Lecture 4

Normal and Hermitian matrices

⊙ Normal matrix:

$A^H A = A A^H$

Lemma ‖ A normal triangular matrix is diagonal

(T) ‖ $A \in \mathbb{C}^n$ is a normal matrix $\Longleftrightarrow$
A is unitarily similar to diagonal matrix
$\exists Q: \quad Q^H A Q = D$ where $D = \mathrm{diag}(\lambda_1, .., \lambda_n)$, $\lambda_i \in \sigma(A)$

Δ $\Longleftarrow$ $Q / Q^H A Q = D$

$A Q = Q D / Q^H$

$A = Q D Q^H$

$A^H A = (Q D Q^H)^H (Q D Q^H) = (Q D^H \underbrace{Q^H)(Q}_{= I} D Q^H) = Q D^2 Q^H$

$A A^H = Q D Q^H (Q D Q^H)^H = Q D \underbrace{Q^H Q}_{= I} D^H Q^H = Q D^2 Q^H$

$\Rightarrow A^H A = A A^H$, A is normal

$\Rightarrow$ A is normal, proove $Q^H A Q = D$

A is normal $\Rightarrow A^H A = A A^H$

Take a Schur form of A: $A = Q R Q^H$, Q is unitary

$A^H A = A A^H$

$(Q R Q^H)^H (Q R Q^H) = (Q R Q^H)(Q R Q^H)^H$

$Q R^H \underbrace{Q^H Q}_{= I} R Q^H = Q R \underbrace{Q^H Q}_{= I} R^H Q^H$

$Q R^H R Q^H = Q R R^H Q^H \quad | \cdot Q$

$Q^H \cdot | \quad Q R^H R = Q R R^H$

$R^H R = R R^H \Rightarrow R$ is normal.

Because R is triangular, from Lemma above it follows that R is diagonal $\triangle$

2) (T) || A is normal $\iff$ $\forall \lambda \in \sigma(A)$ is an eigenvector of A^H

Right eigenvector: x: $Ax = \lambda x$
Left eigenvector x: $x^H A = \lambda x^H$

⊛ Hermitian matrix

A is Hermitian, if $A = A^H$
Any Hermitian matrix is normal
Any normal matrix has a diagonal form:
$\exists Q$: $A = Q D Q^H$, $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_i \in \sigma(A)$

(T1) || 1) Normal matrix with real eigenvalues is Hermitian. 2) Hermitian matrix has real eigenvalues

Δ 2) Take λ: $Au = \lambda u$, $\|u\|_2 = 1$

$$(Au, u) = (\lambda u, u)$$
$$(Au, u) = \lambda$$
$$\lambda = (u, A^H u) = (u, Au) = \overline{(Au, u)} = \overline{\lambda}$$

by definition of inner product

$\lambda = \overline{\lambda} \Rightarrow \lambda \in \mathbb{R}$ Δ

(T2) || Any Hermitian matrix is unitarily similar to a real diagonal matrix
$A = A^H \Rightarrow A = Q D Q^H$, where D is real

(P) || $\forall z \in \mathbb{C}^n$ $(Az, z) \in \mathbb{R} \Rightarrow A = A^H$

Powers of matrices

(T1) || Sequence $\{A^k\} \xrightarrow[k\to\infty]{} 0 \iff \rho(A) < 1$
converges to zero matrix

Spectral radius
$$\rho(A) \overset{\mathrm{def}}{=} \max_{\lambda \in \sigma(A)} |\lambda|$$

T2 || $\lim\limits_{k\to\infty} \|A^k\|^{1/k} = \rho(A)$

T3 || $\sum\limits_{k=0}^{\infty} A^k$ series converges $\Leftrightarrow \rho(A) < 1$

<u>Perturbation analysis</u> анализ возмущений

$Ax = b$, $A \in \mathbb{C}^{n\times n}$

$x = A^{-1}b$ is exact solution

Consider perturbed system

$(A + \delta A)\tilde{x} = b + \delta b$

$\tilde{x}$ is the solution of perturbed system

$\tilde{x}$ is approximate solution to $Ax = b$

$\boxed{\delta x = \tilde{x} - x \text{ is an error}}$

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

Find estimation of error $\|\delta x\|$

Consider initial and perturbed system:

$$(1) \quad Ax = b$$

$$(2) \quad (A+\delta A)(x+\delta x) = b + \delta b$$ Subtract (1) from (2)

$$\cancel{Ax} + A\cdot\delta x + \delta A\cdot x + \delta A\cdot\delta x - \cancel{Ax} = \delta b + \cancel{b} - \cancel{b}$$

$$A\cdot\delta x + \delta A\cdot \underbrace{(x+\delta x)}_{\tilde{x}} = \delta b$$

$$A^{-1}\cdot | \; A\cdot\delta x + \delta A\cdot\tilde{x} = \delta b$$

$$\delta x + A^{-1}\cdot\delta A\cdot\tilde{x} = A^{-1}\delta b$$

$$\delta x = A^{-1}(\delta b - \delta A\cdot\tilde{x})$$

$$\|\delta x\| = \|A^{-1}(\delta b - \delta A\cdot\tilde{x})\| \le \|A^{-1}\|\cdot\|\delta b - \delta A\cdot\tilde{x}\|$$

