Numerical Analysis

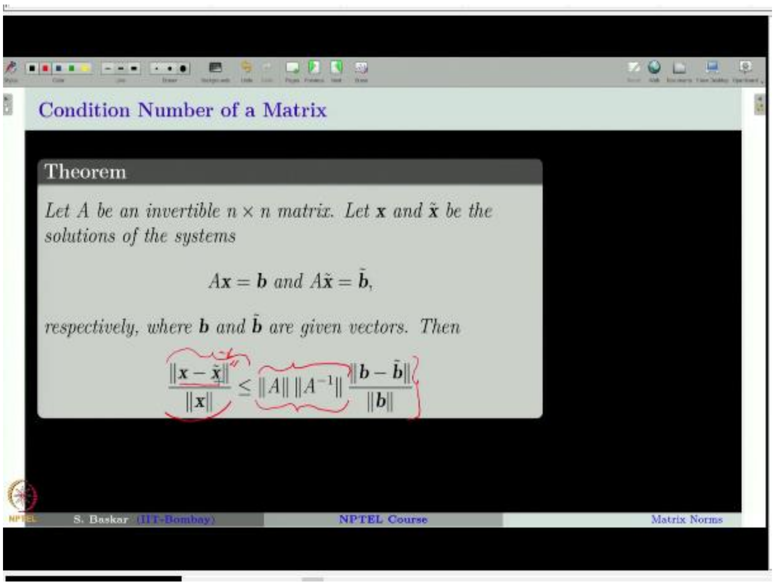
Prof. Dr. S. Baskar
Department of Mathematics
Indian Institute of Technology-Bombay

Lecture-16
Matrix Norms (Condition Number)

Hi, in the last lecture we have started our discussion on matrix norms. We have also introduced an important concept called matrix norm subordinate to vector norms. In this lecture we will continue our discussion on matrix norms where we will define what is mean by condition number of a given matrix and we will also learn how the condition number can be used to study the sensitivity of the matrix that we will be using in solving the linear systems on a computer.

Let us start our discussion with a theorem which will tell us why we have to define condition number of a matrix in a particular form. Let us consider an invertible matrix A which is an $n \times n$ matrix, we will also consider a linear system $Ax = b$.

(Refer Slide Time: 01:23)



Let the solution of this system be x. Now instead of taking the right-hand side vector b we make a small error in it and consider the right-hand side vector with error as $\tilde{b}$. So, we want to solve $Ax = b$, but we land up solving the system $Ax = \tilde{b}$. Now the solution that we obtain from this perturbed system is going to be different from the solution x. So, let us denote this solution from the perturbed system $Ax = \tilde{b}$ as $\tilde{x}$ and now the question is what is the error involved in $\tilde{x}$ when compared to the exact solution x?

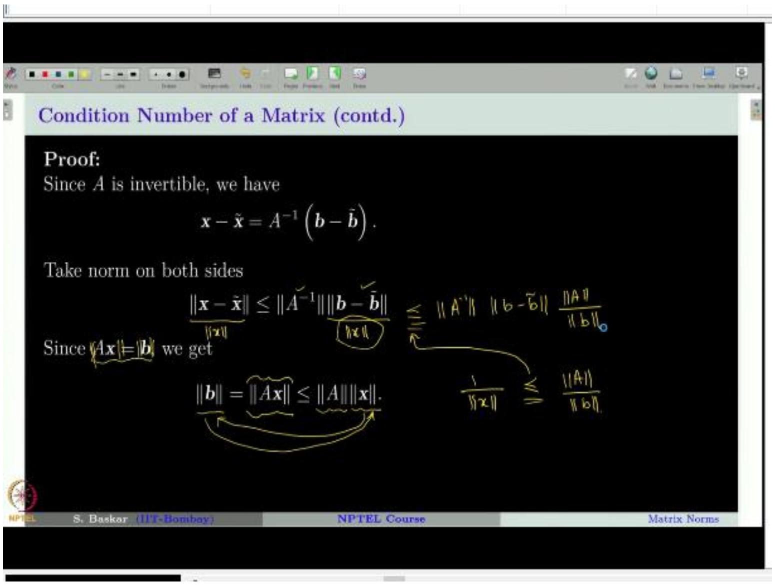
That is the question. The theorem says that the relative error in $\tilde{x}$ when compared to x which is defined as $\frac{||x-\tilde{x}||}{||x||}$. Remember this quantity is obtained for a given vector norm. That is, you have to take a vector norm and with respect to that vector norm we are computing this quantity which can be taken as the relative error in $\tilde{x}$ when compared to x.

