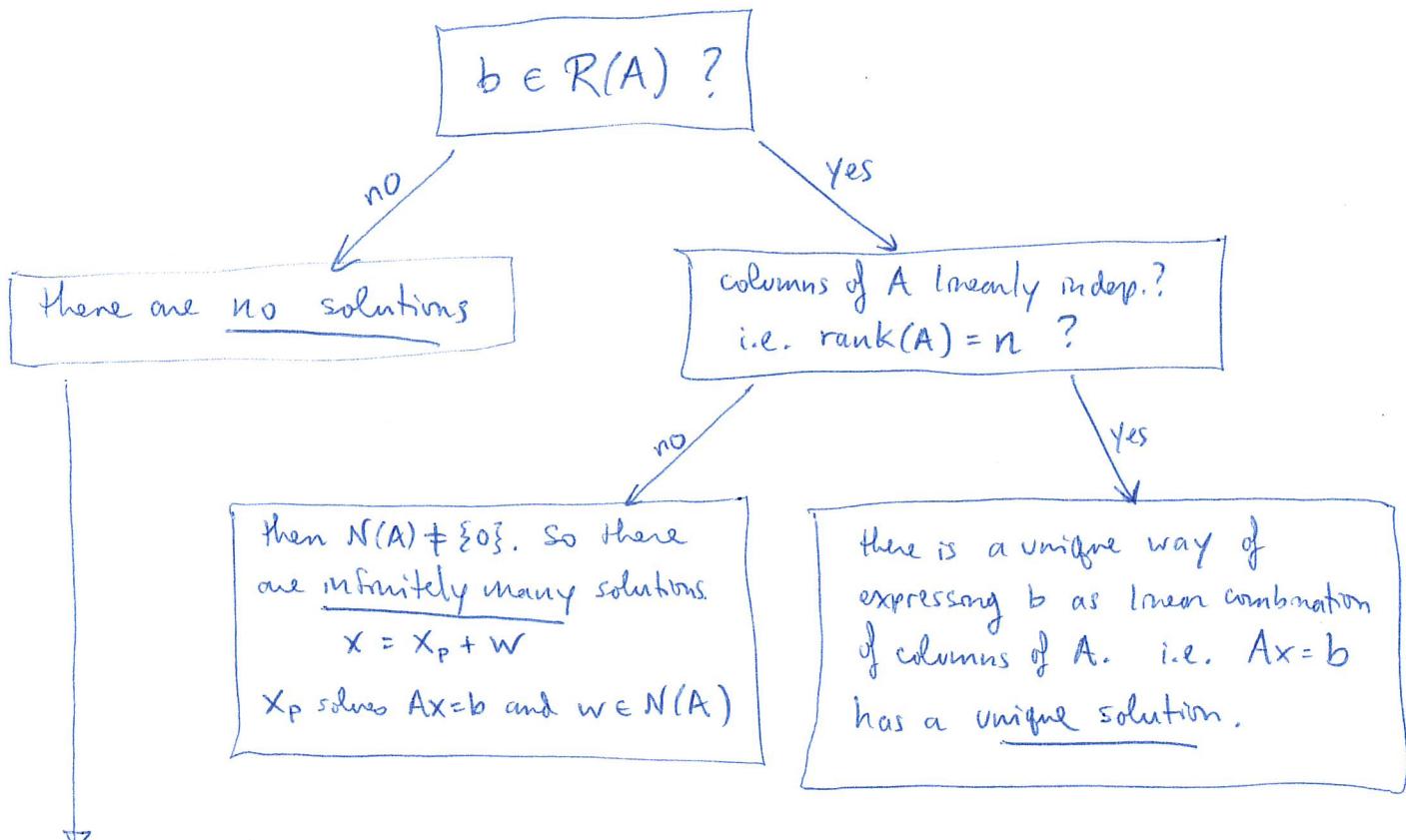


ECE 532 - Lecture 5 - intro to least squares

1

Review of linear equations: $Ax = b$, $A \in \mathbb{R}^{m \times n}$.



what can we do in this case? use least squares.

