Project 1 - Linear Regression and Gradient descent
(Solutions to Theory Questions)

Subgradient

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

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

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

There can be more than one such vector $\boldsymbol{s}$ (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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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