

## ① The bias and variance decomposition:

$\text{مفتون} \rightarrow \text{متغير} \rightarrow \text{متغير} Y = f(x) + \epsilon$   $\rightarrow$   $Y = f(x) + \epsilon$   $\rightarrow$   $Y = f(x) + \epsilon$

$f(x) = B^T X$   $\rightarrow$   $Y = B^T X + \epsilon$   $\rightarrow$   $Y = \hat{f}(x) + \epsilon$

$$\Rightarrow Y = f(x) + \epsilon$$

$x$   $\rightarrow$   $\text{متغير} \rightarrow \text{متغير} \rightarrow \text{متغير}$

$\rightarrow$  random error term

$\hat{Y} \rightarrow$   $\hat{f}(x)$   $\rightarrow$   $\hat{f}(x) + \epsilon$   $\rightarrow$   $\hat{f}(x)$   $\rightarrow$   $\hat{f}(x) + \epsilon$

accuracy  $\hat{Y}$   $\rightarrow$  irreducible error  $\rightarrow$  relate to  $\epsilon$   $\rightarrow$   $\epsilon$   $\rightarrow$   $\epsilon$

reducible error: you can improve  $\hat{f}$  (e.g. different classifier)

$\hat{Y} = f(x) + \epsilon$   $\rightarrow$   $\hat{Y} = f(x) + \epsilon$   $\rightarrow$   $\hat{Y} = f(x) + \epsilon$

وجود دار راه تفاصیل می کند و این درین ویری های غیرقابل اطمینان بودن است.

هستد. اما خطای میتوانیم کاهش دهیم تغییر  $\hat{f}$  است. به عوامل مثل اشکال

مختلف میتوانیم تغییر داده سیم

حال آنکه از این طریق کاهش میتوانیم از تغییر خواهد گذاشت این اسیده است:

$$mse(\hat{Y}) = E(Y - \hat{Y})^2 = E[(f(x) + \epsilon - \hat{f}(x))^2]$$

$$= E(\hat{f}(x) - \hat{f}(x))^2 + E(\varepsilon^2) + \cancel{E(\varepsilon(f(x) - \hat{f}(x)))}$$

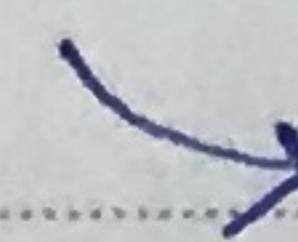
$$\text{var}(\varepsilon) = E(\varepsilon^2) - (E(\varepsilon))^2$$

$$= \underbrace{E(\varepsilon)}_{\text{error}} \underbrace{E(f(x) - \hat{f}(x))}_{\text{bias}}$$

$\text{error} = \text{Total error} - \text{bias}$

$$\Rightarrow E(f(x) - \hat{f}(x))^2 + \text{var}(\varepsilon)$$

أخطاء مترافق



أخطاء مترافق

$$\text{definition} \rightarrow \text{bias}(\hat{f}) = E(\hat{f}) - f$$

أخطاء مترافق في  $\hat{f}$  هي أخطاء مترافق في  $f$  وعوامل أخرى

$$\text{proposition: } \text{mse}(\hat{f}) = \text{bias}(\hat{f})^2 + \text{var}(\hat{f})$$

$$\bar{f} = E(\hat{f})$$

$$\text{mse}(\hat{f}) = E(\hat{f} - f)^2 = E(\hat{f} - \bar{f} + \bar{f} - f)^2$$

$$= E(\hat{f} - \bar{f})^2 + E(\bar{f} - f)^2 + \cancel{E(\hat{f} - \bar{f})(\bar{f} - f)}$$

$$\Rightarrow \text{mse}(\hat{f}) = \text{var}(\hat{f}) + \text{bias}(\hat{f})^2 + \cancel{\underbrace{E[(\hat{f} - \bar{f})(\bar{f} - f)]}}$$

