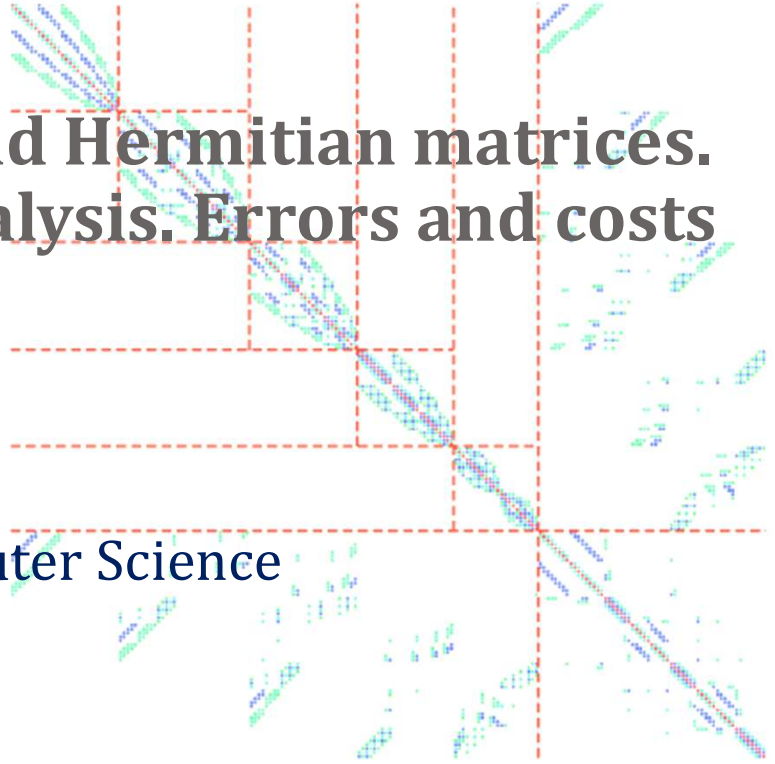
Numerical Methods of Linear Algebra for Sparse Matrices

Lecture 6

Positive definite matrices. Normal and Hermitian matrices. Powers of matrices. Perturbation analysis. Errors and costs



Anna Nasedkina, PhD, Assoc. prof.
Department of Mathematical Modeling
Institute of Mathematics, Mechanics and Computer Science
Southern Federal University

Outline

- Positive definite matrices
- Normal and Hermitian matrices
- Powers of matrices
- Perturbation analysis and condition number
- Errors and costs

Positive definite matrices

- Real positive definite matrix $A \in \mathbb{R}^{n \times n}$

$(Au,u) > 0 \ \forall u \in \mathbb{R}^n, u \neq 0$. Recall that $(Au,u) = u^H Au$

- Real positive semidefinite matrix $A \in \mathbb{R}^{n \times n}$

$(Au,u) \geq 0 \ \forall u \in \mathbb{R}^n$

- Real symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$, $A = A^T$

$(Au,u) > 0 \ \forall u \in \mathbb{R}^n, u \neq 0$

Complex matrix can be positive definite only in the case, when it is Hermitian

(Inner product (Au,u) can be real only when $A = A^H$)

- Hermitian positive definite matrix $A \in \mathbb{C}^{n \times n}$, $A = A^H$

$(Au,u) > 0 \ \forall u \in \mathbb{C}^n, u \neq 0$

- Hermitian positive semidefinite matrix $A \in \mathbb{C}^{n \times n}$, $A = A^H$

$(Au,u) \geq 0 \ \forall u \in \mathbb{C}^n$

Positive definite matrices: theorems

Theorem 1. Complex $A \in \mathbb{C}^{n\times n}$ is positive definite $\Leftrightarrow$ $A = A^H$ and the spectrum $\sigma(A) > 0$ (i.e. all eigenvalues are positive).

Theorem 2. If $A \in \mathbb{R}^{n\times n}$ is real positive definite, then 1) $\exists\, A^{-1}$;

2) $\exists\ \alpha > 0 :\ (Au,u) \geq \alpha \|u\|_2^2\ \ \forall u \in \mathbb{R}^n$

Theorem 3. $\forall A \in \mathbb{C}^{n\times m}$ $A^H A$ is Hermitian positive semidefinite.

