Name: **Student ID:**

CS189: Introduction to Machine Learning

Homework 2 Solutions

Instructions:

- Homework 2 is completely a written assignment, no coding involved.
- Please write (legibly!) or typeset your answers in the space provided. If you choose to typeset your answers, please use this template file (hw2.tex), provided on bCourses announcement page. If there is not enough space for your answer, you can continue your answer on a separate page (for example : You might want to append pages in Questions 6,7,8).
- Submit a pdf of your answers to `https://gradescope.com` under Homework 2. A photograph or scanned copy is acceptable as long as it is clear with good contrast. You should be able to see CS 189/289 on gradescope when you login with your primary e-mail address used in bCourses. Please let us know if you have any problems accessing gradescope.
- While submitting to Gradescope, you will have to select the region containing your answer for each of the question. Thus, write the answer to a question (or given part of the question) at one place only.
- Start early and don't wait until last minute to submit the assignment as the submission procedure might take sometime too.

About the Assignment:

- This assignment tries to refresh the concepts of probability, linear algebra and matrix calculus.
- Questions 1 to 6 are dedicated to deriving fundamental results related to these concepts. You might want to refer your math class textbooks for help.
- Questions 7,8 discuss few applications of these concepts in machine learning.
- Hope you will enjoy doing the assignment !

Problem 1. A target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Shots within the inner circle are given 4 points, shots within the next ring are given 3 points, and shots within the third ring are given 2 points. Shots outside the target are given 0 points.

Let X be the distance of the hit from the center (in feet), and let the p.d.f of X be

$$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value of the score of a single shot?

Solution: The expected value is

$$\begin{aligned} &\int_0^{1/\sqrt{3}} 4\frac{2}{\pi(1+x^2)}dx + \int_{1/\sqrt{3}}^{1} 3\frac{2}{\pi(1+x^2)}dx + \int_1^{\sqrt{3}} 2\frac{2}{\pi(1+x^2)}dx \\ &= \frac{2}{\pi}\left[4\left(\tan^{-1}\frac{1}{\sqrt{3}} - \tan^{-1}0\right) + 3\left(\tan^{-1}1 - \tan^{-1}\frac{1}{\sqrt{3}}\right) + 2\left(\tan^{-1}\sqrt{3} - \tan^{-1}1\right)\right] \\ &= \boxed{\frac{13}{6}} \end{aligned}$$

Problem 2. Assume that the random variable X has the exponential distribution

$$f(x|\theta) = \theta e^{-\theta x} \qquad x > 0, \theta > 0$$

where θ is the parameter of the distribution. Use the method of maximum likelihood to estimate θ if 5 observations of X are $x_1 = 0.9$, $x_2 = 1.7$, $x_3 = 0.4$, $x_4 = 0.3$, and $x_5 = 2.4$, generated i.i.d. (i.e., independent and identically distributed).

Solution: We'll solve the general case for the MLE of an exponential distribution, then plug in the numbers we have.

$$P(X_1, X_2, \ldots, X_n|\theta) = P(X_1|\theta)P(X_2|\theta)\ldots P(X_n|\theta)$$

$$L(\theta|X_1, X_2, \ldots, X_n) = P(X_1|\theta)P(X_2|\theta)\ldots P(X_n|\theta)$$

$$L(\theta|X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\theta) = \theta^n \exp(-\theta \sum_{i=1}^{n} X_i)$$

Finding the log-likelihood:

$$l(\theta|X_1, X_2, \ldots, X_n) = n \ln(\theta) - \theta \sum_{i=1}^{n} X_i$$

Taking the derivative with respect to θ:

$$\frac{\delta l}{\delta \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} X_i = 0$$

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} X_i}$$

Plugging in our values for X_i, we get $\hat{\theta} = 0.88$.

Problem 3. The polynomial kernel is defined to be

$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x^T y} + \mathbf{c})^{\mathbf{d}}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathbf{n}}$, and $c \geq 0$. When we take $d = 2$, this kernel is called the quadratic kernel.

(a) Find the feature mapping $\Phi(\mathbf{z})$ that corresponds to the quadratic kernel.

(b) How do we find the optimal value of d for a given dataset?

