The Australian National University
School of Computing
Liang Zheng

Semester 2, 2022
Theory Assignment 2 of 4

COMP3670/6670: Introduction to Machine Learning

Release Date. 17 August 2022
Due Date. 00:30am, 19 September 2022
Maximum credit. 100

Exercise 1 **Conjectures** 5 credits each

Here are a collection of conjectures. Which are true, and which are false?

- If it is true, provide a formal proof demonstrating so.
- If it is false, give a counterexample, clearly stating why your counterexamples satisfies the premise but not the conclusion.

(No marks for just starting True/False.)

Hint: There's quite a few questions here, but each is relatively simple (the counterexamples aren't very complicated, and the proofs are short.) Try playing around with a few examples first to get an intuitive feeling if the statement is true before trying to prove it.

Let V be a vector space, and let $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ be an inner product over V.

1. Triangle inequality for inner products: For all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in V$, $\langle \mathbf{a}, \mathbf{c} \rangle \leq \langle \mathbf{a}, \mathbf{b} \rangle + \langle \mathbf{b}, \mathbf{c} \rangle$.

   Solution. False. Take $V = \mathbb{R}^2$ with the standard Euclidean dot product, and let $\mathbf{a} = \mathbf{c} = [1, 0]^T$, and $\mathbf{b} = [0, 1]^T$. Then $\mathbf{a} \cdot \mathbf{c} = 1$ but $\mathbf{a} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{c} = 0 + 0 = 0$, and $1 \not\leq 0$.

2. Transitivity of orthogonality: For all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in V$, if $\langle \mathbf{a}, \mathbf{b} \rangle = 0$ and $\langle \mathbf{b}, \mathbf{c} \rangle = 0$ then $\langle \mathbf{a}, \mathbf{c} \rangle = 0$.

   Solution. False. Take same counter example as previous question. Then $\mathbf{a} \cdot \mathbf{b} = 0$ and $\mathbf{b} \cdot \mathbf{c} = 0$ but $\mathbf{a} \cdot \mathbf{c} = 1$.

3. Orthogonality closed under addition: Suppose $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subseteq V$ is a set of vectors, and $\mathbf{x}$ is orthogonal to all of them (that is, for all $i = 1, 2, \ldots n$, $\langle \mathbf{x}, \mathbf{v_i} \rangle = 0$). Then $\mathbf{x}$ is orthogonal to any $\mathbf{y} \in \text{Span}(S)$.

   Solution. True. Suppose $\mathbf{y} \in \text{Span}(S)$. Then there exists coefficeints $c_1, \ldots, c_n$ such that

   $$\mathbf{y} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n$$

   Then,

   $$\langle \mathbf{x}, \mathbf{y} \rangle = \left\langle \mathbf{x}, \sum_i c_i \mathbf{v}_i \right\rangle = \sum_i c_i \langle \mathbf{x}, \mathbf{v}_i \rangle = \sum_i c_i \cdot 0 = 0$$

4. Let $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subseteq V$ be an **orthonormal** set of vectors in V. Then for all **non-zero** $\mathbf{x} \in V$, if for all $1 \leq i \leq n$ we have $\langle \mathbf{x}, \mathbf{v_i} \rangle = 0$ then $\mathbf{x} \notin \text{Span}(S)$.

   Solution. True. Suppose for a contradiction that $x \in \text{Span}(S)$. Then

   $$\mathbf{x} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n$$

   for some coefficients $c_1, \ldots, c_n$. There must exist some coefficient c_j that is non-zero (else $\mathbf{x}$ would be zero.) Then,

   $$\langle \mathbf{x}, \mathbf{v}_j \rangle = \left\langle \sum_i c_i \mathbf{v}_i, \mathbf{v}_j \right\rangle = \sum_i c_i \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \sum_i c_i \delta_{ij} = c_j \neq 0$$

   a contradiction.

5. Let $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subseteq V$ be a set of vectors in V (no assumption of orthonormality). Then for all **non-zero** $\mathbf{x} \in V$, if for all $1 \leq i \leq n$ we have $\langle \mathbf{x}, \mathbf{v}_i \rangle = 0$ then $\mathbf{x} \notin \text{Span}(S)$.