Now the theorem says that this relative error in $\tilde{x}$ will be less than or equal to this quantity times the relative error in the right-hand side vector b. In this we are choosing a vector norm and this quantity is computed using the matrix norm subordinate to the vector norm that we have considered here. So, that is a very important point that we have to keep in mind when we are working with not only this theorem but also whenever you are working with the error analysis of any linear system the matrix norm that we take should be this subordinate matrix norm with respect to the vector norm that is given to us.

In this theorem the vector norm and therefore the subordinate matrix norm can be anything that you choose the result will hold, but in the computations we will generally fix one of the three vector norms that we have introduced in the last class. That is l_1 norm or l_2 norm or l_∞ norm. We will choose one of this norm as per our convenience or requirement and then we have to choose the corresponding matrix norm subordinate to our chosen vector norm.

So, this is how we will work in all the problems that we do from now onwards, but this theorem as I told will hold for any vector norm and the corresponding subordinate norm that we choose. Let us try to prove this inequality. Remember we have to find an estimate for this relative error. So, we will first start with $x - \tilde{x}$.

(Refer Slide Time: 05:14)



Remember from the linear systems we can write $Ax - A\tilde{x} = b - \tilde{b}$. From there we can write $A(x - \tilde{x}) = b - \tilde{b}$. Since A is invertible, therefore we can write $x - \tilde{x} = A^{-1}\left(b - \tilde{b}\right)$. Now let us take norm on both sides. Remember this is the vector norm and this is also the vector norm. Now you recall from our last lecture we have proved an important property for the subordinate matrix norm which says that Ax which is a vector and you take the vector norm on that.

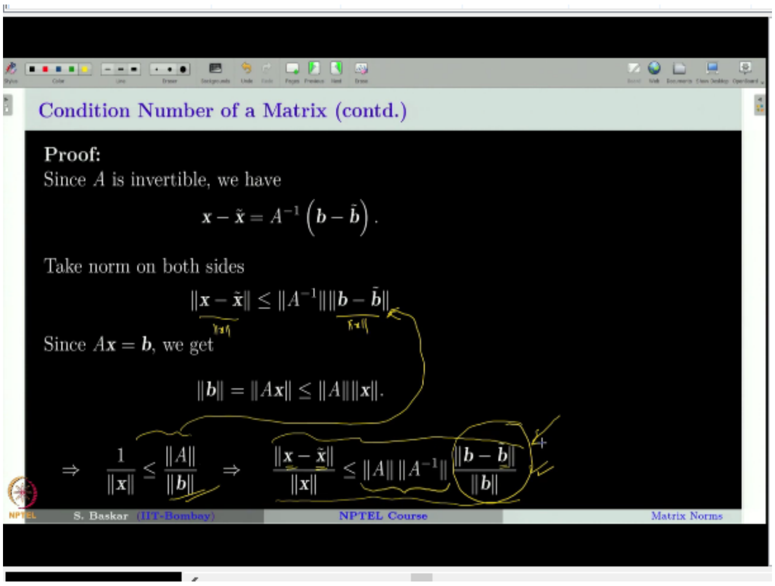
Then you can write this as less than or equal to $||A||$ which is the subordinate matrix norm times $||x||$. We have proved this inequality in the last class. Now you see you can just imagine this as A and this as x and you can apply this inequality to get $||x - \tilde{x}|| \leq ||A^{-1}||$ because you have inverse here into $||b - \tilde{b}||$ that is the vector here.

Now we have this in one hand, see remember we have to prove something for the relative error. So, we have $||x - \tilde{x}||$ now we have to divide both sides by $||x||$ that is what we have to do here now. Now you remember this is less than or equal to $||A||\ ||A^{-1}||\frac{||b-\tilde{b}||}{||b||}$, but here we have x but we want $\tilde{b}$ here. So, we have to somehow dominate this $\frac{x}{b}$ here. How will you do that?

Well, let us go back to our original system $Ax = b$ and take norm on both sides, you will get $||b = ||Ax||$. Now whenever you see this immediately our property should come in your mind and you can write it as $||Ax|| \leq ||A||\ \ ||x||$. That will immediately tell you that $\frac{1}{||x||}$, what I am doing I am just taking this to the left-hand side and that is less than or equal to $||A||$ and this I am bringing to the right-hand side and that will come as $||b||$.

So, this is precisely what we were intended to do. This is $||x||$ divided by $||x||$ was what we had in the previous inequality. Now this part is less than or equal to $\frac{||A||}{||b||}$. Therefore, you can write this is less than or equal to $||A^{-1}||\ ||b - \tilde{b}||$. These are there already. Now instead of $\frac{1}{||x||}$ we will put this that is why I have this less than or equal to sign here. I have $\frac{||A||}{||x||}$. That gives me precisely what we want.

