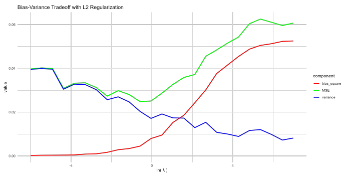
Introduction to Machine Learning

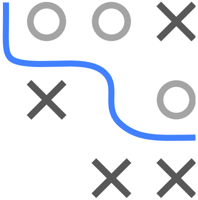
Regularization
Perspectives on Ridge Regression
(Deep-Dive)



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 $\varepsilon \sim (\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, bias of ridge estimator $\hat{\theta}_{\text{ridge}}$ is given by

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

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

BIAS-VARIANCE DECOMPOSITION FOR RIDGE / 2

For the variance of $\hat{\theta}_{\text{ridge}}$, we have

$$\begin{aligned}
\text{Var}(\hat{\theta}_{\text{ridge}}) &= \text{Var}\left((\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right) \qquad \Big| \text{ apply } \text{Var}_u(\boldsymbol{A}\boldsymbol{u}) = \boldsymbol{A}\text{Var}(\boldsymbol{u})\boldsymbol{A}^\top \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \text{Var}(\boldsymbol{y}) \left((\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top\right)^\top \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \text{Var}(\varepsilon) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \sigma^2 \boldsymbol{I}_n \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \\
&= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}
\end{aligned}$$

- $\text{Var}(\hat{\theta}_{\text{ridge}})$ is strictly smaller than $\text{Var}(\hat{\theta}_{\text{OLS}}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ for any $\lambda > 0$, meaning matrix of their difference $\text{Var}(\hat{\theta}_{\text{OLS}}) - \text{Var}(\hat{\theta}_{\text{ridge}})$ is positive definite (bit tedious derivation)
- This further means $\text{trace}\big(\text{Var}(\hat{\theta}_{\text{OLS}}) - \text{Var}(\hat{\theta}_{\text{ridge}})\big) > 0 \, \forall \lambda > 0$

BIAS-VARIANCE DECOMPOSITION FOR RIDGE / 3

Having obtained the bias and variance of the ridge estimator, we can decompose its mean squared error as follows:

$$\text{MSE}(\hat{\theta}_{\text{ridge}}) = \|\text{Bias}(\hat{\theta}_{\text{ridge}})\|_2^2 + \text{trace}\big(\text{Var}(\hat{\theta}_{\text{ridge}})\big)$$

Comparing MSEs of $\hat{\theta}_{\text{ridge}}$ and $\hat{\theta}_{\text{OLS}}$ and using $\text{Bias}(\hat{\theta}_{\text{OLS}}) = 0$ we find

$$\text{MSE}(\hat{\theta}_{\text{OLS}}) - \text{MSE}(\hat{\theta}_{\text{ridge}}) = \underbrace{\text{trace}\big(\text{Var}(\hat{\theta}_{\text{OLS}}) - \text{Var}(\hat{\theta}_{\text{ridge}})\big)}_{>0} - \underbrace{\|\text{Bias}(\hat{\theta}_{\text{ridge}})\|_2^2}_{>0}$$

Since both terms are positive, sign of their diff is *a priori* undetermined. ▸ Theobald 1974 and ▸ Farebrother 1976 prove there always exists some $\lambda^* > 0$ so that

$$\text{MSE}(\hat{\theta}_{\text{OLS}}) - \text{MSE}(\hat{\theta}_{\text{ridge}}) > 0$$

Important theoretical result: While Gauss-Markov guarantuees $\hat{\theta}_{\text{OLS}}$ is best linear unbiased estimator (BLUE), there are biased estimators with lower MSE.

BIAS-VARIANCE IN PREDICTIONS FOR RIDGE

In supervised learning, our goal is typically not to learn an unknown parameter $\boldsymbol{\theta}$, 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:

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

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 σ^2 only appears for predictions as an artifact of the noise in the test sample.