Lab Solutions 06

1. Take the gradient of the ridge regression loss function and set it equal to zero. Get an equation for β_λ in terms of X and y. You may need to use the fact that $b^t b$ is equal to $b^t I_p b$ in this derivation. (Note: Do not yet use the SVD of X here).

The gradient of the loss function is given by:

$$\nabla_b \left[||y - Xb||_2^2 + \lambda \cdot ||b||_2^2\right] = \nabla_b \left[y^t y + b^t (X^t X) b - 2y^t Xb + \lambda b^t I_p b\right] \tag{6.1}$$
$$= 2 \cdot (X^t X)b - 2 \cdot X^t y + 2\lambda \cdot I_p b \tag{6.2}$$

Notice that the gradient of $v^t b$ is equal to v and **not** v^t; that seemed to cause a lot of confusion. Now, setting this equal to zero, we have:

$$0 = 2 \cdot (X^t X)b - 2 \cdot X^t y + 2\lambda \cdot I_p b \tag{6.3}$$
$$X^t y = (X^t X)b + \lambda \cdot I_p b \tag{6.4}$$
$$X^t y = \left((X^t X) + \lambda \cdot I_p\right) b \tag{6.5}$$
$$\left((X^t X) + \lambda \cdot I_p\right)^{-1} X^t y = \beta_\lambda. \tag{6.6}$$

And that is all we need to do. Notice that this is the ordinary least squares solution when $\lambda = 0$.

2. The eigenvalue decomposition of a matrix writes a matrix as $Q^t \Lambda Q$ for a diagonal matrix Λ (the entries are called the matrix eigenvalues) and an orthogonal matrix Q. Unlike the SVD, the eigenvalue decomposition only applies to square matrices and even then does not always exist. Show that the eigenvalues of $X^t X$ are equal to the squared singular values of X.

This comes from just plugging in the SVD:

$$X^t X = (UDV^t)^t UDV^t \tag{6.7}$$
$$= VDU^t UDV^t \tag{6.8}$$
$$= VD^2 V^t \tag{6.9}$$

So, we have an eigen-decomposition in terms of the right singular vectors V and the squared singular values D^2.

3. I want to derive a formula for β_λ in terms of the SVD of X; this is not a long derivation but does require a trick. Take the equation that you have already derived for β_λ; there should be an identity matrix in the formula. consider the SVD of X as UDV^t. Write the identify matrix in the equation as VV^t, plug in the SVD

for X, and simplify. You should be able to factor out some of the terms and are left with something very similar to equation we had for the ordinary leasts squares estimator.

Starting with the answer to question one and plugging in the SVD we have:

$$\begin{aligned}
\beta_\lambda &= \left((X^tX) + \lambda \cdot I_p\right)^{-1} X^t y && (6.10)\\
&= \left(VD^2V^t + \lambda \cdot VV^t\right)^{-1} VDU^t y && (6.11)\\
&= \left(V\left[D^2 + \lambda \cdot I_p\right]V^t\right)^{-1} VDU^t y && (6.12)\\
&= V\left[D^2 + \lambda \cdot I_p\right]^{-1} V^tVDU^t y && (6.13)\\
&= V\left[D^2 + \lambda \cdot I_p\right]^{-1} DU^t y && (6.14)\\
&= V \cdot \text{Diag}\left(\frac{\sigma_1}{\sigma_1^2+\lambda}, \cdots, \frac{\sigma_p}{\sigma_p^2+\lambda}\right) \cdot U^t y && (6.15)
\end{aligned}$$

This is just like fitting ordinary least squares, except that the middle term (which was D^{-1} is now a different diagonal matrix).

4. Understand that the ridge regression is equivalent to fitting ordinary least squares on a new matrix $\bar{X}$ where the singular values have been increased by a factor of λ. Given our argument about the problems with the smallest singular value, does it make sense that this change alleviates the problem of identifiability?

