Read book pages 1-11

Vectors and Matrices

$\mathbb{R}^n = \left\{ \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} : x_i \in \mathbb{R} \right\}$

the set of all ordered n-tuples of real numbers!

$\mathbb{C}^n = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} : x_i \in \mathbb{C} \right\}$

Vector addition: $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1+y_1 \\ \vdots \\ x_n+y_n \end{pmatrix} \in \mathbb{R}^n$

Multiplication by scalar: $\alpha \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{pmatrix} \in \mathbb{R}^n$

$\mathbb{R}^n$ is closed under these operations.

For any $x, y, z \in \mathbb{R}^n$, $\alpha, \beta \in \mathbb{R}$ we have the following properties (verify!):

vector space

$x + (y+z) = (x+y) + z$

$x + y = y + x$

$\exists\, 0$ s.t. $x + 0 = x$ (any x); $0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^n$

$\exists\, (-x)$ s.t. $x + (-x) = 0$

$1 \cdot x = x$

$\alpha(\beta x) = (\alpha\beta)x$

$\alpha(x+y) = \alpha x + \alpha y$

$(\alpha+\beta)x = \alpha x + \beta x$

These are the properties of a vector space

A set that satisfies these properties wrt to the real numbers (complex numbers) as scalars is a real (complex) vector space.

The concept of a vector space is much more general than $\mathbb{R}^n$ and $\mathbb{C}^n$.

In this course we care mainly about $\mathbb{R}^n$ and $\mathbb{C}^n$.

Examples: (from $\mathbb{R}^3$)

$$\begin{pmatrix}1\\2\\3\end{pmatrix}+\begin{pmatrix}4\\5\\6\end{pmatrix}=\begin{pmatrix}5\\7\\9\end{pmatrix},\quad 2\cdot\begin{pmatrix}1\\2\\3\end{pmatrix}=\begin{pmatrix}2\\4\\6\end{pmatrix}$$

$$\begin{pmatrix}1\\2\\3\end{pmatrix}+\begin{pmatrix}0\\0\\0\end{pmatrix}=\begin{pmatrix}1\\2\\3\end{pmatrix},\quad \begin{pmatrix}1\\2\\3\end{pmatrix}+\begin{pmatrix}-1\\-2\\-3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$$

dependent/ independent vectors

Consider the set of vectors

$$\begin{pmatrix}1\\1\\1\end{pmatrix},\ \begin{pmatrix}1\\-1\\0\end{pmatrix},\ \begin{pmatrix}0\\2\\1\end{pmatrix}$$

We have

$$1\cdot\begin{pmatrix}1\\1\\1\end{pmatrix}-1\cdot\begin{pmatrix}1\\-1\\0\end{pmatrix}-1\cdot\begin{pmatrix}0\\2\\1\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$$

there is a nontrivial linear combination of these vectors that equals the zero vector. (trivial $\rightarrow 0\cdot x+0\cdot y+0\cdot z=0$)

We say these vectors are dependent

More generally a set of vectors $x_1, x_2, \cdots, x_m$ is dependent if there are scalars $\alpha_1, \alpha_2, \cdots, \alpha_m$ (not all zero) such that

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m = 0$$

If $\alpha_1 x_1 + \cdots + \alpha_m x_m = 0 \iff \alpha_1 = 0, \cdots, \alpha_m = 0$, we say these vectors are independent.

Consider $\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}$:

$$\alpha_1\begin{pmatrix}1\\0\\0\end{pmatrix} + \alpha_2\begin{pmatrix}0\\1\\0\end{pmatrix} + \alpha_3\begin{pmatrix}0\\0\\1\end{pmatrix} = \begin{pmatrix}\alpha_1\\\alpha_2\\\alpha_3\end{pmatrix} = \begin{pmatrix}0\\0\\0\end{pmatrix} \iff$$

$\alpha_1 = 0$ and $\alpha_2 = 0$ and $\alpha_3 = 0$.

The span of a set of vectors is the set of all linear combinations of those vectors. More formally

$$\text{span}\{x_1, x_2, \cdots, x_k\} \equiv$$

$$\{\alpha_1 x_1 + \cdots + \alpha_k x_k : \text{any } \alpha_1, \alpha_2, \cdots, \alpha_k \in \mathbb{R}\}$$

(scalars from $\mathbb{C}$ for complex vector space)

The span of a set of vectors is a vector space.

basis

A set of vectors forms a basis for a vector space if

i) any vector from the space can be represented by a linear combination of the set of vectors

ii) the set of vectors is independent.

test

Let S be a vector space and $\{x_1, \cdots, x_k\}$ a given set of vectors.

$x_1, \cdots, x_k$ form a basis for S if

$\forall s \in S : s = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k$
(some $\alpha_1, \cdots, \alpha_k$)

and

vectors $x_1, \cdots, x_k$ are independent.

Note that the independence of x_i implies that the sequence (tuple) of scalars $\alpha_1, \cdots, \alpha_k$ is unique.

A basis for a vector space is not unique.

The linear combination

$$s = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k$$

is called the representation / decomposition of s w.r.t. the basis $x_1, \cdots, x_k$.

$\left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$ is a basis for $\mathbb{R}^3$

A particularly useful basis for $\mathbb{R}^3$ is given by

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

These vectors are referred as the canonical or Cartesian basis vectors

The number of vectors of a basis equals the dimension of the vector space. (finite dimensional space)

Consider the set of all polynomials

$$p(x) = \sum_{i=0}^{\infty} a_i x^i \qquad x \in [0,1]$$

Sum of two poly.s is polynomial.
Scalar times poly. gives polynomial.
Also other properties satisfied.

So, set of all polynomials is a vector space.

There is no finite set of polynomials that can span the whole space → infinite dimensional vector space.

Basis : $\{1, x, x^2, x^3, x^4, \cdots\}$

real and complex matrices

$$\mathbb{R}^{m\times n} = \left\{ \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \cdots & & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & \cdots & & a_{mn} \end{pmatrix} \;\middle|\; a_{ij} \in \mathbb{R} \right\}$$

Similarly $\mathbb{C}^{m\times n} = \left\{ (c_{ij}) \mid c_{ij} \in \mathbb{C} \right\}$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \vdots \\ a_{m1} & \cdots & & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix}$$

$$= \begin{pmatrix} (a_{11}+b_{11}) & \cdots & (a_{1n}+b_{1n}) \\ \vdots & & \vdots \\ (a_{m1}+b_{m1}) & \cdots & (a_{mn}+b_{mn}) \end{pmatrix}$$

$$\gamma \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} \gamma a_{11} & \cdots & \gamma a_{1n} \\ \vdots & & \vdots \\ \gamma a_{m1} & \cdots & \gamma a_{mn} \end{pmatrix}$$

The other vector space properties are satisfied too.

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

$$Ax = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} x_1 + \begin{pmatrix} a_{12} \\ \vdots \\ a_{m2} \end{pmatrix} x_2 + \cdots + \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} x_n$$

$B \in \mathbb{R}^{n \times p}$; let $AB = C \in \mathbb{R}^{m \times p}$

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & & \vdots \\ c_{m1} & \cdots & c_{mp} \end{pmatrix} \quad \text{where}$$

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \longrightarrow$$

dot product of ith row of A with jth column of b.

Equivalently, we have for the columns of C

$$C = [c_1 \; c_2 \; \cdots \; c_p]$$

$$= [Ab_1 \; Ab_2 \; \cdots \; Ab_p]$$

c_i, b_i are columns of C and B.

We can also define matrix-matrix product blockwise (also matrix-vector prod.)

$$A = \left(\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1k} & a_{1k+1} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{l1} & \cdots & a_{lk} & a_{lk+1} & \cdots & a_{ln} \\ \hline a_{l+1,1} & \cdots & a_{l+1,k} & a_{l+1,k+1} & \cdots & a_{l+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{m1} & \cdots & a_{mk} & a_{m,k+1} & \cdots & a_{mn} \end{array}\right)$$

$$= \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right)$$

$$B = \left(\begin{array}{ccc|ccc} b_{11} & \cdots & b_{1r} & b_{1r+1} & \cdots & b_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ b_{k1} & \cdots & b_{kr} & b_{kr+1} & \cdots & b_{kp} \\ \hline b_{k+1,1} & \cdots & b_{k+1,r} & b_{k+1,r+1} & & b_{k+1,p} \\ \vdots & & \vdots & \vdots & & \vdots \\ b_{n1} & \cdots & b_{nr} & b_{n,r+1} & & b_{np} \end{array}\right)$$

$$= \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

$$AB = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} A_{11}B_{11}+A_{12}B_{21} & \cdots \\ A_{21}B_{11}+A_{22}B_{21} & \cdots \end{pmatrix}$$

$$= \begin{pmatrix} A_{11}B_{11}+A_{12}B_{21} & A_{11}B_{12}+A_{12}B_{22} \\ A_{21}B_{11}+A_{22}B_{21} & A_{21}B_{12}+A_{22}B_{22} \end{pmatrix}$$

$A \in \mathbb{R}^{m \times n}$, $x, y \in \mathbb{R}^n$, $\alpha, \beta \in \mathbb{R}$

We have $A(\alpha x + \beta y) = \alpha Ax + \beta Ay$

$\rightarrow$ A is linear operator

Examples:

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} (2-1) & (2+1) \\ (1-2) & (1+2) \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ -1 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \left(\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)$$

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

(See also book, pages 1 – 11)

Different ways to represent matrix-vector and matrix-matrix products indicate different implementations which typically differ in performance.

Read book pages 12-13

Solution of (square) linear system

$A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^n$ given $\rightarrow$

Find $x \in \mathbb{R}^n$ such that $Ax = b$.

Examples:

I $\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -2 \end{pmatrix} \rightarrow \begin{cases} 2x_1 + x_2 = 0 \\ x_2 = -2 \end{cases}$

$\rightarrow \begin{cases} 2x_1 = 2 \rightarrow x_1 = 1 \\ x_2 = -2 \end{cases}$

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

Existence & Uniqueness of Solutions

IIa $\begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \rightarrow \begin{cases} 2x_1 + x_2 = 1 \\ 2x_1 + x_2 = 0 \end{cases}$

No solution.

IIb $\begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} \rightarrow \begin{cases} 2x_1 + x_2 = 3 \\ 2x_1 + x_2 = 3 \end{cases}$

solution $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Note that $\begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow$

$\begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \alpha\begin{pmatrix} 1 \\ -2 \end{pmatrix}\right) = \begin{pmatrix} 3 \\ 3 \end{pmatrix} + \alpha\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$

Sol., $\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \alpha\begin{pmatrix} 1 \\ -2 \end{pmatrix}$ is solution for any α.

