On Lloyd’s algorithm: new theoretical insights for clustering in practice

7 Appendix A: a comparison to previous assumptions

In this section, we compare our clusterability assumption with those made previously in two lines of work: the work of [21, 5], and the work of [4]. While the assumptions in [21, 5] were shown to be weaker than many existing probabilistic assumptions, the assumption in [4] was shown to be weaker than some deterministic clusterability assumptions in the literature.

We show that for $d < k$, the assumptions in [21, 5] are stronger than ours, and the assumption in [4] is stronger than ours in general.

Proximity condition in [21] implies center separability in [5]

**Proposition 2.** Theorem 2.2 of [21] only holds for $\epsilon < \frac{n_{\min}}{n}$, and under this constraint, any dataset-solution pair $(X, T_s)$ that satisfies $(d^K_r, \epsilon)$-proximity condition must satisfy $d^A_r$-center separability for the same $d^K_r$.

To prove the proposition, we first show that in the proximity condition of [21], $\epsilon$ must be upper bounded by $\frac{n_{\min}}{n}$, i.e., the number of bad points cannot exceed $n_{\min}$, the size of the smallest cluster. Then we show under this condition, the $(d_r, \epsilon)$-proximity condition implies $d_r$-center separation for the same $d_r$.

The need of an upper bound on $\epsilon$ is not discussed in neither of the work [21]. Here we show for Lloyd’s algorithm to converge to a non-degenerate solution, i.e., finding $k$ non-empty clusters, which is a necessary condition for Theorem 2.2 in [21] to hold, we need $\epsilon < \frac{n_{\min}}{n}$ regardless of how large $d_r$ becomes.

**Lemma 9.** For any fixed $d_r := d + d_s > 0$ and $0 < \delta < \frac{d_r}{6}$, let $(X, T_s)$ satisfy $(d_r, \epsilon)$-clusterability with $\epsilon n \geq n_{\min}$, then there exists $\nu \subseteq \{k\}$ with $\max_{x \in \nu} \|\mu_s - \nu_x\| = \delta$, such that if we apply Lloyd’s algorithm on $(X, \{\nu\})$ until convergence, it returns a degenerate solution.

Proof. Let $(X, T_s)$ be a dataset satisfying $(d_r, \epsilon)$-clusterability with $\epsilon n \geq n_{\min}$. Assume it contains three clusters, $T_1, T_2, T_3$ s.t. $n_2 = n_{\min} = 2$ and both points in $T_2$ are within the $\epsilon n$ bad points, i.e., they don’t satisfy the condition in Definition 1. We can assume the relation between $T_2, T_3$ are symmetric to that of $T_1, T_2$ (which ensures $\mu_3$ is the mean of $T_2$).

We only focus on $T_1, T_2$ (Figure 3) since the case for $T_2, T_3$ is similar. Let both seeds $\nu_1, \nu_2$ fall on the line joining $\mu_1, \mu_2$, and $0 = \|\mu_2 - \nu_2\| \leq \|\mu_1 - \nu_1\| = \delta$ and let $\|\mu_1 - \mu_2\| = d > 6\delta$. Furthermore, $\forall x \in T_2, \|\hat{x} - \mu_2\| = \frac{d}{2} - \frac{\delta}{2}$. So $\|\hat{x} - \nu_2\| \geq \frac{d}{2} - \frac{\delta}{2}$ but $\|\hat{x} - \nu_1\| \leq \frac{d}{2} + \frac{\delta}{3} - \delta < \|\hat{x} - \nu_2\|$. Thus, $x$ in $T_2$ is assigned to $S_1$. Now applying the centroid update, the mean of $S_1$ is $\frac{-\frac{d}{2} + \frac{\delta}{2}}{\frac{3}{2}}$, whose distance to $\hat{x}$ is $\frac{d}{3} - \frac{-\frac{d}{2} + \frac{\delta}{2}}{\frac{3}{2}} = \frac{d^2}{6} + \frac{\delta d}{3}$. This is smaller than $\|\hat{x} - \nu_2\|$, since $d > 6\delta$. Then the clustering assignment does not change, and the same holds for $S_3$, so the algorithm stops and the cluster corresponding to $\mu_2$ vanishes.

