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

Cheng Tang
George Washington University

Claire Monteleoni
George Washington University

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_*) that satisfies (d_r^K, ϵ)-proximity condition must satisfy d_r^A-center separability for the same $d_r^A = d_r^K$.*

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

The need of an upper bound on ϵ 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 $d_{rs} := d_r + d_s > 0$ and $0 < \delta < \frac{d_{rs}}{6}$, let (X, T_*) satisfy (d_r, ϵ)-clusterability with $\epsilon n \geq n_{\min}$. Then there exists seeding $\{\nu\}$ and (X, T_*) with $\max_{s\in[k]} \|\mu_s - \nu_s\| = \delta$, such that if we apply Lloyd's algorithm on $(X, \{\nu\})$ until convergence, it returns a degenerate solution.*

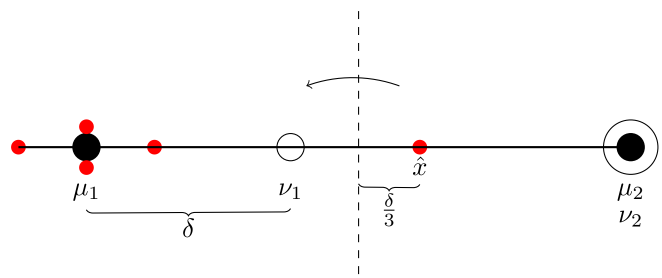

Figure 3: Two clusters in a bad instance with $\epsilon n \geq n_{\min}$: the solid black points are ground-truth centroids; the red points are in T_2, which are closer to ν_1 than to ν_2.

Proof. Let (X, T_*) be a dataset satisfying (d_r, ϵ)-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 ϵn 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 μ_2 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 ν_1, ν_2 fall on the line joining μ_1, μ_2, and $0 = \|\mu_2 - \nu_2\| \leq \|\mu_1 - \nu_1\| = \delta$ and let $\|\mu_1 - \mu_2\| = d > 6\delta$. Furthermore, $\forall x \in T_2$, $\|\hat{x} - \mu_2\| = \frac{d}{2} - \frac{\delta}{3}$. So $\|\hat{x} - \nu_2\| \geq \frac{d}{2} - \frac{\delta}{3}$ 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 $\frac{-4\frac{d}{2}+\frac{\delta}{3}}{5}$, whose distance to $\hat{x}$ is $\frac{\delta}{3} - \frac{-4\frac{d}{2}+\frac{\delta}{3}}{5} = \frac{2d}{5} + \frac{4\delta}{15}$. 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 μ_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_r^K, ϵ)-proximity condition with $\epsilon n < n_{\min}$ implies d_r^A-center separability.

Lemma 10. *If (X, T_*) satisfies (d_r, ϵ)-proximity condition with $\epsilon n < n_{\min}$, then it satisfies d_r-center separability.*

Proof. Since $\epsilon n < n_{\min}$, for any cluster $T_r, r \in [k]$, $\exists x \in T_r$ s.t. x satisfies the proximity condition, i.e., $\|\hat{x} - \mu_r\| \leq \|\hat{x} - \mu_s\| - d_{rs}$, for any $s \neq r$, where $d_{rs} := d_r + d_s$. Since $\|\mu_r - \mu_s\| \geq \|\hat{x} - \mu_s\| - \|\hat{x} - \mu_r\|$, we know all pairs of centroids are separated by at least d_{rs}. □

