

CS221 Fall 2021 Homework 1

SUNet ID: jxiangyu

Name: Xiangyu Liu

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

Problem 1

(a) Expanding the $f(\theta)$ produces the following equation:

$$\begin{aligned} f(x) &= \sum_{i=1}^n w_i (\theta - x_i)^2 \\ &= \theta^2 \sum_{i=1}^n w_i - 2\theta \sum_{i=1}^n w_i x_i + \sum_{i=1}^n w_i x_i^2 \end{aligned} \tag{1}$$

Because w_i is a positive real number, $\sum_{i=1}^n w_i > 0$. Therefore, the quadratic function with respect to θ is a convex parabola. Its vertex is the global minimum, a value of θ that satisfies $\frac{df(x)}{d\theta} = 0$.

$$\begin{aligned} \frac{df(x)}{d\theta} &= \frac{d \sum_{i=1}^n w_i (\theta - x_i)^2}{d\theta} \\ &= \sum_{i=1}^n 2w_i (\theta - x_i) \\ &= 2\theta \sum_{i=1}^n w_i - 2 \sum_{i=1}^n w_i * x_i \end{aligned} \tag{2}$$

When $\frac{df(x)}{d\theta} = 0$, we have:

$$2\theta \sum_{i=1}^n w_i - 2 \sum_{i=1}^n w_i * x_i = 0 \tag{3}$$

Because w_i is a positive real number, $\sum_{i=1}^n w_i > 0$. Therefore, dividing $\sum_{i=1}^n w_i$ on both sides will produce the value:

$$\theta = \frac{\sum_{i=1}^n w_i * x_i}{\sum_{i=1}^n w_i} \tag{4}$$

If some w_i are negative number, then $\sum_{i=1}^n w_i$ can be 0 or negative. Therefore, this quadratic function might not be a convex parabola. and the global minimum value might not exist.

(b) Answer: $f(x) \leq g(x)$.

Proof:

$$\begin{aligned}
 f(x) &= \max_{s \in [-1,1]} \sum_{i=1}^d s x_i \\
 &= \max_{s \in [-1,1]} s(x_1 + x_2 + \dots + x_d) \\
 &= \max_{s \in [-1,1]} (s x_1 + s x_2 + \dots + s x_d)
 \end{aligned} \tag{5}$$

$$\begin{aligned}
 g(x) &= \sum_{i=1}^d \max_{s_i \in [-1,1]} s_i x_i \\
 &= \max_{s_1 \in [-1,1]} s_1 x_1 + \max_{s_2 \in [-1,1]} s_2 x_2 + \dots + \max_{s_d \in [-1,1]} s_d x_d
 \end{aligned} \tag{6}$$

Because the sum of the max is always greater than or equal to the max of the sum, $f(x) \leq g(x)$.

(c) Let V be the expected number of points. Then

$$\begin{aligned}
 V &= \sum_{i=1}^6 \frac{1}{6} * (PointsOfRolling(i) + PointsOfNextRoll(i)) \\
 &= \frac{1}{6} * 0 + \frac{1}{6} * 0 + \frac{1}{6} * (-a + V) + \frac{1}{6} * (b + V) + \frac{1}{6} * V + \frac{1}{6} * V \\
 &= \frac{b-a}{6} + \frac{2V}{3}
 \end{aligned} \tag{7}$$

Then

$$V = \frac{b-a}{2} \tag{8}$$

(d)

$$\begin{aligned}
 \frac{d \ln(L(p))}{dp} &= \frac{1}{L(p)} \frac{dL(p)}{dp} \\
 &= \frac{1}{p^4(1-p)^2} \frac{d(p^6 - 2p^5 + p^4)}{dp} \\
 &= \frac{1}{p^4(1-p)^2} (6p^5 - 10p^4 + 4p^3) \\
 &= \frac{6p^2 - 10p + 4}{p(1-p)^2} \\
 &= \frac{6p-4}{p(p-1)}
 \end{aligned} \tag{9}$$

When $\frac{6p-4}{p(p-1)} = 0$, $\ln(L(p))$ is maximized, thus $L(p)$ is maximized. Therefore $p = \frac{2}{3}$

(e)

$$\begin{aligned}
P(A|B) &= P(B|A) \\
\frac{P(A \cap B)}{P(B)} &= \frac{P(B \cap A)}{P(A)} \\
P(A) &= \frac{P(B \cap A)}{P(A \cap B)} P(B) \\
P(A) &= \frac{P(B \cap A)}{P(A \cap B)} P(B) \\
P(A) &= \frac{P(B \cap A)}{P(B \cap A)} P(B) \\
P(A) &= P(B)
\end{aligned} \tag{10}$$

Now use the equation above to simplify $P(A \cup B)$:

$$\begin{aligned}
P(A \cup B) &= \frac{1}{2} \\
P(A) + P(B) - P(A \cap B) &= \frac{1}{2} \\
2P(A) &= \frac{1}{2} + P(A \cap B) \\
P(A) &= \frac{1}{4} + \frac{P(A \cap B)}{2} > \frac{1}{4}
\end{aligned} \tag{11}$$

(f) The gradient is a d-dimensional vector of the partial derivatives with respect to each w_i :

$$\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \dots, \frac{\partial f(\mathbf{w})}{\partial w_d} \right)^T \tag{12}$$

