LECTURE 1

Sept 29th, 2025

<u>DEF</u>: (Norm) A function $\|\cdot\| : V \to \mathbb{R}$ s.t.

(i) $\|v\| \geq 0 \quad \forall v \in V$ and $\underline{\|v\| = 0 \iff v = 0}$, definiteness

(ii) $\|\lambda v\| = |\lambda| \, \|v\| \quad \forall v \in V, \ \lambda \in \mathbb{R}$

(iii) $\|v + w\| \leq \|v\| + \|w\| \quad \forall w, v \in V$

CONVENTION: $x \in \mathbb{R}^n$ is treated as column vector $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

<u>EXAMPLES</u> (VECTOR NORMS): For $x \in \mathbb{R}^n$

- ℓ^2 NORM: $\|x\|_2 := \sqrt{\sum_{i=1}^{n} |x_i|^2}$ (EUCLIDEAN NORM)

- ℓ^p NORM: $\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad 1 \leq p < \infty$

- ℓ^∞ NORM: $\|x\|_\infty := \max_{1 \leq i \leq n} |x_i| = \lim_{p \to \infty} \|x\|_p$

- For $w = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \in \mathbb{R}_{>0}$, define

$$\|x\|_{p,w} := \left(\sum_{i=1}^{n} w_i |x_i|^p \right)^{1/p} \quad \text{(WEIGHTED p-NORM)}$$

In particular, $p = 2$

$$\|x\|_{2,w} = \sqrt{\sum_{i=1}^{n} w_i |x_i|^2} = \sqrt{x^T \underbrace{W}_{} x}$$

$W = \operatorname{diag}(w_1, \ldots, w_n)$

- <u>MAHALANOBIS</u> <u>NORM</u>: for $A \in \mathbb{R}^{n \times n}$ positive definite, $\|x\|_A := \sqrt{x^T A x}$ is a norm.

$x^T = (x_1 \cdots x_n)$ row vector

- MATRIX NORMS ← These are now fcts $\|\cdot\|: \mathbb{R}^{m\times n} \to \mathbb{R}$

$$\mathbb{R}^{m\times n} \ni X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

DEF: (Matrix Norm) A matrix norm is a norm on $\mathbb{R}^{m\times p}$ (i.e., the properties (i) – (iii) above) with the additional property of being SUBMULTIPLICATIVE; i.e., $\|XY\| \leq \|X\| \, \|Y\|$, $\forall X \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{R}^{n\times p}$.

EXAMPLES (MATRIX NORMS) For $X \in \mathbb{R}^{m\times n}$

- $$\|X\|_{H,p} := \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij}|^p \right)^{1/p}$$ (HÖLDER p-NORM)

WARNING: Hölder ∞-norm is not submultiplicative

$\|X\|_{H,\infty} := \max\limits_{1\leq i\leq m} \max\limits_{1\leq j\leq m} |x_{ij}|$. But $X = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$, $Y = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$

$$XY = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}.$$

- $$\|X\|_F := \|X\|_{H,2} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |x_{ij}|^2} = \sqrt{tr(X^T X)}$$ (FROBENIUS NORM)

Frobenius norm is submultiplicative by Cauchy-Schwarz $\|XY\|_F \leq \|X\|_F \, \|Y\|_F$.

- Linear operator $X: (\mathbb{R}^n, \|\cdot\|_a) \to (\mathbb{R}^m, \|\cdot\|_b)$

$$v \mapsto Xv$$

DEF: The operator norm (or induced norm) is defined as

$$\|X\|_{a,b} := \max_{v\neq 0} \frac{\|Xv\|_b}{\|v\|_a} = \max_{\|v\|_a = 1} \|Xv\|_b = \max_{\|v\|_a \leq 1} \|Xv\|_b$$

(Says how far v can get away from original position after X applied)

b/c $\|\alpha X\| = |\alpha| \|X\|$ and linearity of X

PROP: (Consistency Property) $\|Xv\|_b \leq \|X\|_{a,b} \|v\|_a \quad \forall v \in \mathbb{R}^n, X \in \mathbb{R}^{m\times n}$

· Caveat: Not all norms are consistent.

NOTE: Take $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times p}$. Then $XY \in \mathbb{R}^{m\times p}$; i.e.,
$(\mathbb{R}^p, \|\cdot\|_c) \xrightarrow{Y} (\mathbb{R}^n, \|\cdot\|_b) \xrightarrow{X} (\mathbb{R}^m, \|\cdot\|_a)$. Then

$$\|XY\|_{a,c} \leq \|X\|_{a,b} \|Y\|_{b,c}.$$

DEF (MATRIX p,q-NORM) For $p,q \in [1,\infty]$ and $X \in \mathbb{R}^{m\times n}$, define

$$\|X\|_{p,q} := \max_{0 \neq v \in \mathbb{R}^n} \frac{\|Xv\|_q}{\|v\|_p}$$

Obs: $\|X\|_{1,\infty} = \|X\|_{H,\infty}$

· If $p=q$, we call this the MATRIX p-NORM.

EXAMPLES:

· MATRIX 1-NORM: $\|X\|_1 = \|X\|_{1,1} = \max_{1\leq j\leq n} \sum_{i=1}^{m} |x_{ij}|$ = max of ℓ^1 norms of columns of X

· MATRIX ∞-NORM: $\|X\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |x_{ij}|$ = max of ℓ^1 norms of rows of X

REMEMBER: "∞" looks like rows and "1" looks like column

· MATRIX 2-NORM: $\|X\|_2 \overset{\text{def}}{=} \max_{0\neq v\in\mathbb{R}^n} \frac{\|Xv\|_2}{\|v\|_2}$ ... no closed formula but we have facts about it

Can compute $p=1,2,\infty$ in polynomial time

$=$ largest singular value of X

OBS: For $p \neq 1,2,\infty$, NP-hard.

PF: (closed form formulas for $p = 1, 2, \infty$ norms) ← STUDY FOR TEST 2

- ($\leq$) $$\|X\|_\infty \overset{\text{def}}{=} \max_{0 \neq y \in \mathbb{R}^n} \frac{\|Xy\|_\infty}{\|y\|_\infty} = \max_{\|y\|_\infty = 1} \|Xy\|_\infty$$

$$= \max_{\|y\|_\infty = 1} \left[\max_{1 \leq i \leq m} \left| \sum_{j=1}^{n} x_{ij} y_j \right| \right]$$

Triangle Inequality

$$\leq \max_{\|y\|_\infty = 1} \left[\max_{1 \leq i \leq m} \sum_{j=1}^{n} |x_{ij}| \, |y_j| \right]$$

$\max\limits_{\|y\|_\infty = 1} \iff y_1, \dots, y_n \leq 1$

$$\leq \max_{1 \leq i \leq m} \sum_{j=1}^{n} |x_{ij}| .$$

($\geq$) Need to find some $y \in \mathbb{R}^n$, $\|y\|_\infty = 1$ such that this is satisfied.

Take $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$ s.t. $y_j := \begin{cases} |x_{ij}| / x_{ij}, & \text{for } x_{ij} \neq 0 \\ 0, & \text{for } x_{ij} = 0 \end{cases} = \operatorname{sgn}(x_{ij})$

Then $\|y\|_\infty = 1$ and $\|Xy\|_\infty = \sum_{j=1}^{n} |x_{ij}|$ ■

! LEARN THE OTHER ONES

LECTURE 2

Oct 1st, 2025

$\lim\limits_{k \to \infty} \|v_k\| = \|\lim\limits_{k \to \infty} v_k\|$

FACT: $\|\cdot\| : V \to \mathbb{R}$ is continuous. In fact, for $u, v \in V$,

$$\|u\| = \|u - v + v\| \leq \|u - v\| + \|v\|$$

$\|v\| \leq \|u-v\| + \|u\|$

So, $\big| \|u\| - \|v\| \big| \leq \|u-v\| \Rightarrow \|\cdot\|$ is Lipschitz

FACT: (EQUIVALENCE OF NORMS)

$$\boxed{\text{THM: For any } \underline{\text{finite}} \text{ dimensional space, } \underline{\text{all}} \text{ norms are equivalent.}}$$

DEF: $\|\cdot\|_a : V \longrightarrow \mathbb{R}$ and $\|\cdot\|_b : V \longrightarrow \mathbb{R}$ are equivalent iff $\exists\, c_1, c_2$ s.t. $0 < c_1 < c_2$ and

$$c_1 \|v\|_a \leq \|v\|_b \leq c_2 \|v\|_a \quad \forall v \in V.$$

COR: If $v_k \longrightarrow v$ in V $(\dim V < +\infty)$ in $\|\cdot\|_a$, then $v_k \to v$ in $\|\cdot\|_b$.

NOTE: $v_k \to v$ in $\|\cdot\|_a \Leftrightarrow \|v_k - v\|_a \to 0$ by equivalence of norm and the squeeze thm,

$$\underbrace{c_1 \|v - v_k\|_a}_{\downarrow\, 0} \leq \|v_k - v\|_b \leq \underbrace{c_2 \|v_k - v\|_a}_{\downarrow\, 0}$$

i.e., $\|v_k - v\|_b \to 0$

EXAMPLE: Take $\mathbb{R}^n$ with $\|\cdot\|_2$ and $\|\cdot\|_\infty$. Then

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\, \|x\|_\infty \qquad (c_1 = 1,\ c_2 = \sqrt{n})$$

Pf: $\|x\|_2^2 = \sum_{i=1}^{n} x_i$, $\quad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$. Say $\|x\|_\infty = |x_k|$ for some $k \in \{1, \dots, n\}$

So, since x_k is the largest entry,

$$x_k^2 \leq x_1^2 + \cdots + x_n^2 \leq n x_k^2$$

i.e., $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty$.

NOTE: $\frac{1}{n}\|x\|_1 \leq \|x\|_\infty \leq \|x\|_1$ PROVE THESE!

EXAMPLE: For $\|\cdot\|_F$ and $\|\cdot\|_{H,\infty}$ in $\mathbb{R}^{m\times n}$, we have

$$\|X\|_{H,\infty} \leq \|X\|_F \leq \sqrt{mn}\,\|X\|_{H,\infty}$$

- For $\|X\|_\infty = \max\{\ell^1 \text{ norm of rows}\}$ and $\|X\|_2 = \max_{y\neq 0} \frac{\|Xy\|_2}{\|y\|_2}$.

$$\frac{1}{\sqrt{m}}\|X\|_2 \leq \|X\|_\infty \leq \sqrt{n}\,\|X\|_2$$

- Note

$$\|X\|_2 \leq \|X\|_F \leq \sqrt{n}\,\|X\|_2$$

INNER PRODUCTS

DEF: $\langle \cdot,\cdot \rangle : V \times V \to \mathbb{R}$ such that

(i) $\langle v,v \rangle \geq 0 \quad \forall v \in V$ (Non-negativity)

(ii) $\langle v,w \rangle = 0 \quad \forall w \in V$ iff $v = 0$ (Non-degeneracy)

(iii) $\langle \alpha v + \beta w, \gamma x + \delta y \rangle = \alpha\gamma\langle v,x\rangle + \alpha\delta\langle v,y\rangle + \beta\langle w,x\rangle + \beta\delta\langle w,y\rangle$
$\forall v,w,x,y \in V$ and $\alpha,\beta,\gamma,\delta \in \mathbb{R}$. (Bilinear)

(iii) $\langle v, w\rangle = \langle w, v\rangle \quad \forall v, w \in V$ (symmetry)

<u>DEF</u>: For an inner product $\langle\cdot,\cdot\rangle: V \times V \to \mathbb{R}$, define the <u>induced norm</u> by the inner product as

$$\|v\| := \sqrt{\langle v, v\rangle} \quad \forall v \in V.$$

<u>THM</u>: (CAUCHY-SCHWARZ) If $\|\cdot\|$ is induced by $\langle\cdot,\cdot\rangle$, then

$$|\langle v, w\rangle| \leq \|v\| \, \|w\| \quad \forall v, w \in V.$$

<u>PROP</u>: (PARALLELOGRAM IDENTITY) If for all $v, w \in (V, \|\cdot\|)$ we have

$$\|v+w\|^2 + \|v-w\|^2 = 2\|v\|^2 + 2\|w\|^2,$$

then $\|\cdot\|$ is induced by an inner product. PROVE THIS

<u>EXAMPLE</u>: For $x, y \in \mathbb{R}^n$, the Euclidean inner product is $\langle x, y\rangle = x^T y$

- More generally, for $A \in \mathbb{R}^{n \times n}$ positive definite,

$$\langle x, y\rangle_A := \langle x, Ay\rangle = x^T A y$$

is an inner product.

<u>NOTE</u>: The Euclidean inner product induces the 2-norm

$$\sqrt{\langle x, x\rangle} = \sqrt{x^T x} = \|x\|_2$$

More generally, $\sqrt{\langle x, x\rangle_A} \overset{\text{def}}{=} \sqrt{\langle x, Ax\rangle} = \|x\|_A$ (Mahalanobis norm)

OBS: $\|\cdot\|_\infty$ is not induced by inner product.

EXAMPLE: (Matrix Inner Product) Take $\mathbb{R}^{m\times n}$. Then

$$\langle X, Y\rangle := \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}\, y_{ij} \quad \text{(Standard inner product on } \mathbb{R}^{m\times n})$$

$$= \operatorname{tr}(X^T Y)$$

NOTE: $\sqrt{\langle X, X\rangle} = \sqrt{\operatorname{tr}(X^T X)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2} \overset{\text{def}}{=} \|X\|_F$.

Cauchy-Schwarz: $|\operatorname{tr}(X^T Y)| \leq \|X\|_F \|Y\|_F$.

MATRIX PRODUCTS

1) Inner Product: $x^T y \in \mathbb{R}$

Always rank = 1

2) Outer Product: $xy^T = \begin{pmatrix} x_1y_1 & \cdots & x_1y_n \\ \vdots & \ddots & \vdots \\ x_my_1 & \cdots & x_my_n \end{pmatrix}_{m\times n} \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$.

3) Matrix-Matrix Product: $X \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{R}^{n\times p}$

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 & \cdots & y_p \end{pmatrix} \rightsquigarrow XY = \begin{pmatrix} x_1^T y_1 & \cdots & x_1^T y_p \\ \vdots & & \vdots \\ x_m^T y_1 & & x_m^T y_p \end{pmatrix} \in \mathbb{R}^{m\times p}.$$

4) Take $X \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{R}^{n\times p}$ and write

Matrix product is a sum of rank-1 matrices!

$$X = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}, \quad Y = \begin{pmatrix} y_1^T \\ \vdots \\ y_n^T \end{pmatrix} \rightsquigarrow XY = x_1 y_1^T + x_2 y_2^T + \cdots + x_n y_n^T \in \mathbb{R}^{m\times p}.$$

5) Composition of maps: $X : \mathbb{R}^n \to \mathbb{R}^m$, $X \in \mathbb{R}^{m\times n}$ and $Y = \begin{pmatrix} y_1 & \cdots & y_p \end{pmatrix} \in \mathbb{R}^{n\times p}$. Then $XY = \begin{pmatrix} Xy_1 & \cdots & Xy_p \end{pmatrix}$.

6) Dual composition: $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix} \in \mathbb{R}^{m\times n}$, $XY = \begin{pmatrix} x_1^T Y \\ \vdots \\ x_m^T Y \end{pmatrix}$.

SPECIAL CASES:

(A) $X = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \in \mathbb{R}^{m\times n}$ and $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$. Then

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

This means $\operatorname{im} X = \operatorname{col} X$.

