2.6 Numerical Linear Algebra

For $\underline{\underline{A}}\,\underline{x} = \underline{b}$,

Structure of $\underline{\underline{A}}$:

- full: almost all $\{a_{ij}\}$ are nonzero.
- sparse: almost all $\{a_{ij}\}$ are zero. sparse if more than 90% are zero.

Methods :

- direct: finite (predetermined) number of operation gives answer.
- iterative: method converges asymptotically.
  As $n \to \infty$, solution converges.

Error :

- direct: precision error (machine accuracy)
- iterative: convergence error (depending on algorithm) + precision error

How to decide method :

1. Robustness (stability, convergence)
2. Storage requirement
3. Computational work

Computational work : operation count

- solution of upper triangular matrix

$$\underline{\underline{U}}\,\underline{x} = \underline{b}$$

$$\begin{array}{cccccc} U_{11}x_1+ & U_{12}x_2+ & \cdots+ & U_{1n}x_n & = & b_1 \\ & U_{22}x_2+ & \cdots+ & U_{2n}x_n & = & b_2 \\ & & & \vdots & = & \vdots \\ & & U_{n-1,n-1}x_{n-1}+ & U_{n-1,n}x_n & = & b_{n-1} \\ & & & U_{n,n}x_n & = & b_n \end{array}$$

Back-substitution:

$$x_i = \frac{b_i - \sum_{k=i+1}^{n} U_{ik} x_k}{U_{ii}}, \; i = n, n-1, \ldots, 1$$

For each i,

1	division
$n-i$	addition
1	subtraction
$n-i$	multiplication

$$\begin{aligned} \text{work} &= n + \sum_{i=1}^{n} (n-i) \\ &= n + n^2 - \frac{n(n+1)}{2} \\ &\propto \frac{n^2}{2} \sim \mathcal{O}(n^2) \end{aligned}$$

- Multiplication of two matrices

  For $\underline{\underline{C}} = \underline{\underline{A}}\,\underline{\underline{B}}$

$$\begin{aligned} & i, j \\ & s = 0 \\ & \left(\begin{aligned} & k = 1, \ldots, n \\ & s = s + a_{ik} b_{kj} \end{aligned} \right. \\ & c_{ij} = s \end{aligned}$$

  FLOP = floating-point operations

  $s = s + a_{ik} b_{kj}$: $\mathcal{O}(n^3)$ FLOPS

- Gaussian elimination

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}$$

$$\underline{\underline{J}}_{21}\left(-\frac{a_{21}}{a_{11}}\right) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{aligned} & \underline{\underline{J}}_{41}\left(-\frac{a_{41}}{a_{11}}\right)\underline{\underline{J}}_{31}\left(-\frac{a_{31}}{a_{11}}\right)\underline{\underline{J}}_{21}\left(-\frac{a_{21}}{a_{11}}\right) \\ = \ & \underline{\underline{A}}^{(2)} \\ = \ & \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} & a_{14}^{(2)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & a_{42}^{(2)} & a_{43}^{(2)} & a_{44}^{(2)} \end{bmatrix} \end{aligned}$$

where

$$a_{22}^{(2)} = a_{22} - \frac{a_{12}}{a_{11}} a_{12}$$

Number of operations

$\mathcal{O}(n^3)$ for matrix multiplication

$\frac{1}{2}n^2$ times matrix multiplication

$\rightarrow$ total transformation for GE needs $\sim \frac{1}{2}n^5$.

Cost of sparse matrix multiplication $\underline{\underline{J}}_{ij}\underline{\underline{A}}$ <=> $\mathcal{O}(n)$

Cost of Gaussian elimination <=> $\mathcal{O}(n^3)$

Back-substitution <=> $\mathcal{O}(n^2)$: insignificant to cost of GE

Linear equation solver :

1. Gaussian elimination

   $$\underline{\underline{A}}\,\underline{x} = \underline{b} \rightarrow \underline{\underline{U}}\,\underline{x} = \hat{\underline{b}}$$

2. LU-decomposition

   $$\underline{\underline{A}} \rightarrow \underline{\underline{L}}\,\underline{\underline{U}}$$

Two problems :

1. $\underline{\underline{A}}_i\underline{x} = \underline{b}_i$ where $i = 1, \ldots, m$.
   Both type 1 and 2 are OK.

