Ridge Regression Quiz Questions

Karna Mendonca, Joe Zou, Brian Zhu, Tejasvi Kothapalli

1. OLS: Featurization

We've seen in past assignments that we can use linear regression to model more complex functions, such as polynomials, by **featurizing** our data.

$$X = \begin{bmatrix} -\vec{x_1}^\top - \\ -\vec{x_2}^\top - \\ \vdots \\ -\vec{x_n}^\top - \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \dots & x_n^d \end{bmatrix}$$

$$\begin{bmatrix} -\vec{x_1}^\top - \\ -\vec{x_2}^\top - \\ \vdots \\ -\vec{x_n}^\top - \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_d \end{bmatrix}$$

In the above example:

(a) **What does each row of X, $\vec{x_i}^\top$, represent?**

SOLUTION: *A featurized data point*

(b) **What does each column of X, ϕ_i represent?**

SOLUTION: *A polynomial feature of the data*

2. OLS: Derivation

In OLS Regression, we define our error as the square of our residuals, or the shortest distance between our prediction and the correct value.

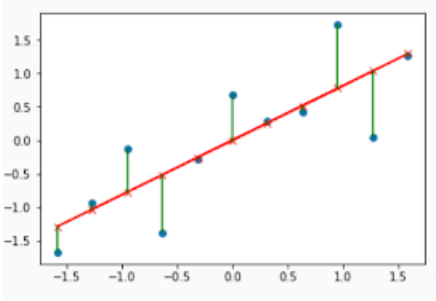

Figure 1: Residuals

(a) **Write an expression for the error for a single data point?**
Hint: how can we represent our prediction and our correct answer using x_i, w, and y_i?

SOLUTION: $(y_i - x_i^\top w)^2$

(b) **We want to gather the collective errors of our data points into a single expression. Write this expression:**
Hint: Can we relate the sum of squares of an element to a vector norm?

SOLUTION: $\|y - Xw\|_2^2$

(c) We can find the optimal weight vector by taking the derivative of the loss function and setting it to 0. For OLS, the derivative of the loss function is:

$$\frac{\partial}{\partial w}\|y - Xw\|_2^2 = 2X^\top Xw - 2X^\top y$$

Set this derivative to 0 and solve for the optimal weight vector.
Hint: Assume $X^\top X$ is invertible.

SOLUTION:

$$2X^\top Xw - 2X^\top y = 0$$
$$X^\top Xw = X^\top y$$
$$\hat{w} = (X^\top X)^{-1}X^\top y$$

3. Overfitting

We've seen that when we have noise in our data, our model can be susceptible to overfitting. Let's try to get an idea of why this happens.

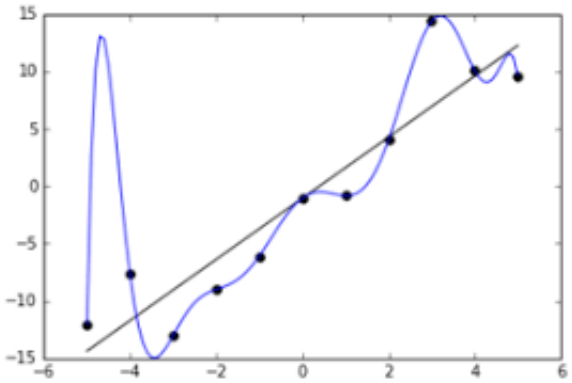

Figure 2: Overfitting

(a) **How does the degree of our predicted function (blue line) compare with the degree of our underlying function?**

SOLUTION: Higher degree

(b) **Why do you think our model predicts an overly complex function?**
Hint: Think about what we're trying to minimize when we're training our model

SOLUTION: In OLS, we're trying to minimize the square of the residuals for each training point. While the blue function doesn't do a great job of predicting points along the line, it manages to go perfectly through every training point, and thus, minimizes training error better.

(c) **How would the weight vector for our predicted and real functions compare?**

SOLUTION: The true function is linear. Thus, the weight vector would only have a two non-zero terms (slope and intercept). The predicted function is high degree, and thus, the weight vector would need to have several terms representing coefficients of the polynomial. Thus, the more complex our model, the higher the magnitude of our weight vector, $\|w\|_2$.