Solution:

(a) (from Wikipedia)

$$K(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} x_i y_i + c \right)^2 = \sum_{i=1}^{n} \left(x_i^2\right) \left(y_i^2\right) + \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\sqrt{2} x_i x_j\right) \left(\sqrt{2} y_i y_j\right) + \sum_{i=1}^{n} \left(\sqrt{2c} x_i\right) \left(\sqrt{2c} y_i\right) + c^2$$

$$\Phi(\mathbf{z}) = \langle x_n^2, \ldots, x_1^2, \sqrt{2} x_n x_{n-1}, \ldots, \sqrt{2} x_n x_1, \sqrt{2} x_{n-1} x_{n-2}, \ldots, \sqrt{2} x_{n-1} x_1, \ldots, \sqrt{2} x_2 x_1, \sqrt{2c} x_n, \ldots, \sqrt{2c} x_1, c \rangle$$

(b) Cross Validation

Def: Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. We say that A is positive definite if $\forall x \in \mathbb{R}^n,\ x^\top A x > 0$. Similarly, we say that A is positive semidefinite if $\forall x \in \mathbb{R}^n,\ x^\top A x \geq 0$.

Problem 4. Let $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$, and let $A \in \mathbb{R}^{n \times n}$ be the square matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

(a) Give an explicit formula for $x^\top A x$. Write your answer as a sum involving the elements of A and x.

(b) Show that if A is positive definite, then the entries on the diagonal of A are positive (that is, $a_{ii} > 0$ for all $1 \leq i \leq n$).

Solution:

(a)

$$x^\top A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

(b) Let $i \in [1, n]$, and let e_i be the i^{th} standard basis vector (that is, the vector of all zeros except for a single 1 in the i^{th} position). Then, by the positive definitness of A, we have $e_i^\top A e_i = a_{ii} > 0$.

Problem 5. Let B be a positive semidefinite matrix. Show that $B+\gamma I$ is positive definite for any $\gamma > 0$.

Solution: Let $x \neq 0$. Then

$$\begin{aligned} x^\top (B+\gamma I)x &= x^\top Bx + x^\top \gamma I x \\ &= x^\top Bx + \gamma \|x\|^2 \\ &> 0 \end{aligned}$$

because $x^\top Bx \geq 0$ (since B is positive semidefinite) and $\|x\|^2 > 0$ (because $x \neq 0$). Hence $B+\gamma I$ is positive definite.

Problem 6 : Derivatives and Norms. Derive the expression for following questions. Do not write the answers directly.

(a) Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$. Derive $\frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial \mathbf{x}}$.

(b) Let $\mathbf{A}$ be a $n \times n$ matrix and $\mathbf{x}$ be a vector in $\mathbb{R}^n$. Derive $\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}}$.

(c) Let $\mathbf{A}$, $\mathbf{X}$ be $n \times n$ matrices. Derive $\frac{\partial \text{Trace}(\mathbf{X}\mathbf{A})}{\partial \mathbf{X}}$.

(d) Let $\mathbf{X}$ be a $m \times n$ matrix, $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$. Derive $\frac{\partial(\mathbf{a}^T\mathbf{X}\mathbf{b})}{\partial \mathbf{X}}$.

(e) Let $\mathbf{x} \in \mathbb{R}^n$. Prove that $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2$. Here $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ and $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$.

Solution:

(a) Let $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$ and $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix}$.

$$\mathbf{x}^T\mathbf{a} = \sum_{i=1}^{n} x_i a_i$$

Taking partial derivative wrt a component, we get

$$\frac{\partial\left(\mathbf{x}^T\mathbf{a}\right)}{\partial x_k} = a_k$$

Placing all partial derivatives into a single vector, we get

$$\frac{\partial\left(\mathbf{x}^T\mathbf{a}\right)}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial\left(\mathbf{x}^T\mathbf{a}\right)}{\partial x_1} \\ \frac{\partial\left(\mathbf{x}^T\mathbf{a}\right)}{\partial x_2} \\ \cdots \\ \frac{\partial\left(\mathbf{x}^T\mathbf{a}\right)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} = \mathbf{a}$$

