Lecture 13

Last time we did

- Cholesky's factorization of a positive definite matrix A as

$$A = LL^t$$ where L is lower triangular

This requires only 50% of the calculation needed for LU decomposition.

If GE has row changes then we can factor A into

$$PA = LU$$ where P is a permutation matrix

We solve $Ax = b$

by $PAx = Pb = b'$

$$LUx = b'$$

Set $y = Ux$

Solve $Ly = b'$

Then solve $Ux = y$

Matrix factorization is useful if we have to solve $Ax = b$ for many different values of b

---

We then turned our attention to round-off errors that can occur while doing GE

We learnt scaled partial pivoting for GE

Today we study Matrix Norms

We first study norms in $\mathbb{R}^n$.

Q. What is a norm?

A. We want to measure "size" of a vector $\bar{a}$ in $\mathbb{R}^n$

Defn A norm assigns to each vector $\bar{a}$ in $\mathbb{R}^n$ a number $\|\bar{a}\|$, called the norm of $\bar{a}$ subject to the following restriction

1) $\|\bar{a}\| \geq 0$ for all $\bar{a} \in \mathbb{R}^n$ and
$\|\bar{a}\| = 0 \iff \bar{a} = 0$

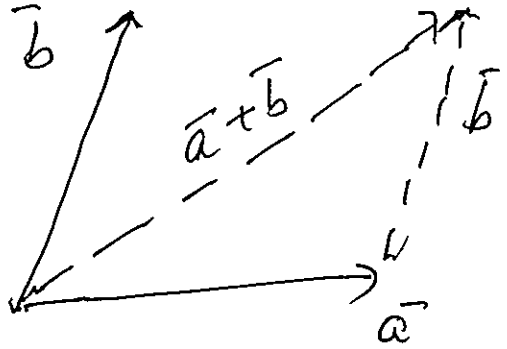
2) $\|r\bar{a}\| = |r| \, \|\bar{a}\|$ for all $\bar{a} \in \mathbb{R}^n$ and $r \in \mathbb{R}$

(3) (Triangle inequality)
$\|\bar{a}+\bar{b}\| \leq \|\bar{a}\| + \|\bar{b}\|$
for all $\bar{a}, \bar{b} \in \mathbb{R}^n$

Note the first restriction forces all n-vectors but the zero vector to have positive "length".

The second restriction states for example that $\bar{a}$ and its negative $-\bar{a}$ have the same length. Also $3\bar{a}$ has length 3 times that of $\bar{a}$

The third restriction is called "triangle-inequality". This is because of the following Geometric picture



Examples of norms

1. Euclidean norm

$$\bar{a} = (a_1, a_2, \ldots, a_n)$$

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

Euclidean norm is also called the 2-norm.

Easy to verify

1) $\|\bar{a}\| \geq 0$ and $= 0$ iff $\bar{a} = 0$

2) $\|r\bar{a}\| = |r| \, \|\bar{a}\|$ for all $r \in \mathbb{R}$

To prove triangle inequality we need the "Cauchy-Schwarz inequality".

Theorem For $\bar{x}, \bar{y} \in \mathbb{R}^n$

$$\bar{x} \cdot \bar{y} \leq \|x\|_2 \, \|y\|_2 .$$

Proof If $\bar{x} = 0$ or $\bar{y} = 0$ then we have nothing to prove.

So assume $\bar{x} \neq 0$ & $\bar{y} \neq 0$

For $\lambda \in \mathbb{R}$

$$0 \leq \|x - \lambda y\|_2^2 = \sum_{i=1}^{n} (x_i - \lambda y_i)^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\lambda \sum_{i=1}^{n} x_i y_i + \lambda^2 \sum_{i=1}^{n} y_i^2$$

$$2\lambda \sum_{i=1}^{n} x_i y_i \leq \|x\|_2^2 + \lambda^2 \|y\|_2^2$$

Notice $\|x\|_2 > 0$ and $\|y\|_2 > 0$

Take $\lambda = \dfrac{\|x\|_2}{\|y\|_2}$

$$2\frac{\|x\|_2}{\|y\|_2}(x\cdot y) \le \|x\|_2^2 + \frac{\|x\|^2}{\|y\|_2^2}\|y\|_2^2$$

$$\| \\ 2\|x_2\|_2^2$$

So $x\cdot y \le \|x\|_2\|y\|_2$.

