review of linear algebra

- a matrix is an array of numbers with m rows & n columns.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$$

$$a_{\underset{\text{row}}{i}\,\underset{\text{column}}{j}}$$

- for two matrices A, B, assuming they have the same dimension, we define

$$A + B = \begin{bmatrix} a_{11}+b_{11} & \cdots & a_{1n}+b_{1n} \\ \vdots & & \vdots \\ a_{m1}+b_{m1} & \cdots & a_{mn}+b_{mn} \end{bmatrix}$$

$$A+B = B+A \qquad A+(B+C) = (A+B)+C$$

- the "transpose" of a matrix is defined as

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & \ddots & & \\ \vdots & & \ddots & \\ a_{1n} & & & a_{mn} \end{bmatrix}$$

$A : m \times n$

$A^T : n \times m$

$$A = \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \qquad A^T = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \qquad (A^T)^T = A.$$

(A^H or A^* is $\overline{(A^T)}$ ← complex conjug.)

- a square matrix is "symmetric" if $A^T = A$.

- example:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \qquad A^T \neq A$$

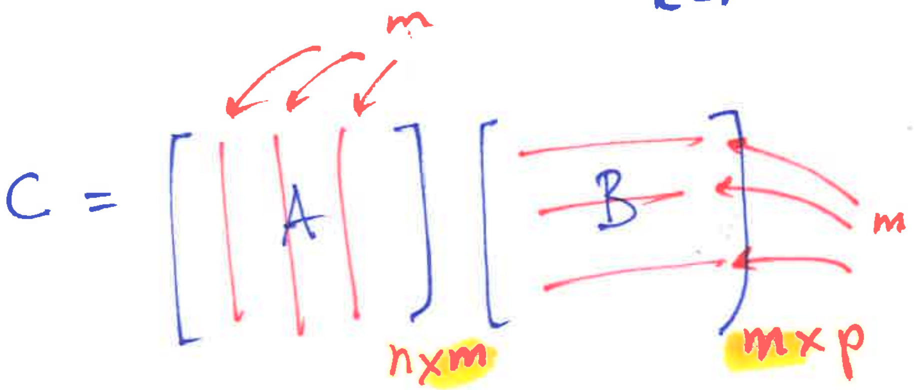
main diagonal

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix} \qquad A^T = A.$$

- if matrix is multiplied by scalar, then all its entries are multiplied.
- if A is an $n \times m$ matrix and B is an $m \times p$ matrix,

then $C = AB$ is an $n \times p$ matrix defined as

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



(# of col. of A = # of rows of B.)

- example:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ -1 & 2 \end{bmatrix}_{3\times \textcircled{2}} \qquad B = \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix}_{\textcircled{2}\times 2}.$$

(BA makes no sense)

$$C = AB = \begin{bmatrix} (1)(3)+(2)(-1) & 4 \\ 5 & 5 \\ -5 & 0 \end{bmatrix}_{3\times 2} = \begin{bmatrix} 1 & 4 \\ 5 & 5 \\ -5 & 0 \end{bmatrix}$$

- in general $AB \neq BA$, but $(AB)C = A(BC)$ always.

- if a matrix has only one column, then it is called a "column vector".

(n-vector)

$$r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} . \quad n \times 1$$

dimension.

- if r, q are two vectors of equal dimension, we define their "inner product" as

$$\langle r, q \rangle = \langle q, r \rangle \quad (= r \cdot q = q \cdot r)$$

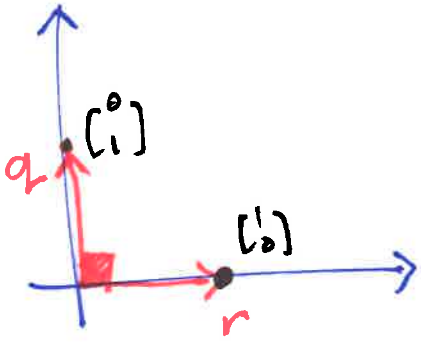
$$= r^T q = q^T r$$

$$= [r_1 \ \ r_2 \ \cdots \ r_n] \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix} = r_1 q_1 + r_2 q_2 + \cdots + r_n q_n$$

$$= \sum_{i=1}^{n} r_i q_i$$

the result of an inner product is always a scalar.

- r and q are "orthogonal" if $r^T q = 0$
- example: $r = \begin{bmatrix}1\\0\end{bmatrix}$ $\quad q = \begin{bmatrix}0\\1\end{bmatrix}$.



$$r^T q = [1 \;\; 0]\begin{bmatrix}0\\1\end{bmatrix} = 1\cdot 0 + 0\cdot 1$$
$$= 0$$

- the "norm" of a vector r is defined as

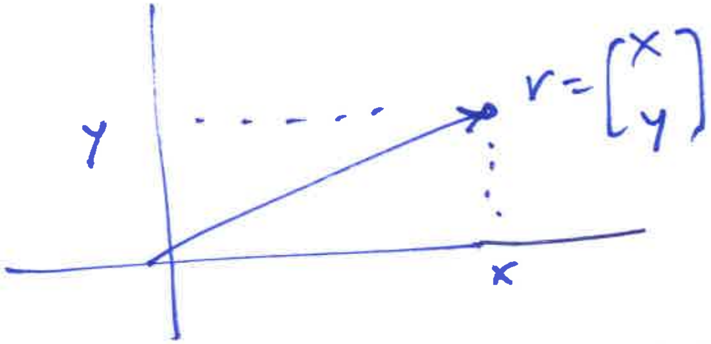
$$\|r\|^2 = r^T r = \sum_{i=1}^{n} r_i^2 = r_1^2 + \cdots + r_n^2$$

$$\left(\|r\| = \sqrt{r_1^2 + \cdots + r_n^2}\right)$$

$$\|r\| = 0 \iff r = \underset{n\times 1}{0} = \begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}$$

- the norm of a vector can be interpreted as its "length".

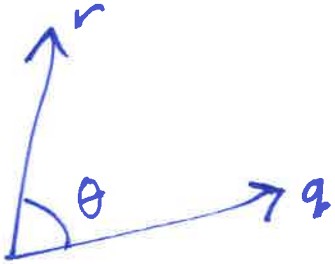
in 2 dimensions



$\|r\| = \sqrt{x^2+y^2}$ = length of vector r.

- we have the following

$$r^T q = \|r\| . \|q\| . \cos(\theta) .$$



$$r^T r = \|r\| \; \|r\| \; \cancelto{1}{\cos 0} = \|r\|^2$$

$$r \perp q \rightarrow r^T q = \|r\| \; \|q\| \; \cancelto{0}{\cos \frac{\pi}{2}} = 0 .$$

- interpretation 1 of matrix multiplication

A: $n \times m$ $\qquad$ B: $m \times p$

$$AB = \begin{bmatrix} \text{---} \; a_1^T \; \text{---} \\ \text{---} \; a_2^T \; \text{---} \\ \vdots \\ \text{---} \; a_n^T \; \text{---} \end{bmatrix}_{n \times \boxed{m}} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix}_{\boxed{m} \times p}$$

$$= \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & & \\ \vdots & & \ddots & \\ a_n^T b_1 & & & a_n^T b_p \end{bmatrix}_{n \times p}$$

(this is how we usually multiply matrices anyway!)

- example : (interp. 1)

$$A = \begin{array}{cc} \left[\begin{array}{cc} 1 & 2 \\ 2 & 1 \\ -1 & 2 \end{array}\right] & \begin{array}{l} a_1^T \\ a_2^T \\ a_3^T \end{array} \end{array} \qquad B = \begin{array}{c} \left[\begin{array}{cc} 3 & 2 \\ -1 & 1 \end{array}\right] \\ \begin{array}{cc} b_1 & b_2 \end{array} \end{array}$$

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

$$= \begin{bmatrix} 1 & 4 \\ 5 & 5 \\ -5 & 0 \end{bmatrix}.$$

- interpretation 2 of matrix multiplication

$A: n \times m \qquad B: m \times p$

$$AB = \underbrace{\begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_m \\ | & | & & | \end{bmatrix}}_{n \times \boxed{m}} \underbrace{\begin{bmatrix} - & b_1^T & - \\ - & b_2^T & - \\ & \vdots & \\ - & b_m^T & - \end{bmatrix}}_{\boxed{m} \times p}$$

$a_1 : n \times \boxed{1}$

$b_1^T : \boxed{1} \times p$

$\parallel$

$a_1 b_1^T : n \times p$

$$= \underbrace{a_1 b_1^T}_{n \times p} + \underbrace{a_2 b_2^T}_{n \times p} + \cdots + \underbrace{a_m b_m^T}_{n \times p}$$

- example: (interp. 2) $A = \begin{array}{c} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ -1 & 2 \end{bmatrix} \\ \begin{array}{cc} a_1 & a_2 \end{array} \end{array}$ $\quad B = \begin{array}{cc} \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix} & \begin{array}{c} b_1^T \\ b_2^T \end{array} \end{array}$

$$AB = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}_{3\times1} \begin{bmatrix} 3 & 2 \end{bmatrix}_{1\times2} + \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \begin{bmatrix} -1 & 1 \end{bmatrix}$$

$$\cancel{= \begin{bmatrix} 1 & 4 \\ 5 & 5 \\ -5 & 0 \end{bmatrix}}.$$

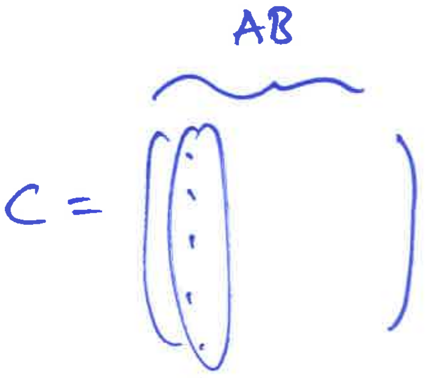
$$= \begin{bmatrix} 3 & 2 \\ 6 & 4 \\ -3 & -2 \end{bmatrix} + \begin{bmatrix} -2 & 2 \\ -1 & 1 \\ -2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 4 \\ 5 & 5 \\ -5 & 0 \end{bmatrix}.$$

- interpretation 3 of matrix multiplication

$$AB = A_{n\times m}\begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix}_{m\times p}$$

$$= \begin{bmatrix} | & | & & | \\ Ab_1 & Ab_2 & \cdots & Ab_p \\ | & | & & | \end{bmatrix}_{n\times p}$$



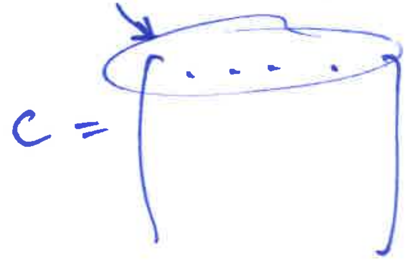
1st col is independent of all col of B except col 1.

- interpretation 4 of matrix multiplication

$$AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_n^T & — \end{bmatrix}_{n\times m} B_{m\times p}$$

$$= \begin{bmatrix} — & a_1^T B & — \\ — & a_2^T B & — \\ & \vdots & \\ — & a_n^T B & — \end{bmatrix}_{n\times p}$$

depends only on 1st rows of A.



- example : (interp. 3)

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ -1 & 2 \end{bmatrix} \qquad B = \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix}$$

$$AB = \left[\overbrace{\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix}}^{Ab_1} \quad \underbrace{\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}}_{Ab_2} \right]$$

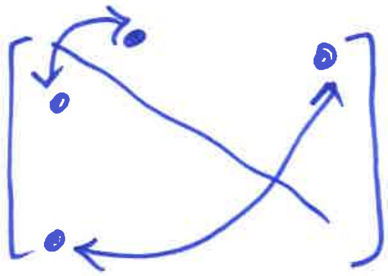
$$= \begin{bmatrix} 1 & 4 \\ 5 & 5 \\ -5 & 0 \end{bmatrix}$$

- example : (interp. 4)

$$AB = \begin{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix} \\ \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix} \\ \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix} \end{bmatrix} \begin{matrix} \} \, a_1^T B \\ \} \, a_2^T B \\ \} \, a_3^T B \end{matrix}$$

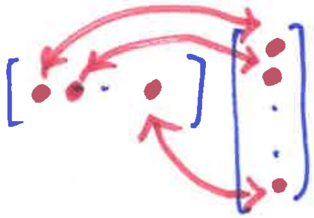
$$= \begin{bmatrix} 1 & 4 \\ 5 & 5 \\ -5 & 0 \end{bmatrix}$$

review of last lecture



- vectors, matrices, symmetric matrices

- inner product $\langle r, q \rangle = r^T q =$



- orthogonality $r^T q = 0$

- norm (or length) of a vector $\|r\|^2 = r^T r$

  $(\|r\| = 0 \iff r = 0)$

- $C = AB$ $\quad A: n \times \textcircled{m} \quad B: \textcircled{m} \times p$

- $AB \neq BA$ in general.

  (BA make not even make sense)

review of last lecture

- $C = AB$ $\qquad A: n \times m \qquad B: m \times p \qquad AB \neq BA.$

- $$AB = \begin{bmatrix} — a_1^T — \\ A \\ — \\ — \end{bmatrix} \begin{bmatrix} | & & | & | \\ b_1 & B & | & | \\ | & & | & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$ interp. 1

$$AB = \begin{bmatrix} | & | & | & | \\ a_1 & A & | & | \\ | & | & | & | \end{bmatrix} \begin{bmatrix} — b_1^T — \\ B \\ — \\ — \end{bmatrix} = \begin{pmatrix} a_1 & — b_1^T — \end{pmatrix} + \begin{pmatrix} \; \end{pmatrix} + \cdots$$ interp. 2

$$AB = A \begin{bmatrix} | & | & | & | \\ b_1 & B & | & | \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} | & & | \\ Ab_1 & \cdots & Ab_p \\ | & & | \end{bmatrix}$$ interp. 3

$$AB = \begin{bmatrix} — a_1^T — \\ A \\ — \\ — \end{bmatrix} B = \begin{bmatrix} — a_1^T B — \\ \vdots \\ — a_n^T B — \end{bmatrix}$$ interp. 4

- the identity matrix $I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$ (square), has the property that for any matrices (& vectors) X, Y

$$I_{m\times m} X_{m\times n} = X_{m\times n}$$

$$Y_{m\times n} I_{n\times n} = Y_{m\times n}$$

(the dimension of I has to be adjusted)

- the square matrix X is the "inverse" of the square matrix A if

$$XA = AX = I$$

then we use notation

$$A^{-1} := X$$

aside: $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^{-1} \neq \begin{bmatrix} 1 & 1/2 \\ 1/3 & 1/4 \end{bmatrix}.$

- example: $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

$$A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

<u>aside</u> (on matrix inverses)

- if A is not square, A^{-1} does not make sense
  (→ we have to consider "pseudo-inverses")

- even if A is square, A^{-1} may still not exist
  (→ a pseudo-inverse exists, however)

- even if A is square & invertible, A^{-1} may be highly unreliable, due to sensitivity to numerical errors.
  (→ the "condition number" of A warns us of this)

- if A is square & invertible with a "reasonable" condition number, then A^{-1} can be $\underbrace{\text{computed}}_{\text{(expensive)}}$ & used.

- if the entries of A are functions of an indendent variable t,

$$A(t) = \begin{bmatrix} a_{11}(t) & a_{12}(t) & \cdots \\ a_{21}(t) & a_{22}(t) & \cdots \\ \vdots & \vdots & \ddots \\ & & & a_{nm}(t) \end{bmatrix}$$

then the derivative of $A(t)$ is defined as

$$\frac{d}{dt} A(t) = \dot{A}(t) = \begin{bmatrix} \dot{a}_{11}(t) & \dot{a}_{12}(t) & \cdots \\ \dot{a}_{21}(t) & \dot{a}_{22}(t) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

- it is not difficult to show that

$$\frac{d}{dt}\left[A(t) \; B(t) \right] = \dot{A}(t) \; B(t) + A(t) \; \dot{B}(t) .$$

- integral of $A(t)$

$$\int^{\tau} A(t)\,dt = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \int^{\tau} a_{ij}(t)\,dt & & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

- Laplace (Fourier) transform

$$\mathcal{L}\{A(t)\} = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \mathcal{L}\{a_{ij}(t)\} & & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

- word of caution:

$$A^2 \neq \begin{bmatrix} a_{ij}^2 \end{bmatrix}, \quad A^{-1} \neq \begin{bmatrix} \frac{1}{a_{ij}} \end{bmatrix}, \quad e^{A} \neq \begin{bmatrix} e^{a_{ij}} \end{bmatrix}$$

- we say that v is a "linear combination" of the vectors $v_1, \dots, v_n$ if there exist coefficients $\underbrace{\beta_1, \dots, \beta_n}_{\in \mathbb{R}}$ such that

$$v = \beta_1 v_1 + \cdots + \beta_n v_n.$$

- example: $v = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ is a linear combination of $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

$$v = 1 \cdot v_1 + 2 \cdot v_2.$$

linear independence

- we say that the vectors $v_1, \dots, v_n$ are "linearly independent" if

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0 \qquad \alpha_1, \dots, \alpha_n \in \mathbb{R}$$

is possible only if $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$ (otherwise the vectors are said to be "linearly dependent")

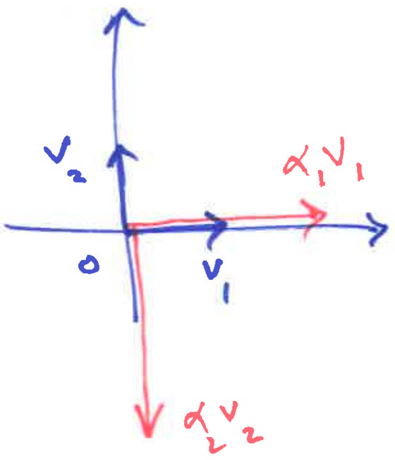
- example: $v_1 = \begin{bmatrix}1\\0\end{bmatrix}, \quad v_2 = \begin{bmatrix}0\\1\end{bmatrix}.$

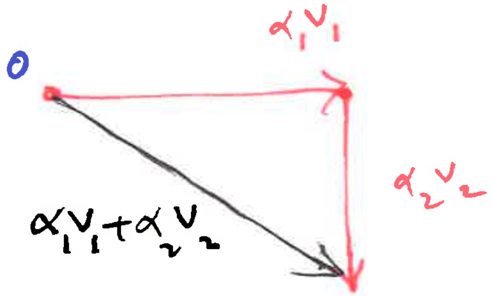
$$\alpha_1 v_1 + \alpha_2 v_2 = \begin{bmatrix}\alpha_1\\0\end{bmatrix} + \begin{bmatrix}0\\\alpha_2\end{bmatrix} = \begin{bmatrix}\alpha_1\\\alpha_2\end{bmatrix}$$

$$= \begin{pmatrix}0\\0\end{pmatrix}.$$

$$\downarrow$$

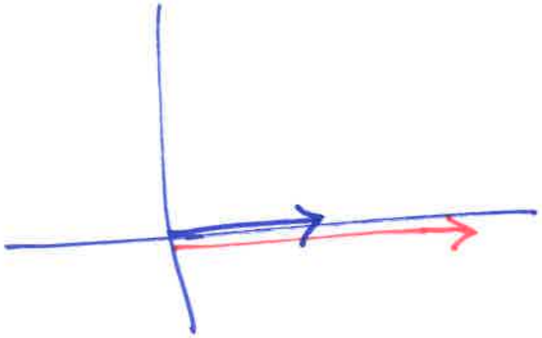
$$\begin{cases}\alpha_1 = 0\\ \alpha_2 = 0\end{cases}$$





linearly indep.

- example: $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$



$$\alpha_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \alpha_2 \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha_1 + 2\alpha_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

$$\alpha_1 + 2\alpha_2 = 0 \rightarrow \begin{cases} \alpha_1 = 2 \\ \alpha_2 = -1 \end{cases}, \quad \begin{cases} \alpha_1 = -4 \\ \alpha_2 = 2 \end{cases}, \ldots$$

linearly dep.

$$v_1 = \frac{1}{2} v_2$$

$$\underbrace{1}_{\alpha_1} v_1 \underbrace{- \frac{1}{2}}_{\alpha_2} v_2 = 0$$

- example: $v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad v_3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$

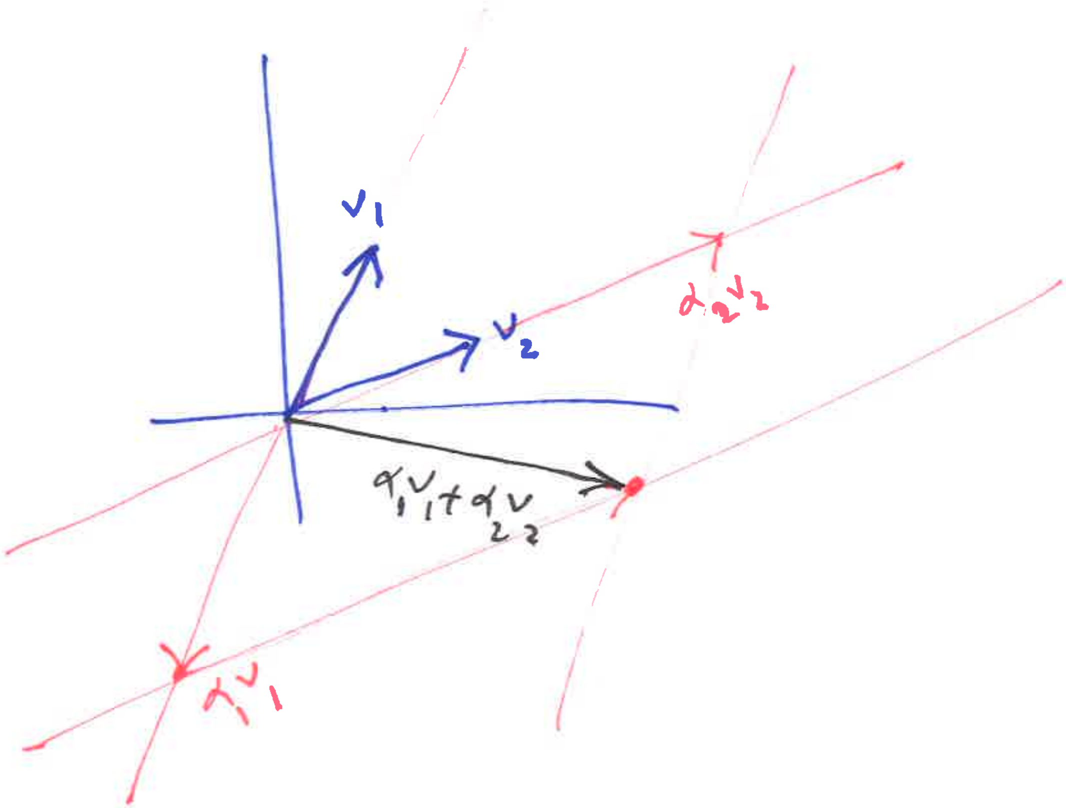
$$1 \cdot v_1 + 2 \cdot v_2 = v_3$$

$$\underbrace{1}_{\alpha_1} \cdot v_1 + \underbrace{2}_{\alpha_2} \cdot v_2 \underbrace{- 1}_{\alpha_3} v_3 = 0$$

linearly dep.

(if you can write some vector as lin comb. of others,
then they are linearly dep.)

- example: $v_1 = \begin{bmatrix}1\\2\end{bmatrix}$ $v_2 = \begin{bmatrix}2\\1\end{bmatrix}$.



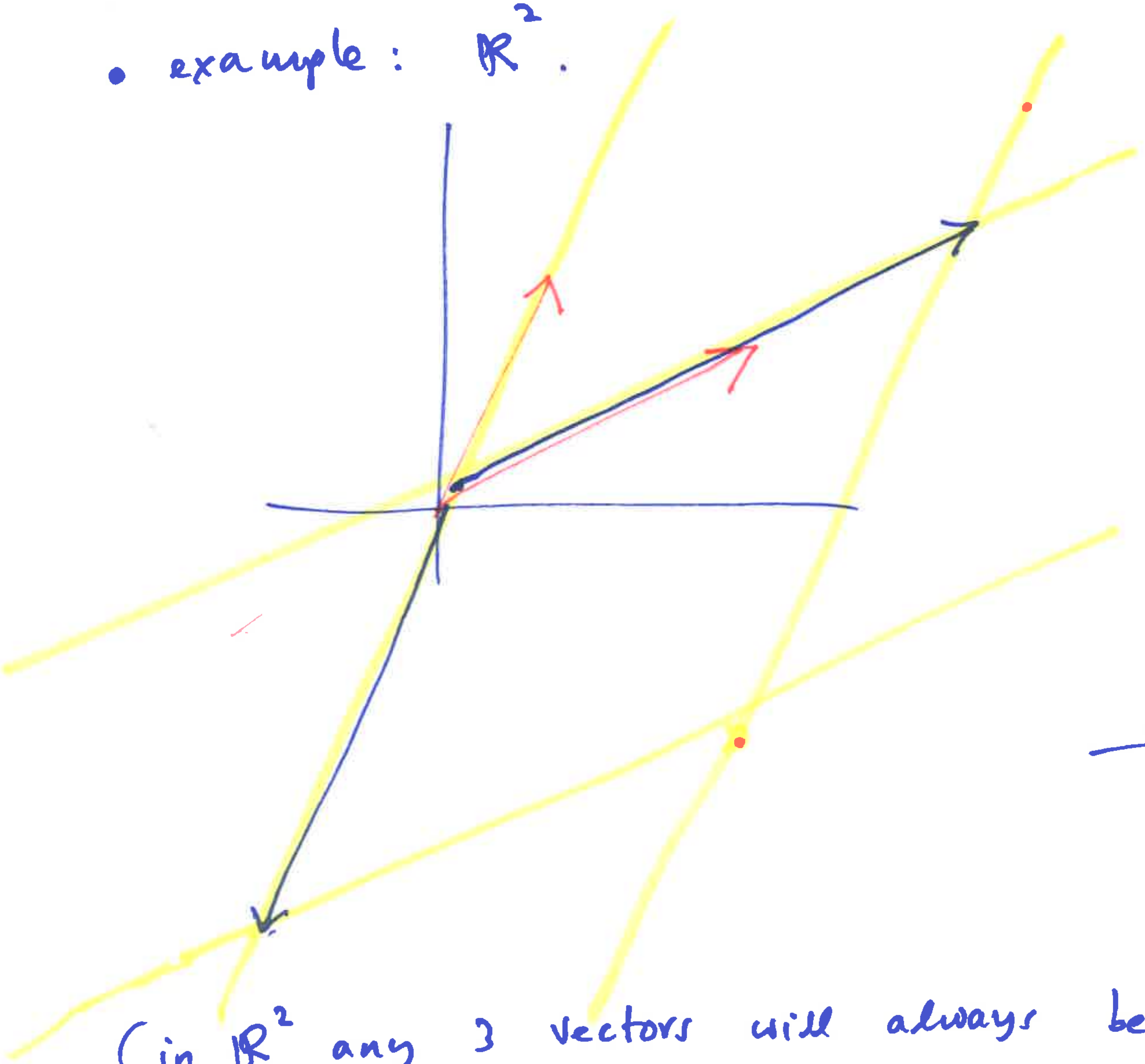
$$\alpha_1\begin{bmatrix}1\\2\end{bmatrix} + \alpha_2\begin{bmatrix}2\\1\end{bmatrix} = \begin{bmatrix}\alpha_1+2\alpha_2\\2\alpha_1+\alpha_2\end{bmatrix} = \begin{pmatrix}0\\0\end{pmatrix}. \rightarrow \begin{cases}\alpha_1+2\alpha_2=0\\2\alpha_1+\alpha_2=0\end{cases}$$

$$\rightarrow \begin{cases}\alpha_1=0\\\alpha_2=0\end{cases}$$

linearly indep.

- the maximum number of linearly independant vectors in a space is called the "dimension" of that space

- example : $\mathbb{R}^2$.



$$v_3 = \alpha_1 v_1 + \alpha_2 v_2$$

$$\alpha_1 v_1 + \alpha_2 v_2 - v_3 = 0$$

$\rightarrow d = 2.$

(in $\mathbb{R}^2$ any 3 vectors will always be linearly dependant)

- if the dimension of a space is n, and if $v_1, \ldots, v_n$ are linearly independent, then every vector in that space can be written as a linear combination of $v_1, \ldots, v_n$.

i.e., for any y there exist $\alpha_1, \ldots, \alpha_n$ such that

$$y = \alpha_1 v_1 + \cdots + \alpha_n v_n . \longrightarrow$$

in this case, the set of vectors $\{v_1, \ldots, v_n\}$ is called a "basis" for this space.

- example : space is R^n

$$v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1}, \quad v_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1}, \quad \ldots, \quad v_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{n\times 1} .$$

$$y = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} = \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_n v_n$$

2-11

- an "orthonormal" basis is a basis in which all vectors are mutually orthogonal & have unit length.

$\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$, $\quad v_i \perp v_j \quad i \neq j \qquad (v_i^T v_j = 0)$,

basis $\qquad \|v_i\| = 1 \quad \forall i \qquad (v_i^T v_i = 1)$

- example: $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$

- example: $w_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \; w_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \rightarrow v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \; v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$

- example: Fourier analysis. (cos & sin functions are the basis elements)

- if $v_1, \ldots, v_n$ are lineary independent in n-dimensional space (i.e., they form a basis), then any vector y is "uniquely" representable as a linear combination of $v_1, \ldots, v_n$.

proof by contradiction:

$$y = \alpha_1 v_1 + \cdots + \alpha_n v_n$$

$$y = \beta_1 v_1 + \cdots + \beta_n v_n$$

$$\alpha_1 v_1 + \cdots + \alpha_n v_n = \beta_1 v_1 + \cdots + \beta_n v_n$$

$$(\alpha_1 - \beta_1) v_1 + \cdots + (\alpha_n - \beta_n) v_n = 0 .$$

from linear independence, we have

$$\begin{cases} \alpha_1 - \beta_1 = 0 \\ \vdots \\ \alpha_n - \beta_n = 0 \end{cases} \to \begin{cases} \alpha_1 = \beta_1 \\ \vdots \\ \alpha_n = \beta_n \end{cases}$$

which is a contradiction.

2-14

fundamental subspaces of a matrix

motivation : consider solving the linear system of equations

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = y_1 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m = y_n \end{cases}$$

$x_1, \ldots, x_m$ unknown

a_{ij} & $y_1, \ldots, y_n$ known

$$\Updownarrow$$

$Ax = y$

A : $n \times m$ matrix

x : $m \times 1$ vector

y : $n \times 1$ vector

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}_{m \times 1}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$$

we know A & y, we seek x.

- we would like to use linear algebra to answer the following questions
  - when does $Ax=y$ have a solution?
  - when is the solution unique?
  - (later: when no solution, what else can we say?)

- very useful to think of $Ax=y$ as

$$Ax = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_m \\ | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

$x_i \in \mathbb{R}$

$$= \enclose{circle}{x_1} \begin{matrix} | \\ a_1 \\ | \end{matrix} + \cdots + \enclose{circle}{x_m} \begin{matrix} | \\ a_m \\ | \end{matrix}$$

(interp 2 of matrix multiplication!)

$$= y$$

i.e., Ax gives linear combination of columns of A (may or may not equal y.)

review of last lecture

- linear independence

$v_1, \cdots, v_n$ linear independence $\iff$ $\underbrace{\alpha_1 v_1 + \cdots + \alpha_n v_n}_{\text{linear combination.}} = 0$ if $\alpha_1 = \cdots = \alpha_n = 0$

- dim. of space = max. # of linear independent vectors
- if $v_1, \cdots, v_n$ are linearly independent in n-dimensional space, then every vector in that space can be written as a linear combination of $\underbrace{v_1, \cdots, v_n}_{\{v_1, \cdots, v_n\} \text{ is a "basis"}}$.

review of last lecture

- orthonormal basis: set of basis vectors $\{v_1, \cdots, v_n\}$ s.t.

$$v_i \perp v_j \qquad i \neq j$$

$$\|v_i\| = 1 \qquad \forall i$$

- uniquness of basis representation of a vector

$$y = \alpha_1 v_1 + \cdots + \alpha_n v_n$$

α_i are unique.

$$y^{(v)} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \longrightarrow$$

in the $\{v_1, \cdots, v_n\}$ basis.

in $\mathbb{R}^n$: $v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \; v_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \; \cdots, \; v_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$

$$y = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \alpha_1 v_1 + \cdots + \alpha_n v_n$$

review of last lecture

- Ax gives linear combination of col's of A

$$Ax = x_1 \begin{matrix} | \\ a_1 \\ | \end{matrix} + \cdots + x_n \begin{matrix} | \\ a_n \\ | \end{matrix} \qquad A : m \times n \qquad A \sim \begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix}$$

- as x varies over all possible vectors in $\mathbb{R}^n$, we call the resulting set of vectors Ax the "column space" of A

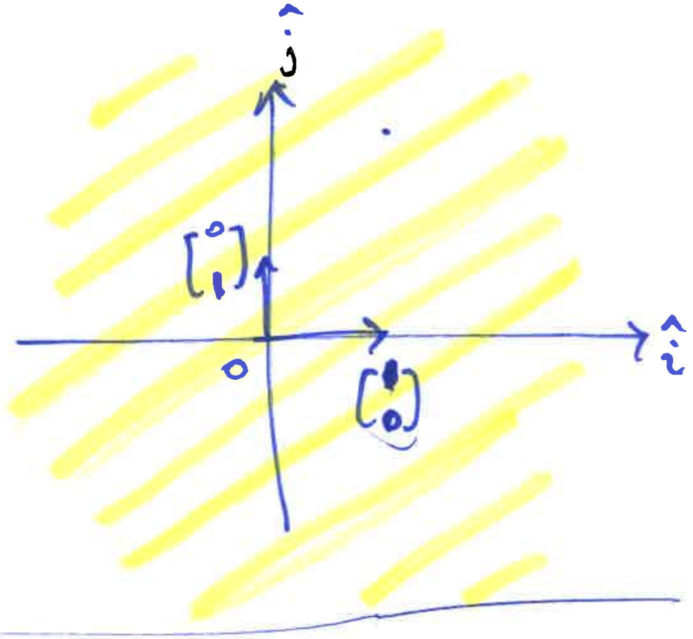
$$\mathcal{C} = \{ y \text{ such that } y = Ax,\ x \in \mathbb{R}^n \}.$$

(also referred to as "range space")

- example: $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

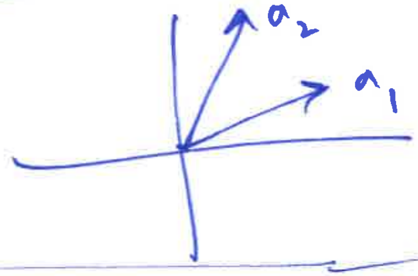
$$Ax = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

$\rightarrow \mathcal{C} = \mathbb{R}^2.$



---

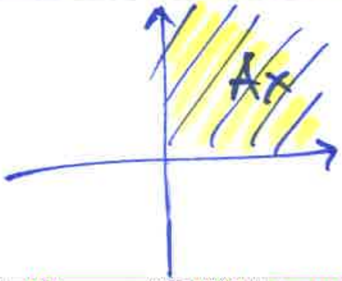
aside 1:



$\rightarrow \mathcal{C} = \mathbb{R}^2$ $\qquad A = (a_1 \ a_2)$

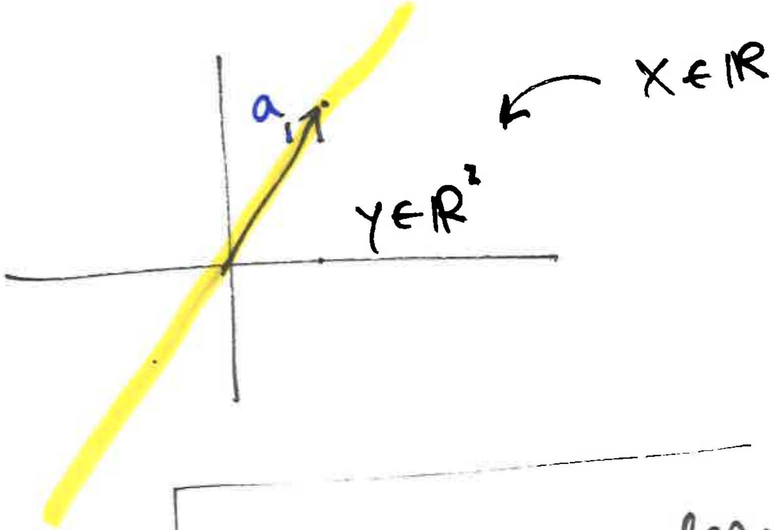
---

aside 2: if $x_1 \geq 0,\ x_2 \geq 0$ $\longrightarrow$



- example: $A = \begin{pmatrix}1\\2\end{pmatrix}$.

$$Ax = \underbrace{\begin{pmatrix}1\\2\end{pmatrix}}_{a_1} x \qquad x\in\mathbb{R}$$



- example: $A = \begin{bmatrix}1&2&3\\2&4&6\end{bmatrix}$

$$Ax = x_1\begin{bmatrix}1\\2\end{bmatrix} + x_2\begin{bmatrix}2\\4\end{bmatrix} + x_3\begin{pmatrix}3\\6\end{pmatrix}$$

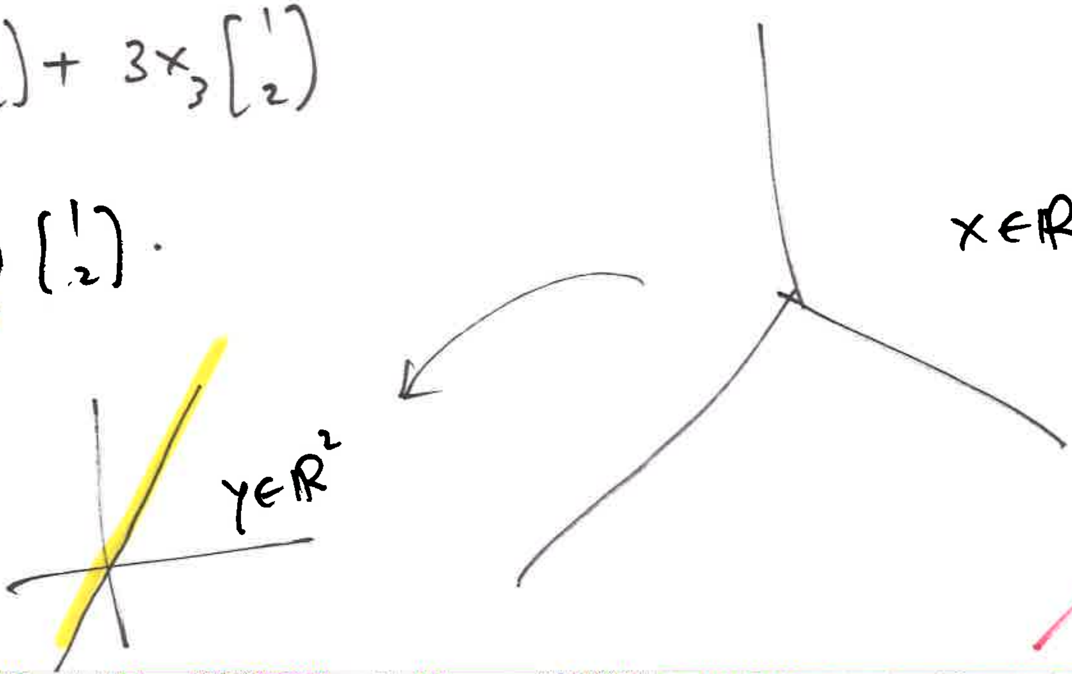
$$= x_1\begin{bmatrix}1\\2\end{bmatrix} + 2x_2\begin{bmatrix}1\\2\end{bmatrix} + 3x_3\begin{bmatrix}1\\2\end{bmatrix}$$

$$= \underbrace{(x_1+2x_2+3x_3)}_{\alpha}\begin{bmatrix}1\\2\end{bmatrix}.$$

$$= \alpha\begin{bmatrix}1\\2\end{bmatrix} \qquad \alpha\in\mathbb{R}$$

in both examples:

$$\mathcal{C} = \left\{\alpha\begin{bmatrix}1\\2\end{bmatrix},\ \alpha\in\mathbb{R}\right\}.$$



- the "right null space" (null space) of A is the set of all vectors x such that $Ax = 0$,

$$N = \{ x \text{ such that } Ax = 0 \}.$$

(also known as the "kernel")

- one interpretation of $Ax = 0$ is that

$$x_1 \overset{|}{\underset{|}{a_1}} + \cdots + x_n \overset{|}{\underset{|}{a_n}} = 0$$

- if the columns of A are linearly independent, then the right null space of A contains only the vector $x = 0$.

- example: $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$Ax = x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \rightarrow \begin{cases} x_1 = 0 \\ x_2 = 0 \end{cases} \rightarrow x = 0$$

$$\rightarrow N = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$$

- example: $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \longrightarrow \mathcal{N} = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\} \curvearrowright \in \mathbb{R}^2$ $\quad (x \in \mathbb{R}^2)$.

$A = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \longrightarrow \mathcal{N} = \{ 0 \} \curvearrowright \in \mathbb{R}$ $\quad (x \in \mathbb{R})$

- example: $A = \begin{bmatrix} 2 & 6 \\ 1 & 3 \end{bmatrix}$

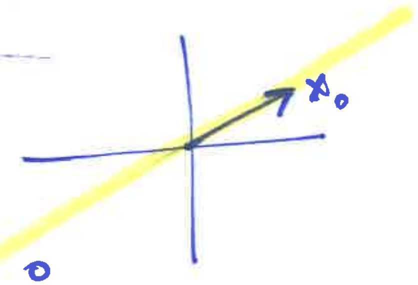
$$x = \begin{bmatrix} -3 \\ 1 \end{bmatrix} \longrightarrow Ax = -3 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 1 \begin{bmatrix} 6 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

$$x = \begin{bmatrix} -3\alpha \\ \alpha \end{bmatrix} \rightarrow Ax = -3\alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \alpha \begin{bmatrix} 6 \\ 3 \end{bmatrix} = \alpha \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

$$\mathcal{N} = \left\{ \alpha \begin{bmatrix} -3 \\ 1 \end{bmatrix}, \alpha \in \mathbb{R} \right\}$$

---

aside: $\boxed{x_0 \in \mathcal{N} \longrightarrow \alpha x_0 \in \mathcal{N}}$



$$Ax_0 = 0 \longrightarrow A(\alpha x_0) = \alpha(\underbrace{Ax_0}_{0}) = \alpha \cdot 0 = 0$$

3.7

- another interpretation of $Ax=0$ is that x is orthogonal to all rows of A

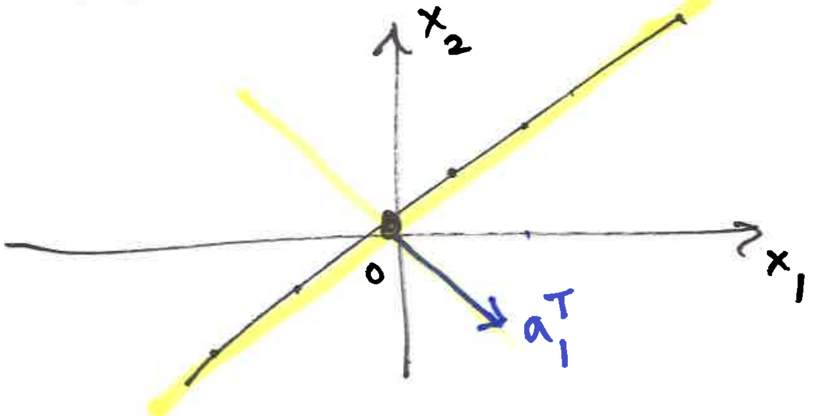
$$\begin{bmatrix} - & a_1^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} \begin{bmatrix} | \\ x \\ | \end{bmatrix} = \begin{bmatrix} a_1^T x \\ \vdots \\ a_m^T x \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \rightarrow \quad a_i^T x = 0 \quad \forall i$$

$$\boxed{\begin{array}{c} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \perp \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ \downarrow \\ N = \left\{ \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \alpha \in \mathbb{R} \right\} \end{array}}$$

- example: $A = [1 \quad -1] \rightarrow Ax = \overbrace{(1 \quad -1)}^{a_1^T} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$

$$x_1 - x_2 = 0 \rightarrow x_2 = x_1$$

$$Ax = a_1^T x = 0$$



- $y^T A$ gives a linear combination of the rows of A

$$y^T A = \begin{bmatrix} y_1 & \cdots & y_m \end{bmatrix} \begin{bmatrix} - & a_1^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} = y_1 a_1^T + \cdots + y_m a_m^T$$

- the "row space" of A is the set of all vectors z, $z^T = y^T A$, as y varies over all possible vectors in $\mathbb{R}^m$.

$$\mathcal{R} = \{ z \text{ such that } z^T = y^T A,\ y \in \mathbb{R}^m \}$$

- the "left null space" of A is the set of all vectors y such that $y^T A = 0^T$.

$$\mathcal{K} = \{ y \text{ such that } y^T A = 0^T \}$$

- if the rows of A are linearly independent, then the left null space of A contains only the vector $y^T = 0^T$.

$$\mathcal{K} = \{ 0 \}.$$

- it can be shown that the number of linearly independent columns of a matrix is equal to its number of linearly independent rows.

rank (A) = number of linearly independent columns of A

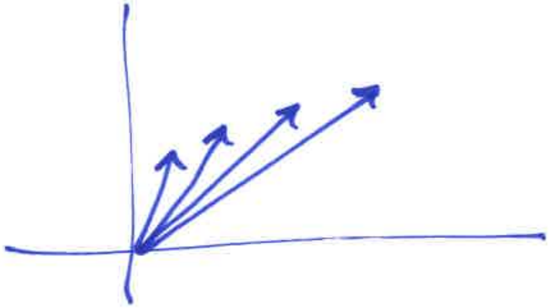
= number of linearly independent rows of A

- example:
$$A = \begin{bmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{bmatrix}$$

# of lin. ind. col. of A = 2 ⟶ rank (A) = 2.

row vectors : $\begin{bmatrix} 1 \\ 5 \end{bmatrix}^T, \begin{bmatrix} 2 \\ 6 \end{bmatrix}^T, \begin{bmatrix} 3 \\ 7 \end{bmatrix}^T, \begin{bmatrix} 4 \\ 8 \end{bmatrix}^T$

# of lin. ind. rows of $A = 2$.



- example: $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \\ 4 & 8 \end{bmatrix}$

$\text{rank}(A) = 1$

---

aside 1:

$\text{rank}(A) \leqslant \min\{m, n\}$ $\qquad A: m \times n$

---

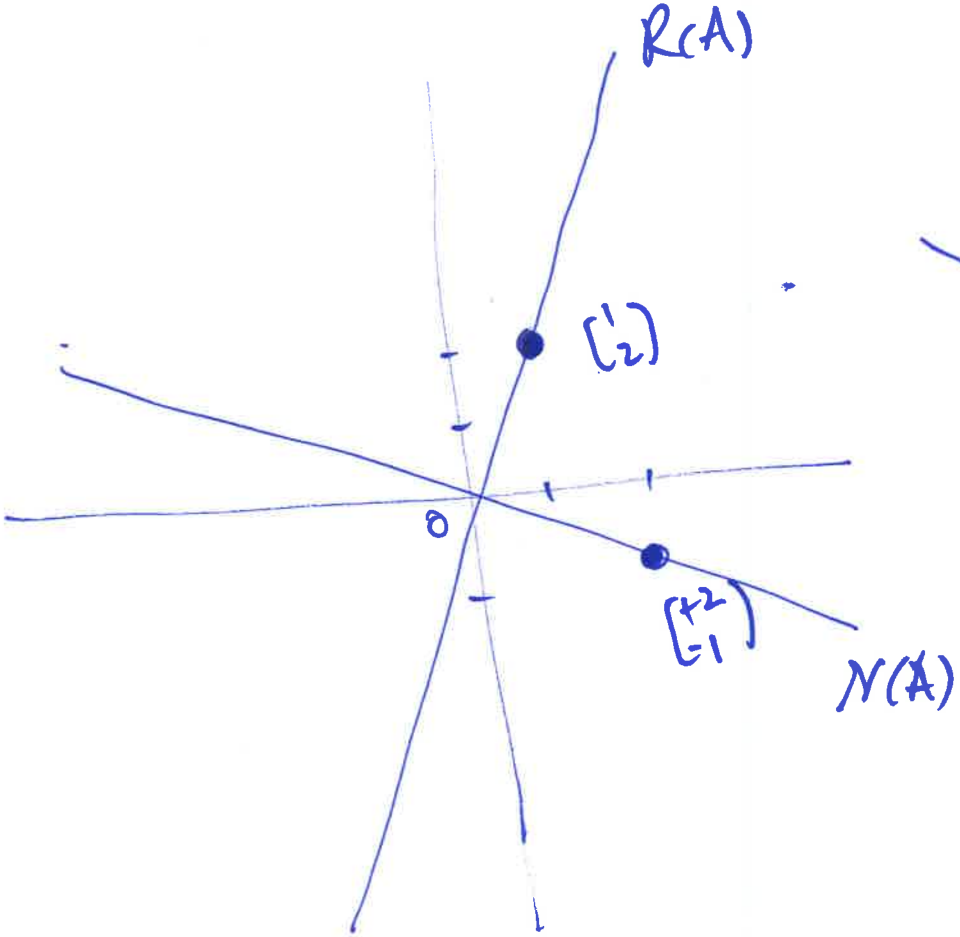
aside 2:

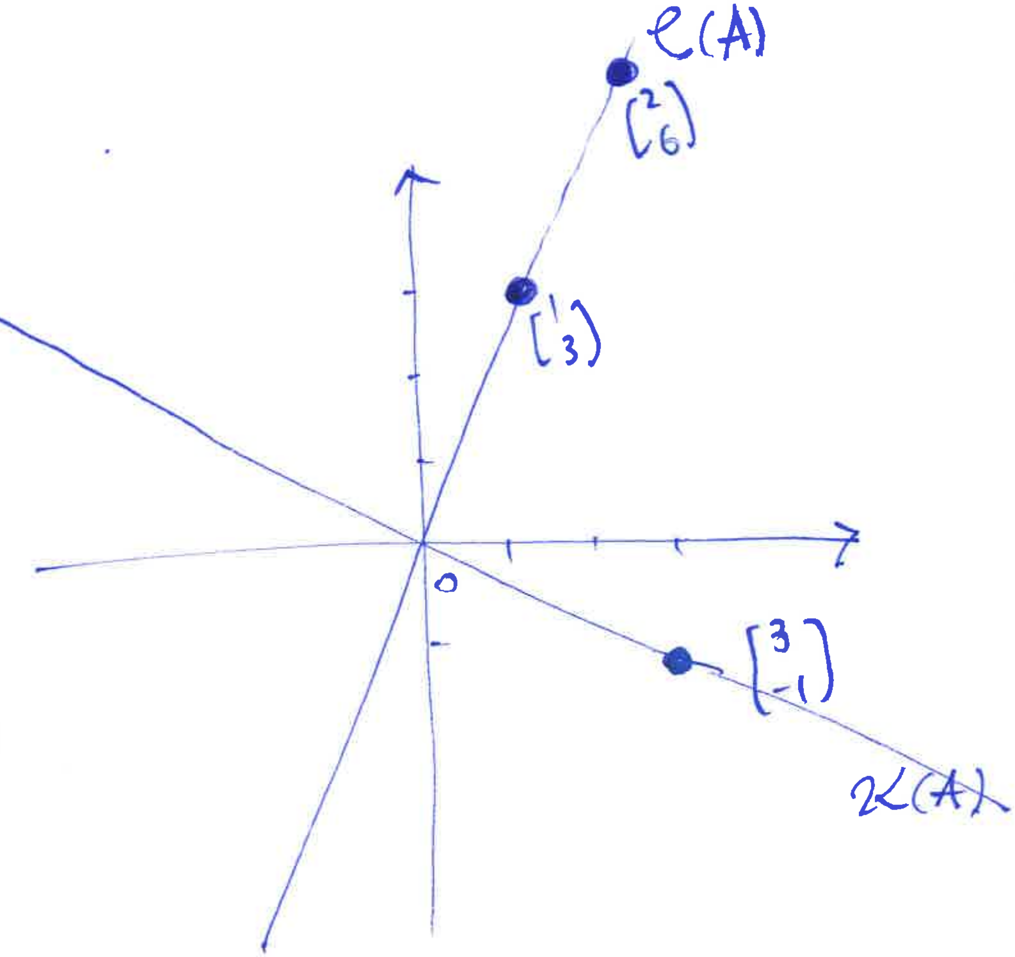
if A is $n \times n$, then

$$\text{rank}(A) = n \iff \mathcal{C}(A) = \mathbb{R}^n \iff \mathcal{N}(A) = \{0\}$$

$$\iff \mathcal{R}(A) = \mathbb{R}^n \iff \mathcal{K}(A) = \{0\}$$

- example: $A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}$





$y^T A$ (gives lin. comb. of rows of A

$Ax = 0$ (such x is orth. to rows. of A)

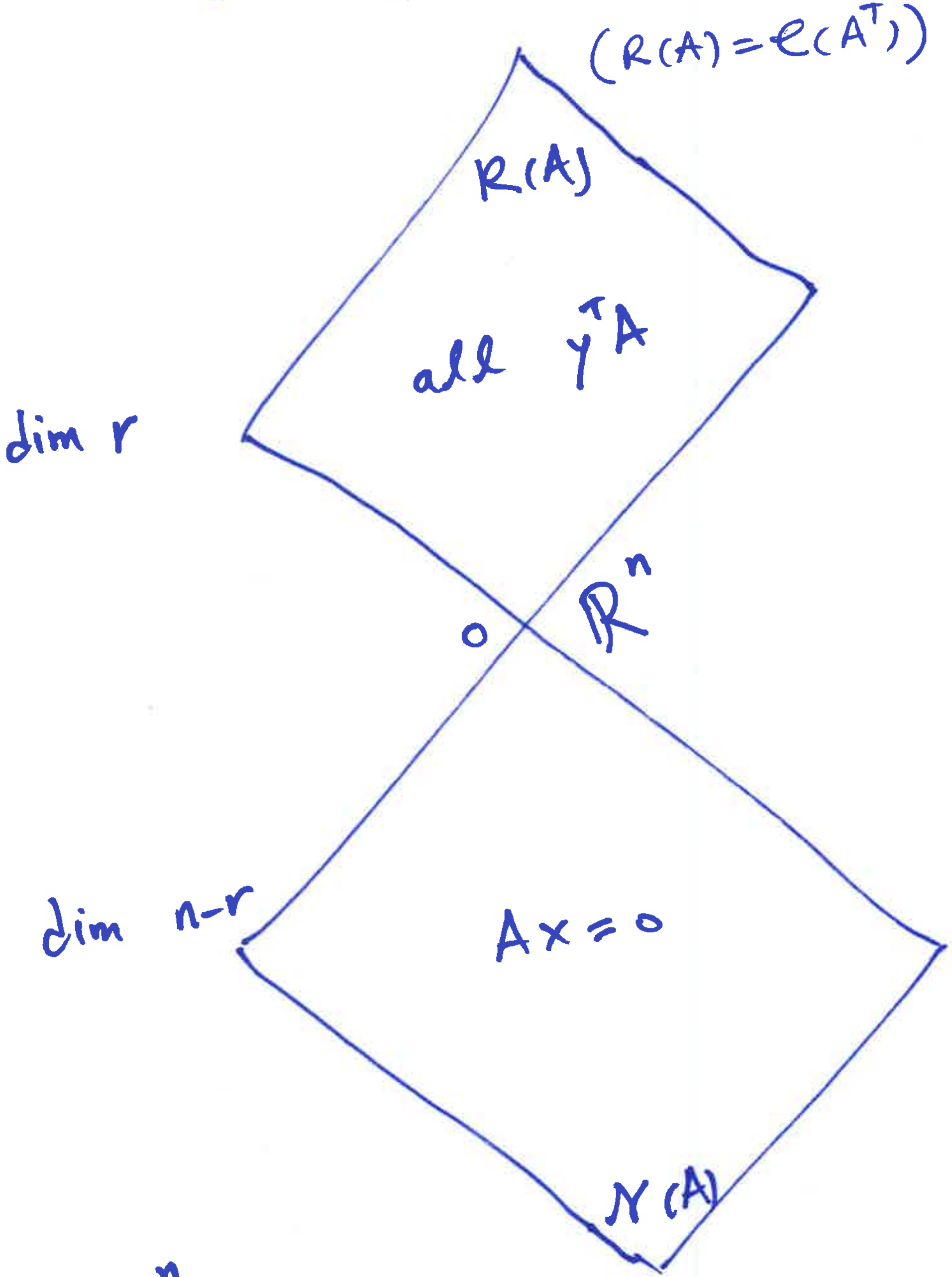
$\boxed{N(A) \perp R(A)}$

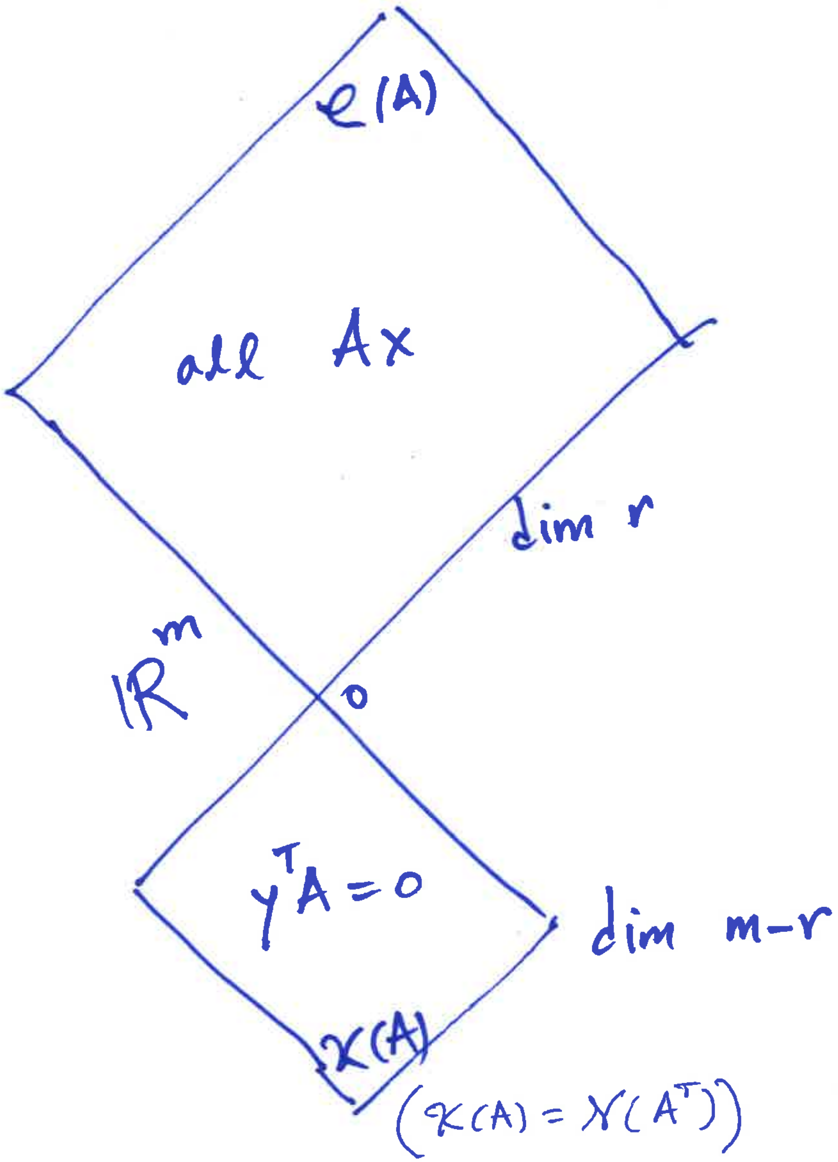
Ax (gives lin. comb. of col. of A

$y^T A = 0$ (such y is orth. to col. of A)

$\boxed{K(A) \perp C(A)}$.

$A = \left[\quad \right]_{m \times n}$ & $\text{rank}(A) = r \leq \min\{m,n\}$.





$x \in \mathbb{R}^n \rightarrow x = x_R + x_N$

$y \in \mathbb{R}^m \rightarrow y = y_C + y_K$

- example: $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $m = n = 2,\ r = 2$

$$\mathcal{R}(A) = \mathbb{R}^2 \qquad \ell(A) = \mathbb{R}^2$$

$$\mathcal{N}(A) = \{0\} \qquad \mathcal{K}(A) = \{0\}$$

- example: $A = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $m = 2,\ n = 1,\ r = 1$

$$\mathcal{R}(A) = \{\gamma\} \qquad \ell(A) = \left\{\alpha \begin{bmatrix} 1 \\ 2 \end{bmatrix}\right\}$$

$$(\dim = r = 1) \qquad (\dim = r = 1)$$

$$\mathcal{N}(A) = \{0\} \qquad \mathcal{K}(A) = \left\{\beta \begin{bmatrix} 2 \\ -1 \end{bmatrix}\right\}$$

$$\left(\dim = n - r = 1 - 1 = 0\right) \qquad \left(\dim = m - r = 2 - 1 = 1\right)$$

- example: $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ $m = n = 2,\ r = 2$

same as $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

- example: $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix}$ $\quad m = 2, \quad n = 3, \quad r = 1$

$\mathcal{R}(A) = \left\{ \gamma \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right\}$ (dim = 1) $\qquad \mathcal{C}(A) = \left\{ \alpha \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$ (dim = 1)

$\mathcal{N}(A) = \left\{ \delta \begin{bmatrix} 3 \\ -3 \\ 1 \end{bmatrix} + \varepsilon \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \right\}$ (dim = 2) $\qquad \mathcal{K}(A) = \left\{ \beta \begin{bmatrix} 2 \\ -1 \end{bmatrix} \right\}$ (dim = 1)

- some terminology $A: m \times n$, $\text{rank}(A) \leq \min\{m,n\}$
  - if $\text{rank}(A) = n$ we say A has "full column rank" (A is also said to be "one-to-one" or "injective")
  - if $\text{rank}(A) = m$ we say A has "full row rank" (A is also said to be "onto" or "surjective".)
  - if $\text{rank}(A) = \min\{m,n\}$ we say A has "full rank"; otherwise it is "rank deficient" ($\text{rank}(A) < \min\{m,n\}$)

- if A has full row rank then $Ax = y$ has a solution for any y

$$\left\{ \begin{array}{l} m \text{ lin. ind. rows} \Rightarrow m \text{ lin. ind. col.} \\ y \in \mathbb{R}^m \end{array} \right. \Rightarrow y \in \mathcal{C}(A) = \mathbb{R}^m$$

- consider $Ax=y$, x unknown, y & A known

we ask 2 questions:

① when does $Ax=y$ have a solution?

② assuming a solution exists, when is it unique?

answer to ①: $Ax = x_1 a_1 + \cdots + x_n a_n = y$

we have a solution if y belongs to the column space of A.

answer to ②: if the right null space of A contains only the vector zero (col. of A are lin. ind.) the solution is unique.

proof by contradiction:
suppose sol. not unique

$$\begin{cases} Ax=y \\ A\tilde{x}=y \end{cases} \rightarrow Ax = A\tilde{x} \rightarrow A(x-\tilde{x}) = 0$$

—> A z = 0

some lin. comb. of col. of A is zero,
which contradicts assumption of lin. ind.

we can show that independence of columns of A is also necessary.

proof by contradiction: suppose col. not ind.

$$\begin{tikzcd}\left\{\begin{array}{l} Az = 0 \\ Ax = y \end{array}\right. \arrow[r] & A(x+z) = y \arrow[d] \\ & A\tilde{x} = y\end{tikzcd}$$

which cotradicts uniquness.

one could ask 3rd question: how can we characterize all solutions when solution is not unique? answer: assume y in C(A), let x0 be a solution, & let F be any matrix s.t. C(F) = N(A). then general solution of Ax = y is given by Fz + x0.

- example: $A = \begin{bmatrix} 1 & 3 \\ 1 & 2 \end{bmatrix}$. $Ax = y$
  - a solution always exists (for any y) because the columns of A "span" the whole space ($\mathcal{C}(A) = \mathbb{R}^2$, $y \in \mathbb{R}^2$).
  - the solution is unique because the columns of A are linearly independent ($\mathcal{N}(A) = \{0\}$, $x \in \mathbb{R}^2$).

- example: $A = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 1 \end{bmatrix}$ $Ax = y$
  - a solution always exists because $\mathcal{C} = \mathbb{R}^2$, $y \in \mathbb{R}^2$.
  - the solution is not unique because the col. of A are lin. dependent ($a_1 = a_2 + a_3 \rightarrow a_1 - a_2 - a_3 = 0$)

$y = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$, solve $Ax = y \rightarrow \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 1 \end{bmatrix} x = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$.

$x_0 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$, $Ax_0 = y$.

$x_1 = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$, $Ax_1 = y$.

there exist many more solutions (infinitely many of them); how do we characterize them all?

$z = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \rightarrow Az = 0 \rightarrow \mathcal{N} = \left\{ \alpha \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \alpha \in \mathbb{R} \right\}$.