(A') $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ and $Y = \begin{pmatrix} y_1^T \\ \vdots \\ y_n^T \end{pmatrix} \in \mathbb{R}^{m\times n}$. Then

$$x^T Y = x_1 y_1^T + \cdots + x_n y_n^T.$$

(B) $D = \operatorname{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n\times n}$, $X \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{R}^{n\times m}$

$$DX = \begin{pmatrix} d_1 x_1^T \\ \vdots \\ d_n x_n^T \end{pmatrix} \quad \text{and} \quad YD = \begin{pmatrix} d_1 y_1 & \cdots & d_n y_n \end{pmatrix}.$$

scales rows

scales columns

REMARK: $\|X\|_1 = \max\{\ell^1 \text{ norm of cols. of } X\}$

$X = (x_1 \cdots x_n)$ $\quad \|X\|_1 = \max_{\|y\|_1=1} \|Xy\|_1 = \max_{\|y\|_1=1} \|y_1 \vec{x}_1 + \cdots + y_n \vec{x}_n\|_1$

$$\leq \max_{\|y\|_1=1} \left(|y_1| \|\vec{x}_1\|_1 + \cdots + |y_n| \|\vec{x}_n\|_1 \right)$$

$$= \left[\max_{\|y\|_1=1} \left(|y_1| + \cdots + |y_n| \right) \right] \left[\max\left(\|x_1\|_1, \ldots, \|x_n\|_1 \right) \right]$$

LECTURE 3

EIGENVALUES & EIGENVECTORS

Oct 6th, 2025

"right eigenvector"

EIGENPROBLEM: $\underbrace{Ax}_{} = \lambda x$, $\quad A \in \mathbb{R}^{n \times n}$, $\quad x \in \mathbb{C}^n \setminus \{0\}$, $\quad \lambda \in \mathbb{C}$

- CONVENTION: Normalize eigenvectors to have $\|x\| = 1$.

DEF: (Eigenspace) $E_\lambda := \{ x \in \mathbb{C}^n : Ax = \lambda x \}$ ← set of all eigenvectors together with zero vector.

(Spectrum) $\lambda(A) := \{ \lambda \in \mathbb{C} : Ax = \lambda x \text{ for some } x \in \mathbb{C}^n \setminus \{0\} \}$

$$= \{ \lambda_1, \ldots, \lambda_n \} \quad \text{(SPECTRUM)}$$

↑ not necessarily distinct

CONVENTION: order of eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$.

$\lambda_1 \equiv \lambda_{max} \equiv$ principal eigenvalue; $|\lambda_1| =: \rho(A)$ [SPECTRAL RADIUS]

Obs: Any $A \in \mathbb{C}^{n \times n}$ has n eigenvalues counted with multiplicity

$\parallel$ ALGEBRAIC MULTIPLICITY

LEFT EIGENVECTORS: $y^T A = \lambda y^T, \quad y \in \mathbb{C}^n \setminus \{0\}$

Transpose

$$A^T y = \lambda y$$

i.e., left-eigenvecs of A are right-eigenvecs of A^T.

<u>NOTE</u>: Left and right eigenvectors are usually distinct, but right and left eigenvalues are the same.

<u>COR</u>: $\lambda(A) = \lambda(A^T)$.

<u>DIAGONALIZATION</u> $A \in \mathbb{C}^{n\times n}$ with (lin. indep.) eigenvectors $x_1, \ldots, x_n \in \mathbb{C}^n$. and eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. Take

$$X = \begin{pmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{pmatrix} \in \mathbb{C}^{n\times n}$$

Then

$$AX = \begin{pmatrix} | & & | \\ \lambda_1 x_1 & \cdots & \lambda_n x_n \\ | & & | \end{pmatrix} = X \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} =: X\Lambda$$

i.e., since the cols. of X are lin. indep., it's invertible:

$$\boxed{A = X \Lambda X^{-1}} \iff \boxed{X^{-1} A X = \Lambda}$$

(EVD Eigenvalue Decomposition)

<u>NOTE</u>: We say A IS DIAGONALIZABLE.

<u>DEF</u>: # of lin. indep. eigenvectors $\equiv$ GEOMETRIC MULTIPLICITY.

SPECTRAL THEOREMS

DEF: (Normal Matrix) $A \in \mathbb{C}^{n\times n}$ is <u>normal</u> iff $A^*A = AA^*$.

$A^* = \text{conjugate transpose} = [\overline{a_{ji}}]_{i,j=1,\dots,n}$.

!

THM: (Complex Spectral Theorem) For $A \in \mathbb{C}^{n\times n}$, TFAE

(i) A is normal; i.e., $A^*A = AA^*$

(ii) $\mathbb{C}^n$ has an orthonormal basis consisting of eigenvectors of A

(iii) There exists $Q \in \mathbb{C}^{n\times n}$, $Q^*Q = \mathbb{1}$, such that $(Q \in U(n))$

$$A = Q \Lambda Q^* \iff \Lambda = Q^* A Q$$

and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_j \in \mathbb{C}$.

COR: (Spectral Theorem for Hermitian matrices) If $A \in \mathbb{C}^{n\times n}$ is <u>Hermitian</u> (i.e., self-adjoint $\iff A^* = A$), then there exists a unitary $Q \in \mathbb{C}^{n\times n}$, $Q^*Q = \mathbb{1}$, and

$$A = Q\Lambda Q^* \iff \Lambda = Q^*AQ = \operatorname{diag}(\underset{\mathbb{R}}{\underset{\cap}{\lambda_1}}, \dots, \underset{\mathbb{R}}{\underset{\cap}{\lambda_n}})$$

REMARK: Self-adjoint (i.e., Hermitian) operators have <u>real</u> eigenvalues.

Cor: (Spectral Theorem for Symmetric Matrices) If $A \in \mathbb{R}^{n\times n}$ is symmetric (i.e., $A^T = A$), then $\exists\, Q \in \mathbb{R}^{n\times n}$ orthogonal (i.e., $Q^TQ = QQ^T = \mathbb{1}$) such that

$$A = Q \Lambda Q^T \quad \leftrightarrow \quad \Lambda = Q^T A Q.$$

DEF: $U(n) := \{ Q \in \mathbb{C}^{n\times n} : Q^*Q = QQ^* = \mathbb{1} \}$ UNITARY

$O(n) := \{ Q \in \mathbb{R}^{n\times n} : Q^TQ = QQ^T = \mathbb{1} \}$ ORTHOGONAL

$SU(n) := \{ Q \in U(n) : \det Q = 1 \}$ SPECIAL UNITARY

$SO(n) := \{ Q \in O(n) : \det Q = 1 \}$ SPECIAL ORTHOGONAL

$GL(n) = \{ A \in \mathbb{C}^{n\times n} : A \text{ invertible} \}$ GENERAL LINEAR

—— // ——

JORDAN CANONICAL FORM

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

$\mu_j \in \{0,1\}$

$$A = X \begin{bmatrix} \lambda_1 & \mu_1 & & 0 \\ & \lambda_2 & \ddots & \\ & & \ddots & \mu_{n-1} \\ 0 & & & \lambda_n \end{bmatrix} X^{-1}$$

Useless because

• Golub–Wilkinson Thm: JCF cannot be computed in finite precision.

Can prove stuff with it though...

Obs: $A^k = (XJX^{-1})^k = XJ^kX^{-1}$ which is nice.

—— // ——

SPECTRAL RADIUS

DEF: The spectral radius of $A \in \mathbb{C}^{n\times n}$ is

$$\rho(A) := \max\{|\lambda| : Ax = \lambda x \text{ for some } x \in \mathbb{C}^n \setminus \{0\}\}$$
$$= |\lambda_{max}| = |\lambda_1|.$$

LEMMA: If $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is consistent (i.e., $\|Ax\| \leq \|A\| \|x\|$) then $\rho(A) \leq \|A\|$ (true for any consistent norm)

Pf: Let $x \in \mathbb{C}^n$ be eigenvect. associated w/ λ_{max}, then

$$\|A\| = \|A\| \|x\| \underset{\text{consistency}}{\geq} \|Ax\| = \|\lambda_{max} x\| = |\lambda_{max}| \|x\|$$
$$= \rho(A) \|x\|$$

■

THM: For any $\varepsilon > 0$, there exists some operator norm

$$\|A\|_\alpha = \max_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}, \quad \|\cdot\|_\alpha : \mathbb{C}^n \to \mathbb{R},$$

such that

Lemma above b/c all operator norms are consistent

$$\rho(A) \leq \|A\|_\alpha \underset{\text{JCF}}{\leq} \rho(A) + \varepsilon.$$

LEMMA: For $A \in \mathbb{C}^{n\times n}$, $A^k \xrightarrow{k\to\infty} 0$ iff $\rho(A) < 1$

PROVE THIS

LECTURE 4

Oct 8th, 2025

GESHGORIN THEOREM

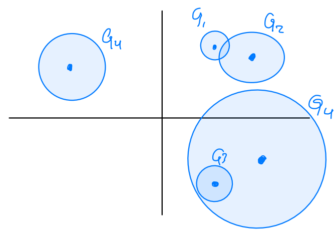
Find estimates as to where the eigenvalues of a matrix are

THM: (Geshgorin) Let $A = (a_{ij})_{i,j=1}^{n} \in \mathbb{C}^{n\times n}$ and define the Geshgorin disks as

ball of radius r_i centered at a_{ii}

$$G_i = B_{r_i}(a_{ii}), \quad i = 1, \dots, n,$$

where $r_i := \sum_{i \neq j} |a_{ij}|$.



Then,

(1) $\mathrm{Spec}(A) \subset \bigcup_{i=1}^{n} G_i$.

(2) The # of eigenvalues of A contained in any (connected) component of $\bigcup_{i=1}^{n} G_i$ is the # of G_i's in that component.

DEF: A matrix $A \in \mathbb{C}^{n\times n}$ is <u>strictly diagonally dominant</u> iff

$$|a_{ii}| > \sum_{i \neq j} |a_{ij}| \quad \forall\, i = 1, \dots, n.$$

LEMMA: Every strictly diagonally dominant matrix is invertible.

PF: Say $A \in \mathbb{C}^{n \times n}$ is SSD and, by contradiction, $A \notin GL(n, \mathbb{C})$.
Then $Ax = 0$ for some $x \neq 0$. Denote $|x_k| := \|x\|_\infty \underset{x \neq 0}{>} 0$.
Look at kth row of Ax:

$$0 = \sum_{j=1}^{n} a_{kj} x_j \qquad \text{i.e.,} \quad a_{kk} x_k = -\sum_{j \neq k} a_{kj} x_j$$

So,

$$|a_{kk}|\cancel{|x_k|} = |a_{kk} x_k| \le \sum_{j \neq k} |a_{kj} x_j| = \sum_{j \neq k} |a_{kj}||x_j|$$

$$\underset{\|x\|_\infty =: |x_k|}{\le} \cancel{|x_k|} \sum_{j \neq k} |a_{kj}|$$

i.e., $|a_{kk}| \le \sum_{j \neq k} |a_{kj}|$ $\longrightarrow$ // $\longleftarrow$ strictly diagonally dominant

Thus, A is invertible.

STUDY THIS PROOF FOR MIDTERM!

PF: (Gershgorin Thm) Let $z \notin \bigcup_{i=1}^{n} G_i$. We want to show that this implies $z \notin \mathrm{Spec}(A)$. But note that $(A - z\mathbb{1})$ is invertible by the definition of Gershgorin discs & lemma above.

MIDTERM: PROVE GESHGORIN THM ⇒ LEMMA

REMARK: $p(z) = x^d + c_{d-1}x^{d-1} + \dots + c_0 = \det(x\mathbb{1} - C)$, where

$$C = \begin{bmatrix} 0 & & & & -c_0 \\ 1 & 0 & & & -c_1 \\ & 1 & \ddots & & \vdots \\ & & \ddots & 0 & \\ 0 & & & 1 & -c_{d-1} \end{bmatrix}$$

QR - DECOMPOSITION (i.e., Gram - Schmidt)

For $A \in \mathbb{C}^{n\times n}$, there exists $Q \in U(n)$, $R = \begin{bmatrix} r_{11} & & * \\ & \ddots & \\ & & r_{nn} \end{bmatrix} \in \mathbb{C}^{n\times n}$ upper triangular s.t. $A = QR$

- FRANCIS QR-ALGORITHM: Use QR-decomposition to find the eigenvalues of a matrix

Set: $A_0 = A$

Step 1: $A_0 = Q_0 R_0$

Step 2: $R_0 Q_0 = A_1$

Step 3: $A_1 = Q_1 R_1$

Step 4: $R_1 Q_1 = A_2$

$\vdots$

NOTE: $A_1 = R_0 Q = Q_0^* A_0 Q_0$ and conjugation preserves the spectrum b/c it's a similarity transformation.

CLAIM: $Q_k \xrightarrow{k\to\infty} Q_*$] $U(n)$ is compact hence converges

$R_k \xrightarrow{k\to\infty} R_*$] Upper triangular

THEN: $A := Q_\star R_\star Q_\star^* = Q_\star \begin{bmatrix} \tilde{r}_{11} & & & * \\ & \tilde{r}_{22} & & \\ & & \ddots & \\ & & & \tilde{r}_{nn} \end{bmatrix} Q_\star^*$

(SCHUR DECOMPOSITION)

NOTE: $\mathrm{Spec}(A) = \mathrm{Spec}(R_\star) = \{\tilde{r}_{11}, \ldots, \tilde{r}_{nn}\}$.

SINGULAR VALUE DECOMPOSITION (SVD) $\leftarrow$ ALWAYS POSSIBLE !

For a matrix $A \in \mathbb{R}^{n \times n}$, the SVD is

$\leftarrow$ if $A \in \mathbb{C}^{n \times n}$, switch "T" $\to$ "*"

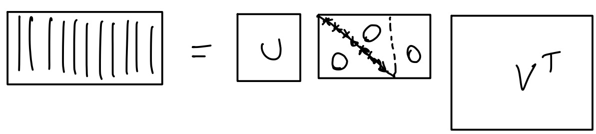
$$A = U \Sigma V^T,$$

$U \in \mathbb{R}^{n \times n}$ unitary, $V \in \mathbb{R}^{n \times n}$ unitary, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$,

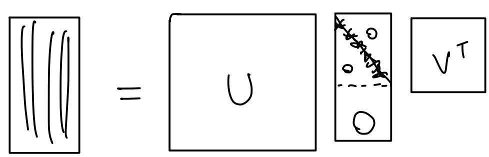
$r := \mathrm{rank}\, A$.

For non-square matrix $A \in \mathbb{R}^{m \times n}$,

$(m < n)$ short and fat:



$(n < m)$ Tall and skinny:



DEF: columns of U $\rightsquigarrow$ left singular vectors

columns of V $\rightsquigarrow$ right singular vectors

diagonal entries of Σ $\rightsquigarrow$ singular values

Singular Values = Eigenvalues of $\sqrt{A^*A}$, i.e., $\sqrt{\text{eigenvals of } A^*A}$

$\Updownarrow$

$\sqrt{AA^*}$

THM: (SVD) Given *any* $A \in \mathbb{R}^{m\times n}$, there exists $U \in O(m)$, and $V \in O(n)$, and

$$\Sigma := \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_{\min\{m,n\}} & \\ \hline & 0 & & 0 \end{array}\right] \in \mathbb{R}^{m\times n},$$

such that

$$A = U \Sigma V^T.$$

Moreover, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, $r := \operatorname{rank} A$.

REMARK: We can write.

$$A = U\Sigma V^T = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T + 0 + 0 \cdots = \sum_{j=1}^{r} \sigma_j u_j v_j^T.$$

$$= \begin{bmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{bmatrix}_{m\times r} \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix}_{r\times r} \begin{bmatrix} | & & | \\ v_1 & \cdots & v_r \\ | & & | \end{bmatrix}_{r\times n}$$

$$=: U_r \Sigma_r V_r^T$$ (CONDENSED SVD $\leftarrow$ Lose some info.)

PF: (Condensed SVD) $A \in \mathbb{R}^{m\times n}$.

$$W = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)} \rightsquigarrow$$ $W^T = W$ thus it is diagonalizable by Spectral Theorem.

$$W = Z \Lambda Z^{-1}$$

Moreover, eigenvals of W are real.