   Solution. Still holds True. Apply Gram-Schmidt to S to obtain a new set of orthonormal vectors S' with $\text{Span}(S) = \text{Span}(S')$. Clearly $\mathbf{x} \notin \text{Span}(S)$ iff $\mathbf{x} \notin \text{Span}(S')$. Repeat previous argument.

6. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a set of **orthonormal** vectors such that $\text{Span}(S) = V$, and let $\mathbf{x} \in V$. Then there is a *unique* set of coefficients $c_1, \ldots, c_n$ such that

$$\mathbf{x} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n$$

   Solution. True. Existence follows directly from the definition of Span. Now, suppose there exists two distinct choices of coefficients $(c_1, \ldots c_n)$ and $(d_1, \ldots, d_n)$ such that

$$\mathbf{x} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n = d_1\mathbf{v}_1 + \ldots + d_n\mathbf{v}_n$$

   Then there must exist some k such that $c_k \neq d_k$ (otherwise the coefficients would not be distinct.) Then,

$$\begin{aligned} &\langle \mathbf{x}, \mathbf{v}_k \rangle \\ = \left\langle \sum_i c_i\mathbf{v}_i, \mathbf{v}_k \right\rangle &= \left\langle \sum_i d_i\mathbf{v}_i, \mathbf{v}_k \right\rangle \\ \sum_i c_i \langle \mathbf{v}_i, \mathbf{v}_k \rangle &= \sum_i d_i \langle \mathbf{v}_i, \mathbf{v}_k \rangle \\ \sum_i c_i\delta_{ik} &= \sum_i d_i\delta_{ik} \\ c_k &= d_k \end{aligned}$$

   a contradiction.

7. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a set of vectors (no assumption of orthonormality) such that $\text{Span}(S) = V$, and let $\mathbf{x} \in V$. Then there is a *unique* set of coefficients $c_1, \ldots, c_n$ such that

$$\mathbf{x} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n$$

   Solution. False. We choose $V = \mathbb{R}^2$ and the set $S = \{\mathbf{v}_1 = [1, 0]^T, \mathbf{v}_2 = [0, 1]^T, \mathbf{v}_3 = [1, 1]^T\}$. Clearly $\text{Span}(S) = \mathbb{R}^2$ (as $[x, y]^T = x\mathbf{v}_1 + y\mathbf{v}_2$). Then

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1\mathbf{v}_1 + 1\mathbf{v}_2 + 0\mathbf{v}_3 = 0\mathbf{v}_1 + 0\mathbf{v}_2 + 1\mathbf{v}_3$$

8. Let $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subseteq V$ be a set of vectors. If all the vectors are pairwise linearly independent (i.e., for any $1 \leq i \neq j \leq n$, then only solution to $c_i\mathbf{v}_i + c_j\mathbf{v}_j = \mathbf{0}$ is the trivial solution $c_i = c_j = 0$.) then the set S is linearly independent.

   Solution. False, counter example $S = \{\mathbf{v}_1 = [1, 0]^T, \mathbf{v}_2 = [0, 1]^T, \mathbf{v}_3 = [1, 1]^T\}$. No vector is a multiple of another vector, so they are all pairwise linearly independent. But the entire set is not, as

$$1\mathbf{v}_1 + 1\mathbf{v}_2 + (-1)\mathbf{v}_3 = \mathbf{0}$$

Exercise 2 **Inner Products induce Norms** 20 credits

Let V be a vector space, and let $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ be an inner product on V. Define $||\mathbf{x}|| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. Prove that $|| \cdot ||$ is a norm.

(Hint: To prove the triangle inequality holds, you may need the Cauchy-Schwartz inequality, $\langle \mathbf{x}, \mathbf{y} \rangle \leq ||\mathbf{x}||||\mathbf{y}||$.)

Solution. We verify the three norm axioms.

