# Lecture 18

Last time we introduced norms on $\mathbb{R}^n$. We also studied matrix norms

<u>Recall</u> a norm on $\mathbb{R}^n$ is a function $\|\ \|$ which associates to each vector $\bar{a}$ in $\mathbb{R}^n$, a real number $\|a\|$ s/t

1) $\|\bar{a}\| \geq 0 \quad \forall \bar{a} \in \mathbb{R}^n$

$$\|\bar{a}\| = 0 \text{ iff } \bar{a} = 0$$

2) for $\alpha \in \mathbb{R}$,

$$\|\alpha \bar{a}\| = |\alpha| \, \|\bar{a}\|$$

3) (triangle inequality)

$$\|\bar{a} + \bar{b}\| \leq \|\bar{a}\| + \|\bar{b}\|$$

---

Examples of norms

$\bar{a} = (a_1, \ldots, a_n)$

$$\|a\|_2 = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$$

$$\|a\|_1 = |a_1| + |a_2| + \cdots + |a_n|$$

$$\|a\|_\infty = \max_{1 \le i \le n} |a_i|$$

---

## Matrix Norms

a matrix norm is a real valued function defined over the set of $n \times n$ matrices such the for any $n \times n$ matrices $A, B$ and real number $\alpha$ we have

(1) $\|A\| \geq 0 \quad \forall A$

and $\|A\| = 0 \iff A = 0$

(2) $\|\alpha A\| = |\alpha| \, \|A\|$

(3) $\|A + B\| \le \|A\| + \|B\|$

(4) $\|AB\| \le \|A\| \, \|B\|$

**Theorem** Let $\| \ \|$ be a norm in $\mathbb{R}^n$

Then for a $n \times n$ matrix $A$

$$\|A\| = \max_{\|x\|=1} \|Ax\|$$

defines a matrix norm on the set of $n \times n$ matrices

**Remark** $$\|A\| = \max_{\|x\|=1} \|Ax\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

Thus we get $\|Ax\| \leq \|A\| \, \|x\|$ for all $x \in \mathbb{R}^n$.

---

**Examples**

1) $$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

2-norm on matrices

"difficult to compute".

(2) $\|A\|_\infty = \max\limits_{\|x\|_\infty = 1} \|Ax\|_\infty$

Theorem

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$$

(3) $\|A\|_1 = \max\limits_{\|x\|_1 = 1} \|Ax\|_1$

Ex

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|$$

---

Remark

If the matrix $A$ is invertible then

$\bar{x} = A^{-1}(Ax)$

So $\|\bar{x}\| = \|A^{-1}(Ax)\| \le \|A^{-1}\| \, \|Ax\|$

Thus $\dfrac{\|\bar{x}\|}{\|A^{-1}\|} \le \|Ax\| \le \|A\| \, \|x\|$

## Reason for studying norm

We want to study errors in computing solutions to linear equation

Solving $Ax = b$

we get an approximate solⁿ $\hat{x}$

error $\bar{e} = \bar{x} - \hat{x}$

This error is unknown to us

"residual error"

$$\bar{r} = A\bar{e} = Ax - A\hat{x} = b - A\hat{x}$$

$$\bar{r} = b - A\hat{x}$$

This we can compute

---

Last time I gave example which showed that the size of the residual $\bar{r}$ is not always a reliable indicator of the size of error $\bar{e}$.

It depends on the size of $A$ and $A^{-1}$

error $\quad \bar{e} = x - \hat{x}$

residual $\quad \bar{r} = A\bar{e} = b - A\hat{x}$

$$\therefore \quad \bar{e} = A^{-1}\bar{r}$$

Recall we have proved that for an invertible matrix $B$

$$\frac{\|\bar{u}\|}{\|B^{-1}\|} \le \|Bu\| \le \|B\| \, \|u\|$$

apply this to $B = A^{-1}$ and $u = \bar{r}$

$$(*) \quad \frac{\|\bar{r}\|}{\|A\|} \le \|e\| = \|A^{-1} r\| \le \|A^{-1}\| \, \|\bar{r}\|$$

