

Introduction to Finite Volume Methods - I
Prof. Ashoke De
Department of Aerospace Engineering
Indian Institute of Technology, Kanpur

Lecture – 40
Error Analysis-III

(Refer Slide Time: 00:13)

Error analysis

- ▶ Let $\delta(\epsilon) = \underline{x}(\epsilon) - \underline{x}$. Then, $\delta \equiv \text{difference}$
 $(A + \epsilon E)\delta(\epsilon) = (b + \epsilon e) - (A + \epsilon E)x = \epsilon (e - Ex)$
 $\delta(\epsilon) = \epsilon (A + \epsilon E)^{-1}(e - Ex)$

$(A + \epsilon E)\delta = \epsilon(e - Ex)$
 $\delta = \epsilon(A + \epsilon E)^{-1}(e - Ex)$
- ▶ $\underline{x}(\epsilon)$ is differentiable at $\epsilon = 0$ and its derivative is
 $x'(0) = \lim_{\epsilon \rightarrow 0} \frac{\delta(\epsilon)}{\epsilon} = A^{-1}(e - Ex)$
- ▶ A small variation $[\epsilon E, \epsilon e]$ will cause the solution to vary by roughly $\epsilon x'(0) = \epsilon A^{-1}(e - Ex)$
- ▶ The relative variation is such that
 $\frac{\|x(\epsilon) - x\|}{\|x\|} \leq \epsilon \|A^{-1}\| \left(\frac{\|e\|}{\|x\|} + \|E\| \right) + O(\epsilon^2)$

$\frac{E}{2} = 1$
 Relative variation
- ▶ Since $\|b\| \leq \|A\| \|x\|$:
 $\frac{\|x(\epsilon) - x\|}{\|x\|} \leq \epsilon \|A\| \|A^{-1}\| \left(\frac{\|e\|}{\|b\|} + \frac{\|E\|}{\|A\|} \right) + O(\epsilon^2)$

INDIAN INSTITUTE OF TECHNOLOGY KANPUR Ashoke De 56

So, welcome to this particular lecture and we will continue our discussion what we have been doing so far. So, these are going to be important, when you actually use the iterative methods to find out the error. Because, the error needs to be I mean decreasing when you go ahead with the iterative process.

(Refer Slide Time: 00:40)

Error analysis

The quantity $\kappa(A) = \|A\| \|A^{-1}\|$ is called the **condition number** of the linear system with respect to the norm $\|\cdot\|$. When using the standard norms $\|\cdot\|_p$, $p = 1, \dots, \infty$, we label $\kappa(A)$ with the same label as the associated norm. Thus,

$$\kappa_p(A) = \|A\|_p \|A^{-1}\|_p \quad \leftarrow p \text{th level}$$

$$\kappa(A) = \|A\| \|A^{-1}\|$$

► **Note:** $\kappa_2(A) = \sigma_{\max}(A) / \sigma_{\min}(A) =$ ratio of largest to smallest singular values of A . Allows to define $\kappa_2(A)$ when A is not square.

define the condition number

► Determinant *is not* a good indication of sensitivity

► Small eigenvalues *do not* always give a good indication of poor conditioning.

Now, there is a quantity, which you have come across is called the condition number; it is kappa A, it is essentially represented at kappa A is the condition number of a matrix A is the magnitude or norm of A into A inverse. So, of a linear system with respect to norm this.

So, when you use the standard norms the norms, 1, 2, 3 to infinity, we say this kappa A with the same level as p condition number for the pth norm of A can be written as magnitude of A at the pth norm and inverse at the pth level. So, this p could, it is a pth level, so p can go from 1, 2 n like that. Now, one can note here, the kappa 2 is the ratio of the largest to a smallest singular values of A. So, it allows to define when A is not square. So, this is a very important note, when A is a not a square matrix, you can actually define the condition number. And as you see this is the ratio of the largest to the smallest singular values of A.

So, which tells you depending on the condition number, the sensitivity of this matrix for the linear solution because, that is what is getting associated with the condition number. So, from the condition number also one can anticipate the convergence level of your linear system. So, when you say the determinant is not good, a good indication of sensitivity. Small eigenvalues do not always give a good indication of poor conditioning. So, that is the point, once you find out the eigenvalues from there, if you try to anticipate whether, the conditioning is good or bad that is not an good indication, but the determinant can be a good indication for getting the condition number.

(Refer Slide Time: 03:14)

Error analysis

Example: Consider, for a large α , the $n \times n$ matrix

$$A = I + \alpha e_1 e_n^T$$

▶ Inverse of A is :

$$A^{-1} = I - \alpha e_1 e_n^T$$

▶ For the ∞ -norm we have

$$\|A\|_\infty = \|A^{-1}\|_\infty = 1 + |\alpha|$$

so that

large for large α → $\kappa_\infty(A) = (1 + |\alpha|)^2$.

▶ Can give a very large condition number for a large α – but all the eigenvalues of A_n are equal to one.