$x_0 + \alpha z = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1-\alpha \\ 1+\alpha \\ \alpha \end{bmatrix}$ is a solution of $Ax = y$ for <u>any</u> value of $\alpha \in \mathbb{R}$.

$$A(x_0 + \alpha z) = Ax_0 + A(\alpha z)$$
$$= Ax_0 + \alpha \cancelto{0}{Az} = Ax_0 = y.$$

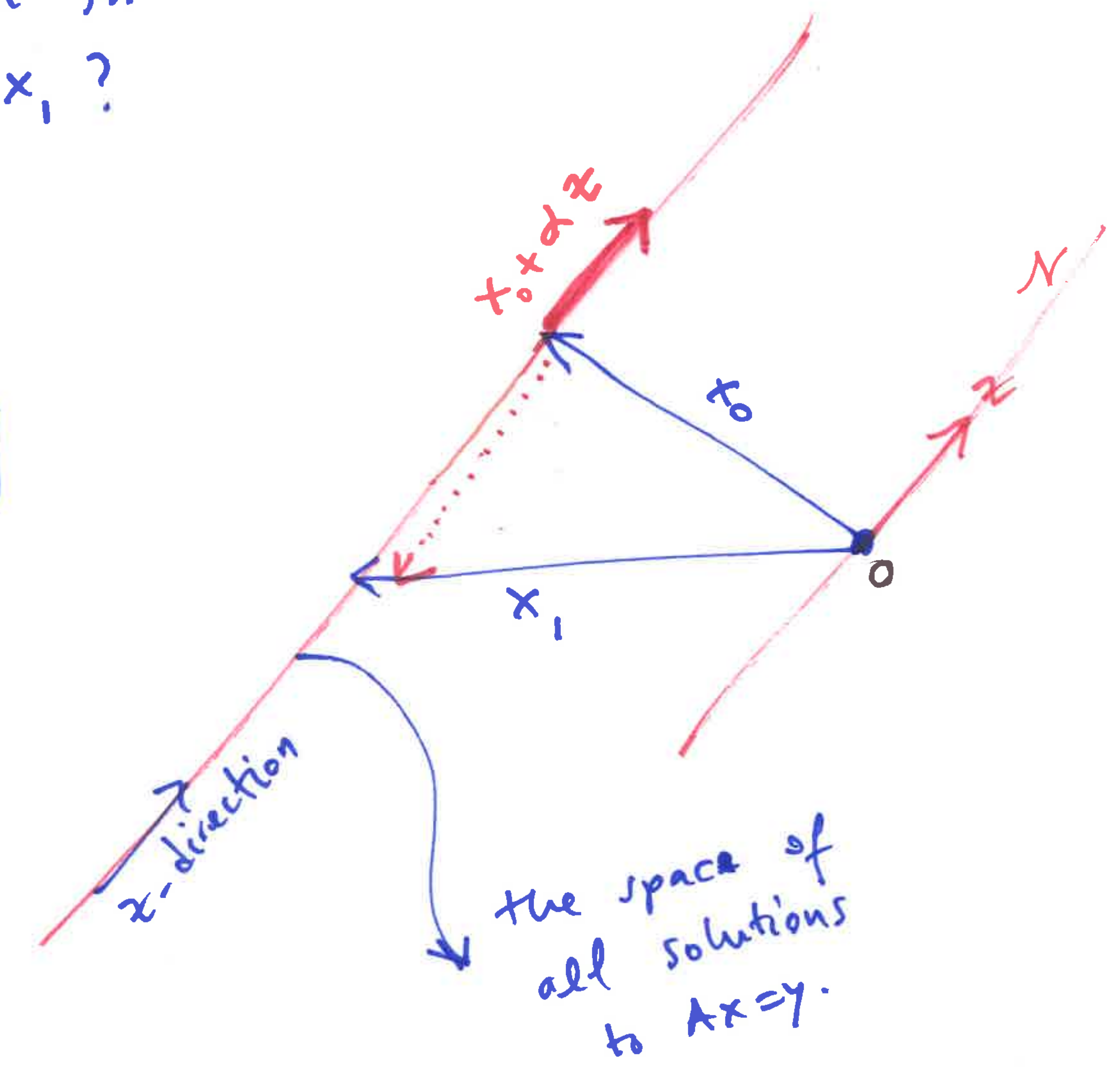
question: is there some α for which $x_0 + \alpha z = x_1$?

answer: yes, try $\alpha = 1$

$$x_0 + 1z = \begin{bmatrix}1\\1\\0\end{bmatrix} + \begin{bmatrix}-1\\1\\1\end{bmatrix}$$

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

$$= x_1 \quad \checkmark$$



---

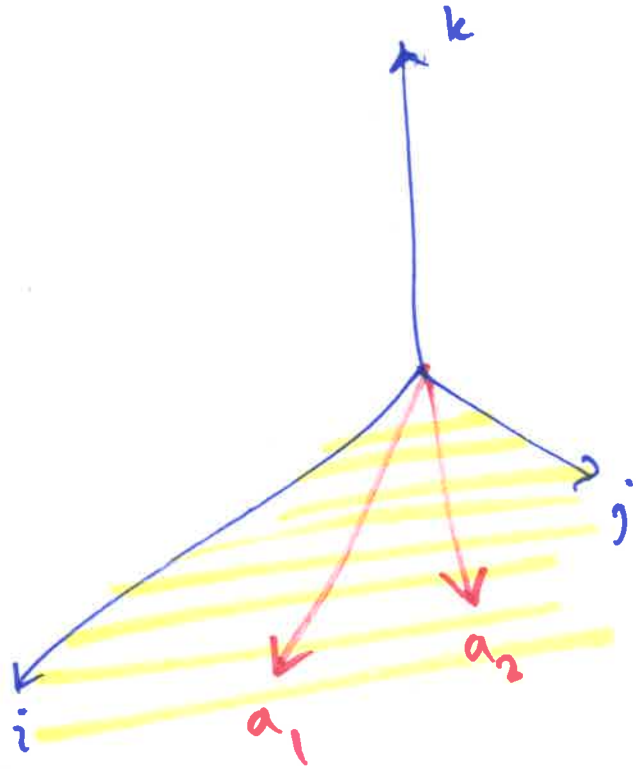
aside:

this has strong resemblance to finding solutions to linear differential equations:

$$\underset{\text{linear differential operator}}{\underset{\downarrow}{\mathcal{D}}}[x(t)] = y(t) \qquad \mathcal{D}\Big[\underbrace{x_p(t)}_{\text{particular solution}} + \underbrace{\sum_{i=1}^{q} \alpha_i x_h(t)}_{\text{spans space of homogeneous solutions.}}\Big] = y(t)$$

- example: $A = \begin{bmatrix} 2 & 3 \\ 1 & 2 \\ 0 & 0 \end{bmatrix}$ $\quad Ax = y \quad x \in \mathbb{R}^2,\ y \in \mathbb{R}^3$

$y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ ☹ $\qquad y = \begin{bmatrix} \alpha \\ \beta \\ 0 \end{bmatrix}$ ☺



$\mathcal{C} \neq \mathbb{R}^3$ $\left(y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \text{ has no solution} \right)$

but, if $y = \begin{bmatrix} \alpha \\ \beta \\ 0 \end{bmatrix}$, $\alpha, \beta \in \mathbb{R}$,

then there is a solution (because $y \in \mathcal{C}$)

and in this case the solution is unique.

- if A is $n \times n$ and $\text{rank}(A) = n$

$$\downarrow$$

$$\mathcal{R}(A) = \mathbb{R}^n$$
$$\mathcal{N}(A) = \{0\}$$

$$\downarrow$$

$Ax = y$ always has unique solution

$$\downarrow$$

A^{-1} exists & $x = A^{-1}y$ is the unique solution for any $y \in \mathbb{R}^n$.

(this creates a one-to-one correspondence between y's & x's)

---

aside:

$$Ax = y \rightarrow \underbrace{\cancel{A^{-1}}(\cancel{A}}_{I}x) = A^{-1}y \longrightarrow Ix = A^{-1}y \rightarrow x = A^{-1}y.$$

review of last lecture

- Ax gives linear combination of col's of A

  $Ax = x_1 a_1 + \cdots + x_n a_n$ $\qquad A : m \times n \qquad A \sim \begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix}$

- col. space $\mathcal{C} = \{ Ax$ as x varies over all vectors in $\mathbb{R}^n \}$

- right null space $\mathcal{N} = \{ x$ such that $Ax = 0 \}$.

  - if columns of A are lin. ind. → $\mathcal{N} = \{0\}$

  - $Ax = 0$ → $\boxed{a_i^T} x = 0 \quad \forall i$

    ↓

    a_i^T is ith row of A

review cont'd

- x^TA gives linear combination of rows of A

$$x^TA = x_1 a_1^T + \cdots + x_m a_m^T \qquad A: m \times n \qquad A \sim \begin{bmatrix} - & a_1^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix}$$

- row space $\mathcal{R} = \{ x^TA$ as x varies over all vectors in $\mathbb{R}^m \}$

- left null space $\mathcal{K} = \{ x$ such that $x^TA = 0^T \}$

  - if rows of A are lin. ind $\rightarrow$ $\mathcal{K} = \{0\}$.
  - $x^TA = 0^T \rightarrow x^T a_i = 0 \quad \forall i$

a_i is ith column of A

review of last lecture

- # of lin. ind. col. of A

  = # of lin. ind. rows of A

  =: rank (A) $\quad (\leq \min\{m,n\})$

(implications for nxn matrices with rank = n)

- consider solving $Ax = y$
  - solvability: $y \in \mathcal{R}(A)$
  - uniqueness: $\mathcal{N}(A) = \{0\}$

- when $y \in \mathcal{R}(A)$, $\mathcal{N}(A) \neq \{0\}$.
  - if x_p is a "particular" solution: $Ax_p = y$
  - if $x_h^{(i)}$ is a "homogeneous" solution: $Ax_h^{(i)} = 0$, $i = 1, \ldots, r$

  $\longrightarrow$ $x_p + \sum_{i=1}^{r} \alpha_i x_h^{(i)}$ is also a solution for any $\{\alpha_i\}$

- question: what if $y \notin \mathcal{C}(A)$? $\boxed{Ax=y \text{ has no solution}}$

if $Ax=y \rightarrow Ax-y=0 \rightarrow \|Ax-y\|=0$

& if $Ax \neq y \rightarrow Ax-y \neq 0 \rightarrow \|Ax-y\| \neq 0$

think of $e=Ax-y$ as error or "mismatch" between Ax & y. the best we can do is choose x such that $\|e\|=\|Ax-y\|$ is as small as possible.

$\downarrow$

least-squares

$\|e\|_2^2 = e_1^2 + \cdots + e_n^2$ is as small as possible by choosing a good x.

$\downarrow$

e^Te

the least-squares problem

- suppose it is desired to find the vector x such that $e = Ax - y$ has minimum length

$$\left(\|e\|_2^2 := \sum_{i=1}^{n} e_i^2 \right)$$

$$\underset{x}{\text{minimize}} \quad \|Ax - y\|_2^2$$

assume that $\mathcal{N}(A) = \{0\}$

$$\underbrace{\text{(col. of } A \text{ lin. ind.)}}$$

$$A \sim \begin{bmatrix} \text{tall} \end{bmatrix}_{m \times n}, \quad m \geq n, \quad \text{rank}(A) = n.$$

A has full column rank

$$A x = x_1 \begin{matrix}|\\a_1\\|\end{matrix} + \cdots + x_n \begin{matrix}|\\a_n\\|\end{matrix} = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix} x$$

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

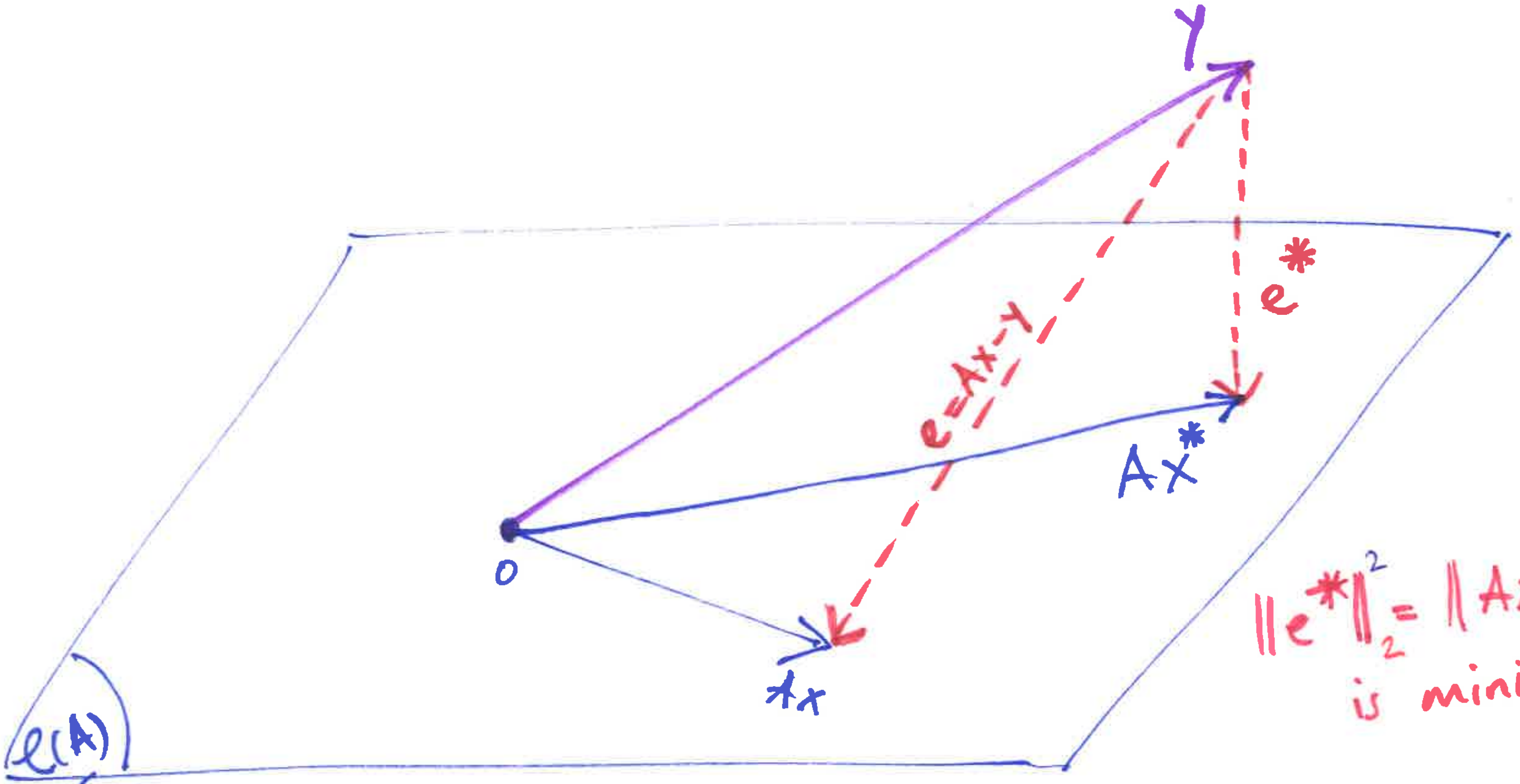
$y \in \mathbb{R}^m$

$\boxed{x \in \mathbb{R}^n}$

$e \in \mathbb{R}^m$

$Ax \in \mathbb{R}^m$

$$e = Ax - y$$



$$\|e^*\|_2^2 = \|Ax^* - y\|_2^2$$
is minimum.

space "spanned" by columns of A

for e^* to have minimum length, we need it to be "orthogonal to all vectors in the column space of A"

$$\Updownarrow$$

$e^* \perp a_1, \ldots, e^* \perp a_n$ ($\{a_i, i=1,\ldots,n\}$ form basis for $\mathcal{C}(A)$)

$$(e^*)^T a_1 = 0, \ldots, (e^*)^T a_n = 0$$

$$-(e^*)^T- \begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix} = [0 \cdots 0] = 0^T$$

($e^* \in \mathcal{K}(A)$)

$$(e^*)^T A = 0^T$$

$$A^T e^* = 0$$

$$A^T(Ax^* - y) = 0$$

$$A^TAx^* - A^Ty = 0$$

$$A^TAx^* = A^Ty \longrightarrow \boxed{x^* = (A^TA)^{-1} A^T y}$$

see comment

aside:

$(A \pm B)^T = A^T \pm B^T$

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

$(A^T)^T = A$

- comment regarding invertibility of A^TA:
  - A^TA is square (as is AA^T)
  - if $\mathcal{N}(A) = \{0\}$ (col's of A are lin. ind.) then (HW problem) A^TA is invertible

* where is x in picture?

- example: $\underbrace{\begin{bmatrix}1\\1\\1\end{bmatrix}}_{A} x = \underbrace{\begin{bmatrix}29\\30\\31\end{bmatrix}}_{Y}$, $\underset{x}{\text{minimize}} \; \|Ax-y\|_2$

$Ax=y$ has no solution $\left(\mathcal{C}(A) = \left\{\begin{bmatrix}\alpha\\\alpha\\\alpha\end{bmatrix}, \alpha \in \mathbb{R}\right\}\right.$, and $y = \begin{bmatrix}29\\30\\31\end{bmatrix} \notin \mathcal{C}(A)$ is not in $\left.\mathcal{C}(A)\right)$

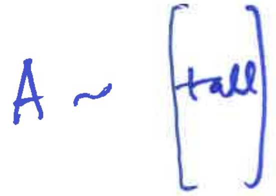
$A^TAx^* = A^Ty$. gives the least-squares solution

$$\begin{bmatrix}1 & 1 & 1\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} x^* = \begin{bmatrix}1 & 1 & 1\end{bmatrix}\begin{bmatrix}29\\30\\31\end{bmatrix}$$

$$3x^* = 29+30+31 \longrightarrow x^* = \frac{29+30+31}{3} = \text{average} !$$ ✓

review of last lecture

- considered $Ax = y$
  - solvability $y \in \mathcal{C}(A)$
  - uniqueness $\mathcal{N}(A) = \{0\}$

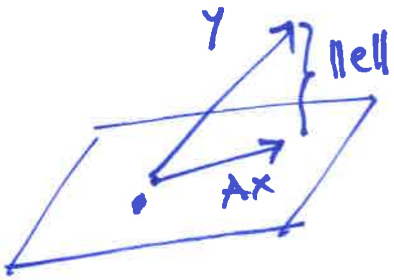


(Ax = y overdetermine
system of equ.)

- what if (A is tall) and $y \notin \mathcal{C}(A)$.
  so that $Ax = y$ has no solution?

↓

then minimize $\|Ax - y\|$
i.e., find x^* such that $\|Ax^* - y\|$ is smallest among all possible $\|Ax - y\|$.



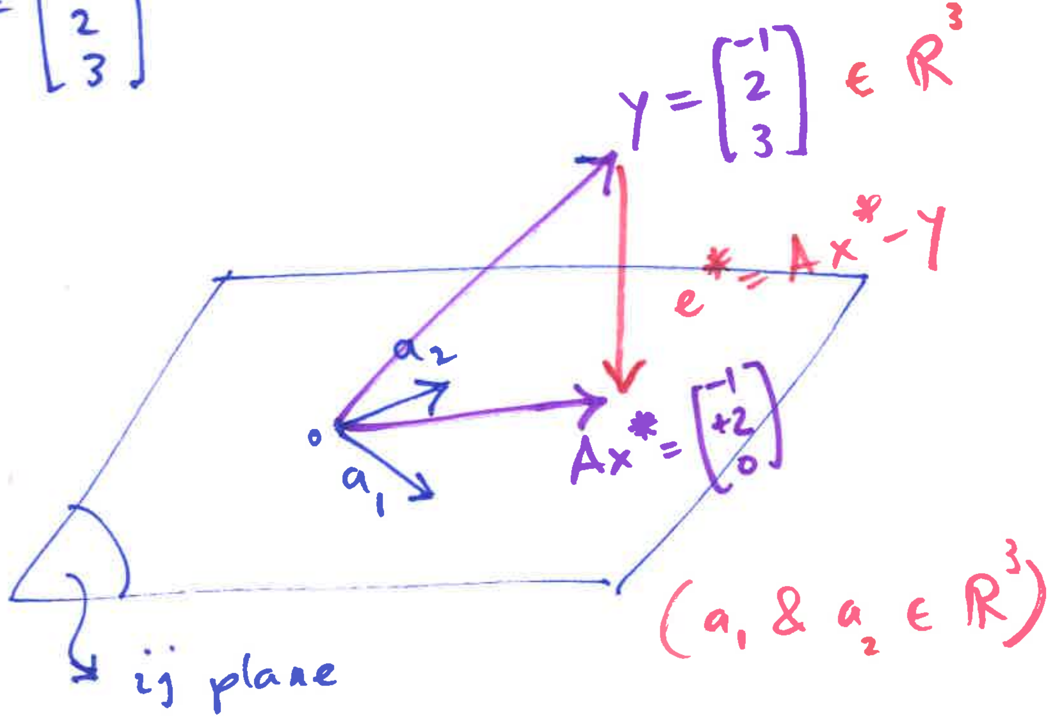
↓

the optimal x^* should be such that
$e^* := Ax^* - y$ is orthogonal to col. space of A.

↓

$$A^T A x^* = A^T y \rightarrow x^* = (A^T A)^{-1} A^T y.$$

- example: $A = \begin{matrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \\ \begin{matrix} a_1 & a_2 \end{matrix} \end{matrix}$, $y = \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix}$



by inspection:

$$Ax^* = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix} = -1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \in \mathbb{R}^3$$

$$x^* = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \in \mathbb{R}^2$$

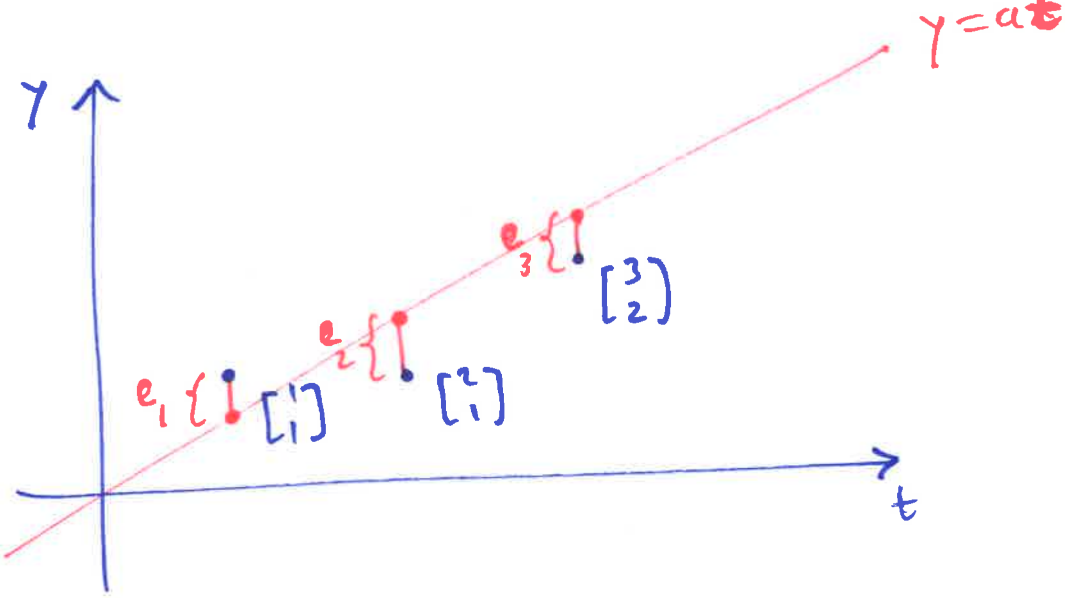
least-squares solution:

$$A^T A x^* = A^T y$$

$$A^T A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A^T y = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

$$\underset{\downarrow \atop I}{\boxed{A^T A}}\, x^* = A^T y \longrightarrow x^* = A^T y = \begin{bmatrix} -1 \\ 2 \end{bmatrix}. \quad ✓$$

- example: find "a" such that the line $y=at$ minimizes $e_1^2+e_2^2+e_3^2$



method 1

$$J(a) = e_1^2+e_2^2+e_3^2$$

$$= \left\| \begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}1\\a.1\end{bmatrix} \right\|^2$$

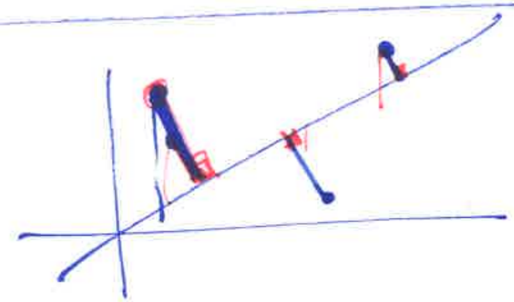
$$+ \left\| \begin{bmatrix}2\\1\end{bmatrix} - \begin{bmatrix}2\\a.2\end{bmatrix} \right\|^2$$

$$+ \left\| \begin{bmatrix}3\\2\end{bmatrix} - \begin{bmatrix}3\\a.3\end{bmatrix} \right\|^2 = (1-a)^2+(1-2a)^2+(2-3a)^2$$

$$\frac{\partial J}{\partial a} = 0 \rightarrow 28a^* - 18 = 0 \rightarrow a^* = \frac{9}{14}.$$

---

aside: $|e_1|+|e_2|+|e_3|.$ or



method 2

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}}_{A} \underbrace{a}_{x} - \underbrace{\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}}_{y}$$

minimize $\|e\|^2 = \|Ax-y\|^2$.

$$a^* = (A^TA)^{-1}A^Ty$$

$$= \left(\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \frac{9}{14}.$$

aside: suppose A were fat

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

$$A^TA = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

↙ rank$(A^TA) = 1$

$$(A^TA)^{-1} \sim \xcancel{\begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} \frac{1}{4-4}}$$

- when A satisfies $N(A)=\{0\}$ (A has lin. ind. col's, & thus A^TA is invertible) then the matrix

$$A^+ := (A^TA)^{-1}A^T$$

is a pseudoinverse of A (also referred to as the left inverse of A)

$$A^+A = (\enclose{circle}{A^TA})^{-1}\enclose{circle}{A^TA}$$

$$= I$$

(AA^+ may or may not be I)

- simple example: $A = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$

$$A^+ = \left(\begin{bmatrix} 2 & 0 \end{bmatrix}\begin{bmatrix} 2 \\ 0 \end{bmatrix}\right)^{-1}\begin{bmatrix} 2 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \end{bmatrix}$$

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

$$AA^+ = \begin{bmatrix} 2 \\ 0 \end{bmatrix}\begin{bmatrix} \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \neq \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

the least-norm problem

- there is a "dual" (mirror image) to the least-squares problem, which is also in the form of an optimization problem

- suppose it is desired to find the vector x of minimum length such that $Ax = y$

$$\begin{array}{ll} \text{minimize} & \|x\|_2^2 \\ \text{subject to} & Ax = y. \end{array}$$

assume that $\mathcal{N}(A) = \{0\}$ (rows of A are lin. ind.)

$$A \sim \Big[\quad \text{fat} \quad \Big]_{m \times n}, \quad m \leq n, \quad \text{rank}(A) = m$$

A has full row rank

---

aside:

① $\mathcal{N}(A) = \{0\} \rightarrow m$ lin. ind. rows $\rightarrow m$ lin. ind. col. $(\in \mathbb{R}^m)$

② $y \in \mathbb{R}^m$

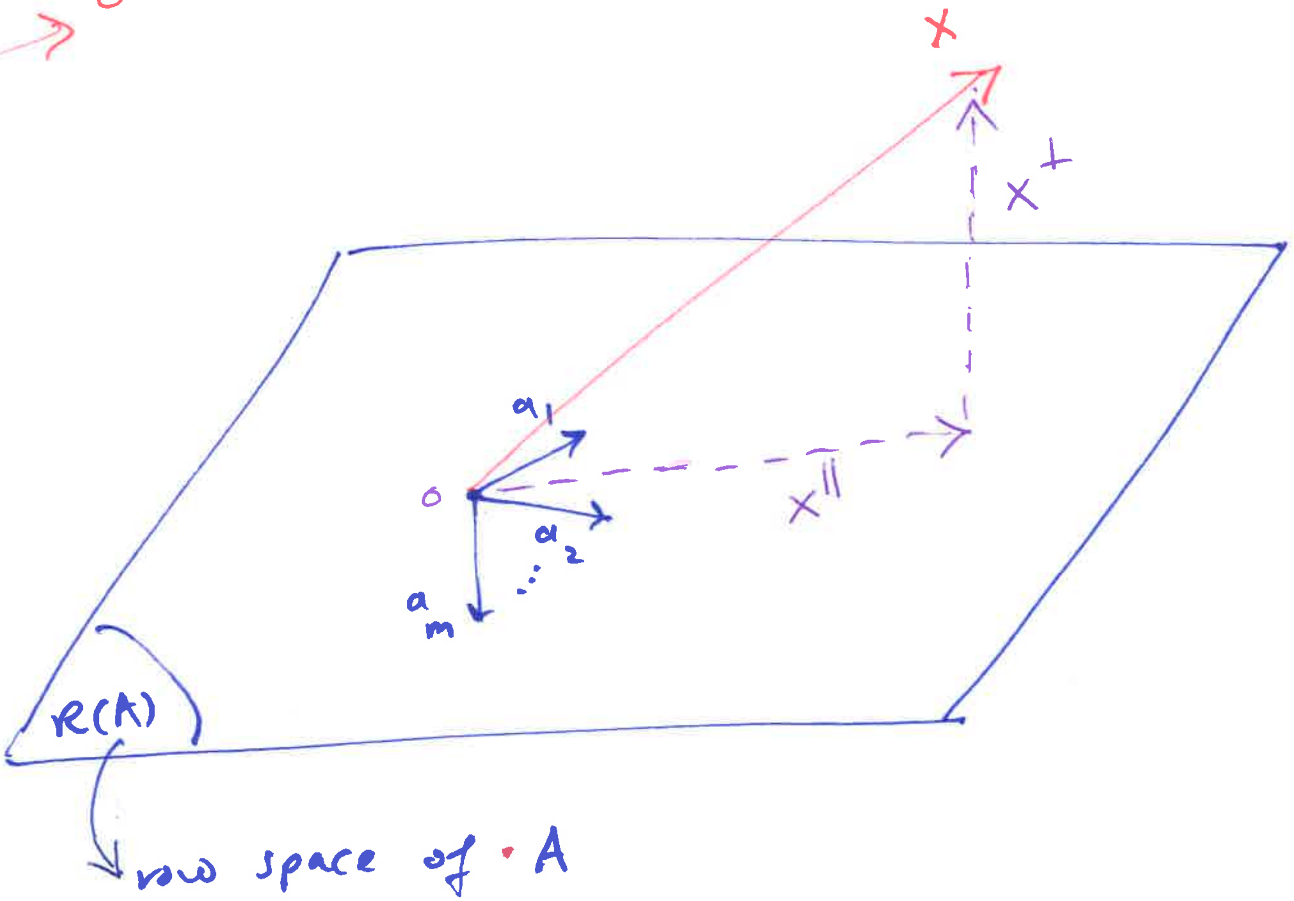
$\}$ ① & ② $\downarrow$ $Ax = y$ always has a solution

$$Ax = y \rightarrow \begin{bmatrix} - & a_1^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ \vdots \\ a_m^T x \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

$$Ax = A(x^{\parallel} + x^{\perp})$$

$$= Ax^{\parallel} + \boxed{Ax^{\perp}} \rightarrow 0$$

$$= y$$



- $Ax^\perp$ is always zero because $x^\perp$ is by definition orthogonal to the rows of A. therefore the component $x^\perp$ does not affect $Ax=y$ and only adds to the length of x.

- therefore the x^* with the minimum norm should be chosen inside the row space of A (i.e., $x^\perp = 0$)

$$x^* = z_1^* a_1 + \cdots + z_m^* a_m = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_m \\ | & & | \end{bmatrix} \begin{bmatrix} z_1^* \\ \vdots \\ z_m^* \end{bmatrix} = A^T z^*$$

in summary $\begin{cases} Ax^* = y \\ x^* = A^T z^* \end{cases} \rightarrow \quad y = Ax^* = A(A^T z^*)$

$$y = \underbrace{AA^T} z^* \longrightarrow z^* = (AA^T)^{-1} y$$

always invertible if A has linearly independent rows. $(\mathcal{N}(A) = \{0\})$

$$\big\downarrow$$

$$x^* = A^T z^*$$

$$\big\downarrow$$

$$\boxed{x^* = A^T (AA^T)^{-1} y}$$

compare with least-squares solution $x^* = (A^TA)^{-1}A^Ty$.

---

aside: AA^T not invertible if A tall

$$A = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \rightarrow AA^T = \begin{bmatrix} 9 & 3 \\ 3 & 1 \end{bmatrix}.$$

- example: $\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}-1\\2\end{bmatrix}.$

only increases length of x without affecting y.

by inspection

$$x^*=\begin{bmatrix}-1\\2\\0\end{bmatrix}$$

$$x^\perp=\begin{bmatrix}0\\0\\\alpha\end{bmatrix},\quad a_1^Tx^\perp=0,\quad a_2^Tx^\perp=0.$$

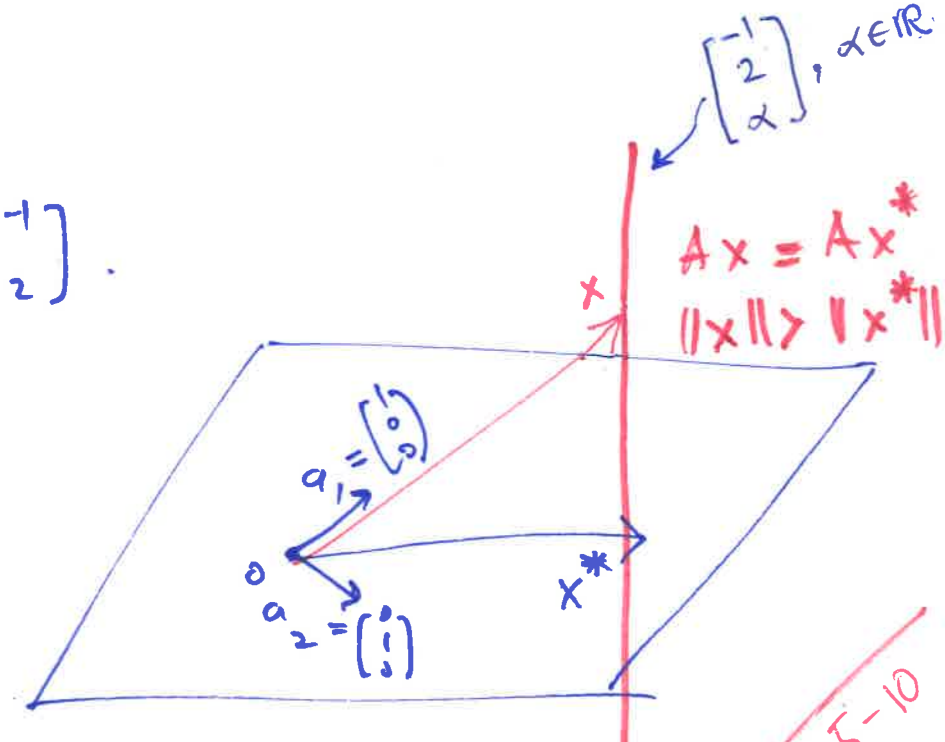
least-norm solution

$$x^*=A^T(AA^T)^{-1}y$$

$$=\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}\overbrace{\left(\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}\right)}^{I}{}^{-1}\begin{bmatrix}-1\\2\end{bmatrix}.$$

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

$$=\begin{bmatrix}-1\\2\\0\end{bmatrix}.$$



- example: $\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} = 1$ find x with minimum length

$$\Updownarrow$$

$$x_1 + x_2 = 1$$

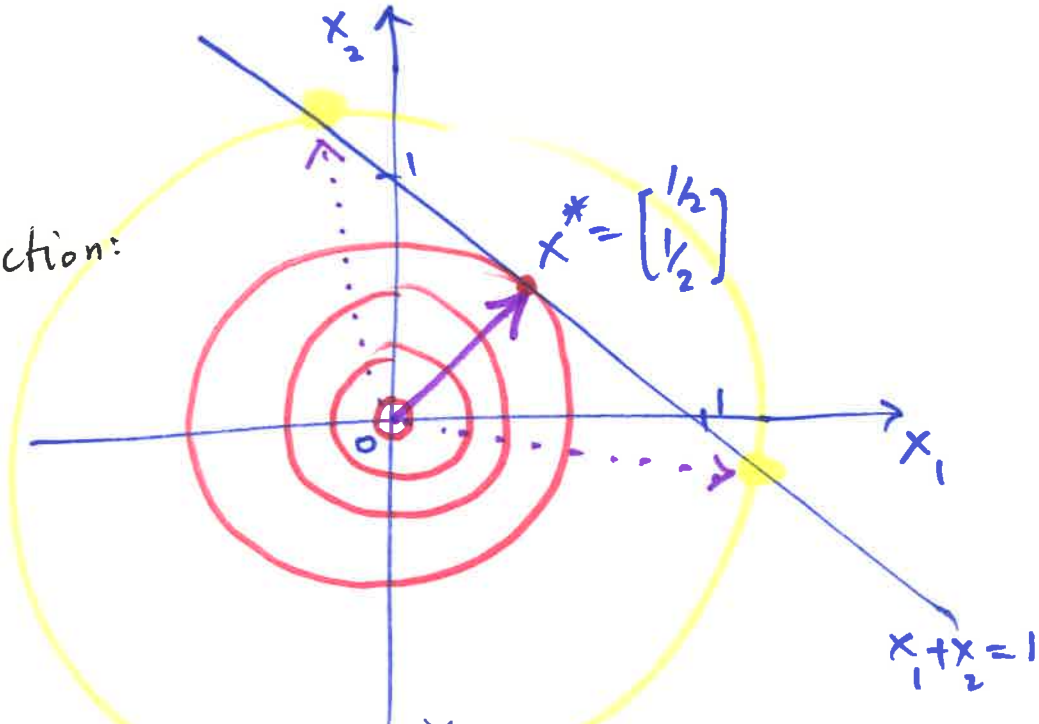
$$A = \begin{bmatrix}1 & 1\end{bmatrix}$$

$$a_1^T = \begin{bmatrix}1 & 1\end{bmatrix}, \quad a_1 = \begin{bmatrix}1\\1\end{bmatrix}$$

$$x^{\parallel} = \begin{bmatrix}\alpha\\\alpha\end{bmatrix}, \quad x^{\perp} = \begin{bmatrix}\beta\\-\beta\end{bmatrix}$$

$$(Ax^{\perp} = 0)$$

by inspection:

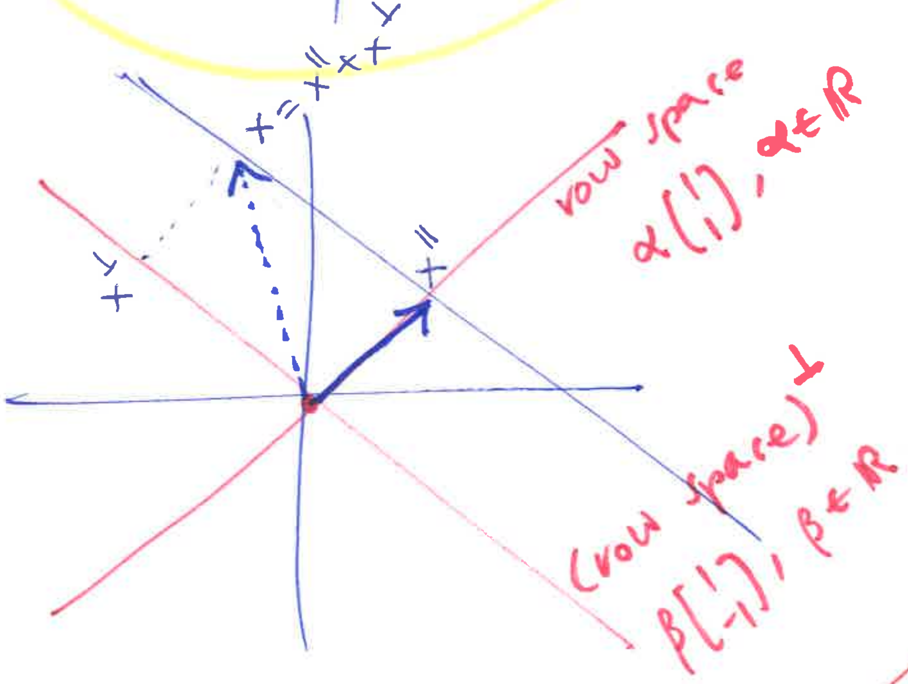


using least-norm:

$$x^* = A^T(AA^T)^{-1}y$$

$$= \begin{bmatrix}1\\1\end{bmatrix}\left(\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix}\right)^{-1} \cdot 1$$

$$= \frac{1}{2}\begin{bmatrix}1\\1\end{bmatrix}.$$



- when A satisfies $x(A) = \{0\}$ (i.e., A has lin. ind. rows & thus AA^T is invertible)

$$A^+ := A^T(AA^T)^{-1}$$

is a pseudoinverse of A.

(also referred to as the right inverse of A)

$$AA^+ = AA^T(AA^T)^{-1}$$
$$= I$$

(A^+A may or may not be I)

- simple example: $A = [2 \;\; 0] \rightarrow A^+ = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix}$.

$$AA^+ = [2 \;\; 0]\begin{bmatrix} 1/2 \\ 0 \end{bmatrix} = 1$$

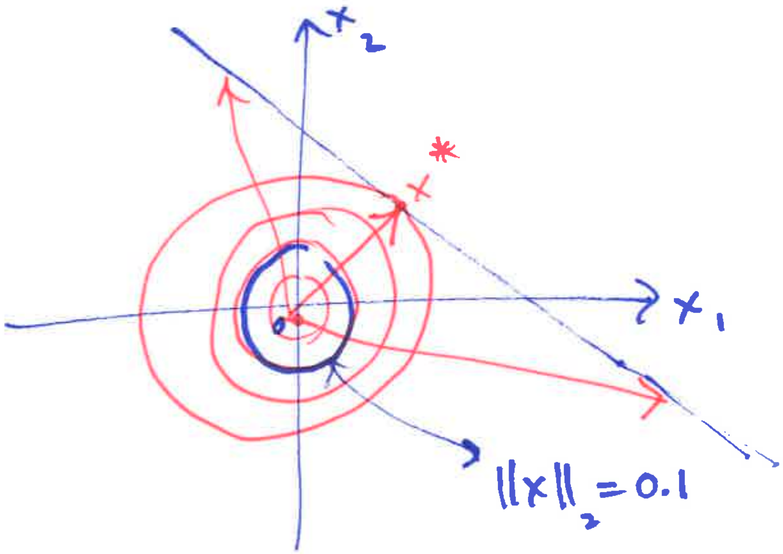
$$A^+A = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix}[2 \;\; 0] = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \neq \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

A function $f : \mathbf{R}^n \to \mathbf{R}$ with $\mathbf{dom}\, f = \mathbf{R}^n$ is called a *norm* if

- f is nonnegative: $f(x) \geq 0$ for all $x \in \mathbf{R}^n$
- f is definite: $f(x) = 0$ only if $x = 0$
- f is homogeneous: $f(tx) = |t| f(x)$, for all $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$
- f satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y)$, for all $x, y \in \mathbf{R}^n$

min. $\|x\|_2 := \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

s.t. $x_1 + x_2 = 1 \rightarrow Ax = y$

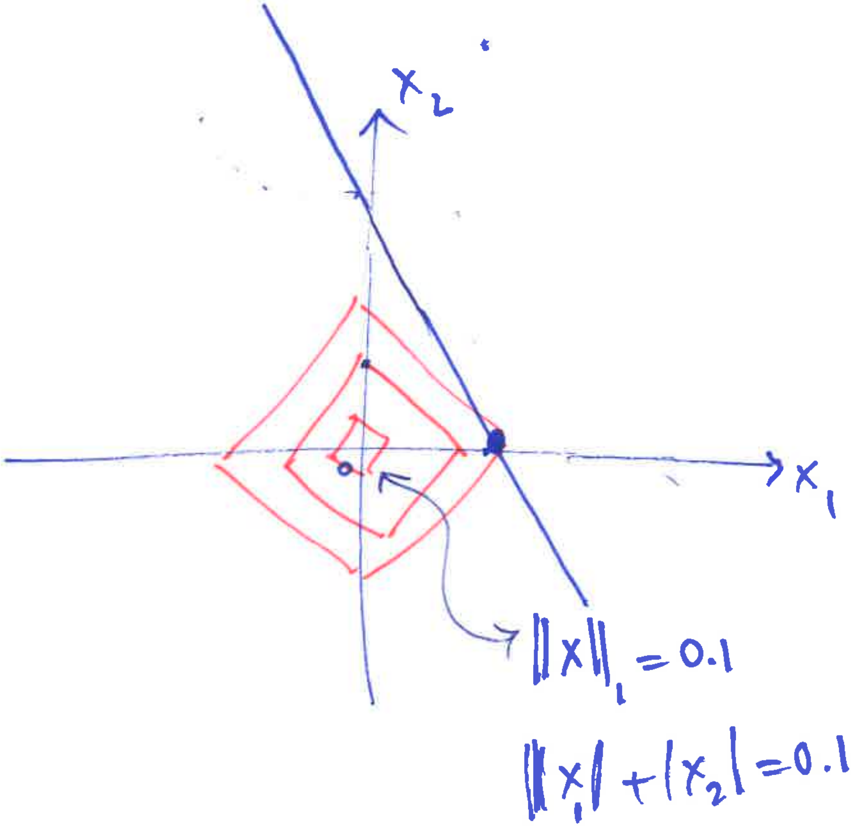


min. $\|x\|_1 := |x_1| + |x_2| + \cdots + |x_n|$

s.t. $Ax = y$

min. $\|x\|_1$

s.t. $2x_1 + x_2 = 1$



$$\|x\|_p = \sqrt[p]{|x_1|^p + \cdots + |x_n|^p}$$

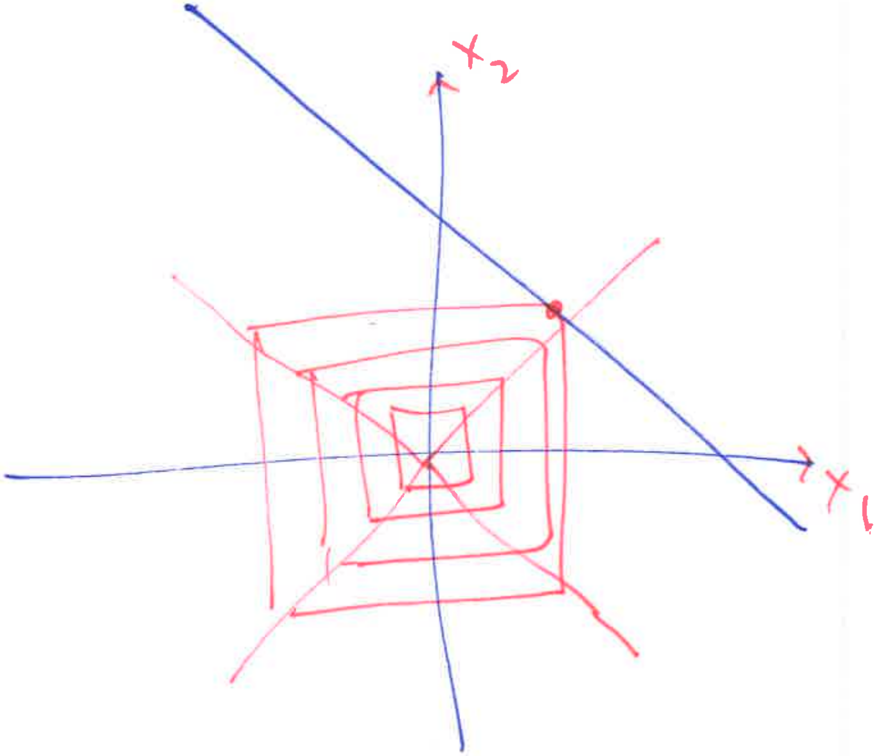
$$= \left(|x_1|^p + \cdots + |x_n|^p \right)^{1/p}$$

$p \to \infty$

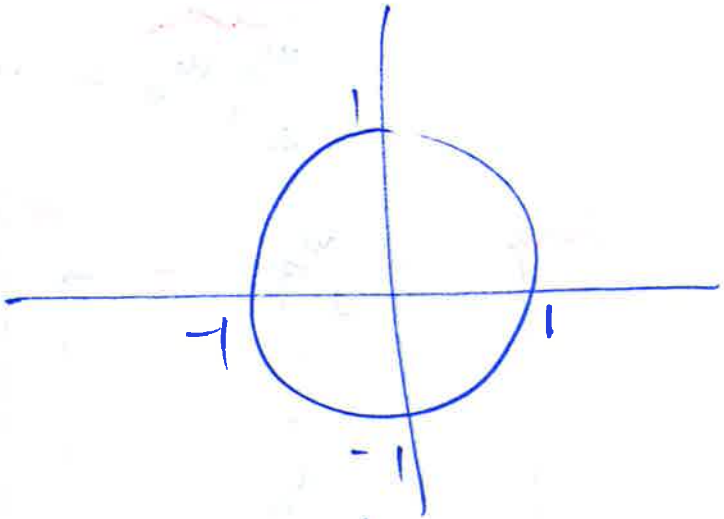
$$= \max_i \{ |x_i| \}$$

$$\|x\|_\infty = \max_i \{ |x_i| \}.$$

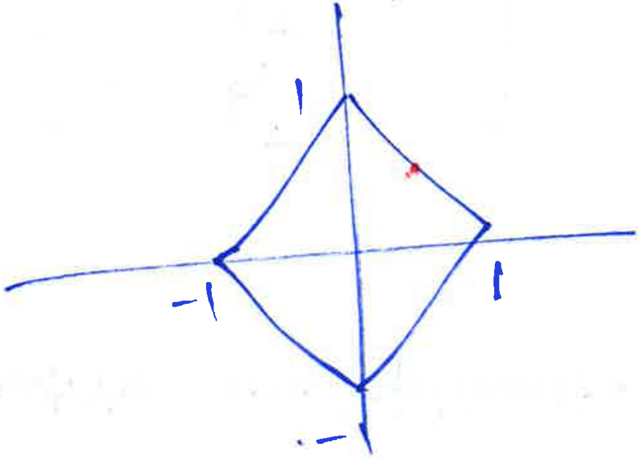
$x = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$ $x = \begin{bmatrix} -0.5 \\ -0.7 \end{bmatrix}$



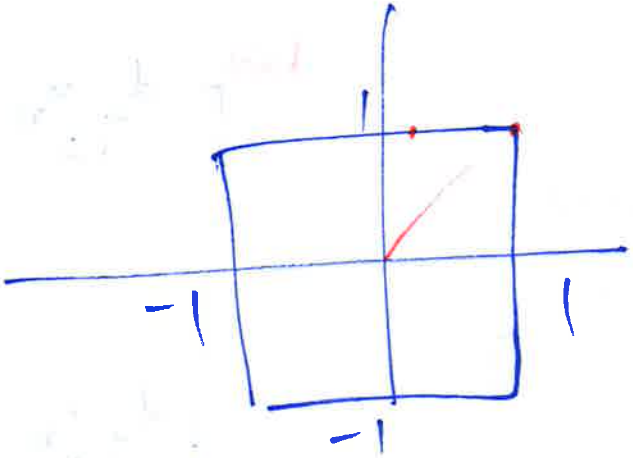
$$\underbrace{\|x\|_2}_{x^T x = 1} = 1$$



$$\|x\|_1 = 1$$



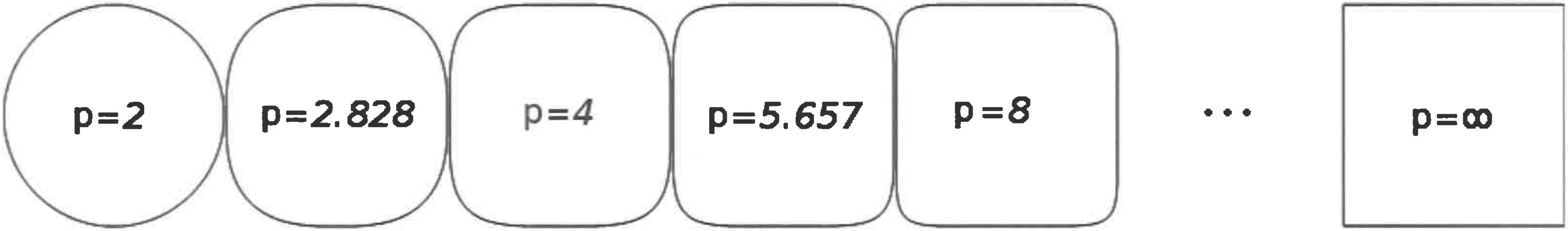
$$\|x\|_\infty = 1$$

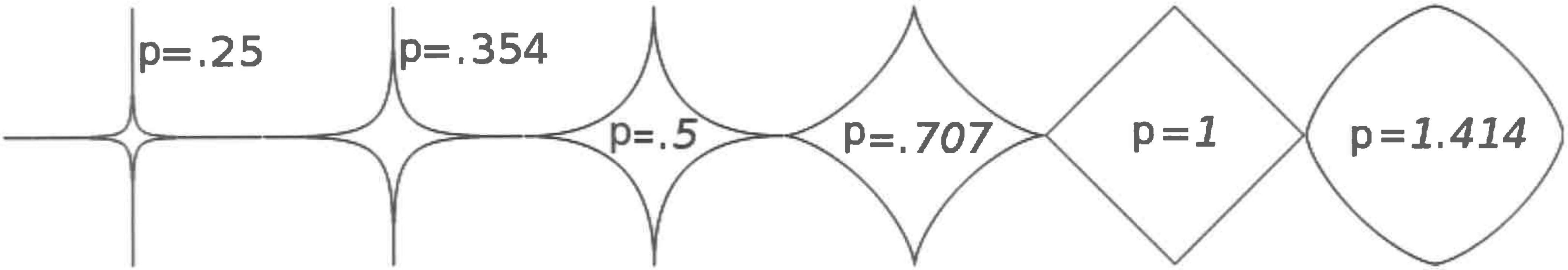


$$\|x\|_p = 1$$

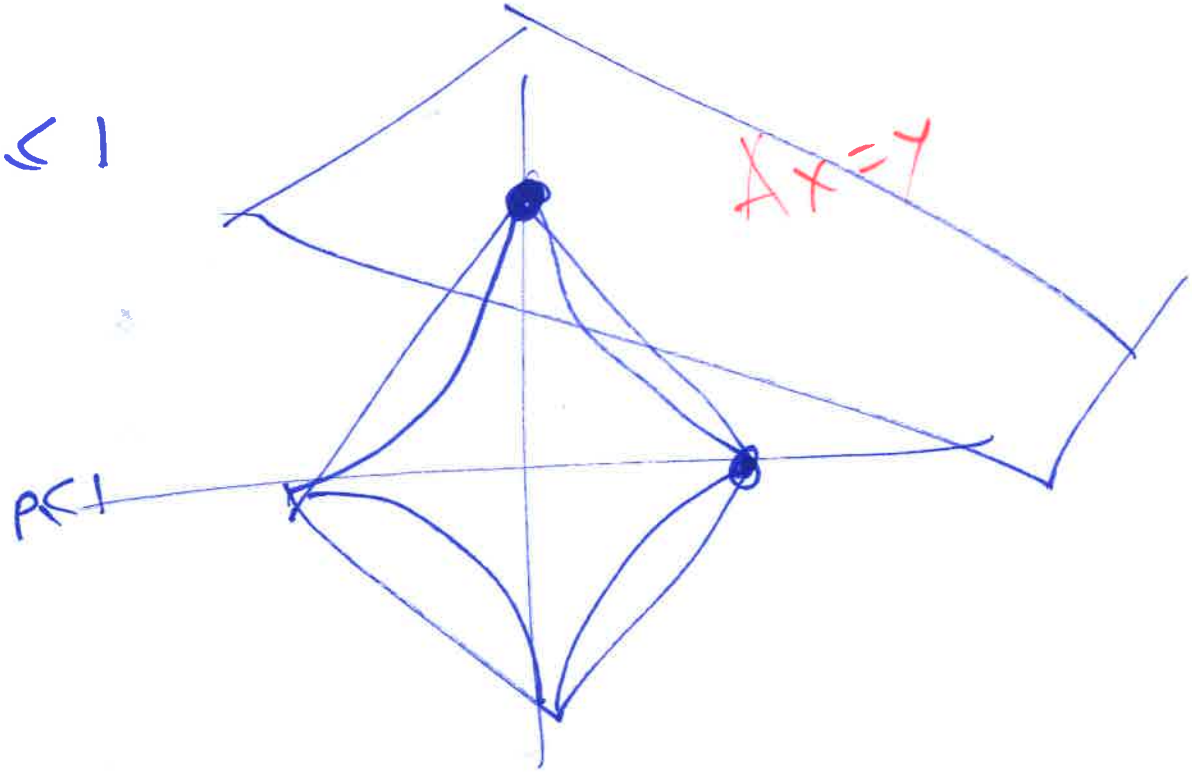
p-norm of vector x:

$$\|x\|_p = \sqrt[p]{|x_1|^p + \cdots + |x_n|^p}, \qquad 0 < p \leqslant \infty$$



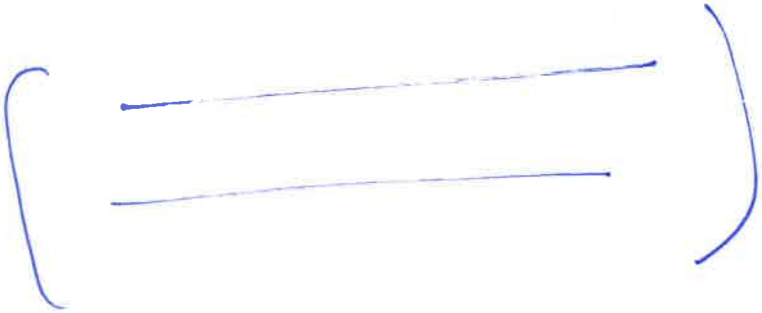

p=.25
p=.354
p=.5
p=.707
p=1
p=1.414

$\|x\|_p$ $\quad p \leq 1$



min. $\|x\|_p$

s.t. $Ax=y$



$$x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

review of last lecture

- finding solution to $Ax=y$ when there are many

minimize $\|x\|_2^2 = \sum_{i=1}^{n} x_i^2$ $\overset{?}{\longrightarrow}$ $\|x\|_p^p = \sum_{i=1}^{n} |x_i|^p$

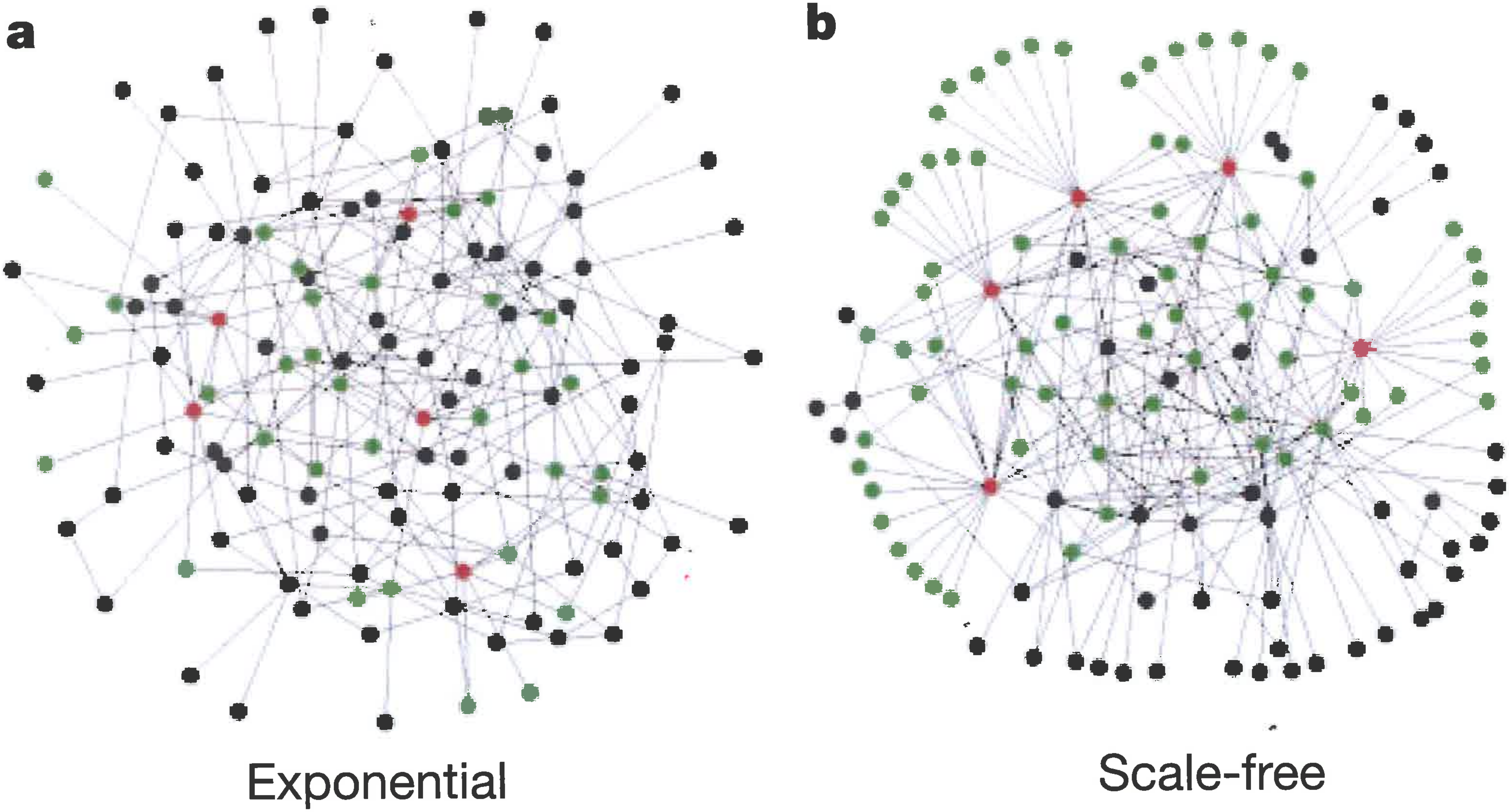
such that $Ax=y$

least-norm problem

$$\downarrow$$

$$x^* = A^T(AA^T)^{-1}y$$

- eigenvalues to analyze/design networks
  - nodes: agents — links: communication or physical interact.



- network → graph characterization → matrix representation
  → eigenvalues & eigenvectors

eigenvalue / eigenvector decomposition (spectral decomposition)

- motivating examples.

– dynamical systems $G(s) = \frac{p(s)}{q(s)} \leftrightarrow \dot{x} = Ax + Bu, \quad y = Cx$

feedback control : move poles of $G(s)$ $\leftrightarrow$ move eigenvalues of A.

$\left\{ \begin{array}{l} \text{transient response} \\ \text{steady-state response} \\ \text{robustness.} \end{array} \right.$

– network analysis.

network $\longrightarrow$ graph characterization

$\longrightarrow$ matrix representation

$\longrightarrow$ eigenvalues & eigenvectors of matrix.

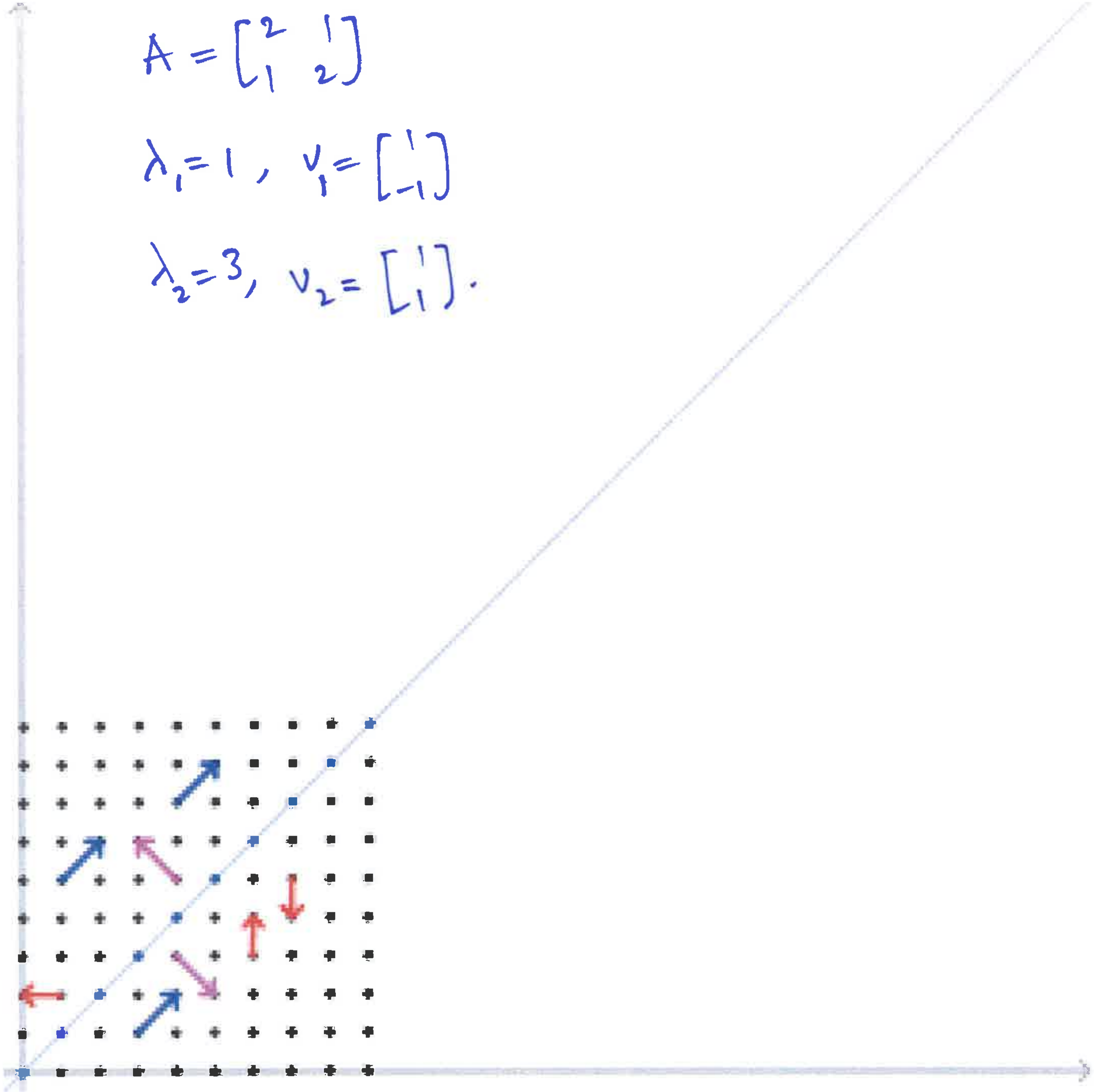
some network examples

- internet / w.w.w
- social networks / consensus networks
- power networks
- sensor networks

– search engines (Google / Yahoo)

– image & audio compression

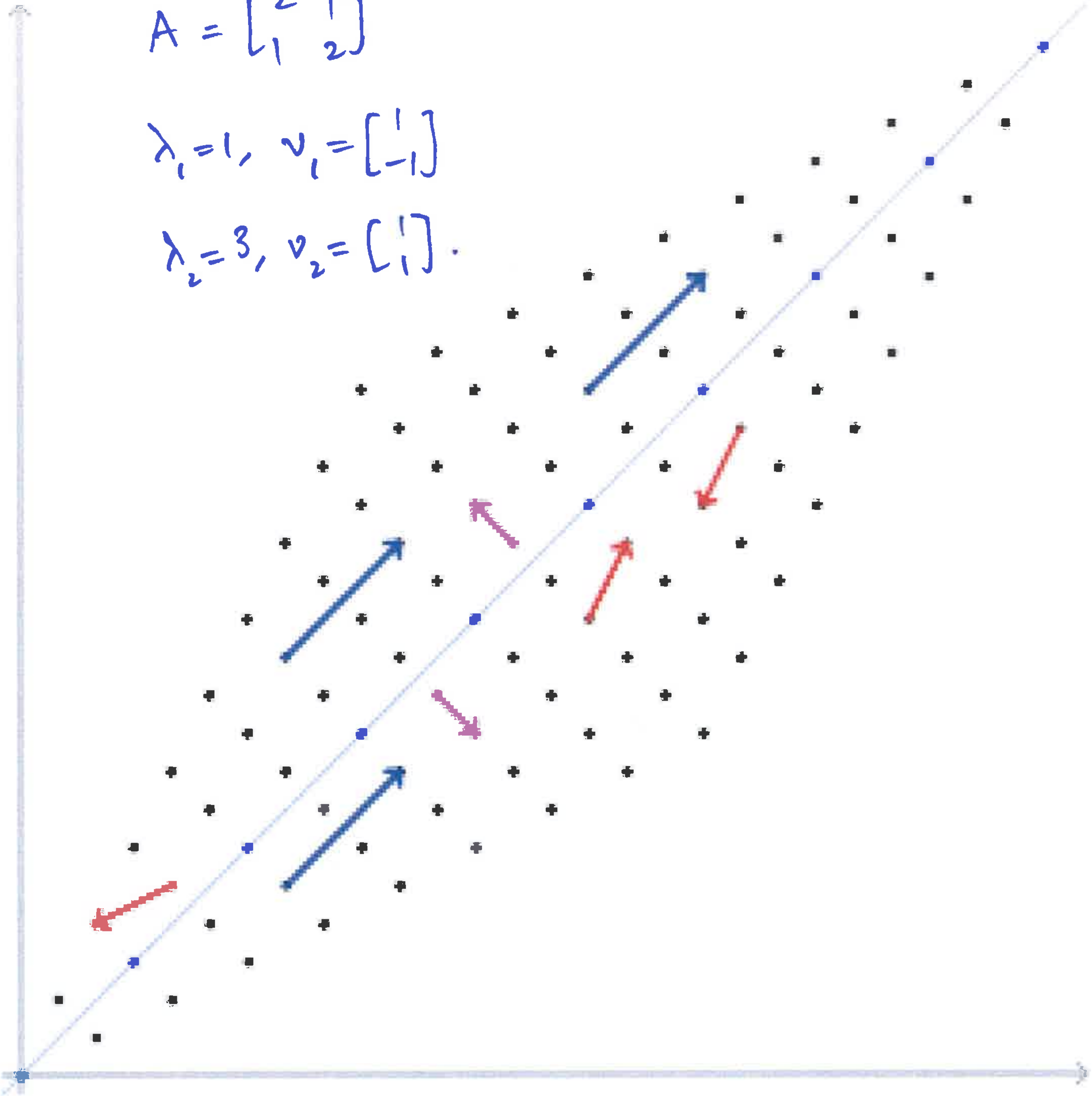
⋮



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

$$\lambda_1 = 1, \quad v_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\lambda_2 = 3, \quad v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$



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

$$\lambda_1 = 1, \ v_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\lambda_2 = 3, \ v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

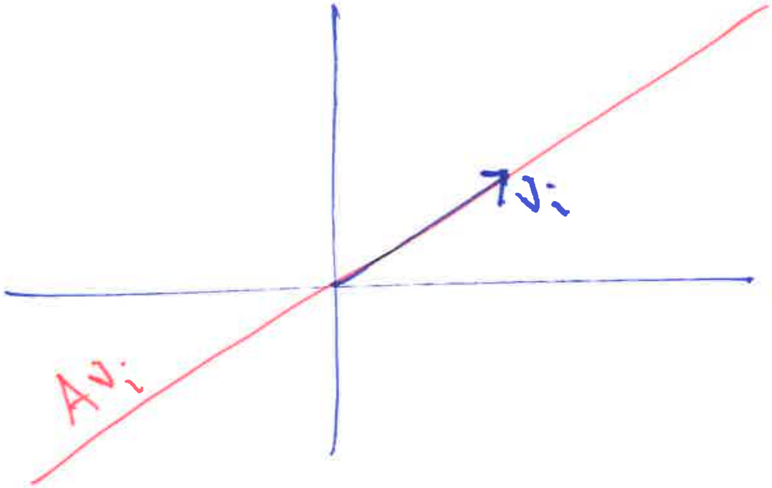
eigenvectors, eigenvalues, & diagonalization

- let A be a square $n \times n$ matrix

  then the vectors $\{v_1, \dots, v_n\}$ and the scalars $\{\lambda_1, \dots, \lambda_n\}$ are called the "eigenvectors" & "eigenvalues" of A if

$$\begin{aligned} Av_1 &= \lambda_1 v_1 \\ &\vdots \\ Av_n &= \lambda_n v_n \end{aligned} \qquad (v_i \neq 0)$$

- this means that the "alignment" of eigenvectors remains "invariant" under multiplication by A



aside:

$$Av = \lambda v$$

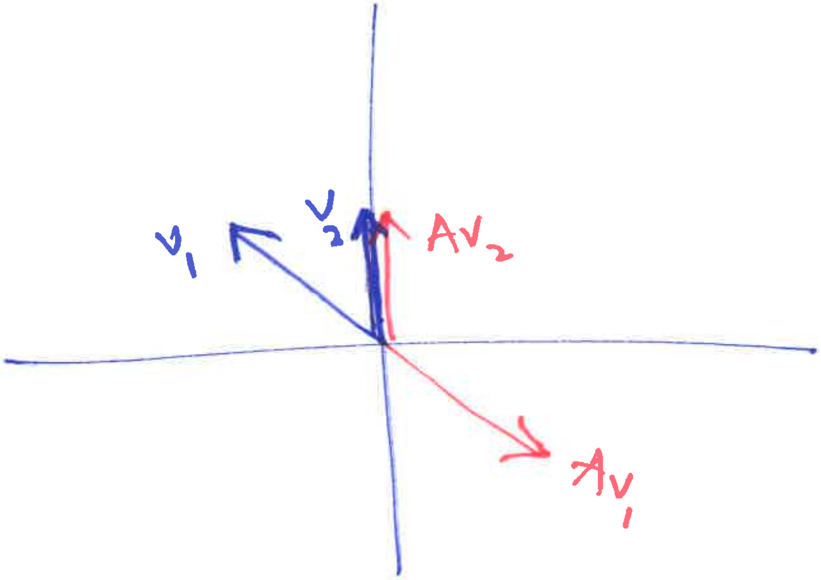
$$A(\alpha v) = \alpha(Av)$$

$$= \alpha(\lambda v) = \lambda(\alpha v)$$

- example: $A = \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix}$ $\quad \begin{cases} \lambda_1 = -1 \\ v_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \end{cases}$ $\quad \begin{cases} \lambda_2 = 1 \\ v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{cases}$

$$Av_1 = \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} -1 \\ 1 \end{bmatrix}}_{v_1} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \underbrace{(-1)}_{\lambda_1} v_1$$

$$Av_2 = \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{v_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \underbrace{(1)}_{\lambda_2} v_2$$



- computing eigenvalues & eigenvectors

$$\underset{(v \neq 0)}{Av = \lambda v} \longrightarrow (\lambda I - A)\, v = 0 \quad ①$$

$$\downarrow$$

$$v \in N(\lambda I - A)$$

② $\det(M) = 0 \longleftarrow$ if $N(M) \neq \{0\}$

① & ② → if λ is an eigenvalue of A, then

$$\det(\lambda I - A) = 0$$

this gives a procedure for finding the eigenvalues of a matrix.

aside:

$Av - \lambda v = 0 \longrightarrow (A - \lambda)v = 0$ ✗

$Av - \lambda I v = 0 \rightarrow (A - \lambda I)v = 0$ ✓

aside:

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

- the eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ are solutions of the polynomial equation ($A : n \times n$)

$$\det(sI - A) = 0$$ (called "characteristic" eqn)

this gives an n^{th} order polynomial in "s" with n roots (perhaps repeated and/or complex-valued).