$$* \cancel{E(\hat{f} - \bar{f}) E(\bar{f} - f)} = 0$$

$$\frac{E(\hat{f}) - E(\bar{f})}{(\bar{f} - f)} = 0$$

$$(\bar{f} - f) = 0$$

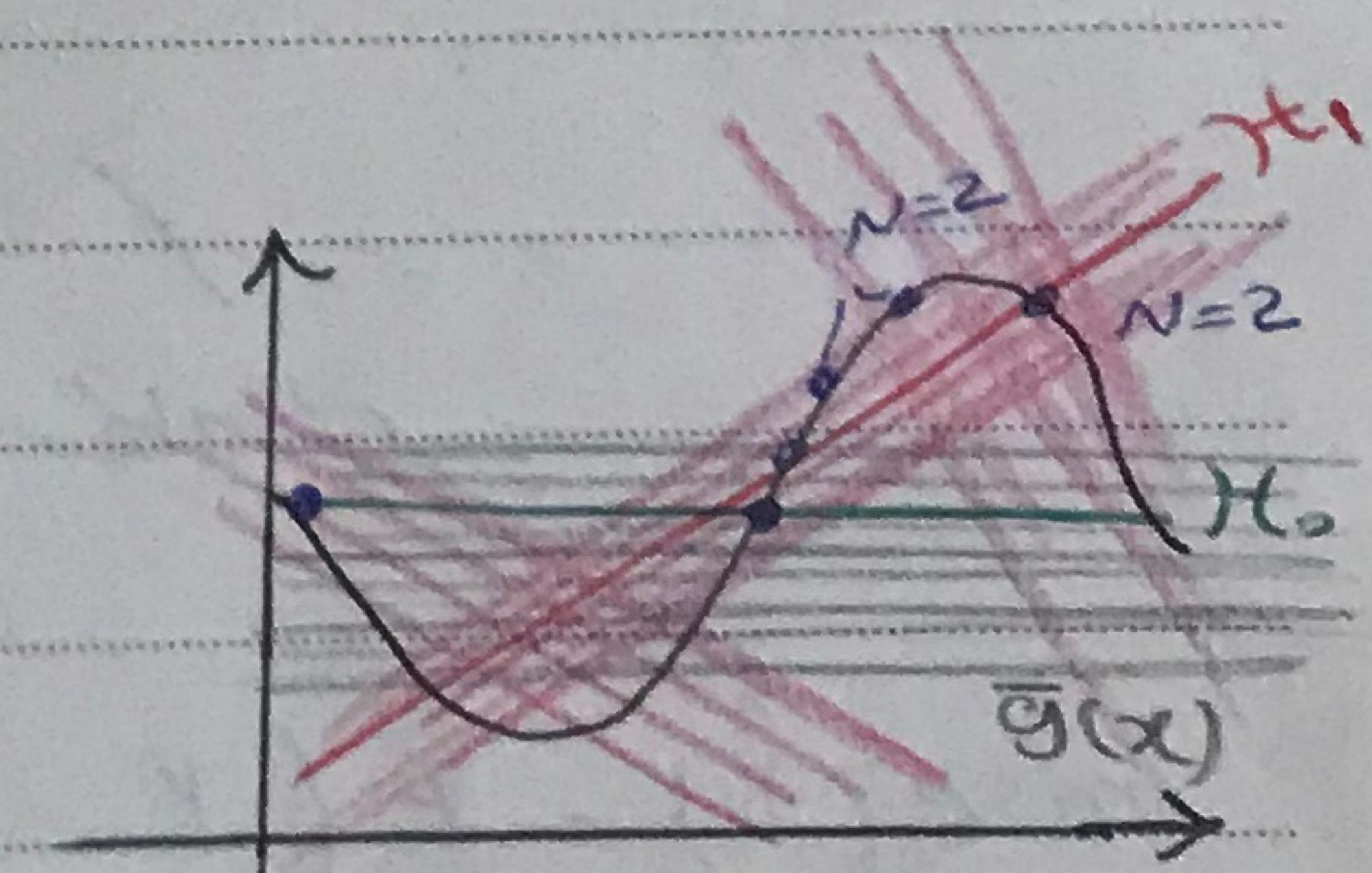
$$\Rightarrow \text{mse}(\hat{f}) = \text{var}(\hat{f}) + \text{bias}(\hat{f})^2$$

