$\underline{\text{Lec 04}}$: Linear Regression (Contd.)

$$\min_{w} \; E(w, D) = \| Xw - y \|^2$$

1-d fn: $(ax - y)^2$ $\qquad \dfrac{d^2 f}{dx^2} \geq 0$

d-dim $\| Ax - y \|^2 \longrightarrow A$ is $\underline{\text{PSD}}$

$$\boxed{w^* = (X^T X)^{-1} X^T y}$$

(pseudoinverse)

$\nabla_w E = 0 \Rightarrow$

$X^T(Xw) - X^T y = 0$

$X^T y$

$$a_{11}x_1 + \cdots + a_{1d}x_d = b_1$$

$$\vdots$$

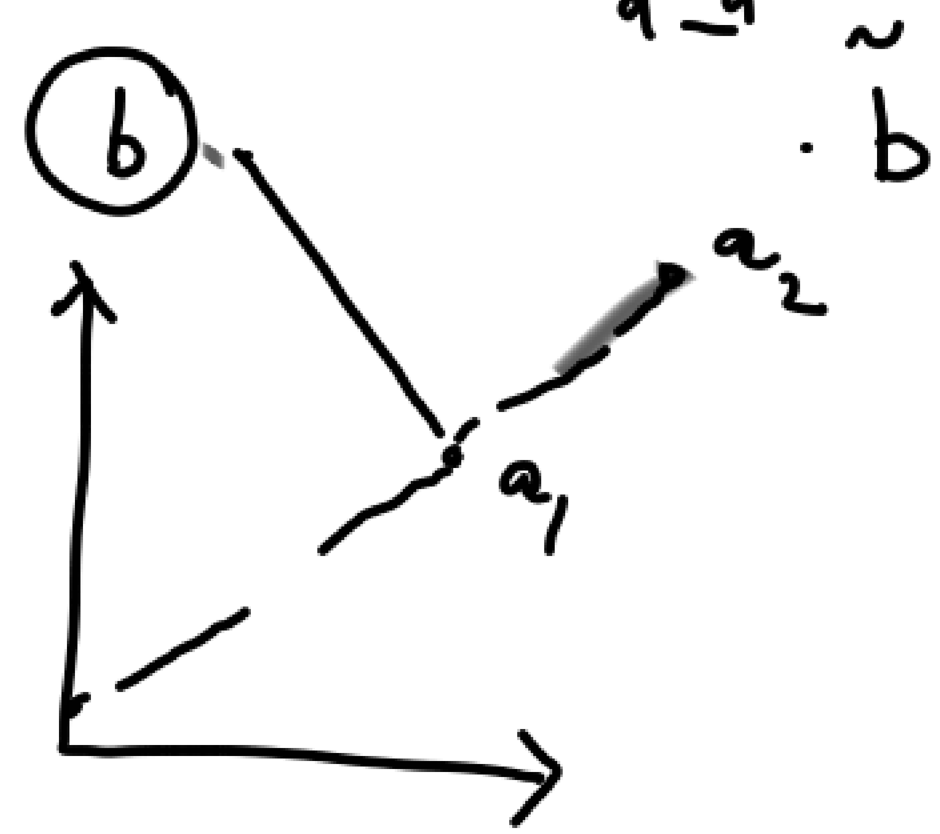
$$a_{n1}x_1 + \cdots + a_{nd}x_d = b_n$$

$$Ax = b \rightarrow$$

$$\begin{bmatrix} | & & | \\ a_1 & \cdots & a_d \\ | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = x_1 \underline{a}_1 + x_2 \underline{a}_2 + \cdots + x_d \underline{a}_d$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} x = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \end{bmatrix} \cdot a_1$$

