LU Factorization

Let

$$A = \begin{bmatrix} a^{(1)} & a^{(2)} & \cdots & a^{(n)} \end{bmatrix} \in \mathbb{C}^{n \times n}.$$

$$A = L \; U$$

$n \times n$ = $n \times n$ lower triangular $n \times n$ upper triangular

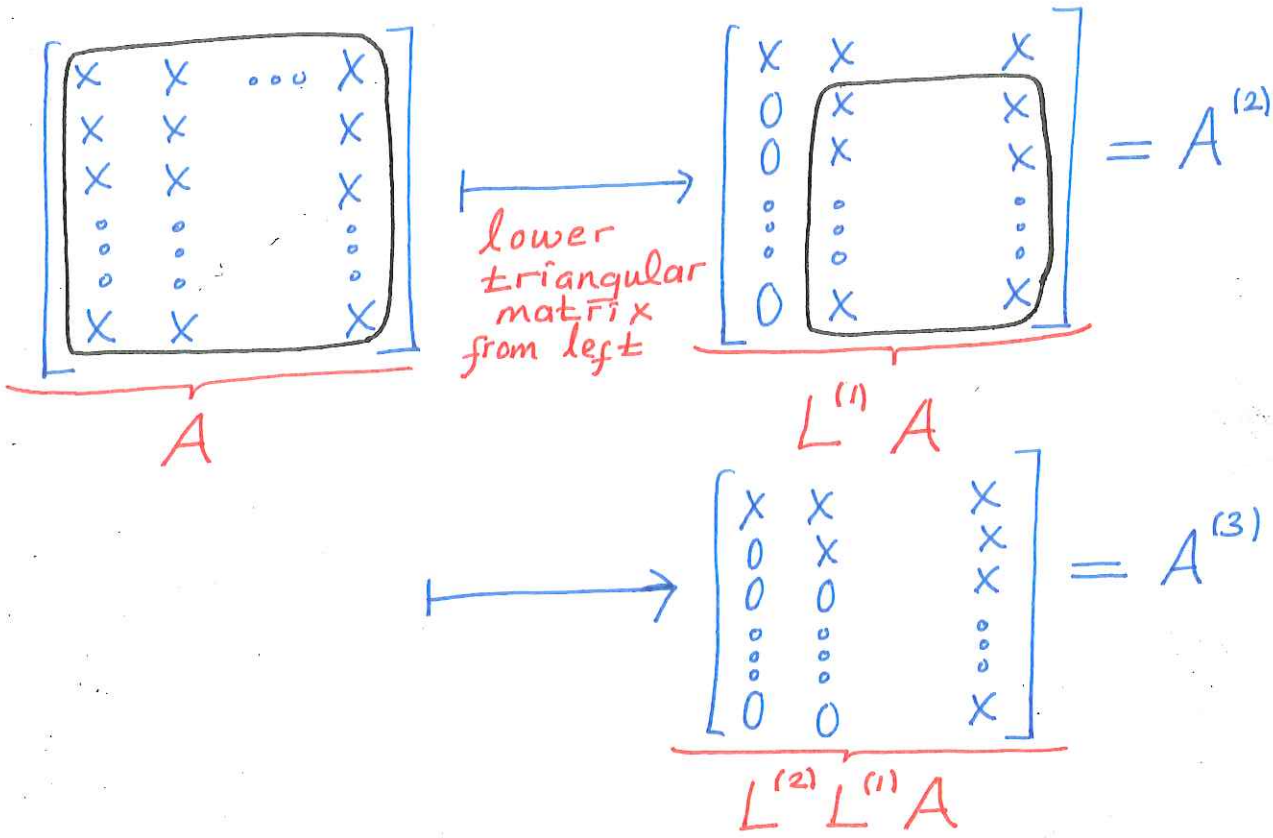
Ex

$$\underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{bmatrix}}_{A} \xrightarrow[r_3 := r_3 - 3r_1]{r_2 := r_2 - 2r_1} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -4 \\ 0 & -4 & -8 \end{bmatrix}$$

$$\xrightarrow{r_3 := r_3 - \frac{4}{3} r_2} \begin{bmatrix} 1 & 2 & 3 \\ 0_2 & -3 & -4 \\ 0_3 & 0_{4/3} & -8/3 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4/3 & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -4 \\ 0 & 0 & -8/3 \end{bmatrix}}_{U}$$

