

Elementary Numerical Analysis
Prof. Rekha P. Kulkarni
Department of Mathematics
Indian Institute of Technology, Bombay

Lecture No. # 29
Gauss-Seidel Method

We are considering iterative methods for solving system of linear equation, n equations in n unknowns, last time we have found a sufficient condition for convergence in the Jacobi method. Today, we are going to look at the Gauss-Seidel method and we will show that the sufficient condition, which we obtained in the Gauss-Seidel method; this will be satisfied, if the sufficient condition in the Jacobi method that is satisfied, and the condition which we get is going to be that coefficient matrix should be diagonally dominant by row.

In fact, if the matrix is diagonally dominant strictly by column, then also the result is going to be true, but I will not prove that result. So, we will first consider the Gauss-Seidel method; obtain sufficient conditions or convergence of it. Then compare this sufficient condition with the sufficient condition which we obtained yesterday for convergence of Jacobi iterates then we will look at a specific example.


If the coefficient matrix a , happens to be upper triangular then we will show that for a system of size n , when you will do n iterates to obtain an exact solution. If the system is upper triangular then we do not need to use these iterative methods, because you can do the back substitution, but this is just to illustrate that; if you apply your Jacobi iterates or Gauss-Seidel iterates, then infinite number of iterates you are going to obtain the exact solution.

Then we will consider a specific example, and if we will see, we may solve some more problems, which are using or which are based on the Newton's method, secant method or fix iteration for solution of non-linear equation. So, this is going to be plan of today's lecture. So, now let us look at the Gauss-Seidel iterate.

(Refer Slide Time: 03:03)

Gauss - Seidel Method

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



We are considering n equations in n unknowns. So, this is the relation satisfied by the exact solution, b_i they are known; the coefficient matrix a is known, hence you know a_{ij} . So, you have got x_i is equal to b_i minus summation over j , j not equal to i , $a_{ij} x_j$ divided by a_{ii} . Our assumption is that, coefficient matrix is invertible and all the diagonal entries they are non-zero. So, that is why we can divide by it.

This sum we split as, summation j goes from 1 to i minus 1 and summation j going from i plus 1 to n , with the understanding that if i is equal to 1, this term will be absent; if j is equal to n or if rather i is equal to n , then this term is going to be absent. So, this is the equation satisfied by the exact solution. Now we are going to define the iterative method, we had done it yesterday already, but let me recall.


(Refer Slide Time: 04:17)

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

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


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

$e_i^{(k)}$



So, this is the Gauss-Seidel iterates $x_i^{(k)}$, the k th iterate will be equal to b_i minus summation j goes from 1 to $i-1$ $a_{ij} x_j^{(k)}$. So, we are going to proceed systematically. So, you would have already calculated $x_1^{(k)}$, $x_2^{(k)}$, $x_{i-1}^{(k)}$; these are the recent values used here, minus summation j is equal to $i+1$ to n $a_{ij} x_j^{(k-1)}$ minus 1, we have come up to x_i . So, x_{i+1} up to x_n you have to use the values from the $k-1$ th iterate, divided by a_{ii} . Take the subtraction, you are going to have $x_i - x_i^{(k)}$ is equal to b_i will get cancelled, minus summation j goes from 1 to $i-1$ a_{ij} by $a_{ii} x_j - x_j^{(k)}$ minus summation j is equal to $i+1$ to n a_{ij} by $a_{ii} x_j - x_j^{(k-1)}$ minus 1. So, this is our $e_i^{(k)}$, this will be $e_j^{(k)}$, this will be $e_j^{(k-1)}$. So, now take the modulus of both the sides and you are going to get modulus of $e_i^{(k)}$ to be less than or equal to α_i .

(Refer Slide Time: 05:35)

$$e_i^{(k)} = - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{(k)} - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} e_j^{(k-1)}$$
$$\text{Let } \alpha_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right|, \quad \beta_i = \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right|,$$
$$\alpha_1 = 0, \quad \beta_n = 0$$
$$|e_i^{(k)}| \leq \alpha_i \|e^{(k)}\|_{\infty} + \beta_i \|e^{(k-1)}\|_{\infty}, \quad i=1, \dots, n$$


So, α_i is, summation j goes from 1 to $i-1$ modulus of a_{ij} by a_{ii} ; β_i is this summation; α_1 is defined to be 0; β_n to be defined to be 0; modulus of $e_j^{(k)}$ will be less than or equal to $\|e^{(k)}\|_{\infty}$, and modulus of $e_j^{(k-1)}$, it will be dominated by $\|e^{(k-1)}\|_{\infty}$. So, this is the relation we obtain for error, in case of the Gauss-Seidel iterates. So, the things are little more complicated than the Jacobi method.

Now, what we have got is modulus of $e_i^{(k)}$; that means, it is the error in the i th component of the k th iterate. On the right hand side, you have got terms which involve $\|e^{(k)}\|_{\infty}$ and $\|e^{(k-1)}\|_{\infty}$. Our aim is to find, $\|e^{(k)}\|_{\infty}$ to be less than or equal to some constant times $\|e^{(k-1)}\|_{\infty}$; that means, the maximum error in the k th iterate should be less than or equal to constant times maximum in the $k-1$ th iterate.