(the roots give all λ's such that $\lambda I - A$ has nontrivial null space) → helps in finding eigenvectors.

- example: $A = \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix}$.

$$sI - A = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} - \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} s+1 & 0 \\ -2 & s-1 \end{bmatrix}.$$

$$\det(sI - A) = (s+1)(s-1) - \cancel{(0)(-2)}^{\,0}$$

← 2nd order polynomial in "s".

$$= (s+1)(s-1)$$

$$= 0 \quad \longrightarrow \quad \begin{cases} \lambda_1 = -1 \\ \lambda_2 = 1 \end{cases}$$

$A v_i = \lambda_i v_i \rightarrow (\lambda_i I - A) v_i = 0$ $\quad$ v_i is in the null space of $\lambda_i I - A$.

$\lambda_1 = -1 \rightarrow \lambda_1 I - A = \begin{bmatrix} 0 & 0 \\ -2 & -2 \end{bmatrix}$

$(\lambda_1 I - A) v_1 = 0 \longrightarrow \begin{bmatrix} 0 & 0 \\ -2 & -2 \end{bmatrix} v_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \longrightarrow v_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$

$\lambda_2 = 1 \rightarrow \lambda_2 I - A = \begin{bmatrix} 2 & 0 \\ -2 & 0 \end{bmatrix}$

$(\lambda_2 I - A) v_2 = 0 \longrightarrow \begin{bmatrix} 2 & 0 \\ -2 & 0 \end{bmatrix} v_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \longrightarrow v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

<u>aside</u>:

$(\lambda I - A) v = 0$

$(\lambda I - A) \alpha v = 0$

- a procedure for finding all eigenvalue/eigenvector pairs $\{(\lambda_1, v_1), \dots, (\lambda_n, v_n)\}$ of an $n \times n$ matrix A is to

① compute the eigenvalues by finding the roots of

$$\det(\lambda I - A) = 0 \longrightarrow \begin{cases} \lambda_1 \\ \vdots \\ \lambda_n \end{cases}$$

② find a vector v_i such that $Av_i = \lambda_i v_i$ (or equivalently $(\lambda_i I - A)\, v_i = 0$). repeat for $i = 1, \dots, n$.

<u>aside</u>: this is <u>not</u> how λ_i & v_i are actually computed for large matrices; need iterative methods

- note that v_i are not unique

  * if v_i is an eigenvector then so is αv_i

  * $A = \alpha \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. $\quad Av = (\alpha I) v = \alpha v$ (will come back to this later)

<u>aside</u>:

$$A = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \quad v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \dots, v_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

$$A = \begin{bmatrix} \lambda_1 & \cdots & \cdots \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{bmatrix}, \quad \begin{bmatrix} \lambda_1 & & 0 \\ \vdots & \ddots & \\ \cdots & \cdots & \lambda_n \end{bmatrix}$$

6-9

diagonalization (using eigenvalues/eigenvectors)

- the equations $Av_i = \lambda_i v_i$, $i = 1, \cdots, n$ can be equivalently rewritten as

$$\underbrace{A \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}}_{T} = \underbrace{\begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}}_{\Lambda}$$

$$\left\{ \begin{array}{l} Av_1 = \lambda_1 v_1 \\ \vdots \\ Av_n = \lambda_n v_n \end{array} \right. \iff AT = T\Lambda$$

left hand side:

$$A\begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ Av_1 & \cdots & Av_n \\ | & & | \end{bmatrix}$$ ①

right hand side:

$$\begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}\begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

$$= \left[\begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}\begin{bmatrix} \lambda_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}\begin{bmatrix} 0 \\ \lambda_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \cdots \quad \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_n \end{bmatrix} \right]$$

$$= \begin{bmatrix} | & & | \\ \lambda_1 v_1 & \cdots & \lambda_n v_n \\ | & & | \end{bmatrix}$$ ②

① = ② $\longleftrightarrow$ $\begin{cases} Av_1 = \lambda_1 v_1 \\ \vdots \\ Av_n = \lambda_n v_n \end{cases}$

$$\left.\begin{aligned} T &= \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} \\ \Lambda &= \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \end{aligned}\right\} \rightarrow \boxed{AT = T\Lambda}$$

the diagonal entries of Λ also referred to as the spectrum of A.

T is an $n \times n$ matrix and under certain conditions (that we will talk about shortly) is invertible.

$$AT = T\Lambda \rightarrow \boxed{T^{-1}AT = \Lambda}$$

$$\searrow \boxed{A = T\Lambda T^{-1}}$$

(hence the term "diagonalization")

- invertibility of T

if all eigenvalues of A are distinct (A has no repeated eigenvalues), or, if A is a symmetric matrix ($A^T = A$)

$$\downarrow$$

A has n eigenvectors that can be chosen to be linearly independent

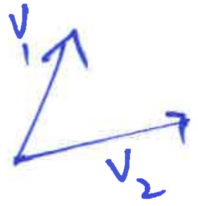
$$\downarrow$$

$$T = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} \text{ is invertible.}$$

- simple example:

$A = \alpha I$. $\left\{ \begin{array}{l} \lambda_1 = \alpha \\ \lambda_2 = \alpha \end{array} \right.$

$$\begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}$$



any pair of lin. ind. vec. will do.

review of last lecture

- eigenvalue/eigenvector decomposition
  - find eigenvalues from $\underbrace{\det(\lambda I - A) = 0}_{\text{characteristic equ. (n^{th} order polynomial)}} \longrightarrow \underbrace{\{\lambda_1, \dots, \lambda_n\}}_{\text{can be complex and/or repeated}}$
  - find eigenvectors from $(\lambda_i I - A) v_i = 0 \longrightarrow \{v_1, \dots, v_n\}$
  - (λ_i, v_i) satisfy $A v_i = \lambda_i v_i \;, \quad i = 1, \dots, n$

$$\Updownarrow$$

$$\boxed{AT = T\Lambda}$$

$$T = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} \rightsquigarrow T^{-1}$$

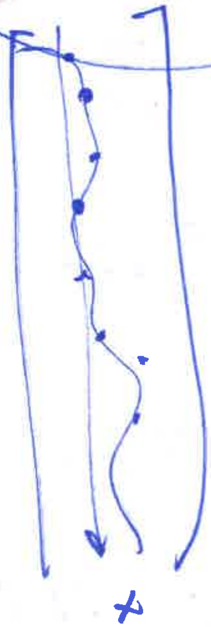
$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

hence "diagonalization" $\longleftarrow$ $\left\{ \begin{aligned} A &= T \Lambda T^{-1} \\ \Lambda &= T^{-1} A T \end{aligned} \right.$

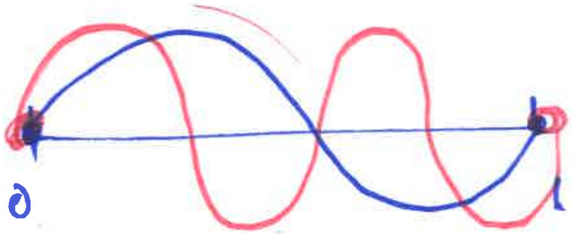
$\frac{d^2}{dx^2}\psi(x)$

$$\underbrace{A}\ \underbrace{v}$$

$$\frac{d^2}{dx^2}\psi(x) = \underbrace{\lambda}\underline{\psi}(x).$$



$\psi($



$$\underbrace{\psi(0)=0} \qquad \underbrace{\psi(1)=1}$$

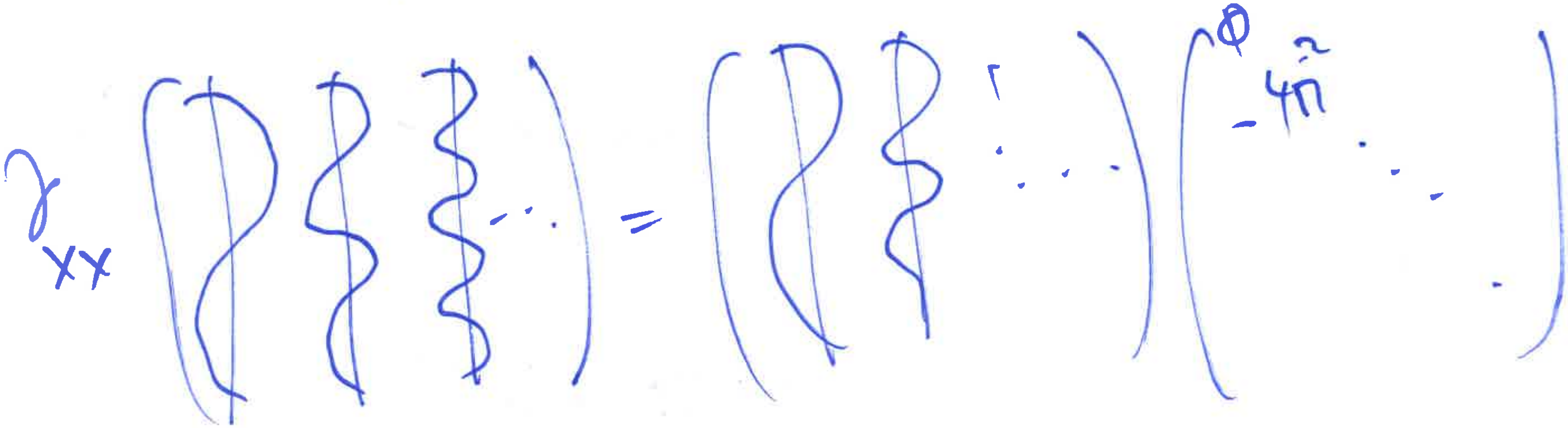
$$\left\{-4\pi^2, -16\pi^2, \ldots, -4.9\pi^2\right\}$$

$$\psi(x) = \sin(2k\pi x)$$

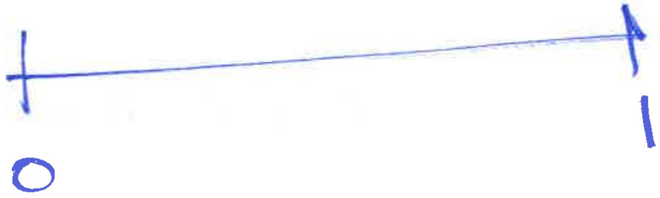
$$\psi'(x) = 2k\pi\cos(2k\pi x)$$

$$\psi''(x) = -(4k^2\pi^2)\sin(2k\pi x) = -4\pi^2k^2\psi(x)$$

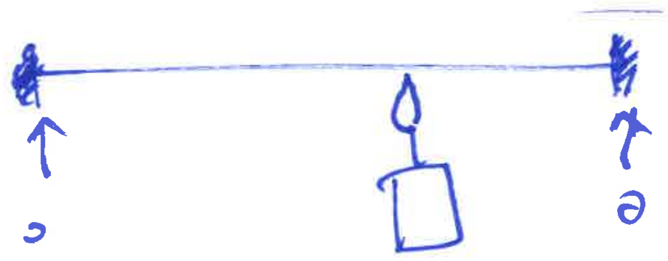
$A \sim \partial_{xx}$



$$\sum c_k \sin(2\pi k x).$$



$\psi(1) = \psi(0) = 0$



$$\ddot{x}(t) + \alpha \dot{x}(t) + f x(t) = g(t)$$

$$A x_p = Y$$

$$A(x_p + \alpha \sum x_h) = Y$$

$$A x_p = Y$$

$$Ax = y$$

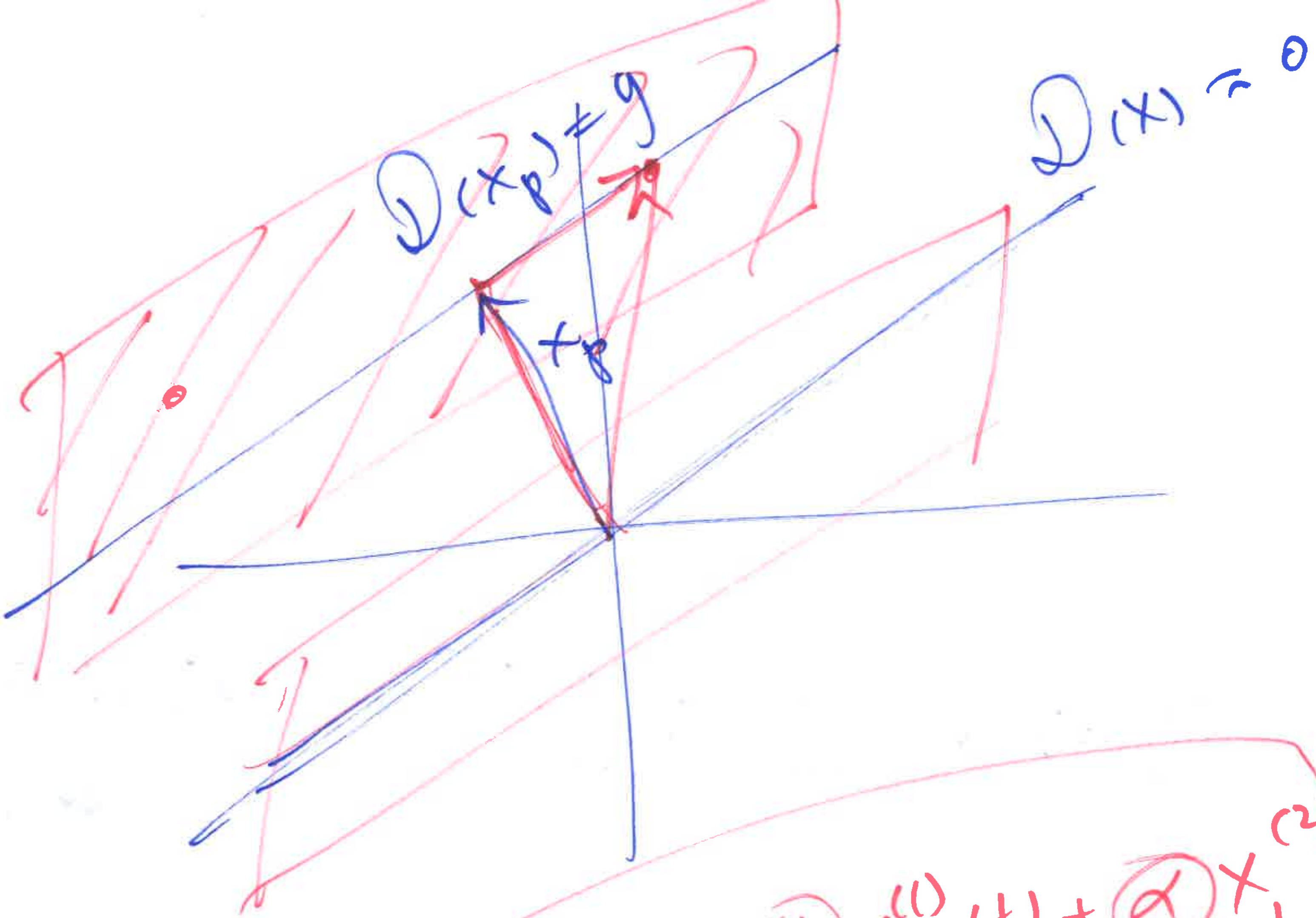
$$x_p + \sum_i \alpha_i x_h^{(i)}$$

$$s^2 + \alpha s + \beta = 0 \longrightarrow s = \begin{cases} c_1 \\ c_2 \end{cases}$$

$$\alpha_1 e^{c_1 t} + \alpha_2 e^{c_2 t}.$$

$$\underline{\mathcal{D}(x)} = \ddot{x} + \alpha \dot{x} + \beta x = 0$$

$$x(t) = x_p + \underbrace{\alpha_1 e^{c_1 t} + \alpha_2 e^{c_2 t}}_{\sum x_h^{(i)}}.$$



$$\boxed{x(t) = x_p(t) + \alpha_1 x_h^{(1)}(t) + \alpha_2 x_h^{(2)}}$$

$x(0)$ $x'(0)$ . . .

- diagonalization is useful in finding higher powers of A

$$AT = T\Lambda \rightarrow A = T\Lambda T^{-1}$$

$$A^m = \underbrace{A \cdots A}_{m \text{ times}} = \underbrace{(T\Lambda \cancel{T^{-1}})(\cancel{T}\Lambda \cancel{T^{-1}}) \cdots (\cancel{T}\Lambda \cancel{T^{-1}})(\cancel{T}\Lambda T^{-1})}_{m \text{ times}}$$

$$= T\Lambda^m T^{-1}$$

$$\Lambda^m = \begin{bmatrix} \lambda_1^m & & \\ & \ddots & \\ & & \lambda_n^m \end{bmatrix}$$

- although computing $\lambda_1, \ldots, \lambda_n$ & $v_1, \ldots, v_n$ is difficult, it only needs to be done once; then A^m can be easily computed for any m. (same applies to any function of A).

- many problems of theoretical interest can be solved using diagonalization (without actual computation of λ_i & v_i)

- application: finding e^A and the "state transition matrix".

$$e^A := I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots$$

$$= T I T^{-1} + T\Lambda T^{-1} + \frac{1}{2!}(T\Lambda \cancel{T^{-1}})(\cancel{T}\Lambda T^{-1}) + \frac{1}{3!}(\quad \cancel{)(} \quad \cancel{)(} \quad) + \cdots$$

$$= T I T^{-1} + T\Lambda T^{-1} + \frac{1}{2!}T\Lambda^2 T^{-1} + \frac{1}{3!}T\Lambda^3 T^{-1} + \cdots$$

$$= T\left(I + \Lambda + \frac{1}{2!}\Lambda^2 + \cdots\right)T^{-1}$$

$$= T \begin{bmatrix} e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_n} \end{bmatrix} T^{-1} = T e^{\Lambda} T^{-1}.$$

trivial to compute

aside:

$$f(A) = T f(\Lambda) T^{-1}$$
$$= T \begin{bmatrix} f(\lambda) & & \\ & \ddots & \\ & & f(\lambda) \end{bmatrix} T^{-1}$$

& similarly $e^{At} = T e^{\Lambda t} T^{-1}$.

state transition matrix corresponding to the differential equ. $\dot{x} = Ax$.

x^3

$$e^{\Lambda} = I + \Lambda + \frac{1}{2!}\Lambda^2 + \frac{1}{3!}\Lambda^3 + \cdots$$

$$= \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} + \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} + \frac{1}{2!}\begin{bmatrix} \lambda_1^2 & & \\ & \ddots & \\ & & \lambda_n^2 \end{bmatrix}$$

$$+ \frac{1}{3!}\begin{bmatrix} \lambda_1^3 & & \\ & \ddots & \\ & & \lambda_n^3 \end{bmatrix} + \cdots$$

$$= \begin{bmatrix} 1+\lambda_1+\frac{1}{2!}\lambda_1^2+\frac{1}{3!}\lambda_1^3 \longrightarrow e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_n} \longleftarrow 1+\lambda_n+\frac{1}{2!}\lambda_n^2+\frac{1}{3!}\lambda_n^3+\cdots \end{bmatrix}$$

- example: $A = \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix}$ find e^{At}.

$v_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, $\lambda_1 = -1$ $\qquad$ $v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\lambda_2 = 1$

$$T = \begin{bmatrix} \underset{\substack{\uparrow \\ v_1}}{\boxed{\begin{matrix} -1 \\ 1 \end{matrix}}} & \underset{\substack{\uparrow \\ v_2}}{\boxed{\begin{matrix} 0 \\ 1 \end{matrix}}} \end{bmatrix} \longrightarrow T^{-1} = \begin{bmatrix} -1 & 0 \\ 1 & 1 \end{bmatrix}.$$

$$\Lambda = \begin{bmatrix} \underset{\lambda_1}{\boxed{-1}} & 0 \\ 0 & \underset{\lambda_2}{\boxed{1}} \end{bmatrix}, \qquad e^{\Lambda} = \begin{bmatrix} e^{-1} & 0 \\ 0 & e^{1} \end{bmatrix}, \qquad e^{\Lambda t} = \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{t} \end{bmatrix}.$$

$$e^{At} = T e^{\Lambda t} T^{-1}$$

$$= \begin{bmatrix} -1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{t} \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} e^{-t} & 0 \\ e^{t} - e^{-t} & e^{t} \end{bmatrix}.$$

<u>aside</u>:

$$e^{A+B} \neq e^{A} e^{B}$$

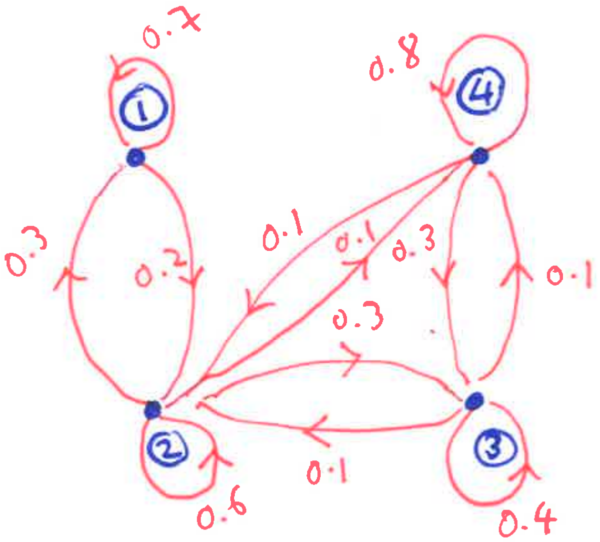
unless $AB = BA$.

$$\dot{x} = Ax$$

- application: propagation of beliefs in a social network

$$x(k) = A\, x(k+1)$$

$$A = \begin{bmatrix} 0.7 & 0.3 & 0 & 0 \\ 0.2 & 0.6 & 0.1 & 0.1 \\ 0 & 0.3 & 0.4 & 0.3 \\ 0 & 0.1 & 0.1 & 0.8 \end{bmatrix}, \quad x(k) = \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \end{bmatrix}$$



the entries of A are positive, between zero & one, and every row of A sums to one

(A is called a "stochastic" matrix)

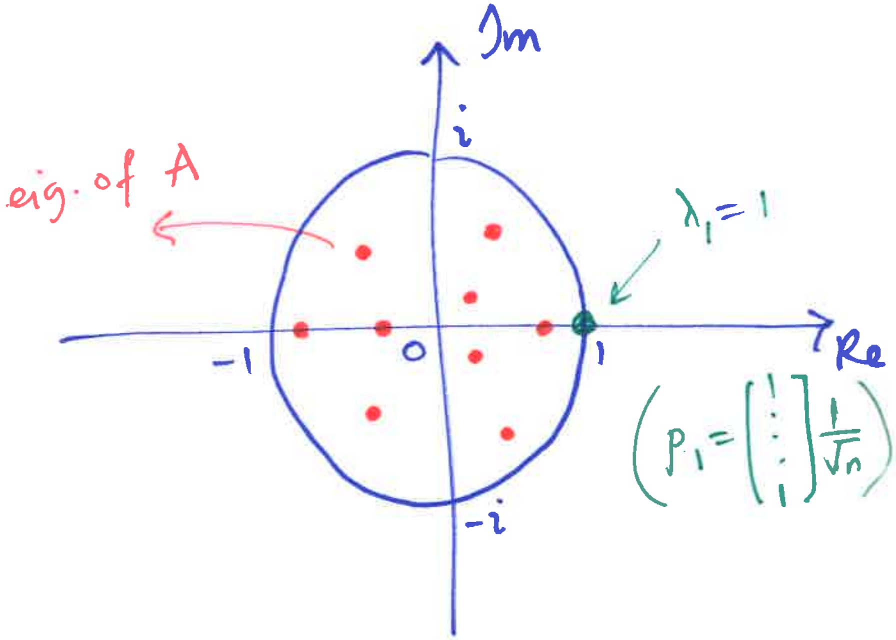
$$\mathbb{1} := \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \qquad A\,\mathbb{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \mathbb{1} = (1)\,\mathbb{1}$$

$(1, \mathbb{1})$ is an eigenvalue/eigenvector pair.

- it can be shown that all eigenvalues of a stochastic matrix A reside inside the open unit disk in the complex plane, except for the eigenvalue at $\lambda = 1$

$$T = \begin{bmatrix} | & | & & | \\ \frac{\mathbb{1}}{\sqrt{n}} & p_2 & \cdots & p_n \\ | & | & & | \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$



$$\left(p_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \frac{1}{\sqrt{n}} \right)$$

$$AT = T\Lambda \longrightarrow A = T\Lambda T^{-1}.$$

$$x(k+1) = A x(k)$$

$$x(1) = A x(0), \quad x(2) = A x(1) = A(A x(0)) = A^2 x(0), \quad \ldots, \quad x(m) = A^m x(0), \ldots$$

$$A^m = T \Lambda^m T^{-1} = T \begin{bmatrix} 1^m & & & \\ & \lambda_2^m & & \\ & & \ddots & \\ & & & \lambda_n^m \end{bmatrix} T^{-1}$$

$$A^m \xrightarrow[m \to \infty]{} T \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} T^{-1}$$

(since $|\lambda_i| < 1$, then $\lambda_i^m \to 0$ as $m \to \infty$ for $i > 1$)

$$= \underbrace{\begin{bmatrix} | & | & & | \\ \frac{1}{\sqrt{n}}\mathbb{1} & p_2 & \cdots & p_n \\ | & | & & | \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}}_{\Lambda^\infty} \underbrace{\begin{bmatrix} - & q_1^T & - \\ - & q_2^T & - \\ & \vdots & \\ - & q_n^T & - \end{bmatrix}}_{T^{-1}}$$

$$= \begin{bmatrix} | & | & & | \\ \frac{1}{\sqrt{n}}\mathbb{1} & 0 & \cdots & 0 \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & q_1^T & - \\ & \vdots & \\ - & q_n^T & - \end{bmatrix} = \frac{1}{\sqrt{n}} \mathbb{1}\, q_1^T$$

$$x_\infty = \lim_{k\to\infty} x(k)$$

$$= \left(\lim_{k\to\infty} A^k\right) x(0)$$

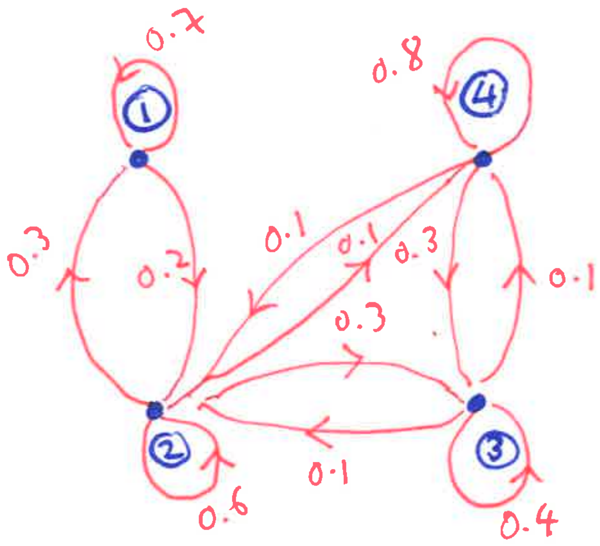
$$= \left(\frac{1}{\sqrt{n}} \mathbb{1} q_1^T\right) x(0)$$

$$= \mathbb{1} \underbrace{\left(\frac{1}{\sqrt{n}} q_1^T x_0\right)}_{\text{scalar}} \longrightarrow x_\infty \sim (\text{scalar}) \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \rightarrow \text{consensus!}$$

- consensus value is given by $\frac{1}{\sqrt{n}} q_1^T x_0$
- q_1^T contains useful information about social agents;

need to compute q_1. (but q_1 comes from T^{-1}, which comes from T ☹)

- application: propagation of beliefs in a social network



$$x(k+1) = A\, x(k)$$

$$A = \begin{bmatrix} 0.7 & 0.3 & 0 & 0 \\ 0.2 & 0.6 & 0.1 & 0.1 \\ 0 & 0.3 & 0.4 & 0.3 \\ 0 & 0.1 & 0.1 & 0.8 \end{bmatrix}, \quad x(k) = \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \end{bmatrix}$$

the entries of A are positive, between zero & one, and every row of A sums to one

(A is called a "stochastic" matrix)

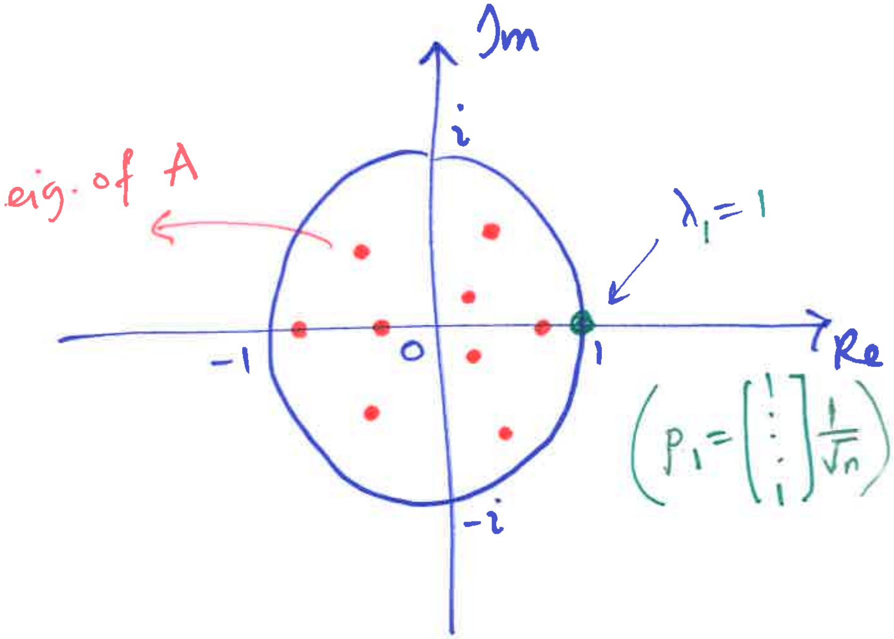
$$\mathbb{1} := \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \qquad A\mathbb{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \mathbb{1} = (1)\,\mathbb{1}$$

$(1, \mathbb{1})$ is an eigenvalue/eigenvector pair.

- it can be shown that all eigenvalues of a stochastic matrix A reside inside the open unit disk in the complex plane, except for the eigenvalue at $\lambda = 1$

$$T = \begin{bmatrix} | & | & & | \\ \frac{\mathbb{1}}{\sqrt{n}} & p_2 & \cdots & p_n \\ | & | & & | \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$



$$AT = T\Lambda \longrightarrow A = T\Lambda T^{-1}.$$

$x(k+1) = Ax(k)$

$x(1)=Ax(0), \quad x(2) = Ax(1) = A(Ax(0)) = A^2x(0), \ \dots, \ x(m) = \boxed{A^m}x(0), \dots$

$$A^m = T\Lambda^m T^{-1} = T\begin{bmatrix} 1^m & & & \\ & \lambda_2^m & & \\ & & \ddots & \\ & & & \lambda_n^m \end{bmatrix} T^{-1}$$

$$A^m \xrightarrow[m\to\infty]{} T\begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} T^{-1}$$

(since $|\lambda_i| < 1$, then $\lambda_i^m \to 0$ as $m \to \infty$ for $i > 1$)

$$= \underbrace{\begin{bmatrix} | & | & & | \\ \frac{1}{\sqrt{n}}\mathbb{1} & p_2 & \cdots & p_n \\ | & | & & | \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}}_{\Lambda^\infty} \underbrace{\begin{bmatrix} — & q_1^T & — \\ — & q_2^T & — \\ & \vdots & \\ — & q_n^T & — \end{bmatrix}}_{T^{-1}}$$

$$= \begin{bmatrix} | & | & & | \\ \frac{1}{\sqrt{n}}\mathbb{1} & 0 & \cdots & 0 \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & q_1^T & — \\ & \vdots & \\ — & q_n^T & — \end{bmatrix} = \underbrace{\frac{1}{\sqrt{n}}\mathbb{1}}_{p_1} q_1^T$$

$$x(\infty) = \lim_{k\to\infty} x(k)$$

$$= \left(\lim_{k\to\infty} A\right)^{k} x(0)$$

$$= \left(\frac{1}{\sqrt{n}}\mathbb{1}\, q_1^T\right) x(0)$$

$$= \mathbb{1}\underbrace{\left(\frac{1}{\sqrt{n}} q_1^T x_0\right)}_{\text{scalar}} \longrightarrow x(\infty) \sim (\text{scalar})\cdot\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix} \rightarrow \text{consensus!}$$

- consensus value is given by $\frac{1}{\sqrt{n}} q_1^T x_0$
- q_1^T contains useful information about social agents;

need to compute q_1. (but q_1 comes from T^{-1}, which comes from T 

- use $T^{-1}T = I$

$$\begin{bmatrix} - & q_1^T & - \\ & \vdots & \\ - & q_n^T & - \end{bmatrix}\begin{bmatrix} | & & | \\ p_1 & \cdots & p_n \\ | & & | \end{bmatrix} = \begin{bmatrix} q_1^T \overset{\frac{\mathbb{1}}{\sqrt{n}}}{\boxed{p_1}} & & * \\ & & \\ * & & \end{bmatrix} = \overbrace{\begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}}^{I}$$

- use $AT = T\Lambda$

$$T^{-1}A = \Lambda T^{-1}.$$

$$\begin{bmatrix} - & q_1^T & - \\ & \vdots & \\ - & q_1^T & - \end{bmatrix} A = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}\begin{bmatrix} - & q_1^T & - \\ & \vdots & \\ - & q_n^T & - \end{bmatrix}$$

$$\Updownarrow$$

$$q_1^T A = \lambda_1 q_1^T$$

$$\Updownarrow$$

$$A^T q_1 = \lambda_1 q_1$$

q_1 is the left eig. vec. of A corresp. to $\lambda_1 = 1$ & equiv. the right eig. vec. of A^T corresp. to $\lambda_1 = 1$.

$$A^{\infty} \longrightarrow \overbrace{\frac{1}{\sqrt{n}}\mathbb{1}}^{p_1}\, q_1^T$$

$$= \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0.\bar{4} & 0.\bar{6} & 0.\bar{2} & 0.\bar{6} \end{bmatrix}$$

$$\begin{bmatrix} | & | & | & | \\ 0.\bar{2} & 0.\bar{3} & 0.\bar{2} & 0.\bar{3} \\ | & | & | & | \end{bmatrix}$$

$$x(\infty) = A^{\infty} x(0) = \mathbb{1}\left(\frac{q_1^T x(0)}{\sqrt{n}}\right)$$

$$= \Big(0.\bar{2}\,(\text{initial belief of agent } 1) + 0.\bar{3}\,(\text{initial belief of agent } 2) + 0.\bar{1}\,(\text{initial belief of agent } 3) + 0.\bar{3}\,(\text{initial belief of agent } 4) \Big) \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

8-6

- the entries of q_1 are positive and can be considered as "averaging weights" and the value of $\frac{1}{\sqrt{n}} q_1^T x(0)$ gives a "weighted average" of the initial beliefs of all agents in the social network.

- furthermore, the entries of q_1 determine the amount of relative "influence" of different social agents in the network.

- note that we never had to compute T or T^{-1} or even find Λ !

```
A =

    0.7000    0.3000         0         0
    0.2000    0.6000    0.1000    0.1000
         0    0.3000    0.4000    0.3000
         0    0.1000    0.1000    0.8000

>> [T,Lambda] = eig(A)

T =
```

$p_1 = \frac{1}{\sqrt{n}} \mathbb{1}$

```
    0.5000   -0.6962   -0.5057   -0.2392
    0.5000   -0.2440    0.4243    0.3616
    0.5000    0.2757    0.6821   -0.8959
    0.5000    0.6163   -0.3146    0.0966

Lambda =
```

$\lambda_1 = 1$

```
    1.0000         0         0         0
         0    0.8051         0         0
         0         0    0.4483         0
         0         0         0    0.2466

>> Q = inv(T)

Q =
```

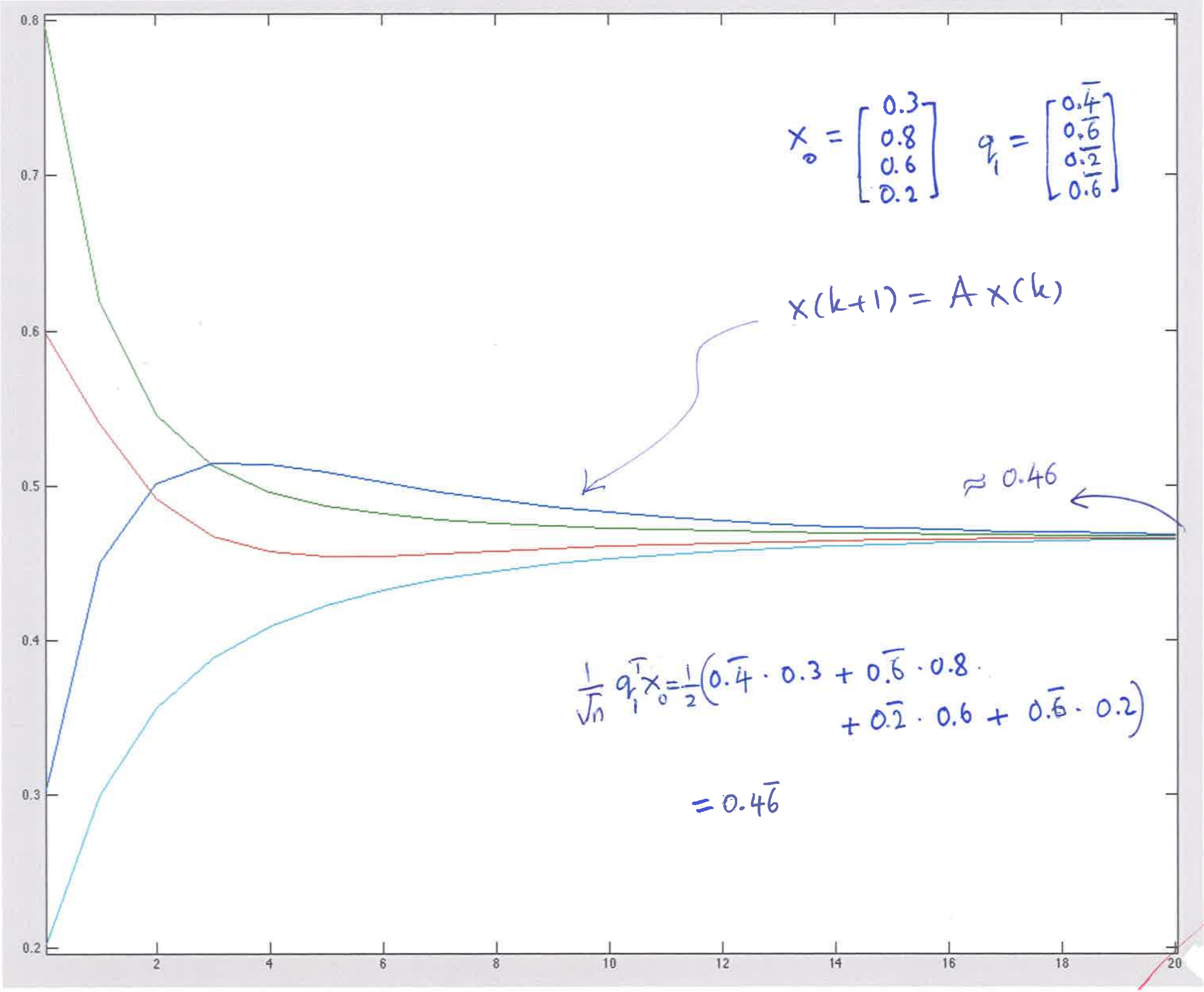
q_1^T

```
    0.4444    0.6667    0.2222    0.6667
   -0.5892   -0.3097    0.1166    0.7823
   -0.5576    0.7019    0.3761   -0.5203
   -0.3578    0.8111   -0.6699    0.2166

>> p1*q1'

ans =

    0.2222    0.3333    0.1111    0.3333
    0.2222    0.3333    0.1111    0.3333
    0.2222    0.3333    0.1111    0.3333
    0.2222    0.3333    0.1111    0.3333

>> A^100
```

$\approx A^\infty$

```
ans =

    0.2222    0.3333    0.1111    0.3333
    0.2222    0.3333    0.1111    0.3333
    0.2222    0.3333    0.1111    0.3333
    0.2222    0.3333    0.1111    0.3333
```

$= p_1 q_1^T$



$$x_0 = \begin{bmatrix} 0.3 \\ 0.8 \\ 0.6 \\ 0.2 \end{bmatrix} \quad q_1 = \begin{bmatrix} 0.\bar{4} \\ 0.\bar{6} \\ 0.\bar{2} \\ 0.\bar{6} \end{bmatrix}$$

$$x(k+1) = A\,x(k)$$

$$\approx 0.46$$

$$\frac{1}{\sqrt{n}} q_1^T x_0 = \frac{1}{2}\Big(0.\bar{4} \cdot 0.3 + 0.\bar{6} \cdot 0.8 \cdot + 0.\bar{2} \cdot 0.6 + 0.\bar{6} \cdot 0.2\Big)$$

$$= 0.4\bar{6}$$

- in this example agent 2 & 4 are the most influential agents in society.

- agent 1 has low influence because it has low connectivity

- agent 3 has low influence because it puts too much value on the opinions of others & has low self confidence. (he is more of a "sink" of ideas than a "source" !)

8-8

symmetric matrices

- a matrix is called "symmetric" if it is equal to its transpose, i.e., $A = A^T$.

unless stated otherwise, in this course we always assume that A has only real entries, $A \in \mathbb{R}^{n \times n}$.

aside: if A has complex entries, then the definition becomes

$$A = A^*, \quad A^* := (\bar{A})^T.$$

"Hermitian"

complex conjugation of all entries.

- eigenvectors of a symmetric matrix, corresponding to two distinct (unequal) eigenvalues, are orthogonal.

$$Av_1 = \lambda_1 v_1 \,, \quad Av_2 = \lambda_2 v_2 \,, \quad \lambda_1 \neq \lambda_2 \implies v_1 \perp v_2 . \qquad (v_1^T v_2 = 0)$$

---

aside: even if A is real, its eig.value & eig.vec may have complex entries.

proof:

$$\begin{aligned} \lambda_1 v_1^T v_2 &= (\lambda_1 v_1)^T v_2 \\ &= (A v_1)^T v_2 \\ &= v_1^T A^T v_2 \\ &= v_1^T A v_2 \\ &= v_1^T (\lambda_2 v_2) \\ &= \lambda_2 v_1^T v_2 \end{aligned}$$

aside: the trick is to start from $v_1^T A v_2$ and then group A once with v_1 & once with v_2.

$$= \vdots$$
$$= v_1^T A^T v_2$$

$$= v_1^T A v_2$$
$$= \vdots$$

$$\to (\lambda_1 - \lambda_2) v_1^T v_2 = 0 \xrightarrow{\lambda_1 \neq \lambda_2} v_1^T v_2 = 0 \to v_1 \perp v_2 .$$

- all eigenvalues of a (real) symmetric matrix are real.

$$\underset{(v \neq 0)}{Av = \lambda v} \to \lambda \in \mathbb{R}.$$

---

aside: $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \to \begin{cases} \lambda_1 = i \\ \lambda_2 = -i \end{cases} \notin \mathbb{R}$, $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \to \begin{cases} \lambda_1 = 1 \\ \lambda_2 = -1 \end{cases} \in \mathbb{R}.$

8-10

proof:

$$\bar{\lambda}\bar{v}^T v = \overline{(\lambda v)}^T v$$

$$= \overline{(Av)}^T v$$

$$= \bar{v}^T \bar{A}^T v$$

$$= \bar{v}^T A v$$

$$= \bar{v}^T (\lambda v)$$

$$= \lambda \bar{v}^T v$$

$$\to (\lambda - \bar{\lambda})\,\bar{v}^T v = 0 \xrightarrow{\boxed{\bar{v}^T v = \|v\|^2 \neq 0}} \lambda - \bar{\lambda} = 0 \to \lambda \in R.$$

$$\bar{v}^T v = \bar{v}_1 v_1 + \cdots + \bar{v}_n v_n$$

$$= |v_1|^2 + \cdots + |v_n|^2$$

$$= \|v\|_2^2$$

$$v \neq 0 \Leftrightarrow \|v\| \neq 0$$

aside:

$\|v\|_2^2 = v_1^2 + \cdots + v_n^2$ if $v \in R^n$

$\|v\|_2^2 = |v_1|^2 + \cdots + |v_n|^2$
$= \bar{v}_1 v_1 + \cdots + \bar{v}_n v_n$ if $v \in \mathbb{C}^n$.

- a matrix is called "orthogonal" or "unitary" if $Q^T = Q^{-1}$. in particular $QQ^T = Q^TQ = I$.

  - the columns (& rows) of an orthogonal matrix are vectors that are mutually orthogonal & have unit length.

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow Q^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = Q^T.$$

$$Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \rightarrow Q^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = Q^T.$$

$$Q = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \rightarrow Q^{-1} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = Q^T.$$

- a symmetric matrix can be diagonalized using an orthogonal transformation, $A = Q \Lambda Q^T$

proof:

assume for simplicity that A has distinct eigenvalues.

all eigenvectors of A are mutually orthogonal, i.e., $v_i^T v_j = 0$, $i \neq j$. furthermore, if we normalize $\{v_i\}$ such that $\|v_i\| = 1$ for all i, then

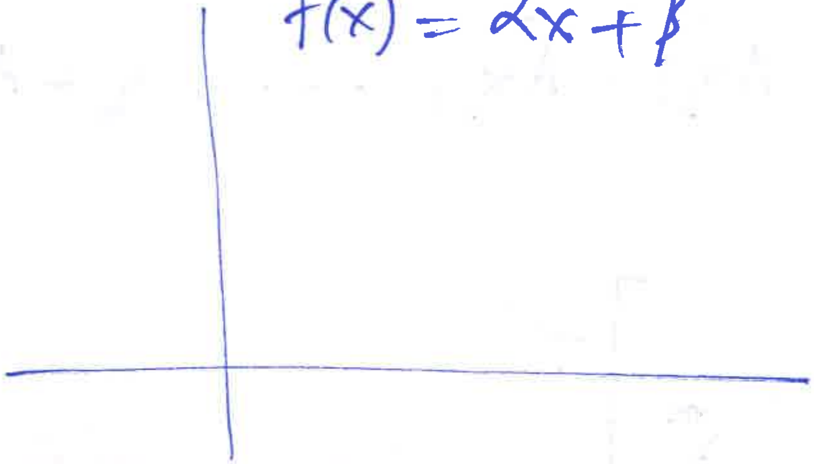
$$Q = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}$$

is an orthogonal matrix.

$$AQ = Q\Lambda \rightarrow A = Q\Lambda Q^{-1} \underset{\substack{\swarrow \\ Q \text{ is} \\ \text{orthogonal}}}{=} Q\Lambda Q^T.$$

8-13

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



$f(x)$

$x \in \mathbb{R}^n$

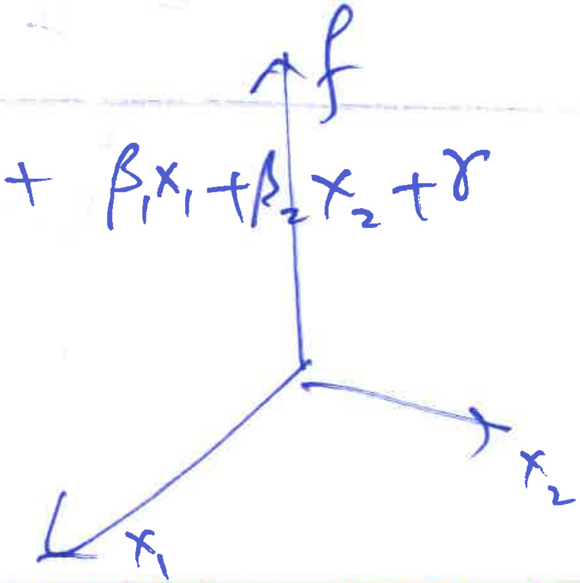
$f(x) \in \mathbb{R}.$

$$f(x) = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n + \beta$$

$$= \alpha^T x + \beta \qquad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}.$$

$$f(x_1, x_2) = \alpha_1 x_1^2 + \alpha_2 x_2^2 + \alpha_3 x_1 x_2 + \beta_1 x_1 + \beta_2 x_2 + \gamma$$

$$= x^T A x + \beta^T x + \gamma$$



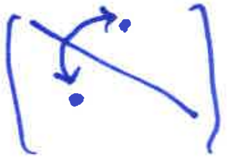
$$A = A_s + A_a$$

$$A_s^T = A_s \qquad A_a^T = -A_a .$$

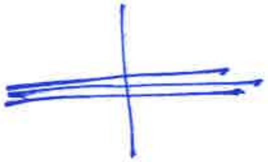
$$x^T A x = x^T A_s x + \underbrace{x^T A_a x}_{0}$$

review of last lecture

- symmetric matrices $A = A^* = \bar{A}^T$ ($A \in \mathbb{R}$, $A = A^T$)



  - $Av_i = \lambda_i v_i$, $Av_j = \lambda_j v_j$, $\lambda_i \neq \lambda_j \rightarrow v_i^T v_j = 0$



  - all eigenvalues are real, $\lambda_i \in \mathbb{R}$ $\forall i$
  - A can be diagonalized using an "orthogonal" matrix, $Q^{-1} = Q^T$.

    $A = Q \Lambda Q^{-1} = Q \Lambda Q^T$.
  - application: multi-variable quadratic functions

    $f(x) = x^T A x + b^T x + c$

- application of symmetric matrices: multivariable quadratic functions can be written using symmetric matrices

$$f(x) = x^T A x + b^T x + c \qquad A \text{ is symmetric}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} b_i x_i + c \qquad (a_{ij}, b_i, c \in \mathbb{R})$$

---

why A symmetric?

$$\boxed{\begin{aligned} A_s &= A_s^T \\ A_a &= -A_a^T \end{aligned}}$$

$$A = A_s + A_a \longrightarrow x^T A x = x^T (A_s + A_a) x$$

$$= x^T A_s x + \underbrace{x^T A_a x}_{=0}$$

another interpretation

$$a_{ij} x_i x_j \sim a_{ji} x_j x_i$$

aside: why not consider the more general case of $q(x) = x^T A x$ where A is not necessarily symmetric?

$$A = \underbrace{\frac{A+A^T}{2}}_{\text{sym.}} + \underbrace{\frac{A-A^T}{2}}_{\text{anti. sym.}}$$

$$\left(\frac{A+A^T}{2}\right)^T = \frac{A+A^T}{2}$$

$$\left(\frac{A-A^T}{2}\right)^T = -\frac{A-A^T}{2}.$$

$$x^T A x = x^T\left(\frac{A+A^T}{2}\right)x + x^T\left(\frac{A-A^T}{2}\right)x$$

$$= \frac{1}{2}\left(x^T A x + x^T A^T x\right) + \frac{1}{2}\left(x^T A x - x^T A^T x\right)$$

$$= \frac{1}{2}\left(x^T(Ax) + \underbrace{(Ax)^T x}_{= x^T(Ax)}\right) + \cancel{\frac{1}{2}\left(x^T(Ax) - \underbrace{(Ax)^T x}_{= x^T(Ax)}\right)} \rightarrow 0 \qquad = x^T A x.$$

a-4

- example:

$$\left.\begin{array}{l} A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \\ \\ A_{sym} = \underbrace{\begin{bmatrix} 1 & \frac{2+3}{2} \\ \frac{2+3}{2} & 4 \end{bmatrix}}_{\frac{A+A^T}{2}} = \begin{bmatrix} 1 & 2.5 \\ 2.5 & 4 \end{bmatrix}. \end{array}\right\} \begin{array}{c} x^T A x \\ = \\ x^T A_{sym} x \end{array}$$

$$q(x) = x^T A x \qquad (A = A^T)$$

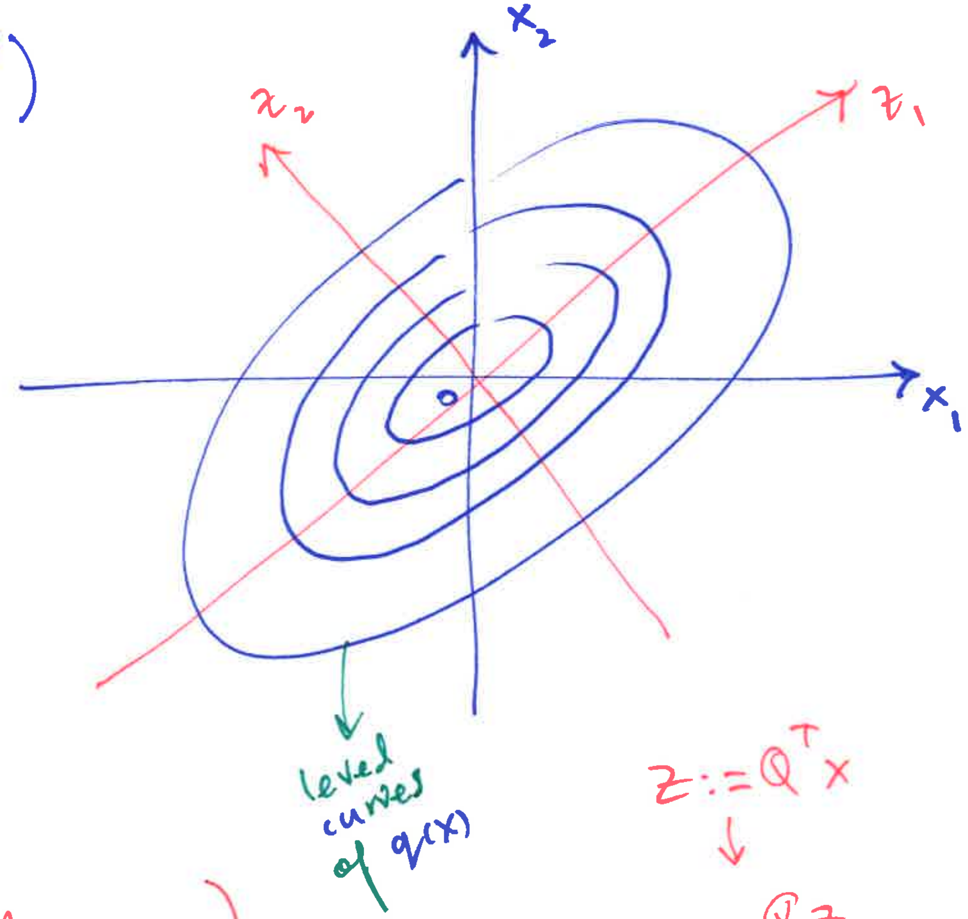
$$= x^T (Q \Lambda Q^T) x$$

$$= (Q^T x)^T \Lambda \left(\boxed{Q^T x}\right) \rightarrow z$$

$$= z^T \Lambda z$$

$$= \lambda_1 z_1^2 + \cdots + \lambda_n z_n^2$$

$$= \sum_{i=1}^{n} \lambda_i z_i^2 =: p(z).$$



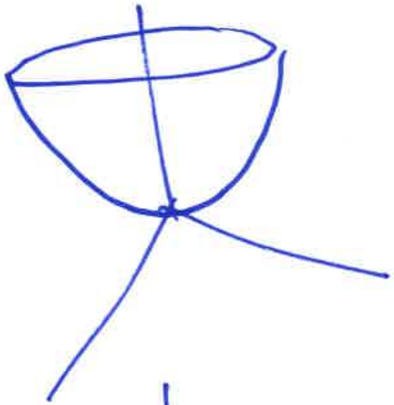
$$z := Q^T x$$
$$\downarrow$$
$$x = Qz$$

(compare with $q(x) = \sum_i \sum_j A_{ij} x_i x_j$)

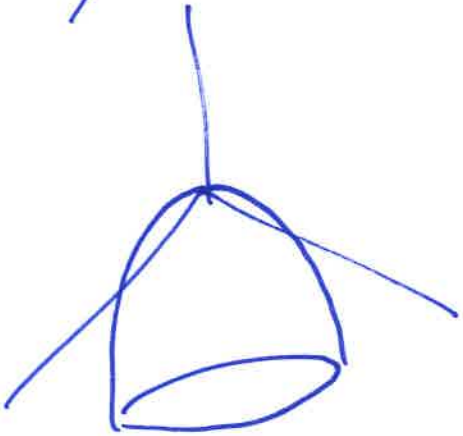
$$\alpha_1 x_1^2 + \alpha_2 x_2^2 + \alpha_{12} x_1 x_2$$

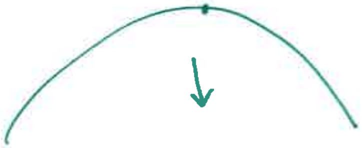
$$\beta_1 z_1^2 + \beta_2 z_2^2$$

- $\lambda_1 \& \lambda_2 > 0$



- $\lambda_1 \& \lambda_2 < 0$

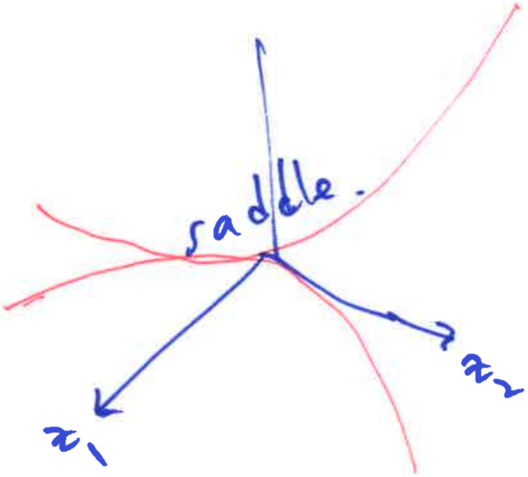


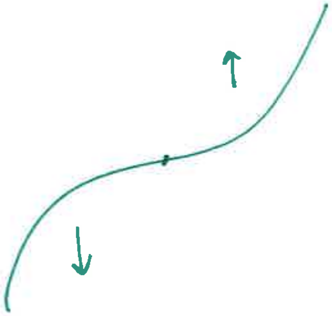


- $\lambda_1 \cdot \lambda_2 < 0$

$p(z) = \lambda_1 z_1^2 + \lambda_2 z_2^2$

$\lambda_1 > 0$
$\lambda_2 < 0$

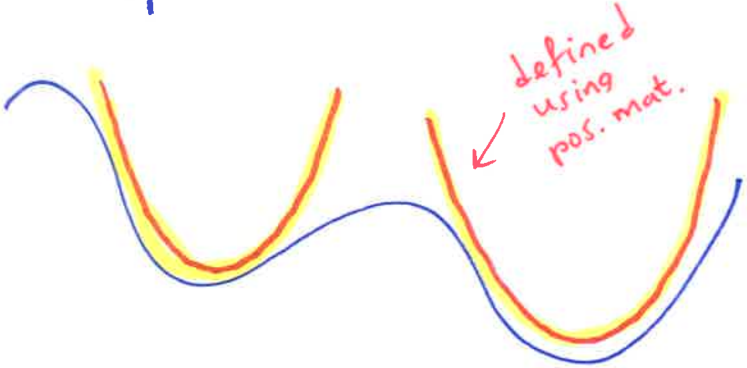


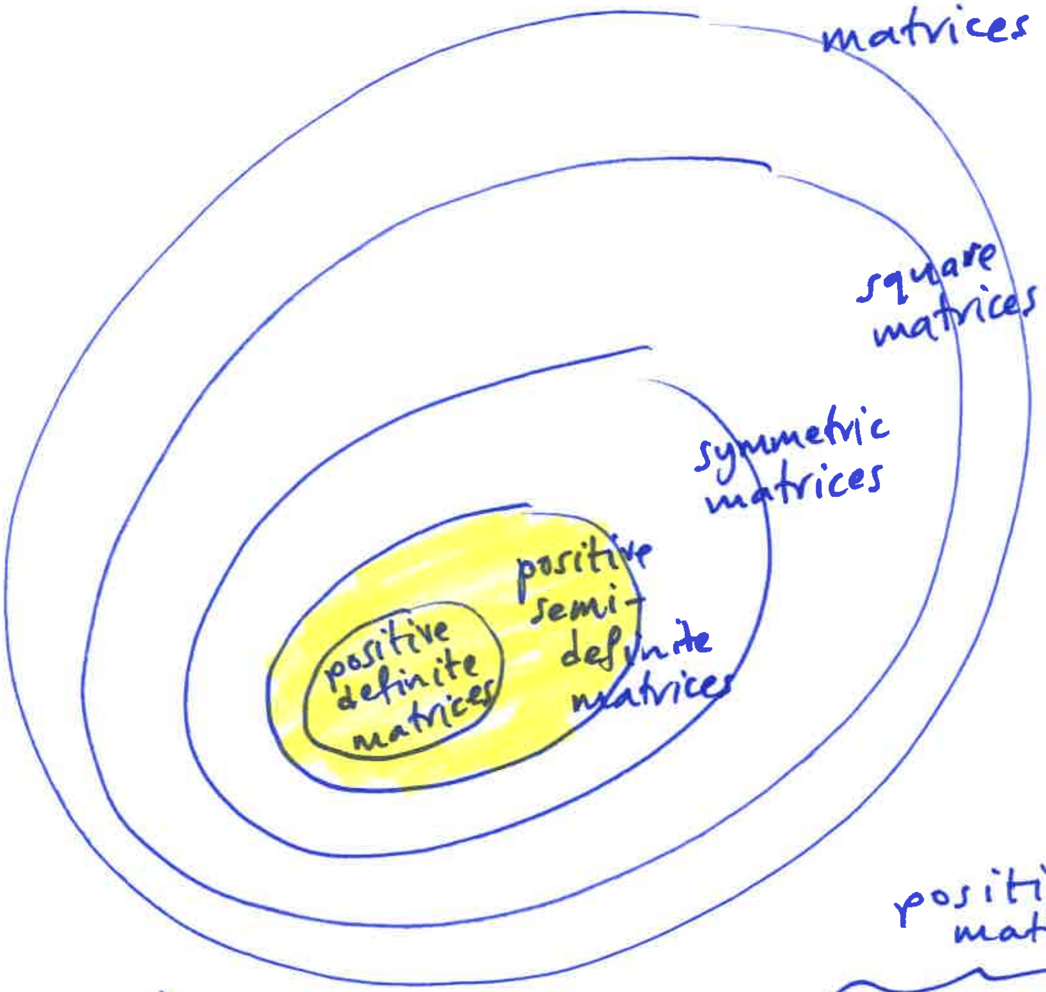


- what if λ_1 or λ_2 is zero?

positive matrices

- motivation
  - multivariable convex optimization





  - optimal control.

minimize $\int_0^T (x^TQx + u^TRu)\,dt$

keep both small

subject to system dynamics.

$(\dot{x} = Ax + Bu)$

$$\overbrace{Q \succ 0,\ R \succ 0}^{\text{positive matrices}}$$

$$\Updownarrow$$

$$x^TQx > 0,\ \forall x,$$

$$u^TRu > 0,\ \forall u.$$

a-7

- a symmetric (real) matrix P is called "positive definite" if $x^T P x > 0$ for any $x \neq 0$. $\xrightarrow{\text{notation}}$ $P \succ 0$

& P is called "positive semi-definite" if $x^T P x \succeq 0$ for any x. $\xrightarrow{\text{notation}}$ $P \succeq 0$

- P is "negative definite" $\Leftrightarrow$ $-P \succ 0$ $\xrightarrow{\text{notation}}$ $P \prec 0$

  P is "negtive semi-definite" $\Leftrightarrow$ $-P \succeq 0$. $\xrightarrow{\text{notation}}$ $P \preceq 0$

---

aside: $P \succ 0$, $P \succeq 0$, $P \prec 0$, $P \preceq 0$

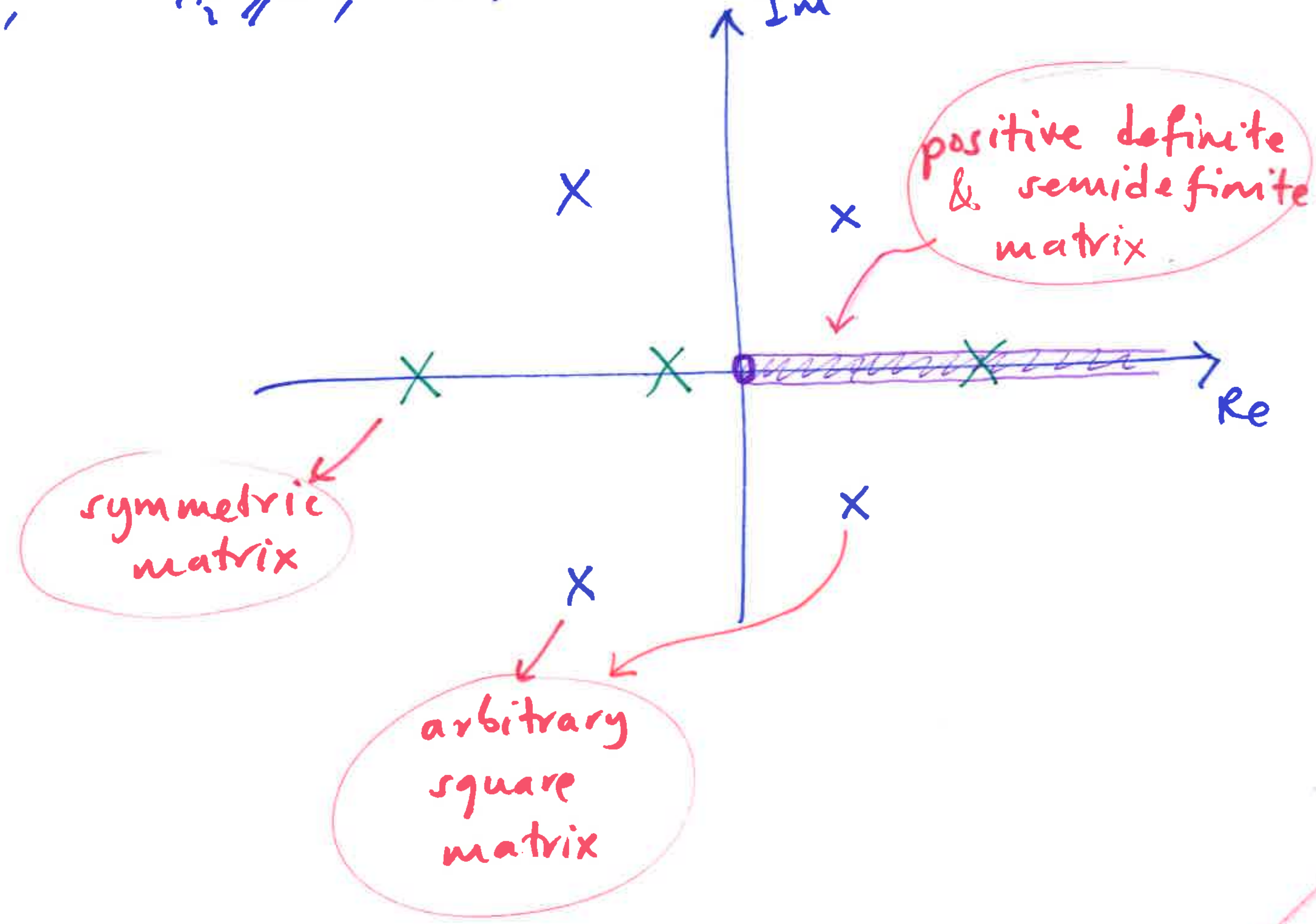
only make sense if P is a square & $\underbrace{\text{symmetric}}_{\text{eigenvalues guaranteed to be real}}$ matrix.