Lemma 9 shows if $\epsilon n \geq n_{\min}$, then in general no matter how good the seeding guarantee is, Lloyd’s algorithm may produce empty clusters. Next, we show $(d^K_r, \epsilon)$-proximity condition with $\epsilon < n_{\min}$ implies $d_A^r$-center separability.

**Lemma 10.** If $(X, T_s)$ satisfies $(d_r, \epsilon)$-proximity condition with $\epsilon n < n_{\min}$, then it satisfies $d_r$-center separability.
Proof. Since \( cn < n_{\min} \), for any cluster \( T_r, r \in [k] \), \( \exists x \in T_r \) s.t. \( x \) satisfies the proximity condition, i.e., 
\[
|x - \mu_r| \leq |x - \mu_s| - d_{rs}, \quad \text{for any } s \neq r, \text{ where } d_{rs} := d_r + d_s.
\]
Since \( |\mu_r - \mu_s| \geq |x - \mu_s| - |x - \mu_r| \), we know all pairs of centroids are separated by at least \( d_{rs} \).

Let \( d_{rs}^*(f) := f \sqrt{\delta_{rs}(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}})} \), with \( f = \Omega(1) \). By Lemma 9 and 10, \((d_{rs}^*(f), \epsilon)\)-clusterability implies our clusterability assumption. Using this relation, we can indirectly compare our clusterability with KK and AS clusterability.

Unlike KK or AS clusterability, which depends on \( \|X - C_r\|_F \), the maximal mean-departure of the entire dataset along one direction, \((d_{rs}^*(f), \epsilon)\)-clusterability depends on \( \sqrt{\phi_{rs}} \). Fix \( T_r \) and the corresponding \( C_r \), since \( |X - C_r| \leq \left\| X - C_r \right\|_F = \sqrt{\phi_{rs}} \leq \sqrt{\text{rank}(X - C_r)} |X - C_r| \leq \sqrt{2k} |X - C_r| \), when \( d \leq k \), our assumption is (a factor of \( \Omega(\sqrt{k}) \)) weaker than KK and may be stronger than AS clusterability.

**Weak-deletion stability** Weak-deletion stability \[4\] captures the intuition that if a dataset has a good clustering solution with respect to the current \( k \), then merging any two clusters in this solution should incur a large \( k \)-means cost.

**Definition 4** (Weak-deletion stability \[4\]). Let \( \{\mu_i, i \in [k]\} \) denote the centers in the optimal \( k \)-means solution, with \( k \)-means cost \( \text{OPT} \). Let \( \text{OPT}^{(i \rightarrow j)} \) denote the cost of the clustering obtained by removing \( \mu_i \) and assigning all its points to \( \mu_j \). Fix \( \delta > 0 \), the dataset satisfies \((1 + \delta)\)-weak-deletion stability if \( \text{OPT}^{(i \rightarrow j)} > (1 + \delta) \text{OPT} \), \( \forall i \neq j \).

In \[4\], weak-deletion stability is shown to be implied by both the clusterability assumption in \[28\] and a special case of the assumption in \[6\]. Here we show it in turn implies the optimal \( k \)-means solution satisfies center separability.

**Proposition 3** If a dataset is \((1 + \delta)\)-weak-deletion stable, then let \( T_r \) be the optimal \( k \)-means solution, we have for all \( r \neq s \in [k] \), \( \|\mu_r - \mu_s\|^2 \geq \frac{\delta_{rs}}{\max\{n_r, n_s\}} \). Furthermore, if \( \delta > f(1 + \frac{1}{\alpha}) \), then \((X, T_r)\) satisfies \( d_{rs}^*(f)\) center separability.

Proof. Let \( r \neq s \) be any pair of indices from \([k]\). Let \( T_{rs} := T_r \cup T_s \), \( \mu_{rs} := m(T_{rs}) \), and let \( \Delta \) denote the increase in \( k \)-means cost by merging \( T_r, T_s \). Then \( \Delta := \phi(T_s) - \phi(T_r) = \phi(T_{rs}) - \phi(T_r) = \phi(\mu_r, T_r) - \phi(\mu_s, T_s) = n_s \|\mu_r - \mu_s\|^2 + n_r \|\mu_r - \mu_s\|^2 \leq (\|\mu_r - \mu_s\|^2 + \|\mu_r - \mu_s\|^2) (n_s + n_r) \). The first equality uses the decomposability of \( k \)-means cost over disjoint sets \( T_r \) and \( T_s \), and the second is by Lemma 4.