★ Typical setup: we'd like to solve $Ax = b$, $A \in \mathbb{R}^{m \times n}$ where $m >> n$ (m much larger than n). Typically, columns of A will be full rank and $b \notin R(A)$; so no solutions.

define the residual $r = Ax - b$.

(2)

* if we can't make $r=0$ ($Ax=b$ has no solutions), instead try to make $\|r\|$ as small as possible.

- often written as:

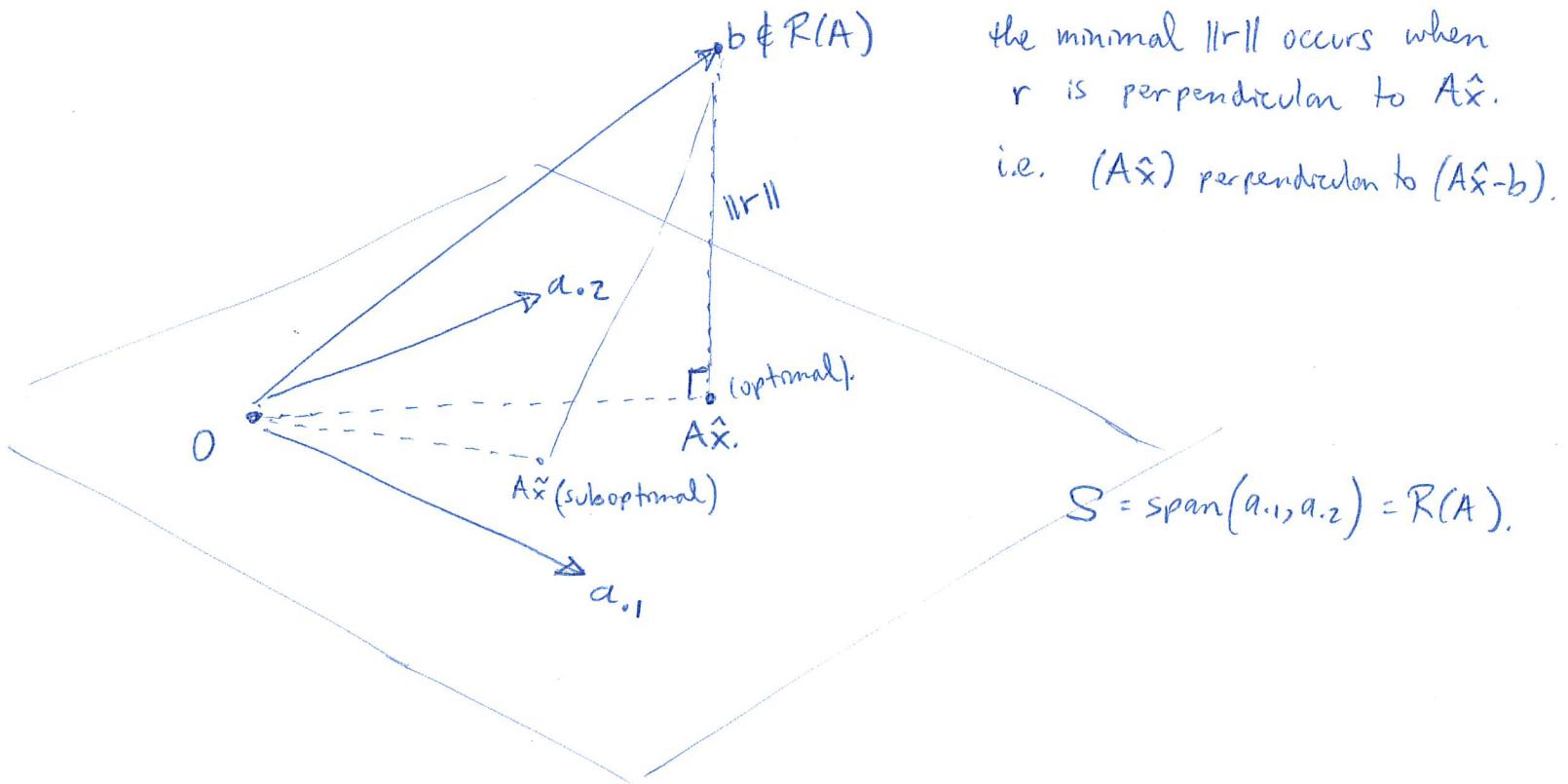
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \|Ax - b\|^2$$

this is not necessary, but it's done by convention so that the square-root cancels.
i.e. $\min x^2 + y^2$ instead of $\min \sqrt{x^2 + y^2}$

geometric intuition. Take the case $m=3, n=2$.