If we can obtain such a relation, then one uses this successively and one gets the maximum error in the k th iterate to be less than or equal to constant; it is raised to k and then multiplied by the error at 0th stage. So, suppose I call that constant to be η . So, I want to get $\|e^{(k)}\|_{\infty}$ to be less than or equal to $\eta^k \|e^{(0)}\|_{\infty}$. $e^{(0)}$ is the starting error. So, that will give us the condition that, if η is less than 1, then you are going to have convergence. The error will tend to 0.

So, now we have got the relation we obtained, we had modulus of $e_i^{(k)}$; that means, k th iterate. On the right hand side, you have got norm $e^{(k)}$ infinity, norm $e^{(k-1)}$ infinity. So, you have got such a relation for i goes from 1,2 up to n . Now, norm $e^{(k)}$ infinity will be achieved for sum value of i . Suppose that value is i is equal to m , then you use that particular equation to obtain the desired estimate.

So, let us look at it. You have got modulus of $e_i^{(k)}$ to be less than or equal to this. Now suppose, the norm $e^{(k)}$ infinity that is maximum of modulus of $e_i^{(k)}$, when i varies between 1 to n . Suppose this is attained, when i is equal to m . So, you have got modulus of $e_m^{(k)}$. So, look at this estimate, when i is equal to m . So, norm $e^{(k)}$ infinity is equal to modulus of $e_m^{(k)}$, which is less than or equal to; now I am looking at i is equal to m .

(Refer Slide Time: 08:47)

$$|e_i^{(k)}| \leq \alpha_i \|e^{(k)}\|_\infty + \beta_i \|e^{(k-1)}\|_\infty, \quad i = 1, \dots, n$$

$$\|e^{(k)}\|_\infty = |e_m^{(k)}|$$

$$\Rightarrow \|e^{(k)}\|_\infty = |e_m^{(k)}| \leq \alpha_m \|e^{(k)}\|_\infty + \beta_m \|e^{(k-1)}\|_\infty$$

$$\Rightarrow \|e^{(k)}\|_\infty \leq \frac{\beta_m}{1 - \alpha_m} \|e^{(k-1)}\|_\infty$$

So, it is α_m times norm $e^{(k)}$ infinity plus β_m times norm $e^{(k-1)}$ infinity. So, take this term on the left hand side and you will get, norm $e^{(k)}$ infinity to be less than or equal to β_m upon $1 - \alpha_m$ into norm of $e^{(k-1)}$ infinity. So, we will be assuming that, α_m is less than 1. So, that $1 - \alpha_m$ will be bigger than 0. It will preserve the inequalities and you are dividing. So, we have to make sure that, $1 - \alpha_m$ is not equal to 0.

So, we have obtained an estimate, it is norm $e^{(k)}$ infinity to be less than or equal to β_m upon $1 - \alpha_m$ norm $e^{(k-1)}$ infinity. What was m ? m was the index when the norm $e^{(k)}$ infinity is equal to modulus of $e_m^{(k)}$. So, $e_m^{(k)}$ is the error in the m th

component. In the error you have got your modulus of $e^{(k)}$ is going to be modulus of $x^{(k)}$; x is the exact solution. So, $x^{(k)} - x^{(k-1)}$; $x^{(k)}$ is something which we compute, but what about $x^{(k-1)}$? If we knew the exact solution, we do not have to do all these things. So that means, for the analysis, it is ok.

We can say that, the norm of $e^{(k)}$ mean infinity that is, when i is equal to m . But as such this bound, in that bound β_m upon $1 - \alpha_m$, we do not know what is m . So, the best we can do is, look at maximum of β_i upon $1 - \alpha_i$; $1 \leq i \leq n$. So, that is going to be our constant η . So, what I want is, $\|e^{(k)}\|_\infty \leq \eta \|e^{(k-1)}\|_\infty$ and for that constant, then we put a condition that, constant should be less than 1 which will give us convergence.

(Refer Slide Time: 12:13)

$$\|e^{(k)}\|_\infty \leq \frac{\beta_m}{1 - \alpha_m} \|e^{(k-1)}\|_\infty \quad \text{for some } m$$

$$\alpha_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right|, \quad \beta_i = \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right|,$$

$$\alpha_1 = \beta_n = 0.$$

$$\eta = \max_{1 \leq i \leq n} \frac{\beta_i}{1 - \alpha_i} < 1.$$

$$\|e^{(k)}\|_\infty \leq \eta \|e^{(k-1)}\|_\infty \leq \dots \leq \eta^k \|e^{(0)}\|_\infty$$

So, we have $\|e^{(k)}\|_\infty \leq \eta \|e^{(k-1)}\|_\infty$. Look at η to be maximum of β_i upon $1 - \alpha_i$; $1 \leq i \leq n$. I do not know what m is, but this m if I take η to be maximum of this quotients, then β_m upon $1 - \alpha_m$ is going to be less than or equal to η . So, we have $\|e^{(k)}\|_\infty \leq \eta \|e^{(k-1)}\|_\infty$ that means, $\|e^{(k)}\|_\infty \leq \eta \|e^{(k-1)}\|_\infty$ will be less than or equal to $\eta \|e^{(k-2)}\|_\infty$.

So, continuing this argument, you get norm ∞ to be less than or equal to η raise to k norm of ∞ . So, if this η is less than 1, then that will guarantee that, norm ∞ will tend to 0 as k tends to infinity. So, this is the sufficient condition for convergence, in the case of Gauss-Seidel iterates. Now, next we want to compare this sufficient condition with the sufficient condition which we had obtained in case of Jacobi iterate.

