Name: _____

Learning as Optimization

5. (14 points) Ben develops a new hypothesis class: $h(x; w_1, w_2) = w_1 x_1 + w_1 x_1^2 + w_2 x_2 + w_2 x_2^2$, where $x = (x_1, x_2)$. He plans to use it for a regression problem on the data set $S_n = \{(x^{(1)}, y^{(1)}), ..., (x^{(n)}, y^{(n)})\}$.

   (a) Ben will use batch gradient descent to compute model parameters w_1, w_2. His loss function is mean squared error (MSE). Derive an update rule for w_1 given the learning rate η.

   Solution:
   Assume, for simplicity, that the batch size is equal to n.

$$L(g, a) = \frac{1}{n} \sum_{i=1}^{n} (w_1 x_1^{(i)} + w_1 x_1^{(i)^2} + w_2 x_2^{(i)} + w_2 x_2^{(i)2} - y^{(i)})^2$$

$$\frac{\delta L}{\delta w_1} = \frac{2}{n} \sum_{i=1}^{n} (w_1 x_1^{(i)} + w_1 x_1^{(i)2} + w_2 x_2^{(i)} + w_2 x_2^{(i)2} - y^{(i)})(x_1^{(i)} + x_1^{(i)2})$$

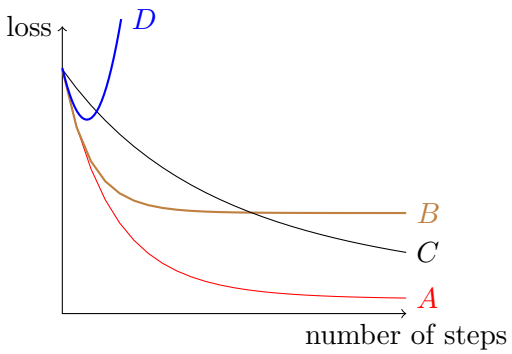
$$w_1 := w_1 - \frac{2\eta}{n} \sum_{i=1}^{n} (w_1 x_1^{(i)} + w_1 x_1^{(i)2} + w_2 x_2^{(i)} + w_2 x_2^{(i)2} - y^{(i)})(x_1^{(i)} + x_1^{(i)2})$$

   (b) Describe the shape of the MSE as a function of w_1 and w_2. How many minima will it have? Assume that the data set S_n is fixed.

   Solution: Paraboloid (all values positive). Single minimum.

Name: _____

(c) Ben tries different settings of the learning rate η. Depending on the setting he obtains different behavior of the gradient descent algorithm. Match each plot (A,B,C,D) to the best fitting description (assume MSE loss).



Learning rate too low (select one):
◯ A ◯ B √ **C** ◯ D

Learning rate about right (select one):
√ **A** ◯ B ◯ C ◯ D

Learning rate too high (select one):
◯ A √ **B** ◯ C ◯ D

Learning rate much too high (select one):
◯ A ◯ B ◯ C √ **D**

Solution: A low learning rate leads to slow decay of the loss– this occurs for C. A good learning rate leads to moderately quick decay to low loss, as in A. A high learning rate can prevent gradient descent from reaching the global minimum of the objective, and can cause it to oscillate between parameter values that give a higher loss value, as in B. A very high learning rate can cause gradient descent to diverge, as in D.

(d) Alyssa suggests using a mean absolute error, instead, defined by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y^{(i)} - h(x^{(i)}, w_1, w_2)\right|$$

What could be an advantage of this approach?

Solution: Robustness to outlier (or high noise). The MSE loss over-penalizes for the samples that have large errors comparing to the smaller ones. The MAE loss equally penalizes all the samples, because the gradient slope is always 1 (except at 0), which favors more majority opinions than MSE.