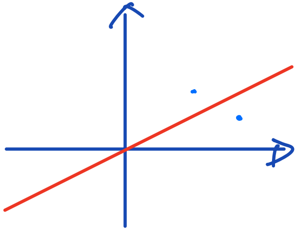
Least-squares

Stephen Boyd and Sanjay Lall
Revision by Amirhossein Afsharrad

EE263
Stanford University

Overdetermined linear equations



consider $y = Ax$ where $A \in \mathbb{R}^{m \times n}$ is (strictly) skinny, *i.e.*, $m > n$

$\operatorname{rank}(A) \leq n < m$

- called *overdetermined* set of linear equations (more equations than unknowns)
- for most y, cannot solve for x $\rightsquigarrow$ for most y, $y \notin \operatorname{range}(A)$

one approach to *approximately* solve $y = Ax$:

- define *residual* or error $r = Ax - y$
- find $x = x_{\text{ls}}$ that minimizes $||r||$

x_{ls} called *least-squares* (approximate) solution of $y = Ax$

Geometric interpretation

LS Problem:

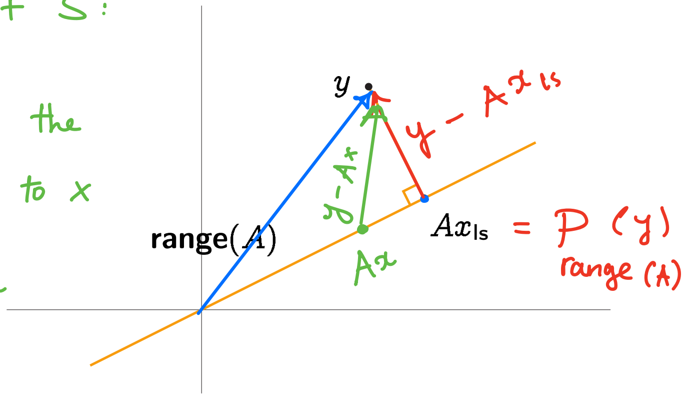
Given $y \in \mathbb{R}^m$, find $x \in \mathbb{R}^n$ to minimize $||Ax - y||$

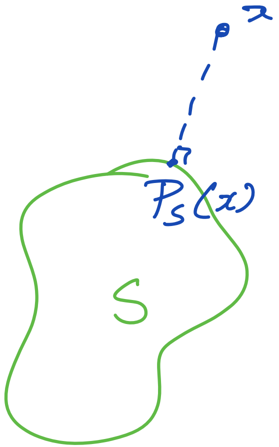
Ax_{ls} is point in **range**(A) closest to y (Ax_{ls} is *projection* of y onto **range**(A))

projection onto a set S:

$y = P_S(x)$ if y is the closest element of S to x

if $x \in S \Rightarrow P_S(x) = x$





Least-squares (approximate) solution

$\|r\| = \|Ax - y\|$

$\|r\|^2 = \|Ax - y\|^2$

$= (Ax-y)^T (Ax - y)$

- assume A is full rank, skinny
- to find $x_{\rm ls}$, we'll minimize norm of residual squared,

$$\|r\|^2 = x^T \underbrace{A^T A}_{P} x - 2y^T Ax + y^T y$$

- set gradient w.r.t. x to zero:

$$\nabla_x \|r\|^2 = 2A^T Ax - 2A^T y = 0$$

- yields the *normal equations*: $A^T Ax = A^T y$
- assumptions imply $A^T A$ invertible, so we have

$$\boxed{x_{\rm ls} = \underbrace{(A^T A)^{-1} A^T}_{\text{famous left inverse}} y}$$

... a very famous formula

$ax^2 \xrightarrow{d/dx} 2ax$

$bx \xrightarrow{d/dx} b$

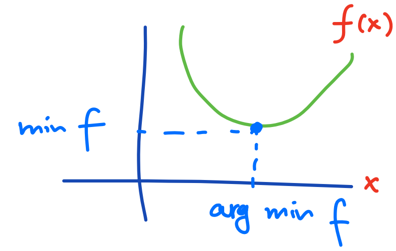
$\nabla_x\, x^T P x = 2Px$

$(P = P^T)$

$\nabla_x\, q^T x = q$

if A is not full rank (A^TA not invertible):

$$x^* = \arg\min_x \| Ax - y \|$$



choose $z \in \text{null}(A) \longrightarrow \tilde{x} = x^* + z$

is another solution

$$\| A\tilde{x} - y \| = \| Ax^* + \underbrace{Az}_{0} - y \|$$

Least-squares (approximate) solution

$$x_{ls} = (A^TA)^{-1}A^Ty$$

- x_{ls} is linear function of y
- $x_{ls} = A^{-1}y$ if A is square
- x_{ls} solves $y = Ax_{ls}$ if $y \in$ **range**(A)