4. Ridge Regression: Derivation

The basic idea of Ridge Regression is to punish more complex models in order to avoid overfitting. We do this by adding a penalty term to our loss function from OLS:

$$\|y - Xw\|_2^2 + \lambda\|w\|_2^2$$

(a) **How does the penalty term $\lambda\|w\|_2^2$ punish overly complex models?**
Hint: Think back to your answer for question 3c. How does complexity of the model affect the norm of the weight vector?

SOLUTION: More complex models will have a "larger" weight vector. By adding the penalty term into our loss function, we add a "tax" on complex models which helps us regularize our prediction.

(b) Let's once again try to derive a closed-form solution by taking the derivative of the loss function and setting it to 0. The derivative of our loss function for Ridge Regression looks like:

$$\frac{\partial}{\partial w}\left(\|y - Xw\|_2^2 + \lambda\|w\|_2^2\right) = 2X^\top Xw - 2X^\top y + 2\lambda w$$

Find the optimal weight vector for Ridge Regression by taking the derivative and setting it to 0.

SOLUTION:

$$2X^\top Xw - 2X^\top y + 2\lambda w = 0$$
$$(X^\top X + \lambda I)w = X^\top y$$
$$\hat{w} = (X^\top X + \lambda I)^{-1}X^\top y$$

5. Ridge Regression: Hyperparameters

Let's think about how our hyperparameter, λ affects our predictions. Below are three examples of Ridge Regression with different values for λ.

