DSC 40A - Homework 2

Due: Wednesday, Jan 24 at 11:59pm

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Homeworks are due to Gradescope by 11:59pm on the due date. You can use a slip day to extend the deadline by 24 hours.

Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should **always explain and justify** your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Homeworks should be written up and turned in by each student individually. You may talk to other students in the class about the problems and discuss solution strategies, but you should not share any written communication and you should not check answers with classmates. You can tell someone how to do a homework problem, but you cannot show them how to do it.

For each problem you submit, you should **cite your sources** by including a list of names of other students with whom you discussed the problem. Instructors do not need to be cited.

This homework will be graded out of 50 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Problem 1. Reflection and Feedback Form

🥑🥑 Make sure to use your UCSD google account to fill out this Reflection and Feedback Form, linked here for additional points on this homework! This form is primarily for your benefit; research shows that reflecting and summarizing knowledge helps you understand and remember it.

Problem 2. Linear Transformations

Suppose we are given a data set $\{d_1, d_2, \ldots, d_n\}$ and know its mean, variance, and standard deviation to be $mean_d$, var_d, and std_d. Consider another data set $\{t_1, t_2, \ldots, t_n\}$, where t_i is a linear transformation of d_i:

$$t_i = f(d_i) = a \cdot d_i + b$$

for each $i = 1, 2, \ldots, n$. Here, a and b are arbitrary constants. Let $mean_t$, var_t, and std_t be the mean, variance, and standard deviation of the transformed data.

a) 🥑🥑🥑 Express $mean_t$ in terms of $mean_d$, a, and b (you may not need all of these).

Solution:

$$\begin{aligned} mean_t &= \frac{1}{n}\sum_{i=1}^{n}(a \cdot d_i + b) \\ &= a \cdot \left(\frac{1}{n}\sum_{i=1}^{n} d_i\right) + b \\ &= a \cdot mean_d + b \end{aligned}$$

b) 🥑🥑🥑 Express var_t in terms of var_d, a, and b (you may not need all of these).

Solution:

$$\begin{aligned} var_t &= \frac{1}{n}\sum_{i=1}^{n}(a \cdot d_i + b - mean_t)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(a \cdot d_i + b - (a \cdot mean_d + b))^2 \\ &= a^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(d_i - mean_d)^2 \\ &= a^2 \cdot var_d \end{aligned}$$

c) Express std_t in terms of std_d, a, and b (you may not need all of these).

Solution:

$$\begin{aligned} std_t &= \sqrt{var_t} \\ &= \sqrt{a^2 \cdot var_d} \\ &= |a| \cdot \sqrt{var_d} \\ &= |a| \cdot std_d. \end{aligned}$$

Problem 3. Quadratic Mean

Suppose we are given a data set of size n with $0 < y_1 \leq y_2 \leq \cdots \leq y_n$.

Define a new loss function by

$$L_Q(h, y) = (h^2 - y^2)^2$$

and consider the empirical risk

$$R_Q(h) = \frac{1}{n}\sum_{i=1}^{n} L_Q(h, y_i).$$

a) Show that $R(h)$ has critical points at $h = 0$ and when h equals the **quadratic mean** of the data, defined as

$$QM(y_1, y_2, \ldots, y_n) = \sqrt{\frac{y_1^2 + y_2^2 + \cdots + y_n^2}{n}}.$$

Solution: We start by finding the critical points of $R_Q(h)$.

$$\begin{aligned}
R_Q'(h) &= \frac{d}{dh}\left(\frac{1}{n}\sum_{i=1}^{n}(h^2 - y_i^2)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{dh}(h^2 - y_i^2)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}2(h^2 - y_i^2)\cdot 2h \\
&= \frac{4h}{n}\sum_{i=1}^{n}(h^2 - y_i^2) \\
&= \frac{4h}{n}\left(nh^2 - \sum_{i=1}^{n}y_i^2\right)
\end{aligned}$$

Let $R_Q'(h) = 0$, then either $\dfrac{4h}{n} = 0$, or $nh^2 - \displaystyle\sum_{i=1}^{n} y_i^2 = 0$.

