1 Bias and Variance of Ridge Regression

Ridge regression solves the regularized least squares problem:

$$\hat{\beta}_\tau = \arg\min_\beta (y - X\beta)^\top (y - X\beta) + \tau\beta^\top\beta$$

with regularization parameter $\tau \geq 0$.

Theorem 1. *Assume that the true model is $y = X\beta^* + \epsilon$ with zero-mean Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and centered features $\frac{1}{N}\sum_i X_i = 0$. Then the expectation and covariance matrix of the regularized solution are given by:*

$$\mathbb{E}[\hat{\beta}_\tau] = S_\tau^{-1} S\beta^*, \quad \mathrm{Cov}[\hat{\beta}_\tau] = S_\tau^{-1} S S_\tau^{-1}\sigma^2$$

where $S = X^\top X$ and $S_\tau = X^\top X + \tau I_D$ are the ordinary and regularized scatter matrices, respectively.

Proof. We will use the Singular Value Decomposition (SVD) of X.

Step 1: Express X using SVD:

$$X = U\Sigma V^\top$$

where U and V are orthogonal matrices and Σ is a diagonal matrix of singular values.

Step 2: Express scatter matrices in terms of SVD:

$$\begin{aligned} S &= X^\top X = V\Sigma^2 V^\top \\ S_\tau &= X^\top X + \tau I = V(\Sigma^2 + \tau I)V^\top \end{aligned}$$

Step 3: Derive closed-form solution for ridge regression:

$$\hat{\beta}_\tau = (X^\top X + \tau I)^{-1} X^\top y = S_\tau^{-1} X^\top y$$

Step 4: Substitute $y = X\beta^* + \epsilon$:

$$\hat{\beta}_\tau = S_\tau^{-1} X^\top (X\beta^* + \epsilon) = S_\tau^{-1} S\beta^* + S_\tau^{-1} X^\top \epsilon$$

Step 5: Calculate expectation:

$$\begin{aligned} \mathbb{E}[\hat{\beta}_\tau] &= \mathbb{E}[S_\tau^{-1} S\beta^* + S_\tau^{-1} X^\top \epsilon] \\ &= S_\tau^{-1} S\beta^* + S_\tau^{-1} X^\top \mathbb{E}[\epsilon] \\ &= S_\tau^{-1} S\beta^* \quad (\text{since } \mathbb{E}[\epsilon] = 0) \end{aligned}$$

Step 6: Calculate covariance:

$$\begin{aligned} \mathrm{Cov}[\hat{\beta}_\tau] &= \mathbb{E}[(\hat{\beta}_\tau - \mathbb{E}[\hat{\beta}_\tau])(\hat{\beta}_\tau - \mathbb{E}[\hat{\beta}_\tau])^\top] \\ &= \mathbb{E}[S_\tau^{-1} X^\top \epsilon\epsilon^\top X S_\tau^{-1}] \\ &= S_\tau^{-1} X^\top \mathbb{E}[\epsilon\epsilon^\top] X S_\tau^{-1} \\ &= S_\tau^{-1} X^\top (\sigma^2 I) X S_\tau^{-1} \\ &= \sigma^2 S_\tau^{-1} X^\top X S_\tau^{-1} \\ &= \sigma^2 S_\tau^{-1} S S_\tau^{-1} \end{aligned}$$

□

Note: When $\tau = 0$, these expressions reduce to ordinary least squares:

$$\mathbb{E}[\hat{\beta}_{\tau=0}] = \beta^*, \quad \mathrm{Cov}[\hat{\beta}_{\tau=0}] = S^{-1}\sigma^2$$

Since $S_\tau \succ S$ (in the positive definite sense), regularization has a shrinking effect on both expectation and covariance.

2 LDA - Derivation from the Least Squares Error

In the lecture, we derived LDA as a generative classifier that fits a Gaussian distribution to the data instances of each class. Assuming for simplicity that the data are centered (i.e., $\sum_{i=1}^{N} X_i = 0$) and the classes are balanced (i.e., $N_1 = N_{-1} = N/2$), this results in the decision rule:

$$\hat{y}_i = \text{sign}(X_i \cdot \hat{\beta}_{LDA}) \quad \text{with} \quad \hat{\beta}_{LDA} = \Sigma^{-1}(\mu_1 - \mu_{-1})^T$$