$(*)$ gives an upper and a lower bound of the relative error $\frac{\|e\|}{\|x\|}$ in terms of

relative residual $\quad \frac{\|\bar{r}\|}{\|\bar{b}\|}$

$$\frac{\|\bar{b}\|}{\|A\|\,\|x\|} \quad \frac{\|\bar{r}\|}{\|b\|} \le \frac{\|e\|}{\|x\|} \le \frac{\|A^{-1}\|\,\|b\|}{\|x\|} \cdot \frac{\|r\|}{\|b\|}$$

(*) for $\hat{x} = 0$ gives

$$\frac{\|b\|}{\|A\|} \le \|x\| \le \|A^{-1}\|\,\|b\|$$

$$\frac{1}{\|A^{-1}\|\,\|A\|} \frac{\|r\|}{\|b\|} \le \frac{\|e\|}{\|x\|} \le \|A^{-1}\|\,\|A\| \frac{\|r\|}{\|b\|}$$

$$\boxed{\mathrm{cond}(A) = \|A\|\,\|A^{-1}\|}$$

condition number of $A$

$$\frac{1}{\mathrm{cond}(A)} \frac{\|r\|}{\|b\|} \le \frac{\|e\|}{\|x\|} \le \mathrm{cond}(A) \frac{\|r\|}{\|b\|}$$

example

$$A = \begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} \frac{101}{4} & -\frac{99}{4} \\ -\frac{99}{4} & \frac{101}{4} \end{bmatrix}$$

$$\|A\|_\infty = 2$$

$$\|A^{-1}\|_\infty = 50$$

So $\text{cond}(A) = 100$

---

Today we do Backward error Analysis and Iterative improvement.

$\boxed{\text{Theorem}}$ Let $A$ be a $n \times n$ invertible matrix.

$$\frac{1}{\text{cond}(A)} = \min \left\{ \frac{\|A-B\|}{\|A\|} \;\middle|\; B \text{ is not invertible} \right\}$$

Pf We only prove $\leq$

So we prove

$$\frac{1}{\text{cond}(A)} \leq \frac{\|A-B\|}{\|A\|}$$ for any non-invertible matrix $B$

$$\text{cond}(A) = \|A\| \, \|A^{-1}\|$$

So we have to prove

$$\frac{1}{\|A^{-1}\|} \leq \|A-B\|$$ $B$ non-invertible

$B$ non-invertible. So $\exists \, x \neq 0$ s/t $Bx = 0$

1) $\|(A-B)x\| \leq \|A-B\| \, \|x\|$

$$\|(A-B)x\| = \|Ax - Bx\| = \|Ax\| \geq \frac{\|x\|}{\|A^{-1}\|}$$

So $\|A-B\| \, \|x\| \leq \frac{\|x\|}{\|A^{-1}\|}$

as $\|x\| \neq 0$ we get $\frac{1}{\|A^{-1}\|} \leq \|A-B\|$

**Corollary** :- If $A$ is invertible and $B$ is a matrix such that

$$\|A-B\| < \frac{1}{\|A^{-1}\|}$$

then $B$ is invertible

---

We use the inequality

$$\frac{1}{\|A^{-1}\|} \leq \|A-B\| \quad \text{where } B \text{ is singular}$$

to estimate $\|A^{-1}\|$ without directly computing $A^{-1}$

**Example** ① $A = \begin{pmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{pmatrix}$

$$B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

B is singular

$$A - B = \begin{pmatrix} 0.01 & -0.01 \\ -0.01 & 0.01 \end{pmatrix}$$

$$\|A-B\|_\infty = 0.02$$

so $\|A^{-1}\|_\infty \geq \dfrac{1}{0.02} = 50$

So $\text{cond}(A) = \|A\| \, \|A^{-1}\| \geq 100$

In this case it is exact.

---

Example 2

A is an invertible upper triangular matrix

Then $\quad \text{cond}(A) \geq \dfrac{\|A\|_\infty}{\min_i |a_{ii}|}$

Pf A invertible & upper triangular.