$\rightarrow$ infinite number of solutions (line)

Note solution is sum of general solution of homogeneous problem plus particular solution of inhomogeneous problem.

It is easy to see that I has a unique solution for any right hand side.

We see that II may have no solution or infinite number of solutions.

We say system II is singular and system I is nonsingular or regular.

What characterizes a singular/nonsingular system?

If $Ax=b$ has solution, we have

$$Ax = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = b$$

(a_i columns of A, x_i coefficients vector x)

So, b is a linear combination of the columns of A. The coefficients x_i define the linear combination

→ $Ax=b$ has at least 1 solution if $b \in \text{span}\{a_1, a_2, \cdots, a_n\}$. The space spanned by the columns of A (also column space or range of A)

→ If the columns of A are independent they span whole space $\mathbb{R}^n$ and we know b has unique decomposition along the columns of A:

$\{a_1, a_2, \cdots, a_n\}$ is a basis for $\mathbb{R}^n$

So, $Ax=b$ has a unique solution.

If the columns of A are dependent, they do not form a basis for $\mathbb{R}^n$ and they do not span $\mathbb{R}^n$. Hence, there are vectors in $\mathbb{R}^n$ that are not a linear combination of the columns of A. For these vectors b, $Ax=b$ has no solution.

If $a_1, \dots, a_n$ are dependent, there is a vector $(x_1\ x_2 \cdots x_n)^T$ such that

$$a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n = 0$$

$$\Rightarrow Ax = 0$$

Analogously $Ax=0$ $(x \neq 0) \Rightarrow$

$\{a_1, a_2, \dots, a_n\}$ dependent.

B is the inverse of A if $AB = BA = \underline{I}$

where I is the identity matrix

$$I \in \mathbb{R}^{n\times n} = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ & & \ddots & \end{pmatrix}^{n\times n}$$

$$I \in \mathbb{R}^{3\times 3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$Ix = x \quad \forall x \in \mathbb{R}^n$

$IA = AI = A \quad \forall A \in \mathbb{R}^{n\times n}$ (same for $\mathbb{C}^n$)

The following statements are equivalent

1) A is nonsingular
2) A^{-1} exists
3) the columns of A are independent
4) the rows of A are independent
5) $Ay = 0 \Rightarrow y = 0$
6) $\det(A) = 0$
7) $Ax = b$ has unique solution for any b

Try to prove these equivalences

Some further definitions:

$\text{null}(A) = \{z : Az = 0\}$ null space or kernel

$\text{range}(A) = \{z : \exists x \in \mathbb{R}^n \text{ such that } Ax = z\}$

(Show that null space of singular matrix A is a vector space)

Read pages 13 - 23

Examples of linear systems

→ read example 1.2.6 about electrical circuits

→ read example 1.2.10 on mass-spring systems

→ read example 1.2.12 on differential equations

Ex 1.2.12: we can prove the approximations for derivatives using Taylor approximations to u (at some point)

(Assume $u(x)$ has sufficiently high derivatives)

$$u(x+h) = u(x) + hu'(x) + \frac{1}{2}h^2u''(x) + \frac{1}{6}h^3u'''(x) + \frac{h^4}{24}u^{(4)}(x) + O(h^5)$$
$$u(x-h) = u(x) - hu'(x) + \frac{1}{2}h^2u''(x) - \frac{1}{6}h^3u'''(x) + \frac{1}{24}h^4u^{(4)}(x) + O(h^5)$$

$$u(x+h) - u(x-h) = 2hu'(x) + \frac{1}{3}h^3u'''(x) + O(h^5)$$

$$u'(x) \equiv \frac{1}{2h}\left(u(x+h) - u(x-h)\right) + O(h^2)$$

$$u(x+h) - 2u(x) + u(x+h) = h^2u''(x) + \frac{1}{12}h^4u^{(4)}(x) + O(h^5)$$

$$u''(x) = \frac{1}{h^2}\left(u(x+h) - 2u(x) + u(x-h)\right) + O(h^2)$$

Consider the differential equation

$$\begin{cases} -u'' + \epsilon u' + du = f & x \in (a,b) \\ u(a) = u_a \quad \text{(given values)} \\ u(b) = u_b \end{cases}$$

We approximate solution $u(x)$ using set of discrete values U_i, $i = 0 \ldots N$, where

U_i approximates the solution $u(x)$ at

$x = a + ih \quad \left(h = \frac{1}{N}(b-a)\right)$

So, $U_i \simeq u(a+ih)$

$U_0 = u(a) = u_a$ and $U_N = u(b) = u_b$ given

Take $a=0$, $b=2$, $d(x) = 1$, $c=0$, $f=1$
$u(0) = 0$, $u(2) = 1$, $h = \frac{1}{10}$

We replace $u''(x_i)$ by $\frac{1}{h^2}(U_{i+1} + 2U_i - U_{i-1})$
etc.
Eqn at point i ($x = ih$)

$$\frac{1}{h^2}(-U_{i+1} + 2U_i - U_{i-1}) + U_i = 1$$

$i = 0$: $U_0 = 0$ (boundary condition)
$i = 1$: $-\frac{1}{h^2}U_0 + \left(1 + \frac{2}{h^2}\right)U_1 - \frac{1}{h^2}U_2 = 1$

$i = 2$: $-1/h^2 U_1 + (1 + 2/h^2)U_2 - 1/h^2 U_3 = 1$

etc

$i = 19$: $-1/h^2 U_{18} + (1 + 2/h^2)U_{19} - 1/h^2 U_{20} = 1$

$U_{20} = u(2) = 1$

We can substitute for U_0 and U_{20} in these equations.

$$\begin{bmatrix} 201 & -100 & & & & & & & & \\ -100 & 201 & -100 & & & & & & & \\ & -100 & 201 & -100 & & & & & & \\ & & & \ddots & & & & & & \\ & & & & & & -100 & 201 & -100 & \\ & & & & & & & -100 & 201 & \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_{18} \\ U_{19} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ \\ 1 \end{bmatrix}$$

The solution represents the amount of a chemical at point x ($u(x)$) in pipe (1D) of water, where the rate of diffusion is 1, there is no convection (no flow) and $d = 1$ represents the rate of production by chemical reaction. $f = 1$ represents a source term.

Once we have a program that computes an approximate solution given the boundary conditions and the functions c, d, and f, we can easily vary these and see the effect on the solution u.

We see that the solution of a system of linear equations can give the (approx) solution to various physical problems.

The mass-spring system is more general than you may think as the displacements in a bar of homogeneous material can be modeled this way. For small displacements most materials obey Hooke's law that the ratio between displacement and force is a constant (the spring constant).

The next question is how to compute the solution to a system of linear equations.

Triangular systems

Read book pages 23 - 32

A central theme in this course will be to reduce a general problem to one for which the solution is easy.

For linear systems this generally means going from a general matrix to an upper or lower triangular system. More precisely, we write the general matrix as a product of a lower and an upper triangular matrix.

Consider

$$\begin{pmatrix} 3 & 0 & 0 \\ 2 & 2 & 0 \\ 2 & 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ -2 \end{pmatrix}$$

$$3x_1 = -3 \rightarrow x_1 = -1$$
$$2x_1 + 2x_2 = 0 \rightarrow 2x_2 = 2 \rightarrow x_2 = 1$$
$$2x_1 + x_2 - x_3 = -2 \rightarrow -x_3 = -2 - 1 + 2 = -1$$
$$x_3 = 1$$

General case:

$$\begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

$$x_1 = b_1 / a_{11}$$
$$x_2 = (b_2 - a_{21} x_1) / a_{22}$$

$$x_i = \left(b_i - \sum_{k=1}^{i-1} a_{ik} x_k \right) / a_{kk}$$

It is clear from the algorithm that we can compute a solution for any right hand side as long as $a_{ii} \neq 0$, $i = 1 \ldots n$.

It is also easy to see that $b = 0 \Rightarrow x = 0$ if $a_{ii} \neq 0$ for $i = 1, \ldots, n$.

We can assess the cost of this algorithm by counting the number of floating point operations.
(we'll count all floating point operations as having equal cost)

$x_1 \rightarrow$ 1 flop
$x_2 \rightarrow$ 3 flops
$x_3 \rightarrow$ 5 "
⋮
$x_n \rightarrow 2n-1$ ($2(n-1)$ products plus subtractions plus 1 division)

$$1+3+5+\cdots+2n-1 = n\cdot(2n-1+1)\cdot\frac{1}{2} = n^2$$

We can rearrange the operations to go from a row oriented algorithm to a column oriented algorithm.

$$x_1 = b_1/a_{11}\,;\quad \begin{pmatrix} b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix} = \begin{pmatrix} b_2^{(1)} \\ \vdots \\ b_n^{(1)} \end{pmatrix} - x_1 \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$$

$$x_2 = b_2^{(2)}/a_{22}\,;\quad \begin{pmatrix} b_3^{(3)} \\ \vdots \\ b_n^{(3)} \end{pmatrix} = \begin{pmatrix} b_3^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix} - x_2 \begin{pmatrix} a_{32} \\ \vdots \\ a_{n2} \end{pmatrix}$$

<u>etc</u>

as soon as x_k is computed we subtract the remainder of the k-th column (which x_k multiplies) from the right hand side.

$$b = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k + a_{k+1} x_{k+1} + \cdots + a_n x_n$$

$$b^{(k+1)} = b - a_1 x_1 - \cdots - a_k x_k$$

This algorithm has the same cost as the row oriented one.

(read sections on upper triangular systems and block algorithms)

Block algorithms are particularly important for efficient (hierarchical) memory use.

Read book pages 32 – 41

Cholesky decomposition for symmetric positive definite (SPD) systems

(Hermitian positive definite in the complex case)

We say a matrix A is symmetric if

$$A = A^T$$

$\left((A^T)_{ij} = a_{ji} \right)$.

We say a complex matrix A is Hermitian if

$$A = A^H$$

where $(A^H)_{ij} = \bar{a}_{ji}$ (complex conjugate transpose)

A matrix A is positive definite if

$x^T A x > 0 \quad \forall x \neq 0$ (real)

$x^H A x > 0 \quad \forall x \neq 0$ (complex)

If A is SPD (HPD) then A is nonsingular (see book)

We now consider the first method to factorize a matrix into a lower triangular and upper triangular factor.

If A is SPD (HPD) then $A = R^T R$ ($R^H R$) for some upper triangular matrix R with strictly positive (real) diagonal coefficients. (proof later)

We can devise an algorithm to compute R from A by looking at the coefficients in $A = R^T R$...