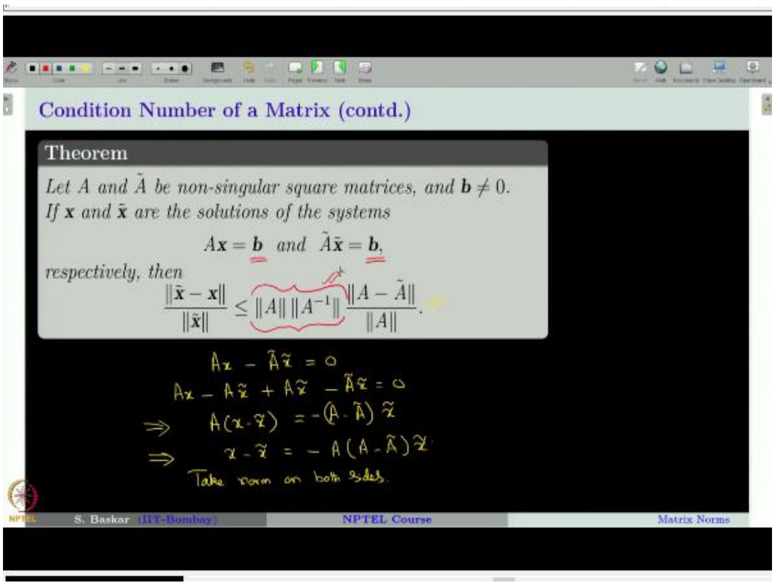
(Refer Slide Time: 09:52)



I will repeat once again what we are doing is we are taking this to the left-hand side and b to the hand side and writing this inequality. From there we are getting what we want by putting this inequality into this. So, remember first we have divided by $||x||$ on both sides and then we put this inequality here to get this. And this is precisely what we want to show.

So, this is an easy proof but the result is very important which tells us that the relative error in the approximate solution when compared to the exact solution is amplified by this factor with respect to the relative error in $\tilde{b}$ when compared to b. That is the error you have committed in your input data. So, remember this theorem is only considering the error in the right-hand side vector, whereas the coefficient matrix is taken exactly. But in practical situations we will have error even in the coefficient matrix.

(Refer Slide Time: 11:17)



So, let us see how the estimate for the relative error in $\tilde{x}$ will look like when compared to the relative error in the coefficient matrix and that is given by this inequality. Here also you can see that this expression is coming as the factor here, in the previous one also, the same expression was there. So, that shows that this quantity that is $||A||\ ||A^{-1}||$ is very important in understanding the relative error in the approximate solution of your linear system when compared to the exact solution.

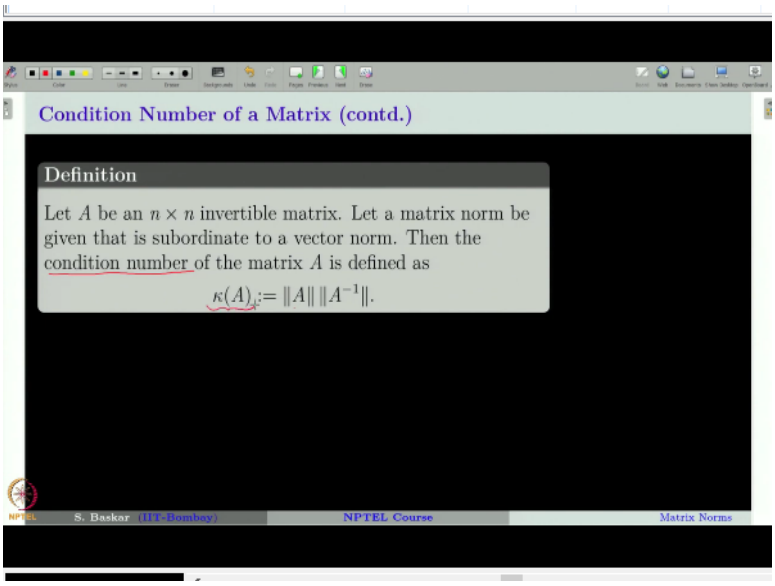
So, now how to prove this theorem? Well, it is not very difficult; we can prove it with a little mathematical manipulation. Let me quickly tell you; you can fill up the gaps; you start with this $Ax - \tilde{A}\tilde{x}$. This is how we have started our proof in the last theorem also, but now you see you do not have any error in your right-hand side vector. You are now considering in this theorem that the right-hand side vector is exact but you have error only in the coefficient matrix.

In the previous case it was the other way round, in fact in practical situations we have error in both these inputs but for the sake of simplicity we are considering these two cases separately and getting the inequalities. You can also combine and get an inequality but that you can do it once you understand these two steps. Therefore, I am only doing these two steps separately. Let us start with this, since the right-hand side vector is same you will have $Ax - \tilde{A}\tilde{x} = 0$.