Let $d^*_{rs}(f) := f\sqrt{\phi_*}(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}})$, 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 clusterabiity, which depends on $\|X - C_*\|$, the maximal mean-departure of the entire dataset along one direction, $(d^*_{rs}(f), \epsilon)$-clusterability depends on $\sqrt{\phi_*}$. Fix T_* and the corresponding C_*, since $\|X - C_*\| \leq \|X - C_*\|_F = \sqrt{\phi_*} \leq \sqrt{rank(X - C_*)}\|X - C_*\| \leq \sqrt{2k}\|X - C_*\|$, when $d \leq k$, our assumption is (a factor of $\Theta(\sqrt{k})$) 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 OPT. Let $OPT^{(i\to j)}$ denote the cost of the clustering obtained by removing μ_i and assigning all its points to μ_j. Fix $\delta > 0$, the dataset satisfies $(1+\delta)$-weak deletion-stability if $OPT^{(i\to j)} > (1+\delta)OPT, \forall i \neq j$.*

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_* be the optimal k-means solution, we have for all $r \neq s \in [k]$, $\|\mu_r - \mu_s\|^2 \geq \frac{\delta\phi_*}{\max\{n_r, n_s\}}$. Furthermore, if $\delta > f(1 + \frac{1}{\alpha})$, then (X, T_*) 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 Δ denote the increase in k-means cost by merging T_s, T_r. Then $\Delta := \phi(T_{rs}) - \phi(T_s) - \phi(T_r) = \phi(\mu_{rs}, T_s) + \phi(\mu_{rs}, T_r) - \phi(\mu_s, T_s) - \phi(\mu_r, T_r) = n_s\|\mu_{rs} - \mu_s\|^2 + n_r\|\mu_{rs} - \mu_r\|^2 \leq (\|\mu_{rs} - \mu_s\|^2 + \|\mu_{rs} - \mu_r\|^2)(n_s + n_r)$. 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 $\mu_{rs} = \frac{\mu_s n_s + \mu_r n_r}{n_s + n_r}$, so μ_{rs} is on the segment joining (a convex combination of) μ_r and μ_s, i.e., $\|\mu_{rs} - \mu_s\| + \|\mu_{rs} - \mu_r\| = \|\mu_s - \mu_r\|$. Hence, $\Delta \leq 2\|\mu_s - \mu_r\|^2(n_s + n_r)$. Since (X, A_*) is $(1+\delta)$-weak-deletion stable, we have $\Delta \geq \delta\phi_*$. Therefore, $2\|\mu_s - \mu_r\|^2 \max\{n_s, n_r\} \geq \delta\phi_*$ and the first statement follows. The second statement follows by substituting the definition of α into the bound. □

8 Appendix B: proofs

Proof of Lemma 1. Suppose $\exists A_r$ s.t. $\forall s \in [k]$, $\|\nu_s - m(A_r)\| > \sqrt{2g+2}\sqrt{\frac{\phi(A)}{|A_r|}}$. Consider $\phi(C_0; A_r) = \sum_{x \in A_r}\|x - C_0(x)\|^2 = \sum_{x\in A_r}\|(C_0(x) - m(A_r)) - (x - m(A_r))\|^2 \geq \sum_{x \in A_r}\frac{\|C_0(x) - m(A_r)\|^2}{2} - \|x - m(A_r)\|^2$. Since all centroids in C_0 are at distance more than $\sqrt{2g+2}\sqrt{\frac{\phi(A)}{|A_r|}}$ away from $m(A_r)$, $\phi(C_0; A_r) > |A_r|\frac{(2g+2)\phi(A)}{2|A_r|} - \sum_{x\in A_r}\|x - m(A_r)\|^2 \geq (g+1)\phi(A) - \phi(A) = g\phi(A) \geq g\phi_{opt}$, where ϕ_{opt} denotes the optimal cost, contradicting the fact that C_0 is a g-approximation solution. □

