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:

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

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

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

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 $\mathbf{x}^T\mathbf{A}^T\mathbf{b} = \mathbf{b}^T\mathbf{A}\mathbf{x}$. Note first that $(\mathbf{b}^T\mathbf{A}\mathbf{x})^T = \mathbf{x}^T\mathbf{A}^T\mathbf{b}$. Further, $dim(\mathbf{x}^T\mathbf{A}^T\mathbf{b}) = (1)\times(1) = dim(\mathbf{b}^T\mathbf{A}\mathbf{x})$. Also, note that the transpose of a 1×1 matrix is the same matrix: $[c]^T = [c]$. It then follows that $\mathbf{x}^T\mathbf{A}^T\mathbf{b} = \mathbf{b}^T\mathbf{A}\mathbf{x}$, completing the proof.

So,

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

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

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

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

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

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

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

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{A}\mathbf{x})^T\mathbf{A}\mathbf{x} = \\
&= \begin{bmatrix} \frac{d}{da_0}(\mathbf{A}\mathbf{x})^T\mathbf{A}\mathbf{x} \\ \frac{d}{da_1}(\mathbf{A}\mathbf{x})^T\mathbf{A}\mathbf{x} \\ \vdots \\ \frac{d}{da_m}(\mathbf{A}\mathbf{x})^T\mathbf{A}\mathbf{x} \end{bmatrix} \\
&= \begin{bmatrix} 2(a_0 + a_1x_0 + \dots + a_mx_0^m) + 2(a_0 + a_1x_1 + \dots + a_mx_1^m) + \ldots + 2(a_0 + a_1x_n + \dots + a_mx_n^m) \\ 2x_0(a_0 + a_1x_0 + \dots + a_mx_0^m) + 2x_1(a_0 + a_1x_1 + \dots + a_mx_1^m) + \ldots + 2x_n(a_0 + a_1x_n + \dots + a_mx_n^m) \\ \vdots \\ 2x_0^m(a_0 + a_1x_0 + \dots + a_mx_0^m) + 2x_1^m(a_0 + a_1x_1 + \dots + a_mx_1^m) + \ldots + 2x_n^m(a_0 + a_1x_n + \dots + 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 & \dots & \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 & \dots & \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 & \dots & \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} & \dots & \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 & \dots & 1 \\ x_0 & x_1 & x_2 & \dots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \dots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^m & x_1^m & x_2^m & \dots & x_n^m \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 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \\
&= 2\mathbf{A}^T\mathbf{A}\mathbf{x}
\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 $\mathbf{b}$.

2 Tikhonov Regression

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

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

is equivalent to

$$\arg\min_x ||\mathbf{Ax} - \mathbf{b}||_2^2 + ||\sigma\mathbf{Ix}||_2^2$$

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 $\mathbf{x}$.

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:

$$\arg\min_x ||\mathbf{Ax} - \mathbf{b}||_2^2 + \lambda^2||\mathbf{Dx}||_2^2$$

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 $\mathbf{x}$, 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 Deriving the Tikhonov Estimator

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

$$\frac{d}{d\mathbf{x}}\left[||\mathbf{Ax}-\mathbf{b}||_2^2+\lambda^2||\mathbf{Dx}||_2^2\right]=0$$

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

$$\frac{d}{d\mathbf{x}}\left[(\mathbf{Ax})^T(\mathbf{Ax})-2\mathbf{b}^T\mathbf{Ax}+\mathbf{b}^T\mathbf{b}+\lambda^2(\mathbf{Dx})^{\mathrm{T}}(\mathbf{Dx})\right]=0$$

From the ridge estimator derivation we already know that

$$\frac{d}{d\mathbf{x}}[(\mathbf{Ax})^{\mathrm{T}}\mathbf{Ax}]=2\mathbf{A}^{\mathrm{T}}\mathbf{Ax}$$
$$\frac{d}{d\mathbf{x}}[(2\mathbf{b}^{\mathrm{T}}\mathbf{Ax}]=2\mathbf{A}^{\mathrm{T}}\mathbf{b}$$

