Today: Linear Least Squares

Start: NLA warm-up

$$\|A\|_2 \triangleq \sup_{\|x\|_2 = 1} \|Ax\|_2$$

$$\|x\|_2 = \sqrt{\sum x_i^2}$$

SVD:

$$U^T U = V^T V = I$$

$$\boxed{A} = \boxed{U}\,\boxed{\Sigma}\,\boxed{V^T} \qquad \sigma_i \geq 0$$

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$$

Claim: $\|A\|_2 = \sigma_1$

$$\|A\|^2 = \sup_{\|x\|^2 = 1} \|Ax\|_2^2$$

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

$$\|U\Sigma V^T x\|_2^2 = x^T V \Sigma \cancel{U^T U} \Sigma V^T x$$

$$x = Vy$$

$$\|Vy\|^2 = y^T \overset{I}{\cancel{V^T V}} y = y^T y = \|y\|^2$$

$$\sup_{\|Vy\|^2 = 1} \|\Sigma \overset{I}{\cancel{V^T V}} y\|^2$$

$$\|\Sigma y\|^2 = \sum \sigma_i^2 y_i^2$$

$$y_1 = 1, \quad y_{2:n} = 0 \Rightarrow \sigma_1^2$$

Linear LS

Model:

$$\varepsilon + a^T x + z = b$$

$$A = \begin{bmatrix} a_1^T & 1 \\ \vdots & \vdots \\ a_m^T & 1 \end{bmatrix}$$

$$\hat{x} = \min_x \|b - Ax\|_2^2$$

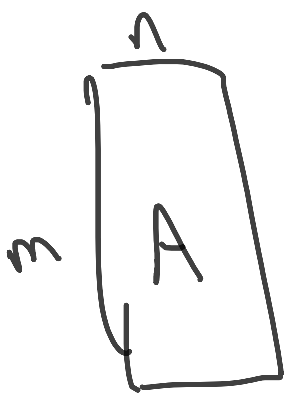
if $\varepsilon \sim N(0,1)$ iid, then

$\hat{x}$ is MLE

$r = A\hat{x} - b$

$r^T A = 0$ Why?

Assume A has <u>full rank</u>



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

$$f(x) = \|Ax - b\|^2$$

$$= x^T A^T A x - 2x^T A^T b + b^T b$$

$$\nabla f(x) = 2A^T A x - 2A^T b$$

$$= 0 \iff \underbrace{A^T(Ax - b)}_{r} = 0$$

$$\overbrace{A^T A}^{H} x = A^T b$$

$$\boxed{A^T A \hat{x} = A^T b \quad \text{normal equations}}$$

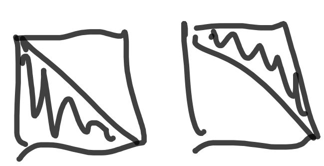
Matrix factorizations for LLS

$$\hat{x} = \underbrace{(A^T A)^{-1} A^T}_{\text{Moore-Penrose pseudoinverse}} b$$

Cholesky

$$(A^T A)\hat{x} = A^T b$$

$$m\,\overset{n}{\boxed{A}} \qquad \underbrace{\boxed{A^T A}}_{O(mn^2)} = \overbrace{L L^T}^{O(n^3)}$$