(Refer Slide Time: 13:44)

Jacobi Method : $\mu = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}|}{|a_{ii}|}$

Gauss-Seidel Method

$$\eta = \max_{1 \leq i \leq n} \frac{\beta_i}{1 - \alpha_i}, \quad \beta_i = \sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|},$$

$$\alpha_i = \sum_{j=i+1}^n \frac{|a_{ij}|}{|a_{ii}|}.$$

Claim: $\mu < 1 \Rightarrow \eta < 1$

$$\mu = \max_i (\alpha_i + \beta_i)$$

So, for the Jacobi iterates, the condition was μ which is maximum of summation j goes from 1 to n modulus of a_{ij} by a_{ii} , maximum taken over i less than or equal to n . This μ should be less than 1. For the Gauss-Seidel method, we have got η should be less than 1, where η is defined in terms of β_i and α_i . Our claim is that, μ less than 1 will imply that η is less than 1.

In fact, we will show that, η is going to be less than or equal to μ . So, I want you to observe that, in β_i , you are taking summation j goes from 1 to i minus 1 modulus of a_{ij} by a_{ii} ; in α_i , summation j is equal to i plus 1 to n modulus of a_{ij} by a_{ii} . If you look at α_i plus β_i , it will be nothing but, summation j goes from 1 to n j not equal to i modulus of a_{ij} by a_{ii} .

So, μ is going to be equal to maximum of α_i plus β_i . So, η is this maximum; μ is this maximum. So, what we will show is, α_i plus β_i minus β_i upon 1 minus α_i is bigger than or equal to 0. Or we want to show that, β_i upon 1 minus

alpha i is less than or equal to alpha i plus beta i and the proof is going to be straight forward. So, here mu is maximum of alpha i plus beta i 1 less than or equal to i less than or equal to n; eta is maximum of beta i upon 1 minus alpha i.


(Refer Slide Time: 15:34)

$$\mu = \max_{1 \leq i \leq n} (\alpha_i + \beta_i), \quad \eta = \max_{1 \leq i \leq n} \frac{\beta_i}{1 - \alpha_i},$$

$$\alpha_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right|, \quad \beta_i = \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right|.$$

Claim: $\mu < 1 \Rightarrow \eta \leq \mu < 1$

Consider $\alpha_i + \beta_i - \frac{\beta_i}{1 - \alpha_i} = \frac{\alpha_i - \alpha_i^2 - \beta_i \alpha_i}{1 - \alpha_i}$

$$= \frac{\alpha_i (1 - (\alpha_i + \beta_i))}{1 - \alpha_i} \geq \frac{\alpha_i}{1 - \alpha_i} (1 - \mu) > 0$$


I consider alpha i plus beta i minus beta i upon 1 minus alpha i ; This will be equal to alpha i minus alpha i square, this beta i and this beta i will get cancelled. So, you will have minus beta i alpha i divided by 1 minus alpha i. So, this is nothing but, alpha i times 1 minus alpha i plus beta i upon 1 minus alpha i ; then mu is maximum of alpha i plus beta i.


And hence, 1 minus alpha i plus beta i will be bigger than or equal to 1 minus mu. So, it is alpha i upon 1 minus alpha i , then mu is less than 1. So, 1 minus mu will be bigger than 0; alpha i is summation j goes from 1 to i minus 1 modulus of a i j by a i i ; mu less than 1 will also imply that, alpha i is also less than 1. So, this number is going to be bigger than 0.

(Refer Slide Time: 17:26)

$$\mu = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}|}{|a_{ii}|} < 1$$

\Rightarrow For $i = 1, \dots, n$, $\sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| < |a_{ii}|$, i.e.,

A is strictly row-diagonally dominant



So, the maximum which you are taking here, this number is bigger than this number. So, that proves that, your η is less than or equal to μ less than 1. So thus, if μ is equal to maximum of this summation less than 1, then we are going to have convergence both in the Jacobi method and Gauss-Seidel method that means, your A should be strictly row diagonally dominant. So, now we have got a sufficient condition which guarantees convergence of both Jacobi iterates and Gauss-Seidel iterates.

As I had mentioned before, we are not claiming that, if the Jacobi iterates converge, then Gauss-Seidel iterates have to converge in the Gauss-Seidel iterates, because you are using the most recent value of your iterate; it is likely to give better results, but when can construct a pathological example, when the Jacobi methods converges, Gauss-Seidel method does not converge.

So, we have compared some sufficient conditions and then we said that, if the sufficient condition in the Jacobi method is satisfied, then the sufficient condition in the case of Gauss-Seidel method also will be satisfied. So, if your A is strictly row diagonally dominant, then both the Jacobi method and the Gauss-Seidel method or these iterates are going to converge to the exact solution. So, let us look at a simple example. We will take 4 by 4 system and we will look at the first few iterates in the Jacobi method and in the Gauss-Seidel method.

(Refer Slide Time: 19:42)

The image shows handwritten mathematical notes on a slide. At the top, a linear system is written as a matrix equation:
$$\begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 3 \end{bmatrix}$$
 To the right of this equation, the text "exact solution" is written in red, followed by a column vector $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$. Below the matrix equation, the text "Jacobi Iterates: $x^{(0)} = \overline{0}$ " is written. Underneath, the formula $x_i^{(1)} = \frac{b_i}{4}, i = 1, 2, 3, 4$ is shown. Finally, the second iteration formulas are given as $x_1^{(2)} = \frac{3 + x_2^{(1)}}{4}$ and $x_2^{(2)} = \frac{2 + x_1^{(1)} + x_3^{(1)}}{4}, \dots$. In the bottom left corner of the slide, there is a logo for NIPTEIL.

