APPM 4600 Project 3: Regularization in Least Squares
Alexey Yermakov, Logan Barnhart, and Tyler Jensen

1 Ridge Regression

1.1 Deriving the Ridge Estimator

The equation for regularized least squares is:

$$\arg\min_x ||\mathbf{Ax} - \mathbf{b}||_2^2 + \gamma||\mathbf{x}||_2^2 \tag{1}$$

Recalling that $||\mathbf{x}||_2^2 = \mathbf{x}^T\mathbf{x}$, we'll rewrite $||\mathbf{Ax} - \mathbf{b}||_2^2 + \gamma||\mathbf{x}||_2^2$:

$$\begin{aligned}
&||\mathbf{Ax} - \mathbf{b}||_2^2 + \gamma||\mathbf{x}||_2^2 = \\
&= (\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}) + \gamma\mathbf{x}^T\mathbf{x} \\
&= ((\mathbf{Ax})^T - \mathbf{b}^T)(\mathbf{Ax} - \mathbf{b}) + \gamma\mathbf{x}^T\mathbf{x} \\
&= (\mathbf{x}^T\mathbf{A}^T - \mathbf{b}^T)(\mathbf{Ax} - \mathbf{b}) + \gamma\mathbf{x}^T\mathbf{x} \\
&= \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b} + \gamma\mathbf{x}^T\mathbf{x}
\end{aligned}$$

Before we proceed further, we'll define what $\mathbf{A}$, $\mathbf{b}$, and $\mathbf{x}$ are. We want a least-squares fit to an m-degree polynomial $p_m(x) = a_0 + a_1 * x + \ldots + a_m * x^m$ where we have n data points $\{x_0, b_0\}$. To have our system be overdetermined, we also assume $n > m$.

$$\mathbf{A} = \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^m \\ 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^n & \ldots & x_n^m \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}$$

Where $dim(\mathbf{A}) = (n+1) \times (m+1)$, $dim(\mathbf{x}) = (m+1) \times (1)$, and $dim(\mathbf{b}) = (n+1) \times (1)$.

We'll now prove $\mathbf{x}^T\mathbf{A}^T\mathbf{b} = \mathbf{b}^T\mathbf{Ax}$. Note first that $(\mathbf{b}^T\mathbf{Ax})^T = \mathbf{x}^T\mathbf{A}^T\mathbf{b}$. Further, $dim(\mathbf{x}^T\mathbf{A}^T\mathbf{b}) = (1) \times (1) = dim(\mathbf{b}^T\mathbf{Ax})$. Also, note that the transpose of a 1×1 matrix is the same matrix: $[c]^T = [c]$. It then follows that $\mathbf{x}^T\mathbf{A}^T\mathbf{b} = \mathbf{b}^T\mathbf{Ax}$, completing the proof.

So,

$$\begin{aligned}
&\mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b} + \gamma\mathbf{x}^T\mathbf{x} = \\
&= \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b} + \gamma\mathbf{x}^T\mathbf{x} \\
&= (\mathbf{Ax})^T(\mathbf{Ax}) - 2\mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b} + \gamma\mathbf{x}^T\mathbf{x}
\end{aligned} \tag{2}$$

Now, we'll show what each of the above values (since each matrix is 1×1, meaning it's a scalar) actually is:

$$
\begin{aligned}
&\mathbf{Ax} = \\
&= \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^m \\ 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^n & \dots & x_n^m \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} \\
&= \begin{bmatrix} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_m x_0^m \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_m x_1^m \\ a_0 + a_1 x_2 + a_2 x_2^2 + \ldots + a_m x_2^m \\ \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_m x_n^m \end{bmatrix}
\end{aligned} \tag{3}
$$

Then, $(\mathbf{Ax})^T\mathbf{Ax}$ is just a simple inner product (we will use the result from 3):

