DSC 140A - Discussion 03

Problem 1.

A subgradient of the absolute loss is:

$$\begin{cases} \operatorname{Aug}(\vec{x}), & \text{if } \operatorname{Aug}(\vec{x}) \cdot \vec{w} - y > 0, \\ -\operatorname{Aug}(\vec{x}), & \text{if} \operatorname{Aug}(\vec{x}) \cdot \vec{w} - y < 0, \\ \vec{0}, & \text{otherwise.} \end{cases}$$

Suppose you are running subgradient descent to minimize the risk with respect to the absolute loss in order to train a function $H(x) = w_0 + w_1 x$ on the following data set:

x	y
1	3
2	5
3	7

Suppose that the initial weight vector is $\vec{w} = (0,0)^T$ and that the learning rate $\eta = 1$. What will be the weight vector after one iteration of subgradient descent?

Solution: $(1,2)^T$

To perform subgradient descent, we need to compute the subgradient of the risk. The main thing to remember is that the subgradient of the risk is the *average* of the subgradient of the loss on each data point.

So to start this problem, calculuate the subgradient of the loss for each of the three points. Our formula for the subgradient of the absolute loss tells us to compute $\operatorname{Aug}(\vec{x}) \cdot w - y$ for each point and see if this is positive or negative. If it is positive, the subgradient is $\operatorname{Aug}(\vec{x})$; if it is negative, the subgradient is $-\operatorname{Aug}(\vec{x})$.

Now, the initial weight vector $\vec{w}$ was conveniently chosen to be $\vec{0}$, meaning that $\operatorname{Aug}(\vec{x}) \cdot \vec{w} = 0$ for all of our data points. Therefore, when we compute $\operatorname{Aug}(\vec{x}) \cdot \vec{w} - y$, we get $-y$ for every data point, and so we fall into the second case of the subgradient formula for every data point. This means that the subgradient of the loss at each data point is $-\operatorname{Aug}(\vec{x})$. Or, more concretely, the subgradient of the loss at each of the three data points is:

- $(-1,-1)^T$
- $(-1,-2)^T$
- $(-1,-3)^T$

This means that the subgradient of the risk is the average of these three:

$$\frac{1}{3}\left(\begin{bmatrix}-1\\-1\end{bmatrix} + \begin{bmatrix}-1\\-2\end{bmatrix} + \begin{bmatrix}-1\\-3\end{bmatrix}\right) = \begin{bmatrix}-1\\-2\end{bmatrix}$$

The subgradient descent update rule says that $\vec{w}^{(1)} = \vec{w}^{(0)} - \eta\vec{g}$, where $\vec{g}$ is the subgradient of the risk. The learning rate η was given as 1, so we have $\vec{w}^{(1)} = \vec{w}^{(0)} - \vec{g} = \vec{0} - \begin{bmatrix}-1\\-2\end{bmatrix} = \begin{bmatrix}1\\2\end{bmatrix}$.

Problem 2.

Consider the function $f(\vec{z}) = f(z_1, z_2) = \max(z_1, z_2)$.

a) Using the definition of the subgradient, check if $(1,1)^T$ is a subgradient at the point $(2,2)^T$.

Solution: The definition says that $\vec{s}$ is a subgradient if

$$f_s(\vec{z}) = f(\vec{z}^{(0)}) + \vec{s} \cdot (\vec{z} - \vec{z}^{(0)}) \leq f(\vec{z}),$$

for all $\vec{z}$.

In this case, we are testing $\vec{s} = (1,1)^T$ and $\vec{z}^{(0)} = (2,2)^T$. So:

$$\begin{aligned} f_s(\vec{z}) &= f(\vec{z}^{(0)}) + \vec{s} \cdot (\vec{z} - \vec{z}^{(0)}) \\ &= f(2,2) + (1,1)^T \cdot ((z_1, z_2) - (2,2)) \\ &= 2 + (1,1)^T \cdot (z_1 - 2, z_2 - 2) \\ &= 2 + (z_1 - 2 + z_2 - 2) \\ &= z_1 + z_2 - 2 \end{aligned}$$

Is $f_s(\vec{z}) = z_1 + z_2 - 2 \leq f(\vec{z})$ for all $\vec{z}$? To check, we can consider the two cases for $f(\vec{z})$: when $z_1 \geq z_2$ and when $z_2 \geq z_1$.

