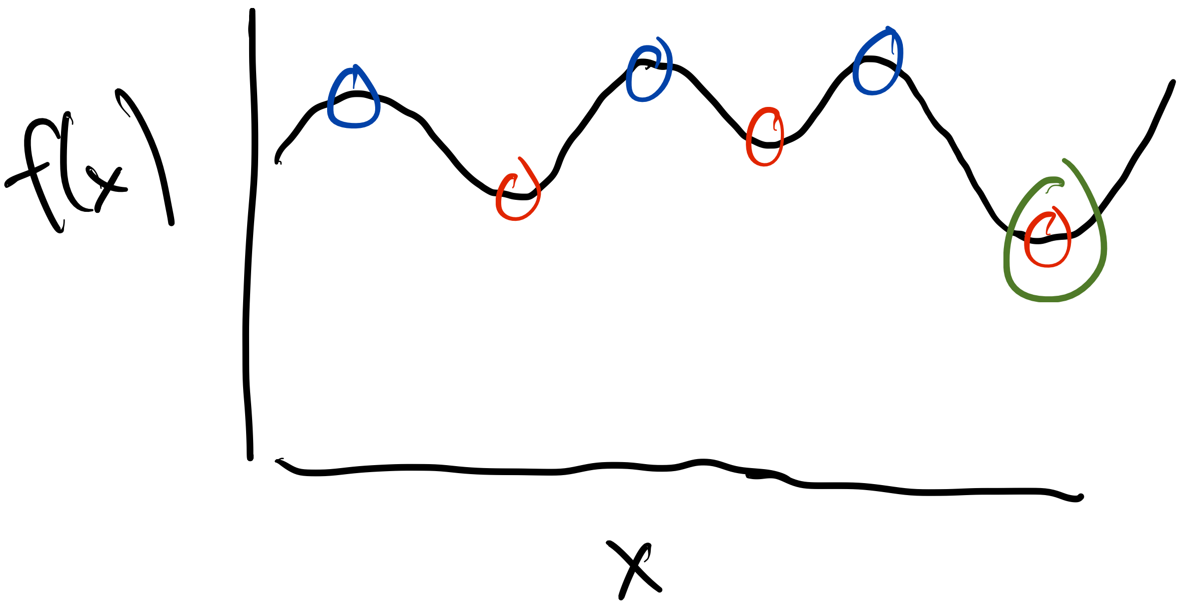
Jan. 23, 2020

$$\min \; f(x) \qquad f: \mathbb{R}^n \to \mathbb{R}$$

$$\text{s.t.} \; \cancel{x \in \Omega} \qquad \Omega = \mathbb{R}^n$$



local mins

global min

convex $\Rightarrow$ local mins are global mins

How do we know if at a local min? (f smooth)

Necessary: $\nabla f(x^*) = 0$

max or a min?

Curvature with Hessian

$H(x) =$ "Hessian at x" $\quad [H(x)]_{ij} = \dfrac{\partial^2}{\partial x_i \partial x_j} f(x)$

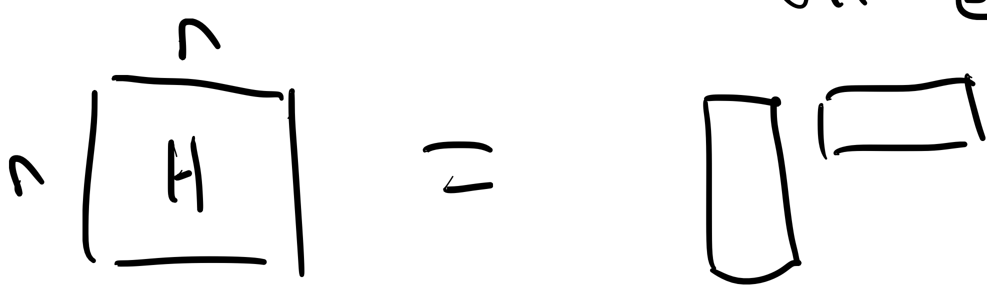
Suppose H is continuous near x^*

<u>Sufficient</u> condition for local min

(1) $\nabla f(x^*) = 0$ + (2) $H(x^*)$ positive def.

(2) $y^T H y > 0 \quad \forall y \in \mathbb{R}^n, \; y \neq 0$

Claim: H is symm. pos. def. (SPD) iff all e-vals > 0



$$y^T H y = y^T X^T X y$$
$$= \|Xy\|_2^2$$

Linear least squares

Model: $b_i \approx a_i^T x + z$

$\uparrow$ outcome $\qquad \uparrow$ features $\qquad$ intercept $\downarrow$

$$\hat{x} = \begin{pmatrix} \hat{x}_1 \\ z \end{pmatrix}$$

Error: $(b - a^T x - z)^2$

$$A = \begin{bmatrix} a_1^T & 1 \\ \vdots & \vdots \\ a_n^T & 1 \end{bmatrix} \qquad \hat{x} = \min_x \|b - Ax\|_2^2$$

Why this error?

