Bias-Variance Decomposition in Ridge Linear Regression

Wen Sun

CS 4780, Fall 2022

1 Ridge Linear Regression with fixed Design

We consider the setting where examples \( \{x_i\}_{i=1}^n \) are fixed (i.e., no randomness on the features), while the regression target \( \{y_i\} \) could be random. We further assume that the regression targets \( y_i \) are generated in the following way:

\[
y_i = (w^*)^\top x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1),
\]

where \( \epsilon_i \) are i.i.d Gaussian noises. We can write everything using matrix and vectors. Denote

\[
X = \begin{bmatrix} x_1, \ldots, x_n \end{bmatrix} \in \mathbb{R}^{d \times n} \quad \text{and} \quad Y = \begin{bmatrix} y_1, \ldots, y_n \end{bmatrix}^\top \in \mathbb{R}^n, \quad \text{and} \quad \epsilon = \begin{bmatrix} \epsilon_1, \ldots, \epsilon_n \end{bmatrix}^\top \in \mathbb{R}^n,
\]

we have:

\[
Y = X^\top w^* + \epsilon.
\]

Ridge LR concerns the following optimization

\[
\hat{w} = \arg \min_w \| X^\top w - Y \|^2_2 + \lambda \| w \|^2_2.
\]

Recall the optimal solution here is

\[
\hat{w} = (XX^\top + \lambda I)^{-1} XY = (XX^\top + \lambda I)^{-1} X(X^\top w^* + \epsilon).
\]

So in this setting, we can think about our dataset \( D = \{x_i, y_i\}_{i=1}^n \) as follows \( D = \{x_i, (w^*)^\top x_i + \epsilon_i\}_{i=1}^n \). Note that the only randomness here is the Gaussian noise. In ML literature, this is called LR w/ fixed design.

We use the following generalization error we introduced in class to model the performance of \( \hat{w} \) from Ridge LR:

\[
\mathbb{E}_\epsilon \sum_{i=1}^n \left( \hat{w}^\top x_i - (w^*)^\top x_i \right)^2.
\]

Here the expectation is with respect to the randomness of the noises since \( \hat{w} \) depends on the noises — recall the dataset is random since it has random Gaussian noises. So we are looking at the squared difference between our prediction \( \hat{w}^\top x_i \) and the best one could get \( (w^*)^\top x_i \) (i.e., the Bayes optimal), summed over the fixed \( n \) examples \( \{x_1, \ldots, x_n\} \) (again in the fixed design setting, the examples \( x_i \) are always fixed, i.e., they are not sampled from some distribution).

2 Bias

In this section, we will derive a specific formulation for bias and show that it is monodically increasing wrt \( \lambda \).

First thing to recall is that \( \hat{w} \) depends on our dataset, i.e., \( \hat{w} = (XX^\top + \lambda I)^{-1} XY \). Since \( Y \) has random noises, \( \hat{w} \) will be a random quantity. So we can compute its expectation.

\[
\mathbb{E}_\epsilon[\hat{w}] = \mathbb{E}_\epsilon \left( XX^\top + \lambda I \right)^{-1} XY = \left( XX^\top + \lambda I \right)^{-1} X \mathbb{E}_\epsilon[Y]
\]

where we use the fact that \( X \) are fixed (i.e., this is the fixed design setting), and the expectation \( \mathbb{E}_\epsilon \) denoting the expectation with respect to the random noise \( \epsilon_i, i \in [1, \ldots, n] \).
Since $Y = X^\top w^* + \epsilon$, and $\mathbb{E}[\epsilon] = 0$, we get:

$$\mathbb{E}[\hat{w}] = \left(XX^\top + \lambda I\right)^{-1} XX^\top w^* = \left(XX^\top + \lambda I\right)^{-1} (XX^\top + \lambda I - \lambda I) w^*$$

$$= \left(XX^\top + \lambda I\right)^{-1} \left(XX^\top + \lambda I\right) w^* - \lambda \left(XX^\top + \lambda I\right)^{-1} w^*$$

$$= w^* - \lambda \left(XX^\top + \lambda I\right)^{-1} w^*.$$ 

Note that the above expression also shows that there is now no randomness in $\mathbb{E}[\hat{w}]$ anymore.

