466/566: Sept 17 - Oct. 3 2019

$\mathcal{X}$ observations / inputs, e.g. $\mathcal{X} \subseteq \mathbb{R}^d$

$\mathcal{Y}$ targets e.g. $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \mathbb{R}^d$

$f : \mathcal{X} \to \mathcal{Y}$ $\quad$ $\text{cost} : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$

$\text{cost}(\hat{y}, y)$, $\hat{y} = f(x)$ $\quad$ X random variable.

instance $x \in \mathcal{X}$, $y \in \mathcal{Y}$ (x, y) $\quad$ Y random variable

$C = \text{cost}(f(X), Y)$

Goal: minimize $E[C]$
$f \in \mathcal{F}$

Classification: (binary) or multi-class

$$\text{cost}(\hat{y}, y) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & y \neq \hat{y} \end{cases}$$

e.g. medical setting

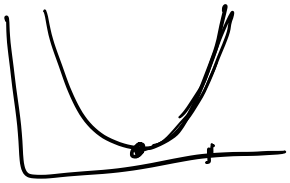
$\hat{y}$: 0 ¬have disease, no test; 1 have disease, test

	y: 0 ¬have disease	y: 1 have disease
0	0	Lawsuit = 1000
1	$C_{lab} = 1$	$C_{lab} = 1$

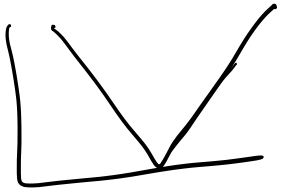
Regression: $y \in \mathbb{R}$

$\text{cost}(\hat{y}, y) = (\hat{y} - y)^2$ (squared error)

$\text{cost}(\hat{y}, y) = |\hat{y} - y|$



sq. err.



abs. error

Classification, 0-1 cost $\quad \mathcal{X} \subseteq \mathbb{R}^d, \quad \mathcal{Y} = \{1, 2, \ldots, k\}$

$$E[C] = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \text{cost}(f(x), y)\, \underbrace{p(x, y)}_{p(y|x)\,p(x)}\, dx$$

$$= \int_{\mathcal{X}} p(x) \underbrace{\left[\sum_{y \in \mathcal{Y}} \text{cost}(f(x), y)\, p(y|x) \right]}_{E[C \mid X = x]} dx$$

$$f^*(x) = \underset{\hat{y} \in \mathcal{Y}}{\text{argmin}} \; E[C \mid X = x] \qquad p(y|x) = \begin{cases} 0.9 \\ 0.1 \end{cases}$$

$$= \underset{\hat{y} \in \mathcal{Y}}{\text{argmin}} \sum_{y \in \mathcal{Y}} \text{cost}(\hat{y}, y)\, p(y|x)$$

$$= \underset{y \in \mathcal{Y}}{\text{argmax}} \; p(y|x)$$

Regressione: Conclusione: $\hat{y} = E[Y|x]$

$$\Rightarrow f^*(x) = E[Y|x]$$

$$E[C] = \int_X \underbrace{\int_y (f(x) - y)^2 p(y|x)\,dy}_{g(f(x))} p(x)\,dx$$

$$g(\hat{y}) = \int_y (\hat{y} - y)^2 p(y|\underline{x})\,dy$$

$$0 = \frac{d\,g(\hat{y})}{d\hat{y}} = \int_y \frac{d\,(\hat{y}-y)^2}{d\hat{y}} p(y|x)\,dy = 2\int_y (\hat{y}-y)\,p(y|x)\,dy$$

$$= 2\hat{y}\underbrace{\int_y p(y|x)\,dy}_{=1} - 2\underbrace{\int_y y\,p(y|x)\,dy}_{E[Y|x]}$$

1. $f(x) = E[Y|x]$ or 2. $f(x) \neq E[Y|x]$

$$E[C] = \int_X p(x) \underbrace{\int_y (E[Y|x] - y)^2 p(y|x)\,dy}_{= \mathrm{Var}[Y|X=x]}\,dx$$

