FUNDAMENTALS

MATRIX - VECTOR MULTIPLICATION

x: n COLUMN VECTOR $\in \mathbb{C}^n$

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

(m: # OF ROWS; n: # OF COLUMNS.)

$$b = Ax$$

b: m COLUMN VECTOR $\in \mathbb{C}^m$

$$b_i = \sum_{j=1}^{n} a_{ij} x_j \qquad i = 1, \ldots, m$$

THIS IS A LINEAR MAP

$$x \longmapsto Ax$$

SINCE

$$A(x+y) = Ax + Ay$$

$$A(\alpha x) = \alpha Ax$$

$x, y \in \mathbb{C}^n$, $\alpha \in \mathbb{C}$

WE CAN WRITE M-V MULTIPLICATION SLIGHTLY DIFFERENTLY.

$$b = \sum_{j=1}^{n} x_j \underbrace{a_j}_{j\text{TH COLUMN OF } A}$$

$$\begin{bmatrix} b \end{bmatrix} = x_1 \begin{bmatrix} a_1 \end{bmatrix} + x_2 \begin{bmatrix} a_2 \end{bmatrix} + \ldots + x_n \begin{bmatrix} a_n \end{bmatrix}$$

b IS A LINEAR COMBINATION OF THE COLUMNS OF A.

MATRIX TIMES A MATRIX.

$A: l \times m$, $C: m \times n$, $B: l \times n$

(1) $b_{ij} = \sum_{k=1}^{m} a_{ik} c_{kj}$ $\left(B = AC\right)$

(2) $b_j = Ac_j = \sum_{k=1}^{m} c_{kj} a_k$

jTH COLUMN OF B | b_j IS A LIN. COMBINATION OF THE COLUMNS OF A WITH COEFFICIENTS c_{kj}

EXAMPLE: OUTER PRODUCT.

$u \in \mathbb{C}^m$

$v \in \mathbb{C}^n$

$$uv^T = \begin{bmatrix} \\ u \\ \\ \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$$

$$= \left[\begin{array}{c|c|c|c} v_1 u & v_2 u & \cdots & v_n u \end{array}\right]$$

DEF: RANGE OF A, RANGE(A) IS THE SET OF VECTORS THAT CAN BE EXPRESSED AS Ax FOR SOME x.

THM: RANGE(A) IS THE SPACE SPANNED BY THE COLUMNS OF A.

DEF: NULL SPACE OF $A \in \mathbb{C}^{m \times n}$, NULL(A), IS THE SET OF VECTORS x THAT SATISFY $Ax = 0$.

DEF: RANK OF A IS THE DIMENSION OF THE COLUMN SPACE OF A.

A MATRIX A $(m \times n)$ IS FULL RANK IF IT HAS THE MAXIMUM POSSIBLE RANK.

(THE LESSER OF m AND n)

IF $m \geq n$ ~~IFF~~ THEN A ~~null~~ MUST HAVE n LINEARLY INDEPENDENT COLUMNS TO BE FULL RANK.

THM: A MATRIX $A \in \mathbb{C}^{m \times n}$ WITH $m \geq n$ HAS FULL RANK IFF IT MAPS NO TWO DISTINCT VECTORS TO THE SAME VECTOR.

INVERSE

A NONSINGULAR OR INVERTIBLE MATRIX IS A SQUARE MATRIX $(m \times m)$ OF FULL RANK.

$\Rightarrow$ COLUMNS PROVIDE A BASIS FOR $\mathbb{C}^m$.

CAN WRITE

$$\underset{\overline{T}}{e_j} = \sum_{i=1}^{m} z_{ij} \underset{\overline{T}}{a_i}$$

UNIT VECTOR WITH 1 IN THE j^{TH} ENTRY

i^{TH} COLUMN OF A.

$$\left[e_1 \,\middle|\, e_2 \,\middle|\, \cdots \,\middle|\, e_m \right] = \underset{\overline{T}}{I} = A \underset{\overline{T}}{Z}$$

$m \times m$ IDENTITY

A^{-1}.

INVERSE

$$A^{-1}A = I$$

↑ INVERSE OF A

THM: FOR $A \in \mathbb{C}^{m \times m}$ THE FOLLOWING ARE EQUIVALENT.

(a) A HAS AN INVERSE A^{-1}

(b) RANK $(A) = m$

(c) RANGE $(A) = \mathbb{C}^m$

(d) NULL $(A) = \{0\}$

(e) 0 IS NOT AN EIGENVALUE OF A

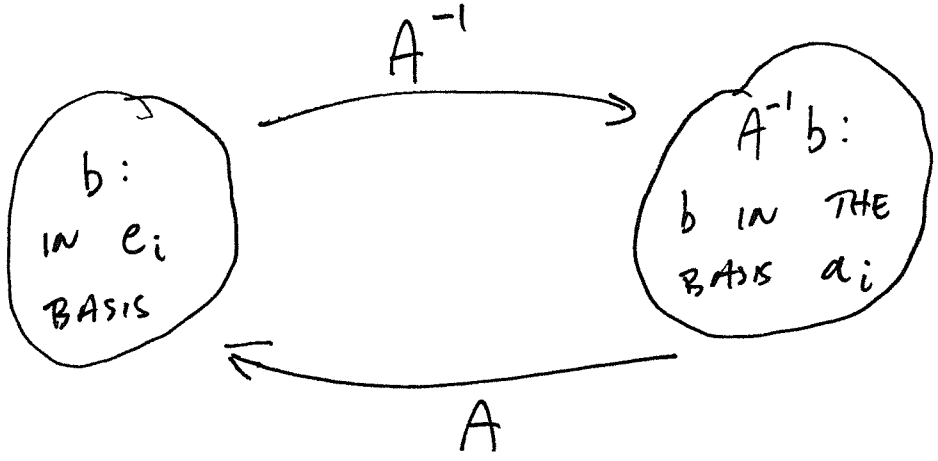
(f) 0 IS NOT A SINGULAR VALUE OF A

(g) $\det(A) \neq 0$

$(Ax = b)$

SINCE $x = A^{-1}b$ IS THE VECTOR OF COEFFICIENTS OF THE UNIQUE LINEAR EXPANSION OF b IN THE BASIS OF THE COLUMNS OF A

MULTIPLICATION BY A^{-1} CAN BE VIEWED AS A CHANGE OF BASIS.



ORTHOGONAL VECTORS AND MATRICES

DEF: ADJOINT OR HERMETIAN CONJUGATE OF $A \in \mathbb{C}^{m \times n}$ WE WRITE AS $A^* \in \mathbb{C}^{n \times m}$

WHERE

$$a^*_{ij} = \overline{a}_{ji}$$

(COMPLEX CONJUGATE OF a_{ji})

EXAMPLE

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}, \quad A^* = \begin{bmatrix} \overline{a}_{11} & \overline{a}_{21} & \overline{a}_{31} \\ \overline{a}_{12} & \overline{a}_{22} & \overline{a}_{23} \end{bmatrix}$$

IF $A = A^*$, WE SAY A IS HERMETIAN.

FOR REAL MATRICES.

$$A^* = A^T$$

IF $A = A^T$ THE MATRIX IS SYMMETRIC.

FOR MATRICES ~~WITH~~ A, B WITH COMPATIBLE DIMENSIONS.

$$(AB)^* = B^* A^*$$

(FOR INVERTIBLE MATRICES $(AB)^{-1} = B^{-1} A^{-1}$)

INNER PRODUCT

· $x, y \in \mathbb{C}^m$

→ $x^* y = \sum_{i=1}^{m} \overline{x}_i y_i$

EUCLIDEAN LENGTH OF A VECTOR

$$\|x\| = \sqrt{x^* x}$$

INNER PRODUCT IS ~~BILEAR~~ BILINEAR.

- $(x_1 + x_2)^* y = x_1^* y + x_2^* y$
- $x^*(y_1 + y_2) = x^* y_1 + x^* y_2$
- $(\alpha x)^*(\beta y) = \bar{\alpha}\beta\, x^* y$

ORTHOGONAL VECTORS

- $x, y \in \mathbb{C}^n$ ARE ORTHOGONAL IF $x^* y = 0$
- SETS OF VECTORS X AND Y ARE ORTHOGONAL IF $x^* y = 0 \quad \forall\ x \in X,\ y \in Y.$

- A SET OF VECTORS S IS ITSELF ORTHOGONAL IF FOR ALL $x, y \in S$ $x \neq y$

$$x^* y = 0.$$

THE
- ~~A~~ SET ~~OF~~ S IS ~~ORTHOR~~ ORTHONORMAL IF WE ALSO HAVE $\|x\| = 1$ FOR EACH $x \in S$

THM: THE VECTOR IN A ORTHOGONAL SET S ARE LINEARLY INDEPENDENT.

COMPONENTS OF A VECTOR

$\{q_1, q_2, \cdots, q_n\}$ IS AN ~~ORTHOGONAL~~ ORTHONORMAL SET AND V IS AN ARBITRARY VECTOR ($ALL \in \mathbb{C}^m$)

DEFINE:

$$r = V - (q_1^* V) q_1 - \cdots - (q_n^* V) q_n$$

r IS ORTHOGONAL TO THE SET $\{q_1, \cdots, q_n\}$ SINCE

$$q_i^* r = q_i^* V - (q_1^* V)(q_i^* q_1) - \cdots - (q_n^* V)(q_i^* q_n)$$

SINCE $q_i^* q_j = 0$ FOR $i \neq j$

$$q_i^* q_i = 1$$

$$q_i^* r = q_i^* V - (q_i^* V)\underbrace{(q_i^* q_i)}_{1} = 0$$

THUS,

$$V = r + \sum_{i=1}^{n} \overbrace{(q_i^* V)}^{\text{COEFFICIENT}} q_i \quad \text{(VECTOR)}$$

$$= r + \sum_{i=1}^{n} \underbrace{(q_i q_i^*)}_{\text{RANK-1}} V$$

IF $\{q_i\}$ IS A BASIS FOR $\mathbb{C}^n$, THEN $n = m$ AND $r = 0$.

THUS,

$$V = \sum_{i=1}^{m} \overbrace{(q_i q_i^*)}^{\text{PROJECTOR}} V$$

UNITARY MATRICES

$Q \in \mathbb{C}^{m \times m}$ IS UNITARY

IF $Q^* = Q^{-1}$.

FOR REAL MATRICES, A MATRIX IS ORTHOGONAL IF $Q^T = Q^{-1}$

$(Q \in \mathbb{R}^{m \times m})$

FOR A UNITARY MATRIX,

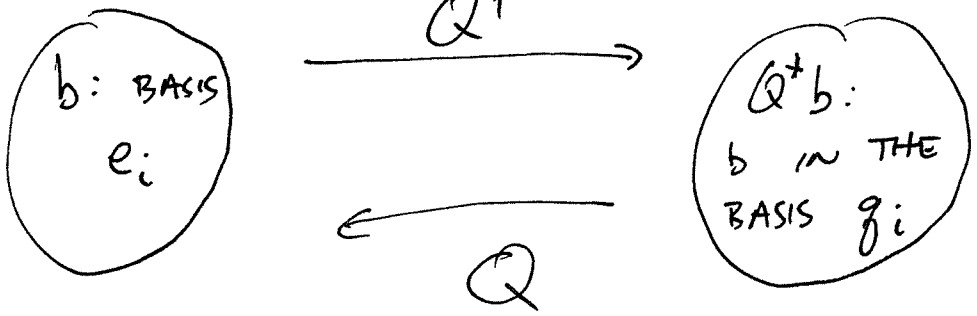
$$Q^* Q = I$$

$$\begin{bmatrix} \text{---} & q_1^* & \text{---} \\ \text{---} & q_2^* & \text{---} \\ & \vdots & \\ \text{---} & q_n^* & \text{---} \end{bmatrix} \begin{bmatrix} | & | & & | \\ q_1 & q_2 & \cdots & q_n \\ | & | & & | \end{bmatrix} = I$$

THIS HOLDS IF

$$q_i^* q_j = \delta_{ij} \quad (\text{KRONECKER DELTA})$$

MULTIPLICATION BY Q^*, Q



OTHER PROPERTIES

- $\|Qx\| = \|x\|$
- $(Qx)^* (Qy) = x^* y$

- HERMITIAN NOT HERMETIAN

- ANN BREW (LIBRARIAN)

  ann.brew@imperial.ac.uk

- LEC. 3 & LEC. 6 (TREFETHEN AND BAU)

VECTOR NORMS

- MEASURE THE "LENGTH" OF A VECTOR.

A NORM IS A FUNCTION

$\|\cdot\| : \mathbb{C}^m \longrightarrow \mathbb{R}$ THAT SATISFIES

(1) $\|x\| \geq 0$ AND $\|x\| = 0$ ONLY IF $x = 0$

(2) $\|x+y\| \leq \|x\| + \|y\|$ (Δ-INEQUALITY)

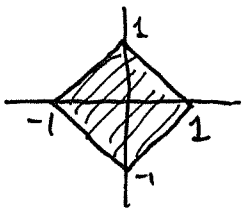
(3) $\|\alpha x\| = |\alpha| \|x\|$

$\forall\ x, y \in \mathbb{C}^m$ AND $\alpha \in \mathbb{C}$

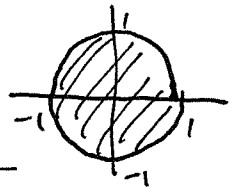
ALREADY SAW THE EUCLIDEAN NORM, BUT THIS IS PART OF A LARGER CLASS OF P-NORMS.

- $\|x\|_1 = \sum_{i=1}^{m} |x_i|$

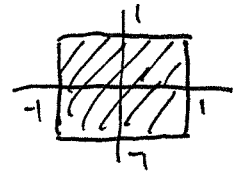
FOR $x \in \mathbb{R}^2$ $\{x \mid \|x\| \leq 1\}$



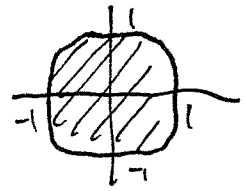
- $\|x\|_2 = \left(\sum_{i=1}^{m} |x_i|^2 \right)^{1/2}$



- $\|x\|_\infty = \max_{1 \leq i \leq m} |x_i|$



- $\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p \right)^{1/p}$



WEIGHTED NORMS

$$\|x\|_W = \|Wx\|$$

W IS A DIAGONAL MATRIX WITH $w_{ii} \neq 0$ $\forall i$.

WEIGHTED 2-NORM

$$\|x\|_w = \left(\sum_{i=1}^{m} |w_{ii} x_i|^2 \right)^{1/2}$$

PROJECTORS

A PROJECTOR IS A SQUARE MATRIX THAT SATISFIES

$$P^2 = P$$

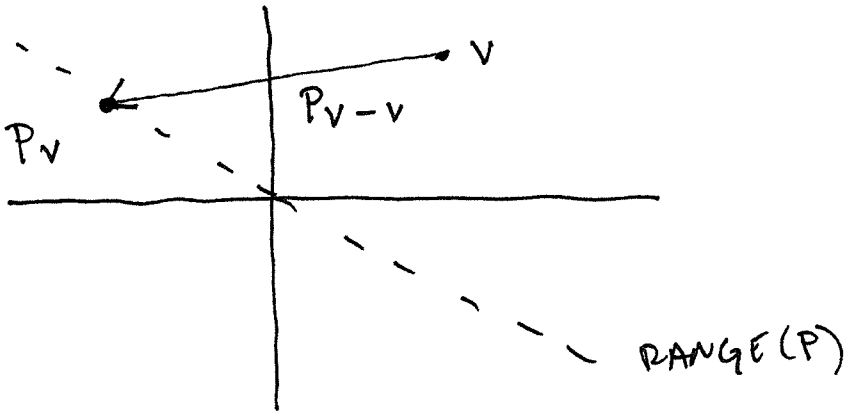
(ALSO SAID TO BE IDEMPOTENT)

- IF $v \in$ RANGE (P)

$\Rightarrow$ $v = Px$ FOR SOME x

$$Pv = P^2 x = Px = v$$

- SUPPOSE THAT $Pv \neq v$



$$P(Pv - v) = P^2 v - Pv = 0$$

$Pv - v \in$ NULL(P).