Say $Wz = \sigma z$, $z = \begin{bmatrix} x \\ y \end{bmatrix}$, $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$.

Then

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \sigma \begin{bmatrix} x \\ y \end{bmatrix} \rightsquigarrow \begin{cases} Ay = \sigma x \\ A^T x = \sigma y \end{cases},$$

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} x \\ -y \end{bmatrix} = \begin{bmatrix} -Ay \\ A^T x \end{bmatrix} = \begin{bmatrix} -\sigma x \\ \sigma y \end{bmatrix} = -\sigma \begin{bmatrix} x \\ -y \end{bmatrix}.$$

So,

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 = 0 \cdots 0 > -\sigma_r \geq -\sigma_{r-1} \geq \cdots \geq -\sigma_1$$

WLOG: $\|z\|^2 = 2 \rightsquigarrow \|z\| = \sqrt{2}$

$$\|z\|^2 = z^T z = [x^T \; y^T] \begin{bmatrix} x \\ y \end{bmatrix} = x^T x + y^T y.$$

Thus, $\begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} x \\ -y \end{bmatrix} = 0$ b/c they they are eigenvects. corresponding to diff. eigenvalues.

$$\begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} x \\ -y \end{bmatrix} = x^T x - y^T y$$

So: $\begin{cases} x^T x + y^T y = 2 \\ x^T x - y^T y = 0 \end{cases} \rightsquigarrow x^T x = 1 = y^T y$ (only solution)

Let $X := [x_1, —, x_r]$, $Y = [y_1, —, y_r]$. Then

$$Z = \begin{bmatrix} X & X \\ Y & -Y \end{bmatrix}.$$

LECTURE 5

Oct 13th, 2025

SINGULAR VALUE DECOMPOSITION

$A \in \mathbb{R}^{m\times n}$ can be decomposed into

$r = \text{rank}(A)$

$$A = \underset{m\times m}{U}\,\underset{m\times n}{\Sigma}\,\underset{n\times n}{V^T} = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_m \\ | & & | \end{bmatrix} \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right] \begin{bmatrix} - & v_1^T & - \\ & \vdots & \\ - & v_n^T & - \end{bmatrix}$$

$$=: \left[\begin{array}{c|c} U_r & U_{m-r} \end{array}\right] \left[\begin{array}{c|c} \Sigma_r & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} V_r^T \\ \hline V_{n-r}^T \end{array}\right]$$

- Condensed SVD: $A = U_r \Sigma_r V_r^T$.

PROPERTIES OF SVD:

(1) $\dim \text{im} A = \text{rank}(A) = r = \#$ nonzero singular values

$\dim \ker A = n - r = \text{nullity}(A)$

(2) columns of U_r = orthonormal basis for $\text{im}(A)$

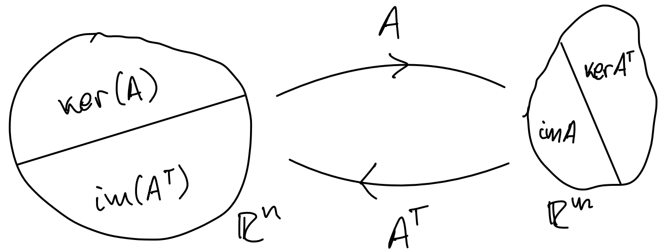
columns of V_{n-r} = orthonormal basis for $\ker(A)$

columns of V_r = orthonormal basis for $\text{im}(A^T)$

columns of U_{m-r} = orthonormal basis for $\ker(A^T)$

(3) FREDHOLM'S ALTERATIVE

$$\begin{cases} \mathbb{R}^m = \operatorname{im} A \oplus \ker A^T \\ \mathbb{R}^n = \operatorname{im} A^T \oplus \ker A \end{cases}$$



Projection Operators: Take $W \subset \mathbb{R}^n$, then any $P^2 = \mathbb{1}$ is a projection. If, in addition, $P^T = P$, then P is an orthogonal projection.

NOTE: $\operatorname{proj}_{\operatorname{im}(A)} = U_r U_r^T$; $\operatorname{proj}_{\ker(A)} = V_{n-r} V_{n-r}^T$; $\operatorname{proj}_{\operatorname{im}(A^T)} = V_r V_r^T$;

$$\operatorname{proj}_{\ker(A^T)} = U_{m-r} U_{m-r}^T$$

SVD AND NORMS

- FROBENIUS NORM $\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2} = \|(\sigma_1, \ldots, \sigma_r)\|_2$
- MATRIX 2-NORM (aka SPECTRAL NORM) $\|A\|_2 = \max\limits_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_1$ ← LARGEST SINGULAR VALUE
$$= \|(\sigma_1, \ldots, \sigma_r)\|_\infty$$
- NUCLEAR NORM $\|A\|_* := \|(\sigma_1, \ldots, \sigma_r)\|_1 = \sigma_1 + \cdots + \sigma_r$
- SCHATTEN p-NORM $\|A\|_{S,p} := \|(\sigma_1, \ldots, \sigma_r)\|_p = \left(\sum_{j=1}^{r} \sigma_j \right)^{1/p}$.
- KY FAN (p,k)-NORM $\|A\|_{S,p,k} := \left(\sigma_1^p + \cdots + \sigma_k^p \right)^{1/p}$, $1 \leq k \leq r$.

- If $m=n$, $|\det(A)| = \sigma_1 \cdots \sigma_n$
- All of these are unitarily invariant:
$$\|Q_1 A Q_2\| = \|A\| \quad \text{for any } Q_1 \in O(m),\ Q_2 \in O(n).$$

—————— // ——————

Solving Linear Systems

Goal: Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, want to find $x \in \mathbb{R}^n$ s.t.
$$Ax = b.$$

Problems: existence $\leftrightarrow$ $b \notin \operatorname{im} A$

uniqueness $\leftrightarrow$ $\ker A \neq \{0\}$

Caveat: from finite precision "errors" or measurement errors, we have error corruption
$$(A + \underset{\text{ERROR}}{\underset{\uparrow}{E}})\,x = b + \underset{\text{ERROR}}{\underset{\uparrow}{r}}$$

- $E \neq 0 \rightsquigarrow$ Total Least Squares
- $E = 0 \rightsquigarrow$ Ordinary Least Squares

Ordinary Least Squares (OLS)

$$Ax \simeq b \longleftrightarrow \min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2$$

← Use 2-norm here b/c it's differentiable

Note: $\|\cdot\|_2$ is convex; thus there is always a solution

Using SVD:

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2 = \min_{x\in\mathbb{R}^n} \|U\Sigma V^T x - b\|_2^2$$

$$= \min_{x\in\mathbb{R}^n} \|U(\Sigma V^T x - U^T b)\|_2^2$$

$\|\cdot\|_2^2$ is unitarily invariant

$$= \min_{x\in\mathbb{R}^n} \|\Sigma V^T x - U^T b\|_2^2$$

Set $y := V^T x$

$c := U^T b$

$$= \min_{y\in\mathbb{R}^n} \|\Sigma y - c\|_2^2$$

same as $\min_{x\in\mathbb{R}^n}$ because V^T is invertible in particular

$$= \min_{y\in\mathbb{R}^n} \left\| \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & 0 & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & 0 & & & \ddots & \\ & & & & & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \right\|_2^2$$

$$= \min_{y\in\mathbb{R}^n} \left\| \begin{pmatrix} \sigma_1 y_1 - c_1 \\ \vdots \\ \sigma_r y_r - c_r \\ -c_{r+1} \\ \vdots \\ -c_n \end{pmatrix} \right\|_2^2$$

$$= (\underbrace{\sigma_1 y_1 - c_1}_{\stackrel{!}{=}0})^2 + \dots + (\underbrace{\sigma_r y_r - c_r}_{\stackrel{!}{=}0})^2 + c_{r+1}^2 + \cdots c_n^2$$

i.e., $\boxed{y_j \stackrel{!}{=} \frac{c_j}{\sigma_j} \quad \forall j = 1, \dots, r}$

$$\boxed{\underline{\text{Upshot}}: \quad x \stackrel{\text{def}}{=} Vy = V \begin{pmatrix} c_1/\sigma_1 \\ \vdots \\ c_r/\sigma_r \\ y_{r+1} \\ \vdots \\ y_n \end{pmatrix}, \quad y_{r+1}, \dots, y_n \in \mathbb{R}^n}$$

ALL SOLUTIONS!

Maybe not unique

MINIMUM NORM LEAST SQUARES (MNOLS)

← Solves the question of uniqueness

$$\min \left\{ \|x\|_2 : x \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \|Ax - b\|_2 \right\}$$

$$\|$$ ← from previous page

$$\min \left\{ \|x\|_2 : x = V \begin{pmatrix} c_1/\sigma_1 \\ \vdots \\ c_r/\sigma_r \\ y_{r+1} \\ \vdots \\ y_n \end{pmatrix},\ y_{r+1}, \dots, y_n \in \mathbb{R}^n \right\}$$

NOTE: $\|x\|_2^2 = \|Vy\|_2^2 = \|y\|_2^2$

$\|\cdot\|_2$ is unitarily invariant $= \left(\frac{c_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{c_r}{\sigma_r}\right)^2 + y_{r+1}^2 + \cdots + y_n^2$

SOLUTION: Set $y_{r+1} = \cdots = y_n \overset{!}{=} 0$. Thus, the general solution to MNOLS is

$$x^+ := V \begin{pmatrix} c_1/\sigma_1 \\ \vdots \\ c_r/\sigma_r \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Recall

$$\begin{cases} c := U^T b \\ x := Vy \end{cases} \quad \text{and} \quad y = \begin{pmatrix} c_1/\sigma_1 \\ \vdots \\ c_r/\sigma_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 1/\sigma_1 & & & & 0 \\ & \ddots & & & \\ & & 1/\sigma_r & & \\ & & & 0 & \\ 0 & & & & \ddots \\ & & & & & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix}$$

$$\Sigma^+ := \operatorname{diag}\left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r}, 0, \dots, 0\right) \in \mathbb{R}^{n \times m}$$

Thus,

$$\boxed{x^+ = V \underbrace{\left[\begin{array}{ccc|c} 1/\sigma_1 & & & \\ & \ddots & & 0 \\ & & 1/\sigma_r & \\ \hline & 0 & & 0 \end{array}\right]}_{=:\Sigma^+} U^T b}$$

$$\left(A^+ := V \Sigma^+ U^T\right)$$

$\overset{!!}{A^+}$ (MOORE – PENROSE PSEUDO – INVERSE)

$$\boxed{\text{Upshot: Solution to OLS is } x^+ = A^+ b \overset{\text{def}}{=} V \Sigma^+ U^T b}$$

REMARK: When $A \in GL(n, \mathbb{R})$, then $A^+ = A^{-1}$
$(m = n = r)$

Math prof (topology) at UChicago

> **THM:** (Moore – Penrose) For any $A \in \mathbb{R}^{m \times n}$, there exists a unique matrix $X \in \mathbb{R}^{n \times m}$ such that
>
> (1) $AXA = A$
>
> (2) $XAX = X$
>
> (3) $(AX)^T = AX$
>
> (4) $(XA)^T = XA$
>
> (This unique X here is given by $X = V \Sigma^+ U^T = A^+$)

PROVE THIS

LECTURE 6

Oct 15^{th}, 2025

LEAST SQUARES: Given $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $\alpha > 0$, want to solve

$$\min_{\|x\| \leq \alpha} \|Ax - b\| .$$

REMARK: if $\alpha < \|A^+ b\|$, this is CONSTRAINED MINIMIZATION and, in this case,

$$\min_{\|x\| \leq \alpha} \|Ax - b\| = \min_{\|x\| = \alpha} \|Ax - b\| \leftarrow$$

Equality constraint minimization
$\Downarrow$
Lagrange Multipliers

$\boxed{\text{LAGRANGIAN:}}$ $\mathcal{L}(x,\mu) = \|Ax - b\|^2 - \mu\left(\|x\|^2 - \alpha^2\right)$

$$0 \overset{!}{=} \nabla_x \mathcal{L}(x,\mu) = 2A^T(Ax - b) + 2\mu x$$

$$\Updownarrow$$

$$(A^T A + \mu \mathbb{1})x = A^T b$$

$$\Rightarrow x = (A^T A + \mu \mathbb{1})^{-1} A^T b$$

STUDY LAGRANGE MULTIPLIERS

SEE PG 55

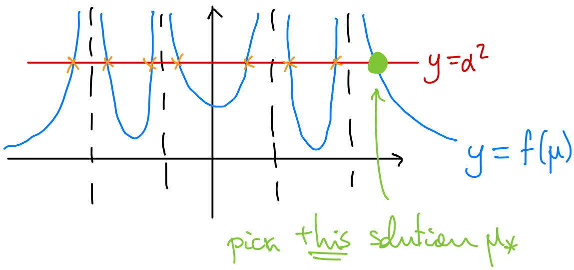
Take SVD
$A = U \Sigma V^T$

$$\alpha^2 = \|x\|^2 = \left\| (A^T A + \mu \mathbb{1})^{-1} A^T b \right\|$$

$c := U^T b$

$$= c^T \Sigma (\Sigma^T \Sigma + \mu \mathbb{1})^{-2} \Sigma^T c$$

$$\Rightarrow \quad \alpha^2 = \|x\|^2 = \underbrace{\sum_{i=1}^{r} \frac{c_i^2 \sigma_i^2}{(\sigma_i^2+\mu)^2}}_{=\, f(\mu)} \quad \Big\} \text{ (SECULAR EQUATION)}$$



- NEWTON - RHAPSON is method to find solutions to $f(\mu) = \alpha^2$.

—— // ——

Matrix Approximation Problem

Given $A \in \mathbb{R}^{n\times n}$, what, for example symmetric, matrix X is the closest to A in $\|\cdot\|_F$?

$$\iff \min_{X \in Sym(n,\mathbb{R})} \|A - X\|_F \;\overset{?}{=}\; \longleftarrow \; X = \frac{A^T + A}{2} \quad \text{(no SVD required)}$$

$$\min_{X^TX = \mathbb{1}} \|A - X\|_F \longleftarrow \|A - X\|_F^2 = tr\left[(A-X)^T(A-X)\right]$$

$$= tr\left(AA^T - X^TA - A^TX + X^TX\right)$$

$$= \underbrace{\cancel{\|A\|_F^2}}_{=n} - 2\,tr(A^TX) + \underbrace{\cancel{\|X\|_F^2}}_{=n}$$

So $\min_{X^TX=\mathbb{1}} \|A - X\|_F \iff -2 \max_{X^TX = \mathbb{1}} tr(A^TX)$

$A = U\Sigma V^T$ (SVD)

Now, $\max\limits_{X^TX=\mathbb{1}} \operatorname{tr}(A^TX) = \max\limits_{X^TX=\mathbb{1}} \operatorname{tr}(V\Sigma^T U^T X)$

$$= \max_{X^TX=\mathbb{1}} \operatorname{tr}(\Sigma^T \underbrace{U^TXV}_{=: Z \in O(n) \Rightarrow |z_{ij}| \le 1})$$

!

$$\operatorname{tr}(\Sigma^T Z) \le \sigma_1 |z_{11}| + \dots + \sigma_n |z_{nn}|$$

$$\le \sigma_1 + \dots + \sigma_n . \quad \max \Leftrightarrow \underset{U^TXV}{\overset{}{Z}} = \mathbb{1} \Rightarrow X = UV^T$$

Upshot: $\boxed{\min\limits_{X^TX=\mathbb{1}} \|A - X\|_F = UV^T, \quad \text{SVD of } A \text{ is } A = U\Sigma V^T}$

Closest rank-r matrix: $\min\limits_{\operatorname{rank} X \le r} \|A - X\|_F$ ($m \times n$)

Polar Decomposition: For $A \in \mathbb{R}^{n\times n}$, $\exists Q \in O(n)$ and $\exists P \succeq 0$ such that $A = QP$. Note that

$$A = U\Sigma V^T = \underbrace{(UV^T)}_{\overset{!!}{=} Q} \underbrace{(V\Sigma V^T)}_{\overset{!!}{=} P}$$

Thus: $\min\limits_{X \succeq 0} \|A - X\|_F = P$

$v^T X v \ge 0 \ \forall x \rightarrow X \succeq 0$

Thm: (ECKART – YOUNG) For $A \in \mathbb{R}^{m\times n}$, write

Eckart was prof of psychology at UChicago and Young his PhD student

$$A = U \left[\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_{\min(m,n)} & \end{array}\right] V^T .$$

Then

$$\min_{\operatorname{rank}(X)\le r} \|A - X\| = \sigma_{r+1}$$

$\|\cdot\|_2$ and $\|\cdot\|_F$

is attained at

$$X := U \begin{bmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ 0 & & & & \ddots \\ & & & & & 0 \end{bmatrix} V^T$$

Pf: ($\|\cdot\|_2$) Suppose by contradiction that $\exists B \in \mathbb{R}^{m\times n}$ s.t. $\operatorname{rank} B \le r$ and $\|A-B\|_2 < \|A-X\|_2 = \sigma_{r+1}$.

