Theory of Computation

Release Date: 18 August 2021
Due Date: 23:59pm, 19 September 2021
Maximum credit: 100

Exercise 1 **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 2 **Vector Calculus Identities** 10+10 credits

1. Let $\mathbf{x}, \mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Prove that $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{a}\mathbf{b}^T\mathbf{x}) = \mathbf{a}^T\mathbf{x}\mathbf{b}^T + \mathbf{b}^T\mathbf{x}\mathbf{a}^T$.

Solution. Compute each component of the derivative.

$$\begin{aligned}\frac{\partial}{\partial x_p}(\mathbf{x}^T\mathbf{a}\mathbf{b}^T\mathbf{x}) &= \frac{\partial}{\partial x_p}\sum_{j,k} x_j(\mathbf{a}\mathbf{b}^T)_{jk}x_k \\ &= \frac{\partial}{\partial x_p}\sum_{j,k} x_j a_j b_k x_k \\ &= \sum_{j,k} a_j b_k \frac{\partial}{\partial x_p} x_j x_k\end{aligned}$$

Note the following:

$$\frac{\partial(x_k x_j)}{\partial x_p} = \begin{cases} 2x_p & j = p = k \\ x_k & j = p \neq k \\ x_j & j \neq p = k \\ 0 & j \neq p \neq k \end{cases}$$

Hence, we can split the sum above,

$$\begin{aligned}\frac{\partial}{\partial x_p}(\mathbf{x}^T\mathbf{a}\mathbf{b}^T\mathbf{x}) = &\sum_{\substack{j,k \\ j=p=k}} a_j b_k \frac{\partial}{\partial x_p} x_j x_k + \sum_{\substack{j,k \\ j=p\neq k}} a_j b_k \frac{\partial}{\partial x_p} x_j x_k \\ &+ \sum_{\substack{j,k \\ j\neq p=k}} a_j b_k \frac{\partial}{\partial x_p} x_j x_k + \sum_{\substack{j,k \\ j\neq p\neq k}} a_j b_k \frac{\partial}{\partial x_p} x_j x_k \\ &= a_p b_p 2x_p + \sum_{k\neq p} a_p b_k x_k + \sum_{j\neq p} a_j b_p x_j\end{aligned}$$

Add the $a_p b_p x_p$ terms back into each summation,

$$\begin{aligned}&= \sum_k a_p b_k x_k + \sum_j a_j b_p x_j \\ &= \left(\sum_k b_k x_k\right) a_p + \left(\sum_j a_j x_j\right) b_p \\ &= (\mathbf{b}^T\mathbf{x})a_p + (\mathbf{a}^T\mathbf{x})b_p \\ &= ((\mathbf{b}^T\mathbf{x})\mathbf{a})_p + ((\mathbf{a}^T\mathbf{x})\mathbf{b})_p \\ &= (\mathbf{b}^T\mathbf{x}\mathbf{a} + \mathbf{a}^T\mathbf{x}\mathbf{b})_p\end{aligned}$$

Note that $(\mathbf{b}^T\mathbf{x}\mathbf{a} + \mathbf{a}^T\mathbf{x}\mathbf{b})_p = (\mathbf{b}^T\mathbf{x}\mathbf{a} + \mathbf{a}^T\mathbf{x}\mathbf{b})_p^T$, and since we want the result to be a row vector to dimensionally match the gradient, we choose the latter.

$$(\mathbf{b}^T\mathbf{x}\mathbf{a} + \mathbf{a}^T\mathbf{x}\mathbf{b})_p^T = (\mathbf{a}^T\mathbf{x}\mathbf{b}^T + \mathbf{b}^T\mathbf{x}\mathbf{a}^T)_p$$

since $\mathbf{b}^T\mathbf{x}$ and $\mathbf{a}^T\mathbf{x}$ are scalars. Hence,

$$\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{a}\mathbf{b}^T\mathbf{x}) = (\mathbf{a}^T\mathbf{x}\mathbf{b}^T + \mathbf{b}^T\mathbf{x}\mathbf{a}^T)$$

2. Let $\mathbf{B} \in \mathbb{R}^{n\times n}, \mathbf{x} \in \mathbb{R}^n$. Prove that $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{B}\mathbf{x}) = \mathbf{x}^T(\mathbf{B} + \mathbf{B}^T)$.