We now prove triangle inequality.

$$\|x+y\|_2^2 = \sum_{i=1}^{n}(x_i+y_i)^2$$

$$= \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2$$

$$\le \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2$$

$$= (\|x\|_2 + \|y\|_2)^2$$

Thus $\|x+y\|_2 \le \|x\|_2 + \|y\|_2$.

1-norm

$\bar{a} = (a_1, a_2, \ldots, a_n)$

~~$\|\bar{a}\|_1 = |a_1| + |a_2| + \cdots$~~

$\|\bar{a}\|_1 = |a_1| + |a_2| + \cdots + |a_n|$

Clearly $\|\bar{a}\|_1 \geq 0$ and $= 0$ iff $\bar{a} = 0$

also $\|r\bar{a}\|_1 = |r| \, \|\bar{a}\|_1$ for all $r \in \mathbb{R}$.

$$\|\bar{a}+\bar{b}\|_1 = |a_1+b_1| + |a_2+b_2| + \cdots + |a_n+b_n|$$

$$\leq |a|_1 + |b_1| + |a_2| + |b_2| + \cdots + |a_n| + |b_n|$$

$$= \|\bar{a}\|_1 + \|\bar{b}\|_1$$

Thus $\|\bar{a}+\bar{b}\|_1 \leq \|\bar{a}\|_1 + \|\bar{b}\|_1$.

∞ - norm

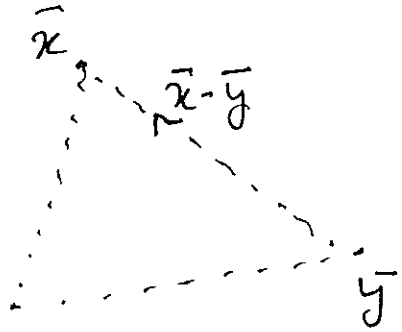
$$\bar{a} = (a_1, a_2, \ldots, a_n)$$

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

One can easily verify that $\|\bar{a}\|_\infty$ is also a norm on $\mathbb{R}^n$

If $\| \ \|$ is a norm in $\mathbb{R}^n$ then we can define distance between two vectors $\bar{x}$ and $\bar{y}$ as

$$\text{dist}(\bar{x}, \bar{y}) = \|\bar{x} - \bar{y}\|$$



Matrix Norm

A matrix norm on the set of all $n \times n$ matrices is a real valued function $\| \ \|$, defined on this set satisfying for all $n \times n$ matrices A, B and all real number α

(i) $\|A\| \geq 0$

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

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

(iii) $\|A+B\| \leq \|A\| + \|B\|$

(iv) $\|AB\| \leq \|A\| \, \|B\|$

Theorem:- If $\|\ \|$ is a norm in $\mathbb{R}^n$

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

is a matrix norm.

