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 - 52

Hessian Modification

So, the starting point is A is not positive definite, ok. So, Hessian modification simply says that I need to perturb the matrix to $A + \Delta A$ such that it becomes positive definite. Now this sounds like a very, very loose thing, right, perturb it, of course I can perturb it to anything I want and, you know, it will become a positive definite matrix, right. So, there is a little bit of care that I have to take. So, you would want to perturb it in such a way that it has as low norm as possible, you can say that. You can make a, by the meaning of a perturbation is it is small compared to the original.

So you can say or rather require that the norm of ΔA be as, if the norm is as low as possible what is the advantage? I have changed my problem as little as possible. Because I can change my A to whatever I want. I can just take all the eigenvalues and replace them by plus 1 everywhere and solve that and feel very good about myself. But I have changed the original problem.

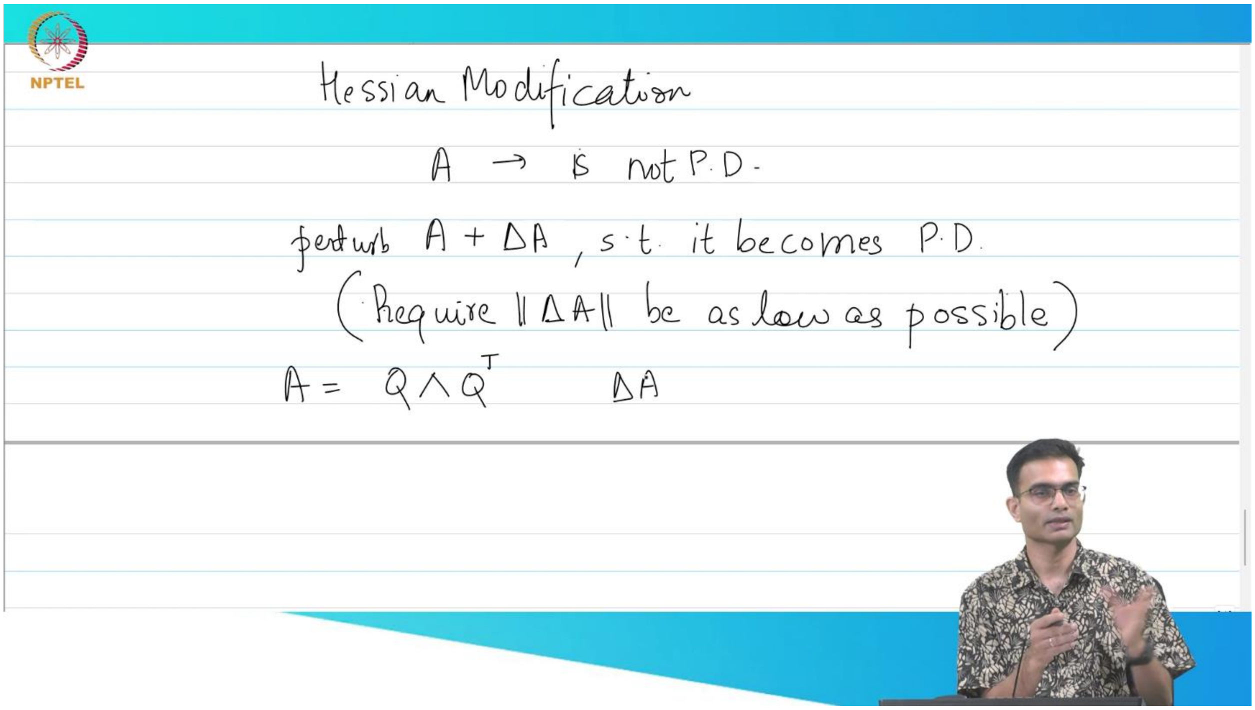
When I go back to my original problem, I will see that the solution does not match. So, I do not want to modify the problem too much. The way for me to ensure it is to say that whatever you do, whatever perturbation you do, make sure it has a low norm, ok. Now can we try to guess how we should perturb this? Do you want to think about it for a second? Kill the largest, so let us talk independent of the method of evaluating positive definite. So, Gershgorin gave us the disks, ok.

But assume that you have enough computing power that any of these three methods A, B, and C are available for you, ok. How would we work it out? How would we work out this minimum norm? What can we start with? So, A, let me start by giving you a hint. So, $A = Q\Lambda Q^T$, ok. Some of these eigenvalues are negative, which is why it is not positive definite. Now, I am adding ΔA. Is there a way you can think of by which this can happen so that I can use this eigenvalue decomposition of A? That is a good idea.