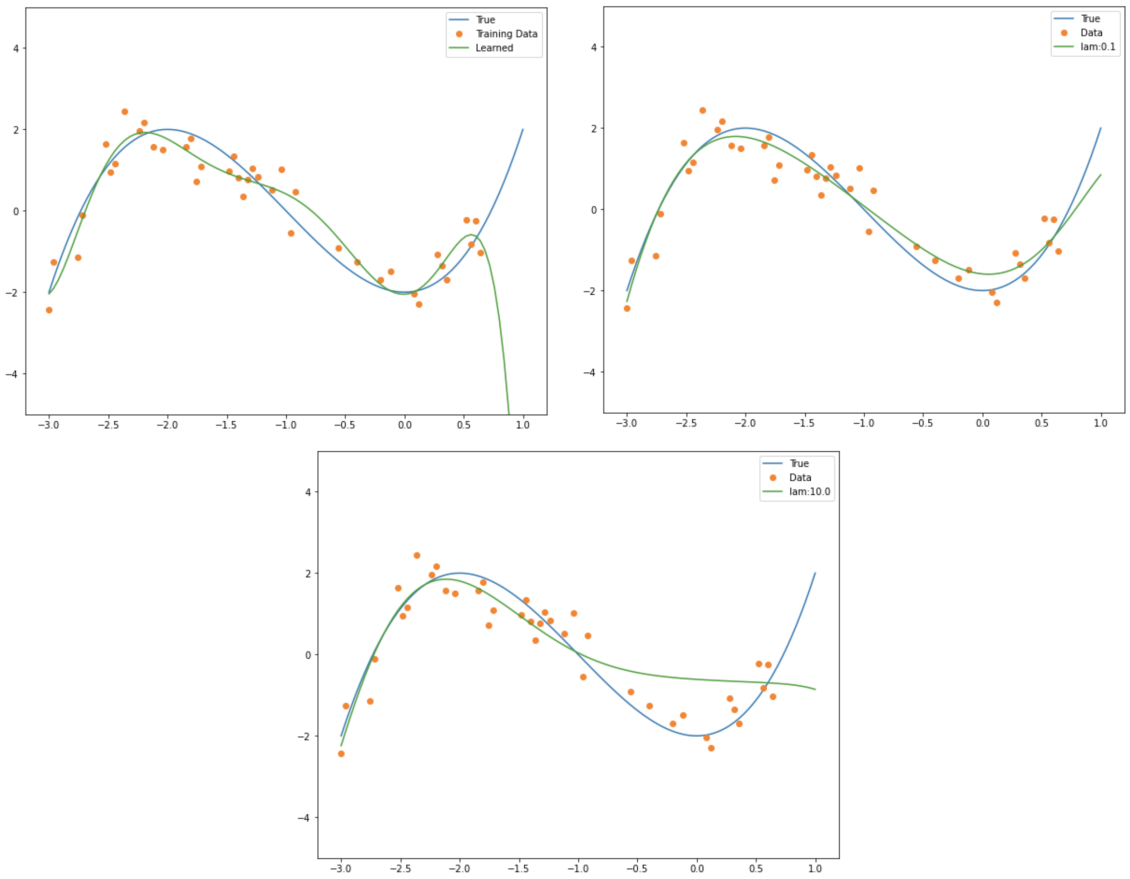

Figure 3: Ridge Regression: $\lambda = \{0, 0.1, 10\}$

Think back to our loss function for Ridge Regression:

$$\|y - Xw\|_2^2 + \lambda\|w\|_2^2$$

(a) **As $\lambda \to 0$, what happens to our predictions?**
SOLUTION: Our predictions become closer to those of OLS.

(b) **As $\lambda \to \infty$, what happens to our predictions?**
SOLUTION: Our weight vector, and thus our predictions, approach 0.

(c) **In a relatively noisy dataset, how would we choose λ? What about in a relatively noise-free dataset?**

SOLUTION: If our data is noisy, we know our model is likely vulnerable to overfitting, so we'd choose a higher value for λ. If our dataset is relatively clean, a higher λ will over-regularize our predictions, and thus we want to choose a lower value. We can also use cross-validation to pick a lambda by splitting our data into training and validation sets.

6. Bias-Variance

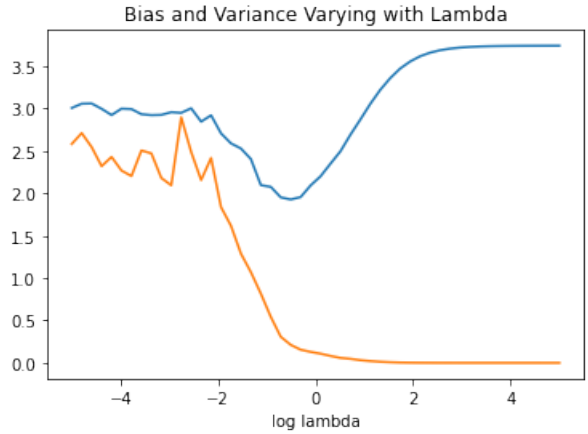

Figure 4: Bias-Variance Tradeoff

Recall that in our toy example, in the notebook, we likened bias and variance to $E[|\hat{w} - w|_2]$ and $Var[|\hat{w}|_2]$, respectively.

(a) **In the plot, which line represents the bias? Which line represents the variance?**
SOLUTION: The blue line is the bias. The orange line is the variance.

(b) **Why is a high bias considered underfitting?**
hint: In general machine learning, high bias implies that the model is likely to miss important patterns. Think about how this relates to the large λ causing a high bias in our case.

SOLUTION: The large λ causes $|\hat{w}|_2$ to be closer to zero. As a result, the large λ is forcing the weights in $\hat{w}$ to be close to zero when it reality the weights may be larger. Thus, the model is underfitting the data by not achieving the function we had hoped to predict.

(c) **Why is high variance considered overfitting?**
hint: High variance generally implies that the model is too sensitive to small changes in the training data. Think about how this relates to the case when λ goes close to zero and the ridge regression essentially becomes OLS.

SOLUTION: When λ is close to zero we are essentially running OLS. In the 'Overfitting of Noise using OLS' section of the notebook it was shown that OLS can overfit to the data. In this case, OLS is too susceptible to the small changes in noise and thus there is a greater spread in the possible predicted weights, $\hat{w}$. This is overfitting because the model fits the training data 'too well' but will not do well on future, unseen data.

7. Overfitting and Eigenvalues

Let's take a look at our solution to OLS again:

$$\hat{w} = (X^\top X)^{-1} X^\top y$$

We can try to learn about the mathematical perspective to overfitting by analyzing the numerical instability in $X^\top X$:

(a) Let's first restrict our data to one dimension. The matrix multiplication $X^T X$ would be $1 \times n$ by $n \times 1$, which outputs a 1×1 matrix, making $X^T X$ a scalar in this case. **What values of $X^T X$ may "mess up" the calculation of $\hat{w}$ in the closed-form solution?**
Hint: What happens if $X^T X$ is 0 or close to 0?

