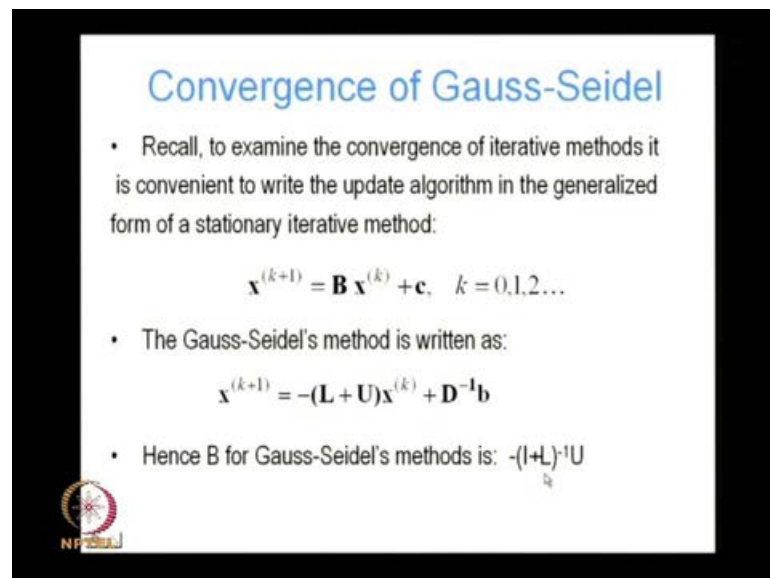


Numerical Methods in Civil Engineering
Prof. Arghya Deb
Department of Civil Engineering
Indian Institute of Technology, Kharagpur

Lecture -9
Iterative Methods-II


Continue our lecture on lecture series on Numerical Methods in Civil Engineering, we are going to continue our discussion on Iterative Methods for solving linear systems.

(Refer Slide Time 00:32)



Convergence of Gauss-Seidel

- Recall, to examine the convergence of iterative methods it is convenient to write the update algorithm in the generalized form of a stationary iterative method:
$$\mathbf{x}^{(k+1)} = \mathbf{B} \mathbf{x}^{(k)} + \mathbf{c}, \quad k = 0, 1, 2, \dots$$
- The Gauss-Seidel's method is written as:
$$\mathbf{x}^{(k+1)} = -(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}$$
- Hence B for Gauss-Seidel's methods is: $-(\mathbf{L} + \mathbf{U})\mathbf{D}^{-1}$



Last time, we looked at iterative methods in general and specifically we looked at, what is probably the simplest iterative method for solving linear systems, Jacobi's method. And we obtained convergence criteria for the Jacobi's method, we figured out under what conditions the repeated iterations using a constant coefficient matrix leads to convergence solutions. We also talked about Gauss Seidel method, which is a variation of the Jacobi method, where instead of continuing with the old iterates for entirely new cycle, we update the new values of the iterates, as they occurred during the solution.

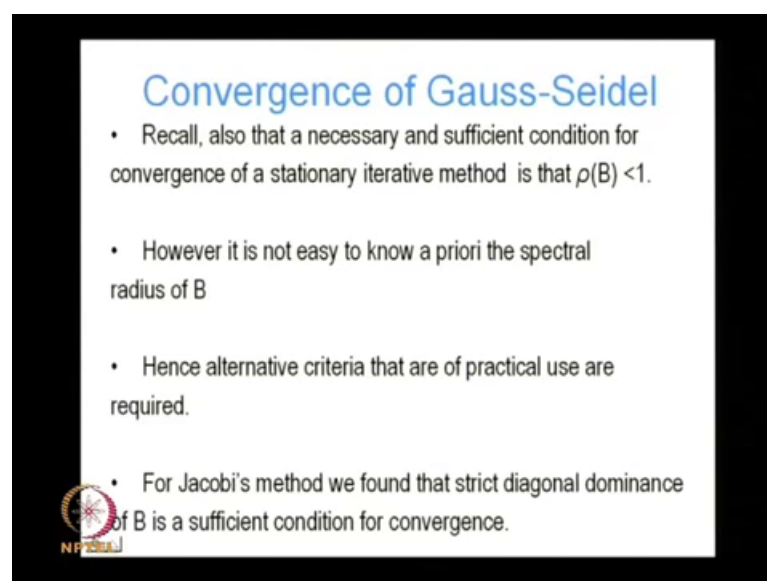
So, we talked about Gauss Seidel, this time we are going to talk about convergence of Gauss Seidel method. Recall, to examine the convergence of iterative methods, it is

convenient to write the update algorithm in the generalized form of a stationary iterative method. Where, x_{k+1} , which is the solution at the $k+1$ th iteration is equal to $Bx_k + c$, where x_k is the solution from the previous iteration plus a constant vector c .

And we continue this iteration for k equal to 0, 1, 2 and so on, until we reach convergence when x_{k+1} is equal to x_k within a certain tolerance and both of them can be taken to be or both of them are close to the true solution within certain error norms. Recall, the Gauss Seidel method in this form, in the form x_{k+1} is equal to $Bx_k + c$ can be written, if we represent the matrix B by $L + U$, where L and U are lower and upper triangular matrices, which are formed from a original coefficient matrix A .


Each of them recall as D is the diagonal matrix, which we defined last time. So, x_{k+1} is equal to $-(L + U)x_k + D^{-1}b$ that is, if we write the Gauss Seidel method in the generalized form of a stationary iterative method. Hence, B for Gauss Seidel method is $-(I + L)^{-1}U$, this also we have seen last time.

(Refer Slide Time 03:35)



Convergence of Gauss-Seidel

- Recall, also that a necessary and sufficient condition for convergence of a stationary iterative method is that $\rho(B) < 1$.
- However it is not easy to know a priori the spectral radius of B
- Hence alternative criteria that are of practical use are required.
- For Jacobi's method we found that strict diagonal dominance of B is a sufficient condition for convergence.

 NPTEL


Also we recall that, we found that, a necessary and sufficient condition for convergence of a stationary iterative method is that, the spectral radius of B be less than 1 that is, the largest Eigen value of B must have now magnitude less than 1. In that case, that is both a necessary as well as a sufficient condition for the convergence of a stationary iterative method.

However, it is not easy to know a priori the spectral radius of B, so it is not always possible to before doing the iterations to dissolve the Eigen value problem and find out the largest Eigen value of B, that is simply too cumbersome and too troublesome and too expensive. Hence, we would like to know, a come up with some alternative criteria that we can use to determine if our stationary iterative method is going to converge a criteria, which does not involve doing a spectral analysis and figuring out the Eigen values of B. For Jacobi's method, we found that, if the matrix B is diagonally dominant, that is a sufficient condition for convergence. That is a sufficient condition, but it is not necessary condition, it is a sufficient condition for convergence.

(Refer Slide Time 05:09)

Convergence of Gauss-Seidel

- To obtain convergence criteria for Gauss-Seidel's method, we recall that for any consistent matrix vector norm e.g. the infinite norm, we have: $\|B_{GS}\|_{\infty} = \max_{|x|=1} |B_{GS}x|$
- Denoting $B_{GS}x = y$, $\|B_{GS}\|_{\infty} = \max_{|x|=1} \frac{|y|_{\infty}}{|x|_{\infty}}$
- Let us suppose that: $|y|_{\infty} = \max_{i=1}^n |y_i| = |y_k|$
- Recall the form of the Gauss-Seidel's update formula $y = -Ly - Ux + D^{-1}c$ and neglecting the constant term $D^{-1}c$ for the time being, we consider the equation $y = -Ly - Ux$



So, we would like to find something similar for Gauss Seidel method, obtain a convergence criteria which is relatively easy to use and which does not required finding the largest Eigen value of B. We recall that, for any consistent matrix vector norm, for

instance the infinite norm we have then, the infinite norm of a matrix is equal to that matrix operating on a vector, that gives me a vector, for instance, if I want to find the infinite norm of B_{GS} , where GS represents the Gauss Seidel, B for Gauss Seidel.