$\triangle\ (A^H Au, u) = (Au, Au) \geq 0\ \ \forall u \in \mathbb{C}^m\ \square$

Theorem 4. $\forall A \in \mathbb{C}^{n\times n}$ there is decomposition of A: $A = H + iS,$ where

$H = \frac{1}{2}(A + A^H),\ S = \frac{1}{2i}(A - A^H);\ H,\ S$ are Hermitian, iS is skew-Hermitian

For eigenvalues $\lambda_i,\ i = \overline{1,n}$ of A holds:

1) $\lambda_{\min}(H) \leq \mathrm{Re}(\lambda_i) \leq \lambda_{\max}(H)$

2) $\lambda_{\min}(S) \leq \mathrm{Im}(\lambda_i) \leq \lambda_{\max}(S)$

$H = \frac{1}{2}(A + A^H)$ is called Hermitian (symmetric in real case) part of A

If $A \in \mathbb{R}^{n\times n},\ u \in \mathbb{R}^n$ then $(Au,u) = (Hu,u)$

The definition of a real positive definite matrix

$(Au,u) > 0\ \forall u \in \mathbb{R}^n, u \neq 0\ \Leftrightarrow$ symmetric part of A is positive definite.

Positive definite matrices: important notes

- Check positive definiteness in Matlab by Cholesky factorization: **chol(A)** or by computing eigenvalues **eig(A)**

Example. Real matrix can be positive definite but nonsymmetric.

Take block-diagonal $A = \begin{pmatrix} 3 & 2 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 2 & 2 \end{pmatrix}$.

Its eigenvalues are $\{2, 5, 2, 2\}$ are all positive.

This matrix is positive definite, but $A \neq A^T$

```
A=rand(3)
```

```
A = 3×3
    0.4942    0.9037    0.6987
    0.7791    0.8909    0.1978
    0.7150    0.3342    0.0305
```

```
chol(A)
```

```
Error using chol
Matrix must be positive definite.
```

```
[R,flag] = chol(A) % flag>0
```

```
R = 0.7030
flag = 2
```

```
eig(A)
```

```
ans = 3×1
    1.7988
   -0.5585
    0.1753
```

Normal matrices: theorems

- Normal matrix $A \in \mathbb{C}^{n\times n}$

$A^H A = AA^H$

Lemma. A normal triangular matrix is diagonal.

Theorem 1. Any normal matrix has a diagonal form :

$A \in \mathbb{C}^{n\times n}$ is normal matrix $\Leftrightarrow$ A is unitarily similar to diagonal matrix:

$\exists Q:\ A = QDQ^H,$ where $D = diag(\lambda_1,...,\lambda_n),\ \lambda_i \in \sigma(A)$

Theorem 2. $A \in \mathbb{C}^{n\times n}$ is normal $\Leftrightarrow$ any eigenvalue $\lambda \in \sigma(A)$ is an eigenvalue of A^H $(\lambda \in \sigma(A^H))$

Hermitian matrices: theorems

- Hermitian matrix $A \in \mathbb{C}^{n\times n}$: $A = A^H$

Proposition. If inner product (Az, z) is real for $\forall z \in \mathbb{C}^n$, then

$A = A^H$ (A is Hermitian)

Lemma. Any Hermitian matrix is normal.

Theorem 1. 1) Normal matrix with real eigenvalues is Hermitian.

2) Hermitian matrix has real eigenvalues.

Theorem 2. Any Hermitian matrix is unitarily similar to

a real diagonal matrix: $A = A^H \Rightarrow A = QDQ^H$, where D is real.

Theorem 3. $\forall A \in \mathbb{C}^{n\times n}$ there is decomposition of A: $A = H + iS$, where

$H = \frac{1}{2}(A + A^H)$, $S = \frac{1}{2i}(A - A^H)$; H, S are Hermitian, iS is skew-Hermitian

For eigenvalues λ_i, $i = \overline{1,n}$ of A holds:

1) $\lambda_{\min}(H) \le \mathrm{Re}(\lambda_i) \le \lambda_{\max}(H)$

2) $\lambda_{\min}(S) \le \mathrm{Im}(\lambda_i) \le \lambda_{\max}(S)$

Powers of matrices

Theorem 1. Sequence of matrix powers $\{A^k\}$ converges to zero matrix:

$\{A^k\} \to 0 \;\Leftrightarrow\; \rho(A) < 1.$

Recall that $\rho(A)$ is the spectral radius of A: $\rho(A) = \max\limits_{\lambda \in \sigma(A)} |\lambda|$.