So, what she is suggesting is that the eigenvalues of ΔA should be in such a way that they bring the negative eigenvalues of A into the positive range, ok. Now, is it possible for you to, ok, let me put it another way. For which matrix can you just look at it and tell the eigenvalues? Diagonal matrix. Diagonal matrix. What is the easiest diagonal matrix you can think of? But if I do identity, what are the eigenvalues? 1. Do we want 1? What do we want? We do not know what we want.

It depends on A. How about if I do this? What are the eigenvalues of this guy? τ. Earlier it was 1, now it is τ. Now, this τ is just a scalar, remember? Can I write identity in a clever way? QQ^T. Very good.

So, what is $A + \Delta A$? Can I group the Q's and the Q^T's in a nice way? So, I am going to get $Q(\Lambda + \tau I)Q^T$, quite easy. So, do I get a value of τ that is required now? Right? So, the eigenvalues that are going to come from here, I am going to get $\lambda_{\min} + \tau$, this is going to be the lowest eigenvalue of this new system, and what should this be? Greater than 0, right. So, should I put it greater than 0? So, greater than 0 can be tricky because why? It is 10^{-15}, is it greater than 0? It is greater. Is it a good idea? Why? The condition number will become insane, right? So, at this point itself, may as well make your life a little bit simpler.



Hessian Modification

$A \rightarrow$ is not P.D.

perturb $A + \Delta A$, s.t. it becomes P.D.

(Require $\|\Delta A\|$ be as low as possible)

$A = Q \Lambda Q^T \qquad \Delta A$

Instead of saying 0, what should you do? Put some threshold, and this threshold you keep in your hand, you can control it, ok. So, this saves you from a bad condition number. So, that tells you that this value of τ is greater than or equal to $\Delta\tau_{\min}$. And all of this exercise of meaning only if $\lambda_{\min}$ is actually less than 0. If it is already positive, there is no point.

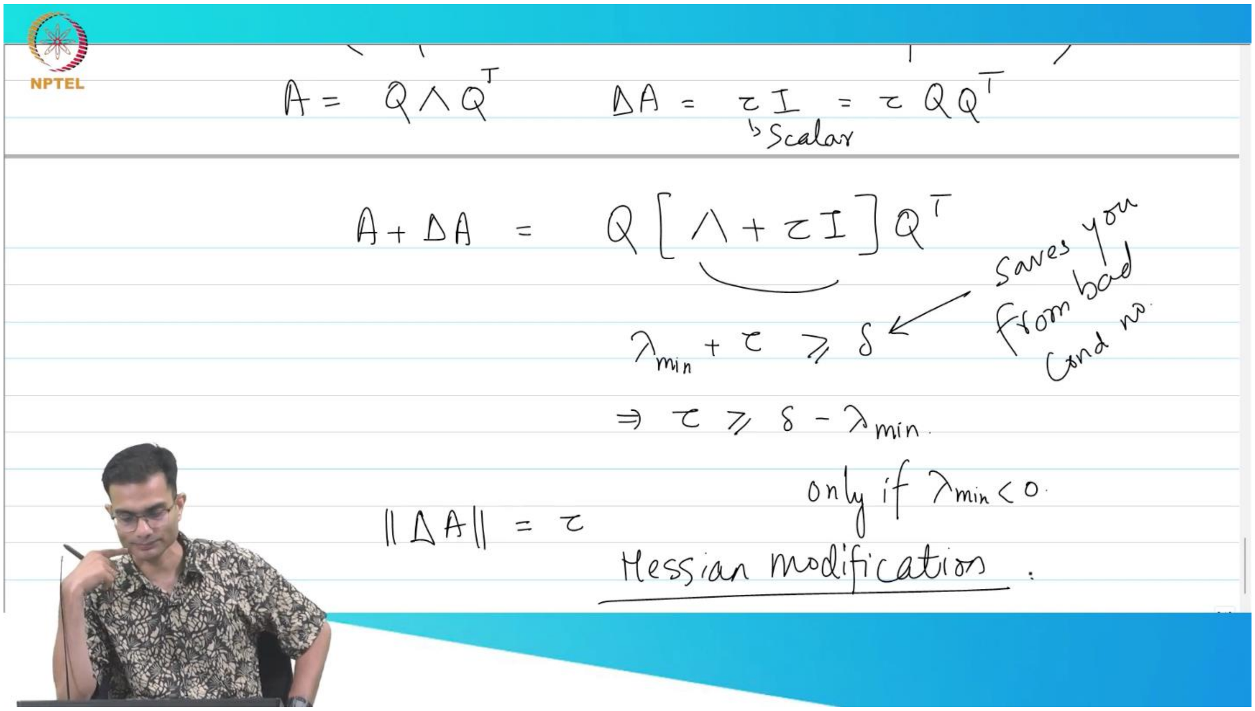
So, this is what is meant by, so ok, and to complete the thing, what is the norm of ΔA? If I were to take your favorite norm, what would be this? What is the norm of the identity matrix? 1. If, for example, take the Frobenius norm, it is 1. So, this is actually just going to be τ, the norm of ΔA. So, now by making ΔA as small as possible, I can say that this is the minimum norm perturbation to the matrix A such that my new Hessian is positive definite. It is not a new Hessian, it is a Hessian modification, it is not the Hessian of any problem, I have just modified it. So, with this Hessian modification in place, yeah.