INDIAN INSTITUTE OF TECHNOLOGY KANPUR Ashoke De 58

One can again check an example, simple example and try to see what it means. You take an large alpha for n by n matrix such that, equals to I plus alpha into epsilon 1 and epsilon 1 n transpose. So, the inverse of A if you find out, it will remain I minus alpha e 1 e to the power transpose n. For the infinity norm we have A infinity norm equals to A inverse infinity norm equals to 1 plus mod alpha so, that is what you get. So, this leads to the calculation of the condition number at the K infinity level which is 1 plus mod alpha and completely square of that.

So, that means, it can provide you back a very large condition number for large value of alpha, but all the eigenvalues of A n are equal to 1. So, this is the check that, what we made a statement here the eigenvalues are not the good indicator for condition number, but once you look at here, the eigenvalues are nicely behaving, but the condition number is absolutely large for large alpha so, that is not a good marker for condition number.

(Refer Slide Time: 05:02)

Error analysis

Rigorous norm-based error bounds

► First need to show that $A + E$ is nonsingular if A is nonsingular and E is small. Begin with simple case:

LEMMA: If $\|E\| < 1$ then $I - E$ is nonsingular and

$$\|(I - E)^{-1}\| \leq \frac{1}{1 - \|E\|}$$

Proof is based on following steps

a) Show: If $\|E\| < 1$ then $I - E$ is nonsingular

b) Show: $(I - E)(I + E + E^2 + \dots + E^k) = I - E^{k+1}$.

c) From which we get:

$$(I - E)^{-1} = \sum_{i=0}^k E^i + (I - E)^{-1} E^{k+1} \rightarrow$$



So, now if you look at more detailed error bound, so one needs to see what happens to A plus E , when it is non singular, if A is non singular and E is small. So, we can take a simple case as in lemma, the magnitude or the mod of E less than 1 then, I minus E is non singular and it satisfied the norm 1 minus E inverse. So, magnitude of that less than equals to 1 by 1 minus E bound. So, this can be proved based on this kind of theorems. So, you have first you can show if mod E than 1 then E is non singular and second step you show I minus E and expand this to the k th level which is identity matrix minus E to the power $k + 1$, then you can from their combining these two you obtain this.

(Refer Slide Time: 06:12)

Error analysis

d) $(I - E)^{-1} = \lim_{k \rightarrow \infty} \sum_{i=0}^k E^i$. We write this as

$$(I - E)^{-1} = \sum_{i=0}^{\infty} E^i$$

e) Finally:

$$\begin{aligned} \|(I - E)^{-1}\| &= \left\| \lim_{k \rightarrow \infty} \sum_{i=0}^k E^i \right\| = \lim_{k \rightarrow \infty} \left\| \sum_{i=0}^k E^i \right\| \\ &\leq \lim_{k \rightarrow \infty} \sum_{i=0}^k \|E^i\| \leq \lim_{k \rightarrow \infty} \sum_{i=0}^k \|E\|^i \\ &\leq \frac{1}{1 - \|E\|} \end{aligned}$$

Proof as lemma =



So, one can actually get and then once you obtain this take the limit at k tends to ∞ , we get $(I - E)^{-1}$, which is $\sum_{i=0}^{\infty} E^i$ the smallest value.

So, the lemma says $\|I - E\|^{-1}$ should be less than $\frac{1}{1 - \|E\|}$. First we showed that $\|E\|$ is small so, magnitude of that less than 1, $(I - E)^{-1}$ is not singular that is the first part. Second part we show that, $(I - E)^{-1}$ the power inverse. So, basically you have expanded this term and you get $\sum_{i=0}^k (I - E)^i$ and when you take the limit this becomes the sum of E to the power ∞ .

And the final step, the magnitude of that equals to $\lim_{k \rightarrow \infty} \sum_{i=0}^k \|E\|^i$ and the limitic sense of k tends to ∞ one can write $\sum_{i=0}^{\infty} \|E\|^i$ then, which is less than $\frac{1}{1 - \|E\|}$ if $\|E\| < 1$, which is nothing, but $\frac{1}{1 - \|E\|}$ so, the it prove the lemma. So, that sense sort of in very strong error bound in terms of computation.

(Refer Slide Time: 07:49)

Error analysis

► Can generalize result:

LEMMA: If A is nonsingular and $\|A^{-1}\| \|E\| < 1$ then $A + E$ is non-singular and

$$\|(A + E)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|E\|}$$

Proof is based on relation $A + E = A(I + A^{-1}E)$ and use of previous lemma.