So, here is the system, you have got 4 equations in 4 unknowns. The diagonal entries are 4, super diagonal and sub diagonal are minus 1. So, it is a tri-diagonal system and it is diagonally dominant. The right hand side is so chosen that, the exact solution is nothing but 1 1 1 1. We are going to choose the initial iterates to be all 0. So, x_0 is going to be a 0 vector.

In the case of Jacobi iterates, you are going to have $x_i^{(1)}$ is equal to b_i divided by 4, i is equal to 1 2 3 4. We are taking all these 0s. So, the $x_i^{(1)}$ s first iterates will be given by $b_i / 4$. When you consider the second iterate in the Jacobi. So, $x_1^{(2)}$. Look at the first equation. Here $x_1^{(2)}$ will be $3 + 1 \times x_2^{(1)}$ divided by 4 then, when you consider $x_2^{(2)}$, this is going to be equal to $2 + x_1^{(1)} + x_3^{(1)}$ and then divided by diagonal entry, that is 4.

So, when we calculate $x_2^{(2)}$, we have already calculated $x_1^{(2)}$. Here if instead of $x_1^{(1)}$, we use $x_1^{(2)}$, the recent value available then that becomes the Gauss-Seidel method. Here what we are going to do is, then you look at $x_3^{(2)}$ that is going to be equal to $2 + x_2^{(1)} + x_4^{(1)}$ divided by 4. Eventhough $x_2^{(2)}$ is available, we are not using it here; that is for the Jacobi iterates.

(Refer Slide Time: 21:51)


$$\begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 3 \end{bmatrix}$$

exact solution $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$.

Jacobi Iterates: $x^{(0)} = 0$

$$x_i^{(1)} = \frac{b_i}{4}, \quad i = 1, 2, 3, 4$$

Gauss-Seidel:

$$x_1^{(1)} = \frac{b_1}{4}, \quad x_2^{(1)} = \frac{b_2 + x_1^{(1)} + x_2^{(0)}}{4}, \quad x_3^{(1)} = \frac{b_3 + x_2^{(1)} + x_4^{(0)}}{4}$$


In the case of Gauss-Seidel iterates, what we do is when you look at, you have the first iterate for the Jacobi; you had $x_i = b_i / 4$, $i = 1, 2, 3, 4$. In the case of Gauss-Seidel, the first component x_1 is going to be $b_1 / 4$; when you calculate x_2 , this x_2 will be equal to $b_2 + x_1 + x_2^{(0)}$ divided by 4. Now, whatever is the recent value of x_1 that we are using for Jacobi, what we did was it was all 0s. So, our x_2 was $b_2 / 4$ whereas, here x_2 will be $b_2 + x_1 / 4$ because, $x_2^{(0)}$ will be 0. When you consider x_3 , for x_3 again x_2 is available; whereas, x_4 we have not yet reached. So, we are forced to use $x_4 = 0$. So, this is the difference between Gauss-Seidel and Jacobi iterates.

(Refer Slide Time: 23:08)


The slide contains the following handwritten content:

$$\begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 3 \end{bmatrix}$$

exact solution $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$.

Jacobi Iterates: $x^{(0)} = \overline{0}$

$$x_1^{(1)} = \frac{3}{4}, x_2^{(1)} = \frac{2}{4}, x_3^{(1)} = \frac{2}{4}, x_4^{(1)} = \frac{3}{4}$$
$$x_1^{(2)} = \frac{3 + \frac{1}{4}}{4} = \frac{7}{8}, x_2^{(2)} = \frac{2 + \frac{3}{4} + \frac{2}{4}}{4} = \frac{13}{16}$$
$$= x_4^{(2)} \qquad = x_3^{(2)}$$



So, let us calculate the iterates. So, the first for the Jacobi iterate. You have got x_0 is equal to 0 vector. So, x_1 will be 3 by 4; x_2 will be 2 by 4 that means, 1 by 2; x_3 will be 2 by 4 and x_4 is equal to 3 by 4. So, you are replacing sort of coefficient matrix by the diagonal matrix. This was for the first iterate.

For the second iterate, x_1 is going to be equal to 3 plus, what we had calculated for x_2 or other it is going to be x_1 is equal to b_1 plus x_2 divided by 4. So, x_1 is 1 by 2. So, that gives you x_1 to be equal to 7 by 8. When you look at second iterate, the second component that is going to be equal to the right hand side 2, then the value of x_1 ; that is 3 by 4 then plus the value of x_3 . So, that is going to be 2 by 4 divided by 4. So, you get it to be 13 by 16 you can check that, you get the same value for x_3 and you get the value x_4 to be 7 by 8.

(Refer Slide Time: 24:56)

Jacobi: $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 3/4 \\ 1/2 \\ 3/4 \end{bmatrix} \rightarrow \begin{bmatrix} 7/8 \\ 13/16 \\ 7/8 \end{bmatrix}$

$x^{(0)}$ $x^{(1)}$ $x^{(2)}$

$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$: exact

NIPTEIL

So, for this particular example what you are getting is, for the Jacobi method, we start with the iterate to be 0 0 0 0, that is our x_0 . Then our next iterate x_1 is going to be 3 by 4, then 1 by 2, 1 by 2 and again 3 by 4. So, this is our x_1 and the next iterate which we get is 7 by 8, then 13 by 16, 13 by 16 and again 7 by 8. So, this is our x_2 . The exact is 1 1 1 1 and then one can continue.