We can find that out, we can find the infinite norm by taking the product of B_{GS} , the coefficient B matrix for Gauss Seidel with any arbitrary vector x such that, the norm of x is not equal to 0 that is, x is not the zero vector. So, B_{GS} times product with x , we find the norm of that, divided by the norm of x and whatever value if we max, maximise that, that is going to be my infinite norms for the matrix. This is our definition of a consistent norm, which we talked about earlier on.

Let us denote $B_{GS} x$ is equal to y , in that case the infinite norm of B_{GS} is equal to maximum of y , infinite norm divided by infinite norm of x and maximise it, that gives me the infinite norm of B_{GS} . Let us suppose that, the infinite norm of y , which is by definition, the maximum of the absolute of all the components, we calculate the absolute value of the components and take the maximum of that over all the components, that by definition is the infinite norm of y .

And let us suppose in that y vector, the k th entry, the k th component gives me the norm of y infinity. So, this is basically the largest entry in the y vector, let us also recall that, the Gauss Seidel's update formula y is equal to minus $L y$ minus $U x$ plus D inverse c and let us neglect the constant term D inverse c for the time being. We consider the equation y is equal to minus $L y$ minus $U x$, why do we ignore the constant term D inverse c , we are interested in the convergence. And when so interested in the convergence, we are interested in the convergence of y , so the constant term does not play big role in the convergence.

(Refer Slide Time 08:14)

Convergence of Gauss-Seidel

- Let us consider in particular the k^{th} equation in this system.

We have in that case:

$$|y|_{\infty} = |y_k| = \left| -\sum_{j=1}^{k-1} L_{kj} y_j - \sum_{j=k+1}^n U_{kj} x_j \right|_{\infty} \leq \left| \sum_{j=1}^{k-1} L_{kj} y_j \right|_{\infty} + \left| \sum_{j=k+1}^n U_{kj} x_j \right|_{\infty}$$

$$= \left| \sum_{j=1}^{k-1} \frac{a_{kj}}{a_{kk}} y_j \right|_{\infty} + \left| \sum_{j=k+1}^n \frac{a_{kj}}{a_{kk}} x_j \right|_{\infty}$$

Since $|y_k| > |y_j| \quad \forall j \neq k$ we have:

$$|y|_{\infty} \leq \left| \sum_{j=1}^{k-1} \frac{|a_{kj}|}{|a_{kk}|} |y_k| \right|_{\infty} + \left| \sum_{j=k+1}^n \frac{|a_{kj}|}{|a_{kk}|} \max_{k=1}^n |y_k| \right|_{\infty} = \sum_{j=1}^{k-1} \frac{|a_{kj}|}{|a_{kk}|} |y|_{\infty} + \sum_{j=k+1}^n \frac{|a_{kj}|}{|a_{kk}|} |x|_{\infty}$$

Let us consider in particular the k th equation in this system, this system I mean, this system, let us consider this simplified system, where we have got rid of the constant vector D inverse c . Because, we said that is not important for convergence for studying the convergence of this method. And we say that, we consider this y is equal to minus L y minus U x and we know that, since the infinite norm of y is given by the k th term in y . So, basically, we find out the k th term on their both on left hand side and the right hand side and the k th term is equal to sigma j equal to 1 to k minus 1 L k j y j minus U k j x j .

So, this is the lower triangular part, this is the upper triangular part and we can see this is L k and U k . Because, we are only considering the k th equation in the system of equations, y k is equal to the k th row of this matrix times the y vector, the k th row of the u matrix times the x vector. So, we can write that like that, just taking the norm on both sides and this is going to be less than this plus ((Refer Time: 09:52)) this, again this minus this is going to be less than this, norm of this plus norm of this.

Recall, we have done this before and this again L k j we can write as a k j by a k k by definition, because a lower triangular matrix is just the original matrix scaled by the D inverse, D inverse times the original matrix. The lower triangular part of the original matrix scaled by D inverse that is, my L , so that is this term and this is my upper

triangular matrix, which is basically the U, which is basically the upper triangular part of a scaled by D inverse, that is why we have this a k k.

So, this is less than that and this is equal to this, while that is equal to that, since we know that, mod of y k is greater than y j, for all j not equal to k, because we have assumed that, y k is the term in y, the entry in y which has the largest absolute value. So, we can write mod of y infinity is lesser than or equal to this term and now, I am going to replace y j by mod of y k. So, since this is y j, but since this is always greater than every j except k, if I replace this by mod of y k, the right hand side is going to be greater than the this this.

Similarly, instead of x j, if I replace it by the maximum component of x, instead of summing it over all the components, I submit all the time with the maximum component of x obviously, that is an upper bound on this. So, if I replace that by that, so but this is nothing but y infinity, that is nothing but, x infinity.

(Refer Slide Time 12:05)

Convergence of Gauss-Seidel

Denoting $\sum_{j=1}^{k-1} \frac{|a_{kj}|}{|a_{kk}|} = s_k$ and $\sum_{j=k+1}^n \frac{|a_{kj}|}{|a_{kk}|} = r_k$ we can write :

$$|y|_{\infty} \leq s_k |y|_{\infty} + r_k |x|_{\infty}$$

$$\therefore \frac{|y|_{\infty}}{|x|_{\infty}} \leq \frac{r_k}{1-s_k} \text{ and } |B_{GS}|_{\infty} \leq \max_i \frac{r_i}{1-s_i}$$

since $\max_{1 \leq i \leq n} \frac{r_i}{1-s_i}$ must be greater than $\frac{r_k}{1-s_k}$

- But from the definition of r_i and s_i , it can be shown that for diagonally dominant matrices, $r_i + s_i < 1$

So, we can write this as, if we write this as x k and write this as r k, we can write this mod of y infinity is lesser than or equal to s k times mod of y infinity plus r k times mod of x infinity So, we get again bringing by mod of y infinity to the left hand side, the

infinite norm of y we have bring it, collect it bring it to the left hand side, we can write this as $\|y\|_\infty$ by $\|x\|_\infty$ is lesser than or equal to r_k divided by $1 - s_k$.

So, $1 - s_k$ $\|y\|_\infty$ is lesser than or equal to r_k times infinite norm of x . So, infinite norm of y divided by infinite norm of x is lesser than or equal to r_k divided by $1 - s_k$. And recall that, we said that, infinite norm of B GS is equal to maximum of r_i by x infinity. So, we can write B GS is lesser than or equal to maximum of r_i by $1 - s_i$, because this is equal to that and we saw that, this was maximum over this, so maximum of r_i by $1 - s_i$, i is equal to 1 to n .

Since maximum of r_i obviously, because this maximum of r_i $1 - s_i$ must be greater than any value r_k $1 - s_k$, so if we replace this on the right and replace this with maximum of that, this has got to be less than this. But, from the definition of r_i and s_i , it can be shown that, for diagonally dominant matrices $r_i + s_i$ is less than 1 , why is that $r_i + s_i$, we denoted r_i to be this and we denoted r_k to be this and s_k to be this.


(Refer Slide Time 14:34)

Convergence of Gauss-Seidel

$$r_i + s_i = \sum_{j=1}^{k-1} \frac{|a_{ij}|}{|a_{ii}|} + \sum_{j=k+1}^n \frac{|a_{ij}|}{|a_{ii}|} = \sum_{j=1}^n \frac{|a_{ij}|}{|a_{ii}|} < 1$$

$\therefore \|B_{GS}\| \leq 1$

- Hence Gauss-Seidel's method is convergent when the B matrix is diagonally dominant.
- Thus both Gauss-Seidel's and Jacobi's method are bound to converge to the true solution with increase in the number of iterations if the B matrix is diagonally dominant.



So, if we write $r_i + s_i$ that is, so just replacing k with i , so that is this plus this, which is equal to, you can see the sum is over j is equal to 1 to $k - 1$ and here the sum is

over k plus 1 to n , so the sum is actually over j is equal to 1 to n , other than when j is not equal to i . So, this is equal to $\text{mod of } a_{ij} \text{ by mod of } a_{ii}$, which is always less than 1 if the matrix is diagonally dominant, that is the definition of diagonally dominance. The sum of the absolute values of the half diagonal terms is going to be less than the absolute value of the diagonal term that is, when matrix is strictly diagonally dominant.