THEOREM 1: Assume that $(A + E)y = b + e$ and $Ax = b$ and that $\|A^{-1}\| \|E\| < 1$. Then $A + E$ is nonsingular and

$$\frac{\|x - y\|}{\|x\|} \leq \frac{\|A^{-1}\| \|A\|}{1 - \|A^{-1}\| \|E\|} \left(\frac{\|E\|}{\|A\|} + \frac{\|e\|}{\|b\|} \right)$$

INDIAN INSTITUTE OF TECHNOLOGY KANPUR Ashoke De 61

So, the thing is that if we can generalise that lemma if A is a non singular and $\|A^{-1}\| \|E\| < 1$ then $A + E$ is also non singular and that satisfied this so, this one can also prove based on the previous one. So, this can be similarly proved like that. So, the theorem says assume $(A + E)y = b + e$ and $Ax = b$ and that $\|A^{-1}\| \|E\| < 1$ then $A + E$ is non singular and it also satisfied magnitude of $\frac{\|x - y\|}{\|x\|}$. So,

this is the normalisation then, you get magnitude of A inverse magnitude of A by 1 minus A inverse E E by A epsilon by b.

(Refer Slide Time: 09:02)

Error analysis

Proof: From $(A + E)y = b + e$ and $Ax = b$ we get $(A + E)(y - x) = e - Ex$. Hence:

$$y - x = (A + E)^{-1}(e - Ex)$$

Taking norms $\rightarrow \|y - x\| \leq \|(A + E)^{-1}\| [\|e\| + \|E\|\|x\|]$
Dividing by $\|x\|$ and using result of lemma

$$\frac{\|y - x\|}{\|x\|} \leq \|(A + E)^{-1}\| [\|e\|/\|x\| + \|E\|]$$

$$\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|E\|} [\|e\|/\|x\| + \|E\|]$$

$$\leq \frac{\|A^{-1}\|\|A\|}{1 - \|A^{-1}\|\|E\|} \left[\frac{\|e\|}{\|A\|\|x\|} + \frac{\|E\|}{\|A\|} \right]$$

Result follows by using inequality $\|A\|\|x\| \geq \|b\| \dots$ QED

INDIAN INSTITUTE OF TECHNOLOGY KANPUR
Ashoke De 62

So, how you prove that? so you consider A plus E y equals to b plus epsilon and A x equals to b So, what we get when you write A plus E y minus x, equals to e minus Ex y minus x equals to A plus E inverse e minus E x so, that is what you get for this. Now, taking the norm y minus x the norm of that must be less than the inequality says A plus E inverse the norm of that and norm of E norm of E norm of x. Now, you divide that by the norm of x, so the inequalities stands and the complete 1 is divided by x and if you do the algebra you can prove that, that this happens.

(Refer Slide Time: 10:00)

Error analysis

Simplification when $e = 0$:

$$\frac{\|x - y\|}{\|x\|} \leq \frac{\|A^{-1}\| \|E\|}{1 - \|A^{-1}\| \|E\|}$$

Simplification when $E = 0$:

$$\frac{\|x - y\|}{\|x\|} \leq \|A^{-1}\| \|A\| \frac{\|e\|}{\|b\|}$$

▶ Slightly weaker form: Assume that $\|E\|/\|A\| \leq \delta$ and $\|e\|/\|b\| \leq \delta$ and $\delta\kappa(A) < 1$ then

$$\frac{\|x - y\|}{\|x\|} \leq \frac{2\delta\kappa(A)}{1 - \delta\kappa(A)}$$

} Linear Algebra Computation

Another common form:

THEOREM 2: Let $(A + \Delta A)y = b + \Delta b$ and $Ax = b$ where $\|\Delta A\| \leq \epsilon\|E\|$, $\|\Delta b\| \leq \epsilon\|e\|$, and assume that $\epsilon\|A^{-1}\|\|E\| < 1$. Then

$$\frac{\|x - y\|}{\|x\|} \leq \frac{\epsilon\|A^{-1}\|\|A\|}{1 - \epsilon\|A^{-1}\|\|E\|} \left(\frac{\|e\|}{\|b\|} + \frac{\|E\|}{\|A\|} \right)$$

INDIAN INSTITUTE OF TECHNOLOGY KANPUR
Ashoke De 63

Now, one can do some sort of an simplification to this kind of system. One immediate simplification is that, you say E is 0, so then norm of x minus y divided by norm of x less than equals to this, so it boils down to a another inequality which is A. So, from the generic case, you can always simplify for the special cases and simplification when capital E 0 x norm x minus y by norm x less than A inverse A e by b. So, that is simplification, which you can do.

Now, if you write in shorten or slightly weaker form and assuming that, norm of E by norm A less than equal to delta and norm of epsilon by b or e by b less than delta then I can write, norm of x minus y divided by norm of x less than 2 delta condition number a by 1 minus delta condition number A. So, it can be also written in another form which leads to second theorem. So, that these are the part where you can find in the any linear algebra calculation or book with linear algebra computation. So, that provides those information very nicely.

So, let A plus delta A y equals to b plus delta b and A x equals to b where norm of delta A less than equals to epsilon norm of E and norm of delta b less than equals to epsilon into e and assume that epsilon norm of A inverse norm of E less than 1 then it satisfied this inequality. These are hard core matrix computing theorem, which typically if one wants to have a very efficient algorithm for the linear system, he needs to keep checking all these details and then put the norm.