From the properties of the norms

$\|Ax\|_p \le \|A\|_p\cdot\|x\|_p$; $\|a+b\| \le \|a\| + \|b\|$

4) $\boxed{\|\delta x\| \le \|A^{-1}\| \cdot (\|\delta b\| + \|\delta A \cdot \tilde{x}\|)}$

estimation of absolute error

Let's derive estimation of relative error:

$$\|\delta x\| \le \|A^{-1}\| \cdot \|\delta b\| + \|A^{-1}\| \cdot \|\delta A\| \cdot \|\tilde{x}\| \quad \Big| \cdot \frac{1}{\|\tilde{x}\|}$$

$$\frac{\|\delta x\|}{\|\tilde{x}\|} \le \|A^{-1}\| \cdot \frac{\|\delta b\|}{\|\tilde{x}\|} + \|A^{-1}\| \cdot \|\delta A\| \quad \Big| \cdot \frac{\|A\|}{\|A\|}$$

$$\boxed{\frac{\|\delta x\|}{\|\tilde{x}\|} \le \underbrace{\|A^{-1}\| \cdot \|A\|}_{\substack{\| \text{def} \\ \text{cond}(A)}} \cdot \left(\frac{\|\delta b\|}{\|\tilde{x}\| \cdot \|A\|} + \frac{\|\delta A\|}{\|A\|} \right)}$$

estimation of relative error

(d) condition number of a matrix

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

Condition number is relative to a norm and

(P) $\boxed{1 \le \text{cond}(A) \le \infty}$

Well-conditioned matrix: $\text{cond}(A) \sim 1$

Ill-conditioned matrix: $\text{cond}(A) \sim \infty$

Consider perturbed system with $\delta A = 0$

We have $A\tilde{x} = b + \delta b$

Estimation of absolute error for $\delta A = 0$

$\boxed{\|\delta x\| \le \|A^{-1}\| \cdot \|\delta b\|}$

Now let's derive estimation of relative error for the case $\delta A = 0$

Consider initial and perturbed system:

$$- \uparrow \begin{array}{ll} (1) & Ax = b \\ (2) & A\tilde{x} = b + \delta b \end{array}$$

Subtract (1) from (2)

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

$A^{-1} | \; A \cdot \delta x = \delta b$

$$\delta x = A^{-1} \cdot \delta b$$

$$\|\delta x\| = \|A^{-1} \cdot \delta b\| \le \|A^{-1}\| \cdot \|\delta b\| \quad (1)$$

$$Ax = b \Rightarrow \|Ax\| = \|b\| \Rightarrow \|b\| \le \|A\| \cdot \|x\| \quad (2)$$

Multiply (1) and (2)

(1) $\|\delta x\| \le \|A^{-1}\| \cdot \|\delta b\|$

(2) $\|b\| \le \|A\| \cdot \|x\|$

$$\Downarrow$$

$$\|\delta x\| \cdot \|b\| \le \underbrace{\|A^{-1}\| \cdot \|A\|}_{\text{cond}(A)} \cdot \|\delta b\| \cdot \|x\| \quad \Big| \cdot \frac{1}{\|x\| \cdot \|b\|}$$

$$\boxed{\frac{\|\delta x\|}{\|x\|} \le \text{cond}(A) \cdot \frac{\|\delta b\|}{\|b\|}}$$

Estimation of relative error for the case $\delta A = 0$

Properties of condition number

1) Condition number is relative to the norm:
$\text{cond}_p(A) = \varkappa_p(A) \overset{def}{=} \|A^{-1}\|_p \cdot \|A\|_p$

2) If A is singular matrix, $\text{cond}(A) = +\infty$
If A is nonsingular, $1 \le \text{cond}(A) < \infty$

3) $\text{cond}(\alpha A) = \text{cond}(A)$, $\alpha \ne 0$

4) $\text{cond}_2(A) = 1 \iff A = \alpha Q$ where Q is unitary ($Q^{-1} = Q^H$)
$\|A\|_2 \overset{def}{=} \sqrt{\rho(AA^H)}$

5) For spectral and Eucledian matrix norms
$\text{cond}(Q^H A Q) = \text{cond}(A)$, where Q is unitary

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

7) $\text{cond}(AB) \le \text{cond}(A) \cdot \text{cond}(B)$

8) $A = A^H \Rightarrow \text{cond}(A) = \left| \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)} \right|$, where λ_{max} and λ_{min} are max and min eigenvalues

Comparison of condition number and determinant

Cond(A) is used to estimate whether the problem is well- or ill-conditioned, and det(A) cannot be used for such purpose.

