COMP3670 2021 Theory Assignment 2

Yuxuan Lin – u6828533@anu.edu.au

September 19, 2021

Introduction to Machine Learning

By turning in this assignment, I agree by the ANU honor code and declare that all of this is my own work.

Exercise 1

(i) *Positive definite*

$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \geq 0$, and if $\sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = 0$ then $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$. Since $\langle \cdot, \cdot \rangle$ is an inner product, from inner product's *positive definiteness*, we have $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0 \iff \boldsymbol{x} = \mathbf{0}$.

Hence, $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \mathbf{0}$, then $\| \cdot \|$ satisfies *positive definite*.

(ii) *Absolutely homogeneous*

Since $\langle \cdot, \cdot \rangle$ is an inner product, from inner product axioms, we have

$$\begin{aligned}
\forall \lambda \in \mathbb{R}, \|\lambda \boldsymbol{x}\| &= \sqrt{\langle \lambda \boldsymbol{x}, \lambda \boldsymbol{x} \rangle} \\
&= \sqrt{\lambda \langle \boldsymbol{x}, \lambda \boldsymbol{x} \rangle} \ldots\ldots \text{(homogeneity in argument 1)} \\
&= \sqrt{\lambda^2 \langle \boldsymbol{x}, \boldsymbol{x} \rangle} \ldots\ldots\ldots \text{(homogeneity in argument 2)} \\
&= |\lambda| \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \\
&= |\lambda| \|\boldsymbol{x}\|
\end{aligned}$$

Hence, $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$, then $\| \cdot \|$ satisfies *absolutely homogeneous*.

(iii) *Triangle inequality*

Since $\langle \cdot, \cdot \rangle$ is an inner product, from inner product axioms, we have

$$\begin{aligned}
\|\boldsymbol{x}+\boldsymbol{y}\| &= \sqrt{\langle \boldsymbol{x}+\boldsymbol{y}, \boldsymbol{x}+\boldsymbol{y} \rangle} \\
&= \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}+\boldsymbol{y} \rangle + \langle \boldsymbol{y}, \boldsymbol{x}+\boldsymbol{y} \rangle} \text{.....................(linearity in argument 1)} \\
&= \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle + \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{y}, \boldsymbol{x} \rangle + \langle \boldsymbol{y}, \boldsymbol{y} \rangle} \text{.............(linearity in argument 2)} \\
&= \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle + \langle \boldsymbol{y}, \boldsymbol{y} \rangle + 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle} \text{.................. (symmetry of arguments)} \\
&= \sqrt{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 + 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle} \\
&\leq \sqrt{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 + 2\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \text{................(Cauchy-Schwarz inequality)} \\
&\leq \sqrt{(\|\boldsymbol{x}\| + \|\boldsymbol{y}\|)^2} \\
&\leq \big| \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \big| \\
&\leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \text{........................ (positive definiteness) : } \|\boldsymbol{x}\|, \|\boldsymbol{y}\| \geq 0
\end{aligned}$$

Hence, $\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$, then $\| \cdot \|$ satisfies *triangle inequality.*

In sum, $\| \cdot \|$ is a norm by fulfilling the three norm axioms.

Exercise 2.1

Since $x, a, b \in R^n$, $x^T a b^T x \in R^{1\times n} \cdot R^{n\times 1} \cdot R^{1\times n} \cdot R^{n\times 1} = R$, $x^T a b^T x \in R$.

Set $f(x) := x^T a b^T x$, for $h > 0$:

$$
\begin{aligned}
f(x+h) &= (x+h)^T a b^T (x+h) \\
&= (x^T + h^T) a b^T (x+h) \quad \cdots \text{(mml-book p.25)} \\
&= (x^T a b^T + h^T a b^T)(x+h) \quad \cdots \text{(distributivity)} \\
&= x^T a b^T x + x^T a b^T h + h^T a b^T x + h^T a b^T h
\end{aligned}
$$

Then, $\nabla_x (f(x)) = \lim_{h \to 0} \dfrac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \dfrac{x^T a b^T h + h^T a b^T x + h^T a b^T h}{h}$

$$
\begin{aligned}
&= x^T a b^T + (a b^T x)^T \quad \cdots \lim_{h \to 0} h^T a b^T = 0 \\
&= x^T a b^T + x^T b a^T \quad \cdots \text{(mml-book p.25)}
\end{aligned}
$$

Since $x^T a \in R^{1\times n} \cdot R^{n\times 1} = R$, $(x^T a)^T = x^T a$, then $x^T a = a^T x$, similarly $x^T b = b^T x$.