In the first case when $z_1 \geq z_2$, we have $f(\vec{z}) = \max(z_1, z_2) = z_1$. Is it true that $z_1 + z_2 - 2 \leq z_1$ for all $z_1 \geq z_2$? No, because if $z_1 = 10$ and $z_2 = 8$, then $f_s(z_1, z_2) = z_1 + z_2 - 2 = 16$, but $f(z_1, z_2) = z_1 = 10$.

Therefore, $\vec{s} = (1,1)^T$ is *not* a subgradient at $(2,2)^T$.

b) Show that $(1,0)^T$ is a valid subgradient at $(2,2)^T$.

Solution: In this case, we're testing $\vec{s} = (1,0)^T$ and $\vec{z}^{(0)} = (2,2)^T$. So:

$$\begin{aligned} f_s(\vec{z}) &= f(\vec{z}^{(0)}) + \vec{s} \cdot (\vec{z} - \vec{z}^{(0)}) \\ &= f(2,2) + (1,0)^T \cdot ((z_1, z_2) - (2,2)) \\ &= 2 + (1,0)^T \cdot (z_1 - 2, z_2 - 2) \\ &= 2 + (z_1 - 2) \\ &= z_1 \end{aligned}$$

It's true that $z_1 \leq \max(z_1, z_2)$ for all z_1, z_2, so $f_s(\vec{z}) = z_1 \leq f(\vec{z})$ for all $\vec{z}$.

c) Show that $(\frac{1}{2}, \frac{1}{2})^T$ *is* a subgradient at $(2,2)^T$.

Solution: We're testing $\vec{s} = (\frac{1}{2}, \frac{1}{2})^T$ and $\vec{z}^{(0)} = (2, 2)^T$. Therefore:

$$\begin{aligned}
f_s(\vec{z}) &= f(\vec{z}^{(0)}) + \vec{s} \cdot (\vec{z} - \vec{z}^{(0)}) \\
&= f(2,2) + (\frac{1}{2}, \frac{1}{2})^T \cdot ((z_1, z_2) - (2,2)) \\
&= 2 + (\frac{1}{2}, \frac{1}{2})^T \cdot (z_1 - 2, z_2 - 2) \\
&= 2 + \frac{1}{2}(z_1 - 2) + \frac{1}{2}(z_2 - 2) \\
&= 2 + \frac{1}{2}z_1 - 1 + \frac{1}{2}z_2 - 1 \\
&= \frac{1}{2}z_1 + \frac{1}{2}z_2 \\
&= \frac{z_1 + z_2}{2}
\end{aligned}$$

Is this always $\leq f(\vec{z})$? Let's break it into two cases: when $z_1 \geq z_2$ and when $z_2 > z_1$.

In the first case, $\max(z_1, z_2) = z_1$, so we need to check if $\frac{z_1 + z_2}{2} \leq z_1$ for all $z_1 \geq z_2$. This is true because the largest z_2 can be in this case is z_1, so:

$$\begin{aligned}
f_s(\vec{z}) &= \frac{z_1 + z_2}{2} \\
&\leq \frac{z_1 + z_1}{2} \\
&= z_1 \\
&= f(\vec{z})
\end{aligned}$$

So $f_s(\vec{z}) \leq f(\vec{z})$ for all $z_1 \geq z_2$.