Proof (i) Clearly $\|A\| \geq 0$

$\|A\| = 0 \Rightarrow \|Ax\| = 0$ for all x with $\|x\| = 1$

$\Rightarrow Ax = 0$ for all x with $\|x\| = 0$

Claim $A = 0$

Otherwise some column (say i^{th} column) is non-zero

$e_i = i^{th}$ co-ordinate vector

$Ae_i \neq 0$

$\uparrow$

i^{th} column

$x = \frac{e_i}{\|e_i\|}$ has norm 1

$Ax = \frac{1}{\|e_i\|} Ae_i \neq 0$

a contradiction

(ii) $\|\alpha A\| = \max_{\|x\|=1} \|\alpha Ax\|$

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

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

$= |\alpha| \|A\|$

(iii) $\|A+B\| = \max_{\|x\|=1} \|(A+B)x\|$

for x with $\|x\|=1$

$\|Ax+Bx\| \leq \|Ax\| + \|Bx\|$

$\leq \max_{\|x\|=1} \|Ax\| + \max_{\|x\|=1} \|Bx\|$

$= \|A\| + \|B\|$

Thus $\max_{\|x\|=1} \|(A+B)(x)\| \leq \|A\| + \|B\|$

So $\|A+B\| \leq \|A\| + \|B\|$

(iv) $\|AB\| = \max_{\|x\|=1} \|ABx\|$

Case(1) $Bx = 0$

$\|ABx\| = 0 \leq \|A\| \|B\|$

Case(2) $Bx \neq 0$

$x = \frac{Bx}{\|Bx\|}$ has norm 1

$\left\| A\left(\frac{Bx}{\|Bx\|}\right) \right\| \leq \|A\|$

$\frac{1}{\|Bx\|} \|ABx\| \leq \|A\|$

$\|ABx\| \leq \|A\| \|Bx\| \leq \|A\| \|B\|$

Thus $\|AB\| \leq \|A\| \|B\|$

Remark :-

for $\bar{z} \neq 0$ $\quad \bar{x} = \dfrac{\bar{z}}{\|\bar{z}\|}$ has norm 1

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

$$\left\| A\left(\frac{\bar{z}}{\|\bar{z}\|}\right) \right\| \leq \|A\|$$

$$\frac{1}{\|\bar{z}\|} \|A\bar{z}\| \leq \|A\|$$

So $\quad \|Az\| \leq \|A\| \|z\|$

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

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1$$ 1-norm

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty$$ ∞-norm

Euclidean norm is very difficult to calculate.

However one can calculate both the 1-norm and the ∞-norm

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

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

$\|A\|_\infty$ sums rows

$\|A\|_1$ sums columns

Example $A = \begin{bmatrix} 3 & 2 & 3 \\ 0 & 4 & 2 \\ -3 & 1 & -2 \end{bmatrix}$

$\|A\|_\infty = \max\{8, 6, 6\} = 8$

$\|A\|_1 = \max\{6, 7, 7\} = 7$

Proof of Theorem

We want to show

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

We first show

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

Let $\bar{x} \in \mathbb{R}^n$ with $\|\bar{x}\|_\infty = 1$

$$\|\bar{x}\|_\infty = \max_{1\le i\le n} |x_i|$$

$$\|Ax\|_\infty = \max_{1\le i\le n} |(Ax)_i|$$

$$= \max_{1\le i\le n} \left| \sum_{j=1}^{n} a_{ij} x_j \right|$$

Now
$$\left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \sum_{j=1}^{n} |a_{ij}||x_j| \leq \sum_{j=1}^{n} |a_{ij}|$$
since $|x_j| \leq 1$

Thus $$\|Ax\|_\infty = \max_{1 \leq i \leq n} \left| \sum_{j=1}^{n} a_{ij} x_j \right|$$
$$\leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$$

Thus $$\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty$$
$$\leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$$

We will now show the opposite inequality.

Let p be an integer with
$$\sum_{j=1}^{n} |a_{pj}| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$$

Let $\bar{x}$ be a vector with components

$$x_j = \begin{cases} 1 & \text{if } a_{pj} \geq 0 \\ -1 & \text{if } a_{pj} < 0 \end{cases}$$

$$\|x\|_\infty = \max_{1 \leq j \leq n} |x_j| = 1$$

$$a_{pj} x_j = |a_{pj}| \quad \text{for } j = 1, 2, \ldots, n$$

$$\|Ax\|_\infty = \max_{1 \leq i \leq n} \left| \sum_{j=1}^{n} a_{ij} x_j \right|$$

$$\geq \left| \sum_{j=1}^{n} a_{pj} x_j \right| = \left| \sum_{j=1}^{n} |a_{pj}| \right|$$

$$= \sum_{j=1}^{n} |a_{pj}|$$

$$= \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$$

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

$$\geq \|Ax\|_\infty \geq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$$

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

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

Frobenius norm

$A = (a_{ij})$

$$\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

One can prove that the Frobenius norm is also a matrix norm

Reasons for studying matrix Norms

Let $\hat{x}$ be computed solution of

$$Ax = b \qquad (*)$$

Let $\bar{x}_0$ be exact solution of $(*)$

Then its error is the difference

$$\bar{e} = \bar{x}_0 - \hat{x}$$

This error is unknown to us since we do not know exact solution $\bar{x}_0$

But we can always compute

"residual error"

$$\bar{r} = A\bar{x}_0 - A\hat{x}$$

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

If $\bar{r} = \bar{0}$ then $\hat{x}$ is exact solution.

One would expect $\bar{r}$ to be small if

$\hat{x}$ is close to x

This is not so as the following example shows

Example

1) $1.01x_1 + 0.99x_2 = 2$

$0.99x_1 + 1.01x_2 = 2$

Unique sol^n $x_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$\hat{x} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ has error $\bar{e} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

but residual error

$\bar{r} = \begin{bmatrix} -0.02 \\ 0.02 \end{bmatrix}$ is small

2) By taking a different right side we can achieve the opposite effect

$$1.01\,x_1 + 0.99\,x_2 = 2$$

$$0.99\,x_1 + 1.01\,x_2 = -2$$

exact answer $x_1 = 100, \quad x_2 = -100$

The approximate solution

$$\hat{x} = \begin{bmatrix} 101 \\ -99 \end{bmatrix}$$ has

small relative error $\bar{e} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

but residual

$$\bar{r} = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$$ is large

---

Thus the residual $\bar{r} = \bar{b} - A\hat{x}$ is not always a reliable indicator of the size of the error $\bar{e} = \bar{x}_0 - \hat{x}$.

