Name: ______________________

Entry no.: ______________________

There are 1 questions for a total of 50 points.

1. The Euclidean k-means problem is defined as follows: Given a set $X \subseteq \mathbb{R}^d$ of n points in d-dimensional Euclidean space, find a set $C \subseteq \mathbb{R}^d$ of k points (called *centers*) such that the following cost function is minimized:

$$\Phi(C, X) = \sum_{x \in X} \min_{c \in C} ||x - c||^2.$$

Assume that there is an approximation algorithm A for the k-means problem that is guaranteed to output a $(1 + \varepsilon)$-approximate solution for any given error parameter $\varepsilon > 0$. That is $A(X, k, \varepsilon)$ outputs a set of centers C' such that $\Phi(C', X) \leq (1 + \varepsilon) \cdot \Phi(C_{OPT}, X)$, where C_{OPT} denotes the optimal k centers. Moreover, the running time of A is $f(n, k, d, \varepsilon)$ for some polynomial function f. Solve the following questions:

(a) (10 points) Show that for any dataset $X \subseteq \mathbb{R}^d$ and any point $p \in \mathbb{R}^d$,

$$\sum_{x \in X} ||x - p||^2 = \sum_{x \in X} ||x - \mu(X)||^2 + |X| \cdot ||\mu(X) - p||^2.$$

Here $\mu(X)$ denotes the mean of points in X. Note that this also shows that the optimal solution for the 1-means problem is the mean of the given points.

(b) (10 points) Show that for any dataset $X \subseteq \mathbb{R}^d$,

$$\sum_{x \in X} ||x - \mu(X)||^2 = \frac{1}{2|X|} \sum_{x \in X} \sum_{y \in X} ||x - y||^2.$$

Here $\mu(X)$ denotes the mean of points in X.

(c) (30 points) Use algorithm A to design another (randomized) algorithm B that runs in time

$$f\left(n, k, O\left(\frac{\log n}{\varepsilon^2}\right), O(\varepsilon)\right) + O\left(nd \cdot \frac{\log n}{\varepsilon^2}\right)$$

and outputs a $(1 + \varepsilon)$ approximate solution for any given $\varepsilon > 0$. Argue correctness and running time of B. (*Hint: use Johnson-Lindenstrauss*)

(*What the above exercise shows is that if f has a very bad dependence on d then there is a way to deal with it using Johnson-Lindenstrauss (JL). This is a useful application of JL-theorem in the context of k-means clustering.*)