(b) Let $\mathbf{A} = [a_{ij}]_{n\times n}$. We can write

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i,j=1}^{n} a_{ij}x_ix_j$$

Taking partial derivative wrt a component, we get

$$\begin{aligned}
\frac{\partial\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)}{\partial x_k} &= \frac{\partial}{\partial x_k}\left(\sum_{i,j=1}^{n} a_{ij}x_ix_j\right) \\
&= \frac{\partial}{\partial x_k}\left(x_1\sum_{j=1}^{n} a_{1j}x_j + x_2\sum_{j=1}^{n} a_{2j}x_j + \cdots + x_k\sum_{j=1}^{n} a_{kj}x_j + \cdots + x_n\sum_{j=1}^{n} a_{nj}x_j\right) \\
&\quad \text{(Use product rule of differentiation, i.e. } (fg)' = f'g + fg'\text{), on each term)} \\
&= x_1a_{1k} + x_2a_{2k} + \cdots + x_ka_{kk} + \sum_{j=1}^{n} a_{kj}x_j + \cdots + x_na_{nk} \\
&= (x_1a_{1k} + x_2a_{2k} + \cdots + x_ka_{kk} + \cdots + x_na_{nk}) + \left(\sum_{j=1}^{n} a_{kj}x_j\right) \\
&= \left(\sum_{i=1}^{n} a_{ik}x_i\right) + \left(\sum_{j=1}^{n} a_{kj}x_j\right) \\
&= \left(k^{th} \text{ column of } \mathbf{A}\right)^T\mathbf{x} + \left(k^{th} \text{ row of } \mathbf{A}\right)^T\mathbf{x}
\end{aligned}$$

Placing all partial derivatives into a single vector, we get

$$\frac{\partial\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)}{\partial\mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$$

(c) Let $\mathbf{A} = [a_{ij}]_{n\times n}$ and $\mathbf{X} = [x_{ij}]_{n\times n}$. We can write

$$\text{Trace}(\mathbf{XA}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}a_{ji}$$

Taking partial derivative wrt a component, we get

$$\begin{aligned}
\frac{\partial\text{Trace}(\mathbf{XA})}{\partial x_{ij}} &= \frac{\partial}{\partial x_{ij}}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}a_{ji}\right) \\
&= a_{ji}
\end{aligned}$$

Placing all partial derivatives into the matrix, we get

$$\frac{\partial\text{Trace}(\mathbf{XA})}{\partial\mathbf{X}} = [a_{ji}]_{n\times n} = \mathbf{A}^T$$

(d) Given that $\mathbf{X}$ is $m \times n$ matrix, $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$. Notice that $\mathbf{a}^T\mathbf{X}\mathbf{b}$ is 1×1 i.e. scalar. We know that Trace of a scalar (i.e. 1×1 matrix) is same as the scalar itself. We can write

$$\mathbf{a}^T\mathbf{X}\mathbf{b} = \text{Trace}\left(\mathbf{a}^T\mathbf{X}\mathbf{b}\right)$$

Note that Trace($\mathbf{CD}$) = Trace($\mathbf{DC}$). We get,

$$\begin{aligned}
\mathbf{a}^T\mathbf{X}\mathbf{b} &= \text{Trace}\left(\mathbf{a}^T\mathbf{X}\mathbf{b}\right) \\
&= \text{Trace}\left(\mathbf{a}^T(\mathbf{X}\mathbf{b})\right) \\
&= \text{Trace}\left((\mathbf{X}\mathbf{b})\mathbf{a}^T\right) \\
&= \text{Trace}\left(\mathbf{X}(\mathbf{b}\mathbf{a}^T)\right)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\frac{\partial\left(\mathbf{a}^T\mathbf{X}\mathbf{b}\right)}{\partial\mathbf{X}} &= \frac{\partial\left(\text{Trace}\left(\mathbf{X}(\mathbf{b}\mathbf{a}^T)\right)\right)}{\partial\mathbf{X}} \\
&\quad \text{(Using result from part (c))} \\
&= (\mathbf{b}\mathbf{a}^T)^T \\
&= \mathbf{a}\mathbf{b}^T
\end{aligned}$$