Now note that \( \mu_{rs} = \frac{n_r \mu_r + n_s \mu_s}{n_r + n_s} \), so \( \mu_{rs} \) is on the segment joining (a convex combination of) \( \mu_r \) and \( \mu_s \), i.e., \( \|\mu_r - \mu_s\| \leq |\mu_r - \mu_s| \leq |\mu_r - \mu_s| \). Hence, \( \Delta \geq 2 \|\mu_r - \mu_s\|^2 (n_s + n_r) \). Since \((X, A_r)\) is \((1 + \delta)\)-weak-deletion stable, we have \( \Delta \geq \delta_{rs} \). Therefore, \( 2 \|\mu_r - \mu_s\|^2 \max\{n_s, n_r\} \geq \delta_{rs} \) and the first statement follows. The second statement follows by substituting the definition of \( \alpha \) into the bound.

---

8 Appendix B: proofs

**Proof of Lemma 1.** Suppose \( \exists A_r \) s.t. \( \forall s \in [k], \|\nu_s - m(A_r)\| > \sqrt{2g} + 2 \sqrt{\phi_{rs}} \), then \( \text{OPT}^{(i \rightarrow j)} > (1 + \delta) \text{OPT} \), \( \forall i \neq j \).

Proof of Lemma 2. \( \|x - \mu_r\| \geq \|x - \nu_s\| - \|\mu_r - \nu_s\| \geq \frac{1}{2} \|x - \nu_s\| - \|\mu_r - \nu_s\| \), since \( x \) is assigned to \( S_r \) by the Voronoi partition induced by \( \nu_s \). Then \( \|x - \mu_r\| \geq \|x - \mu_r + \mu_r - \mu_s + \mu_s - \nu_s\| \geq \|\mu_r - \mu_s\| - \|\nu_s - \mu_s\| - \|\nu_r - \mu_r\| \geq (1 + \gamma_1) \|\mu_r - \mu_s\| \). This implies \( \|x - \mu_r\| \geq (\frac{1}{2} - \gamma_1) \|\mu_r - \mu_s\| - \|\nu_s - \mu_s\| \geq \frac{1}{2} \|\mu_r - \mu_s\| - \|\nu_s - \mu_s\| - \|\nu_r - \mu_r\| \geq (2g + 2) \|\mu_r - \mu_s\|^{-2} \phi_{rs} \). Hence \( \text{OPT}^{(i \rightarrow j)} > (1 + \delta) \text{OPT} \), \( \forall i \neq j \).

Proof of Theorem 1. Fix any \( r, s \neq r, \Delta^1 < \beta^1 \sqrt{\phi_{rs}} \left( \frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}} \right) \leq \beta^1 \sqrt{\phi_{rs}} \left( \frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}} \right) \leq \beta^1 \|\mu_r - \mu_s\| \), hence \( \gamma_1 \leq \frac{\beta^1}{\Delta^1} \), and we can apply Lemma 3 and get \( \rho_{out}^1 \leq \left( \frac{4}{(1-4\gamma^1)r^2} \right) \). Similarly, \( \rho_{out}^1 \leq \left( \frac{4}{(1-4\gamma^1)r^2} \right) \) and \( \rho_{out}^1 < \left( \frac{4}{(1-4\gamma^1)r^2} \right) \). Consider any \( T_s \subset V(\mu_s), s \neq r \). Since \( m(T_s) \subset V(\nu_s) \), by Lemma 2 \( \|m(T_s) - s\| \geq (1 - 4\gamma_1) m(T_s) - \mu_s \). It is easy to check \( \rho_{out}^1 < \left( \frac{8}{(1-4\gamma^1)r^2} \right) \). Applying Lemma 5 with \( R = 1 - 4\gamma_1 \) yields \( \Delta^1 + 1 < \frac{8}{(1-4\gamma^1)r^2} \).
Substituting $w_{\min}$ for every $r$ and applying union bound, we get $Pr(A') \leq m \exp(-2(\frac{4}{3} - 1)^2 w_{\min}^2)$. Now the probability of a cluster $T_r$, not being seeded after $m$ trials is $(1 - p_t)^m \leq \exp(-mp_t)$. Applying union bound, we get $Pr(B^c) \leq k \exp(-mp_{\min})$. Applying union bound again, we get $Pr(A \cap B) \geq 1 - m \exp(-2(\frac{4}{3} - 1)^2 w_{\min}^2) - k \exp(-mp_{\min})$.