(Refer Slide Time: 12:33)

Error analysis

Normwise backward error

Question: By how much do we need to perturb data for an approximate solution y to be the exact solution of the perturbed system?

Normwise backward error for $y \equiv$ (def) smallest ϵ for which

$$(1) \begin{cases} (A + \Delta A)y = b + \Delta b; \\ \|\Delta A\| \leq \epsilon \|E\|; \quad \|\Delta b\| \leq \epsilon \|e\| \end{cases}$$

Denoted by $\eta_{E,e}(y)$.

► y is given (a computed solution). E and e to be selected (most likely 'directions of perturbation for A and b ').

► Typical choice: $E = A, e = b$

☒ Explain why this is not unreasonable



Now, backward norm if you put, so it can tell you how much we can need to perturb the data, so, that we can still exert solution of the perturb system. So, backward error of y is smallest epsilon for which, A plus delta A y equals to b plus delta b and norm of delta A less than equals to delta E , where delta b less than equals to epsilon e . So, y is given at the computed solution and E and small e to be the selected for the perturbation and typical choice is that E is A and small e is b .

(Refer Slide Time: 13:29)

Error analysis

Let $r = b - Ay$ (!) Then we have:

THEOREM 3: $\eta_{E,e}(y) = \frac{\|r\|}{\|E\| \|y\| + \|e\|}$

Normwise backward error is for case $E = A, e = b$:

$$\eta_{A,b}(y) = \frac{\|r\|}{\|A\| \|y\| + \|b\|}$$

☒ Show how this can be used in practice as a means to stop some iterative method which computes a sequence of approximate solutions to $Ax = b$.

☒ Consider the 6×6 Vandermonde system $Ax = b$ where $a_{ij} = j^{2(i-1)}$, $b = A * [1, 1, \dots, 1]^T$. We perturb A by E , with $|E| \leq 10^{-10}|A|$ and b similarly and solve the system. Evaluate the backward error for this case. Evaluate the forward bound provided by Theorem 2. Comment on the results.



Why that is not an unreasonable choice, let us see, you define r equals to b minus Ay then we have $\eta_{E,e}(y)$ equals to norm of r by $\|E\| \|y\| + \|e\|$. So, norm wise

backward error is for case E is equal to A and small e equals to b. So, that shows this expression.

So, one can so, you can find out by considering some system and then estimate this kind of norm or one can take some examples and put it some code in the MATLAB and find out that.

(Refer Slide Time: 14:15)

Error analysis

Componentwise backward error

A few more definitions on norms...

- ▶ A norm is **absolute** $\|x\| = \|x\|$ for all x . (satisfied by all p -norms).
- ▶ A norm is **monotone** if $|x| \leq |y| \rightarrow \|x\| \leq \|y\|$.
- ▶ It can be shown that these two properties are equivalent.

☑ Show: a function which satisfies the first 2 requirements of vector norms (1. $\phi(x) \geq 0$ ($=0$, iff $x = 0$) and 2. $\phi(\lambda x) = |\lambda|\phi(x)$) satisfies the triangle inequality iff its unit ball is convex.

☑ (Continued) Use the above to construct a norm in \mathbb{R}^2 that is **not** absolute.

INDIAN INSTITUTE OF TECHNOLOGY KANPUR Ashoke De 66

Now, if you go by component wise then, some more definitions which would be handy to have. The absolute norm equals to the norm of that, for all x which satisfied the p norm. Secondly, the norm is also monotone, if mod x less than equal to mod y which leads to the norm of x less than equals to norm of y. One can show that, these properties are there are some equivalents to that. So, you can show easily these things.

(Refer Slide Time: 15:00)

Error analysis

Define absolute *matrix* norms in same way. Which of the norms $\|A\|_1$, $\|A\|_\infty$, $\|A\|_2$, and $\|A\|_F$ are absolute?

Recall that for any matrix $f(A) = A + E$ with $|E| \leq \alpha|A|$. For an absolute matrix norm

$$\frac{\|E\|}{\|A\|} \leq \alpha$$

What does this imply?

Analogue of theorem 2:

THEOREM 5 Let $Ax = b$ and $(A + \Delta A)y = b + \Delta b$ where $|\Delta A| \leq \epsilon E$ and $|\Delta b| \leq \epsilon e$. Assume that $\epsilon \|A^{-1}\| \|E\| \leq 1$, where $\|\cdot\|$ is an absolute norm. Then,

$$\frac{\|x - y\|}{\|x\|} \leq \frac{\epsilon \|A^{-1}\| (\|E\| \|x\| + \|e\|)}{1 - \epsilon \|A^{-1}\| \|E\| \|x\|}$$

Now, there are another theorem, which is theorem number 5, which is analogous to theorem 2, which state that, if you have $Ax = b$ and $(A + \Delta A)y = b + \Delta b$, where $|\Delta A| \leq \epsilon E$ or the mod of ΔA less than equals to $A \epsilon E$ or mode of Δb less than equals to ϵe , assume that $\epsilon \|A^{-1}\| \|E\| \leq 1$, where dot is an absolute norm. Then, the linear raised definition would be $\frac{\|x - y\|}{\|x\|} \leq \frac{\epsilon \|A^{-1}\| (\|E\| \|x\| + \|e\|)}{1 - \epsilon \|A^{-1}\| \|E\| \|x\|}$. So, one can also prove this.