(e) First lets prove the left-hand side inequality as follows:

$$\begin{aligned}
\|\mathbf{x}\|_2 &= \sqrt{\sum_{i=1}^{n} x_i^2} \\
&= \sqrt{x_1^2 + x_2^2 \cdots + x_n^2} \\
&\quad \text{(adding positive terms)} \\
&\leq \sqrt{x_1^2 + x_2^2 \cdots + x_n^2 + 2\left(\sum_{1 \leq i < j \leq n} |x_i||x_j|\right)} \\
&= \sqrt{(|x_1| + |x_2| + \cdots + |x_n|)^2} \\
&= |x_1| + |x_2| + \cdots + |x_n| \\
&= \|\mathbf{x}\|_1 \\
\implies \|\mathbf{x}\|_2 &\leq \|\mathbf{x}\|_1
\end{aligned}$$

Lets now prove the right hand side inequality as follows:

$$\begin{aligned}
& \|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n| \\
\implies & \|\mathbf{x}\|_1 = \underbrace{(|x_1|, |x_2|, \ldots, |x_n|)^T}_{\text{call this vector } \mathbf{x}'} \bullet (1, 1, \ldots 1) \\
\implies & \|\mathbf{x}\|_1 = \mathbf{x}'^T \bullet \mathbf{1} \\
& \qquad \text{(Using Cauchy}-\text{Schwarz inequality on the right)} \\
\implies & \|\mathbf{x}\|_1 \leq \|\mathbf{x}'\|_2 \|\mathbf{1}\|_2 \\
& \qquad \text{Note: } \|\mathbf{x}'\|_2 = \|\mathbf{x}\|_2 \text{ and } \|\mathbf{1}\|_2 = \sqrt{n} \\
\implies & \|\mathbf{x}\|_1 \leq \sqrt{n} \|\mathbf{x}\|_2
\end{aligned}$$

Thus, we have shown

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n} \|\mathbf{x}\|_2$$

Problem 7 : Application of Matrix Derivatives.
Let $\mathbf{X}$ be a $n \times d$ data matrix, $\mathbf{Y}$ be the corresponding $n \times 1$ target/label matrix and $\mathbf{\Lambda}$ be the diagonal $n \times n$ matrix containing weight of each example. Expanding them, we have $\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \dots \\ (\mathbf{x}^{(n)})^T \end{bmatrix}$, $\mathbf{Y} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \dots \\ \mathbf{y}^{(n)} \end{bmatrix}$ and $\mathbf{\Lambda} = \text{diag}(\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(n)})$
where $\mathbf{x}^{(i)} \in \mathbb{R}^d$, $\mathbf{y}^{(i)} \in \mathbb{R}$, and $\lambda^{(i)} > 0 \quad \forall \ i \in \{1 \dots n\}$. $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{\Lambda}$ are fixed and known.
In the remaining parts of this question, we will try to fit a weighted linear regression model for this data. We want to find the value of weight vector w which best satisfies the following equation $y^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} + \epsilon^{(i)}$ where ϵ is noise. This is achieved by minimizing the weighted noise for all the examples. Thus, our risk function is defined as follows:

$$\begin{aligned} R[\mathbf{w}] &= \sum_{i=1}^{n} \lambda^{(i)} (\epsilon^{(i)})^2 \\ &= \sum_{i=1}^{n} \lambda^{(i)} (\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)})^2 \end{aligned}$$

(a) Write this risk function $R[\mathbf{w}]$ in matrix notation, i.e., in terms of $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Lambda}$ and $\mathbf{w}$.

(b) Find the value of $\mathbf{w}$, in matrix notation, that minimizes the risk function obtained in Part (a). You can assume that $\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}$ is full rank matrix. Hint: You can use the expression derived in Q-6(b).

(c) What will be the answer for questions in Parts (a) and (b) if you add L_2 regularization (i.e., shrinkage) on $\mathbf{w}$? The L2 regularized risk function, for $\gamma > 0$, is

$$R[\mathbf{w}] = \sum_{i=1}^{n} \lambda^{(i)} (\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)})^2 + \gamma \|\mathbf{w}\|_2^2$$

Hint: You can make use of the result in Q-5.