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

$$\mathbf{Dx}=\begin{bmatrix}-\frac{1}{2} & 0 & \frac{1}{2} & 0 & \dots & 0\\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \vdots\\ \vdots & 0 & \ddots & \ddots & \ddots & 0\\ 0 & \dots & 0 & -\frac{1}{2} & 0 & \frac{1}{2}\end{bmatrix}\begin{bmatrix}a_0\\a_1\\\vdots\\a_n\end{bmatrix}$$

$$=\frac{1}{2}\begin{bmatrix}a_2-a_0\\a_3-a_1\\a_4-a_2\\\vdots\\a_n-a_{n-2}\end{bmatrix}$$

So it follows that

$$(\mathbf{Dx})^{\mathrm{T}}(\mathbf{Dx})=\begin{bmatrix}\frac{a_2-a_0}{2}\\\frac{a_3-a_1}{2}\\\frac{a_4-a_2}{2}\\\vdots\\\frac{a_n-a_{n-2}}{2}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\frac{a_2-a_0}{2}\\\frac{a_3-a_1}{2}\\\frac{a_4-a_2}{2}\\\vdots\\\frac{a_n-a_{n-2}}{2}\end{bmatrix}$$

$$=\left(\frac{(a_2-a_0)^2}{4}+\frac{(a_3-a_1)^2}{4}+\dots\frac{(a_n-a_{n-2})^2}{4}\right)$$

Then it follows that

$$\frac{d}{d\mathbf{x}}\left[(\mathbf{D}\mathbf{x})^{\mathrm{T}}(\mathbf{D}\mathbf{x})\right] = \begin{bmatrix} -\frac{a_2-a_0}{2} \\ \\ -\frac{a_3-a_2}{2} \\ \\ \frac{a_2-a_0}{2} - \frac{a_4-a_2}{2} \\ \vdots \\ \frac{a_{n-2}-a_{n-4}}{2} - \frac{a_n-a_{n-2}}{2} \\ \\ \frac{a_{n-1}-a_{n-3}}{2} \\ \\ \frac{a_n-a_n-2}{2} \end{bmatrix}$$

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

$$\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{x} = \begin{bmatrix} -\frac{1}{2} & 0 & 0 & \dots & 0 \\ 0 & -\frac{1}{2} & 0 & \dots & 0 \\ \frac{1}{2} & 0 & \ddots & \dots & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ \vdots & 0 & \ddots & 0 & -\frac{1}{2} \\ 0 & \dots & 0 & \frac{1}{2} & 0 \\ 0 & \dots & 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{a_2-a_0}{2} \\ \frac{a_3-a_1}{2} \\ \vdots \\ \frac{a_n-a_{n-2}}{2} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{a_2-a_0}{4} \\ \\ -\frac{a_3-a_1}{4} \\ \\ \frac{a_2-a_0}{4} - \frac{a_4-a_2}{4} \\ \\ \vdots \\ \frac{a_{n-2}-a_{n-4}}{4} - \frac{a_n-a_{n-2}}{4} \\ \\ \frac{a_{n-1}-a_{n-3}}{4} \\ \\ \frac{a_n-a_n-2}{4} \end{bmatrix}$$

So, it's true that

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

Finally we can solve for the $\mathbf{x}$ that satisfies

$$\frac{d}{d\mathbf{x}}\left[||\mathbf{A}\mathbf{x}-\mathbf{b}||_2^2 + \lambda^2||\mathbf{D}\mathbf{x}||_2^2\right] = 0$$

or,

$$\begin{aligned}
2\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} - 2\mathbf{A}^{\mathrm{T}}\mathbf{b} + 2\lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{x} &= 0 \\
\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{A}^{\mathrm{T}}\mathbf{b} + \lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{x} &= 0 \\
\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{x} &= \mathbf{A}^{\mathrm{T}}\mathbf{b} \\
(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D})\mathbf{x} &= \mathbf{A}^{\mathrm{T}}\mathbf{b}
\end{aligned}$$

Thus

$$\mathbf{x} = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b}$$

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

$$\arg\min_{x} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2$$

will be solved by

$$\mathbf{x} = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Gamma})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b}$$

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