\#

a. Regression:

$$y = x^T \beta^* + \xi$$

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_2^2 \right.$$

$$Y = X^T \beta^* + \xi$$

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(Y - X^T \beta \right)^2 + \lambda \|\beta\|^2 \Big).$$

$$f(\beta) = \underset{\beta}{\operatorname{argmin}} \left(Y - X^T \beta \right)^T \left(Y - X\beta \right) + \lambda \|\beta\|^2$$

$$\hat{\beta}(\lambda) \sim N\left(\beta, \ldots \text{Var}\right)$$

$$\frac{\partial f(\beta)}{\partial \beta} = -2X^T(Y - X\beta) + 2\lambda\beta$$

$$\boxed{\hat{\beta}_\lambda^{Ridge} = (X^T X + \lambda I_p)^{-1} X^T Y}$$

$$E(\hat{\beta}_\lambda^{ridge} | X) = E\{(I_p + \lambda X^T X)\hat{\beta}^{ls})$$

$$= (I_p + \lambda(X^T X)^{-1})\beta^*$$

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

$$\text{Var}[\hat{\beta}|X] = \sigma^2 (X^T X + \lambda I)^{-1}(X^T X)(X^T X + \lambda I)^{-1}$$

$$= x^T\beta^* - E_D[x^T\beta \ldots$$

$$= x^T\beta^* - E_D\left[x^T(X^TX + \lambda I_p)^{-1}X^TY\right]$$

$$\text{bias} = x^T\beta^* - E_D\left[x^T(X^TX + \lambda I_p)^{-1}X^T(X\beta^* + \xi)\right]$$

$$\text{bias} = x^T\beta^* - E_D\left[x^T(X^TX + \lambda I_p)^{-1}X^TX\beta^* + x^T(X^TX + \lambda I_p)^{-1}X^T\xi\right]$$

$$\text{bias} = x^T\beta^* - E_D\left[x^T(X^TX + \lambda I_p)^{-1}X^TX\beta^*\right.$$

$$\text{bias} = x^T\beta^* - E_D\left[x^T\beta^* + x^T(\lambda I_p)^{-1}X^TX\beta^* + x^T(X^TX)^{-1}X^T\xi + x^T(\lambda I_p)^{-1}X^T\xi\right.$$

$$- x^T(X^TX)^{-1}X^TE_D[\xi]$$

$$\text{bias} = \cancel{x^T\beta^*} - \cancel{x^T\beta^*} - x^T(X^TX)^{-1}X^TE_D[\xi] - x^T(\lambda I_p)^{-1}X^T(X\beta^* + \xi)]$$

$$- E_D\left[x^T(\lambda I_p)^{-1}X^TY\right]$$

$$\text{bias} = - x^T(X^TX)^{-1}X^TE_D[\xi] - E\left[x^T(\lambda I_p)^{-1}X^T(X\beta^* + E[\xi])\right.$$

$$\text{bias} = \underbrace{-x^T(X^TX)^{-1}X^TE_D[\xi]}_{\text{Constant with set sample } x \text{ and given } (X, Y)} - \underbrace{x^T(\lambda I_p)^{-1}X(X\beta^* + E[\xi])}_{f(\lambda) \text{ with } |f(\lambda)| \text{ proportional to } \lambda \text{ value}}$$

Variance

$$E\left[\left(x^T\hat{\beta}(\lambda) - E\left(x^T\hat{\beta}(\lambda)\right)\right)^2\right]$$

$$= E\left[\left(x^T\hat{\beta}(\lambda) - x^T\beta^*\right)^2\right]$$

$$= E\left[\left(x^T(X^TX + \lambda I_p)^{-1}X^TY - x^T\beta^*\right)^2\right]$$

$$= E\left[\left(x^T(X^TX)^{-1}X^TY + x^T(\lambda I_p)^{-1}X^TY - x^T\beta^*\right)^2\right]$$

$$= E\left[\left(x^T\beta^* + x^T(\lambda I_p)^{-1}X^TY - x^T\beta^* + x^T(X^TX)^{-1}X^T\varepsilon\right)^2\right]$$

$$= E\left[\left(x^T(\lambda I_p)^{-1}X^T(X^T\beta^* + \varepsilon) + x^T(X^TX)^{-1}X^T\varepsilon\right)^2\right]$$

$$= \cancel{E}\left[\sigma^2 x^T(X^TX + \lambda I_p)^{-1} x\right].$$

λ is high $\rightarrow$ bias is high, $\rightarrow$ variance is low
underfit

λ is low $\rightarrow$ bias is low, variance is high $\rightarrow$ overfit