2. $f(x) \neq E[Y|x]$

$$(f(x) - y)^2 = \big(\underbrace{f(x) - E[Y|x]}_{} + \underbrace{E[Y|x] - y}_{}\big)^2$$

$$= (f(x) - E[Y|x])^2 \cdot$$

$$+ 2\underbrace{(f(x) - E[Y|x])(E[Y|x] - y)}_{\text{call this } g(x,y)} \cdot$$

$$+ (E[Y|x] - y)^2 \cdot$$

$$E[g(x,Y)] = E\left[(f(x) - E[Y|x])(E[Y|x] - Y) \mid x\right]$$

$$= (f(x) - E[Y|x])\, E\left[E[Y|x] - Y \mid x\right]$$

$$= 0$$

$$E[Y|x] - E[Y|x] = 0$$

$$E[C] = E[(f(X) - Y)^2]$$

$$= E[(f(X) - E[Y|X])^2] + E[(E[Y|X] - Y)^2]$$

$p(y|x)$

reducible error

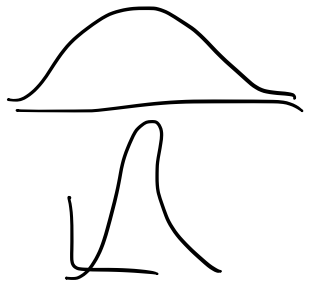
variance of Y

irreducible error.

$f \neq f^*$

$p(y|x_1)$

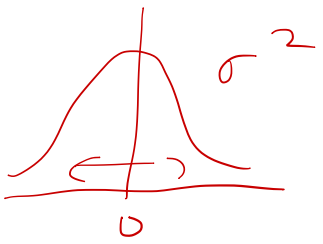
$p(y|x_2)$



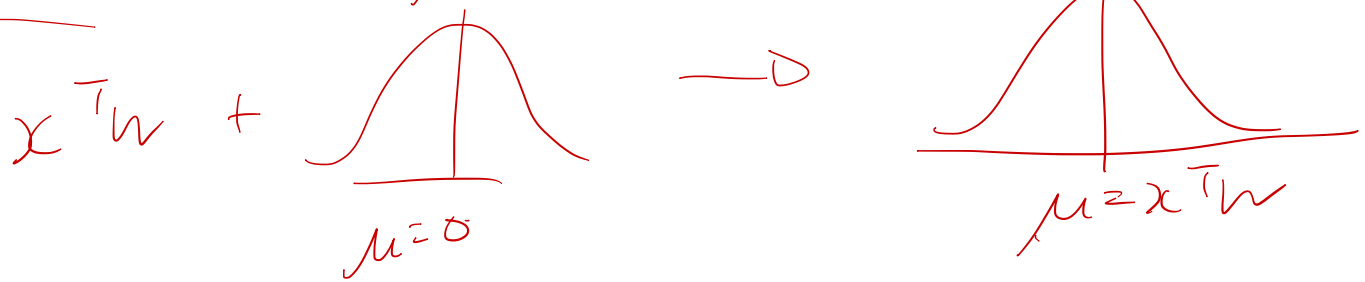
$\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d$ $\quad y \in \mathbb{R}$ $\quad Y = \sum_{j=1}^{d} w_j^* X_j + \varepsilon$

(X, Y)

$\varepsilon \sim \mathcal{N}(\mu = 0, \sigma^2)$



$p(y|x) = \mathcal{N}(\mu = x^T w^*, \sigma^2)$



Fixed $\sigma^2 > 0$

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - x^T w)^2}{2\sigma^2}\right)$$

$$\sum_{i=1}^{n} \log p(y_i|x_i, w) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y - x^T w)^2}{2\sigma^2}$$

$$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; -\sum_{i=1}^{n} \ln p(y_i \mid x_i)$$

$$= \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{(y_i - x_i^T w)^2}{2\sigma^2} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T w)^2$$

$j \in \{1, \dots, d\}$

$$\frac{\partial (y - x^T w)^2}{\partial w_j} = -2(y_i - x^T w) \frac{\partial x^T w}{\partial w_j}$$

$$\frac{\partial x^T w}{\partial w_j} = \frac{\partial \sum_{i=1}^{d} w_i x_i}{\partial w_j} = \frac{\partial (w_1 x_1 + \dots + w_j x_j + \dots + w_d x_d)}{\partial w_j}$$

$$= \frac{\partial w_j x_j}{\partial w_j} = x_j$$

$$\frac{\partial \sum_{i=1}^{n} (y_i - x_i^T w)^2}{\partial w_j} = -2 \sum_{i=1}^{n} (y_i - x_i^T w) x_{ij} = 0$$

