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

Cheng Tang
George Washington University

Claire Monteleoni
George Washington University

Abstract

We provide new analyses of Lloyd's algorithm (1982), commonly known as the k-means clustering algorithm. Kumar and Kannan (2010) showed that running k-SVD followed by a constant approximation k-means algorithm, and then Lloyd's algorithm, will correctly cluster nearly all of the dataset with respect to the optimal clustering, provided the dataset satisfies a deterministic clusterability assumption. This method is viewed as the "Swiss Army knife" for clustering problems, subsuming popular generative models such as Gaussian mixtures. However, it is tailored to high dimensional data, i.e., when $d >> k$.

We analyze Lloyd's algorithm for general d without using the spectral projection, which leads to a weaker assumption in the case $d < k$. Surprisingly, we show that a simple and scalable heuristic that combines random sampling with Single-Linkage serves as a good seeding algorithm for Lloyd's algorithm under this assumption. We then study stopping criteria for Lloyd's algorithm under the lens of clusterability, accompanied by controlled simulations.

1 Introduction

Despite the growing number of new clustering algorithms, many practitioners stick with a few heuristics—Lloyd's algorithm [25] being one of them [18, 34]. However, the current level of theoretical understanding of this algorithm does not match the popularity it enjoys. Lloyd's algorithm is often associated with k-means clustering since Lloyd's update can be viewed as a local descent for the k-means objective [9]. The k-means objective is NP-hard to optimize [26], though various algorithms have been shown to approximate it [20, 3]. On the other hand, additional information generally referred to as the "clusterability" of the dataset has been shown to help with both the design and analysis of clustering algorithms. The argument that, for a practitioner, it only makes sense to perform clustering if a hidden clustering-like structure underlies a dataset, has been used to justify clusterability assumptions [6, 1, 13, 7]. Following the same logic, when one believes such structure exists, it is more reasonable to cast the task of clustering as that of finding the hidden structure rather than optimizing an objective [6]. We study Lloyd's algorithm beyond the scope of k-means clustering, with the goal of uncovering the hidden cluster structure, which we refer to as the "ground-truth," following convention [6].

In computer vision, Lloyd's algorithm is widely used for learning dictionaries [35, 22, 24], where each learned centroid is treated as an dictionary item. In such cases, centroids learned by Lloyd's algorithm are used to represent a dataset, and one usually learns an "overcomplete" dictionary, where $d \ll k$ [23, 8]. Recently, Lloyd's algorithm was also shown to work well empirically for unsupervised feature learning [11], where a whitening step is speculated to be important to its success in this context [11, 33]. We believe characterizing conditions under which Lloyd's algorithm works well can shed light on these applications as well.

2 Preliminaries

Our clustering problem starts with a discrete dataset X, an n by d matrix with each row a data point $x \in X$. We assume X admits one (or more) ground-truth non-degenerate k-clustering $T_* = \{T_s, s \in [k]\}$ [1]. Let $n_s := |T_s|, \forall s \in [k]$, and let $n_{\min} := \min_{s\in[k]} n_s$ and $n_{\max} := \max_{s\in[k]} n_s$. , which partitions X and in addition satisfies $d^*_{rs}(f)$-center separability, defined below.

[1] We say a k-clustering is degenerate if any of its k clusters are empty.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

Mappings Fix a point set Y, we let $m(Y)$ denote the mean of Y. In general, each clustering assignment $A := \{A_s, s \in [k]\}$ induces a unique set of centroids $C = \{m(A_s), s \in [k]\}$. For a ground-truth T_*, we denote the induced centroids by $\mu_s := m(T_s), \forall s \in [k]$. Alternatively, fix a set of k centroids C, we let $C(\cdot)$ denote a mapping $C(x) := \arg\min_{c_r \in C} \|x - c_r\|$. This mapping induces a k-clustering X, i.e., a Voronoi partition of X. We let $V(c_r)$ denote the Voronoi region $\{x \in \mathbb{R}^d, \|x - c_r\| \le \|x - c_s\|, \forall s \neq r\}$.

K-means cost For any subset of points Y, with respect to an arbitrary set of k centroids C, we denote its k-means cost by $\phi(C, Y) := \sum_{y \in Y} \|y - C(y)\|^2$. For a k-clustering $A = \{A_r\}$ of X, we denote its k-means cost with respect to an arbitrary set of k centroids C by $\phi(C, A) := \sum_{r=1}^{k} \phi(C, A_r)$ (or simply $\phi(A)$ when $c_r = m(A_r), \forall c_r \in C, r \in [k]$). We let $\phi_*^r := \phi(\mu_r, T_r)$, and let $\phi_* := \sum_{r=1}^{k} \phi_*^r$ denote the k-means cost of T_* with respect to X.