---

aside: testing whether $x^T P x \geq 0$ is intractable ! what to do ???

- the eigenvalues of a positive definite matrix are strictly positive, $\lambda_i > 0 \quad \forall i$.

  the eigenvalues of a positive semi-definite matrix are non-negative, $\lambda_i \geq 0, \quad \forall i$.



9-9

- example: $P = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$

$$x^TPx = x_1^2 + (x_1-x_2)^2 + (x_2-x_3)^2 + x_3^2 > 0$$

$\forall x \neq 0$

(only zero if
$x_1 = 0$,
$x_2 = x_1$,
$x_3 = x_2$,
$x_3 = 0$, which
means $x = 0$)

$\downarrow$

$P \succ 0$

$\text{eig}(P) = \{0.586, 2, 3.4\}$.

---

aside: $q(x) = \underbrace{\begin{bmatrix} x_1 & x_1-x_2 & x_2-x_3 & x_3 \end{bmatrix}}\underbrace{\begin{bmatrix} x_1 \\ x_1-x_2 \\ x_2-x_3 \\ x_3 \end{bmatrix}} = (Rx)^T(Rx) = x^TR^TRx$.

$$\boxed{P = R^TR}$$

$$\underbrace{\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}}_{x^T}\underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}}_{R^T}\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}}_{R}\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{x}$$

$N(R) = \{0\}$

$(Rx)^T(Rx) > 0$.

$\downarrow$

$P \succ 0$

$(Rx)^T(Rx) = 0$

$\Updownarrow$

$x = 0$

- example: $P = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$

$$x^TPx = (x_1 - x_2)^2 + (x_2 - x_3)^2 \geq 0 \quad \forall x$$

$\searrow$ zero for $x = \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, $\alpha \in \mathbb{R}$.

$\downarrow$

$P \succeq 0$

$\text{eig}(P) = \{0, 1, 3\}$

aside:

$$P = \underbrace{\begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}}_{R^T} \underbrace{\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}}_{R}.$$

$$N(R) = \left\{ \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \alpha \in \mathbb{R} \right\}$$

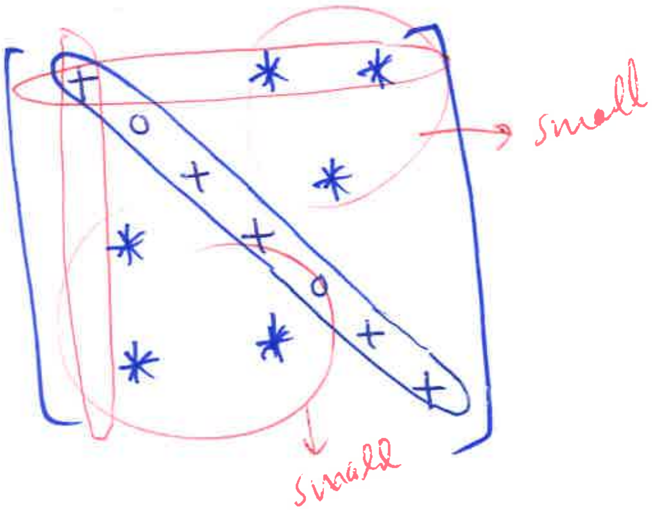
$\rightarrow \mathbb{1}^T P \mathbb{1} = \mathbb{1}^T R^T R \mathbb{1} = 0$

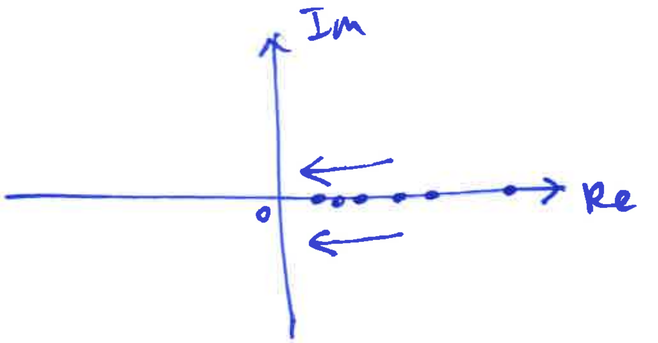
a_ii

- some "obviously" positive matrices
  - diagonal matrix with positive entries
  - "diagonally dominant" matrix with positive diagonal entries
  - matrix written as R^TR (or RR^T)

$$\underbrace{x^T R^T}_{y^T} \underbrace{R x}_{y} = y^T y = \|y\|_2^2 \geq 0$$

$$\|y\|_2^2 = 0 \leftrightarrow y = 0 \leftrightarrow Rx = 0 \leftrightarrow \mathcal{N}(R) \neq 0$$





review of last lecture

- positive matrices
  $(\subset$ symmetric matrices$)$

$$\begin{cases} \text{pos. def. } P>0 : & x^TPx>0 \quad \forall x\neq 0 \\ \text{pos. semi-def. } P\geq 0 : & x^TPx\geq 0 \quad \forall x \end{cases}$$

- $P>0 \overset{HW}{\Longleftrightarrow} \lambda_i>0 \quad \forall i$
  $P\geq 0 \Longleftrightarrow \lambda_i \geq 0 \quad \forall i$

$(P<0 \leftrightarrow x^TPx<0 \leftrightarrow \lambda_i<0)$

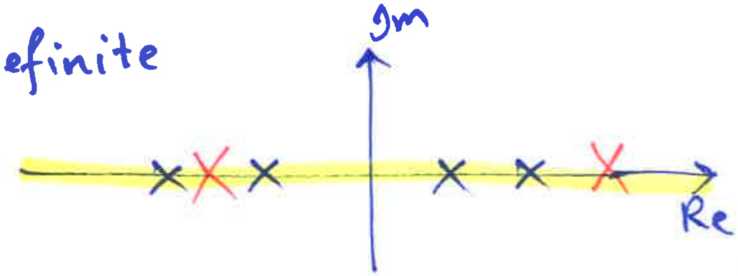
$(P\leq 0 \leftrightarrow x^TPx\leq 0 \leftrightarrow \lambda_i\leq 0)$

$$[x_1 \ \ x_2]\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} = x_1^2+2x_2^2 > 0 \quad x\neq 0$$

$$[x_1 \ \ x_2]\begin{bmatrix}-3 & 0\\ 0 & -4\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} = -3x_1^2-4x_2^2 < 0 \quad x\neq 0$$

$P<0 \leftrightarrow -P>0$

$$[x_1 \ \ x_2]\begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} = -x_1^2+x_2^2 \quad \text{sign indefinite}$$



- the notion of positive (semi-) definiteness defines a "partial ordering" of symmetric matrices

$$M \succeq N \iff M-N \succeq 0$$

$$M \succ N \iff M-N \succ 0$$

<u>aside</u>: $M = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $N = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$

$$M - N = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

<u>not every pair of matrices can be ordered</u>

- some properties of positive. definite matrices
  - every positive definite matrix is invertible, & its inverse is also positive definite $\quad P \succ 0 \leftrightarrow P^{-1} \succ 0$
  - if $M \succeq N \succ 0 \rightarrow N^{-1} \succeq M^{-1} \succ 0$
  - if $M \succ 0, N \succ 0 \longrightarrow rM \succ 0 \ (r>0), \quad M+N \succ 0$

    $NMN \succ 0, \quad MNM \succ 0$

<u>aside</u>:
NM &
MN may
not even
be symmetric

  - if $M \succeq 0 \longrightarrow$ there exists $N \succeq 0$ such that

    $$M = N^2 \quad (\text{or}, \ N := M^{1/2})$$

matrix inner product & norms

- motivation
  - define a geometry for matrices
  - define the "inner product" of matrices, which induces a notion of "angle" between them
  - use inner product to define "size" of, & "distance" between, matrices. in particular, this can be used to approximate one matrix with another.

recall:

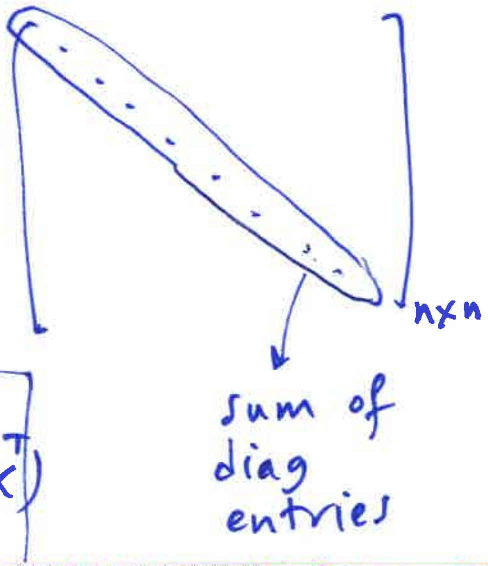
$$\cos\alpha = \frac{\langle x,y\rangle}{\|x\|_2\|y\|_2}$$

- the "trace" of a square matrix is the sum of its diagonal entries

$$\text{trace}(X) = \sum_{i=1}^{n} x_{ii}$$

$X : n \times n$

$X =$



aside 1: trace takes matrix, gives scalar.

aside 2: $\text{trace}(X) = \text{trace}(X^T)$

- we define the inner product of two matrices <u>of the same size</u> as

$$\langle X, Y \rangle := \operatorname{trace}(X^T Y) . \left(= \operatorname{trace}(Y^T X) = \operatorname{trace}(X Y^T) = \operatorname{trace}(Y X^T) \right)$$
$$= \sum_i \sum_j x_{ij} y_{ij}$$

(note that X, Y do not have to be square, but $X^T Y$ is guaranteed to be square, as is $X Y^T$:

$$\begin{cases} X : m \times n \\ Y : m \times n \end{cases} \rightarrow \begin{cases} X^T : n \times m \\ Y^T : n \times m \end{cases} \rightarrow \begin{cases} X^T Y : n \times n \\ X Y^T : m \times m \end{cases} \Bigg)$$

<u>aside</u>:

$\operatorname{trace}(AB)$
$= \operatorname{trace}(BA)$

$A : m \times n$
$B : n \times m$

- this definition of inner product is consistent with that for vectors: if x, y are vectors

$$\langle x, y \rangle = \operatorname{trace}\Big(\underbrace{\begin{bmatrix} \text{---} \; x^T \; \text{---} \end{bmatrix} \begin{bmatrix} | \\ y \\ | \end{bmatrix}}_{\text{scalar}} \Big) = x^T y$$

$$x = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \qquad y = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\mathrm{trace}(x^T y) = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 1+2+3 = 6.$$

$$\mathrm{trace}(y x^T) = \mathrm{trace}\left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} \right)$$

$$= \mathrm{trace} \begin{pmatrix} \textcircled{1} & 2 & 3 \\ 1 & \textcircled{2} & 3 \\ 1 & 2 & \textcircled{3} \end{pmatrix}$$

$$= 6.$$

A function $f : \mathbf{R}^n \to \mathbf{R}$ with $\mathbf{dom}\, f = \mathbf{R}^n$ is called a *norm* if

- f is nonnegative: $f(x) \geq 0$ for all $x \in \mathbf{R}^n$
- f is definite: $f(x) = 0$ only if $x = 0$
- f is homogeneous: $f(tx) = |t| f(x)$, for all $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$
- f satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y)$, for all $x, y \in \mathbf{R}^n$

- the matrix inner product naturally leads to a definition of a matrix norm

$$\|X\|_{\textcircled{F}}^2 := \langle X, X \rangle = \text{trace}(\underbrace{X^T X}_{\text{square}}) = \text{trace}\left(\begin{bmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} \right)$$

"Frobenius" norm of a matrix

$$= \text{trace}\left(\begin{bmatrix} \|x_1\|_2^2 & & * \\ & \ddots & \\ * & & \|x_n\|_2^2 \end{bmatrix} \right)$$

$$= \sum_i \|x_i\|_2^2 \qquad \left(\|x_i\|^2 = \sum_j x_{ji}^2 \right)$$

$$= \sum_i \sum_j x_{ij}^2$$

= sum of squares of all matrix entries

aside:

$$\|X\|_F^2 = 0$$
$$\updownarrow$$
$$X = 0.$$

(this is consistent with definition of vector norm)

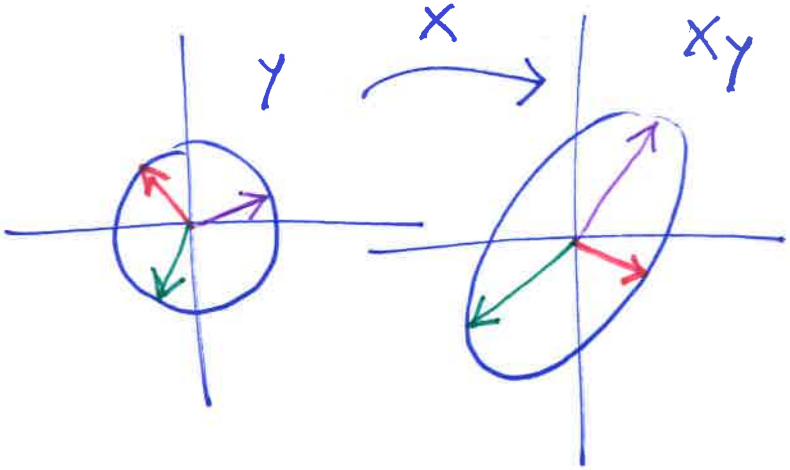
- another matrix norm is the "induced 2-norm", and is defined using the notion of the vector 2-norm

$$\|X\|_{2\text{-ind}} := \max_{y \neq 0} \frac{\|Xy\|_2}{\|y\|_2}$$

$$\left(\|\alpha y\|_2 = |\alpha| \, \|y\|_2 \right)$$

$$= \max_{y \neq 0} \left\| X \frac{y}{\|y\|_2} \right\|_2$$

$$= \max_{\|y\|_2 = 1} \|Xy\|_2$$



- interpretation: maximal amount of amplification of vectors.

- note: $\min_{\|y\|_2=1} \|Xy\|_2$ would <u>not</u> be a norm. (amplifier example)

$$\|X\|_{p\text{-ind}} := \max_{y \neq 0} \frac{\|Xy\|_p}{\|y\|_p}$$

$$\|y\|_p = \sqrt[p]{y_1^p + \cdots + y_n^p}$$

review of last lecture

- matrix inner product:

$$\langle \underbrace{X, Y}_{\text{same dimension}} \rangle = \text{trace}(\underbrace{X^TY}_{\text{square}}) = \text{trace}(Y^TX) = \text{trace}(XY^T) = \text{trace}(YX^T).$$

$$= \sum_{i,j} X_{ij} Y_{ij}$$

aside:
$\text{trace}(XYZ)$
$= \text{trace}(ZXY)$
$= \text{trace}(YZX)$

aside
$\text{trace}\, X$
$= \text{trace}\, X^T$

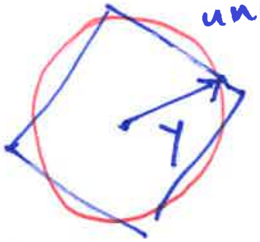
- the inner product induces a norm on matrices

$$\|X\|_F^2 = \text{trace}(X^TX) = \sum_{i,j} X_{ij}^2$$

- another matrix norm is the induced p-norm

$$\|X\|_{p\text{-ind}} = \max_{\|y\|_p = 1} \|Xy\|_p$$

← sweep over all vectors with unit length in p-norm



unit p-ball

for p=2 → use SVD

singular value decomposition (SVD)

- recall that for square matrices, we had

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

$$\Big\downarrow \text{ limitation}$$

A has to be square

$$\Big\downarrow \text{ question}$$

can we have a "similar" decomposition for arbitrary matrices?

$$\Big\downarrow \text{ yes}$$

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

U: orthogonal matrix, $U^{-1} = U^T$

V: orthogonal matrix, $V^{-1} = V^T$

Σ: diagonal matrix, but not necessarily square

- motivation for, and application of, SVD

– solving $Ax=b$, when A is square, invertible,

<u>but</u> "ill-conditioned".

$$\begin{tikzcd} Ax=b \arrow[d] \\ A^{-1}(Ax) = A^{-1}b \arrow[d] \\ x = A^{-1}b. \end{tikzcd}$$

$$A = \begin{bmatrix} 1 & 0 \\ 100 & 1 \end{bmatrix}. \qquad \text{eig}(A) = \{1, 1\}$$

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ -100 & 1 \end{bmatrix}$$

$$\delta A = \begin{bmatrix} 0 & 0.0099 \\ 0 & 0 \end{bmatrix}. \qquad (\|\delta A\|_F = 0.0099 \rightarrow \delta A \text{ "small" perturbation compared to } A, \|A\|_F \approx 100)$$

but we have to know δA that causes trouble!

$$A + \delta A = \begin{bmatrix} 1 & 0.0099 \\ 100 & 1 \end{bmatrix} \qquad \text{eig}(A+\delta A) = \{\underbrace{0.005}_{\text{sign of trouble!!}}, 1.995\}.$$

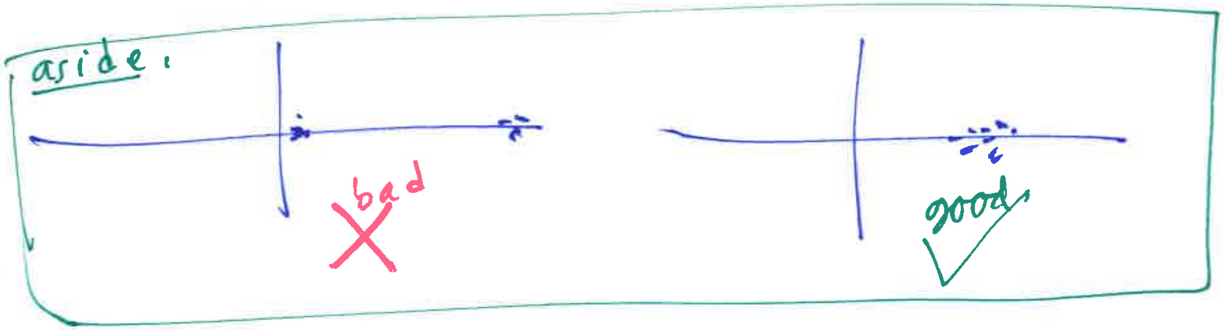
$$(A+\delta A)^{-1} = \begin{bmatrix} 1 & -0.0099 \\ -100 & 1 \end{bmatrix} \cdot \boxed{100}$$

- question: could we have just looked at A and foreseen ill-conditioned-ness?

  - answer: yes, just look at the singular values of A.

    singular values of A = { 100.01, 0.01 }. ← "ratio" important.

aside:

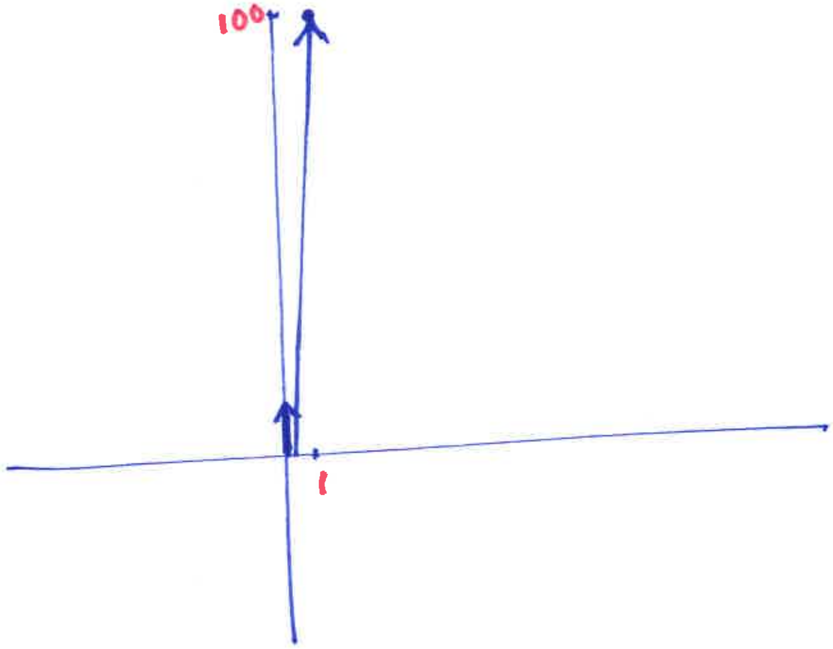


- insight:

$$A = \begin{bmatrix} 1 & 0 \\ 100 & 1 \end{bmatrix}$$

$\varepsilon = 0.01$

columns of A:



- the singular value decomposition of $A \in \mathbb{R}^{m\times n}$ can be written as

$$\underset{m\times n}{\underbrace{A}} = \underset{m\times m}{\underbrace{U}}\;\underset{m\times n\ (\text{same dim. as } A)}{\underbrace{\Sigma}}\;\underset{n\times n}{\underbrace{V^T}}$$

where U & V are orthogonal matrices, $U^TU = I_{m\times m}$, $V^TV = I_{n\times n}$, and Σ is a "diagonal" matrix (not necessarily square) composed of the "singular values" arranged in order of decreasing magnitude

$$\Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right]_{m\times n}$$

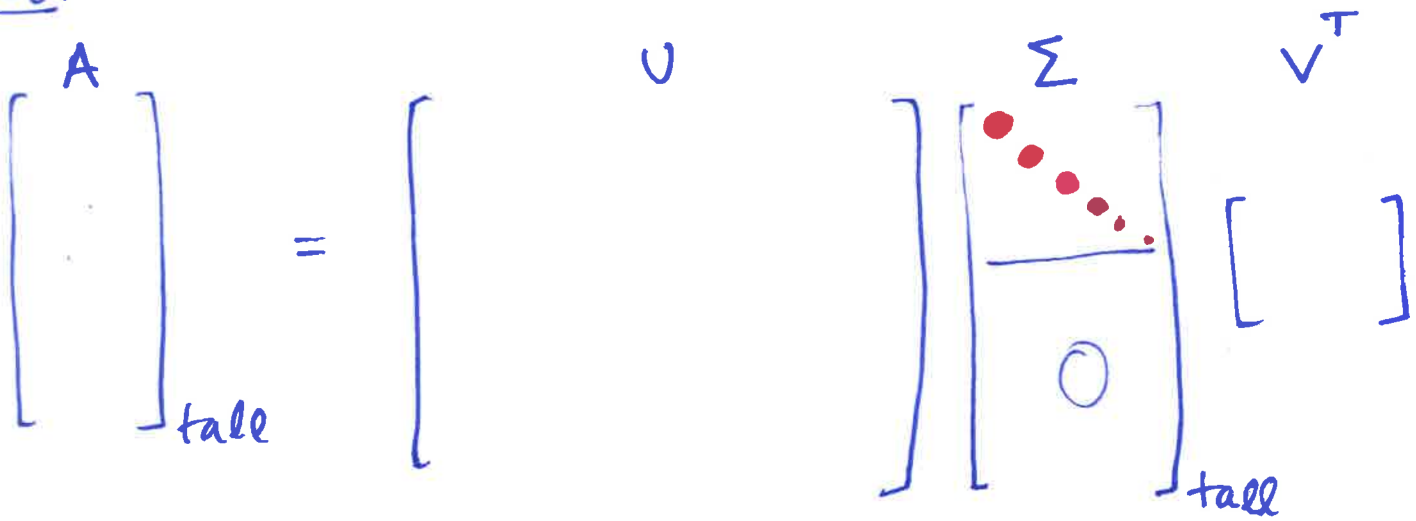
$$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0$$

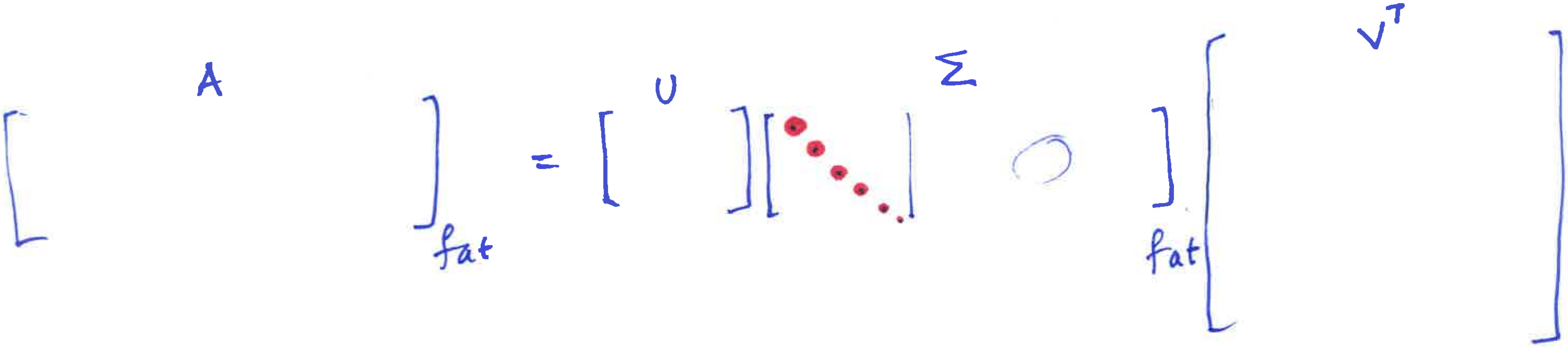
$$\sigma_i := \sqrt{i^{th} \text{ eigenvalue of } A^TA}$$

aside:

$$A^TA \geqslant 0$$
$$\downarrow$$
$$\text{eig}(A^TA) \geqslant 0$$

aside:





- computing the SVD: we perform two eigenvalue decompositions, one for AA^T and the other for A^TA.

$A = U\Sigma V^T$ (A: m×n)

$$\underbrace{AA^T}_{\substack{\text{square}\\ \text{\& symmetric}}} = (U\Sigma \overbrace{V^T)(V}^{I}\Sigma^T U^T) = \underset{\text{orthogonal}}{U}\underbrace{\Sigma\Sigma^T}_{\substack{\text{square}\\ \text{\& digonal}}}U^T = U\left[\begin{array}{ccc|c} \sigma_1^2 & & & \\ & \ddots & & 0 \\ & & \sigma_r^2 & \\ \hline & 0 & & 0 \end{array}\right]_{m\times m} U^T$$

$$\underbrace{A^TA}_{\substack{\text{square}\\ \text{\& symmetric}}} = (V\Sigma^T \overbrace{U^T)(U}^{I}\Sigma V^T) = \underset{\text{orthogonal}}{V}\underbrace{\Sigma^T\Sigma}_{\substack{\text{square}\\ \text{\& diagonal}}}V^T = V\left[\begin{array}{ccc|c} \sigma_1^2 & & & \\ & \ddots & & 0 \\ & & \sigma_r^2 & \\ \hline & 0 & & 0 \end{array}\right]_{n\times n} V^T$$

the nonzero singular values are found from

$$\sigma_i = \sqrt{i^{th} \text{ nonzero eigenvalue of } A^TA}$$
$$= \sqrt{i^{th} \text{ nonzero eigenvalue of } AA^T} > 0$$

note that because A'A and A A' are both symmetric (by construction) they are guaranteed to be diagonalizable

- example: $A = \begin{bmatrix} 1 & 0 \\ 100 & 1 \end{bmatrix}$.

$$AA^T = \begin{bmatrix} 1 & 100 \\ 100 & 10001 \end{bmatrix} \rightarrow AA^T u_i = \lambda_i u_i \rightarrow \begin{cases} u_1 \not\approx \begin{bmatrix} -0.01 \\ -1 \end{bmatrix}, \ \lambda_1 = (100.01)^2 \\ u_2 \not\approx \begin{bmatrix} -1 \\ 0.01 \end{bmatrix}, \ \lambda_2 = (0.01)^2 \end{cases}$$

$$U \not\approx \begin{bmatrix} -0.01 & -1 \\ -1 & 0.01 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 100.01 & 0 \\ 0 & 0.01 \end{bmatrix}.$$

$$A^TA = \begin{bmatrix} 10001 & 100 \\ 100 & 1 \end{bmatrix} \rightarrow A^TA\, v_i = \lambda_i v_i \rightarrow \begin{cases} v_1 \not\approx \begin{bmatrix} -1 \\ -0.01 \end{bmatrix}, \ \lambda_1 = (100.01)^2 \\ v_2 \not\approx \begin{bmatrix} -0.01 \\ 1 \end{bmatrix}, \ \lambda_2 = (0.01)^2 \end{cases}$$

$$V \not\approx \begin{bmatrix} -1 & -0.01 \\ -0.01 & 1 \end{bmatrix}$$

Σ = same as above
($\sigma_1 = \sqrt{\lambda_1}$, $\sigma_2 = \sqrt{\lambda_2}$)

$$A \not\approx \underbrace{\begin{bmatrix} -0.01 & -1 \\ -1 & 0.01 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 100.01 & 0 \\ 0 & 0.01 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} -1 & -0.01 \\ -0.01 & 1 \end{bmatrix}}_{V^T}$$

- use of SVD for computation of "pseudo inverses".

– solving $Ax=b$ when A is tall (over-determined problem) or fat (under-determined problem)

recall

$[\ A\ ]_{\text{tall}} \longrightarrow$ least-squares $\longrightarrow x^*_{ls} = (A^TA)^{-1}A^Tb$

$[\quad A\quad]_{\text{fat}} \longrightarrow$ least-norm $\longrightarrow x^*_{ln} = A^T(AA^T)^{-1}b$

easy to compute from SVD of A.

A^+ pseudo inverse

$A^+ = (A^TA)^{-1}A^T$ if A is tall (& has lin. ind. col.)

$A^+ = A^T(AA^T)^{-1}$ if A is fat (& has lin. ind. rows)

$A^+ = A^{-1}$ if A is square & invertible.

- pseudoinverse of $A = U\Sigma V^T$ $\qquad A \in \mathbb{R}^{m\times n}$

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

$$\Sigma = \left[\begin{array}{c|c} \overset{S}{\begin{matrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_r\end{matrix}} & 0 \\ \hline 0 & 0 \end{array}\right]_{m\times n} \qquad \Sigma^+ := \left[\begin{array}{c|c} \overset{S^{-1}}{\begin{matrix}1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_r\end{matrix}} & 0 \\ \hline 0 & 0 \end{array}\right]_{n\times m}$$

$$(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0)$$

---

aside3:

$$\begin{aligned} A^+ &= (U\Sigma V^T)^+ \\ &= (V^T)^+\Sigma^+U^+ \\ &= (V^T)^{-1}\Sigma^+U^{-1} \\ &= V\Sigma^+U^T. \end{aligned}$$

aside4: if Σ is tall then Σ^+ is fat. & the other way around.

aside1: why?!

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}.$$

aside2: if A^{-1} exists, then

$$A^+ \equiv A^{-1}.$$

- example: $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}_{3\times2}$. find A^+.

$$AA^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \left\{ \begin{array}{ll} u_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, & \lambda_1 = 1 \\ u_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, & \lambda_2 = 1 \\ u_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, & \lambda_3 = 0 \end{array} \right. \rightarrow U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow \left\{ \begin{array}{ll} v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, & \lambda_1 = 1 \\ v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, & \lambda_2 = 1 \end{array} \right. \rightarrow V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$A = \overbrace{I_{3\times3}}^{U} \overbrace{\left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ \hline 0 & 0 \end{array}\right]}^{\Sigma} \overbrace{I_{2\times2}}^{V^T} \qquad \left(\Sigma = \left[\begin{array}{cc} \sqrt{1} & 0 \\ 0 & \sqrt{1} \\ \hline \multicolumn{2}{c}{0} \end{array}\right] \right)$$

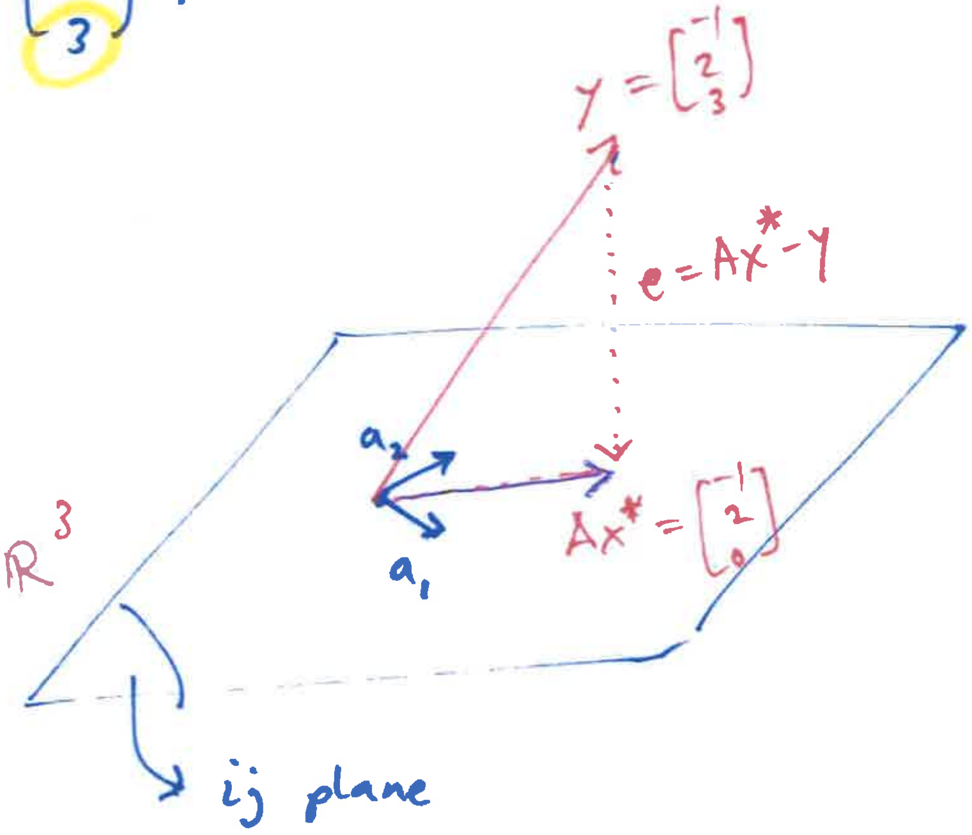
$$A^+ = V\Sigma^+U^T = I_{2\times2}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} I_{3\times3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

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

(agrees with least-squares method of finding the pseudoinverse; see lec. 6)

recall

- example: $A = \begin{array}{cc} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \\ a_1 \quad a_2 \end{array}$ $\quad y = \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix}$.



by inspection

$$x^* = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \in \mathbb{R}^2$$

$$Ax^* = -1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix} \in \mathbb{R}^3$$

least-squares solution

$$A^TAx^* = A^Ty$$

$$A^TA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad A^Ty = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}.$$

$$x^* = \underbrace{(A^TA)^{-1}}_{I} A^Ty = IA^Ty = A^Ty = \begin{bmatrix} -1 \\ 2 \end{bmatrix}.$$

$$A^{\dagger} = (A^TA)^{-1}A^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

recall 11-12

- example: $A = [1 \;\; 1]$. find A^+.

$$AA^T = 1\,(\sqrt{2})^2\,1$$

$$A^TA = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} (\sqrt{2})^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

$$A = \underbrace{1}_{U} \underbrace{\left[\begin{array}{c|c} \sqrt{2} & 0 \end{array}\right]}_{\Sigma} \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}}_{V^T}$$

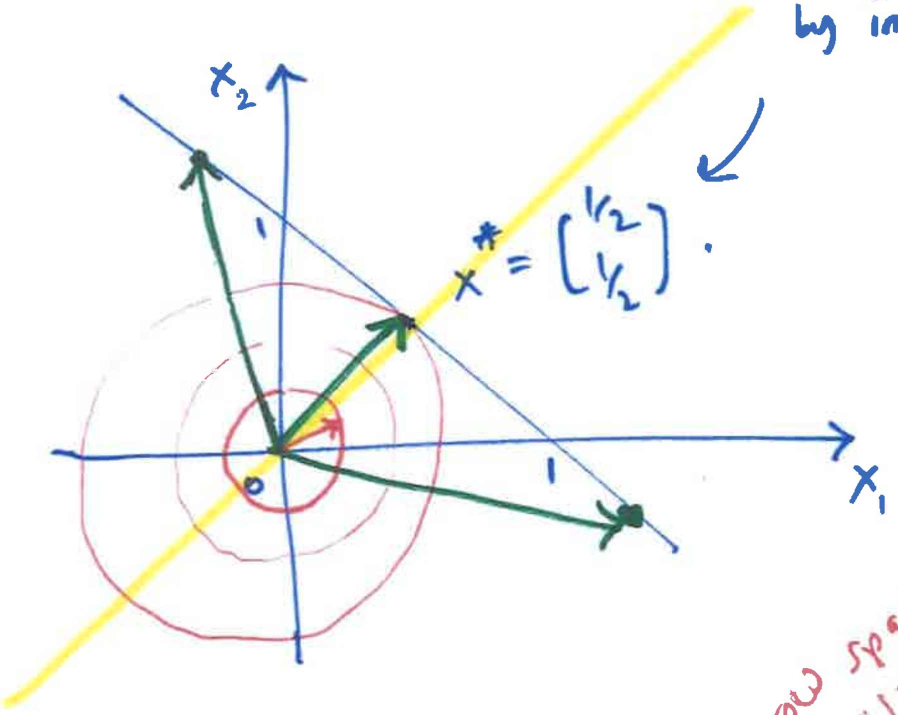
$$A^+ = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 0 \end{bmatrix} 1 = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}.$$

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

(agrees with least-norm method of finding the pseudoinverse; see lec.7) recall

- example: $[1 \quad 1]\begin{bmatrix}x_1\\x_2\end{bmatrix} = 1 \quad \longleftrightarrow \quad x_1 + x_2 = 1$

find x with minimum length.



by inspection

$A = [1 \quad 1]$.

$a_1^T = [1 \quad 1]$, $a_1 = \begin{bmatrix}1\\1\end{bmatrix}$.

$A(x^{\parallel} + x^{\perp})$

$= [1 \quad 1]\left(\begin{bmatrix}\alpha\\\alpha\end{bmatrix} + \begin{bmatrix}\beta\\-\beta\end{bmatrix}\right)$

$= \alpha + \alpha + \cancelto{0}{(\beta - \beta)}$

using least-norm

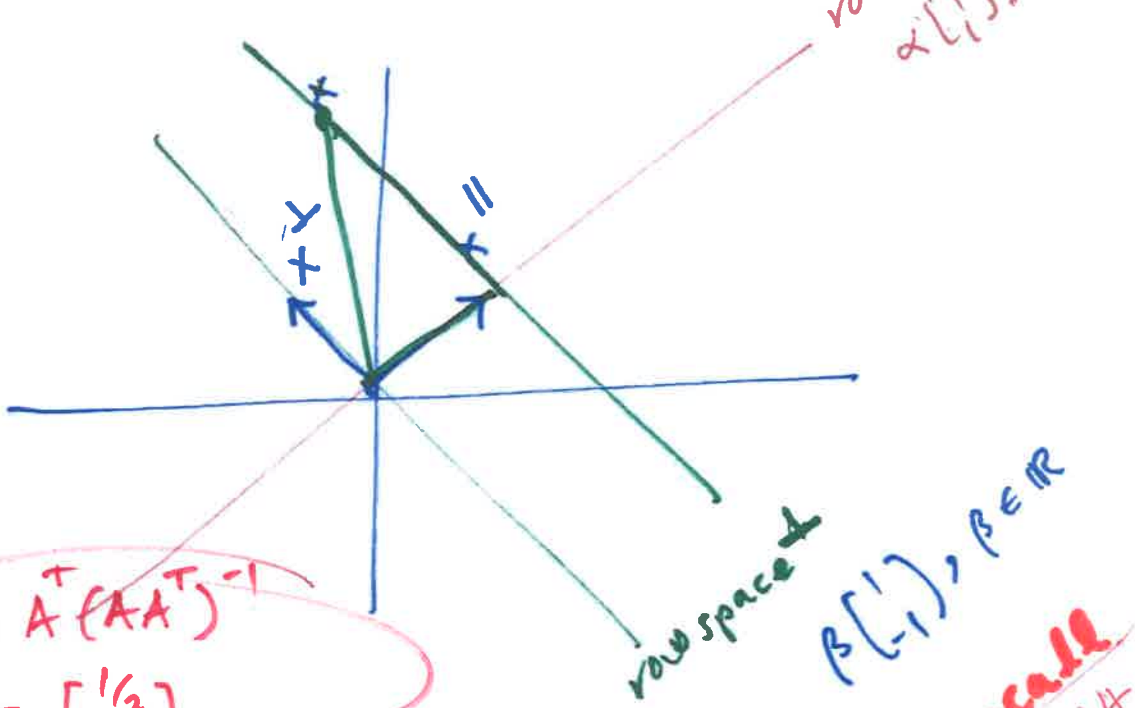
$x^* = A^T(AA^T)^{-1}y$

$= \begin{bmatrix}1\\1\end{bmatrix}\left([1 \quad 1]\begin{bmatrix}1\\1\end{bmatrix}\right)^{-1} \cdot 1$

$= \frac{1}{2}\begin{bmatrix}1\\1\end{bmatrix}$.

$A^+ = A^T(AA^T)^{-1}$

$= \begin{bmatrix}1/2\\1/2\end{bmatrix}$.



recall 11-14

review of last lecture

- for square matrices $A = T \Lambda T^{-1}$

  for non-square matrices $A = U \Sigma V^T$ $U^{-1} = U^T$, $V^{-1} = V^T$

  Σ "diag"

  $A : m \times n \rightarrow \Sigma : m \times n$ $\Sigma = \left[\begin{array}{c|c} \sigma_1 \ldots \sigma_r & 0 \\ \hline 0 & 0 \end{array}\right]$

- singular value decomposition can reveal ill-conditionedness of matries for inversion

  $A \longrightarrow A^{-1}$

  $A + \delta A \longrightarrow (A+\delta A)^{-1} \approx 100 A^{-1}$ has to do with A being "close to singularity".

- $A = U \Sigma V^T$

  $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$

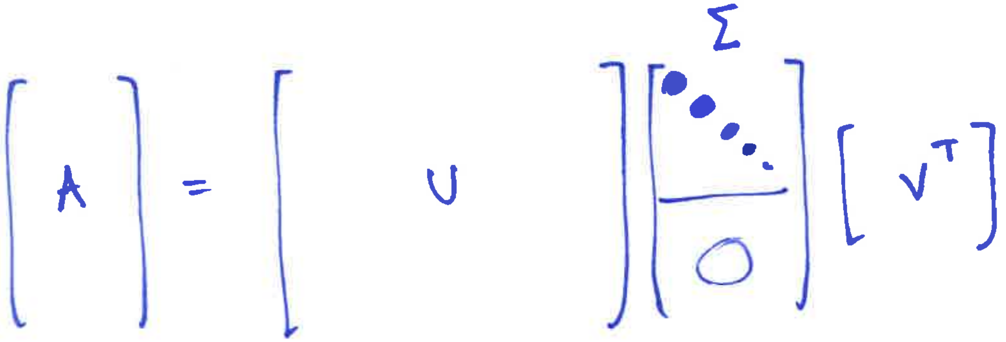
  $\sigma_i = \sqrt{i^{th} \text{ eigenvalue of } A^T A} = \sqrt{i^{th} \text{ eigenvalue of } A A^T} > 0$

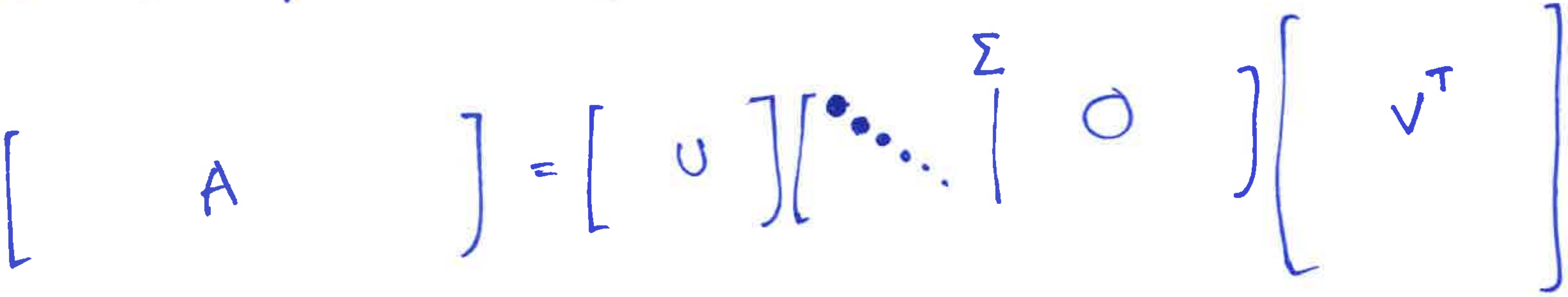
- computation of U & V:

$$AA^T = U\Lambda_U U^T$$

$$A^TA = V\Lambda_V V^T$$

contains squares of singular values





- applications
  - find matrix norms $\|A\|_F$ & $\|A\|_{2\text{-ind}}$
  - image processing & data compression
  - search engines
  - sensitivity analysis & distance from singularity.

review of last lecture

- using SVD to find pseudoinverses

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

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

$$\Sigma = \left[\begin{array}{c|c} S & 0 \\ \hline 0 & 0 \end{array}\right]_{m \times n}$$

$$\Sigma^+ = \left[\begin{array}{c|c} S^{-1} & 0 \\ \hline 0 & 0 \end{array}\right]_{n \times m}$$

$$\begin{aligned} A^+ &= (U \Sigma V^T)^+ \\ &= (V^T)^+ \Sigma^+ U^+ \\ &= (V^T)^{-1} \Sigma^+ U^{-1} \\ &= (V^T)^T \Sigma^+ U^T = V \Sigma^+ U^T \end{aligned}$$

- recall the matrix induced 2-norm

$$\|A\|_{2\text{-ind}} = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2 = 1} \|Ax\|_2$$

we next show that

$$\|A\|_{2\text{-ind}} = \underset{\sigma_{max}}{\underset{\uparrow}{\sigma_1}}(A) .$$

$$\underbrace{\left[\begin{array}{ccc|c} \boxed{\overset{\sigma_{max}}{\sigma_1}} & & & \\ & \ddots & & 0 \\ & & \boxed{\sigma_r} & \\ \hline & 0 & & \boxed{0} \end{array}\right]}_{\Sigma} \rightarrow \sigma_{min}$$

- for any matrix A, we have $\|A\|_{2\text{-ind}} = \sigma_{\max}(A)$.

proof:

$$\|A\|_{2\text{-ind}} = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$

$$= \max_{x\neq 0} \frac{\|U\overbrace{\boxed{\Sigma V^T x}}^{z}\|_2}{\|x\|_2}$$

hw problem $\leftarrow$

$$\boxed{=} \max_{x\neq 0} \frac{\|\Sigma \overbrace{\boxed{V^T x}}^{y}\|_2}{\|x\|_2}$$

$$= \max_{y\neq 0} \frac{\|\Sigma y\|_2}{\|Vy\|_2}$$

$$= \max_{y\neq 0} \frac{\|\Sigma y\|_2}{\|y\|_2}$$

$$= \max_{\|y\|_2=1} \|\Sigma y\|_2$$

(next slide)

aside: hw problem

$$\|Uz\|_2 = \|z\|_2$$

$$V^T x = y$$

$$\boxed{VV^T} x = Vy$$

$$x = Vy$$

$$x \neq 0 \Leftrightarrow y \neq 0$$

$$\frac{\|\Sigma y\|_2}{\|y\|_2} = \left\|\Sigma \boxed{\frac{y}{\|y\|_2}}\right\|_2$$

$$= \max_{\|y\|_2=1} \left\| \begin{bmatrix} \sigma_1 y_1 \\ \vdots \\ \sigma_r y_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\|_2$$

$$= \max_{y_1^2+\cdots+y_n^2=1} \sqrt{\sigma_1^2 y_1^2 + \cdots + \sigma_r^2 y_r^2 + 0\, y_{r+1}^2 + \cdots + 0\, y_n^2}$$

↓

the maximum is achieved for $y^* = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

&

the value of the maximum is $\sigma_1 = \sigma_{\max}(A)$.

↓

$$\|A\|_{2\text{-ind}} = \sigma_{\max}(A) \qquad A = U \left[\begin{array}{ccc|c} \boxed{\sigma_{\max}} & & & \\ & \ddots & & \\ & & \sigma_r & \\ \hline & & & 0 \end{array}\right] V^T$$

recall that $x = Vy$

$$\rightarrow \quad x^* = V y^* = V \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = v_1$$ first column of V.

in summary, the amount of maximum amplification by A is given by σ_1, and the direction of maximum amplification is given by the first row of V^T.

$$A = U \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ \hline & & & 0 \end{array}\right] \begin{bmatrix} — & v_1^T & — \\ & \vdots & \end{bmatrix}$$

- least amplification

$$\min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \min_{\|x\|_2=1} \|Ax\|_2$$ = smallest singular value of A (which will be zero if A has a nonempty null space)

examples of $\|\cdot\|_{2\text{-ind}}$ & SVD computation

- example : $A = aI \qquad a \in \mathbb{R}.$

$$\|A\|_{2\text{-ind}} = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \neq 0} \frac{\|aIx\|_2}{\|x\|_2}$$

$$= \max_{x \neq 0} \frac{\|ax\|_2}{\|x\|_2} = |a| \max_{x \neq 0} \cancel{\frac{\|x\|_2}{\|x\|_2}} = |a|.$$

aside: $a \in \mathbb{R}$

$$\|a\,x\|_2 = |a|\,\|x\|_2$$

$$\downarrow$$

$$\|A\|_{2\text{-ind}} = |a|.$$

$$A^TA = (aI)^T(aI) = a^2 I = \underbrace{I}_{V}\underbrace{(a^2 I)}_{\Lambda_V}\underbrace{I}_{V^T}$$

$$AA^T = \underbrace{I}_{-U}\underbrace{(a^2 I)}_{\Lambda_U}\underbrace{I}_{-U^T}$$

$$\Lambda_V = \Sigma^T\Sigma$$

$$\Lambda_U = \Sigma\Sigma^T$$

$$\Sigma = \begin{bmatrix} |a| & & \\ & \ddots & \\ & & |a| \end{bmatrix}$$

$$\begin{cases} a>0 \rightarrow & A = I\begin{bmatrix} a & & \\ & \ddots & \\ & & a \end{bmatrix} I \\ a<0 \rightarrow & A = (-I)\begin{bmatrix} -a & & \\ & \ddots & \\ & & -a \end{bmatrix} I \end{cases}$$

recall

$A = aI$

in particular, $\|A\|_{2\text{-ind}} = \sigma_{\max}(A) = |a|$.

---

aside: if all singular values are distinct then the SVD is unique up to multiplication of each column of U, & the corresponding row of V^T, by -1.

- example: $A = \Lambda$, where Λ is a diagonal matrix.

$$\|A\|_{2\text{-ind}} = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2 = 1} \|Ax\|_2$$

$$= \max_{\|x\|_2=1} \left\| \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \right\|_2$$

$$= \max_{\|x\|_2=1} \left\| \begin{bmatrix} \lambda_1 x_1 \\ \vdots \\ \lambda_n x_n \end{bmatrix} \right\|_2$$

$$= \max_{x_1^2 + \cdots + x_n^2 = 1} \sqrt{\lambda_1^2 x_1^2 + \cdots + \lambda_n^2 x_n^2}$$

$$= \max_i |\lambda_i|. \longrightarrow$$

this value is achieved for $x^* = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ where the "1" appears in the location of the λ_i with the largest absolute value.

assume for simplicity that all diagonal entries of A are > 0,
(then look at red marks for < 0 entries)

say $\quad A = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$.

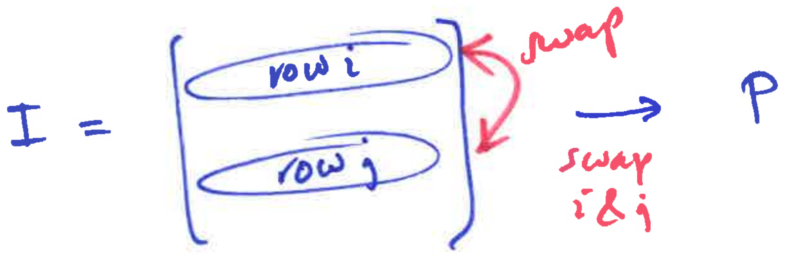
- we will use permutation/elementary matrices, which are found by exchanging the order of the columns & rows of the identity matrix.

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow[1 \,\&\, 2]{\text{swap rows}} \quad P = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

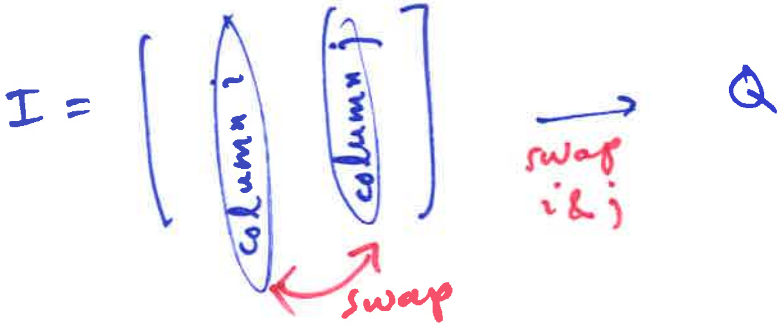
$$PA = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 0 & +2 \\ 1 & 0 \end{bmatrix}$$

- the operations performed on the rows (columns) of I to get P (Q) will be performed on A when we form PA (AQ).

12-11



PA will be the same as A but with rows i & j exchanged.



AQ will be the same as A but with columns i & j exchanged

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \xrightarrow[1 \ \& \ 2]{\text{swap columns}} Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

$$PAQ = \underbrace{\begin{bmatrix} 0 & +2 \\ 1 & 0 \end{bmatrix}}_{PA} \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{Q} = \begin{bmatrix} +2 & 0 \\ 0 & 1 \end{bmatrix}.$$

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

- all permutation matrices are orthagonal matrices:

$$P^T = P^{-1} \qquad Q^T = Q^{-1}.$$

$$\underbrace{P^T (P}_{I} A \underbrace{Q) \, Q^T}_{I} = P^T \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} Q^T$$

$$A = \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{V^T}.$$

- in conclusion, applying this procedure to any diagonal matrix Λ gives that

$$\Sigma = \begin{bmatrix} \text{reordering of the} \\ \text{absolute values of} \\ \text{the entries of } \Lambda \\ \text{in descending order} \end{bmatrix}.$$

$$A = \underset{\swarrow}{P} \, \Sigma \, \underset{\swarrow}{Q}$$

permutation/elementary

(P & Q used to reorder & take absolute values)

12-13

can extend same methodology to other matrices

$$A = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix} = \underset{\underset{U}{\uparrow}}{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}} \underset{\underset{\Sigma}{\uparrow}}{\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}} \underset{\underset{V^T}{\uparrow}}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}} .$$

or

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

$$= \underset{\underset{U}{\uparrow}}{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}} \underset{\underset{\Sigma}{\uparrow}}{\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}} \underset{\underset{V^T}{\uparrow}}{\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}}$$

---

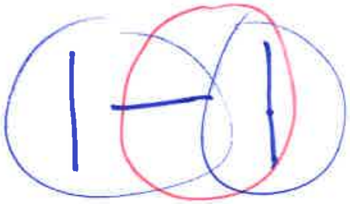
can be generalized to any matrix that has only one nonzero entry on any row & column.

$$\begin{bmatrix} 0 & 7 & 0 \\ -2 & 0 & 0 \\ 0 & 0 & -9 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 7 & \\ & & 2 \end{bmatrix} \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix} .$$

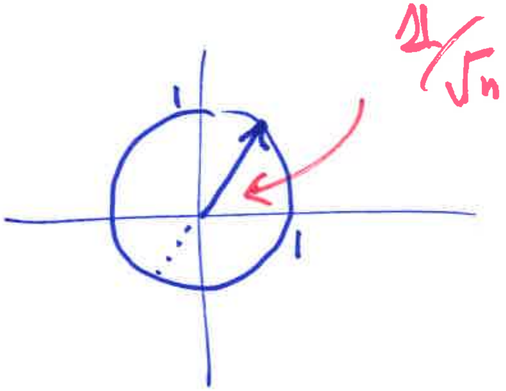
12 – 14

- example: $A = \alpha \mathbb{1}\mathbb{1}^T, \quad \alpha \in \mathbb{R}$ $\qquad \mathbb{1} := \begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix} \in \mathbb{R}^n.$

$$\|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2=1} \|Ax\|_2$$



$$= \max_{\|x\|_2=1} \|(\alpha)\mathbb{1}(\mathbb{1}^T x)\|_2$$



$$= \max_{\|x\|_2=1} |\alpha|\,|\mathbb{1}^T x|\,\|\mathbb{1}\|_2$$

max achieved for $x^* = \pm \mathbb{1}/\sqrt{n}$ ←

$$= |\alpha|\,\left|\mathbb{1}^T \frac{\mathbb{1}}{\cancel{\sqrt{n}}}\right|\,\overset{\cancel{\sqrt{n}}}{\|\mathbb{1}\|_2}$$

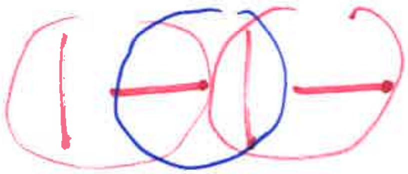
$$= |\alpha|\,|\mathbb{1}^T\mathbb{1}| = |\alpha|\, n.$$

---

aside: $|\mathbb{1}^T x| = \underbrace{\|\mathbb{1}\|}_{\sqrt{n}}\ \underbrace{\|x\|}_{1}\ |\cos\theta| \longrightarrow |\cos\theta| = 1 \leftrightarrow \begin{cases}\theta = 0° \to x^* = \mathbb{1}/\sqrt{n}\\ \theta = 180° \to x^* = -\mathbb{1}/\sqrt{n}\end{cases}$

$A = \alpha \mathbb{1}\mathbb{1}^T$ symmetric matrix $(A^T = A)$

$AA^T = A^TA = A^2 = \alpha^2 \mathbb{1}\mathbb{1}^T \mathbb{1}\mathbb{1}^T = n\alpha^2 \mathbb{1}\mathbb{1}^T$



the singular value decomposition boils down to an eigenvalue decomposition for the matrix $\mathbb{1}\mathbb{1}^T$.

$$H := \mathbb{1}\mathbb{1}^T$$

$$H\mathbb{1} = (\mathbb{1}\mathbb{1}^T)\mathbb{1}$$
$$= \mathbb{1}(\mathbb{1}^T\mathbb{1})$$
$$= n\,\mathbb{1}$$

↓

$\mathbb{1}$ is an eigenvector of H with eigenvalue equal to n.