$\forall j \in \{1, \dots, d\}$

$$-\sum_{i=1}^{n}(y_i - x_i^T w)\,x_{ij} = -\sum_{i=1}^{n} y_i x_{ij} + \sum x_{ij} x_i^T w$$

$$\begin{bmatrix} \sum_{i=1}^{n} y_i x_{i1} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{id} \end{bmatrix}_{d\times 1} \begin{matrix} \leftarrow \frac{\partial}{\partial w_1} \\ \\ \leftarrow \frac{\partial}{\partial w_d} \end{matrix} = \begin{bmatrix} \sum_{i=1}^{n} x_{i1} x_i^T w \\ \vdots \\ \sum_{i=1}^{n} x_{id} x_i^T w \end{bmatrix}_{j=1 \ldots d} = 0$$

$$\Downarrow \quad d\times 1$$

$$\underbrace{\sum_{i=1}^{n} y_i x_i}_{b} - \underbrace{\sum_{i=1}^{n} x_i x_i^T w}_{A} = 0$$

$$b - Aw = 0 \Rightarrow Aw = b$$

If A invertible

$$w = A^{-1} b.$$

$$A = \sum_{i=1}^{n} x_i x_i^T$$

n = # samples
d = # features

$$X = \underset{n \times d}{\begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_n^T \end{bmatrix}} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ x_{21} & \cdots & x_{2d} \\ \vdots & & \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}$$

$$b = \sum_{i=1}^{n} x_i y_i = X^T y$$

$$\underset{n \times 1}{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$A = X^T X = \sum_{i=1}^{n} x_i x_i^T$$

$$X^T = \begin{bmatrix} \vec{x}_1, \cdots, \vec{x}_n \end{bmatrix}$$

$$\underset{n \times d}{X} = \underset{n \times n}{U} \; \underset{n \times d}{\Sigma} \; \underset{d \times d}{V^T}$$

$$\underset{n \times d}{\Sigma} = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_d \\ \hline & 0 & \end{bmatrix} \quad \sigma_d = 0$$

$n - d$ rows of zeros — zero matrix

$$U^T U = I$$

$$U U^T = I$$

$d < n$

$$X^T X = (U \Sigma V^T)^T (U \Sigma V^T) \qquad (AB)^T = B^T A^T$$

$$= V \Sigma^T U^T U \Sigma V^T \qquad U^T U = I$$

$$= V \Sigma^T \underbrace{U^T U}_{=I} \Sigma V^T$$

dxd. dxn nxd dxd.

$$= V \Sigma^T \Sigma V^T$$

$$\Sigma^T \Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix} \quad d \times d$$

$$\Sigma = \left[\begin{array}{ccc} \sigma_1 & & 0 \\ & \ddots & \sigma_d \\ \hline 0 & & \end{array}\right]$$

$$A = X^T X$$

$$= V \Lambda V^T \qquad \Lambda = \Sigma^T \Sigma \qquad \sigma_d^2 = 0$$

$$Aw = V(\Sigma^T \Sigma) \underbrace{V^T w}_{\tilde{w}} = V \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_d^2 \end{bmatrix} \tilde{w}$$

$$= V \begin{bmatrix} \sigma_1^2 \hat{w}_1 \\ \vdots \\ \sigma_d^2 \hat{w}_d \end{bmatrix} = V \begin{bmatrix} \sigma_1^2 \tilde{w}_1 \\ \vdots \\ \sigma_{d-1}^2 \hat{w}_{d-1} \\ 0 \end{bmatrix}$$

$$Aw = b$$

$\min_w \|Aw - b\|_2^2$

$Aw = b$

$\min_w \|Xw - y\|_2^2$

$Xw \neq y$

$\hat{y} \neq y$

$\Sigma_d^{-2} \Sigma^T U$

$d \times n \quad n \times n$

$d \times n$

$\Sigma^T U^T = \Sigma_d U_d^T$

$[\Sigma_d \mid 0]\, U^T = \Sigma_d U_d^T$

$\left[\frac{U_d^T}{U_{d+1:n}^T}\right]$

$$\min_w \sum_{i=1}^{n} (x_i^T w - y_i)^2 = \min_w \underline{\|Xw - y\|_2^2} \leftarrow \|\hat{y} - y\|_2^2$$

$$X = \underset{n \times d}{\begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ x_{21} & \cdots & x_{2d} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} \qquad x_i \in \mathbb{R}^d \qquad x_{ij} \in \mathbb{R} \qquad \text{ugrid } X_{ij} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\|z\|_2 = \sqrt{\sum_{i=1}^{n} z_i^2} \qquad \|z\|_2^2 = \sum_{i=1}^{n} z_i^2$$

$$Xw = \hat{y} \underset{n \times d}{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}}$$

