CS 189 Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk
DIS7

1 Simple Bias-Variance Tradeoff

Consider a random variable X, which has unknown mean μ and unknown variance σ^2. Given n i.i.d. realizations of training points $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ from the random variable, we wish to estimate the mean of X. We will call our estimate of X the random variable $\hat{X}$, which has mean $\hat{\mu}$. There are a few ways we can estimate μ given the realizations of the n samples:

1. Average the n sample points: $\dfrac{x_1 + x_2 + \ldots + x_n}{n}$.
2. Average the n sample points and one sample point of 0: $\dfrac{x_1 + x_2 + \ldots + x_n}{n+1}$.
3. Average the n sample points and n_0 sample points of 0: $\dfrac{x_1 + x_2 + \ldots + x_n}{n+n_0}$.
4. Ignore the sample points: just return 0.

In the parts of this question, we will measure the *bias* and *variance* of each of our estimators. The *bias* is defined to be

$$E[\hat{X} - \mu]$$

and the *variance* is defined to be

$$\mathrm{Var}[\hat{X}].$$

(a) What is the bias of each of the four estimators above?

Solution: Using the linearity of expectation, we write $E[\hat{X} - X]$ as $E[\hat{X}] - E[X] = E[\hat{X}] - \mu$, so we have the following biases.

(a) $E[\hat{X}] = E\left[\dfrac{X_1 + X_2 + \ldots + X_n}{n}\right] = \dfrac{n\mu}{n} \implies \text{bias} = 0$

(b) $E[\hat{X}] = E\left[\dfrac{X_1 + X_2 + \ldots + X_n}{n+1}\right] = \dfrac{n\mu}{n+1} \implies \text{bias} = -\dfrac{1}{n+1}\mu$

(c) $E[\hat{X}] = E\left[\dfrac{X_1 + X_2 + \ldots + X_n}{n+n_0}\right] = \dfrac{n\mu}{n+n_0} \implies \text{bias} = -\dfrac{n_0}{n+n_0}\mu$

(d) $E[\hat{X}] = 0 \implies \text{bias} = -\mu$

(b) What is the variance of each of the four estimators above?

Solution: The two key identities to remember are $\mathrm{Var}[A + B] = \mathrm{Var}[A] + \mathrm{Var}[B]$ (when A and B are independent) and $\mathrm{Var}[kA] = k^2\,\mathrm{Var}[A]$, where A and B are random variables and k is a constant.

(a) $\text{Var}[\hat{X}] = \text{Var}\left[\dfrac{X_1 + X_2 + \ldots + X_n}{n}\right] = \dfrac{1}{n^2}\text{Var}[X_1 + X_2 + \ldots + X_n] = \dfrac{1}{n^2}(n\sigma^2) = \dfrac{\sigma^2}{n}$

(b) $\text{Var}[\hat{X}] = \text{Var}\left[\dfrac{X_1 + X_2 + \ldots + X_n}{n+1}\right] = \dfrac{1}{(n+1)^2}\text{Var}[X_1 + X_2 + \ldots + X_n] = \dfrac{1}{(n+1)^2}(n\sigma^2) = \dfrac{n}{(n+1)^2}\sigma^2$

(c) $\text{Var}[\hat{X}] = \text{Var}\left[\dfrac{X_1 + X_2 + \ldots + X_n}{n+n_0}\right] = \dfrac{1}{(n+n_0)^2}\text{Var}[X_1 + X_2 + \ldots + X_n] = \dfrac{1}{(n+n_0)^2}(n\sigma^2) = \dfrac{n}{(n+n_0)^2}\sigma^2$

(d) $\text{Var}[\hat{X}] = 0$

(c) Suppose we have constructed an estimator $\hat{X}$ from some samples of X. We now want to know how well $\hat{X}$ estimates a fresh (new) sample of X. Denote this fresh sample by X'. Note that X' is an i.i.d. copy of the random variable X.

