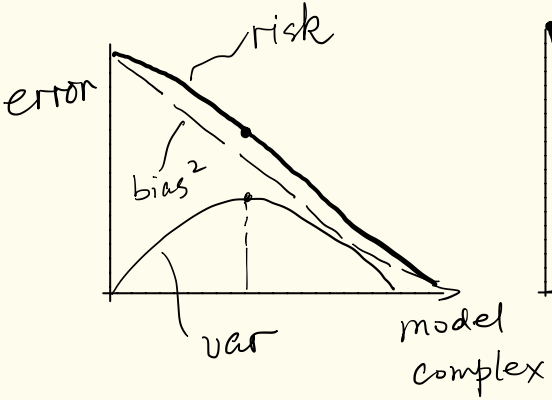
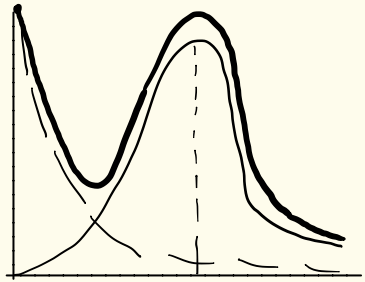
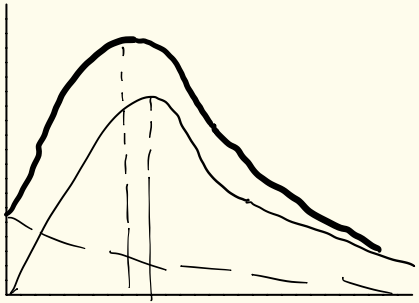
Bias-Variance Tradeoff for Deep Neural Networks

(Yang et al., ICML 2020)



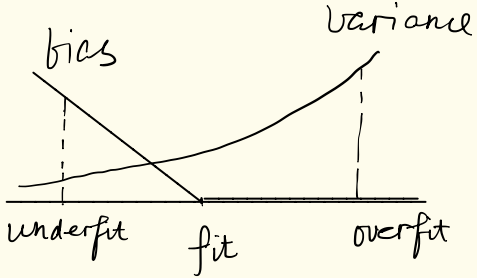




Bias-Variance Tradeoff (Traditional Statistical Learning Result)

Ex1 (Overfit & Underfit linear model)

1. Underfitting



True model:

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{e} = [X_1 \,|\, X_2] \begin{bmatrix} \underset{\sim}{\beta_1} \\ \underset{\sim}{\beta_2} \end{bmatrix} + \underset{\sim}{e}$$

Fitting / Assumed model: $\underset{\sim}{y} = X_1 \underset{\sim}{\beta_1} + \underset{\sim}{\varepsilon}$, w/ LS soln:

$$\hat{\underset{\sim}{\beta}}_1 = (X_1^T X_1)^{-1} X_1^T \underset{\sim}{y}$$

Bias: $E\left[\hat{\underset{\sim}{\beta}}_1\right] = (X_1^T X_1)^{-1} X_1^T \, [X_1 \; X_2] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

$$= (X_1^T X_1)^{-1} X_1^T (X_1 \beta_1 + X_2 \beta_2)$$

$$= \underset{\sim}{\beta_1} + (X_1^T X_1)^{-1} X_1^T X_2 \underset{\sim}{\beta_2} \quad \text{(biased)}$$

Variance: $\mathrm{Var}(\hat{\hat{\beta}}_1) = (X_1^T X_1)^{-1} X_1^T \sigma^2 I X_1 (X_1^T X_1)^{-1} = \sigma^2 (X_1^T X_1)^{-1} = B$

(has smaller var)

$$\mathrm{Var}(\hat{\beta}_{fit}) = \sigma^2 (X^T X)^{-1} = \sigma^2 \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix}^{-1} = \sigma^2 \begin{bmatrix} A & \\ & \end{bmatrix}$$

$$\begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

The upper corner is $\left[X_1^T X_1 - X_1^T X_2 (X_2^T X_2)^{-1} X_2^T X_1\right]^{-1} \overset{def}{=} A$

Block matrix inversion: $\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & x \\ x & x \end{bmatrix}$

We have $A \geq B$ or $A - B \geq 0$ or $v^T (A-B) v \geq 0, \forall v \in \mathbb{R}^n$

Def: M is positive semi-definite (PSD) if $x^T M x \geq 0, \forall x \in \mathbb{R}^n$ ($(M^T x)^T x$)

M ... positive definite (PD) if $x^T M x > 0, \forall x \neq 0$.