System: $\underbrace{X^T X}_{A} \underset{d \times d}{w} = \underbrace{X^T y}_{b}$ $\quad d \times n \;\; n \times d \;\; d \times 1 \;\; d \times 1$

$X = U \Sigma V^T$ where $\Sigma = \underset{n \times d}{\begin{bmatrix} \Sigma_d \\ \hline 0 \end{bmatrix}}$ $\quad \Sigma_d = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_d \end{bmatrix}$ diagonal matrix

$n \times n \;\; n \times d \;\; d \times d$

$U^T U = I$

$V^T V = I$

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$

If A is invertible, A^{-1} is inverse, $w = A^{-1}b$

(i.e. $A^{-1}A = I$, $Aw = b \Rightarrow A^{-1}Aw = A^{-1}b$, $I \Rightarrow w = A^{-1}b$)

$X^TX = (U\Sigma V^T)^T(U\Sigma V^T)$ (d×d, d×n, n×d, d×d)

$= V\Sigma^T U^T U \Sigma V^T = V\underbrace{\Sigma^T\Sigma}V^T = V\Sigma_d^2 V^T$ (d×d, d×d, d×d)

$(X^TX)^{-1} = V\Sigma_d^{-2}V^T$ ($=I$)

$A^{-1} = (X^TX)^{-1} = V\Sigma_d^{-2}V^T$ $\quad (A^{-1}A \overset{?}{=} I)$ check.

$(X^TX)^{-1}X^T = V\Sigma_d^{-2}\underbrace{V^TV}_{=I}\Sigma^T U^T$ $\quad X^T = V\Sigma^T U^T$ (d×d)

$= V\underbrace{\Sigma_d^{-2}\Sigma^T}U^T$ $\quad \begin{bmatrix}\sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_d^{-2}\end{bmatrix}\begin{bmatrix}\sigma_1 & & 0 \\ & \ddots & \\ & & \sigma_d\end{bmatrix}$

$= V\Sigma_d^{-1}U_d^T$ ← X^+ pseudo-inv $\quad \Sigma_d^{-2}$ $\quad \Sigma^T$

$w = V\Sigma_d^{-1}U_d^T y$ ← $\quad U_d$ = first d columns

$X^+X = I$ ✓ check

$w = \underline{X^+} y$ if X full rank $\sigma_1 \geq \dots \sigma_d > 0$

or X not full rank, then $w = X^+ y$ is still solution

$$w = \underset{d\times n}{X^+} y = \left(\sum_{j=1}^{r} \frac{v_j u_j^\top}{\sigma_j} \right)_{d\times n} y_{n\times 1}$$

$r = \text{rank}(X) \leq d$.

$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > \sigma_{r+1} = 0, \dots$

$\sigma_d = 0$

If $\sigma_r = 10^{-6}$

$\Rightarrow$ might cause big w

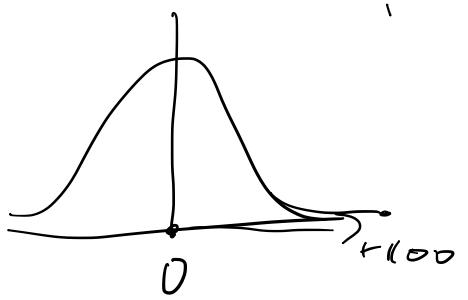
$$\Sigma_d^+ = \begin{bmatrix} \sigma_1^{-1} & & & \\ & \ddots & & \\ & & \sigma_r^{-1} & \\ & & & 0 \\ & & & & 0 \end{bmatrix}$$

---

prior on weights: $w_j \sim \mathcal{N}(\mu = 0, \lambda^{-1}) \quad \forall j \in \{1, \dots d\}$

