DSC 40A - Group Work Session 1
due Friday, July 7 at 11:59pm

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. **One person** from each group should submit your solutions to Gradescope and **tag all group members** so everyone gets credit.

This worksheet won't be graded on correctness, but rather on good-faith effort. Even if you don't solve any of the problems, you should include some explanation of what you thought about and discussed, so that you can get credit for spending time on the assignment.

In order to receive full credit, you must work in a group of two to four students in your assigned discussion section. You can also self-organize a group and meet outside of discussion section for 95 percent credit. You may not do the groupwork alone.

1 Summation Notation

You can often verify for yourself if something is true about summation notation by "expanding" the summation symbol and seeing if the property holds. For instance, suppose we want to see if it is true that

$$\sum_{i=1}^{n} c \cdot x_i = c \sum_{i=1}^{n} x_i$$

We start by "expanding" $\sum_{i=1}^{n} c \cdot x_i$:

$$\sum_{i=1}^{n} c \cdot x_i = cx_1 + cx_2 + cx_3 + \ldots + cx_n$$

Now we see that the c can be factored out:

$$\begin{aligned} &= c(x_1 + x_2 + x_3 + \ldots + x_n) \\ &= c \sum_{i=1}^{n} x_i. \end{aligned}$$

This is a simple proof that the property is true. On the other hand, we can prove that a property doesn't hold in the same way: by expanding both sides and showing that they are not equal.

Problem 1.

Show that $\sum_{i=1}^{n}(x_i + y_i) = \left(\sum_{i=1}^{n} x_i\right) + \left(\sum_{i=1}^{n} y_i\right)$.

Problem 2.

Find a simple expression for $\sum_{i=1}^{n} c$ not involving summation notation. Show that your expression is correct.

2 Minimizers and Maximizers

We've seen that machine learning problems must first be formulated as mathematical problems. Many of these mathematical problems turn out to be optimization problems: finding the value that minimizes or maximizes a function.

For a function of one variable $f(x)$, a value x^* is said to be a **minimizer** of $f(x)$ if

$$f(x^*) \leq f(x) \quad \text{for all } x.$$

Similarly, x^* is said to be a **maximizer** of $f(x)$ if

$$f(x^*) \geq f(x) \quad \text{for all } x.$$

Notice that a function can have multiple minimizers or maximizers. For example, a constant function like $f(x) = 5$ is minimized at all values of x, and it's also maximized at all values of x!

Problem 3.

Should the blank below be filled in with the word *minimizer* or *maximizer* or neither? Prove your result.

If x^* is a minimizer of $f(x)$ then it's a _____ *of $g(x) = 5f(x) + 3$.*

Problem 4.

Should the blank below be filled in with the word *minimizer* or *maximizer* or neither? Prove your result.

If x^* is a minimizer of $f(x)$ then it's a _____ *of $g(x) = -(f(x))^2$.*

3 Empirical Risk Minimization

In class, we've seen how to minimize the empirical risk associated with certain natural loss functions, such as the absolute loss and the squared loss. There are a variety of other possible loss functions we could use instead. This problem explores empirical risk minimization with an alternate choice of loss function.

Problem 5.

In this problem, consider the loss function

$$L(h, y) = \begin{cases} 1, & |y - h| > 1 \\ |y - h|, & |y - h| \leq 1 \end{cases}.$$

a) Consider y to be a fixed number. Plot $L(h, y)$ as a function of h.

b) Suppose that we have the following data:

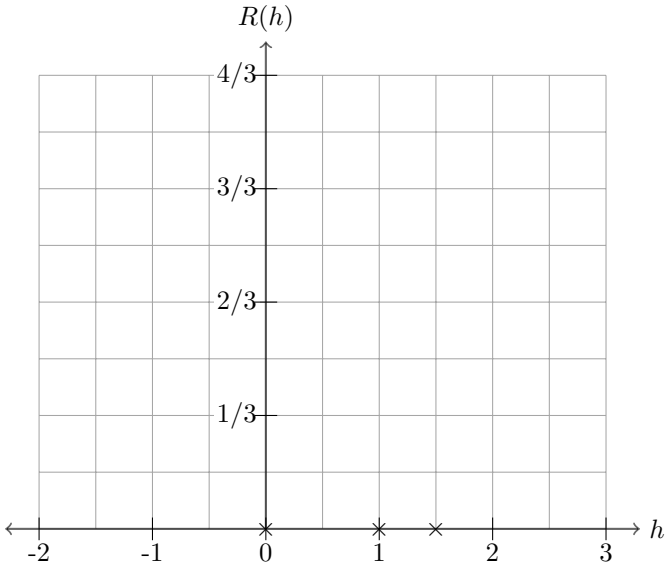
$$y_1 = 0$$
$$y_2 = 1$$
$$y_3 = 1.5$$

Plot the empirical risk

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, y_i)$$

on the domain $[-2, 3]$. It might help to use the grid on the next page; note that the vertical axis tick marks occur in increments of 1/3 while the horizontal axis tick marks are in increments of 1.