1. **Absolutely homogeneous**

$$||\lambda \mathbf{x}|| = \sqrt{\langle \lambda \mathbf{x}, \lambda \mathbf{x} \rangle} = \sqrt{\lambda^2 \langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\lambda^2}\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = |\lambda| ||\mathbf{x}||$$

2. **Positive definiteness**
Follows trivially by positive definiteness of the inner product.

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \geq 0 \text{ as } \langle \mathbf{x}, \mathbf{x} \rangle \geq 0$$

$$\|\mathbf{x}\| = 0 \Leftrightarrow \|\mathbf{x}\|^2 = 0 \Leftrightarrow \langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$$

3. **Triangle Inequality**
This problem is easiest to solve by starting with the triangle inequality $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$, and working towards the Cauchy-Schwartz inequality. We can then reverse the proof.

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{y} \rangle &\leq \|\mathbf{x}\| \, \|\mathbf{y}\| \\
2\langle \mathbf{x}, \mathbf{y} \rangle &\leq 2\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}\sqrt{\langle \mathbf{y}, \mathbf{y} \rangle} \\
\langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle &\leq \langle \mathbf{x}, \mathbf{x} \rangle + 2\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}\sqrt{\langle \mathbf{y}, \mathbf{y} \rangle} + \langle \mathbf{y}, \mathbf{y} \rangle \\
\langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle &\leq (\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} + \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle})^2 \\
\langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle &\leq (\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} + \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle})^2 \\
\|\mathbf{x} + \mathbf{y}\|^2 &\leq (\|\mathbf{x}\| + \|\mathbf{y}\|)^2 \\
\|\mathbf{x} + \mathbf{y}\| &\leq \|\mathbf{x}\| + \|\mathbf{y}\|
\end{aligned}$$

Exercise 3 **General Linear Regression with Regularisation** (10+10+10+5+5 credits)

Let $\mathbf{A} \in \mathbb{R}^{N \times N}, \mathbf{B} \in \mathbb{R}^{D \times D}$ be *symmetric, positive definite* matrices. From the lectures, we can use symmetric positive definite matrices to define a corresponding inner product, as shown below. We can also define a norm using the inner products.

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} &:= \mathbf{x}^T \mathbf{A} \mathbf{y} \\
\|\mathbf{x}\|_{\mathbf{A}}^2 &:= \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} \\
\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{B}} &:= \mathbf{x}^T \mathbf{B} \mathbf{y} \\
\|\mathbf{x}\|_{\mathbf{B}}^2 &:= \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{B}}
\end{aligned}$$

Suppose we are performing linear regression, with a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where for each i, $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. We can define the matrix

$$\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$$

and the vector

$$\mathbf{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N.$$

We would like to find $\boldsymbol{\theta} \in \mathbb{R}^D, \mathbf{c} \in \mathbb{R}^N$ such that $\mathbf{y} \approx \mathbf{X}\boldsymbol{\theta} + \mathbf{c}$, where the error is measured using $\| \cdot \|_{\mathbf{A}}$. We avoid overfitting by adding a weighted regularization term, measured using $||\cdot||_{\mathbf{B}}$. We define the loss function with regularizer:

$$\mathcal{L}_{\mathbf{A},\mathbf{B},\mathbf{y},\mathbf{X}}(\boldsymbol{\theta}, \mathbf{c}) = ||\mathbf{y} - \boldsymbol{X}\boldsymbol{\theta} - \mathbf{c}||_{\mathbf{A}}^2 + ||\boldsymbol{\theta}||_{\mathbf{B}}^2 + \|\mathbf{c}\|_{\mathbf{A}}^2$$

For the sake of brevity we write $\mathcal{L}(\boldsymbol{\theta}, \mathbf{c})$ for $\mathcal{L}_{\mathbf{A},\mathbf{B},\mathbf{y},\mathbf{X}}(\boldsymbol{\theta}, \mathbf{c})$.

HINTS:

- You may use (without proof) the property that a symmetric positive definite matrix is invertible.
- We assume that there are sufficiently many non-redundant data points for $\mathbf{X}$ to be full rank. In particular, you may assume that the null space of $\mathbf{X}$ is trivial (that is, the only solution to $\mathbf{Xz} = \mathbf{0}$ is the trivial solution, $\mathbf{z} = \mathbf{0}$.)
- You may use identities of gradients from the lectures slides, so long as you mention as such.