SOLUTION: If $X^T X$ is 0 or very close to 0, then $(X^T X)^{-1}$, or $\frac{1}{X^T X}$, would end up exploding, since $\frac{1}{x}$ would approach infinity as x approaches 0.

(b) Now, let's generalize to the case where $X^\top X$ is a $d \times d$ matrix. **First, verify that $X^\top X$ is symmetric.**

SOLUTION: $(X^\top X)^\top = X^\top (X^\top)^\top = X^\top X$

(c) Now because $X^\top X$ is symmetric, we know that is also diagonalizable. That is, we can write in the form: $X^\top X = V \Lambda V^{-1}$ where:

$$\Lambda = \begin{bmatrix} \tilde{\lambda}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\lambda}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\lambda}_n \end{bmatrix}$$

How can we write the inverse of $X^\top X$ in terms of V, Λ, and their inverses? Rewrite our solution to OLS in this form.

SOLUTION: $\left(V \Lambda V^{-1}\right)^{-1} = V \Lambda^{-1} V^{-1}$. Thus, we can rewrite our solution to OLS as: $\hat{w} = V \Lambda^{-1} V^{-1} X^\top y$

(d) Note the term Λ^{-1} in $(X^\top X)^{-1}$ takes the form:

$$\Lambda^{-1} = \begin{bmatrix} \frac{1}{\tilde{\lambda}_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\tilde{\lambda}_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{\tilde{\lambda}_n} \end{bmatrix}$$

What happens when one of our eigenvalues $\tilde{\lambda}_j$ is quite small (say, 10^{-10})? How might this affect our solution when there is noise in y?

SOLUTION: The term $\frac{1}{\tilde{\lambda}_j}$ in Λ becomes quite large, to 10^{10}. Thus, a little bit of noise in y can be amplified up to 10^{10} times in our closed form solution.

8. Ridge Regression as a High Pass Filter

(a) Describe how the eigenvalues of $X^\top X + \lambda I$ behave in ridge regression. **What happens when an eigenvalue of $X^\top X$ is much larger than λ? When its much smaller?**

SOLUTION: Let $\tilde{\lambda}_i$ be the ith eigenvalue of $X^\top X$ and let $\tilde{\lambda}'_i$ be the ith eigenvalue of $X^\top X + \lambda I$.

When $\tilde{\lambda}_i \gg \lambda$, $\tilde{\lambda}'_i \approx \tilde{\lambda}_i$.

Conversely, when $\tilde{\lambda}_i \ll \lambda$, $\tilde{\lambda}'_i \approx \lambda$

(b) Recall the high pass filter. In particular, we analyzed a basic RC circuit and found that its transfer function is:

$$H(\omega) = \frac{V_{out}}{V_{in}} = \frac{1}{1 + \frac{1}{j\omega RC}}$$

What happens to the magnitude of H, i.e. $|H(\omega)|$, when ω is very large? How about when ω is very small? For both of these cases, how does the response of H in terms of ω affect V_{out}?

SOLUTION: We see that when ω is very large, i.e. $\omega \to \infty$,

$$H(\infty) = \frac{1}{1 + j\frac{1}{\infty}} = \frac{1}{1 + 0j} = 1 \implies |H(\infty)| = 1$$

So V_{out} has the same magnitude as V_{in}, which means that the signal is preserved.

Conversely, when ω is very small, i.e. $\omega \to 0$,

$$H(0) = \frac{1}{1 + j\frac{1}{0}} = \frac{1}{1 + \infty j}$$

$$|H(0)| = \left|\frac{1}{1 + \infty j}\right| = \frac{1}{|1 + \infty j|} = \frac{1}{\sqrt{1^2 + \infty^2}} \approx \frac{1}{\infty} = 0$$

So the magnitude of V_{out} is 0, or the signal is not preserved.