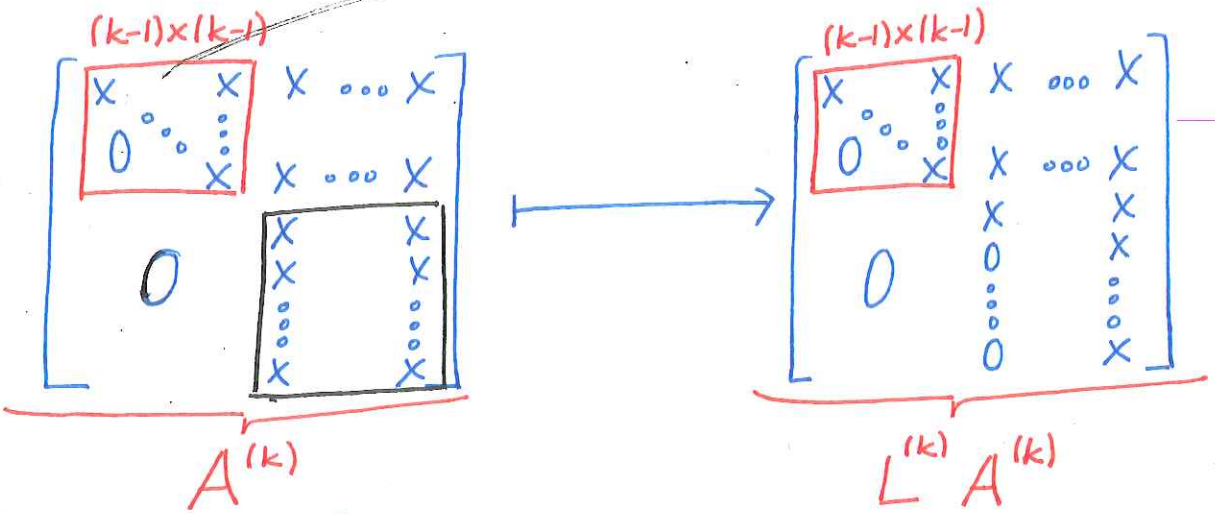
Triangularization by lower triangular matrices



$$L^{(1)} = \begin{bmatrix} 1 & & & \\ -l_{21} & 1 & & 0 \\ -l_{31} & 0 & \ddots & \\ \vdots & \vdots & & 1 \\ -l_{n1} & 0 & & \end{bmatrix}, \quad \begin{aligned} l_{21} &= a_{21} / a_{11} \\ l_{31} &= a_{31} / a_{11} \\ &\vdots \\ l_{n1} &= a_{n1} / a_{11} \end{aligned}$$

$$L^{(2)} = \begin{bmatrix} 1 & & & \\ 0 & 1 & & 0 \\ 0 & -l_{32} & \ddots & \\ \vdots & \vdots & & 1 \\ 0 & -l_{n2} & & \end{bmatrix}, \quad \begin{aligned} l_{32} &= a_{32}^{(2)} / a_{22}^{(2)} \\ &\vdots \\ l_{n2} &= a_{n2}^{(2)} / a_{22}^{(2)} \end{aligned}$$

kth step



$$= \left(L^{(k-1)} \cdots L^{(1)} A \right)$$

$$L^{(k)} = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & -l_{(k+1)k} & \ddots \\ & & \vdots & \\ & & -l_{nk} & \end{bmatrix}, \quad l_{(k+1)k} = a^{(k)}_{(k+1)k} / a^{(k)}_{kk}, \quad \vdots \quad l_{nk} = a^{(k)}_{nk} / a^{(k)}_{kk}$$

$$= I_n - l^{(k)} \left[e^{(k)} \right]^T$$

where

$$l^{(k)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ l_{(k+1)k} \\ \vdots \\ l_{nk} \end{bmatrix}, \quad e^{(k)} - k\text{th column of } I_n$$