1. Find the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c})$.

   Solution.

   $$\begin{aligned}
   &\mathcal{L}(\boldsymbol{\theta}, \mathbf{c}) \\
   &= (\mathbf{y} - (\mathbf{X}\boldsymbol{\theta} + \mathbf{c}))^T \mathbf{A}(\mathbf{y} - (\mathbf{X}\boldsymbol{\theta} + \mathbf{c})) + \boldsymbol{\theta}^T \mathbf{B}\boldsymbol{\theta} + \mathbf{c}^T \mathbf{A}\mathbf{c} \\
   &= \mathbf{y}^T \mathbf{A}\mathbf{y} - \mathbf{y}^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta} + \mathbf{c}) - (\mathbf{X}\boldsymbol{\theta} + \mathbf{c})^T \mathbf{A}\mathbf{y} + (\mathbf{X}\boldsymbol{\theta} + \mathbf{c})^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta} + \mathbf{c}) + \boldsymbol{\theta}^T \mathbf{B}\boldsymbol{\theta} + \mathbf{c}^T \mathbf{A}\mathbf{c}
   \end{aligned}$$

   Note that $\mathbf{y}^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta} + \mathbf{c}) \in \mathbb{R}$, so $(\mathbf{y}^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta} + \mathbf{c}))^T = \mathbf{y}^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta} + \mathbf{c})$, giving $(\mathbf{X}\boldsymbol{\theta} + \mathbf{c})^T \mathbf{A}\mathbf{y} = \mathbf{y}^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta} + \mathbf{c})$.

   $$\begin{aligned}
   &= \mathbf{y}^T \mathbf{A}\mathbf{y} - 2(\mathbf{X}\boldsymbol{\theta} + \mathbf{c})^T \mathbf{A}\mathbf{y} + (\mathbf{X}\boldsymbol{\theta} + \mathbf{c})^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta} + \mathbf{c}) + \boldsymbol{\theta}^T \mathbf{B}\boldsymbol{\theta} + \mathbf{c}^T \mathbf{A}\mathbf{c} \\
   &= \mathbf{y}^T \mathbf{A}\mathbf{y} - 2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{A}\mathbf{y} - 2\mathbf{c}^T \mathbf{A}\mathbf{y} + (\mathbf{X}\boldsymbol{\theta})^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta}) + (\mathbf{X}\boldsymbol{\theta})^T \mathbf{A}\mathbf{c} + \mathbf{c}^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta}) + \mathbf{c}^T \mathbf{A}\mathbf{c} + \boldsymbol{\theta}^T \mathbf{B}\boldsymbol{\theta} + \mathbf{c}^T \mathbf{A}\mathbf{c}
   \end{aligned}$$

   Note that $(\mathbf{X}\boldsymbol{\theta})^T \mathbf{A}\mathbf{c} = ((\mathbf{X}\boldsymbol{\theta})^T \mathbf{A}\mathbf{c})^T = \mathbf{c}^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta})$

   $$\begin{aligned}
   &= \mathbf{y}^T \mathbf{A}\mathbf{y} - 2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{A}\mathbf{y} - 2\mathbf{c}^T \mathbf{A}\mathbf{y} + (\mathbf{X}\boldsymbol{\theta})^T \mathbf{A}(\mathbf{X}\boldsymbol{\theta}) + 2(\mathbf{c}^T \mathbf{A}\mathbf{X})\boldsymbol{\theta} + \mathbf{c}^T \mathbf{A}\mathbf{c} + \boldsymbol{\theta}^T \mathbf{B}\boldsymbol{\theta} + \mathbf{c}^T \mathbf{A}\mathbf{c} \\
   &= \mathbf{y}^T \mathbf{A}\mathbf{y} - 2\boldsymbol{\theta}^T (\mathbf{X}^T \mathbf{A}\mathbf{y}) - 2\mathbf{c}^T \mathbf{A}\mathbf{y} + \boldsymbol{\theta}^T (\mathbf{X}^T \mathbf{A}\mathbf{X})\boldsymbol{\theta} + 2(\mathbf{c}^T \mathbf{A}\mathbf{X})\boldsymbol{\theta} + 2\mathbf{c}^T \mathbf{A}\mathbf{c} + \boldsymbol{\theta}^T \mathbf{B}\boldsymbol{\theta}
   \end{aligned}$$