$\mathbb{1}$ is an eigenvector of H

$q_1 = \mathbb{1}$ ← first eigenvector of H.

aside:

$$Hx = \mathbb{1}\mathbb{1}^T x$$
$$= \mathbb{1}\left(\sum_i x_i\right)$$

$$Hx = \begin{bmatrix} \vdots \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

← all entries equal to $\sum x_i$

aside:

$\mathcal{E} = \{\beta\mathbb{1}, \beta \in \mathbb{R}\}$.

since H is symmetric, its eigenvectors can be chosen orthogonal to each other. since $\mathbb{1}$ is an eigenvector, then every other eigenvector q has to satisfy

$$\mathbb{1}^T q = 0.$$

$$Hq = \mathbb{1}\mathbb{1}^T q = \mathbb{1}\cdot 0 = 0\cdot q$$

---

recall

$q_1 = \mathbb{1}$

$q_1 \perp q_2$

$q_1 \perp q_3$

$q_2 \perp q_3$

for example, in $\mathbb{R}^3$

$$q_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \qquad \mathbb{1}\mathbb{1}^T q_2 = 0 = 0\cdot q_2$$

$$q_3 = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \qquad \mathbb{1}\mathbb{1}^T q_3 = 0 = 0\cdot q_3$$

$$Q = \begin{bmatrix} | & | & | \\ \frac{q_1}{\|q_1\|_2} & \frac{q_2}{\|q_2\|_2} & \frac{q_3}{\|q_3\|_2} \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \end{bmatrix}$$

$$HQ = Q \textcircled{\Lambda} \longrightarrow \begin{bmatrix} n & & \\ & 0 & \\ & & 0 \end{bmatrix}$$

$$H = \underbrace{Q}_{U_H} \underbrace{\begin{bmatrix} n & & \\ & 0 & \\ & & 0 \end{bmatrix}}_{\Sigma_H} \underbrace{Q^T}_{V_H^T}$$

---

recall that $A = \alpha H$.

$$\begin{cases} \text{if } \alpha \geq 0 \rightarrow & A = Q \begin{bmatrix} \alpha n & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} Q^T \\ \text{if } \alpha \leq 0 \rightarrow & A = (-Q) \begin{bmatrix} -\alpha n & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} Q^T \end{cases}$$

in both cases $\begin{cases} \sigma_1 = \|A\|_2 = |\alpha| n \\ \sigma_n = 0 \end{cases}$

- example: $A = ab^T$ (a, b don't have to have same dim.)

aside:

$\mathcal{C} = \{\alpha a, \alpha \in \mathbb{R}\}$

$\mathcal{R} = \{\beta b, \beta \in \mathbb{R}\}$

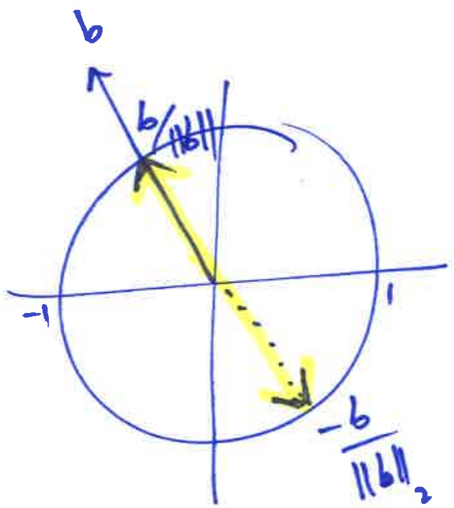
$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

$$= \max_{\|x\|_2=1} \|a(b^Tx)\|_2$$

$$= \max_{\|x\|_2=1} \|a\|_2 \, |b^Tx|$$

$$x^* = \frac{\pm b}{\|b\|_2} \longleftarrow \; = \|a\|_2 \left| b^T \frac{b}{\|b\|_2} \right|$$

$$= \|a\|_2 \, \|b\|_2$$



aside: $|b^Tx| = \underbrace{\|b\|_2}_{\text{fixed}} \; \underbrace{\|x\|_2}_{\text{fixed (in length)}} \; |\cos\theta| \rightarrow \theta^* = \{0^\circ, 180^\circ\}$

$x^* = \frac{b}{\|b\|_2}$ $\qquad$ $x^* = \frac{-b}{\|b\|_2}$

$$AA^T = (ab^T)(ab^T)^T = a\left(b^T b\right)a^T = \|b\|_2^2 \, a a^T.$$

(similar to $\alpha \mathbb{1}\mathbb{1}^T$ from last lecture).

$$aa^T = \overbrace{\begin{bmatrix} | & |\,|\,|\,| \\ \frac{a}{\|a\|_2} & \partial_{n\times(n-1)} \\ | & |\,|\,|\,| \end{bmatrix}}^{U} \overbrace{\begin{bmatrix} \|a\|_2^2 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}}^{\Lambda_U} \underbrace{\begin{bmatrix} - \quad \frac{a^T}{\|a\|_2} \quad - \\ = \quad \partial^T \quad = \end{bmatrix}}_{U^T}$$

matrix ∂ (which is $n\times(n-1)$) has the property that all of its columns are orthogonal to a, or $a^T \partial = 0$, and are also orthogonal to each other.

$$A^TA = (ab^T)^T(ab^T) = ba^Tab^T = \|a\|_2^2\, b b^T$$

$$bb^T = \overbrace{\begin{bmatrix} | & |||| \\ \frac{b}{\|b\|_2} & B_{n\times(n-1)} \\ | & |||| \end{bmatrix}}^{V} \overbrace{\begin{bmatrix} \|b\|_2^2 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}}^{\Lambda_V} \overbrace{\begin{bmatrix} - & \frac{b^T}{\|b\|_2} & - \\ \equiv & B^T & \equiv \end{bmatrix}}^{V^T}$$

B is made of vectors that are orthogonal to b and are orthogonal to each other.

$$\downarrow$$

$$\|b\|_2^2\, aa^T = U \begin{bmatrix} \|a\|_2^2\|b\|_2^2 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} U^T$$

$$\|a\|_2^2\, bb^T = V \begin{bmatrix} \|a\|_2^2\|b\|_2^2 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} V^T$$

$$ab^T = \underbrace{\begin{bmatrix} \frac{a}{\|a\|_2} & \partial \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} \|a\|_2 \, \|b\|_2 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} \frac{b^T}{\|b\|_2} \\ B^T \end{bmatrix}}_{V^T}$$

in particular, $\|ab^T\|_2 = \|a\|_2 \, \|b\|_2$.

check:

$$U \Sigma V^T = U \left[\begin{array}{c|c} \|a\|_2 \|b\|_2 & 0 \\ \hline 0 & O \end{array}\right] \begin{bmatrix} \frac{b^T}{\|b\|_2} \\ B^T \end{bmatrix}$$

$$= U \begin{bmatrix} \|a\|_2 \, b^T \\ 0 \end{bmatrix} = \frac{a}{\cancel{\|a\|_2}} \cancel{\|a\|_2} \, b^T = ab^T.$$

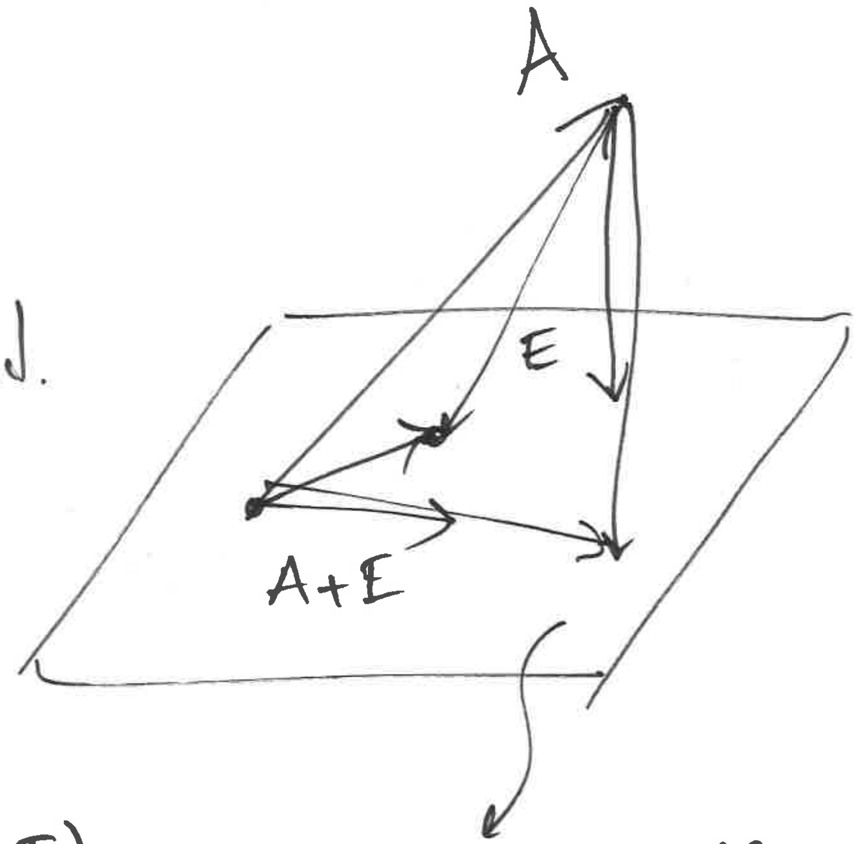
$$\underbrace{A}_{\text{nonsingular}} \longrightarrow \underbrace{A+E}_{\text{singular}}$$

(invertible) (non-invertible)

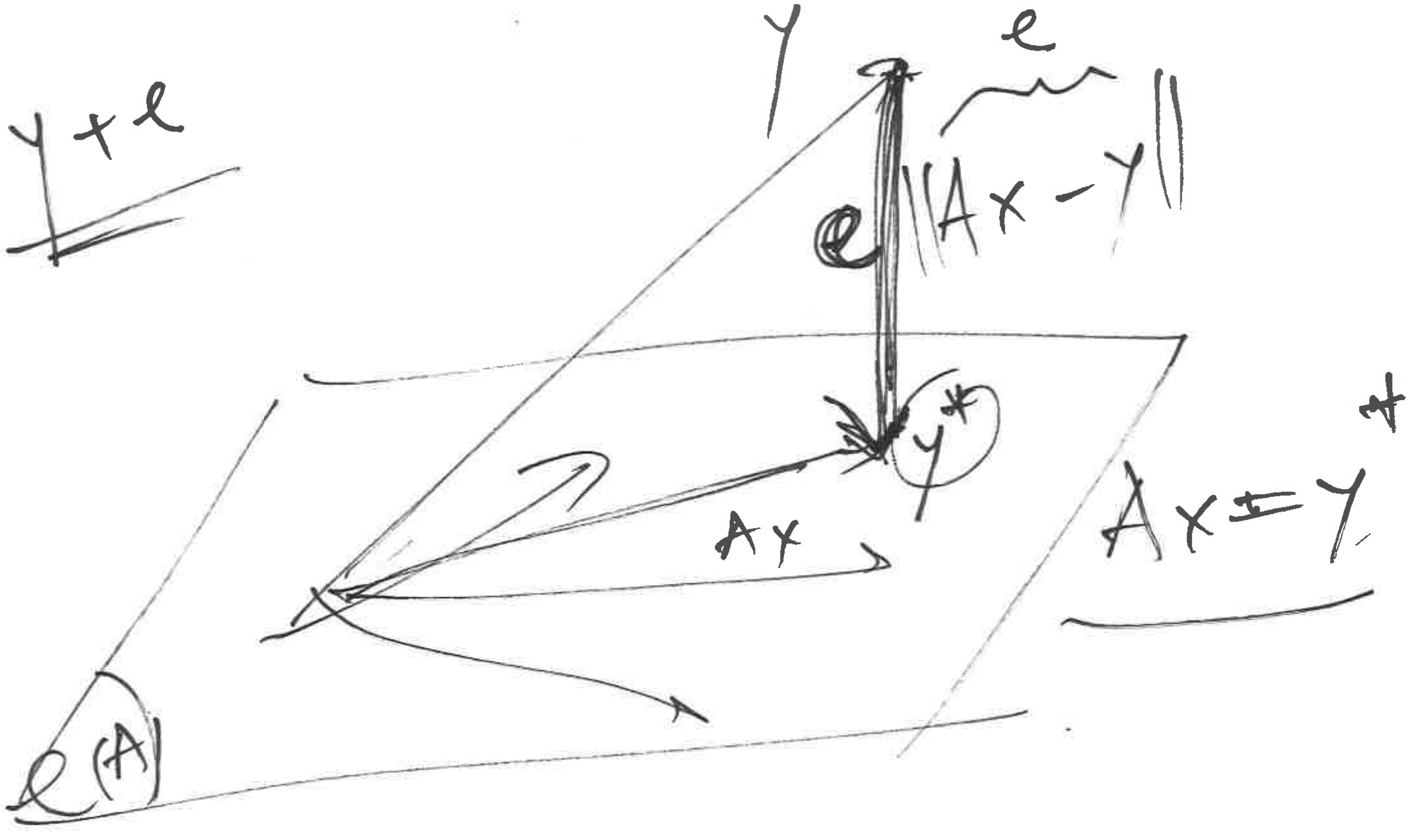
$$\underline{\underline{\|E\|_F}} \qquad \underline{\underline{\|E\|_{2\text{-ind.}}}}$$

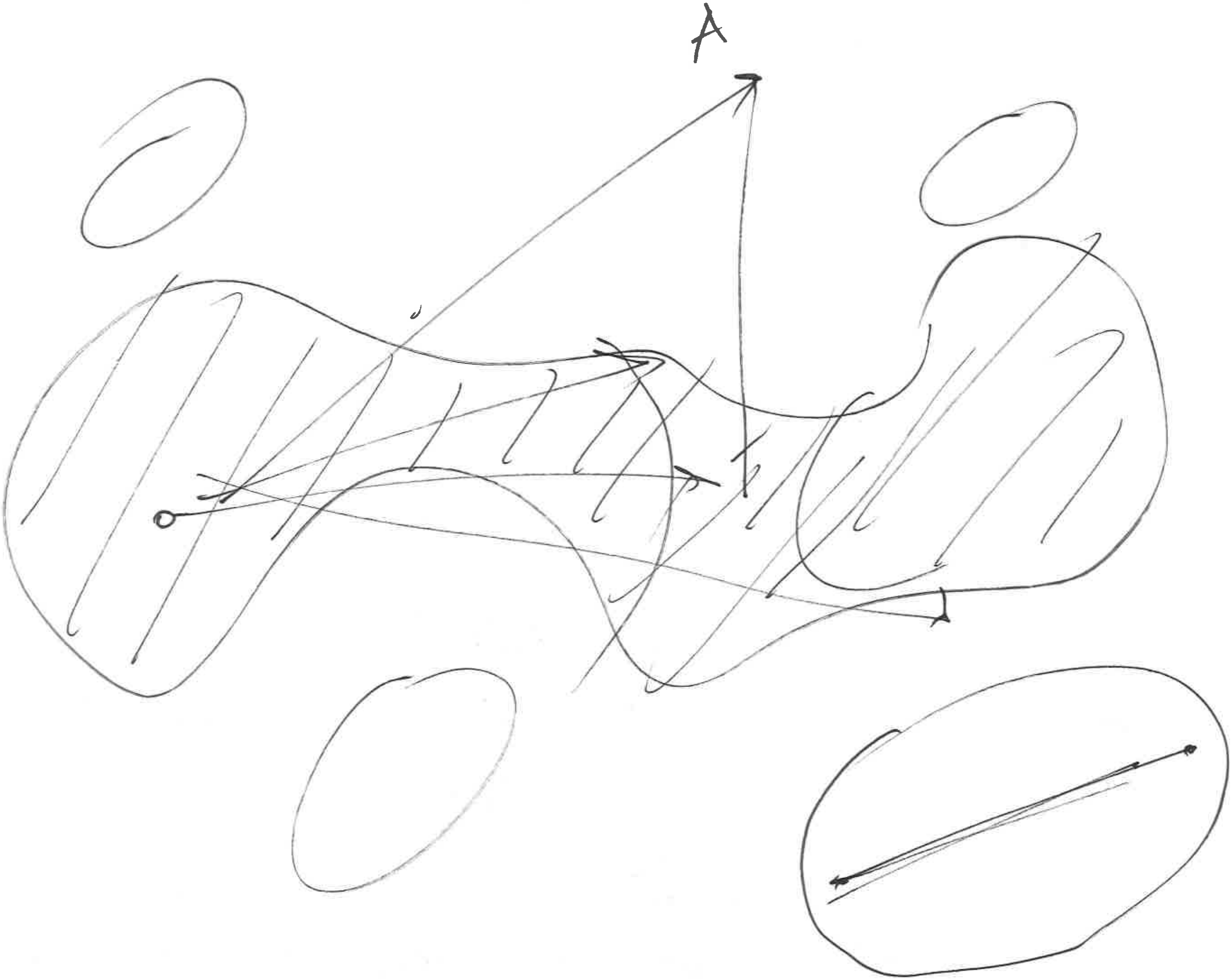
minimize $\|E\|_{2\text{-ind.}}$

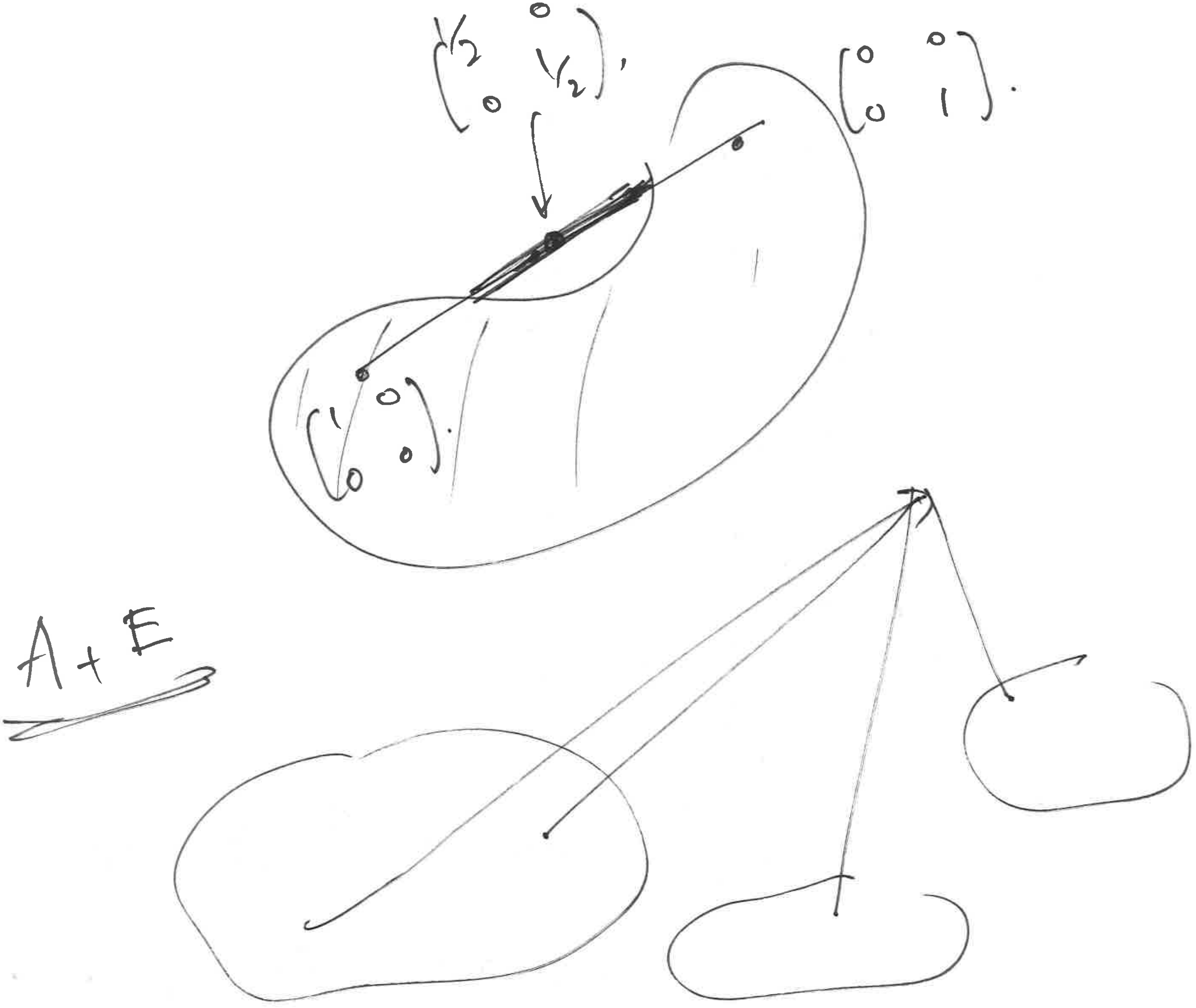
such that $0 \in \operatorname{eig}(A+E)$.



recall least-squares.




A


$\begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$,
$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$.
$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.
A + E

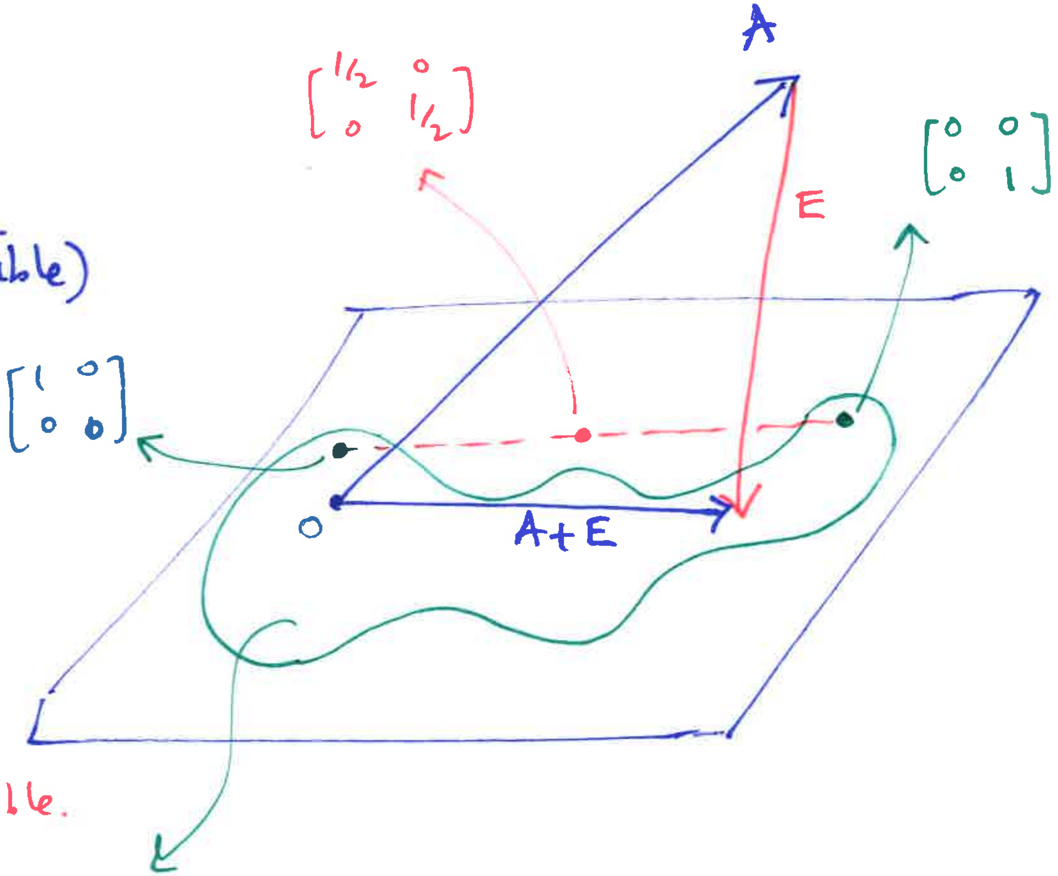
- suppose we'd like to know how close a square invertible matrix is to being noninvertible.

additive perturbation of $A \in \mathbb{R}^{n\times n}$

$$\underbrace{A}_{\substack{\text{nonsingular} \\ \text{(invertible)}}} \longrightarrow \underbrace{A+E}_{\substack{\text{singular} \\ \text{(non-invertible)}}}$$

minimum $\|E\|_{2\text{-ind}}$

such that $\underbrace{0 \in \text{eig}(A+E)}_{\substack{\text{equiv. to} \\ \text{saying noninvertible.}}}$



green set:
set of noninvertible matrices.

- additive matrix perturbations
  - how close is a matrix from being zero ⟶ norm & sigma_max
  - how close is an invertible matrix from being noninvertible?

nonconvex

("hard") problem

& related to sigma_min

$A = \begin{bmatrix} 1 & 0 \\ 100 & 1 \end{bmatrix}.$ A is invertible

$E = -A = \begin{bmatrix} -1 & 0 \\ -100 & -1 \end{bmatrix} \longrightarrow A+E=0$

$E = \begin{bmatrix} -1 & 0 \\ -100 & 0 \end{bmatrix} \longrightarrow A+E = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

$E = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \longrightarrow A+E = \begin{bmatrix} 1 & 0 \\ 100 & 0 \end{bmatrix}.$ $\left(\text{or } E = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}\right)$

$E = \begin{bmatrix} 0 & 0.01 \\ 0 & 0 \end{bmatrix} \longrightarrow A+E = \begin{bmatrix} 1 & 0.01 \\ 100 & 1 \end{bmatrix}.$

$A = \begin{bmatrix} 100 & 100 \\ 100.2 & 100 \end{bmatrix}$

$E = \begin{bmatrix} 0 & 0 \\ -0.2 & 0 \end{bmatrix} \longrightarrow A+E = \begin{bmatrix} 100 & 100 \\ 100 & 100 \end{bmatrix}.$

there is an even "smaller" E!

- example: we want to find the matrix E with smallest norm that makes

$$\begin{bmatrix} 100 & 100 \\ 100.2 & 100 \end{bmatrix} + E \text{ singular.}$$

$$E = \begin{bmatrix} 0 & 0 \\ -0.2 & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 \\ -0.2 \end{bmatrix}}_{a} \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{b^T} \rightarrow \|E\|_2 = \overbrace{\left\| \begin{bmatrix} 0 \\ -0.2 \end{bmatrix} \right\|_2}^{0.2} \cdot \overbrace{\left\| \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\|_2}^{1} = 0.2$$

$$E \neq \begin{bmatrix} 0.05 & -0.05 \\ -0.05 & 0.05 \end{bmatrix} = \underbrace{\begin{bmatrix} 0.05 \\ -0.05 \end{bmatrix}}_{a} \underbrace{\begin{bmatrix} 1 & -1 \end{bmatrix}}_{b^T} \rightarrow \|E\|_2 = 0.05 \left\| \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\|_2 \cdot \left\| \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\|_2$$

$$= 0.05 \left\| \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\|_2^2$$

$$= 0.05 \cdot 2 = 0.1.$$

---

aside:

- any rank-1 matrix can be written as ab^T.

* any non-invertible matrix has a zero singular value

review of last lecture

- eig., det, trace, σ_{max}, σ_{min}, norms, ...

  are all low-dimensional "characterization" of a matrix that has $n \times m$ or n^2 entries/degrees of freedom.

- examples of $\|\cdot\|_{2\text{-ind}}$ & SVD for common matrices.

- matrix with only one nonzero entry in every col. & row

  P & Q elementary (orthogonal)

  $$PAQ = \Sigma \leftrightarrow A = \underbrace{P^T}_{U} \Sigma \underbrace{Q^T}_{V^T} \qquad \& \qquad \|A\|_{2\text{-ind}} = \max_{i,j} |A_{ij}|$$

- rank-one matrix $A = ab^T$

  $$A = \underbrace{\begin{bmatrix} | & ||| \\ \frac{a}{\|a\|_2} & \alpha \\ | & ||| \end{bmatrix}}_{U} \begin{bmatrix} \|a\|_2\|b\|_2 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \underbrace{\begin{bmatrix} - \frac{b^T}{\|b\|_2} - \\ \equiv \beta^T \equiv \end{bmatrix}}_{V^T} \qquad \& \qquad \|A\|_{2\text{-ind}} = \|a\|_2 \|b\|_2$$

  invertible

- find smallest matrix E that makes the matrix $A + E$ singular

- has application in stability of linear dynamical sys. (eigenvalues that pass the imaginary axis & become "unstable").

- other questions:
  - how close is a matrix (not necessarily square), whose columns are linearly independent, to having linearly dependent columns?
  - how close is any given matrix to losing rank?

- suppose $A \in \mathbb{R}^{n\times n}$ has full rank. then

$$\min_{E\in\mathbb{R}^{n\times n}} \left\{ \|E\|_2 \text{ such that } A+E \text{ has rank} < n \right\} = \sigma_{\min}(A)$$

proof: suppose $A+E$ has rank $< n$. then there exists $x \neq 0$, with $\|x\|_2 = 1$, such that $(A+E)x = 0$

$$\begin{aligned}(A+E)x = 0 \rightarrow Ax = -Ex \rightarrow \|Ex\|_2 &= \|Ax\|_2 \\ &\geq \min_{\|y\|_2=1} \|Ay\|_2 \\ &= \sigma_{\min}(A)\end{aligned}$$

$$\|Ex\|_2 \geq \sigma_{\min}(A) \qquad ①$$

$$(\|x\|_2 = 1)$$

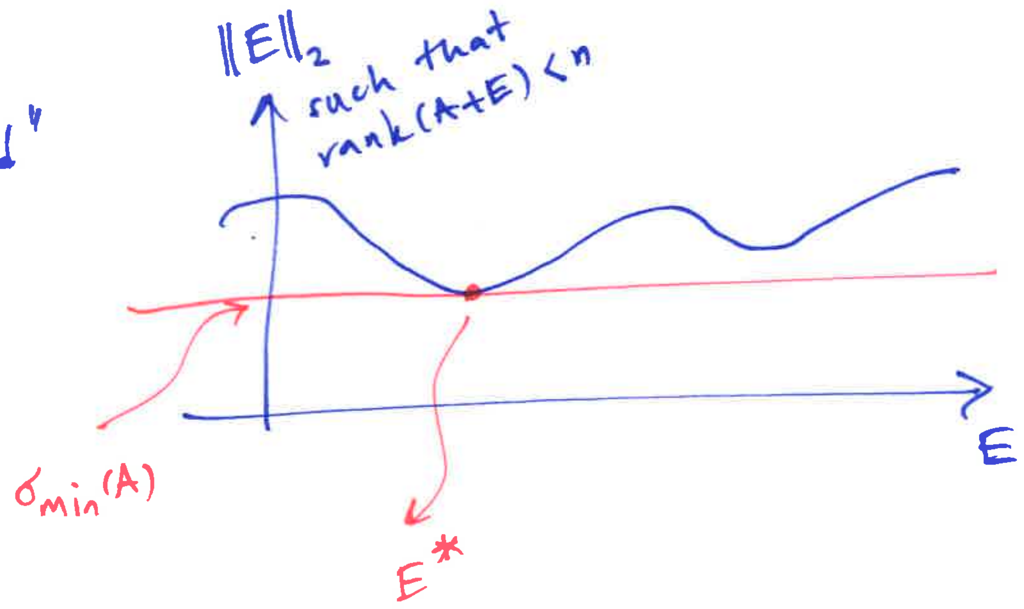
on the other hand

$$\|E\|_2 = \max_{\|y\|_2=1} \|Ey\|_2 \geqslant \|Ex\|_2$$

$$\|E\|_2 \geqslant \|Ex\|_2 \qquad ②$$

① & ② & $\|x\|_2=1$ $\longrightarrow$ $\|E\|_2 \geqslant \|Ex\|_2 = \|Ax\|_2 \geqslant \sigma_{min}(A)$

$\|E\|_{2\text{-ind}}$ is "lower bounded" by $\sigma_{min}(A)$.



now we ~~find~~ propose an E^* that achieves the lower bound

(E^* is such that $\text{rank}(A+E^*) < n$ & $\|E^*\|_2 = \sigma_{min}(A)$)

assume
$A = U \Sigma V^T$
$\sigma_n = \sigma_{min}(A)$

we propose: $E^* = -\sigma_n u_n v_n^T = -U \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \sigma_n \end{bmatrix} V^T$

$$A + E^* = U \Sigma V^T + E^*$$

$$= U\left(\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} - \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \sigma_n \end{bmatrix} \right) V^T = U \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_{n-1} & \\ & & & 0 \end{bmatrix} V^T.$$

rank test:

$x^* = v_n$ ($\|v_n\| = 1$, therefore legit choice for x)

$$(A+E^*)x^* = U \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_{n-1} & \\ & & & 0 \end{bmatrix} V^T v_n = U \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_{n-1} & \\ & & & 0 \end{bmatrix} \begin{bmatrix} - v_1^T - \\ \vdots \\ - v_n^T - \end{bmatrix} v_n$$

$$= U \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_{n-1} & \\ & & & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = 0 \quad \checkmark$$

norm test:

$$\|E^*\|_{2\text{-ind}} = \left\| -U \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \sigma_n \end{bmatrix} V^T \right\|_{2\text{-ind}}$$

$$= \left\| \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \sigma_n \end{bmatrix} \right\|_{2\text{-ind}} = \sigma_n = \sigma_{\min}(A) \quad \checkmark$$

(other method $E^* = -\sigma_n u_n v_n^T \longrightarrow \|E^*\| = \sigma_n \cancelto{1}{\|u_n\|} \cancelto{1}{\|v_n\|} = \sigma_n$).

- example: $A = \begin{bmatrix} 100 & 100 \\ 100.2 & 100 \end{bmatrix}$. what is the smallest perturbation δA such that $A+\delta A$ is rank deficient.

one possibility $\delta A = \begin{bmatrix} 0 & 0 \\ -0.2 & 0 \end{bmatrix} \to \|\delta A\|_2 = 0.2$.

$$A = \underbrace{\begin{bmatrix} 0.7068 & 0.7075 \\ 0.7075 & -0.7068 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 200.1 & 0 \\ 0 & 0.1 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 0.7075 & 0.7068 \\ -0.7068 & 0.7075 \end{bmatrix}}_{V^T}$$

$$(\delta A)^* = -U \begin{bmatrix} 0 & 0 \\ 0 & 0.1 \end{bmatrix} V^T \approx \begin{bmatrix} 0.05 & -0.05 \\ -0.05 & 0.05 \end{bmatrix}. \qquad \|\delta A^*\|_2 = 0.1.$$

- using the SVD, we can show that both of the problems

minimize $\left\{ \|E\|_2 , \|E\|_F^2 \right\}$ → hw.

subject to $\text{rank}(A+E) < n$, $\text{rank}(A) = n$

have the solution $E^* = -\sigma_n u_n v_n^T$.

(& an x^* that gives $(A+E^*)x^* = 0$ is $x^* = v_n$.)

min. $\|E\|_2$

s.t. $(A+E)x = 0$

$\|x\|_2 = 1$

↓

$E^* = -\sigma_n u_n v_n^T$

$x^* = v_n$

- low-rank approximation: let $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank}(A) = r$.

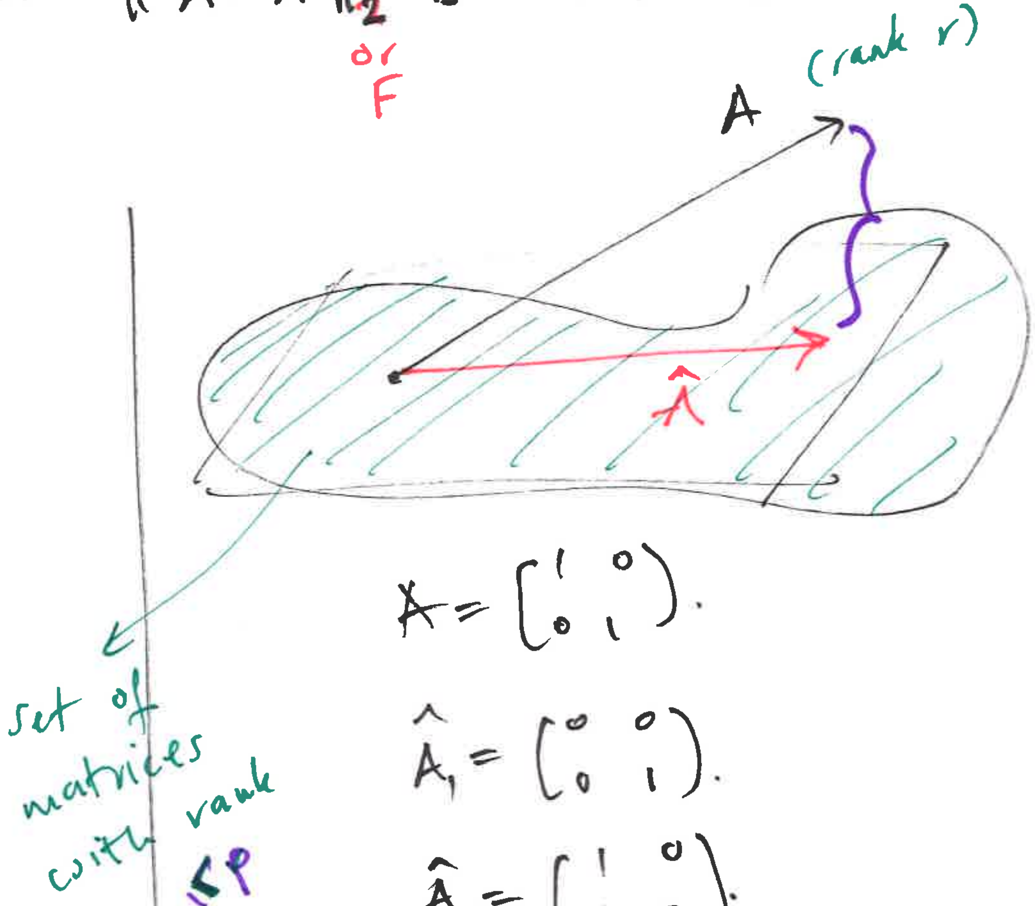
we seek the matrix $\hat{A}$, $\operatorname{rank}(\hat{A}) \le p < r$, such that $\hat{A} \approx A$ in the sense that $\| A - \hat{A} \|_2$ is minimized

or F

the solution is

$$\hat{A} = \sum_{i=1}^{p} \sigma_i u_i v_i^T$$

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



$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

$$\hat{A}_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

$$\hat{A}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

$$\frac{\hat{A}_1 + \hat{A}_2}{2} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

the solution is $\hat{A} = \sum_{i=1}^{p} \sigma_i u_i v_i^T$

$$= U \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_p & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} V^T . \leftarrow \text{"top part"}$$

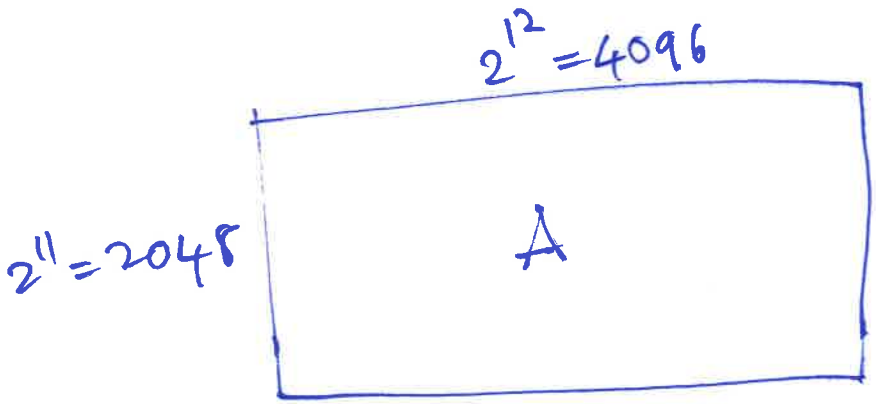
and the corresponding value of $\|A - \hat{A}\|$ is σ_{p+1}.

$$A - \hat{A} = U \begin{bmatrix} 0 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & & & \\ & & & \sigma_{p+1} & & & \\ & & & & \ddots & & \\ & & & & & \sigma_r & \\ & & & & & & 0 \\ & & & & & & & \ddots \\ & & & & & & & & 0 \end{bmatrix} V^T$$

$$\|A - \hat{A}\|_2 = \sigma_{p+1}$$

$$\hat{A} = \underbrace{\begin{pmatrix} | & & | \\ u_1 & \cdots & u_m \\ | & & | \end{pmatrix}}_{m \times p} \underbrace{\begin{pmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_p & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix}}_{p \times p} \underbrace{\begin{pmatrix} - v_1^T - \\ \vdots \\ - v_n^T - \end{pmatrix}}_{p \times n} = \underbrace{\sigma_1 u_1 v_1^T + \cdots + \sigma_p u_p v_p^T}_{m \times n} + \cancel{\cdots}$$

- example : image compression



$2^{12} 2^{11} = 2^{23}$ ↓ size of original matrix

$\hat{A}_1 = \sigma_1 u_1 v_1^T$

$1 + 2^{11} + 2^{12} = 6145$. size of rank-1 approximation.

reduction in data (factor of). $\approx \dfrac{2^{23}}{2^{11} + 2^{12}} \geqslant 1365$.

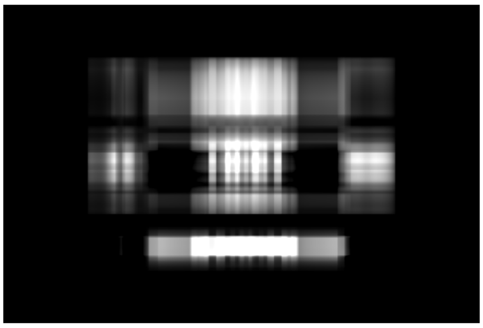
↓ compression factor

$\hat{A}_{10} = \sigma_1 u_1 v_1^T + \cdots + \sigma_{10} u_{10} v_{10}^T$

$(6145) \cdot 10$.

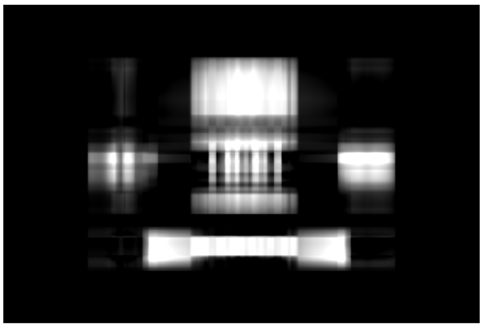
compression factor $\geqslant 136$.

Two singular values (Rank 2 approx):

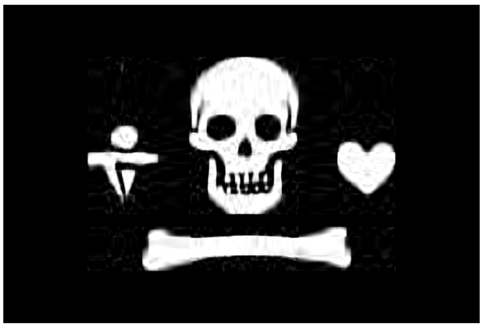


Ten singular values:

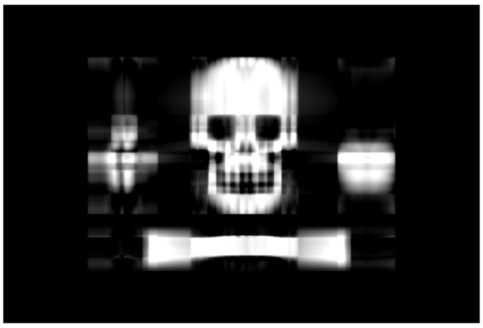
Three singular values:



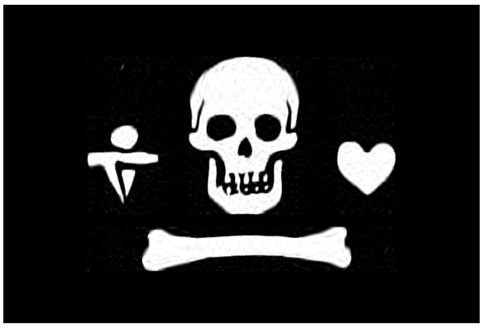
Twenty singular values:



Five singular values:

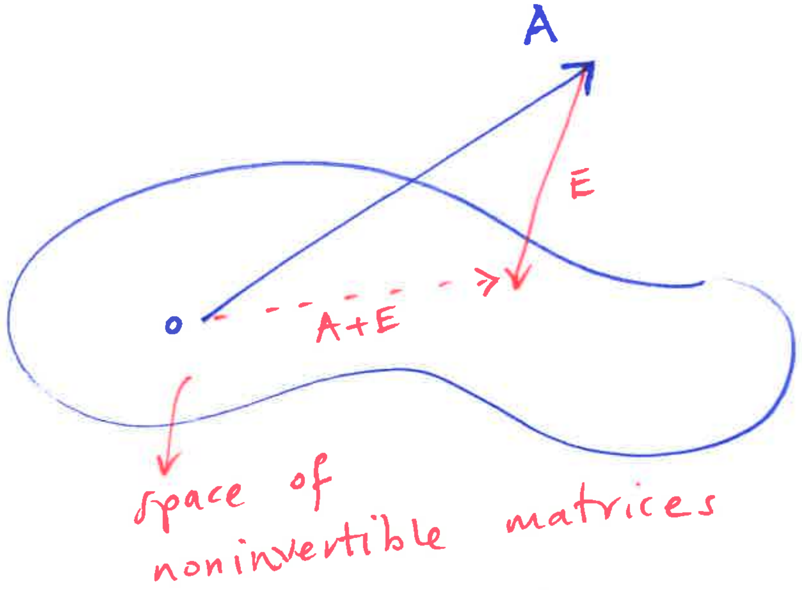


Forty singular values:



review of last lecture

- given an invertible matrix A, what is the "smallest" perturbation E such that A+E is not invertible?

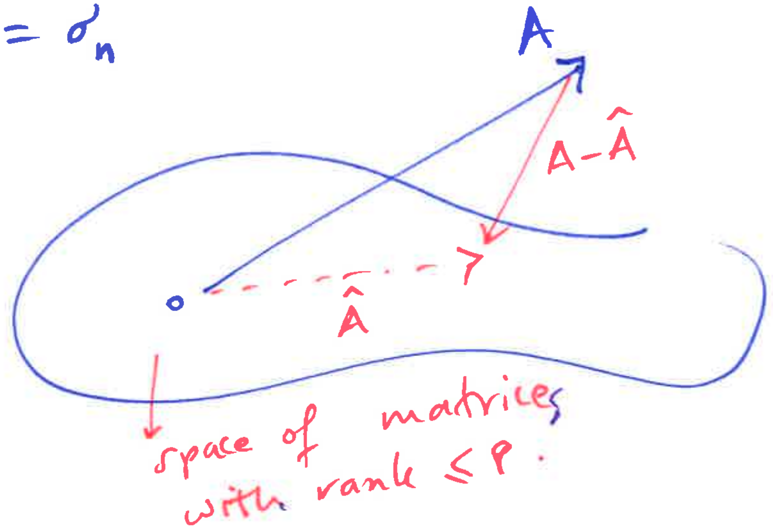


space of noninvertible matrices

$$A = U \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_p & & \\ & & & \ddots & \\ & & & & \sigma_n \end{bmatrix} V^T$$

$$E^* = -U \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \sigma_n \end{bmatrix} V^T \quad \& \quad \|E^*\|_{2\text{-ind}} = \sigma_n$$

- given matrix A of rank r, what is the "best" rank p matrix $\hat{A}$, p<r, that approximates the matrix A (such that $\|A-\hat{A}\|_{2\text{-ind}}$ is minimized)?



space of matrices with rank ≤ p.

$$\hat{A} = U \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_p & & \\ & & & 0 & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} V^T \quad \& \quad \|A-\hat{A}\|_{2\text{-ind}} = \sigma_{p+1}$$

- motivating example: solving $Ax=y$.

$$A = \begin{bmatrix} 100 & 100 \\ 100.2 & 100 \end{bmatrix} \longrightarrow A^{-1} = \begin{bmatrix} -5 & 5 \\ 5.01 & -5 \end{bmatrix}.$$

$$\delta A = \begin{bmatrix} 0 & 0 \\ -0.1 & 0 \end{bmatrix} \longrightarrow (A+\delta A)^{-1} = \begin{bmatrix} -10 & 10 \\ 10.01 & -10 \end{bmatrix}.$$

a 0.1% change in A results in a 100% change in A^{-1}.

let $y = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \longrightarrow x = A^{-1}y = \begin{bmatrix} -10 \\ 10.01 \end{bmatrix}.$

$$x + \delta x = (A+\delta A)^{-1} y = \begin{bmatrix} -20 \\ 20.01 \end{bmatrix}.$$

Note that the “data” is encapsulated in both the matrix A and the vector y. And both are subject to errors/perturbations. In this slide we considered perturbations to A; in the next slide we consider perturbations to y. Ultimately, our aim is to find a bound on dx as a function of the sizes of A, dA, y, dy.

- sensitivity of linear equations to errors in data.

$A \in \mathbb{R}^{n\times n}$, A invertible, $y = Ax \Rightarrow x = A^{-1}y.$

$y \to y+\delta y \Rightarrow x \to x+\delta x$ $\Big($ $x+\delta x = A^{-1}(y+\delta y)$

$$\Downarrow$$

$\delta x = A^{-1}\delta y \Big)$.

$$\|\delta x\|_2 = \|A^{-1}\delta y\|_2 \;\boxed{\leq}\; \underbrace{\|A^{-1}\|_2}\, \|\delta y\|_2 .$$

hw problem

$A = U\Sigma V^T$

$$\boxed{A^{-1} = V\Sigma^{-1}U^T}$$

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_n \end{bmatrix}$$

$\|A^{-1}\|_2 = \dfrac{1}{\sigma_{\min}(A)}$

assuming A is invertible.

(A : full rank

$\downarrow$

all sing. values $\neq 0$.)

- when is $\|\delta x\|_2$ "big"? when is $\|\delta y\|_2$ "big"?

- when is $\|A^{-1}\|_2$ "big"?

suppose we replace A with $10^{-3} \cdot A$:

$\|\delta x\|_2 \leq \|A^{-1}\|_2 \, \|\delta y\|_2$ suggests that the upper bound on $\|\delta x\|_2$ increases 1000 times; but the linear dependence/independence of the columns of A has not changed.

- a better indicator of error, considers "relative" errors in x & y:

$$\frac{\|\delta x\|_2}{\|x\|_2} \leqslant ?\ \frac{\|\delta y\|_2}{\|y\|_2}$$

$$\begin{cases} \delta x = A^{-1}\,\delta y \longrightarrow \|\delta x\|_2 \leqslant \|A^{-1}\|_2\,\|\delta y\|_2 & ① \\ y = A\,x \longrightarrow \|y\|_2 \leqslant \|A\|_2\,\|x\|_2 & ② \end{cases}$$

① $\longrightarrow \dfrac{\|\delta x\|_2}{\|x\|_2} \leqslant \|A^{-1}\|_2\,\dfrac{\|\delta y\|_2}{\|x\|_2}$ ③

② $\longrightarrow \dfrac{1}{\|x\|_2} \leqslant \|A\|_2\,\dfrac{1}{\|y\|_2} \longrightarrow \dfrac{\|\delta y\|_2}{\|x\|_2} \leqslant \|A\|_2\,\dfrac{\|\delta y\|_2}{\|y\|_2}$ ④

③ & ④ $\longrightarrow \dfrac{\|\delta x\|_2}{\|x\|_2} \leqslant \boxed{\|A^{-1}\|_2\,\|A\|_2}\,\dfrac{\|\delta y\|_2}{\|y\|_2}.$

$$\kappa(A) := \|A\|_2 \, \|A^{-1}\|_2 = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}.$$

"condition number."

A is "poorly conditioned" if $\kappa(A)$ is "large."

- recall example: $A = \begin{bmatrix} 100 & 100 \\ 100.2 & 100 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 200.1 & 0 \\ 0 & 0.1 \end{bmatrix}$

$$\kappa(A) = \frac{200.1}{0.1} = 2001$$

aside: if we multiply A by a scalar, its cond. numb. will not change.

- it can be shown

$A \to A + \delta A$

$\delta(A^{-1}) = (A+\delta A)^{-1} - A^{-1}.$

$= -A^{-1}\delta A A^{-1} + \text{h.o.t.}$

$$\frac{\|\delta(A^{-1})\|_2}{\|A^{-1}\|_2} \leq \underbrace{\|A\|_2 \|A^{-1}\|_2}_{\kappa(A)} \frac{\|\delta A\|_2}{\|A\|_2}$$

- in fact, in the most general case, we have

$$\frac{\|\delta x\|_2}{\|x\|_2} \leqslant \underbrace{\|A\|_2 \, \|A^{-1}\|_2}_{\kappa(A).} \underbrace{\left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\delta y\|_2}{\|y\|_2} \right)}$$

(perturbations may occur in both the data y & the matrix A.)

$Ax = y.$

$1 \leqslant \kappa(A)$ (but no upper bound)

review of last lecture

- sensitivity analysis of solution of $Ax=y$ when A is invertible (square & full rank)

$$\frac{\|\delta x\|_2}{\|x\|_2} \leqslant \underbrace{\|A\|_2 \, \|A^{-1}\|_2}_{\kappa(A)} \left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\delta y\|_2}{\|y\|_2} \right)$$

$$\|A\|_2 = \sigma_1 \qquad \|A^{-1}\|_2 = \left\| \begin{bmatrix} 1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_n \end{bmatrix} \right\|_2 = \frac{1}{\sigma_n}$$

$$\kappa(A) := \|A\|_2 \, \|A^{-1}\|_2 = \frac{\sigma_{max}}{\sigma_{min}}$$

$$\Sigma = \begin{bmatrix} \bigcirc^{\uparrow} & & \\ & \ddots & \\ & & {}_{\downarrow}\circ \end{bmatrix}$$

- introduction to optimization.

- consider the problem (with differentiable f)

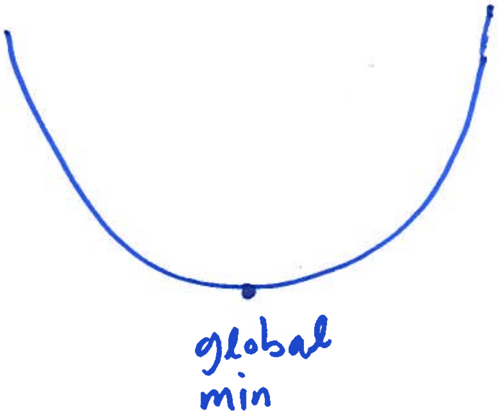
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \qquad\qquad (f(x) \in \mathbb{R})$$

- x^* is a global minimizer if $f(x^*) \leq f(x)$ for all x.
- x^* is a local minimizer if $\underbrace{f(x^*) \leq f(x) \text{ for all } x \in \mathcal{N}}$

neighborhood of x^*

$$\Updownarrow$$

$$f(x^*) \leq f(x^* + \delta x) \text{ for } \delta x \text{ small}$$



- suppose x^* is a local minimizer, $f(x^* + \delta x) \geqslant f(x^*)$.

$$f(x^* + \delta x) = f(x^*) + (\nabla f(x^*))^T \delta x + \text{h.o.t. (in } \delta x)$$

since for small enough δx, this term dominates all these

then necessarily

$$(\nabla f(x^*))^T \delta x \geqslant 0 \quad \text{for all } \delta x \in \mathbb{R}^n$$

- first-order necessary conditions for optimality

$$f(x^*+\delta x) = f(x^*) + (\nabla f(x^*))^T \delta x + \text{h.o.t.}$$

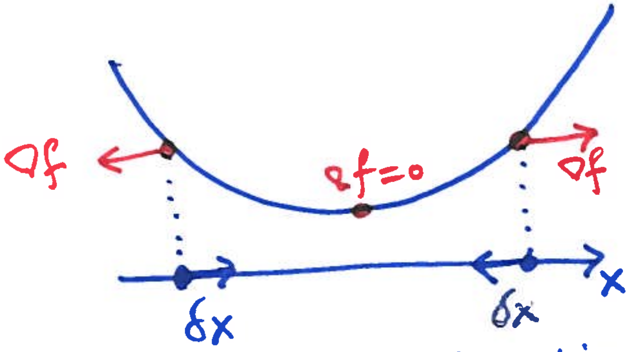
our claim is that $\nabla f(x^*) = 0$. suppose that $\nabla f(x^*) \neq 0$.

then take $\delta x = -\varepsilon \nabla f(x^*)$, $\varepsilon > 0$

$$(\nabla f(x^*))^T \delta x = -\varepsilon \|\nabla f(x^*)\|^2 < 0$$

which contradicts $(\nabla f(x^*))^T \delta x \geq 0$. therefore

if x^* is a local minimizer of f then $\boxed{\nabla f(x^*) = 0}$



(∇f always points in direction of maximal increase in f.)

- second-order necessary conditions for optimality

$$f(x^* + \delta x) = f(x^*) + (\nabla f(x^*))^T \delta x + \frac{1}{2} \delta x^T (\nabla^2 f(x^*)) \delta x + \text{h.o.t.}$$

we know from first-order conditions that $\nabla f(x^*) = 0$. therefore if $\nabla^2 f(x^*)$ not a positive semidefinite matrix then there exists δx such that

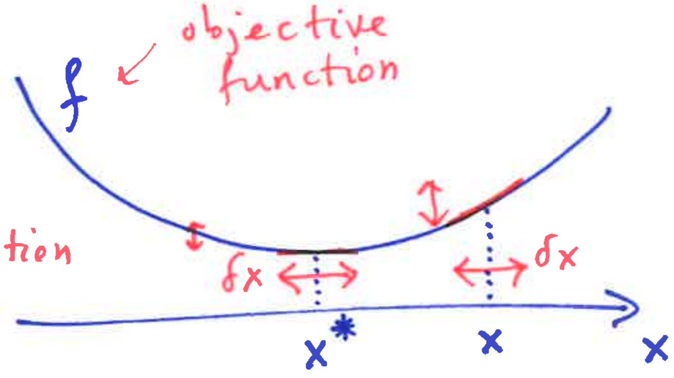
$$\delta x^T (\nabla^2 f(x^*)) \delta x < 0$$

which contradicts $f(x^* + \delta x) \geqslant f(x^*)$. therefore

if x^* is a local minimizer of f then
$$\boxed{\begin{aligned} \nabla f(x^*) &= 0 \\ \nabla^2 f(x^*) &\succcurlyeq 0 \end{aligned}}$$

first- & second-order (necessary) conditions for optimality

(for unconstrained optimization problems)



x: scalar, f: scalar-valued

$$f(x+\delta x) = f(x) + \underbrace{f'(x)\,\delta x}_{\delta f \text{ is first-order variation of } f \text{ at } x} + \overbrace{\tfrac{1}{2} f''(x)\,\delta x^2}^{\delta^2 f \text{ is second-order variation of } f \text{ at } x} + \text{h.o.t.}$$

$$f'(x^*) = 0$$

$$f''(x^*) > 0$$

$$\left\{\begin{aligned} f'(x^*) &= 0 \leftrightarrow \delta f = 0 \quad \forall \delta x \\ f''(x^*) &> 0 \leftrightarrow \delta^2 f > 0 \quad \forall \delta x \end{aligned}\right.$$

x: vector, f: scalar-valued

$$f(x+\delta x) = f(x) + \underbrace{(\nabla f(x))^T \delta x}_{\delta f} + \underbrace{\tfrac{1}{2} (\delta x)^T (\nabla^2 f(x))\, \delta x}_{\delta^2 f} + \text{h.o.t.}$$

$\nabla f(x)$: "gradient" of f at x

$\nabla^2 f(x)$: "Hessian" of f at x.

aside:
gradient of f at x can be defined using <Df(x),dx> = d/dt f(x+tdx) with derivative on right hand side being evaluated at t=0

$$\delta f = \frac{\partial f}{\partial x_1}\delta x_1 + \cdots + \frac{\partial f}{\partial x_n}\delta x_n$$

$$= (\nabla f)^T \delta x$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}_{n\times 1} \qquad \delta x = \begin{bmatrix} \delta x_1 \\ \vdots \\ \delta x_n \end{bmatrix}_{n\times 1}$$

$$\delta^2 f = \frac{1}{2}\sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j}\delta x_i\, \delta x_j$$

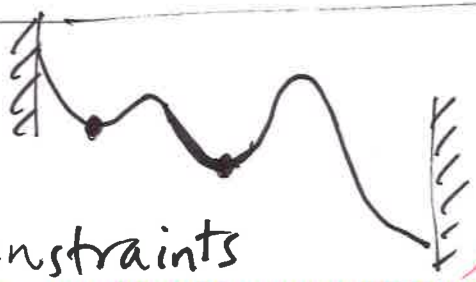
$$= \frac{1}{2}(\delta x)^T(\nabla^2 f)\delta x$$

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & & \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & & \\ & & \ddots & \\ & & & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}_{n\times n}$$

$$\nabla^2 f = (\nabla^2 f)^T$$

$$\begin{cases} \delta f = 0 \quad \forall \delta x \neq 0 \quad \longleftrightarrow \quad \nabla f(x^*) = 0 \\ \delta^2 f > 0 \quad \forall \delta x \neq 0 \quad \longleftrightarrow \quad \nabla^2 f(x^*) > 0 \end{cases}$$

"necessary" cond.



constraints

X: matrix, f: scalar-valued

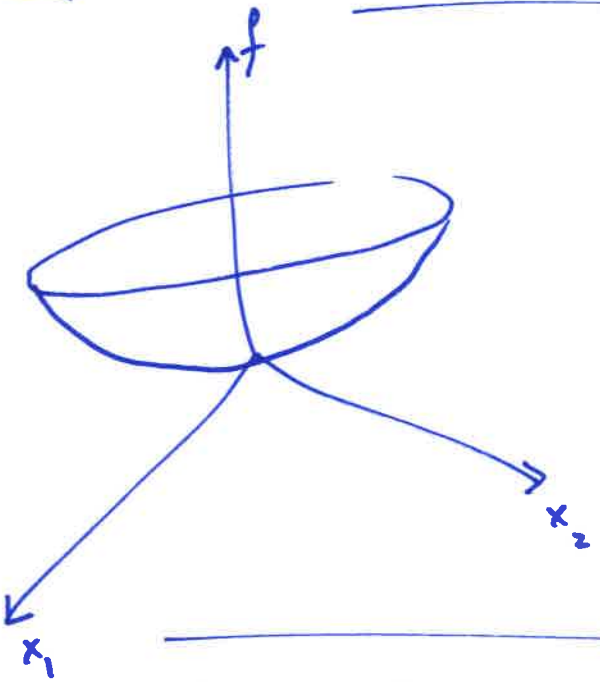
$$f(X+\delta X) = f(X) + \underbrace{\mathrm{tr}((\nabla f)^T \delta X)}_{\delta f} + \underbrace{\tfrac{1}{2}\mathrm{tr}(H(\delta X, X)\,\delta X)}_{\delta^2 f} + \text{h.o.t.}$$

∇f : a matrix $H(\delta X, X)$: a matrix & linear in δX.

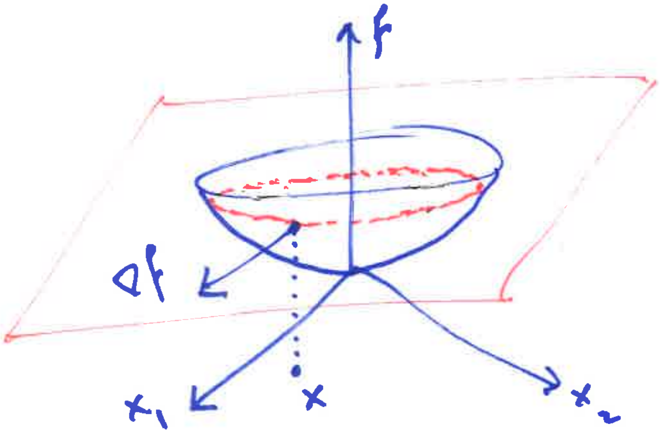
- example: $f(x_1, x_2) = x_1^2 + x_2^2$

$$\nabla f = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \qquad \nabla^2 f = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

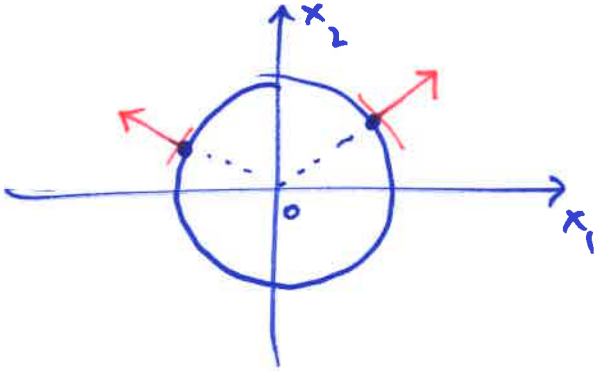
$$\nabla f = 0 \longrightarrow \begin{cases} x_1 = 0 \\ x_2 = 0 \end{cases}$$ unique minimum point of f.



- interpretation of ∇f: the gradient always points in the direction of maximum ascent.



$$\nabla f = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$



- procedure for finding (local) optima of a scalar-valued function, $f(x)$, is as follows:

  - set $x \rightarrow x+\delta x$
  - expand $f(x+\delta x)$, in δx, around x.
    (i.e., write Taylor expansion of f at x in terms of δx)
  - identify the δf & $\delta^2 f$ terms, and extract ∇f & $\nabla^2 f$.
  - set $\nabla f(x)=0$ to find x^*, and check $\nabla^2 f(x^*) \succ 0$.

(when the expression for f is simple or given explicitly in terms of x_i, then direct computation of $\frac{\partial f}{\partial x_i}$ & $\frac{\partial^2 f}{\partial x_i \partial x_j}$ may be easier way of getting ∇f & $\nabla^2 f$, as in example on previous slide.)

(the gradient of f is a column vector of the same dimension as x)

$$J(x+\delta x) = \sum_{i=1}^{m} e^{a_i^T(x+\delta x) + b_i} + c^T(x+\delta x)$$

$$= \sum_{i=1}^{m} e^{a_i^T x + b_i} e^{a_i^T \delta x} + c^T x + c^T \delta x$$

Now recall that $e^{\varepsilon} \approx 1 + \varepsilon + \text{h.o.t.}$ for small ε.

Therefore

$$J(x+\delta x) \approx \sum_{i=1}^{m} e^{a_i^T x + b_i}(1 + a_i^T \delta x + \text{h.o.t.}) + c^T x + c^T \delta x.$$

$$\downarrow$$

$$\delta J = \sum_{i=1}^{m} e^{a_i^T x + b_i} a_i^T \delta x + c^T \delta x$$

$$\downarrow$$

$$\nabla J = \sum_{i=1}^{m} e^{a_i^T x + b_i} a_i + c.$$

- example: least-squares, i.e., minimize $J(x)$

$$J(x) = \|Ax-b\|_2^2 = (Ax-b)^T(Ax-b)$$

where we assume that the columns of A are lin. ind.

$$
\begin{aligned}
J(x+\delta x) &= (Ax + A\delta x - b)^T(Ax + A\delta x - b) \\
&= (Ax-b)^T(Ax-b) + (A\delta x)^T(Ax-b) \\
&\quad + (Ax-b)^T(A\delta x) + (A\delta x)^T(A\delta x) \\
&= \underbrace{(Ax-b)^T(Ax-b)}_{J(x)} + \underbrace{2(Ax-b)^TA\,\delta x}_{\delta J(x)} + \underbrace{(\delta x)^TA^TA\,\delta x}_{\delta^2 J(x)}
\end{aligned}
$$

aside:

$$\boxed{y^Tz = z^Ty}$$

$$\delta J = (\nabla J)^T\delta x = (2A^T(Ax-b))^T\delta x \longrightarrow \nabla J = 2A^T(Ax-b).$$

$$\delta^2 J = \frac{1}{2}(\delta x)^T\nabla^2 J\,\delta x = \frac{1}{2}(\delta x)^T 2A^TA\,\delta x \longrightarrow \nabla^2 J = 2A^TA.$$

(note that here J(x) is a quadratic function of x and therefore J(x+dx) will have at most second-order terms in dx; thus no Taylor expansion is necessary)

$$\nabla J(x^*) = 0 \rightarrow A^TAx^* - A^Tb = 0 \rightarrow A^TAx^* = A^Tb \rightarrow \boxed{x^* = (A^TA)^{-1}A^Tb}$$

assuming A^TA invertible

$$\nabla^2 J(x^*) \overset{?}{\succ} 0 \quad \leftrightarrow \quad \text{is } A^TA \text{ a pos. def. mat. ?}$$

$$\leftrightarrow \quad z^T(A^TA)z > 0 \quad \forall z \neq 0 \text{ ?}$$

$$z^TA^TAz = (Az)^T(Az) = \|Az\|_2^2 \geq 0 \rightarrow \left\{ \begin{array}{l} \|Az\|_2^2 = 0 \\ \|Az\|_2^2 > 0 \end{array} \right.$$

$$\|Az\|_2^2 = 0 \leftrightarrow Az = 0$$

$\downarrow$ this contradicts the lin. ind. of col. of A

$$z^TA^TAz > 0 \quad \forall z \rightarrow A^TA \succ 0$$

& in particular, A^TA is invertible.

- example: weighted least-squares

$$J(x) = \| W(Ax-b) \|_2^2$$

$$W = \begin{bmatrix} \omega_1 & & \\ & \ddots & \\ & & \omega_n \end{bmatrix} \succ 0 \qquad e = Ax - b = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$$J(x) = \omega_1^2 e_1^2 + \cdots + \omega_n^2 e_n^2$$

$$J(x) = (Ax-b)^T W^2 (Ax-b) \longrightarrow \boxed{x^* = (A^T W^2 A)^{-1} A^T W^2 b.}$$

<u>aside</u>: this can be generalized to the case where W is not necessarily diagonal & is only restricted to being a pos. def. mat.

- example: a more general form of the sensor fusion problem

$$y = Cx + v$$

$$\underset{\substack{\uparrow \\ \text{measurements}}}{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}} = \begin{bmatrix} & & \\ & C_{n \times m} & \\ & & \end{bmatrix} \underset{\substack{\uparrow \\ \text{unknowns}}}{\begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}} + \underset{\substack{\uparrow \\ \text{measurement errors}}}{\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}}$$

$$\mathcal{E}\{v_i\} = 0$$

$$\mathcal{E}\{v_i\} = \sigma_i^2$$

find K such that $\hat{x} = Ky$ and $\mathcal{E}\{(x-\hat{x})^2\}$ is minimized.

$$\boxed{\hat{x} = (C^T Q^{-1} C)^{-1} C^T Q^{-1} y}$$

$$Q = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix}$$

aside:

---

the special case we considered last lecture corresponds to $c = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

review of last lecture

- (necessary) conditions for optimality $\begin{cases} \nabla J(x^*) = 0 \\ \nabla^2 J(x^*) \succ 0 \end{cases}$

  $J : \mathbb{R}^n \to \mathbb{R}$ if x vector variable

  $J : \mathbb{R}^{m \times n} \to \mathbb{R}$ if X matrix variable

- finding ∇J & $\nabla^2 J$ $\begin{cases} \text{expand } J(x+\delta x) \text{ in } \delta x \text{ around } x \\ \text{extract } \delta J \text{ \& } \delta^2 J \end{cases}$

↓

(if J explicitely given in terms of x_i, then ∇J & $\nabla^2 J$ can be computed directly)

$\delta J = (\nabla J)^T \delta x$ if x vector

$\delta J = \mathrm{tr}((\nabla J)^T \delta X)$ if X matrix

$\delta^2 J = \frac{1}{2} (\delta x)^T \nabla^2 J \, \delta x$ if x vector

$\delta^2 J = \frac{1}{2} \mathrm{tr}(\text{quadratic in } \delta X)$ if X matrix.

- ∇J orthogonal to level curve of J and belongs to same space as x.

(don't confuse tangent plane $\delta z = (\nabla J)^T \delta x$, i.e., $z - z_0 = (\nabla J)^T (x - x_0)$, with ∇J.)

- example: system identification.

$$x(k+1) = Ax(k) + Bu(k) + v(k)$$

"mismatch" $v(k)$ unknown. we would like to find A & B

assuming x & u are known, for all time.

$$x(k+1) - Ax(k) - Bu(k) = v(k).$$

$$\begin{aligned} J(A,B) = & \left\| x(1) - Ax(0) - Bu(0) \right\|_2^2 \\ & + \left\| x(2) - Ax(1) - Bu(1) \right\|_2^2 \\ & + \qquad \vdots \\ & + \left\| x(K) - Ax(K-1) - Bu(K-1) \right\|_2^2 \end{aligned}$$

aside: the Frobenius norm has the following property

$$\|Z\|_F^2 = \left\| \begin{bmatrix} | & & | \\ z_1 & \cdots & z_n \\ | & & | \end{bmatrix} \right\|_F^2 = \|z_1\|_2^2 + \cdots + \|z_n\|_2^2 .$$

$$J = \left\| \begin{bmatrix} | & & | \\ x(1) - Ax(0) - Bu(0) & \cdots & x(K) - Ax(K-1) - Bu(K-1) \\ | & & | \end{bmatrix} \right\|_F^2$$

$$= \left\| \underbrace{\begin{bmatrix} | & & | \\ x(1) & \cdots & x(K) \\ | & & | \end{bmatrix}}_{\Psi^T} - \underbrace{\begin{bmatrix} A & B \end{bmatrix}}_{X^T} \underbrace{\begin{bmatrix} | & & | \\ x(0) & & x(K-1) \\ | & \cdots & | \\ u(0) & & u(K-1) \\ | & & | \end{bmatrix}}_{\Phi^T} \right\|_F^2$$

$$= \|\Psi^T - X^T \Phi^T\|_F^2$$

$$= \|\Psi - \Phi X\|_F^2 .$$

where the optimization variable is the matrix

$$X = \begin{bmatrix} A^T \\ B^T \end{bmatrix},$$

and Φ & Ψ are fully known.

$$J(X) = \| \Phi X - \Psi \|_F^2$$

$$J(X+\delta X) = tr([\Phi X - \Psi + \Phi \delta X]^T [\Phi X - \Psi + \Phi \delta X])$$

$$\delta J = tr([\Phi X - \Psi]^T \Phi \delta X) + tr([\Phi \delta X]^T [\Phi X - \Psi])$$

$$= \boxed{2}\, tr(\boxed{[\Phi X - \Psi]^T \Phi}\, \delta X)$$

$$= tr(\{2 \Phi^T [\Phi X - \Psi]\}^T \delta X)$$

$$\nabla J = 2 \Phi^T (\Phi X - \Psi)$$

18-4

$$\delta J = 0 \quad \longrightarrow \quad X^* = (\Phi^T \Phi)^{-1} \Phi^T \Psi .$$

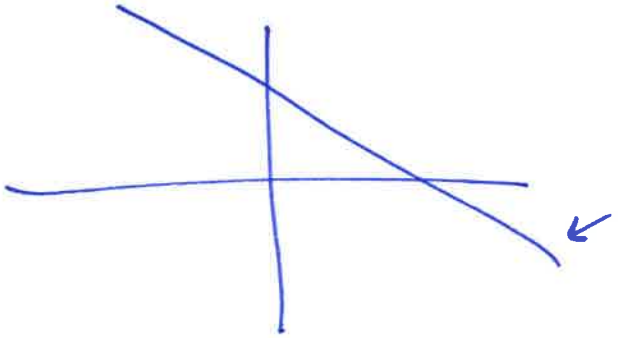
$$\delta^2 J = tr\left((\delta X)^T \Phi^T \Phi \, \delta X \right) > 0$$

conditions for strict positive-definitness of $\Phi^T \Phi$.