After $(n-1)$ steps

$$L^{(n-1)} \cdots L^{(2)} L^{(1)} A = U$$

$$A = \underbrace{\left[L^{(1)}\right]^{-1}\left[L^{(2)}\right]^{-1} \cdots \left[L^{(n-1)}\right]^{-1}}_{L} \cdot U$$

Observe

$$L^{(k)} = I_n - \ell^{(k)}\left[e^{(k)}\right]^T$$

$$\left[L^{(k)}\right]^{-1} = I_n + \ell^{(k)}\left[e^{(k)}\right]^T$$

since

$$L^{(k)}\left[L^{(k)}\right]^{-1} = \left(I_n - \ell^{(k)}\left[e^{(k)}\right]^T\right)\left(I_n + \ell^{(k)}\left[e^{(k)}\right]^T\right)$$

$$= I_n - \ell^{(k)}\left[e^{(k)}\right]^T + \ell^{(k)}\left[e^{(k)}\right]^T - \ell^{(k)}\underbrace{\left[e^{(k)}\right]^T \ell^{(k)}}_{0}\left[e^{(k)}\right]^T$$

$$= I_n$$

Hence,

$$L = \underbrace{\left(I_n + \ell^{(1)}\left[e^{(1)}\right]^T\right)\left(I_n + \ell^{(2)}\left[e^{(2)}\right]^T\right) \cdots \left(I_n + \ell^{(n-1)}\left[e^{(n-1)}\right]^T\right)}_{I_n + \ell^{(1)}\left[e^{(1)}\right]^T + \ell^{(2)}\left[e^{(2)}\right]^T + \ell^{(1)}\left[e^{(1)}\right]^T \ell^{(2)}\left[e^{(2)}\right]^T}$$

$$= \underbrace{\left(I_n + \ell^{(1)}[e^{(1)}]^T + \ell^{(2)}[e^{(2)}]^T\right)\left(I_n + \ell^{(3)}[e^{(3)}]^T\right)}\cdots\left(I_n + \ell^{(n-1)}[e^{(n-1)}]^T\right)$$

$$I_n + \ell^{(1)}[e^{(1)}]^T + \ell^{(2)}[e^{(2)}]^T + \ell^{(3)}[e^{(3)}]^T$$

$$+ \ell^{(1)}\underbrace{[e^{(1)}]^T\ell^{(3)}}_{0}[e^{(3)}]^T + \ell^{(2)}\underbrace{[e^{(2)}]^T\ell^{(3)}}_{0}[e^{(3)}]^T$$

$$= I_n + \ell^{(1)}[e^{(1)}]^T + \cdots + \ell^{(n-1)}[e^{(n-1)}]^T$$

$$= \begin{bmatrix} 1 & & & 0 \\ \ell_{21} & 1 & & \\ \ell_{31} & \ell_{32} & 1 & \\ \vdots & \vdots & & \ddots \\ \ell_{n1} & \ell_{32} & & 1 \end{bmatrix}$$

due to $\ell^{(1)}[e^{(1)}]^T$ due to $\ell^{(2)}[e^{(2)}]^T$

Pseudocode (LU Factorization)

Input : $A \in \mathbb{C}^{n\times n}$

Output : A lower triangular $L \in \mathbb{C}^{n\times n}$
an upper triangular $U \in \mathbb{C}^{n\times n}$
such that $A = LU$.

