Last time

Numerical sensitivity of basic operations

Basic representations (why we won't deal with it directly)

Today - start with numerical linear algebra

Solving $Ax=b$

- vector & matrix norms
- condition number $\kappa(A)$
- direct methods
  - Gaussian elimination
  - LU decomposition
  - Cholesky decomposition

Norms (vector)

$\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$ (why $\mathbb{C}$?)

→ difference between $\mathbb{C}$ & $\mathbb{R}$

satisfies

positive definiteness : $\|x\| \geq 0 \ \forall x \in \mathbb{C}^n$ & $\|x\| = 0 \Leftrightarrow x = 0$

homogeneity : $\|\alpha x\| = |\alpha| \|x\| \ \forall \alpha \in \mathbb{C}$ & $x \in \mathbb{C}^n$

triangle inequality : $\|x+y\| \leq \|x\| + \|y\|$

Example

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad x_i \in \mathbb{C} \Rightarrow x \in \mathbb{C}^n$$

p-norm

$$\|x\|_p = \left(\sum_{i=1}^{n} x_i^p \right)^{1/p}$$

popular $p = 1, 2, \infty$
(in probability $3, 4, 2+\varepsilon$)

$p = \infty$ (max or sup norm)

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \dots |x_n|\}$$

Idea $\lim_{p \to \infty} \|x\|_p \to \max|\{x_i\}|$

$M = \max|\{x_i\}|$ so $|x_i| \le M \ \forall i$

$$\left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \le \left(\sum_{i=1}^{n} M^p \right)^{1/p} = \left(n M^p \right)^{1/p} = n^{1/p} M$$

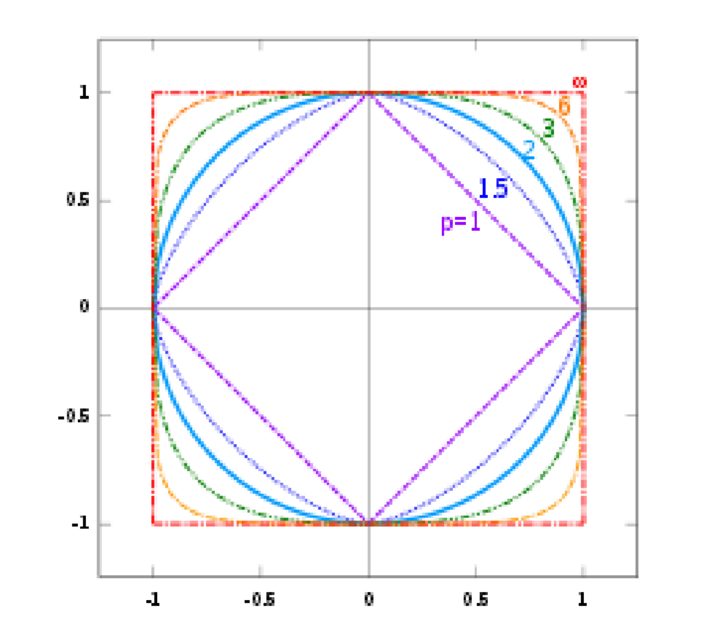
since there is an i st $|x_i| = M$

$$\left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \ge \left(M^p \right)^{1/p}$$

$$M \le \|x\|_p \le n^{1/p} M \qquad p \to \infty$$

$n^{1/p} \to 1$ if $n \ge 1$

Geometric visualization of norms



Note: norms for $p<1$ are not convex

& $p=0$ is usually defined as

# of nonzero entries

Matrix norms

$\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$

- positive definiteness $\|A\| \geq 0 \quad \forall A \in \mathbb{C}^{n\times n}$ &
  $\|A\| = 0 \iff A = 0$

- homogeneity $\|\alpha A\| = |\alpha| \cdot \|A\| \quad \forall \alpha \in \mathbb{C}^n$
  $A \in \mathbb{C}^{n\times n}$

- triangle inequality $\|A+B\| \leq \|A\| + \|B\|$
  $\forall A, B \in \mathbb{C}^{n\times n}$

- submultiplicativity $\|A \cdot B\| \leq \|A\| \cdot \|B\| \quad \forall A, B \in \mathbb{C}^{n\times n}$

Claim:

For every vector norm $\|\cdot\|_*$ on $\mathbb{C}^n$

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

determines a matrix norm on $\mathbb{C}^{n\times n}$

Proof: Positive Definitness: if $A \neq 0$ $\exists x' \in \mathbb{C}^n$ s.t. $\|x'\|=1$ & $Ax' \neq 0$

so $\|A\|_* \geq \|Ax\|_* \geq \|Ax'\|_* > 0$

if $A=0$ $Ax=0 \Rightarrow \|A\|_*=0$

homogeneity: $\forall x$ $\|\alpha Ax\|_* = |\alpha| \|Ax\|_*$

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

Triangle inequality

$$\|A+B\|_* = \max_{\|x\|=1} \|(A+B)x\|_*$$
$$= \max_{\|x\|=1} \|Ax+Bx\|_*$$
$$\leq \max_{\|x\|=1} \|Ax\|_* + \max_{\|x\|=1} \|Bx\|_*$$
$$\leq \max_{\|x\|=1} \|Ax\|_* + \max_{\|y\|=1} \|By\|_*$$
$$= \|A\|_* + \|B\|_*$$

Submultiplicativity

$$\|Ax\| \le \|A\| \, \|x\|$$

$$\|A \cdot B\|_* = \max_{\|x\|=1} \frac{\|ABx\|_*}{\|x\|_*}$$

$$= \max_{\|x\| \neq 0 \; \|Bx\|_* \neq 0} \left(\frac{\|ABx\|_*}{\|Bx\|_*} \cdot \frac{\|Bx\|_*}{\|x\|_*} \right)$$

$$\le \max_{\|Bx\|_* \neq 0} \frac{\|ABx\|_*}{\|Bx\|_*} \max_{\|x\|_* \neq 0} \frac{\|Bx\|_*}{\|x\|_*}$$

$$\le \max_{y \neq 0} \frac{\|Ay\|_*}{\|y\|_*} \cdot \max_{\|x\| \neq 0} \frac{\|Bx\|_*}{\|x\|_*}$$

ce $\|Bx\|_* = 0 \rightsquigarrow \|ABx\|_* = 0$

$$= \|A\|_* \cdot \|B\|_*$$

Examples

1-norm $\|A\|_1 = \max_j \left(\sum_{i=1}^{n} |a_{ij}| \right)$ (Proof) ?

2-norm (spectral norm)

$$\|A\|_2 = \sqrt{\max_j \lambda_j (A^T A)}$$

Frobenius norm

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

Supremenum norm

$$\|A\|_\infty = \max_i \left(\sum_j |a_{ij}| \right)$$

Nuclear norm (Schatten norm)

$$\|A\|_* = \sum_j \sqrt{\lambda_j (A^T A)}$$

Why different norms

- Computation different complexity
- all "equivalent" (to some extent)

$$\frac{1}{\sqrt{n}} \|A\|_F \le \|A\|_2 \le \|A\|_F$$

$$\frac{1}{\sqrt{n}} \|A\|_1 \le \|A\|_2 \le \sqrt{n} \|A\|_1$$

$$\frac{1}{\sqrt{n}} \|A\|_\infty \le \|A\|_2 \le \sqrt{n} \|A\|_\infty$$

Can be good for fixed n (dimension)

Sensitivity of $Ax=b$

<u>Question</u>: Given $Ax=b$, how does the solution x change if we perturb A &/or b

$$(A+\Delta A)(x+\Delta x) = b+\Delta b$$

<u>Goal</u>: Estimate $\frac{\|\Delta x\|}{\|x\|}$ ⇝ vector norm

Sensitivity / condition number

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

Theorem

Let A be an invertible matrix (A^{-1} exists)

① if $\Delta A = 0$ (only care about Δb)

$$\frac{\|\Delta x\|}{\|x\|} \le K(A) \frac{\|\Delta b\|}{\|b\|}$$

② if $\Delta A \neq 0$, let I denote the identity matrix ; $\|A^{-1}\| \|\Delta A\| < 1$

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{K(A)}{1 - K(A)\frac{\|\Delta A\|}{\|A\|}} \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right)$$

Interpretation of ① - know system measurement error

Proof of ①

$Ax = b$ $\qquad$ $A\underbrace{\tilde{x}}_{x+\Delta x} = \tilde{b} = b + \Delta b$

Upper bound

$$A(\tilde{x} - x) = \Delta b$$

$$A(x + \Delta x - x) = \Delta b$$

$$A \Delta x = \Delta b$$

$$\Delta x = A^{-1} \Delta b \Rightarrow \|\Delta x\| \le \|A^{-1}\| \|\Delta b\|$$

Lower bound

$$Ax = b \Rightarrow \|b\| = \|Ax\| \le \|A\| \cdot \|x\|$$

$$\|x\| \ge \frac{\|b\|}{\|A\|} \Leftrightarrow \frac{1}{\|x\|} \le \frac{\|A\|}{\|b\|} \quad (*)$$