(d) What role does the regularization (i.e., shrinkage) play in fitting the regression model and how ? You can observe the difference in expressions for $\mathbf{w}$ obtained in Parts (c) and (d), and argue.

Solution:

(a) We can re-write the risk function $R[\mathbf{w}]$ in matrix notation as follows

$$\begin{aligned} R[\mathbf{w}] &= \sum_{i=1}^{n} \lambda^{(i)} (\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)})^2 \\ &= (\mathbf{X}\mathbf{w} - \mathbf{Y})^T \mathbf{\Lambda} (\mathbf{X}\mathbf{w} - \mathbf{Y}) \end{aligned}$$

(b) We minimize the risk function, i.e.,

$$\begin{aligned}
&\min_{\mathbf{w}} \; R[\mathbf{w}] \\
\implies &\min_{\mathbf{w}} \; (\mathbf{X}\mathbf{w}-\mathbf{Y})^T\mathbf{\Lambda}(\mathbf{X}\mathbf{w}-\mathbf{Y}) \\
\implies &\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-(\mathbf{X}\mathbf{w})^T\mathbf{\Lambda}\mathbf{Y}-\mathbf{Y}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} \\
&\qquad \text{(Note that } \mathbf{\Lambda}^T=\mathbf{\Lambda}\text{)} \\
\implies &\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-(\mathbf{X}\mathbf{w})^T\mathbf{\Lambda}\mathbf{Y}-(\mathbf{Y}^T\mathbf{\Lambda}\mathbf{X})\mathbf{w} \\
\implies &\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-\mathbf{w}^T(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})-(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w} \\
\implies &\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w} \qquad (1)
\end{aligned}$$

For taking derivative w.r.t. $\mathbf{w}$, we use results from Q-6(a),6(b) as follows

$$\begin{aligned}
\implies &\frac{\partial}{\partial \mathbf{w}}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w}+\mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y}-2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w}\right) \\
&= (\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}+(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})^T)\mathbf{w}+0-2\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \\
&= 2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})\mathbf{w}-2\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \\
&\text{(Make derivative equal to zero and solve)} \\
&= 0 \\
\implies &(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})\mathbf{w}=\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \\
&\text{(Given that } (\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}) \text{ is full rank and thus invertible)} \\
\implies &\mathbf{w}=(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y}
\end{aligned}$$

Notice that we have found that the stationary point of the quadratic (equation (1)). To prove that it is indeed the minimum, we need to show the matrix in the square term is positive semi-definite. That is, we will show that

$$\implies \mathbf{z}^T(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})\mathbf{z}\geq 0 \;\; \forall \mathbf{z}\in\mathbb{R}^n$$

Let $\mathbf{y}=\mathbf{X}\mathbf{z}$, then we get

$$\begin{aligned}
\implies &\mathbf{z}^T(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})\mathbf{z} \\
&= \mathbf{y}^T\mathbf{\Lambda}\mathbf{y} \\
&= \sum_{i=1}^{n}\lambda^{(i)}y_i^2 \\
&\text{(Given that } \lambda^{(i)}>0\text{)} \\
&\geq 0
\end{aligned}$$

Thus, our solution is

$$\mathbf{w}=(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y}$$

(c) For the risk function with a regularizer term, we have

$$\begin{aligned}
&\implies R[\mathbf{w}] \\
&\implies \sum_{i=1}^{n} \lambda^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)})^2 + \gamma\|\mathbf{w}\|_2^2 \\
&\implies \sum_{i=1}^{n} \lambda^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)})^2 + \gamma\mathbf{w}^T\mathbf{w} \\
&\implies (\mathbf{X}\mathbf{w} - \mathbf{Y})^T\mathbf{\Lambda}(\mathbf{X}\mathbf{w} - \mathbf{Y}) + \gamma\mathbf{w}^T\mathbf{w} \qquad (2)
\end{aligned}$$