Here, μ_1 and μ_{-1} are the class means:

$$\mu_{-1} = \frac{1}{N_{-1}} \sum_{i:y_i=-1} X_i \quad \text{and} \quad \mu_1 = \frac{1}{N_1} \sum_{i:y_i=1} X_i$$

and Σ is the shared covariance matrix of the two clusters (also known as "within-class covariance"):

$$\Sigma = \frac{1}{N} \left[\sum_{i:y_i=-1} (X_i - \mu_{-1})^T (X_i - \mu_{-1}) + \sum_{i:y_i=1} (X_i - \mu_1)^T (X_i - \mu_1) \right]$$

Thanks to our simplifying assumptions, we don't have to deal with the intercept parameter, because $\hat{b} = 0$ under these conditions. Recall also that centering and balanced classes imply that $\mu_1 + \mu_{-1} = 0$.

Theorem 2. *An equivalent decision rule arises from minimizing the squared loss:*

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2 \implies \hat{\beta}_{OLS} = \tau \Sigma^{-1}(\mu_1 - \mu_{-1})^T$$

where $\tau > 0$ is a constant which doesn't alter the sign of $X_i \cdot \hat{\beta}_{OLS}$ and therefore leads to the same predictions $\hat{y}_i$.

Proof. To derive the expression for $\hat{\beta}_{OLS}$, we set the derivative of the loss with respect to β to zero:

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2 \stackrel{!}{=} 0 \tag{1}$$

After some algebraic manipulations, we arrive at the expression:

$$\Sigma \cdot \beta + \frac{1}{4}(\mu_1 - \mu_{-1})^T (\mu_1 - \mu_{-1}) \cdot \beta = \frac{1}{2}(\mu_1 - \mu_{-1})^T \tag{2}$$

The derivation steps from equation (1) to equation (2) are as follows:

1. Start with equation (1):
$$\frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2 = 0$$

2. Expand the squared term:
$$\frac{\partial}{\partial \beta} \sum_{i=1}^{N} \left(y_i^{*2} - 2y_i^* X_i \cdot \beta + (X_i \cdot \beta)^2 \right) = 0$$

3. Apply the derivative:
$$\sum_{i=1}^{N} (-2y_i^* X_i + 2(X_i \cdot \beta) X_i) = 0$$

4. Rearrange:
$$\sum_{i=1}^{N} (X_i \cdot \beta) X_i = \sum_{i=1}^{N} y_i^* X_i$$

5. Factor out β:
$$\left(\sum_{i=1}^{N} X_i X_i^T\right)\beta = \sum_{i=1}^{N} y_i^* X_i$$

6. Recognize that $\sum_{i=1}^{N} X_i X_i^T$ is the definition of Σ (scaled by N):
$$N\Sigma\beta = \sum_{i=1}^{N} y_i^* X_i$$

7. Split the sum based on y_i^* values:
$$N\Sigma\beta = \sum_{i:y_i^*=1} X_i - \sum_{i:y_i^*=-1} X_i$$

8. Use the definitions of μ_1 and μ_{-1}:
$$N\Sigma\beta = N_1\mu_1 - N_{-1}\mu_{-1}$$

9. Apply the balanced classes assumption ($N_1 = N_{-1} = N/2$):
$$N\Sigma\beta = \frac{N}{2}(\mu_1 - \mu_{-1})$$

10. Divide both sides by N:
$$\Sigma\beta = \frac{1}{2}(\mu_1 - \mu_{-1})$$

11. Add and subtract $\frac{1}{4}(\mu_1 - \mu_{-1})^T(\mu_1 - \mu_{-1})\beta$ to the left-hand side:
$$\Sigma\beta + \frac{1}{4}(\mu_1 - \mu_{-1})^T(\mu_1 - \mu_{-1})\beta = \frac{1}{2}(\mu_1 - \mu_{-1})^T$$

This gives us equation (2). By noticing that $(\mu_1 - \mu_{-1}) \cdot \beta = \tau'$ for some scalar τ', we can bring the second term of the left-hand side to the right-hand side and obtain the desired result:
$$\Sigma \cdot \beta = \left(\frac{1}{2} - \frac{\tau'}{4}\right)(\mu_1 - \mu_{-1})^T \implies \hat{\beta}_{OLS} = \tau\Sigma^{-1}(\mu_1 - \mu_{-1})^T$$
with $\tau = \frac{1}{2} - \frac{\tau'}{4}$. $\square$