COMPLEMENTARY PROJECTOR.

IF P IS A PROJECTOR,

$I - P$ IS ALSO A PROJECTOR.

$$(I-P)^2 = I^2 - 2P + P^2 = I - P$$

$I-P$ IS THE COMPLEMENTARY PROJECTOR OF P.

· IF $Pu = 0$ WE HAVE

$$(I-P)u = u$$

$$\Rightarrow \text{RANGE}(I-P) \supseteq \text{NULL}(P)$$

· ALSO HAVE $\text{RANGE}(I-P) \subseteq \text{NULL}(P)$ SINCE

$$(I-P)u = u - Pu \in \text{NULL}(P)$$

$$\Rightarrow \text{RANGE}(I-P) = \text{NULL}(P)$$

SINCE WE ~~CAN~~ ~~WRITE~~ ~~INSTEAD~~ ALSO HAVE $P = I - (I-P)$,

$$\text{NULL}(I-P) = \text{RANGE}(P)$$

THUS, $\text{NULL}(I-P) \cap \text{NULL}(P) = \{0\}$

AND, $\text{RANGE}(P) \cap \text{RANGE}(I-P) = \{0\}$

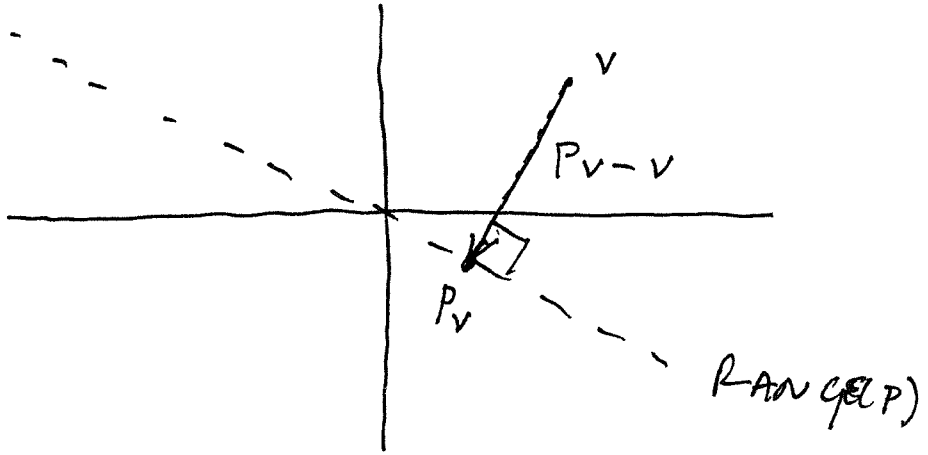
A PROJECTOR SEPARATES $\mathbb{C}^n$ INTO TWO SPACES.

IN FACT, SUPPOSE SUBSPACES, $S_1, S_2 \subseteq \mathbb{C}^n$ s.t. $S_1 \cap S_2 = \{0\}$ AND $S_1 + S_2 = \mathbb{C}^n$, THEN $\exists$ PROJECTOR P s.t. $\text{RANGE}(P) = S_1$ AND $\text{NULL}(P) = S_2$

P PROJECTS ONTO S_1 ALONG S_2.

ORTHOGONAL PROJECTORS

AN ORTHOGONAL PROJECTOR IS ONE THAT PROJECTS ONTO S_1 ALONG S_2 WHERE S_1 AND S_2 ARE ORTHOGONAL.



CAN CONSTRUCT ORTHOGONAL PROJECTORS FROM SETS OF ORTHONORMAL VECTORS.

LET $\{q_1, \cdots, q_n\}$ A SET OF n ORTHONORMAL VECTORS IN $\mathbb{C}^m$

LET

$$\hat{Q} = \left[q_1 \middle| q_2 \middle| \cdots \middle| q_n \right]$$

HAVE SEEN THAT $v \in \mathbb{C}^m$

$$v = r + \sum_{i=1}^{n} (q_i q_i^*) v$$

r: COMPONENT OF v ORTHOGONAL TO THE COLUMN SPACE OF $\hat{Q}$

$\sum_{i=1}^{n} (q_i q_i^*) v$: COMPONENT OF v IN THE COLUMN SPACE OF $\hat{Q}$

THUS THE map

$$V \longmapsto \underbrace{\sum_{i=1}^{n} (q_i q_i^*) v}_{y}$$

IS AN ORTHOGONAL PROJECTION ONTO RANGE $(\hat{Q})$.

$$y = \sum_{i=1}^{n} (q_i q_i^*) v = \underbrace{\hat{Q}\hat{Q}^*}_{P} v$$

THE COMPLEMENTARY PROJECTOR IS $I - \hat{Q}\hat{Q}^*$

AN IMPORTANT CASE $n = 1$

$$P = qq^* \quad \text{(RANK-1 ORTHOGONAL PROJECTOR)}$$

$$P_\perp = I - qq^* \quad \text{(RANK } m-1\text{)}$$

QR FACTORIZATION

$A = QR$ (FULL QR)

- A: $m \times n$
- Q: $m \times m$ UNITARY MATRIX ($Q^{-1} = Q^*$)
- R: $m \times n$ UPPER TRIANGULAR MATRIX

$r_{ij} = 0 \quad i > j$

$R = \begin{bmatrix} \triangle \\ 0 \end{bmatrix}$

$A = \hat{Q}\hat{R}$ (REDUCED QR)

- $\hat{Q}$: $m \times n$ MATRIX USE COLUMNS FORM AN ORTHONORMAL SET OF VECTORS.
- $\hat{R}$: $n \times n$ UPPER TRIANGULAR

WHY DO THIS?

(1) SOLVE LINEAR SYSTEMS

$Ax = b$

A: $m \times m$, FULL RANK.

(a) $QRx = b$ (PERFORM QR)

(b) MULTIPLY BY Q^*

$$\underbrace{Q^*Q}_{I} Rx = \underbrace{Q^*b}_{y}$$

(c) USE BACKWARDS SUBSTITUTION TO SOLVE $Rx = y$

LAST ROW

$$r_{mm} x_m = y_m$$

$$x_m = y_m / r_{mm}$$

PENULTIMATE ROW

$$r_{m-1\,m-1} x_{m-1} + r_{m-1\,m} x_m = y_{m-1}$$

(2) LEAST SQUARE PROBLEMS.

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

FIND x s.t.

$\|Ax - b\|_2$ is MINIMIZED.

THE SOLUTION IS GIVEN BY

$$Ax = \hat{Q}\hat{Q}^* b$$

$$\hat{Q}\hat{R}x = \hat{Q}\hat{Q}^* b$$

MULTIPLY BY $\hat{Q}^*$

$$\Rightarrow \hat{R}x = \hat{Q}^* b$$

SOLVE FOR x USING BACKWARDS SUBSTITUTION.

COMPUTING QR FACTORIZATION.

- CLASSICAL GRAM-SCHMIDT.
- MODIFIED GRAM-SCHMIDT.
- HOUSEHOLDER TRIANGULARIZATION.

USING THE MATRIX A $(m \times n)$ WE CAN CONSTRUCT SUCCESSIVE SPACES SPANNED BY ITS COLUMNS.

$$\langle a_1 \rangle \subseteq \langle a_1, a_2 \rangle \subseteq \langle a_1, a_2, a_3 \rangle \subseteq \cdots$$

$\langle \ldots \rangle$ SPACE SPANNED BY THE VECTORS IN THE BRACKETS.

IDEA BEHIND QR:

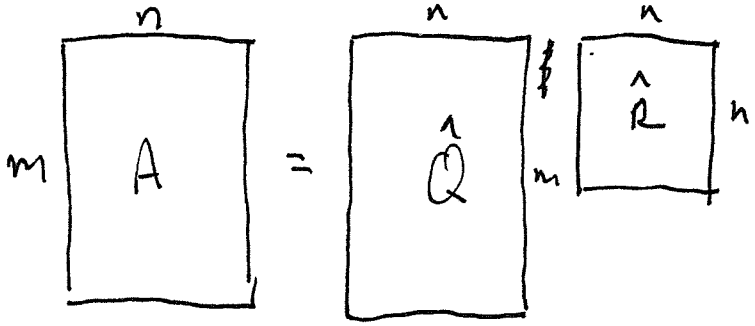
CONSTRUCT SEQUENCE OF ORTHONORMAL VECTORS $q_1, q_2, \ldots$ THAT SPAN THE SUCCESSIVE SPACES.

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} = \begin{bmatrix} q_1 & q_2 & \cdots & q_n \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & & \\ & 0 & \ddots & \\ & & & r_{nn} \end{bmatrix}$$

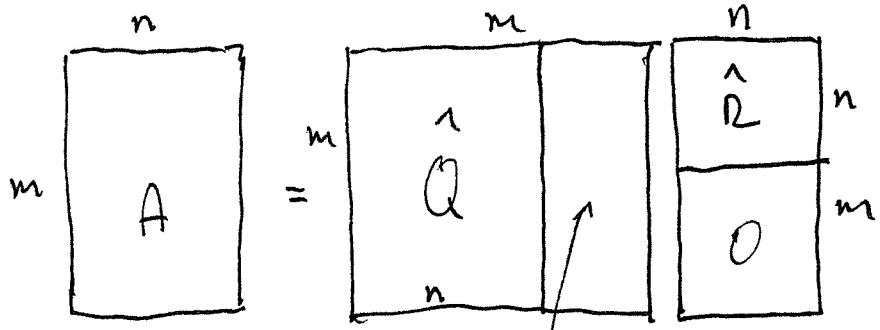
A $(m \times n)$ $\quad$ $\hat{Q}$ $(m \times n)$ $\quad$ $\hat{R}$ $(n \times n)$

THE REDUCED QR FACTORIZATION

FOR $m \geq n$



FULL QR $(m \geq n)$



Q

ORTHONORMAL

ADDITIONAL VECTORS THAT IF A IS FULL RANK SPAN RANGE(A)$^\perp$ $\{$ OR NULL$(A^*)\}$ DONE TO MAKE Q UNITARY.

GRAM-SCHMIDT ORTHOGONALIZATION

GIVEN $a_1, a_2, \cdots, a_n$ FIND

$q_1, q_2, \cdots, q_n$ AND r_{ij}

BY SUCCESSIVE ORTHOGONALIZATION

FIND $q_j \in \langle a_1, a_2, \dots, a_j \rangle$

THAT IS ORTHOGONAL TO

$q_1, q_2, \dots, q_{j-1}$

1. $v_j = a_j - (q_1^* a_j) q_1 - (q_2^* a_j) q_2 - \dots - (q_{j-1}^* a_j) q_{j-1}$

2. $q_j = v_j / \|v_j\|_2$

BY THE FACT THAT $A = \hat{Q}\hat{R}$

WE KNOW

$$q_1 = a_1 / r_{11}$$

$$q_2 = \frac{a_2 - r_{12} q_1}{r_{22}}$$

$$\vdots$$

$$q_n = \frac{a_n - \sum_{i=1}^{n-1} r_{in} q_i}{r_{nn}}$$

BY COMPARING THE TWO APPROACHES WE SEE THAT

$$r_{ij} = q_i^* a_j$$

AND $|r_{jj}| = \| a_j - \sum_{i=1}^{j-1} r_{ij} q_i \|_2$

CLASSICAL GRAM-SCHMIDT (CGS) ALGORITHM

for $j = 1$ to n

$v_j = a_j$

for $i = 1$ to $j-1$

$r_{ij} = q_i^* a_j$

$v_j = v_j - r_{ij} q_i$

END FOR

$r_{jj} = \|v_j\|_2$

$q_j = v_j / r_{jj}$

END FOR

CAN USE CGS TO PROVE EXISTENCE & UNIQUENESS (FULL RANK A)

(TREFETHEN & LAU p. 51)

BUT TURNS OUT TO BE UNSTABLE FOR NUMERICAL PURPOSES.

CGS PROJECTORS

GS CAN ALSO BE WRITEN INTERMS OF PROJECTORS.

$$q_1 = \frac{P_1 a_1}{\| P_1 a_1 \|}, \quad q_2 = \frac{P_2 a_2}{\| P_2 a_2 \|}$$

$$q_n = \frac{P_n a_n}{\| P_n a_n \|}$$

P_j IS AN ORTHOGONAL PROJECTOR AN $m \times m$ MATRIX THAT PROJECTS $\mathbb{C}^m$ ORTHOGONALLY ONTO $\langle q_1, q_2, \cdots, q_{j-1} \rangle$

LET $\hat{Q}_{j-1} = [q_1 | q_2 | \cdots | q_{j-1}]$

THEN

$$P_j = I - \hat{Q}_{j-1} \hat{Q}_{j-1}^*$$

AT EACH STEP, CGS COMPUTES

$$v_j = P_j a_j \qquad (*)$$

MODIFIED GS (MGS)

DECOMPOSE P_j INTO $j-1$ PROJECTORS OF RANK $m-1$

$$P_j = P_{\perp q_{j-1}} \cdots P_{\perp q_2} P_{\perp q_1}$$

WHERE

$$P_{\perp q_j} = I - q_j q_j^*$$

THUS, (*) CAN BE WRITTEN AS

$$v_j = P_{\perp q_{j-1}} \cdots P_{\perp q_2} P_{\perp q_1} q_j$$

AT EACH STEP MGS APPLIES $P_{\perp q_i}$ TO ALL COLUMNS.

MGS ALGORITHM

```
for i = 1 to n
    v_i = a_i
END FOR

for i = 1 to n
    r_ii = ||v_i||_2
    q_i = v_i / r_ii
    for j = i+1 to n
        r_ij = q_i^* v_j
        v_j = v_j - r_ij q_i
    END FOR
END FOR
```

MGS CAN BE THOUGHT OF AS TRIANGULAR ORTHOGONALIZATION.

· FIRST STEP CAN BE VIEWED AS

$$\begin{bmatrix} v_1 \mid v_2 \mid \cdots \mid v_n \end{bmatrix} \underbrace{\begin{bmatrix} \frac{1}{r_{11}} & -\frac{r_{12}}{r_{11}} & \cdots & -\frac{r_{1n}}{r_{11}} \\ & 1 & & 0 \\ & & \ddots & \\ & & & 1 \end{bmatrix}}_{R_1}$$

$$= \begin{bmatrix} q_1 \mid v_2^{(2)} \mid \cdots \mid v_n^{(2)} \end{bmatrix}$$

RIGHT MULTIPLY BY R_1

SECOND STEP: RIGHT MULTIPLY BY R_2

$$R_2 = \begin{bmatrix} 1 & 1 & & & 1 \\ & 1/r_{22} & -\frac{r_{23}}{r_{22}} & \cdots & -\frac{r_{2n}}{r_{22}} \\ & 0 & 1 & & 0 \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

DOING THIS n TIMES

$$\underbrace{A R_1 R_2 \cdots R_n}_{\hat{R}^{-1}} = \hat{Q}$$

TRIANGULAR ORTHOGONALIZATION USING TRIANGULAR MATRICES TO REDUCE A TO A MATRIX WITH ORTHONORMAL COLUMNS.

· R_i'S ARE NEVER FORMED EXPLICITLY BUT THEY GIVE US A CLEAR WAY OF THINKING ABOUT WHAT IS HAPPENING AT EACH STEP.

3. HOUSEHOLDER TRIANGULARIZATION

. ORTHOGONAL TRIANGULARIZATION.

TRIANGULARIZE A USING ORTHOGONAL MATRICES.

$$\underbrace{Q_{m-1} \cdots Q_2 Q_1}_{Q^t} A = R$$

FORM FULL QR OF A.

TRIANGULARIZE BY INTRODUCING ZEROS.

$$\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix} \xrightarrow{Q_1} \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & x & x \\ 0 & x & x \end{bmatrix} \xrightarrow{Q_2} \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & x \end{bmatrix}$$

A $\qquad$ $Q_1 A$ $\qquad$ $Q_2 Q_1 A$

$$\xrightarrow{Q_3} \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & 0 \end{bmatrix}$$

$$Q_3 Q_2 Q_1 A$$

FIND Q_k (UNITARY) TO INTRODUCE ZEROS BELOW THE kTH DIAGONAL WHILE PRESERVING THE ZEROS PREVIOUSLY INTRODUCED.