Proof of Lemma 2. $\|x - \mu_s\| \geq \|x - \nu_s\| - \|\mu_s - \nu_s\| \geq \frac{1}{2}\|\nu_s - \nu_r\| - \|\mu_s - \nu_s\|$, since x is assigned to S_r by the Voronoi partition induced by $\{\nu\}$. Since $\|\nu_r - \nu_s\| = \|\nu_r - \mu_r + \mu_r - \mu_s + \mu_s - \nu_s\| \geq \|\mu_r - \mu_s\| - \|\nu_s - \mu_s\| - \|\nu_r - \mu_r\| \geq (1 - 2\gamma_t)\|\mu_r - \mu_s\|$. This implies $\|x - \mu_s\| \geq (\frac{1}{2} - \gamma_t)\|\mu_r - \mu_s\| - \|\mu_s - \nu_s\| \geq (\frac{1}{2} - 2\gamma_t)\|\mu_r - \mu_s\|$. For 2, $\|x - \mu_r\| \leq \|\mu_r - \nu_r\| + \|x - \nu_r\| \leq \|\mu_r - \nu_r\| + \|x - \nu_s\| \leq \|\mu_r - \nu_r\| + \|x - \mu_s\| + \|\mu_s - \nu_s\|$. Note the first statement also implies $(\frac{1}{2\gamma_t} - 2)\|\mu_l - \nu_l\| \leq \|x - \mu_s\|, \forall l \in [k]$. Our result follows after rearrangement. □

Proof. Combining the assumption on $\|\mu_r - \mu_s\|$ with the first statement of Lemma 2, we get $\rho^s_{out} n_s(\frac{1}{2} - 2\gamma_t)^2 y^2\frac{\phi_*}{n_s} \leq \sum_{r\neq s}|T_s \cap S_r|(\frac{1}{2} - 2\gamma_t)^2\|\mu_l - \mu_s\|^2 = \sum_{r\neq s}\sum_{A_i \in T_s\cap S_r}(\frac{1}{2} - 2\gamma_t)^2\|\mu_l - \mu_s\|^2 \leq \sum_{r\neq s}\sum_{A_i\in T_s\cap S_r}\|A_i - \mu_s\|^2 \leq \phi_*$. So $\rho^s_{out} \leq \frac{4}{(1-4\gamma_t)^2 y^2}$. Similarly, $\rho^s_{in} n_s(\frac{1}{2} - 2\gamma_t)^2 y^2\frac{\phi_*}{n_s} = \sum_{r\neq s}\rho^s_{in}(r) n_s(\frac{1}{2} - 2\gamma_t)^2 y^2\frac{\phi^s_*}{n_s} \leq \sum_{r\neq s}|S_s \cap T_r|\|\mu_s - \mu_l\|^2 \leq \phi_*$, implying $\rho^s_{in} \leq \frac{4}{(1-4\gamma_t)^2 y^2}$. □

Proof of Theorem 1. Fix any r, $\forall s \neq r$, $\Delta^t_r < \beta_t\sqrt{\phi_*}(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}}) = \beta_t\frac{1}{f}f\sqrt{\phi_*}(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}}) \leq \beta_t\frac{1}{f}\|\mu_r - \mu_s\|$. Hence $\gamma_t \leq \frac{\beta_t}{f} \leq \max\{\frac{\gamma}{8}, \frac{64}{f^2}\} < \frac{1}{8}$,we can apply Lemma 3 and get $\rho^r_{in} < \frac{4}{(1-4\gamma_t)^2 f^2}$, and $\rho^r_{out} < \frac{4}{(1-4\gamma_t)^2 f^2}$. Consider any $T_s \subset V(\mu_s), s \neq r$. Since $m(T_s \cap S_r) \subset V(\nu_r)$, by Lemma 2 $\|m(T_s \cap S_r) - \mu_s\| \geq (1 - 4\gamma_t)\|m(T_s \cap S_r) - \mu_r\|$. It's easy to check $\rho^r_{in} + \rho^r_{out} < \frac{1}{2}$. Applying Lemma 5 with $R = 1 - 4\gamma_t$ yields $\Delta^{t+1}_r \leq \frac{8}{(1-4\gamma_t)^2 f}\frac{\sqrt{\phi_*}}{\sqrt{n_r}} \leq$