Note

$$\|A-X\|_2 = \left\| \begin{bmatrix} 0 & & & & 0 \\ & \ddots & & & \\ & & 0 & & \\ & & & \sigma_{r+1} & \\ 0 & & & & \ddots \\ & & & & & \sigma_{\min(m,n)} \end{bmatrix} \right\|_2 = \sigma_{r+1} .$$

Moreover, $\operatorname{rank} B + \underbrace{\operatorname{nullity} B}_{\substack{\| \\ \dim \ker B \,\ge\, n-r}} = n$

Let $w \in \ker B$, $w \neq 0$. Then $Bw = 0$ and

$$\|Aw\|_2 = \|(A-B)w\|_2 \le \|A-B\|_2 \, \|w\|_2 < \sigma_{r+1}\|w\|_2 - 1 .$$

Let $w \in \operatorname{span}\{v_1, \dots, v_{r+1}\} =: W$. Then

$$w = \alpha_1 v_1 + \dots + \alpha_{r+1} v_{r+1} = \underbrace{\begin{bmatrix} v_1 & \cdots & v_{r+1} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{r+1} \end{bmatrix}}_{\overset{!!}{=}\; V_{r+1}\alpha}$$

So, $$\|Aw\|_2^2 = \|U \Sigma V^T V_{r+1} \alpha\|^2$$

$$= \left\| \Sigma \begin{bmatrix} I_{r+1} \\ 0 \end{bmatrix} \alpha \right\|^2$$

$$= \left\| \begin{bmatrix} \sigma_1 & & 0 \\ 0 & \ddots & \sigma_{r+1} \\ & 0 & \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{r+1} \end{bmatrix} \right\|^2$$

$$= \sum_{j=1}^{r+1} \sigma_j^2 \alpha_j^2$$

$$\ge \sigma_{r+1}^2 \overbrace{\sum_{j=1}^{r+1} \alpha_j^2}^{= \|w\|^2} \implies \|Aw\| \ge \sigma_{r+1} \|w\|$$

But $\underbrace{\ker B}_{\dim = n-r} \cap \underbrace{\operatorname{span}\{v_1, \dots, v_{r+1}\}}_{\dim = r+1} \ne \{0\}$ ✓

$> n \implies$ intersection is non-empty.

MIDTERM: PROVE FOR $\|\cdot\|_F$ $\longrightarrow$ ∎

TOTAL LEAST SQUARES

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

$(A+E)x = b + r$, where x, E, r are unknowns

possible errors from measurements, etc.

SOLUTION

$$\text{Consider } \min_{(A+E)x=b+r} \left(\|E\|_F^2 + \|r\|_2^2 \right) = \min_{(A+E)x=b+r} \| [E, r] \|_F^2$$

$$\underbrace{Ax - b}_{\substack{\| \\ [A,b]\begin{bmatrix} x \\ -1 \end{bmatrix} \\ \| \\ C \in \mathbb{R}^{m \times (n+1)}}} + \underbrace{Ex - r}_{\substack{\| \\ [E,r]\begin{bmatrix} x \\ -1 \end{bmatrix} \\ \| \\ \mathcal{F}}} = 0$$

$$= \min_{(C+\mathcal{F})z=0} \|\mathcal{F}\|_F^2$$

and $z := \begin{bmatrix} x \\ -1 \end{bmatrix} \neq 0$

$$\begin{cases} \text{nullity}(C+F) \geq 1 \\ \text{rank}(C+F) \leq n \end{cases}$$

$$= \min_{\text{rank}(C+\mathcal{F}) \leq n} \|\mathcal{F}\|_F^2$$

$$C \overset{\text{def}}{=} [A, b] \overset{\text{SVD}}{=} U \Sigma V^T = U \begin{bmatrix} \sigma_1 & & 0 \\ 0 & \ddots & \sigma_{n+1} \\ & 0 & \end{bmatrix} V^T$$

$$= \min_{\text{rank}(C+\mathcal{F}) \leq n} \| C - (C+\mathcal{F}) \|_F^2$$

$$= \operatorname{rank}_{\text{rank}(X) \leq n} \| C - X \|_F^2 \quad \leftarrow \text{APPLY ECKHART–YOUNG}$$

LECTURE 7

Oct 20th, 2025

FLOATING POINT ARITHMETIC

$F_n := \{2^n : n \in \mathbb{N}\}$.

Problem: Computer system w/ n bits can only represent 2^n numbers

TYPES OF ERRORS: • ABSOLUTE ERROR: $\|x - x_*\|$ ← Depends on units

• RELATIVE ERROR: $\dfrac{\|x - x_*\|}{\|x\|}$

• POINTWISE ERROR: Set $y \in \mathbb{R}^d$, $y_i := \begin{cases} \dfrac{x_i - x_i^*}{x_i}, & x_i \neq 0 \\ 0, & x_i = 0 \end{cases}$ (Preferred)

IEEE-754 STANDARD: What #'s to store so that our system is useful.

- Store the sign ± on the 1st bit
- Always set the decimal place to be after the first #.

$$a = \pm a_1 \cdot 2^0 + a_2 \cdot 2^{-1} + a_3 \cdot 2^{-2} + \cdots + a_n \cdot 2^{-n+1}$$

$a_1, \ldots, a_n \in \{0,1\}$ (MANTISSA)

$fl : \mathbb{R} \to F_n$ (floating pt. representation)

THM: (KAHAN) For $x, y \in \mathbb{R}$, $fl(x) = x(1+\varepsilon_0)$, $|\varepsilon_0| \leq \varepsilon_{machine}$.

$fl(x \pm y) = (x \pm y)(1 + \varepsilon_1)$, $|\varepsilon_1| \leq \varepsilon_{machine}$;

$fl(x \cdot y) = (x \cdot y)(1 + \varepsilon_2)$, $|\varepsilon_2| \leq \varepsilon_{machine}$;

$fl(x/y) = (x/y)(1 + \varepsilon_3)$, $|\varepsilon_3| \leq \varepsilon_{machine}$;

DEF: $\varepsilon_{machine} := \inf \{ x \in \mathbb{R} : x > 0, fl(1+x) \neq 1 \}$.

$\pm$	a_0	a_1	$\cdots$	a_k	e_1	$\cdots$	e_l

$$n = 1 + k + l$$

$n = 32 \longrightarrow l = 8, \; k = 23$ (SINGLE PRECISION)

$n = 64 \longrightarrow l = 11, \; k = 52$ (DOUBLE PRECISION)

SINGLE: 7 decimal digits $\quad 10^{-38} \longleftrightarrow 10^{38}$

$\varepsilon_{mach} = 2^{-23} \approx 1.2 \cdot 10^{-7} \quad \begin{matrix} \wr\wr & & \wr\wr \\ 2^{-128} & \longleftrightarrow & 2^{128} \end{matrix}$

DOUBLE: 16 decimal digits $\quad 10^{-308} \longleftrightarrow 10^{308}$

$\varepsilon_{machine} = 2^{-52} \simeq 2.2 \cdot 10^{-16} \quad \begin{matrix} \wr\wr & & \wr\wr \\ 2^{-1024} & \longleftrightarrow & 2^{1024} \end{matrix}$

E.g.: Standard precision for telephones / voice $\rightsquigarrow$ G-117

Deep learning $\rightsquigarrow$ half-precision (16-bit)

reduced precision (8-bit, 4-bit)

—————— // ——————

Floating Point Errors: $\pm a_1 a_2 a_3 a_4 \cdot 10^e$, $-99 \leq e \leq 99$, $a_1, \dots, a_4 \in \{0, 1, \dots, 9\}$

1) Rounding Error

Storage: $fl(1.112 \cdot 10^5) = 1.111 \cdot 10^5$

Arithmetic: $fl(\underbrace{1.111 \cdot 10^1 \cdot 1.111 \cdot 10^2}_{\text{Actually: } 1.234\underbrace{321}_{\text{lost all of these}} \cdot 10^3}) = 1.234 \cdot 10^3$

2) Overflow / Underflow

When the # is way too large for the FP# system ← overflow

e.g.: $fl(1.000 \cdot 10^{55} \cdot 1.000 \cdot 10^{55})$ overflow

When the # is too small for FP# (i.e., too close to zero) ← underflow

e.g.: Compute $\|x\|_2$.

Don't do $\sqrt{x_1^2 + \cdots + x_n^2}$. Instead do: Step 1: $\hat{x} = \begin{cases} x / \|x\|_\infty, & \text{if } \|x\|_\infty \neq 0 \\ 0, & \text{if } \|x\|_\infty = 0 \end{cases}$

Step 2: $\|x\|_2 = \|x\|_\infty \|\hat{x}\|_2$.

3) Cancellation Error

Subtracting two #'s that are very close together

LECTURE 8

Oct 22^{nd}, 2025

ERROR: $\|x - \hat{x}\|_\infty \leq \varepsilon = \underbrace{10^{-15}}$ $\Rightarrow$ we know that $\hat{x}$ is accurate up to 14 decimal places.

computed quantity

Need to know this error bound

$\boxed{\text{ERROR BOUNDS}}$ depend on
- condition # of problem
- stability of algorithm

* CONDITIONING OF PROBLEMS

- Consider the linear problem $Ax = b$, $A \in GL(n, \mathbb{R})$

⚠ Actually, because of floating pt. errors $\rightsquigarrow$ A is actually $A + \Delta A$, b is actually $b + \Delta b$

computed qty.

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

Goal: Bound the relative error $\dfrac{\|\overbrace{\hat{x} - x}^{=: \Delta x}\|}{\|x\|} \leq ?$

PERTURBATION ANALYSIS: Toy example
$A\hat{x} = b + \Delta b$
$Ax = b.$

$\Delta x := \hat{x} - x$

$$\|b\| = \|Ax\| \leq \|A\|\,\|x\|$$

$$\Rightarrow \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|} \quad (2)$$

$\cancel{Ax} + A\Delta x = \cancel{b} + \Delta b$

$\Rightarrow A\,\Delta x = \Delta b$

$\Rightarrow \Delta x = A^{-1}\Delta b$

$\Rightarrow \|\Delta x\| \leq \|A\|^{-1}\,\|\Delta b\| \quad (1)$

$(1)\cdot(2) \rightsquigarrow \dfrac{\|\Delta x\|}{\|x\|} \leq \underbrace{\boxed{\|A\|\,\|A^{-1}\|}}_{\text{CONDITION NUMBER}} \dfrac{\|\Delta b\|}{\|b\|}$

relative error in solution

relative error in input

DEF: Condition Number $\kappa_2(A) := \|A\|_2 \|A^{-1}\|_2$

$$\kappa_F(A) := \|A\|_F \|A^{-1}\|_F$$

$$\kappa_*(A) = \|A\|_F \|A^{-1}\|_2 \,, \quad \text{etc. ...}$$

p-norms

(HW02)

REMARK:
$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)\left[\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right]}{1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}} \quad \text{in general.}$$

$\frac{1}{1-x} = 1 + x + x^2 + \dots$

Only need to know this up to order of magnitude

ASIDE: SVD implies that every matrix can be written as the product of a rotation, scaling, and another rotation: $A = U\Sigma V^*$.

SPECTRAL CONDITION NUMBER: For $A \in GL(n,\mathbb{R})$

$$\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{max}(A)}{\sigma_{min}(A)} \; ; \quad \kappa(A) = +\infty \text{ if } A \notin GL(n,\mathbb{R})$$

DEF: $\kappa_2(A) := \dfrac{\sigma_{max}(A)}{\sigma_{min}(A)} = \|A\|_2 \|A^+\|_2$

Always exists for all matrices $A \in \mathbb{R}^{m\times n}$

pseudo-inverse

Similarly, $\kappa_{\text{generalized}}(A) := \|A\| \|A^+\|$ } Exists $\forall$ rectangular matrices $A \in \mathbb{R}^{m \times n}$

REMARK: All of the above condition #'s are for the problem $Ax = b$.

We have different condition #'s for other problems; e.g., $Ax = \lambda x$, $\min \|Ax - b\|$, $\min c^T x$ s.t. $Ax \leq b$, etc

WELL-POSED PROBLEM
- Solution exists
- Solution is unique
- Continuous dependency on initial conditions

$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

well-posed

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

ill-posed

Condition number is a continuous measure to how close you are from an ill-posed problem. The bigger the worse!

NOTE: $\kappa(A) = 5 \Rightarrow$ At each operation, we lose $\sqrt{5}$ digits of precision in the answer.

DEF: Given a problem, the <u>condition number</u> of the problem is $1 \Big/ \left(\text{normalized distance to the nearest ill-posed problem} \right)$.

NOTE: $Ax = b$ $\quad A \in GL(n, \mathbb{R}) \leftarrow$ well-posed region

$A \in M_{ill} := \{ X : \det X = 0 \} \leftarrow$ ill-posed region

$$\operatorname{dist}(A, M_{ill}) = \min_{\det X = 0} \|A - X\|_2 = \min_{\operatorname{rank} X \leq n-1} \|A - X\|_2$$

Eckhart–Young

$$\overset{\text{Eckhart-Young}}{=} \sigma_n(A)$$

$$= \frac{1}{\sigma_1(A^{-1})}.$$

Normalize the distance: $\dfrac{\operatorname{dist}(A, M_{ill})}{\|A\|_2} = \dfrac{1}{\|A\|_2} \cdot \dfrac{1}{\sigma_1(A^{-1})}$

$$= 1/\kappa_2(A).$$

<u>CONDITION NUMBER AS A DERIVATIVE</u>: Consider

(LINEAR SYSTEM)

$$f : GL(n, \mathbb{R}) \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$$
$$(A, b) \longmapsto A^{-1} b = x$$

(MIN. NORM LEAST SQUARES)

$$f : \mathbb{R}^{m \times n} \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$$
$$(A, b) \longmapsto A^{+} b = x$$

etc... <u>Upshot</u>: Every problem can be seen as functions (mod measure 0 sets)

DEF: $$\kappa_f(A) := \lim_{\delta \to 0} \sup_{\text{Rel. Error}(a) \le \delta} \frac{\text{Rel Error}(f(a))}{\text{Rel Error}(a)}$$

$$= \lim_{\delta \to 0} \sup_{\|\Delta a\| < \delta} \frac{\|\Delta f(a)\| / \|f(a)\|}{\|\Delta a\| / \|a\|}$$

$$= \frac{\|Df(a)\|}{\|f(a)\| / \|a\|}.$$

Easier to compute but assumes problem can be written as a diffable fct.