(note that $\Phi^T \Phi > 0$ would insure that $(\Phi^T \Phi)^{-1}$ exists).

equality constrained optimization

- motivating examples:
  - recall the least-norm problem



$$\text{minimize } \|x\|_2^2$$

$$\text{subject to } Ax = y$$

  - recall the estimation problem $\begin{cases} y_1 = x + v_1 \\ y_2 = x + v_2 \end{cases}$, $\hat{x} = k_1 y_1 + k_2 y_2$

$$\begin{array}{ll} \text{minimize} & \mathcal{E}\{(x-\hat{x})^2\} \\ \text{subject to} & \mathcal{E}\{\hat{x}\} = x \end{array} \longrightarrow \begin{array}{ll} \text{minimize} & k_1^2\sigma_1^2 + k_2^2\sigma_2^2 \\ \text{subject to} & k_1 + k_2 = 1. \end{array}$$

$$\left(\begin{array}{lll} \text{generalize to} & \text{minimize} & k_1^2\sigma_1^2 + \cdots + k_n^2\sigma_n^2 \\ n \text{ measurements} & \text{subject to} & k_1 + \cdots + k_n = 1. \end{array} \right)$$

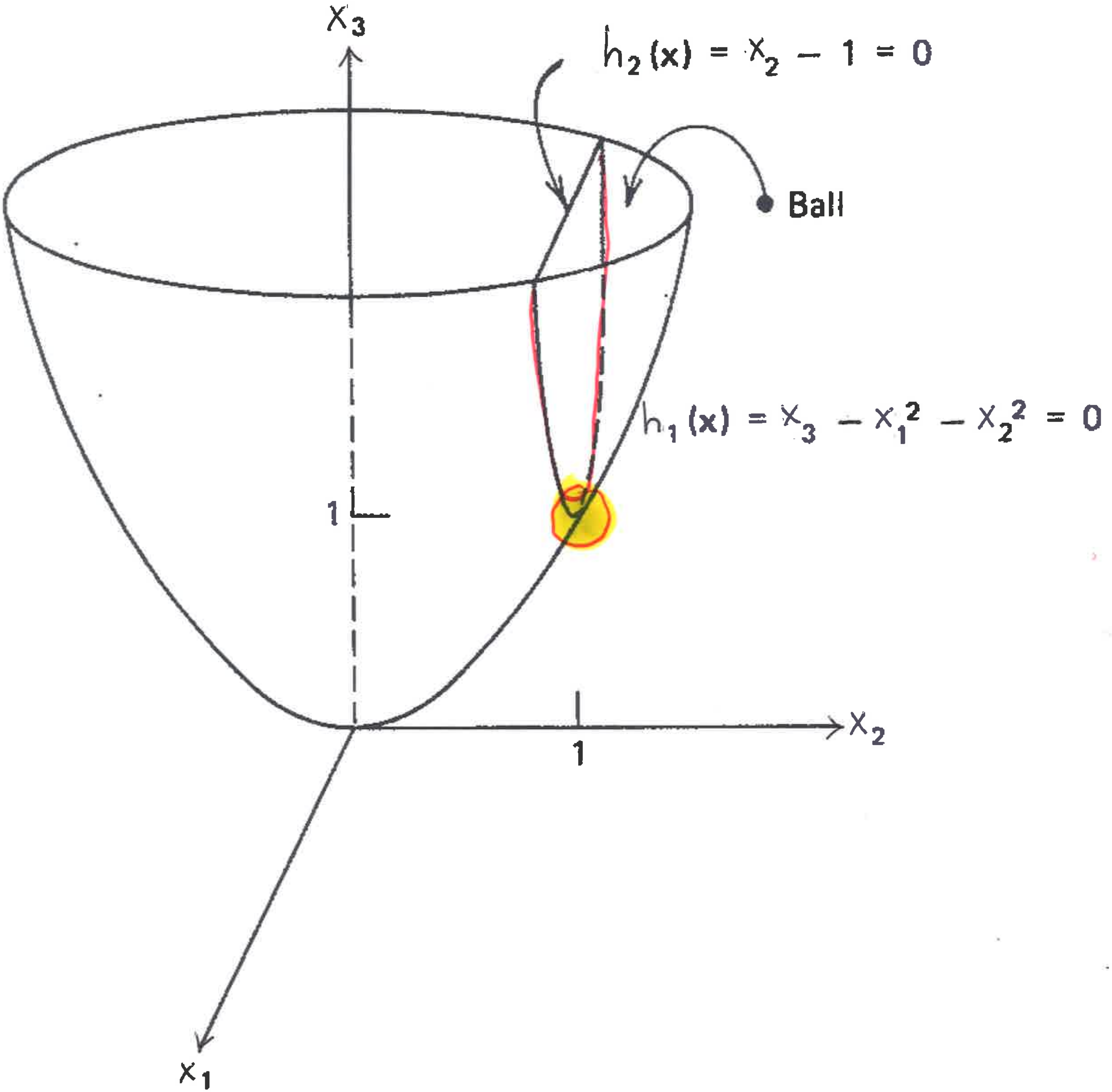
- in general an equality constrained optimization problem can be written as

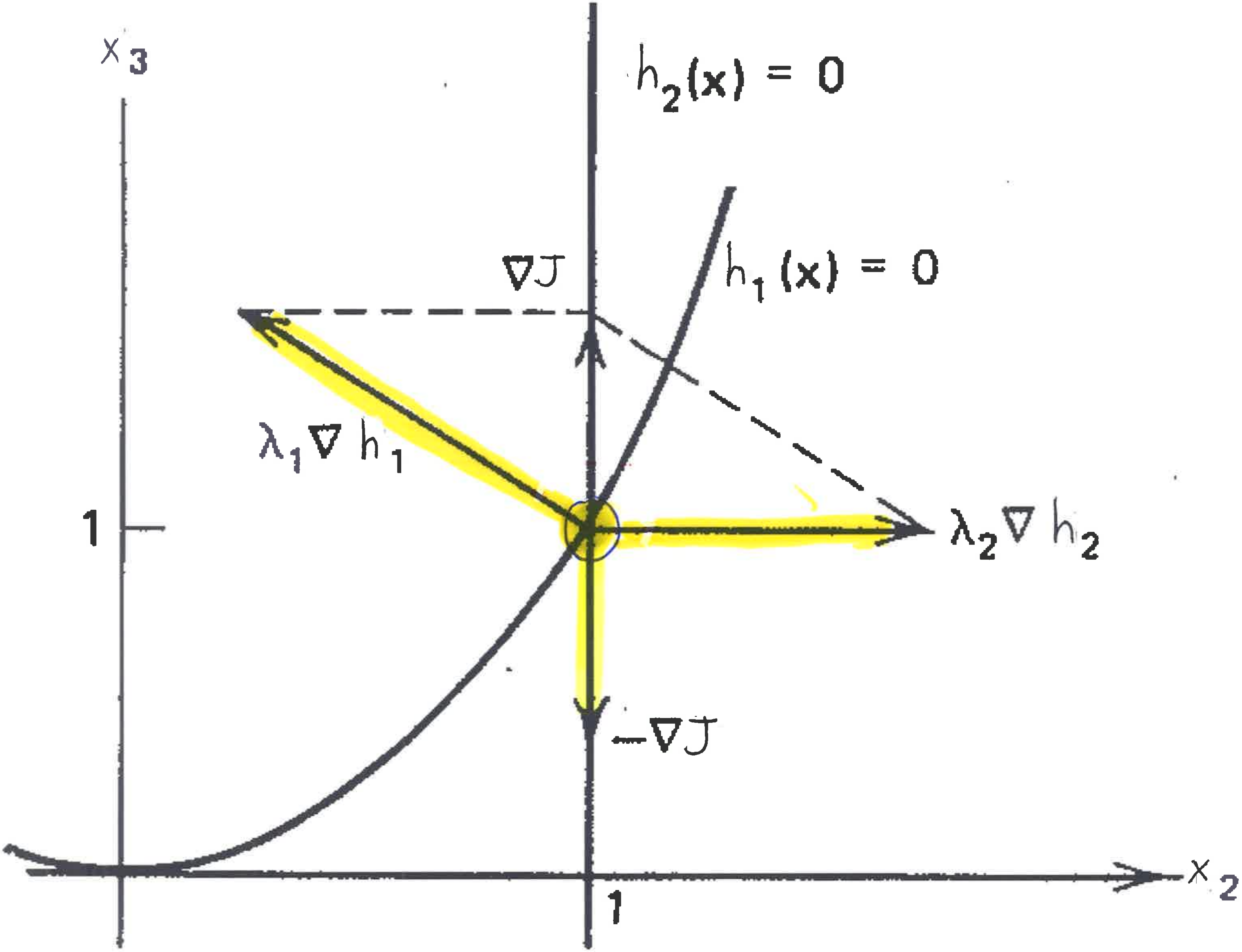
$$\begin{aligned} &\text{minimize} && J(x) \\ &\text{subject to} && h_1(x) = 0 \\ &&& \vdots \\ &&& h_m(x) = 0 \end{aligned}$$

we assume differentiability of $J(x), h_1(x), \ldots, h_m(x)$.

- example: the "ball in the bowl" problem

$$\begin{aligned} &\text{minimize} && x_3 \\ &\text{subject to} && x_3 = x_1^2 + x_2^2 \\ &&& x_2 = 1 \end{aligned} \quad \longrightarrow \quad \begin{aligned} &\text{minimize} && J(x) = x_3 \\ &\text{subject to} && -x_1^2 - x_2^2 + x_3 = 0 \\ &&& -x_2 + 1 = 0 \end{aligned}$$


X_3
$h_2(x) = X_2 - 1 = 0$
Ball
$h_1(x) = X_3 - X_1^2 - X_2^2 = 0$
1
1
X_2
X_1


x_3
$h_2(x) = 0$
∇J
$h_1(x) = 0$
$\lambda_1 \nabla h_1$
1
$\lambda_2 \nabla h_2$
$-\nabla J$
x_2
1

from a physics point of view, a force balance takes place

- $\sum$ forces $= 0$.
- the constraint surface $h_i(x) = 0$ applies a force in the direction of $\nabla h_i(x)$. (normal to the surface)
- the weight of the ball acts in the direction $-\nabla J(x)$ ($-\nabla J(x)$ is the direction of maximal decrease in J.)

the optimal x must satisfy

$$\nabla J(x) + \lambda_1 \nabla h_1(x) + \lambda_2 \nabla h_2(x) = 0$$

$$h_1(x) = 0, \quad h_2(x) = 0$$

for some λ_1 & λ_2, x.

$$J = x_3 \rightarrow \nabla J = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$h_1 = -x_1^2 - x_2^2 + x_3 \rightarrow \nabla h_1(x) = \begin{bmatrix} -2x_1 \\ -2x_2 \\ 1 \end{bmatrix}$$

$$h_2 = x_2 - 1 \rightarrow \nabla h_2(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{cases} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \lambda_1 \begin{bmatrix} -2x_1 \\ -2x_2 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \\ h_1(x) = -x_1^2 - x_2^2 + x_3 = 0 \\ h_2(x) = x_2 - 1 = 0 \end{cases}$$

5 equations
&
5 unknowns

↓

$$x^* = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \; \& \; \lambda^* = \begin{bmatrix} -1 \\ -2 \end{bmatrix}.$$

λ has dimension m

18-9

- for a general constrained optimization problem

$$\text{minimize} \quad J(x)$$
$$\text{subject to} \quad h_i(x) = 0 \qquad i = 1, \cdots, m$$

the necessary optimality conditions are

$$\begin{cases} \nabla J(x) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x) = 0 \\ h_i(x) = 0, \quad i = 1, \cdots, m. \end{cases}$$

- we can define "Lagrangian"

$$\mathcal{L}(x,\lambda) := J(x) + \sum_{i=1}^{m} \lambda_i h_i(x)$$

$$\begin{cases} \nabla_x \mathcal{L} = \nabla J(x) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x) = 0 \\ \nabla_\lambda \mathcal{L} = \begin{bmatrix} h_1(x) \\ \vdots \\ h_m(x) \end{bmatrix} = 0 \longrightarrow \begin{cases} h_1(x) = 0 \\ \vdots \\ h_m(x) = 0 \end{cases} \end{cases}$$

review of last lecture

- necessary optimality conditions for

$$\begin{array}{lll} \text{minimize} & J(x) & \\ \text{subject to} & h_i(x) = 0 & i=1,\dots,m \end{array}$$

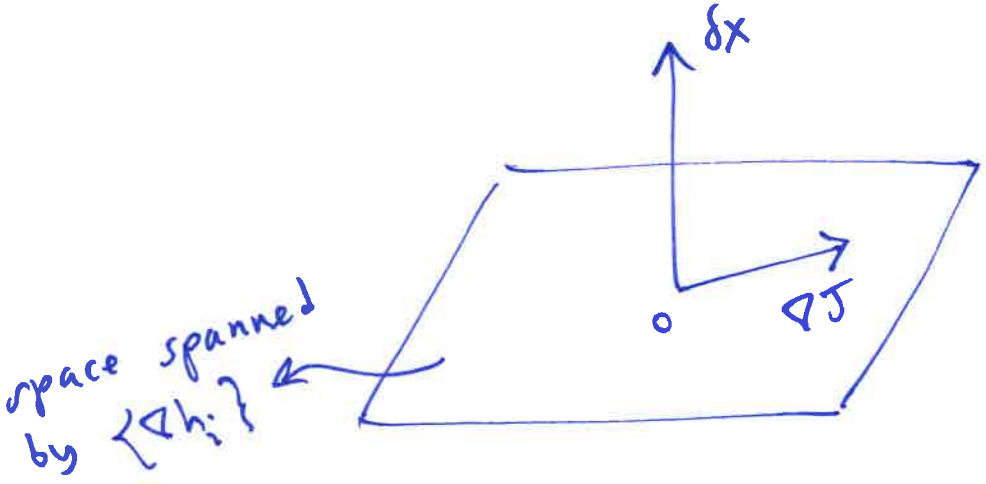
are given by

$$\begin{cases} \nabla J(x) + \sum_i \lambda_i \nabla h_i(x) = 0 \\ h_i(x) = 0 \qquad i=1,\dots,m \end{cases}$$

Lagrange multipliers $\lambda \in \mathbb{R}^m$

- define Lagrangian $\mathcal{L}(x,\lambda) = J(x) + \sum_{i=1}^{m} \boxed{\lambda_i}\, h_i(x)$

$$\begin{cases} \nabla_x \mathcal{L} = \nabla J + \sum \lambda_i \nabla h_i = 0 \\ \nabla_\lambda \mathcal{L} = \begin{bmatrix} h_1(x) \\ \vdots \\ h_m(x) \end{bmatrix} = 0 \end{cases}$$



- derivation of necessary optimality conditions for the problem

$$\begin{aligned} &\text{minimize} \quad J(x) \\ &\text{subject to} \quad h_i(x) = 0, \qquad i=1,\dots,m \end{aligned}$$

(h_i : scalar-valued)

objective: $J(x+\delta x) = J(x) + (\nabla J(x))^T \delta x + \text{h.o.t.}$

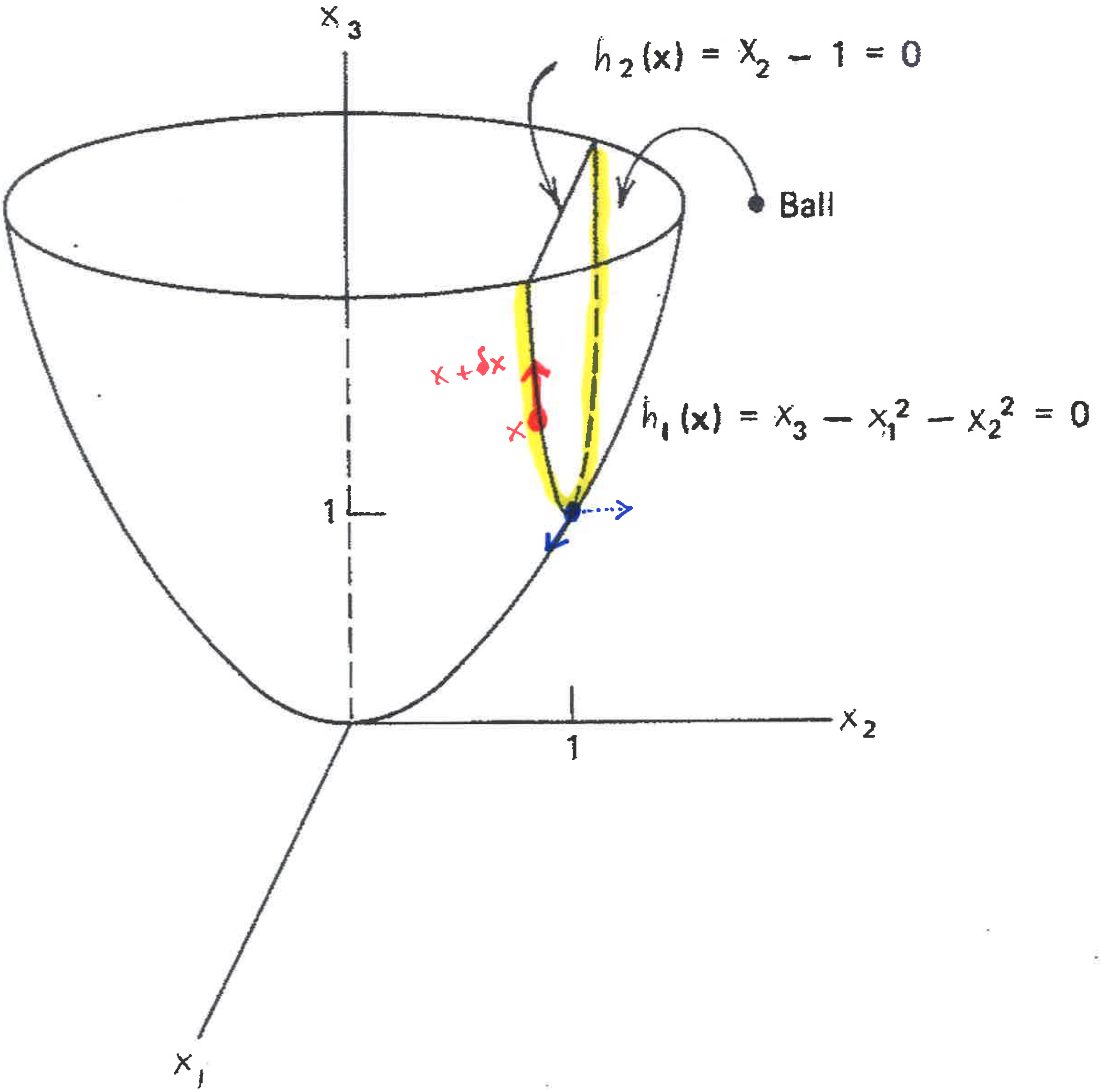
constraints: $\underbrace{h_i(x+\delta x)}_{=0} = \underset{0}{\cancel{h_i(x)}} + \underbrace{(\nabla h_i(x))^T \delta x}_{=0} + \text{h.o.t.} \qquad i=1,\dots,m$

necessary conditions for optimality of x

① δx must be such that $(\nabla h_i(x))^T \delta x = 0$.

(i.e., the perturbations δx must respect the constraints)

② $(\nabla J(x))^T \delta x = 0$.

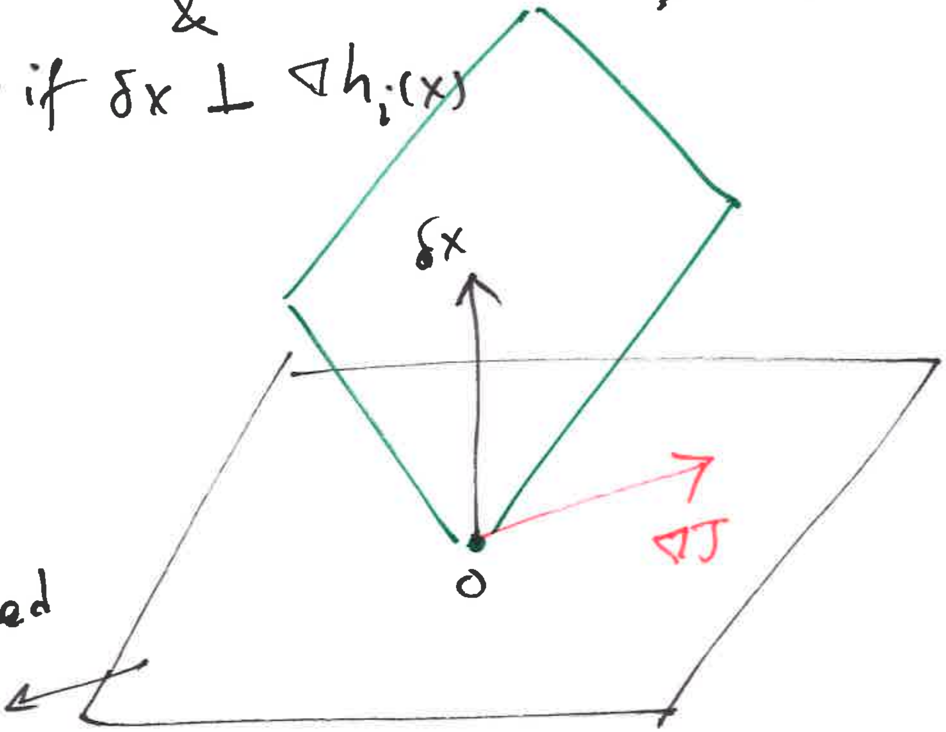

x_3
$h_2(x) = x_2 - 1 = 0$
Ball
$x + \delta x$
x
$h_1(x) = x_3 - x_1^2 - x_2^2 = 0$
1
1
x_2
x_1

① & ② → if x is the minimal point, then

$(\nabla J(x))^T \delta x = 0$ for all δx that satisfy

$(\nabla h_i(x))^T \delta x = 0$, i.e.,

$$\left\{\begin{array}{l}\text{if } x \text{ is optimal}\\ \quad \& \\ \text{if } \delta x \perp \nabla h_i(x)\end{array}\right. \longrightarrow \delta x \perp \nabla J(x).$$



space spanned by $\nabla h_i(x)$, $i=1,\dots,m$

(we decompose the x-space to the subspace spanned by $\{\nabla h_i(x)\}_{i=1,\dots,m}$ & its complement)

$\nabla J(x)$ belongs to the space spanned by the vectors $\{\nabla h_i(x)\}_{i=1}^{m}$.

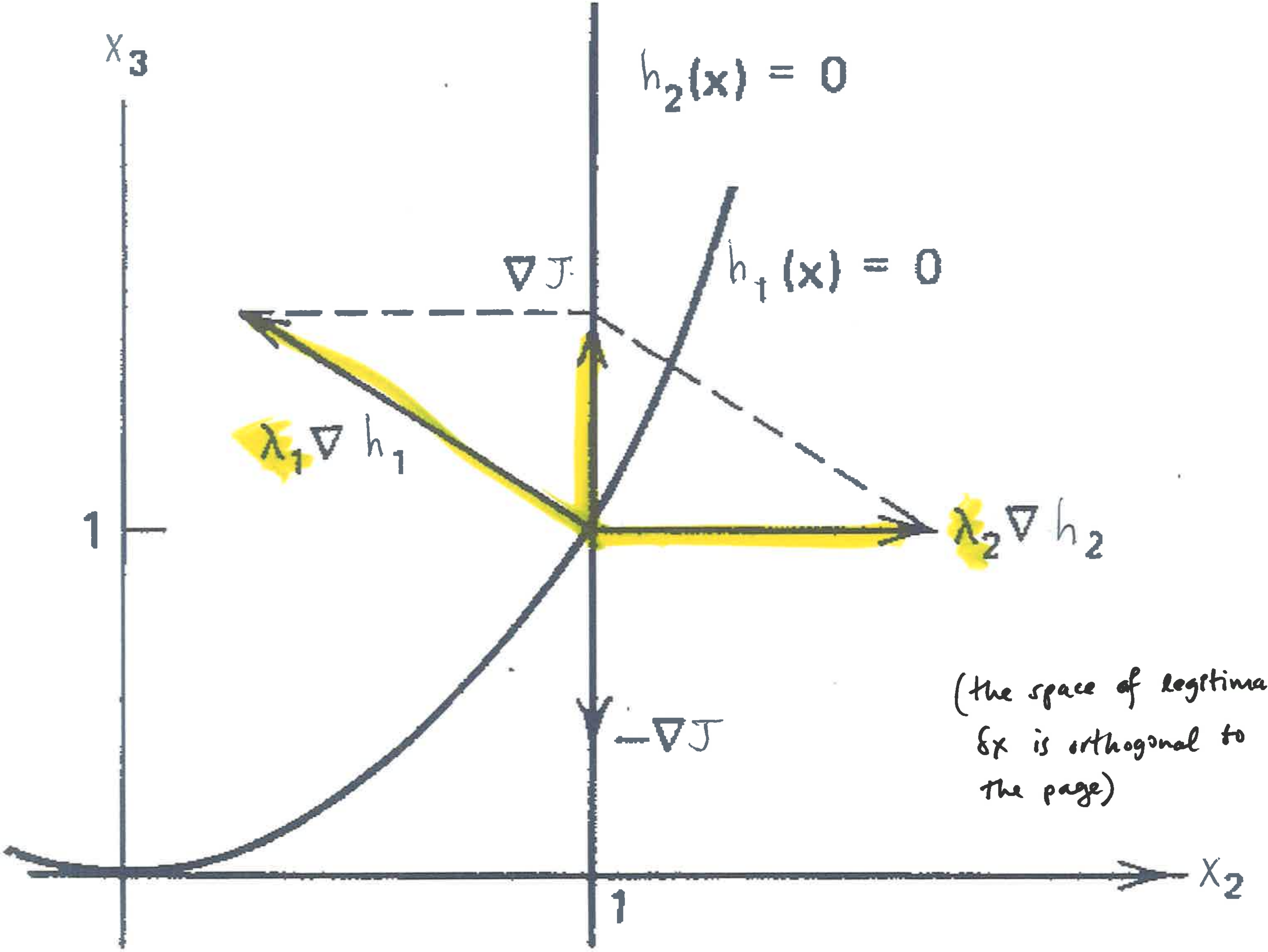
$$\downarrow$$

$$\nabla J(x) = \sum_{i=1}^{m} (-\lambda_i) \nabla h_i(x)$$

$$\downarrow$$

$$\nabla J(x) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x) = 0$$

$\{\lambda_i\}_{i=1}^{m}$ are called the "Lagrange multipliers."



(the space of legitimate δx is orthogonal to the page)

- example: minimize $k_1^2\sigma_1^2 + k_2^2\sigma_2^2$
  subject to $k_1 + k_2 = 1$

$k = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$

$h(k) = k_1 + k_2 - 1 = 0$

$$\mathcal{L} = J + \lambda h = k_1^2\sigma_1^2 + k_2^2\sigma_2^2 + \lambda(k_1 + k_2 - 1)$$

$$\nabla_k \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial k_1} \\ \frac{\partial \mathcal{L}}{\partial k_2} \end{bmatrix} = \begin{bmatrix} 2k_1\sigma_1^2 + \lambda \\ 2k_2\sigma_2^2 + \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \longrightarrow \begin{cases} k_1 = -\frac{\lambda}{2\sigma_1^2} \\ k_2 = -\frac{\lambda}{2\sigma_2^2} \end{cases}$$

$$\nabla_\lambda \mathcal{L} = k_1 + k_2 - 1 = 0$$

3 equ. & 3 unknowns

$$k_1 + k_2 - 1 = -\frac{\lambda}{2\sigma_1^2} - \frac{\lambda}{2\sigma_2^2} - 1 = 0$$

$$-\frac{\lambda}{2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) = 1 \rightarrow \lambda = -2\,\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

$$k_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \quad k_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

- example: minimize $\|x\|_2^2$ subject to $Ax = b$ $\quad A: m\times n \quad n \geq m$, rows of A are lin. ind.

$$\begin{bmatrix} - a_1^T - \\ \vdots \\ - a_m^T - \end{bmatrix} x = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \longrightarrow \left\{ \begin{array}{l} a_1^T x = b_1 \\ \vdots \\ a_m^T x = b_m \end{array} \right. \longrightarrow \left\{ \begin{array}{l} h_1(x) := a_1^T x - b_1 = 0 \\ \vdots \\ h_m(x) := a_m^T x - b_m = 0 \end{array} \right.$$

$$\mathcal{L}(x,\lambda) = x^T x + \lambda_1 (a_1^T x - b_1) + \cdots + \lambda_m (a_m^T x - b_m)$$

$$= x^T x + \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{bmatrix}^T \begin{bmatrix} a_1^T x - b_1 \\ \vdots \\ a_m^T x - b_m \end{bmatrix}$$

$$= x^T x + \lambda^T (Ax - b)$$

to find $\nabla_x \mathcal{L}$ we set $x \to x + \delta x$ and find the first-order variation $\delta \mathcal{L}$.

$$\mathcal{L}(x+\delta x, \lambda) = (x+\delta x)^T (x+\delta x) + \lambda^T (Ax + A\delta x - b)$$

$$\delta \mathcal{L} = (\delta x)^T x + x^T \delta x + \lambda^T A \delta x = 2x^T \delta x + \lambda^T A \delta x$$

$$= (2x^T + \lambda^T A)\, \delta x = (2x + A^T \lambda)^T \delta x$$

<u>aside</u>:

$$\delta^2 \mathcal{L} = \delta x^T \delta x$$

$$\nabla_x^2 \mathcal{L} = 2I \succ 0$$

$$\nabla_x \mathcal{L} = 2x + A^T \lambda = 0 \longrightarrow x = -\tfrac{1}{2} A^T \lambda. \quad ①$$

$$\nabla_\lambda \mathcal{L} = Ax - b = 0 \longrightarrow Ax = b \quad ②$$

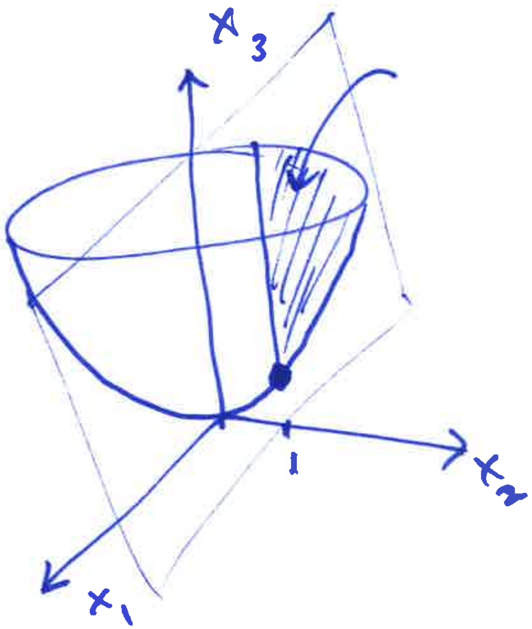
$$① \ \& \ ② \rightarrow Ax = -\tfrac{1}{2} AA^T \lambda = b \rightarrow \lambda = -2(AA^T)^{-1} b \quad ③$$

$$① \ \& \ ③ \rightarrow x = -\tfrac{1}{2} A^T \left(-2(AA^T)^{-1} b\right)$$

$$= A^T (AA^T)^{-1} b.$$

19-7

optimization with inequality constraints



- recall "ball & bowl" problem

$$\begin{aligned} &\text{minimize} && x_3 \\ &\text{subject to} && x_2 \geq 1 \\ & && x_3 \geq x_1^2 + x_2^2 \end{aligned}$$

$$\begin{aligned} &\text{minimize} && J(x) \\ &\text{subject to} && f_i(x) \leq 0 \qquad i = 1, \ldots, m \end{aligned}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$J(x) = x_3$$

$$f_1(x) = x_1^2 + x^2 - x_3 \leq 0$$

$$f_2(x) = 1 - x_2 \leq 0$$

- KKT (Karush-Kuhn-Tucker) optimality conditions

$$x \in \mathbb{R}^n$$

$$\nabla J(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) = 0$$

$$f_i(x) \leq 0 \qquad i=1,\dots,m$$

$$\lambda_i \geqslant 0 \qquad i=1,\dots,m$$

$$\lambda_i f_i(x) = 0 \qquad i=1,\dots,m$$

- main idea of proof:

$$f_i(x) \leq 0 \quad \leftrightarrow \quad f_i(x) + s_i^2 = 0 \qquad s_i \in \mathbb{R}.$$

(slack variables)

$$\hat{x} := \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ \hline s_1 \\ \vdots \\ s_m \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} x_1 \\ \vdots \\ x_n \end{matrix}}\right\} x \\ \left.\vphantom{\begin{matrix} s_1 \\ \vdots \\ s_m \end{matrix}}\right\} s \end{matrix}$$

$$\hat{J}(\hat{x}) := J(x)$$

$$\hat{f}_i(\hat{x}) := f_i(x) + s_i^2 .$$

minimize $\hat{J}(\hat{x})$

subject to $\hat{f}_i(\hat{x}) = 0 \qquad i=1,\dots,m$

$$\hat{\mathcal{L}}(\hat{x},\lambda) = \hat{J}(\hat{x}) + \sum_{i=1}^{m} \lambda_i \hat{f}_i(\hat{x})$$

$$= J(x) + \sum_{i} \lambda_i \left(f_i(x) + s_i^2\right)$$

$$\downarrow$$

$$\nabla_x \mathcal{L} = \nabla_x J(x) + \sum_{i=1}^{m} \lambda_i \nabla_x f_i(x) = 0$$

$$\nabla_\lambda \mathcal{L} = \begin{bmatrix} f_1(x) + s_1^2 \\ \vdots \\ f_m(x) + s_m^2 \end{bmatrix} = 0$$

$$\nabla_s \mathcal{L} = \begin{bmatrix} 2\lambda_1 s_1 \\ \vdots \\ 2\lambda_m s_m \end{bmatrix} = 0$$

<u>aside</u>:

$$\nabla_x^2 \mathcal{L} = \nabla_x^2 J + \sum \lambda_i \nabla_x^2 f_i$$

$$\nabla_\lambda^2 \mathcal{L} = 0$$

$$\nabla_s^2 \mathcal{L} = 2\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix}$$

from $\lambda_i s_i = 0$ we conclude that either $\lambda_i = 0$, or, $s_i = 0$.

if $s_i = 0 \rightarrow f_i(x) = 0 \longrightarrow$ we have $\lambda_i f_i(x) = 0$.

if $s_i \neq 0 \rightarrow \lambda_i = 0 \longrightarrow$ again we get $\lambda_i f_i(x) = 0$.

---

the optimality conditions become

$$\nabla J + \sum_i \lambda_i \nabla f_i = 0$$

$$f_i(x) = \cancel{-s_i^2} \leq 0 \qquad i=1, \cdots, m$$

$$\lambda_i f_i(x) = 0 \qquad i=1, \cdots, m$$

$$\lambda_i \geq 0 \qquad i=1, \cdots, m$$

review of last lecture

- optimization with inequality constraints

$$\text{minimize} \quad J(x)$$
$$\text{subject to} \quad f_i(x) \leq 0 . \qquad i=1,\dots,m$$

KKT conditions for optimality (nec. cond.)

$$\nabla J(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) = 0$$

$$f_i(x) \leq 0 \qquad i=1,\dots,m$$

$$\lambda_i \geq 0 \qquad i=1,\dots,m$$

$$\lambda_i f_i(x) = 0 \qquad i=1,\dots,m$$

KKT cond. are nec. & suff. if optimization problem is "convex"!
→ see book by Boyd.

- basic idea: $f_i(x) \leq 0 \iff f_i(x) + \underset{\text{slack variable}}{s_i^2} = 0$

20-1

- ball & bowl example revisited.

$$\begin{aligned} \text{min.} \quad & x_3 \\ \text{s.t.} \quad & x_1^2 + x_2^2 - x_3 \leq 0 \\ & x_2 \geq 1 \end{aligned}$$

$$\mathcal{L} = x_3 + \lambda_1 (x_1^2 + x_2^2 - x_3) + \lambda_2 (1 - x_2)$$

KKT conditions

$$\nabla_x \mathcal{L} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \lambda_1 \begin{bmatrix} 2x_1 \\ 2x_2 \\ -1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} = 0 \longrightarrow \begin{cases} 2\lambda_1 x_1 = 0 & \longrightarrow x_1 = 0 \\ 2\lambda_1 x_2 - \lambda_2 = 0 & \longrightarrow \lambda_2 = 2x_2 \\ 1 - \lambda_1 = 0 & \longrightarrow \lambda_1 = 1 \end{cases}$$

start here.

$$f_1(x) = x_1^2 + x_2^2 - x_3 \leq 0, \quad f_2(x) = 1 - x_2 \leq 0 \longrightarrow x_2 \geq 1 \xrightarrow[\lambda_2 = 2x_2]{\text{since}} \lambda_2 \geq 2 \geq 0$$

$$\lambda_1 \geq 0, \quad \lambda_2 \geq 0$$

$$\begin{cases} \cancel{\lambda_1 = 0} \\ \text{or} \\ f_1(x) = 0 \end{cases}, \quad \begin{cases} \cancel{\lambda_2 = 0} \\ \text{or} \\ f_2(x) = 0 \end{cases}$$

$$f_1(x) = 0 \longrightarrow x_3 = 1$$

$$f_2(x) = 0 \longrightarrow x_2 = 1 \xrightarrow[\lambda_2 = 2x_2]{\text{since}} \lambda_2 = 2$$

$$x^* = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \begin{matrix} \checkmark \\ \checkmark \\ \checkmark \end{matrix}, \quad \lambda^* = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{matrix} \checkmark \\ \checkmark \end{matrix}$$

(solution is the same as that found in case of equality constraints)

- example: solve the constrained optimization problem

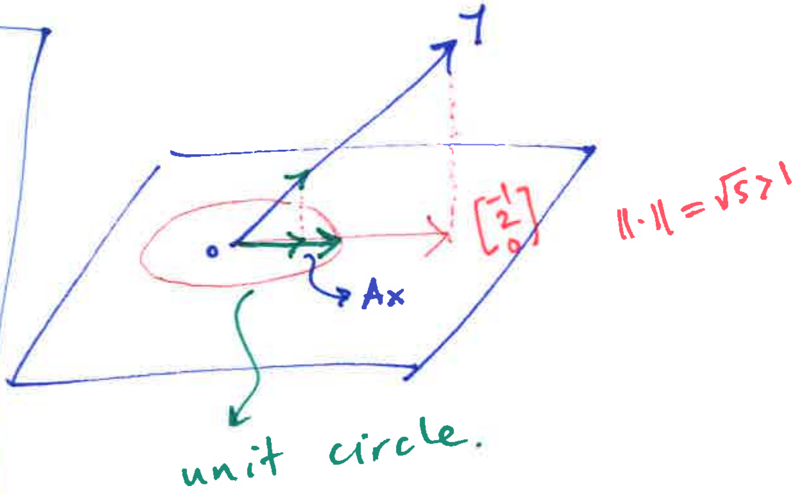
$$\begin{aligned} &\text{minimize} && \|Ax-y\|_2^2 \\ &\text{subject to} && \|x\|_2^2 \leq 1. \end{aligned} \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix}.$$

<u>aside</u>: if $y = 0.1 \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix}$, then x^* would be given by standard least squares, & the constraint would be satisfied with strict inequality.

<u>aside</u>:

$$Ax = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix}$$

$$\|x\| \leq 1$$
$$\Updownarrow$$
$$\|Ax\| \leq 1$$



$$\begin{aligned} &\text{minimize} && J(x) := \|Ax-y\|_2^2 \\ &\text{subject to} && f(x) := \|x\|_2^2 - 1 \leq 0 \end{aligned}$$

$$\mathcal{L} = \|Ax-y\|_2^2 + \lambda(\|x\|_2^2 - 1)$$

KKT conditions:

$$\begin{cases} \nabla\mathcal{L} = 2A^T(Ax-y) + 2\lambda x = 0 \\ \|x\|_2^2 \leqslant 1 \\ \lambda \geqslant 0 \\ \lambda f(x) = 0 \rightarrow \begin{cases} \lambda = 0 \\ \text{or} \\ f(x) = 0. \end{cases} \end{cases}$$

$$\nabla\mathcal{L} = 0 \rightarrow (A^TA + \lambda I)x = A^Ty. \rightarrow x = (A^TA + \lambda I)^{-1}A^Ty$$

$$A^TA = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow (A^TA + \lambda I)^{-1} = \frac{1}{1+\lambda}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\rightarrow (A^TA + \lambda I)^{-1}A^Ty = \frac{1}{1+\lambda}\begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

$$\|x\|_2^2 \leqslant 1 \rightarrow \|(A^TA + \lambda I)^{-1}A^Ty\|_2^2 \leqslant 1$$

20-4

$$\|x\|_2^2 \leqslant 1 \rightarrow \frac{1}{(1+\lambda)^2}(1+4) \leqslant 1 \rightarrow (1+\lambda)^2 \geqslant 5.$$

$$\downarrow$$

$\lambda = 0$ is impossible

$$\lambda \neq 0 \rightarrow f(x) = 0 \rightarrow \|x\|_2^2 = 1 \rightarrow \frac{1}{(1+\lambda)^2} 5 = 1 \rightarrow (1+\lambda)^2 = 5$$

$$\rightarrow 1+\lambda = \pm\sqrt{5} \rightarrow \begin{cases} \lambda = \sqrt{5}-1 > 0 \\ \cancel{\lambda = -\sqrt{5}-1 < 0} \end{cases}$$

$(\lambda \geqslant 0)$

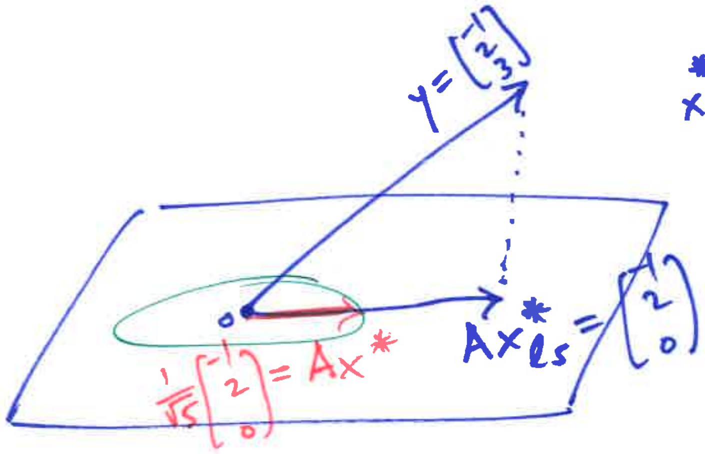
$$\overset{*}{x} = (A^TA + \lambda I)^{-1} A^T y$$

$$= \frac{1}{1+\lambda} \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

$$= \frac{1}{\boxed{\sqrt{5}}} \begin{bmatrix} -1 \\ 2 \end{bmatrix}.$$

$\downarrow$ normalizing factor

(its positivity ensures that x^* is aligned with $x^*_{\text{least-squares}}$)



$$x_{ls}^* = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

aside: what would have happened if $y = 0.1\begin{bmatrix}-1\\2\\3\end{bmatrix}$:

$$\|x\|_2^2 \leq 1 \rightarrow \frac{1}{(1+\lambda)^2}\,0.01\,(1+4) \leq 1 \rightarrow (1+\lambda)^2 \geq 0.05$$

($\lambda = 0$ is not impossible)

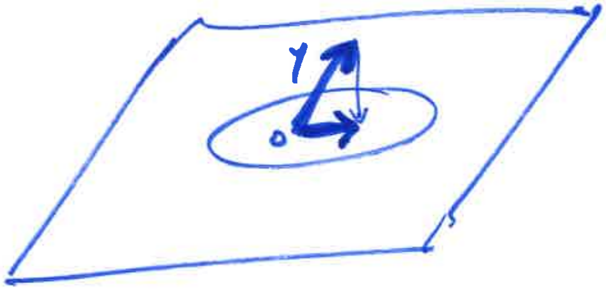
$$\lambda \neq 0 \rightarrow f(x) = 0 \rightarrow \|x\|_2^2 = 1 \rightarrow \frac{0.05}{(1+\lambda)^2} = 1 \rightarrow (1+\lambda)^2 = 0.05$$

$$\rightarrow 1+\lambda = \pm\, 0.1\sqrt{5} \quad \times$$

($\because \lambda$ always negative)

$$\lambda = 0 \rightarrow x^* = (A^TA + \underbrace{\lambda I}_{0})^{-1}A^Ty = (A^TA)^{-1}A^Ty.$$

(just as in case of unconstrained least_squares)



min. $\|x\|_2^2$

s.t. $Ax = y$.

---

min. $\|Ax-y\|_2$

---

lin. prog. ←

min. $\|x\|_1 = |x_1| + \cdots + |x_n|$ (compress. sensing)

s.t. $Ax = y$.

---

min. $\|Ax-y\|_1 = |e_1| + \cdots + |e_n|$

---

min. $\|x\|_\infty = \max_i \{|x_i|\}$

s.t. $Ax = y$

---

min. $\|Ax-y\|_\infty = \max_i \{|e_i|\}$

Linear programming

- motivating example : the "dairy" problem

① 22 gallons of milk per week

② 2 gallons of milk → 1 kilogram of butter

③ 3 gallons of milk → 1 gallon of ice cream

④ refrigerator space (to store butter) is unlimited

⑤ freezer space (to store ice cream) is 6 gallons

⑥ 1 hour of work → 4 gallons of ice cream

⑦ 1 hour of work → 1 kilogram of butter

⑧ total hours of work per week is 6

⑨ profit for ice cream is \$ 5 per gallon

⑩ profit for butter is \$4 per kilogram

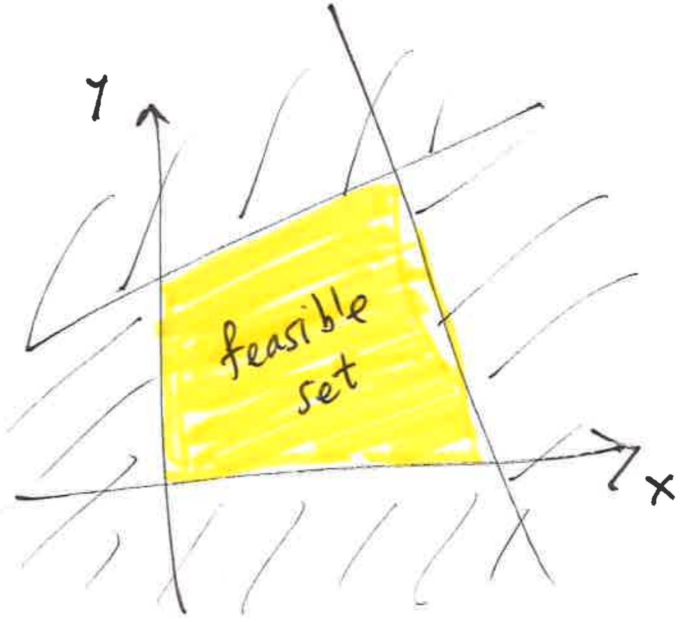
how much butter & ice cream should we make to maximize profit?

---

x: gallons of ice cream

y: kilograms of butter

$J(x,y)$: total profit.



$$\underset{x,y}{\text{maximize}} \quad J(x,y) := 5x + 4y$$ ← ⑨ & ⑩ (profit)

subject to

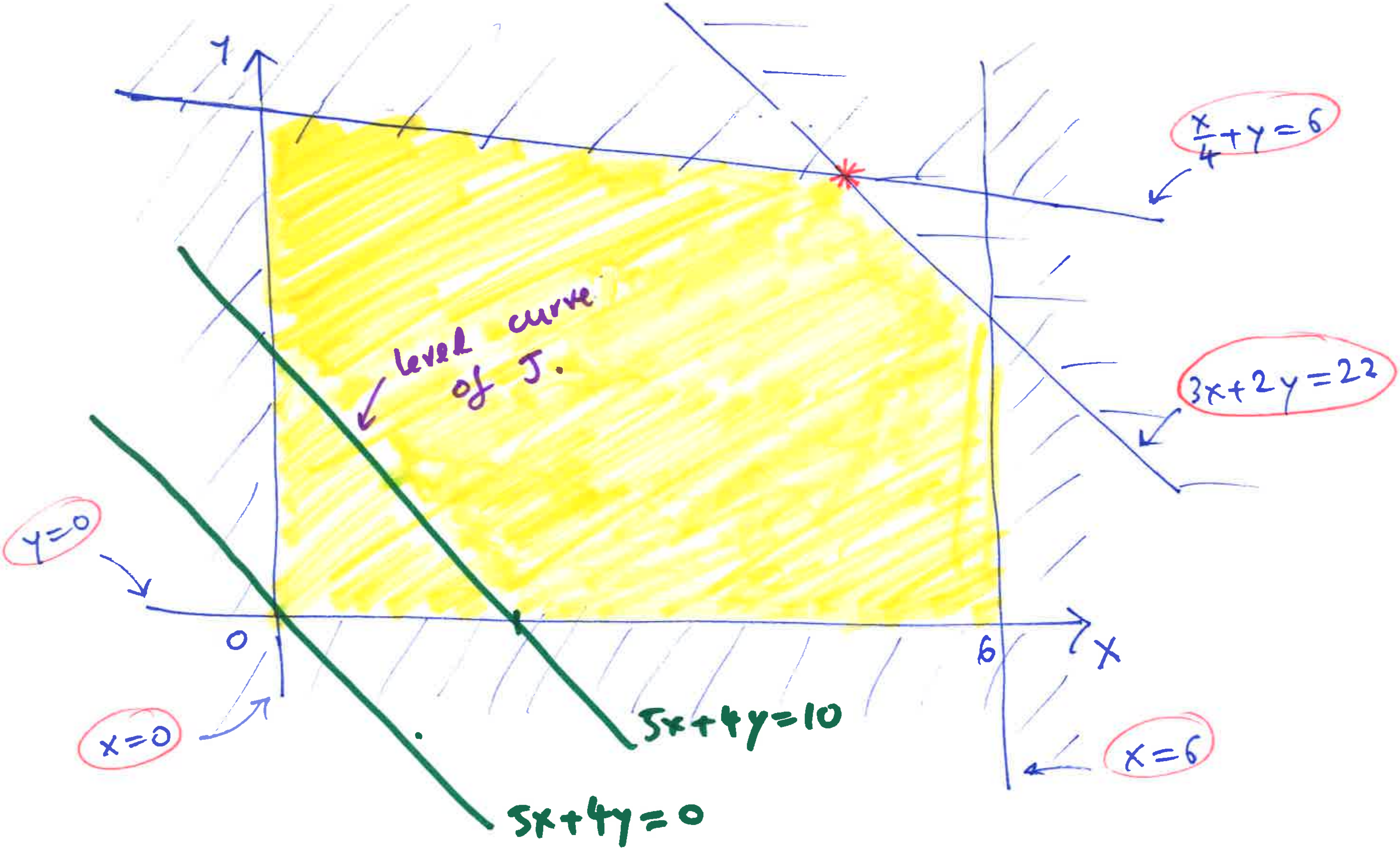
$$3x + 2y \leq 22$$ ← ①,②,③ (resource constraints)

$$x \leq 6$$ ← ~~④~~, ⑤ (storage constraints)

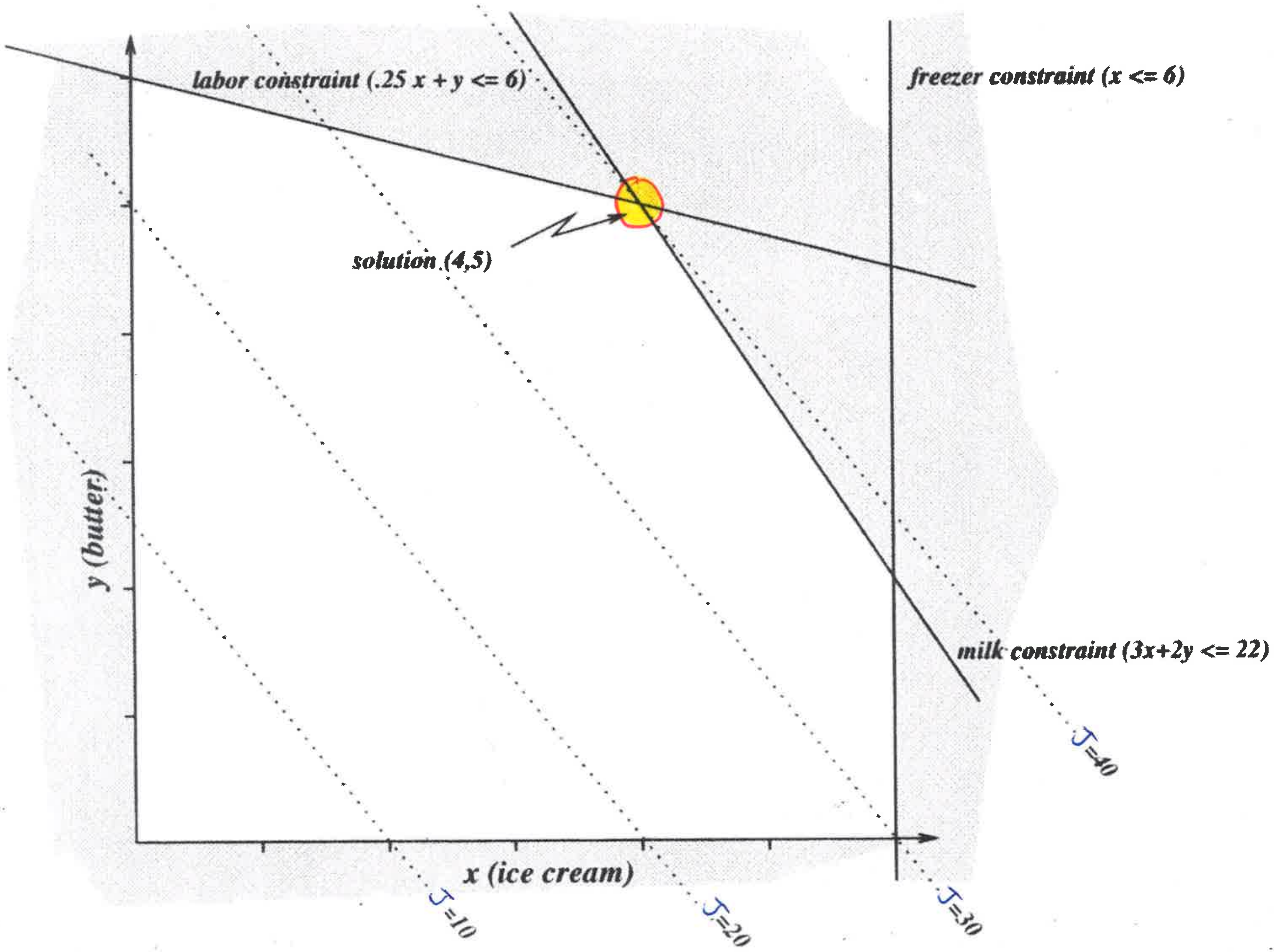
$$\frac{x}{4} + y \leq 6$$ ← ⑥,⑦,⑧ (time constraints)

$$x \geq 0, \quad y \geq 0$$ (feasibility)

21-2



the optimization problem seeks the point in the feasible (yellow) set that maximizes J.


labor constraint (.25 x + y <= 6)
freezer constraint (x <= 6)
solution (4,5)
y (butter)
milk constraint (3x+2y <= 22)
x (ice cream)
J=10
J=20
J=30
J=40

- maximum profit corresponds to the highest level curve that is still (barely) in the feasible set

$J^* = 40$, $x^* = 4$, $y^* = 5$. (doesn't have to be integer-valued)

- the solution (almost) always occurs at a corner (vertex)

- the graphical description allows easy answers to questions such as
  ~ example: how increase/decrease in ice cream & butter prices affect the optimal solution? (corresponds to changes in the slope of level curves)
  ~ example: would be a good idea to buy extra milk at say \$1 per gallon?

- graphical method is very difficult if we have more than 2 variables (say we make yogurt & cheese too!)

- the "dairy" problem can be rewritten as

$$\text{minimize} \quad J(x) = c^T z \qquad\qquad z = \begin{bmatrix} x \\ y \end{bmatrix}.$$

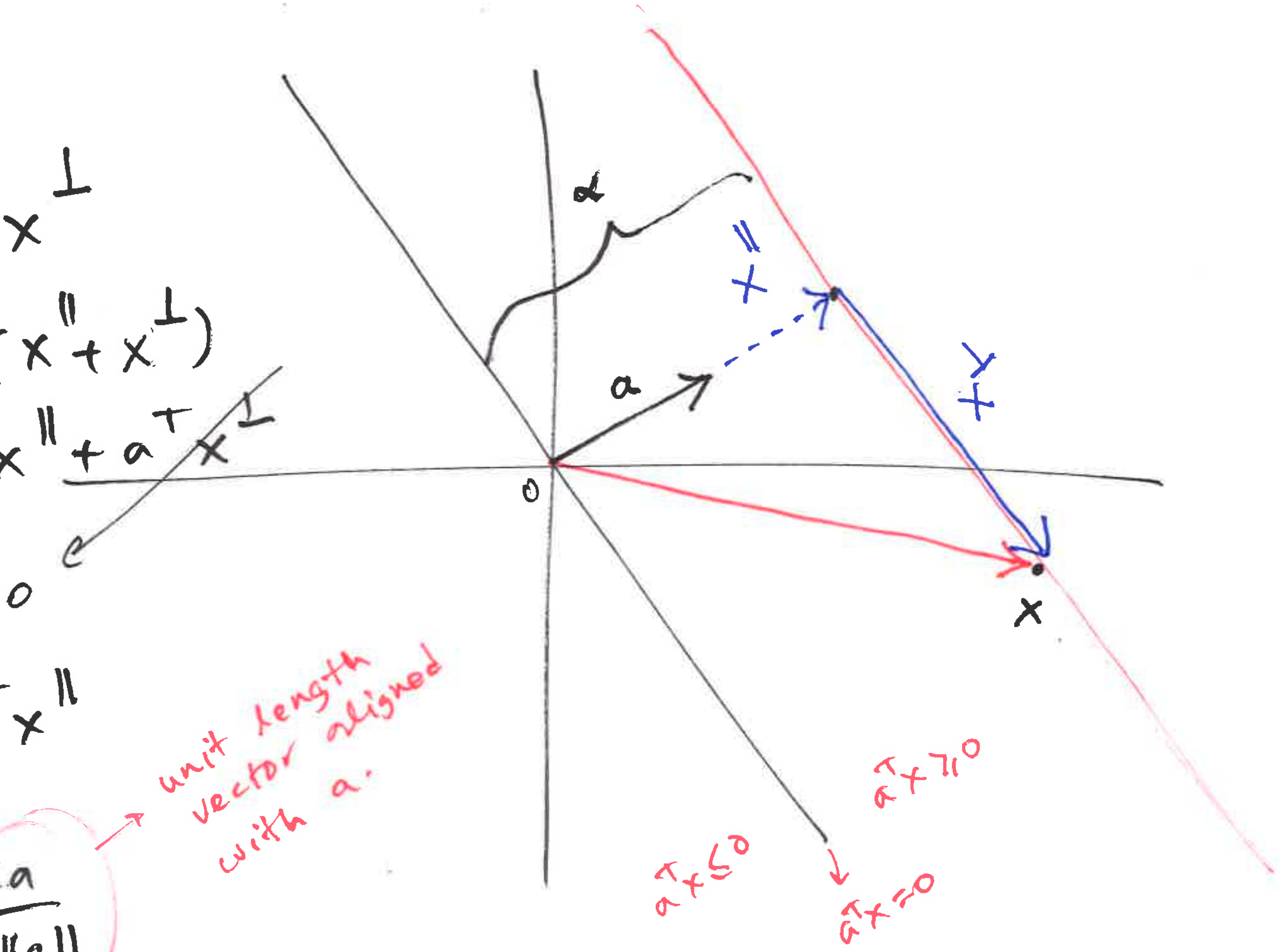
$$\text{subject to} \quad Az \leq b$$

with $c = \begin{bmatrix} -5 \\ -4 \end{bmatrix}.$

$$A = \begin{bmatrix} 3 & 2 \\ 1 & 0 \\ 1/4 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} \qquad\qquad b = \begin{bmatrix} 22 \\ 6 \\ 6 \\ 0 \\ 0 \end{bmatrix}$$

this is an example of a "linear program".

- what does the set $\{x \mid a^T x = 0\}$ look like?
- what does the set $\{x \mid a^T x = b\}$ look like? $(b \in \mathbb{R})$



$$x = x^{\parallel} + x^{\perp}$$

$$a^T x = a^T(x^{\parallel} + x^{\perp})$$
$$= a^T x^{\parallel} + a^T x^{\perp}$$

$$a^T x = ?$$

$$a^T x = a^T x^{\parallel}$$

$$x^{\parallel} = \alpha \frac{a}{\|a\|_2}$$

→ unit length vector aligned with a.

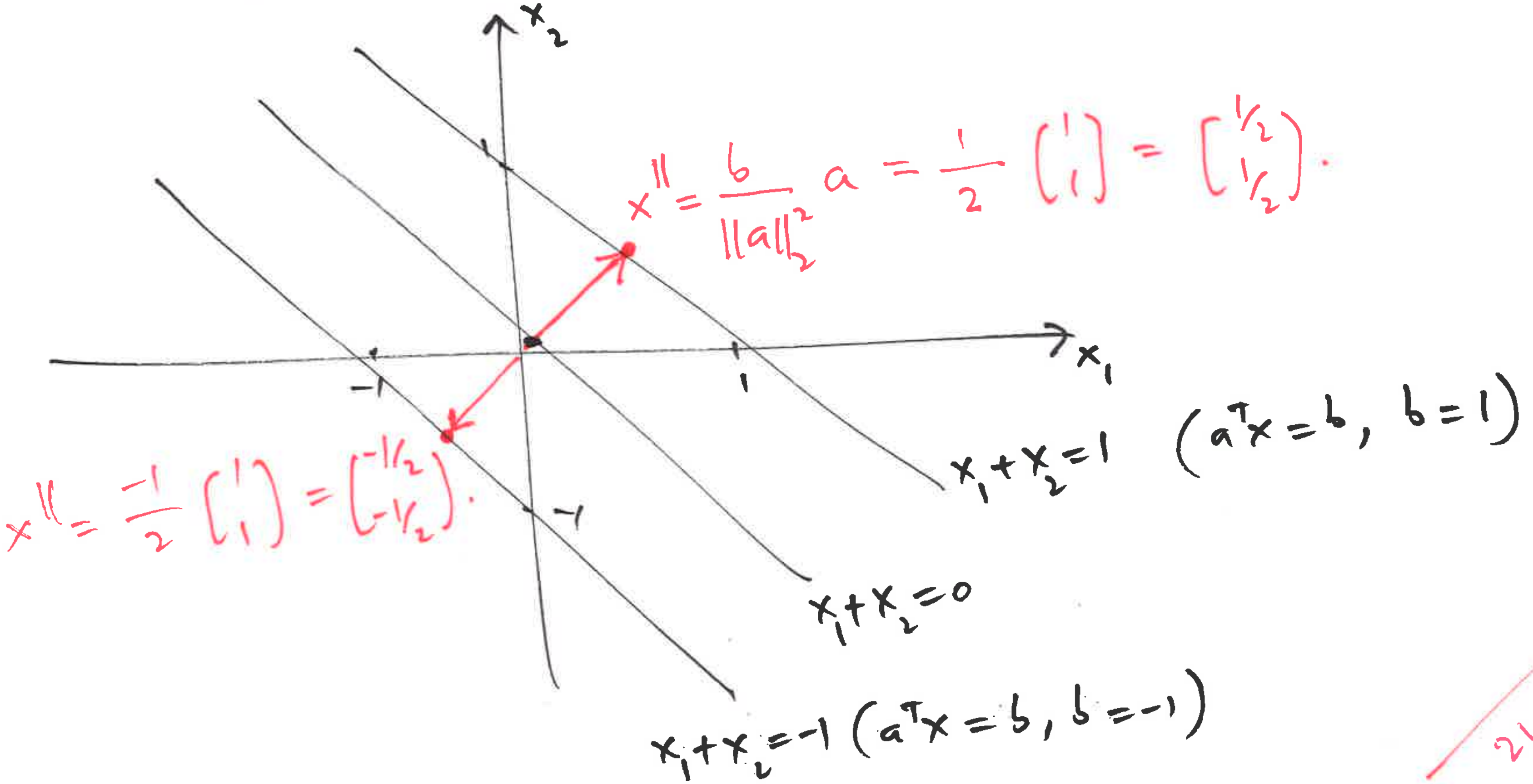
$$a^T x = b \rightarrow a^T x^{\parallel} = b \rightarrow a^T\left(\alpha \frac{a}{\|a\|_2}\right) = b$$

↓

$$\alpha \frac{a^Ta}{\|a\|_2} = b \rightarrow \alpha \frac{\|a\|_2^2}{\|a\|_2} = b \rightarrow \boxed{\alpha = \frac{b}{\|a\|_2}}$$

$$x^{\|} = \alpha \frac{a}{\|a\|_2} = \frac{b}{\|a\|_2^2}\, a \rightarrow \boxed{x^{\|} = \frac{b}{\|a\|_2^2}\, a}$$

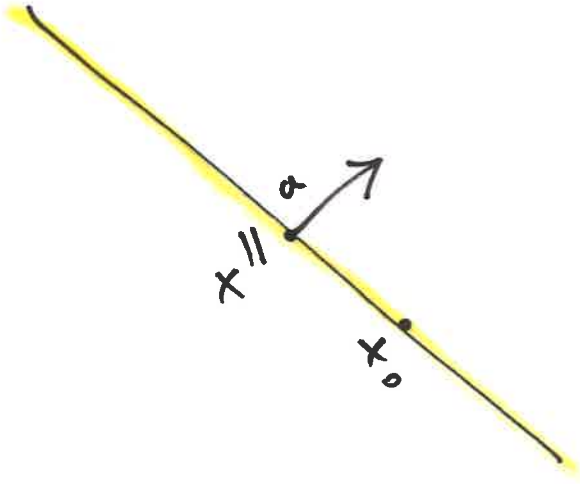
- example: $a=\begin{bmatrix}1\\1\end{bmatrix} \rightarrow a^Tx = x_1 + x_2$



$$x^{\|} = \frac{b}{\|a\|_2^2}\, a = \frac{1}{2}\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}1/2\\1/2\end{pmatrix}.$$

$$x^{\|} = \frac{-1}{2}\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}-1/2\\-1/2\end{pmatrix}.$$

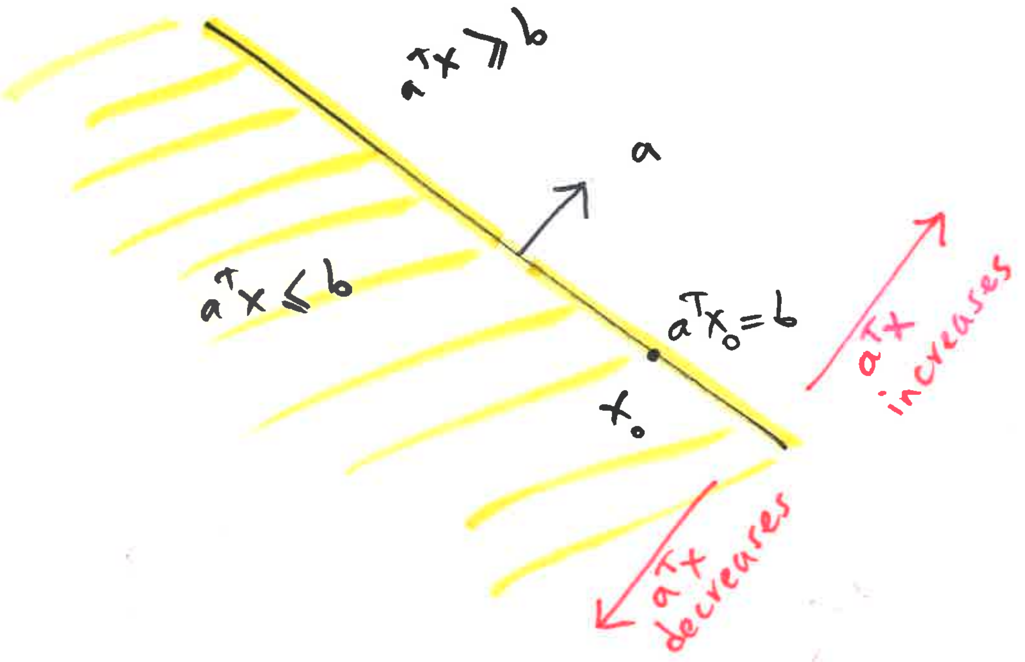
- hyperplane : set of points described by $\{x \mid a^T x = b\}$, $a \neq 0$.



$a^T x_0 = b$, $a^T x'' = b$.

aside:
in $\mathbb{R}^2$ → line
in $\mathbb{R}^3$ → plane
in $\mathbb{R}^4$ → 3D "plane"!

- half space : set points described by $\{x \mid a^T x \leq b\}$.