2. $\underline{\underline{A}}_o \underline{x} = \underline{b}_i$
   LU-decomposition is superior.

   $$\underline{\underline{L}}\,\underline{\underline{U}}\,\underline{x} = \underline{b}_i$$

   (a) $\underline{\underline{L}}\,\underline{z} = \underline{b} \rightarrow$ solve lower triangular set.

   (b) $\underline{\underline{U}}\,\underline{x} = \underline{z} \rightarrow$ solve upper triangular set $\rightarrow \underline{x}$.

Theorem Let $\underline{\underline{A}} \in \Re^{n\times n}$ and let $\underline{\underline{A}}_k$ be $\Re^{k\times k}$ matrix formed by the intersection of first k rows and k columns of $\underline{\underline{A}}$. If $\det(\underline{\underline{A}}_k) \neq 0, k = 1, \ldots, n-1$, then a unique $\underline{\underline{L}}$ exists with $L_{ij} = m_{ij}$ and $m_{ii} = 1$ and a unique $\underline{\underline{U}}$ exists with $U_{ij} = u_{ij}$.

Proof Suppose the theorem holds for $n = k-1$.

$$\underline{\underline{A}}_k = \begin{bmatrix} \underline{\underline{A}}_{k-1} & \underline{b} \\ \underline{c}^T & a_{kk} \end{bmatrix} \in \Re^{k\times k}$$

$$\underline{\underline{L}}_k = \begin{bmatrix} \underline{\underline{L}}_{k-1} & \underline{0} \\ \underline{m}^T & 1 \end{bmatrix}$$

$$\underline{\underline{U}}_k = \begin{bmatrix} \underline{\underline{U}}_{k-1} & \underline{u} \\ \underline{0}^T & u_{kk} \end{bmatrix}$$

Find $\underline{m}, \underline{u}$, and u_{kk}.

$$\underline{\underline{L}}_k \underline{\underline{U}}_k = \begin{bmatrix} \underline{\underline{L}}_{k-1}\underline{\underline{U}}_{k-1} & \underline{\underline{L}}_{k-1}\underline{u} \\ \underline{m}^T \underline{\underline{U}}_{k-1} & \underline{m}^T \underline{u} + u_{kk} \end{bmatrix} = \underline{\underline{A}}_k$$

Equate

1. $\underline{\underline{A}}_{k-1} = \underline{\underline{L}}_{k-1}\underline{\underline{U}}_{k-1}$: OK
2. $\underline{b} = \underline{\underline{L}}_{k-1}\underline{u} \rightarrow$ solve for $\underline{u}$.
3. $\underline{c}^T = \underline{m}^T \underline{\underline{U}}_{k-1} \rightarrow$ solve for $\underline{m}$.
4. $\underline{m}^T \underline{u} + u_{kk} = a_{kk} \rightarrow$ solve for u_{kk}.

Pivoting When does GE breaks down? $a_{kk}^{(k)} = 0$

Pivoting: To prevent problems with zero pivot. Search the column and move the row with largest figure top in the unfinished part during GE.

- Partial pivoting

$$|a_{rk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|$$

- Full pivoting

$$|a_{rk}^{(k)}| = \max_{\substack{k \le i \le n \\ k \le j \le n}} |a_{ij}^{(k)}|$$

If matrix is singular, zero pivot moves to $a_{nn}^{(n)}$ position.

When we don't have to pivot?