EXAMPLE: (CANCELLATION ERROR) Take $f: \mathbb{R}^2 \to \mathbb{R}$, $(a,b) \mapsto a-b$.

$$Df\left(\begin{bmatrix} a \\ b \end{bmatrix}\right) = [1, -1].$$

Thus, $$\kappa_f\left(\begin{bmatrix} a \\ b \end{bmatrix}\right) = \frac{\|[1,-1]\|_\infty}{|a-b| / \|[a,b]\|_\infty} = \frac{1}{|a-b| / \max(a,b)}$$

$$= \frac{\max(a,b)}{|a-b|} \implies$$ Subtraction is an ill-conditioned problem (Cancellation Error)

REMARK: $\text{RelError}(f(a)) \lesssim \kappa_f(a)\, \text{RelError}(a)$.

error bound we wanted at the beginning

CONDITION NUMBER OF SVD:

$$f: \mathbb{R}^{m\times n} \longrightarrow O(m) \times \mathbb{R}^{\min(m,n)} \times O(n)$$
$$A \longmapsto (U, \sigma_1, \ldots, \sigma_{\min(m,n)}, V)$$

- CRUDE CONDITION # ESTIMATION: Suppose $A = X \Lambda X^{-1}$ is diagonalizable.

WRONG!, but useful...

$$(A + \Delta A) = X(\Lambda + \Delta\Lambda) X^{-1}$$

$$\|\Delta A\| = \|X \Delta\Lambda X^{-1}\| \leq \underbrace{\|X\| \|X^{-1}\|}\|\Delta\Lambda\|$$
$$= \underbrace{\kappa(X)}_{\text{can be arbitrarily large}} \|\Delta\Lambda\|$$

$\Downarrow$

Eigenval. problems are, in general, ill-conditioned

Except when...

$$A^T A = A A^T \Rightarrow X = Q \in O(n)$$

$$\Rightarrow \|\Delta A\| \leq \underbrace{\kappa(Q)}_{=1 \ \forall Q \in O(n)} \|\Delta\Lambda\|$$

* SVD: $A = U\Sigma V^T \Rightarrow \|\Delta A\| = \|\Delta\Sigma\|$

$\Downarrow$

SVD can always be computed to machine precision

LECTURE 9

Oct 27th, 2025

* ALGORITHM STABILITY

Model problems as a function $f : \underset{\text{input}}{X} \to \underset{\text{output}}{Y}$

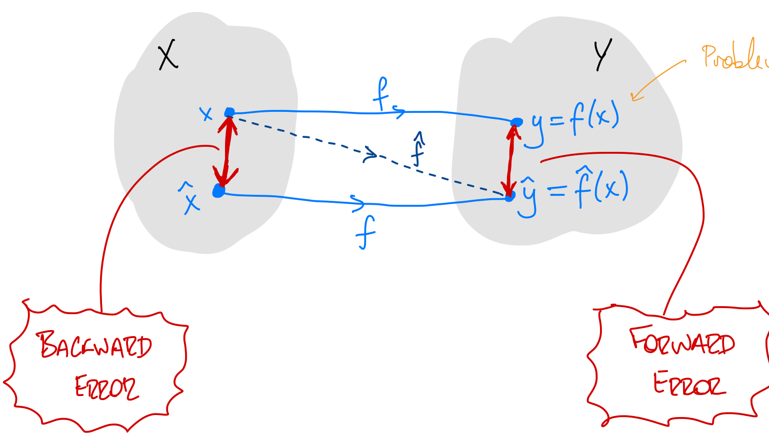
Approximation b/c of finite-precision arithmetic $\hat{f} \approx f$.

EXAMPLE : (SVD)

$$f : \mathbb{R}^{m\times n} \longrightarrow O(m) \times \mathbb{R}^{m\times n} \times O(n)$$

$$(\hat{U}, \hat{\Sigma}, \hat{V}^T) \overset{\hat{f}}{\longleftarrow} A \longmapsto (U, \Sigma, V^T)$$

DEF :
- FORWARD ERROR $\to \| f(x) - \hat{f}(x) \|$
- BACKWARD ERROR $\to$ Assume $f(\hat{x}) = \hat{f}(x)$ then the backward error is $\| x - \hat{x} \|$



Problem = don't know what the "true" full precision f is...

<u>REMARK</u>: $\begin{pmatrix}\text{FORWARD}\\\text{ERROR}\end{pmatrix} \lesssim \begin{pmatrix}\text{BACKWARD}\\\text{ERROR}\end{pmatrix} \cdot \begin{pmatrix}\text{CONDITION}\\\text{NUMBER}\end{pmatrix}$

- For the SVD map above

$$\begin{pmatrix}\text{forward}\\\text{error}\end{pmatrix} = \|\hat{U} - U\| + \|\hat{\Sigma} - \Sigma\| + \|\hat{V} - V^T\|$$

<u>PROBLEM</u>: Don't know what U, Σ, V^T are...

$$\begin{pmatrix}\text{backward}\\\text{error}\end{pmatrix} = \|\hat{U}\hat{\Sigma}\hat{V}^T - A\|$$

Can compute ✓

$$\boxed{\text{SMALL BACKWARD ERROR} \equiv \text{BACKWARD STABLE}}$$

<u>EXAMPLE</u>: (Linear System)

$$\begin{aligned} f: GL(n,\mathbb{R}) \times \mathbb{R}^n &\longrightarrow \mathbb{R} \\ (A,b) &\longmapsto x := A^{-1}b \end{aligned}$$

The approximate version

(Gaussian Elimination)

$$\begin{aligned} \hat{f}: GL(n,\mathbb{R}) \times \mathbb{R}^n &\longrightarrow \mathbb{R} \\ (A,b) &\longmapsto \hat{x} \end{aligned}$$

$$\begin{pmatrix}\text{forward}\\\text{error}\end{pmatrix} = \|\hat{x} - A^{-1}b\|$$

<u>PROBLEM</u>: We don't know precisely what $A^{-1}b$ is...

$$\begin{pmatrix}\text{backward}\\\text{error}\end{pmatrix} = \|Ax - b\|$$

<u>RESIDUAL</u> ✓
(can compute)

Obs:

$$\frac{\|\hat{x}-x\|}{\|x\|} \le \frac{\kappa(A)\,\frac{\|\Delta A\|}{\|A\|}}{1-\kappa(A)\,\frac{\|\Delta A\|}{\|A\|}} .$$

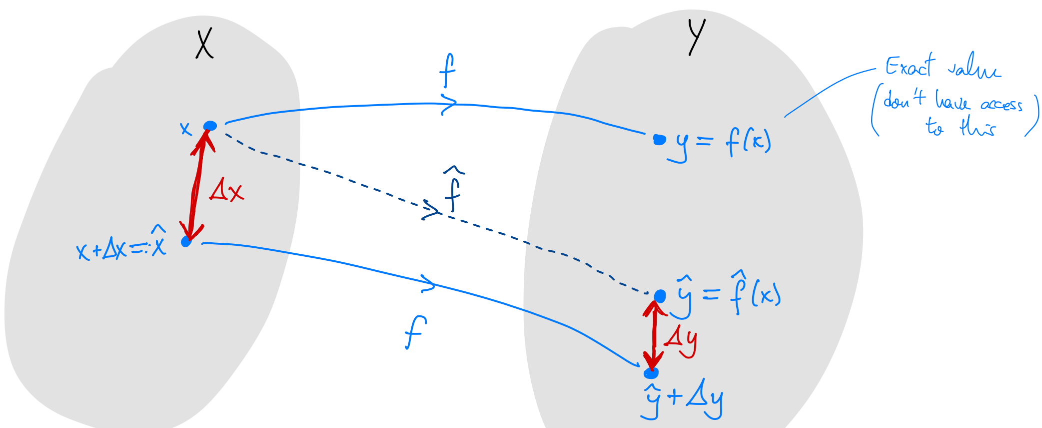
← Assume no error in b

$\frac{1}{1-x} \approx x$

Thm:

$$\underbrace{\frac{\|\Delta x\|}{\|x\|}}_{\text{forward error}} \lesssim \underbrace{\kappa(A)}_{\text{Condition number}}\ \underbrace{\frac{\|\Delta A\|}{\|A\|}}_{\text{Backward error}} .$$

Less naive version (w/out assuming f is onto)



$$\boxed{\text{Mixed Forwar Backward Stability} \equiv \text{Numerical Stability}}$$

Small $\|\Delta x\|$ and $\|\Delta y\|$ $\Rightarrow$ $\hat{f}$ is numerically stable

!

* QR FACTORIZATION

$$\underset{\in\,\mathbb{R}^{m\times n}}{A} = \underset{\in\, O(m)}{Q}\;\underset{\in\,\mathbb{R}^{m\times n}}{R} \rightarrow R \text{ is upper triangular } \; r_{ij} = 0 \;\; \forall\, i > j$$

$$= \begin{bmatrix} | & | & \cdots & | \\ | & | & & | \end{bmatrix}_{m\times m} \begin{bmatrix} \ddots & & * \\ & \ddots & \\ 0 & & \ddots \end{bmatrix}_{m\times n}$$

1) FULL COLUMN-RANK QR (tall-and-skinny matrices)

$A \in \mathbb{R}^{m\times n}$, $\operatorname{rank}(A) = n$

$$\overset{m\times n}{A} = \overset{m\times m}{Q}\,\overset{m\times n}{R} = Q \begin{bmatrix} R_1 \\ \hline 0 \end{bmatrix}, \quad R_1 \in GL(n,\mathbb{R}) \text{ and } R_1 \text{ is upper triangular}$$

(Gram-Schmidt)

Obs: Needed full-col. rank so that Gram-Schmidt is well-defined

2) CONDENSED FULL COLUMN-RANK QR:

$$\overset{m\times n}{A} = \overset{m\times m}{Q}\,\overset{m\times n}{R} = Q \begin{bmatrix} R_1 \\ \hline 0 \end{bmatrix}, \quad R_1 \in GL(n,\mathbb{R}) \text{ and } R_1 \text{ is upper triangular}$$

(Gram-Schmidt)

$$= \left[\overset{n}{Q_1} \;\vdots\; \overset{m\times n}{Q_2}\right] \begin{bmatrix} R_1 \\ \hline 0 \end{bmatrix} = \underbrace{\overset{m\times n}{Q_1}\,\overset{n\times n}{R_1}}_{\text{Condensed QR factorization}}$$

3) RANK-RETAINING QR (~~Steiner's Replacement~~ Theorem)

BAD

Given $A \in \mathbb{R}^{m\times n}$, $\operatorname{rank}(A) = r$.

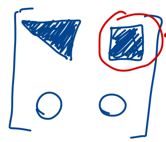
> ⚠ Permutation Matrix
> $\Pi^{-1}\Pi = \mathbb{1}$

$$A\underbrace{\Pi}_{\substack{\text{permutation matrix}\\ \text{to permute cols.}}} = \Big[\underbrace{A_1}_{\substack{\text{linearly}\\ \text{independent}\\ \text{columns}}} \;\Big\vdots\; \underbrace{A_2}_{\substack{\text{linearly}\\ \text{dependent}\\ \text{columns}}}\Big]$$

Better Algorithm:

$$R^{r\times r} \qquad \in \mathbb{R}^{r\times(n-r)}$$

$$A\Pi = \underbrace{Q}_{\in O(m)} \underbrace{\begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix}}$$



Trash can

Benefit = gives you the rank by default.

Use this to get (4):

$$\begin{bmatrix} R^T \\ S^T \end{bmatrix} =: Q_2 \begin{bmatrix} R_2 \\ 0 \end{bmatrix}$$

$$A = Q \begin{bmatrix} R^T & 0 \\ 0 & 0 \end{bmatrix} \underbrace{Q_2 \Pi_2^T}_{\text{orthogonal}}$$

lower triang.

4) COMPLETE ORTHOGONAL FACTORIZATION:

Given: $A \in \mathbb{R}^{m\times n}$, $\operatorname{rank}(A) = r$

$L \in \mathbb{R}^{r\times r}$ lower-triangular ($L \in GL(r, \mathbb{R})$)

$$A = \underbrace{Q}_{\in O(m)} \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} \underbrace{Z^T}_{\in O(n)}$$

Obs: Can be done in finitely-many steps (unlike SVD, EVD, Schur, etc)

LECTURE 10

Oct 29, 2025

QR - DECOMPOSITION

QR DECOMPOSITION

! WLOG, assume square matrix b/c we can append cols/rows to make rect → square

- Given $A \in \mathbb{R}^{n \times n}$, A full-rank, $A = QR$.

$$[a_1 \cdots a_n] = A = [q_1 \cdots q_n] \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & & \vdots \\ 0 & & \ddots & \\ & & & r_{nn} \end{bmatrix}$$

Thus:

$$\begin{cases} a_1 = r_{11} q_1 \\ a_2 = r_{12} q_1 + r_{22} q_2 \\ \vdots \\ a_n = r_{1n} q_1 + \cdots + r_{nn} q_n \end{cases}$$

Have QR decomp. $\Updownarrow$ Solution to system

Note: (i) From the 1st eq. in the system

$$\|a_1\| = r_{11} \|q_1\| = r_{11}$$

← no problem here with a_1 being zero b/c we assumed A is full-rank.

$$\Rightarrow a_1 = \|a_1\| q_1, \text{ i.e., } q_1 = \frac{a_1}{\|a_1\|}$$

(ii) $a_2^T q_1 = r_{12} q_1^T q_1 + r_{22} \underbrace{q_2^T q_1}_{= 0} \Rightarrow r_{12} = a_2^T q_1$

$= 0$ b/c $Q \in O(n)$

So, $r_{22}\|q_2\| = \|a_2 - (a_2^T q_1) q_1\| \implies r_{22} = \|a_2 - (a_2^T q_1) q_1\|$

Thus: $$q_2 = \frac{a_2 - (a_2^T q_1) q_1}{\|a_2 - (a_2^T q_1) q_1\|}$$

(iii) Repeat

⋮

(k) $$a_k = \sum_{j=1}^{k} r_{jk} q_j$$

$\rightarrow r_{jk} = q_j^T a_k, \ j = 1, \dots, k$

$\rightarrow q_k = \frac{1}{r_{kk}} \left[a_k - \sum_{j=1}^{k-1} r_{jk} q_j \right]$

$$r_{kk} = \left\| a_k - \sum_{j=1}^{k-1} r_{jk} q_j \right\| \neq 0$$ b/c A is full-rank

NOTE: We made the choice to make $\{r_{kk}\}_{k=1}^{n}$ all positive. But could have chosen them to be negative too...

Useful to make some algorithms more stable

⚠ IF A is NOT FULL-RANK: rank $A = r < n$

KEY: Permute the linearly dependent cols. to the end

$$A = [a_1 \ \cdots \ a_n] \rightsquigarrow A\Pi = [A_1, A_2] = Q \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix}$$

STUDY FOR MIDTERM: HOW TO GET S!

REMARK: Gram-Schmidt is UNSTABLE

$$A = U\Sigma V^T, \quad A = Q\Lambda Q^T, \quad A = QR$$

Orthogonal or unitary

$$A = Q\begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} Z^T, \quad A = LR, \quad A = R^T R$$

Types of decomposition blocks:

- Diagonal: Σ, Λ
- Triangular: L, R
- Orthogonal (or unitary...): U, V, Q, Z
- Structured matrices, tridiagonals, Toeplitz, etc...

* BACK SUBSTITUTION

STUDY PROBLEM ON THIS FOR TEST

Note:

$$\begin{bmatrix} r_{11} & & * \\ & r_{22} & \\ & & \ddots \\ & & & r_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \overset{\text{Common}}{\underset{\text{sense}}{\rightsquigarrow}} x_n = \frac{b_n}{r_{nn}}.$$

Iterate as back substitution

$$x_{n-1} = \frac{b_{n-1} - r_{nn}/b_n}{r_{n-1,n-1}}, \quad \dots, \quad x_1 = \dots$$

<u>DEF</u>: Let $A \in \mathbb{R}^{m \times n}$ with $\operatorname{rank} A = r$. We can decompose A as

$$A = GH, \quad G \in \mathbb{R}^{m \times r}, \quad H \in \mathbb{R}^{r \times n}.$$

This is called <u>RANK-RETAINING DECOMPOSITION</u>

WILL APPEAR ON MIDTERM

<u>PROVE THIS</u>!