So, if the matrix is strictly diagonally dominant, $r_i + s_i$ will be less than 1, which means that, B_{GS} must be less than 1. Hence, again similar to Jacobi's method, we see that Gauss Seidel method is convergent when the B matrix is diagonally dominant, thus both Gauss Seidel and Jacobi's method are bound to converge to the true solution with increase in the number of iterations if the B matrix is diagonally dominant. If my B matrix, which in case of Gauss Seidel is given by this, which for Gauss Seidel is given by this. If this matrix is diagonally dominant then, I am guaranteed that, my Gauss Seidel iteration is going to converge after a certain number of iterations.

(Refer Slide Time 16:41)

Rate of Convergence

- However recall that the rate of convergence varies depending on the factor $R = -\log(\rho_B)$ the asymptotic rate of convergence, which again depends on the spectral radius of B .
- The rate of convergence depends on how fast the error $|\mathbf{x}^{(k)} - \mathbf{x}|$ goes to zero as k , the number of iterations, increases
- To get an estimate for the error in $\mathbf{x}^{(k)}$ we use the relation:

$$\mathbf{x}^{(k)} - \mathbf{x} = -\mathbf{B}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) + \mathbf{B}(\mathbf{x}^{(k)} - \mathbf{x})$$
 [Recall $\mathbf{x}^{(k)} = \mathbf{B}\mathbf{x}^{(k-1)} + \mathbf{c}$; $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{c}$]

So, that is convergence, however what is important is also the rate of convergence, something might converge, but it may turn out that, it converges so slowly, where it might be practically useless. So, something converges after 1 million iterations that is not going to be useful for any practical computational purposes. So, we are also interested

not only in the fact that, the iterative method converges, it must converge in a reasonable number of iterations.

Therefore, the rate of convergence is also very important and we recall that, the rate of convergence depends on this factor R , which is equal to $-\log \rho(B)$, which we call the asymptotic rate of convergence, $\rho(B)$ being the spectral radius of B . The rate of convergence depends on, how fast the error at iteration k , which I denote as $\|x_k - x\|$ goes to 0 as k , the number of iterations increases. So, suppose at iteration k , I had certain error which is given by this, I would like to know the rate of...

If it converges fast then, it would converge to it, the error would go down fast as the number of iterations increases. To get an estimate of the error in x_k , we use the following relation, $x_k - x$ is equal to, recall that x_k is equal to $Bx_{k-1} + c$ that is, standard for a stationary iterative method. So, we can write $x_k - x$ as $Bx_{k-1} + c - x$, basically $Bx_{k-1} + c - Bx + c$ cancels out, we have $Bx_{k-1} - Bx$.

And recall that, x is equal to $Bx + c$, so we are recovering this equation, so basically we are writing this equation in a slightly different form, this equation I am rewriting it in a slightly different form, why is x is equal to $Bx + c$. Because, at convergence after it has converged, x_{k+1} is almost equal to x_k . So, x_k is equal to $Bx + c$, where x is the true solution, x_k is the converged solution, once it has converged, x_k is equal to $Bx + c$, so we use this equation to get an estimate for the error.

(Refer Slide Time 19:26)

Error estimate


If $\|B\| = \beta < 1$, we get :

$$\begin{aligned} \|x^{(k)} - x\| &\leq \|B\| \|x^{(k)} - x^{(k-1)}\| + \|B\| \|x^{(k)} - x\| \\ &= \beta \|x^{(k)} - x^{(k-1)}\| + \beta \|x^{(k)} - x\| \end{aligned}$$

Hence, $(1 - \beta) \|x^{(k)} - x\| = \beta \|x^{(k)} - x^{(k-1)}\|$

$$\therefore \|x^{(k)} - x\| \leq \frac{\beta}{(1 - \beta)} \|x^{(k)} - x^{(k-1)}\| \quad (*)$$

- Thus only when $\beta \leq 0.5$ can we be assured that the iterate $x^{(k)}$ is sufficiently close to the true solution x i.e. its difference from the true solution is less than the magnitude of the last correction

 We can then stop the iteration.

Let us see, what we do if mod of B is equal to beta less than 1, so if the infinite norm of B is suppose beta and suppose it is less than 1, which we know is necessary for convergence. We get mod of $x^{(k)} - x$, taking the norm on both sides, norm of $x^{(k)} - x$ is lesser than or equal to norm of B norm of $x^{(k)} - x^{(k-1)}$ plus norm of B $x^{(k)} - x$. So, just taking the norm of both sides and using the fact that, norm of this operating on that is lesser than or equal to norm of this times the norm of that.

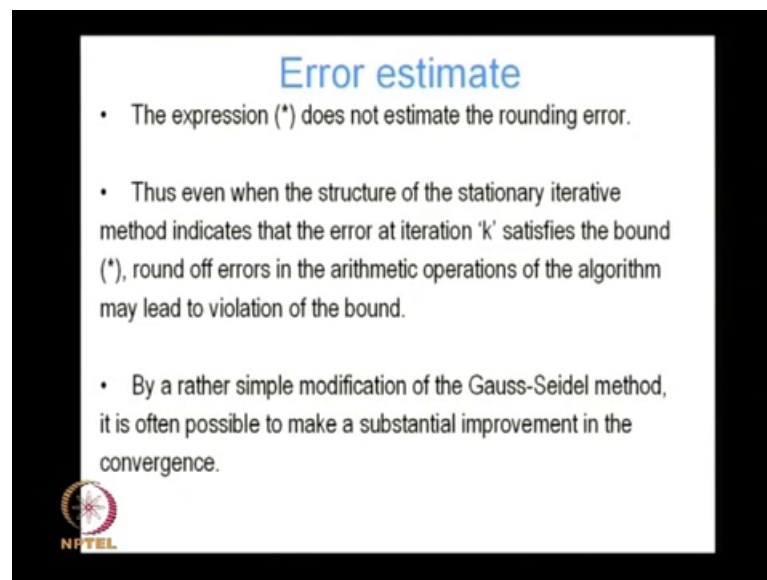
And using the fact that, norm of B is equal to beta, we get beta times norm of $x^{(k)} - x^{(k-1)}$ plus beta times norm of $x^{(k)} - x$. Again bringing this term to the left hand side, we have $1 - \beta$ norm of $x^{(k)} - x$ is equal to beta times norm of $x^{(k)} - x^{(k-1)}$. Therefore, norm of $x^{(k)} - x$ is lesser than or equal to beta by $1 - \beta$ norm of $x^{(k)} - x^{(k-1)}$. We will see that, only when beta is lesser than or equal to 0.5, this term is going to be less than 1, this term is going to be less than 1, beta by $1 - \beta$ is going to be less than 1 only if beta is lesser than or equal to half.

Then, we can be assured that means, this term is less than 1 so that means, that the iterate $x^{(k)}$ is sufficiently close to the true solution. That is, its difference from the true solution, norm of $x^{(k)} - x$, the difference of $x^{(k)}$ with the true solution is less than or equal to the magnitude of the last correction, $x^{(k)} - x^{(k-1)}$, if

this factor is less than 1 that means, this difference is less than to the magnitude of the last correction.


So, in that case, once we reach, once we satisfy that criteria, we can stop our iteration, we say that, our solution x_k has almost converged, because the change x_k is closer to the true solution than the change from the last iteration value, so that is when it has converged.

(Refer Slide Time 22:13)



Error estimate

- The expression (*) does not estimate the rounding error.
- Thus even when the structure of the stationary iterative method indicates that the error at iteration 'k' satisfies the bound (*), round off errors in the arithmetic operations of the algorithm may lead to violation of the bound.
- By a rather simple modification of the Gauss-Seidel method, it is often possible to make a substantial improvement in the convergence.