(Refer Slide Time: 26:23)

$$\begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 3 \end{bmatrix}$$

exact solution $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

Gauss-Seidel Iterates: $x^{(0)} = 0$

$$x_1^{(1)} = \frac{3}{4}, \quad x_2^{(1)} = \frac{2 + x_1^{(1)} + x_2^{(0)}}{4} = \frac{11}{16}$$

$$x_3^{(1)} = \frac{2 + x_2^{(1)} + x_4^{(0)}}{4} = \frac{2 + \frac{11}{16}}{4} = \frac{43}{64}$$

NIPTEIL

Now let us look at the Gauss-Seidel iterates. For the Gauss-seidel, x_1 is 3 by 4 whereas, when you calculate x_2 , you are going to use the value to be x_1 , the recent

value; whereas, for $x \geq 0$, you have to choose it to be you had to take it to be 0 only. So, you get 11 by 16. In case of $x \leq 1$, again $x \geq 1$ is we have calculated. So, you will use that value; whereas, $x \leq 0$ you have to take it to be 0 and then you get these values.

(Refer Slide Time: 27:06)

Jacobi: $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{x^{(0)}} \rightarrow \begin{bmatrix} 3/4 \\ 1/2 \\ 3/4 \end{bmatrix}_{x^{(1)}} \rightarrow \begin{bmatrix} 7/8 \\ 13/16 \\ 7/8 \end{bmatrix}_{x^{(2)}}$

$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$: exact

Gauss-Seidel: $\begin{bmatrix} 3/4 \\ 11/16 \\ 43/64 \\ * \end{bmatrix} \leftarrow x^{(1)}$

So, in the case of Gauss-Seidel method, your x^1 vector is going to be 3 by 4, then 11 by 16, 43 by 64 and then something you will get here. So, this is x^1 . So, here it looks like that, the Gauss-Seidel iterates are going to converge faster. The exact solution was 1 1 1 1. So, if you use the recent value, then the result, the error will be small. So, this was just an illustrative example and as I have been mentioning that, whatever we consider in this course, it is not for hand computations.

It all makes sense, when you are going to use big systems using a computer. So, here one can when has to write a program, it will not be difficult to write a program for Jacobi method or Gauss-Seidel iterates. You have to give some stopping criteria and then you can see that, it will give you an approximate solution in contrast to the direct solution, but if n is big, then one may have to resort to indirect methods because, the direct method may be too expensive.

So, now I want to quickly tell you or show you that, if the coefficient matrix is upper triangular matrix, then in n iterates you are going to reach the exact solution. So, now we look at a x is equal to b , Jacobi method and A is assumed to be an upper triangular matrix.

So, in case of Jacobi method, this is our formula that, $x_i^{(k)}$ is equal to b_i minus sum j not equal to i divided by a_{ii} , i goes from 1, 2 up to n .


(Refer Slide Time: 29:28)

Jacobi Method

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

A : upper triangular, $a_{ij} = 0$ if $i > j$.

$$x_n^{(1)} = \frac{b_n}{a_{nn}} : \text{exact value}$$

 NIPTRIL

A is upper triangular that means, a_{ij} will be 0, if i is bigger than j . If we take the initial approximation to be all 0s or infact , whatever approximation you take. So, when you have got A to be upper triangular, then your first equation is going to be $a_{11}x_1 + a_{1n}x_n = b_1$. Your second equation will be $a_{22}x_2 + a_{2n}x_n = b_2$ and your last equation will be $a_{nn}x_n = b_n$. So, in the last equation, there are no x_1, x_2, \dots, x_{n-1} . So, whatever are your initial iterate or initial approximation, it just does not come into picture.

(Refer Slide Time: 30:09)

A: upper triangular

$$\begin{aligned} a_{11}x_1 + \dots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{nn}x_n &= b_n \end{aligned}$$

The image shows a whiteboard with handwritten equations. A hand is visible at the bottom right, holding a black marker. In the bottom left corner, there is a circular logo with a star and the text 'NPTEL'.

So, in the case of Jacobi method, put i is equal to n then it tells you that, x_{n+1} is going to be b_n upon a_{nn} . So, whatever is your initial approximation for the n th component, you have already obtained an exact value. So, we are starting with some approximation x_{10}, x_{20}, x_{n0} . We calculate the Jacobi iterate, you are going to get these iterates x_{11}, x_{21}, x_{n1} ; of which, the last component x_{n1} is going to be the exact value.

(Refer Slide Time: 31:09)

Jacobi Method

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

A: upper triangular, $a_{ij} = 0$ if $i > j$.

