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:

$$\arg\min_{\boldsymbol{\beta}} ||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 + \gamma||\boldsymbol{\beta}||_2^2 \tag{1}$$

Recalling that $||\boldsymbol{\beta}||_2^2 = \boldsymbol{\beta}^T\boldsymbol{\beta}$, we'll rewrite $||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 + \gamma||\boldsymbol{\beta}||_2^2$:

$$\begin{aligned}
&||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 + \gamma||\boldsymbol{\beta}||_2^2 = \\
&= (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta} \\
&= ((\mathbf{X}\boldsymbol{\beta})^T - \mathbf{y}^T)(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta} \\
&= (\boldsymbol{\beta}^T\mathbf{X}^T - \mathbf{y}^T)(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta} \\
&= \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y} + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta}
\end{aligned}$$

Before we proceed further, we'll define what $\mathbf{X}$, $\mathbf{y}$, and $\boldsymbol{\beta}$ are. We want a least-squares fit to an m-degree polynomial $p_m(x) = \beta_0 + \beta_1 * x + \ldots + \beta_m * x^m$ where we have n data points $\{x_0, y_0\}$. To have our system be overdetermined, we also assume $n > m$.

$$\mathbf{X} = \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^m \\ 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^n & \ldots & x_n^m \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Where $dim(\mathbf{X}) = (n+1) \times (m+1)$, $dim(\boldsymbol{\beta}) = (m+1) \times (1)$, and $dim(\mathbf{y}) = (n+1) \times (1)$.

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

So,

$$\begin{aligned}
&\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y} + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta} = \\
&= \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y} + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta} \\
&= (\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\boldsymbol{\beta}) - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y} + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta}
\end{aligned} \tag{2}$$

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

$$\begin{aligned}
&\mathbf{X}\boldsymbol{\beta} = \\
&= \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^m \\ 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^n & \dots & x_n^m \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \\
&= \begin{bmatrix} \beta_0 + \beta_1 x_0 + \beta_2 x_0^2 + \dots + \beta_m x_0^m \\ \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \dots + \beta_m x_1^m \\ \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \dots + \beta_m x_2^m \\ \vdots \\ \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \dots + \beta_m x_n^m \end{bmatrix}
\end{aligned} \tag{3}$$

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

$$\begin{aligned}
&(\mathbf{X}\boldsymbol{\beta})^T\mathbf{X}\boldsymbol{\beta} = \\
&= \begin{bmatrix} \beta_0 + \beta_1 x_0 + \beta_2 x_0^2 + \dots + \beta_m x_0^m \\ \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \dots + \beta_m x_1^m \\ \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \dots + \beta_m x_2^m \\ \vdots \\ \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \dots + \beta_m x_n^m \end{bmatrix}^T \times \begin{bmatrix} \beta_0 + \beta_1 x_0 + \beta_2 x_0^2 + \dots + \beta_m x_0^m \\ \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \dots + \beta_m x_1^m \\ \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \dots + \beta_m x_2^m \\ \vdots \\ \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \dots + \beta_m x_n^m \end{bmatrix} \\
&= (\beta_0 + \beta_1 x_0 + \dots + \beta_m x_0^m)^2 + (\beta_0 + \beta_1 x_1 + \dots + \beta_m x_1^m)^2 \\
&+ \dots + (\beta_0 + \beta_1 x_n + \dots + \beta_m x_n^m)^2
\end{aligned} \tag{4}$$

$$\begin{aligned}
&2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} = \\
&= 2 \times \begin{bmatrix} y_0 & y_1 & \dots & y_n \end{bmatrix} \times \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^m \\ 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \\
&= 2 \times \begin{bmatrix} y_0 + y_1 + y_2 + \dots + y_n \\ y_0 x_0 + y_1 x_1 + y_2 x_2 + \dots + y_n x_n \\ y_0 x_0^2 + y_1 x_1^2 + y_2 x_2^2 + \dots + y_n x_n^2 \\ \vdots \\ y_0 x_0^m + y_1 x_1^m + y_2 x_2^m + \dots + y_n x_n^m \end{bmatrix}^T \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \\
&= 2(\beta_0(y_0 + y_1 + \dots + y_n) + \beta_1(y_0 x_0 + y_1 x_1 + \dots + y_n x_n) + \dots + \beta_m(y_0 x_0^m + y_1 x_1^m + \dots + y_n x_n^m))
\end{aligned} \tag{5}$$

$$\begin{aligned}
&\gamma\boldsymbol{\beta}^T\boldsymbol{\beta} = \\
&\text{This is a simple inner product multiplied by a scalar} \\
&= \gamma\beta_0^2 + \gamma\beta_1^2 + \dots + \gamma\beta_m^2
\end{aligned} \tag{6}$$