 NPTEL

The expression star however, does not take into account in round off error, so we are assuming that, everything is being done in infinite precision when we write that equation, so does not take into account rounding error. Thus, even when the structure of the stationary iterative method indicates that, the error at iteration k satisfies this bound ((Refer Time: 22:38)) even when we think that it satisfied this bound, round off errors in the arithmetic operations of the algorithm may lead to violation of the bound. Next we talk about slight modification to the Gauss Seidel method, which substantially improves the convergence of Gauss Seidel algorithm.

(Refer Slide Time 23:04)


SOR algorithm

- Recall that the Gauss Seidel algorithm is given by:

$$x_i^{(k+1)} = \frac{-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} + b_i}{a_{ii}}, \quad i = 1, \dots, n$$

Denoting $x_i^{(k+1)} = x_i^{(k)} + r_i^{(k)}$ for the i^{th} equation, we can write :

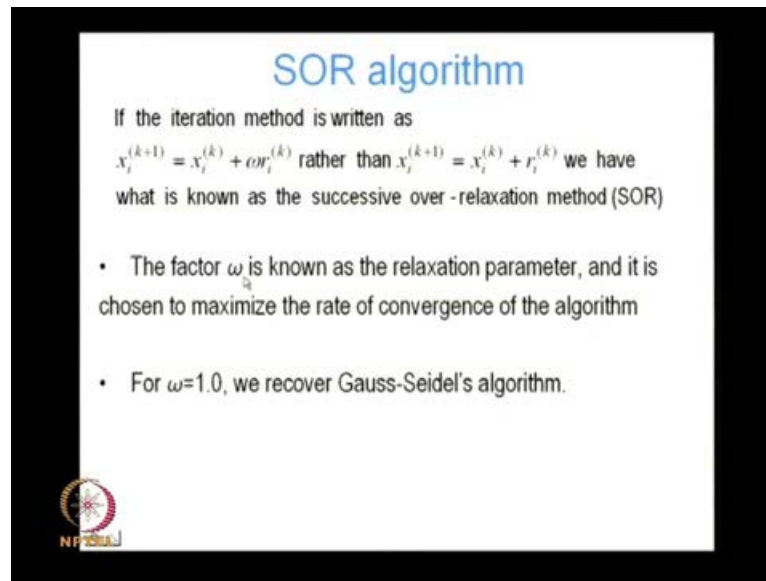
$$x_i^{(k+1)} = \frac{-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} + b_i}{a_{ii}} + x_i^{(k)}$$

$$r_i^{(k)} = \frac{-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} + b_i}{a_{ii}}$$


That is, recall that the Gauss Seidel algorithm is given by x_i^{k+1} is equal to this, where this is the lower triangular path comprising the values of the elements, which I have already been found from the new iteration. The values of the elements, which I have already been found if continuing iteration, this involves the values, which I using from the old iteration and this is the right hand side. So, we recall this is the Gauss Seidel algorithm then, we denote x_i^{k+1} is equal to $x_i^k + r_i^k$ for the i^{th} equation.

In that case, we can write x_i^{k+1} is equal to, this part remains the same, but this part instead of summing from j equal to $i+1$ to n , now I am going to sum from j is equal to i to n . So, I am adding an extra term here, what is that extra term, the term for j is equal to i , which is nothing but, $a_{ii}x_i^k$. $a_{ii}x_i^k$ divided by a_{ii} that means, I am adding a minus x_i^k and I am compensating for that by adding another x_i^k . So, I rewrite this like this and since we have denoted x_i^{k+1} is equal to $x_i^k + r_i^k$, so this whole thing becomes my r_i^k .

(Refer Slide Time 24:46)



SOR algorithm

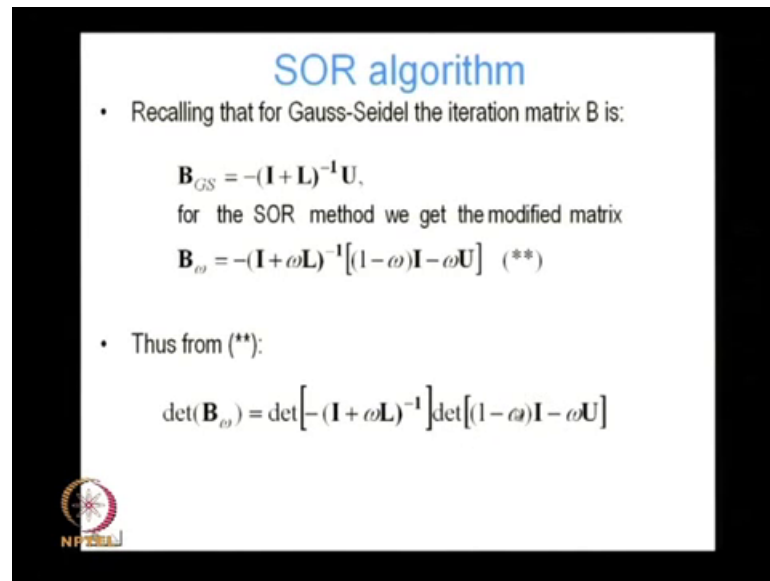
If the iteration method is written as
 $x_i^{(k+1)} = x_i^{(k)} + \omega r_i^{(k)}$ rather than $x_i^{(k+1)} = x_i^{(k)} + r_i^{(k)}$ we have
what is known as the successive over-relaxation method (SOR)

- The factor ω is known as the relaxation parameter, and it is chosen to maximize the rate of convergence of the algorithm
- For $\omega=1.0$, we recover Gauss-Seidel's algorithm.

If the iteration method is written as $x_i^{(k+1)} = x_i^{(k)} + \omega r_i^{(k)}$, so recall what is Gauss Seidel, Gauss Seidel says that, $x_i^{(k+1)} = x_i^{(k)} + r_i^{(k)}$. So now, we are saying that, we are not going to use full Gauss Seidel, we are going to use $x_i^{(k+1)} = x_i^{(k)} + \omega r_i^{(k)}$, ω is a scalar, some scalar times $r_i^{(k)}$. So, we are just changing the Gauss Seidel algorithm, instead of using the full $r_i^{(k)}$, which we recover when ω is equal to 1, we are going to use some fraction, either we will see what are the liable values of ω .


But, for the time being, let us suppose that, we are multiplying this by scalar times $r_i^{(k)}$, instead of just taking $r_i^{(k)}$ then, we have what is known as the successive over relaxation method or the SOR method. And the factor ω is known as the relaxation parameter and it is chosen to maximize the rate of convergence of the algorithm. So, pure Gauss Seidel, $x_i^{(k+1)} = x_i^{(k)} + r_i^{(k)}$, SOR method instead of taking $r_i^{(k)}$, we are multiplying $r_i^{(k)}$, we are scaling the $r_i^{(k)}$ s with the scalar ω . And we are going to choose this scalar ω so that, we get maximum rate of convergence of the algorithm.

(Refer Slide Time 26:30)



SOR algorithm

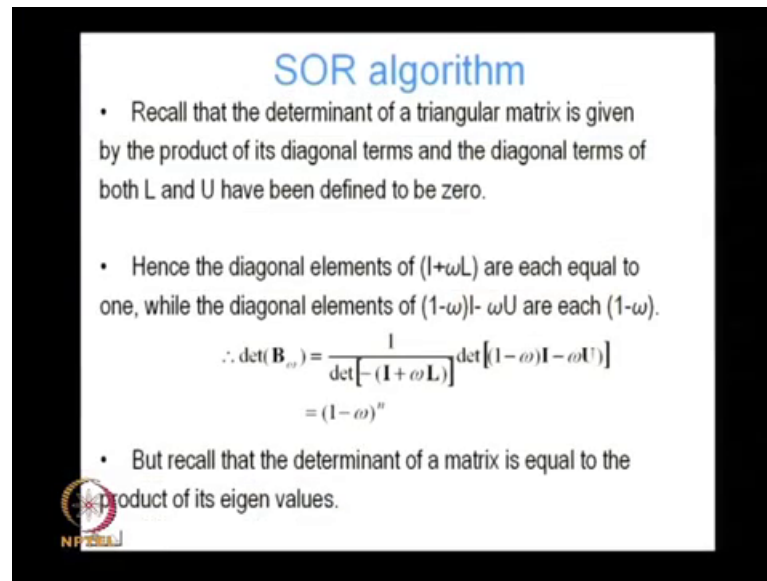
- Recalling that for Gauss-Seidel the iteration matrix B is:
$$\mathbf{B}_{GS} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U},$$
for the SOR method we get the modified matrix
$$\mathbf{B}_{\omega} = -(\mathbf{I} + \omega\mathbf{L})^{-1}[(1 - \omega)\mathbf{I} - \omega\mathbf{U}] \quad (**)$$
- Thus from (**):
$$\det(\mathbf{B}_{\omega}) = \det[-(\mathbf{I} + \omega\mathbf{L})^{-1}] \det[(1 - \omega)\mathbf{I} - \omega\mathbf{U}]$$