Characterization of (X, T_*) Three properties of the dataset-solution pair (X, T_*) are useful to our analysis. We use $p_{\min} := \min_{r \in [k]} \frac{n_r}{n}$ to characterize the fraction of the smallest cluster in T_* to the entire dataset. We use $\alpha := \min_{r \neq s} \frac{n_r}{n_s}$ to characterize the level of cluster balance in T_* ($0 < \alpha \le 1$ always holds; $\alpha = 1$ when the ground-truth is perfectly balanced). We use $w_r := \frac{\frac{\phi_*^r}{n_r}}{\max_{x \in T_r} \|x - \mu_r\|^2}$ to characterize the ratio between average and maximal "spread" of cluster T_r, and we let $w_{\min} := \min_{r \in [k]} w_r$.

Algorithm-related notation We analyze Lloyd's algorithm (Algorithm 1) using different seeding procedures. In analyzing the t-th iteration of Lloyd's update, we let $\{\nu_r, r \in [k]\}$ denote the set of centroids C_{t-1} and $\Delta_s^t := \|\mu_s - \nu_s\|$, and we let $\gamma_t := \max_{s, r \neq s} \frac{\Delta_s^t}{\|\mu_r - \mu_s\|}$. We let $S := \{S_r, r \in [k]\}$ denote the clustering S_t. Fix T_s, let $\rho_{in}^s := \frac{\sum_{r \neq s} |T_r \cap S_s|}{n_s}$, $\rho_{out}^s := \frac{\sum_{r \neq s} |T_s \cap S_r|}{n_s}$, i.e., $\rho_{in}^s + \rho_{out}^s$ captures the fraction of misclassification in S_s with respect to T_s (we use the word "misclassification" for clustering error following [21]).

Algorithm 1 Lloyd's algorithm

1: (Seeding) Select an initial set of k centroids C_0
2: (Lloyd's updates)
3: **while** Lloyd's algorithm has not converged or the stopping criterion is not met **do**
4: $\quad S_t \leftarrow \{V(\nu_r) \cap X, \nu_r \in C_{t-1}, r \in [k]\}$
5: $\quad C_t \leftarrow \{m(S_r), S_r \in S_t, r \in [k]\}$
6: **end while**

Our goal and clusterability assumption We aim to show that under a sufficient clusterability assumption Algorithm 1 finds a k-clustering of X, $S = \{S_r, r \in [k]\}$, such that the *clustering distance* between S and T_*, defined as the sum of symmetric set differences, $d(T_*, S) := \sum_{r \in [k]} \min_{\pi:[k] \to [k]} |S_{\pi(r)} \triangle T_r|$, is small ($\pi$ is a permutation of $[k]$). The clusterability assumption our analysis relies on is a special realization of the following assumptions.

Definition 1 ((d_{rs}-center separability)**.** *A dataset-solution pair (X, T_*) satisfies d_{rs}-center separability if $\forall r \in [k], s \neq r$, $\|\mu_r - \mu_s\| \ge d_{rs}$, where d_{rs} is a distance measure, a function, of pairwise clusters $T_r, T_s, r \neq s$.*

In particular, we require (X, T_*) to satisfy $d_{rs}^*(f)$-center separability with $d_{rs}^*(f) := f\sqrt{\phi_*}(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}})$.

2.1 Related work

Most existing analyses study upper and lower bounds on the time complexity of Lloyd's algorithm [17, 2, 16, 31]. For performance guarantees, Ostrovsky et al. [28] modified the Lloyd's update and showed that when combined with a k-means++ seeding [3], a Lloyd-like algorithm finds a $(1+\epsilon)$-approximation to the k-means objective on well-clusterable instances. On the other hand, Kumar and Kannan [21] generalized the assumptions of mixture models [19, 32, 14] and proposed a deterministic analog, which they show is weaker than that in [28]. Under this assumption they showed the **k-SVD + constant k-means approximation + Lloyd's update** scheme efficiently and correctly clusters all but a $k^2\epsilon$-fraction of points with respect to the ground-truth.

The clusterability assumption introduced in [21] carries a geometric intuition. It demands that any two clusters r and s in the ground-truth must be sufficiently separated from each other, where the degree of separation is measured by the difference between the (projected) intra and inter cluster distances, in units of $d_r^K + d_s^K$, with

$$d_r^K := \frac{ck}{\sqrt{n_r}} \|X - C\|$$

In the equation above, $\|\cdot\|$ denotes the spectral norm, and we abuse the notation of the set of k-centroids, C, by using it to represent a n by d matrix, whose i-th row $C_i = C(X_i)$. Formally, their assumption requires the following:

Definition 2 ((d_r^K, ϵ)-proximity condition [21])**.** *A dataset-solution pair (X, T_*) satisfies (d_r^K, ϵ)-proximity condition if at least a $(1-\epsilon)$-fraction of points in X satisfy, $\forall s, \forall x \in T_s$, $\|\hat{x} - \mu_r\| - \|\hat{x} - \mu_s\| \ge d_r^K + d_s^K$ for any $r \neq s$, where $\hat{x}$ is the projection of x onto the line joining μ_r and μ_s.*