$$Ax = b \Rightarrow x_1 a_{.1} + x_2 a_{.2} = b.$$

$$\Rightarrow x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$



(3)

optimal solution to least squares problem:

$$\hat{x} = \arg \min_x \|Ax - b\|^2$$

happens when $\hat{r} = A\hat{x} - b$ is perpendicular to ~~any~~ $R(A)$.
(perpendicular to every vector in $R(A)$)

Aside: the set $S^\perp = \{w \in \mathbb{R}^n \mid w^T x = 0 \text{ for all } x \in S\}$
"S perp"

is a subspace! Why? Use the definition:

- $0^T x = 0$ (zero is in S^\perp),
- if $w_1^T x = 0$ for all $x \in S$ and $w_2^T x = 0$ for all $x \in S$,
then $(w_1 + w_2)^T x = \underbrace{w_1^T x}_0 + \underbrace{w_2^T x}_0 = 0$ whenever $x \in S$,
- if $w^T x = 0$ for all $x \in S$, then $(\alpha w)^T x = \underbrace{\alpha(w^T x)}_0 = 0$.

so \hat{x} is optimal if $\hat{r} \in R(A)^\perp$

$$\Rightarrow w^T (A\hat{x} - b) = 0 \quad \text{for every } w \in R(A). \text{ i.e. whenever } w = Ax \text{ for some } x.$$

$$\Rightarrow (Ax)^T (A\hat{x} - b) = 0 \quad \text{for all } x.$$

$$\Rightarrow x^T (A^T A \hat{x} - A^T b) = 0 \quad \text{for all } x.$$

$$\Rightarrow \boxed{A^T A \hat{x} = A^T b}$$

These are called the normal equations

(4)

so if $\hat{x} = \arg \min_x \|Ax - b\|^2$

then $\underbrace{A^T A \hat{x}}_{n \times n} = \underbrace{A^T b}_{n \times 1}$ $\left\{ \begin{array}{l} \text{this is an } n \times n \\ \text{system of linear} \\ \text{equations.} \end{array} \right.$

We proved that if $\hat{x} = \arg \min_x \|Ax - b\|^2$, then $A^T A \hat{x} = A^T b$.

what about the converse? what if we find some \hat{x} such that $A^T A \hat{x} = A^T b$... does it follow that \hat{x} minimizes $\|Ax - b\|^2$?

[Yes, but we didn't prove it yet! To see why, just replace our normal equations by something that's always true, like $\hat{x} = \hat{x}$].

Proof of the converse: Suppose $A^T A \hat{x} = A^T b$. Let's compute the residual for some other candidate point x .

$$\begin{aligned} r &= Ax - b \\ &= \underbrace{A\hat{x} - b}_{\hat{r}} + A(x - \hat{x}) \\ &= \underbrace{\hat{r}}_{\cdot} + A(x - \hat{x}) \end{aligned}$$

Note: $A^T \hat{r} = \hat{r}^T A = 0$
 (These are precisely the
 normal equations!)

$$\begin{aligned} \|r\|^2 &= \|\hat{r} + A(x - \hat{x})\|^2 \\ &= (\hat{r} + A(x - \hat{x}))^T (\hat{r} + A(x - \hat{x})) \\ &= \hat{r}^T \hat{r} + \underbrace{\hat{r}^T A}_{0} (x - \hat{x}) + (x - \hat{x})^T \underbrace{A^T \hat{r}}_{0} + (x - \hat{x})^T A^T A (x - \hat{x}) \\ \|r\|^2 &= \|\hat{r}\|^2 + \underbrace{\|A(x - \hat{x})\|^2}_{\geq 0} \end{aligned}$$

