CS221 Problem Workout Solutions

Week 1

1) Problem 1: Gradient and Gradient Descent

(i) Let $\phi(x) : \mathbb{R} \mapsto \mathbb{R}^d$, $\mathbf{w} \in \mathbb{R}^d$. Consider the following objective function (a.k.a. loss function).

$$\text{Loss}(x, y, \mathbf{w}) = \begin{cases} 1 - 2(\mathbf{w} \cdot \phi(x))y & \text{if } (\mathbf{w} \cdot \phi(x))y \leq 0 \\ (1 - (\mathbf{w} \cdot \phi(x))y)^2 & \text{if } 0 < (\mathbf{w} \cdot \phi(x))y \leq 1 \\ 0 & \text{if } (\mathbf{w} \cdot \phi(x))y > 1, \end{cases}$$

where $y \in \mathbb{R}$. Compute the gradient $\nabla_{\mathbf{w}} \text{Loss}(x, y, \mathbf{w})$.

Solution We apply the rules to compute the gradient for each case separately, leading to the following piece-wise function for the gradient.

$$\nabla_{\mathbf{w}} \text{Loss}(x, y, \mathbf{w}) = \begin{cases} -2\phi(x)y & \text{if } (\mathbf{w} \cdot \phi(x))y \leq 0 \\ -2(1 - (\mathbf{w} \cdot \phi(x))y)\phi(x)y & \text{if } 0 < (\mathbf{w} \cdot \phi(x))y \leq 1 \\ 0 & \text{if } (\mathbf{w} \cdot \phi(x))y > 1 \end{cases} \tag{1}$$

(ii) Write out the Gradient Descent update rule for some function $\text{TrainLoss}(\mathbf{w}) : \mathbb{R}^d \mapsto \mathbb{R}$.

Solution $\mathbf{w} := \mathbf{w} - \eta \nabla_{\mathbf{w}} \text{TrainLoss}(\mathbf{w})$, where η is the step size.

(iii) Let $d = 2$ and $\phi(x) = [1, x]$. Consider the following loss function.

$$\text{TrainLoss}(\mathbf{w}) = \frac{1}{2}\Big(\text{Loss}(x_1, y_1, \mathbf{w}) + \text{Loss}(x_2, y_2, \mathbf{w})\Big). \tag{2}$$

Compute $\nabla_w \text{TrainLoss}(\mathbf{w})$ for the following values of $x_1, y_1, x_2, y_2, \mathbf{w}$.

$$\mathbf{w} = \left[0, \frac{1}{2}\right],$$
$$x_1 = -2, \quad y_1 = 1,$$
$$x_2 = -1, \quad y_2 = -1.$$

Solution

$$\begin{aligned}\nabla_w \text{TrainLoss}(\mathbf{w}) &= \frac{1}{2}\nabla_{\mathbf{w}}\Big(\text{Loss}(x_1, y_1, \mathbf{w}) + \text{Loss}(x_2, y_2, \mathbf{w})\Big) \\ &= \frac{1}{2}\nabla_{\mathbf{w}}\text{Loss}(x_1, y_1, \mathbf{w}) + \frac{1}{2}\nabla_{\mathbf{w}}\text{Loss}(x_2, y_2, \mathbf{w})\end{aligned}$$

For each of the terms above, we plug in the expression for the gradient computed in part (i) above.

Term one. Note that $\phi(x_1) = [1, -2]$. Since $(\mathbf{w} \cdot \phi(x_1))y_1 = -1$, we consider the first piece (Case 1) in the gradient expression (Equation 1). We have

$$\begin{aligned}\nabla_{\mathbf{w}}\text{Loss}(x_1, y_1, \mathbf{w}) &= -2\phi(x_1)y_1 \\ &= [-2, 4].\end{aligned} \tag{3}$$

Term two. Note that $\phi(x_2) = [1, -1]$. Similarly, $(\mathbf{w} \cdot \phi(x_2))y_2 = \frac{1}{2}$ taking us to Case 2 so

$$\begin{aligned}\nabla_{\mathbf{w}}\text{Loss}(x_2, y_2, \mathbf{w}) &= -2(1 - (\mathbf{w} \cdot \phi(x_2))y_2)\phi(x_2)y_2 \\ &= [1, -1].\end{aligned} \tag{4}$$