Recall again, that for Gauss Seidel, our iteration matrix B is given by B GS is equal to minus i plus L inverse U, which we have seen several times. For the SOR method, we get a modified B matrix, B omega which we denote by B omega, which is equal to minus i plus omega L inverse 1 minus omega i minus omega U. I have not actually shown this, but I am asking you to take this for granted, it can be shown quite easily that, instead of taking of $x_i^k + 1$ is equal to $x_i^k + r_i^k$, I replace $x_i^k + 1$ is equal to $x_i^k + \omega r_i^k$ then, my B matrix changes from this matrix to this matrix.

So, if we use the SOR method, my B matrix changes from this to this, from this if we take determinant of both sides, what do we have, determinant of B omega is equal to determinant of this times determinant of that and recall the determinant of A inverse is equal to 1 by determinant of A.

(Refer Slide Time 27:49)



SOR algorithm

- Recall that the determinant of a triangular matrix is given by the product of its diagonal terms and the diagonal terms of both L and U have been defined to be zero.
- Hence the diagonal elements of $(I + \omega L)$ are each equal to one, while the diagonal elements of $(I - \omega)I - \omega U$ are each $(1 - \omega)$.

$$\therefore \det(\mathbf{B}_{SOR}) = \frac{1}{\det[-(I + \omega L)]} \det[(1 - \omega)I - \omega U]$$
$$= (1 - \omega)^n$$

- But recall that the determinant of a matrix is equal to the product of its eigen values.

NPTTEL

So, we are going to use that determinant of B inverse, therefore is equal to 1 by determinant of this part and trans the determinant of 1 minus omega I minus omega U, which is this part. But, recall that the determinant of a triangular matrix, which we have seen several times before, is given by the product of the diagonal terms, product of the terms on the principle diagonal. And by definition, we have defined L and U to have 0s on their principle diagonal.

So obviously, the determinant of both L and U are 0, hence the diagonal elements of I plus omega L, this matrix, this term is not going to contribute anything. So, on the diagonal, are each equal to 1, so this is going to be a triangular matrix, because this is a lower triangular matrix. To this, we are adding a the identity matrix, so the result is going to remain a lower triangular matrix. And for a lower triangular matrix, the determinant is the product of the diagonal terms and since the diagonal terms of L are 0, the diagonal terms of I plus omega L are nothing but, 1.

So, this determinant is going to be 1 and the diagonal elements of 1 minus omega I minus omega U, similarly again U has got 0s on its diagonal. So, on it is diagonal of this entire matrix, we have 1 minus omega, each diagonal term is equal to 1 minus omega. So, this determinant is going to be 1 minus omega to the power n, where n is the size of

the matrix. So, this gives me determinant of B_ω is equal to $1 - \omega^n$, but let us go back, take a step back and recall that, determinant of a matrix is equal to the product of its Eigen values.

(Refer Slide Time 30:07)

SOR algorithm

- Hence:
 $\det(\mathbf{B}_\omega) = \lambda_1 \lambda_2 \dots \lambda_n$ where $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigen values of \mathbf{B}_ω
Hence, $\lambda_1 \lambda_2 \dots \lambda_n = (1 - \omega)^n$
 $\therefore \max_i \lambda_i \geq |1 - \omega|$ and since for convergence, $\max_i \lambda_i \leq 1$, we have:
 $1 \geq \max_i \lambda_i \geq |1 - \omega| \Rightarrow |1 - \omega| \leq 1 \Rightarrow 0 < \omega < 2$
- Thus the SOR method only converges for $0 < \omega < 2$.
- It can be shown that if the coefficient matrix A is symmetric and positive definite then the SOR method always converges when $0 < \omega < 2$ (necessary as well as sufficient condition)

We have seen that before, therefore what do we have, we have determinant of B_ω , which is equal to $1 - \omega^n$ must be equal to $\lambda_1 \lambda_2 \dots \lambda_n$, where $\lambda_1 \lambda_2 \dots \lambda_n$ are the Eigen values of B_ω . Therefore, the maximum Eigen value must be greater than or equal to $1 - \omega$ why, because this product of all these Eigen values is equal to $1 - \omega^n$.

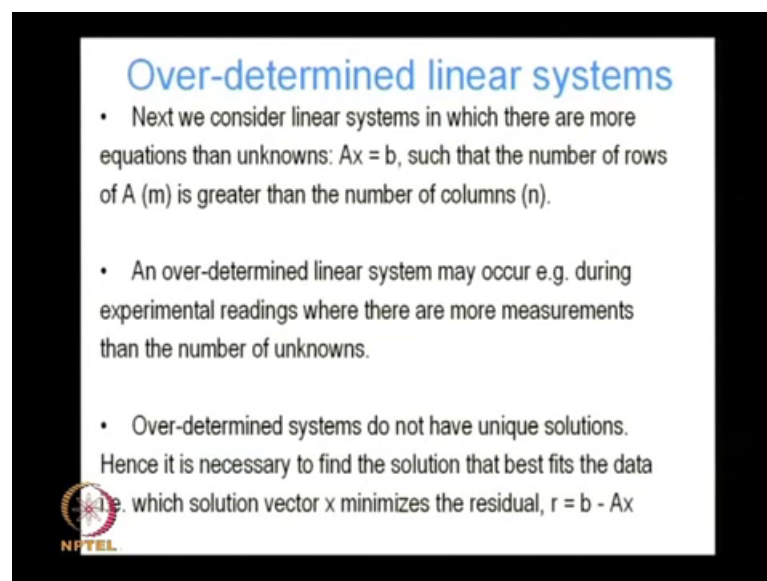
Therefore, the maximum value must be greater than or equal to $1 - \omega$, I think a minute about it that should be clear, this is the product of all the Eigen values is this. So, it is clear that, its maximum value must be greater than the n th root of this, so maximum λ_i must be greater than or equal to $|1 - \omega|$. And recall also that, for convergence what is our requirement, the necessary and sufficient condition for convergence is that, the maximum Eigen value of B has to be less than 1.

So, maximum λ_i over i must be lesser than or equal to 1, we therefore have this

condition, maximum λ over i is greater than or equal to 1. But, this is again greater than or equal to $1 - \omega$, which mod of $1 - \omega$, which implies mod of $1 - \omega$ must be lesser than or equal to 1 which means that, ω must lie between 0 and 2. That is what do we see that, the SOR method, the Successive Over Relaxation method is only going to converge and ω lies between 0 and 2.

So, this value of ω , we cannot choose any arbitrary value of ω and hope that, we are going to converge. Only if we choose ω equal lying between 0 and 2, are we going to get convergence and it can be shown that, this the coefficient matrix A is symmetric and positive definite then, the SOR method always convergence when ω is bias between 0 and 2, so it is a necessary as well as a sufficient condition.

