Assignment Ten
ECE 4200

- Provide credit to **any sources** other than the course staff that helped you solve the problems. This includes **all students** you talked to regarding the problems.
- You can look up definitions/basics online (e.g., wikipedia, stack-exchange, etc)
- **The due date is 11/29/2020, 23.59.59 eastern time**.
- Submission rules are the same as previous assignments.

Problem 1. (10 points). Suppose W is a $k \times d$ matrix, where each entry of W is picked independently from the set $\{-\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}}\}$. In other words, for each i, j,

$$\Pr\left(W_{ij} = -\frac{1}{\sqrt{k}}\right) = \Pr\left(W_{ij} = \frac{1}{\sqrt{k}}\right) = \frac{1}{2}.$$

1. Let $\overrightarrow{x} \in \mathbb{R}^d$. If we pick W with this distribution, show that

$$\mathbb{E}\left[\|W\overrightarrow{x}\|_2^2\right] = \|\overrightarrow{x}\|_2^2.$$

2. Just like the Gaussian matrix we considered in the class, we might as well take a random matrix W designed like this for JL transform. What is an advantage of this matrix over the Gaussian matrix?

Problem 2. (10 points). Suppose $d = 1$. Come up with a set of n real numbers, and an initial set of k distinct cluster centers such that the k-means algorithm **does not converge** to the best solution of the k-means clustering problem. You can choose any value of n, and k that you want! (Hint: small n, k are easier to think about.)

Problem 3. (15 points). Let $C = \left\{\bar{x}_1 \cdots \bar{x}_{|C|}\right\}$ be a cluster where $\bar{x}_i \in \mathbb{R}^d$. Let

$$c_{av} = \frac{1}{|C|}\sum_{\bar{x}_i \in C} \bar{x}_i$$

Prove that for any $c \in \mathbb{R}^d$,

$$\sum_{\bar{x}_i \in C} \|\bar{x}_i - c\|_2^2 \geq \sum_{\bar{x}_i \in C} \|\bar{x}_i - c_{av}\|_2^2$$

(Hint: $\bar{x}_i - c = \bar{x}_i - c_{av} + c_{av} - c$)

Problem 4. (30 points). Please see attached jupyter notebook.