Derive a general expression for the expected squared error $E[(\hat{X} - X')^2]$ in terms of σ^2 and the bias and variance of the estimator $\hat{X}$. Similarly, derive an expression for the expected squared error $E[(\hat{X} - \mu)^2]$. Compare the two expressions and comment on the differences between them, if any.

Solution: Since $\hat{X}$ is a function of X, we conclude that the random variables $\hat{X}$ and X' are independent of each other. Now we provide two ways to solve the first problem.

Method 1: In this method, we use the trick of adding and subtracting a term to derive the desired expression:

$$\begin{aligned}
E[(\hat{X} - X')^2] &= E[(\hat{X} - \mu + \mu - X')^2] \\
&= E[(\hat{X} - \mu)^2 + \underbrace{E[(\mu - X')^2]}_{=\text{Var}(X')=\sigma^2} \\
&= E[(\hat{X} - \mu)^2 + \sigma^2 \\
&= E[(\hat{X} - E[\hat{X}] + E[\hat{X}] - \mu)^2 + \sigma^2 \\
&= \underbrace{E[(\hat{X} - E[\hat{X}])^2}_{=\text{Var}(\hat{X})} + \underbrace{(E[\hat{X}] - \mu)^2}_{=\text{bias}^2} + 2\underbrace{E[(\hat{X} - E[\hat{X}]) \cdot (E[\hat{X}] - \mu)]}_{=0} + \sigma^2
\end{aligned}$$

Method 2: In this method, we make use of the definition of variance. We have

$$\begin{aligned}
E[(\hat{X} - X')^2] &= E[\hat{X}^2] + E[X'^2] - 2E[\hat{X}X'] \\
&= (\text{Var}(\hat{X}) + (E[\hat{X}])^2) + (\text{Var}(X') + (E[X'])^2) - 2E[\hat{X}X'] \\
&= ((E[\hat{X}])^2 - 2E[\hat{X}X'] + (E[X'])^2) + \text{Var}(\hat{X}) + \underbrace{\text{Var}(X')}_{=\text{Var}(X)} \\
&= (E[\hat{X}] - \underbrace{E[X']}_{=E[X]=\mu})^2 + \text{Var}(\hat{X}) + \text{Var}(X)
\end{aligned}$$

$$= \underbrace{(E[\hat{X}] - \mu)^2}_{=\text{bias}^2} + \text{Var}(\hat{X}) + \sigma^2$$

The first term is equivalent to the bias of our estimator squared, the second term is the variance of the estimator, and the last term is the irreducible error.

Now let's do $E[(\hat{X} - \mu)^2]$.

$$\begin{aligned}
E[(\hat{X} - \mu)^2] &= E[\hat{X}^2] + E[\mu^2] - 2E[\hat{X}\mu] \qquad (1)\\
&= (\text{Var}(\hat{X}) + E[\hat{X}]^2) + (\text{Var}(\mu) + E[\mu]^2) - 2E[\hat{X}\mu] \qquad (2)\\
&= (E[\hat{X}]^2 - 2E[\hat{X}\mu] + E[\mu]^2) + \text{Var}(\hat{X}) + \text{Var}(\mu) \qquad (3)\\
&= (E[\hat{X}] - E[\mu])^2 + \text{Var}(\hat{X}) + \text{Var}(\mu) \qquad (4)\\
&= (E[\hat{X}] - \mu)^2 + \text{Var}(\hat{X}). \qquad (5)
\end{aligned}$$

Notice that these two expected squared errors resulted in the same expressions except for the σ^2 in $E[(\hat{X} - X')^2]$. The error σ^2 is considered "irreducible error" because it is associated with the noise that comes from sampling from the distribution of X. This term is not present in the second derivation because μ is a fixed value that we are trying to estimate.

(d) For the following parts, we will refer to expected total error as $E[(\hat{X} - \mu)^2]$. It is a common mistake to assume that an unbiased estimator is always "best." Let's explore this a bit further.