λ big, λ^{-1} small

λ small, λ^{-1} big



$$\min_w -\sum_{i=1}^{n} \ln p(y_i|x_i,w) - \ln p(w)$$

prior

$$p(w_j) = \frac{1}{\sqrt{2\pi\lambda^{-1}}} \exp\left(-\frac{(w_j-0)^2\lambda}{2}\right)$$

drop constants

$$-\ln p(w_j) = \boxed{-\ln \frac{1}{\sqrt{2\pi\lambda^{-1}}}} + \frac{\lambda w_j^2}{2} \quad \forall j$$

$$p(\vec{w}) = \prod_{j=1}^{d} p(w_j) \Rightarrow \ln p(w) = \sum_{j=1}^{d} \ln p(w_j)$$

$$\text{MAP}: \operatorname*{argmin}_w \sum_{i=1}^{n} (x_i^T w - y_i)^2 + \frac{\lambda}{2} \underbrace{\sum_{j=1}^{d} w_j^2}_{\|w\|_2^2}$$

$$\equiv \operatorname*{argmax}_w p(D|w)p(w)$$

Gaussian prior

Gaussian $p(y|x)$

Procedure: ① take log ② drop constants ③ min negative instead of max

MLE $\equiv$ min negative log likelihood, MAP: min $-$ log of posterior

$$w_{MLE} = (X^T X)^{-1} X^T y$$

$$w_{MAP} = (X^T X + \lambda I)^{-1} X^T y$$

$d \times d$

or $w_{MLE} = X^+ y$

$$\lambda I = \begin{bmatrix} \lambda & & 0 \\ & \lambda & \\ 0 & & \ddots \\ & & & \lambda \end{bmatrix}$$

$$X^T X + \lambda I = V \Sigma_d^2 V^T + \lambda \underbrace{I}_{VV^T} = V \Sigma_d^2 V^T + V(\lambda I) V^T$$

$$= V(\Sigma_d^2 + \lambda I) V^T$$

$$w_{MLE} = \sum_{j=1}^{r} \frac{v_j u_j^T}{\sigma_j} y$$

σ_j ← could be small

$$w_{MAP} = \sum_{j=1}^{d} \sigma_j \frac{(v_j u_j^T)}{\sigma_j^2 + \lambda} y$$

$\sigma_j^2 + \lambda$ ≠ 0

$$(X^T X + \lambda I)^{-1} X^T$$

$$(\Sigma_d^2 + \lambda I)^{-1} \Sigma_d$$

Exercise: If $\lambda > 0$, is $(X^T X + \lambda I)$ invertible

X → rescale $2X$ $A^2 = AA$ ($n \times d$ $n \times d$)

$\Sigma\Sigma$

$Xw \approx y$ $\quad 2Xw \approx y$ $\quad \Sigma^T\Sigma = \Sigma^2$ $\quad \Sigma^3$?

$\downarrow$ $\quad w = \frac{1}{2} w_{best}$

w_{best}

$2X \frac{1}{2} w_{best} = Xw_{best}$

$\lambda = 1$

λ should be $\sqrt{2}$

Σ_d $d \times d$ $\quad d \times (n-d)$

$$\Sigma^T\Sigma = \begin{bmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_d \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_d & \\ 0 & & & \end{bmatrix}$$

$d \times d$ $\quad (n-d) \times d$

$$= \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix} = \Sigma_d^2$$

$w_{MLE}(\mathcal{D})$

↑ random variable

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

↑ random variable because its a sample

$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

$\varepsilon_i \sim N(0, \sigma^2)$

$$w_{MLE}(\mathcal{D}) = (X^T X)^{-1} X^T y$$

unbiased?

w = true underlying weight

ε noise in targets

$$\underset{n\times 1}{y} = \underset{n\times d}{X} w + \underset{n\times 1}{\varepsilon}$$

$$E[w_{MLE}(\mathcal{D})] = E[(X^T X)^{-1} X^T y]$$

$$= E[(X^T X)^{-1} X^T (Xw + \varepsilon)]$$

$$= E[(X^T X)^{-1} X^T X w] + E[(X^T X)^{-1} X^T \varepsilon]$$

$$= E[w] + E[(X^T X)^{-1} X^T] E[\varepsilon]$$

X^+ ; $E[\varepsilon] = 0$

independent RVs A, B: $E[AB] = E[A]E[B]$

$$= w$$

$$w_{MLE}(\mathcal{D}) = w + X^+ \varepsilon$$

$$\mathrm{Cov}(w_{MLE}(\mathcal{D})) = E\left[(w_{MLE}(\mathcal{D}) - w)(w_{MLE}(\mathcal{D}) - w)^T\right]$$