$$y = Az \rightarrow x_{ls} = \underbrace{(A^TA)^{-1}}\underbrace{A^TA}z = z$$

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

$$= A^{-1}\underbrace{A^{-T}A^T}_{I}$$

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

$$= A^{-1}$$

Least-squares (approximate) solution

for A skinny and full rank, the *pseudo-inverse* of A is

$$\boxed{A^\dagger = (A^\mathsf{T} A)^{-1} A^\mathsf{T}}$$

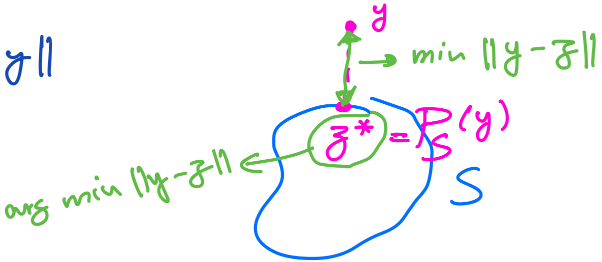
- for A skinny and full rank, $A^\dagger$ is a *left inverse* of A

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

- if A is not skinny and full rank then $A^\dagger$ has a different definition

Projection on $\textbf{range}(A)$

$$P_S(y) = \arg\min_{z \in S} \|z - y\|$$



$Ax_{\rm ls}$ is (by definition) the point in $\textbf{range}(A)$ that is closest to y, $i.e.$, it is the *projection* of y onto $\textbf{range}(A)$

$$Ax_{\rm ls} = \mathcal{P}_{\textbf{range}(A)}(y)$$

$$By$$
$$\downarrow$$
$$A(A^TA)^{-1}A^T$$

- the projection function $\mathcal{P}_{\textbf{range}(A)}$ is linear, and given by

$$\mathcal{P}_{\textbf{range}(A)}(y) = Ax_{\rm ls} = A(A^TA)^{-1}A^Ty$$

- $A(A^TA)^{-1}A^T$ is called the *projection matrix* (associated with $\textbf{range}(A)$)

P : proj matrix if $P = P^T$, $P^2 = P$ $\left\{ (I-P)^2 = \underbrace{P^2}_{P} + I - 2P = I - P \right.$

$$P^2 = A\cancel{(A^TA)^{-1}}\cancel{A^TA}(A^TA)^{-1}A^T = P$$

Orthogonality principle

optimal residual

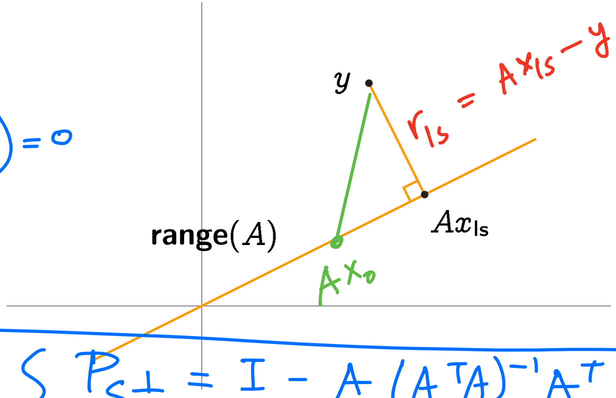
$$r = Ax_{\text{ls}} - y = \overbrace{(\underbrace{A(A^\top A)^{-1}A^\top}_{P} - I)}^{P - I}y$$

is orthogonal to **range**(A):

$$\underbrace{\langle r, Az\rangle}_{r^\top Az} = \underbrace{y^\top(A(A^\top A)^{-1}A^\top - I)^\top}_{r^\top}\underbrace{Az} = 0$$

for all $z \in \mathbb{R}^n$

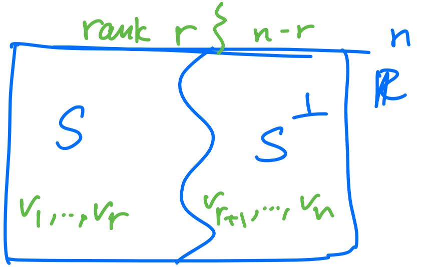
$$y^\top\left(A\,\cancel{(A^\top A)}^{-1}\cancel{(A^\top A)}z - Az\right) = 0$$