$$
\begin{aligned}
&(\mathbf{Ax})^T\mathbf{Ax} = \\
&= \begin{bmatrix} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_m x_0^m \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_m x_1^m \\ a_0 + a_1 x_2 + a_2 x_2^2 + \ldots + a_m x_2^m \\ \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_m x_n^m \end{bmatrix}^T \times \begin{bmatrix} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_m x_0^m \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_m x_1^m \\ a_0 + a_1 x_2 + a_2 x_2^2 + \ldots + a_m x_2^m \\ \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_m x_n^m \end{bmatrix} \\
&= (a_0 + a_1 x_0 + \ldots + a_m x_0^m)^2 + (a_0 + a_1 x_1 + \ldots + a_m x_1^m)^2 \\
&+ \ldots + (a_0 + a_1 x_n + \ldots + a_m x_n^m)^2
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
&2\mathbf{b}^T\mathbf{Ax} = \\
&= 2 \times \begin{bmatrix} b_0 & b_1 & \dots & b_n \end{bmatrix} \times \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^m \\ 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} \\
&= 2 \times \begin{bmatrix} b_0 + b_1 + b_2 + \ldots + b_n \\ b_0 x_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \\ b_0 x_0^2 + b_1 x_1^2 + b_2 x_2^2 + \ldots + b_n x_n^2 \\ \vdots \\ b_0 x_0^m + b_1 x_1^m + b_2 x_2^m + \ldots + b_n x_n^m \end{bmatrix}^T \times \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} \\
&= 2(a_0(b_0 + b_1 + \ldots + b_n) + a_1(b_0 x_0 + b_1 x_1 + \ldots + b_n x_n) + \ldots + a_m(b_0 x_0^m + b_1 x_1^m + \ldots + b_n x_n^m))
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
&\gamma\mathbf{x}^T\mathbf{x} = \\
&\text{This is a simple inner product multiplied by a scalar} \\
&= \gamma a_0^2 + \gamma a_1^2 + \ldots + \gamma a_m^2
\end{aligned} \tag{6}
$$

Great! Now lets take the derivates of 4, 5, and 6 with respect to $\mathbf{x}$ to get the derivative of 1 with respect to $\mathbf{x}$. In effect, we'll get a vector where the i-th element is the derivative of 1 with respect to a_i.

First, let's take the derivatives of 4:

$$
\begin{aligned}
&\frac{d}{d\mathbf{x}}(\mathbf{Ax})^T\mathbf{Ax} = \\
&= \begin{bmatrix} \frac{d}{da_0}(\mathbf{Ax})^T\mathbf{Ax} \\ \frac{d}{da_1}(\mathbf{Ax})^T\mathbf{Ax} \\ \vdots \\ \frac{d}{da_m}(\mathbf{Ax})^T\mathbf{Ax} \end{bmatrix} \\
&= \begin{bmatrix} 2(a_0 + a_1x_0 + \cdots + a_mx_0^m) + 2(a_0 + a_1x_1 + \cdots + a_mx_1^m) + \ldots + 2(a_0 + a_1x_n + \cdots + a_mx_n^m) \\ 2x_0(a_0 + a_1x_0 + \cdots + a_mx_0^m) + 2x_1(a_0 + a_1x_1 + \cdots + a_mx_1^m) + \ldots + 2x_n(a_0 + a_1x_n + \cdots + a_mx_n^m) \\ \vdots \\ 2x_0^m(a_0 + a_1x_0 + \cdots + a_mx_0^m) + 2x_1^m(a_0 + a_1x_1 + \cdots + a_mx_1^m) + \ldots + 2x_n^m(a_0 + a_1x_n + \cdots + a_mx_n^m) \end{bmatrix} \\
&= 2 \times \begin{bmatrix} a_0\sum_{i=0}^n 1 + a_1\sum_{i=0}^n x_i + a_2\sum_{i=0}^n x_i^2 + \ldots + a_m\sum_{i=0}^n x_i^m \\ a_0\sum_{i=0}^n x_i + a_1\sum_{i=0}^n x_i^2 + a_2\sum_{i=0}^n x_i^3 + \ldots + a_m\sum_{i=0}^n x_i^{m+1} \\ \vdots \\ a_0\sum_{i=0}^n x_i^m + a_1\sum_{i=0}^n x_i^{m+1} + a_2\sum_{i=0}^n x_i^{m+2} + \ldots + a_m\sum_{i=0}^n x_i^{2m} \end{bmatrix} \\
&= 2 \times \begin{bmatrix} \sum_{i=0}^n 1 & \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \cdots & \sum_{i=0}^n x_i^m \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \cdots & \sum_{i=0}^n x_i^{m+1} \\ \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \sum_{i=0}^n x_i^4 & \cdots & \sum_{i=0}^n x_i^{m+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^n x_i^m & \sum_{i=0}^n x_i^{m+1} & \sum_{i=0}^n x_i^{m+2} & \cdots & \sum_{i=0}^n x_i^{2m} \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \\
&= 2 \times \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_0 & x_1 & x_2 & \ldots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^m & x_1^m & x_2^m & \ldots & x_n^m \end{bmatrix} \times \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^m \\ 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \\
&= 2\mathbf{A}^T\mathbf{Ax}
\end{aligned} \tag{7}
$$