It has a condition number 1. Condition number 1, correct. That is good. What has that got to do with the norm? Condition number and norm are different. If you are thinking of the definition, what is the norm of ΔA^{-1}? $1/\tau$, so the condition number is 1.

What qualifies as a good, so this is not an approximation of the Hessian, this is a Hessian modification. We are modifying it in the least norm way to arrive at a matrix which is positive definite and then we continue the rest of the procedure. So, if you look at the very first step over here that we have, maybe summarize the Newton method, right. If B_k is not positive definite in step 2, you are suggesting that I chuck B_k and I pick any other, yeah, why is that a good idea? So, that is a good question. So, how did we come upon this expression of $-B_k^{-1}\nabla f_k$? Does anyone remember? We took the second-order Taylor expansion of f and then we minimized it and we came upon exactly this expression.

So, if I choose exactly this expression, I am coming to a minima of a quadratic expression. If I replace this B_k by any other matrix, that is no longer a minima. So, that is why I want to be as close to the original Hessian as possible so that I get the property of that second-order convergence and approximation of Taylor. If I replace this, it will be a legitimate descent direction, I agree, but it may not be the best descent direction. But that said, like I said at the start of the class, there are a full family of methods which talk about what is the best way to perturb this Hessian so that I am as close to the quadratic approximation and I am, you know, getting good properties of convergence.

So, this is just one of many and this is the simplest way. There are much more sophisticated ways also. So, just a descent direction is not good enough because you saw in the case of steepest descent versus CG. Just a descent direction got you a certain rate of convergence. When we moved to CG, you know, you got much faster convergence, right? So, just a descent direction, it works, but you can do better.



$A = Q \Lambda Q^T$ $\Delta A = \tau I = \tau Q Q^T$

↳ Scalar

$A + \Delta A = Q\left[\underbrace{\Lambda + \tau I}\right]Q^T$

$\lambda_{min} + \tau \geq \delta$ ← Saves you from bad Cond no.

$\Rightarrow \tau \geq \delta - \lambda_{min}$.

only if $\lambda_{min} < 0$.

$\|\Delta A\| = \tau$

Hessian modification.

That is the motivation. $\sqrt{3} \times \tau$, why root? Oh, you will get a root n, correct, correct. With the Frobenius norm, you would get $\sqrt{n}$, with the 2 norm what would you get? 1, right? So, okay. Let us just make it, okay.

Yeah, question. How do you know $\lambda_{\min}$? You come to these three methods. So, you would... It does, right, because in fact, I would do the disk theorem in this case because it is saying that your τ should be $\delta - \lambda_{\min}$. So, your question is how do I know $\lambda_{\min}$, right? To find out $\lambda_{\min}$, what should I do? I look at the circle which is to the left, I mean which is coming closest to the origin. So, let us say for example, let us take a good example, right.

So, all my disks were like this. In this case, my $\lambda_{\min}$ for example will be somewhere in this disk. It's positive definite. Now, what happened? Let us say one disk became like this.

And I have this over here. So, I've not actually computed $\lambda_{\min}$. I'm saying $\lambda_{\min}$ lives in this disk. Now, what do I need to do? It may be positive. Yeah. So this is because I am putting less effort, I am also getting a rough answer.

I am saying that it lives in this disk. Where in this disk? I do not know. So, you are right. The eigenvalue could be here.

$\lambda_{\min}$ could be this. Right? So I am losing this information because I am virtually putting in no effort. But this disk, if I move by this amount to the right, then I am assured that $\lambda_{\min}$ is greater than this. So, I mean I am trading off accuracy for computation power. That is why this method is particularly useful if you are looking at a big data problem. This becomes very attractive. It is simple enough to just modify.

And so, this one may ask that you know I am perturbing the Hessian. How is it going to work out because I am actually changing the problem? Now, what happens is that if, again, I am going to state this without proof, if your function is a well-behaved function in the sense that it is twice continuously differentiable, Lipschitz continuous, all of that, then as you approach the solution point, the Hessian becomes positive. So, as you approach the solution, it turns out that the Hessian becomes positive definite and this is proven in the book. So, as I approach the true solution, I actually do not need to do Hessian modifications.