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix} = \underbrace{\begin{pmatrix} r_{11} & 0 & & \\ r_{12} & r_{22} & & \\ r_{13} & r_{23} & r_{33} & \\ \vdots & \vdots & & \ddots \\ r_{1n} & r_{2n} & \cdots & r_{nn} \end{pmatrix}}_{R^T} \underbrace{\begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ 0 & r_{22} & r_{23} & \cdots & r_{2n} \\ & & r_{33} & \cdots & r_{3n} \\ & & & \ddots & \vdots \\ & & & & r_{nn} \end{pmatrix}}_{R}$$

(note $a_{ij} = a_{ji}$)

$$a_{11} = r_{11}^2 + 0 + \cdots + 0 \rightarrow r_{11} = \sqrt{a_{11}}$$

$$a_{1j} = r_{11} r_{1j} \rightarrow r_{1j} = a_{1j}/r_{11} \quad \text{for } j = 2 \ldots n$$

After this, first row of R (column of R^T) is known.

$$a_{22} = r_{12}^2 + r_{22}^2 \rightarrow r_{22}^2 = \sqrt{a_{22} - r_{12}^2}$$

$$a_{2j} = r_{12} r_{1j} + r_{22} r_{2j} \rightarrow r_{2j} = (a_{2j} - r_{12} r_{1j})/r_{22}, \; j = 3 \ldots n$$

More generally, assume we have computed the first k rows of R. We compute row $k+1$ as follows.

$$a_{k+1\,k+1} = \sum_{i=1}^{k} r_{i\,k+1}^2 + r_{k+1\,k+1}^2 \rightarrow r_{k+1\,k+1} = \left(a_{k+1\,k+1} - \sum_{i=1}^{k} r_{i\,k+1}^2\right)^{1/2}$$

$$a_{k+1\,j} = \sum_{i=1}^{k} r_{i\,k+1} r_{ij} + r_{k+1\,k+1} r_{k+1\,j} \rightarrow r_{k+1\,j} = \left(a_{k+1\,j} - \sum_{i=1}^{k} r_{i\,k+1} r_{ij}\right)/r_{k+1\,k+1}$$

$$\text{for } j = k+2 \ldots n$$

Cost of Cholesky factorization?

Work for k-th row:

a) $r_{kk} = \left(a_{kk} - \sum_{i=1}^{k-1} r_{ik}^2\right)^{1/2} \rightarrow 2(k-1)+1$ flop

b) $r_{kj} = \left(a_{kj} - \sum_{i=1}^{k-1} r_{ik} r_{ij}\right)/r_{kk} \rightarrow (n-k)(2(k-1)+1)$

Step k: $(n-k+1)(2k-1)$

Total: $\sum_{k=1}^{n} (n-k+1)(2k-1)$

$$= \underbrace{n \sum_{k=1}^{n} (2k-1)}_{\text{I}} - \underbrace{\sum_{k=1}^{n} (k-1)(2k-1)}_{\text{II}}$$

I: $n\left(1+3+5+\cdots+2n-1\right) = n \cdot \frac{1}{2}n \cdot 2n = n^3$

II: let $f_m = \sum_{k=1}^{m} (k-1)(2k-1)$

$$\Rightarrow f_m - f_{m-1} = 2m^2 - 3m + 1$$

$$f_m = am^3 + bm^2 + cm + d$$
$$f_{m-1} = a(m-1)^3 + b(m-1)^2 + c(m-1) + d$$
$$= am^3 - 3am^2 + 3am - a + bm^2 - 2bm + b + cm - c + d$$

$$f_m - f_{m-1} = 3am^2 - 3am + a + 2bm - b + c$$
$$= 3am^2 + (-3a+2b)m + a - b + c$$
$$= 2m^2 - 3m + 1$$

$$\rightarrow \begin{cases} 3a = 2 \rightarrow a = 2/3 \\ -3a + 2b = -3 \rightarrow b = -1/2 \\ a - b + c = 1 \rightarrow c = -1/6 \end{cases}$$

$f_1 = 0 \Rightarrow a+b+c+d = 0 \Rightarrow d = 0$

$$\text{I} + \text{II} = n^3 - \frac{2}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n \text{ flop}$$

$n=1 \rightarrow 1$, $n=2 \rightarrow 5$, etc.

Now, we are ready to solve SPD linear systems.

Solve $Ax = b$, $A = R^T R$, R upper triangular

$R^T R\, x = b$, substitute $y = Rx$

$$\begin{cases} \text{Solve } R^T y = b \\ \text{Solve } R x = y \end{cases}$$

Example: $\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -4 \\ 9 \end{pmatrix}$

$$\begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -4 \\ 9 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} -4 \\ 9 \end{pmatrix} \rightarrow y_1 = -4,\ y_2 = 1$$

$$\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -4 \\ 1 \end{pmatrix} \rightarrow x_2 = 1,\ x_1 = -2$$

(verify answer)

Alternative algorithm. We reduce an $n \times n$ Cholesky factorization to an $(n-1)\times(n-1)$ Cholesky factorization.

$$\begin{pmatrix} a_{11} & b^T \\ b & \hat{A} \end{pmatrix} = \begin{pmatrix} r_{11} & 0 \\ s & \hat{R}^T \end{pmatrix}\begin{pmatrix} r_{11} & s^T \\ 0 & \hat{R} \end{pmatrix}$$

$$a_{11} = r_{11}^2 \rightarrow r_{11} = \sqrt{a_{11}}$$
$$r_{11} s^T = b^T \rightarrow s^T = b^T / r_{11} = b^T / \sqrt{a_{11}}$$

$$ss^T + \hat{R}^T\hat{R} = \hat{A} \rightarrow \hat{R}^T\hat{R} = \hat{A} - ss^T \quad \text{(Chol. fact.)}$$

We have reduced the problem now to a Cholesky factorization of size $(n-1)\times(n-1)$.

$$ss^T = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{n-1} \end{pmatrix}(s_1\, s_2 \cdots s_{n-1}) = \begin{pmatrix} s_1^2 & s_1 s_2 & s_1 s_3 & \cdots & s_1 s_{n-1} \\ s_2 s_1 & s_2^2 & & \cdots & \\ \vdots & & & & \\ s_{n-1}s_1 & s_{n-1}s_2 & & \cdots & s_{n-1}^2 \end{pmatrix}$$

(outer product)

Let c_n indicate the cost of an $n \times n$ Cholesky factorization.

$$c_n = n + n(n-1) + c_{n-1} = n^2 + c_{n-1}$$

(Note that for the step $\hat{A} - ss^T$ we only need to update the coefficients on and above the diagonal $\rightarrow$
$2 \cdot (n-1 + n-2 + \cdots + 1) = 2 \cdot (n-1) \cdot \frac{1}{2} \cdot n$)

$$c_n - c_{n-1} = n^2$$

$$c_n = an^3 + bn^2 + cn + d$$
$$c_{n-1} = an^3 - 3an^2 + 3an - a + bn^2 - 2bn + b + cn - c + d$$

$$c_n - c_{n-1} = 3an^2 + (-3a + 2b)n + a - b + c$$

$$\begin{cases} 3a = 1 & \rightarrow a = 1/3 \\ -3a + 2b = 0 & \rightarrow b = 1/2 \\ a - b + c = 0 & \rightarrow c = 1/6 \end{cases}$$

$$c_1 = 1 \Rightarrow d = 0$$

$$c_n = 1/3\, n^3 + 1/2\, n^2 + 1/6\, n$$

<u>Block-wise Cholesky factorization.</u>

$$\begin{pmatrix} A_{11} & B \\ B^T & \hat{A} \end{pmatrix} = \begin{pmatrix} R_{11}^T & 0 \\ S^T & \hat{R}^T \end{pmatrix} \begin{pmatrix} R_{11} & S \\ 0 & \hat{R} \end{pmatrix}$$

$A_{11} = R_{11}^T R_{11}$ (Cholesky fact.)

$B = R_{11}^T S$ (Solve lower triangular systems)

$\hat{R}^T \hat{R} = \hat{A} - S^T S$ (Outer product update + Chol. factorization)

If A is SPD, then A is nonsingular

$x^T A x > 0$ if A SPD
Assume A singular. Then $Ax = 0$ for some $x \neq 0$. $Ax = 0 \Rightarrow x^T A x = 0$. This leads to a contradiction.

If A is SPD, then every principal submatrix is also SPD.

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & A_{22} & A_{23} \\ A_{13}^T & A_{23}^T & A_{33} \end{pmatrix}.$$ Consider the principal submatrix A_{22}:

Take $x = \begin{pmatrix} 0 \\ x_2 \\ 0 \end{pmatrix}$.

$x^T A x = x_2^T A_{22} x_2 > 0$ (A is SPD)

So, A_{22} must also be SPD.

(A principal submatrix is any matrix that remains after removing some set of rows and the corresponding (same indices) set of columns.)

Note the special case where $A_{22} \in \mathbb{R}^{1\times 1}$ → all diagonal coefficients of A are positive and real.

If A is SPD, then all the eigenvalues of A are positive.

($Ax = \lambda x$ → λ is eigenvalue, x is eigenvector)
Let x be an eigenvector.
$x^T A x = x^T(\lambda x) = \sum_{i=1}^{n} \lambda x_i^2$.

So, if $\lambda < 0$ then $x^T A x < 0$. Contradiction

* symmetric

A* matrix, A, has a factorization $A = R^T R$, where R is upper triangular with positive and real diagonal coefficients, if and only if A is SPD (HPD in the complex case).

i) Let $A = R^T R$ (with R as above).
Consider $x^T A x = x^T R^T R x = y^T y$ with $y = Rx$. $y^T y = \sum_i y_i^2 > 0$ if $y \neq 0$.
Since the diagonal coefficients of R are nonzero, R is nonsingular and $Ry \neq 0$ (if $y \neq 0$).

So, for any nonzero y, $y^T A y > 0$.
Hence, A is SPD.

ii) Let A be SPD. We will prove the existence of the Cholesky factorization by induction.
$A \in \mathbb{R}^{1\times 1} \rightarrow A = [a_{11}]$ with $a_{11} > 0 \Rightarrow$
$R = [r_{11}]$ with $r_{11} = \sqrt{a_{11}}$.

Next, we show that if the Cholesky factorization exists for every SPD matrix $A \in \mathbb{R}^{(n-1)\times(n-1)}$, then it exists for every SPD matrix $A \in \mathbb{R}^{n\times n}$.

