Solution 1: K-Means Clustering (Final Exam 2018, Problem 6)

Given the dataset $\mathcal{D} = \{0, 0.5, 0.5+\Delta, 1.5+\Delta\}$, where $\Delta \geq 0$ and considering $K = 2$ clusters, $\mathcal{B}_1$ and $\mathcal{B}_2$, are the two clusters with means μ_1 and μ_2 respectively, we solve the problem subparts using Lloyd's algorithm.

a) Letting $\Delta = 0.5$ and initializing the means to $\mu_1[0] = 1, \mu_2[0] = 2$, we find the following clustering:

- $\mathcal{B}_1[0] = \{0, 0.5, 1\}$
- $\mathcal{B}_2[0] = \{2\}$.

Continuing the Lloyd's algorithm, we find the new means:

- $\mu_1[1] = \frac{0+0.5+1}{3} = 0.5$
- $\mu_2[1] = 2$.

Following the changed means, we again find the new clustering:

- $\mathcal{B}_1[1] = \{0, 0.5, 1\}$
- $\mathcal{B}_2[1] = \{2\}$.

Since the cluster memberships have converged, we stop.

b) To find the condition on Δ to shift the convergence, we first consider the case where $\Delta < 1$. The clustering at the first iteration will be identical to before. However, the new means would be as follows:

- $\mu_1[1] = \frac{0+0.5+0.5+\Delta}{3} = \frac{1}{3} + \frac{\Delta}{3}$
- $\mu_2[1] = \frac{3}{2} + \Delta$.

Here note that the distance from the closest point in $\mathcal{B}_1$ to the only point in $\mathcal{B}_2$ is equal to $||\frac{3}{2}+\Delta-\frac{1}{2}-\Delta|| = 1$, which always is greater than the distance from $\mu_1[1]$ to the point $\frac{1}{2}+\Delta$, i.e., $||\frac{1}{2}+\Delta-\frac{1}{3}-\frac{\Delta}{3}|| = \frac{1}{6}+\frac{2\Delta}{3} < 1$ for $0 \leq \Delta < 1$. Hence, this case can never yield a change in final clustering.

Therefore, we now consider the case where $\Delta = 1 + \epsilon$, where $\epsilon > 1$ is a small constant. Then, the initial clustering will be:

- $\mathcal{B}_1[0] = \{0, 0.5\}$
- $\mathcal{B}_1[0] = \{0.5 + \Delta, 1.5 + \Delta\}$.

The corresponding means would then become:

- $\mu_1[1] = \frac{0+0.5}{2} = 0.25$
- $\mu_2[1] = \frac{0.5+\Delta+1.5+\Delta}{2} = 1 + \Delta$.

Then notice that the distance from $\mu_1[1]$ to the closest point in $\mathcal{B}_1$ to $\mathcal{B}_2$ is equal to $||\frac{1}{2}+\Delta-\frac{1}{4}|| = \frac{1}{4}+\Delta > \frac{5}{4}$, whereas the distance from $\mu_2[1]$ to the same point is $||1 + \Delta - \frac{1}{2} - \Delta|| = \frac{1}{2}$, which is less than the former in this case and hence there would be no further updates the the clusters. Therefore, for this case, i.e., $\Delta > 1$, we are able to find a change in clustering.