Compute the expected squared error for each of the estimators above.

Solution: Adding the previous two answers gives us

(a) $\dfrac{\sigma^2}{n}$

(b) $\dfrac{1}{(n+1)^2}(\mu^2 + n\sigma^2)$

(c) $\dfrac{1}{(n+n_0)^2}(n_0^2\mu^2 + n\sigma^2)$

(d) μ^2

(e) Demonstrate that the four estimators are each just special cases of the third estimator, but with different instantiations of the hyperparameter n_0.

Solution: The derivation for the third estimator works for *any* value of n_0. The first estimator is just the third estimator with n_0 set to 0:

$$\frac{x_1 + x_2 + \ldots + x_n}{n + n_0} = \frac{x_1 + x_2 + \ldots + x_n}{n + 0} + \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

The second estimator is just the third estimator with n_0 set to 1:

$$\frac{x_1 + x_2 + \ldots + x_n}{n + n_0} = \frac{x_1 + x_2 + \ldots + x_n}{n + 1}.$$

The last estimator is the limiting behavior as n_0 goes to ∞. In other words, we can get arbitrarily close to the fourth estimator by setting n_0 very large:

$$\lim_{n_0\to\infty} \frac{x_1 + x_2 + \ldots + x_n}{n + n_0} = 0.$$

(f) What happens to bias as n_0 increases? What happens to variance as n_0 increases?

Solution:

One reason for increasing the samples of n_0 is if you have reason to believe that X is centered around 0. In increasing the number of zeros we are injecting more confidence in our belief that the distribution is centered around zero. Consequently, in increasing the number of ”fake” data, the variance decreases because your distriubtion becomes more peaked. Examining the expressions for bias and variance for the third estimator, we can see that larger values of n_0 result in decreasing variance ($\frac{n}{(n+n_0)^2}\sigma^2$) but potentially increasing bias ($\frac{n_0\mu}{n+n_0}$). Hopefully you can see that there is a trade-off between bias and variance. Using an unbiased estimator is not always optimal nor is using an estimator with small variance always optimal. One has to carefully trade-off the two terms in order to obtain minimum squared error.

(g) Say that $n_0 = \alpha n$. Find the setting for α that would minimize the expected total error, assuming you secretly knew μ and σ. Your answer will depend on σ, μ, and n.

Solution: First, we write our expression for the total error in terms of α.

$$\frac{1}{(n+n_0)^2}(n_0^2\mu^2 + n\sigma^2)$$

$$\frac{1}{(n+\alpha n)^2}((\alpha n)^2\mu^2 + n\sigma^2)$$

$$\frac{1}{(1+\alpha)^2}\frac{1}{n^2}(\alpha^2 n^2\mu^2 + n\sigma^2)$$

$$\frac{1}{(1+\alpha)^2}(\alpha^2\mu^2 + \frac{\sigma^2}{n})$$

$$\frac{\alpha^2}{(1+\alpha)^2}\mu^2 + \frac{1}{(1+\alpha)^2}\frac{\sigma^2}{n}$$

Now take the derivative with respect to α and set it equal to 0.

$$\frac{2\alpha}{(1+\alpha)^3}\mu^2 - \frac{2}{(1+\alpha)^3}\frac{\sigma^2}{n} = 0.$$

$$\frac{2\alpha}{(1+\alpha)^3}\mu^2 = \frac{2}{(1+\alpha)^3}\frac{\sigma^2}{n}$$

$$2\alpha\mu^2 = 2\frac{\sigma^2}{n}$$

$$\alpha = \frac{\sigma^2}{n\mu^2}.$$

(h) For this part, let's assume that we had some reason to believe that μ *should be small* (close to 0) and σ *should be large*. In this case, what happens to the expression in the previous part?