$\frac{8}{(1-4\frac{\beta_t}{f})^2 f}\frac{\sqrt{\phi_*}}{\sqrt{n_r}} = \frac{8f}{(f-4\beta_t)^2}\frac{\sqrt{\phi_*}}{\sqrt{n_r}}$. Since $0 \le 4\frac{\beta_t}{f} \le \max\{4\gamma\frac{f}{8}\frac{1}{f}, 4\frac{64}{f}\frac{1}{f}\} \le \frac{1}{2}$ for any β_t and γ in the range we consider, we can always upper bound $\frac{8}{(1-4\frac{\beta_t}{f})^2 f}$ using $\frac{8}{(1-4\max\{4\gamma\frac{f}{8}\frac{1}{f},4\frac{64}{f}\frac{1}{f}\})^2 f}$.

If $\frac{f}{16} \le \gamma\frac{f}{8}$, then $0 \le 4\beta_t \le \frac{\gamma f}{2} < \frac{f}{2}$. So $(f-4\beta_t)^2 \ge (f-\frac{\gamma f}{2})^2 = f^2(1-\frac{\gamma}{2})^2$ and the latter is lower bounded by $32^2\frac{1}{4}$. Thus, $\Delta_r^{t+1} \le \frac{f}{32}\frac{\sqrt{\phi_*}}{\sqrt{n_r}} \le \frac{\gamma}{2}\frac{f}{8}\frac{\sqrt{\phi_*}}{\sqrt{n_r}}$.

If $\gamma\frac{f}{8} < \frac{f}{16}$, then we can similarly get $\Delta_r^{t+1} \le \frac{128}{9f}$. Finally, if $\frac{128}{9f} \ge \gamma\frac{f}{8}$, we get $(f-4\beta_t)^2 \ge f^2(1-\frac{64}{f^2})^2 > \frac{15^2}{16^2}f^2$. So $\Delta_r^{t+1} \le \frac{8f}{\frac{15^2}{16^2}f^2}\frac{\sqrt{\phi_*}}{\sqrt{n_r}} \le \frac{128}{9f}$. □

Proof of Theorem 2. By Lemma 1, after seeding $\Delta_r^1 < \frac{f}{8}\sqrt{\frac{\phi_*}{n_r}}, \forall r \in [k]$. Applying Theorem 1, running Lloyd's algorithm until convergence, we will obtain a solution s.t. $\Delta_r^T \le \frac{128}{9f}\sqrt{\frac{\phi_*}{n_r}}$. Then applying Lemma 3 with $\gamma_T \le \frac{128}{9f^2}$, we get $\rho_{in}^r + \rho_{out}^r \le \frac{8}{(1-\frac{4*128}{9f^2})^2 f^2} \le \frac{81}{8f^2}, \forall r \in [k]$. Hence, $d(T_*, S_T) := \sum_{r\in[k]}|S_r \triangle T_r| = \sum_r(\rho_{in}^r + \rho_{out}^r)n_r \le \frac{81}{8f^2}\sum_r n_r = \frac{81}{8f^2}n$. □

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_{SL} 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_{SL}, 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_{SL} has the same number of connected components as that of $G_{\max}$, G_{SL} must have exactly the same k connected components as those of $G_{\max}$, thus each component V_{SL}^r of G_{SL} 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 $\nu_i \in T_r$, then $\|\nu_i - \mu_r\|$ satisfies the bound in A with high probability. Let $q := \|\nu_i - \mu_r\|^2$, we have $0 \le q \le \max_{x\in T_r}\|x-\mu_r\|^2$ and $E[q|\nu_i \in T_r] = \frac{\sum_{x\in T_r}\|x-\mu_r\|^2}{n_r} = \frac{\phi_*^r}{n_r}$. Then applying Hoeffding's bound, we get conditioning on the event $\{\nu_i \in T_r\}$,

$$Pr(q - Eq \ge (\frac{f}{4}-1)\frac{\phi_*^r}{n_r}) \le \exp(-\frac{2[(\frac{f}{4}-1)\frac{\phi_*^r}{n_r}]^2}{(\max_{x\in T_r}\|x-\mu_r\|^2)^2})$$