APPROACH

$$Q_k = \begin{bmatrix} I & 0 \\ 0 & F \end{bmatrix}$$

I IS $(k-1)\times(k-1)$ IDENTITY

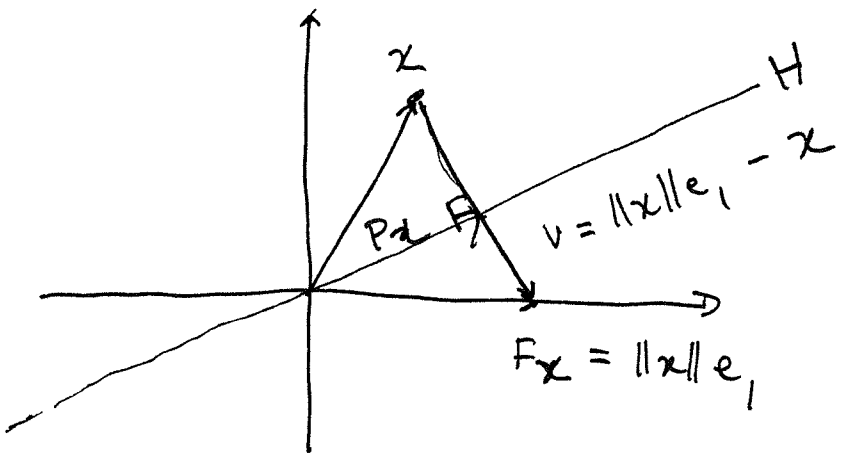
F IS $(m-k+1)\times(m-k+1)$ UNITARY MATRIX CALL HOUSEHOLDER REFLECTOR.

FOR THE k^{TH} STEP

THE $k,\dots,m$ ENTRIES OF THE k^{TH} COLUMN WE WRITE AS $x \in \mathbb{C}^{m-k+1}$

$$x = \begin{bmatrix} x \\ x \\ x \\ x \\ x \end{bmatrix} \xrightarrow{F} Fx = \begin{bmatrix} \|x\| \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$= \|x\| e_1$$



F MAPS A VECTOR ON ONE SIDE OF H TO ITS MIRROR IMAGE.

THE PROJECTOR OF x ONTO H IS

$$P = I - \frac{vv^*}{v^*v}$$

TO REFLECT ABOUT H WE NEED TO GO TWICE AS FAR IN THE DIRECTION OF v.

$$F = I - 2\frac{vv^*}{v^*v}$$

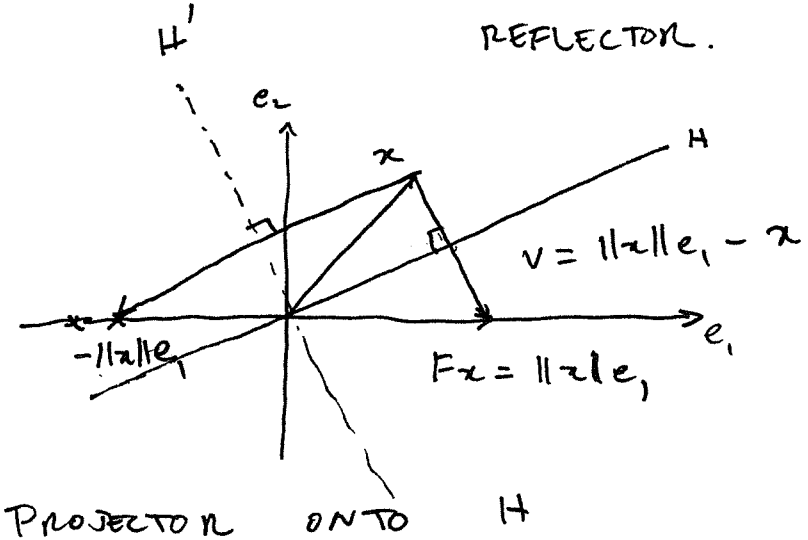
FULL RANK AND UNITARY.

HOUSEHOLDER TRIANGULARIZATION

$$Q_n \cdots Q_2 Q_1 A = R$$

$$Q_k = \begin{bmatrix} I & 0 \\ 0 & F \end{bmatrix}$$

↑ HOUSEHOLDER REFLECTOR.



PROJECTOR ONTO H

$$P = I - qq^* \quad \text{WHERE} \quad q = \frac{v}{\|v\|}$$

TO GO TWICE THE DISTANCE IN q

$$F = I - 2qq^* = I - 2\frac{vv^*}{v^*v}$$

TURNS OUT THERE ARE TWO CHOICES.

WHILE MATHEMATICALLY SPEAKING BOTH CHOICES ARE EQUALLY GOOD, NUMERICALLY WE'D LIKE TO CHOOSE THE ~~REFLECT~~ REFLECTION FURTHEST AWAY FROM x.

$$v = -\operatorname{sign}(x_1)\|x\| e_1 - x$$

OR

$$v = \operatorname{sign}(x_1)\|x\| e_1 + x$$

TAKING ALSO $\operatorname{sign}(0) = 1$

~~Hobes~~

HOUSEHOLDER ALGORITHM

for $k = 1$ to n

$x = A_{k:m,k}$ ← VECTOR OF THE k THROUGH m ENTRIES OF THE kTH COLUMN

$$v_k = \text{sign}(x_1)\|x\|_2 e_1 + x$$

$$v_k = v_k / \|v_k\|_2$$

$$A_{k:m,k:n} = A_{k:m,k:n} - 2v_k(v_k^* A_{k:m,k:n})$$

END. FOR

A IS REDUCED TO UPPER TRIANGULAR FORM WITHOUT EVER EXPLICITLY COMPUTING Q.

HOWEVER, WE KNOW HOW TO APPLY Q (AND Q^*) ONCE THE v_k'S ARE KNOWN

SINCE

$$Q^* = Q_n \cdots Q_2 Q_1$$

$$Q = Q_1 Q_2 \cdots Q_n$$

RATHER THAN COMPUTING Q FROM THE Q_k'S THEN APPLYING Q_1 THE SEQUENCE CAN BE APPLIED DIRECTLY

IMPLICIT CALCULATION OF Q^*b

for $k = 1$ to n

$$b_{k:m} = b_{k:m} - 2v_k(v_k^* b_{k:m})$$

END FOR

IMPLICIT CALCULATION OF ~~Q^*~~ $Q x$

for $k = n$ DOWN TO 1

$$x_{k:m} = x_{k:m} - 2 v_k (v_k^* x_{k:m})$$

END FOR

TO FORM Q WE RECOGNIZE THAT

$$Q I = Q$$

AND PERFORM THE IMPLICIT CALCULATION

$$Q e_i = q_i \quad \text{FOR ALL } i = 1, \ldots, m$$

EXAMPLE: LEGENDRE POLYNOMIALS

$P_n(x)$ ON $[-1, 1]$

POLYNOMIALS THAT ARE ORTHOGONAL WITH RESPECT TO

THE INNER PRODUCT

$$(f, g) = \int_{-1}^{1} \overline{f(x)} g(x) \, dx$$

FOR $P_n(x)$ AND $P_m(x)$

$$\int_{-1}^{1} \overline{P_n(x)} P_m(x) \, dx = \frac{2}{2n+1} \delta_{nm}$$

THE FIRST FEW ARE

$$P_0(x) = 1$$

$$P_1(x) = x$$

$$P_2(x) = \frac{3x^2}{2} - \frac{1}{2}$$

$$P_3(x) = \frac{5x^3}{2} - \frac{3}{2} x$$

$$\vdots$$

THEY CAN BE GENERATED BY THE CONTINUOUS VERSION OF QR APPLIED TO THE "MATRIX" WHOSE "COLUMNS" ARE THE MONOMIALS

$$A = \left[\, 1 \;\middle|\; x \;\middle|\; x^2 \;\middle|\; x^3 \;\middle|\; \cdots \;\middle|\; x^{n-1} \,\right]$$

USING A CONTINUOUS VERSION OF GRAM-SCHMIDT REPLACING

$q_i^* v_j$ BY $\underbrace{\int_{-1}^{1} \overline{q_i(x)}\, v_j(x)\, dx}_{r_{ij}}$

AND

$\|v\|_2$ BY $\underbrace{\left(\int_{-1}^{1} \overline{v(x)} v(x)\, dx \right)^{1/2}}_{r_{ii}}$

WE OBTAIN

$$A = \left[\, q_0(x) \;\middle|\; q_1(x) \;\middle|\; \cdots \;\middle|\; q_{n-1}(x) \,\right] \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \\ & & r_{nn} \end{bmatrix}$$

$P_n(x)$ IS RELATED TO $q_n(x)$ THROUGH

$$P_n(x) = \frac{q_n(x)}{q_n(1)}$$

DISCRETE LEGENDRE POLYNOMIALS

THE CONTINUOUS FORMULATION CAN BE MADE DISCRETE BY EVALUATING THE MONOMIALS AT m EQUALLY SPACED POINTS.

$$x_i = \frac{2}{m}(i-1) - 1, \quad i = 1, \ldots, m$$

THIS TURNS A INTO AN $m \times n$ VANDERMONDE MATRIX

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & & x_m^{n-1} \end{bmatrix}$$

BY PERFORMING THE STANDARD QR, THE RESULTING COLUMNS OF Q, q_k, WILL APPROXIMATE $q_k(x_i)$ SINCE THE DISCRETE INNER PRODUCT IS PROPORTIONAL TO AN APPROXIMATION OF THE CONTINUOUS ONE

$$f^* g = \sum_{i=1}^{m} \bar{f}_i g_i$$

$$\approx \sum_{i=1}^{m} \bar{f}(x_i) g(x_i)$$

$$\approx \frac{n}{2} \int_{-1}^{1} \bar{f}(x) g(x) \, dx$$

THE APPROXIMATE VALUES OF $P_{mk}(x_i)$ IS

$$P_k(i) = \frac{q_k(i)}{q_k(m)} \quad \left(\begin{matrix}\text{MATLAB}\\ \text{NOTATION}\end{matrix}\right)$$

LEAST SQUARES PROBLEMS

$A \in \mathbb{C}^{m \times n} \quad m \geq n$

$b \in \mathbb{C}^m$

FIND $x \in \mathbb{C}^n$ SUCH THAT

$\|\underbrace{b - Ax}_{r}\|_2$ IS MINIMIZED

r, RESIDUAL.

THEOREM

GIVEN THE LEAST SQUARES PROBLEM, A VECTOR

x MINIMIZES $\|r\|_2$

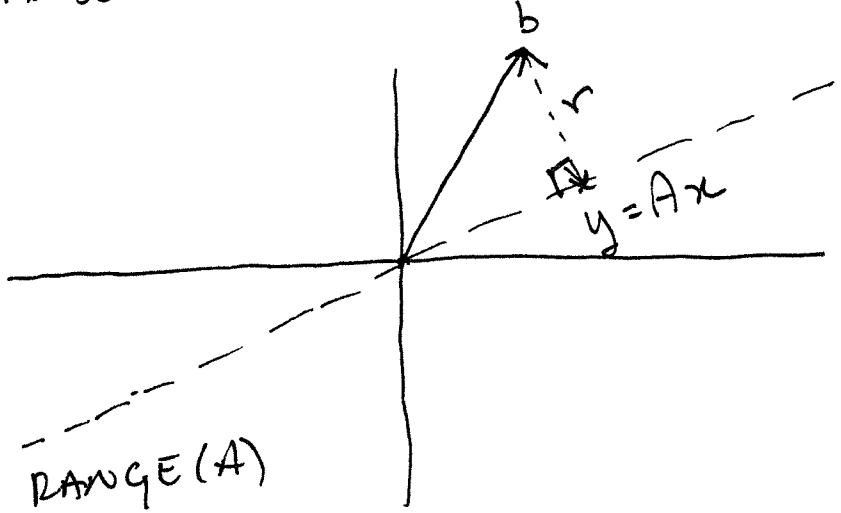
IFF $r \perp \text{RANGE}(A)$, THAT IS

(i) $A^* r = 0$

(ii) $A^* A x = A^* b$ $\left(\begin{matrix}\text{NORMAL}\\ \text{EQUATIONS}\end{matrix}\right)$

(iii) $Pb = Ax$

WHERE P IS THE ORTHOGONAL PROJECTOR ONTO THE RANGE(A)



IF A IS FULL RANK

THEN x IS UNIQUE.

USING REDUCED QR, WE CAN CONSTRUCT THE PROJECTOR P AS

$$P = \hat{Q}\hat{Q}^*$$

SINCE THE COLUMNS OF $\hat{Q}$ SPAN THE RANGE OF A.

THUS, FROM (iii)

$$Pb = \hat{Q}\hat{Q}^* b = Ax = \hat{Q}\hat{R}x$$

LEFT-MULTIPLYING BY $\hat{Q}^*$

WE HAVE

$$\hat{R}x = \hat{Q}^* b$$

THIS CAN BE SOLVED USING BACKWARDS SUBSTITUTION.

EXAMPLE: POLYNOMIAL FITTING.

GIVEN $x_1, \ldots, x_m$ AND $y_1, \ldots, y_m$. FIND

$$p(x) = c_0 + c_1 x + c_2 x^2 + \ldots c_{n-1} x^{n-1}$$

SUCH THAT

$$\sum_{i=1}^{m} |p(x_i) - y_i|^2$$

IS MINIMIZED.

WE CAN SET THIS UP AS A LEAST SQUARES PROBLEM.

$$b = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \quad \text{"}x\text{"} = \begin{bmatrix} c_0 \\ \vdots \\ c_{n-1} \end{bmatrix}$$

(UNKNOWNS)

$$A = \begin{bmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ \vdots & \vdots & & \\ 1 & x_m & & x_m^{n-1} \end{bmatrix}$$

IF x_i ARE EQUALLY SPACED, THEN $\hat{Q}$ ARE RELATED TO THE DISCRETE LEGENDRE POLYNOMIALS ON THE INTERVAL $[x_1, x_m]$.

THUS $\hat{Q}\hat{Q}^* b$ IS THE PROJECTION ONTO THE SPACE

SPANNED BY THE DISCRETE LEGENDRE POLYNOMIALS.

ANALYSING SCHEMES

FOR QR, WE ENCOUNTERED THREE DIFFERENT ALGORITHMS FOR ~~REALM~~ PERFORMING THE SAME COMPUTATION.

WHAT MAKES ONE APPROACH BETTER THAN THE OTHER?

(1) ONE COULD BE FASTER THAN THE OTHERS.

(2) ONE METHOD COULD BE MORE ROBUST THAN THAN THE OTHERS.

THE FIRST PROPERTY IS CHARACTERIZED BY THE OPERATION COUNT.

THE SECOND IS RELATED TO THE STABILITY OF THE ALGORITHM, OR HOW SENSITIVE IT IS TO PERTURBATIONS.

OPERATION COUNT

- COUNT THE NUMBER OF "FLOPS" (FLOATING OPERATIONS) THAT THE ALGORITHM REQUIRES IN THE LIMIT OF VERY LARGE MATRICES.

· THE OPERATIONS

$+, -, \times, \div, \sqrt{\ }$

ALL COST ONE FLOP

FOR EACH REAL NUMBER

THE OPERATION COUNT IS A CLASSICAL MEASURE OF THE COST. IN PRACTICE, THERE ARE ALSO OTHER IMPORTANT FACTORS.

(i) MOVEMENT OF DATA IN MEMORY.

(ii) PARALLEL VS. SERIAL COMPUTATION.

(iii) OTHER PROGRAMS RUNNING ON THE COMPUTER

⋮

OPERATION COUNT FOR MGS.

THE GRAM-SCHMIDT ALGORITHMS REQUIRE $\sim 2mn^2$ FLOPS TO COMPUTE A QR FACTORIZATION OF AN $m \times n$ MATRIX OF REAL NUMBERS.

