Name: ______________________________

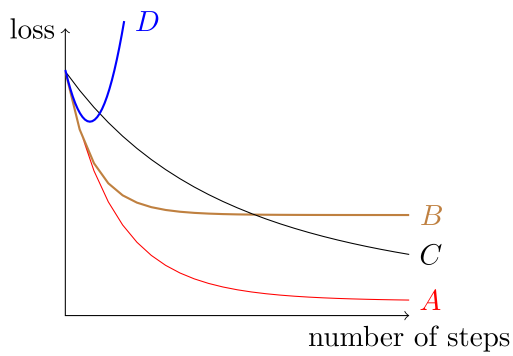
Learning as Optimization

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.

   (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.

(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

(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?