$w_{MLE}(\mathcal{D}) - w$: $d \times 1$; $(w_{MLE}(\mathcal{D}) - w)^T$: $1 \times d$; $E[w_{MLE}(\mathcal{D})]^T$

$$= E\left[w_{MLE}(\mathcal{D}) w_{MLE}(\mathcal{D})^T\right] - w w^T$$

$$w_{MLE}(\mathcal{D}) = w + X^+ \varepsilon$$

$$= E\left[(w + X^+\varepsilon)(w + X^+\varepsilon)^T\right] - w w^T$$

$$= E[w w^T] + \underbrace{E[X^+\varepsilon w^T]}_{E[X^+]E[\varepsilon]w^T = 0} + \underbrace{E[w(X^+\varepsilon)^T]}_{w E[\varepsilon^T] E[X^+]^T = 0} + \underbrace{E[X^+\varepsilon(X^+\varepsilon)^T]}_{-ww^T}$$

$$= ww^T + E[X^+\varepsilon\varepsilon^T X^{+T}] - ww^T$$

$$= E[X^+\varepsilon\varepsilon^T X^{+T}]$$

$$= E[E[X^+\varepsilon\varepsilon^T X^{+T} | X]]$$

$$= E[X^+ E[\varepsilon\varepsilon^T | X] X^{+T}]$$

$$= \sigma^2 E[X^+ X^{+T}] = \sigma^2 E\left[\sum_{i=1}^{r} v_j v_j^T / \sigma_j^2\right]$$

Recall

$$E[A] = E[E[A|B]]$$

$$E[\varepsilon\varepsilon^T | X] = E[\varepsilon\varepsilon^T]$$

$$= E[\varepsilon\varepsilon^T] - E[\varepsilon]E[\varepsilon]^T$$

$$= \mathrm{cov}(\varepsilon) = \sigma^2 I$$

$$w_{MAP}(\mathcal{D}) = (X^T X + \lambda I)^{-1} X^T y$$

$$E[w_{MAP}(\mathcal{D})] = E[(X^T X + \lambda I)^{-1} X^T (Xw + \varepsilon)]$$

$$= \underbrace{E[(X^T X + \lambda I)^{-1} X^T X]}_{\neq I} w + 0$$

$$\neq w$$

$\lambda = 0$, unbiased MLE

n large, close to unbiased

$$X^T X = \sum_{i=1}^{n} x_i x_i^T$$

$$X^T X + \lambda I = \sum_{i=1}^{n} x_i x_i^T + \lambda I$$

$$\frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

$$\Updownarrow$$

$$w_{MLE}(\mathcal{D}) = \left(\frac{1}{n} X^T X\right)^{-1} \left(\frac{1}{n} X^T y\right)$$

$$w_{MAP}(\mathcal{D}) = \left(\frac{1}{n} X^T X + \frac{\lambda}{n} I\right)^{-1} \left(\frac{1}{n} X^T y\right)$$

$$\text{Bias}\left(w_{MAP}(\mathcal{D})\right)^2 = \| E[w_{MAP}(\mathcal{D})] - w \|_2^2$$

$$\text{Bias}\left(w_{MLE}(\mathcal{D})\right)^2 = \| \underbrace{E[w_{MLE}(\mathcal{D})]}_{w} - w \|_2^2$$

$$= \| w - w \|_2^2 = 0$$

Assume $Y = f(X) + \varepsilon$ $\quad \varepsilon \sim N(\mu = 0, \sigma^2)$

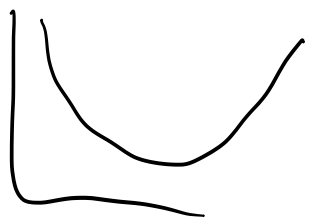
$\hat{f}(x) = x^T \hat{w}$ $\quad$ w true parameters for f