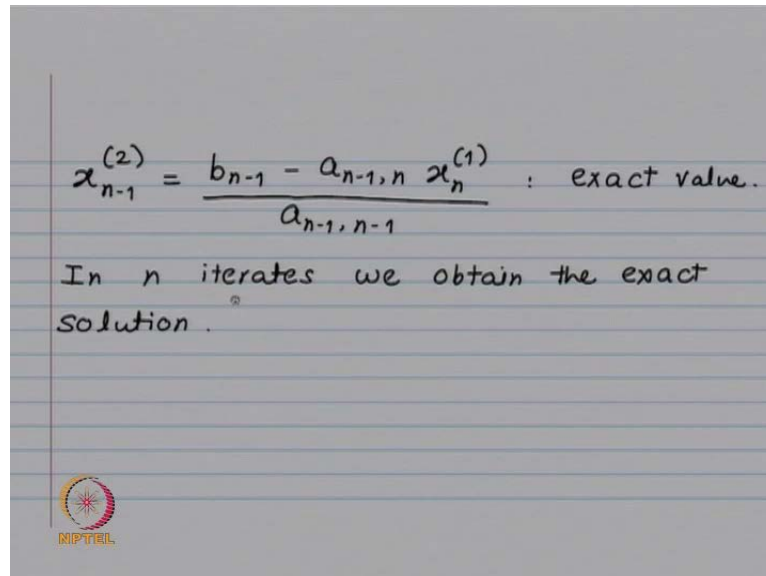
$$x_n^{(1)} = \frac{b_n}{a_{nn}} : \text{exact value}$$

The image shows handwritten notes on lined paper. At the bottom left, there is a circular logo with a star and the text 'NPTEL'.

If your matrix is upper triangular, now when you go to the next iterate, then what will happen will be the n th component is already exact; then n minus first component will


become exact. So, like that when you will do n iterates, in the end the nth iterate is going to be equal to the exact solution.

(Refer Slide Time: 32:28)



$$x_{n-1}^{(2)} = \frac{b_{n-1} - a_{n-1,n} x_n^{(1)}}{a_{n-1,n-1}} : \text{exact value.}$$

In n iterates we obtain the exact solution.




So, we have $x_{n-1}^{(2)}$ is equal to $b_{n-1} - a_{n-1,n} x_n^{(1)}$ divided by this n th component is already exact. So, now in the second iterate $n-1$ th component also is exact. When you consider n iterate, then we obtain the exact solution. This is about the iterative solution of system of linear equations. So, what we have considered is, we looked at solution of non-linear equations and now iterative solution of system of linear equations.

Our next topic is going to be approximate solution of differential equations, but before we start that topic what I want to do is, solve some problems which are for the Newton's method, **secant** method and bisection method based on truths. We are going to look at some of the problems based on those and there are some results about the norms. So, we are also going to look at those problems. In the lectures we have seen that, we had left something to be done in the tutorial. So, we are going to look at truth problems also.

(Refer Slide Time: 34:12)

Q. Let $g : [a, b] \rightarrow [a, b]$ be continuously differentiable and $M = \max_{x \in [a, b]} |g'(x)| < 1$.
Let c be the unique fixed point of g in $[a, b] : g(c) = c$. Let $x_0 \in [a, b]$ and $x_{n+1} = g(x_n), n = 0, 1, 2, \dots$
Show that

$$|x_{n+1} - c| \leq \frac{M}{1-M} |x_{n+1} - x_n|$$



Now, first let us look at the first problem. So, the problem is, g is from a to b . It is a continuously differentiable function and M is maximum of modulus of g dash x belonging to a to b ; this is less than 1. Under this condition, g is going to have a unique fixed point in the interval a to b . So, I call that fixed point to be c . We look at the Picard iterates. So, it starts with x_0 in a to b and define x_{n+1} is equal to g of x_n , then we want to show that modulus of x_{n+1} minus c is less than or equal to m upon $1 - m$ modulus of x_{n+1} minus x_n .

What is the importance of this example or this problem? When we will prove the convergence of the Picard's iteration method? We had modulus of x_{n+1} minus c to be less than or equal to m times modulus of x_n minus c . Continuing the argument, we got modulus of x_{n+1} minus c to be less than or equal to m raise to $n+1$ modulus of x_0 minus c , since m is less than 1; m raise to $n+1$ tends to 0. So, that proves x_n converging to c .

So, now I know that, if some conditions are satisfied then this Picard's iterates are going to converge. Now in practice, I want to know where to stop that. If I require some accuracy, say the error should be less than 10^{-6} , then how many iterates I should calculate. I cannot calculate modulus of x_n minus c . c is the fixed point, which we do not know because we are trying to find the approximation.

Unless I am considering an illustrative example, the modulus of c minus x_n I cannot calculate, but what I can compare is, the distance between the 2 successive iterates. So, modulus of x_{n+1} minus x_n , now suppose, the 2 successive iterates modulus of x_{n+1} minus x_n , if they are small whether that will mean that, modulus of x_{n+1} minus c also is small. So, the answer is yes and that is where, this example shows or this is the result which we are going to show. That modulus of x_{n+1} minus c is going to be less than or equal to M upon $1 - M$ modulus of x_n plus 1 minus x_n .

So, M is less than 1. Suppose M is equal to half, then this is going to be equal to half. We will have M upon $1 - M$ to be equal to 1, then if modulus of x_{n+1} minus x_n is less than 10^{-6} ; that also guarantees that, x_{n+1} minus c will be less than 10^{-6} . This is something one can compute; this gives you the stopping criteria and proof of this result is not difficult. So, let us prove this result.