Solution: The value of α can be quite large, since the solution has a small value of μ in the denominator. In mathematical terms, we could write the limit as μ goes to 0.

$$\lim_{\mu \to 0} \frac{\sigma^2}{n\mu^2} = \infty.$$

(i) In the previous part, we assumed there was reason to believe that μ *should be small*. Now let's assume that we have reason to believe that μ is not necessarily small, but *should be close to some fixed value* μ_0.

In terms of X and μ_0, how can we define a new random variable X' such that X' is expected to have a small mean? Compute the mean and variance of this new random variable.

Solution: Shift the random variable X by the constant guess μ_0 to get the random variable $X' = X - \mu_0$. Let's calculate the mean and variance of this new random variable.

$$E[X'] = E[X - \mu_0] = E[X] - \mu_0 = \mu - \mu_0 \approx 0.$$

The last line ($\mu - \mu_0 \approx 0$) comes from the assumption that μ is close to μ_0.

We can also calculate the variance of X'.

$$\text{Var}[X'] = \text{Var}[X - \mu_0] = \text{Var}[X] - \text{Var}\mu_0 = \text{Var}[X] - 0 = \text{Var}[X].$$

This is a useful step to understand the relation between X and X', but not necessary for a full solution to the question asked.

(j) Draw a connection between α in this problem and the regularization parameter λ in the ridge-regression version of least-squares.

What does this problem suggest about choosing a regularization coefficient and handling our data-sets so that regularization is most effective? This is an open-ended question, so do not get too hung up on it.

Solution: The key lesson is another reminder that regularization reduces variance, at the cost of increasing bias, by forcing solutions towards zero. But the bias-variance trade-off is not always the same. If we first center our data around some prior, so that the model is supposed to be close to zero anyways, then we can use larger values of α or λ and reduce variance considerably for a small cost in bias. It may also be instructive to realize that regularization can be thought of as adding "fake" training data which is uniformly zero.

2 The Ridge Regression Estimator

Recall the ridge regression estimator for $\lambda > 0$,

$$\widehat{\theta}_\lambda := \arg\min_\theta \|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2.$$

Let

$$X = UDV^\top = \sum_i d_i u_i v_i^\top$$

be the singular value decomposition of X. Here U and V are orthogonal matrices, meaning that $U^\top U = I$ and $V^\top V = I$. D is a diagonal matrix.

(a) Show that the optimal weight vector $\widehat{\theta}_\lambda$ can be expressed in the form

$$\widehat{\theta}_\lambda = V\Sigma U^\top y$$

where Σ is a diagonal matrix with $\Sigma_{ii} = \frac{d_i}{d_i^2 + \lambda}$. Equivalently, we can write $\widehat{\theta}_\lambda$ as

$$\widehat{\theta}_\lambda = \sum_{i=1}^d \frac{d_i}{d_i^2 + \lambda} v_i \langle u_i, y\rangle.$$

Solution: By taking the gradient of the objective, we have that $\widehat{\theta}_\lambda$ has to satisfy

$$X^\top(X\widehat{\theta}_\lambda - y) + \lambda\widehat{\theta}_\lambda = 0,$$

so $\widehat{\theta}_\lambda = (X^\top X + \lambda I)^{-1}X^\top y$. In terms of the SVD of X, this expression is equal to

$$\widehat{\theta}_\lambda = (VDU^\top UDV^\top + \lambda I)^{-1}VDU^\top y = V(D^2 + \lambda I)^{-1}V^\top VDU^\top = V\Sigma U^\top y,$$

where Σ is a diagonal matrix with $\Sigma_{ii} = \frac{d_i}{d_i^2 + \lambda}$. Equivalently, we can write this as

$$\widehat{\theta}_\lambda = \sum_{i=1}^d \frac{d_i}{d_i^2 + \lambda} v_i \langle u_i, y\rangle.$$

(b) Show that

$$\|\widehat{\theta}_\lambda\|_2^2 = \sum_{i:d_i>0} \left(\frac{d_i}{d_i^2 + \lambda}\right)^2 (u_i^\top y)^2.$$