(Refer Slide Time: 15:53)

Error analysis

In addition, equality achieved to order ϵ for infinity norm.

Implication:

$$\lim_{\epsilon \rightarrow 0} \sup \left\{ \frac{\|\Delta x\|_\infty}{\epsilon \|x\|_\infty} : (A + \Delta A)(x + \Delta x) = b + \Delta b \right\}$$

equals $\text{rcond}_{E,e}(A, x) \equiv \frac{\|A^{-1}\| (\|E\| \|x\| + \|e\|)_\infty}{\|x\|_\infty}$

Cond. number depends on x (i.e. on right-hand side b)

Case $E = |A|$

$e = 0$ yields:

$$\text{rcond}(A, x) \equiv \frac{\|A^{-1}\| \|A\| \|x\|_\infty}{\|x\|_\infty}$$

Componentwise relative condition number :

$$\text{rcond}(A) \equiv \|A^{-1}\| \|A\|_\infty$$

useful formula for condition number

Redo example seen after Theorem 3, (6×6 Vandermonde system) using componentwise analysis.

Now, in addition to that, finding that infinity norm, the implication is that if you take the limit of that whole theorem you get back this condition system. And the condition

number, recondition number depends on x and all these things and the relative condition number would be define that A inverse magnitude, A magnitude and the norm of that at the infinity level. So, this would be an very useful formula for condition number, and where you need to take care of all these while you will be programming your linear solver.

(Refer Slide Time: 16:45)

Error analysis

Componentwise backward error for $y \equiv$ is the smallest ϵ for which

$$(2) \begin{cases} (A + \Delta A)y = b + \Delta b; \\ |\Delta A| \leq \epsilon E; \quad |\Delta b| \leq \epsilon e \end{cases}$$

Denoted by $\omega_{E,e}(y)$.

THEOREM 4 [Oettli-Prager] Let $r = b - Ay$ (residual). Then

$$\omega_{E,e}(y) = \max_i \frac{|r_i|}{(E|y| + e)_i}$$

Zero denominator case: $0/0 \equiv 0$ and nonzero/ $0 \equiv \infty$


INDIAN INSTITUTE OF TECHNOLOGY KANPUR
Ashoke De 69

And the final component twice backward error for y equivalent is the smallest epsilon then, A plus delta y, y equals to b plus delta b and the magnitude less than this.

And the theorem says that, if r is b minus A y which is the residual then, w E y satisfied the maximum of this calculation.

(Refer Slide Time: 17:13)

Error analysis

Example of ill-conditioning: The Hilbert Matrix

- Notorious example of ill conditioning.

$$H_n = \begin{pmatrix} \frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n-1} & \frac{1}{2n} \end{pmatrix} \quad \text{i.e.,} \quad h_{ij} = \frac{1}{i+j-1}$$

- For $n = 5$ $\kappa_2(H_n) = 4.766.. \times 10^5$.
- Let $b_n = H_n(1, 1, \dots, 1)^T$.
- Solution of $H_n x = b$ is $(1, 1, \dots, 1)^T$.
- Let $n = 5$ and perturb $h_{5,1} = 0.2$ into 0.20001 .
- New solution: $(0.9937, 1.1252, 0.4365, 1.865, 0.5618)^T$

Now, you can have an example of the ill condition system, but we may skip that.

(Refer Slide Time: 17:17)

Error analysis

Estimating condition numbers.

Let A, B be two $n \times n$ matrices with A nonsingular and B singular. Then

$$\frac{1}{\kappa(A)} \leq \frac{\|A - B\|}{\|A\|}$$

Proof: B singular $\rightarrow \exists x \neq 0$ such that $Bx = 0$.

$$\begin{aligned} \|x\| &= \|A^{-1}Ax\| \leq \|A^{-1}\| \|Ax\| = \|A^{-1}\| \|(A - B)x\| \\ &\leq \|A^{-1}\| \|A - B\| \|x\| \end{aligned}$$

Divide both sides by $\|x\| \times \kappa(A) = \|x\| \|A\| \|A^{-1}\|$ ➤ result. QED.

Now, this is an estimation of the condition numbers if, you have A and B two n by n matrices and A is non singular and B is singular so, that means, you have 1 singular matrix and 1 non singular matrix then, one can prove 1 by condition number is less than equals to A minus B norm divided by A norm. So, B is singular for x not equals to 0, such that Bx equals to 0. Now, the norm of x is A inverse Ax that, norm which is essentially less than equals to norm of A inverse and norm of Ax .