<u>LEMMA</u>: (1) $G^T G$ is invertible

(2) HH^T is invertible

(3) $\operatorname{im}(A) = \operatorname{im}(G)$

(4) $\ker(A^T) = \ker(G^T)$

(5) $\ker(A) = \ker(H)$

(6) $\operatorname{im}(A^T) = \operatorname{im}(H^T)$

STUDY FOR TEST 1

<u>EXAMPLE</u>: $A = U \Sigma V^T \xrightarrow[\text{SVD}]{\text{Condensed}} A = \overbrace{U_r \Sigma_r}^{G} \overbrace{V_r^T}^{H}$, $\quad A = \underbrace{U_r}_{G} \underbrace{\Sigma_r V_r^T}_{H}$

$$U \Sigma V^T = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r \\ V_{n-r} \end{bmatrix}$$

REMARK: Can solve least norm least squares with this:

$$x^+ \in \operatorname{argmin}\left\{ \|x\| : x \in \operatorname{argmin} \|Ax-b\| \right\},$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^n$. Write $\begin{cases} \mathbb{R}^n = \ker A \oplus \operatorname{im} A^T \\ \mathbb{R}^m = \operatorname{im} A \oplus \ker A^T \end{cases}$

So, $\underset{\mathbb{R}^m}{\underset{\cap}{b}} = \underset{\ker A^T}{\underset{\cap}{b_0}} + \underset{\operatorname{im} A}{\underset{\cap}{b_1}}$, moreover $b_0 \perp b_1$ (Fredholm's Alternative).

Thus,

$$\|Ax-b\|^2 = \|Ax - b_0 - b_1\|^2 \overset{b_0 \perp b_1}{=} \|Ax - b_1\|^2 + \|b_0\|^2.$$

So, the $x \in \mathbb{R}^n$ that solves this is the x that $x \overset{A}{\longmapsto} b$. So, $Ax - b_1 = 0$.

Use rank-preserving decomposition: $A = GH$. Then, by the lemma above (**PROVE IT!**) $\operatorname{im} A = \operatorname{im} G$, hence

$$Ax = b_1 \rightsquigarrow GHx = Gs \quad \text{for some } s \in \mathbb{R}^r$$

b/c $b_0 \in \ker A^T = \ker G^T$

$$G^TGHx = G^TGs = G^Tb_1 = G^Tb_1 + \overset{=0}{\cancel{G^Tb_0}} = G^Tb_1$$

$$\overset{G^TG \text{ invertible}}{\iff} Hx = (G^TG)^{-1}G^Tb_0$$

Decompose $x = \underset{\ker A}{\underset{\cap}{x_0}} + \underset{\operatorname{im} A^T = \operatorname{im} H^T}{\underset{\cap}{x_1}}$, $x_0 \perp x_1$, then $\|x\|^2 = \|x_0\|^2 + \|x_1\|^2$. $\implies x_1 = H^Ty$ for some y

i.e., $Hx = \underset{=0}{\underbrace{Hx_0}} + Hx_1 = Hx_1 \iff Hx_1 = Hx = (G^TG)^{-1}G^Tb_0$

$$\iff \quad x_1 = H^T(HH^T)^{-1}(G^TG)^{-1}G^T b_e$$

$$\| \quad$$

$$x^+ \in \operatorname{argmin}\{\|x\| : x \in \operatorname{argmin}\|Ax-b\|\}$$

$$\| \quad$$

$$A^+ b$$

MEMORIZE

$$\boxed{\begin{array}{l}\underline{\text{Upshot:}} \text{ We can write the pseudo-inverse as:} \\ \qquad A^+ = H^T(HH^T)^{-1}(G^TG)^{-1}G^T\end{array}}$$

---

* Full-Rank Least Squares

- Given: $A \in \mathbb{R}^{m \times n}$, rank $A = n$ ← more cols. than rows
- WTF: $\min \|Ax - b\|$

Note: $\|Ax-b\|^2 \overset{A = QR}{=} \left\| Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - b \right\|^2$

$$= \left\| \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - Q^T b \right\|^2 \qquad Q^T b = \begin{bmatrix} c \\ d \end{bmatrix}$$

$$= \left\| \begin{bmatrix} R_1 x \\ 0 \end{bmatrix} - \begin{bmatrix} c \\ d \end{bmatrix} \right\|^2$$

$$= \left\| \begin{bmatrix} R_1 x - c \\ -d \end{bmatrix} \right\|^2$$

$$= \| R_1 x - c \|^2 + \| d \|^2$$

Back Substitution $\longrightarrow R_1 x = c$

Upshot: Solver to the full-rank least squares | **QR METHOD**

1. Form QR decomposition $A = QR$
2. Set $Q^T b = \begin{bmatrix} c \\ d \end{bmatrix}$
3. $R_1 x = c$ solve through back substitution

Output x.

(Very stable numerically)

- **NORMAL EQUATION METHOD** (frowned upon in numerical analysis)

(to solve full-rank least squares)

$$Ax = b \overset{\text{Rewrite}}{\rightsquigarrow} \boxed{A^T A x = A^T b}$$

Always has a solution

WARNING: This is ILL-CONDITIONED $\left(\kappa(A^T A) \simeq \kappa(A)^2\right)$ and UNSTABLE

Cannot store this... $fl(1+\delta^2) = 1$

$$A = \begin{bmatrix} 1 & 1 \\ \delta & 0 \end{bmatrix} \rightarrow A^T A = \begin{bmatrix} 1+\delta^2 & 1 \\ 1 & 1 \end{bmatrix}$$

end up with $A^T A$ singular

REMARK: Usually $m \gg n$.

- Condition # of QR METHOD:

$$\frac{\|\hat{x}-x\|}{\|x\|} \lesssim 2\gamma_{QR}\,\kappa(A)\,\varepsilon + \gamma_{QR}\,\kappa(A)^2 \frac{\|b-Ax\|_2}{\|A\|\,\|x\|}$$

- Condition # of NORMAL EQUATION METHOD:

$$\frac{\|\hat{x}-x\|}{\|x\|} \lesssim \gamma_{NE}\,\kappa(A)^2 \left(1 + \frac{\|b\|}{\|A\|\,\|x\|}\right)\varepsilon$$

REMARK: Normal Equation is better when $m \ggg n$ (e.g. $n=3$ and $m=100{,}000$) b/c then we only have to solve a 3×3 system. Actually, better than forming the QR-decomposition.

- AUGMENTED SYSTEM METHOD:

$$A^TAx = A^Tb$$

keep the nice structures (sparsity, tridiagonality, etc) of A.

$$\begin{cases} r := Ax-b \\ A^Tr = 0 \end{cases} \rightsquigarrow \begin{bmatrix} \mathbb{1} & A \\ A^T & 0 \end{bmatrix}\begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

(AUGMENTED SYSTEM)

PROBLEM: Increased dimensions by a lot

LECTURE 11

Nov 3^{rd}, 2025

QR Decomposition $\rightsquigarrow$ Solve norm constraint least squares $\begin{cases} \min \|Ax-b\| \\ \text{s.t. } \|x\| \leq \alpha \end{cases}$.

$\leftarrow$ quadratic constraint

$\rightsquigarrow$ Solve linearly constrained least squares

$$\begin{cases} \min \|Ax-b\| \\ \text{s.t. } C^Tx = d \end{cases}, \quad \begin{matrix} A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m \\ C \in \mathbb{R}^{n\times p}, d \in \mathbb{R}^p. \end{matrix}$$

Write $C^Tx = d := \begin{cases} c_1^Tx = d_1 \\ \vdots \\ c_p^Tx = d_p \end{cases}$

<u>LAGRANGE</u> <u>MULTIPLIERS</u>: $\mathcal{L}(x, \overbrace{\lambda_1, \ldots, \lambda_p}^{\lambda}) := \|Ax-b\|^2 + 2\sum_{i=1}^{p} \lambda_i (c_i^Tx - d_i)$

$$\|$$

$$\mathcal{L}(x,\lambda) = \|Ax-b\|^2 + 2\lambda^T(C^Tx+d)$$

1) Differentiate w.r.t. x and set to zero:

$$0 \overset{!}{=} \nabla_x \mathcal{L}(x,\lambda) = 2(A^TAx - A^Tb + C\lambda)$$

2) Differentiate w.r.t. λ and set to zero $\leftarrow$ Don't waste time b/c this only gives back the constraints

$$0 \overset{!}{=} \nabla_\lambda \mathcal{L}(x,\lambda) = C^Tx = 0$$

- <u>METHOD 1</u>: Rewrite as matrix $\begin{bmatrix} A^TA & C \\ C^T & 0 \end{bmatrix}\begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} A^Tb \\ 0 \end{bmatrix}$.

Problem: $\kappa(A^TA) \simeq \kappa(A)^2$. Good: Analytically

- METHOD 2: 1) $\hat{x} \in \operatorname{argmin} \|Ax-b\|$

2) $C^T(A^TA)^{-1}C\lambda = C^Tx - d$ to get λ_*

3) $A^TAx = A^Tb - C\lambda_*$ → $x = \underbrace{(A^TA)^{-1}A^Tb}_{=\hat{x}} - \underbrace{(A^TA)^{-1}C\lambda}_{=\xi}$

This works because of METHOD 1 and $A^TAx = A^Tb$ bad conditioning...

METHOD 2 can be implemented without forming any M^TM and requires only two QR's:

ALGORITHM FOR METHOD 2 (Suppose A is full-rank)

- INPUT: A, b, C, d

- STEP 1: Compute $A = Q\begin{bmatrix} R \\ 0 \end{bmatrix}$, $R \in GL(n,\mathbb{R})$

- STEP 2: Solve $\min \|Ax-b\|$ to get $\hat{x}$ by using QR from step 1.

- STEP 3: Form $W = R^{-T}C$ by doing p back substitutions to solve the following linear systems: good for triangular systems

$$R^Tw_i = c_i, \quad i = 1, \dots, p.$$

- STEP 4: Compute $W = Q_1R_1$

- STEP 5: Form $\eta := C^Tx - d$

- STEP 6: Solve $R_1^T R_1 \lambda = \eta$ using two back substitutions as follows:

good for triangular systems

$$\begin{cases} R_1^T u = \eta \\ R_1 \lambda = u \end{cases}$$

- STEP 7: Set $x = \hat{x} - \underbrace{(R^T R)^{-1} C \lambda} =: \xi$ but don't actually compute the inverse. Compute ξ using back substitutions in the following triangular systems

$$\begin{cases} R^T \zeta = C\lambda \\ R \zeta = \xi \end{cases}$$

dummy variable

— OUTPUT: x.

REMARK: Solving $\begin{cases} \min \|Ax - b\| \\ \text{s.t. } C^T x = b \end{cases}$ involves n variables of x and the constraint should decrease the # of possible solutions.

But, solving the system from METHODS 1 & 2 actually increase the dimensionality of the problem:

$$\begin{bmatrix} A^TA & C \\ C^T & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} A^Tb \\ d \end{bmatrix}.$$

n+p variables...

(METHOD 3 below fixes this)

- METHOD 3 :

1) Compute the QR of the feasible set

$$\underset{n\times p}{C} := \underset{n\times n}{Q_2} \begin{array}{c} p \\ \begin{bmatrix} R_2 \\ 0 \end{bmatrix} \end{array} \begin{array}{l} p \\ n-p \end{array}$$

2) Then, the constraint takes the analytic form

$$\begin{cases} R_2^T u = d \\ Q_2^T x = \begin{bmatrix} u \\ v \end{bmatrix} \end{cases}$$

$v \in \mathbb{R}^{n-p}$

3) Consider $\| b - Ax \| = \| b - \underbrace{A \overbrace{Q_2 Q_2^T}^{= \mathbb{1}}}_{= \tilde{A} = AQ_2} x \| = \| b - \tilde{A} \underbrace{Q_2^T x}_{= \begin{bmatrix} u \\ v \end{bmatrix}} \|$

$$= \left\| b - \tilde{A} \begin{bmatrix} u \\ v \end{bmatrix} \right\|$$

$$= \left\| b - [\tilde{A}_1 \ \tilde{A}_2] \begin{bmatrix} u \\ v \end{bmatrix} \right\|$$

$$= \left\| b - \tilde{A}_1 u - \tilde{A}_2 v \right\| .$$

ALGORITHM FOR METHOD 3

- INPUT : A, b, C, d

  - STEP 1 : Compute $C = Q_2 R_2$

  - STEP 2 : Compose $\tilde{A} = AQ_2$

- **Step 3**: Solve $R_2^T u = d$ using back substitution to get $\hat{u}$
- **Step 4**: Solve $\hat{v} \in \arg\min \| (b - \tilde{A}_1 \hat{u}) - \tilde{A}_2 v \|$ ← unconstrained least squares
- **Step 5**: Set $x := Q \begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix}$.

- **Output**: x.

Remark: Method 3 has less unknowns. Also,

$$\kappa_2(\tilde{A}_2) \leq \kappa_2(\tilde{A}) = \kappa(A).$$

κ_2 measures "how linearly independent a set of vectors are"

Computing QR Decomposition

1) **Gram-Schmidt**: $A R_1^{-1} R_2^{-1} \cdots R_{n-1}^{-1} = Q$

$$A = Q \underbrace{R_{n-1} R_{n-2} \cdots R_1}_{=R} = QR$$

BAD: Condition # of upper triangular can be arbitrarily large

$\Rightarrow$ Gram-Schmidt is very numerically unstable

2) Do the opposite: find an upper-triangular using orthogonal matrices

$$Q_{n-1}^T \cdots Q_2^T Q_1^T A = R$$

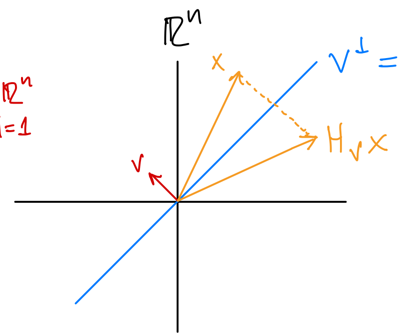
$$A = \underbrace{Q_1 Q_2 \cdots Q_{n-1}}_{\substack{\shortparallel \\ Q}} R = QR.$$

$\Rightarrow$ Well-conditioned because condition # of orthogonal matrices is always $\equiv 1$.

[Alston Householder got his Math PhD at UChicago ☺]

* HOUSEHOLDER REFLECTORS

$v \in \mathbb{R}^n$
$\|v\| = 1$



$v^\perp = \{x \in \mathbb{R}^n : x^T v = 0\}$ hyperplane

$H_v \in O(n, \mathbb{R})$

$$\boxed{H_v = \mathbb{1} - 2vv^T \in O(n, \mathbb{R})}$$

HOUSEHOLDER REFLECTOR

NOTE: (i) $H_v H_v = H_v^2 = \mathbb{1}$ (H_v is an INVOLUTION)

(ii) $H_v^T H_v = \mathbb{1} = H_v H_v^T$ ($H_v \in O(n, \mathbb{R})$)

REMARK: If $v \in \mathbb{R}^n$ not normalized, take $H_v = \mathbb{1} - \frac{vv^T}{v^T v}$.