Solution: First, we have that

$$\|\widehat{\theta}_\lambda\|_2^2 = y^\top U\Sigma V^\top V\Sigma U^\top y$$

$$
\begin{aligned}
&= y^\top U \Sigma^2 U^\top y \\
&= \sum_{1 \le i \le d} \left(\frac{d_i}{d_i^2 + \lambda} \right)^2 \langle u_i, y \rangle^2 \\
&= \sum_{i : d_i > 0} \left(\frac{d_i}{d_i^2 + \lambda} \right)^2 \langle u_i, y \rangle^2 .
\end{aligned}
$$

(c) Recall the least-norm least-squares solution is $\widehat{\theta}_{LN,LS}$ from Discussion Section 6. Show that if $\widehat{\theta}_{LN,LS} = 0$, then $\widehat{\theta}_\lambda = 0$ for all $\lambda > 0$. *Hint*: Recall that in Discussion 6 we showed that $\widehat{\theta}_{LN,LS} = \sum_{i:d_i>0} d_i^{-1} \langle u_i, y \rangle v_i$. This shows that in the case where the least-norm least square solution is zero, the ridge regression solution is also zero.

Solution: If the least-norm least-squares solution is 0, then

$$\sum_{i:d_i>0} d_i^{-1} \langle u_i, y \rangle v_i = 0,$$

which means that $\langle u_i, y \rangle = 0$ for each i where $d_i > 0$, because the v_i's are linearly independent. Hence, $\widehat{\theta}_\lambda = 0$ by plugging into the formula for $\widehat{\theta}_\lambda = 0$.

(d) Show that if $\widehat{\theta}_{LN,LS} \neq 0$, then the function $f(\lambda) = \|\widehat{\theta}_\lambda\|_2^2$ is strictly decreasing and strictly positive on $(0, +\infty)$.

Solution: If $\widehat{\theta}_\lambda \neq 0$, then at least one of the terms $\langle u_i, y \rangle^2$ is strictly greater than zero. Thus,

$$\|\widehat{\theta}_\lambda\|_2^2 = \sum_{i:d_i>0} \left(\frac{d_i}{d_i^2 + \lambda} \right)^2 \langle u_i, y \rangle^2 ,$$

is a nonnegative linear combination of terms $\left(\frac{d_i}{d_i^2 + \lambda} \right)^2$, which are positive and strictly decreasing in λ.

(e) Show that

$$\lim_{\lambda \to 0^+} \widehat{\theta}_\lambda \to \widehat{\theta}_{LN,LS} .$$

Note that just because the limit of the ridge-regression objective as $\lambda \to 0^+$ is the least-squares objective, this does not immediately guarantee that the limit of the ridge solution is the least-squares solution.

Solution: Start with the form

$$\widehat{\theta}_\lambda = \sum_{i=1}^{d} \frac{d_i}{d_i^2 + \lambda} v_i \langle u_i, y \rangle .$$

Since limits commute with sums, we have

$$\lim_{\lambda\to 0^+}\widehat{\theta}_\lambda = \sum_{i=1}^{n} v_i\langle u_i, y\rangle \cdot \left(\lim_{\lambda\to 0^+}\frac{d_i}{d_i^2+\lambda}\right)$$

Now, we have to consider the cases where $d_i = 0$ and $d_i \geq 0$. (The singular values d_i cannot be negative).

$$\lim_{\lambda\to 0^+}\frac{d_i}{d_i^2+\lambda} = \begin{cases} 0 & d_i = 0, \\ d_i^{-1} & d_i > 0. \end{cases}.$$

Thus,

$$\lim_{\lambda\to 0^+}\widehat{\theta}_\lambda = \sum_{i:d_i>0} d_i^{-1} v_i\langle u_i, y\rangle,$$

which we have shown above is the least-norm solution.

