DSC 40A - Group Work Session 2
due Monday, April 15th 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 for at least 50 minutes in your assigned discussion section. You can also self-organize a group and meet outside of discussion section for 80 percent credit. You may not do the groupwork alone.

1 Limits

When studying the theoretical foundations of data science, we often say that some theorem or formula holds true "in the limit" as some value grows to infinity. For example, you may recall such language when you learned about the Central Limit Theorem in DSC 10, which said that as we collect more and more samples of a fixed sample size from a population, the distribution of sample means approached a normal distribution with mean population mean and standard deviation $\frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$. We even saw an example in Lecture 3, where as $p \to \infty$, the constant minimizer h^* of the empirical risk $\frac{1}{n}\sum_{i=1}^{n} |y_i - h|^p$ approached the midrange of $y_1, y_2, ..., y_n$. Here, we'll work through a few problems that involve the idea of a limit, as limits will continue to appear in this class and in the future.

Consider a dataset of n numbers $y_1, y_2, ..., y_n$ with mean M and standard deviation S. Suppose we introduce k new values to the dataset, $y_{n+1}, y_{n+2}, ..., y_{n+k}$, all of which are equal to 25.

Let the new mean and standard deviation of all $n + k$ values be M' and S', respectively.

Problem 1.

Find M' in terms of M, n, k, and S. (You may not need to use all of these variables in your answer.)

Problem 2.

Evaluate the following limit:

$$\lim_{k \to \infty} M'$$

That is, as we introduce more and more values equal to 25, what does the mean of the resulting dataset approach?

Problem 3.

Evaluate the following limit:

$$\lim_{k \to \infty} S'$$

That is, as we introduce more and more values equal to 25, what does the standard deviation of the resulting dataset approach? *Hint: You can answer this question conceptually, without first finding the value of S'. If we add more and more values equal to 25, what happens to the spread of the dataset?*

2 Pythagorean Means

Recall the geometric mean from Groupwork 1. It turns out that the geometric mean is one of three means, collectively known as the "Pythagorean means." These comprise of the arithmetic mean, the geometric mean, and the harmonic mean. For an arbitrary dataset, $y_1, ..., y_n$, **where each** $y_i > 0$, they are defined as follows:

- Arithmetic mean: $\frac{1}{n}\sum_{i=1}^{n} y_i$.
- Geometric mean: $\left(\prod_{i=1}^{n} y_i\right)^{1/n}$.
- Harmonic mean: $\frac{n}{\sum_{i=1}^{n} \frac{1}{y_i}}$.

Each mean is a valid measure of center and has different uses and applications over the others. For example, the arithmetic mean is the best choice when aggregating information about counts or frequencies, whereas the harmonic mean is the best choice for averaging multiple rates, or frequencies (like speeds, as you saw in this example in DSC 10).

Problem 4.

For the following dataset, compute its arithmetic mean, geometric mean, and harmonic mean.

$$1, 2, 2, 4, 8$$

Problem 5.

In the above example, you may have noticed that the arithmetic mean $\geq$ geometric mean $\geq$ harmonic mean. This inequality is true in general, for any dataset of positive numbers $y_1, y_2, ..., y_n$, and is referred to as the AM-GM-HM inequality.

Use the fact that the AM-GM inequality holds true to prove the GM-HM inequality. That is, assuming that $\frac{1}{n}\sum_{i=1}^{n} y_i \geq \left(\prod_{i=1}^{n} y_i\right)^{1/n}$ is true, prove that:

$$\left(\prod_{i=1}^{n} y_i\right)^{1/n} \geq \frac{n}{\sum_{i=1}^{n} \frac{1}{y_i}}$$

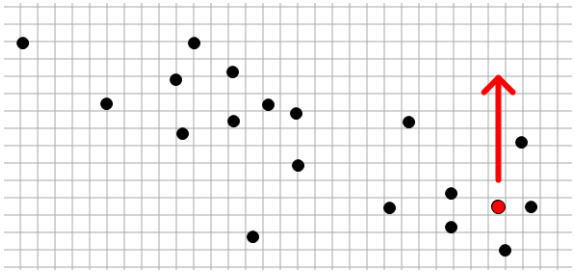
Hint: Start by assuming the AM-GM inequality holds true, and define $z_i = \frac{1}{y_i}$. *Then, try and re-write the left side of the inequality to look like* $\frac{n}{\sum_{i=1}^{n} \frac{1}{y_i}}$.

3 Visualizing Changes in the Data

The problems in this section will help you visualize how changes in the data affect the regression line. Assume all data is in the first quadrant (positive x and y coordinates).

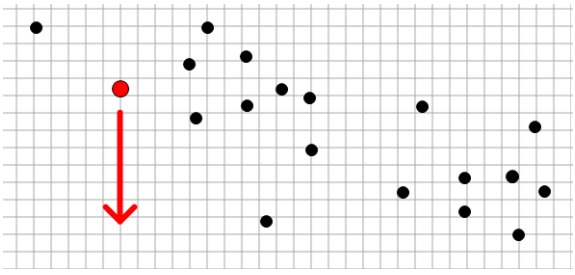
Problem 6.

For the data set shown below, how will the slope and intercept of the regression line change if we move the red point in the direction of the arrow?



Problem 7.

For the data set shown below, how will the slope and intercept of the regression line change if we move the red point in the direction of the arrow?



Problem 8.

Suppose we transform a data set of $\{(x_i, y_i)\}$ pairs by doubling each y-value, creating a transformed data set $\{(x_i, 2y_i)\}$. How does the slope of the regression line fit to the transformed data compare to the slope of the regression line fit to the original data? Can you prove your answer from the formula for the slope of the regression line?

Problem 9.

Suppose we transform a data set of $\{(x_i, y_i)\}$ pairs by doubling each x-value, creating a transformed data set $\{(2x_i, y_i)\}$. How does the slope of the regression line fit to the transformed data compare to the slope of the regression line fit to the original data? Can you prove your answer from the formula for the slope of the regression line?

Problem 10.

Compare two different possible changes to the data set shown below.

- Move the red point down c units.
- Move the blue point down c units.

Which move will change the slope of the regression line more? Why?