"$\sim$" MEANS

$$\lim_{m,n \to \infty} \frac{\#\text{ OF FLOPS}}{2mn^2} = 1$$

FOR MGS, WHEN $m \& n$ ARE LARGE THE WORK IS DOMINATED BY WHAT IS DONE IN THE INNER LOOP

(i) $r_{ij} = q_i^* v_j$

INNER PRODUCT OF TWO VECTORS IN $\mathbb{R}^m$

m : $\times$

$m-1$: $+$

WHICH GIVE $2m-1$ FLOPS.

(ii) $v_j = v_j - r_{ij} q_i$

SUBTRACTION OF TWO VECTORS IN $\mathbb{R}^m$

m : $-$

SCALAR TIMES A VECTOR IN $\mathbb{R}^m$

m : $\times$

THIS REQUIRES $2m$ FLOPS.

THUS EACH ITERATION OF THE INNER LOOP REQUIRE $\sim 4m$ FLOPS.

$$\# \text{ OF FLOPS} \sim \sum_{i=1}^{n} \sum_{j=i+1}^{n} 4m$$

$$= 4m \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1$$

$$\sim 4m \sum_{i=1}^{n} i \qquad \left(\int_1^n x \, dx \right)$$

$$\sim 4m \frac{n^2}{2}$$

$$\sim 2mn^2$$

OPERATION COUNT FOR HOUSE HOLDER.

WORK IS DOMINATED BY

$$A_{k:m,k:n} = A_{k:m,k:n} - (2v_k)\left(v_k^* A_{k:m,k:n}\right)$$

(3) (2) (1)

(1) $v_k^* A_{k:m,k:n}$

$(n-k)$ INNER PRODUCTS OF VECTORS IN $\mathbb{R}^{m-k}$

OF FLOPS $\sim 2(n-k)(m-k)$

(2) $(2v_k)(\quad)$

· OUTER PRODUCT OF TWO VECTORS IN $\mathbb{R}^{m-k}$ AND $\mathbb{R}^{n-k}$

OF FLOPS $\sim (m-k)(n-k)$

(3) SUBTRACTION OF TWO $(m-k)\times(n-k)$ MATRICES

OF FLOPS $\sim (m-k)(n-k)$

TOTAL COST

$$\sim \sum_{k=1}^{n} 4(n-k)(m-k)$$

$$= 4\left(\sum_{k=1}^{n} nm - k(n+m) + k^2\right)$$

$$= 4\sum_{k=1}^{n} nm - 4\sum_{k=1}^{n} k(n+m) + 4\sum_{k=1}^{n} k^2$$

$$\sim 4mn^2 - 4(n+m)\frac{n^2}{2} + \frac{4n^3}{3}$$

$$= 4mn^2 - 2n^3 - 2mn^2 + \frac{4n^3}{3}$$

$$= 2mn^2 - \frac{2n^3}{3}$$

HOUSEHOLDER REQUIRES FEWER FLOPS TO COMPUTE QR.

MATRIX NORMS

- CAN THINK OF A MATRIX AS AN mn LENGTH VECTOR

- APPLY USUAL VECTOR NORMS TO THIS VECTOR.

- MUST SATISFY

(1) $\|A\| \geq 0$ AND $\|A\| = 0$ ONLY IF $A = 0$

(2) $\|A+B\| \leq \|A\| + \|B\|$

b) $\|\alpha A\| = |\alpha| \|A\|$

FROBENIUS NORM

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

CAN SHOW:

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

- INDUCED MATRIX NORM.

FOR AN $m \times n$ MATRIX A AND VECTOR NORMS $\|\cdot\|_{(n)}$ AND $\|\cdot\|_{(m)}$ ON THE DOMAIN AND RANGE

$$\|A\|_{(m,n)} = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_{(m)}}{\|x\|_{(n)}}$$

$$= \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_{(n)} = 1}} \|Ax\|_{(m)}$$

- MEASURES THE MAXIMUM FACTOR BY WHICH A WILL "STRETCH" A VECTOR.

EXAMPLES

(i) $\|A\|_2$, $A \in \mathbb{R}^{2\times 2}$

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

ALL UNIT2 VECTOR IN $\mathbb{R}^2$ CAN BE WRITTEN AS

$$x = (\cos\theta, \sin\theta) \qquad \theta \in [0, 2\pi)$$

$$Ax = \begin{bmatrix} \cos\theta + 2\sin\theta \\ 2\sin\theta \end{bmatrix}$$

$$\|Ax\|_2^2 = \cos^2\theta + 4\sin\theta\cos\theta + 8\sin^2\theta$$

TO FIND MAX. VALUE.

$$0 = \frac{\partial}{\partial\theta}\|Ax\|_2^2 = 14\sin\theta\cos\theta + 4(\cos^2\theta - \sin^2\theta)$$

$$0 = 7\sin 2\theta + 4\cos 2\theta$$

$$\Rightarrow \tan 2\theta = -\frac{4}{7}$$

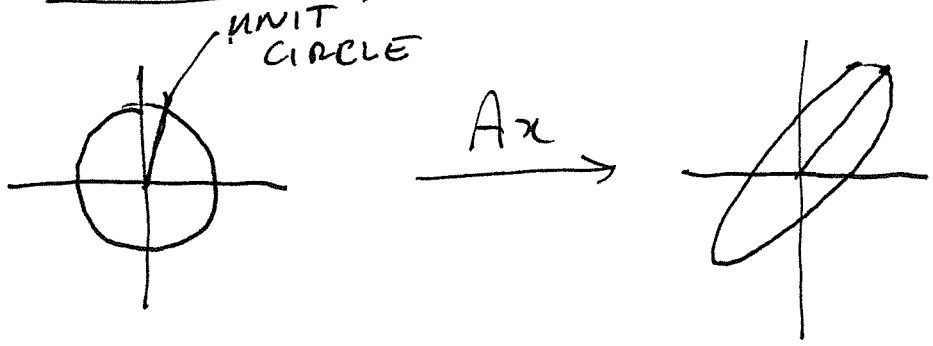
$$2\theta = \tan^{-1}\left(-\frac{4}{7}\right) + n\pi \qquad (n \in \mathbb{Z})$$

$$\theta = \frac{1}{2}\tan^{-1}\left(-\frac{4}{7}\right) + n\frac{\pi}{2}$$

MAXIMUM ~~NOT~~ FOR $n = 1$ FOR EXAMPLE, WHICH GIVES.

$$\|A\|_2 = 2.9208\ldots$$

GRAPHICALLY,



(ii) 2-NORM OF A DIAGONAL MATRIX

$$D = \begin{bmatrix} d_1 & & 0 \\ & d_2 & \\ & & \ddots \\ 0 & & d_m \end{bmatrix}$$

- STRETCHES & COMPRESSES ALONG THE DIRECTIONS e_i

THUS, $\|D\|_2 = \max_i |d_i|$.

OFTEN DIFFICULT TO FIND EXACT VALUES, SO WE LOOK FOR BOUNDS.

HÖLDER INEQUALITY

$$|x^* y| \leq \|x\|_p \|y\|_q$$

FOR $1 = \frac{1}{p} + \frac{1}{q}$

CAUCHY-SCHWARZ

$p = q = 2$

$$|x^* y| \leq \|x\|_2 \|y\|_2$$

EXAMPLE: 2-NORM OF AN OUTER PRODUCT.

$$A = uv^*$$

$$\|Ax\|_2 = \|u \overbrace{v^* x}^{\text{SCALAR}}\|_2$$

$$= |v^* x| \|u\|_2$$

$$\leq \|u\|_2 \|v\|_2 \|x\|_2$$

$$\Rightarrow \|A\|_2 \leq \|u\|_2 \|v\|_2$$

BOUNDS FOR $\|AB\|$

$A: l \times m$

$B: m \times n$

$$\|ABx\|_{(l)} \leq \|A\|_{(l,m)} \|Bx\|_{(m)}$$

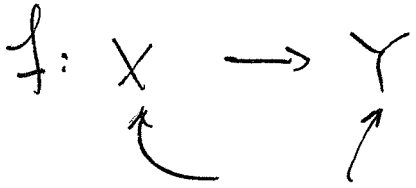
$$\leq \|A\|_{(l,m)} \|B\|_{(m,n)} \|x\|_{(n)}$$

$$\|AB\|_{(l,n)} \leq \|A\|_{(l,m)} \|B\|_{(m,n)}$$

CONDITIONING AND CONDITION NUMBER

- CONDITIONING PERTAINS TO THE PERTURBATION BEHAVIOUR OF THE MATHEMATICAL PROBLEM.

- A "PROBLEM" AS A FUNCTION



NORMED VECTOR SPACES.

f CAN BE NONLINEAR.

- USUALLY CONCERNED WITH THE BEHAVIOUR AT A PARTICULAR VALUE OF $x \in X$.

PROBLEM (INSTANCE)

- WELL-CONDITIONED PROBLEM
  ALL SMALL PERTURBATIONS OF x LEAD TO SMALL CHANGES IN $f(x)$.

- ILL-CONDITIONED PROBLEM

  SMALL CHANGES IN x

  GIVE LARGE CHANGES IN

  $f(x)$.

MEASURED VIA THE CONDITION

NUMBER.

CONDITION NUMBER

ABSOLUTE CONDITION NUMBER

INTRODUCE AN INFINITESIMAL PERTURBATION δx TO x.

$$\delta f = f(x+\delta x) - f(x)$$

ABSOLUTE CONDITION NUMBER

$$\hat{\kappa} = \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|}$$

FOR DIFFERENTIABLE $f(x)$

CAN FIND

$$J(x), \quad J_{ij}(x) = \frac{\partial f_i}{\partial x_j}$$

JACOBIAN

FOR INFINITESIMAL QUANTITIES

$$\delta f = J(x)\delta x$$

$$\hat{\kappa} = \sup_{\delta x} \frac{\|J(x)\,\delta x\|}{\|\delta x\|}$$

$$= \|J(x)\|$$

RELATIVE CONDITION NUMBER

$$\kappa = \sup_{\delta x} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right)$$

FOR DIFFERENTIABLE f

$$\kappa = \sup \|J(x)\| \Big/ \frac{\|f(x)\|}{\|x\|}$$

FOR FLOATING POINT NUMBERS INTRODUCE RELATIVE ERRORS, SO

κ IS USED IN NUMERICAL ANALYSIS.

$\kappa = 1 - 10^2$ WELL-CONDITIONED

$\kappa > 10^6$ ILL-CONDITIONED.

EXAMPLES (i) $f : x \longmapsto \sqrt{x}$

$$J = \frac{df}{dx} = \frac{1}{2} x^{-1/2}$$

$$\kappa = \frac{\left| \frac{1}{2} x^{-1/2} \right|}{\frac{|x^{1/2}|}{|x|}} = \frac{1}{2}$$

(ii) FINDING THE ROOTS OF A POLYNOMIAL GIVEN THE COEFFICIENTS.

CONSIDER

$$x^2 - 2x + 1 = (x-1)^2$$

ROOTS ARE $x = 1$ TWICE.

NOW TAKE

$$x^2 - 2x + 0.9999$$

$$= (x - 0.99)(x - 1.01)$$

ROOTS ARE $x = 0.99$ AND $x = 1.01$

RELATIVE CHANGE OF 10^{-4} IN COEFF GIVES ~~#~~ RELATIVE CHANGE OF 10^{-2} IN THE ROOTS.

IN FACT, $\delta r = C\sqrt{\delta c}$

(δr: CHANGE IN ROOTS; C: CONST.; δc: CHANGE IN COEFFICIENTS)

THIS GIVES

$$\kappa = \infty$$

THIS PROBLEM IS ILL-CONDITIONED.

A MORE COMPLEX EXAMPLE

WITH WILKENSON POLYNOMIAL

$$p(x) = \prod_{i=1}^{20} (x - i)$$

CONDITION NUMBER OF MATRIX VECTOR MULTIPLICATION

Fix $A \in \mathbb{C}^{m \times n}$, $f_A : x \to Ax$

$$\kappa = \sup_{\delta x} \frac{\|A(x+\delta x) - Ax\|}{\|Ax\|} \Big/ \frac{\|\delta x\|}{\|x\|}$$

$$= \sup_{\delta x} \frac{\|A\,\delta x\|}{\|\delta x\|} \Big/ \frac{\|Ax\|}{\|x\|}$$

$$= \|A\| \Big/ \frac{\|Ax\|}{\|x\|}$$

IF $A \in \mathbb{C}^{m \times m}$ AND NONSINGULAR

$$\|x\| = \|A^{-1}Ax\|$$

$$\leq \|A^{-1}\| \|Ax\|$$

$$\Rightarrow \frac{\|x\|}{\|Ax\|} \leq \|A^{-1}\|$$

$$\Rightarrow \kappa \leq \|A\| \|A^{-1}\|$$

THEOREM: LET $A \in \mathbb{C}^{m \times m}$ AND NONSINGULAR AND CONSIDER $Ax = b$. THE PROBLEM OF COMPUTING b GIVEN x HAS

$$\kappa = \|A\| \frac{\|x\|}{\|b\|} \leq \|A\| \|A^{-1}\|$$

THE PROBLEM OF COMPUTING x GIVEN b HAS

$$\kappa = \|A^{-1}\| \frac{\|b\|}{\|x\|} \leq \|A\| \|A^{-1}\|.$$

THE CONDITION NUMBER OF A

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

CONDITION NUMBER OF A SYSTEM, $Ax = b$

- b IS FIXED
- $f: A \rightarrow x$

HOW DO PERTURBATIONS IN A, δA, CHANGE x?

$$(A + \delta A)(x + \delta x) = b$$

USE THE FACT THAT $Ax = b$ AND IGNORING $\delta A \delta x$, WE HAVE

$$\delta A x + A \delta x = 0$$

$$\delta x = -A^{-1} \delta A x$$

$$\|\delta x\| \leq \|A^{-1}\| \|\delta A\| \|x\|$$

CONDITION NUMBER

$$\kappa = \sup_{\delta A} \frac{\|\delta x\|}{\|x\|} \Big/ \frac{\|\delta A\|}{\|A\|}$$

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

EQUALITY IF

$$\|A^{-1} \delta A x\| = \|A^{-1}\| \|\delta A\| \|x\|$$

CAN ALWAYS BE FOUND.

$$\kappa = \kappa(A).$$

THEOREM: LET b BE FIXED AND CONSIDER $x = A^{-1}b$, WHERE A IS SQUARE AND NON-SINGULAR. THE CONDITION NUMBER OF THIS PROBLEM W.R.T. PERTURBATIONS IN A IS

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

FLOATING POINT NUMBERS AND ARITHMATIC

A COMPUTER MUST USE A DISCRETE REPRESENTATION OF $\mathbb{R}$.

(a) THERE MUST BE A LARGEST AND SMALLEST POSITIVE NUMBER.

ON DOUBLE PRECISION MACHINE

$$N_{MAX} \approx 1.79 \times 10^{308}$$

$$N_{MIN} \approx 2.23 \times 10^{-308}$$

THIS IS TYPICALLY NOT THE ISSUE.

(b) GAPS BETWEEN ADJACENT NUMBERS.

ON DOUBLE PRECISION MACHINE.

$[1, 2]$

$$1, 1 + 2^{-52}, 1 + 2 \times 2^{-52}, \ldots, 2$$

NEXT INTERVAL

$[2, 4]$

$$2, 2 + 2^{-51}, 2 + 2 \times 2^{-51}, \ldots, 4$$

IN GENERAL, THE INTERVAL $[2^j, 2^{j+1}]$ IS REPRESENTED AS 2^j TIMES THE NUMBERS REPRESENTING THE INTERVAL $[1, 2]$.

IN THE FLOATING POINT REPRESENTATION THE GAPS BETWEEN ~~SUCCESSIVE~~ ADJACENT NUMBERS SCALE WITH THEIR SIZE.

CALL SET OF FLOATING POINT NUMBERS $\mathbb{F} \subset \mathbb{R}$.

CALL ϵ ("MACHINE EPSILON") IS THE RESOLUTION OF THE FLOATING POINT NUMBERS. AND IS HALF THE DISTANCE ~~BET~~ BETWEEN 1 AND THE ADJACENT NUMBER.

