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:

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

where ϵ_i are i.i.d Gaussian noises. We can write everything using matrix and vectors. Denote $X = [x_1, \dots, x_n] \in \mathbb{R}^{d\times n}$ and $Y = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$, and $\epsilon = [\epsilon_1, \dots, \epsilon_n]^\top \in \mathbb{R}^n$, we have:

$$Y = X^\top w^\star + \epsilon.$$

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

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

So in this setting, we can think about our dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ as follows $\mathcal{D} = \{x_i, (w^\star)^\top x_i + \epsilon_i\}_{i=1}^n$. 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:

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

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 $(w^\star)^\top x_i$ (i.e., the Bayes optimal), summed over the fixed n examples $\{x_1, \dots, x_n\}$ (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 λ.

First thing to recall is that $\hat{w}$ depends on our dataset, i.e., $\hat{w} = \left(XX^\top + \lambda I\right)^{-1} XY$. 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 ϵ_i, $i \in [1, \dots, n]$.

Since $Y = X^\top w^\star + \epsilon$, and $\mathbb{E}_\epsilon[\epsilon] = 0$, we get:

$$\begin{aligned}\mathbb{E}_\epsilon[\hat{w}] &= \left(XX^\top + \lambda I\right)^{-1} XX^\top w^\star = \left(XX^\top + \lambda I\right)^{-1} (XX^\top + \lambda I - \lambda I) w^\star \\ &= \left(XX^\top + \lambda I\right)^{-1} \left(XX^\top + \lambda I\right) w^\star - \lambda \left(XX^\top + \lambda I\right)^{-1} w^\star \\ &= w^\star - \lambda \left(XX^\top + \lambda I\right)^{-1} w^\star.\end{aligned}$$

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

Now we define the bias as follows,

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

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

$$\begin{aligned}\text{bias} &= \sum_{i=1}^n \lambda^2 \left((w^\star)^\top \left(XX^\top + \lambda I\right)^{-1} x_i\right)^2 \\ &= \lambda^2 \sum_i (w^\star)^\top \left(XX^\top + \lambda I\right)^{-1} x_i x_i^\top \left(XX^\top + \lambda I\right)^{-1} (w^\star) \\ &= \lambda^2 (w^\star)^\top \left(XX^\top + \lambda I\right)^{-1} \sum_i [x_i x_i^\top] \left(XX^\top + \lambda I\right)^{-1} (w^\star) \\ &= \lambda^2 (w^\star)^\top \left(XX^\top + \lambda I\right)^{-1} XX^\top \left(XX^\top + \lambda I\right)^{-1} (w^\star) \qquad \text{(we used } \sum_i x_i x_i^\top = XX^\top\text{)}\end{aligned}$$

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

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 $i \in [1, \dots, d]$, i.e., $XX^\top + \lambda I = U(\Sigma + \lambda I)U^\top$.

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

$$\begin{aligned}\text{bias} &= \lambda^2 (w^\star)^\top U (\Sigma + \lambda I)^{-1} U^\top U \Sigma U^\top U (\Sigma + \lambda I)^{-1} U^\top w^\star \\ &= \lambda^2 (w^\star)^\top U (\Sigma + \lambda I)^{-1} \Sigma (\Sigma + \lambda I)^{-1} U^\top w^\star \qquad \text{we used } UU^\top = U^\top U = I \\ &= \lambda^2 (w^\star)^\top U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1+\lambda)^2} & 0 & 0\dots \\ 0 & \frac{\sigma_2}{(\sigma_2+\lambda)^2} & 0\dots \\ \dots & \dots & \dots \\ 0, & \dots & \frac{\sigma_d}{(\sigma_d+\lambda)^2} \end{bmatrix} U^\top w^\star \qquad \text{since } \Sigma \text{ and } \Sigma + \lambda I \text{ are diagonal} \\ &= (w^\star)^\top U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1/\lambda+1)^2} & 0 & 0\dots \\ 0 & \frac{\sigma_2}{(\sigma_2/\lambda+1)^2} & 0\dots \\ \dots & \dots & \dots \\ 0, & \dots & \frac{\sigma_d}{(\sigma_d/\lambda+1)^2} \end{bmatrix} U^\top w^\star\end{aligned}$$

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 $\frac{\sigma_i}{(\sigma_i/\lambda+1)^2} \to 0$. 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 $\frac{\sigma_i}{(\sigma_i/\lambda+1)^2} \to \sigma_i$. This means that for expression we had for bias approaches to:

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

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

Monotonicity of Bias Note that Bias is monotonically increasing as λ increases.