Let A be $\begin{pmatrix} a_{11} & b^T \\ b & \hat{A} \end{pmatrix}$.

We carry out the first step of the outer product algorithm:

$$\begin{pmatrix} a_{11} & b^T \\ b & \hat{A} \end{pmatrix} = \begin{pmatrix} r_{11} & 0 \\ s & \hat{R}^T \end{pmatrix} \begin{pmatrix} r_{11} & s^T \\ 0 & \hat{R} \end{pmatrix}$$

where $r_{11} = \sqrt{a_{11}}$ and $s = r_{11}^{-1} b$.

We have $\hat{A} = ss^T + \hat{R}^T \hat{R} \Leftrightarrow \hat{R}^T \hat{R} = \hat{A} - ss^T$.
Now if $\hat{A} - ss^T$ is positive definite (its obviously symmetric), by the

induction hypothesis the Cholesky factorization

$$\hat{R}^T\hat{R} = \hat{A} - ss^T \quad (\in \mathbb{R}^{(n-1)\times(n-1)})$$

exists, and hence the Cholesky factorization

$R^TR = A$ ([redacted] above) exists.

So, we must show that

$$\tilde{x}^T(\hat{A} - ss^T)\tilde{x} > 0 \quad \text{for any } \tilde{x} \neq 0 \ (\in \mathbb{R}^{n-1}).$$

Let $x_1 = -a_{11}^{-1} b^T\tilde{x}$. Then

$$\begin{pmatrix} x_1 \\ \tilde{x} \end{pmatrix}^T \begin{pmatrix} a_{11} & b^T \\ b & \hat{A} \end{pmatrix} \begin{pmatrix} x_1 \\ \tilde{x} \end{pmatrix} = x_1^2 a_{11} + x_1 b^T\tilde{x} + \tilde{x}^T b x_1 + \tilde{x}^T\hat{A}\tilde{x}$$

$$= x_1^2 a_{11} + 2x_1 b^T\tilde{x} + \tilde{x}^T\hat{A}\tilde{x}$$

$$= a_{11}^{-1}(b^T\tilde{x})^2 - 2a_{11}^{-1}(b^T\tilde{x})^2 + \tilde{x}^T\hat{A}\tilde{x}$$

$$= \tilde{x}^T\hat{A}\tilde{x} - a_{11}^{-1}(b^T\tilde{x})^2 = \tilde{x}^T(\hat{A} - ss^T)\tilde{x}.$$

Hence, $\tilde{x}^T(\hat{A} - ss^T)\tilde{x} = \begin{pmatrix} x_1 \\ \tilde{x} \end{pmatrix}^T \begin{pmatrix} a_{11} & b^T \\ b & \hat{A} \end{pmatrix} \begin{pmatrix} x_1 \\ \tilde{x} \end{pmatrix} > 0$

(since A is positive definite).

Banded Matrices

We say a matrix A is banded with semi bandwidth b if $a_{ij} = 0$ for $|i-j| > b$ (and at least one coefficient $a_{ij} \neq 0$ with $|i-j| = b$.

From the construction of the Cholesky factorization (1st. algorithm) we can see that the factor R has the same semi bandwidth.

What does this mean for the cost of the Cholesky factorization?

Gaussian Elimination and LU decomposition

We now consider the systematic solution of general, nonsingular linear systems.

The main idea is similar to what we have done for SPD systems. We will decompose/factorize a matrix into a product of a lower triangular matrix and an upper triangular matrix.

$$Ax = b \rightarrow LUx = x \quad (A = LU)$$

where L: lower triangular
U: upper "

If P is nonsingular, then

$$Ax = b \text{ if and only if } PAx = Pb$$

(show this). The linear systems $Ax=b$ and $PAx = Pb$ are called equivalent. So, the solution is not changed by left-multiplication with a nonsingular matrix.

We now show how to solve $Ax=b$ by systematically taking such products with special matrices P.

For simplicity we assume that the leading principal submatrices are all nonsingular.

We will discuss what to do when this is not the case later.

$A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \vdots \\ a_{31} & a_{32} & & \\ \vdots & \vdots & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \quad x = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Note that M_1 is always nonsingular

$$M_1 = \begin{pmatrix} 1 & & & \\ -m_{21} & 1 & & \\ -m_{31} & 0 & 1 & \\ \vdots & & & \ddots \\ -m_{n1} & 0 & \cdots & 0 \; 1 \end{pmatrix} \quad \text{where} \quad m_{j1} = +a_{j1}/a_{11} \quad j = 2 \ldots n$$

($a_{11} \neq 0$, it's the 1×1 leading principal submatrix)

$$M_1 A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots \\ 0 & a_{22} - \frac{a_{21}}{a_{11}} a_{12} & a_{23} - \frac{a_{21}}{a_{11}} a_{13} & \cdots \\ 0 & a_{32} - \frac{a_{31}}{a_{11}} a_{12} & a_{33} - \frac{a_{31}}{a_{11}} a_{13} & \cdots \\ \vdots & & & \\ 0 & & & \end{pmatrix}$$

$A^{(1)} = M_1 A$ then

$$\begin{cases} a^{(1)}_{1j} = a_{1j} & j = 1, \ldots, n \\ a^{(1)}_{j1} = 0 & j = 2, \ldots, n \\ a^{(1)}_{ij} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} & i, j = 2, n \end{cases}$$

$$A^{(1)} = \begin{pmatrix} a^{(1)}_{11} & a^{(1)}_{12} & \cdots & a^{(1)}_{1n} \\ 0 & a^{(1)}_{22} & \cdots & a^{(1)}_{2n} \\ 0 & \vdots & & \vdots \\ 0 & a^{(1)}_{n2} & \cdots & a^{(1)}_{nn} \end{pmatrix}$$

We now repeat the process for the $(n-1)\times(n-1)$ submatrix $\begin{pmatrix} a^{(1)}_{22} & \cdots & a^{(1)}_{2n} \\ \vdots & & \vdots \\ a^{(1)}_{n2} & \cdots & a^{(1)}_{nn} \end{pmatrix}$

First, we prove that $a^{(1)}_{22} \neq 0$
The argument will hold for all subsequent steps in an analogous way.

By assumption the 2×2 leading principal submatrix $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ of A is nonsingular

$$\begin{pmatrix} 1 & 0 \\ -m_{21} & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22}^{(1)} \end{pmatrix}$$

If $a_{22}^{(1)} = 0$ then the matrix on the right is singular whereas both factors on the left are nonsingular. Can this happen?

A, B nonsingular. Assume AB singular.

$\rightarrow \exists\, x \neq 0 : (AB)x = 0 \Rightarrow A(Bx) = 0$

B nonsingular $\Rightarrow Bx \neq 0$, but then
A " $\Rightarrow A(Bx) \neq 0$
Contradiction!
Hence AB also nonsingular.

So, $\begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22}^{(1)} \end{pmatrix}$ nonsingular and $a_{22}^{(1)} \neq 0$.

$$M_2 = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & -m_{32} & 1 & & \\ \vdots & -m_{42} & 0 & 1 & \\ 0 & -m_{n2} & 0 & \cdots & 0\,1 \end{pmatrix} \quad \text{where} \quad m_{j2} = \frac{a_{j2}^{(1)}}{a_{22}^{(1)}}, \quad j = 3, 4, \cdots, n$$

$$M_2 A^{(1)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \\ 0 & 0 & a_{33}^{(1)} - a_{32}^{(1)}/a_{22}^{(1)}\, a_{23} & \cdots \\ \vdots & 0 & a_{43}^{(1)} - a_{42}^{(1)}/a_{22}^{(1)}\, a_{23} & \\ \vdots & & \vdots & \end{pmatrix}$$

$A^{(2)} = M_2 A^{(1)}$ $\quad a_{1j}^{(2)} = a_{1j}^{(1)} = a_{1j} \quad j = 1, \cdots, n$

$a_{2j}^{(2)} = a_{2j}^{(1)} \quad j = 1, \cdots, n$

$(a_{21}^{(1)} = 0)$

$a_{j1}^{(2)} = 0 \quad j = 2, \cdots, n$

$a_{j2}^{(2)} = 0 \quad j = 3, \cdots, n$

$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}} a_{2j}^{(1)}$

$i, j = 3, \cdots, n$

Schematically:

$A^{(1)}$

$$\left(\begin{array}{cc|ccc} x & x & x & x & x \\ 0 & x & x & x & x \\ \hline 0 & x & x & x & x \\ \vdots & \vdots & & & \\ 0 & x & x & x & x \end{array}\right) \longrightarrow$$

$A^{(2)}$

$$\left(\begin{array}{cc|ccc} x & x & x & x & x \\ 0 & x & x & x & x \\ \hline 0 & 0 & * & * & * \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & * & * & * \end{array}\right)$$

unchanged

set to zero — updated

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -m_{32} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} =$$

$$\begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} \\ 0 & 0 & a_{33}^{(2)} \end{pmatrix} \longrightarrow \text{nonsingular} \longrightarrow a_{33}^{(2)} \neq 0$$

Hence we can repeat the process again.

$$M_3 = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & -m_{43} & 1 & & \\ & & -m_{53} & 0 & 1 & \\ & & \vdots & & & \ddots \\ & & -m_{n3} & 0 & \cdots & 1 \end{pmatrix}$$ etc

where $m_{j3} = a_{j3}^{(2)} / a_{33}^{(2)}$

$$A^{(k)} = M_k A^{(k-1)} \qquad (A^{(0)} = A)$$

$$M_k = \left(\begin{array}{c|c|c} I_{k-1} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & 0 \\ \hline 0 \cdots 0 & 1 & 0 \\ \hline & -m_{k+1\,k} & \\ 0 & -m_{k+2\,k} & I_{n-k} \\ & \vdots & \\ & -m_{n\,k} & \end{array}\right)$$

where $m_{jk} = a_{jk}^{(k-1)} / a_{kk}^{(k-1)}$

$M_k A^{(k-1)}$ sets the k-th column below the diagonal coefficient in $A^{(k-1)}$ to zero and updates the lower right submatrix.

Step 1 : $Ax = b \rightarrow M_1 Ax = M_1 b \rightarrow A^{(1)} x = b^{(1)}$

Step 2 : $M_2 A^{(1)} x = M_2 b^{(1)} \rightarrow A^{(2)} x = b^{(2)}$

etc

$b^{(1)} = M_1 b$

$$\begin{pmatrix} 1 & & & \\ -m_{21} & 1 & & \\ -m_{31} & 0 & 1 & \\ \vdots & \vdots & 0 & \ddots \\ -m_{n1} & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 - m_{21} b_1 \\ b_3 - m_{31} b_1 \\ \vdots \\ b_n - m_{n1} b_1 \end{pmatrix}$$

etc

Note that $A^{(n)}$ is upper triangular and therefore

$$A^{(n)} x = b^{(n)}$$

is easy to solve.