(Refer Slide Time 32:38)



Over-determined linear systems

- Next we consider linear systems in which there are more equations than unknowns: $Ax = b$, such that the number of rows of A (m) is greater than the number of columns (n).
- An over-determined linear system may occur e.g. during experimental readings where there are more measurements than the number of unknowns.
- Over-determined systems do not have unique solutions. Hence it is necessary to find the solution that best fits the data i.e. which solution vector x minimizes the residual, $r = b - Ax$

NPTEL

So, that was our discussion about the Gauss Seidel method and the variation of the Gauss Seidel method. And before moving on to something else, this is a good time to talk about certain linear systems, which are also amenable to iterative solutions. For instance, systems you may specifically that we want to talk about systems, which are over determined, where we have number of equations. The number of equations is more than the number of unknowns, which is an over determined system.

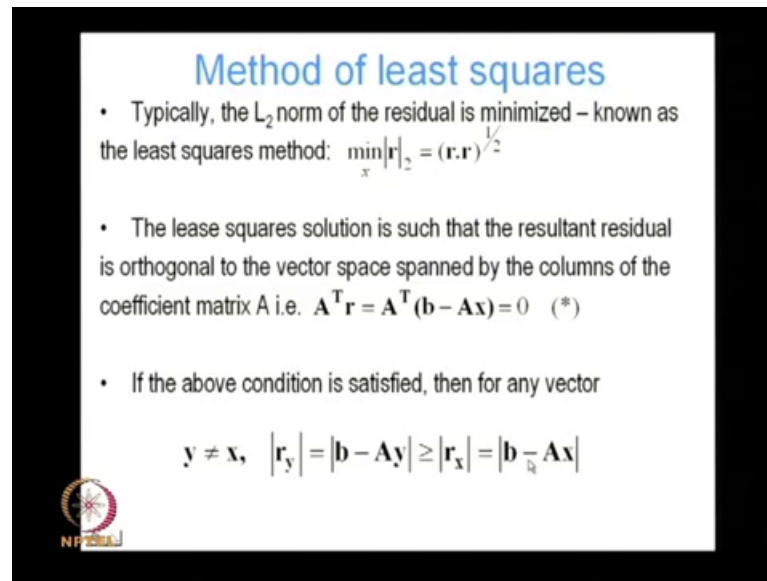
For instance, that means, basically we are considering a system $Ax = b$ such that, the number of rows of A m is greater than the number of columns n . An over determined linear system may occur during experimental readings when there are more measurements than the number of unknowns. If we are trying to find some values x and we are setting certain parameters, which are governed by the entries in A and we are looking at certain results.

And from those results and those parameters, we want to find out the values of our unknown variables x , we perform a large number of experiments, more experiments than there are unknowns. So, in that case, we come up with an over determined system and over determined systems do not have unique solutions, you can understand, you have more equations than unknowns. So, there are no unique solutions and which is in that case, it becomes necessary to find the best solution.

Best solution means, something which best fits the data, suppose we do a whole number of experiments and we have a lot of parameters, we vary a lot of parameters and we find the lot of results. But, our independent variables, the number of independent variables are less than the number of reading's that we have taken. So, we want to find out, what values of the independent variables, what are the values of independent variables, which best fit are experimental data, which best fit our values b , which we have obtained as a consequent to our experiment.

So, it is necessary to find the solution that best fits the data, so these are suppose, our experimental readings and A represents the parameters we have varied and vary the parameters, we have found our experimental results. And now, we are interested in finding the independent values of the independent variables, which best fit our experimental results b . So, basically we are interested in minimizing the residual, residual we define as $b - Ax$. So, again this becomes an iteration, it can be solved iteratively, we try with successive values of x , find out what is the residual and continue iterating till you get a relatively small or acceptable value of the acceptably small value of the residual.


(Refer Slide Time 36:04)



Method of least squares

- Typically, the L_2 norm of the residual is minimized – known as the least squares method: $\min_x \|r\|_2 = (\mathbf{r} \cdot \mathbf{r})^{1/2}$
- The least squares solution is such that the resultant residual is orthogonal to the vector space spanned by the columns of the coefficient matrix A i.e. $\mathbf{A}^T \mathbf{r} = \mathbf{A}^T (\mathbf{b} - \mathbf{A}\mathbf{x}) = 0$ (*)
- If the above condition is satisfied, then for any vector

$$y \neq x, \quad \|r_y\| = \|\mathbf{b} - \mathbf{A}y\| \geq \|r_x\| = \|\mathbf{b} - \mathbf{A}x\|$$



But, it turns out that, instead of doing that, if we try to minimize the L_2 norm of the residual, we can come up with the more systematic solution. So, basically we find the norm of the residual in the L_2 sense, basically we take the r dotted with r , we find the scalar and take the square root of that, which gives me that two norm. And then, if we try to find the value of x , which minimizes the value of the residual, we get what is known as the least squares solution.

We can see why it is called least squares, it is the value of x which minimizes the square root of the inner product of r with r . The least squares solution is such that, the resultant residual is orthogonal to the vector space spanned by the columns of the coefficient matrix. Basically, $\mathbf{A}^T \mathbf{r}$ is equal to 0, we call we defined r is equal to $\mathbf{b} - \mathbf{A}x$, if we take the projection on A , $\mathbf{A}^T (\mathbf{b} - \mathbf{A}x)$ means, we are taking the projection of this vector $\mathbf{b} - \mathbf{A}x$ on the space, that is spanned by the columns of A .

That linear space which is spanned by the columns of A , which has got basis vectors, which are given by the each individual column of A . If we take the projection of $\mathbf{b} - \mathbf{A}x$ on that space, which is spanned by the columns of A then, we get 0. So, basically the residual has zero component in the space that is spanned by the columns of A , so it is orthogonal to the space that is, to the vector space that is spanned by the columns of the

coefficient matrix A.

If the above condition is satisfied, we can show that, for any vector y, which is not equal to x, x being the vector which minimizes the residual. For any vector y, which is not equal to x, the residual which we now denote by r y, which is b minus A y instead of A x. Y is different from x, x minimizes the residual that, in that case, b minus A y is always going to be greater than b minus A x. So, if we find x such that, x satisfies this condition that, any vector y will give me a larger residual or if this is sort of obvious. Because, we are finding x, which minimizes the residual, so any vector y is going to give me a larger residual.

(Refer Slide Time 39:15)

Method of least squares

- This is proved as follows:

$$r_y = (b - Ax) + (Ax - Ay) = r_x + A(x - y)$$

$$\therefore r_y \cdot r_y = r_x \cdot r_x + r_x \cdot A(x - y) + A(x - y) \cdot r_x + A(x - y) \cdot A(x - y)$$

$$= r_x \cdot r_x + A^T r_x \cdot (x - y) + (x - y) \cdot A^T r_x + (x - y) \cdot A^T A(x - y)$$

But from (*), $A^T r_x = 0$

$$\therefore r_y \cdot r_y = (x - y) \cdot A^T A(x - y) + r_x \cdot r_x$$

$$\therefore |r_y|^2 = |r_x|^2 + |A(x - y)|^2 \geq |r_x|^2$$

That is intuitively clear that, it can be proved quite easily also, so just let us go through the proof ones, r y is equal to, let us recall how we defined r y, r y is equal to b minus A y. So, r y is equal to b minus, instead of writing b minus A y, I write it as b minus A x plus A x minus A y, but this is equal to r x. My residual for my least squares solution x, b minus A x is equal to r x plus A x minus y, I take dot product in a product on both sides.