Substituting $w_{\min}$ for every r and applying union bound, we get $Pr(A^c) \le m\exp(-2(\frac{f}{4}-1)^2 w_{\min}^2)$. Now the probability of a cluster T_r not being seeded after m trials is $(1-p_r)^m \le \exp(-mp_r)$. Applying union bound, we get $Pr(B^c) \le k\exp(-mp_{\min})$. Applying union bound again, we get $Pr(A \cap B) \ge 1 - m\exp(-2(\frac{f}{4}-1)^2 w_{\min}^2) - k\exp(-mp_{\min})$. □

Proof of Lemma 8. Let $\pi(i) = \pi(j) = r$. Then $\|\nu_i - \nu_j\| \le \|\nu_i - \mu_r\| + \|\nu_j - \mu_r\| \le 2\frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^r}{n_r}}$. Let $\pi(p) = t, \pi(q) = s$. Then $\|\nu_p - \nu_q\| \ge \|\mu_t - \mu_s\| - \|\nu_p - \mu_t\| - \|\nu_q - \mu_s\| \ge f\sqrt{\phi_*}(\frac{1}{\sqrt{n_t}} + \frac{1}{\sqrt{n_s}}) - \frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^t}{n_t}} - \frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^s}{n_s}} > \frac{f}{2}\sqrt{\phi_*}(\frac{1}{\sqrt{n_t}} + \frac{1}{\sqrt{n_s}})$. On the other hand, recall $\alpha := \min_{r\ne s}\frac{n_r}{n_s}$, we get $\sqrt{\frac{1}{n_r}} \le \min\{\frac{1}{\sqrt{\alpha n_t}}, \frac{1}{\sqrt{\alpha n_s}}\}$, so $2\sqrt{f}\sqrt{\frac{\phi_*^r}{n_r}} \le \sqrt{f\phi_*^r}(\frac{1}{\sqrt{\alpha n_t}} + \frac{1}{\sqrt{\alpha n_s}})$. Since $f > \frac{1}{\alpha}$, we get $\sqrt{f}\sqrt{\frac{\phi_*^r}{n_r}} \le \frac{f}{2}\sqrt{\phi_*^r}(\frac{1}{\sqrt{n_t}} + \frac{1}{\sqrt{n_s}})$. Substituting this bound on $\|\nu_i - \nu_j\|$ and comparing it with the lower bound on $\|\nu_p - \mu_q\|$ completes the proof. □

Proof of Theorem 4. Consider $A \cap B$. Under this event, we know that the ground-truth T_* induces a non-degenerate k-clustering of $\{\nu_i, i \in [m]\}$, which we denote by $\{V_r^*, r \in [k]\}$ with $V_r^* := T_r \cap \{\nu_i, i \in [m]\}, \forall r \in [k]$. In addition, Lemma 8 implies the bipartite edge sets E_{in}^* and E_{out}^* induced by $\{V_r^*, r \in [k]\}$ satisfies $\forall e_1 \in E_{in}^*, e_2 \in E_{out}^*$, $w(e_1) < w(e_2)$. Thus, by Lemma 6 if we apply Single-Linkage on $G_0 = (\cup_{r\in[k]}V_r^*, \emptyset)$ until k components remain, each returned connected component S_r corresponds to exactly one cluster V_r^*. In addition, with the seeding guarantee by event A, $\forall r \in [k], \|m(V_r^*) - \mu_r\| \le \frac{1}{|V_r^*|}\sum_{\nu_i\in V_r^*}\|\nu_i - \mu_r\| \le \frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^r}{n_r}}$. Noting $Pr(A\cap B) \ge 1 - m\exp(-2(\frac{f}{4}-1)^2 w_{\min}^2) - k\exp(-mp_{\min})$ by Lemma 7 and $m(V_r^*) = \nu_r^*$ completes the proof. □