2. Overfitting:

True model: $\underset{\sim}{y} = X_1 \underset{\sim}{\beta}_1 + \underset{\sim}{e}$

Fitting/Assumed model: $\underset{\sim}{y} = \underbrace{[X_1 \,|\, X_2]}_{X} \begin{bmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{\beta}_2 \end{bmatrix} + \underset{\sim}{\varepsilon}$, w/

LS soln: $\hat{\underset{\sim}{\beta}} = (X^T X)^{-1} X^T \underset{\sim}{y}$

Bias: $E[\hat{\underset{\sim}{\beta}}] = (X^T X)^{-1} X^T (X_1 \underset{\sim}{\beta}_1 + X_2 \underset{\sim}{0}) = (X^T X)^{-1} X^T X \begin{bmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{0} \end{bmatrix}$

$= \begin{bmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{0} \end{bmatrix}$ (unbiased (for $\underset{\sim}{\beta}_1$) even overfit!)

Variance: $\mathrm{Var}(\hat{\underset{\sim}{\beta}}) = \sigma^2 (X^T X)^{-1} = \sigma^2 \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_1^T X_2 \end{bmatrix}^{-1}$ (has larger variance.)

$\mathrm{Var}(\hat{\underset{\sim}{\beta}}_{fit}) = \sigma^2 (X_1^T X_1)^{-1}$

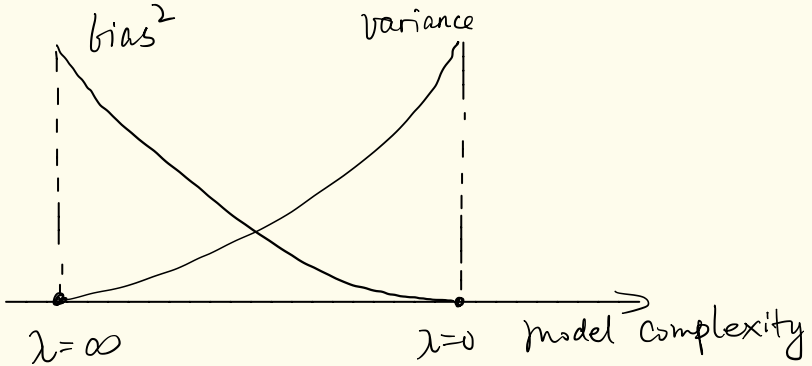
Ex 2 (Ridge Regression) $\hat{\beta} = \arg\min_{\beta} \left[\| y - X\beta \|_2^2 + \lambda \|\beta\|_2^2 \right]$.

$X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^{n \times 1}$, $\beta \in \mathbb{R}^{p \times 1}$, X full column rank, $\lambda \geq 0$

Assume true model: $y = X\beta + e$

Claims: 1. $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$ (HW)

2.



Bias: $E[\hat{\beta}] = E[(X^TX+\lambda I)^{-1}X^T y] = (X^TX+\lambda I)^{-1}X^TX\beta$

$R = V\Lambda V^T$

Eigen analysis Review

characteristic

$X^TX = R$ (symmetric)

$(X^TX)^T = X^TX = R^T$

1. $R = V\Lambda V^T$

$$= \left[\underset{\sim}{v}_1 \cdots \underset{\sim}{v}_p\right]\begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p\end{bmatrix}\begin{bmatrix}v_1^T \\ \vdots \\ v_p^T\end{bmatrix}$$

$$= \left[\lambda_1 \underset{\sim}{v}_1 \cdots \lambda_n \underset{\sim}{v}_p\right]\begin{bmatrix}v_1^T \\ \vdots \\ v_p^T\end{bmatrix}$$

$$= \sum_{i=1}^{p} \lambda_i \underset{\sim}{v}_i \underset{\sim}{v}_i^T$$

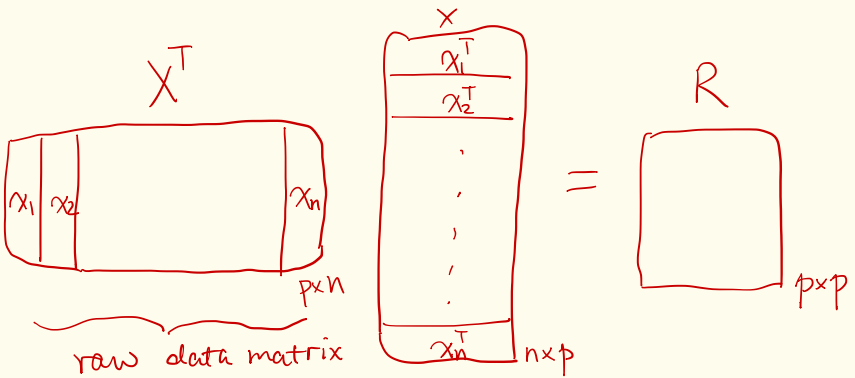
2. $RV = V\Lambda = [v_1 \cdots v_p]\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_p\end{bmatrix}$