Therefore, $R_Q(h)$ has critical points when:

$$h = 0 \text{ and } h = \sqrt{\frac{y_1^2 + y_2^2 + \cdots + y_n^2}{n}} = QM(y_1, y_2, \cdots, y_n).$$

b) Recall from single-variable calculus the **second derivative test**, which says that for a function f with critical point at x^*,

- if $f''(x^*) > 0$, then x^* is a local minimum, and
- if $f''(x^*) < 0$, then x^* is a local maximum.

Use the second derivative test to determine whether each critical point you found in part (a) is a maximum or minimum of $R_Q(h)$.

Solution: For a critical point h to be a minimum, we must show that $R_Q''(h) > 0$. For a critical point to be a maximum, we must show that $R_Q''(h) < 0$.

We start by rewriting the first derivative and taking the second derivative:

$$\begin{aligned}
R_Q'(h) &= \frac{4}{n}\left(nh^3 - h\sum_{i=1}^{n}y_i^2\right) \\
R_Q''(h) &= \frac{4}{n}\left(\frac{d}{dh}\left(nh^3\right) - \frac{d}{dh}\left(h\sum_{i=1}^{n}y_i^2\right)\right) \\
&= \frac{4}{n}\left(3nh^2 - \sum_{i=1}^{n}y_i^2\right) \\
&= 12h^2 - \frac{4}{n}\sum_{i=1}^{n}y_i^2
\end{aligned}$$

When $h = 0$, $R_Q''(h) = -\frac{4}{n}\sum_{i=1}^{n} y_i^2 < 0$ because $0 < y_1 \leq y_2 \leq \cdots \leq y_n$. Therefore $h = 0$ is a local maximum.

When h is the quadratic mean, $h^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2$, so

$$\begin{aligned}
R_Q''(h) &= 12h^2 - \frac{4}{n}\sum_{i=1}^{n} y_i^2 \\
&= \frac{12}{n}\sum_{i=1}^{n} y_i^2 - \frac{4}{n}\sum_{i=1}^{n} y_i^2 \\
&= \frac{8}{n}\sum_{i=1}^{n} y_i^2 \\
&> 0 \text{ because } 0 < y_1 \leq y_2 \leq \cdots \leq y_n.
\end{aligned}$$

Therefore, $h = QM(y_1, y_2, \ldots, y_n)$ is a local minimum.

c) Show that the quadratic mean always falls between the smallest and largest data values, which is a property that any reasonable prediction should have. This amounts to proving the inequality

$$y_1 \leq QM(y_1, y_2, \ldots, y_n) \leq y_n.$$

Solution: We start by proving $y_1 \leq QM(y_1, y_2, \ldots, y_n)$.

Given $0 < y_1 \leq y_2 \leq \cdots \leq y_n$, then we know $y_1 \leq y_i$ for $i = 1, 2, \ldots, n$. This means $y_1^2 \leq y_i^2$ for $i = 1, 2, \ldots, n$.

Therefore,

$$\begin{aligned}
QM(y_1, y_2, \ldots, y_n) &= \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n}} \\
&\geq \sqrt{\frac{\sum_{i=1}^{n} y_1^2}{n}} \\
&= \sqrt{\frac{n \cdot y_1^2}{n}} \\
&= \sqrt{y_1^2} \\
&= y_1
\end{aligned}$$

This shows $QM(y_1, y_2, \ldots, y_n) \geq y_1$.

The second part of the inequality can be proved similarly. We are given that $0 < y_1 \leq y_2 \leq \cdots \leq y_n$, which implies $y_n \geq y_i$ for $i = 1, 2, \ldots, n$. This means $y_n^2 \geq y_i^2$ for $i = 1, 2, \ldots, n$.

Therefore,

$$\begin{aligned}
QM(y_1, y_2, \ldots, y_n) &= \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n}} \\
&\leq \sqrt{\frac{\sum_{i=1}^{n} y_n^2}{n}} \\
&= \sqrt{\frac{n \cdot y_n^2}{n}} \\
&= \sqrt{y_n^2} \\
&= y_n
\end{aligned}$$

This shows $QM(y_1, y_2, \ldots, y_n) \leq y_n$.

Problem 4. Mean as a minimizer

Suppose we have a cluster of one dimensional data points $x_1, x_2, \ldots, x_n \in \mathbb{R}$, we want to choose a value x^* such that it minimizes the total squared cost:

$$L(x^*) = \sum_{i=1}^{n} (x_i - x^*)^2$$

Prove that the mean of data points, $\overline{x} = \sum_i x_i/n$ is the minimizer of $L(x^*)$.