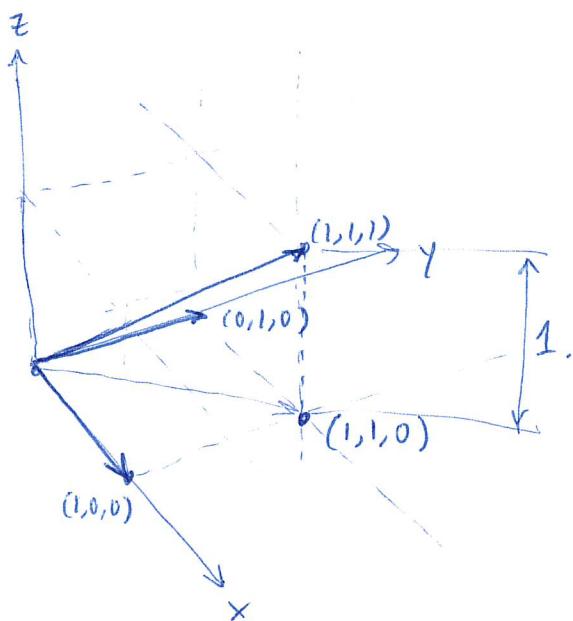
Pythagorean theorem!

so $\|r\|^2$ can't be any smaller than $\|\hat{r}\|^2$, i.e. \hat{x} is optimal ■

(5)

Examplefind x that minimizes

$$\left\| \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\|^2$$

1) draw a picture:so closest point is $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ which is $\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

minimum residual has length of 1.

2) solve normal equations:

$$A^T A \hat{x} = A^T b.$$

$$\Rightarrow \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{\text{Matrix } A^T A} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

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

$$\Rightarrow \begin{pmatrix} x = 1 \\ y = 1 \end{pmatrix}$$

(6)

So far we have seen that:

$$\left\{ \begin{array}{l} \hat{x} \text{ is a solution to} \\ \min_x \|Ax - b\|^2 \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} \hat{x} \text{ is a solution} \\ \text{to } A^T A \hat{x} = A^T b \end{array} \right\}.$$

* fact: $A^T A \hat{x} = A^T b$ always has a solution.

so $A^T b \in R(A^T A)$ no matter what A looks like.

We will prove this later on.

* fact: $A^T A$ is invertible (i.e. $A^T A \hat{x} = A^T b$ has a unique solution) if and only if A has linearly independent columns.
(i.e. $\text{rank}(A) = n$, or $N(A) = \{0\}$).

Proof: observe that if $Ax = 0$ for some x , then $A^T A x = 0$ also.

Similarly, if $A^T A x = 0$, then $x^T A^T A x = 0$, which is the same as $\|Ax\|^2 = 0$, which implies $Ax = 0$ (property of norm!).

so $Ax = 0 \Leftrightarrow A^T A x = 0$. This implies in particular that $N(A) = N(A^T A)$, so if $N(A) = \{0\}$, then $N(A^T A) = \{0\}$ and vice versa. □

Conclusion: if A has lin. indep. columns, solution to LS problem $\min \|Ax - b\|^2$ is $\hat{x} = (A^T A)^{-1} A^T b$.

otherwise, $\hat{x} = \hat{x}_1 + w$ \nwarrow any element in $N(A)$.
 \uparrow any solution to $A^T A \hat{x} = A^T b$