Similar to part (b), we can minimize it as follows

$$\begin{aligned}
&\min_{\mathbf{w}} \; R[\mathbf{w}] \\
\implies &\min_{\mathbf{w}} \; (\mathbf{X}\mathbf{w} - \mathbf{Y})^T\mathbf{\Lambda}(\mathbf{X}\mathbf{w} - \mathbf{Y}) + \gamma\mathbf{w}^T\mathbf{w} \\
&\qquad \text{(Using similar calculation as in part(b), we obtain)} \\
\implies &\min_{\mathbf{w}} \; \mathbf{w}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}\mathbf{w} + \mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y} - 2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w} + \gamma\mathbf{w}^T\mathbf{w} \\
\implies &\min_{\mathbf{w}} \; \mathbf{w}^T(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X} + \gamma\mathbf{I})\mathbf{w} + \mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y} - 2(\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y})^T\mathbf{w}
\end{aligned}$$

This expression is very similar to equation (1) from part(b), thus we similarly obtain derivative and get

$$(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X} + \gamma\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{\Lambda}\mathbf{Y}$$

We are given that $\gamma > 0$ and in part(b) we showed that $\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}$ is positive semi-definite. From Q-5, we know that $(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X} + \gamma\mathbf{I})$ is positive definite and hence invertible. This also means our stationary point is indeed the minima.
Thus, our solution is

$$\mathbf{w} = (\mathbf{X}^T\mathbf{\Lambda}\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^T\mathbf{\Lambda}\mathbf{Y} \qquad (3)$$

(d) In regularization, we have essentially added a penalty on $\mathbf{w}$ as shown in equation (2). This prevents the magnitude of $\mathbf{w}$ from becoming large. This is reflected in our solution in equation (3), where γ appears in the inverse. Thus, larger the value of γ is, smaller the magnitude of $\mathbf{w}$ will be.
This prevents over-fitting on the training data, where the hyperparameter γ is obtained using cross-validation.

Problem 8: Classification. Suppose we have a classification problem with classes labeled $1, \ldots, c$ and an additional doubt category labeled as $c+1$. Let the loss function be the following:

$$\ell(f(x) = i, y = j) = \begin{cases} 0 & \text{if } i = j \quad i, j \in \{1, \ldots, c\} \\ \lambda_r & \text{if } i = c+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

where λ_r is the loss incurred for choosing doubt and λ_s is the loss incurred for making a misclassification. Note that $\lambda_r \geq 0$ and $\lambda_s \geq 0$.
Hint : The risk of classifying a new datapoint as class $i \in \{1, 2, \ldots, c+1\}$ is

$$R(\alpha_i|x) = \sum_{j=1}^{j=c} \ell(f(x) = i, y = j) P(\omega_j|x)$$

(a) Show that the minimum risk is obtained if we follow this policy: (1) choose class i if $P(\omega_i|x) \geq P(\omega_j|x)$ for all j and $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s$, and (2) choose doubt otherwise.

(b) What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$?

Solution:

(a) Define $\lambda_{ij} = \ell(f(x) = i, y = j)$. The risk of classifying a new datapoint as class i is

$$R(\alpha_i|x) = \sum_j \lambda_{ij} P(\omega_j|x) = \lambda_s(1 - P(\omega_i|x)),$$

and the risk of classifying the new datapoint as doubt is

$$R(\alpha_{c+1}|x) = \lambda_r \sum_j P(\omega_j|x) = \lambda_r.$$

For choosing doubt to be better than choosing any of the classes, the ratio of the risks must satisfy

$$1 > \frac{R(\alpha_{c+1}|x)}{R(\alpha_i|x)} = \frac{\lambda_r}{\lambda_s(1 - P(\omega_i|x))} \implies P(\omega_i|x) < 1 - \frac{\lambda_r}{\lambda_s}$$

for all i. This means that any particular i for which $P(\omega_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$ should not be assigned doubt. In this case, the class to choose must be

$$\arg \min_{1 \leq i \leq c} R(\alpha_i|x) = \arg \min_{1 \leq i \leq c} \lambda_s(1 - P(\omega_i|x)) = \arg \max_{1 \leq i \leq c} P(\omega_i|x),$$

as required.

(b) If $\lambda_r = 0$, then doubt will always be assigned, since for all i, $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s = 1$ is not satisfied unless $P(\omega_i|x) = 1$.

If $\lambda_r > \lambda_s$, then doubt will never be assigned, since for all i, $P(\omega_i|x) \geq 0 > 1 - \lambda_r/\lambda_s$ always holds.