Proof of Lemma 8. Let $\pi(i) = \pi(j) = r$. Then $|\nu_i - \nu_j| \leq |\nu_i - \mu_j| + |\nu_j - \mu_j| \leq 2\sqrt{\frac{2}{\alpha_n}}$. Let $\pi(p) = t, \pi(q) = s$. Then $|\nu_p - \nu_q| \geq |\mu_s - \mu_t| - |\nu_p - \mu_t| - |\nu_q - \mu_s| \geq f(\sqrt{\phi_{s}(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}})} - \sqrt{\frac{2}{\alpha_n}}) - \sqrt{\frac{2}{\alpha_n}}$. On the other hand, recall $\alpha := \min_{r \in \mathbb{N}} \frac{\alpha_n}{m_{r}}$, we get $\sqrt{\frac{2}{\alpha_n}} \leq \min(1, \frac{1}{\sqrt{\alpha_n}})$, so $2\sqrt{\frac{2}{\alpha_n}} \leq \sqrt{\phi_{s}(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}})}$. Substituting this bound on $|\nu_i - \nu_j|$ and comparing it with the lower bound on $|\nu_p - \nu_q|$ completes the proof.

Proof of Theorem 2. By Lemma 1, after seeding $\Delta^k < \frac{f}{\sqrt{\alpha_n}}, \forall \alpha \in [k]$. Applying Theorem 1, running Lloyd’s algorithm until convergence, we will obtain a solution s.t. $\Delta^k \leq \frac{128}{\sqrt{\alpha_n}}$. Then applying Lemma 3 with $\gamma \leq \frac{128}{\sqrt{\alpha_n}}$, we get $\rho_{in}^* + \rho_{out}^* \leq \frac{128}{\sqrt{\alpha_n}} \frac{f^2}{\gamma^2} \leq \frac{81}{\sqrt{\alpha_n}} \forall \alpha \in [k]$. Hence, $d(T_s, T_r) := \sum_{r \in [k]} |S_r \Delta T_r| = \sum_{r}(\rho_{in}^* + \rho_{out}^*) n_r \leq \frac{81}{\sqrt{\alpha_n}} \sum_{r} n_r = \frac{81}{\sqrt{\alpha_n}} n$.

Proof of Lemma 6. Consider the graph $G_{\max}$ obtained by adding all edges in $E_{in}$ to $G_0$. Clearly, $G_{\max}$ has $k$ connected components, where each component corresponds to a vertex cluster $V^*_r$ for some $r \in k$. Adding any more edges from $E_{out}$ to $G_{\max}$ will reduce the number of components to $k - 1$. Furthermore, any $e \in E_{out}$ can only be added to $G_{\max}$ after all edges in $E_{in}$ are added. This means the algorithm must stop before any edges in $E_{out}$ are added. This in turn implies the final solution $G_{\max}$, if not equal to $G_{\max}$, can be obtained by removing edges in $G_{max}$. Since removing edges can only maintain or disconnect existing connected components and $G_{\max}$ has the same number of connected components as that of $G_{\max}$, $G_{\max}$ must have exactly the same $k$ connected components as those of $G_{\max}$, thus each component $V_{\max}^r$ of $G_{\max}$ corresponds to exactly one cluster $V^*_r$ for some $r$.