$S = \text{range}(A)$

$P_S(y) = A(A^TA)^{-1}A^T$

$P_{S^\perp} = I - A(A^TA)^{-1}A^T$



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

$$y = \underbrace{\alpha_1 v_1 + \cdots + \alpha_r v_r}_{y_S \,\in S} + \underbrace{\alpha_{r+1} v_{r+1} + \cdots + \alpha_n v_n}_{y_{S^\perp} \,\in S^\perp}$$

$y = y_S + y_{S^\perp}$

$y_S = P_S(y)$

$y_{S^\perp} = P_{S^\perp}(y) = y - P_S(y)$

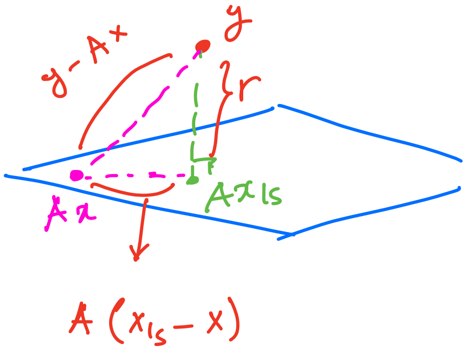
Completion of squares

$a^T b = 0 \rightarrow \|a-b\|^2 = \|a\|^2 + \|b\|^2 - \underbrace{\cancel{2a^Tb}}_{0}$

$Ax - y = Ax - Ax_{ls} + Ax_{ls} - y$

since $r = \overbrace{Ax_{\mathsf{ls}} - y}^{r} \perp \overbrace{A(x - x_{\mathsf{ls}})}^{Az}$ for any x, we have

$$\begin{aligned} \|Ax - y\|^2 &= \|\overbrace{(Ax_{\mathsf{ls}} - y)}^{r} \overset{\perp}{+} \overbrace{A(x - x_{\mathsf{ls}})}^{Az}\|^2 \\ &= \underbrace{\|Ax_{\mathsf{ls}} - y\|^2}_{\text{constant wrt } x} + \|A(x - x_{\mathsf{ls}})\|^2 \end{aligned}$$



this shows that for $x \neq x_{\mathsf{ls}}$, $\|Ax - y\| > \|Ax_{\mathsf{ls}} - y\|$

Least-squares estimation

many applications in inversion, estimation, and reconstruction problems have form

$$\boxed{y = Ax + v}$$

could be rounding errors

- x is what we want to estimate or reconstruct
- y is our sensor measurement(s)
- v is an unknown *noise* or *measurement error* (assumed small)
- ith row of A characterizes ith sensor

Least-squares estimation

least-squares estimation: choose as estimate $\hat{x}$ that minimizes

$$||A\hat{x} - y||$$

$i.e.$, deviation between

- what we actually observed (y), and
- what we would observe if $x = \hat{x}$, and there were no noise ($v = 0$)

least-squares estimate is just $\hat{x} = (A^\mathsf{T} A)^{-1} A^\mathsf{T} y$

BLUE property

$B = \begin{bmatrix} 1000 & -2000 \\ \cdot\cdot & \end{bmatrix}$ $B = \begin{bmatrix} 1 & 1 & \cdots \\ -1 & 0.5 & \end{bmatrix}$

suppose A full rank, skinny, and we have linear measurement with noise

$$y = Ax + v$$

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

consider a *linear estimator* of form $\hat{x} = By$

- B is called *unbiased* if $\hat{x} = x$ whenever $v = 0$
  - no estimation error when there is no noise
  - equivalent to left inverse property $BA = I$
- estimation error of unbiased linear estimator is

$$x - \hat{x} = \underbrace{x - B(Ax}_{0} + v) = \boxed{-Bv}$$

- so we'd like B 'small' and $BA = I$

$\hat{x} = By = B(Ax + \cancel{v}) = BAx = x$, $v = 0$

$BA = I$

$\Downarrow$

B left inverse of A

BLUE property

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 4 \times 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

if sensors 1, 2 are perfect ?

fact: $A^\dagger = (A^\mathsf{T} A)^{-1} A^\mathsf{T}$ is the *smallest* left inverse of A, in the following sense:

for any B with $BA = I$, we have

$$\sum_{i,j} B_{ij}^2 \geq \sum_{i,j} A_{ij}^{\dagger 2}$$

$$y = \begin{bmatrix} U_{2\times 2} \\ V_{2\times 2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$B = \begin{bmatrix} U^{-1} & 0 \end{bmatrix}$$

$$BA = \begin{bmatrix} U^{-1} & 0 \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} = I$$

$i.e.$, least-squares provides the *best linear unbiased estimator* (BLUE)