We have proved that $\nabla_x (x^T a b^T x) = a^T x b^T + b^T x a^T$.

Exercise 2.2

Since $x \in R^n$, $B \in R^{n\times n}$, $x^T B x \in R^{1\times n} \cdot R^{n\times n} \cdot R^{n\times 1} = R$, $x^T B x \in R$

Set $f(x) := x^T B x$, for $h > 0$:

$$f(x+h) = (x+h)^T B (x+h)$$
$$= (x^T + h^T) B (x+h) \quad \cdots \text{(mml-book p.25)}$$
$$= (x^T B + h^T B)(x+h) \quad \cdots \text{(distributivity)}$$
$$= x^T B x + x^T B h + h^T B x + h^T B h$$

Then,
$$\nabla_x (f(x)) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{x^T B h + h^T B x + h^T B h}{h}$$

$$= x^T B + (Bx)^T \quad \cdots \lim_{h\to 0} h^T B = 0$$
$$= x^T B + x^T B^T$$
$$= x^T (B + B^T)$$, we proved that

$$\nabla_x (x^T B x) = x^T (B + B^T) .$$

Exercise 3

(i) *Symmetric*

Since $\boldsymbol{A}$, $\boldsymbol{B}$ are symmetric matrices, then $\boldsymbol{A} = \boldsymbol{A}^\top$, $\boldsymbol{B} = \boldsymbol{B}^\top$.

$$\begin{aligned}(p\boldsymbol{A} + q\boldsymbol{B})^\top &= (p\boldsymbol{A})^\top + (q\boldsymbol{B})^\top \ldots \text{(mml-book p.25)} \\ &= p\boldsymbol{A}^\top + q\boldsymbol{B}^\top \ldots\ldots \text{(mml-book p.26)} \\ &= p\boldsymbol{A} + q\boldsymbol{B}\end{aligned}$$

Hence, $p\boldsymbol{A} + q\boldsymbol{B} = (p\boldsymbol{A} + q\boldsymbol{B})^\top$, then $p\boldsymbol{A} + q\boldsymbol{B}$ satisfies *symmetry.*

(ii) *Positive definiteness*

$\forall \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\}$:

$$\begin{aligned}\boldsymbol{x}^\top (p\boldsymbol{A} + q\boldsymbol{B})\boldsymbol{x} &= (p\boldsymbol{x}^\top \boldsymbol{A} + q\boldsymbol{x}^\top \boldsymbol{B})\boldsymbol{x} \ldots \text{(distributivity)} \\ &= p\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + q\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} \ldots \text{(distributivity)}\end{aligned}$$

Since $\boldsymbol{A}$, $\boldsymbol{B}$ are positive definite matrices, then $\forall \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\} : \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} > 0$, and notice that $p, q > 0$, we derive $p\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + q\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} > 0$.

Hence, $\forall \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\} : \boldsymbol{x}^\top (p\boldsymbol{A} + q\boldsymbol{B})\boldsymbol{x} > 0$, then $p\boldsymbol{A} + q\boldsymbol{B}$ satisfies *positive definiteness.*

In sum, $p\boldsymbol{A} + q\boldsymbol{B}$ is symmetric and postive definite.

Exercise 4

From the given equations we derive $\|\boldsymbol{x}\|^2_{\boldsymbol{A}} = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, $\|\boldsymbol{x}\|^2_{\boldsymbol{B}} = \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}$. Then we can transform the loss function with regularizer as below

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{c}) &= \|\boldsymbol{y} - \boldsymbol{X\theta} - \boldsymbol{c}\|^2_{\boldsymbol{A}} + \|\boldsymbol{\theta}\|^2_{\boldsymbol{B}} + \|\boldsymbol{c}\|^2_{\boldsymbol{A}} \\ &= (\boldsymbol{y} - \boldsymbol{X\theta} - \boldsymbol{c})^\top \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{X\theta} - \boldsymbol{c}) + \boldsymbol{\theta}^\top \boldsymbol{B\theta} + \boldsymbol{c}^\top \boldsymbol{A c} \\ &= (\boldsymbol{y}^\top - (\boldsymbol{X\theta})^\top - \boldsymbol{c}^\top)\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{X\theta} - \boldsymbol{c}) + \boldsymbol{\theta}^\top \boldsymbol{B\theta} + \boldsymbol{c}^\top \boldsymbol{A c} \ldots \text{(mml-book p.25)} \\ &= (\boldsymbol{y}^\top \boldsymbol{A} - (\boldsymbol{X\theta})^\top \boldsymbol{A} - \boldsymbol{c}^\top \boldsymbol{A})(\boldsymbol{y} - \boldsymbol{X\theta} - \boldsymbol{c}) + \boldsymbol{\theta}^\top \boldsymbol{B\theta} + \boldsymbol{c}^\top \boldsymbol{A c} \ldots \text{(distributivity)} \\ &= \boldsymbol{y}^\top \boldsymbol{A y} - \boldsymbol{y}^\top \boldsymbol{A X\theta} - \boldsymbol{y}^\top \boldsymbol{A c} - (\boldsymbol{X\theta})^\top \boldsymbol{A y} + (\boldsymbol{X\theta})^\top \boldsymbol{A X\theta} + (\boldsymbol{X\theta})^\top \boldsymbol{A c} \\ &\quad - \boldsymbol{c}^\top \boldsymbol{A y} + \boldsymbol{c}^\top \boldsymbol{A X\theta} + \boldsymbol{c}^\top \boldsymbol{A c} + \boldsymbol{\theta}^\top \boldsymbol{B\theta} + \boldsymbol{c}^\top \boldsymbol{A c}\end{aligned}$$