Proof. To prove Lemma 7, we first show without any assumption, if we sample $X$ i.i.d. uniformly at random, then for each target cluster $T_r$, if $\nu \in T_r$, then $|\nu - \mu_r|$ satisfies the bound in $A$ with high probability. Let $q := |\nu - \mu_r|^2$, we have $0 \leq q \leq \max_{x \in T_r} ||x - \mu_r||^2$ and $E[q | \nu \in T_r] = \sum_{x \in T_r} \mathbb{E}[|x - \mu_r|^2] = \frac{\phi_n^2}{n_r}$. Then applying Hoeffding’s bound, we get conditioning on the event $\{\nu \in T_r\}$,

$$Pr(q - Eq \geq \frac{f}{4} - 1 \frac{\phi_n^2}{n_r}) \leq \exp(-\frac{2[(\frac{f}{4} - 1)^2 \frac{\phi_n^2}{n_r}]^2}{\max_{x \in T_r} ||x - \mu_r||^2})$$
\( C_*(x)^2 \) = 4\( \phi_*(C_* \) is the set of optimal centroids). □

**proof of Corollary 1.** We first find a sufficient condition for Algorithm 1 to have a \( 1 + \epsilon \)-approximation. Note, as in the proof of Theorem 1, the approximation guarantee is upper bounded by \((\frac{1}{1-4\epsilon^2})^2\), where \( \gamma \leq \frac{\sqrt{2}}{2} \). So to have a \( 1 + \epsilon \)-guarantee, it suffices to have \((\frac{1}{1-4\epsilon^2})^2 \leq 1 + \epsilon\), which holds if \( f = \Omega(\frac{1}{\epsilon}) \). Now we find a sufficient condition for the success probability to be at least \( 1 - \delta \). It suffices to require that 

\[
\begin{align*}
\frac{1}{2m_{\min}} \log \left( \frac{2e}{ \epsilon} \right) - \frac{1}{2m_{\min}} \log \left( \frac{2e}{ \epsilon} \right)
\end{align*}
\]

imposing an additional constraint on \( f \). Taking log on both sides and rearrange, we get \((\frac{2}{q} - 1)^2 \geq \frac{1}{2m_{\min}} \log \left( \frac{2e}{ \epsilon} \right)\). Thus, it is sufficient for a \( 1 + \epsilon \)-approximation to hold with probability at least \( 1 - \delta \) if \( f = \Omega\left( \sqrt{\log \left( \frac{2e}{ \epsilon} \right)} \right), \) and we choose \( m \) to be in the interval \([\frac{1}{2m_{\min}} \log \left( \frac{2e}{ \epsilon} \right), \frac{1}{2m_{\min}} \log \left( \frac{2e}{ \epsilon} \right) + \frac{1}{2} \exp(2(\frac{2}{q} - 1)^2w_{\min}^2)].\) □

9 Appendix C: details on the generation of synthetic data

The clusterability of each dataset is controlled by three parameters \((\epsilon, \alpha, \nu)\), where \( \epsilon \in [0, 1] \) controls the fraction of outliers, i.e., those far away from any center, \( \alpha \in [0, 1] \) controls the degree of centroid separation (the centroids become more separated as \( \alpha \) increases), \( \nu \in [0, \infty) \) controls the degree of balance of the cluster sizes in the ground-truth clustering (\( \nu \) is the symmetric Dirichlet prior for the multinomial distribution; the higher \( \nu \) is, the more balanced the cluster sizes will likely be). Note the parametrization of clusterability here does not correspond exactly to our clusterability assumption, but incorporates more parameters. In particular, fixing dimension \( d \), number of clusters \( k \), \( k = 2d \), and total number of points \( n \), we first fix the \( 2d \) vertices of a \( d \)-dimensional cross-polytope as our ground-truth centroids. Then we generate the numbers of points for each cluster, \( n_1, \ldots, n_k \), such that \( \sum_{i=1}^{k} n_i = n \), where \( n_i \) is sampled from a multinomial distribution with parameter \( \theta \), with \( u \) characterizing the sparsity of \( \theta \). Then for each centroid, we generate a set of \( k - 1 \) linear constraints based on our center separability condition (i.e., \( k - 1 \) hyperplanes cutting through the lines that join this centroid to the \( k - 1 \) other centroids) and parameter \( \alpha \).

References