MSE = mean squared error

$$MSE(\hat{w}) = E[\|\hat{w} - w\|_2^2]$$

$\hat{w}$ is an RV

$$= \underbrace{\|E[\hat{w}] - w\|_2^2}_{Bias(\hat{w})^2} + Var(\hat{w})$$



more generally: mse for one sample (x, y)

$$E[(\hat{f}(x) - f(x))^2] = \underbrace{(E[\hat{f}(x)] - f(x))^2}_{Bias} + Var(\hat{f}(x))$$

$$MSE(\hat{f}) = E[(\hat{f}(X) - f(X))^2]$$

$$E[(Y - \hat{f}(X))^2] = E[(\hat{f}(X) - f(X))^2] + \sigma^2$$

reducible ↙ $\quad$ ↙ irreducible

Objective function $c: \mathbb{R}^d \to \mathbb{R}$ $\min_w c(w)$

Starting simple : $d=1$

$$c(w) = \sum_{n=0}^{\infty} \frac{c^{(n)}(w_0)}{n!} (w-w_0)^n$$

$\nabla c(w) = 0$

Second-order

Locally around w_0 $\frac{d}{dw} c(w) = 0$

Start at w_0

$$c(w) \approx \hat{c}_0(w) = c(w_0) + (w-w_0) c'(w_0) + \frac{1}{2}(w-w_0)^2 c''(w_0)$$

$$\frac{d\hat{c}_0(w)}{dw} = 0 + c'(w_0) + (w-w_0) c''(w_0) = 0$$

$$\Rightarrow w_1 = w_0 - \frac{c'(w_0)}{c''(w_0)}$$

$w_1 \neq \min_w c(w)$

$c''(w_t) > 0$

$$\therefore w_{t+1} = w_t - \frac{c'(w_t)}{c''(w_t)}$$

eventually $w_t \to$ stationary point of c

First-order: $\hat{c}_0(w) = c(w_0) + (w - w_0)\, c'(w_0) + \frac{1}{2\eta}(w - w_0)^2$

$\Rightarrow w_{t+1} = w_t - \eta\, c'(w_t)$

stepsize.

$\frac{1}{2\eta}(w - w_0)^2 + O((w - w_0)^2)$

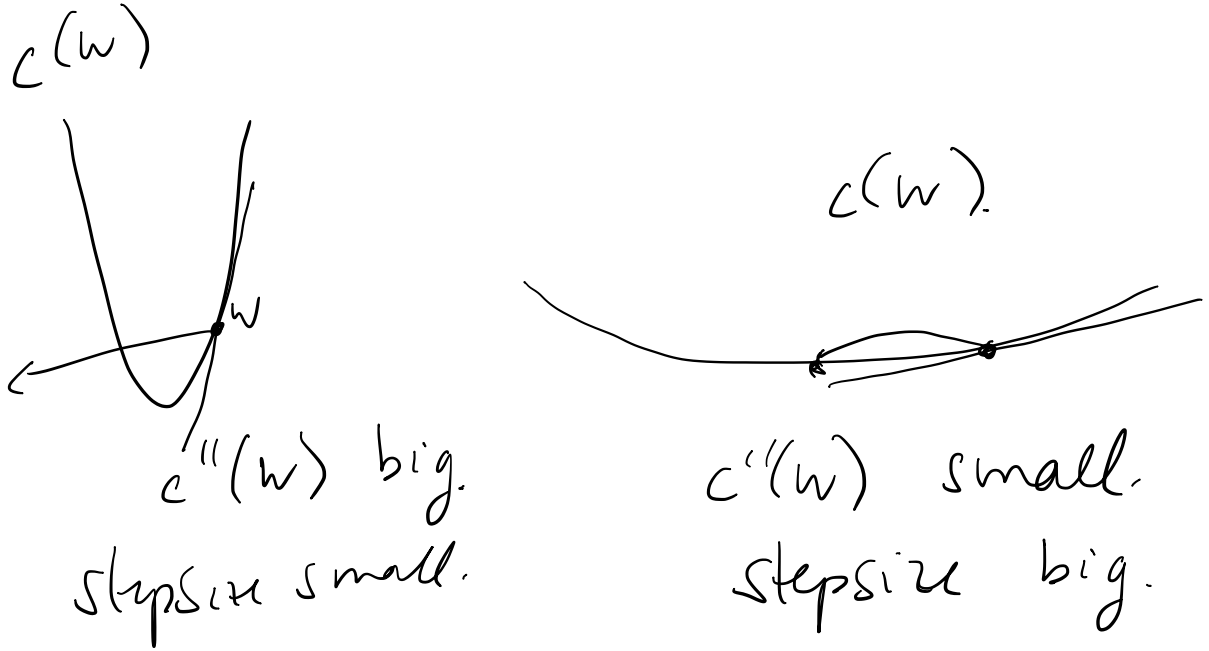
↳ approximates $\frac{c''(w_0)}{2}(w - w_0)^2$