Now we define the bias as follows,

$$\text{bias} := \sum_{i=1}^n (\mathbb{E}[\hat{w}]^\top x_i - (w^*)^\top x_i)^2 = \sum_{i=1}^n ((\mathbb{E}[\hat{w}] - w^*)^\top x_i)^2$$

Since we have shown that $\mathbb{E}[\hat{w}] - w^* = -\lambda \left(XX^\top + \lambda I\right)^{-1} w^*$, plug in this into the Bias term, we get:

$$\text{bias} = \lambda^2 \sum_{i=1}^n (w^*)^\top \left(XX^\top + \lambda I\right)^{-1} x_i (w^*)^\top \left(XX^\top + \lambda I\right)^{-1} x_i$$

$$= \lambda^2 \sum_{i=1}^n (w^*)^\top \left(XX^\top + \lambda I\right)^{-1} \sum_{i=1}^n x_i x_i^\top \left(XX^\top + \lambda I\right)^{-1} (w^*)$$

$$= \lambda^2 (w^*)^\top \left(XX^\top + \lambda I\right)^{-1} XX^\top \left(XX^\top + \lambda I\right)^{-1} (w^*) \quad \text{(we used} \sum_{i=1}^n x_i x_i^\top = XX^\top)$$

Denote the eigendecomposition of $XX^\top$ as $XX^\top = U \Sigma U^\top$, where $\Sigma$ is a diagonal matrix $\text{diag}(\sigma_1, \ldots, \sigma_d)$, where $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_d \geq 0$, and $U$ are orthonormal matrices.

One fact is that for $XX^\top + \lambda I$, we can easily verify that its eigenvectors are columns of $U$, and its eigenvalues are $\sigma_i + \lambda$ for $i \in [1, \ldots, d]$, i.e., $XX^\top + \lambda I = U (\Sigma + \lambda I) U^\top$.

Using eigendecomposition, we can express the bias term using eigenvalues:

$$\text{bias} = \lambda^2 (w^*)^\top U (\Sigma + \lambda I)^{-1} U^\top U \Sigma U^\top U (\Sigma + \lambda I)^{-1} U^\top w^*$$

$$= \lambda^2 (w^*)^\top U (\Sigma + \lambda I)^{-1} \Sigma (\Sigma + \lambda I)^{-1} U^\top w^* \quad \text{we used} \quad U U^\top = U^\top U = I$$

$$= \lambda^2 (w^*)^\top U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1 + \lambda)^2} & 0 & \ldots & 0 \\ 0 & \frac{\sigma_2}{(\sigma_2 + \lambda)^2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0, & \ldots & \ldots & \frac{\sigma_d}{(\sigma_d + \lambda)^2} \end{bmatrix} U^\top w^*$$

$$= (w^*)^\top U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1 + \lambda)^2} & 0 & \ldots & 0 \\ 0 & \frac{\sigma_2}{(\sigma_2 + \lambda)^2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0, & \ldots & \ldots & \frac{\sigma_d}{(\sigma_d + \lambda)^2} \end{bmatrix} U^\top w^*$$

Ok, the above the form for Bias that we would like to analyze a bit.

**Case 1: when $\lambda \to 0$** In this case, we note that element in the diagonal matrix $\frac{\sigma_i}{(\sigma_i + \lambda)^2}$ go to 0. This means that our bias term will approach to zero as well. Namely, when $\lambda = 0$, we do not have bias.
Case 2: when $\lambda \to +\infty$. In this case, we get $\frac{\sigma_i}{(\sigma_i + \lambda)^2} \to \sigma_i$. This means that for expression we had for bias approaches to:

$$
\lim_{\lambda \to +\infty} \text{bias} = (w^*)^T U \Sigma U^T w^* = (w^*)^T X X^T w^* = \sum_{i=1}^{n} (x_i^T w^*)^2.
$$

This indeed makes a lot of sense since when $\lambda \to +\infty$, Ridge linear regression will return $\hat{\omega} \to 0$ which means that we always gonna predict zero, which in turn means that $\mathbb{E}_{\epsilon} \hat{\omega} \to 0$. So in this case, we have large bias.

Monotonicity of Bias Note that Bias is monotonically increasing as $\lambda$ increases.

3 Variance

Here we will give an explicit formulation for the variance and show that it is monodically decreasing.

Recall that $\hat{\omega}$ is a random vector and we have calculated its expectation as $\mathbb{E}_{\epsilon} \hat{\omega} = (XX^T + \lambda I)^{-1} XX^T w^*$. We abuse notation a little bit to write it as $\mathbb{E}[\hat{\omega}]$ below.

We define the form of variance as follows:

$$
\text{Var} := \mathbb{E}_{\epsilon} \left( (\mathbb{E}[\hat{\omega}] - \hat{\omega})^T x_i \right)^2 = \mathbb{E}_{\epsilon} \left( (\mathbb{E}[\hat{\omega}] - \hat{\omega})^T XX^T (\mathbb{E}[\hat{\omega}] - \hat{\omega}) \right)
$$