This whole process is called Gaussian Elimination (without pivoting).

Now $M_{n-1} M_{n-2} \cdots M_1 A = A^{(n)} = U \Rightarrow$

$$A = M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1} U$$

Verify $M_k = I - m_k e_k^T$

where $e_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ k-th canonical basis vector

$$m_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ m_{k+1\,k} \\ \vdots \\ m_{n\,k} \end{pmatrix}$$

$$C^{m \times n} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} (b_1 \cdots b_n)$$

$$c_{ij} = a_i b_j$$

$$m_k e_k^T = \left(\begin{array}{c|c|c} 0 \cdots 0 & 0 & 0 \cdots 0 \\ \vdots & \vdots & \vdots \\ 0 \cdots 0 & 0 & 0 \cdots 0 \\ \hline 0 \cdots 0 & 0 & 0 \cdots 0 \\ \vdots \quad \vdots & m_{k+1\,k} & 0 \cdots 0 \\ 0 \cdots 0 & \vdots & \vdots \\ & m_{n\,k} & 0 \cdots 0 \end{array}\right)$$

Now it is easy to verify that

$M_k = I - m_k e_k^T$ and $M_k^{-1} = I + m_k e_k^T$

$$M_k^{-1} M_k = (I + m_k e_k^T)(I - m_k e_k^T)$$

$$= I - m_k e_k^T + m_k e_k^T - m_k \underbrace{e_k^T m_k} e_k^T$$

$$= I$$

Note $e_k^T m_k$ is the k-th coefficient in m_k, which is zero.

By construction M_k (and M_k^{-1}) are always nonsingular

Now it is easy to verify that

$M_k = I - m_k e_k^T$ and $M_k^{-1} = I + m_k e_k^T$

$M_k^{-1} M_k = (I + m_k e_k^T)(I - m_k e_k^T)$

$= I - m_k e_k^T + m_k e_k^T - m_k \underbrace{e_k^T m_k}_{} e_k^T$

$= I$ (check $M_k M_k^{-1}$)

Note $e_k^T m_k$ is the k-th coefficient in m_k, which is zero.

By construction M_k (and M_k^{-1}) are always nonsingular

$M_k^{-1} M_{k+j}^{-1} = (I + m_k e_k^T)(I + m_{k+j} e_{k+j}^T)$

$= I + m_k e_k^T + m_{k+j} e_{k+j}^T + m_k \underbrace{e_k^T m_{k+j}}_{m_{k,k+j} = 0} e_{k+j}^T$

$= I + m_k e_k^T + m_{k+j} e_{k+j}^T$

$M_1^{-1} M_2^{-1} M_3^{-1} \cdots = I + m_1 e_1^T + m_2 e_2^T + m_3 e_3^T + \cdots$

$$= \begin{pmatrix} 1 & & & 0 \\ m_{21} & 1 & & \\ m_{31} & m_{32} & 1 & \\ \vdots & \vdots & & \ddots \\ m_{n1} & m_{n2} & & 1 \end{pmatrix}$$

So, construction of factor $L = M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1}$ is trivial!

$A = LU$ where $\begin{cases} L = M_1^{-1} \cdots M_{n-1}^{-1} = I + \sum_{i=1}^{n-1} m_i e_i^T \\ U = A^{(n-1)} = M_{n-1} M_{n-2} \cdots M_1 A \end{cases}$

So, at end of Gaussian elimination:

Solve $A^{(n-1)} x = b^{(n-1)} \leftrightarrow Ux = b^{(n-1)}$

(solve by back substitution)

$b^{(n-1)} = M_{n-1} M_{n-2} \cdots M_1 b = L^{-1} b$

Solve for other right hand side

$Ax = f \iff LUx = f \iff Ux = L^{-1} f$
(result from G.E.)

$LUx = f$

substitute $y = Ux$ (some unknown vector)

i) Solve $Ly = f \rightarrow y = L^{-1} f$

ii) Solve $Ux = y \rightarrow x = U^{-1} y = U^{-1} L^{-1} f = A^{-1} f$

Example

$$\begin{pmatrix} 1 & 1 & -1 & 1 \\ 2 & 4 & 0 & 3 \\ 1 & -1 & -4 & 0 \\ 0 & 2 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 22 \\ -13 \\ 9 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & & \\ -2 & 1 & & \\ -1 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 1 \\ 2 & 4 & 0 & 3 \\ 1 & -1 & -4 & 0 \\ 0 & 2 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 & 1 \\ 0 & 2 & 2 & 1 \\ 0 & -2 & -3 & -1 \\ 0 & 2 & -1 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} 4 \\ 14 \\ -17 \\ 9 \end{pmatrix}$$

$$\begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 1 & 1 & \\ 0 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 1 \\ 0 & 2 & 2 & 1 \\ 0 & -2 & -3 & -1 \\ 0 & 2 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 & 1 \\ 0 & 2 & 2 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -3 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 4 \\ 14 \\ -3 \\ -5 \end{pmatrix}$$

$$\begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & -3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 1 \\ 0 & 2 & 2 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 & 1 \\ 0 & 2 & 2 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 4 \\ 14 \\ -3 \\ 4 \end{pmatrix}$$

---

a) $$\begin{pmatrix} 1 & 1 & -1 & 1 \\ 0 & 2 & 2 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 14 \\ -3 \\ 4 \end{pmatrix} \rightarrow \begin{cases} x_4 = 4 \\ x_3 = 3 \\ 2x_2 + 6 + 4 = 14 \rightarrow x_2 = 2 \\ x_1 + 2 - 3 + 4 = 4 \rightarrow x_1 = 1 \end{cases} \rightarrow \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

b) $$\begin{pmatrix} 1 & & & \\ 2 & 1 & & \\ 1 & -1 & 1 & \\ 0 & 1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 1 \\ 0 & 2 & 2 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 & 1 \\ 2 & 4 & 0 & 3 \\ 1 & -1 & -4 & 0 \\ 0 & 2 & -1 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 1 & & & \\ 2 & 1 & & \\ 1 & -1 & 1 & \\ 0 & 1 & 3 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 22 \\ -13 \\ 9 \end{pmatrix} \rightarrow \begin{cases} y_1 = 4 \\ y_2 = 14 \\ y_3 = -3 \\ y_4 = 4 \end{cases}$$

$Ux = y \rightarrow$ see (a)

What if one of the leading principle submat.s of A is singular

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \\ 0 & -1 & 1 & -1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & & & \\ -2 & 1 & & \\ +1 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \\ 0 & -1 & 1 & -1 \end{pmatrix}$$

$$= \left(\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \\ \hline 0 & 1 & 2 & 0 \\ 0 & -1 & 1 & -1 \end{array}\right) \quad \underline{\text{stuck}}$$

swap row 3 and row 2 to get nonzero pivot

(what if whole column is zero?)

↳ does not change solution

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$ (permutation matrix)

note $PP = I$ $P^TP = I$

&

$$\left(\begin{array}{c|ccc} 1 & 0 & 1 & 0 \\ \hline 0 & 1 & 2 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & -1 & 1 & -1 \end{array}\right) \quad \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & -1 & 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 3 & -1 \end{pmatrix}$$

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

swap rows 2 & 3

$$Ax = \begin{pmatrix}1\\1\\1\\1\end{pmatrix} \rightarrow \begin{pmatrix}1&&&\\-2&1&&\\1&\cdot&1&\\0&\cdot&\cdot&1\end{pmatrix}\begin{pmatrix}1\\1\\1\\1\end{pmatrix} = \begin{pmatrix}1\\-1\\2\\1\end{pmatrix} \rightarrow \begin{pmatrix}1\\2\\-1\\1\end{pmatrix}$$

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

compute corresponding rhs'

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

$$\begin{pmatrix}1&0&1&0\\0&1&2&0\\0&0&-2&1\\0&0&0&1/2\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix} = \begin{pmatrix}1\\2\\-1\\3/2\end{pmatrix}$$

$1/2\, x_4 = 3/2 \rightarrow x_4 = 3$

$\rightarrow -2x_3 + 3 = -1 \rightarrow x_3 = 2$

$x_2 + 2x_3 = 2 \rightarrow x_2 = -2$

$x_1 + x_3 = 1 \rightarrow x_1 = -1$

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

$$\left(\begin{pmatrix}1&&&\\-1&1&&\\2&0&1&\\0&-1&-3/2&1\end{pmatrix}\begin{pmatrix}1&0&1&0\\0&1&2&0\\0&0&-2&1\\0&0&0&1/2\end{pmatrix} = \begin{pmatrix}1&0&1&0\\-1&1&1&0\\2&0&0&1\\0&-1&1&-1\end{pmatrix}\right)$$ swapped

---

$$\begin{pmatrix}1&0&0&0\\-1&1&0&0\\+2&0&1&0\\0&-1&-3/2&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix} = \begin{pmatrix}1\\1\\1\\1\end{pmatrix} \rightarrow \begin{cases}y_1 = 1\\y_2 = 2\\y_3 = -1\end{cases}$$

$-2 + 3/2 + y_4 = 1$

$y_4 = 3/2$

$$Ax = \begin{pmatrix}1\\2\\3\\4\end{pmatrix} \rightarrow PA = Pb \rightarrow LU = \begin{pmatrix}1\\3\\2\\4\end{pmatrix}$$

$$\begin{pmatrix}1&&&\\-1&1&&\\2&0&1&\\0&-1&-3/2&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix} = \begin{pmatrix}1\\3\\2\\4\end{pmatrix} \rightarrow \begin{cases}y_1 = 1\\y_2 = 4\\y_3 = 0\\y_4 = 8\end{cases}$$

$$\begin{pmatrix}1&0&1&0\\&1&2&0\\&&-2&1\\&&&1/2\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix} = \begin{pmatrix}1\\4\\0\\8\end{pmatrix} \begin{cases}x_4 = 16\\x_3 = +8\\x_2 = 20 - 12\\x_1 = 4 - 7\end{cases}$$

So, we may need to change rows → partial pivoting

For accuracy we pivot if any coefficient in pivot column (below diagonal) is larger than pivot itself (in absolute value).