(c) Compare the behavior of the eigenvalues of $X^\top X + \lambda I$ to the behavior of $H(\omega)$ in the high pass filter. **How are they similar?**

SOLUTION: We see that both "systems" preserve or discard a value based on the size of an input. In the case of the eigenvalues, each eigenvalue serves as an input for whether it is preserved in ridge regression, while in the high-pass filter, ω is the input that decides whether the output voltage is preserved. In both cases, when the input is large, the output is preserved, and when the input is small, the output is essentially discarded.

9. Fake Data Perspective of Ridge Regression

Let us augment our data with fake data points such that:

$$\hat{X} = \begin{bmatrix} X \\ aI \end{bmatrix}$$

$$\hat{y} = \begin{bmatrix} \vec{y} \\ 0 \end{bmatrix}$$

(a) Given that our original X matrix is $n \times d$ and y vector has length n. **What is the dimension of $\hat{X}$, $\hat{y}$, and the learned $\hat{w}$?**

SOLUTION: $\hat{X}$ is $(n+d) \times d$, y has length $n+d$, and $\hat{w}$ has length d.

(b) **Find the OLS solution using the augmented data and simplify**
Hint: Recall that the OLS closed-form solution is $\hat{w} = (X^T X)^{-1} X^T y$

SOLUTION:

$$\hat{w} = (\hat{X}^T \hat{X}) \hat{X}^T \hat{y}$$

$$\hat{w} = (\begin{bmatrix} X^T aI \end{bmatrix} \begin{bmatrix} X \\ aI \end{bmatrix})^{-1} \begin{bmatrix} X^T aI \end{bmatrix} \begin{bmatrix} \vec{y} \\ 0 \end{bmatrix}$$

$$\hat{w} = (X^T X + a^2 I)^{-1} X^T y$$

(c) Recall that the closed-form solution for ridge regression is $\hat{w} = (X^T X + \lambda I)^{-1} X^T y$. **What value of a should we pick for augmenting our data so that we get the same solution as we would from ridge regression?**
Answer in terms of λ

SOLUTION: We can set $a = \sqrt{\lambda}$

10. Fake Features Perspective of Ridge Regression

Let us augment our data with fake features such that:

$$\hat{X} = \left[X\sqrt{\lambda}I\right]$$

(a) Given that our original X matrix is $n \times d$ and y vector has length n. **What is the dimension of $\hat{X}$, y, and the learned $\hat{w}$?**

SOLUTION: $\hat{X}$ is $n \times (d+n)$, y has length n, and $\hat{w}$ has length $d+n$.

(b) Let us represent the weight vector for the augmented $\hat{X}$ matrix as $\begin{bmatrix}\hat{w}\\ \hat{\epsilon}\end{bmatrix}$ where $\hat{w}$ are the weights associated with the original features and $\hat{\epsilon}$ are weights associated with the fake features. **What are the sizes of $\hat{w}$ and $\hat{\epsilon}$?**

SOLUTION: $\hat{w}$ has length d and $\hat{\epsilon}$ has length n.

(c) Notice that the augmented $\hat{X}$ matrix is wide, so we must use the minimum norm solution. **Show that the minimum-norm solution will net us the same solution for $\hat{w}$ as the alternative Ridge Regression solution.**
Hint: Recall that the minimum norm closed-form solution is $\hat{w} = X^T(XX^T)^{-1}y$ and the alternative closed-form solution to ridge regression is $\hat{w} = X^T(XX^T)^{-1}y$

SOLUTION:

$$\begin{bmatrix}\hat{w}\\ \hat{\epsilon}\end{bmatrix} = \hat{X}^T(\hat{X}\hat{X}^T)^{-1}\vec{y}$$

$$\begin{bmatrix}\hat{w}\\ \hat{\epsilon}\end{bmatrix} = \begin{bmatrix}X^T\\ \sqrt{\lambda}I\end{bmatrix}(\left[X\sqrt{\lambda}I\right]\begin{bmatrix}X^T\\ \sqrt{\lambda}I\end{bmatrix})^{-1}\vec{y}$$

$$\begin{bmatrix}\hat{w}\\ \hat{\epsilon}\end{bmatrix} = \begin{bmatrix}X^T\\ \sqrt{\lambda}I\end{bmatrix}(XX^T + \lambda I)^{-1}\vec{y}$$

$$\hat{w} = X^T(XX^T + \lambda I)^{-1}\vec{y}$$