Topic and contents

UNSW, School of Mathematics and Statistics

MATH2089 – Numerical Methods

Week 03 – Linear Systems, Norms, LU Factorization

- MATLAB M-files
  - `vecnorms.m`
  - `errex1.m`
  - `matnorms.m`
  - `luex1.m`
  - `luex2.m`
  - `LUsolvers.m`

Systems of linear equations

Systems of linear equations (or linear systems) arise in

- statistics (linear regression, least squares approximation)
- solving partial differential equations numerically in civil/mechanical engineering problems
- signal processing, electrical engineering (electrical networks), chemical engineering (balancing chemical reactions) etc.

Calculations are done on computers using floating point arithmetic

- Accuracy: Effects of finite precision
- Efficiency: Time and storage

Systems of Linear Equations

- System $A\mathbf{x} = \mathbf{b}$, $\quad A \in \mathbb{R}^{m\times n}$, $\quad \mathbf{x} \in \mathbb{R}^n$, $\quad \mathbf{b} \in \mathbb{R}^m$

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}$$

  - m linear equations in n variables
  - each row is one of m equations
  - all m equations must be satisfied simultaneously
- Exploit structure of A to solve in ways which are
  - numerically stable – limit effects of errors in the data
  - efficient – time (flops) and memory
- Systems are
  - Well-determined: $m = n$, same number equations, variables
  - Over-determined: $m > n$, more equations than variables
  - Under-determined: $m < n$, fewer equations than variables

A motivational example

Example (Good/bad matrices)

Let

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 + 10^{-10} \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 2 \\ 2 & 4 + 10^{-10} \end{bmatrix},$$

In solving $A\mathbf{x}_1 = \mathbf{b}$ and $B\mathbf{x}_2 = \mathbf{b}$, which of the two computed solutions $\mathbf{x}_1$ and $\mathbf{x}_2$ is more accurate?

In order to answer this question properly, we need to introduce norms of vectors and matrices.

Vector Norms

Measuring the magnitude of a quantity

- $\alpha \in \mathbb{R}$, magnitude $|\alpha| = \begin{cases} \alpha & \text{if } \alpha \geq 0; \\ -\alpha & \text{if } \alpha < 0. \end{cases}$
- $\mathbf{x} \in \mathbb{R}^n$, different measures of magnitude $\|\mathbf{x}\|$

Definition (Vector norm)

Vector norm on $\mathbb{R}^n$ is a function $\|\cdot\|$ from $\mathbb{R}^n$ to $\mathbb{R}$ satisfying

1. $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$
2. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (Triangle inequality)
3. $\|\alpha \mathbf{x}\| = |\alpha| \, \|\mathbf{x}\|$ for all $\alpha \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^n$

- p-norms defined by, for $p \geq 1$,

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

Vector p-norms

- 1-norm
$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$
- 2-norm
$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}$$
- ∞-norm or maximum norm
$$\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,n} |x_i|$$

- $\|\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{x}$, $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$ (cf. dot product $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i$)
- MATLAB function `norm`, see `vecnorms.m`
  - 2-norm `norm(x) = sqrt(x'*x)` (default)
  - 1-norm `norm(x,1) = sum(abs(x))`
  - ∞-norm `norm(x,Inf) = max(abs(x))`

Vector norms – examples

Example (Vector norms)

Find the 1, 2 and infinity norms of the vector

$$\mathbf{v} = (-1, 2, -3, 2)^T$$

Solution

- *1-norm*
$$\|\mathbf{v}\|_1 = \sum_{i=1}^{4} |v_i| = 1 + 2 + 3 + 2 = 8.$$
- *2-norm*
$$\|\mathbf{v}\|_2^2 = \sum_{i=1}^{4} |v_i|^2 = (-1)^2 + 2^2 + (-3)^2 + 2^2 = 18 \Longrightarrow \|\mathbf{v}\|_2 = \sqrt{18}.$$
- *infinity norm*
$$\|\mathbf{v}\|_\infty = \max_{i=1,2,3,4} |v_i| = \max\{1, 2, 3, 2\} = 3.$$

