Problem 1. Suppose \( n \) data points \( S \subset \mathbb{R}^d \) are partitioned into \( k \) subsets \( S_1, S_2, \ldots, S_k \); let \( \mu_j := \text{mean}(S_j) \) for each \( j \in [k] \). Let \( A \in \mathbb{R}^{n \times d} \) be the data matrix (whose rows are the points in \( S \)), and let \( M \in \mathbb{R}^{n \times d} \) be the matrix whose \( i \)-th row is \( \mu_j^\top \) when the \( i \)-th data point belongs to \( S_j \). Let \( \Pi \) denote the orthogonal projector to the top \( k \)-dimensional (uncentered) PCA subspace \( W \) for \( S \); let \( \tilde{S} \subset W \) be the \( \Pi \)-projected data points (i.e., the rows of \( A\Pi \)).

Assume, for some \( c_0 \geq 1 \), that \( C \subset W \) is a set of representatives with \( k \)-means cost on \( \tilde{S} \) at most \( c_0 \) times optimal (among representatives in \( W \)):

\[
\text{cost}(\tilde{S}, \tilde{C}) \leq c_0 \cdot \min_{C \subset W: |C|=k} \text{cost}(\tilde{S}, C).
\]

Instructions: Do at least two parts among (a), (b), and (c); and also do part (d).

(a) Modify the proof from the \( k \)-means notes to deduce that for each \( i \in [k] \),

\[
\|\mu_i - \tilde{C}(\mu_i)\|^2_2 \leq \frac{1}{|S_i|} \left( 1 + \sqrt{\frac{c_0 \cdot \|A - M\Pi\|_F^2}{\|A\Pi - M\|_F^2}} \right)^2 \|A\Pi - M\|_F^2.
\]

You can just explain, precisely, how the argument from the notes should be modified. Explain why \( \|A - M\Pi\|_F/\|A\Pi - M\|_F \) is at most one (assuming the denominator is non-zero).

(b) Prove that there is an absolute constant \( c \geq 1 \) such that

\[
\|A\Pi - M\|_F^2 \leq c \cdot \min \left\{ \|A - M\|_F^2, k\|A - M\|_F^2 \right\}.
\]

(c) Suppose each \( x \in S_i \) is actually a random vector with distribution \( N(\mu_i, \sigma^2 I) \). We'll use this assumption to prove a bound on \( \|A - M\|_F^2 \). (Here, \( M \) remains the matrix whose rows are the \( \mu_i^\top \).) Prove that with probability 0.99, \( \|A - M\|_F^2 \leq O(\sigma^2(n + d)) \). Since this is fairly routine, you can just give a rough sketch of the argument.

\( \text{Hint:} \) There are many ways to do this, but one simple way is to use \( \chi^2 \) tail bounds together with a covering argument.

(d) Continuing from (c), assume further that \( |S_i| = n/k \) for each \( i \in [k] \). Deduce an upper-bound on \( \max_{i \in [k]} \|\mu_i - \tilde{C}(\mu_i)\|^2_2 \) that holds with probability 0.99. It should be stated in terms of \( c_0, d, k, n, \text{ and } \sigma \). Here, you can use the results from parts (a), (b), and (c), as well as big-\( O \) notation.

Problem 2. Let \( T \) be an undirected tree on \( n \) vertices \( V \). Consider the metric space \((V, \rho)\), where \( \rho \) is the shortest path metric on \( T \) (i.e., \( \rho(x, y) = \text{length of shortest path from } x \text{ to } y \)). Note that because \( T \) is a tree, the shortest path from \( x \) to \( y \) is actually the only path!

Instructions: Do at least two parts among (a), (b), and (c).

(a) Show how to construct an embedding \( f: (V, \rho) \to \ell_1^n \) with no distortion.

\( \text{Hint:} \) Remove a single leaf \( v \in T \), recursively construct the embedding for \( T \setminus \{v\} \) (into \( \ell_1^{n-1} \)), and then modify the embedding to also work with \( v \).
(b) Show how to find subtrees $T'$ and $T''$ of $T$ such that: (i) $T'$ and $T''$ share a single vertex $v_0 \in T$ and no edges, (ii) $T = T' \cup T''$, and (iii) $\max\{|T'|, |T''|\} \leq 1 + 3n/4$. (It is possible to do this with $1 + 2n/3$ instead of $1 + 3n/4$.)

(Here, $|\cdot|$ denotes number of vertices.)

The following lemma is helpful:

**Lemma 1.** For any tree $T$ on $n$ vertices, there is a special vertex $v_0 \in T$ such that $T \setminus \{v_0\}$ is a forest of trees, each of which has at most $n/2$ vertices.

You can use the lemma without proof, and also assume a blackbox algorithm for finding this special vertex $v_0 \in T$.

(c) Show how to construct an embedding $f: (V, \rho) \to \ell^d_{\infty}$ with no distortion and $d = O(\log n)$.

**Hint:** Use divide and conquer, as suggested in (b). Here is a rough sketch of a possible approach.

1. If $n$ is smaller than some absolute constant, just let $f$ be the Fréchet embedding.
2. Otherwise, obtain subtrees $T'$ and $T''$ as in (b), and recursively construct embeddings $f': T' \to \ell^k_{\infty}$ and $f'': T'' \to \ell^k_{\infty}$.
3. Modify $f'$ and $f''$ so that the special vertex $v_0 \in T' \cap T''$ gets mapped to the same point (so $f'(v_0) = f''(v_0)$). (Can this be done without incurring any distortion?)
4. Construct a new embedding $f: T \to \ell^{k+1}_{\infty}$:

   $$f(x) := \begin{cases} 
   (f'(x), \rho(x, v_0)) & \text{if } x \in T', \\
   (f''(x), -\rho(x, v_0)) & \text{if } x \in T''.
   \end{cases}$$

If you follow this sketch, fill in the details, and prove that it is correct.