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

First, let's take the derivatives of 4:

$$
\begin{aligned}
&\frac{d}{d\boldsymbol{\beta}}(\mathbf{X}\boldsymbol{\beta})^T\mathbf{X}\boldsymbol{\beta} = \\
&= \begin{bmatrix} \frac{d}{d\beta_0}(\mathbf{X}\boldsymbol{\beta})^T\mathbf{X}\boldsymbol{\beta} \\ \frac{d}{d\beta_1}(\mathbf{X}\boldsymbol{\beta})^T\mathbf{X}\boldsymbol{\beta} \\ \vdots \\ \frac{d}{d\beta_m}(\mathbf{X}\boldsymbol{\beta})^T\mathbf{X}\boldsymbol{\beta} \end{bmatrix} \\
&= \begin{bmatrix} 2(\beta_0 + \beta_1 x_0 + \cdots + \beta_m x_0^m) + 2(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_1^m) + \ldots + 2(\beta_0 + \beta_1 x_n + \cdots + \beta_m x_n^m) \\ 2x_0(\beta_0 + \beta_1 x_0 + \cdots + \beta_m x_0^m) + 2x_1(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_1^m) + \ldots + 2x_n(\beta_0 + \beta_1 x_n + \cdots + \beta_m x_n^m) \\ \vdots \\ 2x_0^m(\beta_0 + \beta_1 x_0 + \cdots + \beta_m x_0^m) + 2x_1^m(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_1^m) + \ldots + 2x_n^m(\beta_0 + \beta_1 x_n + \cdots + \beta_m x_n^m) \end{bmatrix} \\
&= 2 \times \begin{bmatrix} \beta_0 \sum_{i=0}^n 1 + \beta_1 \sum_{i=0}^n x_i + \beta_2 \sum_{i=0}^n x_i^2 + \ldots + \beta_m \sum_{i=0}^n x_i^m \\ \beta_0 \sum_{i=0}^n x_i + \beta_1 \sum_{i=0}^n x_i^2 + \beta_2 \sum_{i=0}^n x_i^3 + \ldots + \beta_m \sum_{i=0}^n x_i^{m+1} \\ \vdots \\ \beta_0 \sum_{i=0}^n x_i^m + \beta_1 \sum_{i=0}^n x_i^{m+1} + \beta_2 \sum_{i=0}^n x_i^{m+2} + \ldots + \beta_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 & \ldots & \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 & \ldots & \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 & \ldots & \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} & \ldots & \sum_{i=0}^n x_i^{2m} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_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} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \\
&= 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}
\end{aligned}
\tag{7}
$$

Secondly, let's take the derivatives of 5:

$$
\begin{aligned}
&\frac{d}{d\boldsymbol{\beta}}2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} = \\
&= \begin{bmatrix} \frac{d}{da_0}2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} \\ \frac{d}{da_1}2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} \\ \vdots \\ \frac{d}{da_m}2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} \end{bmatrix} \\
&= 2 \times \begin{bmatrix} y_0 + y_1 + \ldots + y_n \\ y_0x_0 + y_1x_1 + \ldots + y_nx_n \\ y_0x_0^2 + y_1x_1^2 + \ldots + y_nx_n^2 \\ \vdots \\ y_0x_0^m + y_1x_1^m + \ldots + y_nx_n^m \end{bmatrix} \\
&= 2 \times \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_0 & x_1 & x_2 & \ldots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^m & x_1^m & x_2^m & \ldots & x_n^m \end{bmatrix} \times \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \\
&= 2\mathbf{X}^T\mathbf{y}
\end{aligned}
\tag{8}
$$

Lastly, let's take the derivatives of 6:

$$
\begin{aligned}
&\frac{d}{d\boldsymbol{\beta}}\gamma\boldsymbol{\beta}^T\boldsymbol{\beta} = \\
&= \begin{bmatrix} \frac{d}{d\beta_0}\gamma\boldsymbol{\beta}^T\boldsymbol{\beta} \\ \frac{d}{d\beta_1}\gamma\boldsymbol{\beta}^T\boldsymbol{\beta} \\ \vdots \\ \frac{d}{d\beta_m}\gamma\boldsymbol{\beta}^T\boldsymbol{\beta} \end{bmatrix} \\
&= \gamma \times \begin{bmatrix} 2\beta_0 \\ 2\beta_1 \\ 2\beta_2 \\ \vdots \\ 2\beta_m \end{bmatrix} \\
&= 2 \times \gamma \times \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \\
&= 2\gamma\mathbf{I}\boldsymbol{\beta}
\end{aligned}
\tag{9}
$$

