

Problem Set 2, Sept 19, 2024 (Solutions to Theory Questions)

1 MAE Subgradient (Exercise 6)

We recall below the definition of a subgradient seen in Lecture 2.

Definition (Subgradient). A subgradient of a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ at a point \mathbf{w} is any vector $\mathbf{s} \in \mathbb{R}^d$ such that

$$\forall \mathbf{z}, f(\mathbf{z}) \geq f(\mathbf{w}) + \mathbf{s}^\top (\mathbf{z} - \mathbf{w}). \quad (1)$$

There can be more than one such vector \mathbf{s} (or none, for general nonconvex functions) at points where f is not differentiable. The set of all subgradients, so the vectors satisfying property (1), is denoted as

$$\partial f(\mathbf{w}) = \{\mathbf{s} \mid \mathbf{s} \in \mathbb{R}^d \text{ such that } \forall \mathbf{z}, f(\mathbf{z}) \geq f(\mathbf{w}) + \mathbf{s}^\top (\mathbf{z} - \mathbf{w})\}.$$

In this exercise, we ask you to derive the expression of a subgradient of the MAE loss $\mathcal{L}(\mathbf{w}) : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\mathcal{L}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N |y_n - \mathbf{x}_n^\top \mathbf{w}|$, which is not differentiable due to the presence of the absolute value function. You are therefore looking for a subgradient vector \mathbf{s} of the combined function such that

$$\mathbf{s} \in \partial \mathcal{L}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \partial |y_n - \mathbf{x}_n^\top \mathbf{w}|.$$

Note that we can write each summand of $\mathcal{L}(\mathbf{w})$ as $h(q_n(\mathbf{w}))$, where $h : \mathbb{R} \rightarrow \mathbb{R}$, $h(e) := |e|$ and $q_n : \mathbb{R}^2 \rightarrow \mathbb{R}$, $q_n(\mathbf{w}) := y_n - \mathbf{x}_n^\top \mathbf{w}$. As given in the annotated notes of Lecture 2, we can use the **chain-rule for subgradients** for $h(q(\mathbf{w}))$, when the outer function h is not differentiable and q is differentiable. Then, any vector

$$\mathbf{s} \in \partial h(q_n(\mathbf{w})) \cdot \nabla q_n(\mathbf{w})$$

is a subgradient of $h(q_n(\mathbf{w}))$, where we can pick any element of $\partial h(q_n(\mathbf{w}))$ and multiply it with $\nabla q_n(\mathbf{w})$. We immediately see that $\nabla q_n(\mathbf{w}) = -\mathbf{x}_n$.

Regarding ∂h , we saw in Lecture 2 that the set of subgradients of $h = |e|$ at a point e is

$$\partial h(e) = \begin{cases} -1, & e < 0, \\ [-1, 1], & e = 0, \\ 1, & e > 0. \end{cases}$$

Then, a possible subgradient of h at a point e is for example given by

$$\text{sign}(e) := \begin{cases} -1, & e < 0, \\ 0, & e = 0, \\ 1, & e > 0, \end{cases}$$

where we selected a single value in the interval $[-1, 1]$ from $\partial h(0)$ (namely the value 0).

The expression of $\mathbf{s} \in \partial h(q_n(\mathbf{w})) \cdot \nabla q_n(\mathbf{w})$ therefore is

$$\mathbf{s} = \underbrace{\text{sign}(y_n - \mathbf{x}_n^\top \mathbf{w})}_{\in \partial h(q_n(\mathbf{w}))} \cdot \underbrace{(-\mathbf{x}_n)}_{= \nabla q_n(\mathbf{w})}.$$

We can then write a subgradient for the entire loss by summing up the subgradients we found for each \mathcal{L}_n , so

$$-\frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \cdot \text{sign}(y_n - \mathbf{x}_n^\top \mathbf{w}) \in \partial \mathcal{L}(\mathbf{w}).$$

Finally, we can rewrite this using a more compact notation (which will be useful for your Python implementation):

$$= -\frac{1}{N} \mathbf{X}^\top \cdot \text{sign}(\mathbf{e}),$$

where $\mathbf{e} := \mathbf{y} - \mathbf{X} \cdot \mathbf{w}$ and sign applied element-wise to \mathbf{e} , and \mathbf{X} is the matrix collecting all datapoints as its rows.