Combining the terms,

$$\begin{aligned}\nabla_{\mathbf{w}}\text{TrainLoss}(\mathbf{w}) &= \frac{1}{2}\Big([-2, 4] + [1, -1]\Big) \\ &= \left[-\frac{1}{2}, \frac{3}{2}\right].\end{aligned} \tag{5}$$

(iv) Perform two iterations of Gradient Descent to minimize the objective function $\text{TrainLoss}(\mathbf{w}) = \frac{1}{2}\Big(\text{Loss}(x_1, y_1, w) + \text{Loss}(x_2, y_2, w)\Big)$ with values for x_1, y_1, x_2, y_2 as above. Use initialization $\mathbf{w}^0 = \left[0, \frac{1}{2}\right]$ and step size $\eta = \frac{1}{2}$.

Solution Note that we have already computed $\nabla_{\mathbf{w}}\text{TrainLoss}(\mathbf{w})$ at the initialization point $\mathbf{w}^0$ in the question above.

$$\begin{aligned}\mathbf{w}^1 &= \mathbf{w}^0 - \eta\nabla_{\mathbf{w}}\text{TrainLoss}(\mathbf{w}) \text{ at } \mathbf{w}^0 \\ &= \left[0, \frac{1}{2}\right] - \left(\frac{1}{2}\right) \underbrace{\left(\frac{1}{2}\right)[-1, 3]}_{\text{From part (iii) above}} \\ &= \left[\frac{1}{4}, -\frac{1}{4}\right].\end{aligned}$$

Now we need to compute $\nabla_{\mathbf{w}}\text{Loss}(x_1, y_1, \mathbf{w})$ and $\nabla_{\mathbf{w}}\text{Loss}(x_2, y_2, \mathbf{w})$ at the new iterate $\mathbf{w}^1$.
We repeat the process we did for (iii) by applying the piece-wise defined gradient (Equation 1) to the two points, this time setting $\mathbf{w} = \mathbf{w}^1$.

Term one. Since $(\mathbf{w}^1 \cdot \phi(x_1))y_1 = \frac{3}{4}$, we have $\nabla_{\mathbf{w}}\text{Loss}(x_1, y_1, \mathbf{w}) = -2(1 - (\mathbf{w}^1 \cdot \phi(x_1))y_1)\phi(x_1)y_1 = [-\frac{1}{2}, 1]$. Note that we are now in Case 2 with respect to the piecewise definition of the gradient (Equation 1). When computing $\nabla_{\mathbf{w}}\text{Loss}(x_1, y_1, \mathbf{w})$ at $\mathbf{w}^0$, we were in Case 1.

Term two. $(\mathbf{w}^1 \cdot \phi(x_2))y_2 = -\frac{1}{2}$ taking us to Case 1, so $\nabla_{\mathbf{w}}\text{Loss}(x_2, y_2, \mathbf{w}) = -2\phi(x_2)y_2 = [2, -2]$.

Hence,

$$\begin{aligned}
\mathbf{w}^2 &= \mathbf{w}^1 - \eta \nabla_{\mathbf{w}}\text{TrainLoss}(\mathbf{w}) \text{ at } \mathbf{w}^1 \\
&= \left[\frac{1}{4}, -\frac{1}{4}\right] - \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\left[-\frac{1}{2}, 1\right] + [2, -2]\right) \\
&= \left[-\frac{1}{8}, 0\right].
\end{aligned}$$

2) Problem 2: Gradient computation

(i) Let $\phi(x) : \mathbb{R} \mapsto \mathbb{R}^d$, $\mathbf{w} \in \mathbb{R}^d$, and $f(x, \mathbf{w}) = \mathbf{w} \cdot \phi(x)$. Consider the following loss function.

$$\text{Loss}(x, y, \mathbf{w}) = \frac{1}{2} \max\{2 - (\mathbf{w} \cdot \phi(x))y, 0\}^2. \tag{6}$$

Compute its gradient $\nabla_{\mathbf{w}} \text{Loss}(x, y, \mathbf{w})$.