$L \leftarrow I_n$
for $k = 1, 2, \cdots, n-1$
  for $j = k+1, \cdots, n$
    $\ell_{jk} \leftarrow a_{jk} / a_{kk}$
    $A(j, k{:}n) \leftarrow A(j, k{:}n) - \ell_{jk} A(k, k{:}n)$
  end
end
$U \leftarrow A$

(2) (1)

# of flops

① $\sim (n-k)$

② $\sim (n-k)$

Total # of flops

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

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

$$= \sum_{l=1}^{n-1} 2l^2 = 2\frac{(n-1)\,n\,(2n-1)}{6}$$

$$\sim 2n^3/3$$

W8 - P2

Condition Numbers

$f : V \to W$

V, W - vector spaces with norms

$$\kappa = \lim_{\delta \to 0^+} \sup_{\substack{\delta x \in V \\ 0 < \|\delta x\| \leq \delta}} \frac{\|f(x+\delta x) - f(x)\|}{\|\delta x\|}$$

κ: absolute condition number

$$\tilde{\kappa} = \lim_{\delta \to 0^+} \sup_{\substack{\delta x \in V \\ 0 < \|\delta x\| \leq \delta}} \frac{\|f(x+\delta x) - f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|}$$

$\tilde{\kappa}$: relative condition number

$$= \frac{\|x\|}{\|f(x)\|} \kappa$$

Examples

① Matrix-vector product

Using 2-norms on $\mathbb{C}^{m \times n}$ and $\mathbb{C}^m$

$f : \mathbb{C}^{m \times n} \to \mathbb{C}^m$

$f(A) = Ab$ for a given $b \in \mathbb{C}^n$

$$\kappa = \lim_{\delta \to 0^+} \sup_{0 < \|\delta A\|_2 \leq \delta} \frac{\|(A+\delta A)b - Ab\|_2}{\|\delta A\|_2}$$

$$= \lim_{\delta \to 0^+} \sup_{0 < \|\delta A\|_2 \leq \delta} \frac{\|\delta A\, b\|_2}{\|\delta A\|_2}$$

For all $\delta A \in \mathbb{C}^{m \times n}$

$$\frac{\|\delta A\, b\|_2}{\|\delta A\|_2} \leq \frac{\|\delta A\|_2 \, \|b\|_2}{\|\delta A\|_2} = \|b\|_2$$

Choosing $\delta A = \delta \dfrac{b b^*}{\|b\|^2}$

$$\frac{\left\| \left(\delta \frac{b b^*}{\|b\|^2} \right) b \right\|_2}{\underbrace{\left\| \delta \frac{b b^*}{\|b\|^2} \right\|}_{\delta}} = \|b\|_2$$

Hence,

$$K = \|b\|_2$$

$$\tilde{K} = \frac{\|b\|_2 \, \|A\|_2}{\|Ab\|_2}$$

If $A \in \mathbb{C}^{n \times n}$ and invertible, $b = A^{-1} A b$

$$\|b\|_2 \leq \|A^{-1}\|_2 \, \|Ab\|_2,$$

so

$$\tilde{K} \leq \|A^{-1}\|_2 \, \|A\|_2$$

Linear Systems

Solution x of

$$Ax = b$$

as a function of $b \in \mathbb{C}^n$ (and a given invertible A) $\in \mathbb{C}^{n\times n}$

$$x : \mathbb{C}^n \to \mathbb{C}^n \qquad \left(\begin{array}{c}\text{Using 2-norms}\\ \text{on } \mathbb{C}^n\end{array}\right)$$
$$x(b) = A^{-1}b$$

$$\kappa = \lim_{\delta \to 0^+} \sup_{0 < \|\delta b\|_2 \le \delta} \frac{\|A^{-1}(b+\delta b) - A^{-1}b\|_2}{\|\delta b\|_2}$$

$$= \lim_{\delta \to 0^+} \sup_{0 < \|\delta b\|_2 \le \delta} \frac{\|A^{-1}\delta b\|_2}{\|\delta b\|_2} = \|A^{-1}\|_2$$

$$\|A^{-1}\|_2$$

$$\tilde{\kappa} = \frac{\|b\|_2}{\|A^{-1}b\|_2}\kappa = \frac{\|b\|_2\,\|A^{-1}\|_2}{\|A^{-1}b\|_2}$$