So all diagonal entries of A are non-zero.

without loss of generality say

$$|a_{11}| = \min_i |a_{ii}|$$

$$B = A - \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & 0 & & 0 \\ & & & \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

note $\quad$ B is singular $\left( \because \text{B is upper triangular } \& \ b_{11} = 0 \right)$

$$\|A - B\|_\infty = |a_{11}|$$

So $\quad \|A^{-1}\| \geq \dfrac{1}{|a_{11}|}$

## Perturbations of linear systems of equations

If the linear system $Ax = b$ derives from a practical problem, we must expect the coefficients of this system to be subject to error either because they result from other calculations or from physical measurement.

Hence assuming for example the RHS is accurate we are in effect solving the system $\hat{A}\hat{x} = b$ instead of $Ax = b$ where $A = \hat{A} + E$, the matrix $E$ contains the errors in the coefficients, Even if all calculations are carried out exactly we only have a solution of $\hat{A}\hat{x} = b$ rather than $Ax = b$

Now $x = A^{-1} b$

$$= A^{-1} \hat{A} \hat{x}$$

$$= A^{-1} (A + \hat{A} - A) \hat{x}$$

$$= \hat{x} + A^{-1} (\hat{A} - A) \hat{x}$$

Now $\hat{A} - A = -E$

So $x = \hat{x} - A^{-1} E \hat{x}$

So $\|x - \hat{x}\| \leq \|A^{-1}\| \|E\| \|\hat{x}\|$

$$= \|A^{-1}\| \|A\| \frac{\|E\|}{\|A\|} \|\hat{x}\|$$

Thus $\boxed{\dfrac{\|x - \hat{x}\|}{\|\hat{x}\|} \leq \operatorname{cond}(A) \dfrac{\|E\|}{\|A\|}}$

Thus if the coefficients of the linear system $Ax=b$ are known to be accurate only to about $10^{-s}$ (relative to size of $A$) and $\text{cond}(A) \simeq 10^{t}$

Then there is no point in calculating the solution to a relative accuracy of $10^{t-s}$.

---

Quite loosely we say the linear system $Ax=b$ is "ill-conditioned" if cond (A) is "large".

## Iterative improvement of solution

Let $e = x - \hat{x}^{(1)}$ be the (unknown) error in the approximate solution $\hat{x}^{(1)}$ for $Ax = b$

$$A e = r = b - A \hat{x}^{(1)} \qquad (*)$$

Let $\hat{e}^{(1)}$ be approximate solution of $(*)$. Then $\hat{e}^{(1)}$ need not equal $e$

But at the very least $\hat{e}^{(1)}$ gives an indication of the size of $e$

If $\dfrac{\|\hat{e}^{(1)}\|}{\|\hat{x}^{(1)}\|} \approx 10^{-3}$ then we

conclude that the first 3 decimals

of $\hat{x}^{(1)}$ agree with that of the exact answer $x$.

We would then also expect $\hat{e}^{(1)}$ to be an accurate approximation to $e$.

So $\hat{x}^{(2)} = \hat{x}^{(1)} + \hat{e}^{(1)}$ to be a better approximation to $x$ than $\hat{x}^{(1)}$

We can now, if necessary compute the new residual

$$r = b - A\hat{x}^{(2)}$$

and solve $Ae = r$ to obtain a new correction $\hat{e}^{(2)}$ and a new approximation $\hat{x}^{(3)} = \hat{x}^{(2)} + \hat{e}^{(2)}$

The number of places in agreement in the successive approximation $\hat{x}^{(1)}, \hat{x}^{(2)}, \ldots$ as well as an examination of the successive residuals should give an indication of the accuracy of these approximate solutions.

One normally carries out this iteration until

$$\frac{\| \hat{e}^{(k)} \|}{\| \hat{x}^{(k)} \|} \approx 10^{-t} \quad \text{if } t$$

decimal places are carried during the calculation.

# Iteration steps increase with cond(A)

If cond(A) is "very large" the correction $\hat{e}^{(1)}, \hat{e}^{(2)}, \ldots,$ may never decrease in size

## Remark

For the success of iterative improvement it is absolutely mandatory that residuals be computed as accurately as possible. If, as is usual, floating pt arithmetic is used, the residual should always be calculated in double-precision arithmetic.