Proof of Theorem 5. Consider each cluster S_r in the final solution. Its k-means cost, by definition, is $\phi(\{m(S_r)\}, S_r) \le \phi(\{\mu_r\}, S_r) = \phi(\{\mu_r\}, S_r \cap T_r) + \phi(\{\mu_r\}, \cup_{s\ne r}S_r \cap T_s)$. By Theorem 4 and our assumption on center separation, $\gamma \le \frac{\sqrt{f}}{2f} < \frac{1}{4}$, we can apply Lemma 2 to get $\phi(\{\mu_r\}, \cup_{s\ne r}S_r \cap T_s) = \sum_{s\ne r}\sum_{x\in S_r\cap T_s}\|x-\mu_r\|^2 \le \sum_{s\ne r}\sum_{x\in S_r\cap T_s}\frac{1}{(1-4\gamma)^2}\|x-\mu_s\|^2$, by Lemma 2. Since $f > 16$, we get $\frac{1}{(1-4\gamma)^2} \le 4$. Summing over all $r \in [k]$, $\phi(\{S_r, r \in [k]\}) \le \sum_r\phi(\{\mu_r\}, S_r \cap T_r) + \sum_r\frac{1}{(1-4\gamma)^2}\sum_{s\ne r}\sum_{x\in S_r\cap T_s}\|x-\mu_s\|^2 \le 4(\sum_r\phi(\{\mu_r\}, S_r \cap T_r) + \sum_r\sum_{s\ne r}\sum_{x\in S_r\cap T_s}\|x - \mu_s\|^2) = 4\{\sum_r(\sum_{x\in S_r\cap T_r}\|x-\mu_r\|^2 + \sum_{s\ne r}\sum_{x\in S_r\cap T_s}\|x-\mu_s\|^2)\} = 4\{\sum_r\sum_{x\in S_r}\|x -$

$C_*(x)\|^2\} = 4\phi_*$ (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 $(\frac{1}{1-4\gamma})^2$, where $\gamma \leq \frac{\sqrt{f}}{2f}$. So to have a $1+\epsilon$-guarantee, it suffices to have $(\frac{1}{1-4\frac{\sqrt{f}}{2f}})^2 \leq 1+\epsilon$, which holds if $f = \Omega(\frac{1}{\epsilon^2})$. Now we find a sufficient condition for the success probability to be at least $1-\delta$. It suffices to require that $m\exp(-2(\frac{f}{4}-1)^2 w_{\min}^2) \leq \frac{\delta}{2}$ and $k\exp(-mp_{\min}) \leq \frac{\delta}{2}$. So we need $\frac{1}{p_{\min}}\log\frac{2k}{\delta} \leq m \leq \frac{\delta}{2}\exp(2(\frac{f}{4}-1)^2 w_{\min}^2)$. Note for this inequality to be possible, we also need $\frac{\delta}{2}\exp(2(\frac{f}{4}-1)^2 w_{\min}^2) \geq \frac{1}{p_{\min}}\log\frac{2k}{\delta}$, imposing an additional constraint on f. Taking log on both sides and rearrange, we get $(\frac{f}{4}-1)^2 \geq \frac{1}{2w_{\min}}\log(\frac{\frac{2}{\delta}\log\frac{2k}{\delta}}{p_{\min}})$. Thus, it is sufficient for a $1+\epsilon$-approximation to hold with probability at least $1-\delta$ if $f = \Omega\left(\sqrt{\log(\frac{\frac{1}{\delta}\log\frac{k}{\delta}}{p_{\min}})} + \frac{1}{\epsilon^2}\right)$, and we choose m to be in the interval $[\frac{1}{p_{\min}}\log\frac{2k}{\delta}, \frac{\delta}{2}\exp(2(\frac{f}{4}-1)^2 w_{\min}^2)]$. □