Exploiting $\|b\|_2 \le \|A\|_2\,\|A^{-1}b\|_2$

$$\tilde{\kappa} \le \|A\|_2\,\|A^{-1}\|_2$$

| W 8 - P3 |

Backward Stability

$f : V \longrightarrow W$ exact problem

$\hat{f} : V \longrightarrow W$ computed solution by the algorithm

The algorithm is called backward stable if

$$\hat{f}(x) = f(x + \delta x)$$

$$\exists\ \delta x \in V \quad \text{s.t.} \quad \frac{\|\delta x\|}{\|x\|} = O(\epsilon_{mach})$$

for all $x \in V$, $x \neq 0$.

Reminder

$g(\epsilon_{mach}) = O(h(\epsilon_{mach}))$

means there exists a constant c s.t.

$$|g(\epsilon_{mach})| \leq c\,|h(\epsilon_{mach})|$$

for all ϵ_{mach} close to 0.

e.g. $2\epsilon_{mach} = O(\epsilon_{mach})$

$2\epsilon_{mach} = O(\sqrt{\epsilon_{mach}})$

$g(\epsilon_{mach}) = o(h(\epsilon_{mach}))$

means $\lim_{\epsilon_{mach} \to 0^+} \frac{g(\epsilon_{mach})}{h(\epsilon_{mach})} = 0$

e.g. $2\epsilon_{mach} = o(\sqrt{\epsilon_{mach}})$

Examples

① Addition in IEEE floating point arithmetic

Suppose $x, y \in \mathbb{R}$ is representable in IEEE arith.

$$x \oplus y = fl(x+y)$$
$$= (x+y)(1+\epsilon)$$

$\exists \epsilon$ s.t. $|\epsilon| \leq \epsilon_{mach}$

(as $\{fl(x+y) - (x+y)\} / (x+y) \leq \epsilon_{mach}$)

Hence,

$$x \oplus y = (x+\delta x) + (y+\delta y)$$

where $\delta x = x\epsilon$ ~~ϵ_{mach}~~ $\delta y = y\epsilon$ ~~ϵ_{mach}~~

$$\left\| \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} \right\|_2 = |\epsilon| \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\|_2$$
$$\leq \epsilon_{mach} \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\|_2$$

Addition in IEEE arith. is backward stable.

② Inner Product

$$f(x) = x^T y$$
$$= x_1 y_1 + x_2 y_2$$

for a given $y \in \mathbb{R}^2$, $y \neq 0$.

$$\hat{f}(x) = (x_1 \otimes y_1) \oplus (x_2 \otimes y_2)$$

$$\hat{f}(x) = \{(x_1 y_1)(1+\epsilon_1)\} \oplus \{(x_2 y_2)(1+\epsilon_2)\}$$
$$= \left[\{(x_1 y_1)(1+\epsilon_1)\} + \{(x_2 y_2)(1+\epsilon_2)\}\right](1+\epsilon_3)$$
$$= (x_1 y_1)(1+\epsilon_1)(1+\epsilon_3) + x_2 y_2 (1+\epsilon_2)(1+\epsilon_3)$$
$$= (x_1 + \delta x_1) y_1 + (x_2 + \delta x_2) y_2$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are s.t. $|\epsilon_1|, |\epsilon_2|, |\epsilon_3| \leq \epsilon_{mach}$

$$\delta x_1 := x_1(1+\epsilon_1)(1+\epsilon_3) - x_1$$
$$= (\epsilon_1 + \epsilon_3) x_1 + O(\epsilon_{mach}^2) x_1$$
$$\delta x_2 := x_2(1+\epsilon_2)(1+\epsilon_3) - x_2$$
$$= (\epsilon_2 + \epsilon_3) x_2 + O(\epsilon_{mach}^2) x_2$$