Explanation later

In step k we must swap rows k and k+j ($j \geq 1$)

$$P_k = \begin{pmatrix} e_1^T \\ \vdots \\ e_{k-1}^T \\ e_{k+j}^T \\ e_{k+1}^T \\ \vdots \\ e_k^T \\ \vdots \end{pmatrix} \qquad \text{note } P_k^{-1} = P_k$$

$$A^{(1)} = M_1 A^{(0)} \quad \cdots \rightarrow \quad A^{(k-1)} = M_{k-1} A^{(k-2)}$$

$$A^{(k)} = \underbrace{M_k P_k M_{k-1} M_{k-2} \cdots M_1}_{\rightarrow\ M_1^{-1} M_2^{-1} \cdots M_{k-1}^{-1} P_k M_k^{-1}} A$$

$$= (I + m_1 e_1^T + \cdots + m_{k-1} e_{k-1}^T) P_k (I + m_k e_k^T)$$

$$U = M_{n-1} \cdots M_k P_k M_{k-1} \cdots M_1 A$$

$$A = LU \rightarrow L = M_1^{-1} \cdots M_{k-1}^{-1} P_k M_k^{-1} \cdots M_{n-1}^{-1}$$

It turns out we can assume all pivoting done in advance and elementary elimination matrices "adapted" for that case. → move all perm. mat.s to right.

P_2 swaps rows 2 and 3

Note that permutation matrix at step k, P_k, leaves rows and columns 1 -- k-1 unchanged.

3x3 matrix : $U = M_2 P_2 M_1 P_1 A$

$\rightarrow U = M_2 \hat{M}_1 P_2 P_1 A \rightarrow U = M_2 \hat{M}_1 (PA)$

with $L = \hat{M}_1^{-1} M_2^{-1} \rightarrow LU = PA$

$P_2 M_1 = \hat{M}_1 P_2 \iff P_2 M_1 P_2 = \hat{M}_1$

$\rightarrow P_2 (I - m_1 e_1^T) P_2 = I - \underbrace{P_2 m_1}_{\hat{m}_1} \underbrace{e_1^T P_2}_{e_1^T}$

$= I - \hat{m}_1 e_1^T$

Note that $\hat{m}_1$ is just m_1 with coefficients 2 and 3 swapped.

So, $\hat{M}_1$ is elementary elimination matrix for matrix A with rows 2 and 3 swapped.

More generally $P_k M_i = \hat{M}_i P_k$

where $\hat{M}_i = P_k M_i P_k = I - (P_k m_i) e_i^T$

$M_3 P_3 M_2 P_2 M_1 P_1 A = M_3 P_3 M_2 \hat{M}_1 P_2 P_1 A =$

$M_3 \hat{M}_2 \hat{\hat{M}}_1 P_3 P_2 P_1 A$, $\hat{\hat{M}}_1 = I - \underline{P_3 P_2 m_1} e_1^T$

$\rightarrow LU = PA$, $Ax = b \rightarrow PAx = Pb$

Solve $LUx = Pb$ (standard way)

Pb reorders right hand side.

Cost of LU (ignore rhs and solve)

$$A^{(k)} = \left(I - m_k e_k^T\right) A^{(k-1)}$$

$$\begin{pmatrix} I_k & 0 \\ -m_{k+1} & \\ \vdots & I \\ -m_n & \end{pmatrix} \begin{pmatrix} 0 & \square \\ 0 & \text{BR}^{n-k} - m_k e_k^T R^{n-k} \end{pmatrix}$$

cost $n-k$ div.s
$(n-k)^2 - 2$ $\Big\}$ $2(n-k)^2 + (n-k)$

$$C_n = C_{n-1} + 2(n-1)^2 + (n-1)$$
$$= C_{n-1} + 2n^2 - 4n + n + 2 - 1$$

$$C_n = C_{n-1} + 2n^2 - 3n + 1 \quad \rightarrow \quad C_n - C_{n-1} = 2n^2 - 3n + 1$$

$$C_n = a_n^3 + bn^2 + cn + d$$

$$C_{n-1} = a(n^3 - 3n^2 + 3n - 1) = an^3 - 3an^2 \ldots$$
$$+ b(n^2 - 2n + 1)$$
$$+ c(n-1)$$
$$+ d$$

$$C_n - C_{n-1} = +n^2 3a - 4\,3na + a$$
$$+ 2bn - b$$
$$+ c$$

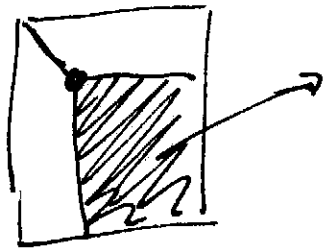
$$\begin{cases} 3a = 2 \\ -3a + 2b = -3 \\ -b + c = 1 \end{cases}$$

$a = 2/3$ $\quad d = ?$ (init. cond.)

$b = -1/2$

$c = 3/2$

part. pivoting <u>almost</u> always good enough.
when not: <u>complete pivoting</u> (rows + col.s)



search for largest absolute coeff.

P represents all row exchanges
Q " " column "

$$PAQ = LU$$

row changes — column changes

$P, Q, \rightarrow P^T P = I$

$Q^T Q = I$

(ex. of orthogonal mat.s)

Better even than compl. piv. is using QR fact. with Householder transf. / reflections
(see overdetermined systems)

$Ax = b$

software: $A \rightarrow P, L, U$ s.t. $PA = LU$ $(PAx = Pb)$

1) $\hat{b} = Pb$
2) $Ly = \hat{b}$ (solve)
3) $Ux = y$ (solve)

all pivoting in advance

$A \rightarrow P, Q, L, U$ s.t. $PAQ = LU$

$PAQ\tilde{x} = Pb$

1) $\hat{b} = Pb$ 2) $Ly = \hat{b}$ (solve) 3) $U\tilde{x} = y$

4) $x = Q\tilde{x}$

$y = L^{-1}\hat{b} \rightarrow \tilde{x} = U^{-1}L^{-1}\hat{b} \rightarrow$
$x = QU^{-1}L^{-1}\hat{b} = QU^{-1}L^{-1}Pb$

$$x = Q U^{-1} L^{-1} P b$$

$$LU = PAQ \rightarrow U^{-1} L^{-1} = Q^T A^{-1} P^T$$

$$x = Q Q^T A^{-1} P^T P b = \underline{\underline{A^{-1} b}}$$

(what we need)

Block-wise LU / GE

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

→ note special case where $A_{21} \in \mathbb{R}^{(n-1)\times 1}$

$$\begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{pmatrix}$$

$$= \begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 - A_{21}A_{11}^{-1} b_1 \end{pmatrix}$$

$S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ (Schur compl.)

note x_2 sol. "de coupled" from x_1

$$S x_2 = b_2 - A_{21}A_{11}^{-1} b_1 \rightarrow x_2$$

$$A_{11} x_1 = b_1 - A_{12} x_2$$

(note how all the indices match)

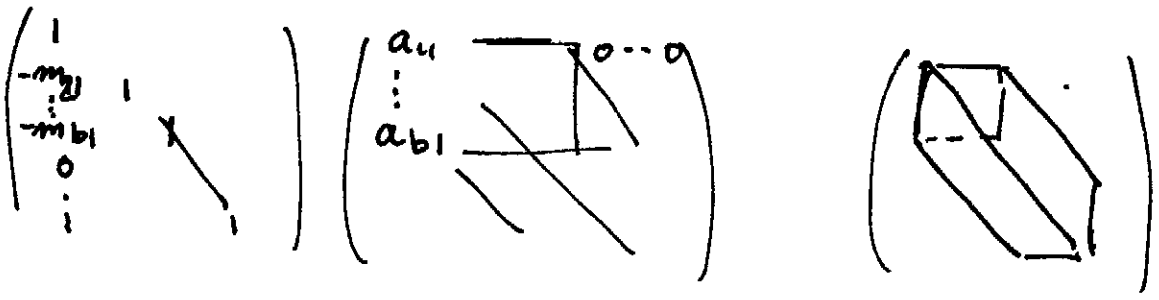
block inverse

$$\begin{pmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{pmatrix}\begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ 0 & \underbrace{A_{22} - A_{21}A_{11}^{-1}A_{12}}_{S} \end{pmatrix}$$

$$A^{-1} = U^{-1}L^{-1}$$

$$U^{-1} = \begin{pmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ 0 & S^{-1} \end{pmatrix}$$



verify that LU of banded system
(pivoting only within band) remains banded
(bandwidth grows by at most factor 2)
no pivoting ⇒ same semi bandwidth

So far we have not worried about accuracy of our computations. How to measure?

Solve $Ax = b$ using floating point $\rightarrow \hat{x}$
Exact solution : x

error : $\hat{x} - x$

How to measure size of error ?

Ex.

1) $\max_i |\hat{x}_i - x_i|$
2) $\sum_i |\hat{x}_i - x_i|$
3) $\left(\sum_i |\hat{x}_i - x_i|^2\right)^{1/2}$

} different measures for different purposes

We call such functions to measure the size of a vector norms.

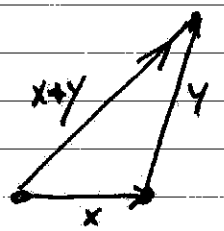
A function is a norm if it satisfies the following properties:

$\|x\| \in \mathbb{R} \geq 0$

$\|x\| = 0 \iff x = 0$

$\|\alpha x\| = |\alpha| \, \|x\| \quad (\alpha \in \mathbb{R})$

$\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)



An important property of norms that follows immediately from the triangle inequality is continuity.

A function $f: \mathbb{R}^n \to \mathbb{R}$ is lipschitz continuous with constant k if

$$|f(x) - f(y)| \leq k \|x - y\|$$

(for all x, y or in some neighborhood)

scalar case $|f(x) - f(y)| \leq k|x - y|$

Taking $f(x) \equiv \|x\|$ we get

a) $\|x\| = \|y + x - y\| \leq \|y\| + \|x - y\| \Rightarrow$

$\|x\| - \|y\| \leq \|x - y\|$

b) $\|y\| = \|x + y - x\| \leq \|x\| + \|x - y\| \Rightarrow$

$\|y\| - \|x\| \leq \|x - y\|$

From (a) and (b) it follows that

$$|\|x\| - \|y\|| \leq \|x - y\|$$

So, $\|\cdot\|$ is lipschitz continuous with constant $k = 1$.

①

$\|x\|_\infty = \max_i |x_i|$ → 1) $\|x\| = \max_i |x_i| \geq 0$

$\max_i |x_i| = 0 \Rightarrow x_j = 0 \ \forall j \Rightarrow x = 0$ (and v.v.)