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:

$$\begin{aligned}\frac{d}{d\boldsymbol{\beta}}(\ (\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\boldsymbol{\beta}) - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y} + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta}) &= 0\\ 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^T\mathbf{y} + 2\gamma\mathbf{I}\boldsymbol{\beta} &= 0\\ \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^T\mathbf{y} + \gamma\mathbf{I}\boldsymbol{\beta} &= 0\\ \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{I}\boldsymbol{\beta} - \mathbf{X}^T\mathbf{y} &= 0\\ (\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I})\boldsymbol{\beta} - \mathbf{X}^T\mathbf{y} &= 0\\ (\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I})\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{y}\\ \boldsymbol{\beta} &= (\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned} \tag{10}$$

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

2 Tikhonov Regression

We can further generalize the loss function used for Ridge Regression if we notice that:

$$\arg\min_x ||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 + \gamma||\boldsymbol{\beta}||_2^2$$

is equivalent to

$$\arg\min_x ||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 + ||\sigma\mathbf{I}\boldsymbol{\beta}||_2^2$$

where $\sigma = \sqrt{\gamma}$. We can generalize this by replacing $\sigma\mathbf{I}$ with various other *weight matrices* to ’focus’ the penalization on certain qualities or terms of $\boldsymbol{\beta}$.

For example, we can use forward differences to estimate the derivative of a vector by using the matrix:

$$\mathbf{D} = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \dots & 0\\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \vdots\\ \vdots & 0 & \ddots & \ddots & \ddots & 0\\ 0 & \dots & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

Note that $\mathbf{D}$ is $(n-2)\times n$. If we put this in our loss function, it becomes:

$$\arg\min_x ||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 + \lambda^2||\mathbf{D}\boldsymbol{\beta}||_2^2$$

Where λ is just a general weighting constant to control how much we want to penalize that last term. Now instead of just penalizing the magnitude of $\boldsymbol{\beta}$, we are penalizing its derivative - or in other words, forcing the resulting polynomial to be smooth. This will obviously change the solution of our estimator to solve the above loss function.

2.1 Exploring the Ridge Estimator

With our derivation of the Ridge Estimator, we can now implement and test the Ridge Estimator on some data. All code is in python 3, any random data was generated by using numpy's random number generator with seed 50. We will now construct some data that we will test our Ridge Estimator implementation on. We take 20 random samples from the line $y = 3x + 2$ with Gaussian noise from a standard normal distribtution. We now randomly sample 10 of these data points and will use them to fit/train our ridge regression model on, we will refer to these points as our training data, and the remaining 10 points we will call our validation data and will be used to measure the accuracy of our model using Residual Sum of Squares (also known as Sum of Squared Errors)

For this data we will try to find the line of best fit, that is find m and c in $y = mx + b$. We begin by trying $\gamma = 0$. This case reduces ridge regression to just normal linear regression/ordinary least squares. We also try $\gamma = 0.1$, this is the first real test of ridge regression. RSS compared below

From this we see that we achieved a lower RSS with $\gamma = 0.1$. That means that we have already improved over ordinary least squares! Ridge Regression has already shown to be an improvement. But we just tried some random γ value, are there other values of γ where the performance is even better? There's no analytic way to test this, so we try a whole lot of γ's and see what happens.

From this graph we see that our lowest RSS's occur for γ's from $[0, 10]$. We now try those γ's.

We achieve a min RSS of at $\gamma = 1$. Thus ridge regression gives us a model that performs about better than what we had with ordinary least square. That's quite an exciting result. Let's see if this translates to other functions/models.

We now consider $y = x^2$. We begin by taking 20 random samples from the interval $[-5, 5]$ and then splitting the data into our training data and validation data as we did for $y = 3x + 2$.