   Now, we can take the gradient with respect to $\boldsymbol{\theta}$, using the identity $\nabla_{\mathbf{x}}(\mathbf{x}^T \mathbf{A}\mathbf{x}) = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$ and $\nabla_{\mathbf{x}}(\mathbf{w}^T \mathbf{x}) = \nabla_{\mathbf{x}}(\mathbf{x}^T \mathbf{w}) = \mathbf{w}^T$.

   $$\begin{aligned}
   &\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c}) \\
   &= 0 - 2(\mathbf{X}^T \mathbf{A}\mathbf{y})^T - 0 + \boldsymbol{\theta}^T((\mathbf{X}^T \mathbf{A}\mathbf{X}) + (\mathbf{X}^T \mathbf{A}\mathbf{X})^T) + 2(\mathbf{c}^T \mathbf{A}\mathbf{X})^T + 0 + \boldsymbol{\theta}^T(\mathbf{B} + \mathbf{B}^T) \\
   &= -2\mathbf{y}^T \mathbf{A}\mathbf{X} + 2\boldsymbol{\theta}^T (\mathbf{X}^T \mathbf{A}\mathbf{X}) + 2\mathbf{X}^T \mathbf{A}\mathbf{c} + 2\boldsymbol{\theta}^T \mathbf{B}
   \end{aligned}$$

2. Let $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c}) = \mathbf{0}$, and solve for $\boldsymbol{\theta}$. If you need to invert a matrix to solve for $\boldsymbol{\theta}$, you should prove the inverse exists.

   Solution. Set the gradient to zero, and solve for $\boldsymbol{\theta}$.

   $$\begin{aligned}
   \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c}) = -2\mathbf{y}^T \mathbf{A}\mathbf{X} + 2\boldsymbol{\theta}^T (\mathbf{X}^T \mathbf{A}\mathbf{X}^T) + 2\mathbf{X}^T \mathbf{A}\mathbf{c} + 2\boldsymbol{\theta}^T \mathbf{B} &= \mathbf{0} \\
   \boldsymbol{\theta}^T (\mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B}) &= \mathbf{y}^T \mathbf{A}\mathbf{X} - \mathbf{X}^T \mathbf{A}\mathbf{c}
   \end{aligned}$$

   At this point, we need to show that $\mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B}$ is invertible. First, note that $\mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B}$ is symmetric, as

   $$(\mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B})^T = \mathbf{X}^T \mathbf{A}^T (\mathbf{X}^T)^T + \mathbf{B}^T = \mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B}$$

   Also note that $\mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B}$ is positive definite, as

   $$\mathbf{w}^T (\mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B})\mathbf{w} = (\mathbf{X}\mathbf{w})^T \mathbf{A}(\mathbf{X}\mathbf{w}) + \mathbf{w}^T \mathbf{B}\mathbf{w} = \|\mathbf{X}\mathbf{w}\|_{\mathbf{A}}^2 + \|\mathbf{w}\|_{\mathbf{B}}^2 \geq 0$$

   with equality $\|\mathbf{X}\mathbf{w}\|_{\mathbf{A}}^2 + \|\mathbf{w}\|_{\mathbf{B}}^2 = 0$ iff $\mathbf{X}\mathbf{w} = \mathbf{0}$ and $\mathbf{w} = \mathbf{0}$ iff $\mathbf{w} = \mathbf{0}$ (as the null space of $\mathbf{X}$ is trivial.) Hence, we have that $\mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B}$ is symmetric positive definite and therefore also invertible. Hence, we can write

   $$\begin{aligned}
   \boldsymbol{\theta}^T &= (\mathbf{y}^T \mathbf{A}\mathbf{X} - \mathbf{X}^T \mathbf{A}\mathbf{c})(\mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B})^{-1} \\
   \boldsymbol{\theta} &= (\mathbf{X}^T \mathbf{A}\mathbf{X} + \mathbf{B})^{-T} (\mathbf{X}^T \mathbf{A}\mathbf{y} - \mathbf{c}^T \mathbf{A}\mathbf{X})
   \end{aligned}$$

3. Find the gradient $\nabla_{\mathbf{c}}\mathcal{L}(\boldsymbol{\theta}, \mathbf{c})$.

   We now compute the gradient with respect to $\mathbf{c}$.