Hint: $R(h)$ is made up of several line segments. What is the slope of each line segment?



Note: You should be able to do this by hand without using technology. After you've done this, https://www.desmos.com/calculator/g2opj3owgeclick this link to check your work using Desmos, an online graphing calculator.

c) Suppose that we are interested in finding the typical price of an avocado using this loss function. To do so, we have gathered a data set of n avocado prices, $y_1, \ldots, y_n$, and we found the price h^* which minimized the empirical risk (a.k.a, average loss), $R(h) = \frac{1}{n}\sum_{i=1}^{n} L(h, y_i)$.

Unfortunately, a flat tax of c dollars has been imposed on avocados since we performed our analysis, increasing every price in our data set by c.

Is it true that $h^* + c$ is a minimizer of R when we use the new prices, $(y_1 + c), (y_2 + c), \ldots, (y_n + c)$? Explain why or why not by explaining how the graph of R changes.

d) Given avocado prices $\{1/4, 1/2, 3/4, 7/8, 9/8\}$, find a minimizer of R. Provide some justification for your answer.

Hint: you don't need to plot R or do any calculation to find the answer.

4 Gradient Descent

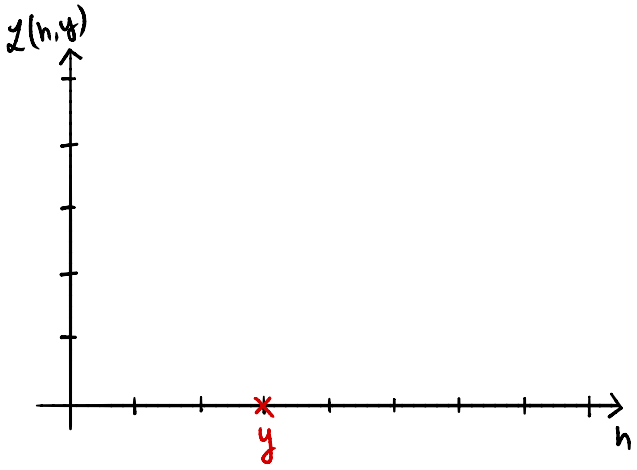
Gradient descent is an algorithm used to minimize differentiable functions. In this class, we will primarily use it to minimize empirical risk, though its use is much more broad. In this problem, we'll explore a new loss function and see how our initial prediction impacts how much our prediction changes with one iteration.

Problem 6.

Consider a new loss function,

$$L(h, y) = \begin{cases} (h-y)^2, & h < y \\ (h-y)^3 & h \geq y \end{cases}.$$

a) Fix an arbitrary value of y. On the axes below, draw the graph of $L(h, y)$ as a function of h.



b) For any data set $y_1, \ldots, y_n$, the empirical risk, $R(h) = \frac{1}{n}\sum_{i=1}^{n} L(h, y_i)$ will be differentiable and convex (also known as concave up). This means gradient descent is guaranteed to be able to find its minimum value. It might take many iterations or only a few. We won't do the whole algorithm, just one iteration for practice.

Suppose our data set is $\{2, 3, 6, 10\}$. Perform one iteration of gradient descent by hand on the empirical risk function $R(h)$ for this data set, starting with an initial prediction of $h_0 = 3$ and using a step size of $\alpha = \frac{1}{10}$. Calculate h_1, the prediction after the first iteration.

c) For the same data set $\{2, 3, 6, 10\}$, suppose instead we did one iteration of gradient descent on $R(h)$ but starting at a different initial prediction, $h_0 = 7$. Using the same step size of $\alpha = \frac{1}{10}$, calculate h_1, the prediction after the first iteration, for this new starting point.

d) Compare your answers to the previous parts and notice that the prediction moves only a little bit when we start at $h_0 = 3$, but it moves a lot when we start at $h_0 = 7$. Can you explain why that happens by looking at the loss function we're using?