FOR DOUBLE PRECISION

$$\epsilon = 2^{-53} \simeq 1.11 \times 10^{-16}$$

~~WE KNOW~~ AS A ~~RE~~SULT

$$\forall\ x \in \mathbb{R} \quad \exists\ x' \in \mathbb{F}$$

$$\text{S.T.} \quad |x - x'| \leq \epsilon |x|$$

LET $fl : \mathbb{R} \to \mathbb{F}$ BE THE FUNCTION THAT ROUNDS $x \in \mathbb{R}$ TO THE NEAREST FLOATING POINT.

FLOATING POINT AXIOM I (FPA I)

$$\forall\ x \in \mathbb{R} \quad \exists\ \epsilon' \text{ WITH } |\epsilon'| \leq \epsilon \quad \text{S.T.} \quad fl(x) = x(1 + \epsilon')$$

Floating Point Arithmatic

$+, -, \times, \div$ on $\mathbb{R}$

have analogues

$\oplus, \ominus, \otimes, \oslash$ on $\mathbb{F}$

constructed s.t.

$$x \circledast y = fl(x \circledast * y)$$

For $x, y \in \mathbb{F}$

where $*$ is $+, -, \times$, or $\div$.

Fundamental Axiom of Floating Point Arithmatic (FPA II)

$\forall\, x, y \in \mathbb{F}$ $\exists\, \epsilon'$ with $|\epsilon'| \leq \epsilon$ s.t.

$$x \circledast y = (x * y)(1 + \epsilon')$$

Stability

- Stability ~~pertains~~ pertains to the perturbation behaviour of the algorithm used to solve the problem on a computer.
- Algorithm: $\tilde{f}: X \to Y$ between same spaces as the problem.

Fix: (i) Problem f

(ii) Floating pt. computer

(iii) An algorithm for f

(iv) implementation of the algorithm.

$x \in X$ IS ROUNDED $x' = fl(x)$ THEN SUPPLIED TO THE PROGRAM.

THE PROGRAM IS RUN AND THE RESULT IS $\tilde{f}(x) \in Y$.

DESPITE THE COMPLEXITY, WE CAN MAKE CLEAN STATEMENTS ABOUT $\tilde{f}(x)$ USING FPA I & II.

ACCURACY

- ABSOLUTE ERROR

$$\| \tilde{f}(x) - f(x) \|$$

- RELATIVE ERROR

$$\frac{\| \tilde{f}(x) - f(x) \|}{\| f(x) \|}$$

AN ALGORITHM IS ACCURATE IF FOR EACH $x \in X$

$$\frac{\| \tilde{f}(x) - f(x) \|}{\| f(x) \|} = O(\epsilon)$$

ERROR IS ON THE ORDER OF MACHINE EPSILON.

MORE PRECISELY, $\exists$ CONSTANT C S.T. $\forall x \in X$

$$\frac{\| \tilde{f}(x) - f(x) \|}{\| f(x) \|} \leq C\epsilon$$

AS $\epsilon \to 0$.

STABILITY

AN ALGORITHM $\tilde{f}$ FOR PROBLEM f IS

STABLE IF FOR EACH $x \in X$

$$\frac{\| \tilde{f}(x) - f(\tilde{x}) \|}{\| f(\tilde{x}) \|} = O(\epsilon)$$

FOR SOME $\tilde{x} \in X$

$$\frac{\| \tilde{x} - x \|}{\| x \|} = O(\epsilon)$$

A STABLE ALGORITHM GIVES NEARLY THE RIGHT ANSWER TO NEARLY THE RIGHT QUESTION.

BACKWARD STABILITY

THE ALGORITHM IS BACKWARD STABLE IF FOR EACH $x \in X$

$$\tilde{f}(x) = f(\tilde{x}) \quad \text{FOR SOME}$$

$\tilde{x} \in X$ WITH

$$\frac{\| \tilde{x} - x \|}{\| x \|} = O(\epsilon)$$

A BACKWARD STABLE ALGORITHM GIVES EXACTLY THE RIGHT ANSWER TO NEARLY THE RIGHT QUESTION.

STABILITY

AN ALGORITHM $\tilde{f}$ IS STABLE IF FOR EACH $x \in X$

$$\frac{\| \tilde{f}(x) - \tilde{f}(\tilde{x}) \|}{\| f(\tilde{x}) \|} = O(\epsilon)$$

FOR SOME $\tilde{x} \in X$

$$\frac{\| \tilde{x} - x \|}{\| x \|} = O(\epsilon)$$

ACCURACY OF A BACKWARD STABLE ALGORITHM.

THM: SUPPOSE A BACKWARD STABLE ALGORITHM $\tilde{f}$ IS APPLIED TO SOLVE A

PROBLEM $f : X \to Y$ WITH CONDITION NUMBER κ ON A COMPUTER SATISFYING EPA I & II. THEN THE RELATIVE ERROR SATISFIES

$$\frac{\| \tilde{f}(x) - f(x) \|}{\| f(x) \|} = O(\kappa(x)\epsilon)$$

PROOF: SINCE $\tilde{f}$ IS BACKWARD STABLE, $\tilde{f}(x) = f(\tilde{x})$ FOR SOME $\tilde{x}$ WITH $\frac{\| \tilde{x} - x \|}{\| x \|} = O(\epsilon)$.

DEFINE

$$\kappa_\epsilon(x) = \sup_{\substack{\delta x \\ \frac{\| \delta x \|}{\| x \|} \le \epsilon}} \frac{\frac{\| \delta f \|}{\| f \|}}{\frac{\| \delta x \|}{\| x \|}}$$

THUS,

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|}$$

$$\leq \kappa_\epsilon(x) \frac{\|\tilde{x} - x\|}{\|x\|}$$

TAKING THE LIMIT AS $\epsilon \to 0$

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \# = O(\kappa(x)\epsilon)$$

BACKWARDS ERROR ANALYSIS

INVESTIGATE ACCURACY USING THE CONDITIONING AND STABILITY.

IF OUR ALGORITHM IS BACKWARDS STABLE THEN

THE ERROR IS REFLECTED BY THE CONDITION NUMBER.

FORWARD ERROR ANALYSIS

KEEP A RUNNING TALLY OF THE ~~ALG~~ ERROR COMMITTED AT EACH STEP OF THE ALGORITHM.

STABILITY ~~=~~ FLOATING POINT OPERATIONS

$\oplus$, $\ominus$, $\otimes$, $\oslash$ ARE ALL BACKWARDS STABLE.

FOR SUBTRACTION:

DATA: x_1, x_2

SOLUTION: $f(x_1, x_2) = x_1 - x_2$

ALGORITHM: $\tilde{f}(x_1, x_2) = fl(x_1) \ominus fl(x_2)$

BY FPA I:

$$fl(x_1) = x_1(1+\epsilon_1)$$

$$fl(x_2) = x_2(1+\epsilon_2)$$

WITH $|\epsilon_1|, |\epsilon_2| \leq \epsilon$

BY FPA II:

$$fl(x_1) \ominus fl(x_2) = \left(fl(x_1) - fl(x_2)\right)(1+\epsilon_3)$$

WITH $|\epsilon_3| \leq \epsilon$

COMBINING THESE RESULTS.

$$fl(x_1) \ominus fl(x_2)$$

$$= \left[x_1(1+\epsilon_1) - x_2(1+\epsilon_2)\right] \cdot (1+\epsilon_3)$$

$$= x_1(1+\epsilon_1)(1+\epsilon_3) - x_2(1+\epsilon_2)(1+\epsilon_3)$$

$$= \overbrace{x_1}^{\tilde{x}_1}(1+\epsilon_4) - \overbrace{x_2}^{\tilde{x}_2}(1+\epsilon_5)$$

WITH $|\epsilon_4|, |\epsilon_5| \leq 2\epsilon + O(\epsilon^2)$

THUS, $\tilde{f}(x_1, x_2) = \tilde{x}_1 - \tilde{x}_2$

$$= f(\tilde{x}_1, \tilde{x}_2)$$

$$\frac{|\tilde{x}_{1,2} - x_{1,2}|}{|x_{1,2}|} = O(\epsilon)$$

INNER PRODUCT

$$x, y \in \mathbb{R}^m, \qquad \alpha = x^T y$$

ALGORITHM:

$$\tilde{\alpha} = \left(\sum_{i=1}^{m} \right) fl(x_i) \otimes fl(y_i)$$

THIS IS BACKWARDS STABLE.

$$= \left(\sum_{i=1}^{m} \right) x_i(1+\epsilon_i) \otimes y_i(1+\delta_i)$$

WITH $|\epsilon_i|, |\delta_i| \leq \epsilon$

$$= \left(\sum_{i=1}^{m} \right) \left[x_i(1+\epsilon_i) \otimes y_i(1+\delta_i) \right](1+\gamma_i)$$

WITH $|\gamma_i| \leq \epsilon$

$$= \left(\sum_{i=1}^{m} \right) \left[x_i(1+\epsilon_i)\, y_i(1+\tilde{\epsilon}_i) \right]$$

WITH $|\tilde{\epsilon}_i| \leq 2\epsilon + O(\epsilon^2)$

TAKE THE FIRST TWO TERMS

$$x_1(1+\epsilon_1)\, y_1(1+\tilde{\epsilon}_1) \oplus x_2(1+\epsilon_2) y_2(1+\tilde{\epsilon}_2)$$

$$= \left[x_1(1+\epsilon_1) y_1(1+\tilde{\epsilon}_1) + x_2(1+\epsilon_2) y_2(1+\tilde{\epsilon}_2) \right](1+\beta_1)$$

WITH $|\beta_1| \leq \epsilon$

$$= x_1(1+\epsilon_1) y_1(1+\hat{\epsilon}_1) + x_2(1+\epsilon_2)^{y_2}(1+\hat{\epsilon}_2)$$

WITH $|\hat{\epsilon}_1|, |\hat{\epsilon}_2| \leq 3\epsilon + O(\epsilon^2)$

FOLLOWING THE SAME PROCESS WITH THE REMAINING TERMS

$$\tilde{\alpha} = \sum_{i=1}^{n} \underbrace{x_i(1+\epsilon_i)}_{\tilde{x}_i} \underbrace{y_i(1+\hat{\delta}_i)}_{\tilde{y}_i}$$

$$|\hat{\delta}_i| \leq (n+1)\epsilon + O(\epsilon^2)$$

WE HAVE THAT

$$\tilde{\alpha}(x,y) = \alpha(\tilde{x},\tilde{y})$$

FOR $\frac{\|\tilde{x}-x\|}{\|x\|} = O(\epsilon)$ AND

$$\frac{\|\tilde{y}-y\|}{\|y\|} = O(\epsilon).$$

STABLE BUT NOT B. STABLE

$$f(x) = x + 1$$

$$\tilde{f}(x) = fl(x) \oplus 1$$

CAN SHOW

$$\tilde{f}(x) = x(1+\epsilon_3) + 1 + \epsilon_2$$

UNSTABLE ALGORITHM

USING $\det(\lambda I - A) = 0$ TO FIND THE EIGENVALUES OF A.

WHAT DOES STABILITY MEAN IN THE CONTEXT OF LINEAR ALGEBRA COMPUTATIONS? ①

PROBLEM: QR FACTORIZATION OF A.

INPUT OR DATA: A

OUTPUT OR SOLUTION: Q, R

WITH $QR = A$ AND Q IS UNITARY AND R UPPER TRIANGULAR

ALGORITHM: HOUSEHOLDER ON A COMPUTER SATISFYING FPA I & II.

INPUT: A, OUTPUT: $\tilde{Q}, \tilde{R}$

IF BACKWARDS STABLE ② THEN $\tilde{Q}, \tilde{R}$ ARE SOLUTIONS TO THE PROBLEM WITH INPUT $A + \delta A$ FOR SOME $\frac{\|\delta A\|}{\|A\|} = O(\epsilon)$

SINCE THIS IS A SOLUTION WE KNOW

$$\tilde{Q}\tilde{R} = A + \delta A$$

SO IN OUR EXPERIMENT, WE WERE MEASURING $\frac{\|\delta A\|}{\|A\|}$.

TYPICALLY THIS ERROR IS CALLED BACKWARD ERROR OR THE RESIDUAL.

THM: LET THE QR FACTORIZATION $A = QR$ OF $A \in \mathbb{C}^{m \times n}$ BE COMPUTED BY HOUSEHOLDER ON A COMPUTER SATISFYING FPAI 1 & II. THEN WE HAVE

$$\tilde{Q}\tilde{R} = A + \delta A \qquad \frac{\|\delta A\|}{\|A\|} = O(\epsilon)$$

FOR SOME $\delta A \in \mathbb{C}^{m \times n}$

WHAT IS MEANT BY $\tilde{R}$ AND $\tilde{Q}$?

$\tilde{R}$ IS THE TRIANGULAR MATRIX THAT GETS COMPUTED.

$\tilde{Q}$ IS THE EXACTLY UNITARY MATRIX FORMED USING THE COMPUTED VECTORS $\tilde{v}_k$.

TYPICALLY WE WOULD LIKE TO FACTOR OUR MATRIX TO DO SOMETHING, E.G. SOLVE A LINEAR SYSTEM.

IS BACKWARD STABILITY ENOUGH OR DO WE NEED ACCURATE Q AND R?

SOLVING $Ax = b$ USING QR.

ALGORITHM

1. $QR = A$ — DONE USING HOUSEHOLDER.

2. $y = Q^* b$ — CONSTRUCT $Q^* b$ USING IMPLICIT MULTIPLICATION.

3. $x = R^{-1} y$ SOLVING THE TRIANGULAR SYSTEM USING BACKWARD SUBSTITUTION.

WE'VE ALREADY DISCUSSED THE BACKWARD STABILITY OF STEP 1.

AND THAT IT OUTPUTS

$\tilde{Q}$, $\tilde{R}$.

STEP 2 IS ALSO BACKWARD STABLE. USING $\tilde{Q}$ GIVEN BY STEP 1. THIS MEANS.

(i) $(\tilde{Q} + \delta Q)\tilde{y} = b \quad \|\delta Q\| = O(\epsilon)$

STEP 3 IS ALSO BACKWARD STABLE

(ii) $(\tilde{R} + \delta R)\tilde{x} = \tilde{y}$

WITH $\frac{\|\delta R\|}{\|R\|} = O(\epsilon)$.

PROVING THESE RESULTS IS VERY TEDIOUS!!! SEE LECTURE 17 IN TREFETHEN & BAU.

THEOREM: THE QR ALGORITHM TO SOLVE $Ax = b$ IS BACKWARD STABLE, SATISFYING

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

FOR SOME $\frac{\|\Delta A\|}{\|A\|} = O(\epsilon)$.

PROOF: COMBINING (i) & (ii)

$$b = (\tilde{Q} + \delta Q)(\tilde{R} + \delta R)\tilde{x}$$

$$= [\tilde{Q}\tilde{R} + (\delta Q)\tilde{R} + \tilde{Q}(\delta R) + (\delta Q)(\delta R)]\tilde{x}$$

FROM STEP 1. WE KNOW

$$\tilde{Q}\tilde{R} = A + \delta A$$

$$b = A\tilde{x} + [\underbrace{\delta A + (\delta Q)\tilde{R} + \tilde{Q}(\delta R) + (\delta Q)(\delta R)}_{\Delta A}]\tilde{x}$$

SHOW EACH TERM IS $O(\epsilon)$ TO GET RESULT.