Combined

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|\Delta b\|}{\|b\| \,/\, \|A\|}$$

upper bound

lower bound

$$= \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\Delta b\|}{\|b\|}$$

$$= \kappa(A) \frac{\|\Delta b\|}{\|b\|}$$

Proof of ②

$$(A+\Delta A)(x+\Delta x) = b + \Delta b$$

$$Ax + \Delta A x + A\Delta x + \Delta A \Delta x = b + \Delta b$$

we can cancel them since $Ax = b$

$$(\Delta A)x + A\Delta x + (\Delta A)(\Delta x) = \Delta b$$

$$(A+\Delta A)\Delta x = \Delta b - (\Delta A)x$$

→ Assume $A+\Delta A$ is invertible

$$\Delta x = (A+\Delta A)^{-1} \cdot (\Delta b - \Delta A x)$$

→ take the norm

$$\|\Delta x\| \leq \|(A+\Delta A)^{-1}\| \cdot \|\Delta b - (\Delta A)x\|$$

$$\leq \|(A+\Delta A)^{-1}\| \cdot (\|\Delta b\| + \|(\Delta A)x\|)$$

triangle ineq

$$\leq \|(A+\Delta A)^{-1}\| \cdot (\|\Delta b\| + \|(\Delta A)\| \cdot \|x\|)$$

Using ✱ $\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$

$$\Rightarrow \boxed{\frac{\|\Delta x\|}{\|x\|} \leq \|(A+\Delta A)^{-1}\| \cdot \left(\|A\| \cdot \frac{\|\Delta b\|}{\|b\|} + \|\Delta A\|\right)}$$

✱

Express in terms of A & $\kappa(A)$

$$A + \Delta A = A(I + A^{-1}(\Delta A))$$ recall $(XY)^{-1} = Y^{-1}X^{-1}$

$$\|(A+\Delta A)^{-1}\| \le \|(I + A^{-1}(\Delta A))^{-1} \cdot A^{-1}\|$$

$$\le \|(I + A^{-1}(\Delta A))^{-1}\| \cdot \|A^{-1}\| \quad (*)$$

Use series expansion

$$\|(I+A)^{-1}\| \le \frac{1}{1 - \|A\|}$$ (just like for scalars)

$$\|(I + A^{-1}(\Delta A))^{-1}\| \le \frac{1}{1 - \|A^{-1}(\Delta A)\|}$$

$$\le \frac{1}{1 - \|A^{-1}\| \cdot \|\Delta A\|}$$

Putting it together

$$\|(A+\Delta A)^{-1}\| \overset{(*)}{\le} \|(I + A^{-1}(\Delta A))^{-1}\| \, \|A^{-1}\|$$

$$\le \frac{\|A\|^{-1}}{1 - \|A^{-1}\| \cdot \|\Delta A\|}$$

Insert into (*)

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \, \|\Delta A\|} \left(\|A\| \cdot \frac{\|\Delta b\|}{\|b\|} + \|\Delta A\| \right)$$

$$= \frac{\kappa(A)}{1 - \|A^{-1}\| \cdot \|\Delta A\|} \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right)$$

$$= \frac{\kappa(A)}{1 - \kappa(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \cdot \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right)$$

$$\|A^{-1}\| \cdot \|\Delta A\| = \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\Delta A\|}{\|A\|}$$

Example 1

Line example from week 1

$$A_1 = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 0.98 \end{pmatrix}$$