Secondly, let's take the derivatives of 5:

$$
\begin{aligned}
&\frac{d}{d\mathbf{x}} 2\mathbf{b}^T\mathbf{A}\mathbf{x} = \\
&= \begin{bmatrix} \frac{d}{da_0} 2\mathbf{b}^T\mathbf{A}\mathbf{x} \\ \frac{d}{da_1} 2\mathbf{b}^T\mathbf{A}\mathbf{x} \\ \vdots \\ \frac{d}{da_m} 2\mathbf{b}^T\mathbf{A}\mathbf{x} \end{bmatrix} \\
&= 2 \times \begin{bmatrix} b_0 + b_1 + \ldots + b_n \\ b_0 x_0 + b_1 x_1 + \ldots + b_n x_n \\ b_0 x_0^2 + b_1 x_1^2 + \ldots + b_n x_n^2 \\ \vdots \\ b_0 x_0^m + b_1 x_1^m + \ldots + b_n x_n^m \end{bmatrix} \\
&= 2 \times \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_0 & x_1 & x_2 & \ldots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^m & x_1^m & x_2^m & \ldots & x_n^m \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \\
&= 2\mathbf{A}^T\mathbf{b}
\end{aligned} \tag{8}
$$

Lastly, let's take the derivatives of 6:

$$
\begin{aligned}
&\frac{d}{d\mathbf{x}} \gamma\mathbf{x}^T\mathbf{x} = \\
&= \begin{bmatrix} \frac{d}{da_0} \gamma\mathbf{x}^T\mathbf{x} \\ \frac{d}{da_1} \gamma\mathbf{x}^T\mathbf{x} \\ \vdots \\ \frac{d}{da_m} \gamma\mathbf{x}^T\mathbf{x} \end{bmatrix} \\
&= \gamma \times \begin{bmatrix} 2a_0 \\ 2a_1 \\ 2a_2 \\ \vdots \\ 2a_m \end{bmatrix} \\
&= 2 \times \gamma \times \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \\
&= 2\gamma\mathbf{I}\mathbf{x}
\end{aligned} \tag{9}
$$

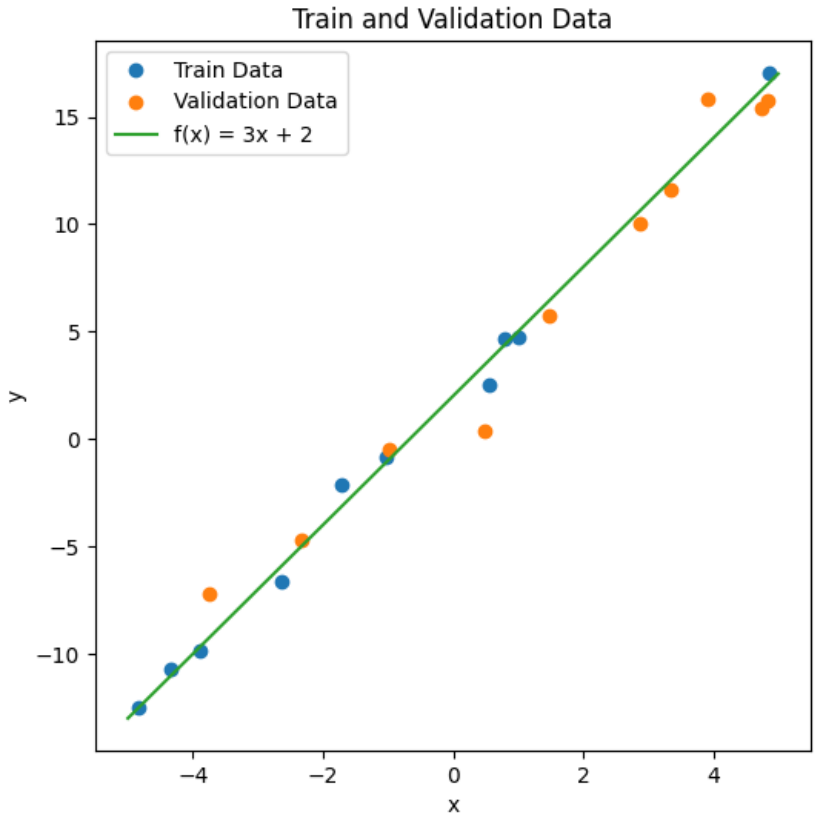
Now we can recall 2 and note that to find the minimum of the least squares equation given by 1 we can find where the derivative of 2 is equal to zero:

$$
\begin{aligned}
\frac{d}{d\mathbf{x}}(\,(\mathbf{Ax})^T(\mathbf{Ax}) - 2\mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b} + \gamma\mathbf{x}^T\mathbf{x}) &= 0 \\
2\mathbf{A}^T\mathbf{Ax} - 2\mathbf{A}^T\mathbf{b} + 2\gamma\mathbf{Ix} &= 0 \\
\mathbf{A}^T\mathbf{Ax} - \mathbf{A}^T\mathbf{b} + \gamma\mathbf{Ix} &= 0 \\
\mathbf{A}^T\mathbf{Ax} + \gamma\mathbf{Ix} - \mathbf{A}^T\mathbf{b} &= 0 \\
(\mathbf{A}^T\mathbf{A} + \gamma\mathbf{I})\mathbf{x} - \mathbf{A}^T\mathbf{b} &= 0 \\
(\mathbf{A}^T\mathbf{A} + \gamma\mathbf{I})\mathbf{x} &= \mathbf{A}^T\mathbf{b} \\
\mathbf{x} &= (\mathbf{A}^T\mathbf{A} + \gamma\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b}
\end{aligned}
\tag{10}
$$

We have now arrived at the equation for Ridge Regression! We note that there is a typo in the project description for the equation $E_{ridge} = (\mathbf{A}^T\mathbf{A} + \gamma\mathbf{I})^{-1}\mathbf{A}^T$, where there is no trailing $\mathbf{b}$.

1.2 Exploring the Ridge Estimator

Now that we have derived the Ridge Estimator, we are now ready to implement and test the Ridge Estimator on some data. All code is in python 3, numpy's random generator with seed 50 was used to generate any random data. We begin by taking 20 random samples from the line $y = 3x + 2$ on the interval $[-5, 5)$ with Gaussian noise from a standard normal $\sim N(0, 1)$ distribution. We now randomly sample 10 of these data points and will use them to build the line of best fit (find m and c in $y = mx + b$) for this sample using ridge regression. We will refer to these data points as training data, and the remaining 10 points will be referred to as our validation data and will be used to measure the accuracy of the model using Residual Sum of Squares(also known as Sum of Squared Errors).



$$RSS = \sum_{n=1}^{N} (y_{valid,n} - y_{predict,n})^2 \tag{11}$$

where y_{valid} is the y values of the validation data, and $y_{predict}$ are the predicted y values for the validation data using the ridge regression model trained from our training data, and N is the total number of samples that we are predicting for(In this case, 10).

From (?) we know that $E_{ridge} = (\mathbf{A}^T\mathbf{A} + \gamma\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b}$. The A matrix for this model is

$$\mathbf{A} = \begin{bmatrix} 1 & x_{train,1} \\ 1 & x_{train,2} \\ 1 & x_{train,3} \\ \vdots & \vdots \\ 1 & x_{train,10} \end{bmatrix} \mathbf{b} = \begin{bmatrix} y_{train,1} \\ y_{train,2} \\ y_{train,3} \\ \vdots \\ y_{train,10} \end{bmatrix} \mathbf{x}* = \begin{bmatrix} b \\ m \end{bmatrix}$$

We first start with $\gamma = 0$. In that case, the ridge estimator reduces to $E_{ridge} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$ which is just the normal equation for standard non penalized least squares. We thus try $\gamma = 0.1$ as well, for our first real "ridge" regression. -We get the following RSS

γ	0	0.1
RSS		

We see that we have a lower RSS with $\gamma = 0.1$ rather than $\gamma = 0$. Thus we have already produced a result better than non penalized least squares! We have already found a situation where ridge regression is more accurate than non penalized regression. Now the question is are there any other γ's where ridge regression performs better than standard least squares? What is the best γ? There is no analytic way to determine the best γ, so we try a lot of γ's and see what happens. First we try $\gamma = 0.1, 0.2, 0.3, 0.4, \ldots 10,000$.

From the graph above we see that the γ's with the best RSS are in the interval $[0, 10]$. Thus we now try $\gamma = 0, 0.01, 0.02, 0.03, 0.04, \ldots, 10$

We found minimum RSS to occur at $\gamma = 1$.

We next move on to using ridge regression for higher degree polynomials. We consider the model

$$y = ax^5 + bx^4 + cx^3 + dx^2 + ex + f \tag{12}$$

and we will fit $y = x^2$ using the same process as we did for finding the line of best fit. We sample 20 points on the interval $[-5, 5)$, take 10 to be our training data, and 10 to be our validation data, with Gaussian noise from a standard normal added in. We get the following plot