Solution Note that $\text{Loss}(x, y, \mathbf{w})$ can be written as the following piecewise defined function using the definition of max.

$$\text{Loss}(x, y, \mathbf{w}) = \begin{cases} \frac{1}{2}(2 - (\mathbf{w} \cdot \phi(x)y))^2 & \text{if } 2 - (\mathbf{w} \cdot \phi(x))y \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Using the chain rule, we get that the gradient is:

$$\nabla_{\mathbf{w}} \text{Loss}(x, y, \mathbf{w}) = \begin{cases} -(2 - \mathbf{w} \cdot \phi(x)y)\phi(x)y & \text{if } 2 - \mathbf{w} \cdot \phi(x)y \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

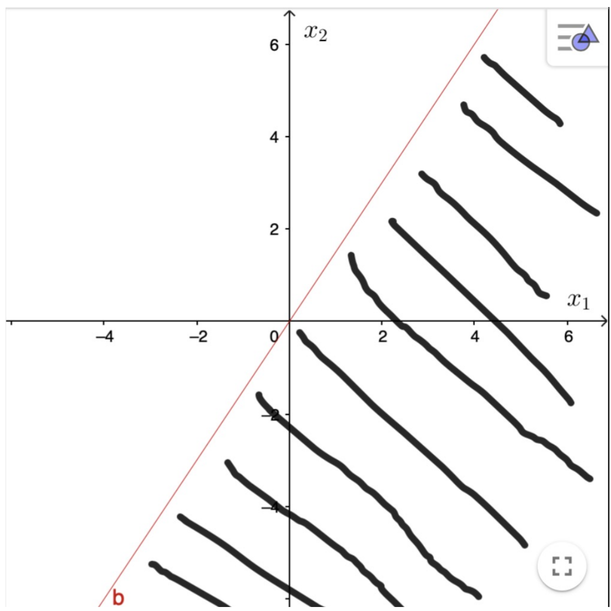
3) **Problem 3: Vector visualization**

Recall that we can visualize a vector $\mathbf{w} \in \mathbb{R}^d$ as a point in d-dimensional space. Let us now visualize some vectors in 2 dimensions on pen and paper.

(i) Consider $\mathbf{x} \in \mathbb{R}^2$. Draw the line (i.e. the "decision boundary") that separates between vectors having a positive dot product with weights $\mathbf{w} = [3, -2]$ and those having a negative dot product. Shade the part of the 2D plane that contains vectors satisfying $\mathbf{w} \cdot \mathbf{x} > 0$.

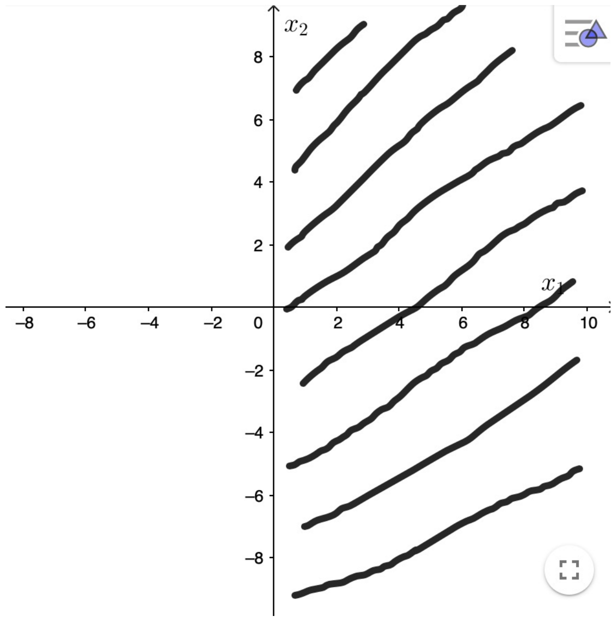
Hint: It might help to write out the expression for the dot product and seeing the relation between x_1 and x_2 that leads to a positive dot product. You could also use the geometric interpretation of the dot product.