Now what you do is you add and subtract $A\tilde{x}$ that is Ax which is already there $-A\tilde{x}$ which you newly insert into this equation then you add the same quantity and then you take $\tilde{A}\tilde{x}$ which is already there equal to 0. That can be written as $A(x - \tilde{x}) = A - \tilde{A}$, there is a minus sign here because you are taking to the other side and $\tilde{x}$. That implies since A is invertible, $x - \tilde{x} = -A(A - \tilde{A})\tilde{x}$.

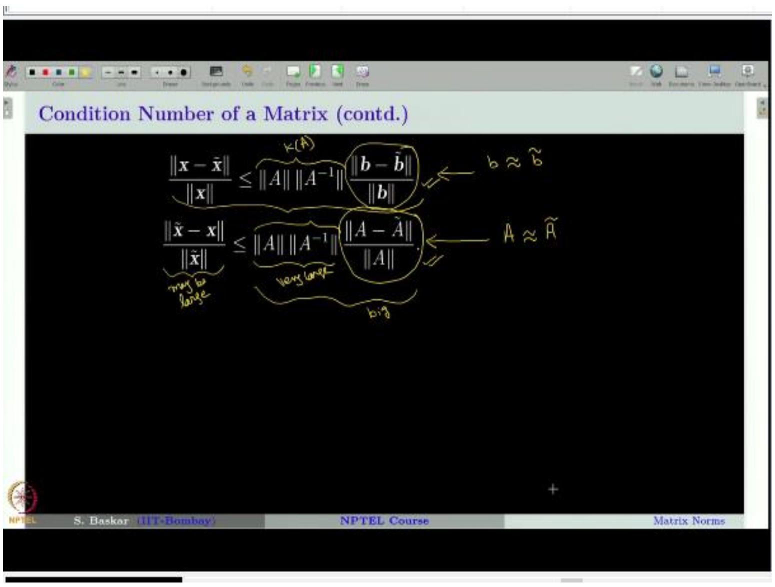
Now you take norm on both sides. The remaining part of the proof will go more or less in a similar way as we did in the previous theorem. I will leave it to you as an exercise to complete this. Now you see the important point as I mentioned this quantity seems to be very important in quantifying at least as an upper bound for the relative error.

(Refer Slide Time: 15:16)



Therefore, we will take this quantity and give the name as condition number of the matrix A. So, we will define $||A||\ ||A^{-1}||$ as the condition number of a given $n \times n$ invertible matrix. We generally use the notation $\kappa(A)$ for this.

(Refer Slide Time: 15:37)



So, let me summarize. When we have the error only on the right-hand side vector that is b is taken as $\tilde{b}$ with some error then the relative error in the corresponding solution of the linear system will have such an estimate and similarly if you have the error in the coefficient matrix that is instead of A if you take an approximation to A which is denoted by $\tilde{A}$ then the corresponding relative error in $\tilde{x}$ when compared to x is given by this estimate.

In both the case the amplification factor is nothing but the condition number $\kappa(A)$. Now you can see that if $\kappa(A)$ is very small for a given matrix A. Then if you assume that this relative error, equivalently this relative error, is very small then the upper bound is going to be very small. Generally, it is fair to assume that this relative error in the input data is very small because that comes as most probably the rounding error on your computer or it may be some small error that we may make somehow.

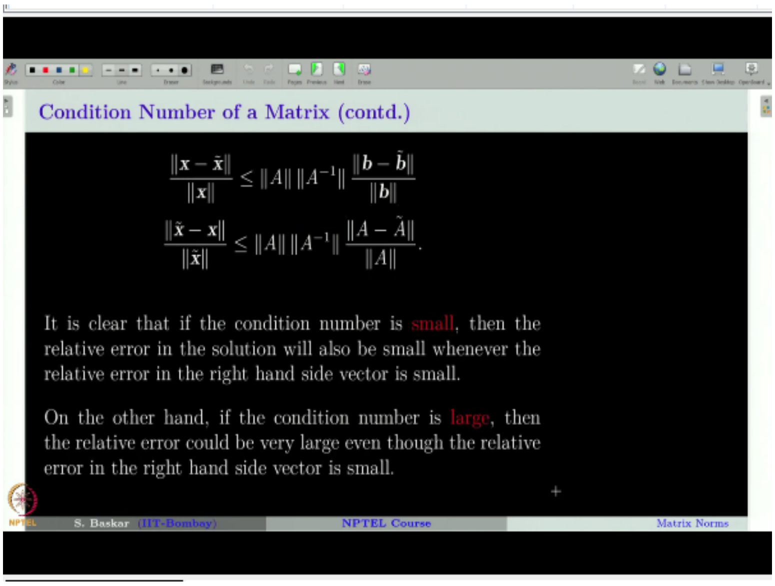
Therefore, these relative errors may be assumed to be very small. Therefore, if the condition number of your coefficient matrix is very small then the relative error in the approximate solution will also be very small. On the other hand, if the condition number is very large, so large that this product becomes a big number. Then what happens? It gives us a possibility that your relative error in the approximate solution may also be very large.