SINCE $\tilde{Q}\tilde{R} = A + \delta A$

$$\Rightarrow \tilde{R} = \tilde{Q}^*(A + \delta A)$$

$$\frac{\|\tilde{R}\|}{\|A\|} \le \|\tilde{Q}^*\| \frac{\|A + \delta A\|}{\|A\|}$$

$$\le 1 + \frac{\|\delta A\|}{\|A\|} = O(1)$$

THUS

$$\frac{\|\delta Q\,\tilde{R}\|}{\|A\|} \le \|\delta Q\| \frac{\|\tilde{R}\|}{\|A\|} = O(\epsilon)$$

$$\frac{\|\tilde{Q}(\delta R)\|}{\|A\|} \le \frac{\|\delta R\|}{\|\tilde{R}\|} \|\tilde{Q}\| \frac{\|\tilde{R}\|}{\|A\|} = O(\epsilon)$$

$$\frac{\|\delta Q\ \delta R\|}{\|A\|} \leq \|\delta Q\| \frac{\|\delta R\|}{\|\tilde{R}\|} \frac{\|\tilde{R}\|}{\|A\|} = O(\epsilon^2)$$

Thus $$\frac{\|\Delta A\|}{\|A\|} = O(\epsilon) .$$

GAUSSIAN ELIMINATION

EXAMPLE

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}$$

$$L_1 A = \begin{bmatrix} 1 & & & \\ -2 & 1 & & \\ -4 & & 1 & \\ -3 & & & 1 \end{bmatrix} \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & 3 & 5 & 5 \\ & 4 & 6 & 8 \end{bmatrix}$$

$$L_2 L_1 A = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & -3 & 1 & \\ & -4 & & 1 \end{bmatrix} \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & 2 & 4 \end{bmatrix}$$

$$L_3 L_2 L_1 A = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & -1 & 1 \end{bmatrix} \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

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

WHAT ABOUT L? IF $A = LU$

$$L^{-1} = L_3 L_2 L_1$$

$$\Rightarrow L = L_1^{-1} L_2^{-1} L_3^{-1}$$

$$L_1^{-1} = \begin{bmatrix} 1 & & & \\ -2 & 1 & & \\ -3 & & 1 & \\ -4 & & & 1 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & & 1 & \\ 3 & & & 1 \end{bmatrix}$$

FOR L_k^{-1} JUST REPLACE ~~VALUE~~ VALUES BELOW l_{kk} BY -1 TIMES THE VALUE.

$$L = L_1^{-1} L_2^{-1} L_3^{-1}$$

$$= \begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & 3 & 1 & \\ 3 & 4 & 1 & 1 \end{bmatrix}$$

TO GET L_1, SIMPLY NEED TO COMPILE ALL THE SUBDIAGONAL ENTRIES OF L_k^{-1} IN ONE NON-ZERO MATRIX!

RECALL:

MGS WE SAID WAS TRIANGULAR ORTHOGONALIZATION

HOUSEHOLDER WE SAID WAS ORTHOGONAL TRIANGULARIZATION.

GAUSSIAN ELIMINATION IS TRIANGULAR TRIANGULARIZATION.

GENERAL FORMULA

LET x_k BE THE kTH COLUMN OF THE MATRIX AT THE kTH STEP.

$$x_k = \begin{bmatrix} x_{1k} \\ \vdots \\ x_{kk} \\ x_{k+1,k} \\ \vdots \\ x_{mk} \end{bmatrix}$$

WOULD LIKE L_k SUCH THAT

$$L_k x_k = \begin{bmatrix} x_{1k} \\ \vdots \\ x_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

TO DO THIS, WE MUST HAVE

$$l_{jk} = \frac{x_{jk}}{x_{kk}} \quad \text{for } j = k+1, \ldots, m$$

SO THAT

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

THEN CAN WRITE

$$L_k = I - l_k e_k^*$$

WHERE

$$l_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ l_{k+1,k} \\ \vdots \\ l_{mk} \end{bmatrix}$$

AND e_k IS THE VECTOR WITH THE kTH ENTRY 1 AND 0 ELSEWHERE.

SINCE, $e_k^* l_k = 0$

$$(I - l_k e_k^*)(I + l_k e_k^*) = I$$

THUS,

$$L_k^{-1} = I + l_k e_k^*$$

$$L_k^{-1} L_{k+1}^{-1} = (I + l_k e_k^*)(I + l_{k+1} e_{k+1}^*)$$

$$= I + l_k e_k^* + l_{k+1} e_{k+1}^*$$

SINCE $e_k^* l_{k+1} = 0$.

THUS,

$$L = L_1^{-1} L_2^{-1} \cdots L_{m-1}^{-1}$$

$$= \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & l_{32} & \ddots & \\ \vdots & \vdots & \ddots & 1 \\ l_{m1} & l_{m2} & l_{m,m-1} & \end{bmatrix}$$

GAUSSIAN ELIMINATION ALGORITHM

$A \in \mathbb{C}^{m \times m}$ (SQUARE)

$U = A$

$L = I$

for $k = 1$ to $m-1$

  for $j = k+1$ to m

    $l_{jk} = u_{jk} / u_{kk}$

    (*) $u_{j,k:m} = u_{j,k:m} - l_{jk}\, u_{k,k:m}$

  END FOR

END FOR

OPERATION COUNT

WORK DOMINATED BY LINE (*)

CONSISTS OF:

(i) $m-k+1$: $\times$

(ii) $m-k+1$: $-$

$2(m-k+1)$ FLOPS EVERY TIME (*) IS REACHED.

$$\#\text{ OF FLOPS} \sim \sum_{k=1}^{m-1} \sum_{j=k+1}^{m} 2(m-k+1)$$

$$= \sum_{k=1}^{m-1} 2(m-k+1) \underbrace{\sum_{j=k+1}^{m} 1}_{m-k}$$

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

$$\sim \sum_{k=1}^{m-1} 2m^2 - 4mk + 2k^2$$

$$\sim 2m^3 - \frac{4m^3}{2} + \frac{2}{3}m^3$$

$$\#\text{ OF FLOPS} \sim \frac{2}{3}m^3$$

GAUSSIAN ELIMINATION WITH PIVOTING

SUPPOSE $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$

STANDARD GAUSSIAN ELIMINATION WILL FAIL AT THE FIRST STEP.

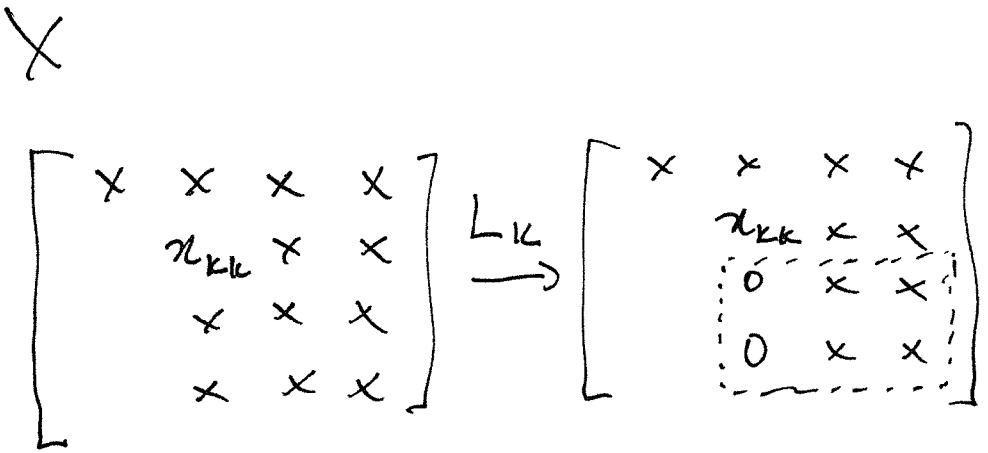
ALSO, PROBLEMATIC IF

$$A = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix}$$

SINCE DIVIDING BY 10^{-20} COULD LEAD TO NUMERICAL INSTABILITIES.

PIVOTS

AT STEP k HAVE MATRIX

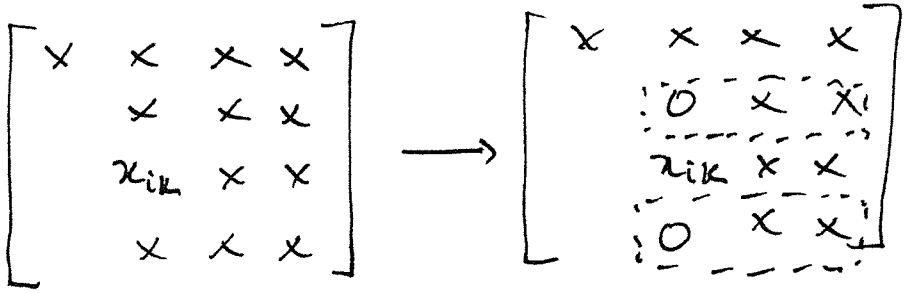


ENTRY x_{kk} IS CALLED THE PIVOT.

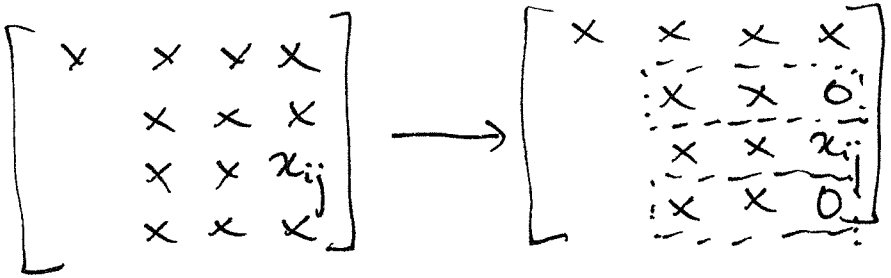
EVERY ENTRY OF $X_{k+1:m,\,k:m}$ IS SUBTRACTED BY THE PRODUCT OF AN ENTRY IN ROW k AND AN ENTRY IN COLUMN k DIVIDED BY x_{kk}.

IF THE GOAL IS TO INTRODUCE ZEROS IN A COLUMN, THE CHOICE OF x_{kk} IS NOT UNIQUE.

TAKE x_{ik} AS PIVOT:



CAN EVEN TAKE x_{ij}



- FREE TO CHOOSE ANY ENTRY IN $X_{k:m,k:m}$ AS THE PIVOT

- FOR STABILITY, CHOOSE THE LARGEST ENTRY IN $X_{k:m,k:m}$.

- CHOOSING ENTRY OTHER THAN x_{kk} WILL DESTROY THE INDUCED TRIANGULAR STRUCTURE

- ADD ADDITIONAL STEP OF PERMUTING THE ROWS AND COLUMNS TO MOVE CHOSEN PIVOT INTO x_{kk}'S POSITION.

COMPLETE PIVOTING

- ALLOW ALL ENTRIES OF $X_{k:m,k:m}$ TO BE A PIVOT.

- # OF FLOPS = $O(m^3)$

PARTIAL PIVOTING

- CHOOSE PIVOT FROM THE kTH COLUMN

- ONLY ROW EXCHANGES ARE INVOLVED.

- # OF FLOPS = $O(m^2)$

IDEA:

$$\begin{bmatrix} x & x & x & x \\ & x & x & x \\ & x_{ik} & x & x \\ & x & x & x \end{bmatrix} \xrightarrow{P_{ik}} \begin{bmatrix} x & x & x & x \\ & x_{ik} & x & x \\ & x & x & x \\ & x & x & x \end{bmatrix}$$

SELECT THE PIVOT. EXCHANGE ROWS.

$$\xrightarrow{L_k} \begin{bmatrix} x & x & x & x \\ & x_{ik} & x & x \\ & 0 & x & x \\ & 0 & x & x \end{bmatrix}$$

ELIMINATION STEP.

EXAMPLE

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}$$

STEP 1: $P_1 A$

$$\begin{bmatrix} & & 1 & \\ & 1 & & \\ 1 & & & \\ & & & 1 \end{bmatrix} A = \begin{bmatrix} 8 & 7 & 9 & 5 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{bmatrix}$$

STEP 2: $L_1 P_1 A$

$$\begin{bmatrix} 1 & & & \\ -1/2 & 1 & & \\ -1/4 & & 1 & \\ -3/4 & & & 1 \end{bmatrix} P_1 A = \begin{bmatrix} 8 & 7 & 9 & 5 \\ & -1/2 & -3/2 & -3/2 \\ & -3/4 & -5/4 & -5/4 \\ & 7/4 & 9/4 & 17/4 \end{bmatrix}$$

STEP 3: $P_2 L_1 P_1 A$

$$\begin{bmatrix} 1 & & & \\ & 1 & & 1 \\ & & 1 & \\ & & & \end{bmatrix} L_1 P_1 A =$$

EVENTUALLY OBTAIN DECOMPOSITION THAT MAY BE WRITTEN AS

$$\begin{bmatrix} & & 1 & \\ & & & 1 \\ & 1 & & \\ 1 & & & \end{bmatrix} A = \begin{bmatrix} 1 & & & \\ 3/4 & 1 & & \\ 1/2 & -2/7 & 1 & \\ 1/4 & -3/7 & 1/3 & 1 \end{bmatrix}$$

P L

$$\times \begin{bmatrix} 8 & 7 & 9 & 5 \\ & 7/4 & 9/4 & 17/4 \\ & & -6/7 & -2/7 \\ & & & 2/3 \end{bmatrix}$$

U

$PA = LU$ IS THE TRUE LU DECOMPOSITION.

INTERPRETATION!

1. PERMUTE ROWS OF A ACCORDING P.
2. APPLY GAUSSIAN ELIMINATION WITHOUT PIVOTING TO PA.

WHY IS THIS THE CASE?

WE DID:

$$L_3 P_3 L_2 P_2 L_1 P_1 A = U$$

THESE OPERATIONS CAN BE REORDERED

$$L_3 P_3 L_2 P_2 L_1 P_1 = (L_3' L_2' L_1')(P_3 P_2 P_1)$$
$$L^{-1} \qquad P$$

WHERE

$$L_3' = L_3 \qquad L_2' = P_3 L_2 P_3^{-1}$$

$$L_1' = P_3 P_2 L_1 P_2^{-1} P_3^{-1}$$

PERMUTATION MATRICES PRESERVE LOWER TRIANGULAR FORM.

LU DECOMPOSITION

FOR $A \in \mathbb{C}^{m \times m}$ WE HAVE WITH PARTIAL PIVOTING.

$$L_{m-1} P_{m-1} \cdots L_2 P_2 L_1 P_1 A = U$$

THIS CAN ALSO BE WRITTEN AS

$$\underbrace{(L'_{m-1} \cdots L'_2 L'_1)}_{L^{-1}} \overbrace{(P_{m-1} \cdots P_2 P_1)}^{P} A = U$$

$$L'_k = P_{m-1} \cdots P_{k+1} L_k P_{k+1}^{-1} \cdots P_{m-1}^{-1}$$

THUS

$$PA = LU$$

ALGORITHM FOR LU

$U = A$

$L = I$

$P = I$

for $k = 1$ to $m-1$

- SELECT $i \geq k$ to MAXIMIZE $|u_{ik}|$ } SELECT PIVOT
- $u_{k,k:m} \longleftrightarrow u_{i,k:m}$
- $l_{k,1:k-1} \longleftrightarrow l_{i,1:k-1}$
- $P_{k,:} \longleftrightarrow P_{i,:}$ } EXCHANGE OF ROWS.
- for $j = k+1$ to m
  - $l_{jk} = u_{jk} / u_{kk}$

$$u_{j,k:m} = u_{j,k:m} - l_{jk}\, u_{k,k:m}$$

END FOR

END FOR.

OPERATION COUNT

OF FLOPS $\sim \frac{2}{3} m^3$

STABILITY OF LU

CONSIDER AGAIN

$$A = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix}$$

STANDARD GAUSSIAN ~~ELIMINA~~ ELIMINATION LEADS TO LARGE ENTRIES OF L AND U.

THIS AMPLIFICATION IS AT THE HEART OF THE INSTABILITY

THM: $A = LU$ FOR NONSINGULAR $A \in \mathbb{C}^{m \times m}$ IS COMPUTED BY STANDARD GAUSSIAN ELIMINATION ON A COMPUTER SATISFYING FPA I AND FPA II. IF NO ZERO-PIVOTS ARE ENCOUNTERED THEN COMPUTED $\tilde{L}$ AND $\tilde{U}$ SATISFY

$$\tilde{L}\tilde{U} = A + \delta A \qquad \frac{\|\delta A\|}{\|L\|\|U\|} = O(\epsilon)$$

FOR SOME $\delta A \in \mathbb{C}^{m \times m}$

BACKWARD STABILITY IF

$$\|L\|\|U\| = O(\|A\|)$$

BUT IF $\|L\|\|U\| \neq O(\|A\|)$

CAN EXPECT BACKWARD INSTABILITY.