2) $\|x\|_\infty = 0$ iff $x = 0$

3) $\|\alpha x\|_\infty = \max_i |\alpha x_i| = |\alpha| \max_i |x_i| = |\alpha| \|x\|$

4) $\|x+y\|_\infty = \max_i |x_i + y_i|$ → obtained for j:

$|x_j + y_j| \geq |x_i + y_i|$ (all i)

$\forall i \ |x_i + y_i| \leq |x_j + y_i| \leq |x_j| + |y_i| \leq$

$\max_k |x_k| + \max_k |y_k| = \|x\|_\infty + \|y\|_\infty$

② $\|x\|_1$ do yourself.

③ $\|x\|_2 = \left(\sum x_i^2\right)^{1/2}$ (or $(x^T x)^{1/2}$)

①+② obvious

③ obvious

④ $\|x + \tau y\|_2^2 = \sum_i (x_i + \tau y_i)^2 = \sum_i x_i^2 + 2\tau x_i y_i + \tau^2 y_i^2$

$= \tau^2 \sum y_i^2 + \tau \sum 2 x_i y_i + \sum x_i^2 \geq 0$

$\left(\sum 2 x_i y_i\right)^2 \leq 4 \left(\sum y_i^2\right)\left(\sum x_i^2\right) \Rightarrow$

$\sum x_i y_i \leq \|y\|_2 \, \|x\|_2$

Using $\sum x_i y_i \leq \|x\|_2 \cdot \|y\|_2$ we get

$$\|x+y\|_2^2 = \sum_i (x_i+y_i)^2 = \sum_i (x_i^2 + 2x_i y_i + y_i^2)$$

$$= \sum_i x_i^2 + 2\sum_i x_i y_i + \sum_i y_i^2$$

$$\leq \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 = (\|x\|_2 + \|y\|_2)^2$$

Therefore,

$$\|x+y\|_2 \leq \|x\|_2 + \|y\|_2$$

More generally, we define p-norm(3)

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

$$\|x\|_1 = \sum_i |x_i|$$

$$\|x\|_\infty = \lim_{p\to\infty} \|x\|_p = \max_i |x_i|$$

Matrix norms

As we saw earlier $\mathbb{R}^{m\times n}$ is a vector space.

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

$\|A\| \in \mathbb{R} \geq 0$ and $\|A\| = 0 \iff A = 0$

$\|\alpha A\| = |\alpha| \, \|A\|$ $(\alpha \in \mathbb{R})$

$\|A+B\| \leq \|A\| + \|B\|$

(same as for vectors)

Often (as in the book) we consider one extra property that distinguishes certain "nicer" norms.

$\|AB\| \leq \|A\| \cdot \|B\|$

Called consistency property or submultiplicative property

One important method to define matrix norms is to derive them from a particular vector norm. (induced matrix norm)

Since a matrix defines a "transformation" on vectors, it makes sense to measure the "size" of a matrix ~~to~~ in terms of this transformation.

$$\|A\|_\alpha \equiv \max_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}$$

Note that on the right hand side only vector norms are used.

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

Let $\|A\|_\alpha$ be induced matrix norm.

$\|A\|_\alpha \geq 0$ obvious

$\|A\|_\alpha = 0$ if $A = 0$ (zero matrix) also obvious

$A \neq 0 \Rightarrow \exists\, a_{ij} \neq 0$

Then $\|Ae_j\|_\alpha \neq 0$ as Ae_j is not a zero vector and vector norm is proper norm.

So, $\|A\|_\alpha = 0 \iff A = 0$

$$\|\gamma A\|_\alpha \equiv \max_{x \neq 0} \frac{\|\gamma A x\|_\alpha}{\|x\|_\alpha} = |\gamma| \max_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}$$

$$= |\gamma| \, \|A\|_\alpha$$

$$\|A+B\|_\alpha \equiv \max_{x \neq 0} \frac{1}{\|x\|_\alpha} \left(\|(A+B)x\|_\alpha \right)$$

$$= \max_{x \neq 0} \|x\|_\alpha^{-1} \|Ax + Bx\|_\alpha$$

$$\leq \max_{x \neq 0} \|x\|_\alpha^{-1} \left(\|Ax\|_\alpha + \|Bx\|_\alpha \right)$$

$$\leq \max_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha} + \max_{y \neq 0} \frac{\|By\|_\alpha}{\|y\|_\alpha}$$

$$= \|A\|_\alpha + \|B\|_\alpha$$

$$\|AB\|_\alpha \equiv \max_{x \neq 0} \frac{\|ABx\|_\alpha}{\|x\|_\alpha} = \max_{x \neq 0} \frac{\|A(Bx)\|_\alpha}{\|x\|_\alpha}$$

$$\overset{*}{\leq} \max_{x \neq 0} \frac{\|A\|_\alpha \, \|Bx\|_\alpha}{\|x\|_\alpha} = \|A\|_\alpha \, \|B\|_\alpha$$

* $\|A\|_\alpha \equiv \max_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}$

$\Rightarrow$

$\|A\|_\alpha \geq \frac{\|Ay\|_\alpha}{\|y\|_\alpha}$

$\Rightarrow$

$\|A\|_\alpha \|y\|_\alpha \geq \|Ay\|_\alpha$

(any y)

Frobenius norm: $\|A\|_F = \left(\sum_{ij} |a_{ij}|^2\right)^{1/2}$

Clearly $\|A\|_F \geq 0$ and $\|A\|_F = 0 \iff A = 0$

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

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

$$\|A+B\|_F^2 = \sum_{ij} (a_{ij}+b_{ij})^2 = \sum_{ij} \left(a_{ij}^2 + 2a_{ij}b_{ij} + b_{ij}^2\right)$$

Again consider $\|A+\tau B\|_F^2 =$

$$\sum_{ij} a_{ij}^2 + \sum_{ij} 2\tau a_{ij} b_{ij} + \sum_{ij} b_{ij}^2 \tau^2 =$$

$$\tau^2 \left(\sum_{ij} b_{ij}^2\right) + \tau \cdot 2 \sum_{ij} a_{ij} b_{ij} + \sum_{ij} a_{ij}^2 \geq 0$$

(quadratic in τ)

At most 1 root $\rightarrow$ discriminant ≤ 0

$$4\left(\sum_{ij} a_{ij} b_{ij}\right)^2 - 4\left(\sum_{ij} b_{ij}^2\right)\left(\sum_{ij} a_{ij}^2\right) \leq 0 \implies$$

$$\left(\sum_{ij} a_{ij} b_{ij}\right)^2 \leq \|A\|_F^2 \|B\|_F^2 \iff \sum_{ij} a_{ij} b_{ij} \leq \|A\|_F \|B\|_F$$

$$\rightarrow \|A+B\|_F^2 = \sum_{ij} \left(a_{ij}^2 + 2a_{ij}b_{ij} + b_{ij}^2\right) \leq \|A\|_F^2 + 2\|A\|_F \|B\|_F + \|B\|_F^2$$

$$= \left(\|A\|_F + \|B\|_F\right)^2$$

$$AB = C$$

$$\|AB\|_F^2 = \|C\|_F^2 = \sum_{ij} c_{ij}^2 = \sum_{ij} \left(\sum_k a_{ik} \cdot b_{kj} \right)^2$$

$$\leq \sum_{ij} \left(\sum_k a_{ik}^2 \cdot \sum_k b_{kj}^2 \right) =$$

$$\sum_{ik} a_{ik}^2 \cdot \sum_{jk} b_{kj}^2 = \|A\|_F^2 \; \|B\|_F^2$$

$$\longrightarrow$$

$$\|AB\|_F \leq \|A\|_F \, \|B\|_F$$

We are now ready to consider the accuracy of the computed solution

2 types of error:
a) data error
b) computational error

- both are unavoidable in numerical computation with floating point numbers

- a good algorithm should keep (b) at modest level for problems that are not very sensitive (see below)

2 types of influence on error in final solution:
a) sensitivity of problem to changes in initial data (also conditioning)

b) sensitivity of problem to errors made during the computation (stability)

- (a) is a problem characteristic that we cannot change (in general)

- (b) is an algorithm characteristic that we should always check/analyze
An algorithm that greatly amplifies small errors as it proceeds is useless.
This is, of course, not a black or white issue, but one with shades of gray for most problems

In most cases we are interested in relative error and therefore the influence of sensitivity and stability on relative error.

In naive approach (and often useful) we consider at each step of an algorithm for hypothetical problem the worst case influence on final answer and relative error.

Typically bound on relative error quickly explodes, even for algorithms that behave well for almost all problems.

Problem: analysis does not distinguish between influence of computational error and the sensitivity of the problem.

Even if propagation of error in algorithm is very good (accumulation and amplification are small), the relative error in final answer caused by small accumulated error may be large if problem is sensitive.

Alternative: proceed in 2 steps

1) Consider computed solution as exact solution of related problem

   and bound distance between related problem and original problem

2) Use sensitivity of problem to bound effect of solving related problem instead of original problem. (perturbation analysis)

→ Backward Error Analysis

A good algorithm (stable) will compute the solution for a nearby problem. So, small bound on distance between original and related problem.

Sensitivity of problem then determines effect on final answer.

Example: sum coefficients of vector in original order

sum (x_1, x_2, x_3) : $\rightarrow (x_1 + x_2) + x_3$

$y = 0$; $y = y + y_1$; $y = y + y_2$; $y = y + y_3$;

Consider $x_1 = 1$, $x_2 = 10^{-20}$, $x_3 = -1$

rounded answer of $x_1 + x_2 = 1$: $fl(x_1 + x_2)$

then $fl(x_1 + x_2) - 1 = 0$ $\rightarrow$ large relative error

However, consider the following analysis, where we assume each calculation has small relative error (IEEE arithmetic)

each $|\varepsilon_i| \leq \varepsilon_{mach}$ the machine round off

$fl(x_1 + x_2) = (x_1 + x_2)(1 + \varepsilon_1)$

$fl(fl(x_1 + x_2) + x_3) = ((x_1 + x_2)(1 + \varepsilon_1) + x_3)(1 + \varepsilon_2)$

$\rightarrow (x_1 + x_1\varepsilon_1 + x_1\varepsilon_2) + (x_2 + x_2\varepsilon_1 + x_2\varepsilon_2) + (x_3 + x_3\varepsilon_2) + H.O.T.$

So, result is exact for slightly different problem.

problem (instance) is ill-conditioned

The summation of coefficients of vector $(1, 10^{-20}, -1)$ is very sensitive to small changes in input.