Likewise, in the second case when $z_2 > z_1$, we have $\max(z_1, z_2) = z_2$. Here,

$$\begin{aligned}
f_s(\vec{z}) &= \frac{z_1 + z_2}{2} \\
&\leq \frac{z_2 + z_2}{2} \\
&= z_2 \\
&= f(\vec{z})
\end{aligned}$$

So in both cases, $f_s(\vec{z}) \leq f(\vec{z})$, and $(\frac{1}{2}, \frac{1}{2})^T$ is a subgradient at $(2, 2)^T$.

Problem 3.

The absolute loss of a linear predictor is

$$\ell_{\text{abs}}(\text{Aug}(x) \cdot \vec{w}, y) = |\,\text{Aug}(x) \cdot \vec{w} - y|.$$

We can write this as a piecewise function:

$$\ell_{\text{abs}}(\text{Aug}(x) \cdot \vec{w}, y) = \begin{cases} \text{Aug}(x) \cdot \vec{w} - y, & \text{if } \text{Aug}(x) \cdot \vec{w} - y > 0, \\ y - \text{Aug}(x) \cdot \vec{w}, & \text{if } \text{Aug}(x) \cdot \vec{w} - y < 0, \\ 0, & \text{if } \text{Aug}(x) \cdot \vec{w} = y. \end{cases}$$

This loss function is not differentiable at $\text{Aug}(x) \cdot \vec{w} = y$, but has a well-defined gradient everywhere else.

a) What is the gradient of the absolute loss with respect to $\vec{w}$ when $\text{Aug}(x) \cdot \vec{w} - y > 0$?

Solution: When $\text{Aug}(x) \cdot \vec{w} - y > 0$, we fall into the first case of the piecewise definition of the absolute loss. In this case, the loss is $\text{Aug}(x) \cdot \vec{w} - y$, so the gradient of the loss with respect to $\vec{w}$ is just $\text{Aug}(x)$.

b) What is the gradient of the absolute loss with respect to $\vec{w}$ when $\text{Aug}(x) \cdot \vec{w} - y < 0$?

Solution: When $\text{Aug}(x) \cdot \vec{w} - y < 0$, we fall into the second case of the piecewise definition of the absolute loss. In this case, the loss is $y - \text{Aug}(x) \cdot \vec{w}$, so the gradient of the loss with respect to $\vec{w}$ is $-\text{Aug}(x)$.

c) Optional: Show that $\vec{0}$ is a subgradient of the absolute loss at $\text{Aug}(x) \cdot \vec{w} = y$.

Solution: Showing that $\vec{0}$ is a subgradient of the absolute loss at $\text{Aug}(x) \cdot \vec{w} = y$ is maybe more straightforward than it seems, though it requires making sure that we know how to set up the problem correctly.

The definition of subgradient says that a vector $\vec{s}$ is a subgradient of $f(\vec{z})$ at $\vec{z}^{(0)}$ if $f_s(\vec{z}) = f(\vec{z}^{(0)}) + \vec{s} \cdot (\vec{z} - \vec{z}^{(0)}) \leq f(\vec{z})$ for all $\vec{z}$.

In this case, the role of $f(\vec{z})$ is played by the absolute loss function. Let's assume that $\vec{x}$ and y are fixed, and define $f(\vec{w}) = \ell_{\text{abs}}(\text{Aug}(x) \cdot \vec{w}, y)$.

The role of $f_s(\vec{z})$ is then played by

$$f_s(\vec{w}) = f(\vec{w}^{(0)}) + \vec{s} \cdot (\vec{w} - \vec{w}^{(0)}).$$

We're testing $\vec{s} = \vec{0}$ to see if it makes $f_s(\vec{w}) \leq f(\vec{w})$ for all $\vec{w}$. We're also assuming that $\vec{w}^{(0)}$ is such that $\text{Aug}(x) \cdot \vec{w}^{(0)} = y$. Note that this is exactly where the absolute loss is zero; in other words, $f(\vec{w}^{(0)}) = 0$. So:

$$f_s(\vec{w}) = 0 + \vec{0} \cdot (\vec{w} - \vec{w}^{(0)}) = 0 + 0 = 0$$

Here we used the fact that $\vec{0}$ dotted with anything is zero. So we've found that $f_s(\vec{w})$ is simply 0 for all $\vec{w}$.