Errors in vectors

- $\bar{\mathbf{x}} \in \mathbb{R}^n$ approximation to $\mathbf{x} \in \mathbb{R}^n$
- Absolute error
$$\|\bar{\mathbf{x}} - \mathbf{x}\|$$
- Relative error for $\mathbf{x} \neq \mathbf{0}$
$$\rho_{\mathbf{x}} = \frac{\|\bar{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|}$$
- Largest component of $\mathbf{x}$ has k significant figures $\iff$
$$\frac{\|\bar{\mathbf{x}} - \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} < 0.5 \times 10^{-k}$$
- Number of significant figures k
$$k \approx -\log_{10}\left(2\frac{\|\bar{\mathbf{x}} - \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}\right)$$

Errors in vectors – example

Example (Errors in vectors)

$\mathbf{x} = (-3.641,\ 0.7843)^T$, approximation $\bar{\mathbf{x}} = (-3.633,\ 0.7915)^T$
Find the absolute and relative errors using the infinity norm, and the estimate number of significant figures

Solution (MATLAB M-file `errex1.m`)

Matrix norms

Definition (Matrix norm)

Matrix norm is a scalar function $\|\cdot\|$ defined on $\mathbb{R}^{m\times n}$ satisfying

1. $\|A\| \geq 0$ for all $A \in \mathbb{R}^{m\times n}$ and $\|A\| = 0 \iff A = 0$.
2. $\|A + B\| \leq \|A\| + \|B\|$ for all $A, B \in \mathbb{R}^{m\times n}$ (Triangle inequality).
3. $\|\alpha A\| = |\alpha|\ \|A\|$ for all $\alpha \in \mathbb{R}$ and $A \in \mathbb{R}^{m\times n}$.

Definition (Consistent matrix norms)

Matrix norms are consistent $\iff$

$$\|AB\| \leq \|A\|\ \|B\| \text{ for all } A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times \ell}.$$

Subordinate matrix norms

- Subordinate matrix norm defined in terms of vector norms

$$\|A\|_p = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

- Common subordinate matrix norms

$$\begin{aligned} \|A\|_1 &\equiv \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \\ \text{matrix 1-norm} &= \text{maximum absolute column sum} \end{aligned}$$

$$\begin{aligned} \|A\|_\infty &\equiv \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}| \\ \text{matrix } \infty\text{-norm} &= \text{maximum absolute row sum.} \end{aligned}$$

- Subordinate matrix norms satisfy for $A \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$

$$\|A\mathbf{x}\|_p \le \|A\|_p \|\mathbf{x}\|_p$$

Eigenvalues and Eigenvectors

Definition (Eigenvalue and eigenvector)

For a square matrix $A \in \mathbb{R}^{n\times n}$, $\lambda \in \mathbb{C}$ is an eigenvalue with corresponding non-zero eigenvector $\mathbf{v} \in \mathbb{C}^n \iff$

$$A\mathbf{v} = \lambda\mathbf{v}$$

- Eigenvalues satisfy the characteristic equation $\det(A - \lambda I) = 0$
- Eigenvalues distinct $\implies$ eigenvectors linearly independent
- Difficulties may arise with multiple eigenvalues
- $A \in \mathbb{R}^{n\times n} \implies$ eigenvalues λ either real or occur in complex conjugate pairs
- For a real symmetric ($A^T = A$) matrix, the eigenvalues are all real, and can choose eigenvalues to form an orthonormal basis for $\mathbb{R}^n$
- MATLAB `eig` calculates eigenvalues and eigenvectors

Matrix 2-norm via eigenvalues of A^TA

Let $A \in \mathbb{R}^{m\times n}$

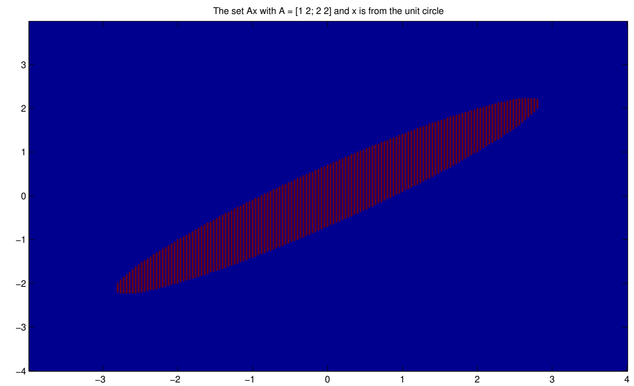
- $A^TA \in \mathbb{R}^{n\times n}$ is symmetric: $(A^TA)^T = A^TA$
  - Uses $(UV)^T = V^TU^T$, $(U^T)^T = U$
- Let $\lambda, \mathbf{v} \neq \mathbf{0}$ be an eigenvalue, eigenvector pair of A^TA

$$A^TA\mathbf{v} = \lambda\mathbf{v} \implies \mathbf{v}^TA^TA\mathbf{v} = \lambda\mathbf{v}^T\mathbf{v} \implies \lambda = \frac{\|A\mathbf{v}\|_2^2}{\|\mathbf{v}\|_2^2} \ge 0$$