Hint: One way to do this is by calculus, that is, to take the derivative $L'(x^*)$ and set to zero. Another way is to show: $\sum_{i=1}^{n}(x_i - x^*)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - x^*)^2$. Since $n(\overline{x} - x^*)^2$ is non-negative, we have $\overline{x}$ as the minimizer of $L(x)$.

Solution:

$$\begin{aligned}
\sum_i (x_i - x^*)^2 &= \sum_i (x_i - \overline{x} + \overline{x} - x^*)^2 \\
&= \sum_i (x_i - \overline{x})^2 + n(\overline{x} - x^*)^2 + (\overline{x} - x^*) \cdot \underbrace{(\sum_i x_i - n\overline{x})}_{=0} \\
&= \sum_i (x_i - \overline{x})^2 + n(\overline{x} - x^*)^2
\end{aligned}$$

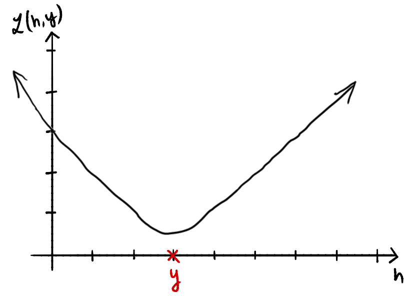
Problem 5. Huber Loss

The *Huber loss* is a mixture between the square loss and the absolute loss. It is defined piecewise as follows:

$$L_{\text{hub}}(h, y) = \begin{cases} |h - y|, & |h - y| > 1 \\ \frac{1}{2}(h - y)^2 + \frac{1}{2}, & |h - y| \leq 1 \end{cases}$$

a) Fix an arbitrary value of y. Draw the graph of $L_{\text{hub}}(h, y)$ as a function of h. You should notice that $L_{\text{hub}}(h, y)$ is minimized at y.

Solution:



b) What is the derivative of L_{hub} with respect to h? Your answer should also be a piecewise function.

Solution: We break L_{hub} into pieces and find the slope in each part. When $h - y < -1$, the function looks like $y - h$ and the derivative with respect to h is -1. Likewise, when $h - y > 1$, the function looks like $h - y$ and the derivative with respect to h is 1. In the middle, when $|h - y| \leq 1$, the function looks like $\frac{1}{2}(h-y)^2 + \frac{1}{2}$, and the derivative with respect to h is $h - y$. Hence

$$\frac{dL_{\text{hub}}}{dh}(h) = \begin{cases} -1, & h - y < -1 \\ 1, & h - y > 1 \\ h - y, & \text{otherwise} \end{cases}$$

Problem 6. The Behavior of Gradient Descent

Sometimes in optimization, it is difficult or impossible to find a closed-form solution for the objective function, such as the formulae for w_0 and w_1 in least squares regression. In such cases, one can use an iterative method such as gradient descent to find the optimal solution (or get close to it).

Recall the iterative step/update rule in gradient descent:

$$h_i = h_{i-1} - \alpha \cdot \frac{dR}{dh}(h_{i-1})$$

where α is a chosen learning rate.

a) Before we begin gradient descent, let's check whether the algorithm is guaranteed to work. Using the definition of convexity, show whether gradient descent is guaranteed to find the optimal solution (i.e. the minimum) for the function,

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = x^2.$$

Solution: Recall the definition of a convex function:

$$f : \mathbb{R} \to \mathbb{R}, \quad f(ta + (1-t)b) \leq tf(a) + (1-t)f(b), \quad \forall a, b \in \mathbb{R}, \forall t \in [0, 1].$$

We can show whether gradient descent is guaranteed to find the optimal solution if the function is convex (and differentiable). To show that $f(x) = x^2$ is convex using the definition, we can show that the LHS $\leq$ RHS by showing that RHS $-$ LHS ≥ 0.