Hence,

$$\hat{f}(x) = f(x + \delta x)$$

for some $\delta x = \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix}$ such that

$$\|\delta x\|_2 = \left\| \begin{bmatrix} (\epsilon_1+\epsilon_3)x_1 + O(\epsilon_{mach}^2)x_1 \\ (\epsilon_2+\epsilon_3)x_2 + O(\epsilon_{mach}^2)x_2 \end{bmatrix} \right\|_2 \leq 2\epsilon_{mach}\|x\|_2 + O(\epsilon_{mach}^2)\|x\|_2$$

$$\Rightarrow \|\delta x\|/\|x\| = O(\epsilon_{mach}).$$

Backward Error Analysis

Relative error

$$\widetilde{E} := \frac{\| \hat{f}(x) - f(x) \|}{\| f(x) \|}$$

Now suppose the algorithm is backward stable, i.e.

$$\hat{f}(x) = f(x + \widetilde{\delta x})$$

$$\exists\, \widetilde{\delta x} \quad \| \widetilde{\delta x} \| \leq c\, \epsilon_{mach} \| x \|$$

where c is a constant independent of ϵ_{mach}.

$$\widetilde{E} \leq \underbrace{\sup_{0 < \| \delta x \| \leq \delta} \frac{\| f(x+\delta x) - f(x) \| / \| f(x) \|}{\| \delta x \| / \| x \|}}_{\widetilde{\kappa}_\delta} \cdot \frac{\| \delta x \|}{\| x \|}$$

for $\delta := c\, \epsilon_{mach} \| x \|$

$$\widetilde{E} \leq \widetilde{\kappa}_\delta \frac{\| \widetilde{\delta x} \|}{\| x \|} \implies \boxed{\widetilde{E} = \widetilde{\kappa}_\delta \cdot O(\epsilon_{mach})}$$

Example

Inner product in $\mathbb{R}^n$

$$f : \mathbb{R}^n \to \mathbb{R}$$

$$f(x) = x^T y$$

$$= x_1 y_1 + \cdots + x_n y_n$$

for a given $y \in \mathbb{R}^n$.

$$\tilde{K}_\delta = \sup_{0 < \|\delta x\| \leq \delta} \frac{|(x+\delta x)^T y - x^T y| \,/\, |x^T y|}{\|\delta x\| \,/\, \|x\|}$$

$$= \sup_{0 < \|\delta x\| \leq \delta} \frac{|\delta x^T y| \; \|x\|}{\|\delta x\| \; |x^T y|}$$

Observe

$$|\delta x^T y| \leq \|\delta x\| \; \|y\| \qquad \forall \, \delta x$$

Specifically, for $\delta x_* := (\delta) \cdot y \,/\, \|y\|$

$$|\delta x_*^T y| = \delta \cdot \|y\| = \|\delta x_*\| \; \|y\|.$$

Hence

$$\tilde{K}_\delta = \frac{\|y\| \; \|x\|}{|x^T y|} = \frac{1}{|\cos \theta|}$$

where θ is the angle between x and y.

Computed in IEEE arith.

$$\hat{f}(x) = (x_1 \otimes y_1) \oplus \cdots \oplus (x_n \otimes y_n)$$

It can be shown that (as we have done for the $\mathbb{R}^2$ case above)

$$\hat{f}(x) = f(x + \tilde{\delta x})$$

$$\exists\, \delta x \text{ s.t. } \frac{\|\tilde{\delta x}\|}{\|x\|} = n\,\epsilon_{mach} + O(\epsilon_{mach}^2)$$

It follows that

$$\frac{|\hat{f}(x) - f(x)|}{|f(x)|} \leq \frac{1}{|\cos\theta|}\{n\,\epsilon_{mach} + O(\epsilon_{mach}^2)\}$$

$$= \frac{1}{|\cos\theta|} O(\epsilon_{mach})$$