(f) In light of the above, why do you think that people describe the ridge regression as "controlling the complexity" of the solution $\widehat{\theta}_\lambda$?

Solution: We see that increasing the ridge parameter λ shrinks the norm of $\widehat{\theta}_\lambda$, and that even as $\lambda \to 0^+$, $\widehat{\theta}_\lambda$ picks out the least-norm least-squares solution.

In addition we know that adding a ridge parameter helps lower the variance of our model (in exchange for bias). High variance is caused by d_i being very close to 0, which causes $\frac{1}{d_i}$ to be large and highly variable depending on our data; small shifts in d_i cause drastic shifts in its reciprocal. With a ridge parameter we see the $\frac{1}{d_i}$ in the summation is replaced by $\frac{d_i}{d_i^2+\lambda}$, where λ is not close to 0. This new fraction becomes close to 0 when d_i is close to 0 and is much more stable; minor shifts in d_i won't cause this new value to vary drastically. This helps reduce unstable, high variance "complex" weights.

3 The Bias-Variance Tradeoff for Ridge Regression

Recall the statistical model for ridge regression from lecture. We have a set of sample points $\{x_i, y_i\}_{i=1}^n$ and Gaussian noise z_i. Our model follows, where the rows of X are x_i.

$$Y = Xw^* + z$$

Throughout this problem, you may assume $X^\top X$ is invertible. Recall both least-squares estimators we studied.

$$w_{\text{ols}} = \min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2$$

$$w_{\text{ridge}} = \min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_2^2$$

(a) Write the solution for $w_{\text{ols}}, w_{\text{ridge}}$. No need to derive it.

Solution:

$$w_{\text{ols}} = (X^\top X)^{-1} X^\top y$$

$$w_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top y$$

(b) Let $\widehat{w} \in \mathbb{R}^d$ denote any estimator of w_*. In the context of this problem, an estimator $\widehat{w} = \widehat{w}(X, Y)$ is any function which takes the data X and a realization of Y, and computes a guess of w_*.

Define the MSE (mean squared error) of the estimator $\widehat{w}$ as

$$\text{MSE}(\widehat{w}) := E\left\|\widehat{w} - w_*\right\|_2^2 .$$

Above, the expectation is taken with respect to the randomness inherent in z. Define $\widehat{\mu} := E\widehat{w}$. Show that the MSE decomposes as

$$\text{MSE}(\widehat{w}) = \left\|\widehat{\mu} - w_*\right\|_2^2 + \text{Tr}(\text{Cov}(\widehat{w})) .$$

Hint: Expectation and trace commute, so $E[\text{Tr}(A)] = \text{Tr}(E[A])$ for any square matrix A.

Solution:

$$\begin{aligned}
E\left\|\widehat{w} - w_*\right\|_2^2 &= E\left\|(\widehat{w} - \widehat{\mu}) - (w_* - \widehat{\mu})\right\|_2^2 \\
&= E\left\|\widehat{w} - \widehat{\mu}\right\|^2 - 2E(\widehat{w} - \widehat{\mu})(w_* - \widehat{\mu}) + E\left\|w_* - \widehat{\mu}\right\|_2^2 \\
&= E\left\|\widehat{w} - \widehat{\mu}\right\|^2 + \left\|w_* - \widehat{\mu}\right\|_2^2 \\
&= E\text{Tr}((\widehat{w} - \widehat{\mu})(\widehat{w} - \widehat{\mu})^\top) + \left\|w_* - \widehat{\mu}\right\|_2^2 \\
&= \text{Tr}(E(\widehat{w} - \widehat{\mu})(\widehat{w} - \widehat{\mu})^\top) + \left\|w_* - \widehat{\mu}\right\|_2^2 \\
&= \text{Tr}(\text{Cov}(\widehat{w})) + \left\|w_* - \widehat{\mu}\right\|_2^2 .
\end{aligned}$$

(c) Show that

$$E[w_{\text{ols}}] = w_*, \qquad E[w_{\text{ridge}}] = (X^\top X + \lambda I_d)^{-1} X^\top X w_* \,.$$

That is, w_{ols} is an *unbiased* estimator of w_*, whereas w_{ridge} is a *biased* estimator of w_*.