It will not tell that the relative error in the approximate solution will be surely large but it gives us a possibility that this may be large. So, that is a kind of caution that this condition number will give us. If your condition number is very large then the computation of the solution of a linear system may be very sensitive to even small error in your input data. That is what the message we are getting through this analysis.

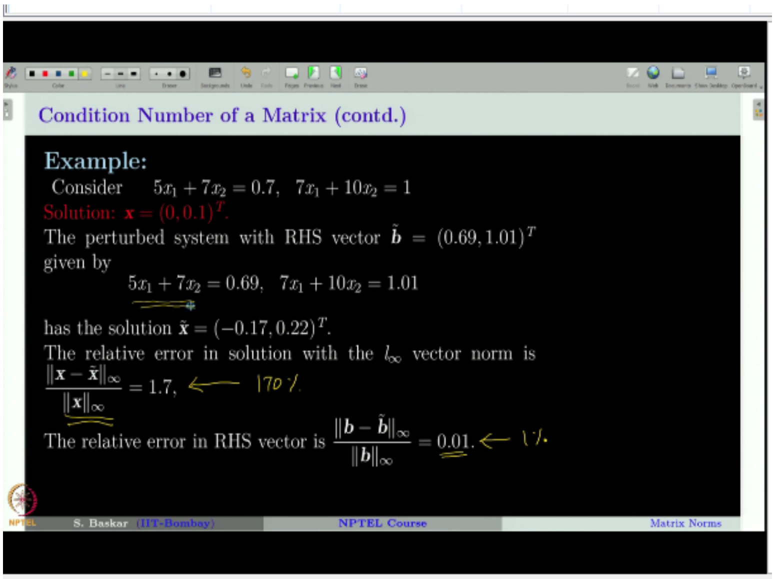
Here we should note very importantly that the condition number of a matrix depends on the subordinate norm that we use. That is, we will first decide what vector norm we will be using and from there we will have a subordinate norm corresponding to that vector norm and the condition number is then calculated with respect to that matrix norm. Generally, we will use infinite norm because that is the easiest norm.

Although it is not physically realistic, but it is easy to handle infinite norm rather than the l_2 norm. So, in all our analysis we will fix l_∞ norm and compute the condition number of a given matrix.

(Refer Slide Time: 19:20)



(Refer Slide Time: 19:21)



Let us take an example, consider this system where the right-hand side vector is 0.7 and 1. The corresponding solution is 0.1. Now what we will do is, we will make a slight perturbation in the right hand side vector; thereby we will consider $\tilde{b}$ as our hand side which is given by 0.69, 1.01. So, this is slightly different from the right-hand side vectors given here. So, we have made a slide error in our right-hand side vector.

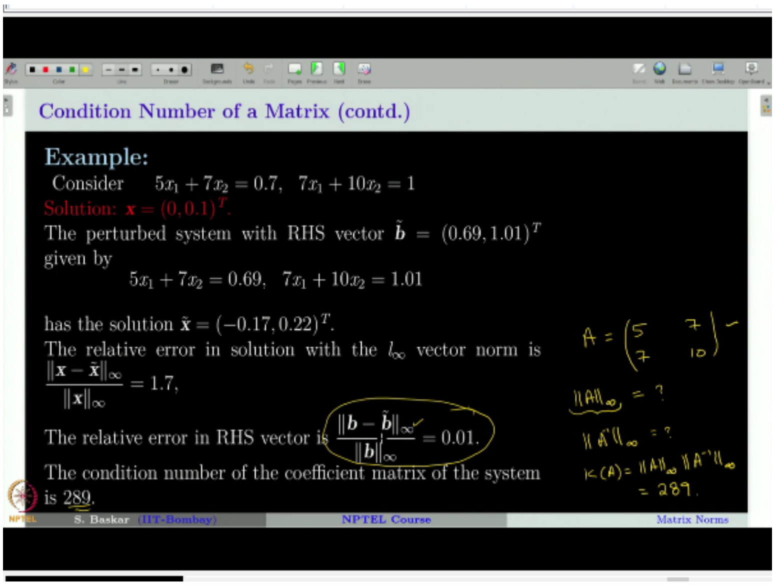
Therefore, we are considering this system, but we are actually intended to solve this system and get this solution. But we are actually solving this system and therefore we are getting its solution which is given by -0.17 and 0.22. Let us compute the relative error in $\tilde{x}$ when compared to x and see how it looks like. Now coming to precise computation like this we have to choose one particular norm to do this computation.