We know that $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{B} \in \mathbb{R}^{D\times D}$, $\boldsymbol{X} \in \mathbb{R}^{N\times D}$, $\boldsymbol{y} \in \mathbb{R}^{N}$, $\boldsymbol{\theta} \in \mathbb{R}^{D}$, $\boldsymbol{c} \in \mathbb{R}^{N}$.

Since $\boldsymbol{A}$, $\boldsymbol{B}$ are symmetric matrices, then $\boldsymbol{A} = \boldsymbol{A}^\top$, $\boldsymbol{B} = \boldsymbol{B}^\top$.

For $\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{X}\boldsymbol{\theta}$, by dot product we derive $\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{X}\boldsymbol{\theta} \in \mathbb{R}$, then

$$\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{X}\boldsymbol{\theta} = (\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{X}\boldsymbol{\theta})^\top = \boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A}^\top \boldsymbol{y} = (\boldsymbol{X}\boldsymbol{\theta})^\top \boldsymbol{A}\boldsymbol{y}$$

For $\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{c}$, by dot product we derive $\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{c} \in \mathbb{R}$, then

$$\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{c} = (\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{c})^\top = \boldsymbol{c}^\top \boldsymbol{A}^\top \boldsymbol{y} = \boldsymbol{c}^\top \boldsymbol{A}\boldsymbol{y}$$

For $(\boldsymbol{X}\boldsymbol{\theta})^\top \boldsymbol{A}\boldsymbol{c}$, by dot product we derive $(\boldsymbol{X}\boldsymbol{\theta})^\top \boldsymbol{A}\boldsymbol{c} \in \mathbb{R}$, then

$$(\boldsymbol{X}\boldsymbol{\theta})^\top \boldsymbol{A}\boldsymbol{c} = ((\boldsymbol{X}\boldsymbol{\theta})^\top \boldsymbol{A}\boldsymbol{c})^\top = \boldsymbol{c}^\top \boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{c}^\top \boldsymbol{A}\boldsymbol{X}\boldsymbol{\theta}$$

Use above three equations to simplify the loss function equation as below

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

1. Use two useful identities for computing gradients, equation (5.104) and equation (5.105) from **mml-book** p.158, plus what we have proved in **Exercise 2**, we derive

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{X}\boldsymbol{\theta}) = \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{X}$$

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}\boldsymbol{\theta}) = \boldsymbol{\theta}^\top(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} + (\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X})^\top) = \boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A}^\top \boldsymbol{X} = 2\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}$$

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{c}) = (\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{c})^\top = \boldsymbol{c}^\top \boldsymbol{A}^\top \boldsymbol{X} = \boldsymbol{c}^\top \boldsymbol{A}\boldsymbol{X}$$

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^\top \boldsymbol{B}\boldsymbol{\theta}) = \boldsymbol{\theta}^\top(\boldsymbol{B} + \boldsymbol{B}^\top) = \boldsymbol{\theta}^\top(\boldsymbol{B} + \boldsymbol{B}) = 2\boldsymbol{\theta}^\top \boldsymbol{B}$$

Hence, the gradient of the loss function with respect to $\boldsymbol{\theta}$ is

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

2. Let $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{c}) = 0$, we derive

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

For the next, we will prove that $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}$ is invertible, by proving it is a symmetric positive definite matrix.

Symmetry

$$(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B})^\top = \boldsymbol{X}^\top\boldsymbol{A}^\top\boldsymbol{X} + \boldsymbol{B}^\top = \boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}$$

Hence, it is symmetric.