  - So eigenvalues of A^TA are real and non-negative
- For $A \in \mathbb{R}^{m\times n}$

$$\|A\|_2 = \max_{\mathbf{v}\neq\mathbf{0}} \frac{\|A\mathbf{v}\|_2}{\|\mathbf{v}\|_2} = \max_{i=1,\ldots,n} \sqrt{\lambda_i(A^TA)}$$

$\lambda_i(A^TA) \ge 0, i = 1, \ldots, n$ are eigenvalues of A^TA



The set $\{A\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$ with $A = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$. The norm $\|A\|_2$ is the radius of the red set.

Matrix norms

- Frobenius norm

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

  - Not the same as $\|A\|_2$

- MATLAB function norm

```
% Assume the matrix A is defined
norm(A)        % default is 2-norm, norm(A, 2)
norm(A,1)      % 1-norm, max col sum, max(sum(abs(A)))
norm(A,'inf') % Inf-norm, max row sum, max(sum(abs(A),2))
norm(A,'fro') % Frobenius norm, sqrt(sum(sum(A.*A)))
```


Matrix norms – examples

Example (Matrix norms)

Find the 1, 2, infinity and Frobenius norms of

$$A = \begin{pmatrix} 1 & -3 & 2 \\ -2 & 2 & 4 \end{pmatrix}$$

Solution (MATLAB M-file `matnorms.m`)

- $\|A\|_1 = \max\{3, 5, 6\} = 6$
- $\|A\|_2$

$$A^T A = \begin{pmatrix} 5 & -7 & -6 \\ -7 & 13 & 2 \\ -6 & 2 & 20 \end{pmatrix}$$

  Eigenvalues $\lambda_i(A^T A) = 0, 14, 24 \Longrightarrow \|A\|_2 = \sqrt{24} = 4.8990$
- $\|A\|_\infty = \max\{6, 8\} = 8$
- $\|A\|_F^2 = 1^2 + (-3)^2 + 2^2 + (-2)^2 + 2^2 + 4^2 = 38$
  $\Longrightarrow \|A\|_F = \sqrt{38} = 6.1644$

Condition numbers

Definition (Condition number)

For $A \in \mathbb{R}^{n\times n}$, A nonsingular: Condition number

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

- A nonsingular $\iff$ det $(A) \neq 0 \iff A^{-1}$ exists
- Consistent matrix norm

$$I = AA^{-1} \Longrightarrow 1 = \|AA^{-1}\| \leq \|A\| \, \|A^{-1}\| \Longrightarrow \kappa(A) \geq 1$$

- $\kappa(\alpha I) = 1$, $\alpha \in \mathbb{R}$, $\alpha \neq 0$
- For a real symmetric matrix, using 2-norm

$$\kappa_2(A) = \|A\|_2 \, \|A^{-1}\|_2 = \frac{|\lambda_1|}{|\lambda_n|}$$

  $|\lambda_1| \geq, \ldots, \geq |\lambda_n|$ are magnitudes of eigenvalues of A

Ill-conditioned matrices

- Condition number $\kappa(A) \geq 1$
- A ill-conditioned $\iff$ $\kappa(A)$ large
  - What is large?
  - Large $\iff$ $\kappa > 1/\epsilon \approx 10^{16}$ ($\epsilon =$ relative machine precision)
- Reciprocal condition number rcond(A)

$$0 < \text{rcond}(A) \equiv \frac{1}{\kappa(A)} \leq 1$$

  - Well conditioned $\iff$ rcond(A) close to 1
  - Badly conditioned $\iff$ rcond(A) close to ϵ
- MATLAB functions
  - `cond(A)`, uses 2-norm
  - `cond(A, p)` uses p-norm
  - `rcond(A)` estimate of $1/\kappa(A)$ using 1-norm