Stepsize for 1st: η.

Stepsize for 2nd: $\frac{1}{c''(w_t)}$

Recall: 2nd.

$w_{t+1} = w_t - \frac{c'(w_t)}{c''(w_t)}$



$c''(w)$ big. stepsize small.

$c''(w)$ small. stepsize big.

$$\nabla C(w) = \begin{bmatrix} \frac{\partial C}{\partial w_1}(w) \\ \frac{\partial C}{\partial w_2}(w) \\ \vdots \\ \frac{\partial C}{\partial w_d}(w) \end{bmatrix}$$

$$H_{C(w)} = \begin{bmatrix} \frac{\partial^2 C}{\partial w_1^2}(w) & \cdots & \frac{\partial^2 C(w)}{\partial w_1 \partial w_d} \\ \frac{\partial^2 C}{\partial w_2 \partial w_1}(w) & & \\ \vdots & & \\ \frac{\partial^2 C}{\partial w_d \partial w_1}(w) & & \frac{\partial^2 C(w)}{\partial w_d^2} \end{bmatrix}$$

$$H_{C(w)}[i,j] = \frac{\partial C(w)}{\partial w_i \partial w_j}$$

$$\nabla C(w) = 0$$

$$C(w) \approx \hat{C}_0(w) = C(w_0) + \nabla C^T(w_0)(w - w_0) + \frac{1}{2}(w - w_0)^T H_{C(w_0)}(w - w_0)$$

$$\sum_{j=1}^{d} \frac{\partial C(w_0)}{\partial w_j}(w_j - w_0[j])$$

$$\|w - w_0\|^2_{H_{C(w_0)}}$$

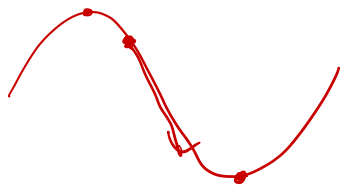
$$c(w) \approx \hat{c}_0(w) = c(w_0) + \nabla c(w_0)^T (w - w_0) + \frac{1}{2\eta^{-1}} (w - w_0)^T (w - w_0)$$

First order

Now $H_{c(w_0)} \approx \frac{1}{\eta} I$, $\eta > 0$

$$\Rightarrow 0 = \nabla \hat{c}_0(w) = \nabla c(w_0) + \eta^{-1}(w - w_0)$$

$$\Rightarrow w = w_0 - \eta \nabla c(w_0)$$



How pick η? Why do first order?

Cost of one iteration in 1st order (in terms of n and d)

$$c(w) = \sum_{i=1}^{n} c_i(w) \quad \text{e.g} \quad c_i(w) = \frac{1}{2}(x_i^T w - y_i)^2$$

$$\nabla c_i(w) = (x_i^T w - y_i) x_i$$

$O(d)$

Compute $\nabla c(w_0) = \sum_{i=1}^{n} \nabla c_i(w_0)$

costs $O(nd)$

Cost of 2nd-order:

$$w_t = w_{t-1} - H^{-1}_{c(w_{t-1})} \nabla c(w_{t-1})$$

costs $O(nd)$

$d \times d$

$$H_{c(w)} = \sum_{i=1}^{n} H_{c_i(w)}$$

Exercise: Show it costs $O(nd^2)$

Computing inverse: about $O(d^3)$

2nd-order Cost: $O(nd^2 + d^3)$

Superlinear, takes 2 iterations

First order: $O(nd)$

Linear, take 15 iterations

Big d, bad for 2nd order

Small d worth it to do 2nd order

Huge n ⟶ neither, we should do "stochastic gradient descent