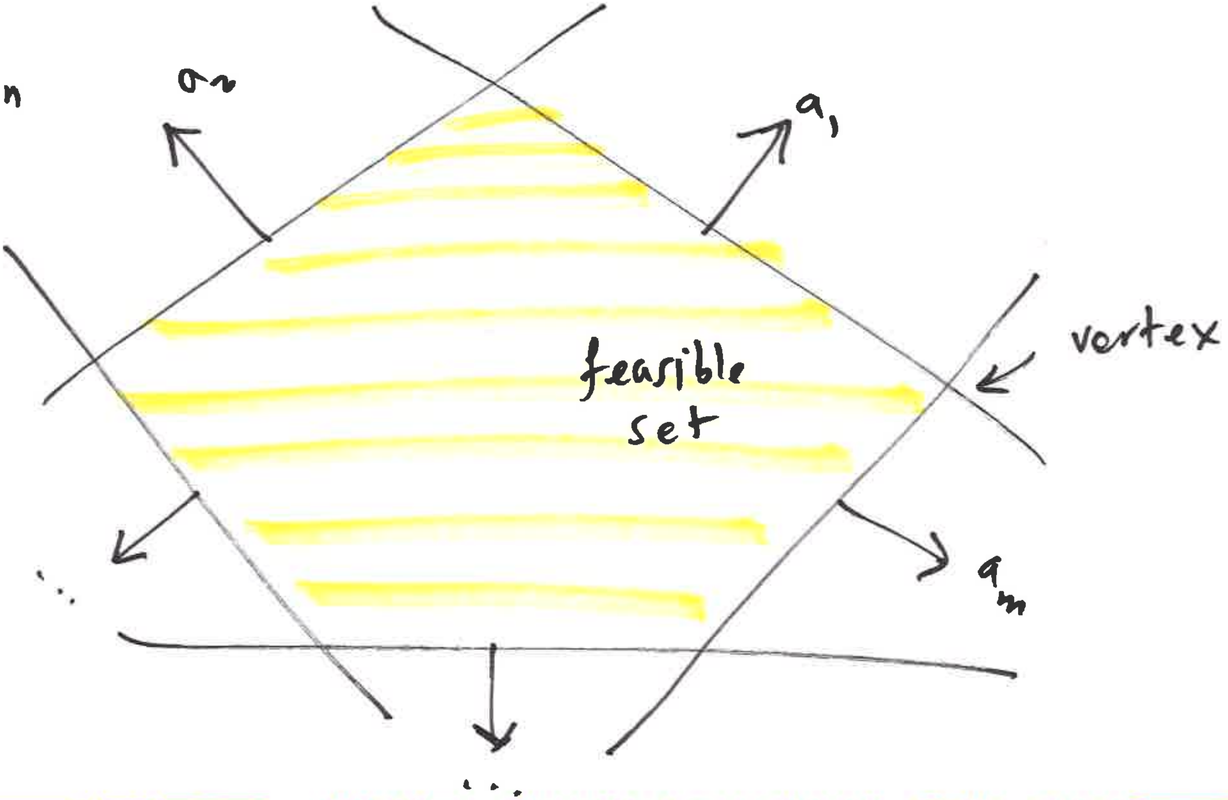
aside.
$J(x) = a^T x$
$\nabla J = a$
direction of maximal increase in objective

- polyhedra (polytope) : solution set of finitely many linear inequalities & linear equlities.

$$Ax \leqslant b \qquad Gx = h$$

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} \rightarrow Ax = \begin{bmatrix} a_1^T x \\ \vdots \\ a_m^T x \end{bmatrix} \leqslant \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \longrightarrow \left\{ \begin{array}{l} a_1^T x \leqslant b_1 \\ \vdots \\ a_m^T x \leqslant b_m \end{array} \right.$$

a polyhedron is the intersection of a finite number of halfspaces & hyperplanes.



21-9

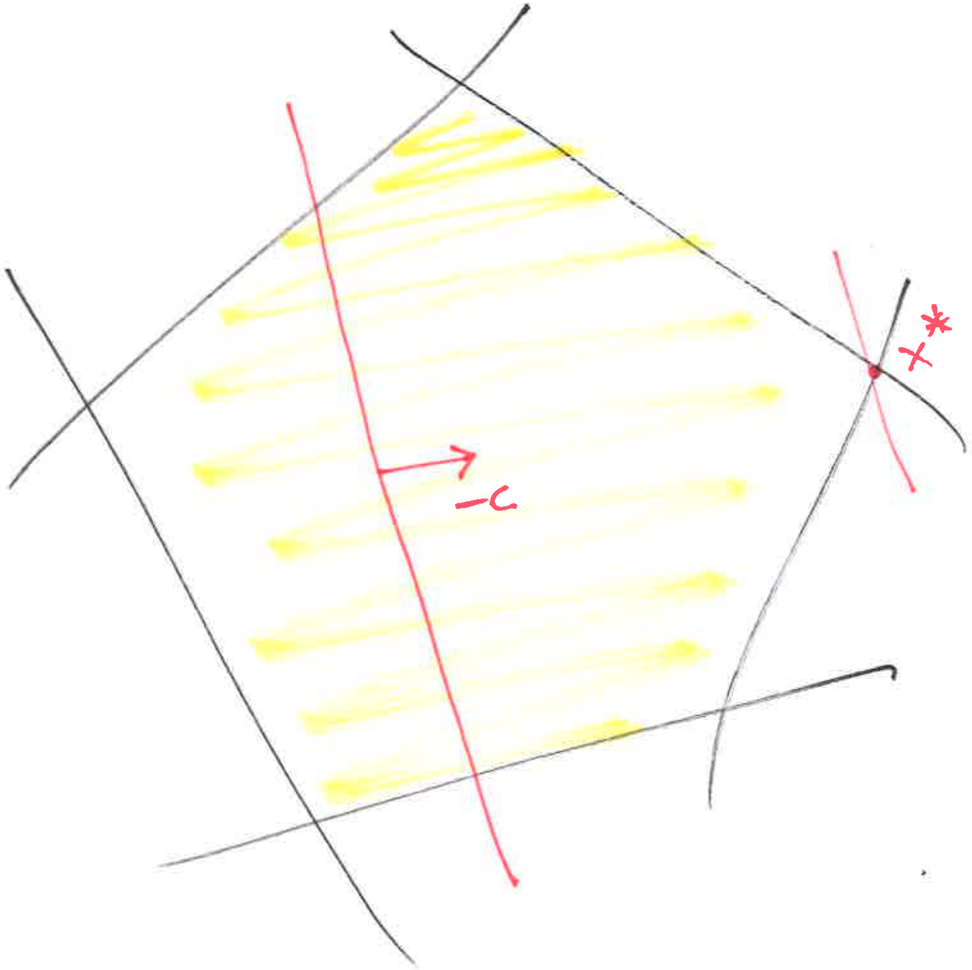
- example: add the constraint $x_3 = 5$ to the "dairy" problem

$$A = \begin{bmatrix} 3 & 2 & 0 \\ 1 & 0 & 0 \\ 1/4 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 22 \\ 6 \\ 6 \\ 0 \\ 0 \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$x_3 = 5 \quad \rightarrow \quad G = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}, \quad h = 5$$

- linear program (LP)

$$\begin{aligned} \text{minimize} \quad & c^T x + d \\ \text{subject to} \quad & Ax \leq b \\ & Gx = h \end{aligned}$$



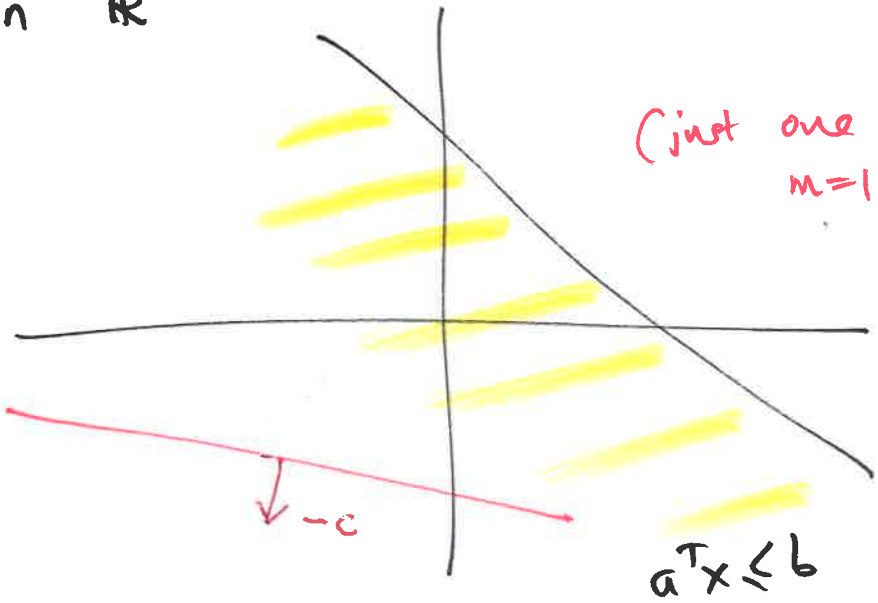
aside:

$f(x) = c^T x$

$\downarrow$

$\nabla f(x) = c.$

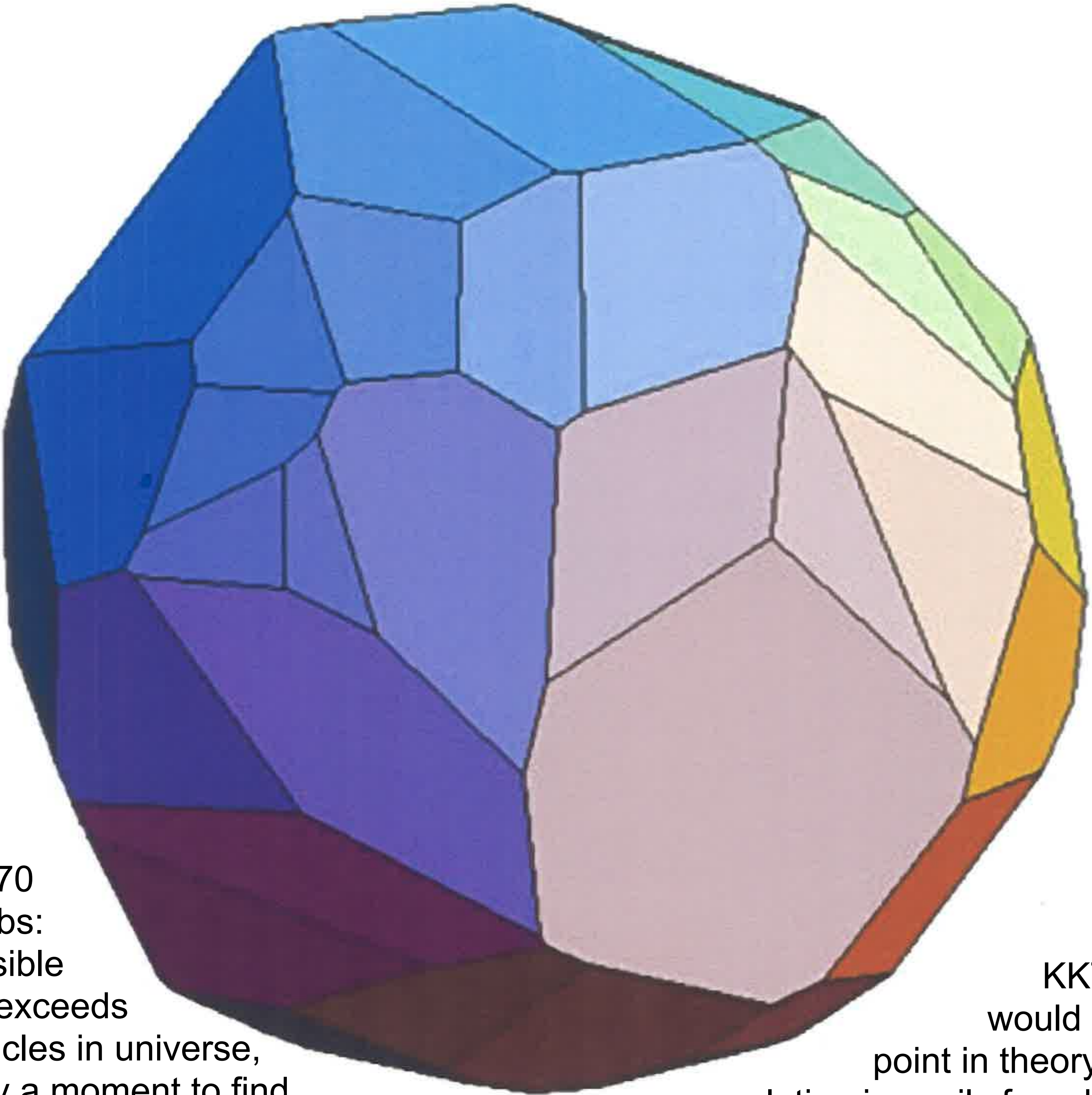
- if we have too few constraints, then $J = -\infty$.
- example: in $\mathbb{R}^2$



(just one constraint)
$m=1$

- in $\mathbb{R}^n$ we need the intersection of n hyperplanes to create a vertex.
- how many vertices for a problem in $\mathbb{R}^n$ & with m inequality constraints? (some may be infeasible)

$$\left\{ \begin{array}{l} n=10 \\ m=20 \end{array} \right. \rightarrow \binom{m}{n} \approx 200{,}000. \qquad \binom{20}{10} = \frac{20!}{10!\,10!}$$



finding best
assignment of 70
people to 70 jobs:
number of possible
configurations exceeds
number of particles in universe,
but it takes only a moment to find
optimum by posing problem as LP

KKT conditions
would give optimal
point in theory; in practice
solution is easily found numerically

- applications of LP include:
  - the diet problem:

x_j: amount of food j

F_{ij}: amount of nutrient i in food j

r_i: minimum daily requirement of nutrient i

c_j: cost of serving unit amount of food j

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Fx \geq r \\ & x \geq 0 \end{array} \xrightarrow[\text{put in standard LP form}]{\text{rewrite to}} \begin{array}{l} -Fx \leq -r \\ -x \leq 0 \end{array}$$

$$\downarrow$$

$$\underbrace{\begin{bmatrix} -F \\ -I \end{bmatrix}}_{A} x \leq \underbrace{\begin{bmatrix} -r \\ 0 \end{bmatrix}}_{b}$$

Kantorovich.

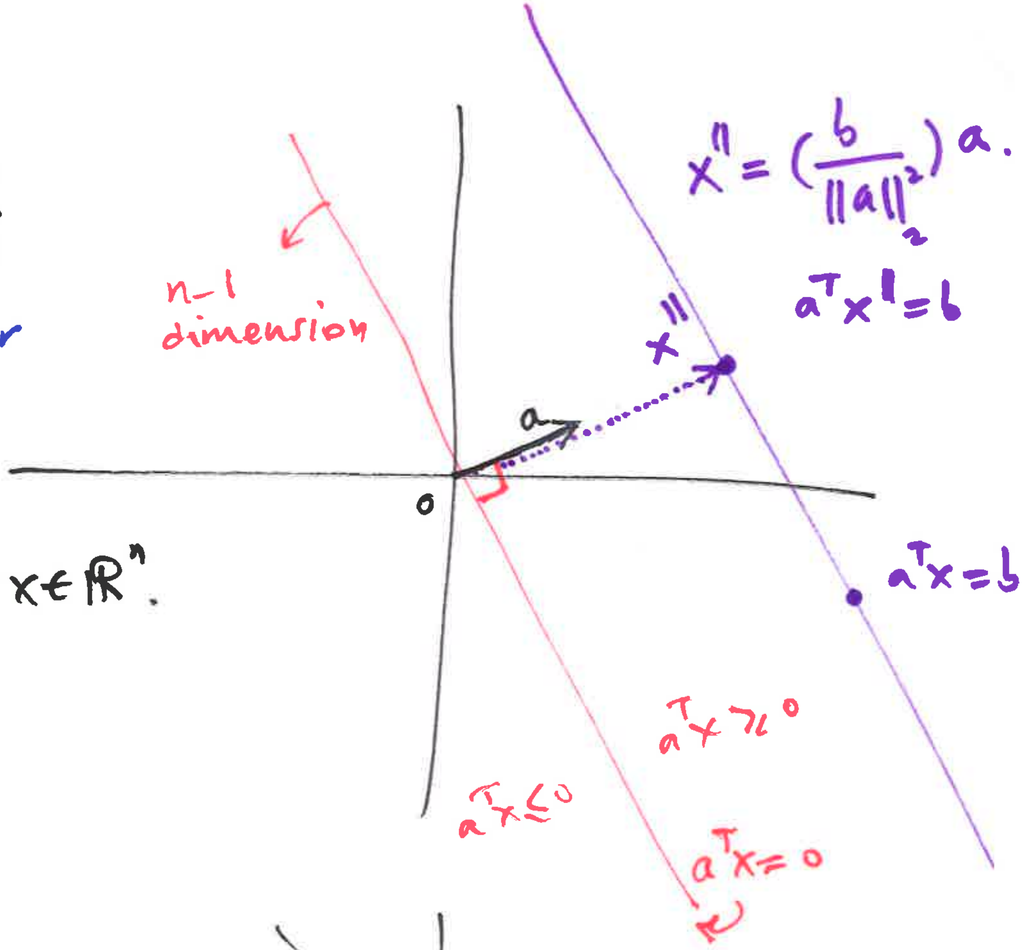
- load balancing (in parallel computing).
- resource allocation (similar to the "dairy" problem).
- network flow (movement of commodities from where they are produced to where they are consumed; UPS delivery, ...)
- "least-norm" problems

min. $\|Ax-b\|_1$	min. $\|x\|_1$ s.t. $Ax=b$.
min. $\|Ax-b\|_\infty$	min. $\|x\|_\infty$ s.t. $Ax=b$

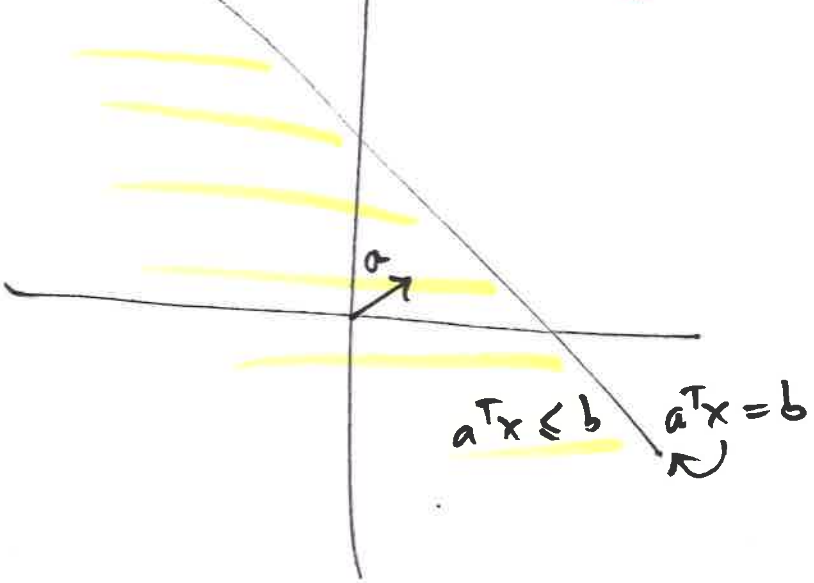
(least-norm problems often don't look that way at first!)

review of last lecture

- hyperplane: $\{x \mid a^T x = \underset{\substack{\uparrow \\ \text{scalar}}}{b}\}$

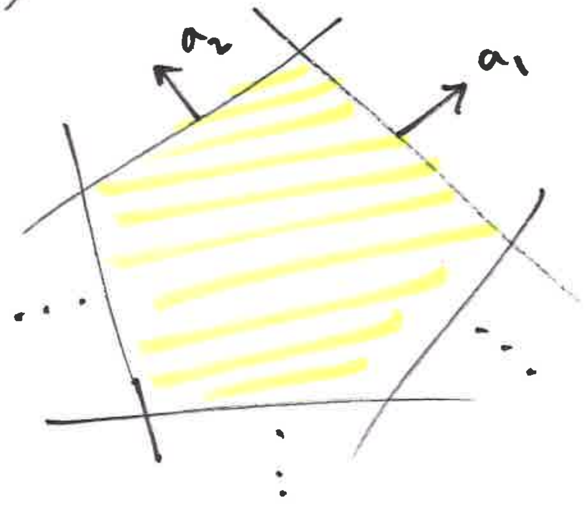


- halfspace: $\{x \mid a^T x \leq b\}$



- polyhedron: $Ax \preceq b$ ↙ vector $\qquad Gx = h$

$(a_i^T x \leq b_i) \qquad (g_i^T x = h_i)$



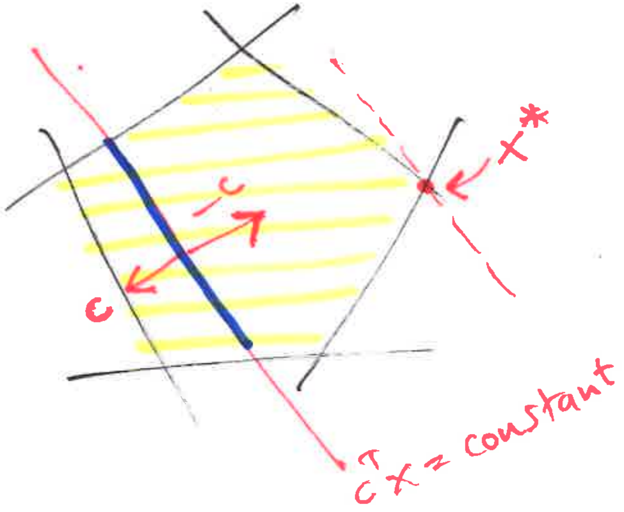
- linear program (LP):

$$\begin{array}{ll} \text{minimize} & c^T x + d \\ \text{subject to} & Ax \preceq b \\ & Gx = h \end{array}$$

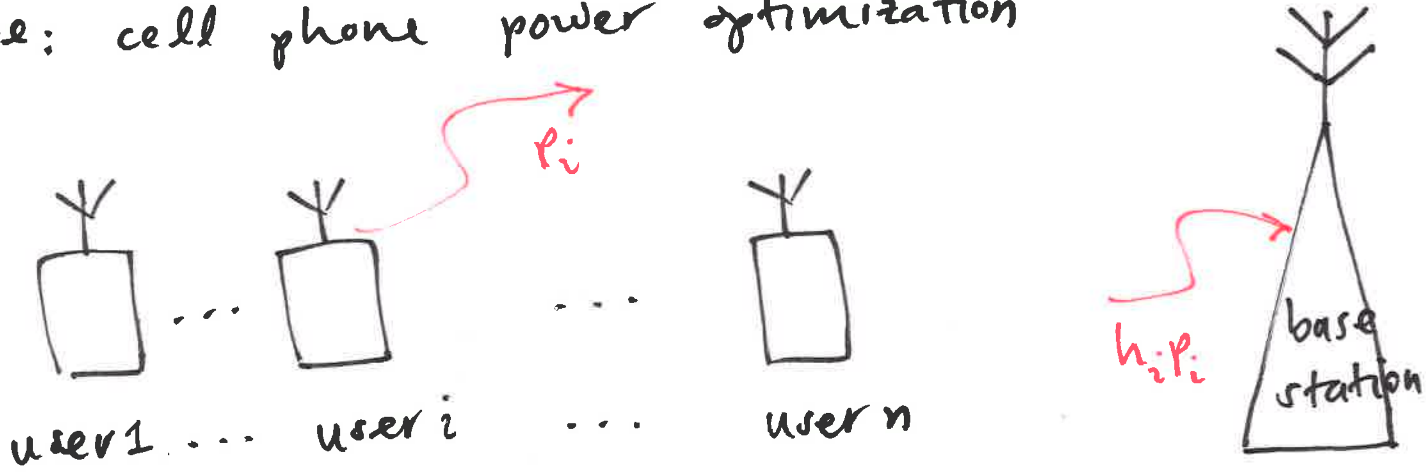
aside: $A_1 x \preceq b_1$

$A_2 x \preceq b_2$

$$\overset{A}{\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}} x \preceq \overset{b}{\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}}$$



- example: cell phone power optimization



- user i transmits with power p_i.

- attenuation factor h_i (for p_i)
  (actual power received at base station for user i is $h_i p_i$)

- for reliable communication, the signal-to-interference ratio for user i must exceed γ_i.

- interference for user i from all other users is $\sum_{j \neq i} h_j p_j$

it is desired to find p_i, $i=1,\dots,n$, that achieve reliable communication, but minimize total power usage.

$$\begin{array}{ll} \text{minimize} & p_1 + \cdots + p_n \\ \text{subject to} & p_i \geqslant 0 \quad i=1,\dots,n \\ & \dfrac{h_i p_i}{\sum_{j\neq i} h_j p_j} \geqslant \gamma_i \quad i=1,\dots,n \end{array}$$

$$p := \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} \qquad c := \mathbb{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \longrightarrow c^T p = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} = p_1 + \cdots + p_n .$$

$$p_i \geqslant 0 \quad \forall i \rightarrow \underbrace{\begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & & 1 \end{bmatrix}}_{I} \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} \geqslant 0 \longrightarrow \underbrace{-I}_{A_1} p \leqslant \underbrace{0}_{b_1} .$$

$$h_i p_i \geqslant \gamma_i \sum_{j\neq i} h_j p_j \rightarrow -h_i p_i + \underbrace{\gamma_i \sum_{j\neq i} h_j p_j}_{\gamma_i (h_1 p_1 + \cdots + h_{i-1} p_{i-1} + h_{i+1} p_{i+1} + \cdots + h_n p_n)} \leqslant 0$$

$$\begin{bmatrix} +\gamma_i h_1 & \cdots & +\gamma_i h_{i-1} & -h_i & +\gamma_i h_{i+1} & \cdots & +\gamma_i h_n \end{bmatrix} \begin{bmatrix} p_1 \\ \vdots \\ p_i \\ \vdots \\ p_n \end{bmatrix} \leqslant 0$$

$$\begin{bmatrix} -h_1 & \gamma_1 h_2 & \cdots & & \gamma_1 h_n \\ \gamma_2 h_1 & -h_2 & \cdots & & \gamma_2 h_n \\ \vdots & & \ddots & & \vdots \\ \gamma_i h_1 & \cdots & -h_i & & \gamma_i h_n \\ \vdots & & & \ddots & \vdots \\ \gamma_n h_1 & \cdots & \cdots & & -h_n \end{bmatrix} \leftarrow A_2 \begin{bmatrix} p_1 \\ \vdots \\ p_i \\ \vdots \\ p_n \end{bmatrix} \leqslant \underset{b_2}{0} \rightarrow A_2 p \leqslant b_2$$

22-5

minimize $\quad c^T p$

subject to $\quad A_1 p \leq 0$

$\qquad\qquad\quad A_2 p \leq 0$

$$\left.\begin{matrix} A_1 p \leq 0 \\ A_2 p \leq 0 \end{matrix}\right\} \quad \underbrace{\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}}_{A} p \leq \underbrace{\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}}_{b}$$

- example : $\quad$ maximize $\quad x_2 - x_1 \quad \longrightarrow \quad$ minimize $\quad x_1 - x_2$

subject to $\quad 3x_1 = x_2 - 5$

$\qquad\qquad\quad x_1 \leq 0$

$\qquad\qquad\quad |x_2| \leq 2 .$

$c := \begin{bmatrix} 1 \\ -1 \end{bmatrix} . \qquad G := \begin{bmatrix} 3 & -1 \end{bmatrix} \qquad h := -5$

$$|x_2| \leq 2 \longrightarrow -2 \leq x_2 \leq 2 \longrightarrow \left\{ \begin{matrix} x_2 \leq 2 \\ x_2 \geq -2 \end{matrix} \right. \longrightarrow \left\{ \begin{matrix} x_2 \leq 2 \\ -x_2 \leq 2 \end{matrix} \right.$$

$$\begin{cases} x_1 \leq 0 \\ x_2 \leq 2 \\ -x_2 \leq 2 \end{cases} \longrightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix} \longrightarrow A := \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \quad b = \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix}.$$

- example: minimize $c^T x$
  subject to $\|x\|_\infty \leq 1$.

aside:

$$\|x\|_\infty = \max_i |x_i|.$$

$$x = \begin{pmatrix} 2 \\ -5 \\ 3 \end{pmatrix} \rightarrow \|x\|_\infty = 5$$

$$\|x\|_\infty = \text{maximum} \{ |x_1|, \ldots, |x_n| \}$$

$$\downarrow$$

$$\|x\|_\infty \leq 1 \iff \text{maximum} \{ |x_1|, \ldots, |x_n| \} \leq 1$$

$$\Updownarrow$$

$$|x_1| \leq 1, \ldots, |x_n| \leq 1$$

$$\Updownarrow$$

$$-1 \leq x_1 \leq 1, \ldots, -1 \leq x_n \leq 1.$$

$$\left\{\begin{array}{l} x_1 \leq 1 \\ -x_1 \leq 1 \\ \vdots \\ x_n \leq 1 \\ -x_n \leq 1 \end{array}\right. \longrightarrow \quad A := \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & \cdots & \cdots & 0 & 1 \\ 0 & \cdots & \cdots & 0 & -1 \end{bmatrix} \qquad b := \mathbb{1}.$$

- example: minimize $\|Ax - b\|_\infty$.

- example: minimize $\|Ax-b\|_\infty$

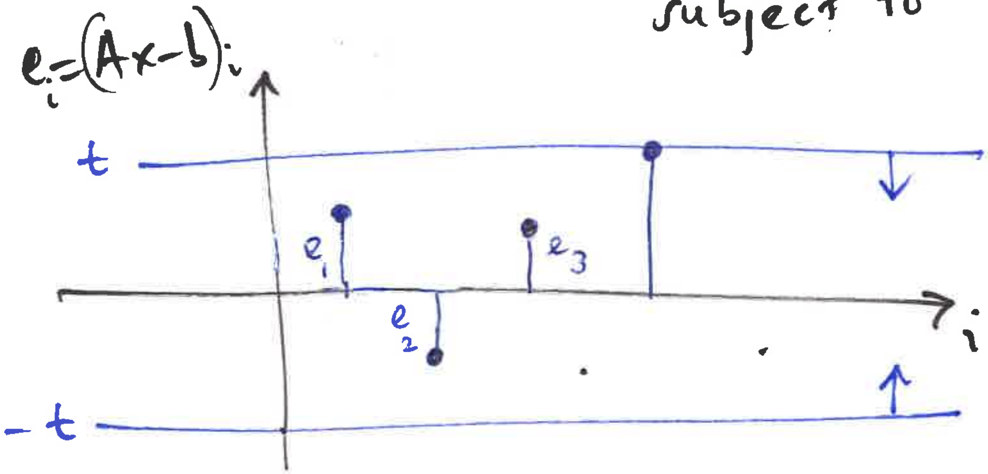
aside:
$\|e\|_\infty = \max_i |e_i|$.

$\leftrightarrow$

minimize t
subject to $\|Ax-b\|_\infty = t$.

$\leftrightarrow$

minimize t
subject to $\|Ax-b\|_\infty \leq t$

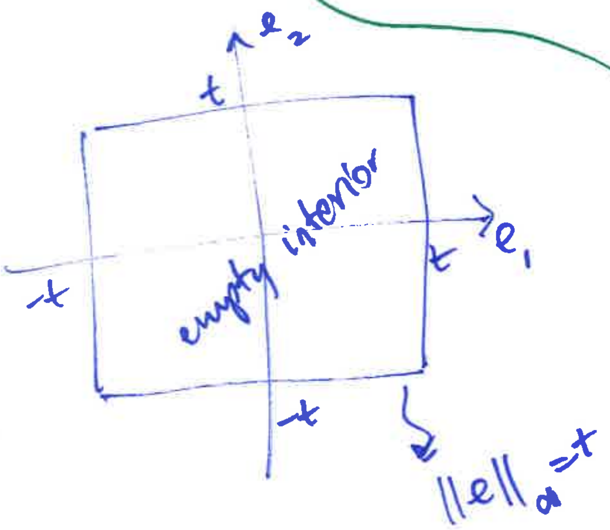


$$\|Ax-b\|_\infty \leq t \leftrightarrow \text{maximum}\{|\underbrace{(Ax-b)_1}_{e_1}|, \dots, |\underbrace{(Ax-b)_m}_{e_m}|\} \leq t$$

$$\leftrightarrow |(Ax-b)_1| \leq t, \dots, |(Ax-b)_m| \leq t$$

$$\leftrightarrow -t \leq (Ax-b)_1 \leq t, \dots, -t \leq (Ax-b)_m \leq t$$

$$\leftrightarrow -t\mathbb{1} \leq Ax-b \leq t\mathbb{1}$$



22-9

new vector of optimization variables $\hat{x} = \begin{bmatrix} x \\ t \end{bmatrix}$.

$$\underset{\begin{bmatrix} x \\ t \end{bmatrix}}{\text{minimize}} \quad \begin{bmatrix} 0^T & 1 \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix}$$

$$c := \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

zero is of same dim. as x.

$$\text{subject to} \quad \begin{bmatrix} A & -\mathbb{1} \\ -A & -\mathbb{1} \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} \leq \begin{bmatrix} b \\ -b \end{bmatrix}$$

22-10

review of last lecture

- examples of turning optimization problems into LPs.

$$\frac{h_i x_i}{\sum_{j\neq i} h_j x_j} \leq \gamma_i \quad \longrightarrow \quad h_i x_i - \gamma_i \sum_{j\neq i} h_j x_j \leq 0$$

$$\|x\|_\infty \leq t \quad \longleftrightarrow \quad \max.\, \{|x_1|, \dots, |x_n|\} \leq t$$

$$\longleftrightarrow \quad |x_1| \leq t, \dots, |x_n| \leq t \quad \longleftrightarrow \quad -t \leq x_1 \leq t, \dots, -t \leq x_n \leq t.$$

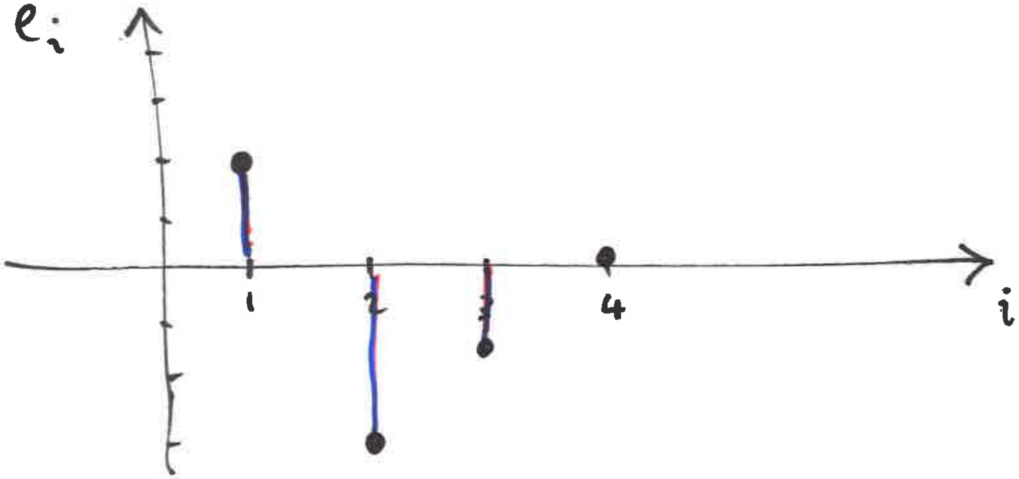
$$\longleftrightarrow \quad -t\mathbb{1} \leq x \leq t\mathbb{1}$$

$$\min.\, \|Ax-b\|_\infty \quad \longleftrightarrow \quad \begin{array}{ll} \min. & t \\ \text{s.t.} & \|Ax-b\|_\infty = t \end{array}$$

$$\longleftrightarrow \quad \begin{array}{ll} \min. & t \\ \text{s.t.} & \|Ax-b\|_\infty \leq t \end{array} \quad \longleftrightarrow \quad \begin{array}{ll} \min. & t \\ \text{s.t.} & -t\mathbb{1} \leq Ax-b \leq t\mathbb{1} \end{array}$$

aside: $e := Ax - b$, $\|e\|_2$ vs $\|e\|_\infty$.

$$\left.\begin{array}{l} x \in \mathbb{R}^n \\ A \in \mathbb{R}^{m\times n} \\ b \in \mathbb{R}^m \end{array}\right\} \to e \in \mathbb{R}^m$$

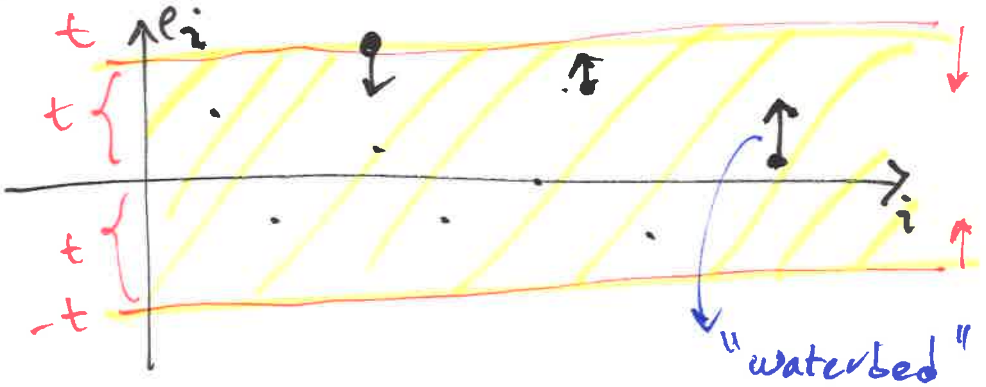


suppose $e = \begin{bmatrix} 2 \\ -3 \\ -1 \\ 0 \end{bmatrix}$

$\downarrow$

$$\left\{\begin{array}{l} \|e\|_2 = \sqrt{2^2 + (-3)^2 + (-1)^2 + 0} \\ \|e\|_\infty = 3 \end{array}\right.$$

minimize $\|Ax - b\|_\infty = \|e\|_\infty$



"waterbed" effect

$|e_i| \leq t \quad \forall i$

- example:

$$\begin{array}{ll} \text{minimize} & \|x\|_\infty \\ \text{subject to} & Ax=b. \end{array}$$

$$\downarrow$$

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \|x\|_\infty \leq t \\ & Ax=b \end{array}$$

relaxed from equality (which will always be satisfied)

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -t\,\mathbb{1} \leq x \leq t\,\mathbb{1} \\ & Ax=b \end{array}$$

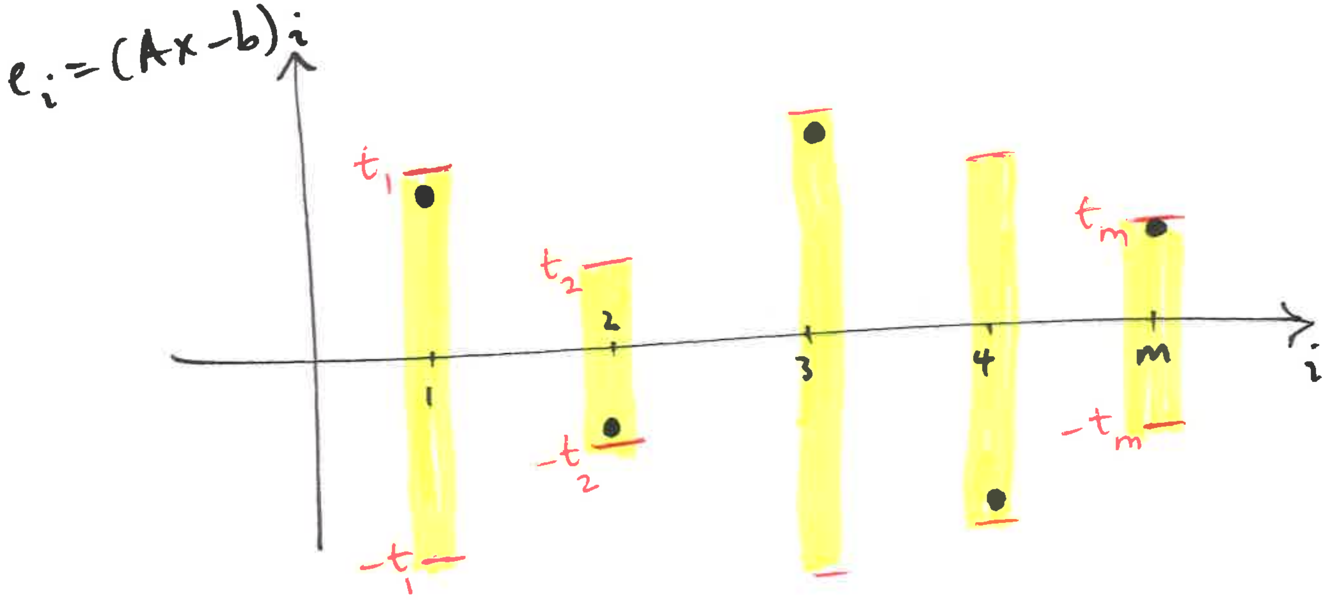
$$\downarrow$$

$$x - t\,\mathbb{1} \leq 0$$
$$-x - t\,\mathbb{1} \leq 0$$

$$\downarrow$$

$$\begin{bmatrix} +I & -\mathbb{1} \\ -I & -\mathbb{1} \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

- example: minimize $\|Ax-b\|_1$



aside 1:

$$\|e\|_1 = |e_1| + \cdots + |e_m|$$

aside 2:

if $Ax-b \geqslant 0$

↓

$$\|Ax-b\|_1 = \mathbb{1}^T(Ax-b) = \underbrace{(\mathbb{1}^T A)}_{c^T} x - \underbrace{\mathbb{1}^T b}_{d}$$

$$\|Ax-b\|_1 = \underbrace{|(Ax-b)_1|}_{t_1} + \cdots + \underbrace{|(Ax-b)_m|}_{t_m}$$

$$\text{min. } \|Ax-b\|_1 \quad \longleftrightarrow \quad \begin{aligned} &\text{min.} && t_1 + \cdots + t_m \\ &\text{s.t.} && |(Ax-b)_i| = t_i \qquad i=1,\ldots,m \end{aligned}$$

$$\Updownarrow$$

$$\begin{array}{ll} \text{min.} & t_1 + \cdots + t_m \\ \text{s.t.} & |(Ax-b)_i| \leq t_i \qquad i=1,\ldots,m \end{array}$$

$$|(Ax-b)_i| \leq t_i \iff -t_i \leq (Ax-b)_i \leq t_i$$

$$\iff -\begin{bmatrix} t_1 \\ \vdots \\ t_m \end{bmatrix} \leq \begin{bmatrix} (Ax-b)_1 \\ \vdots \\ (Ax-b)_m \end{bmatrix} \leq \begin{bmatrix} t_1 \\ \vdots \\ t_m \end{bmatrix}$$

$$\iff -t \leq Ax-b \leq t.$$

$$\Updownarrow$$

↕

$$\begin{array}{ll} \underset{\left[\begin{smallmatrix} x \\ t \end{smallmatrix}\right]}{\text{minimize}} & \mathbb{1}^T t \\ \text{subject to} & -t \leq Ax - b \leq t \end{array}$$

$$\hat{x} = \begin{bmatrix} x \\ t \end{bmatrix} \in \mathbb{R}^{n+m}$$

↓

$$\begin{array}{ll} \text{minimize} & \begin{bmatrix} 0^T & \mathbb{1}^T \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} \\ \text{subject to} & \begin{bmatrix} A & -I \\ -A & -I \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} \leq \begin{bmatrix} b \\ -b \end{bmatrix}. \end{array}$$

---

aside: Compare problem on top of page with the "min $\|Ax-b\|_\infty$" problem:

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -t\,\mathbb{1} \leq Ax - b \leq t\,\mathbb{1}. \end{array}$$

- example:

$$\left.\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax=b \end{array}\right\} \sim \text{compressed sensing}$$

$$\updownarrow$$

$$\begin{array}{lll} \text{minimize} & t_1+\cdots+t_n & \\ \text{subject to} & |x_i|=t_i & i=1,\dots,n \\ & Ax=b & \end{array}$$

$$\updownarrow$$

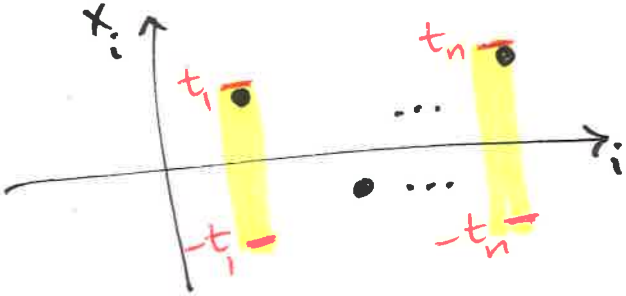
$$\begin{array}{lll} \text{minimize} & t_1+\cdots+t_n & \\ \text{subject to} & |x_i|\le t_i & i=1,\dots,n \\ & Ax=b & \end{array}$$

$$\updownarrow$$

$$\begin{array}{ll} \text{minimize} & \mathbb{1}^T t \\ \text{subject to} & -t\le x\le t \\ & Ax=b. \end{array}$$

<u>aside</u>: if x is "sparse," in the sense that it has a lot of zeros, then A can be very fat & you can still find the actual x

(in this scenario, x would be an actual unknown and A,b would represent the measuremen of x.)



- geometric interpretation

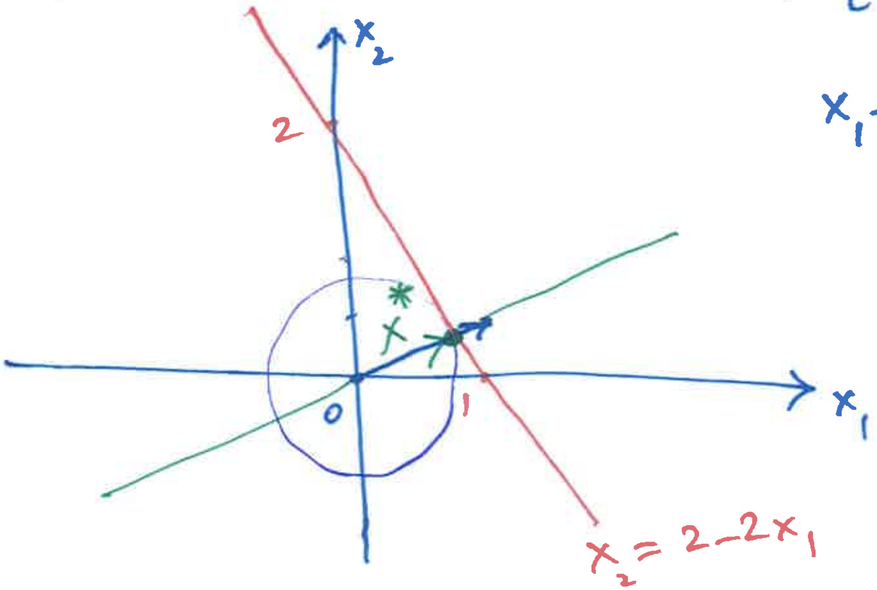
recall least-norm

minimize $\|x\|_2^2$

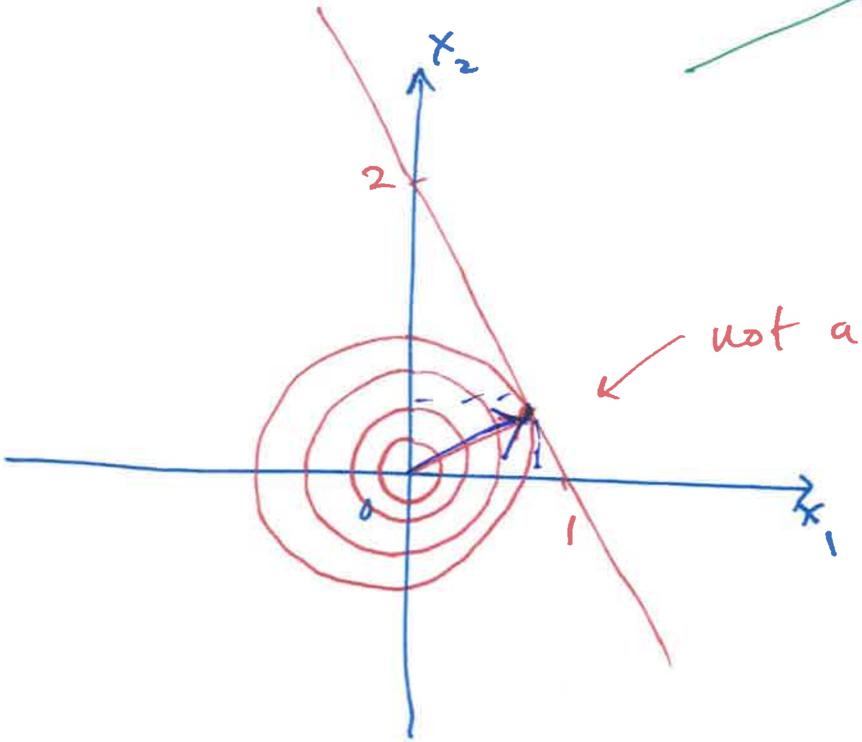
subject to $1 = \begin{bmatrix} 1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$x_1 + \frac{x_2}{2} = 1$$

$$x_2 = 2 - 2x_1$$

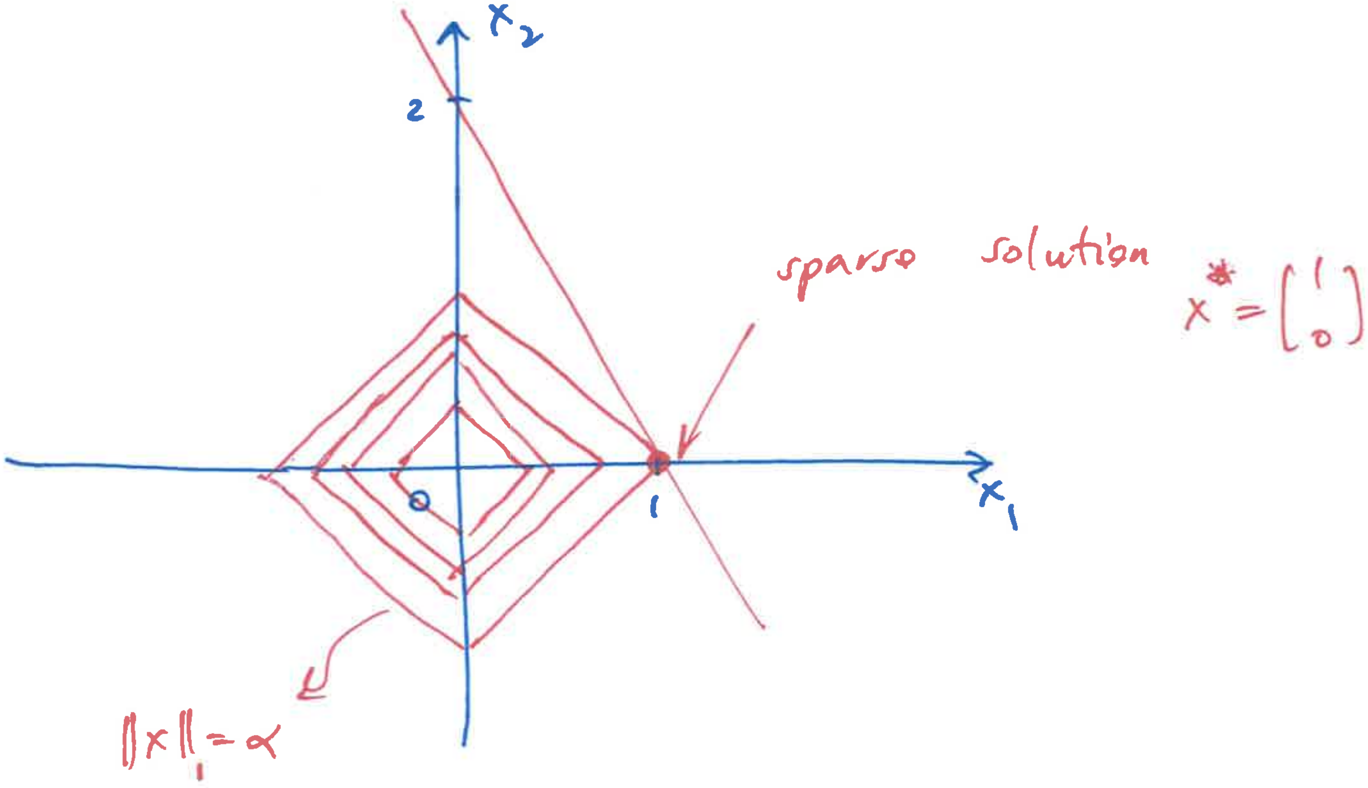


not a sparse solution



minimize $\|x\|_1$

subject to $1 = \begin{bmatrix} 1 & \frac{1}{2} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



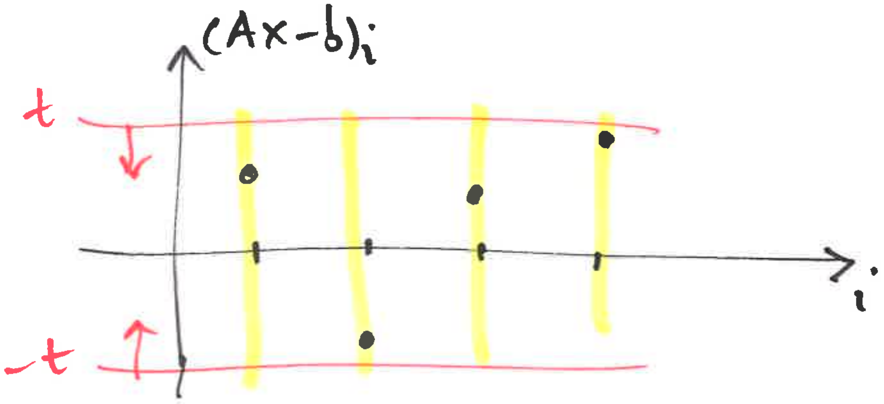
$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

review of last lecture

- minimize $\|Ax-b\|_\infty$ $\longrightarrow$ minimize $t \nearrow \in \mathbb{R}.$

  subject to $-t\mathbb{1} \leqslant Ax-b \leqslant t\mathbb{1}$

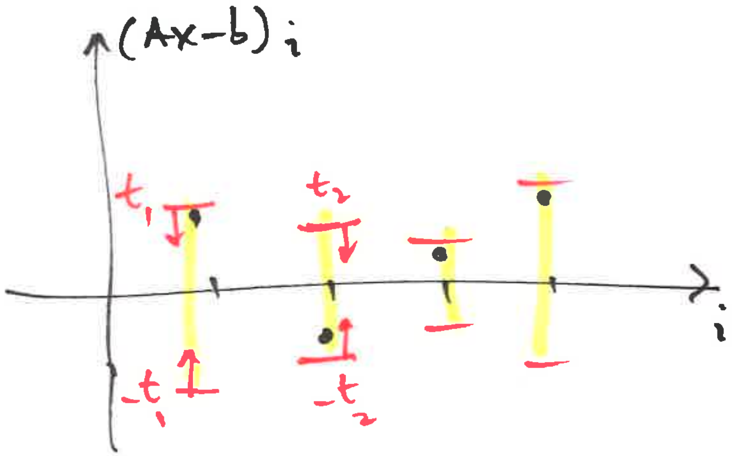
  $\hookrightarrow$ minimization over $\begin{bmatrix} x \\ t \end{bmatrix} \in \mathbb{R}^{n+1}$



- minimize $\|Ax-b\|_1$ $\longrightarrow$ minimize $\mathbb{1}^T t \nearrow \in \mathbb{R}^m$

  subject to $-t \leqslant Ax-b \leqslant t$

  $\hookrightarrow$ minimization over $\begin{bmatrix} x \\ t \end{bmatrix} \in \mathbb{R}^{n+m}$



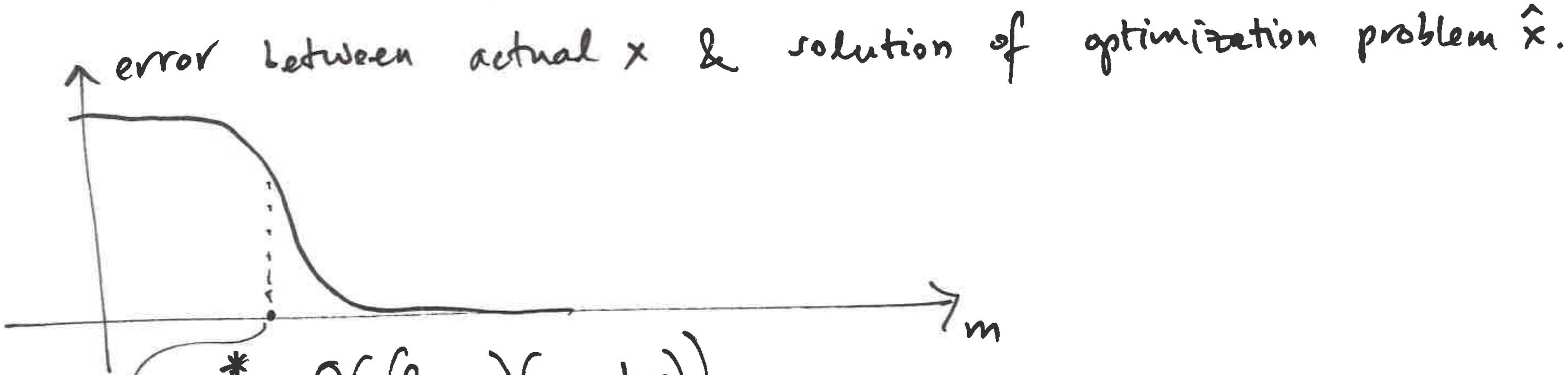
- sparsity-promoting properties of the 1-norm → compressive sensing.

- compressive/compressed sensing

$$\text{minimize} \quad \|x\|_1 = \sum_i |x_i| \qquad x \in \mathbb{R}^n, \quad A \in \mathbb{R}^{m \times n}, \quad m \leq n$$

$$\text{subject to} \quad Ax = y$$

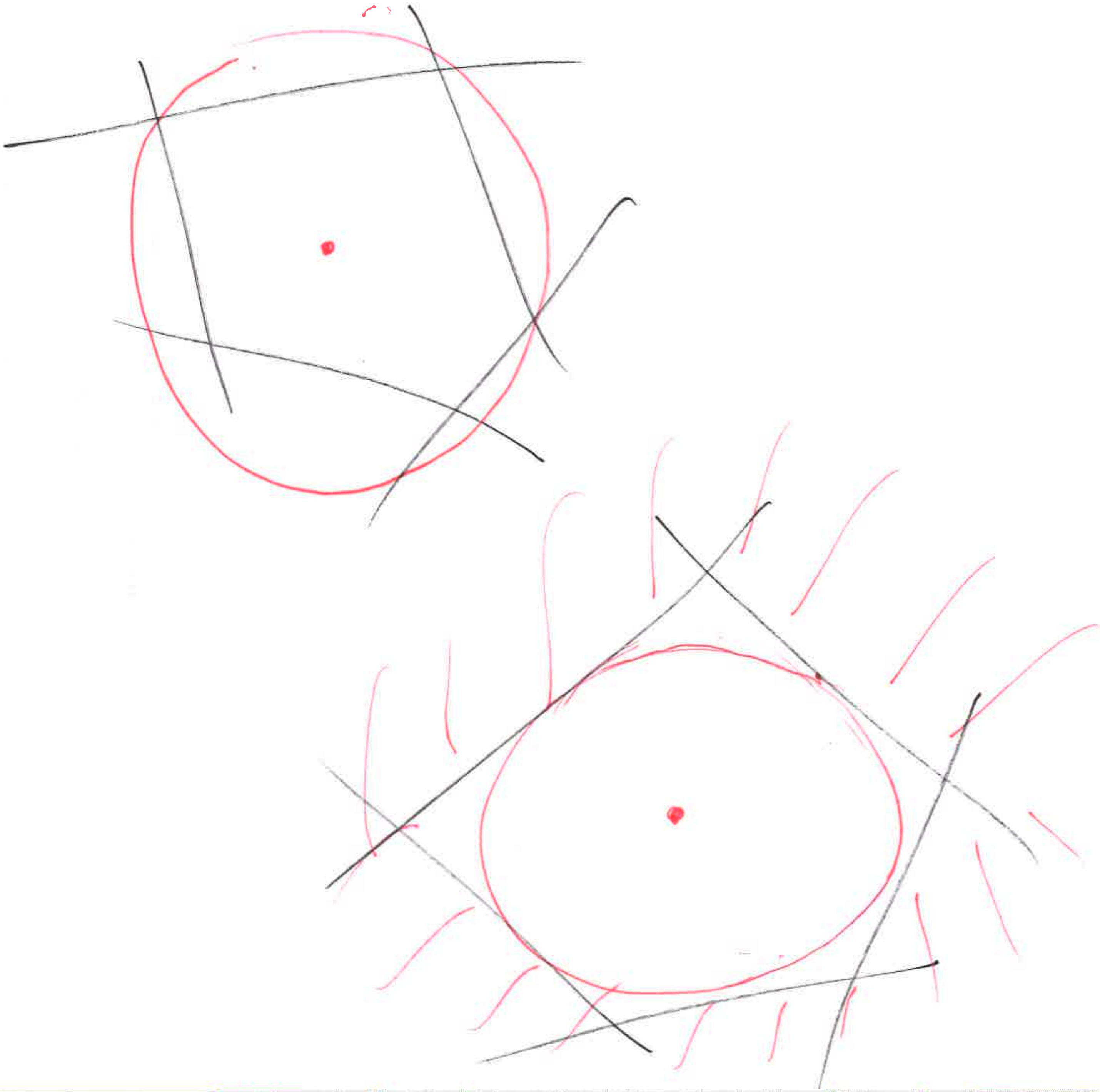
(interesting when A very fat)



$$m^* \sim O\left((\log n)(\text{card } x)\right)$$

$$n = 10^6, \quad \text{card } x \approx 10^3 \quad \longrightarrow \quad m \sim O(6000)$$

so instead of 10^6 measurements we can get away with 6000, a 166-fold reduction

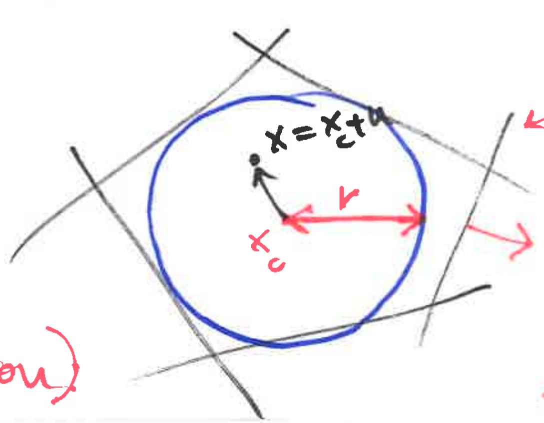


- Chebyshev center of a polyhedron

this is defined as the point in the polyhedron $P=\{x \mid a_i^T x \le b_i\ ,\ i=1,\dots,m\}$ which is the center of the largest inscribed ball,

$$B(x_c,r)=\{x_c+u \mid \|u\|_2 \le r\}.$$

(x_c: "deepest" point inside the polyhedron)

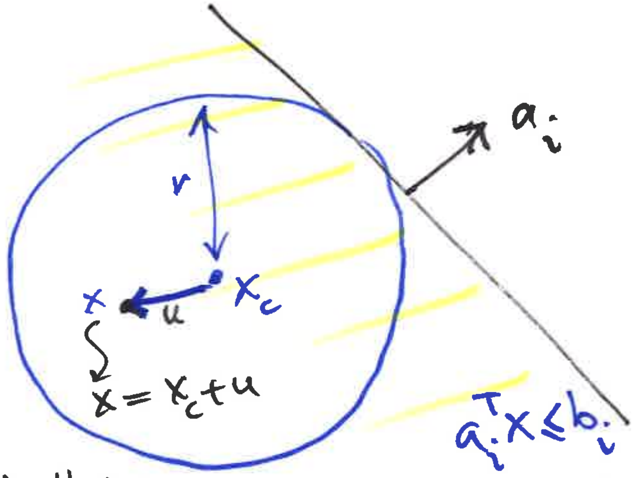


24-3

- example: given A, b, find x_c and r.

objective: maximize r

constraint: ball is inside the polyhedron



i^{th} constraint: $a_i^T x \le b_i$ for all $x \in B(x_c, r)$

$$\Updownarrow$$

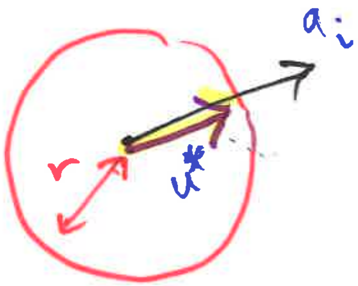
$a_i^T(x_c+u) \le b_i$ for all u with $\|u\|_2 \le r$

$$\Updownarrow$$

$$\left\{ \underset{\|u\|_2 \le r}{\text{maximum}} \; a_i^T(x_c+u) \right\} \le b_i$$

$$\Updownarrow$$

$$a_i^T x_c + \underbrace{\left\{ \underset{\|u\|_2 \leq r}{\text{maximum}} \quad a_i^T u \right\}}_{u^* = r \frac{a_i}{\|a_i\|_2}} \leq b_i$$



$$\left(a_i^T u = \|a_i\| \, \|u\| \cos\theta \right)$$

$$\Updownarrow$$

$$a_i^T x_c + a_i^T \left(r \frac{a_i}{\|a_i\|_2} \right) \leq b_i$$

$$\Updownarrow$$

$$a_i^T x_c + \|a_i\|_2 \, r \leq b_i$$

now, repeat this procedure for every $i = 1, \dots, m$

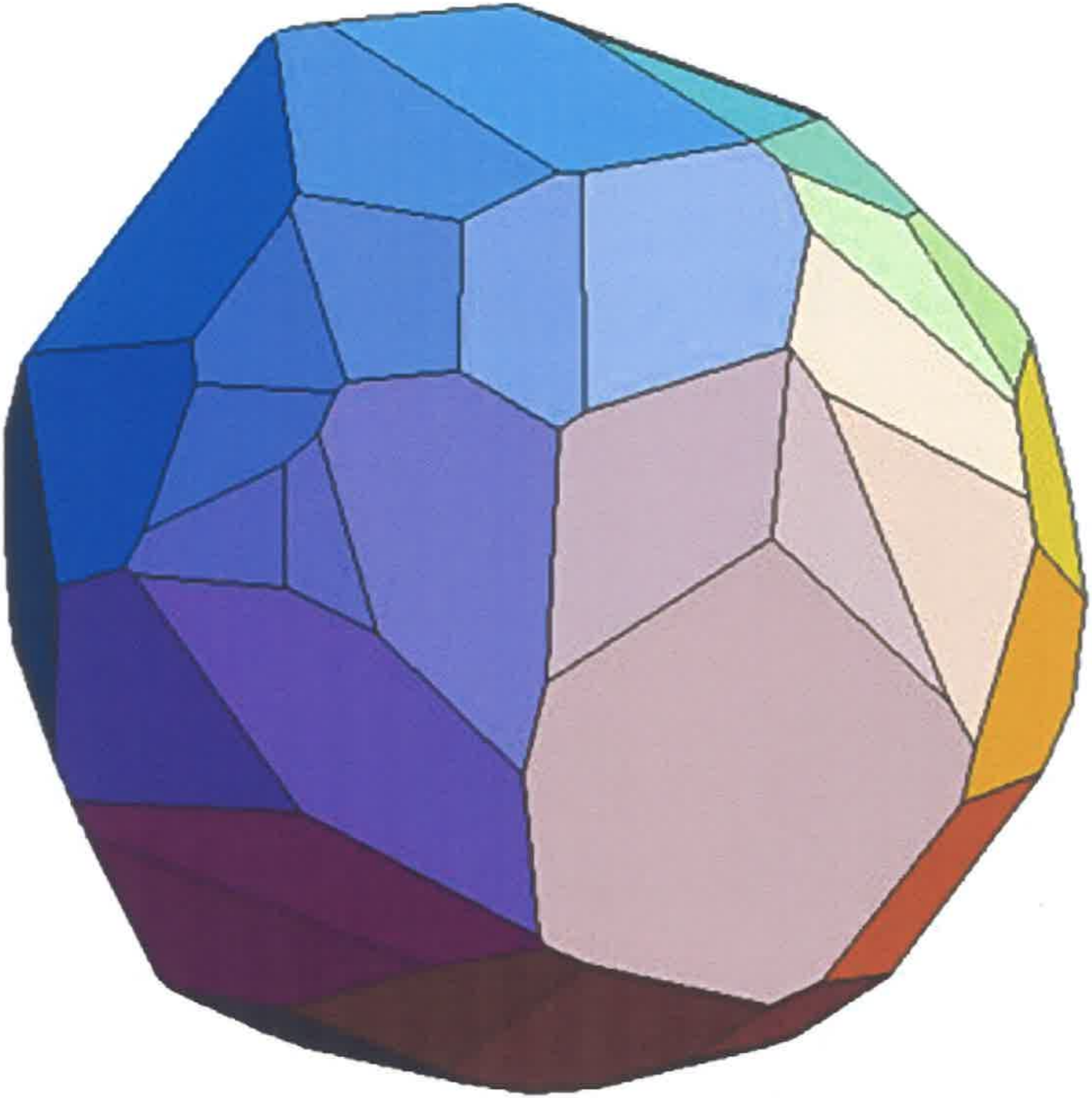
$$\underbrace{\begin{bmatrix} — a_1^T — & \|a_1\|_2 \\ \vdots & \vdots \\ — a_m^T — & \|a_m\|_2 \end{bmatrix}}_{A} \begin{bmatrix} x_c \\ \hline r \end{bmatrix} \leqslant \underbrace{\begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}}_{b} .$$

the formulation of the optimization as an LP is given by

$$\begin{array}{ll} \text{maximize} & r \\ \text{subject to} & A \begin{bmatrix} x_c \\ r \end{bmatrix} \leqslant b . \end{array}$$

$$\downarrow$$

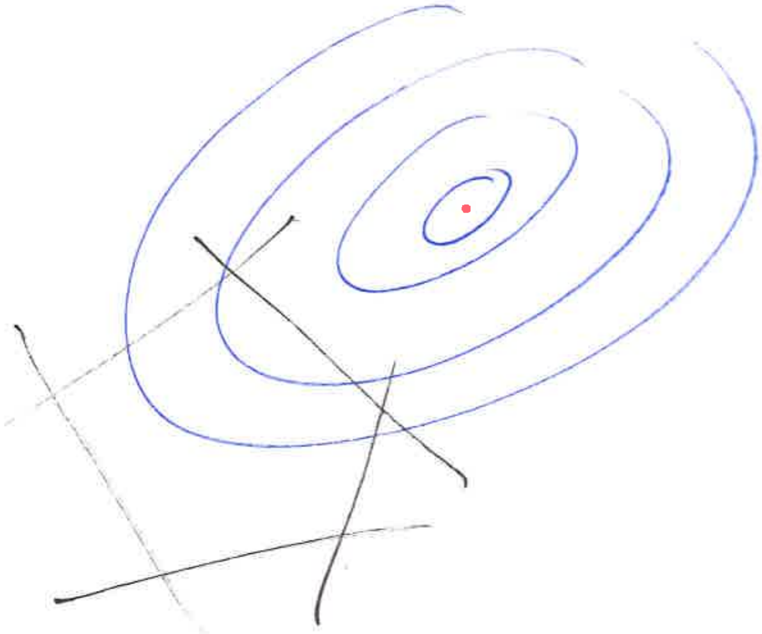
$$\begin{array}{ll} \text{minimize} & -\begin{bmatrix} 0^T & 1 \end{bmatrix} \begin{bmatrix} x_c \\ r \end{bmatrix} \\ \text{subject to} & A \begin{bmatrix} x_c \\ r \end{bmatrix} \leqslant b . \end{array}$$



- quadratic program (QP)

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^TPx + q^Tx + r \\ \text{subject to} & Ax \leq b \\ & Gx = h \end{array}$$

we need $P \succeq 0$.