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

$$\bar{r} = A\bar{x}_0 - A\hat{x}$$
$$= A(\bar{x}_0 - \hat{x})$$
$$= A\bar{e}$$

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

Remark If a matrix A is invertible

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

$$\|\bar{x}\| \leq \|A^{-1}\| \|Au\|$$

Thus $$\frac{\|\bar{x}\|}{\|A^{-1}\|} \leq \|Au\| \leq \|A\| \|x\|$$

Apply this to $\bar{e} = A^{-1}\bar{r}$ we get

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

This gives an upper and lower bound

On the relative error $\frac{\|\bar{e}\|}{\|\bar{x}_0\|}$ in terms of relative residual $\frac{\|\bar{r}\|}{\|\bar{b}\|}$

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

So

$$\frac{\|b\|}{\|A\| \|x_0\|} \frac{\|\bar{r}\|}{\|\bar{b}\|} \leq \frac{\|\bar{e}\|}{\|x_0\|} \leq \|A^{-1}\| \|b\| \frac{\|\bar{r}\|}{\|\bar{x}_0\| \|\bar{b}\|}$$

If we apply $(*)$ to $\hat{x}=0$ we get $\bar{e} = \bar{x}_0$

So $\frac{\|\bar{b}\|}{\|\bar{A}\|} \leq \|\bar{x}_0\| \leq \|A^{-1}\| \|\bar{b}\|$

$$\frac{1}{\|A^{-1}\| \|b\|} \leq \frac{1}{\|\bar{x}_0\|} \leq \frac{\|\bar{A}\|}{\|\bar{b}\|}$$

So $\frac{1}{\|A\| \|A^{-1}\|} \frac{\|\bar{r}\|}{\|\bar{b}\|} \leq \frac{\|\bar{e}\|}{\|\bar{x}_0\|} \leq \|\bar{A}^{-1}\| \|A\| \frac{\|r\|}{\|b\|}$

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

is called condition number

Thus

$$\frac{1}{\text{cond}(A)} \frac{\|\bar{r}\|}{\|\bar{b}\|} \leq \frac{\|\bar{e}\|}{\|\bar{x}_0\|} \leq \text{cond}(A) \frac{\|\bar{r}\|}{\|\bar{b}\|}$$

Remark

$$\text{Cond}(A) \geq 1$$

$$I = A^{-1} A$$

$$\|I\| \leq \|A^{-1} A\| \leq \|A^{-1}\| \|A\|$$

Note $$\|I\| = \max_{\|x\|=1} \|Ix\|$$

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

$$= 1$$

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

Example

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

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

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

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

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

Theorem Let A be a $n \times n$ invertible matrix

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

Proof We only prove $\leq$

So we prove

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

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

So we have to prove

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

B non-invertible

B non-invertible. So $\exists \, \bar{x} \neq 0$ such that

$$Bx = 0$$

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

2) $\|(A-B)x\| = \|Ax - Bx\|$

$$= \|Ax\|$$

$$\geq \frac{\|x\|}{\|A^{-1}\|} \longrightarrow$$

$\|x\| = \|A^{-1}Ax\|$
$\leq \|A^{-1}\| \, \|Ax\|$

So $\|A-B\| \, \|x\| \geq \dfrac{\|x\|}{\|A^{-1}\|}$

As $\|\bar{x}\| \neq 0$ we get

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

Corollary If A is invertible and B is a matrix such that $\|A-B\| < \dfrac{1}{\|A^{-1}\|}$

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 ② $A = \begin{pmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{pmatrix}$

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

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

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

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

(Infact this is exact!)

Example A is an invertible upper triangular matrix.

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

Proof A is invertible and upper triangular

So all diagonal entries of A are non-zero.

Without loss of generality say

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

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

Note B is singular ($\because$ B is upper triangular and $b_{11} = 0$)

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

$$\text{So } \|A^{-1}\| \geq \frac{1}{|a_{11}|}$$