stable unstable

$$\kappa_1(A_1) = \kappa_F(A_1) = \kappa_\infty(A_1) = 2$$

$$\kappa_2(A_1) = 1$$

$$\kappa_1(A_2) = \kappa_\infty(A_2) = 3.96 \cdot 10^4$$

$$\kappa_2(A_2) = 3.94 \cdot 10^4$$

$$\kappa_F(A_2) = 3.92 \cdot 10^4$$

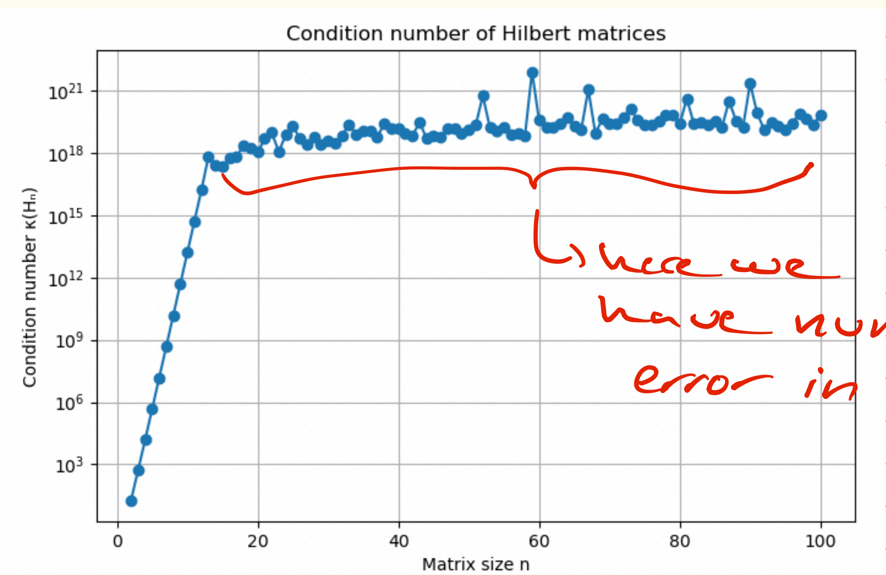
Example 2

Hilbert matrix $\quad H_n = \left(\frac{1}{i+j-1} \right)_{ij} \in \mathbb{R}^{n \times m}$

$(i,j = 1 \dots n)$

$$2\times 2 \quad H_2 = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}$$

(κ_2)



↳ here we have numerical error in calculation

Direct Solution Methods

- Gaussian elimination
- LU factorization
- pivoting (growth of pivots)
- Cholesky decomposition

Gaussian Elimination

Solving square linear systems

n equations, n unknowns

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + \cdots \qquad + a_{2n}x_n &= b_2 \\
\vdots \qquad\qquad\qquad \vdots \quad &\quad \vdots \\
a_{n1}x_1 + \cdots \qquad + a_{nn}x_n &= b_n
\end{aligned}
$$

$$
\underbrace{\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}}_{A}
\underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{x}
=
\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}
$$

If A is invertible $\left\{ \begin{bmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{bmatrix} \right\}_i$ forms a basis of $\mathbb{R}^n$

Look for coefficients such that

$$
\begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix} x_1 + \begin{bmatrix} a_{12} \\ \vdots \\ a_{n2} \end{bmatrix} x_2 + \cdots + \begin{bmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{bmatrix} x_n = \overbrace{\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}}^{\text{given vector}}
$$

coordinates of b in terms of $\left\{ \begin{bmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{bmatrix} \right\}_i$

Cost of solving $Ax=b$

- computational complexity : # of operations
  $\{+, -, \cdot, :\}$
- error analysis (in terms of $K(A)$)

* For which matrices is the problem "easy" (for both complexity & error analysis)

Gaussian Elimination Review

$\omega = 2.21\ldots$

Case 1: compute $A^{-1} \sim O(n^\omega)$ (or $O(n^3)$ practically)

We do not need A^{-1}
→ solve it "half way"

$$A = \begin{bmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{bmatrix} \qquad b = \begin{bmatrix} -1 \\ -7 \\ -6 \end{bmatrix}$$

if we do row operations on A & b
⇒ x does not change

Idea: zero out columns below

$$A = \begin{bmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{bmatrix} \qquad b = \begin{bmatrix} -1 \\ -7 \\ -6 \end{bmatrix}$$

$\Downarrow$ write it this way

$$\left[\begin{array}{ccc|c} -3 & 2 & -1 & -1 \\ 6 & -6 & 7 & -7 \\ 3 & -4 & 4 & -6 \end{array}\right]$$

$$\begin{array}{c} x \\ y \\ \end{array} \left[\begin{array}{ccc|c} -3 & 2 & -1 & -1 \\ 6 & -6 & 7 & -7 \\ 3 & -4 & 4 & -6 \end{array}\right] \longrightarrow \begin{array}{c} x \\ \\ z \end{array} \left[\begin{array}{ccc|c} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 3 & -4 & 4 & -6 \end{array}\right]$$

$y = y + 2x$ $\qquad\qquad$ $z \leftarrow z + x$

$$\left[\begin{array}{ccc|c} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & -2 & 3 & -7 \end{array}\right] \qquad \begin{array}{c} \\ y \\ z \end{array} \left[\begin{array}{ccc|c} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & -2 & 3 & -7 \end{array}\right]$$

$z \leftarrow z - y$

$$\left[\begin{array}{ccc|c} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & 0 & -2 & 2 \end{array}\right] \rightsquigarrow \text{upper triangular}$$

$-2x_3 = 2 \Rightarrow x_3 = -1$

$-2x_2 + 5x_3 = -9 \Rightarrow -2x_2 - 5 = -9 \Rightarrow x_2 = 7$

$-3x_1 + 2x_2 - x_3 = 2 \Rightarrow -3x_1 + 4 + 1 = 2 \Rightarrow x_1 = 1$

Back substitution