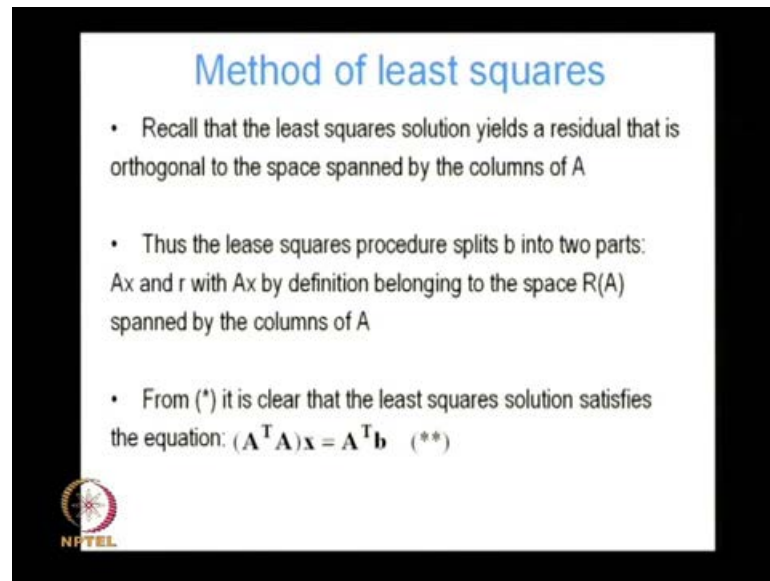
So, r y dotted with r y is equal to r x dotted with r x plus r x dotted with this, plus A x minus y dotted with that, plus A x minus y dotted with A x minus y, where A x minus y I

am treating it as a vector. A matrix operating on a vector after all gives me always a vector, so this is a vector dotted with that. So, this is nothing but that, this is we can write $r \cdot x$ dotted with $A \cdot x - y$ as $A^T r \cdot x$ dotted with $x - y$. Similarly, $A \cdot x - y$, I can write it $x - y$ dotted with $A^T r \cdot x$, you can verify with this is always true, this is our transpose is defined.

The mathematical definition of transpose is like this, so this you can verify and this again we can write, again we do the same trick, we bring $x - y$ $A^T r \cdot x - y$. So, these are identical equal, but let us look at this, $A^T r \cdot x$ is equal to 0 by definition why, because that is what we have here, that is exactly what we have here. So, this term goes to 0, this term goes to 0, because $A^T r \cdot x$ goes to 0 and we are left with $r \cdot x$ dotted with $r \cdot x + x - y$ $A^T r \cdot x - y$.


And this is then, we rewrite this as the square of the norm of $r \cdot y$ is equal to the square of the norm of $r \cdot x$ term and then, we again treat this as a vector $x - y$ $A^T r \cdot x - y$, I can write as $A \cdot x - y$ dotted with $A \cdot x - y$. So, it is as if I am taking the dot product in a product of a vector $A \cdot x - y$ and that I can write as $A \cdot x - y$ whole square. Since this is not going to be 0 for x not equal to y then, we can say that, $r \cdot y$ square is always going to be greater than $r \cdot x$ square. Because, this term is not going to be 0, this is something intuitive, but this is just a proof for that.

(Refer Slide Time 42:15)



Method of least squares

- Recall that the least squares solution yields a residual that is orthogonal to the space spanned by the columns of A
- Thus the least squares procedure splits b into two parts: Ax and r with Ax by definition belonging to the space $R(A)$ spanned by the columns of A
- From (*) it is clear that the least squares solution satisfies the equation: $(A^T A)x = A^T b$ (**)



Recall that, the least squares solution yields a residual that is orthogonal to the space spanned by the columns of A . So, the residual $b - Ax$ is orthogonal to the space to the vector space spanned by the columns of A . So, I consider each column of A , they are linearly independent and they form a basis for the space. Suppose, a dimension of A is n by n , so they form a basis for the vector space of dimension n and R the residual is nothing but, $b - Ax$, it is we have taken the projection on to the space spanned by the columns of A .

We have taken the projection of b onto the space that is, spanned by the columns of A and whatever is left is the residual. Thus, the least squares procedure splits b into two parts Ax and r , with Ax by definition belonging to the space $R(A)$, why does Ax belong by definition to the space $R(A)$. Because, after all what is Ax , Ax is nothing but, a linear combination of the columns of A . If you just write it out, if you expand it, Ax is nothing but, the column 1 of A times x_1 plus column 2 of A times x_2 and so on and so forth.

So, Ax is nothing but, a linear combination of the columns of A and therefore, it must belong to the space $R(A)$, which is spanned by the columns of A . And then, we have r , which is orthogonal to the space, from this ((Refer Time: 44:06)) expression it is clear that, the least square solution satisfies, this satisfies this criterion $A^T(Ax - r) = 0$ is equal

to $A^T b$. Just expanding this out, $A^T b$ is equal to $A^T A x$, this is equal to 0, so $A^T b$ is equal to $A^T A x$, so that is what we have written out here.

And therefore, we can get the least square solution by actually solving this system, if it is at all solvable. If this system is solvable, so what will make this system solvable, if this matrix is invertible, if $A^T A$ is non singular, we can find the least square solution by solving this system.

(Refer Slide Time 45:05)

Method of least squares

- Thus the n components of x can be found from the solution of eqn (**), provided $A^T A$ ($n \times n$) is non singular
- It can be shown that $A^T A$ is non singular only if A has full rank
If A has full rank, $x \neq 0 \Rightarrow Ax \neq 0$
 $\therefore x \neq 0 \Rightarrow |Ax|^2 = (Ax)^T Ax = x^T (A^T A)x > 0$
- Hence $A^T A$ is positive definite and hence non-singular.
- Hence when A has full rank the least squares solution is unique and is given by: $x = A^+ b = (A^T A)^{-1} A^T b$

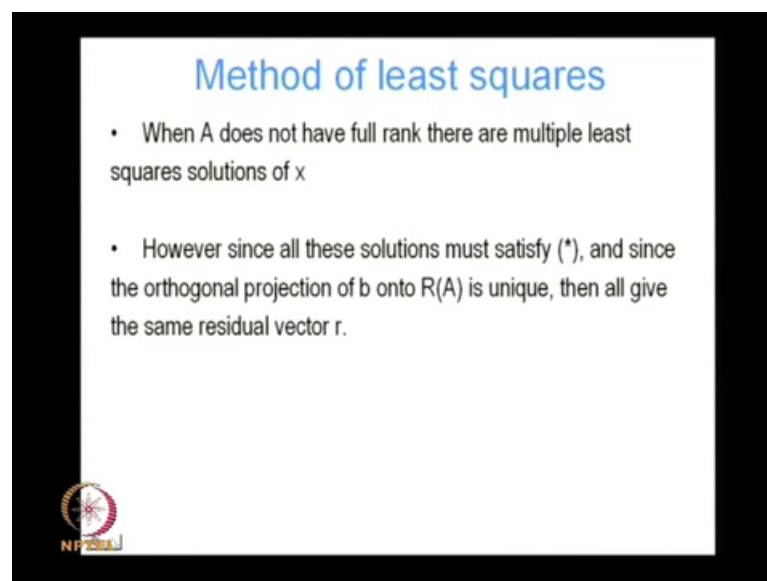
Then, the n components of x can be found from solving that system, $A^T A x = A^T b$, which we just looked at only $A^T A x = A^T b$. It can be shown that, $A^T A$ is non singular only if, A has full rank meaning that, if A is non singular then, it only then will $A^T A$ be non singular why is that. If A has full rank then, for any x not equal to 0 implies that, Ax cannot be equal to 0. Because, Ax is equal to 0 is going to give me x not equal to 0, only when A is singular.

So, if A has full rank, if A is non singular when x not equal to 0 implies, Ax is not equal to 0, hence x not equal to 0 implies, norm of Ax square. Since Ax is not equal to 0, norm of Ax square must be greater than 0 and norm of Ax square if I think of this as a vector,

it is nothing but, that vector transpose multiplied by itself. So, $x^T A x$, which I can write as $x^T A^T A x$ must be greater than 0. What does this mean, that means that, if A is non singular $A^T A$ is positive definite, because this is the criteria for positive definiteness.

What is the criteria for positive definiteness, any vector x^T times the product with the matrix times that vector product with that same vector is going to be greater than 0, so this means that, $A^T A$ is positive definite, hence non singular. Thus, when A has full rank, the least squares solution is unique and is given by x is equal to $A^T A^{-1} b$, just by inverting that matrix, just by inverting that matrix we can find $x = A^T A^{-1} b$.