Now we want to try and fit our data to the model $y =$. We try a similar approach as we did before, testing

2.2 Deriving the Tikhonov Estimator

As we saw when deriving the ridge estimator, to find this estimator we want to solve

$$\frac{d}{d\boldsymbol{\beta}}\left[||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 + \lambda^2||\mathbf{D}\boldsymbol{\beta}||_2^2\right] = 0$$

or, if we expand that similar to the loss function for ridge estimation,

$$\frac{d}{d\boldsymbol{\beta}}\left[(\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\boldsymbol{\beta}) - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y} + \lambda^2(\mathbf{D}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{D}\boldsymbol{\beta})\right] = 0$$

From the ridge estimator derivation we already know that

$$\frac{d}{d\boldsymbol{\beta}}[(\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}] = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$
$$\frac{d}{d\boldsymbol{\beta}}[(2\mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}] = 2\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

So what is $\frac{d}{d\boldsymbol{\beta}}[(\mathbf{D}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}]$? Well,

$$\mathbf{D}\boldsymbol{\beta} = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \dots & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} \beta_2 - \beta_0 \\ \beta_3 - \beta_1 \\ \beta_4 - \beta_2 \\ \vdots \\ \beta_n - \beta_{n-2} \end{bmatrix}$$

So it follows that

$$(\mathbf{D}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{D}\boldsymbol{\beta}) = \begin{bmatrix} \frac{\beta_2-\beta_0}{2} \\ \frac{\beta_3-\beta_1}{2} \\ \frac{\beta_4-\beta_2}{2} \\ \vdots \\ \frac{\beta_n-\beta_{n-2}}{2} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \frac{\beta_2-\beta_0}{2} \\ \frac{\beta_3-\beta_1}{2} \\ \frac{\beta_4-\beta_2}{2} \\ \vdots \\ \frac{\beta_n-\beta_{n-2}}{2} \end{bmatrix}$$

$$= \left(\frac{(\beta_2 - \beta_0)^2}{4} + \frac{(\beta_3 - \beta_1)^2}{4} + \dots \frac{(\beta_n - \beta_{n-2})^2}{4} \right)$$

Then it follows that

$$\frac{d}{d\boldsymbol{\beta}} \left[(\mathbf{D}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{D}\boldsymbol{\beta}) \right] = \begin{bmatrix} -\frac{\beta_2-\beta_0}{2} \\ \\ -\frac{\beta_3-\beta_2}{2} \\ \\ \frac{\beta_2-\beta_0}{2} - \frac{\beta_4-\beta_2}{2} \\ \vdots \\ \frac{\beta_{n-2}-\beta_{n-4}}{2} - \frac{\beta_n-\beta_{n-2}}{2} \\ \\ \frac{\beta_{n-1}-\beta_{n-3}}{2} \\ \\ \frac{\beta_n-\beta_n-2}{2} \end{bmatrix}$$

This initially doesn't seem like anything useful, but note that

$$\mathbf{D}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta} = \begin{bmatrix} -\frac{1}{2} & 0 & 0 & \dots & 0 \\ 0 & -\frac{1}{2} & 0 & \dots & 0 \\ \frac{1}{2} & 0 & \ddots & \dots & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ \vdots & 0 & \ddots & 0 & -\frac{1}{2} \\ 0 & \dots & 0 & \frac{1}{2} & 0 \\ 0 & \dots & 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{\beta_2-\beta_0}{2} \\ \frac{\beta_3-\beta_1}{2} \\ \vdots \\ \frac{\beta_n-\beta_{n-2}}{2} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{\beta_2-\beta_0}{4} \\ \\ -\frac{\beta_3-\beta_1}{4} \\ \\ \frac{\beta_2-\beta_0}{4} - \frac{\beta_4-\beta_2}{4} \\ \\ \vdots \\ \frac{\beta_{n-2}-\beta_{n-4}}{4} - \frac{\beta_n-\beta_{n-2}}{4} \\ \\ \frac{\beta_{n-1}-\beta_{n-3}}{4} \\ \\ \frac{\beta_n-\beta_n-2}{4} \end{bmatrix}$$

So, it's true that

$$\frac{d}{d\boldsymbol{\beta}}\left[(\mathbf{D}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{D}\boldsymbol{\beta})\right] = 2\mathbf{D}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}$$

Finally we can solve for the $\boldsymbol{\beta}$ that satisfies

$$\frac{d}{d\boldsymbol{\beta}}\left[||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 + \lambda^2||\mathbf{D}\boldsymbol{\beta}||_2^2\right] = 0$$

or,

$$\begin{aligned} 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta} &= 0 \\ \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^{\mathrm{T}}\mathbf{y} + \lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta} &= 0 \\ \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta} &= \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D})\boldsymbol{\beta} &= \mathbf{X}^{\mathrm{T}}\mathbf{y} \end{aligned}$$

Thus

$$\boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Generally speaking, a loss function with weight matrix $\mathbf{\Gamma}$ of the form

$$\arg\min_x ||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 + ||\Gamma\boldsymbol{\beta}||_2^2$$

will be solved by

$$\boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

but weight matrices can be customized to such a degree that it's best to verify this property for each case.