Solution. Compute each component of the derivative.

$$\begin{aligned}\frac{\partial}{\partial x_p}(\mathbf{x}^T\mathbf{B}\mathbf{x}) &= \frac{\partial}{\partial x_p}\sum_k x_k(\mathbf{B}\mathbf{x})_k \\ &= \frac{\partial}{\partial x_p}\sum_k x_k \sum_j B_{kj}x_j \\ &= \sum_{j,k} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p}\end{aligned}$$

Note the following:

$$\frac{\partial(x_k x_j)}{\partial x_p} = \begin{cases} 2x_p & p = k = j \\ x_k & p = j \neq k \\ x_j & p = k \neq j \\ 0 & p \neq k, p \neq j \end{cases}$$

Hence, we can split the sum above,

$$\begin{aligned}\frac{\partial}{\partial x_p}(\mathbf{x}^T\mathbf{B}\mathbf{x}) &= \sum_{\substack{j,k \\ p=k=j}} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p} + \sum_{\substack{j,k \\ p=j\neq k}} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p} \\ &+ \sum_{\substack{j,k \\ p=k\neq j}} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p} + \sum_{\substack{j,k \\ p\neq k, p\neq j}} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p} \\ &= B_{pp}2x_p + \sum_{\substack{k \\ p\neq k}} B_{kp}x_k + \sum_{\substack{j \\ p\neq j}} B_{pj}x_j\end{aligned}$$

Add the $B_{pp}x_p$ terms back into each summation,

$$\begin{aligned}&= \sum_k x_k B_{kp} + \sum_j x_j B_{pj} \\ &= (\mathbf{x}^T\mathbf{B})_p + \sum_j x_j(\mathbf{B}^T)_{jp} \\ &= (\mathbf{x}^T\mathbf{B})_p + (\mathbf{x}^T\mathbf{B}^T)_p \\ &= \mathbf{x}^T(\mathbf{B}+\mathbf{B}^T)_p\end{aligned}$$

Hence,

$$\nabla_x(\mathbf{x}^T\mathbf{B}\mathbf{x}) = \mathbf{x}^T(\mathbf{B}+\mathbf{B}^T)$$

Exercise 3 **Properties of Symmetric Positive Definiteness** 10 credits

Let $\mathbf{A}, \mathbf{B}$ be symmetric positive definite matrices. [1] Prove that for any $p, q > 0$ that $p\mathbf{A} + q\mathbf{B}$ is also symmetric and positive definite.

Solution. Let $\mathbf{A}, \mathbf{B}$ be symmetric positive definite. Then $p\mathbf{A} + q\mathbf{B}$ is symmetric, as

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

Also, $p\mathbf{A} + q\mathbf{B}$ is positive definite, as

$$\mathbf{w}^T(p\mathbf{A}+q\mathbf{B})\mathbf{w} = p\mathbf{w}^T\mathbf{A}\mathbf{w} + q\mathbf{w}^T\mathbf{B}\mathbf{w} \geq 0$$

[1] A matrix is *symmetric positive definite* if it is both symmetric and positive definite.

as $\mathbf{w}^T\mathbf{A}\mathbf{w} \geq 0$ and $\mathbf{w}^T\mathbf{B}\mathbf{w} \geq 0$. Now, clearly if $\mathbf{w} = \mathbf{0}$ then $\mathbf{w}^T(p\mathbf{A} + q\mathbf{B})\mathbf{w} = 0$. Conversly, if $\mathbf{w}^T(p\mathbf{A} + q\mathbf{B})\mathbf{w} = 0$ then $p\mathbf{w}^T\mathbf{A}\mathbf{w} + q\mathbf{w}^T\mathbf{B}\mathbf{w} = 0$. The sum of two non-negative terms is zero only when both terms are zero, hence $p\mathbf{w}^T\mathbf{A}\mathbf{w} = 0$. Since $p > 0$, we have $\mathbf{w}^T\mathbf{A}\mathbf{w} = 0$ which is true iff $\mathbf{w} = \mathbf{0}$ as $\mathbf{A}$ is positive definite. Hence, $p\mathbf{A} + q\mathbf{B}$ is symmetric positive definite.

Exercise 4 General Linear Regression with Regularisation (10+10+10+10+10 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. From the previous question, we can also define a norm using the inner products.

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

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:

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

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

For this question:

- 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 $\mathbf{X}\mathbf{z} = \mathbf{0}$ is the trivial solution, $\mathbf{z} = \mathbf{0}$.)

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

   Solution.

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

   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 $(\mathbf{X}\boldsymbol{\theta} + \mathbf{c})^T\mathbf{A}\mathbf{y} = \mathbf{y}^T\mathbf{A}(\mathbf{X}\boldsymbol{\theta} + \mathbf{c})$.

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

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

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

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

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

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 $\boldsymbol{\theta}$.

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

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

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

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

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

with equality $\|\mathbf{X}\mathbf{w}\|_{\mathbf{A}} = 0$ iff $\mathbf{X}\mathbf{w} = \mathbf{0}$ iff $\mathbf{w} = \mathbf{0}$ (as the null space of $\mathbf{X}$ is trivial.) Hence, we have that $\mathbf{X}^T\mathbf{A}\mathbf{X}$ is symmetric positive definite, and hence so is $\mathbf{X}^T\mathbf{A}\mathbf{X}+\mathbf{B}$ (by the previous question) 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.

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

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.

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

$\mathbf{A}+\mathbf{B}$ is symmetric positive definite by the previous question, and is in particular invertible.

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

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 4.2 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.

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