I use the straightforward Newton method which gets me... So, I mean that we just get that for free. Yeah, that is what we push the disk. Whichever disk has the leftmost point, I move that to the right. No, I am only pushing one disk to the right.

Okay, correct. I am pushing all of the disks to the right. So what? Okay, right. So, what he is saying is why push all of the disks by an amount τ to the right? Why not just push the offending disk to the right? So, what would that correspond to? For what kind of a ΔA would that correspond to? τ is also possible. So, what he is saying is that instead of pushing everything by $\tau \times I$, have a diagonal matrix where the non-zero entries are only for the disks which are flowing, overflowing onto the left, push them to the right, right. And then your, when you look at $A + \Delta A$, you would have all the disks are now on the right, some of the disks are not modified.

And some disks are modified, that is what you are suggesting, yeah. So, there is a full family and range of solutions possible over here and they all give varying amounts of accuracy. That is a good idea, right? In that case, what would be the norm of ΔA? Frobenius norm, Frobenius norm would be lesser and probably the two-norm may also be lesser. We can compute that. Okay, so I think with that I will bring this to an end because the next thing, I do not want to start quasi-Newton method and then have to stop in five minutes.

So, in the next class, we will look at Quasi-Newton methods. The advantage of a Quasi-Newton method is that Quasi-Newton methods sit between Newton methods and steepest descent methods. Quasi-Newton methods give you super-linear convergence, rate of convergence. So, it is not as good as quadratic, but it is better than linear rate of convergence. So, we will look at that tomorrow in detail, right.

That is in your hand. So, if the modification is in your hand because of this value of δ over here. If you choose because you are modifying the smallest eigenvalue, if δ ended up being very small, then your condition number is very high. So, do not make δ small. Make it larger so that the eigenvalue, the lowest eigenvalue of $A + \Delta A$ is not very, very low.

That you can do by choosing δ. Again, there is a trade-off. The more you modify it, the further you are going away from the original problem. So, you have to have this trade-off in mind. Yeah, but keep in mind the nuances of this Gershgorin disk theorem. It can sometimes give you some misleading interpretation if you do not look at it carefully.

It is only saying the eigenvalue is somewhere in the disk. It is giving you no information of where it is in the disk. So, even though part of the disk may be in the left half, there is no guarantee that the eigenvalue will be there. It can be anywhere in there. So, you may be doing more work for no reason, but you cannot know it unless you spend order. If your Hessian is positive definite, then you will simply use this Newton direction, you will not do any modification.

Correct. Correct, you have to each time determine whether or not it is positive definite. Not using the disk method, using whichever method. The disk method is one example of three methods, right. Now, if you have an inaccurate method, then you are likely to go off course.

So, you again there, you can have a mix. Sometimes you could do this method, sometimes you could do that method, depending on how much accuracy you want. You could do that also. The difficulty there is that if you end up, if it is not a descent direction and you did not check for it, then actually you ended up going in an ascend direction. So, you made your problem worse. So, whether or not it is a descent direction should be checked at each iteration, right.

In fact, if you do not check whether it is a descent direction, your step 3 will actually never converge. In the backtracking line search, when you are either doing sufficient decrease or curvature condition, all of that assumes that the p_k is a descent direction, but actually your p_k is an ascend direction and you did not check it. So your step three itself will give you a problem. So you have to check this every time. How you check it? You could vary the accuracy at each step, but you must check it each time.

Yeah, question. Yes, so okay, two responses. When I am changing it, when I am changing the Hessian, I am trying to make sure that the norm of the change is as low as possible, okay. So I am changing the original problem, right. Now, as I mentioned a few minutes ago, as I actually approach my true solution, the true stationary point, if the function f has all of these nice properties, it turns out that the Hessian is positive definite. So you will end up not requiring the modification.

And so that is why you end up in the correct solution, okay. If that were not the case, we would be in real trouble. Then you would have solved a different problem altogether. So this is somewhere from between your starting point and your solution, somewhere in the middle, you have to do all of these tricks. Okay, if you have no more questions, you can hand us your feedback chips.

See you tomorrow. The rate will still be quadratic, but because it is positive definite. Yes, the modified Hessian is still positive definite, but it is not completely. So, maybe you could say that the rate of the coefficient of the rate of convergence now changes. The rate will still be quadratic.

I mean, you may be going at a quadratic rate to a wrong answer in between. Towards the end, you will get back to it.