   Solution.

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}, \mathbf{c}) \\
&= \mathbf{y}^T\mathbf{A}\mathbf{y} - 2\boldsymbol{\theta}^T(\mathbf{X}^T\mathbf{A}\mathbf{y}) - 2\mathbf{c}^T\mathbf{A}\mathbf{y} + \boldsymbol{\theta}^T(\mathbf{X}^T\mathbf{A}\mathbf{X})\boldsymbol{\theta} + 2(\mathbf{c}^T\mathbf{A}\mathbf{X})\boldsymbol{\theta} + 2\mathbf{c}^T\mathbf{A}\mathbf{c} + \boldsymbol{\theta}^T\mathbf{B}\boldsymbol{\theta} \\
&\nabla_{\mathbf{c}}\mathcal{L}(\boldsymbol{\theta}, \mathbf{c}) \\
&= -2(\mathbf{A}\mathbf{y})^T + 2(\mathbf{A}\mathbf{X}\boldsymbol{\theta})^T + 2\mathbf{c}^T(\mathbf{A} + \mathbf{A}^T) \\
&= -2\mathbf{y}^T\mathbf{A} + 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{A} + 4\mathbf{c}^T\mathbf{A}
\end{aligned}$$

4. Let $\nabla_{\mathbf{c}}\mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$, and solve for $\mathbf{c}$. If you need to invert a matrix to solve for $\mathbf{c}$, you should prove the inverse exists.

   Solution.

$$\begin{aligned}
\nabla_{\mathbf{c}}\mathcal{L}(\boldsymbol{\theta}, \mathbf{c}) = -2\mathbf{y}^T\mathbf{A} + 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{A} + 4\mathbf{c}^T\mathbf{A} &= \mathbf{0} \\
\mathbf{c}^T(2\mathbf{A}) &= \mathbf{y}^T\mathbf{A} - \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{A}
\end{aligned}$$

   $2\mathbf{A}$ is symmetric positive definite, and is in particular invertible.

$$\begin{aligned}
\mathbf{c}^T &= (\mathbf{y}^T\mathbf{A} - \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{A})(2\mathbf{A})^{-1} \\
\mathbf{c} &= \frac{1}{2}\mathbf{A}^{-T}(\mathbf{A}\mathbf{y} - \mathbf{A}\mathbf{X}\boldsymbol{\theta})
\end{aligned}$$

5. Show that if we set $\mathbf{A} = \mathbf{I}, \mathbf{c} = \mathbf{0}, \mathbf{B} = \lambda\mathbf{I}$, where $\lambda \in \mathbb{R}$, your answer for 3.2 agrees with the analytic solution for the standard least squares regression problem with L2 regularization, given by

$$\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\mathbf{y}.$$

   Solution.

$$\begin{aligned}
\boldsymbol{\theta} &= (\mathbf{X}^T\mathbf{A}\mathbf{X} + \mathbf{B})^{-T}(\mathbf{X}^T\mathbf{A}\mathbf{y} - \mathbf{c}^T\mathbf{A}\mathbf{X}) \\
&= (\mathbf{X}^T\mathbf{I}\mathbf{X} + \lambda\mathbf{I})^{-T}(\mathbf{X}^T\mathbf{I}\mathbf{y} - \mathbf{0}^T\mathbf{A}\mathbf{X}) \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-T}\mathbf{X}^T\mathbf{y} \\
&= \left(\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^T\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left((\mathbf{X}^T\mathbf{X})^T + (\lambda\mathbf{I})^T\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$