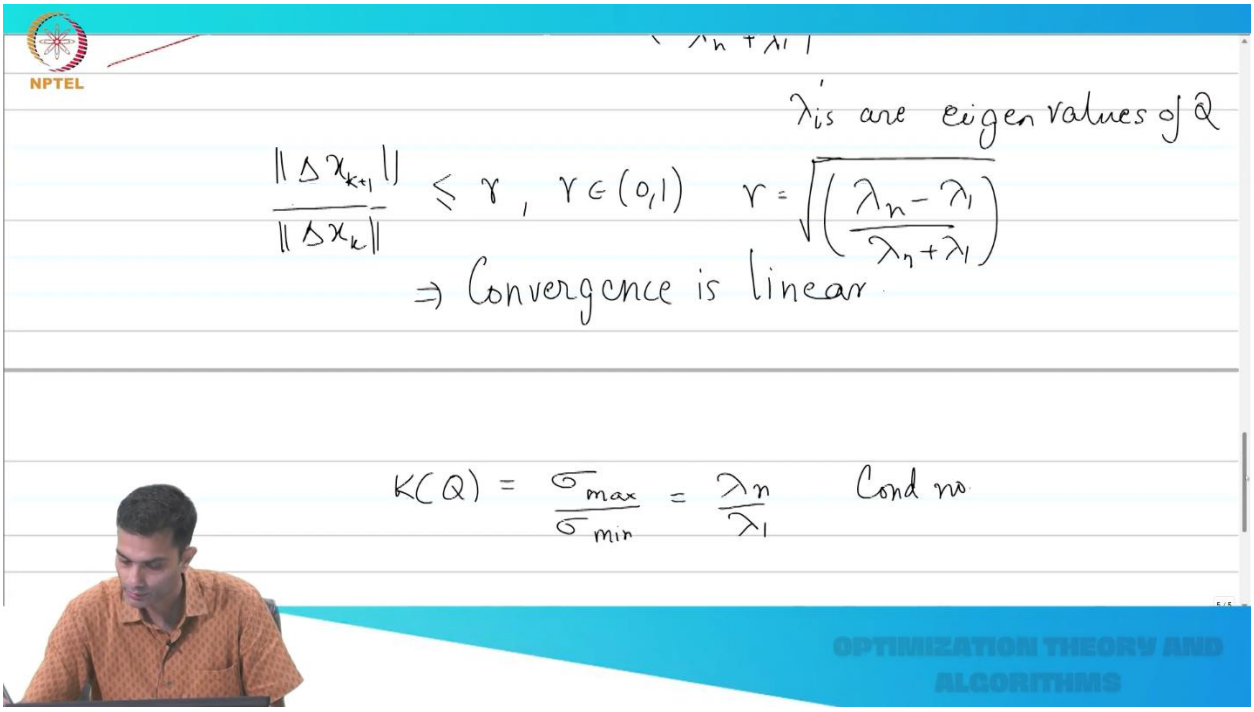
So you would have the same eigenvalues. So that was actually a special case. If I have circular contours, right? That means the bowl is actually like a circle in n, I mean it is a bowl in n dimension circular in cross section. No matter where you start from, you are going straight into the solution. So you can see that over here, $\lambda_n = \lambda_1$, therefore what? It becomes 0, right? So the distance between x_{k+1} and x^* becomes 0 in one shot and this theorem is telling us that.

So the curve and on the contrary, if I have a very squished cost, I mean a very squished bowl so to speak, right? Like a very very elliptical thing. Then λ_1 and λ_n are going to be very different. So this r term is now going to become larger and larger, right? So that means you may take many steps to arrive at it. No matter how many steps you take, the rate at which these steps are going to go is linear.

That is clear from this proof. Okay, yeah that is a good point. So this was proved under some very restrictive conditions. The first restrictive condition was I assumed quadratic cost function. I assumed convexity, well that is okay.

I assumed exact line search, right. So the question is will this work in general for inexact line search? Will it work for non-quadratic cost function? As you will find out in the research literature people have proved that linear convergence happens even with inexact line searches. But we won't do it here. Those proofs get more and more involved.

Yes. Yeah. Luenberger's proof is in the line of this, but there are several steps over here. So we have just stated the final result over here. But other people have generalized this and shown that even if you do it inexactly, you will still get linear convergence. It's not that linear will become sublinear or something like that. So, this expression that we have over here λ_n and λ_1 which could also be written in terms of σ_n and σ_1.



When we were doing a review of linear algebra we had encountered these two numbers and what was that? When did we encounter σ, the ratio of σ_n and σ_1? The condition number, right? So, the condition number of Q is simply the maximum singular value by the minimum singular value which is simply

$$\kappa(Q) = \frac{\sigma_n}{\sigma_1} = \frac{\lambda_n}{\lambda_1}$$

So intuitively what we are sort of getting from here is that the worse the condition number. Worse means a worse condition number is a high condition number. The best condition number possible is actually for the identity matrix which is 1. You cannot have a condition number less than 1. So the higher the condition number the bigger is this ratio λ_n to λ_1 and this constant over here is going to get larger and larger.

As it gets larger and larger that means you may have to make more and more steps to reach here. So this is the story with how the gradient descent family of methods work. They converge and they converge at a linear rate. This convergence we have proven under a restricted set of assumptions but the class nodes have proofs for the other methods as well.