Here the expectation $\mathbb{E}_{\epsilon}$ is associated with the random vector $\hat{\omega}$ and we used the fact that $\sum_{i} x_i x_i^T = XX^T$ again. Denote $\text{tr}(A)$ as the trace of a matrix $A$. Recall that we have already had the formulation for both $\hat{\omega}$ and $\mathbb{E}[\hat{\omega}]$, so:

$$
\mathbb{E}[\hat{\omega}] - \hat{\omega} = (XX^T + \lambda I)^{-1} XX^T w^* - (XX^T + \lambda I)^{-1} X (X^T w^* + \epsilon)
$$

$$
= -(XX^T + \lambda I)^{-1} X \epsilon
$$

$$
\text{Var} = \mathbb{E}_{\epsilon} \left[ \epsilon^T X^T (XX^T + \lambda I)^{-1} XX^T (XX^T + \lambda I)^{-1} X \epsilon \right]
$$

$$
= \mathbb{E}_{\epsilon} \text{tr} \left( \epsilon^T X^T (XX^T + \lambda I)^{-1} XX^T (XX^T + \lambda I)^{-1} X \epsilon \right)
$$

$$
= \mathbb{E}_{\epsilon} \text{tr} \left( \epsilon \epsilon^T (XX^T + \lambda I)^{-1} XX^T (XX^T + \lambda I)^{-1} X \right)
$$

$$
= \text{tr} \left( XX^T (XX^T + \lambda I)^{-1} XX^T (XX^T + \lambda I)^{-1} X \right)
$$

$$
= \text{tr} \left( XX^T (XX^T + \lambda I)^{-1} XX^T (XX^T + \lambda I)^{-1} \right)
$$

Plug in the Eigendecomposition of $XX^T$ (and $XX^T + \lambda I$) into the above formulation, we get:

$$
\text{Var} = \text{tr} \left( U \Sigma U^T U (\Sigma + \lambda I)^{-1} U^T U \Sigma U^T U (\Sigma + \lambda I)^{-1} U^T \right)
$$

$$
= \text{tr} \left( U \Sigma U^T U (\Sigma + \lambda I)^{-1} U^T \right)
$$

$$
= \text{tr} \left( U^T U (\Sigma + \lambda I)^{-1} U \Sigma U^T U (\Sigma + \lambda I)^{-1} U^T \right)
$$

$$
= \text{tr} \left( (\Sigma + \lambda I)^{-1} U \Sigma (\Sigma + \lambda I)^{-1} U^T \right)
$$

$$
= \text{tr} \left( (\Sigma + \lambda I)^{-1} (\Sigma + \lambda I)^{-1} \right)
$$

$$
= \sum_{i=1}^{d} \frac{\sigma_i^2}{(\sigma_i + \lambda)^2},
$$
where the last equality uses the fact that $\Sigma(\Sigma + \lambda I)^{-1} \Sigma(\Sigma + \lambda I)^{-1}$ as a whole is a diagonal matrix with entries being $\sigma_i^2/(\sigma_i + \lambda)^2$.

**Case 1:** when $\lambda \to +\infty$. In this case we have $\sigma_i^2/(\sigma_i + \lambda)^2 \to 0$, which means that $\text{Var} \to 0$. This makes a lot of sense since when $\lambda \to +\infty$, we always have $\hat{\omega} \to 0$, which means that there is not too much randomness on $\hat{\omega}$ (it just converges to the zero vector in the limit).

**Case 2:** when $\lambda \to +0$. In this case, we have $\sigma_i^2/(\sigma_i + \lambda)^2 \to 1$, which means that $\text{Var} \to d$.

**Monotonicity of $\lambda$.** Note that when $\lambda$ increases, our variance decreases.

### 4 The Bias-Variance Decomposition

Now we can put everything together. For our ultimate generalization error, following what we did in class, we have:

$$E \epsilon_n \sum_{i=1}^{n} (\hat{w}^\top x_i - (w^*)^\top x_i)^2 = E \epsilon_n \sum_{i=1}^{n} (\hat{w}^\top x_i - E[\hat{w}]^\top x_i + E[\hat{w}]^\top x_i - (w^*)^\top x_i)^2$$

$$= \sum_i E \epsilon (\hat{w}^\top x_i - E[\hat{w}]^\top x_i)^2 + \sum_i E \epsilon (E[\hat{w}]^\top x_i - (w^*)^\top x_i)^2$$

$$= \text{Variance} + \text{Bias} = \sum_{i=1}^{d} \sigma_i^2/(\sigma_i + \lambda)^2 + (w^*)^\top U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1 + 1)^2} & 0 & \cdots \\ 0 & \frac{\sigma_2}{(\sigma_2 + 1)^2} & \cdots \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \frac{\sigma_d}{(\sigma_d + 1)^2} \end{bmatrix} U^\top w^*$$

**Q: why don’t we have the noise term here?**

Since Variance is monodically decreasing while Bias is monotonically increasing, there must exist a sweep spot for $\lambda$ that minimizes the sum of these two terms. The above formulation allows us in theory to calculate that (just take the derivative with respect to $\lambda$, set it to zero, and solve for $\lambda$). Of course in practice we cannot calculate this sweep spot for $\lambda$ since we do not know $w^*$ and $U$ and $\sigma_i$. So in practice, we use techniques like Cross Validation to select the best $\lambda$. 