(Refer Slide Time: 38:20)

Solution: $g \in C^1[a, b]$, $g(c) = c$, $|g'(x)| \leq M < 1$


$$x_{n+1} - c = g(x_n) - g(c)$$

$$= g'(d_n)(x_n - c)$$

$$\Rightarrow |x_{n+1} - c| \leq M |x_n - c|$$

$$\leq M |x_n - x_{n+1}| + M |x_{n+1} - c|$$

$$\Rightarrow |x_{n+1} - c| \leq \frac{M}{1 - M} |x_n - x_{n+1}|$$

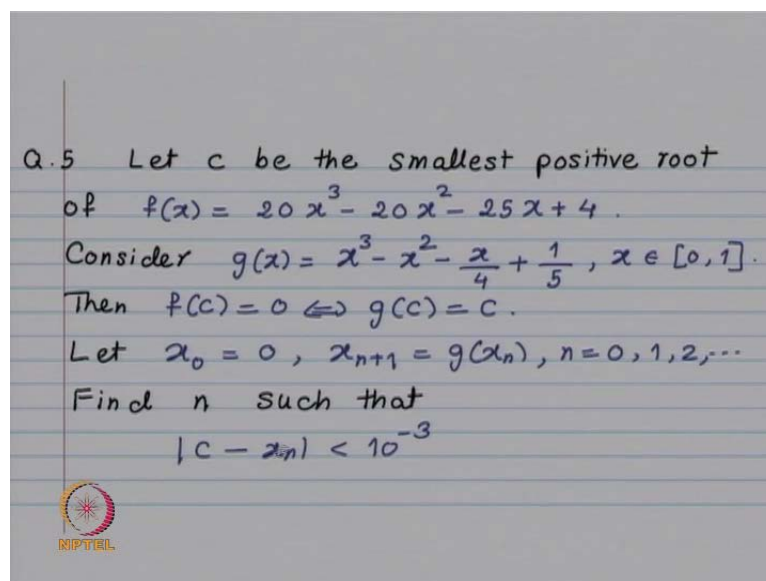


So, our assumption is, g is continuously differentiable on interval a, b ; c is the unique fixed point; g of c is equal to c ; modulus of g dash x is less than or equal to M less than 1, that is our assumption. Then x_{n+1} minus c will be equal to, x_{n+1} is g of x_n ; c is g of c ; g of x_n minus g of c . Using the mean value theorem, this will be equal to g dash d_n x_n minus c where, d_n lies between x_n and c ; modulus of g dash d_n will be less than or equal to M .

So, you will have modulus of $x_{n+1} - c$ to be less than or equal to M times modulus of $x_n - c$. Now, here I add and subtract x_{n+1} . So, I write $x_n - c$ as $x_n - x_{n+1} + x_{n+1} - c$, use the triangle inequality. So, it is less than or equal to M times modulus of $x_n - x_{n+1}$ plus M times modulus of $x_{n+1} - c$.

Take this term on the other side and then you will get, modulus of $x_{n+1} - c$ to be less than or equal to M divided by $1 - M$ modulus of $x_n - x_{n+1}$. So, this is the error in the successive iterates, this is going to be the error in the n plus first iterate and we have related these 2 errors. Now, the next problem which we are going to do is, we will try to find that, how big n I should choose in a particular situation. So, that is going to be our next problem.

(Refer Slide Time: 40:26)



So, we look at c to be the smallest positive root of $f(x)$, this calculating the 0 of a function f ; we relate it to calculating a fixed point of another function. So, if I define $g(x)$ is equal to $x^3 - x^2 - x/4 + 1/5$, then $f(c) = 0$, if and only if, $g(c) = c$; we are going to show this. Take x_0 to be equal to 0 and x_{n+1} is equal to $g(x_n)$. So, these are our Picard's iterate. We want to find n such that modulus of $c - x_n$ is going to be less than 10^{-3} .

We are first going to look at, how the 0s of f are going to be to. So, it is a cubic equation. So, it is going to have 3 roots; we look at the smallest positive root. So that, its

approximation we are considering, we will show that $f(c)$ is equal to 0, if and only if, $g(c)$ is equal to c and finally, find n which will imply that, the error is less than 10^{-3} .

(Refer Slide Time: 41:49)

Solution : $f(x) = 20x^3 - 20x^2 - 25x + 4$
 $= 20x^3 - 20x^2 - 5x + 4 - 20x$
 $g(x) = x^3 - x^2 - \frac{x}{4} + \frac{1}{5}$
 $= \frac{f(x)}{20} + x$
 $f(c) = 0 \Leftrightarrow g(c) = c$

So, this is our $f(x)$; this $25x$ I split as $5x$ and $20x$, then $g(x)$ is our function x^3 minus x^2 minus x by 4 plus 1 by 5 . So, $g(x)$ is nothing but, $f(x)$ divided by 20 and then plus x . This is our $g(x)$ when I divide by $f(x)$ by 20 . I will have x^3 then minus x^2 minus x by 4 and then plus 1 by 5 . So, that is our $g(x)$ and then I am dividing by 20 . So, this will be $g(x) = f(x)/20 + x$ or equivalently $g(x) = f(x)/20 + x$ that gives us $f(c) = 0$, if and only if, $g(c) = c$.

Now, we will want to look at how the 0 s of f are situated, f is a cubic polynomial. So, it is going to have 3 roots; it can happen that, all the 3 roots are real or there will be 1 real root and 1 pair of complex roots. Now if I can find points, like suppose I have got value of f at 2 points; if it is of opposite sign, then there is at least one 0 that can be more than one 0 s, but there is at least one 0 . So, that is why we will try to find 3 intervals, in which there are going to be f has 0 s.