Solution $\mathbf{w} \cdot \mathbf{x} = 3x_1 - 2x_2 > 0$

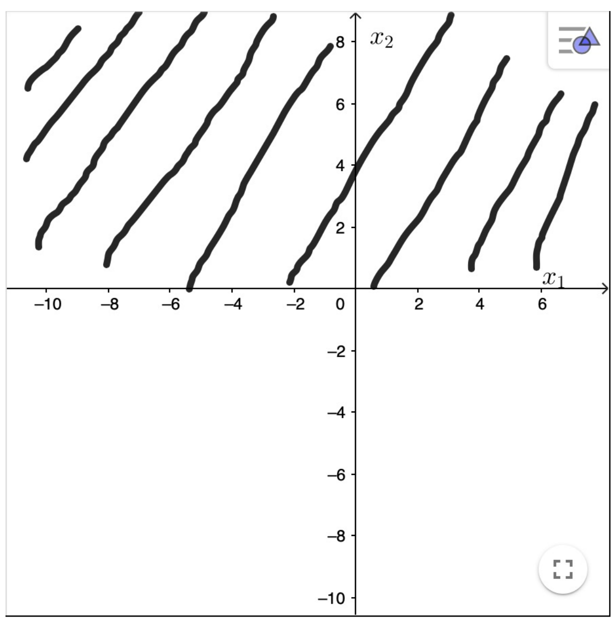


(ii) Repeat the above for $\mathbf{w} = [2, 0]$ and $\mathbf{w} = [0, 2]$.

Solution When $\mathbf{w} = [2, 0]$, $\mathbf{w} \cdot \mathbf{x} = 2x_1 > 0$

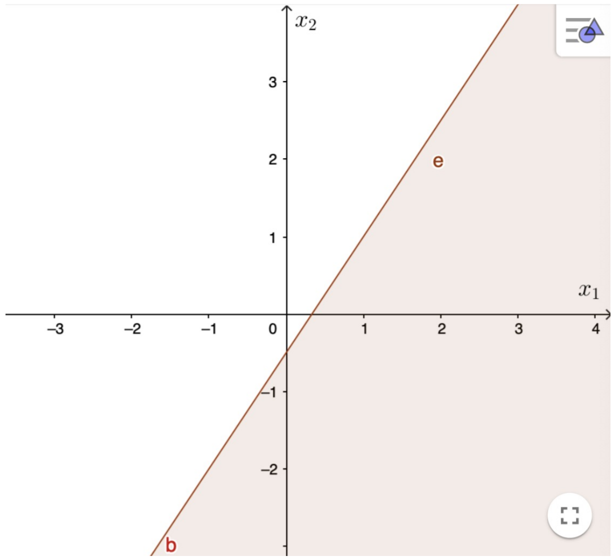


When $\mathbf{w} = [0, 2]$, $\mathbf{w} \cdot \mathbf{x} = 2x_2 > 0$



(iii) A small twist: visualize the set of vectors where $\mathbf{w} \cdot \mathbf{x} \geq 1$ for $\mathbf{w} = [3, -2]$.

Solution $\mathbf{w} \cdot \mathbf{x} = 3x_1 - 2x_2 \geq 1$, so $3x_1 - 2x_2 - 1 \geq 0$



Note that we get a line that is parallel to the one in (i) but shifted by a certain amount.

(iv) Consider the following element-wise inequality notation. For two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$,

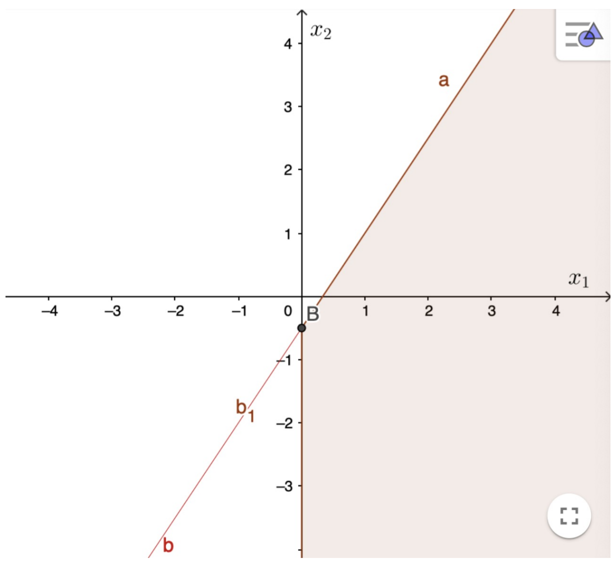
$$\mathbf{a} \leq \mathbf{b} \iff a_i \leq b_i \ \forall i = 1, 2, \ldots d. \tag{9}$$

Suppose we have a matrix $A \in \mathbb{R}^{2\times 2}$ and a vector $\mathbf{b} \in \mathbb{R}^2$ as follows.

$$A = \begin{bmatrix} 3 & -2 \\ 2 & 0 \end{bmatrix}, \mathbf{b} = [1, 0]. \tag{10}$$

Visualize the set of vectors where $A\mathbf{x} \geq \mathbf{b}$. Hint: A matrix vector product is a collection of dot products, and the above set can be obtained by the intersection of two of the sets constructed in the previous questions.