Consider a given w_x , where $1 \leq x \leq d$, the partial derivative with respect to w_x is:

$$\begin{aligned}
\frac{\partial f(\mathbf{w})}{\partial w_x} &= \frac{\partial(\sum_{i=1}^n \sum_{j=1}^n (\mathbf{a}_i^T \mathbf{w} - \mathbf{b}_j^T \mathbf{w})^2) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2}{\partial w_x} \\
&= \frac{\partial \sum_{i=1}^n \sum_{j=1}^n ((\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w})^2}{\partial w_x} + \frac{\lambda}{2} \frac{\partial \|\mathbf{w}\|_2^2}{\partial w_x} \\
&= \frac{\partial \sum_{i=1}^n \sum_{j=1}^n ((\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w})^2}{\partial w_x} + \frac{\lambda}{2} \frac{\partial \sum_{k=1}^d w_k^2}{\partial w_x} \\
&= \frac{\partial \sum_{i=1}^n \sum_{j=1}^n ((\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w})^2}{\partial w_x} + \frac{\lambda}{2} \frac{\partial \sum_{k=1}^d w_k^2}{\partial w_x} \\
&= \sum_{i=1}^n \sum_{j=1}^n \frac{\partial((\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w})^2}{\partial w_x} + \frac{\lambda}{2} \sum_{k=1}^d \frac{\partial w_k^2}{\partial w_x}
\end{aligned} \tag{13}$$

Let's examine the first term in the equation:

$$\sum_{i=1}^n \sum_{j=1}^n \frac{\partial((\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w})^2}{\partial w_x} = \sum_{i=1}^n \sum_{j=1}^n \frac{2(\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w} \partial(\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w}}{\partial w_x} \tag{14}$$

For all w_i , where $i \neq x$, $\frac{\partial w_i}{\partial w_x} = 0$.

Hence:

$$\frac{\partial(\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w}}{\partial w_x} = (\mathbf{a}_i[x] - \mathbf{b}_j[x]) \tag{15}$$

Where $\mathbf{a}_i[x]$ denotes the x -th element in vector \mathbf{a}_i , and $\mathbf{b}_j[x]$ denotes the x -th element in vector \mathbf{b}_j .

This results in:

$$\sum_{i=1}^n \sum_{j=1}^n \frac{2(\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w} \partial(\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w}}{\partial w_x} = \sum_{i=1}^n \sum_{j=1}^n 2(\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w} (\mathbf{a}_i[x] - \mathbf{b}_j[x]) \tag{16}$$

Similarly,

$$\begin{aligned}\frac{\lambda}{2} \sum_{k=1}^d \frac{\partial w_k^2}{\partial w_x} &= \frac{\lambda}{2} \frac{\partial w_x^2}{\partial w_x} \\ &= \frac{\lambda}{2} 2w_x \\ &= \lambda w_x\end{aligned}\tag{17}$$

Therefore,

$$\frac{\partial f(\mathbf{w})}{\partial w_x} = \lambda w_x + 2(\mathbf{a}_i[x] - \mathbf{b}_j[x]) \sum_{i=1}^n \sum_{j=1}^n (\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w}\tag{18}$$

Finally,

$$\begin{aligned}\nabla f(\mathbf{w}) = & \begin{bmatrix} \lambda w_1 + 2(\mathbf{a}_i[1] - \mathbf{b}_j[1]) \sum_{i=1}^n \sum_{j=1}^n (\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w}, \\ \lambda w_2 + 2(\mathbf{a}_i[2] - \mathbf{b}_j[2]) \sum_{i=1}^n \sum_{j=1}^n (\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w} \\ \dots \\ \lambda w_d + 2(\mathbf{a}_i[d] - \mathbf{b}_j[d]) \sum_{i=1}^n \sum_{j=1}^n (\mathbf{a}_i^T - \mathbf{b}_j^T) \mathbf{w} \end{bmatrix},\end{aligned}\tag{19}$$

Problem 2

- (a) To determine a rectangle only requires to determine the 2 diagonal vertices of a rectangle. Hence, on a $n \times n$ grid, there are $O(n^2)$ ways to place the top left vertex, and $O(n^2)$ to place the bottom right vertex. Therefore, there are $O(n^4)$ unique rectangles. To place 4 rectangles, there are $O((n^4)^4) = O(n^{16})$ ways.
- (b) Let $Cost(i, j)$ be the minimal cost of the path that goes through position (i, j) from position $(1, 1)$. Because we can only step down or to the right, $Cost(i, j)$ can be represented by:

$$Cost(i, j) = \text{Min}(Cost(i - 1, j), Cost(i, j - 1)) + c(i, j) \quad (20)$$

By using Dynamic programming, we can start from position $(1, 1)$ and calculate the $Cost(i, j)$ for its neighbor on the right and below. Then repeat this operation to calculate the neighbors for each position, of which the $Cost(i, j)$ has been determined, until the entire grid has been scanned.

The runtime is $O(n^2)$.

Problem 3

- (a) Because it predicts likelihood of loan default, this algorithm has likely encoded, contains, or potentially exacerbates bias against people of a certain race or ethnicity. Because the economical status in the U.S. is unequal, Black people tend to have lower income than their white counterparts. Thus the algorithm will incorrectly predict that they will default a loan.
- (b) Even though it is to analyze historical documents and literature, deanonymizing texts, including code, will raise safety or security concerns, because it can open security vulnerabilities. It allows any hostile forces to identify the anonymous person and perform hostile acts to that person.
- (c) Such development of the facial recognition technology uses information that could be deduced about individuals that they have not consented to share, because the celebrities did not give permission for their images to be used in the dataset.
- (d) Such model can potentially severely damage the environment. Because it enables users to identify the plants in the wild, users will then know the worth of such plant. It can potentially result in significant deforestation.