So, one can rewrite norm of A inverse and norm of A minus B x. So, if I write down that inequality, it is A inverse norm of A inverse A minus B delta x. So, if you divide by both sides the norm of x into condition number it gets you this results. So, that shows the proof for that thing.

(Refer Slide Time: 18:28)

Error analysis

Example:

$\text{let } A = \begin{pmatrix} 1 & 1 \\ 1 & 0.99 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$

non-singular *singular*

Then $\frac{1}{\kappa_1(A)} \leq \frac{0.01}{2} \Rightarrow \kappa_1(A) \geq 200.$

► It can be shown that (Kahan)

$$\frac{1}{\kappa(A)} = \min_B \left\{ \frac{\|A - B\|}{\|A\|} \mid \det(B) = 0 \right\}$$

INDIAN INSTITUTE OF TECHNOLOGY KANPUR Ashoke De 72

Again you take an example, let us say A equals to 1 1 1 0.99 and B has 1 1 1 0. So, this case it is singular and this is a non singular system. Then, I have to show or one has to show that 1 by condition number is less than that or other condition number of A greater than equals to 200.

So, you can show that, this is the minimum of this and find out those numbers.

(Refer Slide Time: 19:10)

Error analysis

Estimating errors from residual norms

Let \tilde{x} an approximate solution to system $Ax = b$ (e.g., computed from an iterative process). We can compute the residual norm:

$$\|r\| = \|b - A\tilde{x}\|$$

Question: How to estimate the error $\|x - \tilde{x}\|$ from $\|r\|$?

- One option is to use the inequality

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}$$

- We must have an estimate of $\kappa(A)$.

$$\begin{aligned} A\tilde{x} &= b \\ \|r\| &= \|b - A\tilde{x}\| \end{aligned}$$

Now, it comes down to the stage where, now you get the errors from the residual norms. Residual means when you are solving $Ax = b$, the residual vector would be $b - Ax$. And this information is quite handy, while you will be dealing with different kind of iterative solvers where, you need to check the convergence level and this error becomes or slowly reducing, so that you get close to the exact solution. And how do you estimate that error, from the residuals norm let us say x prime is an approximate solution what we are trying to find out that Ax equals to b and x would be the exact solution.

So, let us assume x prime is the approximate solution and then that satisfied the equation of the solution to the system $Ax = b$ then it should satisfy $Ax = b$. And then, we can compute the norm of the residual which would be norm of $b - Ax$ prime ok. So, how to estimate this, from this one point is to use the inequality of the norm of $x - \tilde{x}$ divided by $\|x\|$ less than the condition number of A and divide by the norm of r norm of b . So, this plays an crucial role and along with that, the condition number which also comes into the picture to find out that system.

(Refer Slide Time: 21:03)

Error analysis

Proof of inequality.

$$\begin{aligned} Ax &= b \\ A\tilde{x} &= \tilde{b} \end{aligned}$$

First, note that $A(x - \tilde{x}) = b - A\tilde{x} = r$. So:

$$\|x - \tilde{x}\| = \|A^{-1}r\| \leq \|A^{-1}\| \|r\|$$

Also note that from the relation $\tilde{b} = Ax$, we get

$$\Rightarrow \|b\| = \|Ax\| \leq \|A\| \|x\| \rightarrow \|x\| \geq \frac{\|b\|}{\|A\|}$$

Therefore,

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{\|A^{-1}\| \|r\|}{\|b\|/\|A\|} = \kappa(A) \frac{\|r\|}{\|b\|} \quad \blacksquare$$

Show that

$$\frac{\|x - \tilde{x}\|}{\|x\|} \geq \frac{1}{\kappa(A)} \frac{\|r\|}{\|b\|}$$



And as the proof follows you see we have now, the Ax equals to b and the approximate solution satisfied Ax binary to b , then one can have Ax binary x prime equals to b minus Ax prime which is r . So, the norm of x minus x prime is going to be the norm of A inverse r , which is when you put that equality to inequality it is norm of a inverse norm of r . So, from the relation of b equals to Ax you get back, the norm of b equals to norm of Ax , which is again less than norm of A and norm of x . So, what one can write norm of x from here, greater than equals to norm of b divided by norm of A .

So, you just use this simple relationship of the norms for vector and the matrix. So, now, I can write for norm of x minus x , divided by x , which would be less than A inverse norm of A inverse and norm of r divided by norm of b by norm of A . So, which nothing but the condition number of r and b so, one can show that it is greater than this factor.

(Refer Slide Time: 22:47)

Error analysis

THEOREM 6 Let A be a nonsingular matrix and \tilde{x} an approximate solution to $Ax = b$. Then for any norm $\|\cdot\|$,

$$\|x - \tilde{x}\| \leq \|A^{-1}\| \|r\|$$

$r =$ residual vector

In addition, we have the relation

$$\frac{1}{\kappa(A)} \frac{\|r\|}{\|b\|} \leq \frac{\|x - \tilde{x}\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}$$

in which $\kappa(A)$ is the condition number of A associated with the norm $\|\cdot\|$.