$(1, 10^{-10}, -1) \rightarrow 10^{-10}$ rel. change in output $\approx 10^{10}$

rel. change in input $\approx \dfrac{\|(0, (10^{-20} - 10^{-10}), 0)^T\|_\infty}{\|(1, 10^{-20}, 1)^T\|_\infty} \approx 10^{-10}$

Forward (naive) error analysis:

$$|\text{computed answer} - \text{correct answer}| =$$

$$|x_1 + x_2 + x_3 + x_1\varepsilon_1 + x_1\varepsilon_2 + x_2\varepsilon_1 + x_2\varepsilon_2 + x_3\varepsilon_2 - (x_1 + x_2 + x_3)|$$

$$= |x_1(\varepsilon_1 + \varepsilon_2) + x_2(\varepsilon_1 + \varepsilon_2) + x_3\varepsilon_2|$$

$$\leq |x_1|\cdot(|\varepsilon_1| + |\varepsilon_2|) + |x_2|(|\varepsilon_1| + |\varepsilon_2|) + |x_3||\varepsilon_3|$$

$$\leq |x_1| 2\varepsilon_{mach} + |x_2| 2\varepsilon_{mach} + |x_3|\varepsilon_{mach}$$

$$= (2|x_1| + 2|x_2| + |x_3|)\varepsilon_{mach}$$

$$|\text{relative error}| \leq \frac{(|x_1| + |x_2| + |x_3|) 2\varepsilon_{mach}}{|x_1 + x_2 + x_3|}$$

This may be large if $|x_1 + x_2 + x_3|$ is small relative to $|x_1| + |x_2| + |x_3|$.

(For this simple problem it is clear that source of potentially large relative error is due to problem sensitivity, but forward error analysis itself does not reveal this)

Backward error analysis:

1) "computed problem" ~~(x_1(ε_1+ε_2), x_2(ε_1+ε_2), x_3(ε_2))^T~~

sum $(x_1 + x_1(\varepsilon_1 + \varepsilon_2), x_2 + x_2(\varepsilon_1 + \varepsilon_2), x_3 + x_3(\varepsilon_2))^T$

"original problem" sum $(x_1, x_2, x_3)^T$

distance in ~~max~~-norm:

$$\|(x_1(\varepsilon_1 + \varepsilon_2), x_2(\varepsilon_1 + \varepsilon_2), x_3\varepsilon_2)^T\|_\infty \leq$$

$$2\varepsilon_{mach}\|(x_1, x_2, x_3)^T\|_\infty$$

→ only slightly perturbed problem

Changes of $O(\varepsilon_{mach})$ in problem are unavoidable.

2) Conditioning or sensitivity of problem

As already shown, problem is very sensitive to small (relative) changes in input.

Not that changes of $O(\varepsilon_{mach})$ in 1st or 3rd coefficient of vector also lead to large relative changes in output.

↓

a "smarter" addition algorithm does not cure this problem.

Conditioning of linear system

Original problem: $A, b : \rightarrow x = A^{-1}b$
or solution of $Ax = b$

Perturbed problem: $(A+E)(x+e) = (b+\delta)$

perturbed input: $(A+E), (b+\delta)$
output: $x+e$

Relative error: $\frac{\|x+e-x\|}{\|x\|} = \frac{\|e\|}{\|x\|}$

$(A+E)(x+e) = b+\delta \Leftrightarrow Ax + Ex + Ae + Ee = b + \delta$

$\rightarrow Ae = \delta - Ex \Rightarrow e = A^{-1}\delta - A^{-1}Ex \Rightarrow$

$\|e\| = \|A^{-1}\delta - A^{-1}Ex\| \leq \|A^{-1}\delta\| + \|A^{-1}Ex\|$

$\Rightarrow \|e\| \leq \|A^{-1}\| \cdot \|\delta\| + \|A^{-1}\| \cdot \|E\| \cdot \|x\|$

$\Rightarrow \frac{\|e\|}{\|x\|} \leq \|A^{-1}\| \cdot \frac{\|\delta\|}{\|x\|} + \|A^{-1}\| \cdot \|E\| =$

$\|A^{-1}\| \cdot \|A\| \frac{\|\delta\|}{\|A\| \cdot \|x\|} + \|A^{-1}\| \cdot \|A\| \frac{\|E\|}{\|A\|}$

[illegible]

From $Ax = b \rightarrow \|Ax\| = \|b\| \Rightarrow$

$\|b\| \leq \|A\| \cdot \|x\| \rightarrow$ substitute above

$\frac{\|e\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \left(\frac{\|\delta\|}{\|b\|} + \frac{\|E\|}{\|A\|} \right)$

We call the number $\|A\| \|A^{-1}\|$ the condition number, often written $\kappa(A)$.

The condition number times the relative error in the input bounds the relative error in the output.

Example:
A exact, but b accurate to 3 digits $\Rightarrow$

$\frac{\|\delta b\|}{\|b\|} < 10^{-2}$ (expected of this size)

Let $\kappa(A) = 10^4$. Then bound on relative error: $\frac{\|e\|}{\|x\|} \leq 100$ (typically useless!)

$\Rightarrow$ we need more accurate b

Example:
A, b known exactly but not machine representable $\rightarrow$ relative error $O(\varepsilon_{mach})$

relative error in solution: $\frac{\|e\|}{\|x\|} \leq \kappa(A) 2\varepsilon_{mach}$

Note that the condition number depends on the chosen norm.

The relation between relative error in solution and relative change in input holds for any & consistent norm

(more precisely any consistent matrix norm with a consistent vector norm)

Backward error for linear system

Define residual $r = b - A\tilde{x}$

$\rightarrow A\tilde{x} = b - r$

So, if $\|r\| \ll \|b\|$ then $b - r$ is a small perturbation of b and $\tilde{x}$ is the exact solution of nearby problem.

Also $\left(A + \frac{r\tilde{x}^T}{\tilde{x}^T\tilde{x}}\right)\tilde{x} = A\tilde{x} + r = b$

$(A+E)\tilde{x} = b$ where $E = \frac{r\tilde{x}^T}{\tilde{x}^T\tilde{x}}$.

If $\|E\| \ll \|A\|$ then $\tilde{x}$ solution to nearby problem. This will generally be the case if $\|r\|$ small, although $\|\tilde{x}\|$ plays a role too.

$A^{-1}r = A^{-1}b - \tilde{x} = e$ (error)

$\|e\| \leq \|A^{-1}\| \cdot \|r\| = \kappa(A) \cdot \frac{\|r\|}{\|A\|}$

$\frac{\|e\|}{\|\tilde{x}\|} \leq \kappa(A) \cdot \frac{\|r\|}{\|A\| \cdot \|\tilde{x}\|}$, also $\frac{\|e\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}$

So, if residual sufficiently small and $\kappa(A)$ not too large, the relative error is small. If $\kappa(A)$ large, small residual may not mean much.

$(A+E)\tilde{x} = b \iff r = E\tilde{x} \Rightarrow \|r\| \leq \|E\| \cdot \|\tilde{x}\|$

$\frac{\|r\|}{\|A\| \cdot \|\tilde{x}\|} \leq \frac{\|E\|}{\|A\|}$

So a (relatively) large residual shows that algorithm used is not (backward) stable. Approx. solution is not solution of nearby problem.

Solving $Gy = b$

Let G be lower/upper triangular non-singular matrix. Let $\hat{y}$ be computed solution of forward/backward substitution. Then the following holds:

$$(G + \delta G)\hat{y} = b,$$

where $|\delta G| \leq 2n\varepsilon_{mach}|G| + O(\varepsilon_{mach}^2)$

$$|G| = \left(|g_{ij}|\right)_{i,j=1\ldots n}$$

$$\rightarrow \|\delta G\|_\infty \leq 2n\varepsilon_{mach}\|G\|_\infty + O(\varepsilon_{mach}^2)$$

Compute LU decomposition of A without pivoting during the computation (but possibly before).

$$A + E = \hat{L}\hat{U}$$

$u \equiv \varepsilon_{mach}$

where $|E| \leq 2nu|\hat{L}||\hat{U}| + O(u^2)$

and $\|E\|_\infty \leq 2nu\|\hat{L}\|_\infty\|\hat{U}\|_\infty + O(u^2)$

If we use this LU decomposition to solve $Ax = b$. Then computed $\hat{x}$ satisfies

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

where $|\delta A| \leq 6nu|\hat{L}||\hat{U}|$

and $\|\delta A\|_\infty \leq 6nu\|\hat{L}\|_\infty\|\hat{U}\|_\infty + O(u^2)$

So, stability of Gaussian elimination and solving by LU decomposition depends on

$$\|\hat{L}\|_\infty \text{ and } \|\hat{U}\|_\infty$$

This immediately shows importance of pivoting if the pivot is smaller than other coefficients in same column (below diagonal).

If a multiplier is large, $m_{ij} = \frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}}$, then $\hat{L}$ has large coefficient!

$$a_{ik}^{(j)} = a_{ik}^{(j-1)} - \frac{a_{ij}^{(j-1)} \cdot a_{jk}^{(j-1)}}{a_j^{(j-1)}}$$

Updating with large multiplier in general will create large coefficients in $\hat{U}$ as well.

Also, worse, growth of coefficients in

$$A^{(j-1)} \to A^{(j)}$$

can lead to significant growth in coefficients of $\hat{U}$.

Partial pivoting makes all multipliers (m_{ij}) less than or equal to 1. (good)

However, growth in coefficients of $\hat{U}$ still possible.
Worst case $O(2^{n-1})$ → basically, adding without amplifying can still grow

coefficients by factor 2 in each step.

Very rare, and partial pivoting generally method of choice.

Complete pivoting → $\|\hat{U}\| / \|A\| = O(n)$

For Cholesky decomposition $A = R^T R$ we have that $\|\hat{R}^T\|_F \|\hat{R}\|_F$ cannot be large relative to $\|A\|_F$ → unconditionally backward stable.

Alternative route to stability is iterative refinement.

$A = \hat{L}\hat{U} + E$

Solve $A\tilde{x}_1 = b$ (by LU) ; $r_1 = b - A\tilde{x}_1$

since error $e_1 = A^{-1} r_1$, we can ~~do~~ solve $\hat{L}\hat{U}\tilde{e}_1 = r_1$; $\tilde{x}_2 = x_1 + \tilde{e}_1$
$r_2 = b - A\tilde{x}_2$

etc.

If $\|r_k\|$ sufficiently small we have small backward error. Note that when $\|r\|$ is sufficiently small, r is dominated by round-off and iterative refinement does not improve solution.

Iterative refinement is not guaranteed to work if factorization is sufficiently bad. Also, expensive if many steps are needed.