1. Diagonally dominant matrix

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

2. Symmetric and positive definite matrix

$$\underline{\underline{A}}^T = \underline{\underline{A}}, \quad \underline{x}^T \underline{\underline{A}}\, \underline{x} > 0 \ \forall \underline{x} \in \Re^n$$

Theorem A diagonally dominant matrix $\underline{\underline{A}}$ satisfies

1. Each princial minor of $\underline{\underline{A}}$ is diagonally dominant
2. $\underline{\underline{A}}$ is non-singular.

Sparse matrix :

- Banded structure
- Unstructured

Pivoting of banded matrix

Partial pivoting is OK

Full pivoting is not OK (It destroys the band structure).

2.7 Error Analysis for Linear Systems

Residual vector Residual vector $\underline{r} = \underline{b} - \underline{\underline{A}}\,\underline{x}$

When $\underline{r} = 0 \rightarrow \underline{\underline{A}}\,\underline{x} = \underline{b}$

When $\|\underline{r}\| \ll 1 \rightarrow \underline{\underline{A}}\,\underline{x} \cong \underline{b}$

Example For the following matrix $\underline{\underline{A}}$ and vector $\underline{b}$

$$\underline{\underline{A}} = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 0.8642 \\ 0.1440 \end{bmatrix}$$

The solution and residual vector are

$$\underline{x} = \begin{bmatrix} 0.9911 \\ -0.4870 \end{bmatrix}, \quad \underline{r} = \begin{bmatrix} 10^{-8} \\ -10^{-8} \end{bmatrix}$$

It looks reasonable. But, the exact solution is

$$\underline{x}_{\text{exact}} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

During Gaussian elimination

$$\left[\begin{array}{cc|c} 1.2969 & 0.8648 & 0.8642 \\ 0.2161 & 0.1441 & 0.1440 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{cc|c} 1.2969 & 0.8648 & 0.8642 \\ 0 & 10^{-8} & -2 \times 10^{-8} \end{array}\right]$$

Perturbation analysis for $\underline{\underline{A}}\,\underline{x} = \underline{b}$

Consider

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

Then

$$\underline{\underline{A}}\delta\underline{x} = \delta\underline{b} \rightarrow \delta\underline{x} = \underline{\underline{A}}^{-1}\delta\underline{b} \rightarrow \|\delta\underline{x}\| \leq \|\underline{\underline{A}}^{-1}\|\|\delta\underline{b}\|$$

$$\underline{\underline{A}}\,\underline{x} = \underline{b} \rightarrow \|\underline{b}\| \leq \|\underline{\underline{A}}\|\|\underline{x}\| \rightarrow \frac{1}{\|\underline{x}\|} \leq \frac{\underline{\underline{A}}}{\|\underline{b}\|}$$

Then,

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \left[\|\underline{\underline{A}}\|\|\underline{\underline{A}}^{-1}\|\right]\frac{\|\delta\underline{b}\|}{\|\underline{b}\|}$$

$[\|\underline{\underline{A}}\|\|\underline{\underline{A}}^{-1}\|]$ is the condition number and bounds relative error in $\underline{x}$ wrt relative error in $\underline{b}$. Condition number has the following relation.

$$\kappa(\underline{\underline{A}}) = \|\underline{\underline{A}}\|\|\underline{\underline{A}}^{-1}\| \sim \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}$$

Example For the example above

$$\underline{\underline{A}}^{-1} = 10^8 \begin{bmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{bmatrix}$$

$$\|\underline{\underline{A}}\|_\infty = \max_{i=1,2} \sum_{j=1}^{2} |a_{ij}| = 2.1617$$

$$\|\underline{\underline{A}}^{-1}\|_\infty = 1.5130 \times 10^8$$

Then,

$$\kappa(\underline{\underline{A}}) = 3 \times 10^8$$

For your reference, $\lambda_{\text{max}} = 1.4410, \lambda_{\text{min}} =\sim 10^{-8}$.