Positive definiteness

$\forall \boldsymbol{y} \in V\backslash\{\boldsymbol{0}\}$:

$$\boldsymbol{y}^\top(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B})\boldsymbol{y} = \boldsymbol{y}^\top(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X})\boldsymbol{y} + \boldsymbol{y}^\top\boldsymbol{B}\boldsymbol{y} \ldots \text{(distributivity)}$$

Since $\|\boldsymbol{X}\boldsymbol{y}\|^2_{\boldsymbol{A}} \geq 0$, we derive

$$\|\boldsymbol{X}\boldsymbol{y}\|^2_{\boldsymbol{A}} = (\boldsymbol{X}\boldsymbol{y})^\top\boldsymbol{A}(\boldsymbol{X}\boldsymbol{y}) = (\boldsymbol{y}^\top\boldsymbol{X}^\top)\boldsymbol{A}(\boldsymbol{X}\boldsymbol{y}) = \boldsymbol{y}^\top(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X})\boldsymbol{y} \geq 0$$

Since the only solution to $\boldsymbol{X}\boldsymbol{y} = \boldsymbol{0}$ is the trivial solution $\boldsymbol{y} = \boldsymbol{0}$, but as assumption, $\boldsymbol{y} \neq \boldsymbol{0}$, thus $\boldsymbol{X}\boldsymbol{y} \neq \boldsymbol{0}$. From the postive definiteness of norm, $\|\boldsymbol{X}\boldsymbol{y}\|^2_{\boldsymbol{A}} > 0$. Hence, $\boldsymbol{y}^\top(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X})\boldsymbol{y} > 0$.

Also, since $\boldsymbol{B}$ is a positive definite matrix, from definition we have $\boldsymbol{y}^\top\boldsymbol{B}\boldsymbol{y} > 0$. Then, $\forall \boldsymbol{y} \in V\backslash\{\boldsymbol{0}\} : \boldsymbol{y}^\top(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B})\boldsymbol{y} = \boldsymbol{y}^\top(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X})\boldsymbol{y} + \boldsymbol{y}^\top\boldsymbol{B}\boldsymbol{y} > 0$.

In sum, we have proved that $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}$ is is a symmetric positive definite matrix, thus it is invertible.

Return to the calculation of $\boldsymbol{\theta}$, we can now derive

$$\boldsymbol{\theta} = (\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B})^{-1}(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{c})$$

3. Similar as **Exercise 4.1**, use two useful identities for computing gradients, equation (5.104) and equation (5.105) from **mml-book** p.158, plus what we have proved in **Exercise 2**, we derive the gradient of the loss function as below

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

Hence, the gradient of the loss function with respect to $\boldsymbol{c}$ is

$$\nabla_{\boldsymbol{c}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{c}) = -2\boldsymbol{y}^\top \boldsymbol{A} + 2\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A} + 4\boldsymbol{c}^\top \boldsymbol{A}$$

4. Let $\nabla_{\boldsymbol{c}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{c}) = 0$, we derive

$$\begin{aligned}
-2\boldsymbol{y}^\top \boldsymbol{A} + 2\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A} + 4\boldsymbol{c}^\top \boldsymbol{A} &= 0 \\
\boldsymbol{c}^\top \boldsymbol{A} &= \frac{1}{2}(\boldsymbol{y}^\top \boldsymbol{A} - \boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A}) \\
(\boldsymbol{c}^\top \boldsymbol{A})^\top &= (\frac{1}{2}(\boldsymbol{y}^\top \boldsymbol{A} - \boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{A}))^\top \\
\boldsymbol{A}^\top \boldsymbol{c} &= \frac{1}{2}(\boldsymbol{A}^\top \boldsymbol{y} - \boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{\theta}) \\
\boldsymbol{A}\boldsymbol{c} &= \frac{1}{2}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{A}\boldsymbol{X}\boldsymbol{\theta})
\end{aligned}$$

Note that $\boldsymbol{A}$ is a symmetric positive definite matrix, so $\boldsymbol{A}$ is invertible. Then

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

Since $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, we derive

$$\boldsymbol{c} = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

5. Set $\boldsymbol{A} := \boldsymbol{I}, \boldsymbol{c} := \boldsymbol{0}, \boldsymbol{B} := \lambda \boldsymbol{I}, \lambda \in \mathbb{R}$ for our calculation of $\boldsymbol{\theta}$ in **Exercise 4.2**, we derive

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

Hence, my answer agrees with the analytic solution for the standard least squares regression problem with L2 regularization $\boldsymbol{\theta} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.