9 Appendix C: details on the generation of synthetic data

The clusterability of each dataset is controlled by three parameters (ϵ, α, u), 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 α increases), $u \in [0,\infty)$ controls the degree of balance of the cluster sizes in the ground-truth clustering (u is the symmetric Dirichlet prior for the multinomial distribution; the higher u 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[4], $k=2d$, 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, $n_1,\dots,n_k$, such that $\sum_{i=1}^k n_i = n$, where n_i is sampled from a multinomial distribution with parameter θ, with u characterizing the sparsity of θ. 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 α.

[4]In the experiments, we restrict our attention to the case $d < k$.

References

[1] Margareta Ackerman and Shai Ben-David. Clusterability: A theoretical study. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, pages 1–8, 2009.

[2] David Arthur, Bodo Manthey, and Heiko Röglin. Smoothed analysis of the k-means method. *J. ACM*, 58(5):19, 2011.

[3] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035, 2007.

[4] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a PTAS for k-median and k-means clustering. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 309–318, 2010.

[5] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, pages 37–49, 2012.

[6] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2009, New York, NY, USA, January 4-6, 2009*, pages 1068–1077, 2009.

[7] Shai Ben-David. Computational feasibility of clustering under clusterability assumptions. *CoRR*, abs/1501.00437, 2015.

[8] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012.

[9] Léon Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, pages 585–592, 1994.

[10] Adam Coates and Andrew Y. Ng. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade - Second Edition*, pages 561–580. 2012.

[11] Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 215–223, 2011.

[12] Douglas R. Cutting, Jan O. Pedersen, David R. Karger, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992*, pages 318–329, 1992.

[13] Amit Daniely, Nati Linial, and Michael Saks. Clustering is difficult only when it does not matter. *CoRR*, abs/1205.4891, 2012.

[14] Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science, FOCS '99, 17-18 October, 1999, New York, NY, USA*, pages 634–644, 1999.

[15] Sanjoy Dasgupta. Lecture 4 — hierarchical clustering. CSE 291: Unsupervised learning, 2008.

[16] Sariel Har-Peled and Bardia Sadri. How fast is the k-means method? *Algorithmica*, 41(3):185–202, 2005.

[17] Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering (extended abstract). In *Proceedings of the Tenth Annual Symposium on Computational Geometry, Stony Brook, New York, USA, June 6-8, 1994*, pages 332–339, 1994.

[18] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[19] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. *SIAM J. Comput.*, 38(3):1141–1156, 2008.

[20] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for k-means clustering. *Comput. Geom.*, 28(2-3):89–112, 2004.

[21] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 299–308, 2010.

[22] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[23] Michael S. Lewicki and Terrence J. Sejnowski. Learning overcomplete representations. *Neural Comput.*, 12(2):337–365, February 2000.

[24] Fei-Fei Li and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 524–531, 2005.

[25] S. Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, Mar 1982.

[26] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. In *Proceedings of the 3rd International Workshop on Algorithms and Computation*, WALCOM '09, pages 274–285, Berlin, Heidelberg, 2009. Springer-Verlag.

[27] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[28] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *J. ACM*, 59(6):28, 2012.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[30] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North*

Carolina, USA, April 26-30, 2010, pages 1177–1178, 2010.

[31] Andrea Vattani. k-means requires exponentially many iterations even in the plane. *Discrete & Computational Geometry*, 45(4):596–616, 2011.

[32] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860, 2004.

[33] Alon Vinnikov and Shai Shalev-Shwartz. K-means recovers ICA filters when independent components are sparse. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 712–720, 2014.

[34] Ulrike von Luxburg, Robert C. Williamson, and Isabelle Guyon. Clustering: Science or art? In *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, pages 65–80, 2012.

[35] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3360–3367, 2010.