So, that takes you to the theorem number 6, which is let A be a non singular matrix and x prime an approximate solution to $Ax = b$. Then for any norm $\|x - x\prime\|$, the norm of that less than equals to $\|A^{-1}\|$ and the r norm.

The r is the residual vector. Now, we have the relationship $\frac{1}{\kappa(A)} \frac{\|r\|}{\|b\|} \leq \frac{\|x - \tilde{x}\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}$. So, which so, $\kappa(A)$ is the condition number and A is associated with the norm.

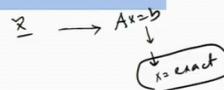
(Refer Slide Time: 23:49)

Error analysis

Iterative refinement

- Define residual vector:

$$r = b - A\tilde{x}$$



- We have seen that: $x - \tilde{x} = A^{-1}r$, i.e., we have

$$x = \tilde{x} + A^{-1}r$$

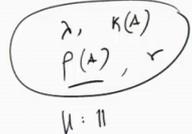
- Idea: Compute r accurately (double precision) then solve

$$A\delta = r$$

... and correct \tilde{x} by

$$\tilde{x} := \tilde{x} + \delta$$

... repeat if needed.



Now, when you go to the iterative solution procedure because finally, you are solving $Ax = b$ and slowly approaching to the exact solution of A through iterative procedure.

What you need to do, you can say that, x prime as we said it could be the approximate solution to this particular system then, the residual vector would be r equals to b minus x prime and we have seen that x minus x prime equals to A inverse minus r . So, we have x equals to x prime plus A inverse r .

Now, what is the idea? The idea behind this process is that, you compute r accurately, that means: precision wise accuracy, then you solve for a δ equals to r So, if r tends to 0, then δ should approaches towards the solution and the correct \bar{x} as \tilde{x} plus δ .

So, what happens if residual vector tends to 0, that means, solution is going towards the convergence then this δ would be also tending to 0. Then the solution the approximate solution will reach towards the exact solution and this is what it is done in the iterative process. And when you do that, actually you have some terms which are associated with A inverse and all these calculations that is taken care of through all these and while doing those things the properties and the theorems that we have discussed they become important.

(Refer Slide Time: 25:56)

Error analysis

ALGORITHM : 1. Iterative refinement

Do {

1. Compute $r = b - A\tilde{x}$
2. Solve $A\delta = r$
3. Compute $\tilde{x} := \tilde{x} + \delta$

} while $\|\delta\| \geq \epsilon \|\tilde{x}\|$

$Ax = b$
 $r = b - Ax \rightarrow 0$

→ exact solution

Why does this work? Model: each solution gets m digits at most because of the conditioning: For example 3 digits. At the first iteration, the error is roughly $\approx 0.001 \times \|b\|$.

► Second iteration: error in δ is roughly $0.001 \times \|r\|$. (but now $\|r\|$ is much smaller than $\|b\|$).

etc ..

INDIAN INSTITUTE OF TECHNOLOGY KANPUR
Ashoke De 77

One has to look at λ , one has to look at the condition number of A , one has to look at the spectral radius of A all these parameters, residual vectors, norm so becomes important. Now, what is that refinement algorithm if one has to see through the pseudo algorithm? You compute first the residual vector b minus A x prime then, you solve for A

delta equals to r. So, now, if you see your original system was $Ax = b$ and your r is $b - Ax$. So, theoretically when you get the exact solution this should approach towards 0.

Now, what you are solving here, the approximated solution \tilde{x} delta equals to r then you once you get delta you update your approximated solution by $\tilde{x} + \delta$. While delta or the norm of delta will remain within this bound, which is greater than equals to epsilon and mod of \tilde{x} . So, this is where so, why it works actually. So, you get some digit, at most because of the conditioning for example, 3 digits at the first iteration, the error is roughly, this second iteration the error delta would be like that, but now residual or the norm of r is much smaller than norm of b .

So, that is why slowly, while you keep doing this process, you move towards the exact solution and you can reduce the computational over rate by not doing the direct approach.

(Refer Slide Time: 27:39)

Error analysis

Iterative refinement - Analysis

► Assume residual is computed exactly. Backward error analysis:

$$(A + F_k)\delta_k = r_k \quad \rightarrow \quad x_{k+1} = x_k + (A + F_k)^{-1}r_k$$

So: $r_{k+1} = b - Ax_{k+1} = \dots = F_k(A + F_k)^{-1}r_k \rightarrow$

$$\|r_{k+1}\| \leq \|F_k\| \|(A + F_k)^{-1}\| \|r_k\|$$

A previous result showed that if $\|F_k\| \|A^{-1}\| < 1$ then

$$\|F_k\| \|(A + F_k)^{-1}\| \leq \frac{\|F_k\| \|A^{-1}\|}{1 - \|F_k\| \|A^{-1}\|}$$