(1) seem reasonable

(2) computationally very nice

(3) statistically interpretable

$$b_i = a_i^T x + \varepsilon_i \qquad \varepsilon_i \sim \text{i.i.d. } N(0,1)$$

$$b_i - a_i^T x = \varepsilon \qquad f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$$

$$\text{Likelihood}(b;\, x, A) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \sum_{i=1}^{m} (b_i - a_i^T x)^2\right)$$

Gauss-Markov Thm:

$$b_i = a_i^T x + \varepsilon_i$$

- $\mathbb{E}(\varepsilon_i) = 0$
- $\mathrm{Var}(\varepsilon_i) < \infty$
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$

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

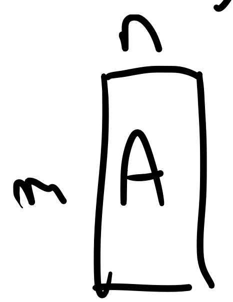
(1) $\hat{x}$ is unbiased $\mathbb{E}(\hat{x}) - \hat{x} = 0$

(2) $\hat{x}$ has least variance amongst <u>all</u> linear unbiased estimators

(BLUE)

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

$A^T r \approx 0$

A  full rank

$$f(x) = \|Ax - b\|_2^2 = (Ax-b)^T(Ax-b)$$

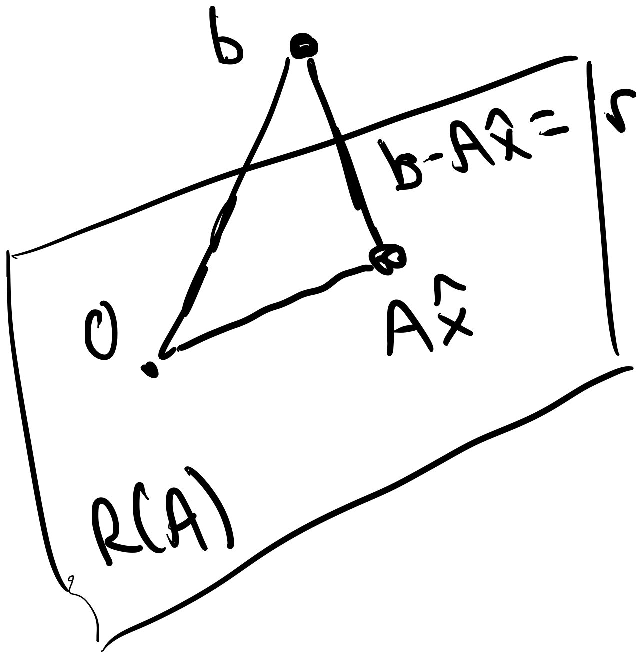
$$= x^T A^T A x - 2x^T A^T b + b^T b$$

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

$$\nabla f(\hat{x}) = 0 \iff A^T(\underbrace{A\hat{x} - b}_{r}) = 0$$

$$\boxed{A^TA\hat{x} = A^Tb \quad \text{normal equations}}$$

$$H(x) = A^TA$$



$$\min r \iff b - Ax \perp R(A)$$

$$R(A) = \{Az\}$$

$$z^T A^T (b - Ax) = 0 \quad \forall z$$

$$z^T \underbrace{(A^T b - A^T A x)}_{=0} = 0 \quad \forall z$$

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

Matrix factorizations for LLS

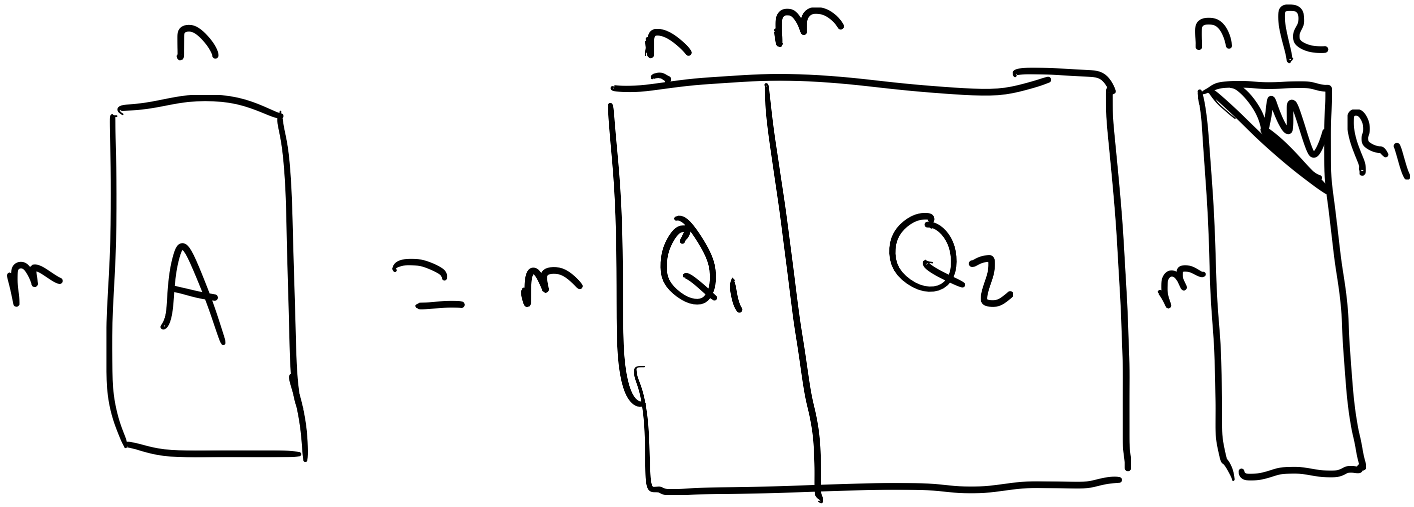
QR

$$\left.\begin{array}{l} \underset{m\times n}{A} = \underset{m\times n}{Q}\;\underset{n\times n}{\begin{bmatrix} * & * \\ 0 & * \end{bmatrix}}_{R} \\[1em] Q^T Q = I \end{array}\right\} O(mn^2)$$

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

$$R^{-T}\, R^T \cancel{Q^T Q} R \hat{x} = R^{-T} R^T Q^T b \;\Rightarrow\; \left.\underbrace{R \hat{x}}_{O(n^2)} = Q^T b \right\} O(mn)$$

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



$$z^T S S^T z = z^T z$$

$$\uparrow$$

Fact: for any orthog. S, S is $m \times m$ $\quad \|S^T z\|_2^2 = \|z\|_2^2$ for all $z \in \mathbb{R}^m$