STEP 1 : Want to eliminate the 1^{st} column by

$$Q^TA = \begin{bmatrix} r & \\ 0 & \\ \vdots & * \\ 0 & \end{bmatrix} \rightsquigarrow Q^T \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \overset{!}{=} \begin{bmatrix} r \\ 0 \\ \vdots \\ 0 \end{bmatrix} = r e_1$$

pick the "+" one here, but can also pick "–" if it's better.

$$\rightsquigarrow r = \pm \sqrt{a_1^2 + \cdots + a_n^2} = \|a\|$$

STEP 2 : Take

$$H_v a \overset{\text{def}}{=} (\mathbb{1} - 2vv^T)a = \pm \|a\| e_1 .$$

Since H_v is an involution,

$$a = H_v^2 a = \pm \|a\| (e_1 - 2v \underbrace{v^T e_1}_{=v_1})$$

i.e.,
$$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \underbrace{\pm\|a\|}_{=: \alpha} \begin{bmatrix} 1 - 2v_1^2 \\ -2v_1v_2 \\ \vdots \\ -2v_1v_n \end{bmatrix}$$

$$\boxed{\Rightarrow v_1 = \pm \sqrt{\frac{1}{2}\left(1 - \frac{a_1}{\alpha}\right)} \; ; \; v_j = - \frac{a_i}{2\alpha v_1} , \; i = 2, \ldots, n .}$$

STEP 3: So far

$$H_1^T A = \begin{bmatrix} \alpha & * & * & * & * \\ 0 & * & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & * & * & * & * \end{bmatrix} \text{ as desired.}$$

RESULT
$$\boxed{H_{n-1}^T H_{n-2}^T \cdots H_1^T A = R}$$

$=: \hat{A}_2$ and take $\hat{A}_2$ and apply Steps 1&2 above to $\hat{A}_2$

$$\begin{bmatrix} \alpha_2 & * & * & * & * \\ 0 & * & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & * & * & * & * \end{bmatrix} \longrightarrow \text{take this as } \hat{A}_3 \text{ and repeat}$$

LECTURE 12

Nov 10th, 2025

Givens Rotations & QR Decomposition

WTF: An orthogonal matrix that gives us QR

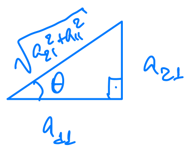
$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} r & * \\ 0 & * \end{bmatrix} \qquad \left.\begin{array}{l} \gamma := \cos\theta \\ \sigma := \sin\theta \end{array}\right\} \Rightarrow \gamma^2 + \sigma^2 = 1$$

$$\begin{bmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \overset{!}{=} \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix}$$

$$\Rightarrow \quad -\sigma a_{11} + \gamma a_{21} \overset{!}{=} 0 \quad \leftrightarrow \quad \gamma^2 a_{21}^2 = \sigma^2 a_{11}^2$$

$$= (1-\gamma^2)\, a_{11}^2$$

$$\Rightarrow \quad \gamma = \pm \frac{a_{11}}{\sqrt{a_{21}^2 + a_{11}^2}}$$



and so $\sigma = \pm \dfrac{a_{21}}{\sqrt{a_{21}^2 + a_{11}^2}}$.

Moreover, $r_{11} = \pm\sqrt{a_{21}^2 + a_{11}^2}$

GENERALIZE:

$$\begin{array}{c c} & \begin{array}{ccccccc} & & i\text{th} & & j\text{th} & & \end{array} \\ \begin{array}{c} \\ \\ i\text{th} \\ \\ j\text{th} \\ \\ \end{array} & \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & \boxed{\gamma} & & \boxed{\sigma} & & \\ & & & \ddots & & & \\ & & \boxed{-\sigma} & & \boxed{\gamma} & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \end{array} \begin{bmatrix} * \\ \vdots \\ * \\ \alpha \\ * \\ \vdots \\ * \\ \beta \\ * \\ \vdots \\ * \end{bmatrix} \begin{array}{c} \\ \\ \\ i\text{th} \\ \\ \\ \\ j\text{th} \\ \\ \\ \\ \end{array} = \begin{bmatrix} * \\ \vdots \\ * \\ \sqrt{\alpha^2+\beta^2} \\ * \\ \vdots \\ * \\ 0 \\ * \\ \vdots \\ * \end{bmatrix} \begin{array}{c} \\ \\ \\ i\text{th} \\ \\ \\ \\ j\text{th} \\ \\ \\ \\ \end{array}$$

other entries are just zero

$$\underbrace{\qquad\qquad\qquad}_{= G_{ij}(\theta)}$$

$\boxed{\text{Doing this, we can introduce zeros anywhere we need on the vector on the right}}$

REMARK: Don't really care about θ itself, as long as we get the zeros on the right.

REMARK: To do QR on an $n \times n$ matrix, we need $O(n^2)$ Givens rotations (cf $O(n-1)$ from Householder reflectors).

REMARK: Very useful for matrices that are almost in upper triangular, for ex.:

$$\begin{bmatrix} * & * & \cdots & * \\ & * & \cdots & * \\ & 0 & \ddots & \vdots \\ & & \odot & * \\ * & * & \cdots \odot & * \end{bmatrix}$$

$\sqrt{\alpha^2+\beta^2}$

becomes 0

becomes zero

becomes zero

(SEE NOTES)!

COMPLETE ORTHOGONAL FACTORIZATION: Combine Householder with Givens to enforce linear independence!

$$\underbrace{H_{n-1}\cdots H_1}_{Q^T} \underbrace{\begin{array}{cccc} A_1 & A_2 & \cdots & A_n \\ \downarrow & \downarrow & & \downarrow \end{array} \begin{bmatrix} * & * & \cdots & * \\ * & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & * \end{bmatrix}}_{=:A} \underbrace{\Pi_1 \cdots \Pi_{n-1}}_{=\Pi} = \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix}$$

$\|A_1\| > \cdots > \|A_j\| = \|A_r\| = \cdots$

Now, redo the above with

$$\hat{H}_{r-1}\cdots\hat{H}_1 \begin{bmatrix} R^T \\ S^T \end{bmatrix} = \begin{bmatrix} L^T \\ 0 \end{bmatrix} \Rightarrow A = Q \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} Z^T,$$

$$Z = \Pi Q$$

Gaussian Elimination (GE) $\leftarrow$ Way to get LU factorization

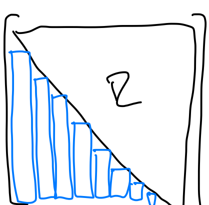
LU Factorization:

$$\mathbb{R}^{n\times n} \ni A = LU, \quad U = \begin{bmatrix} u_{11} & \cdots & u_{1n} \\ & \ddots & \vdots \\ 0 & & u_{nn} \end{bmatrix} \leftarrow \text{(upper triangular)}$$

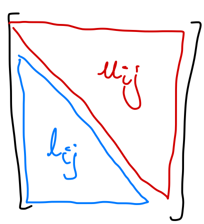
$$L = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ (l_{ij}) & & \ddots \\ & & & 1 \end{bmatrix}$$

$\leftarrow$ Row-Echelon operations to get reduced form (unit lower triangular)

Remark: Can store QR and LU decompositions without having to use more memory by



Householder reflectors



Already know the diagonal of L is just 1's.

NOTE: $H_v = \mathbb{1} - 2vv^T$ (rank-1 perturbation of the $\mathbb{1}$)

$$\boxed{M_i = \mathbb{1} - m e_i^T, \quad m \in \mathbb{R}^n}$$ (Gaussian elimination Matrix)

$$M_1 a = \begin{bmatrix} \gamma \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \gamma e_1 \rightsquigarrow (\mathbb{1} - m e_i^T) a = \gamma e_1$$

can be big/small b/c M is not orthogonal (unlike H_v)

$\Downarrow$

Gaussian elimination is unstable.

So: $(\mathbb{1} - m_1 e_1^T) a = \gamma e_1 \overset{M_1 a}{=}$, $a = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$,

$$\begin{aligned} &\parallel \\ &a - e_1^T a\, m_1 \quad \text{where } a_1 \neq 0 \\ &\parallel \\ &a_1 - a_1 m_1 \end{aligned}$$

big assumption

If $a_1 \neq 0$, $m_1 = \frac{1}{a_1}(a - \gamma e_1) = \begin{bmatrix} 0 \\ a_2/a_1 \\ \vdots \\ a_n/a_1 \end{bmatrix}$.

Choose $\gamma = a_1$

$-l_{i1} := -\frac{a_i}{a_1}$

$$M_1 = \begin{pmatrix} 1 & & & 0 \\ -a_2/a_1 & 1 & & \\ \vdots & & \ddots & \\ -a_n/a_1 & 0 & & 1 \end{pmatrix}.$$

Ex: $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$, $(a_{11} \neq 0)$

$$M_1 A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{m2}^{(2)} & \cdots & a_{mn} \end{bmatrix}$$

$\parallel$

$A^{(2)} \in \mathbb{R}^{(m-1)\times(m-1)}$

$$M_2 M_1 A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & & & \\ \vdots & & A^{(2)} & \\ 0 & & & \end{bmatrix}$$

$$M_2 = \mathbb{1} - m_2 e_2^T, \quad m_2 = \begin{bmatrix} 0 \\ 1 \\ -a_{32}^{(2)}/a_{22}^{(2)} \\ \vdots \\ -a_{m2}^{(2)}/a_{22}^{(2)} \end{bmatrix} =: \begin{bmatrix} 0 \\ 1 \\ -l_{32} \\ \vdots \\ -l_{m2} \end{bmatrix},$$

$$l_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}.$$

$\vdots$ Iterate

$$a_{kk}^{(k)} \neq 0, \quad k = 1, 2, \ldots, m-1$$

$$M_k = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & -l_{mk} & & & 1 \end{pmatrix}, \quad l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$$

$\vdots$

$$\underbrace{M_{m-1} M_{m-2} \cdots M_1}_{\substack{\parallel \\ L^{-1}}} A = \begin{bmatrix} * & * & * \\ & \ddots & * \\ 0 & & * \end{bmatrix} =: U$$

$$\Rightarrow L^{-1} = M_1^{-1} \cdots M_{m-2}^{-1} M_{m-1}^{-1}$$

Note: M_k^{-1} is easily computed b/c

$$M_k = \mathbb{1} - m_k e_k^T \iff M_k^{-1} = \mathbb{1} + m_k e_k^T$$

$$M_k^{-1} = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & l_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & l_{mk} & & & 1 \end{bmatrix}$$

DEF: Matrices with $a_{kk}^{(k)} \neq 0$ for all $k = 1, \ldots, m$ is called

STRONGLY NONSINGULAR MATRIX. ← dense set in the space of matrices

$\Leftrightarrow$ $\det A^{(k)} \neq 0$, $A^{(k)}$ k-minor

$\Leftrightarrow$ $A = LU$ with L unit lower triangular and U upper triangular

CLAIM: A sufficient condition for strongly nonsingular:

⚠ CHECK

A is strictly diagonally dominant

$\Downarrow$

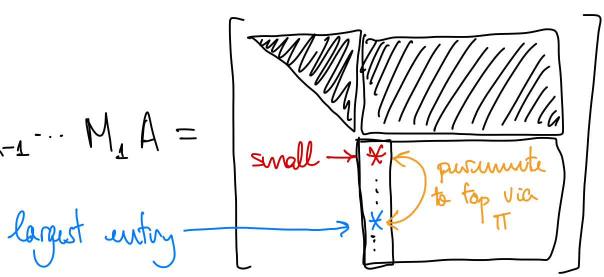
A is strongly nonsingular

⚠ **WARNING:** Dividing by $a_{kk}^{(k)}$ very small is BAD (fl. pt. arithmetic)

$\Rightarrow$ Pivoting

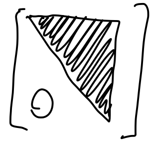
PIVOTING: Partial Pivoting $\rightsquigarrow$ Gaussian elimination with Partial Pivoting = GEPP.

$$\Pi_k M_k M_{k-1} \cdots M_1 A =$$



$$\boxed{|a_{kk}^{(k)}| = \max_{k+1 \le i \le m} |a_{ik}^{(k)}|}$$

Actual LU: Partial Pivoting

$$M_m \Pi_m M_{m-1} \Pi_{m-1} \cdots M_2 \Pi_2 M_1 \Pi_1 A =$$  $= U$

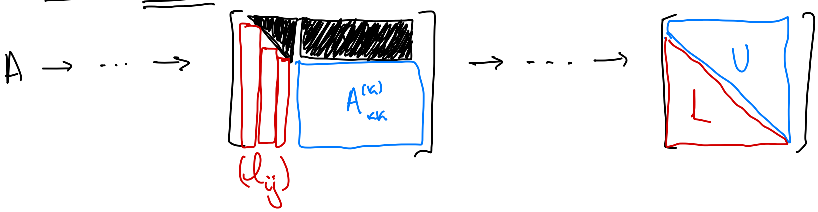
READ NOTES

$\Rightarrow \Pi A = LU$, $\Pi = \mathbb{1}$ only for strongly nonsingular matrices A

LECTURE 13

Nov 12, 2025

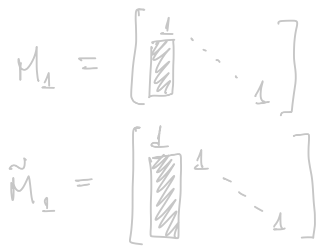
LU with Partial Pivoting: (GEPP) $\quad M_k := \mathbb{1} - m_k e_k^T$



$$M_2 \Pi_2 M_1 \Pi_1 A = U$$

$$\shortparallel$$

$$M_2 \underbrace{(\Pi_2 M_1 \Pi_2^T)}_{\tilde{M}_1} \underbrace{\Pi_2 \Pi_1}_{\Pi} A = U$$



$$\shortparallel$$

$$\underbrace{M_2 \tilde{M}_1}_{L^{-1}} \Pi A = U$$

$$\shortparallel$$

$$L^{-1} \Pi A = U \longrightarrow \boxed{\Pi A = LU}$$

• **NOTE:**

$$L = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ (l_{ij}) & & 1 \end{bmatrix}, \quad |l_{ik}| = \left| \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \right| \leq 1$$

$$\Rightarrow |l_{ij}| \leq 1 \quad \forall i,j.$$

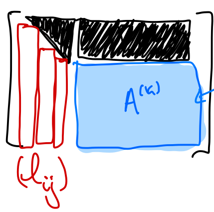
• **NOTE:** $|a_{kk}^{(k)}| = \max\limits_{k \leq i \leq n} |a_{ik}^{(k)}|$ ← From pivoting

(GROWTH FACTOR) $\gamma_{GEPP} := \max\limits_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n \\ 1 \leq k \leq n}} \dfrac{|a_{ij}^{(k)}|}{|a_{ij}|}$] Can be as large as 2^{n-1}, but, in practice, this seen hardly often.

$\boxed{\text{cf. } \gamma_{QR} = 1}$

Improve (i.e., decrease growth factor) by doing COMPLETE PIVOTING

LU WITH COMPLETE PIVOTING: GECP

$A \to \cdots \to$ 

Choose pivot from all of $A^{(k)}$ (instead of just the 1^{st} column)

$$\Pi A \underbrace{\hat{\Pi}}_{} = LU$$

permutes the column to do the full pivoting

$$\gamma_{GECP} = \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n \\ 1 \leq k \leq n}} \frac{|a_{ij}^{(k)}|}{|a_{ij}|} \leq \underbrace{c n^{\frac{1}{2}} n^{\frac{1}{4} \log n}}_{\text{Much better}}$$

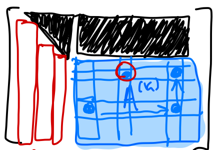
- <u>Note</u>: GECP $\rightarrow$ $\boxed{\Pi A \hat{\Pi} = LU}$ $= \begin{bmatrix} L_1 & 0 \\ L_2 & 1 \end{bmatrix} \begin{bmatrix} U_1 & U_2 \\ 0 & 0 \end{bmatrix}$

Full LU

(RANK - RETAINING FACTORIZATION) $\longrightarrow$ $\boxed{= \underbrace{\begin{bmatrix} L_1 \\ L_2 \end{bmatrix}}_{n \times r} \underbrace{\begin{bmatrix} U_1 & U_2 \end{bmatrix}}_{r \times n}}$

Condensed LU

- <u>Remark</u>: Rook Pivoting (between GEPP and GECP)



(l_{ij})

"Rook" style search

<u>PROBLEM FOR QUIZ 2</u>: Find eigenvalues of finite difference matrix with bdry condition.