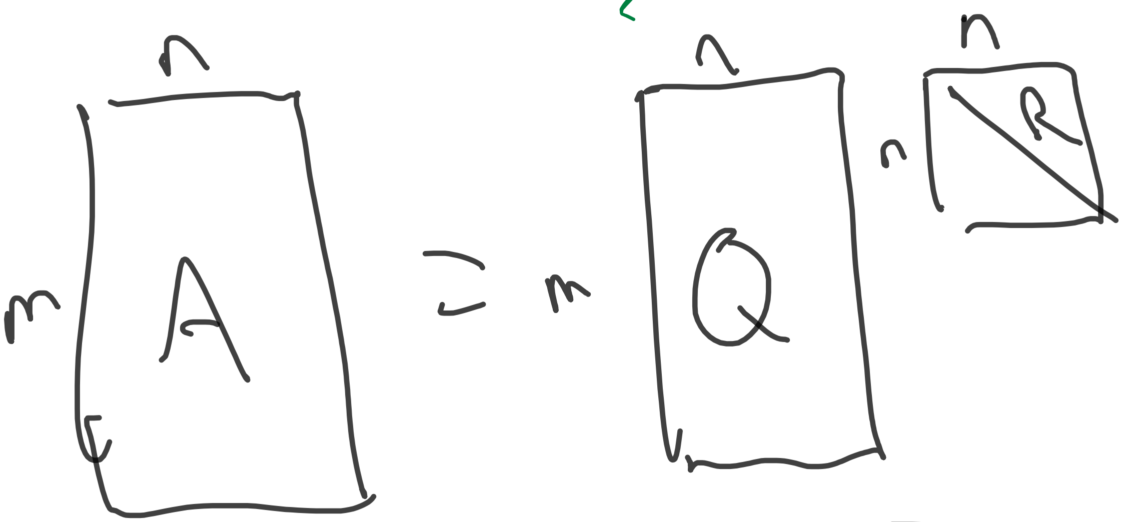


$$\underbrace{L L^T}_{O(n^2)} \hat{x} = \underbrace{A^T b}_{O(mn)}$$

$$\hat{x} = L^{-T} L^{-1} A^T b$$

QR

$O(mn^2)$



$$Q^T Q = I$$

$$A^T A = R^T Q^T Q R = R^T R$$

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

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

$$\hat{x} = R^{-1} Q^T b$$

SVD

$$\underset{m \times n}{A} = \underbrace{\underset{m \times n}{U} \; \underset{n \times n}{\Sigma} \; \underset{n \times n}{V^T}}_{O(mn^2)}$$

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$$

$$U^T U = V^T V = I$$

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

$$\cancel{V} \Sigma^{\cancel{2}} V^T \hat{x} = \cancel{V} \cancel{\Sigma} U^T b$$

$$\Sigma V^T \hat{x} = U^T b$$

$$\hat{x} = V \Sigma^{-1} U^T b$$

$O(mn^2)$ time

$$A = \begin{bmatrix} 3+\varepsilon_1 & 1 \\ \vdots & \vdots \\ 3+\varepsilon_n & 1 \end{bmatrix} \quad b = \begin{bmatrix} 6+2\varepsilon_1 \\ \vdots \\ 6+2\varepsilon_n \end{bmatrix}$$

$\varepsilon_i \sim N(0, 1e^{-7})$ Lots of high quality solutions

$$\hat{x} = [2 \;\; 0]$$

$$A\hat{x} = b$$

$$A[1 \;\; 3] \approx b$$

$$[1/3 \;\; 5] \approx b$$

Bias-variance and ill-conditioning

prediction error

$$\approx \underbrace{\text{bias error}}_{\text{model}} + \underbrace{\text{variance error}}_{\text{sensitivity}}$$

$$A = \begin{bmatrix} A_{tr} \\ A_{te} \end{bmatrix} \qquad b = \begin{bmatrix} b_{tr} \\ b_{te} \end{bmatrix}$$

$\exists\ \bar{x}$ s.t. $A\bar{x} = b + r$ $\quad r^T A = 0$

observe A_{tr}, $b_{tr} + e_{tr}$

$$\hat{x} = \min_{x} \| A_{tr} x - (b_{tr} + e_{tr}) \|^2$$

$$\|A\hat{x} - b\|_2^2$$

model residual

$$= \|A\hat{x} - A\bar{x} + r\|^2$$

$$= \|A(\hat{x} - \bar{x})\|^2 + \|r\|^2$$

$$+ \cancel{2r^T A(\hat{x} - \bar{x})} \quad 0$$

variance

model error (bias)

$$\hat{x} = A_{tr}^{+}(b_{tr} + e_{tr})$$

$$\bar{x} = A^{+}(b + r) \approx A_{tr}^{+}(b_{tr} + r_{tr})$$

$$\approx \|A A_{tr}^{+}(e_{tr} - r_{tr})\|^2$$

$$\leq \left[\|A\| \|A_{tr}^{+}\| (\|e\|_2 + \|r_{tr}\|_2)\right]^2$$

Recap

$$\|A\hat{x} - b\| \le (1 + \|A\| \|A_{tr}^{+}\|) \|r\|$$
$$+ \|A\| \|A_{tr}^{+}\| \|e\|$$

No error?

$$(1 + \|A\| \|A_{tr}^{+}\|) \|r\|$$

can control (e.g. scaling)

inherent ill-conditioning

can't control

Ill-conditioning

- can be inherent in data
- lots of near-optimal solutions

Regularization helps us pick a solution (next Tues)