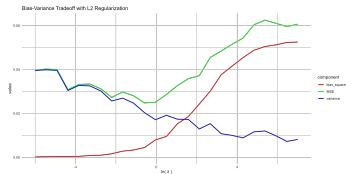


# Introduction to Machine Learning

## Regularization

## Bias-Variance Decomposition for Ridge Regression (Deep-Dive)



### Learning goals

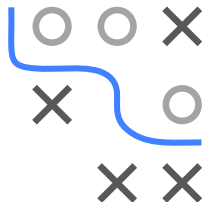
- Bias-Variance trade-off for ridge regression

# BIAS-VARIANCE DECOMPOSITION FOR RIDGE

For a linear model  $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \varepsilon$  with fixed design

$\mathbf{X} \in \mathbb{R}^{n \times p}$  and  $\varepsilon \sim (\mathbf{0}, \sigma^2 \mathbf{I}_n)$ , bias of ridge estimator  $\hat{\boldsymbol{\theta}}_{\text{ridge}}$  is given by

$$\begin{aligned}\text{Bias}(\hat{\boldsymbol{\theta}}_{\text{ridge}}) &:= \mathbb{E}[\hat{\boldsymbol{\theta}}_{\text{ridge}} - \boldsymbol{\theta}] = \mathbb{E}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}] - \boldsymbol{\theta} \\ &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\theta} + \varepsilon)] - \boldsymbol{\theta} \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \underbrace{\mathbb{E}[\varepsilon]}_{=0} - \boldsymbol{\theta} \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta} \\ &= \left[ (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1} \right] \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}\end{aligned}$$



- Last expression shows bias of ridge estimator only vanishes for  $\lambda = 0$ , which is simply (unbiased) OLS solution
- It follows  $\|\text{Bias}(\hat{\boldsymbol{\theta}}_{\text{ridge}})\|_2^2 > 0$  for all  $\lambda > 0$

# BIAS-VARIANCE IN PREDICTIONS FOR RIDGE

In supervised learning, our goal is typically not to learn an unknown parameter  $\theta$ , but to learn a function  $f(\mathbf{x})$  that can predict  $y$  given  $\mathbf{x}$ .

The bias and variance of predictions  $\hat{f} := \hat{f}(\mathbf{x}) = \hat{\theta}_{\text{ridge}}^\top \mathbf{x}$  is obtained as:

$$\begin{aligned}\text{Bias}(\hat{f}) &= \mathbb{E}[\hat{f} - f] = \mathbb{E}[\hat{\theta}_{\text{ridge}}^\top \mathbf{x} - \theta^\top \mathbf{x}] = \mathbb{E}[\hat{\theta}_{\text{ridge}} - \theta]^\top \mathbf{x} \\ &= \text{Bias}(\hat{\theta}_{\text{ridge}})^\top \mathbf{x} \\ \text{Var}(\hat{f}) &= \text{Var}(\hat{\theta}_{\text{ridge}}^\top \mathbf{x}) = \mathbf{x}^\top \text{Var}(\hat{\theta}_{\text{ridge}}) \mathbf{x}\end{aligned}$$

The MSE of  $\hat{f}$  given a fresh sample  $(y, \mathbf{x})$  can now be decomposed as

$$\text{MSE}(\hat{f}) = \mathbb{E}[(y - \hat{f}(\mathbf{x}))^2] = \text{Bias}^2(\hat{f}) + \text{Var}(\hat{f}) + \sigma^2$$

This decomposition is similar to the statistical inference setting before, however, the irreducible error  $\sigma^2$  only appears for predictions as an artifact of the noise in the test sample.