Solution $A\mathbf{x} = [3x_1 - 2x_2, 2x_1] \geq [1, 0]$, so it's the intersection of $3x_1 - 2x_2 \geq 1$ and $x_1 \geq 0$



4) **Problem 4: More gradient computations**

(i) Compute the gradient of the loss function below.

$$\text{Loss}(x, y, \mathbf{w}) = \sigma(-(\mathbf{w} \cdot \phi(x))y), \tag{11}$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is the logistic function.

Solution Let $z = (-\mathbf{w} \cdot \phi(x))y$, then $\text{Loss}(x, y, \mathbf{w}) = \sigma(z) = (1 + \exp(-z))^{-1}$. Applying the chain rule, we get

$$\begin{aligned} \nabla_{\mathbf{w}} \text{Loss}(x, y, \mathbf{w}) &= \frac{\partial \sigma(z)}{\partial z} \nabla_{\mathbf{w}} z & (12) \\ &= -(1 + \exp(-z))^{-2} \exp(-z) y \phi(x) & (13) \\ &= -(1 + \exp(-z))^{-1} \left(\frac{\exp -z}{1 + \exp(-z)} \right) y \phi(x) & (14) \\ &= -\sigma(z)(1 - \sigma(z)) y \phi(x). & (15) \end{aligned}$$

Plugging in the expression for z gives us the final expression.

$$\nabla_{\mathbf{w}} \text{Loss}(x, y, \mathbf{w}) = -\sigma(-(\mathbf{w} \cdot \phi(x))y)(1 - \sigma(-(\mathbf{w} \cdot \phi(x))y)) y \phi(x). \tag{16}$$

(ii) Suppose we have the following loss function.

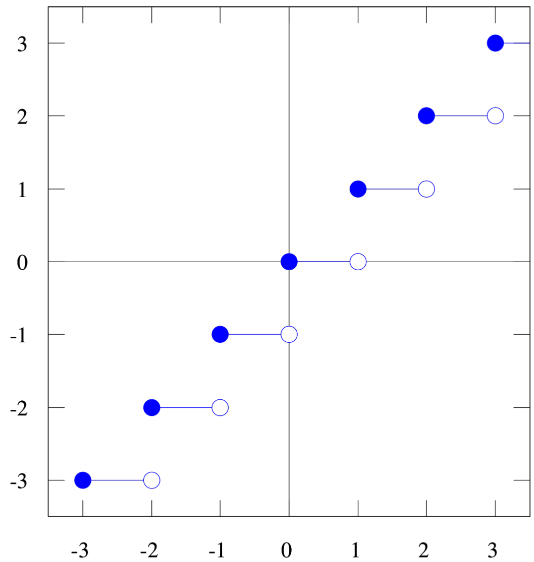
$$\text{Loss}(x, y, \mathbf{w}) = \max\{1 - \lfloor (\mathbf{w} \cdot \phi(x))y \rfloor, 0\}, \tag{17}$$

where $\lfloor a \rfloor$ returns a rounded down to the nearest integer. Determine what the gradient of this function looks like, and whether gradient descent is suitable to optimize this loss function.

Solution

$$\text{Loss}(x, y, \mathbf{w}) = \begin{cases} 1 - \lfloor (\mathbf{w} \cdot \phi(x))y \rfloor & \text{if } \lfloor (\mathbf{w} \cdot \phi(x))y \rfloor \leq 1, \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

If we draw the plot for the floor function, we can see that its derivative is 0 (the lines are flat and the slope is 0) almost everywhere.



Thus, when applying chain rule to find the gradient of $\text{Loss}(x, y, \mathbf{w})$, the computed gradient will also be 0 almost everywhere, so gradient descent is not suitable to optimize this function as the iterates would not move from the point of initialization.