PIVOTING

THM APPLIES REPLACING

A BY PA.

WITH PARTIAL PIVOTING

WE KNOW

$$\|L\| = O(1)$$

THEN FOR BACKWARD

STABILITY, WE MUST HAVE

$$\|U\| = O(\|A\|).$$

THIS ξ CAN BE MEASURED

BY A GROWTH FACTOR.

$$\rho = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|}$$

SO. THEN ~~WHERE~~ WE HAVE

$$\|U\| = O(\rho\|A\|)$$

THM: $PA = LU$ BE COMPUTED

USING GAUSSIAN ELIMINATION

WITH PIVOTING ON A COMPUTER

SATISFYING FPA I AND FPA II.

THEN $\tilde{L}$, $\tilde{U}$, AND $\tilde{P}$ SATISFY

$$\tilde{L}\tilde{U} = \tilde{P}A + \delta A$$

$$\frac{\|\delta A\|}{\|A\|} = O(\rho\epsilon)$$

FOR SOME $\delta A \in \mathbb{C}^{m\times m}$ WHERE ρ

IS THE GROWTH FACTOR.

EXAMPLE:

$$A = \begin{bmatrix} 1 & & & & 1 \\ -1 & 1 & & & 1 \\ -1 & -1 & 1 & & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

AFTER EACH STEP k, THE $k+1$ THROUGH m ENTRIES IN THE LAST COLUMN WILL BE 2^k.

$$U = \begin{bmatrix} 1 & & & & 1 \\ & 1 & & & 2 \\ & & 1 & & 4 \\ & & & 1 & 8 \\ & & & & 16 \end{bmatrix}$$

ENTRY u_{mm} IS 2^{m-1}

THUS, $\rho = 2^{m-1}$

THIS IS HUGE FOR LARGE MATRICES!

EIGEN VALUES AND EIGENVECTORS

$A \in \mathbb{C}^{m \times m}$ (SQUARE)

$x \in \mathbb{C}^m$ EIGENVECTOR AND

$\lambda \in \mathbb{C}$ CORRESPONDING EIGENVALUE

IF

$$Ax = \lambda x$$

EIGENSPACE: SET OF THE EIGENVECTORS

SPECTRUM: SET OF ALL EIGENVALUES.

WHY COMPUTE THIS

1. E-VALS AND E-VECS ENCODE INFORMATION ABOUT A.

2. ARISE IN SAY, STABILITY CALCULATIONS.

3. USE E-VECS AS A BASIS TO SOLVE A LINEAR SYSTEM.

EIGENVALUE DECOMPOSITION

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

X: NON-SINGULAR $m \times m$ WITH x_k (EIGENVECTOR k) IN THE KTH COLUMN.

Λ: DIAGONAL MATRIX WITH

$\Lambda_{kk} = \lambda_k$, THE kTH EIGENVALUE.

THIS DECOMPISITION IS NOT ALWAYS POSSIBLE!

GEOMETRIC MULTIPLICITY OF λ

NUMBER OF E-VECS CORRESPONDING TO EIGENVALUE λ.

ALGEBRAIC MULTIPLICITY

CHARACTERISTIC POLYNOMIAL OF A

$$P_A(z) = \det(zI - A)$$

THE ROOTS OF $P_A(z)$ ARE THE EIGENVALUES.

$$P_A(z) = (z - \lambda_1)(z - \lambda_2)\cdots(z - \lambda_m)$$

THE ALGEBRAIC MULTIPLICITY IS THE NUMBER OF TIMES λ IS A ROOT OF $P_A(z)$.

A E-VAL IS SIMPLE IF IT HAS ALGEBRAIC MULTIPLICITY 1.

THM: THE ALGEBRAIC MULTIPLICITY OF λ IS AT LEAST AS GREAT AS ITS GEOMETRIC MULTIPLICITY.

DEFECTIVE MATRICES AND EIGENVALUES

- AN EIGENVALUE IS DEFECTIVE IF ITS ALGEBRAIC MULTIPLICITY EXCEEDS ITS GEOMETRIC MULTIPLICITY.
- A MATRIX IS DEFECTIVE IF IT HAS AT LEAST ONE DEFECTIVE E-VAL.

THM: AN $m \times m$ MATRIX A IS NON-DEFECTIVE IFF IT HAS AN EIGENVALUE DECOMPOSITION

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

SIMILARITY TRANSFORMATION

FOR $X \in \mathbb{C}^{m \times m}$ NONSINGULAR THEN THE MAP $A \mapsto X^{-1}AX$ IS A SIMILARITY TRANSFORMATION OF A.

MATRICES A & B ARE SIMILAR IF $\exists\ X$ S.T. $B = X^{-1}AX$.

THM: IF X IS NON-SINGULAR THEN A AND $X^{-1}AX$ HAVE SAME CHARACTERISTIC POLYNOMIAL, EIGENVALUES, AND MULTIPLICITIES.

SCHUR FACTORIZATION

$$A = Q\,T\,Q^*$$

Q: UNITARY MATRIX

T: UPPER-TRIANGULAR.

THM: EVERY SQUARE MATRIX HAS A SCHUR FACTORIZATION.

SINCE THIS IS A SIMILARITY TRANSFORMATION, A & T HAVE THE SAME E-VALS.

SINCE T IS TRIANGULAR THE E-VALS ARE THE DIAGONAL ENTRIES!

UNITARY DIAGONALIZATION

MATRIX A IS UNITARY DIAGONALIZABLE IF $\exists$ UNITARY Q S.T.

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

THM: A HERMITIAN MATRIX ($A = A^*$) IS UNITARY DIAGONALIZABLE AND ITS E-VALS ARE REAL.

EIGENVALUE ALGORITHMS

ALREADY DISCUSSED THAT A ROOT FINDING ALGORITHM FOR $P_A(z)$ WOULD BE UNSTABLE.

THIS CONNECTION WITH ROOT FINDING, HOWEVER, TELLS US THAT WE SHOULD NOT BE ABLE TO FIND THE EIGENVALUES IN A FINITE NUMBER OF STEPS.

ANY EIGENVALUE SOLVER MUST BE ITERATIVE!

COMPUTE SCHUR FACTORIZATION ITERATIVELY.

GENERATE A SEQUENCE OF UNITARY Q_j SUCH THAT

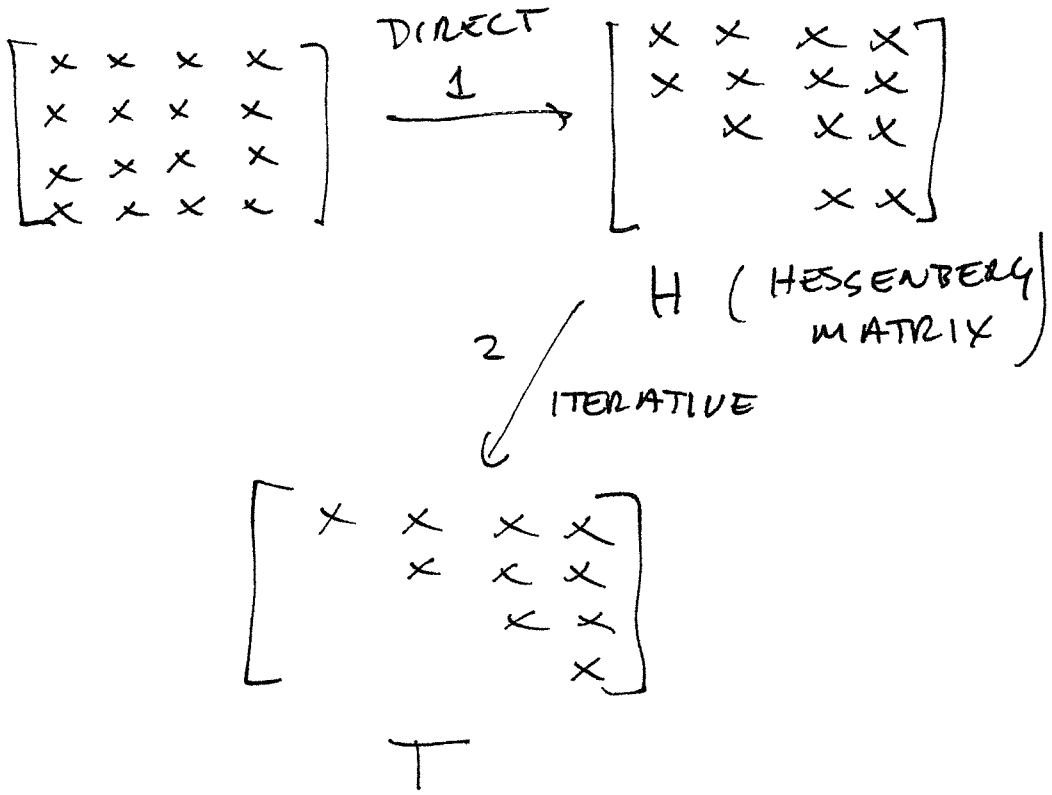
$$Q_j^* \cdots Q_2^* Q_1^* A Q_1 Q_2 \cdots Q_j$$

CONVERGES TO T AS

$j \to \infty$.

IF A IS HERMITIAN, T WILL BE DIAGONAL.

TWO PHASES OF E-VAL COMPUTATION



HESSENBERG MATRIX

- PHASE 1 OF THE EIGENVALUE COMPUTATION.

$$H = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ & x & x & x \\ & & x & x \end{bmatrix}$$

- FOR HERMITIAN MATRICES THE ASSOCIATED HESSENBERG MATRIX WILL BE TRIDIAGONAL

$$\Delta = \begin{bmatrix} x & x & & \\ x & x & x & \\ & x & x & x \\ & & x & x \end{bmatrix}$$

TO FORM H FROM A WE'LL NEED TO USE A SIMILARITY TRANSFORMATION TO PRESERVE THE E-VALS.

WE CAN USE HOUSEHOLDER REFLECTORS TO CONSTRUCT H FROM A.

$$\begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix} \xrightarrow{Q_1^*} \left[\begin{array}{cccc} x & x & x & x \\ \hdashline x & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \end{array}\right]$$

A $\qquad$ $Q_1^* A$

$$\left[\begin{array}{c|ccc} x & x & x & x \\ \hdashline x & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \end{array}\right]$$

Q_1

$Q_1^* A Q_1$

FOR LARGE m

OF FLOPS $\sim 4m(m-k)$

FOR (ii)

~~THUS FOR m TIMES~~

$$\#\text{ OF FLOPS} \sim 4\sum_{k=1}^{m-2} m(m-k)$$

$$\sim 2m^3$$

TOTAL COST $\sim \frac{10}{3}m^3$

IF A IS A HERMITIAN MATRIX, THE WORK IS REDUCED TO $\sim \frac{4}{3}m^3$ DUE TO

(a) SYMMETRY

(b) SPARSITY.

THIS ALGORITHM IS BACKWARD STABLE.

RAYLEIGH QUOTIENT AND INVERSE ITERATION

- CLASSICAL E-VAL ALGORITHMS.
- FROM THIS POINT ON, RESTRICT TO REAL, SYMMETRICAL MATRICES.

AND TAKE $\|\cdot\| = \|\cdot\|_2$

- EIGENVALUES ARE ALL REAL, $\lambda_1, \ldots, \lambda_m$
- ORTHONORMAL E-VECS $q_1, \ldots, q_m$.

RAYLEIGH QUOTIENT

$$r(x) = \frac{x^T A x}{x^T x}$$

NOTICE IF x IS AN E-VEC THEN $r(x) = \lambda$ WHERE λ IS THE E-VAL.

VIEW $x \in \mathbb{R}^m$ AS VARIABLE AND $r : \mathbb{R}^m \longrightarrow \mathbb{R}$.

TAKE PARTIAL DERIVATIVE OF r W.R.T. x_j

$$\frac{\partial r}{\partial x_j} = \frac{1}{x^T x} \frac{\partial}{\partial x_j} (x^T A x)$$

$$- \frac{(x^T A x)}{(x^T x)^2} \frac{\partial}{\partial x_j} (x^T x)$$

$$= \frac{2}{x^T x} (Ax - r(x)x)_j$$

COLLECT THE PARTIAL DERIVATIVES FOR $j = 1, \ldots, m$ INTO ONE m VECTOR

$$\nabla r(x) = \frac{2}{x^T x} (Ax - r(x)x)$$

THUS, $\nabla r = 0$ IF x IS AN E-VEC.

E-VECTORS ARE STATIONARY POINTS OF $r(x)$.

WHEN WE RESTRICT TO $\|x\| = 1$, $r(x)$ IS A CONTINUOUS FUNCTION ON THE UNIT SPHERE.

IF q_J IS AN E-VEC, SINCE $\nabla r(q_J) = 0$ AND r IS SMOOTH, TAYLOR EXPANSION REVEALS

$$r(x) - r(q_J) = O\left(\|x - q_J\|^2\right)$$

AS $x \to q_J$

RAYLEIGH QUOTIENT GIVES QUADRATICALLY ACCURATE ESTIMATE OF THE E-VAL.

POWER ITERATION

METHOD TO FIND E-VEC ~~ASSOCIATE~~ ASSOCIATED WITH THE LARGEST E-VAL.

START WITH $v^{(0)}$, $\|v^{(0)}\| = 1$

AND APPLY A AND NORMALIZE AT EACH ITERATION.

$$v^{(1)} = \frac{A v^{(0)}}{\|A v^{(0)}\|}$$

$$v^{(2)} = \frac{A v^{(1)}}{\|A v^{(1)}\|}$$

REPEATING THE ~~RE~~ PROCESS $m-2$ TIMES WE OBTAIN THE HESSENBERG FORM

$$Q_{m-2}^* \cdots Q_2^* Q_1^* A Q_1 Q_2 \cdots Q_{m-2}$$

ALGORITHM: HOUSEHOLDER REDUCTION TO HESSENBERG FORM.

For $k = 1$ to $m-2$

$$x = A_{k+1:m,\,k}$$

$$v_k = \text{sign}(x_1)\|x\|_2 e_1 + x$$

$$v_k = v_k / \|v_k\|_2$$

(i) $A_{k+1:m,\,k:m} = A_{k+1:m,\,k:m} - 2v_k(v_k^* A_{k+1:m,\,k:m})$

(ii) $A_{1:m,\,k+1:m} = A_{1:m,\,k+1:m} - 2\left(A_{1:m,\,k+1:m} v_k\right) v_k^*$

END FOR

OPERATION ~~COST~~ COUNT

WORK DOMINATED BY (i) AND (ii)

- OPERATION COUNT FOR (i) IS EXACTLY THAT FOR HOUSEHOLDER TRIANGULARIZATION.

$\#$ OF FLOPS $\sim \frac{4}{3} m^3$

- WORK FOR (ii)

AND KEEP GOING!

WRITE $v^{(0)}$ AS A LIN. COMB. OF THE E-VECS OF A.

$$v^{(0)} = a_1 q_1 + a_2 q_2 + \cdots + a_m q_m$$

SINCE $v^{(k)}$ IS A MULTIPLE OF $A^k v^{(0)}$, WE HAVE FOR SOME c_k

$$v^{(k)} = c_k A^k v^{(0)}$$

$$= c_k \left(a_1 \lambda_1^k q_1 + a_2 \lambda_2^k q_2 + \cdots + a_m \lambda_m^k q_m \right)$$

SUPPOSE OUR E-VALS ARE INDEXED SUCH THAT

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_m|$$

$$v^{(k)} = c_k \lambda_1^k \left(a_1 q_1 + a_2 (\lambda_2/\lambda_1)^k q_2 + \cdots + a_m (\lambda_m/\lambda_1)^k q_m \right)$$

THUS,

$$\| v^{(k)} - (\pm q_1) \| = O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right)$$

AND USING THE RAYLEIGH QUOTIENT $\lambda^{(k)} = r(v^{(k)})$ WE HAVE

$$|\lambda^{(k)} - \lambda_1| = O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^{2k} \right)$$

POWER ITERATION ALGORITHM

$v^{(0)}$ = SOME VECTOR WITH $\|v^{(0)}\| = 1$.