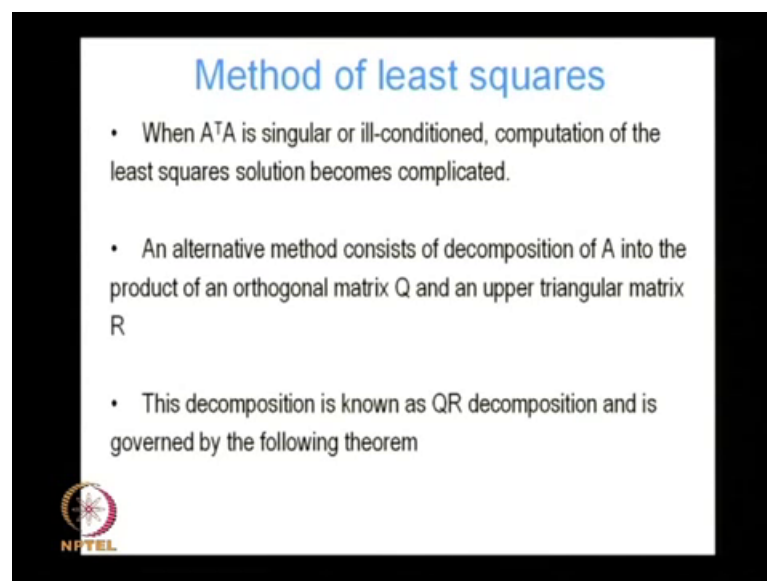
(Refer Slide Time 47:22)



However, when A does not have full rank, the multiple least squares solution of x when then, we cannot solve that equation, $A^T A$ is becoming singular. When $A^T A$ does not, A does not have full rank means, $A^T A$ is singular then, we will going to have non unique least square solutions. However, you must keep in mind, all of these solutions must satisfy the minimum residual criteria, and since, the orthogonal projection of b onto $R(A)$ is unique, they all going to give the same residual vector r .


What is the residual vector r , again it is b , after we take the projection of b onto the space that is spanned by A , whatever is left is the residual vector r . Since the vector b is known, the orthogonal projection of b onto $R A$ is also known, the residual vector is also going to be the same. So, respective of the fact that, we have non unique x , large number of least square solution x , which satisfies least square criteria, they are all going to get the same residual vector r .

(Refer Slide Time 48:50)



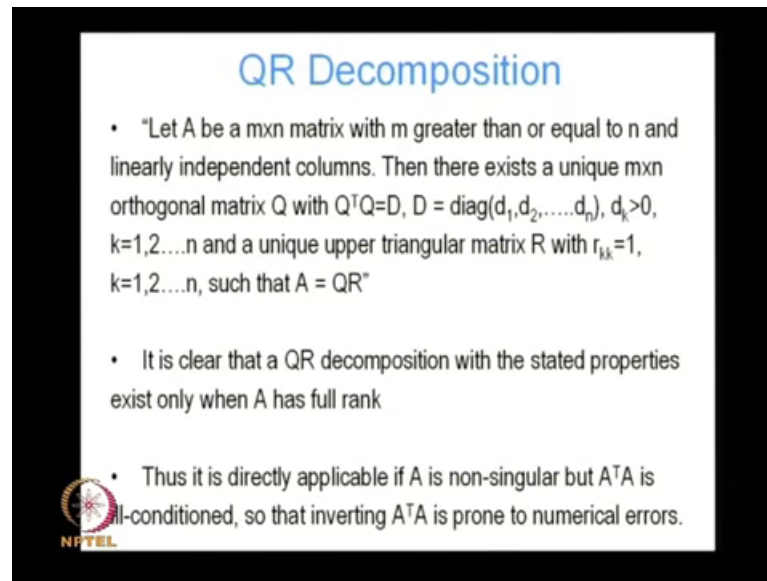
Method of least squares

- When $A^T A$ is singular or ill-conditioned, computation of the least squares solution becomes complicated.
- An alternative method consists of decomposition of A into the product of an orthogonal matrix Q and an upper triangular matrix R
- This decomposition is known as QR decomposition and is governed by the following theorem

 NPTEL

When $A^T A$ is singular or ill conditioned, computation of the least square solution becomes complicated, you can see that, becomes harder to invert this matrix. If $A^T A$ is ill conditioned or is near singular, it becomes hard to invert that matrix and the computation of the least squares solution becomes complicated. An alternative method consists of the decomposition of A into the product of an orthogonal matrix Q and upper triangular matrix R . This decomposition is known as QR decomposition, it is very important in numerical analysis and it is governed by the following theorem.

(Refer Slide Time 49:38)



The slide is titled "QR Decomposition" in blue text. It contains three bullet points. The first bullet point states: "Let A be a mxn matrix with m greater than or equal to n and linearly independent columns. Then there exists a unique mxn orthogonal matrix Q with $Q^T Q = D$, $D = \text{diag}(d_1, d_2, \dots, d_n)$, $d_k > 0$, $k=1, 2, \dots, n$ and a unique upper triangular matrix R with $r_{kk}=1$, $k=1, 2, \dots, n$, such that $A = QR$ ". The second bullet point states: "It is clear that a QR decomposition with the stated properties exist only when A has full rank". The third bullet point states: "Thus it is directly applicable if A is non-singular but $A^T A$ is ill-conditioned, so that inverting $A^T A$ is prone to numerical errors." In the bottom left corner of the slide, there is a small circular logo with a star and the text "NPTEL" below it.

Let A be a m by n matrix with m greater than or equal to n, recall we are talking about over determined systems currently, so m can be greater than or equal to n. And let us suppose for the timing, that A is non singular that, it is columns are linearly independent, so it has got full rank and it is non singular. Then, there exists a unique m by n orthogonal matrix Q, orthogonal matrix meaning, each columns is orthogonal with every other column, each column if I take a dot with every other column, it gives me 0.

So, each column is orthogonal, orthogonal matrix Q with Q transpose Q is equal to a diagonal matrix and such that, each term on the diagonal is greater than 0. So, what this theorem says that, if A is a non singular matrix or size m by n, we can always split it up into a product of an orthogonal matrix Q and an upper triangular matrix R. A unique upper triangular matrix R, which has got 1 on all it is diagonal terms and this decomposition is known as the QR decomposition.

QR decomposition exists with a state of properties only when A has got full rank, thus it is directly applicable if A is non singular, but A transpose A is ill conditioned so that, A transpose A is prone to numerical errors. So, if A transpose A here is ill conditioned, inverting that is a problem, so we cannot invert it directly. So, instead of when A transpose A is ill conditioned, we can do a QR decomposition. It turns out that, the


original QR theorem it assumes that, A is non singular, but it turns out that, even if A is singular, it is possible to modify the QR decomposition to solve for the least square solution, we are going to talk about that later on.

(Refer Slide Time 52:03)

QR Decomposition

- However a modified form of the decomposition can also be used to find the least squares solution in case A does not possess full rank and consequently $A^T A$ is singular.
- Considering first A with full rank, using QR decomposition, $A^T(b-Ax) = 0$ can be written as: $R^T Q^T (b - Ax) = 0$
- Since R is non-singular (its determinant is one) this condition becomes: $Q^T (b - Ax) = 0 \Rightarrow Q^T b = Q^T Ax = Q^T QRx = DRx$

Hence to find x we need to solve the system $x = R^{-1} D^{-1} Q^T b$



However, modified form of the decomposition can also be used to find the least squares solution in case A does not possess full rank and consequently, A transpose A is singular. So, neither of these two cases, whether A transpose A is ill conditioned, when solving that system A transpose A inverse operating on A transpose b is going to be hard, is going to be prone to numerical errors either in that case. Or in the situation, when A transpose A is straight away singular, either in those two cases I can use my QR decomposition to solve my problem.

So, we are going to continue with our discussion in QR decomposition next class and following that, we were going to talk about some iterative methods for solving the Eigen value problem. So, Eigen value problem in civil engineering and for that matter, any field of engineering, everyone has to encounter at some time at the Eigen value problem, because it is too important, it governs the physical behaviour of so many systems.

So, it is very important that, we know of easy to use efficient algorithms for obtaining the

Eigen values of matrices, particularly for symmetric positive definite matrices, which are commonly encountered while modelling physical systems.

Thank you.