► So : process will converge if (suff. condition)

$$\|F_k\| \|A^{-1}\| \leq \gamma < \frac{1}{2}$$


INDIAN INSTITUTE OF TECHNOLOGY KANPUR
Ashoke De 78

Now, once you do the refinement analysis, so what we have done that residual is computed exactly, so A plus so the backward error if you want to estimate A plus F_k delta k equals to r_k . So, at the second level of refinement x_{k+1} would be $x_k + (A + F_k)^{-1}r_k$. So, one can see that r_{k+1} is $b - Ax_{k+1}$ and similarly if you write. So, the convergence is this sufficient condition for the convergence that one has to achieve or these are the similar kind of prove that we have done.

(Refer Slide Time: 28:25)

Error analysis

Important: Iterative refinement won't work when the residual r consists mostly of noise:

$$\delta = A^{-1}\text{noise}$$

➤ However, see section 2.5 of text for iterative refinement in single precision

Heuristic: If $\epsilon = 10^{-d}$, and $\kappa_{\infty}(A) \approx 10^q$ then each iterative refinement step will gain about $d - q$ digits.

📌 A matlab experiment [do operations indicated]

```
1. >> n = 6;           2. >> A = hilb(n);
3. >> b = A*ones(n,1); 4. >> A \ b
....
5.>> B = A;           6. >> B(6,1)=B(6,1)+1.E-06;
7. >> x = B \ b       ....
8. >> xex = ones(n,1); 9. >> norm(xex-x,2)
```

INDIAN INSTITUTE OF TECHNOLOGY KANPUR Ashoke De 79

So, the now what is important here is that, iterative refinement would not work when the residual are consist mostly of noise; that means, delta equals to a inverse noise. However, when you heuristic is that epsilon is 10 to the power minus d and condition number of these. So, the iterative refinement we gain by this much of digits. So, these are the things one can calculate.

(Refer Slide Time: 28:53)

Error analysis

```
....
10. >> res = b - A*x;   11. >> x = x + B \ res ;
12. >> norm(xex-x,2)    ...
13. >> res = b - A*x;   14. >> x = x + B \ res;
15. >> norm(xex-x,2)    ...
```

repeat a couple of times..

Observation: we gain about 3 digits per iteration.

📌 Read Section 3.5 of text

INDIAN INSTITUTE OF TECHNOLOGY KANPUR Ashoke De 80

And these are the observation what one can obtain.

(Refer Slide Time: 28:58)

Solution of linear systems

Background: Linear systems

The Problem: A is an $n \times n$ matrix, and b a vector of \mathbb{R}^n . Find x such that:

$$Ax = b$$

► x is the **unknown vector**, b the **right-hand side**, and A is the **coefficient matrix**

Example:

$$\begin{cases} 2x_1 + 4x_2 + 4x_3 = 6 \\ x_1 + 5x_2 + 6x_3 = 4 \\ x_1 + 3x_2 + x_3 = 8 \end{cases} \text{ or } \begin{pmatrix} 2 & 4 & 4 \\ 1 & 5 & 6 \\ 1 & 3 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \\ 8 \end{pmatrix}$$

☒ Solution of above system ?

INDIAN INSTITUTE OF TECHNOLOGY KANPUR Ashoke De 81

Now, the linear system the background is that, you have A n by n system or with the b vector.

Vector b right hand side, so you have to solve, so this is a system for example, you get 2 x 1 x 2 and you get A x equals to b.

(Refer Slide Time: 29:17)

Solution of linear systems

► **Standard mathematical solution by Cramer's rule:**

$$x_i = \det(A_i) / \det(A)$$

A_i = matrix obtained by replacing i -th column by b .

► **Note:** This formula is useless in practice beyond $n = 3$ or $n = 4$.

Three situations:

1. The matrix A is nonsingular. There is a unique solution given by $x = A^{-1}b$.
2. The matrix A is singular and $b \in \text{Ran}(A)$. There are infinitely many solutions.
3. The matrix A is singular and $b \notin \text{Ran}(A)$. There are no solutions.

INDIAN INSTITUTE OF TECHNOLOGY KANPUR Ashoke De 82

So, A get you the coefficient matrix and standard mathematical solution by Cramer's rule will get you like that. And the situations the matrix a is non singular, you get a solution if matrix A is singular, there are infinitely many solution of matrix A is singular there are no solutions. So, this is where, you can see when you are trying to solve A x equals to b

system and these are the properties one has to keep in mind. So, now, once we talk about all these properties and these are the basic important properties, one has to know regarding a linear system or the matrix.

And once we know all these properties then, it will be easier to talk about or discuss about the linear system and how we get a solution procedure and the linear solver, different kind of linear solver. This actually concludes the content of the finite volume 1 and it will give you an jump start or get you started with a finite volume method to going on. And I hope you have enjoyed this lecture series and continue in the next part of this finite volume series.

Thank you very much.