Is $f_s(\vec{w}) = 0 \leq f(\vec{w})$ for all $\vec{w}$? Yes, because the absolute loss is always non-negative, and so $f(\vec{w}) \geq 0$. Therefore, $\vec{0}$ is a subgradient of the absolute loss.

Problem 4.

Using the definition, show that the function $f(\vec{w}) = a\vec{x} \cdot \vec{w} - b$ is convex as a function of $\vec{w}$, where $a, b \in \mathbb{R}$ and $\vec{x}, \vec{w} \in \mathbb{R}^d$.

Solution:

$$\nabla f(\vec{w}) = a\vec{x}$$
$$H_f(\vec{w}) = \mathbf{0}_{d,d} \succeq 0$$

In the Hessian, $H_f(\vec{w})$, all values are 0. We can see that the eigenvalues will also be 0. Therefore, $f(\vec{w})$ is convex.

Problem 5.

Let $f_1(\vec{w})$ and $f_2(\vec{w})$ be convex functions from $\mathbb{R}^d$ to $\mathbb{R}$. Define

$$f(\vec{w}) = \max\{f_1(\vec{w}), f_2(\vec{w})\}.$$

Show that $f(\vec{w})$ is convex.

Solution:

Take $t \in [0, 1]$.

$$\begin{aligned} f(t\vec{w}_1 + (1-t)\vec{w}_2) &= \max\{f_1(t\vec{w}_1 + (1-t)\vec{w}_2), f_2(t\vec{w}_1 + (1-t)\vec{w}_2)\} \\ &\leq \max\{tf_1(\vec{w}_1) + (1-t)f_1(\vec{w}_2), tf_2(\vec{w}_1) + (1-t)f_2(\vec{w}_2)\} \quad \text{(due to convexity of } f_1, f_2\text{)} \\ &\leq \max\{tf_1(\vec{w}_1), tf_2(\vec{w}_1)\} + \max\{(1-t)f_1(\vec{w}_2), (1-t)f_2(\vec{w}_2)\} \\ &= t\max\{f_1(\vec{w}_1), f_2(\vec{w}_1)\} + (1-t)\max\{f_1(\vec{w}_2), f_2(\vec{w}_2)\} \\ &= tf(w_1) + (1-t)f(w_2) \end{aligned}$$

Problem 6.

Recall that the Perceptron loss is:

$$L_{\text{perc}}(\vec{w}, \vec{x}, y) = \begin{cases} 0, & \text{if } \operatorname{sign}(\vec{w} \cdot \vec{x}) = y \text{ (correctly classified)}, \\ |\vec{w} \cdot \vec{x}|, & \text{if } \operatorname{sign}(\vec{w} \cdot \vec{x}) \neq y \text{ (misclassified)}. \end{cases}$$

Using the trick that $-y\,\vec{w} \cdot \vec{x} = |\vec{w} \cdot \vec{x}|$ in the case of misclassification, this can be be written in the equivalent form:

$$L_{\text{perc}}(\vec{w}, \vec{x}, y) = \max\{0, -y\,\vec{w} \cdot \vec{x}\}$$

Argue that the perceptron loss is convex as a function of $\vec{w}$.

Solution: We'll argue that the loss is the maximum of two convex functions, which (by Problem 2 above) is convex.

Let $f_1(\vec{w}) = 0$ and $f_2(\vec{w}) = -y\,\vec{w} \cdot \vec{x}$. Recognize that

$$L_{\text{perc}}(\vec{w}, \vec{x}, y) = \max\{f_1(\vec{w}),\ f_2(\vec{w})\}$$

f_2 is convex by the result of Problem 1 (with $a = y$ and $b = 0$). f_1 is constant, and trivially convex. Therefore L_{perc} is convex by the result of Problem 2.