3 Variance

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

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

We define the form of variance as follows:

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

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

$$\begin{aligned}
\mathbb{E}[\hat{w}] - \hat{w} &= (XX^\top + \lambda I)^{-1} XX^\top w^\star - (XX^\top + \lambda I)^{-1} X (X^\top w^\star + \epsilon) \\
&= -(XX^\top + \lambda I)^{-1} X \epsilon
\end{aligned}$$

$$\begin{aligned}
\text{Var} &= \mathbb{E}_\epsilon \left[\epsilon^\top X^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} X \epsilon \right] \\
&= \mathbb{E}_\epsilon \operatorname{tr} \left(\epsilon^\top X^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} X \epsilon \right) \\
&= \mathbb{E}_\epsilon \operatorname{tr} \left(\epsilon \epsilon^\top X^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} X \right) && \text{fact: } \operatorname{tr}(AB) = \operatorname{tr}(BA) \\
&= \operatorname{tr} \left(\mathbb{E}_\epsilon[\epsilon \epsilon^\top] X^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} X \right) \\
&= \operatorname{tr} \left(X^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} X \right) && \text{since } \epsilon \sim \mathcal{N}(0, I_{n \times n}) \\
&= \operatorname{tr} \left(XX^\top (XX^\top + \lambda I)^{-1} XX^\top (XX^\top + \lambda I)^{-1} \right) && \text{fact: } \operatorname{tr}(AB) = \operatorname{tr}(BA)
\end{aligned}$$

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

$$\begin{aligned}
\text{Var} &= \operatorname{tr} \left(U \Sigma U^\top U (\Sigma + \lambda I)^{-1} U^\top U \Sigma U^\top U (\Sigma + \lambda I)^{-1} U^\top \right) \\
&= \operatorname{tr} \left(U \Sigma (\Sigma + \lambda I)^{-1} \Sigma (\Sigma + \lambda I)^{-1} U^\top \right) \\
&= \operatorname{tr} \left(U^\top U \Sigma (\Sigma + \lambda I)^{-1} \Sigma (\Sigma + \lambda I)^{-1} \right) && \text{fact: } \operatorname{tr}(AB) = \operatorname{tr}(BA) \\
&= \operatorname{tr} \left(\Sigma (\Sigma + \lambda I)^{-1} \Sigma (\Sigma + \lambda I)^{-1} \right) && \text{fact:} U^\top U = I \\
&= \sum_{i=1}^{d} \sigma_i^2 / (\sigma_i + \lambda)^2,
\end{aligned}$$

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 Var $\to 0$. This makes a lot of sense since when $\lambda \to +\infty$, we always have $\hat{w} \to \mathbf{0}$, which means that there is not too much randomness on $\hat{w}$ (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 Var $\to d$.

Monotonicity of λ Note that when λ 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:

$$
\begin{aligned}
&\mathbb{E}_\epsilon \sum_{i=1}^n \left(\hat{w}^\top x_i - (w^\star)^\top x_i\right)^2 = \mathbb{E}_\epsilon \sum_{i=1}^n \left(\hat{w}^\top x_i - \mathbb{E}[\hat{w}]^\top x_i + \mathbb{E}[\hat{w}]^\top x_i - (w^\star)^\top x_i\right)^2 \\
&= \sum_i \mathbb{E}_\epsilon \left(\hat{w}^\top x_i - \mathbb{E}[\hat{w}]^\top x_i\right)^2 + \sum_i \mathbb{E}_\epsilon \left(\mathbb{E}[\hat{w}]^\top x_i - (w^\star)^\top x_i\right)^2 \\
&= \text{Variance} + \text{Bias} = \sum_{i=1}^d \sigma_i^2/(\sigma_i + \lambda)^2 + (w^\star)^\top U \begin{bmatrix} \frac{\sigma_1}{(\sigma_1/\lambda+1)^2} & 0 & 0\dots \\ 0 & \frac{\sigma_2}{(\sigma_2/\lambda+1)^2} & 0\dots \\ \dots & \dots & \dots \\ 0, & \dots & \frac{\sigma_d}{(\sigma_d/\lambda+1)^2} \end{bmatrix} U^\top w^\star
\end{aligned}
$$

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 λ 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 λ, set it to zero, and solve for λ). Of course in practice we cannot calculate this sweep spot for λ since we do not know $w^\star$ and U and σ_i. So in practice, we use techniques like Cross Validation to select the best λ.