Example: Sine target

$$f: [-1, 1] \rightarrow \mathbb{R}$$

$$f(x) = \sin \pi x$$

$$N = 2$$



در دریل رایی پایه سیوال از این کرد:

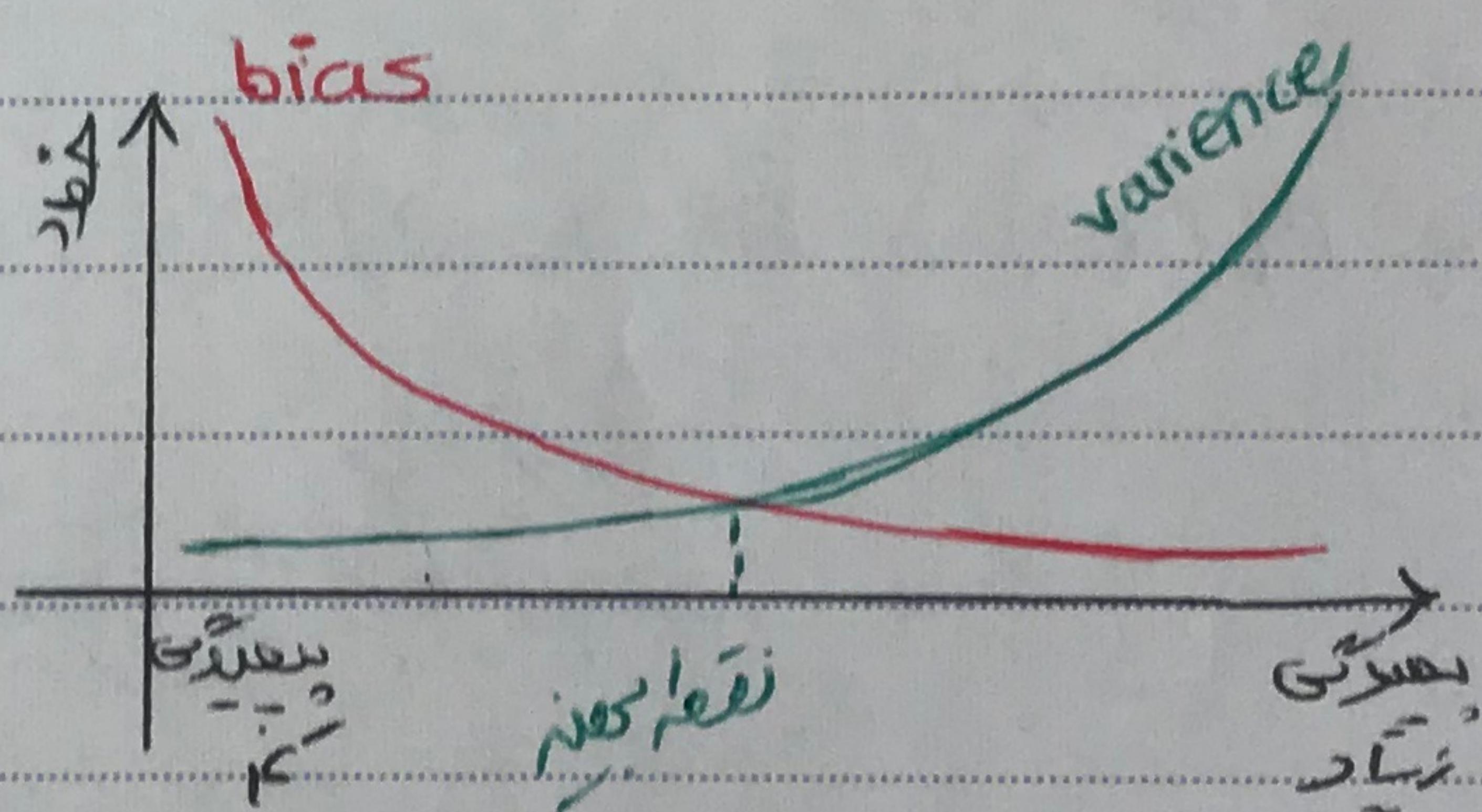
$$H_0: h(x) = b$$

$$H_1: h(x) = ax + b$$

سوال: آنچه عدل بمحض اینست:

بخطای وقتی در این مورد بالد bias آن زیاد است و در این مورد رگرسیون bias آن کم است.  
اگر شود، از طرف رگرسیون وقتی مدلی پیشنهاد بشه واریانس آن کم است و در این مورد مدلی بالد  
واریانس آن کم است می‌باشد. در مدل  $H_0$  مقدار bias زیاد است واریانس کم و در مدل  $H_1$   
مقدار واریانس زیاد است و bias کم. صفاتی اینکه در مدل زیاد است:

از پارامتر ببعیدیتی عدل، واریانس آن زیاد است و پارامتر ببعیدیتی bias زیاد است  
و زیاد است. هر دو باعث بالد نتیجه مقدار خطای آن می‌شود و باعث بیشتر مغقول بین این دو مقدار وجود  
علتی نداشت و انتظار از مدلها این است که ببعیدیتی زیادی را نشان داشته و نه خیلی کم نشان باشد.



حال وقتی ridge اسکوئر لینم این عدل حفاظتی را بعیدیتی regularize نمی‌کند رایی مثال از ridge اسکوئر لینم این عدل حفاظتی را بعیدیتی regularize نمی‌کند.

راستک متن و نتیجی را نمی‌نماییم (تفصیل شون) و در این عدل مدار که  $B_1$  را مدار  $B$  هایی به تسبیل می‌کند

زیاده را وارد می‌کند و این نتیجی واریانس را افزایش می‌دهد. از این طوره ridge و lasso

unbias estimator نماین linear regression و bias estimator

"MLE" : مکالمه احتمالی

Given  $D = \{x[1], x[2], \dots, x[n]\}$

جیسے دارم / این دیتاها در مرا برآمد خونگی اور جیسیت ہو رنپر بستے مکالمہ مانی خونگی کراس کریں

$D \sim P(x; \theta)$  or  $P_{\theta}(x)$  or  $P(x)$  تقریبی است / بر روی جیسیت کام اسٹ :