$$\|Ax - b\|_2^2 = \|Q^T(Ax - b)\|_2^2 = \left\| \begin{pmatrix} R_1 x \\ 0 \end{pmatrix} - \begin{pmatrix} Q_1^T b \\ Q_2^T b \end{pmatrix} \right\|_2^2$$

$$= \underbrace{\|R_1 x - Q_1^T b\|_2^2}_{x = R_1^{-1} Q_1^T b} + \underbrace{\|Q_2^T b\|_2^2}_{\|r\|_2^2}$$

SVD

$$O(mn^2)\left\{\ \boxed{A} = \boxed{U}\ \overset{\Sigma}{\boxed{\diagdown}}\ \boxed{V^T}\right.$$

$$U^TU = I = V^TV = VV^T$$

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

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$$

$$> 0$$

$$A^TA\hat{x} = A^Tb \qquad V^T\cancel{V}\Sigma^T\cancel{U^TU}\Sigma V^T\hat{x} = V^T\cancel{V}\Sigma^TU^Tb$$

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

$$\Sigma V^T\hat{x} = U^Tb \qquad \hat{x} = \underbrace{V}_{O(n^2)}\underbrace{\Sigma^{-1}}_{O(n)}\underbrace{U^Tb}_{O(mn)}$$

$$\text{temp} \sim \text{chirps} \cdot x_1 + x_0$$

$$A = \begin{bmatrix} 10 & 1 \\ 12 & 1 \\ 8 & 1 \\ 15 & 1 \end{bmatrix} \qquad \hat{x}_1 = 2 \qquad \hat{x}_0 = 50$$

peeps

$$\text{temp} \sim \text{peeps} \cdot x_2 + \text{chirps} \cdot x_1 + x_0$$

$$\bar{A} = \begin{bmatrix} 13 & 10 & 1 \\ 14.9 & 12 & 1 \\ 11.2 & 8 & 1 \\ 18.1 & 15 & 1 \end{bmatrix}$$

$$\text{peeps} \sim \text{chirps} + 3$$

$$x_2 \approx x_1 + 3x_0$$

linear correlation

$$\bar{A}\begin{pmatrix} 0.01 \\ 2 \\ 50 \end{pmatrix} \approx \bar{A}\begin{pmatrix} 1 \\ 1 \\ 47 \end{pmatrix}$$

lots of near-opt solutions

$\bar{A}$ is ill-conditioned

Bias - variance tradeoff

prediction error =

bias error + variance error

model: $Ax \approx b$ actual: $Ax = b + r$ $A^T r = 0$

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

observe:

A_{tr}, $b_{tr} + e$

w/ random

Model fit: $\hat{x} = \min_x \| A_{tr} x - (b_{tr} + e) \|_2^2$

$\| A\hat{x} - b \|_2^2$ vs. $\| r \|_2^2$?

$$Ax = b + r \qquad A(\hat{x} - x) + r = A\hat{x} - b$$

$$\|A\hat{x} - b\|_2^2 = \|A(\hat{x} - x) + r\|_2^2$$

$$= \underbrace{\|r\|_2^2}_{\text{bias}} + \underbrace{\|A(\hat{x} - x)\|_2^2}_{\text{variance}}$$

$$\hat{x} = A_{tr}^{+}(b_{tr} + e), \quad x = A_{tr}^{+}(b_{tr} + r_{tr})$$

$$\|A(\hat{x} - x)\|_2 = \|AA_{tr}^{+}(e - r_{tr})\|_2$$

$$\leq \|A\| \, \|A_{tr}^{+}\| \left(\|e\| + \|r_{tr}\|\right)$$