FOR $k = 1, 2, \ldots$

$$w = A v^{(k-1)}$$

$$v^{(k)} = w / \|w\|$$

$$\lambda^{(k)} = (v^{(k)})^T A v^{(k)}$$

END FOR ONCE "CONVERGED"

- POWER ITERATION IS RESTRICTED TO FINDING THE ~~LARGEST~~ ~~E-VA~~ EIGENVECTOR ASSOCIATED WITH THE LARGEST EIGENVALUE.

- POWER ITERATION HAS A LINEAR CONVERGENCE WITH RATE BASED ON $|\lambda_2/\lambda_1|$

INVERSE ITERATION

FOR $\mu \in \mathbb{R}$ ~~AS~~ NOT AN E-VAL OF A,

- THE E-VECS OF THE MATRIX $(A - \mu I)^{-1}$ ARE THE SAME AS A.

- THE E-VALS OF $(A - \mu I)^{-1}$ ARE $\{(\lambda_j - \mu)^{-1}\}$ WHERE λ_j ARE THE EVALS OF A.

CAN CHOOSE μ CLOSE TO λ_J, SUCH THAT $(\lambda_J - \mu)^{-1}$ IS MUCH LARGER THAN $(\lambda_j - \mu)^{-1}$ $\forall\, j \neq J$.

INVERSE ITERATION ALGORITHM

$v^{(0)}$ = SOME VECTOR $\|v^{(0)}\| = 1$

for $k = 1, 2, \ldots$

SOLVE $(A - \mu I) w = v^{(k-1)}$

$v^{(k)} = w / \|w\|$

$\lambda^{(k)} = (v^{(k)})^T A (v^{(k)})$

THOUGH CONVERGENCE IS STILL LINEAR, ITS RATE CAN BE CONTROLLED BY CHOOSING μ TO BE CLOSE TO THE E-VAL OF INTEREST.

BASED ON THE CONVERGENCE OF THE POWER ITERATION,

$$\|v^{(k)} - (\pm q_J)\| = O\left(\left|\frac{\mu - \lambda_J}{\mu - \lambda_K}\right|^k\right) \quad ④$$

$$|\lambda^{(k)} - \lambda_J| = O\left(\left|\frac{\mu - \lambda_J}{\mu - \lambda_K}\right|^{2k}\right)$$

RAYLEIGH QUOTIENT ITERATION

COMBINE OUTPUT OF THE RAYLEIGH QUOTIENT WITH INVERSE ITERATION, I.E. SET $\mu = \lambda^{(k)}$ AT EACH ITERATION.

ALGORITHM

$v^{(0)}$ = SOME VECTOR $\|v^{(0)}\| = 1$

$\lambda^{(0)} = (v^{(0)})^T A v^{(0)}$

for $k = 1, 2 \ldots$

SOLVE $(A - \lambda^{(k-1)} I) w = v^{(k-1)}$

$$v^{(k)} = w / \|w\|$$

$$\lambda^{(k)} = (v^{(k)})^T A v^{(k)}$$

THIS SIMPLE MODIFICATION GREATLY ACCELERATES CONVERGENCE.

SUPPOSE WE HAVE $v^{(k)}$ AND $\lambda^{(k)}$ SUFFICIENTLY CLOSE TO q_J AND λ_J.

$$\|v^{(k+1)} - q_J\| = O(|\lambda^{(k)} - \lambda_J| \, \|v^{(k)} - q_J\|)$$

IF $\|v^{(k)} - q_J\| \leq \delta$ THEN

FROM RAYLEIGH QUOTIENT THAT

$$|\lambda^{(k)} - \lambda_J| \leq O(\delta^2)$$

$$\|v^{(k+1)} - q_J\| \leq O(\delta^3)$$

THUS WE SEE THE FOLLOWING CONVERGENCE PATTERN EMERGE.

$\|v^{(k)} - q_J\|$ $\qquad$ $|\lambda^{(k)} - \lambda_J|$

$\delta \xrightarrow{\text{R.Q.}} O(\delta^2)$

$\downarrow$ I.I. $\swarrow$

$O(\delta^3) \xrightarrow{\text{R.Q.}} O(\delta^6)$

$\downarrow \swarrow$

$O(\delta^9) \longrightarrow O(\delta^{18})$

$\vdots \quad \swarrow$

OPERATION COUNTS

- POWER ITERATION
  - MATRIX - VEC. MULTIPLICATION AT EACH ITERATION
    $\Rightarrow O(m^2)$
- INVERSE ITERATION
  - NEED TO INVERT SAME LINEAR SYSTEM AT EACH STEP.
  - 1 $O(m^3)$ COMPUTATION
  - $N_{ITER} \times O(m^2)$
- RAYLEIGH QUOTIENT

  INVOLVES A NEW LINEAR SYSTEM AT EVERY ITERATION.

$\Rightarrow O(m^3)$ AT EACH ITERATION.

NOW WE SEE THE REASON FOR "PRETREATING" A.

IF A IS TRIDIAGONAL EACH METHOD IS $O(m)$.

THESE METHODS GIVE US ONE E-VAL AND ONE E-VEC. HOW DO WE OBTAIN THE SPECTRUM AND EIGEN SPACE OF A?

QR ALGORITHM

"PURE" QR ALGORITHM

$$A^{(0)} = A$$

FOR $k = 1, 2, \ldots$

$$Q^{(k)} R^{(k)} = A^{(k-1)}$$

$$A^{(k)} = R^{(k)} Q^{(k)}.$$

AS $k \to \infty$ $A^{(k)} \to T$ (SCHUR FORM)

OR IF A IS HERMITIAN

$$A^{(k)} \to \Lambda.$$

$A^{(k)}$ IS SIMILAR TO $A^{(k-1)}$ SINCE

$$R^{(k)} = \left(Q^{(k)}\right)^T A^{(k-1)}$$

THUS,

$$A^{(k)} = \left(Q^{(k)}\right)^T A^{(k-1)} Q^{(k)}.$$

THIS STILL SEEM IMPOSSIBLY SIMPLE!

HOW / WHY DOES IT WORK?

UNNORMALIZED SIMULTANEOUS ITERATION

- APPLY POWER ITERATION TO SEVERAL VECTORS AT ONCE.

UNNORMALIZED SIMULTANEOUS ITERATION

- APPLY POWER ITERATION TO SEVERAL VECTORS SIMULTANEOUSLY.

- TAKE

$$V^{(0)} = \left[V_1^{(0)} \,\middle|\, V_2^{(0)} \,\middle|\, V_3^{(0)} \,\middle|\, \cdots \,\middle|\, V_n^{(0)} \right] \quad m \times n$$

- SET $V^{(k)} = A^k V^{(0)}$

- CAN PROVE THAT THE SPACE

$$\langle A^k V_1^{(0)}, \ldots, A^k V_n^{(0)} \rangle$$

CONVERGES TO THE SPACE SPANNED BY THE DOMINANT E-VECS OF A

$$\langle q_1, \ldots, q_n \rangle$$

THUS IF

$$\hat{Q}^{(k)} \hat{R}^{(k)} = V^{(k)}$$

THEN WE HAVE THAT THE COLUMNS OF $\hat{Q}^{(k)}$ CONVERGE TO $q_1 \cdots q_n$.

SIMULTANEOUS ITERATION

SINCE ALL $V_j^{(k)}$ CONVERGE TO q_1, $V^{(k)}$ IS ILLCONDITIONED

REMEDY : COMPUTE

$\hat{Q}^{(k)} \hat{R}^{(k)} = V^{(k)}$ AT EACH ITERATION AND USE $\hat{Q}^{(k)}$ INSTEAD OF $V^{(k)}$ AT THE NEXT ITERATION.

SIMULTANEOUS ITERATION ALGORITHM

PICK $\hat{Q}^{(0)} \in \mathbb{R}^{m \times n}$

for $k = 1, 2, \ldots$

$$Z = A \hat{Q}^{(k-1)}$$

$$\hat{Q}^{(k)} \hat{R}^{(k)} = Z$$

IT TURNS OUT SIMULTANEOUS ITERATION AND THE QR ALGORITHM ARE EQUIVALENT.

TAKE $n = m$, REDUCED → FULL QR

SIMULTANEOUS ITERATION

$$\underline{Q}^{(0)} = I$$

$$Z = A \underline{Q}^{(k-1)}$$

$$\underline{Q}^{(k)} R^{(k)} = Z .$$

$$A^{(k)} = (\underline{Q}^{(k)})^T A \underline{Q}^{(k)}$$

"PURE" QR ALGORITHM

$$A^{(0)} = A$$

$$A^{(k-1)} = Q^{(k)} R^{(k)}$$

$$A^{(k)} = R^{(k)} Q^{(k)}$$

$$\underline{Q}^{(k)} = Q^{(1)} Q^{(2)} \cdots Q^{(k)}$$

FOR BOTH,

$$\underline{R}^{(k)} = R^{(k)} R^{(k-1)} \cdots R^{(1)}$$

THM: SIMULTANEOUS ITERATION WITH $\underline{Q}^{(0)} = I$ AND QR ALGORITHM GENERATE IDENTICAL SEQUENCES $\underline{R}^{(k)}$, $\underline{Q}^{(k)}$ AND $A^{(k)}$

WHERE

$$A^k = \underline{Q}^{(k)} \underline{R}^{(k)}$$

AND

$$A^{(k)} = (\underline{Q}^{(k)})^T A \underline{Q}^{(k)}$$

WE SEE NOW WHY QR ALGORITHM WORKS!

- $\underline{Q}^{(k)}$ CONTAINS OUR APPROXIMATION OF THE EIGENVECTORS.
- THE DIAGONAL ENTRIES OF $A^{(k)}$ ARE THE RAYLEIGH QUOTIENTS.

WE CAN MODIFY THE "PURE" QR ALGORITHM TO BE "PRACTICAL".

"PRACTICAL" QR ALGORITHM

$$(Q^{(0)})^T A^{(0)} Q^{(0)} = A$$

for $k = 1, 2, \ldots$

PICK SHIFT $\mu^{(k)}$

$$Q^{(k)} R^{(k)} = A^{(k-1)} - \mu^{(k)} I$$

$$A^{(k)} = R^{(k)} Q^{(k)} + \mu^{(k)} I$$

IF OFF-DIAGONAL $A^{(k)}_{j,j+1}$ IS CLOSE TO ZERO

SET $A_{j,j+1} = A_{j+1,j} = 0$

TO OBTAIN

$$\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$$

APPLY SAME ALGORITHM TO A_1, A_2 SEPARATELY.

CONNECTION TO INVERSE ITERATION

$$A^k = \underline{Q}^{(k)} \underline{R}^{(k)}$$

TAKE THE INVERSE

$$A^{-k} = (\underline{R}^{(k)})^{-1} (\underline{Q}^{(k)})^T$$

USING THE SYMMETRY OF A^{-1}.

$$A^{-k} = \underline{Q}^{(k)} (\underline{R}^{(k)})^{-T}$$

TAKE

$$P = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}$$

SINCE $P^2 = I$

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

$$A^{-k} P = \left[\underline{Q}^{(k)} P \right] \left[P \left(\underline{R}^{(k)} \right)^{-T} P \right]$$

THIS IS THE SIMULTANEOUS ITERATION OF A^{-1} WITH INITIAL MATRIX P.

FIRST COLUMN OF $\underline{Q}^{(k)} P$ IS ~~THAT~~ THE LAST COLUMN OF $\underline{Q}^{(k)}$

THIS CAN BE VIEWED AS INVERSE ITERATION APPLIED TO e_m.

THIS CAN THEREFORE BE ACCELERATED USING SHIFTS!

$\mu^{(k)}$.

$$A^{(k-1)} - \mu^{(k)} I = Q^{(k)} R^{(k)}$$

$$A^{(k)} = R^{(k)} Q^{(k)} + \mu^{(k)} I$$

THIS RETAINS

$$A^{(k)} = \left(Q^{(k)}\right)^{T} A^{(k-1)} Q^{(k)}$$

AND CONSEQUENTLY

$$A^{(k)} = \left(\underline{Q}^{(k)}\right)^{T} A \, \underline{Q}^{(k)}.$$

CHOICE OF SHIFTS

(1) RAYLEIGH QUOTIENT OF THE LAST COLUMN OF $\underline{Q}^{(k)}$.

$$\mu^{(k)} = (q_m^{(k)})^T A\, q_m^{(k)}$$

THIS WILL GIVE A CUBIC CONVERGENCE TO THE E-VEC AND E-VAL.

THIS COMES FOR FREE SINCE IT IS THE m,m ENTRY OF $A^{(k)}$.

$$A_{mm}^{(k)} = e_m^T A^{(k)} e_m$$

$$= e_m^T \underline{Q}^{(k)T} A \underline{Q}^{(k)} e_m$$

$$= q_m^{(k)T} A\, q_m^{(k)}$$

SETTING $\mu^{(k)} = A_{mm}^{(k)}$ IS KNOWN AS THE RAYLEIGH QUOTIENT SHIFT.

(2) ~~WILKO~~ WILKINSON SHIFT

SET B (2×2) AS THE LOWER RIGHT MOST SUBMATRIX OF $A^{(k)}$

$$B = \begin{bmatrix} a_{m-1} & b_{m-1} \\ b_{m-1} & a_m \end{bmatrix}$$

SET $\mu^{(k)}$ AS THE EIGENVALUE OF B CLOSEST TO a_m.

$$\mu^{(k)} = a_m - \frac{\text{sign}(\delta)\, b_{m-1}^2}{\left(|\delta| + \sqrt{\delta^2 + b_{m-1}^2}\right)}$$

WHERE

$$\delta = \frac{a_{m-1} - a_m}{2}$$

STABILITY

FOR $A \in \mathbb{R}^{m \times m}$ REAL, SYMMETRIC, AND TRIDIAGONAL.

QR ALGORITHM ON A COMPUTER SATISFYING FPA I & II GIVES

$$\tilde{Q}\tilde{\Lambda}\tilde{Q}^T = A + \delta A$$

$$\frac{\|\delta A\|}{\|A\|} = O(\epsilon)$$

FOR SOME $\delta A \in \mathbb{R}^{m \times m}$

QUESTION

(i) $(A^T)^{-1} = (A^{-1})^T$

THIS IS TRUE

$(A^{-1}A)^T = I$

$A^T(A^{-1})^T = I$

THUS, $(A^{-1})^T = (A^T)^{-1}$

(ii) $(A^{-1})^k = (A^k)^{-1}$

$(A^{-1} \cdots A^{-1})(A \cdots A) = I$

k TIMES

$(A^{-1})^k A^k = I$

THUS,

$(A^{-1})^k = (A^k)^{-1}$.

UNSYMMETRIC E-VAL PROBLEM

MANY OF THE METHODS USED IN THE SYMMETRIC E-VAL PROBLEM CARRY OVER TO THE UNSYMMETRIC ONE.

- POWER METHOD. (ITERATION)
- INVERSE ITERATION.
- "PURE" QR
- "PRACTICAL" QR.

BUT, MANY OF THE IMPORTANT DETAILS (CHOOSING SHIFTS, ETC) BECOME MORE COMPLICATED.

ITERATIVE METHODS

- WITH DIRECT METHODS OPERATION COUNT TO SOLVE A LINEAR ~~SYM~~ SYSTEM IS $\mathcal{O}(m^3)$

- INSTEAD GENERATE A SEQUENCE THAT CONVERGES TO THE SOLUTION (HOPEFULLY) VERY QUICKLY.

- & EACH ITERATION COSTS MATRIX-VECTOR MULTIPLICATION

TOTAL: $\mathcal{O}(m^2) \times N_{ITER}$

- ALLOWS USERS TO SOLVE LINEAR SYSTEMS ~~DEFINED~~ USING IMPLICIT MATRIX VECTOR MULTIPLICATION.

- KRYLOV SUBSPACE METHODS.