Sensitivity of Linear Systems

- Linear system $A\mathbf{x} = \mathbf{b}$, A nonsingular
- How do errors in data A, $\mathbf{b}$ affect computed solution?
- Perturbed system, parameter $\eta \in \mathbb{R}$

$$(A + \eta\, C)\mathbf{x}(\eta) = \mathbf{b} + \eta\, \mathbf{c},$$

  $C \in \mathbb{R}^{n\times n}$, $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{x}(0) = \mathbf{x}$
- Relative error in input data

$$\rho_A = \eta \frac{\|C\|}{\|A\|}, \qquad \rho_b = \eta \frac{\|\mathbf{c}\|}{\|\mathbf{b}\|}$$

  - Input errors $\rho_A \geq \epsilon$, $\rho_b \geq \epsilon$
- Relative error in solution

$$\frac{\|\mathbf{x}(\eta) - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(A)[\rho_A + \rho_b] + O(\eta^2),$$

- relative error in solution $\leq \kappa(A) \times$ relative error in data

Accuracy of computed solution – example

Example (Accuracy of computed solution)

A symmetric matrix A is known exactly and has ordered eigenvalues

$$1004.2, \quad 999.8, \quad \ldots, \quad 0.0034, \quad 0.0005,$$

while the right-hand-side vector $\mathbf{b}$ is only measured to an accuracy of 6 significant figures.

1. Explain why A is nonsingular
2. Estimate the condition number of A
3. What is the relative error in the inputs A, $\mathbf{b}$
4. Estimate the relative error of the computed solution to $A\mathbf{x} = \mathbf{b}$
5. What confidence do you have in the computed solution?

Accuracy of computed solution – solution

Solution (Accuracy of computed solution)

A motivational example

Example

Suppose a matrix A is given. There are 10^6 input vectors $\mathbf{b}_k$. What is the most effective way to solve 10^6 linear systems $A\mathbf{x}_k = \mathbf{b}_k$?

Diagonal, triangular and permutation matrices

- Coefficient matrix $A \in \mathbb{R}^{m\times n}$
- D is diagonal $\iff$ $D_{ij} = 0$ for all $i \neq j$
  - All elements not on the main diagonal are zero
  - MATLAB command `diag`
- L is lower triangular $\iff$ $L_{ij} = 0$ for all $j > i$
  - All elements above the main diagonal are zero
  - L unit lower triangular $\iff$ L lower triangular and $L_{ii} = 1$ for all i
  - MATLAB command `tril`
- U is upper triangular $\iff$ $U_{ij} = 0$ for all $i > j$
  - All elements below the main diagonal are zero
  - U unit upper triangular $\iff$ U upper triangular and $U_{ii} = 1$ for all i
  - MATLAB command `triu`
- P is permutation matrix $\iff$ $P = [\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_n}]^T$
  $(i_1, \ldots, i_n)$ permutation of $(1, \ldots, n)$, $\mathbf{e}_i \in \mathbb{R}^n$ is ith unit vector
  - P is obtained by permuting rows of identity matrix I
  - P permutation matrix $\Longrightarrow PP^T = I$.

Matrix factorizations

- Square system: $A \in \mathbb{R}^{n\times n}$
- Express A as a product of matrices with special structure
- $A_k \in \mathbb{R}^{k\times k}$ leading principal sub-matrix of A
  - $(A_k)_{ij} = A_{ij}, \quad i, j = 1, \ldots, k$
  - MATLAB `Ak = A(1:k,1:k)`

Proposition (LU factorization)

Leading principal sub-matrices A_k non-singular for $k = 1, \ldots, n \Longrightarrow$ there exist $L \in \mathbb{R}^{n\times n}$, L unit lower triangular and $U \in \mathbb{R}^{n\times n}$, U upper triangular:

$$A = LU$$

- MATLAB `[L, U] = lu(A)`
- LU factorization does not exist for all non-singular A

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \ell_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}$$

  - $u_{11} = 0$ contradicts $\ell_{21}u_{11} = 1$
  - Leading principal sub-matrix $A_1 = 0$ is singular

LU factorization – Example

- Use row operations of the form $R_i \leftarrow R_i - L_{ij}R_j$ to make elements (i,j) zero for $i > j$.
- Continue until get upper triangular (row-echelon form) U

Example (LU factorization)

Calculate the LU factorization of $A = \begin{pmatrix} 1 & 2 & 3 \\ -1 & -3 & 2 \\ 2 & 4 & 5 \end{pmatrix}$.