- example: $$\begin{array}{ll} \text{minimize} & \|Ax-b\|_2^2 = (Ax-b)^T(Ax-b) \\ \text{subject to} & l \leq x \leq u \end{array}$$

$$= x^TA^TAx - 2b^TAx + b^Tb$$

$P = 2A^TA$

$q^T = -2b^TA$

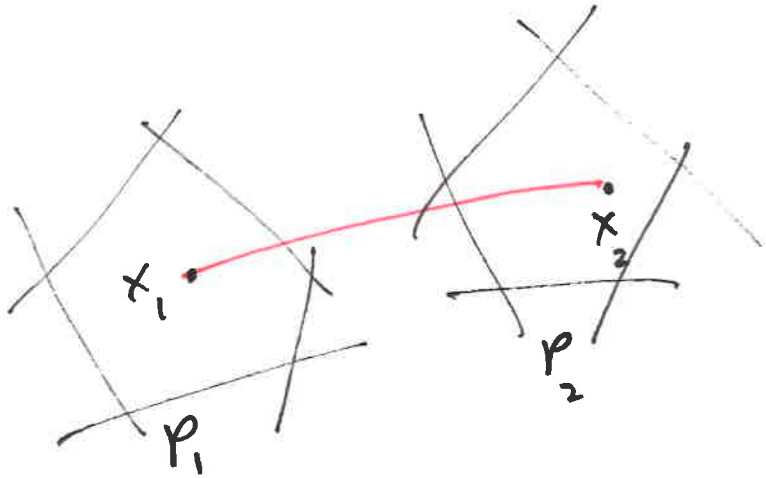
$r = b^Tb$.

- example: find the distance between two given polyhedra

$$P_1 = \{x \mid A_1 x \leq b_1\}, \quad P_2 = \{x \mid A_2 x \leq b_2\}$$

$$\mathrm{dist}(P_1, P_2) = \min.\{\|x_1 - x_2\|_2 \mid x_1 \in P_1,\ x_2 \in P_2\}.$$

$$\begin{array}{lll} \underset{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}{\text{minimize}} & & \|x_1 - x_2\|_2^2 \\ \text{subject} & \text{to} & A_1 x_1 \leq b_1 \\ & & A_2 x_2 \leq b_2 \end{array}$$



$$\|x_1 - x_2\|_2^2 = (x_1 - x_2)^T (x_1 - x_2) = \begin{bmatrix} x_1^T & x_2^T \end{bmatrix} \begin{bmatrix} I \\ -I \end{bmatrix} \begin{bmatrix} I & -I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} x_1^T & x_2^T \end{bmatrix} \underbrace{2 \begin{bmatrix} I & -I \\ -I & I \end{bmatrix}}_{P} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

- example : portfolio optimization

consider a portfolio with n stocks held over a period of time T. assume

x_i : fraction of portfolio invested in stock i

p_i : rate of return of asset i

$$\left(= \frac{\text{price at time } T - \text{price at time } 0}{\text{price at time } 0} \right)$$

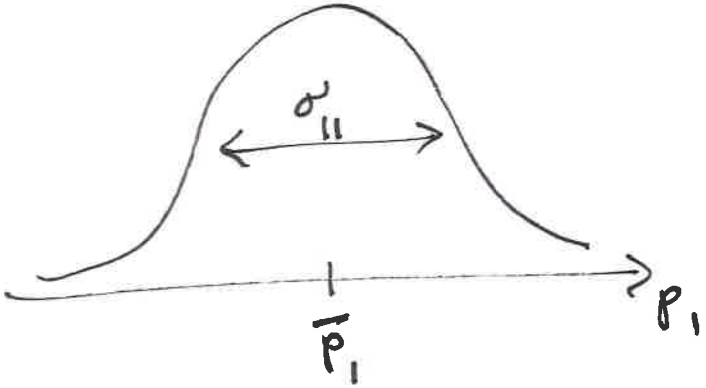
r : overall rate of return on portfolio.

furthermore, we assume that the rate of return p_i for stock i is a random variable such that

$$\mathcal{E}\{r\} = \underset{\substack{\downarrow \\ \text{mean vector}}}{\bar{p}} , \quad \mathcal{E}\{(p-\bar{p})(p-\bar{p})^T\} = \Sigma \; \overset{\text{covariance matrix}}{\swarrow}$$

aside:

$$p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} \qquad \mathcal{E}\{p\} = \begin{bmatrix} \mathcal{E}\{p_1\} \\ \vdots \\ \mathcal{E}\{p_n\} \end{bmatrix} = \bar{p}$$



$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \\ \sigma_{21} & \ddots & & \sigma_{ij} \\ \vdots & & \ddots & \\ & & & \sigma_{nn} \end{bmatrix}$$

"correlation" between i & j

$$\rho_{ij} := \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

$$-1 \leq \rho_{ij} \leq 1$$

($\sigma_{ii} =: \sigma_i^2$ is the variance of variable i) therefore, for portfolio $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, the return rate

$r = p_1 x_1 + \cdots + p_n x_n = p^T x$ is a random variable

with $\mathcal{E}\{r\} = \bar{r}$, $\mathcal{E}\{(r-\bar{r})^2\} = \sigma_r^2$.

find the optimal portfolio that minimizes the return variance ("risk" or "uncertainty" of portfolio) subject to a minimum acceptable return rate r_{min}.

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \text{risk} = \sigma_r^2 \\ \text{subject to} & \bar{r} \geq r_{min} \\ & x \geq 0, \ \mathbb{1}^T x = 1 \end{array}$$

$$\bar{r} = \mathcal{E}\{p^T x\} = \left(\mathcal{E}\{p\}\right)^T x = (\bar{p})^T x .$$

$$\begin{aligned}
\sigma_r^2 = \mathcal{E}\{(r-\bar{r})^2\} &= \mathcal{E}\{(p^T x - (\bar{p})^T x)^2\} = \mathcal{E}\{[(p-\bar{p})^T x]^2\} \\
&= \mathcal{E}\{[x^T (p-\bar{p})][(p-\bar{p})^T x]\} \\
&= x^T \mathcal{E}\{(p-\bar{p})(p-\bar{p})^T\} x \\
&= x^T \Sigma x .
\end{aligned}$$

$$\downarrow$$

$$\begin{array}{ll}
\text{minimize} & x^T \Sigma x \\
\text{subject to} & (\bar{p})^T x \geq r_{\min} \\
& x \geq 0 , \quad \mathbb{1}^T x = 1 .
\end{array}$$

- another formulation

determines how much risk I am willing to take.

$$\begin{aligned} \text{minimize} \quad & -(\bar{p})^T x + \gamma\, x^T \Sigma x . \\ \text{subject to} \quad & x \geqslant 0, \quad \mathbb{1}^T x = 1 . \end{aligned}$$

- yet another formulation

$$\begin{aligned} \text{minimize} \quad & -\bar{p}^T x \\ \text{subject to} \quad & x^T \Sigma x \leq s_{max} \\ & x \geqslant 0, \quad \mathbb{1}^T x = 1 \end{aligned}$$

not a linear constraint

(but can still be solved after some tricks that are beyond the scope of this class).

- recall the diet problem:

x_j : amount of food j

F_{ij} : amount of nutrient i in food j

r_i : minimum daily requirement of nutrient i

c_j : cost of serving unit amount of food j

$$\begin{array}{ll}\text{minimize} & c^T x \\ \text{subject to} & Fx \geq r \\ & x \geq 0\end{array} \xrightarrow[\text{put in standard LP form}]{\text{rewrite to}} \begin{array}{l} -Fx \leq -r \\ -x \leq 0 \end{array}$$

$$\downarrow$$

$$\underbrace{\begin{bmatrix}-F \\ -I\end{bmatrix}}_{A} x \leq \underbrace{\begin{bmatrix}-r \\ 0\end{bmatrix}}_{b}$$

Kantorovich.

- example: diet problem with uncertainty

$$\begin{array}{ll} \text{minimize} & \mathcal{E}\{c^T x\} + \gamma \operatorname{var}\{c^T x\} \\ \text{subject to} & Fx \succeq r, \quad Ax = b, \quad x \succeq 0 \end{array}$$

with $\mathcal{E}\{c\} = \bar{c}$, $\mathcal{E}\{(c-\bar{c})(c-\bar{c})^T\} = \Sigma$

c is a random variable with mean c & variance Σ.

$$\operatorname{var}(c^T x) = \mathcal{E}\{(c^T x - \bar{c}^T x)(\cdots)^T\} = \mathcal{E}\{x^T (c-\bar{c})(c-\bar{c})^T x\}$$

$$= x^T \Sigma x.$$

γ is the risk-aversion parameter

$$\begin{array}{ll} \text{minimize} & \bar{c}^T x + \gamma x^T \Sigma x \\ \text{subject to} & Fx \succeq r, \quad Ax = b, \quad x \succeq 0. \end{array}$$

clustering

(Stephen Boyd, "Vectors, Matrices, & Least Squares".)

- suppose we have N vectors, $x_1, \dots, x_N$, and we'd like to cluster them into k groups with vectors in every cluster "similar" to each other

$x_i \in \mathbb{R}^n \quad i=1,\dots,N$, seek k clusters.

- applications:
  - customer segmentation

$$x_i \sim \begin{bmatrix} \$ \\ \vdots \\ \$ \end{bmatrix}_{n\times 1}$$

group customers with similar purchasing patterns.

- survey response

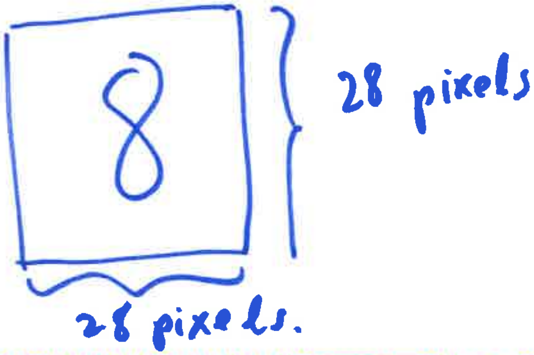
strongly disagree	disagree	neutral	agree	strongly agree
(-2)	(-1)	(0)	(1)	(2)

$x_i \sim$ response vector to n questions

- energy use pattern

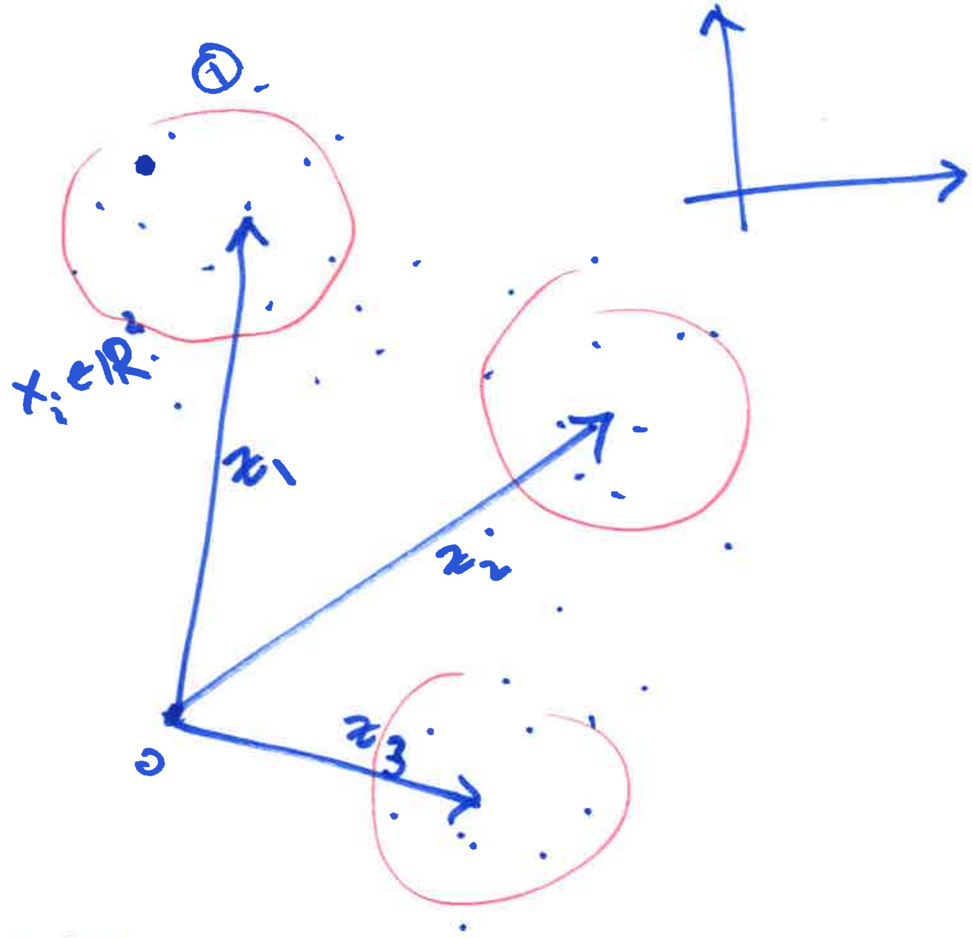
$x_i \sim$ average electricity use over n time periods.

- image clustering



- special case $n=2$

- in reality n really large
- clusters not so well separated
- number of clusters not clear



- we use vector c to characterize clustering of data

$$c = \begin{bmatrix} 3 \\ 1 \\ 1 \\ 1 \\ 2 \end{bmatrix} \qquad G_1 = \{2,3,4\} \qquad G_2 = \{5\} \qquad G_3 = \{1\}.$$

c_i indicates cluster to which x_i belongs

z_j denotes center of cluster j

$$J = \frac{1}{N}\left(\|x_1 - z_{c_1}\|^2 + \|x_2 - z_{c_2}\| + \cdots + \|x_N - z_{c_N}\|^2 \right)$$

would have obvious solution z_{c_i} = x_i if we were allowed N clusters!

$$\min. \ \frac{1}{N}\sum_{i=1}^{N} \|x_i - y_i\|^2$$

s.t. the set $\{y_1, \ldots, y_N\}$ has only k distinct values.

once found, will give $\{z_1, \ldots, z_k\}$.

- a heuristic method

step1 assuming that the z_j are known, we find group assignments

$$\|x_i - z_{c_i}\|^2 = \min_{j=1,\ldots,k} \|x_i - z_j\|^2$$

step 2 assuming clusters are known, we find new z_i

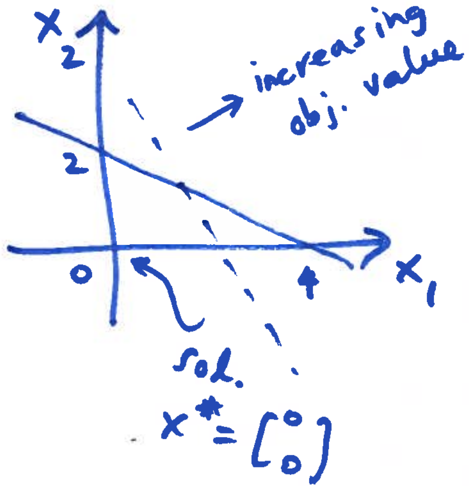
$$J_j = \frac{1}{N} \sum_{i \in G_j} \| x_i - z_j \|^2$$

$$z_j = \frac{1}{\underbrace{|G_j|}_{\text{number of vectors in cluster } j}} \sum_{i \in G_j} x_i .$$

- the iteration between step 1 & step 2 is known as the "k-means" algorithm.

- example: use KKT conditions to solve the LP

$$\begin{aligned} \text{minimize} \quad & 2x_1 + x_2 \\ \text{subject to} \quad & x_1 \geq 0, \quad x_2 \geq 0 \\ & x_1 + 2x_2 \leq 4 \end{aligned}$$



$$\mathcal{L}(x,\lambda) = 2x_1 + x_2 + \lambda_1(-x_1) + \lambda_2(-x_2) + \lambda_3(x_1 + 2x_2 - 4)$$

KKT

$$\left\{ \begin{array}{l} \nabla_x \mathcal{L} = \begin{bmatrix} 2 - \lambda_1 + \lambda_3 \\ 1 - \lambda_2 + 2\lambda_3 \end{bmatrix} = 0 \rightarrow \begin{cases} 2 - \lambda_1 + \lambda_3 = 0 & \textcircled{1} \\ 1 - \lambda_2 + 2\lambda_3 = 0 & \textcircled{2} \end{cases} \qquad \begin{array}{l} \lambda_1, \lambda_2, \lambda_3 \geq 0 \\ x_1 \geq 0, x_2 \geq 0, x_1 + 2x_2 \leq 4 \end{array} \\ \\ \lambda_1 x_1 = 0 \quad \textcircled{3} \qquad \lambda_2 x_2 = 0 \quad \textcircled{4} \qquad \lambda_3(x_1 + 2x_2 - 4) = 0 \quad \textcircled{5} \end{array} \right\}$$

let's try some points in the feasible set (we can also use knowledge that solution occurs at vertex)

$*$ $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$:

$$x_1 = 1 \xrightarrow{\textcircled{3}} \lambda_1 = 0 \xrightarrow{\textcircled{1}} \lambda_3 = -2$$

$$x_2 = 1 \xrightarrow{\textcircled{4}} \lambda_2 = 0 \xrightarrow{\textcircled{2}} \lambda_3 = -1/2$$

inconsistent & negative values of λ_3

$$\begin{tikzcd} * \quad x = \begin{bmatrix} 4 \\ 0 \end{bmatrix} : & \begin{matrix} x_1 = 4 \\ x_2 = 0 \end{matrix} \arrow[r, "③"] & \lambda_1 = 0 \arrow[r, "①"] & \lambda_3 = -2 \arrow[d, bend right, "②"'] \\ & & & \lambda_2 = -3 \end{tikzcd}$$

negative λ_2, λ_3

$$\begin{tikzcd} * \quad x = \begin{bmatrix} 0 \\ 2 \end{bmatrix} : & \begin{matrix} x_1 = 0 \\ x_2 = 2 \end{matrix} \arrow[r, "④"] & \lambda_2 = 0 \arrow[r, "②"] & \lambda_3 = -1/2 \arrow[r, "①"] & \lambda_1 = 3/2 \end{tikzcd}$$

negative λ_3

$$\begin{tikzcd} & & & \lambda_1 = 2 \\ * \quad x = \begin{bmatrix} 0 \\ 0 \end{bmatrix} : & \begin{matrix} x_1 = 0 \\ x_2 = 0 \end{matrix} \arrow[r, "⑤"] & \lambda_3 = 0 \arrow[ur, "①"] \arrow[dr, "②"'] & & \left\{ \begin{matrix} x_1^* = x_2^* = 0 \\ \lambda_1^* = 1, \ \lambda_2^* = 2, \ \lambda_3^* = 0 \end{matrix} \right. \\ & & & \lambda_2 = 1 \end{tikzcd}$$

(this is not an efficient method for solving LPs; we'll talk about efficient numerical schemes next)

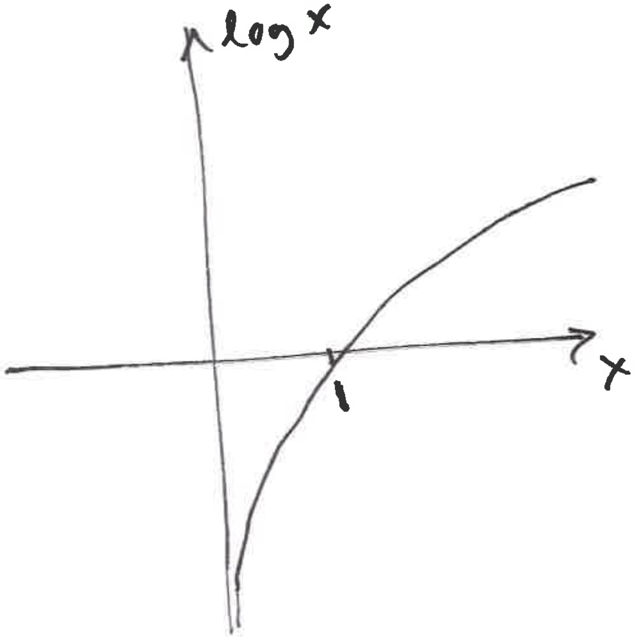
- numerical methods for solving LPs.
  - simplex method.
  - interior point method

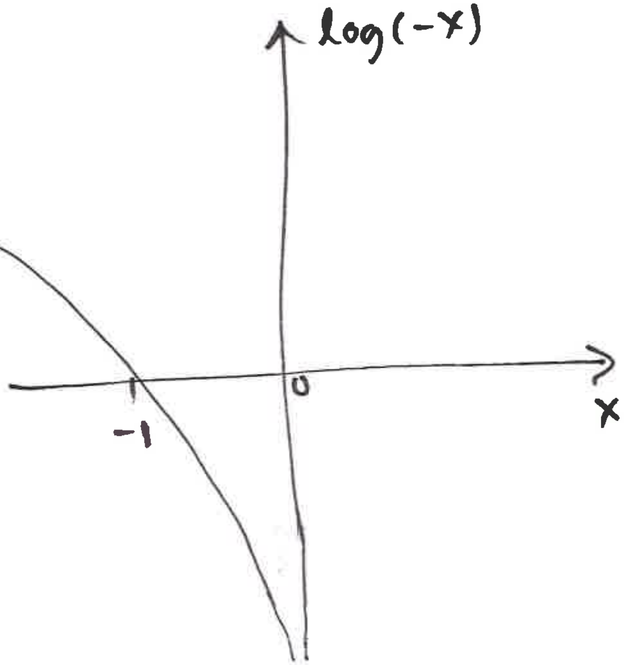
$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \boxed{Ax \leq b}. \end{array} \longrightarrow \begin{array}{c} a_1^T x \leq b_1 \\ \vdots \\ a_m^T x \leq b_m \end{array}$$

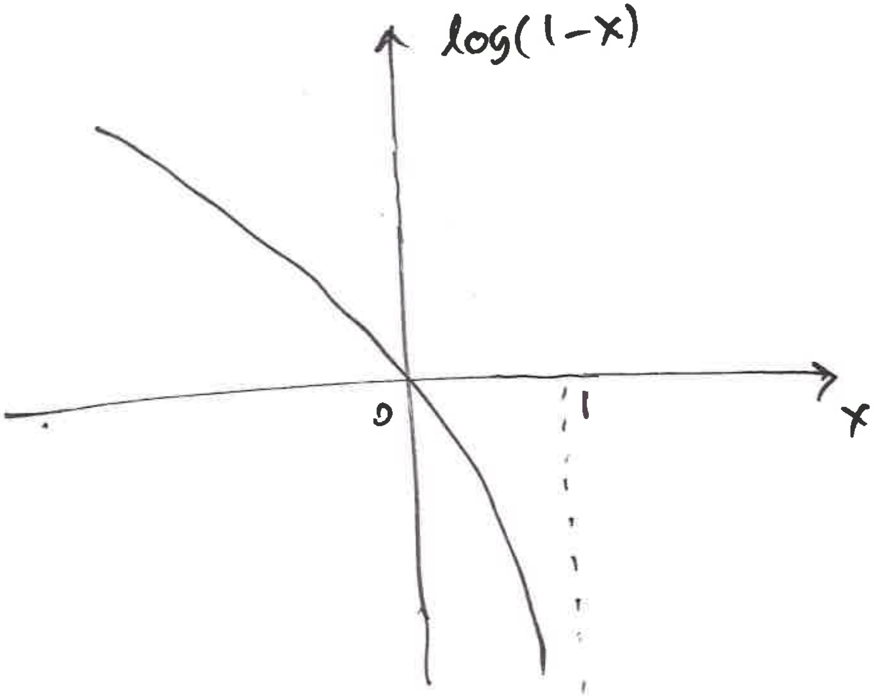
$$\text{minimize} \quad c^T x - \frac{1}{t} \sum_{i=1}^{m} \underbrace{\log(b_i - a_i^T x)}_{\text{"log barrier"}}$$

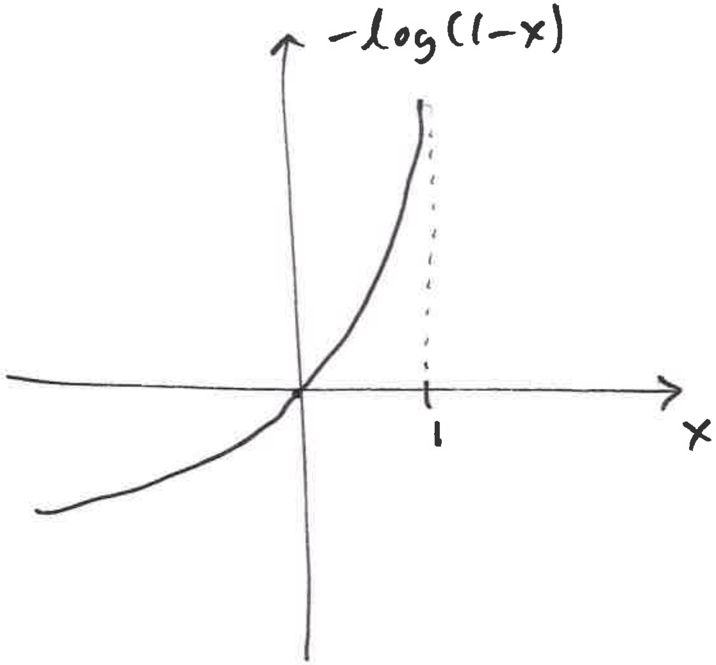
$t \to \infty$.

suppose $x \in \mathbb{R}$, $a=1$, $b=1$ $\rightarrow$ $x \leqslant 1.$ $\rightarrow$ $1-x \geqslant 0$

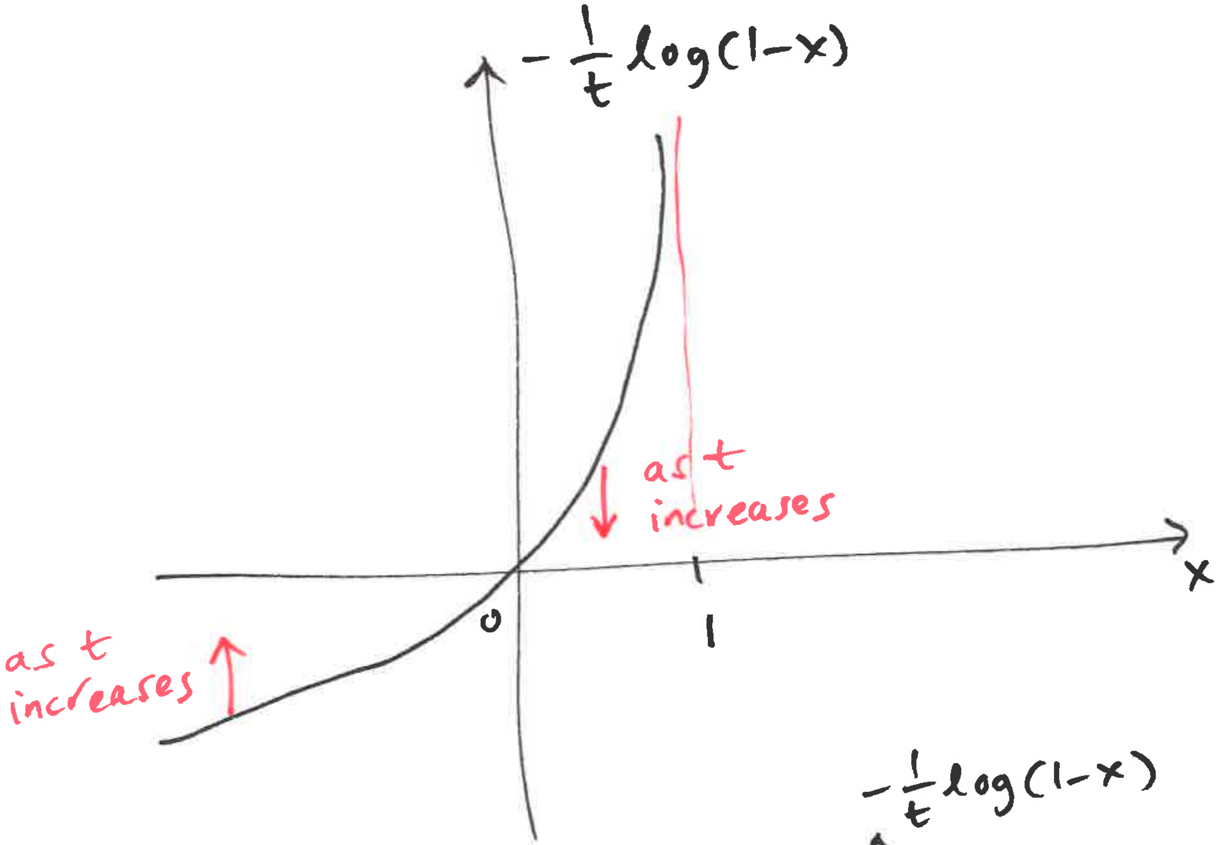






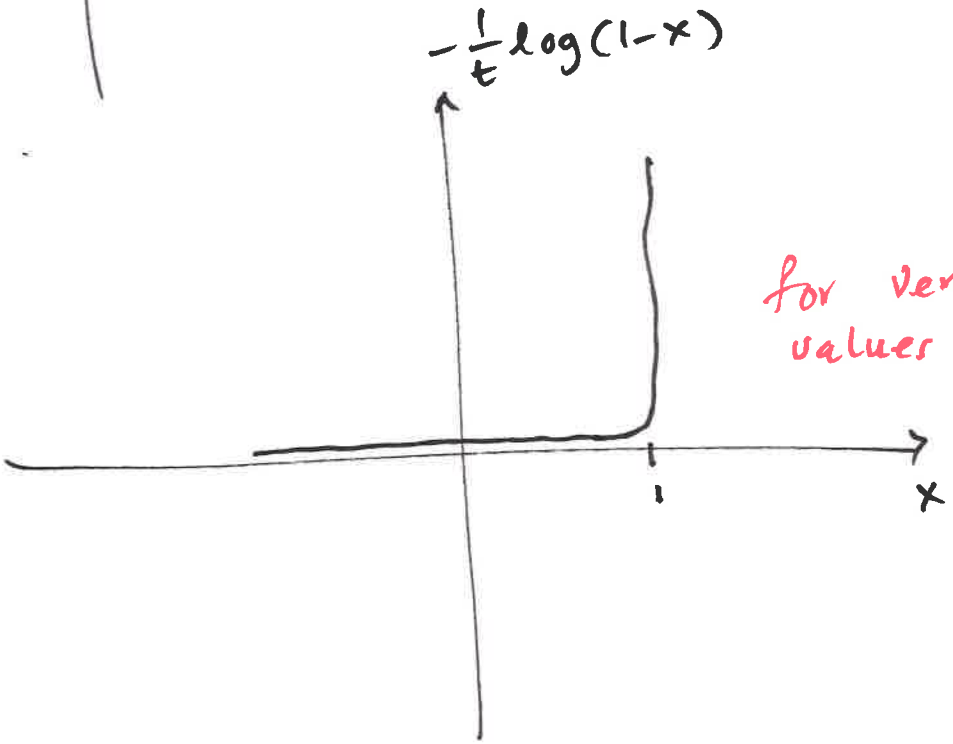


24-8



as t increases

as t increases



for very large values of t.

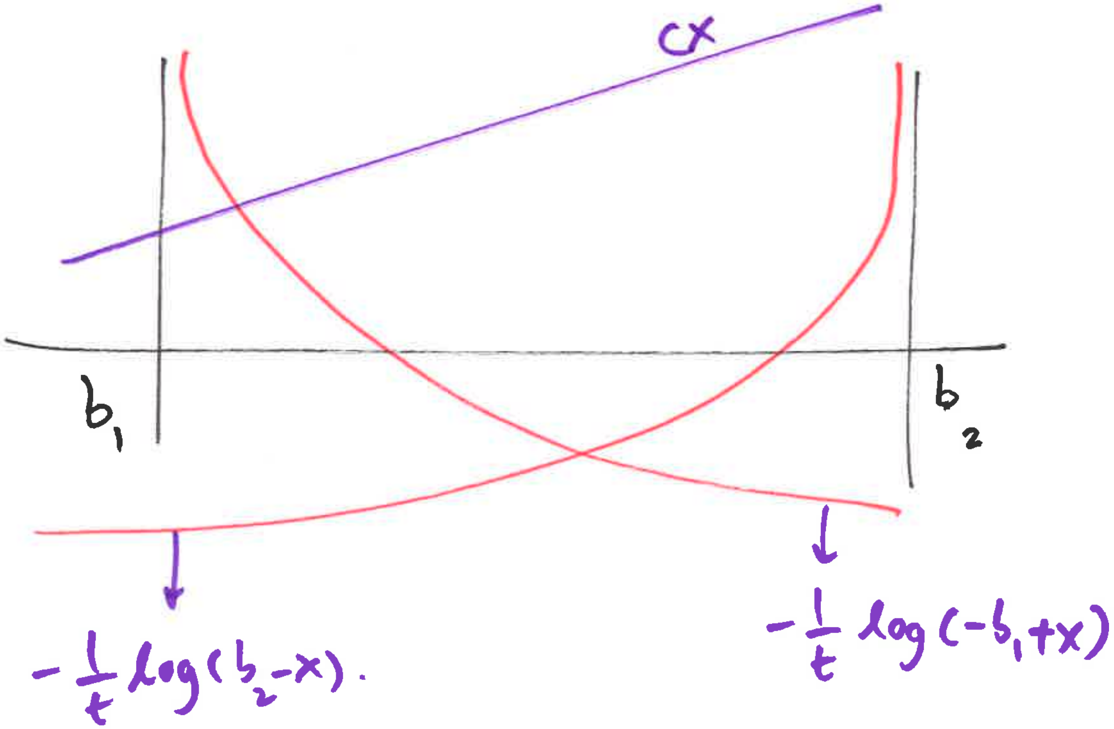
24-9

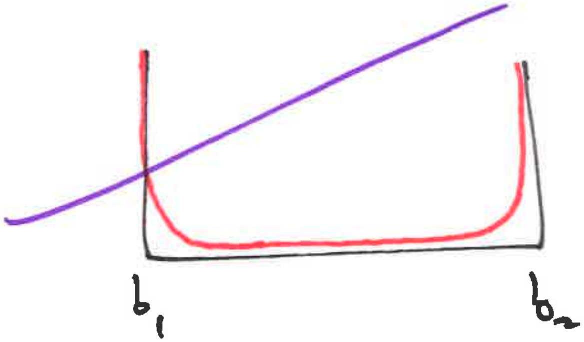
for example, suppose we want to solve

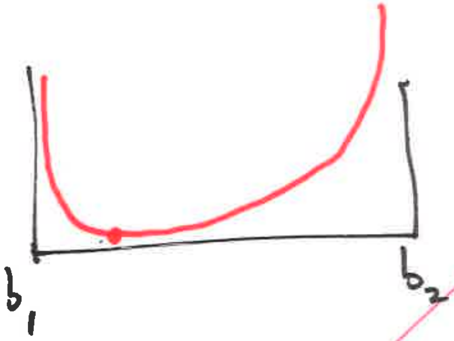
$$\begin{array}{ll} \text{minimize} & cx \\ \text{subject to} & b_1 \leq x \leq b_2 \end{array}$$

$$\downarrow$$

$$\text{minimize} \quad cx - \frac{1}{t}\log(-b_1 + x) - \frac{1}{t}\log(b_2 - x).$$







- interior point algorithm to solve LPs

1 begin with some guess in the feasible set.

2 using variational methods, find the gradient & Hessian of the objective

$$J(x) = c^T x - \frac{1}{t} \sum_{i=1}^{m} \log(b_i - a_i^T x)$$

$$\Big(\nabla J = c + \frac{1}{t} \sum_{i=1}^{m} \frac{1}{b_i - a_i^T x} a_i$$

$$\nabla^2 J = \frac{1}{t} \sum_{i=1}^{m} \left(\frac{1}{b_i - a_i^T x} \right)^2 a_i a_i^T \Big) \rightarrow$$

use $\log(1-x)$

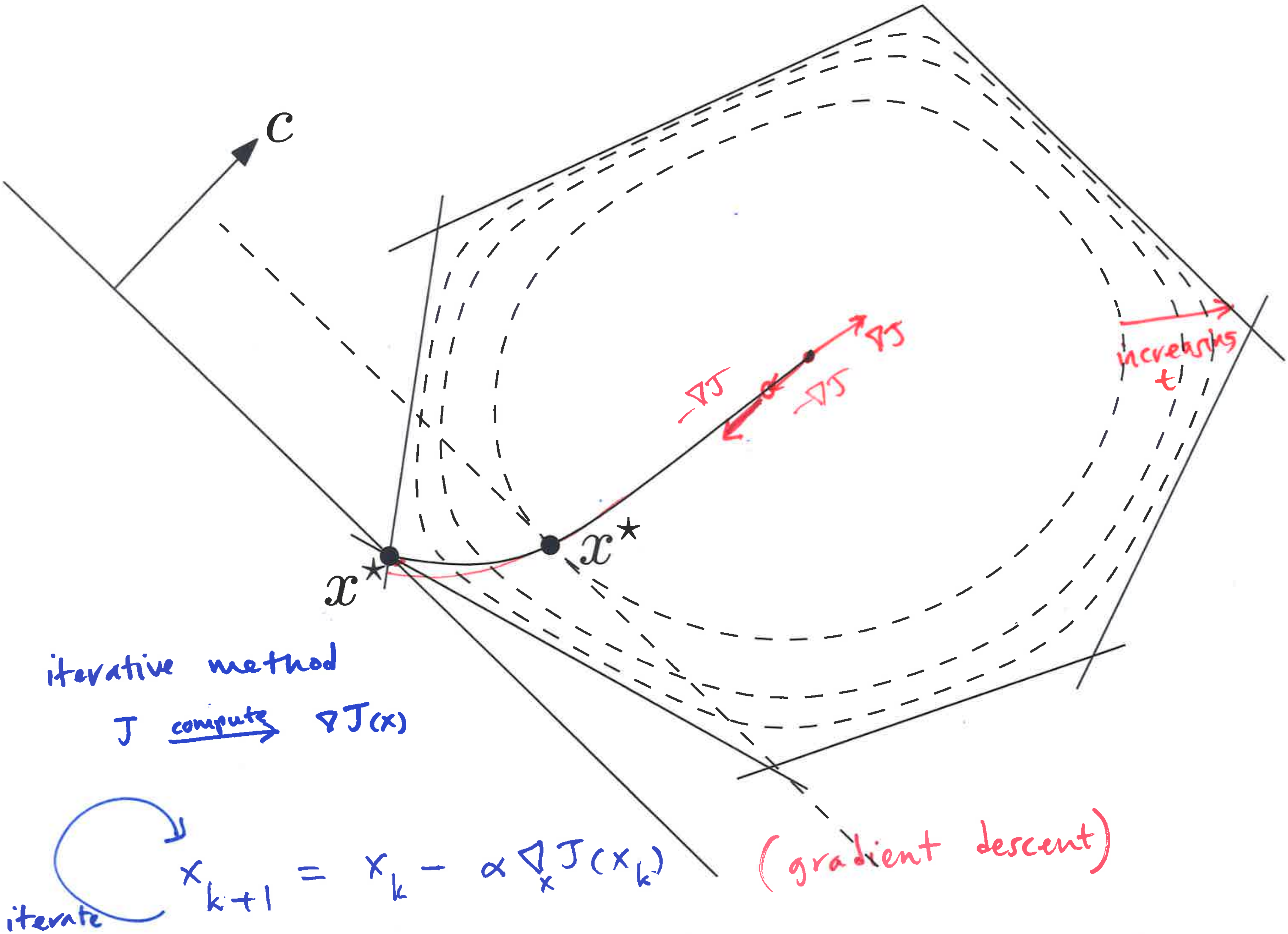
$\approx -x - \frac{1}{2}x^2 + \text{h.o.t.}$

to update the estimate of the optimal $x \rightarrow x_{k+1} = x_k - \alpha \nabla_x J(x_k)$

3 increase t.

4 go back to step 2 & repeat untill x changes very little.

24-11



iterative method

$J \xrightarrow{\text{compute}} \nabla J(x)$

$$x_{k+1} = x_k - \alpha \nabla_x J(x_k)$$

iterate

(gradient descent)

review of last lecture

- solving LPs using "$\underbrace{\text{interior point}}$" method.

because we move inside the feasible set toward the solution.

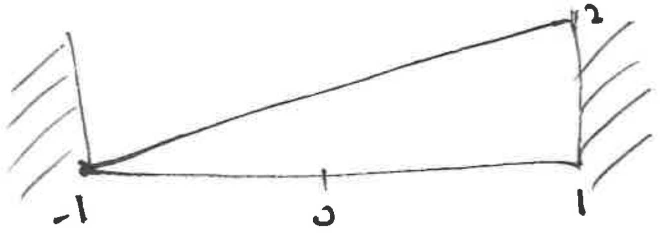
replace $\begin{array}{ll}\text{minimize} & c^Tx \\ \text{subject to} & Ax\le b\end{array}$

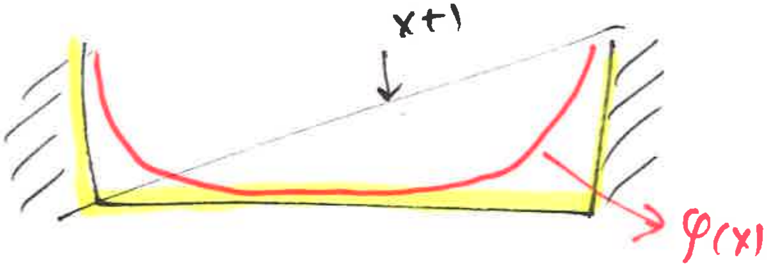
↓

with minimize $c^Tx + \frac{1}{t}\varphi(x)$ → logarithmic barrier

$$\frac{1}{t}\sum_i \log(b_i - a_i^Tx).$$

$$\begin{array}{ll}\min. & x+1 \\ \text{s.t.} & -1\le x\le 1\end{array} \longrightarrow \min.\quad x+1\overbrace{-\frac{1}{t}\log(1-x)-\frac{1}{t}\log(1+x)}^{\frac{1}{t}\varphi(x)}$$



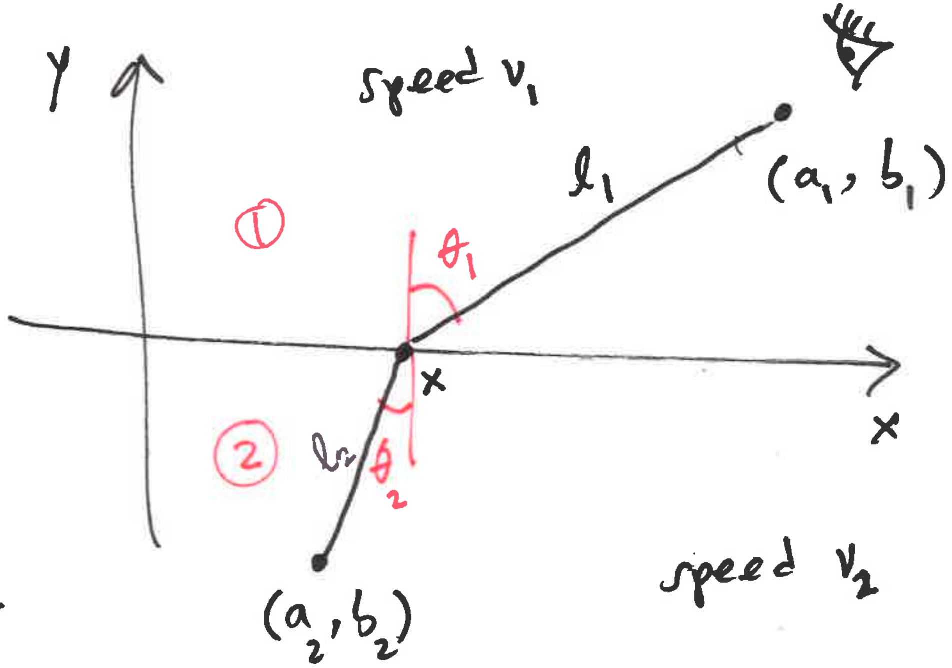


25/2

an introduction to optimization

- example; light refraction & Snell's law

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2}$$



Fermat's principle: light reaching the detector follows the path of minimum time from the source.

$$T(x) = \frac{l_1}{v_1} + \frac{l_2}{v_2} = \frac{\sqrt{(a_1-x)^2 + b_1^2}}{v_1} + \frac{\sqrt{(x-a_2)^2 + b_2^2}}{v_2}$$

15-8

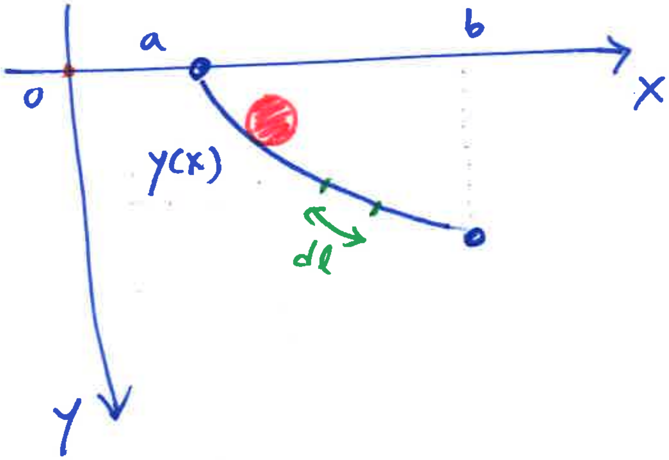
$$\frac{dT}{dx} = -\frac{1}{V_1}\frac{a_1 - x}{\sqrt{(a_1-x)^2 + b_1^2}} + \frac{1}{V_2}\frac{x - a_2}{\sqrt{(x-a_2)^2 + b_2^2}} = 0$$

$$\rightarrow \quad -\frac{1}{V_1}\sin\theta_1 + \frac{1}{V_2}\sin\theta_2 = 0$$

$$\rightarrow \quad \frac{\sin\theta_1}{\sin\theta_2} = \frac{V_1}{V_2} .$$

15-9

- example 2: (Brachistochrone) given two fixed points in the vertical plane, find the path between them such that a particle sliding without friction takes the shortest possible time to travel from one to the other.



$$T = \int dt \qquad dt = \frac{d\ell}{v}$$

$$\frac{mv^2}{2} = mgy \qquad (m=1,\ g=\tfrac{1}{2})$$

$$\rightarrow v = \sqrt{y} \quad \rightarrow \quad dt = \frac{d\ell}{\sqrt{y}}$$

$$d\ell = \sqrt{dx^2+dy^2} = \sqrt{1+\left(\frac{dy}{dx}\right)^2}\ dx = \sqrt{1+(y')^2}\ dx$$

$$\boxed{T(y(x)) = \int_a^b \frac{\sqrt{1+(y'(x))^2}}{\sqrt{y(x)}}\ dx}$$

how to determine "optimal" y(x) that minimizes T ?

$$T(y+\delta y) = \int_a^b \frac{\sqrt{1+(y'+\delta y')^2}}{\sqrt{y+\delta y}}\,dx$$

("calculus of variations")

$$\approx \int_a^b \frac{\sqrt{1+(y')^2+2y'\delta y'}}{\sqrt{y+\delta y}}$$

infinite dimensional problem!

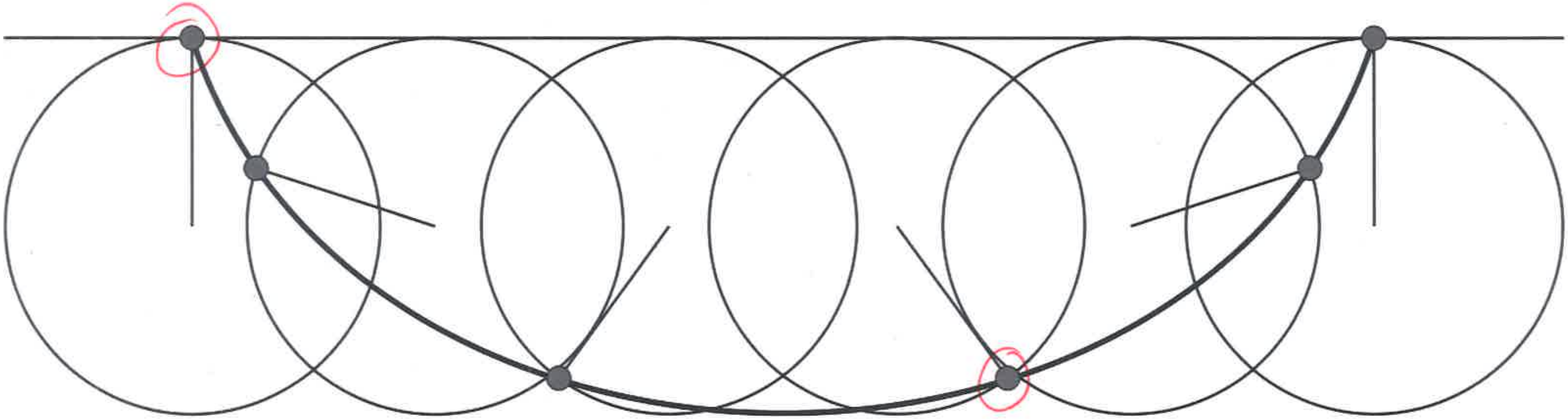
$$= \cdot \cdot \cdot \cdot$$

$$T(y+\delta y) \approx T(y) + \underbrace{\delta T(y)}_{=0}$$

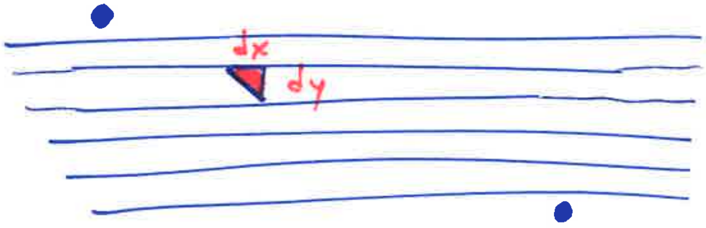
$$\int (\ldots)\,\delta y$$

$$\boxed{2yy'' + y'^2 + 1 = 0}$$

$$\boxed{\begin{aligned} x(\theta) &= a + c(\theta - \sin\theta) \\ y(\theta) &= c(1-\cos\theta) \end{aligned}}$$

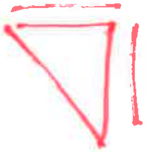


Johann Bernoulli's solution



$$\frac{\sin\theta_i}{v_i} = \text{const}$$

$$\frac{\sin\theta}{\sqrt{y}} = K$$



$$\frac{dx}{\sqrt{dx^2+dy^2}} = \sin\theta \longrightarrow \frac{dx^2}{dx^2+dy^2} = K^2 y$$

$$\left(\frac{dy}{dx}\right)^2+1 = \frac{1}{K^2 y} \longrightarrow y\left((y')^2+1\right) = \text{const}$$

$$\boxed{y'\left(2yy''+y'^2+1\right) = 0}$$

review of last lecture

- sensitivity analysis of solution of $Ax=y$ when A is invertible (square & full rank)

$$\frac{\|\delta x\|_2}{\|x\|_2} \leqslant \underbrace{\|A\|_2 \, \|A^{-1}\|_2}_{\kappa(A)} \left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\delta y\|_2}{\|y\|_2}\right)$$

$$\|A\|_2 = \sigma_1 \qquad \|A^{-1}\|_2 = \left\|\begin{bmatrix} 1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_n \end{bmatrix}\right\|_2 = \frac{1}{\sigma_n}$$

$$\kappa(A) := \|A\|_2 \, \|A^{-1}\|_2 = \frac{\sigma_{max}}{\sigma_{min}}$$

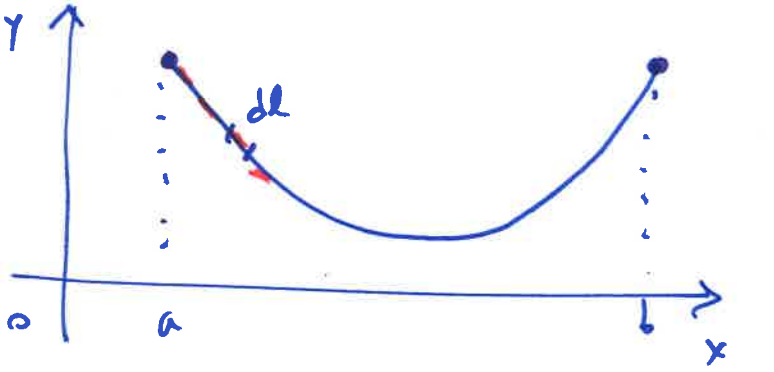
$$\Sigma = \begin{bmatrix} 0^{\uparrow} & & \\ & \ddots & \\ & & {}_{\downarrow}0 \end{bmatrix}$$

- introduction to optimization.

- example 3: catenary

$$J(y) = \int_a^b y(x)\, dl(x)$$

$$= \int_a^b y\sqrt{(y')^2+1}\; dx$$



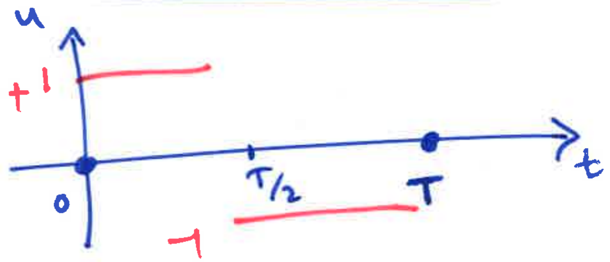
↓

$$y(x) = c \cosh \frac{x}{c}$$

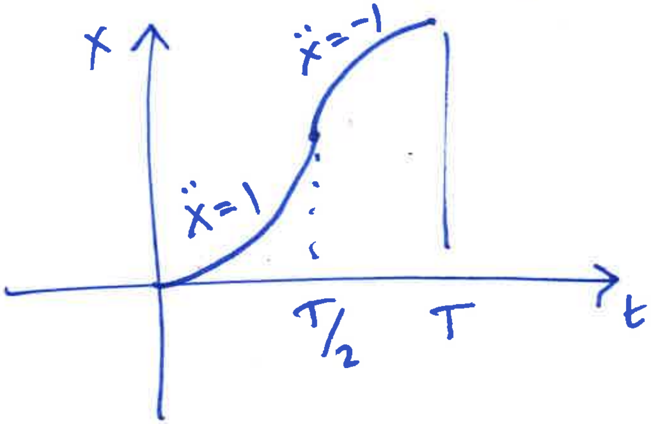
- example 4: $\ddot{x} = u$

x: car's location.

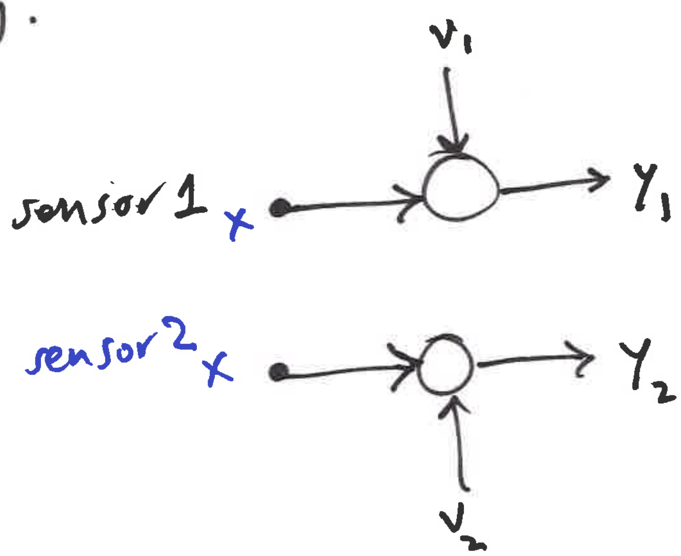
acceleration



$u \in [-1,1]$.



- example: consider a system made of two sensors, each making a measurement of a constant but unknown quantity.



$$y_1 = x + v_1$$
$$y_2 = x + v_2$$

measurement noise

v_1 & v_2 are independent of each other.

desired: design a data processing algorithm that combines y_1 & y_2 to produce an "optimal" estimate $\hat{x}$ of x.

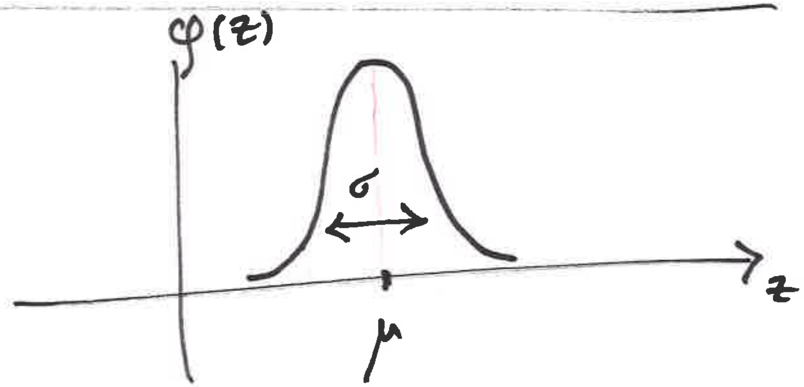
- suppose we seek a linear estimate

$$\hat{X} = k_1 Y_1 + k_2 Y_2 = \begin{bmatrix} k_1 & k_2 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

we search for "optimal" values of k_1 & k_2.

- Gaussian random variable

$$\varphi(z) \sim \frac{1}{\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$



mean $\quad \mu = \mathcal{E}\{z\} \approx \frac{1}{n}\sum_{i=1}^{n} z_i \quad$ most likely value of z

variance $\quad \sigma^2 = \mathcal{E}\{(z-\mu)^2\} \approx \frac{1}{n}\sum_{i=1}^{n} (z_i-\mu)^2 \quad$ the amount of randomness/uncertainty in z.

- assume v_1 & v_2 are Gaussian, with zero mean $\mathcal{E}\{v_1\}=0$, $\mathcal{E}\{v_2\}=0$, and variance $\mathcal{E}\{v_1^2\}=\sigma_1^2$, $\mathcal{E}\{v_2^2\}=\sigma_2^2$.

v_1 & v_2 are Gaussian

$$\Big\downarrow \begin{aligned} y_1 &= x+v_1 \\ y_2 &= x+v_2 \end{aligned}$$

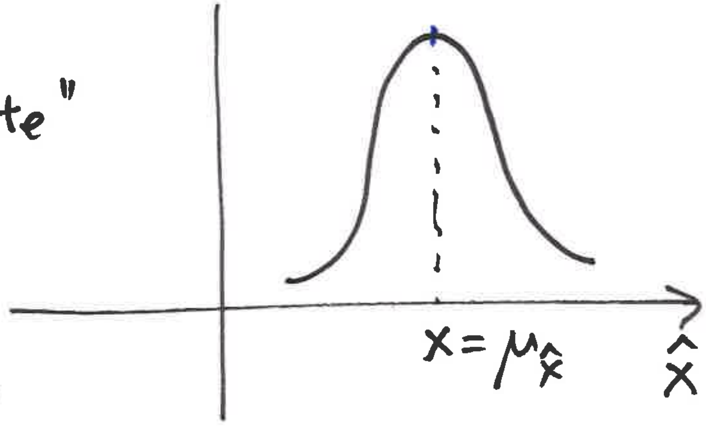
y_1 & y_2 are Gaussian

$$\Big\downarrow$$

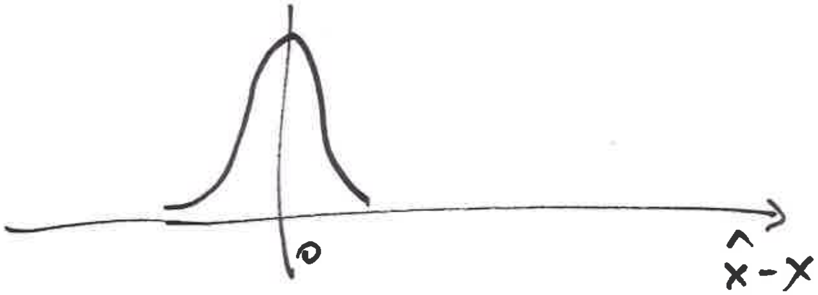
$\hat{x} = k_1 y_1 + k_2 y_2$ is Gaussian.

- our objective is to find $\hat{x}$ such that

① $\mu_{\hat{x}} = \mathcal{E}\{\hat{x}\} = x$ "unbiased estimate"



② $\sigma_{\hat{x}}^2 = \mathcal{E}\{(\underbrace{x-\hat{x}}_{e})^2\}$ is minimized.



$$\hat{x} = k_1 y_1 + k_2 y_2$$

$$\begin{aligned} \mathcal{E}\{\hat{x}\} &= \mathcal{E}\{k_1 y_1 + k_2 y_2\} \\ &= k_1 \mathcal{E}\{y_1\} + k_2 \mathcal{E}\{y_2\} \\ &= k_1 \mathcal{E}\{x+v_1\} + k_2 \mathcal{E}\{x+v_2\}. \end{aligned}$$

$$= k_1 (x + \mathcal{E}\{v_1\}) + k_2 (x + \mathcal{E}\{v_2\})$$

$$= (k_1 + k_2)\, x + k_1 \cancelto{0}{\mathcal{E}\{v_1\}} + k_2 \cancelto{0}{\mathcal{E}\{v_2\}}.$$

$$= (k_1 + k_2)\, x$$

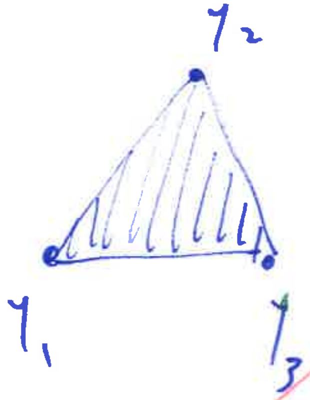
$$\mathcal{E}\{\hat{x}\} = (k_1 + k_2)\, x$$

from objective ① (unbiased estimate $\mathcal{E}\{\hat{x}\} = x$)

$$(k_1 + k_2)\, x = x \rightarrow \boxed{k_1 + k_2 = 1}$$

$$\hat{x} = k_1 Y_1 + k_2 Y_2$$

$$= k_1 Y_1 + (1 - k_1)\, Y_2$$



define $e = x - \hat{x}$. then k_1 is "optimal" if $\mathcal{E}\{e^2\}$ is minimum.

from objective ②

$$
\begin{aligned}
e &= x - \hat{x} \\
&= x - \left[k_1(x+v_1) + (1-k_1)(x+v_2) \right] \\
&= x - \left[\cancel{k_1 x} + k_1 v_1 + x - \cancel{k_1 x} - k_1 v_2 \right] \\
&= -k_1 v_1 - (1-k_1) v_2
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{E}\{e^2\} &= \mathcal{E}\{ k_1^2 v_1^2 + (1-k_1)^2 v_2^2 + 2k_1(1-k_1) v_1 v_2 \} \\
&= k_1^2 \underbrace{\mathcal{E}\{v_1^2\}}_{\sigma_1^2} + (1-k_1)^2 \underbrace{\mathcal{E}\{v_2^2\}}_{\sigma_2^2} + 2k_1(1-k_1) \underbrace{\mathcal{E}\{v_1 v_2\}}_{?}
\end{aligned}
$$

if v_1 & v_2 are independent, then

$$\mathcal{E}\{v_1 v_2\} = \cancelto{0}{\mathcal{E}\{v_1\}} \; \cancelto{0}{\mathcal{E}\{v_2\}} . = 0$$

$$\mathcal{E}\{e^2\} = k_1^2 \sigma_1^2 + (1-k_1)^2 \sigma_2^2$$

$$=: J$$

$$\frac{\partial J}{\partial k_1} = 2k_1 \sigma_1^2 + 2(1-k_1)(-1)\sigma_2^2 = 0 \longrightarrow \boxed{k_1 = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}}$$

$$k_2 = 1-k_1 = 1 - \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} \longrightarrow \boxed{k_2 = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}}$$

16-9

$$\boxed{\hat{x} = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\, y_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\, y_2}$$

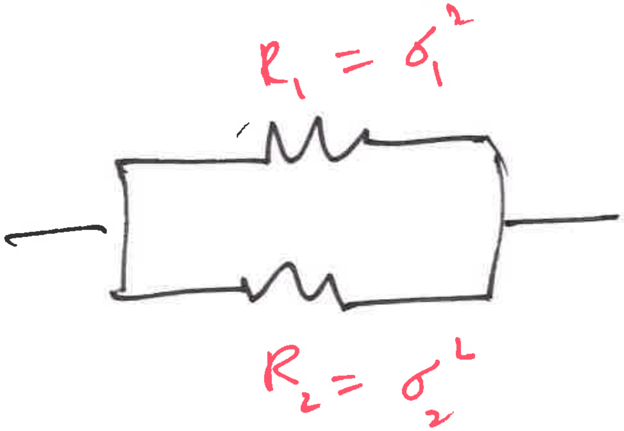
- if sensor 1 much better than sensor 2 → $\sigma_1 \approx 0$ $\Longrightarrow$ $\hat{x} \approx y_1$
- if the sensors are equally / bad → $\sigma_1 \approx \sigma_2$

$$\Longrightarrow \quad \hat{x} = \frac{y_1 + y_2}{2}.$$

$$\mathcal{E}\{e^2\}_{\text{optimal}} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} = \begin{cases} \dfrac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\,\sigma_1^2 \leqslant \sigma_1^2 \\[2ex] \dfrac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\,\sigma_2^2 \leqslant \sigma_2^2 \end{cases}$$

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{R_1 + R_2}{R_1 R_2}$$

$$R = \frac{R_1 R_2}{R_1 + R_2}$$



review of last lecture

- linear estimation problem

$$y_1 = x + v_1 \qquad \hat{x} = Ky = \begin{bmatrix} k_1 & k_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \underset{K.}{\text{minimize}} \ \mathcal{E}\{(x-\hat{x})^2\}$$

$$y_2 = x + v_2$$

$$\hat{x} = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} y_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} y_2$$

$$\mathcal{E}\{(x-\hat{x})^2\} = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2+\sigma_2^2} \leqslant \min\{\sigma_1^2, \sigma_2^2\}.$$

- (necessary) conditions for optimality $\begin{cases} \nabla J(x^*) = 0 \\ \nabla^2 J(x^*) > 0 \end{cases}$