و ہدف این اسٹ  $\theta$  را تغییر بخوبیں

Example : W/P output

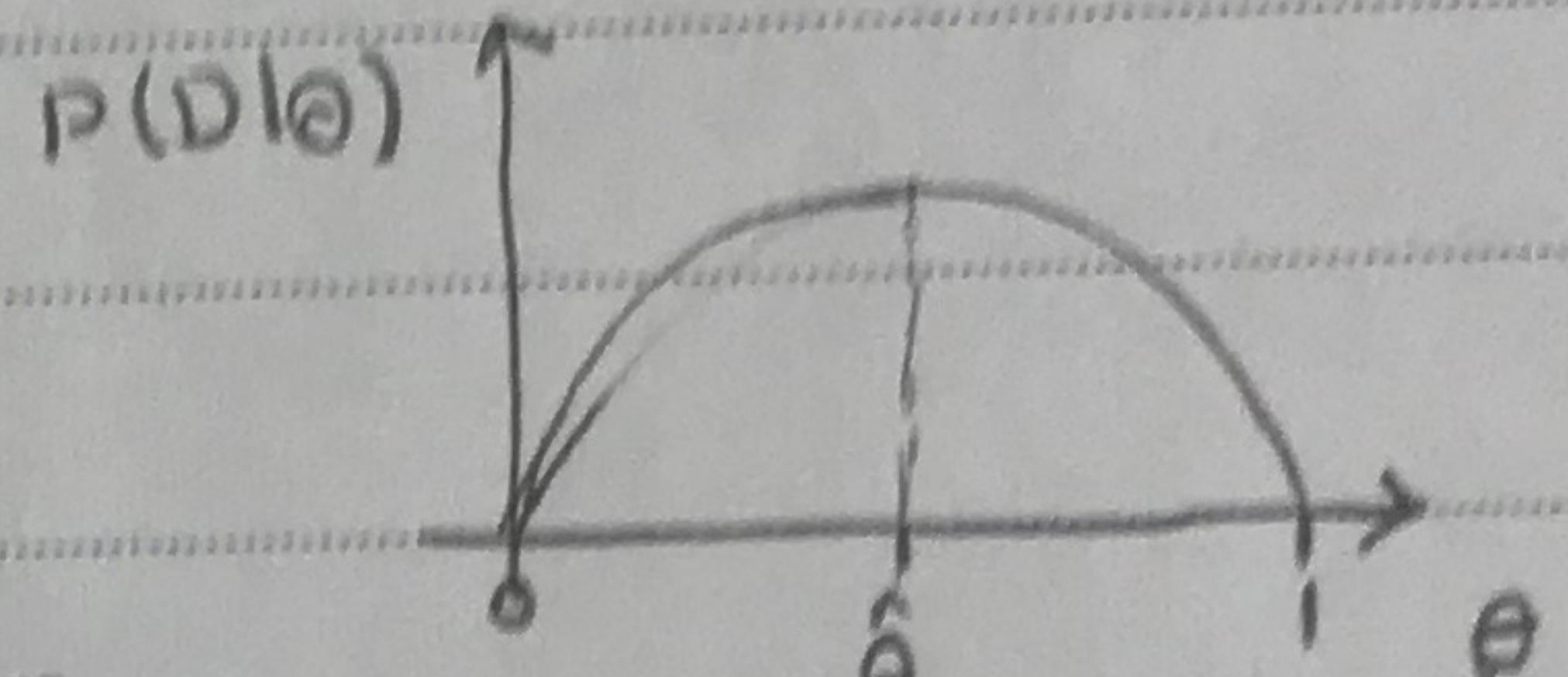
Data =  $\{W, P, P, W, W\}$   $P(X=W) = \theta$

$x[1], x[2], \dots, x[5]$

$L(\theta) = P(D|\theta) = P(W, P, P, W, W; \theta) = P(W)P(P)P(P)P(W)P(W)$

عواید  $\theta$

$$\Rightarrow P(D|\theta) = \theta^3(1-\theta)^2$$



یعنی اون  $\theta$  ایسا است کہ اعمال دقچ دیکارہ مانیں ہند.

$$\hat{\theta}_{ML} = \arg \max_{\theta} L(\theta)$$

Example : n coin ; m head  $L(\theta) = P(D|\theta) = \theta^m(1-\theta)^{n-m}$

$$\hat{\theta}_{ML} = \arg \max_{\theta} L(\theta) \Rightarrow \log L(\theta) = m \log \theta + (n-m) \log (1-\theta)$$

$$L'(\theta) = \frac{m}{\theta} - \frac{n-m}{1-\theta} \Rightarrow \hat{\theta}_{ML} = \frac{m}{n}$$

مشتق بازی صفر ہے

$$\frac{\partial^2 L}{\partial \theta^2} = -\frac{m}{\theta^2} - \frac{n-m}{(1-\theta)^2} < 0 \rightarrow \frac{m}{n} \text{ سے ایسا ہے}$$

Subject:

Year. Month. Date. ( )

## Example : MLE - For univariate Gaussian

$$X \sim N(\mu, \sigma^2)$$

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad \text{suppose } \sigma^2 \text{ is known}$$

we want estimate  $\hat{\mu}_{ML}$

$$D = \{x[1], x[2], \dots, x[n]\}$$

$$L(\mu) = P(D; \mu) \stackrel{iid}{=} \prod_{i=1}^n \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x[i] - \mu)^2\right)$$

$$\ell(\mu) = \log L(\mu) = \sum_{i=1}^n n \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x[i] - \mu)^2\right)$$