(Refer Slide Time: 43:54)

$$f(x) = 20x^3 - 20x^2 - 25x + 4$$

$$f(-1) = -11, \quad f(0) = 4, \quad f(1) = -21, \quad f(2) = 34$$

f has roots in

$$(-1, 0), \quad (0, 1), \quad (1, 2)$$

$$c \in (0, 1)$$


So, we will look at value of f at, say minus 1, 0, 1 and 2. So, f of minus 1 you can check that it is minus 11, f of 0 is 4. So, f minus 1 and f of 0 they are of opposite sign; it has at least 1 root in the interval minus 1 to 0; f of 1 is going to be minus 21. So, in the interval 0 to 1, there is at least 1 root and f of 1 is minus 21, f of 2 is 34. So, f 1 and f 2 are opposite signs. So, there has to be at least 1 root in 1 to 2. Now f can have at most 3 roots. So, in each interval, there is going to be exactly 1 root and we are concentrating on c which belongs to 0 to 1, the smallest positive root of our function f .

(Refer Slide Time: 44:59)


$$g(x) = x^3 - x^2 - \frac{x}{4} + \frac{1}{5}, \quad x \in [0, 1]$$

$$g'(x) = 3x^2 - 2x - \frac{1}{4}$$

We want to find $M = \max_{x \in [0, 1]} |g'(x)|$

$$g''(x) = 6x - 2 = 0 \Rightarrow x = \frac{2}{3}$$

x	0	$\frac{2}{3}$	1
$g'(x)$	$-\frac{1}{4}$	$-\frac{1}{4}$	$\frac{3}{4}$

$$M = \frac{3}{4}$$


Now, $g(x)$ is our function given by this, you want to calculate. So, for a fixed point iterate, we need to calculate the maximum of modulus of $g'(x)$ over an appropriate interval

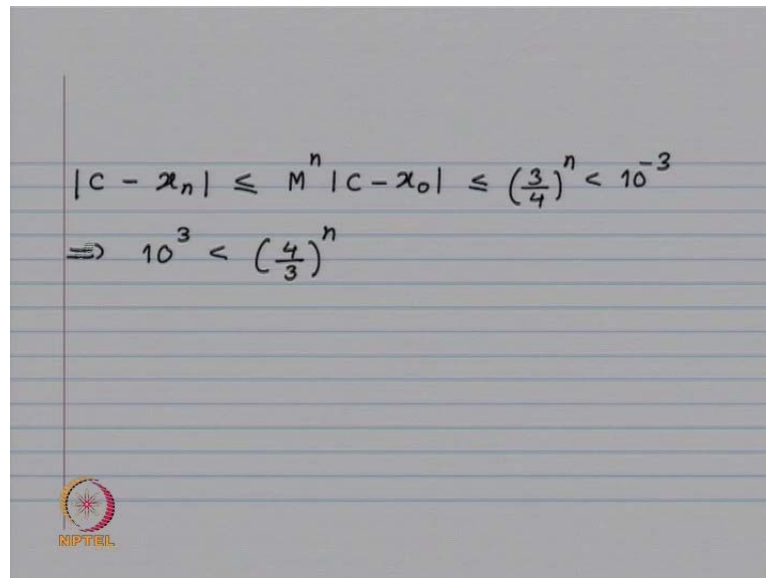
because, that is our sufficient condition; that modulus of $x^n - c$ is going to be less than or equal to M^n into modulus of $x_0 - c$. So, we need to calculate M , which is maximum of modulus of $g'(x)$ over interval, in our case it is going to be interval 0 to 1.

G is a polynomial function. So, we can calculate its derivative. When you want to find absolute value of a continuous function, you have to compare the values at the 2 end points and at the critical point. The critical points are those, points at which either derivative vanishes or derivative does not exist. In our case, g is a polynomial. So, we will have to only look at the points, where the derivative vanishes.

Now, hopefully one gets only finitely many such points. So, one compares the values at those points and decides which is the absolute maximum and which is the absolute minimum. One did not consider the second derivative, the second derivative test tells you only about the local maximum and local minimum. So, now we are going to look at the derivative, the value at the 2 end points and the value where the derivative vanishes; that means, the second derivative should be equal to 0.

So, $g(x)$ is given by $x^3 - x^2 - x + 1$; its derivative will be $3x^2 - 2x - 1$; the second derivative will be given by $6x - 2$. So, this will be 0 when x is equal to $\frac{1}{3}$. So, we need to compare the value at the 2 end points 0 and 1, and at the critical point $\frac{1}{3}$; these values are 1 , $\frac{1}{27}$, 1 which will give us M to be equal to 1 .

(Refer Slide Time: 47:50)


$$|c - x_n| \leq M^n |c - x_0| \leq \left(\frac{3}{4}\right)^n < 10^{-3}$$
$$\Rightarrow 10^3 < \left(\frac{4}{3}\right)^n$$

And then modulus of c minus x_n will be less than or equal to 3 by 4 raise to n and now one thing here we do not know c ; x_0 is 0 , but we know that, c is going to be in the interval 0 to 1 . So, M dominates modulus of c by 1 . So, this should be less than 10 raise to minus 3 . So, that gives you condition that, your n should be such that 10 cube is less than 4 by 3 raise to n . So, in our next lecture, we are going to consider some more problems and then we will start the new topic, approximate solution of differential equations. So, **thank you.**