<u>Remark</u>: Gauss - Jordan Elimination (UNSTABLE)

$$A \rightarrow \cdots \rightarrow U \rightarrow \cdots \longrightarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

(Reduced Row Echelon Form)

$\equiv$

(Hermite Form)

Gives very large entries in L, which is bad.

CHOLESKY DECOMPOSITION (POSITIVE DEFINITE MATRICES)

GOAL: Have $L = U^T$ in LU decomposition.

$$\boxed{A = R^T R} \Rightarrow x^T A x = x^T R^T R x = (Rx)^T (Rx) = \| Rx \|^2 > 0$$

Unique if $r_{ii} > 0$ $\forall i = 1, \dots, n$

not necessarily unique

FACT: $A = R^T R$ where $R \in GL(n)$ upper triangular $\Leftrightarrow$ A is symmetric positive definite

Pf: ($\Rightarrow$) Clear, as above.

($\Leftarrow$) Want to write $A = R^T R := F F^T$, $F := R^T$.

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} f_{11} & & & 0 \\ f_{21} & f_{22} & & \\ \vdots & \vdots & \ddots & \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix} \begin{bmatrix} f_{11} & f_{21} & \cdots & f_{n1} \\ & f_{22} & \cdots & f_{n2} \\ & & \ddots & \vdots \\ 0 & & & f_{nn} \end{bmatrix}$$

Pick this

$$a_{11} = f_{11}^2 \rightsquigarrow f_{11} = \pm\sqrt{a_{11}} = \sqrt{a_{11}}$$

$$a_{12} = f_{11} f_{21} = \sqrt{a_{11}}\, f_{21} \rightsquigarrow f_{21} = \frac{a_{12}}{\sqrt{a_{11}}}$$

$$a_{1i} = f_{11} f_{i1} \longrightarrow f_{i1} = \frac{a_{1i}}{\sqrt{a_{11}}}, \quad i = 2, \dots, n$$

$$a_{22} = f_{21}^2 + f_{22}^2 \rightsquigarrow f_{22} = \sqrt{a_{22} - f_{21}^2} \in \mathbb{R}$$

$\leftarrow$ Sylvester's Thm

$\vdots$

keep comparing entries

$\vdots$

$$\boxed{a_{kk}^2 = f_{k1}^2 + f_{k2}^2 + \cdots + f_{kk}^2} \Rightarrow f_{kj}^2 \leq a_{kk}$$

$$|f_{kj}| \leq \sqrt{a_{kk}}$$

FACT: Every diagonal entry of an $A \succ 0$ is positive because $a_{kk} = e_k^T A e_k > 0$

CHOLESKY IS VERY STABLE $\Leftarrow$ NO GROWTH FACTOR IN CHOLESKY $\longleftarrow$ $|f_{kj}| \leq \sqrt{a_{kk}}$

$$f_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} f_{kj}^2}\,, \quad f_{ik} = \frac{a_{kk} - \sum_{j=1}^{k-1} f_{kj} f_{ij}}{f_{kk}}$$

$\blacksquare$

REMARK: SQUARE-ROOT-FREE CHOLESKY

$$A = LDL^T = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ (l_{ij}) & & \ddots & \\ & & & 1 \end{bmatrix} \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix} \begin{bmatrix} 1 & & (l_{ij}) \\ & 1 & \\ 0 & & \ddots & \\ & & & 1 \end{bmatrix}$$

REMARK: Another way to get Cholesky

$$A = FF^T = [f_1, \ldots, f_n] \begin{bmatrix} f_1^T \\ \vdots \\ f_n^T \end{bmatrix} = f_1 f_1^T + \cdots + f_n f_n^T,$$

$$f_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ * \\ * \\ \vdots \\ * \end{bmatrix} \rightsquigarrow f_1 = \frac{1}{\sqrt{a_{11}}} a_1$$

So, rewrite

$$\begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \boxed{A_2} & \\ 0 & & & \end{bmatrix} = A^{(2)} := A - f_1 f_1^T = f_2 f_2^T + \cdots + f_n f_n^T$$

Now,

$$S := [f_1, e_2, \dots, e_n] = \begin{bmatrix} f_{11} & & 0 \\ f_{21} & 1 & \\ \vdots & & \ddots \\ f_{n1} & 0 & & 1 \end{bmatrix}$$

$$A = S \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \boxed{A_2} & \\ 0 & & & \end{bmatrix} S^T$$

$$\Rightarrow x \neq 0, \quad \begin{bmatrix} 0 \\ x \end{bmatrix}^T \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \boxed{A_2} & \\ 0 & & & \end{bmatrix} \begin{bmatrix} 0 \\ x \end{bmatrix} = y^T S^{-1} A S^{-T} y$$

$$= (S^{-T} y) A (S^{-T} y) > 0$$

$$\parallel$$

$$x^T A_2 x \qquad \Rightarrow A_2 \succ 0.$$

Thus, $f_2 = \dfrac{1}{\sqrt{a_{22}^{(2)}}} \begin{bmatrix} 0 \\ a_2^{(2)} \end{bmatrix}$

$$A^{(3)} := A^{(2)} - f_2 f_2^T = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & & \boxed{A_3^{\succ 0}} \end{bmatrix}$$

LECTURE 13

Solve $Ax_1 = b_1, \dots, Ax_p = b_p \rightsquigarrow AX \overset{(1)}{=} B$, $X = [x_1, \dots, x_p]_{n\times p}$

e.g. compute A^{-1} $(AX = \mathbb{1}) \Leftrightarrow (Ax_j = e_j)$ $\quad B = [b_1, \dots, b_p]_{m\times p}$

Find: $A^{-1}B$. Solve (1) and multiply it to B with only one QR or LU or Cholesky, etc.

$$\mathbb{R}^{n\times n} \ni A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \leftarrow \text{Can generalize the decompositions from before to block matrices}$$

- Block Elimination $M_1 := \mathbb{1} - U_1 V_1^T$, $U_1 = \begin{bmatrix} 0 \\ A_{21}A_{11}^{-1} \end{bmatrix}$, $V_1 = \begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix}$

$$\rightsquigarrow M_1 A = \begin{bmatrix} \mathbb{1} & 0 \\ -A_{21}A_{11}^{-1} & \mathbb{1} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

Schur Complement of A_{11} in A

$$\Rightarrow L_1 := M_1^{-1} = \begin{bmatrix} \mathbb{1} & 0 \\ A_{21}A_{11}^{-1} & \mathbb{1} \end{bmatrix}$$

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}$$

Block LU: ⚠ $L_{11}, -, L_{22}$, $U_{11}, -, U_{22}$ can be any matrices (not necessarily upper/lower Δ)

Block Cholesky when $A \succ 0$, $A^T = A$

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} R_{11}^T & 0 \\ R_{12}^T & R_{22}^T \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

$R_{11}, -, R_{22}$ don't need to be upper Δ !

$$\Rightarrow \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

$$\rightsquigarrow \boxed{\det A = (\det A_{11})(\det A_{22} - A_{21} A_{11}^{-1} A_{12})}$$

<u>SHERMAN - WOODBURY - MORRISON FORMULA:</u>

$$(A + BC)^{-1} = A^{-1} - A^{-1} B(\mathbb{1} + CA^{-1}B)^{-1} CA^{-1}.$$

LECTURE 14

ITERATIVE METHODS

Nov 19th, 2025

GOAL: Find x s.t. $Ax = b$ $\rightsquigarrow$ $\lim_{k\to\infty} x^{(k)} = x_* = A^{-1}b$.

1) SPLITTING METHODS: $A = M - N$

$$(M-N)\,x = b \rightsquigarrow Mx = Nx + b$$

$x^{(0)}$, for $k = 1,2,3,\dots,$ $\quad M\, x^{(k+1)} = N x^{(k)} + b$

easy to invert

2) Semi-iterative methods (next time)

3) Krylov Subspace Methods

REMARK: Cayley-Hamilton $\rightsquigarrow$ $A \in GL(n,\mathbb{R})$

$p(x) \in \mathbb{R}[x]$

$$p(A) = a_0 \mathbb{1} + a_1 A + a_2 A^2 + \dots + a_d A^d = 0$$

$$\Rightarrow \quad A^{-1} = -\frac{a_1}{a_0}\mathbb{1} - \frac{a_2}{a_0} A - \dots - \frac{a_d}{a_0} A^{d-1}$$

$$x_* := A^{-1} b = -\frac{a_1}{a_0} b - \frac{a_2}{a_0} Ab - \dots - \frac{a_d}{a_0} A^{d-1} b$$

$$\Rightarrow \quad x_* = A^{-1} b \in \operatorname{span}\underbrace{\{b, Ab, \dots, A^{d-1}b\}}_{\text{KRYLOV SUBSPACE} \;\equiv\; K_d(A,b)}$$

$k = \#\{\text{distinct ev's of } A\} \rightsquigarrow x^{(k)} = x_* = A^{-1} b$

1) Splitting Method: $A := M - N$, $A \in GL(n, \mathbb{R})$

$$\Longrightarrow \quad M x^{(k+1)} = N x^{(k+1)} - b$$

Error: $e^{(k)} = x^{(k)} - \underbrace{x_*}$ $\leftarrow x_* := A^{-1}b \Rightarrow Mx_* = Nx_* + b$

Thus: $M(x^{(k+1)} - x_*) = N(x^{(k)} - x_*)$

$$\rightsquigarrow \quad e^{(k+1)} = \underbrace{M^{-1}N}_{=B} e^{(k)} \quad \rightsquigarrow \quad \boxed{e^{(k+1)} = B e^{(k)}}$$

Take: $\|\cdot\|$ consistent (e.g., ∞) $\Rightarrow \|e^{(k+1)}\| \leq \|B\| \, \|e^{(k)}\|$

By induction: $e^{(k+1)} = Be^{(k)} = B^2 e^{(k-1)} = \cdots = B^{k+1} e^{(0)}$

$$\Rightarrow \|e^{(k+1)}\| \leq \|B^{k+1}\| \, \|e^{(0)}\| \leq \|B\|^{k+1} \, \|e^{(0)}\| .$$

THM: $\lim\limits_{k\to\infty} x^{(k)} \overset{\text{def}}{=} \lim\limits_{k\to\infty} B^k e^{(0)} = 0 \quad \forall e^{(0)} \in \mathbb{R}^n$

$$\Updownarrow$$

$$\lim_{k\to\infty} B^k = 0$$

$$\Updownarrow \quad \text{Jordan Canonical Form}$$

$$\rho(B) < 1$$

CONVERGENCE RATES

$\frac{1}{n} \xrightarrow{\text{sublinear}} 0, \quad \frac{1}{2^n} \xrightarrow{\text{linear}} 0, \quad \frac{1}{2^{2^n}} \xrightarrow{\text{quadratic}} 0$

linear = "exponential"

DEF:

$$\text{RATE OF CONVERGENCE} \equiv \limsup_{k\to\infty} \frac{\| e^{(k+1)} \|}{\| e^{(k)} \|^p} \qquad \begin{cases} p = 1 \ (\text{linear}) \\ p = 2 \ (\text{quadratic}) \end{cases}$$

$$\text{PSEUDO-RATE OF CONVERGENCE} \equiv \|B\| \qquad \leftarrow \text{b/c } \|e^{(k+1)}\| \leq \|B\| \, \|e^{(k)}\|$$

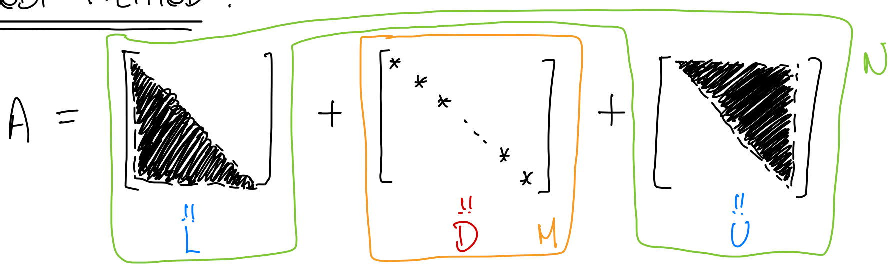
$$\text{PSEUDO-PSEUDO RATE OF CONVERGENCE} \equiv \rho(B)$$

$$\text{AVERAGE RATE OF CONVERGENCE OVER } k\text{-STEPS} \equiv \|B^k\|^{1/k}$$

$$\lim_{k\to\infty} \|B^k\|^{1/k} = \rho(B)$$

---

* JACOBI METHOD:



$$M = D, \quad N = -(L+U) \Rightarrow \boxed{\begin{array}{l} M x^{(k)} = (L+U)\, x^{(k)} + b \\ B_{\text{Jacobi}} = D^{-1}(L+U) \end{array}}$$

Jacobi doesn't work if $D \equiv 0$... so it's only good strictly diag. dom. matrices (e.g. >0)

> **THM:** If A is strictly diagonally dominant, then $\|B_J\| < 1$

ALGORITHM: (JACOBI METHOD) For $j = 1, \dots, n,$ ← easily parallelizable

$$x_j^{(k+1)} := \frac{1}{a_{jj}} \left[b_j - \sum_{l=1}^{j-1} a_{jl} x_j^{(k)} - \sum_{l=j+1}^{n} a_{jl} x_j^{(k)} \right], \quad k = 1, 2, 3, \dots$$

---

* GAUSS - SEIDEL

$$A = \underbrace{\underset{L}{\begin{bmatrix} * & & \\ * & \ddots & \\ * & * & \ddots \end{bmatrix}} + \underset{D}{\begin{bmatrix} * & & & \\ & * & & \\ & & \ddots & \\ & & & * \end{bmatrix}}}_{M} + \underbrace{\underset{U}{\begin{bmatrix} \ddots & * & * \\ & \ddots & * \\ & & \ddots \end{bmatrix}}}_{N}$$

$$\Rightarrow \quad (L + D)\, x^{(k+1)} = U x^{(k)} + b$$

$$B_{GS} = (L + D)^{-1} U$$

> **THM:** If A is strictly diagonally dominant, then $\|B_{GS}\| < 1$

ALGORITHM: (GAUSS - SEIDEL) For $j = 1, \dots, n,$

cf. Jacobi

$$x_j^{(k+1)} = \frac{1}{a_{jj}} \left[b_j - \sum_{l=1}^{j-1} a_{jl} x_j^{(k+1)} - \sum_{l=j+1}^{n} a_{jl} x_j^{(k)} \right]$$

← not parallelizable

← Twice as fast than Jacobi

THM: Under mild-conditions $\rho(B_{GS}) = \rho(B_J)^2$

——— // ———

* SOR

ALGORITHM (SOR): For $j = 1, 2, \dots, n$,

$$x_j^{(k+1)} = \frac{\omega}{a_{jj}} \left[b_j - \sum_{\ell=1}^{j-1} a_{\ell j} \overset{GS}{x_j^{(k+1)}} - \sum_{\ell=j+1}^{n} a_{\ell j} x_j^{(k)} \right] + \underline{(1-\omega)\, x_j^{(k)}}$$

previous vec. contains useful info.

$0 < \omega < 2$ (EXTRAPOLATION)

$$\omega = \frac{2}{1 + \sqrt{1 + \rho(B_J)^2}} \rightsquigarrow \boxed{B_\omega = \left(\frac{1}{\omega} D + L\right)^{-1} \left[\left(\frac{1}{\omega} - 1\right) D - U \right]}$$