Theorem 2. $\lim\limits_{k\to\infty} \|A^k\|^{1/k} = \rho(A).$

Theorem 3. Series of matrix powers $\sum\limits_{k=0}^{\infty} A^k$ converges $\;\Leftrightarrow\; \rho(A) < 1.$

Perturbation analysis

Consider a linear system $Ax = b,\ A \in \mathbb{C}^{n\times n},\ b \in \mathbb{C}^n$

$x \in \mathbb{C}^n$ is the vector of unknowns

$x = A^{-1}b$ will be *exact* solution

Consider a pertubed system $(A + \delta A)\tilde{x} = b + \delta b$

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

$\tilde{x}$ will be *approximate* solution to initial system $Ax = b$

$\delta x = \tilde{x} - x$ is an *error,* hence $\tilde{x} = x + \delta x$

Let us find estimation of the error $\|\delta x\|$

Estimation of absolute error

Consider initial and pertubed systems:

1) $Ax = b$

2) $(A + \delta A)(x + \delta x) = b + \delta b$

$Ax + A \cdot \delta x + \delta A \cdot (x + \delta x) = b + \delta b$

Subtract (1) from (2):

$Ax + A \cdot \delta x + \delta A \cdot (x + \delta x) - Ax = b + \delta b - b,\ x + \delta x = \tilde{x}$

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

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

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

$\|\delta x\| = \left\|A^{-1} \cdot (\delta b - \delta A \cdot \tilde{x})\right\| \le \left\|A^{-1}\right\| \cdot \left\|\delta b - \delta A \cdot \tilde{x}\right\| \le \left\|A^{-1}\right\| \cdot \left(\|\delta b\| + \left\|\delta A \cdot \tilde{x}\right\|\right)$, that follows from the properties of the norm: $\|Ax\|_p \le \|A\|_p \cdot \|x\|_p$, $\|a + b\| \le \|a\| + \|b\|$

Hence $\|\delta x\| \le \left\|A^{-1}\right\| \cdot \left(\|\delta b\| + \left\|\delta A \cdot \tilde{x}\right\|\right)$ is the estimation of absolute error

Estimation of relative error

From the estimation of absolute error $\|\delta x\| \le \|A^{-1}\| \cdot \left(\|\delta b\| + \|\delta A \cdot \tilde{x}\| \right)$ let us derive the estimation of relative error

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

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

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

$$\frac{\|\delta x\|}{\|\tilde{x}\|} \le cond(A) \cdot \left(\frac{\|\delta b\|}{\|\tilde{x}\| \cdot \|A\|} + \frac{\|\delta A\|}{\|A\|} \right)$$ is the estimation of relative error

where $cond(A) = \|A^{-1}\| \cdot \|A\|$ is the condition number of the matrix

Estimation of relative error

$$\frac{\|\delta x\|}{\|\tilde{x}\|} \le cond(A) \cdot \left(\frac{\|\delta b\|}{\|\tilde{x}\| \cdot \|A\|} + \frac{\|\delta A\|}{\|A\|} \right)$$

When $\delta A = 0$

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

$\frac{\|\delta x\|}{\|\tilde{x}\|} \le cond(A) \cdot \frac{\|\delta b\|}{\|b\|}$ is the estimation of relative error with respect to *approximate* solution.

In a similar way it can be derived that

$\frac{\|\delta x\|}{\|x\|} \le cond(A) \cdot \frac{\|\delta b\|}{\|b\|}$ is the estimation of relative error with respect to *exact* solution

Condition number

- Condition number of the matrix $A \in \mathbb{C}^{n\times n}$

$$cond(A) = \left\|A^{-1}\right\| \cdot \|A\|$$

Condition number is relative to the matrix norm $cond_p(A) = \left\|A^{-1}\right\|_p \cdot \|A\|_p$

Properties of condition number

1) $1 \le cond(A) \le +\infty$

2) For singular matrix $cond(A) = +\infty$, otherwise $1 \le cond(A) < +\infty$

3) $cond(\alpha A) = cond(A)$ for $\forall \alpha \neq 0,\ \alpha \in \mathbb{C}$

4) For spectral and Eucledian matrix norms $cond(Q^H AQ) = cond(A)$, where Q is unitary

5) $cond(A^{-1}) = cond(A)$

6) $cond(AB) \le cond(A) \cdot cond(B)$

7) $A = A^H \Rightarrow cond(A) = \left|\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}\right|$, where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are maximal and minimal eigenvalues of A

Condition number and determinant

Condition number is used to estimate whether the problem is well- or ill-conditioned, and determinant cannot be used for such purpose.