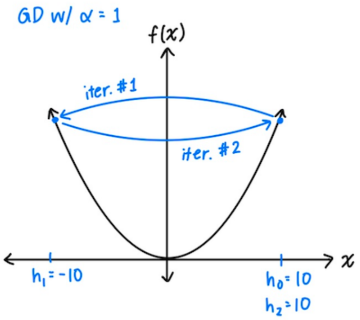
$$\begin{aligned}
\text{LHS: } f(ta+(1-t)b) &= (ta+(1-t)b)^2 \\
&= t^2a^2 + (1-t)^2b^2 + 2t(1-t)ab \\
\text{RHS: } tf(a)+(1-t)f(b) &= ta^2 + (1-t)b^2 \\
\text{RHS - LHS: } tf(a)+(1-t)f(b) - f(ta+(1-t)b) &= t(1-t)a^2 + [(1-t)-(1-t)^2]b^2 - 2t(1-t)ab \\
&= t(1-t)a^2 + t(1-t)b^2 - 2t(1-t)ab \\
&= t(1-t)(a-b)^2 \\
&\geq 0 \quad (\text{since } t, 1-t, (a-b)^2 \geq 0)
\end{aligned}$$

Thus, $f(x) = x^2$ is convex by definitions.

b) In lecture, you may have encountered the phrase, ”given an appropriate step size, α” or ”small enough α”. Let’s experiment with what this might mean. Suppose one chooses a starting prediction of $h_0 = 10$. Run two iterations of gradient descent with $\alpha = 1$ and $\alpha = 0.5$ on the function $f(x) = x^2$. Draw a graph depicting the results of the two gradient descent algorithms and explain which α is a better choice of a learning rate over the other.

Solution: To run gradient descent, one needs to find the gradient of $f(x) = x^2$. $\nabla f(x) = \frac{df}{dx}(x) = 2x$. Now, let’s run gradient descent with $\alpha = 1$.

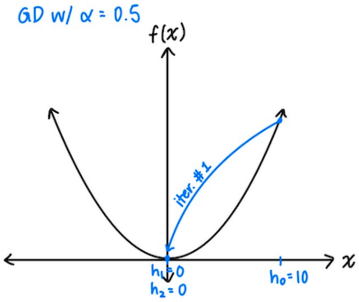
$$\begin{aligned}
h_0 &= 10 \\
h_1 &= h_0 - \alpha \cdot 2(h_0) \\
&= 10 - 2(10) \\
&= -10 \\
h_2 &= h_1 - \alpha \cdot 2(h_1) \\
&= -10 - 2(-10) \\
&= 10
\end{aligned}$$



As we can see, $\alpha = 1$ leads us to oscillate between the values of -10 and 10, and not progressing towards a minimum value.

Next, we implement two iterations of gradient descent with $\alpha = 0.5$.

$$\begin{aligned}
h_0 &= 10 \\
h_1 &= h_0 - \alpha \cdot 2(h_0) \\
&= 10 - 0.5 \cdot 2(10) \\
&= 0 \\
h_2 &= h_1 - \alpha \cdot 2(h_1) \\
&= 0 - 0.5 \cdot 2(0) \\
&= 0
\end{aligned}$$



Using $\alpha = 0.5$ led us to converge after one step! Clearly, $\alpha = 0.5$ is a better choice for the learning rate because it converges quickly which means it requires less iterations (and less computation), and $\alpha = 1$ fails to converge even with an infinite number of iterations.

c) Find the optimal solution using calculus methods (i.e., find the critical points) for the following function,

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x_1, x_2) = 3x_1^2 + 0.5x_2^2.$$

Suppose one runs gradient descent on f with a small enough α. Write the update rule for f. Explain whether x_1 or x_2 reaches their respective minimum value at a faster rate than the other.

Solution:

$$\begin{aligned}
f(x_1, x_2) &= 3x_1^2 + 0.5x_2^2 \\
\nabla f(x_1, x_2) &= \begin{bmatrix} 6x_1 \\ x_2 \end{bmatrix} \\
\begin{bmatrix} 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} 6x_1 \\ x_2 \end{bmatrix}
\end{aligned}$$

There is one critical point, $x_1 = 0, x_2 = 0$. Using the second derivative test, we can show that this point is the minimum. Alternatively, f is convex in the direction of x_1 and x_2, so the optimal solution of a convex function is always a minimum.

The update rule for f is

$$h_i = h_{i-1} - \alpha \begin{bmatrix} 6 \\ 1 \end{bmatrix} \cdot h_{i-1}.$$

Since x_1 changes at a faster rate than x_2 ($6 > 1$), one would expect x_1 to reach its minimum value of 0 at a faster rate than x_2, for a small enough α.