3. $Rv_i = \lambda_i v_i$, $i = 1, \ldots, p$

V: orthogonal / orthonormal matrix:

① all vectors are orthogonal, has unit norm, i.e., $\underset{\sim}{v}_i^T \underset{\sim}{v}_j = \begin{cases} 0, & \forall i \neq j, \\ 1, & \forall i = j. \end{cases}$

② $V^T = V^{-1}$.



$$X^T X = R$$

$$R = X^T X = \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^T \quad : \text{ sample correlation matrix}$$

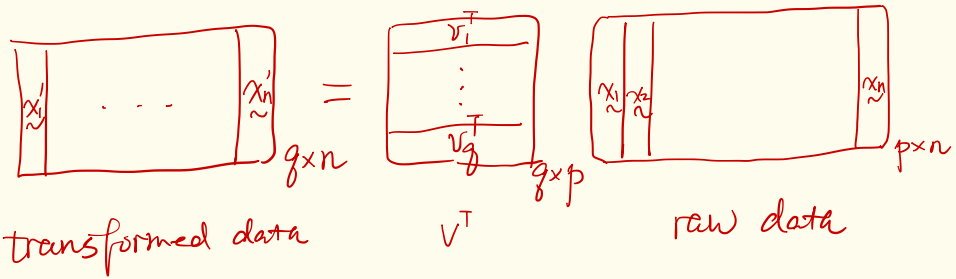
$\mathbb{E}[\underset{\sim}{x}\underset{\sim}{x}^T]$: correlation matrix

$$C = \sum_{i=1}^{n} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})^T \quad : \text{ sample covariance matrix}$$

$\mathbb{E}[(\underset{\sim}{x} - \mathbb{E}\underset{\sim}{x})(\underset{\sim}{x} - \mathbb{E}\underset{\sim}{x})^T]$: covariance matrix

Principal Component Analysis (PCA), usually $q \ll p$;

Karhunen-Loeve Transform (KLT), $q = p$.



Bias: $\mathbb{E}[\hat{\beta}] = (X^T X + \lambda I)^{-1} X^T X \beta$

$= (V \Lambda V^T + \lambda V V^T)^{-1} V \Lambda V^T \beta$

$= \left[V(\Lambda + \lambda I) V^T\right]^{-1} V \Lambda V^T \beta$

$= V (\Lambda + \lambda I)^{-1} V^T V \Lambda V^T \beta$

$= V (\Lambda + \lambda I)^{-1} \Lambda V^T \beta$

$= V \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\lambda} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\lambda_p}{\lambda_p+\lambda} \end{bmatrix} V^T \beta$

$= V \left(I - \begin{bmatrix} \frac{\lambda}{\lambda_1+\lambda} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\lambda}{\lambda_p+\lambda} \end{bmatrix} \right) V^T \beta$

$= \beta - \underbrace{\sum_{i=1}^{p} \frac{1}{\lambda_i/\lambda + 1} v_i v_i^T \beta}_{\neq 0} = \begin{cases} \beta, & \lambda = 0 \text{ (unbiased)} \\ 0, & \lambda = \infty \text{ (biased)} \end{cases}$

$\begin{bmatrix} \lambda_1+\lambda & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p+\lambda \end{bmatrix}_{p \times p}$

Variance: $Var(\hat{\beta}) = \underbrace{(X^TX+\lambda I)^{-1}X^T}\ \underbrace{Cov(y)}_{\sigma^2 I}\ X(X^TX+\lambda I)^{-1}$

$$= \underbrace{V(\Delta+\lambda I)^{-1}V^T}\ \underbrace{X^TX}_{V\Delta V^T}\ V(\Delta+\lambda I)^{-1}V^T\cdot\sigma^2$$

$$= V\begin{bmatrix} \frac{\lambda_1}{(\lambda_1+\lambda)^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\lambda_p}{(\lambda_p+\lambda)^2} \end{bmatrix} V^T\ \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+\lambda)^2}\ \underset{\sim}{v_i}\ \underset{\sim}{v_i}^T = \begin{cases} 0, & \lambda=\infty, \ (\text{zero variation}) \\ \sigma^2(X^TX)^{-1}, & \lambda=0. \ (\text{LS variance}) \end{cases}$$