- well-conditioned matrix

$cond(A) \sim 1$

- ill-conditioned matrix

$cond(A) \sim +\infty$

Example

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

It is easy to show that $\det(A) = \alpha^n$

For $|\alpha| < 1$ $\det(A)$ is small, but for $|\alpha| \geq 1$ $\det(A)$ is large.

However, $cond(A) = 1$ for any norm,

thus $A = \alpha I$ is a well-conditioned matrix.

Numerical errors: three types

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

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

$\frac{1}{3} \to fl(\frac{1}{3})$ floating-point representation 0.33333...

2) Algorithmic errors, or trancation errors, which appear because the applied algorithm is not exact

Example. Consider a series expansion, such as Teylor series

$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k$. Suppose that the function $f(x)$ is infinitely differentiable at point a, but in a computer $f(x) \simeq \sum_{k=0}^{100} \frac{f^{(k)}(a)}{k!}(x-a)^k$

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

Example. Input data x is stored as $fl(x)$ in a computer

$fl(x) = x(1+\varepsilon)$, where $|\varepsilon| \le \varepsilon_c$ (machine precision epsilon)

Numerical errors and solution

Let X be an input data, $R(X)$ is the problem solution.

Consider the problem of finding the solution of a linear system

$Mx = b,\ M \in \mathbb{C}^{n\times n},\ b \in \mathbb{C}^n$

Here $X = M,\ b$ is the input data, $R(x) = M^{-1}b$ is the solution.

Now let us consider real input data, stored in computer $X + \delta X,$ as there can be input error δX. Suppose that we apply the algorithm A to solve the problem.

In reality we can deal with different solutions:

- $R(x)$ is the exact solution for exact input data X
- $R_A(x)$ is the solution by algorithm A for exact input data X
- $R(x+\delta x)$ is the exact solution for real input data $X+\delta X$
- $R_A(x+\delta x)$ is the solution by algorithm A for real input data $X+\delta X$

Estimation of numerical errors

- $\|R_A(x) - R(x)\|$ is the absolute algorithmic error for exact solution or
- $\|R_A(x+\delta x) - R(x+\delta x)\|$ is the absolute algorithmic error for approximate solution
- $\dfrac{\|R_A(x) - R(x)\|}{\|R(x)\|}$ is the relative algorithmic error for exact solution
- $\|R(x+\delta x) - R(x)\|$ is the absolute propagated error or
- $\|R_A(x+\delta x) - R_A(x)\|$ is the absolute propagated error for solution by algorithm
- $\dfrac{\|R(x+\delta x) - R(x)\|}{\|R(x)\|}$ is the relative propagated error

Note that in practice δx is unavoidable, and all sorts of errors are combined.

In computer we cannot find $R(x)$, but only $R_A(x+\delta x)$.

Error of finding $R_A(x+\delta x)$ instead of $R(x)$ will be:

$$\|R_A(x+\delta x) - R(x)\| = \|R_A(x+\delta x) + R(x+\delta x) - R(x+\delta x) - R(x)\| \le$$

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

Conditioning of the problem and algorithm stability

Absolute propagated error shows how sensitive the problem is to small changes or errors in input data.

- well-conditioned problem: small error in input data leads to a small absolute propagated error $\|\delta x\| < \varepsilon \Rightarrow \|R(x+\delta x) - R(x)\| < \varepsilon$
- ill-conditioned problem: small error in input data leads to a large absolute propagated error $\|\delta x\| < \varepsilon \Rightarrow \|R(x+\delta x) - R(x)\| > \varepsilon$
- stable algorithm for a well-conditioned problem: small error in input data leads to a small absolute propagated error for solution obtained by algorithm $\|\delta x\| < \varepsilon \Rightarrow \|R_A(x+\delta x) - R_A(x)\| < \varepsilon$
- unstable algorithm: otherwise

Dangerous situation is to apply a stable algorithm to an ill-conditioned problem, because the error $\|R_A(x+\delta x) - R_A(x)\|$ can be small, but such answer will be wrong!

Computational costs

- flops (floating-point operations per second) is a measure of computer performance, indicating the number of operations a processor can perform per second

Examples

1) $Ax + y,\ A \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^n$ requires $2mn$ flops.

2) $lu(A),\ A \in \mathbb{R}^{n\times n}$ requires $\frac{2}{3}n^3$ flops.

The aim is to choose an algorithm that requires minimal flops.

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

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

Properties.

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

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