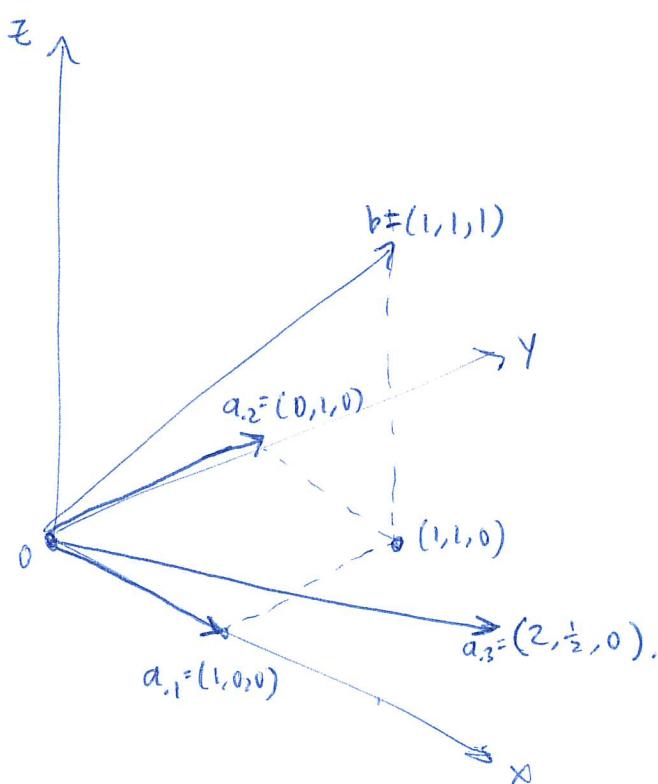
7

Makes sense that a LS problem could have multiple solutions. Even if $Ax = b$ has no solutions, if $w \in N(A)$

$$\text{then } \|A(\hat{x}+w) - b\|^2 = \|A\hat{x} + Aw - b\|^2 = \|A\hat{x} - b\|^2.$$

so there may be multiple \hat{x} that solve $\min \|Ax - b\|^2$, and all achieve the same residual.

Our previous example:



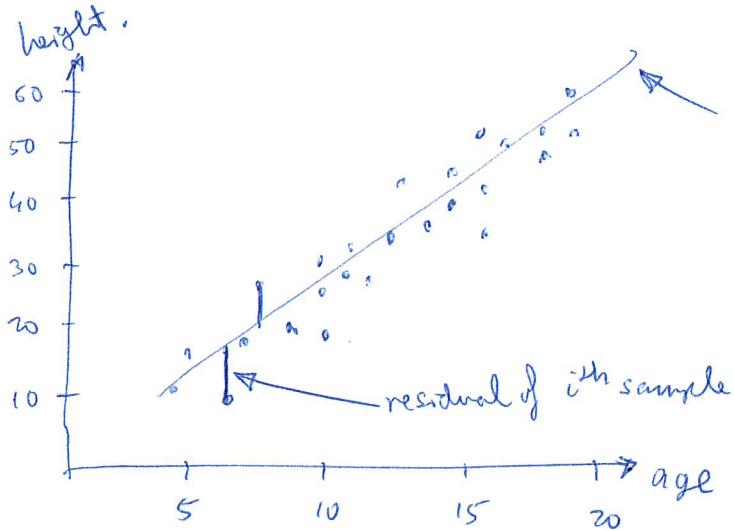
add a new column: $\begin{pmatrix} z \\ y \\ 0 \end{pmatrix}$.

Now there are multiple ways of producing $\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$, which is still the closest point to $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ because we haven't changed $R(A)$. So this LS problem is degenerate. (has many solutions).

Example : Linear regression.

Suppose we have data from an experiment to characterize the age of a tree as a function of its height (for a particular species). Collect data: (x_i, y_i)

$$\begin{cases} x_i = \text{age} & i=1,2,\dots,N \\ y_i = \text{height} \end{cases}$$



suspect model of the form
 $y_i = P x_i + q$
 can explain the data well.

1) Write equations:

$$\underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}}_A \underbrace{\begin{bmatrix} P \\ q \end{bmatrix}}_x = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix}}_b$$

Note: A will always have full rank unless all the x_i are identical...

2) solve least squares problem, $\min_x \|Ax - b\|^2$,

which gives us $\hat{x} = \begin{bmatrix} \hat{P} \\ \hat{q} \end{bmatrix}$.

solution to our LS problem.

3) predict future heights based on age: $y_{\text{est}} = \hat{P} \hat{x}_{\text{new}} + \hat{q}$

estimated ↑ Large of a new
height ~~height~~ tree.