Example

Consider $A = \alpha I$, $A \in \mathbb{C}^{n \times n}$, $\alpha \in \mathbb{C}$

$$\det(A) = \alpha^n$$

For $|\alpha| < 1$ det(A) is small, ~ 0

For $|\alpha| \geq 1$ det(A) is large.

But cond(A) = 1 for $\forall$ norm, i.e. $A = \alpha I$ is a well-conditioned matrix.

Errors and costs

There are three types of numerical errors:

1) Round-off errors, which appear because of finite precision of computer arithmetics

Ex. $\pi \to fl(\pi)$ floating-point representation of π in computer

$\frac{1}{3} \to fl(\frac{1}{3})$ 0.333...

2) Algorithmic errors, or truncation errors, which appear because applied algorithm is inexact.

Ex. Expanction in series for a function

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k$$ Taylor Series

For $a=0$ are called Maclarin series

The function $f(x)$ is infinitely differentiable at a, but in computer:

$$f(x) \simeq \sum_{k=0}^{10} \cdots$$

3) Propagated errors, which are the errors that spread during computations

Ex. Input data $x \to fl(x)$ in computer

$fl(x) = x(1+\varepsilon)$, $|\varepsilon| \leq \varepsilon_c$ - machine ε precision

Estimation of errors for the solution of the problem

x is input data

$R(x)$ is the problem solution

Consider the problem: find solution to a linear system $Az = b$, $A \in \mathbb{C}^{n\times n}$, $b \in \mathbb{C}^n$

In this problem,

$x = A, b$ is input data

$R(x) = A^{-1}b$ is solution

Now let's consider the following:

- x is input data
- $x + \delta x$ is real input data in computer (δx is input error)

8) • $R(x)$ is exact solution for exact input data

• $R_A(x)$ is solution by algorithm (A) for exact input data

• $R(x+\delta x)$ is exact solution for real input data

• $R_A(x+\delta x)$ is solution by algorithm A for real input data.

• Absolute algorithmic error

$\|R_A(x) - R(x)\|$

• Relative algorithmic error

$\dfrac{\|R_A(x) - R(x)\|}{\|R(x)\|}$

• Absolute propagated error

$\|R(x+\delta x) - R(x)\|$

• Relative propagated error

$\dfrac{\|R(x+\delta x) - R(x)\|}{\|R x)\|}$

Note that in practice δx is unavoidable, and all sorts of errors are combined in computer. In computer, we cannot find $R(x)$, but only $R_A(x+\delta x)$.

Let's estimate an error of finding $R_A(x+\delta x)$ instead of $R(x)$.

$$\|R_A(x+\delta x) - R(x)\| = \|R_A(x+\delta x) + R(x+\delta x) - R(x+\delta x) - R(x)\| \le \underbrace{\|R_A(x+\delta x) - R(x+\delta x)\|}_{\text{algorithmic error (1)}} + \underbrace{\|R(x+\delta x) - R(x)\|}_{\text{absolute propagated error (2)}}$$

Absolute propagated error (2) shows how sensitive the problem is to initial data

- <u>well-conditioned problem</u> $x \to R(x)$ $\boxed{\text{хорошо обусловленная}}$

$$\|\delta x\| < \varepsilon \Rightarrow \|R(x+\delta x) - R(x)\| < \varepsilon$$

if small change of δx leads to a small change of absolute propagated error

- <u>ill-conditioned problem</u> (otherwise) $\boxed{\text{плохо обусловленная}}$

$$\|\delta x\| < \varepsilon \Rightarrow \|R(x+\delta x) - R(x)\| > \varepsilon$$

large change of absolute propagated error for small change of δx

- <u>stable algorithm</u> (устойчивый алгоритм) for a well-conditioned problem

$$\|\delta x\| < \varepsilon \Rightarrow \underbrace{\|R_A(x+\delta x) - R_A(x)\|}_{\text{propagated error for solution by algorithm}} < \varepsilon$$

- <u>unstable algorithm</u> (otherwise) неустойчивый алгоритм

<u>Dangerous situation</u>: apply a stable algorithm to an ill-conditioned problem, because $\|R_A(x+\delta x) - R_A(x)\|$ can be small, but such answer is wrong.

<u>Computational costs</u>

- <u>flops</u> is the number of floating-point operations per second.

flops is used to measure computer performance. The aim is to choouse an algorithm that require min flops

10) Ex: matrix-by-vector multiplication

1) $Ax+y$, $A \in \mathbb{R}^{m \times n}$, $x, y \in \mathbb{R}^n$

$2mn$ flops

2) $lu(A)$, $A \in \mathbb{R}^{n \times n}$

$\frac{2}{3}n^3$ flops

ⓓ Costs of algorithm are $O(f(n))$ flops, where n is the size of the problem

(P) $O(f(n)) + O(g(n)) = O(\max\{f(n), g(n)\})$

$O(f(n) \cdot g(n)) = O(f(n)) \cdot O(g(n))$

Example. If part 1 of algorithm is $O(n^2)$ flops and part 2 is $O(n)$ flops, the total sum is $O(n^2)$ flops.