$\cdot\, a_2$ $\quad \cdot\, b$



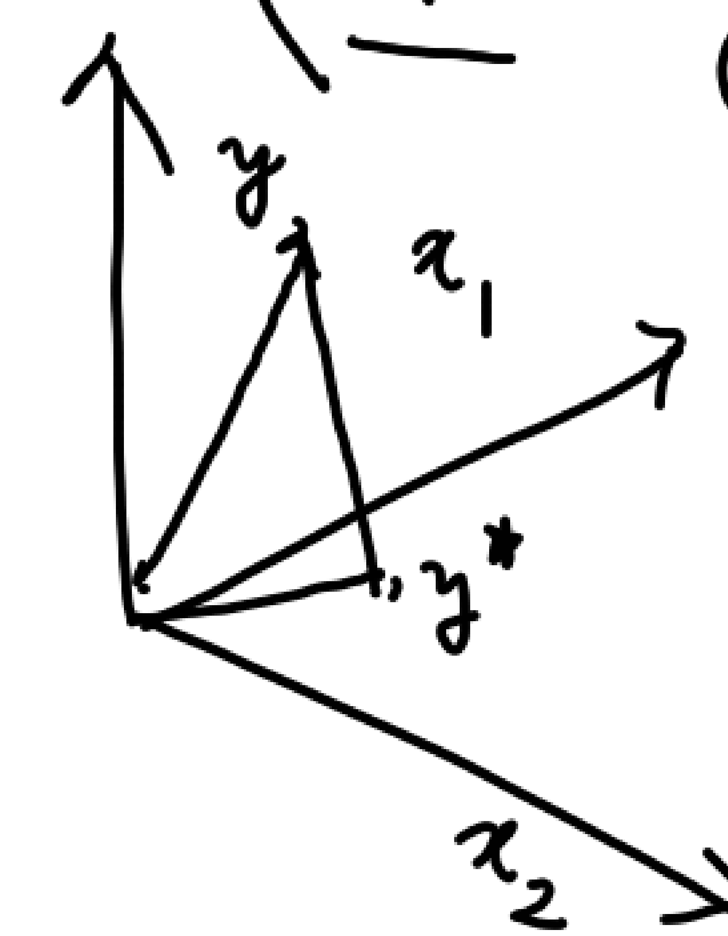
Case 1: $\min \|Xw - y\|^2$

$$Xw^* = y$$

Case 2: can't find w^*

$$X^T X w - X^T y = 0$$

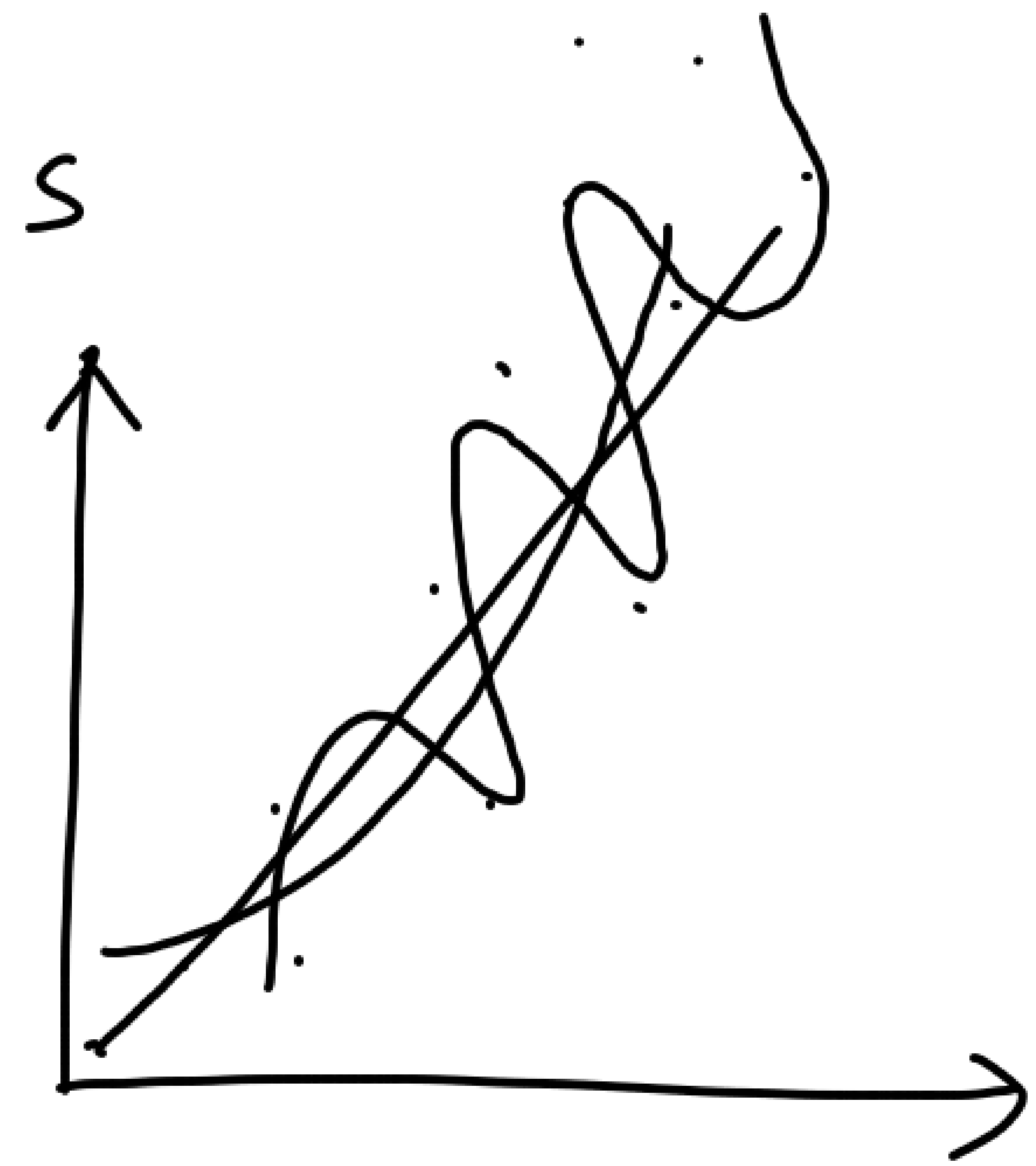
$$\Rightarrow X^T(\underbrace{Xw^*}_{\to\, y^*} - y) = 0$$