As I told you it is very easy to handle l_∞ norm. Therefore, I will take the relative error in $\tilde{x}$ with respect to the infinite norm. Remember infinite norm of a vector is defined as maximum of the components of the vector, its absolute value. So, this is how it comes. Now if you take $||x - \tilde{x}||_\infty$ that is going to be 0 - of -1.75. Therefore, its first component is 0.17.

And the second one is 0.1 – 0.22 that will give us 0.12 and therefore this is 0.17. Similarly infinite norm of x you can see that it is 0.1. So, the relative error in $\tilde{x}$ when compared to x is given by 1.7. Now let us see what is the relative error in the right-hand side vector? The right-hand side vector relative error is 0.01. Now you see in percentage you have made only 1% error in the input data.

That gave us 170% error in the solution. So, that seems to be very bad. Therefore, the coefficient matrix in this system is very sensitive to the input data. Let us see what is the condition number of this coefficient matrix.

(Refer Slide Time: 22:48)

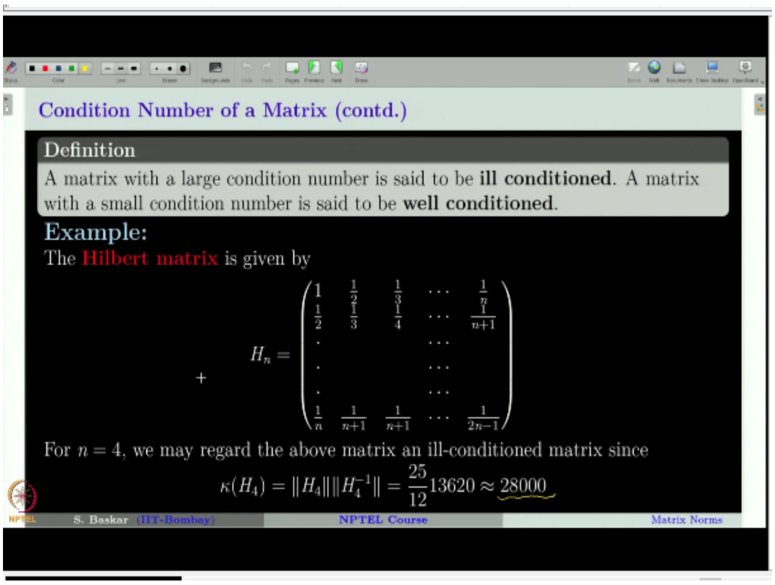


The condition number of the coefficient matrix is 289. How I computed this number? Well, I have taken the coefficient matrix as 5, 7, 7, 10 and now I have to find the subordinate matrix norm of A. What is that subordinate matrix norm? That is subordinate to the infinite norm. In the last class I have given the formula for this. So, you go back to our previous lecture and get the formula for this and compute this.

And also, you find the inverse of this matrix and similarly you can find the subordinate matrix norm of the inverse of this matrix and find this number also and then multiply both of them because the condition number is nothing but $||A||$. Here we are taking infinite norm therefore you have to use that formula into A^{-1} again infinite norm of A^{-1} that will happen to be 289.

So, you can see that the condition number of this matrix is pretty large, it is 289 and therefore when you are working with a matrix whose condition number is very large then you have to be very, very careful because even if you make small error in your input here, we made an error in the right-hand side vector. Similarly, you may make an error in the coefficient matrix also, accordingly it will also magnify the relative error in the approximate solution.

(Refer Slide Time: 24:39)

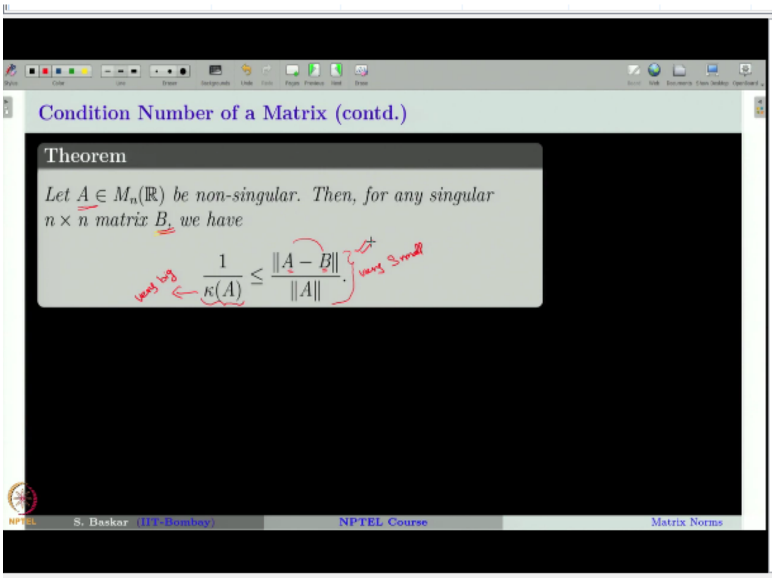


A matrix with a large condition number is said to be ill conditioned; whereas a matrix with small condition number is said to be well conditioned. Such concepts were also introduced in the case of evaluating a function, a C^1 function at some point in one of our previous classes. So, here the words large and small is in general very difficult for us to quantify how large it is and how small it is.