ستقتصر على مسأله مقدر واربطة

$$\ell'(\mu) = +\frac{1}{\sigma^2} \sum_{i=1}^n (x[i] - \mu) = \sum_{i=1}^n \frac{1}{\sigma^2} (x[i] - \mu) = 0$$

$$\Rightarrow \hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n x[i] = \bar{x}$$

$$\text{رسالة منطقية} \rightarrow \frac{\partial^2 \ell}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0$$

2

تحقيق مقدمة اقصى

## Linear regression - probabilistic view:

دليلاً لـ  $F(x)$  وعوامله باختلاف الرأي بايدرهاوس،

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon \xrightarrow{\text{error term}} \text{دليلاً لـ} \varepsilon \sim N(0, \sigma^2)$$

Subject:

Year. Month. Date. ( )

حال اگر ایک ایکس کا تغییر از  $F(x)$  راستہ باشد۔

: ایک پیسپاہ ایک ایکس و ایک بیسٹ کا سلسلہ کسی از جوین سے متفاہ نہیں کیا۔

$E \rightarrow$  random

$x \rightarrow$  suppose fix

$B \rightarrow$  not random

$y \rightarrow Y = F(x) + \epsilon$ , random  $\Rightarrow Y$  is random

$\hat{B} \rightarrow$  random

$\sigma^2 \rightarrow$  not random

حال اگر یہ فرم  $Y | X$  کا ایکہ نیم باقیتی توزع آن

$\text{Env}(0, \sigma^2)$

وہیانہ

یہ میتھے میانسی  $(F(x))$

. ایک

یہ تابع تابع اس کے میانسی خود

$\sigma$

$\Rightarrow$  Likelihood Function:

$$L(B) = P(D|B) = P(Y[1], Y[2], \dots, Y[n] | \vec{x}, B)$$

$$\xrightarrow{\text{iid}} \prod_{i=1}^n P(Y[i] | x[i], B) = \prod_{i=1}^n \left( \frac{1}{\sqrt{2\pi}\sigma} \right) e^{-\frac{1}{2\sigma^2} (y[i] - B^T x[i])^2}$$

$$= \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (y[i] - B^T x[i])^2 \right)$$

$$\ell(B) = \log L(B) = -n \log \sqrt{2\pi} \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (y[i] - B^T x[i])^2$$

فیصلہ

نبایان اگر معرفت نیم  $(\text{Env}(0, \sigma^2))$  کو دیں تو اس کا تجھے سارے کوئی تجھے سارے

$$\boxed{\frac{\partial}{\partial B} L = \hat{B}_{MSE} = (X^T X)^{-1} X^T Y}$$

تفنین ML یا باقیابانی مارکوف متغیر کوئنستیت  $\sigma^2$  بدانندگان آزاد است.

\* پیش‌روشنگی برای تفین ام به تفین بینی  $\hat{\theta}_{MAP}$   $\hat{\theta}_{ML}$

$\hookrightarrow$  maximum a posterior

$$D = \{x[1], x[2], \dots, x[n]\}^{e_{MAP}}$$

هدف عاریه  $P(\theta|D)$  است و تفین  $\theta$  برای روش گرایی است.

$$\hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta|D)$$

حال برای عاریه احتمال بین از جمله بزرگ استفاده می‌نماییم:

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\text{marginal likelihood}} \propto P(\theta)P(D|\theta)$$

Example: MAP For univariate Gaussian:

$$x_i | \mu \sim N(\mu, \sigma^2) \quad ; \text{ suppose } \sigma^2 \text{ fixed.}$$

$$\prod_{i=1}^n P(x[i] | \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x[i]-\mu)^2}{2\sigma^2}} \rightarrow P(D|\mu)$$

$$P(\mu) \sim N(\theta, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mu-\theta)^2}{2}}$$

$$\hat{\mu}_{MAP} = \arg \max_{\mu} P(\mu|D) \propto P(\mu)P(D|\mu)$$

or

$$\hat{\mu}_{MAP} = \arg \max_{\mu} \log P(\mu|D) \propto \log P(\mu) + \log P(D|\mu).$$