Solution (MATLAB `luex1.m`)

- *Row operations $R_2 \leftarrow R_2 - (-1)R_1$ and $R_3 \leftarrow R_3 - (2)R_1$ give*

$$U = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -1 & 5 \\ 0 & 0 & -1 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}, \quad A = LU.$$

LU factorization

Proposition (LU factorization)

A non-singular $\Longrightarrow$ there exists $L \in \mathbb{R}^{n\times n}$, L unit lower triangular, $U \in \mathbb{R}^{n\times n}$, U upper triangular and permutation matrix $P \in \mathbb{R}^{n\times n}$: $PP^T = I$

$$PA = LU$$

- Pre-multiplying A by P to get PA reorders rows (equations) of A
  - Row operation of swapping/interchanging rows
  - Does not change solution to linear system
  - Example

$$PA = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 2 \\ 3 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$

- Post-multiplying A by P to get AP reorders columns (variables) of A
- MATLAB
  - `[L, U, P] = lu(A)`
  - `[L, U, p] = lu(A, 'vector')`
  - `P*A` same as `A(p,:)`

LU factorization – Example

Example (LU factorization with row swap)

Calculate the LU factorization of $A = \begin{pmatrix} 1 & 2 & 3 \\ -1 & -2 & 2 \\ 2 & 3 & 5 \end{pmatrix}$.

Solution (MATLAB `luex2.m`)

- *Row operations $R_2 \leftarrow R_2 - (-1)R_1$ and $R_3 \leftarrow R_3 - (2)R_1$ give*

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 5 \\ 0 & -1 & -1 \end{pmatrix}$$

- *Swapping rows 2 and 3: $R_2 \leftrightarrow R_3$ gives*

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, U = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -1 & -1 \\ 0 & 0 & 5 \end{pmatrix}, L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, PA = LU$$

Pivoting

- Working on sub-matrix in rows $j, j+1, \dots, n$
- Need pivot element $a_{jj} \neq 0$
- Swap rows to get non-zero pivot element
- Numerical stability $\Longrightarrow$ pivot element as large as possible
- Partial pivoting choose largest magnitude element in column

$$|a_{\hat{i}j}| = \max_{i=j,\dots,n} |a_{ij}|$$

  - Only need to swap rows/equations
- Complete pivoting choose largest magnitude element in sub-matrix

$$|a_{\hat{i}\hat{j}}| = \max_{\substack{i=j,\dots,n \\ \ell=j,\dots,n}} |a_{i\ell}|$$

  - Need to swap both rows/equations and columns/variables

$$PAQ^TQ\mathbf{x} = P\mathbf{b} \iff PAQ^T\mathbf{y} = P\mathbf{b}, \quad \mathbf{y} = Q\mathbf{x}$$

Solving $A\mathbf{x} = \mathbf{b}$ using LU factorization

- Factorization: $PA = LU$
- $A\mathbf{x} = \mathbf{b} \Longrightarrow PA\mathbf{x} = P\mathbf{b} \Longrightarrow LU\mathbf{x} = P\mathbf{b}$
- Forward substitution: Solve $L\mathbf{y} = P\mathbf{b} = \hat{\mathbf{b}}$

$$y_1 = \hat{b}_1, \qquad y_i = \hat{b}_i - \sum_{j=1}^{i-1} y_i\hat{b}_i, \quad i = 2, \dots, n$$

- Back-substitution: Solve $U\mathbf{x} = \mathbf{y}$

$$x_n = y_n/U_{nn}, \qquad x_i = \big(y_i - \sum_{j=i+1}^{n} x_jy_j\big)/U_{nn}, \quad i = n-1, \dots, 1$$

Counting flops in solving $A\mathbf{x} = \mathbf{b}$ using LU factorization

- Factorization $\frac{2n^3}{3} + O(n^2)$ flops
- Forward substitution $\frac{n^2}{2} + O(n)$ flops
- Back-substitution $\frac{n^2}{2} + O(n)$ flops
- Total solve $\frac{2n^3}{3} + O(n^2)$ flops
- Several RHS $\mathbf{b}_k$, $k = 1, \dots, K$ with $K \ll n$ and one factorization $\Longrightarrow$ same total flops $\frac{2n^3}{3} + O(n^2)$
- Matlab `LUsolvers.m`