Because it depends on many factors; more importantly it depends on the computational power. A more powerful computer can handle a very bad matrices relatively better than a small computer. Also, it depends on what kind of applications that we are working in. Therefore, in general we cannot quantify what is large and what is small. Let us recall a very famous matrix called Hilbert matrix which comes quite often in many mathematical models.

Just with $n = 4$ one can see that the condition number of the Hilbert matrix H_4 is something like 28,000. Now you can imagine, if a matrix of condition number is amplifying the error in the solution so much what will happen to such matrices. This really tells us how serious it is for us to understand and then work with the matrices. Just going blindly with matrices and their computation is very dangerous on a computer.

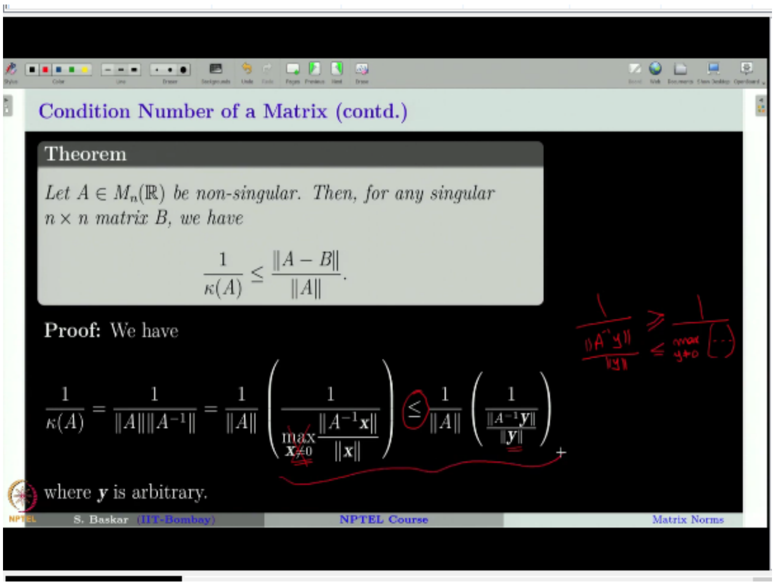
(Refer Slide Time: 26:41)



Before ending this lecture let us prove an important theorem which tells us how one can identify a bad matrix that is a matrix with large condition number. Generally, it may not be possible for us to use this theorem in practical situations. However, it is very interesting from the mathematical point of view. The theorem says that let us take a non-singular matrix A then for any singular matrix B you have 1 by condition number of A is less than or equal to the relative error in B when compared to A.

Therefore, if A is very close to B you can see that the relative error is going to be very small and that will make the reciprocal of the condition number to be very small. It means the condition number is very big. So, this theorem says that if you are working with a matrix which may be non-singular, but if it is very close to a singular matrix in this sense then the condition number of your original non-singular matrix will be very large. That is what the theorem says.

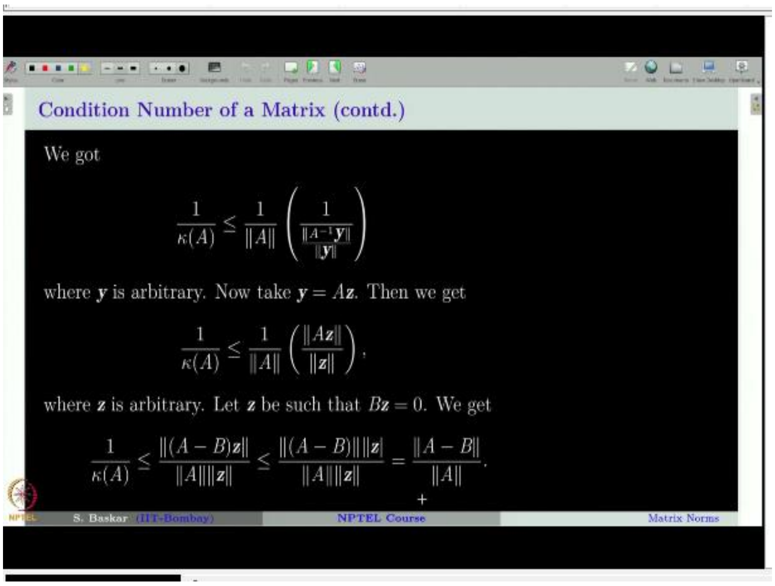
(Refer Slide Time: 28:09)



Let us quickly go through the proof of this theorem. You take the left-hand side $\frac{1}{\kappa(A)}$. I am just putting the definition of the condition number. Now I will fix this term and apply the subordinate matrix norm definition only for the second one that is $||A^{-1}||$. If you recall from our last lecture this is a definition of the subordinate matrix norm and now, I will simply take instead of this maximum I will take an arbitrary vector y.