Regression model with basis functions

$$\hat{y}_i = W_0 + W_1 x_i + \left(W_2 x_i^2 + w_3 x_i^3 + \cdots + w_m x_i^m \right)$$

1-D data : x_i's scalars



$$y^* = \phi w^*$$

$$\phi_0(x_i) = 1$$

$$\phi_1(x_i) = x_i$$

$$\vdots$$

$$\phi_m(x_i) = x_i^m$$

$$\hat{y}_i = \sum_{j=0}^{m} w_j \phi_j(x_i) \qquad m >> d$$

d-dim

$$\hat{y}_i = \sum_{j=0}^{m} \phi_j(x_j) w_j$$

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}$$

$$\Phi = \begin{bmatrix} \phi_0(x_1) & \cdots & \phi_m(x_1) \\ \vdots & & \\ \phi_0(x_n) & \cdots & \phi_m(x_n) \end{bmatrix}_{n \times (m+1)} \qquad W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix}_{(m+1)\times 1}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \operatorname{argmin} \| \Phi W - y \|^2$$

$$w^* = (\phi^T \phi)^{-1} \phi^T y$$

$$y_i = \underline{w}^T x_i + \epsilon_i$$

Probabilistic model of linear regression

noisy linear model

$$D = \{(x_i, y_i)\}_{i=1}^{n}$$

parameters: w

noise: $\epsilon_i \sim N(0, \sigma^2)$ i.i.d.

$\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$

- Estimate w from this probabilistic model
- Maximum likelihood estimation (MLE)

MLE

A set of independent and identically distributed (i.i.d.) observations $\{y_1, \dots, y_n\}$ are generated by a probabilistic model parametrized by θ

$$y_j \sim \underbrace{P(y \mid \theta)}_{\text{Likelihood}} \qquad P(y;\theta)$$

Log likelihood: $\log\left(P(y|\theta)\right)$

- Mathematically nicer
- Numerical advantage

$$P(\underline{y}|\underline{x},\theta)$$

$$= \prod_{i=1}^{n} P(y_i|x_i,w)$$

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} L(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{n} \log P(y_j|\theta)$$

Coin toss: Toss a coin n times, each is a binary RV with Bernoulli dist

$$P(y_j|\theta) = \theta^{y_j}(1-\theta)^{1-y_j}$$

$$L(\theta) = \log P(y|\theta) = \log\left(\prod_{j=1}^{n} P(y_j|\theta)\right)$$

$$= \sum_{j=1}^{n} \log(P_j|\theta) = \sum_{j=1}^{n}\left(y_j \log\theta + (1-y_j)\log(1-\theta)\right)$$

$$\Rightarrow \boxed{\hat{\theta}_{MLE} = \frac{1}{n}\sum_{j=1}^{n} y_j}$$

MLE for regression

$$y_j = w^T x_j + \epsilon_j \sim N(0, \sigma^2)$$

$$y_j \sim N(w^T x_j, \sigma^2)$$

$$P(y_j \mid x_j, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_j - w^T x_j)^2}{2\sigma^2} \right\}$$

$$L(w) = \text{const.} - \sum_{j=1}^{n} \frac{(y_j - w^T x_j)^2}{2\sigma^2}$$

$$\arg\max_{w} L(w)$$

$$w^*_{MLE} = \arg\min_{w} \sum_{j=1}^{n} (y_j - w^T x_j)^2$$

$$\arg\min \| Xw - y \|^2$$

$$w^* = (X^TX)^{-1}X^Ty$$

Algorithmic way to optimize

Gradient Descent

- $w \leftarrow w_0$
- Repeat until convergence $\rightarrow$ $\|\nabla E(w)\| < \epsilon$

$$w \leftarrow w - \eta \nabla_w E$$

learning rate

$$\|Xw - y\|^2 = E$$