7.6 Semantics of \( k \)-means clustering

Suppose the input data set \( S \subseteq \mathbb{R}^d \) is partitioned into \( k \) subsets \( S_1, S_2, \ldots, S_k \); associate each \( S_i \) with a representative \( \mu_i \in \mathbb{R}^d \). When is a near-optimal solution to \( k \)-means clustering on \( S \) close to the representatives \( M := \{ \mu_1, \mu_2, \ldots, \mu_k \} \)?

**Theorem 7.6.** Let \( S, S_1, S_2, \ldots, S_k \) and \( M = \{ \mu_1, \mu_2, \ldots, \mu_k \} \) be as described above. Furthermore, let \( \hat{C} \subseteq \mathbb{R}^d \) be a set of representatives that satisfies, for some \( b \geq 1 \),

\[
\text{cost}(S, \hat{C}) \leq b \cdot \min_{C \subseteq \mathbb{R}^d | |C| = k} \text{cost}(S, C).
\]

Then for each \( i \in [k] \),

\[
\| \mu_i - \hat{C}(\mu_i) \|_2^2 \leq \frac{(1 + \sqrt{b})^2}{|S_i|} \sum_{j=1}^{k} \text{cost}(S_j, \mu_j).
\]

**Proof.** Observe that for any vectors \( a, b \in \mathbb{R}^d \) and any \( \varepsilon \in (0, 1) \), Cauchy-Schwarz and the AM/GM inequality implies

\[
\| a + b \|_2^2 \geq \| a \|_2^2 - 2\| a \|_2 \| b \|_2 + \| b \|_2^2 \geq -(1 - \varepsilon) \| a \|_2^2 + (1 - \varepsilon) \| b \|_2^2.
\]

Thus, for any \( i \in [k] \),

\[
\begin{align*}
\text{cost}(S_i, \hat{C}) &= \sum_{x \in S_i} \| x - \hat{C}(x) \|_2^2 \\
&= \sum_{x \in S_i} \| x - \mu_i + \mu_i - \hat{C}(x) \|_2^2 \\
&\geq \sum_{x \in S_i} \left( (1 - \varepsilon) \| \mu_i - \hat{C}(x) \|_2^2 - \left( \frac{1}{\varepsilon} - 1 \right) \| x - \mu_i \|_2^2 \right) \\
&\geq (1 - \varepsilon) |S_i| \| \mu_i - \hat{C}(\mu_i) \|_2^2 - \left( \frac{1}{\varepsilon} - 1 \right) \sum_{x \in S_i} \| x - \mu_i \|_2^2 \\
&= (1 - \varepsilon) |S_i| \| \mu_i - \hat{C}(\mu_i) \|_2^2 - \left( \frac{1}{\varepsilon} - 1 \right) \text{cost}(S_i, \mu_i).
\end{align*}
\]

This implies

\[
\| \mu_i - \hat{C}(\mu_i) \|_2^2 \leq \frac{1}{(1 - \varepsilon)|S_i|} \left( \text{cost}(S_i, \hat{C}) + \left( \frac{1}{\varepsilon} - 1 \right) \text{cost}(S_i, \mu_i) \right).
\]

We trivially have \( \text{cost}(S_i, \hat{C}) \leq \text{cost}(S, \hat{C}) \). Since

\[
\text{cost}(S, \hat{C}) \leq b \cdot \text{cost}(S, M) \leq b \cdot \sum_{j=1}^{k} \text{cost}(S_j, \mu_j),
\]

we have

\[
\| \mu_i - \hat{C}(\mu_i) \|_2^2 \leq \frac{1}{(1 - \varepsilon)|S_i|} \left( \frac{1}{\varepsilon} - 1 + b \right) \sum_{j=1}^{k} \text{cost}(S_j, \mu_j).
\]

Plugging-in \( \varepsilon := (\sqrt{b} - 1)/(b - 1) \) (or \( \varepsilon := 1/2 \) when \( b = 1 \)) proves the claim. \( \square \)
To interpret Theorem 7.6, suppose for simplicity that the $\mu_1, \mu_2, \ldots, \mu_k$ partition $S$ into equal size clusters $S_1, S_2, \ldots, S_k$, with $|S_j| = |S|/k$. Let $\text{rmse}^2 := \sum_{j=1}^{k} \frac{\text{cost}(S_j, \mu_j)}{|S|}$ be the average squared distance from a point $x \in S$ to its associated representative among the $\mu_j$. Theorem 7.6 implies that for each $\mu_j$, there is a representative $\hat{C}(\mu_j)$ from $\hat{C}$ within distance $(1 + \sqrt{b})\sqrt{k}\text{rmse}$. 