

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*

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

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

(ii) *Absolutely homogeneous*

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

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

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

(iii) *Triangle inequality*

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

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

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

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

Exercise 2.2

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

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

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

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

$$= x^T B + (Bx)^T \quad \dots \lim_{h \rightarrow 0} h^T B = 0$$

$$= x^T B + x^T B^T$$

$$= x^T (B + B^T) \text{ , we proved that}$$

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

Exercise 2.1

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

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

$$\begin{aligned} f(x+h) &= (x+h)^T a b^T (x+h) \\ &= (x^T + h^T) a b^T (x+h) \quad \dots \text{(mm1-book p.25)} \\ &= (x^T a b^T + h^T a b^T) (x+h) \quad \dots \text{(distributivity)} \\ &= x^T a b^T x + x^T a b^T h + h^T a b^T x + h^T a b^T h \end{aligned}$$

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

$$= x^T a b^T + (a b^T x)^T \quad \dots \lim_{h \rightarrow 0} h^T a b^T h = 0$$

$$= x^T a b^T + x^T b a^T \quad \dots \text{(mm1-book p.25)}$$

Since $x^T a \in \mathbb{R}^{1 \times n} \cdot \mathbb{R}^{n \times 1} = \mathbb{R}$, $(x^T a)^T = x^T a$, 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 3

(i) *Symmetric*

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

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

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

(ii) *Positive definiteness*

$\forall \mathbf{x} \in V \setminus \{\mathbf{0}\}$:

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

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

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

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

Exercise 4

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Symmetry

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

Hence, it is symmetric.

Positive definiteness

$\forall \mathbf{y} \in V \setminus \{\mathbf{0}\}$:

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

Since $\|\mathbf{X} \mathbf{y}\|_A^2 \geq 0$, we derive

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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