You can see from the previous answer that we have made the adjustment:

$$\bar{\sigma_k} \rightarrow \frac{\sigma_k^2 + \lambda}{\sigma_k} \tag{6.16}$$

5. Let's somewhat switch gears here and consider a specific example problem. Let $p = 2$ and assume that the first column of X (X_1) can be written as:

$$X_1 = \alpha + X_2, \quad \alpha \in \mathbb{R}^n \tag{6.17}$$

Where α is a small noise vector. So, X_1 and X_2 are very similar to one another. Write an equation for the value Xb, factoring in terms of α and X_2 (there should not be any X_1 left in the equation). Then, assume that we have data generated by:

$$y = X_2 + \text{noise} \tag{6.18}$$

Where the noise is not too large. Convince yourself that all of the following values

of b produce a reasonable estimate for $\widehat{y} = Xb$:

$$b = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{6.19}$$

$$b = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{6.20}$$

$$b = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \tag{6.21}$$

$$b = \begin{bmatrix} -100 \\ 101 \end{bmatrix} \tag{6.22}$$

What do you think is the approximate value of β_λ for ridge regression for a small value of λ assuming the noise vector and α are also both small?

The value of Xb will be equal to just $X_1 \cdot b_1 + X_2 \cdot b_2$ (writing out the linear regression explicitly). Simplifying we have:

$$\begin{aligned} Xb &= X_1 \cdot b_1 + X_2 \cdot b_2 && (6.23)\\ &= (\alpha + X_2) \cdot b_1 + X_2 \cdot b_2 && (6.24)\\ &= b_1 \cdot \alpha + X_2 \cdot (b_1 + b_2) && (6.25)\\ &\approx X_2 \cdot (b_1 + b_2) && (6.26) \end{aligned}$$

Since we want:

$$Xb \approx y = X_2 + \text{noise} \tag{6.27}$$

And solution with $b_1 + b_2 = 1$ will do the trick. Therefore all of the examples work. For ridge regression, all of the solutions are equally predictive (more or less), so pick the vector that has the smallest size: $b = [0.5, 0.5]$.

6. For the previous question, can you guess a plausible value for V_p (the last left singular) of the matrix X? Look back at handout 5, question 3, for a hint.

We know that adding a multiple of V_p should not change the predictions very much; here the V_p that does that is the vector $[1, -1]$ (it adds something to b_1, subtracts it from b_2 and keeps the sum the same). Keeping in mind that $||V_p||_2 = 1$ we see that a good guess is:

$$V_p = \begin{bmatrix} +2^{-1/2} \\ -2^{-1/2} \end{bmatrix} \tag{6.28}$$

You could also have the signs switched (there's no way to know from the problem

which is most likely).

7. In the following final questions, we are going to consider a dataset where:

$$X^t X = 1_p \cdot n. \tag{6.29}$$

There is nothing you need to compute here, but make sure that you understand: (a) why it makes sense to include the n on the right-hand side, and (b) why it makes sense that we say in this case that the columns of X are uncorrelated.

Consider just the first element of $X^t X$, its the size of the first column of X: $X_1^t X_1 = ||X_1||_2^2$. It makes sense that this increases with the sample size. For (b), the assumption can be read as saying that the columns of X are perpendicular to one another. This geometric property is similar to the probabilistic idea of uncorrelated.

8. Taking $X^t X = 1_p \cdot n$, what is the value of the ordinary least squares estimator? Can you explain exactly what a particular component β_k is in terms of an inner product?

Plugging into the formula we have for the OLS estimator, this is just:

$$\beta = (X^t X)^{-1} X^t y = X^t y. \tag{6.30}$$

So each value of β_k is equal to just $X_k^t y$, the inner product of the k-th column of X with y.

9. Again taking $X^t X = 1_p \cdot n$, write the ridge regression vector β_λ as a function of just n, λ, and the OLS solution β.

Here, we now have:

$$\begin{aligned} \beta_\lambda &= (X^t X + \lambda I_p)^{-1} X^t y && (6.31)\\ &= ((1+\lambda) \cdot I_p)^{-1} X^t y && (6.32)\\ &= \frac{1}{1+\lambda} \cdot X^t y && (6.33)\\ &= \frac{1}{1+\lambda} \cdot \beta && (6.34)\\ & && (6.35) \end{aligned}$$

So when the columns of X are uncorrelated, the ridge regression just scales the OLS solution by a factor of $(1+\lambda)^1$