Then what happens? I will just forget this maximum; therefore, I will have less than or equal to. Remember $||A^{-1}y||$ divided by $||y||$ is less than or equal to maximum of this y not equal to 0 of the same quantity. Therefore 1 by this will be 1 by this, but the inequality is reversed. That is what I have written here.

(Refer Slide Time: 29:30)

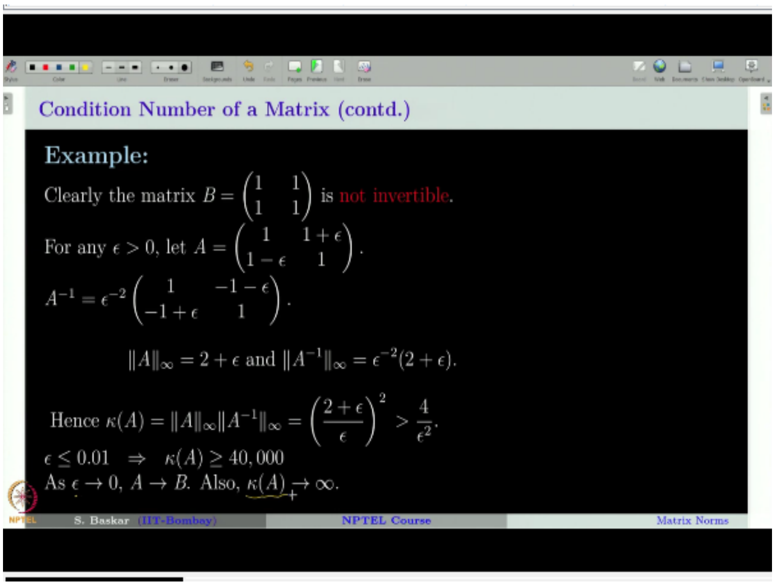


So, we got this inequality for our condition number. Let us see how to go ahead with that. I will just put $y = Az$ just to make this term to look little better here. Thereby I have $\frac{1}{||A||}$ into this goes to the numerator. I have Az divided by $A^{-1}y$ is now written as z. So, this holds for any z that is what we have seen. Now you take a non-zero z such that $Bz = 0$. How is that possible? Well, you go back to the theorem and see B is a singular matrix.

Therefore, this is possible. That is you can find a non-zero vector z, such that $Bz = 0$. Once you have such a z than remember this inequality holds for any z, I will in particular take my z such that $Bz = 0$. Therefore, I can write this term as $||Az - Bz||$. There is no harm in writing it because this is 0. So, I will write $\frac{1}{\kappa(A)}$ is less than equal to $||(A - B)z||$ divided by this. This is there.

Now I will use wherever you see $||Ax||$ kind of expression you should immediately remember our property of the subordinate matrix norm, you put that here, this is less than or equal to $||A - B||\ ||z||$. Now z, z gets canceled and you got what you want and this is what the theorem says.

(Refer Slide Time: 31:30)



Let us see quickly an example for this. Let us take the matrix B which is equal $to \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, which is not an invertible matrix. Now let us perturb the matrix B little bit by adding ϵ to some of its terms and call this as A. As long as ϵ is greater than 0 this is a non-singular matrix therefore you can find A^{-1} for this and now what is $||A||$ with respect to the subordinate matrix norm to l_∞ norm.

That is given by this quantity; you just go back to our previous lecture. Recall what is the formula for this, come back and apply it to this matrix you will get this answer and similarly $||A^{-1}||$ is given by this. Therefore, the condition number of the matrix *A* is given like this which is in our case is given by this expression and that can be greater than $\frac{4}{\epsilon^2}$.

Now you see if you take ϵ very small you can see that the condition number of the matrix *A* is going to be very large. So, that is what is given in the previous theorem also, we are just cross checking the theorem that we have proved just now in this example just to be more precise let us take ϵ to be something like 0.01 and that leads to your condition number of the matrix *A* as something like 40,000.

You can also see that your matrix *A* is pretty close to the singular matrix *B* whenever ϵ is very small. That is what is written here and consequently if you just take ϵ going to 0 you can see that the condition number of *A* will in fact tend to infinity.

So, this tells us that if you are working with a matrix which is very close to singular matrix, in the sense that is defined in the previous theorem, then you are likely to have very sensitive computation for your input data. So, in this lecture we have understood what is mean by condition number of a matrix and how to learn the sensitivity of the matrix in computing solution of a linear system in terms of condition number of a matrix. Thanks for your attention.