Subject:

Year. Month. Date. ( )

$$\Rightarrow \ell(P(\theta | D)) \propto \left( -\frac{(\mu - \theta)^2}{\sigma^2} \right) = \sum_{i=1}^n \frac{(x[i] - \mu)^2}{\sigma^2}$$

$$-\frac{1}{\sigma^2} \times n(\mu - \theta) + \frac{1}{\sigma^2} \sum_{i=1}^n (x[i] - \mu) = 0.$$

$$\Rightarrow -\mu + \theta + \frac{1}{\sigma^2} \sum_{i=1}^n x[i] - \frac{n\mu}{\sigma^2} = 0 \Rightarrow \mu(1 + \frac{n}{\sigma^2}) = \theta + \frac{1}{\sigma^2} \sum_{i=1}^n x[i]$$

$$\Rightarrow \hat{\mu} = \frac{\sigma^2}{n + \sigma^2} \theta + \frac{n}{n + \sigma^2} \bar{x}$$

$$\hat{\theta}_{MAP} = \frac{n}{n + \sigma^2} \bar{x} + \frac{\sigma^2}{n + \sigma^2} \theta$$

NLE  $\nwarrow$  Prior  $\searrow$

$$\text{if } n=0 \Rightarrow \hat{\theta}_{MAP} = \theta$$

$$\text{if } n=\infty \Rightarrow \hat{\theta}_{MAP} = \bar{x}$$

$$\theta \quad \hat{\theta}_{MAP} \quad \bar{x}$$

Combination

■ Bayesian interpretation of linear regression

## ② Ridge Regression

$$\hat{\beta}_{ridge} = \arg \min_{\beta} \left( \sum_{i=1}^n (\gamma[i] - \beta^T x[i])^2 + \lambda \sum_{i=1}^p \beta_i^2 \right)$$

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

↑  
↓

$$Y | X, \beta \sim N(\beta^T x, \sigma^2)$$

## MAP Estimate

Likelihood =  $P(y[1], \dots, y[n] | x, \beta) = \exp\left(-\frac{1}{\sigma^2} \sum_{i=1}^n (y[i] - \beta^T x[i])^2\right)$

Prior =  $P(\beta)$

$\beta$   $\sim$   $N(0, T^2)$  معرفی شده اند  $\beta_1, \dots, \beta_P$  دارای توزیع نرمال

$$\Rightarrow P(\beta) = P(\beta_1) P(\beta_2) \dots P(\beta_P) = \left(\frac{1}{\sqrt{\pi T^2}}\right)^P \times \exp\left(-\frac{1}{\sigma^2} \sum_{i=1}^P (\beta_i - 0)^2\right)$$

$$P(B|D) \propto P(B) P(D|B) \Rightarrow \log P(B|D) \propto \log P(B) + \log P(D|B)$$

$$= \left(-\frac{1}{\sigma^2} \sum_{i=1}^n (\beta_i)^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (y[i] - \beta^T x[i])^2\right)$$

$$\xrightarrow{x \sim N(0, T^2)} \left( \sum_{i=1}^n (y[i] - \beta^T x[i])^2 + \frac{\sigma^2}{T^2} \sum_{i=1}^P \beta_i^2 \right) \quad \begin{array}{l} \text{min-f(x)} \\ \text{max f(x)} \end{array}$$

$$= \underset{\beta}{\operatorname{arg\,min}} \left( \sum_{i=1}^n (y[i] - \beta^T x[i])^2 + \underbrace{\frac{\sigma^2}{T^2} \sum_{i=1}^P \beta_i^2}_{\lambda \text{ complexity}} \right)$$

و هدف حل مسئله ridge باشد اند  $\beta$  را تولید کنید و prior  $\beta$  را درست کنید

MAP estimate اور یعنی  $\hat{\beta}$  را برابر با  $\hat{\beta}_{LS}$  نمایی کنید

بروآفع در برخورد overfitting با MAP estimate