Solution: For OLS,

$$\begin{aligned} w_{\text{ols}} &= (X^\top X)^{-1} X^\top Y \\ &= (X^\top X)^{-1} X^\top (X w_* + z) \\ &= w_* + (X^\top X)^{-1} X^\top z \,. \end{aligned}$$

Hence, since $E[z] = 0$, $E[w_{\text{ols}}] = w_*$.

Similarly,

$$\begin{aligned} w_{\text{ridge}} &= (X^\top X + \lambda I_d)^{-1} X^\top Y \\ &= (X^\top X + \lambda I_d)^{-1} X^\top (X w_* + z) \\ &= (X^\top X + \lambda I_d)^{-1} X^\top X w_* + (X^\top X + \lambda I_d)^{-1} X^\top z \,, \end{aligned}$$

and therefore $E[w_{\text{ridge}}] = (X^\top X + \lambda I_d)^{-1} X^\top X w_*$.

(d) Let $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_d$ denote the d eigenvalues of the matrix $X^\top X$ arranged in non-increasing order. First, argue that the smallest eigenvalue, γ_d, is positive (i.e. $\gamma_d > 0$). Then, show that

$$\text{Tr}(\text{Cov}(w_{\text{ols}})) = \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i}, \qquad \text{Tr}(\text{Cov}(w_{\text{ridge}})) = \sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i + \lambda)^2} \,.$$

Finally, use these formulas to conclude that

$$\text{Tr}(\text{Cov}(w_{\text{ridge}})) < \text{Tr}(\text{Cov}(w_{\text{ols}})) \,.$$

Hint: For the ridge variance, consider writing $X^\top X$ in terms of its eigendecomposition $U \Sigma U^\top$.

Solution: For OLS, we simply compute

$$\begin{aligned} \text{Tr}(E(X^\top X)^{-1} X^\top z z^\top X (X^\top X)^{-1}) &= \sigma^2 \text{Tr}((X^\top X)^{-1} X^\top X (X^\top X)^{-1}) \\ &= \sigma^2 \text{Tr}((X^\top X)^{-1}) \\ &= \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i} \,. \end{aligned}$$

For ridge, writing $X^\top X = U \Sigma U^\top$, observe that

$$\begin{aligned} (X^\top X + \lambda I_d)^{-1} &= U(\Sigma + \lambda I_d)^{-1} U^\top \\ (X^\top X + \lambda I_d)^{-1} X^\top X &= U(\Sigma + \lambda I_d)^{-1} \Sigma U^\top \,. \end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{Tr}(E(X^\top X+\lambda I_d)^{-1}X^\top zz^\top X(X^\top X+\lambda I_d)^{-1}) &= \sigma^2\mathrm{Tr}((X^\top X+\lambda I_d)^{-1}X^\top X(X^\top X+\lambda I_d)^{-1})\\ &= \sigma^2\mathrm{Tr}(U(\Sigma+\lambda I_d)^{-1}\Sigma(\Sigma+\lambda I_d)^{-1}U^\top)\\ &= \sigma^2\mathrm{Tr}(\Sigma(\Sigma+\lambda I_d)^{-2})\\ &= \sigma^2\sum_{i=1}^{d}\frac{\gamma_i}{(\gamma_i+\lambda)^2}\,.\end{aligned}$$

The inequality $\mathrm{Tr}(\mathrm{Cov}(w_{\mathrm{ridge}})) < \mathrm{Tr}(\mathrm{Cov}(w_{\mathrm{ols}}))$ holds because $(\gamma_i+\lambda)^2 > \gamma_i^2$ for all $1 \le i \le d$.