Note how the proximity condition becomes stronger as c, k becomes larger. Subsequent work [5] introduced a related center separation condition[2], and reduced the linear dependence of d_r^K on k by a factor of $\sqrt{k}$. Formally, they require $\forall r \in [k], s \neq r$

$$\|\mu_r - \mu_s\| \geq d_r^A + d_s^A \tag{1}$$

with

$$d_r^A := \frac{c\sqrt{k}}{\sqrt{n_r}}\|X - C\|$$

The assumption can be viewed as a deterministic analog of the earlier work on learning mixtures of Gaussians (or other distributions), where the mean separation between two Gaussians is measured by their maximal standard deviations [19, 32, 14]; the spectral norm can be viewed as an empirical counterpart of maximal standard deviation. The reduction on the dependence of k in d_r^A exploited the property of the spectral subspace and the spectral norm, and relied on an additional "ball k-means" pruning step before the iterative application of Lloyd's updates.

2.2 Our contributions

In Section 3, we analyze the original Lloyd's algorithm in general dimension without using the spectral projection, which leads to a weaker assumption than [21, 5] when $d < k$. In Section 4, we devise a clustering algorithm for Lloyd's algorithm that satisfies the seeding requirement for Lloyd's algorithm in Section 3. In fact, it achieves constant k-means approximation on its own under our clusterability assumption. Notably, its performance guarantee does not depend on the data size, making it highly scalable for large datasets.

We next elaborate on how we achieved the improvement over [21, 5] in Section 3. Both [21] and [5] focused on the case $d >> k$, and used the spectral norm as a measure of cluster separation. It is possible to directly extend the analysis of [21] for general d, without using the spectral projection. In this case, the seeding guarantee in Lemma 5.1 of [21] becomes (assuming the data size is larger than both d and k):

$$|\mu_r - \nu_r| \leq 20\sqrt{2\max\{d, k\}}\frac{\|X - C\|}{\sqrt{n_r}}$$

For $d < k$, the original statement, which depends on k, is recovered up to a constant factor $\sqrt{2}$. Then using the proximity condition with possibly larger constant c, one can obtain the convergence of Lloyd's algorithm for general d, without the spectral projection. However, in this case,

$$\|X - C\|_F \leq \sqrt{2k}\|X - C\| < 2k\|X - C\|$$

[2]In fact, we show (in the Appendix) that the proximity condition implies the center separation assumption.

As a result, d_r^K and d_r^A both become rather large compared to $\|X - C\|_F$, which notably is the square root of the k-means cost of C on X. This leads us to address the question of whether analogous results using Frobenius norm as a separation measure can be achieved.

We found the answer to be positive. To adapt the result of [21] to ours using $\|X - C\|_F$, we modified three parts of their analysis in Section 3: 1). Our Lemma 1 extends the seeding lemma in [21] to be compatible with $\|X - C\|_F$. 2). Lemma 3 shows a small distance to the ground-truth centroids implies that the misclassification error is small. Both follow smoothly by distilling the analysis in Theorem 3.1 of [5]. 3). The more interesting part is to show the other direction: a small misclassification error implies a small distance to the ground-truth centroids. Both [21] and [5] rely on the relation stated in Fact 1.3 of [5], which has a tight $\sqrt{k}$ dependence (see discussion in [5]). To mitigate this dependence on k, [5] exploited the property of the k-SVD subspace and spectral norm, and added the "ball k-means" step to achieve a reduction on the dependence of k in step 2). Instead, we directly eliminate the dependence on k without modifying the original Lloyd's algorithm, via the decomposability of k-means objective, i.e., the global k-means objective is the sum of the k-means cost of each of the individual clusters (our Lemma 5). As a result, our assumption, formally stated as Definition 1, is of the form (1) with d_r^A substituted by d_r^*, with

$$d_r^* = \frac{c}{\sqrt{n_r}}\|X - C\|_F$$

This is weaker than d_r^K, d_r^A for the same constant c in the case $d < k$. The assumptions in [21, 5] and our work all lead to similar iteration-wise convergence result as in our Theorem 1, implying the geometric convergence of Lloyd's algorithm upon a good initialization.

3 Analysis of global convergence

We present our analysis of Lloyd's algorithm in a way that corresponds to the seeding, clustering assignment, and centroid update steps of Algorithm 1. This proof framework builds on and simplifies that of [21, 5]. The proofs of Lemmas 1,2, 3, Theorem 1 and Theorem 2 are similar to those in [21, 5]; we move them to the Appendix.

The seeding phase For the seeding phase, we show using any g-approximate k-means algorithms, the distance between the seeds and the mean of any k-clustering $\{A_r, r \in [k]\}$ can be bounded.

Lemma 1. *Given a dataset X, and let $C_0 = \{\nu_1, \ldots, \nu_k\}$ be the set of centroids produced by a g-*

approximate k-means algorithm, then for any clustering of X, denoted by $A := \{A_r, r \in [k]\}$, we have $\forall A_r$, $\exists \nu_r$ s.t. $\|\nu_r - m(A_r)\| \leq \sqrt{2g+2}\sqrt{\frac{\phi(A)}{|A_r|}}$.

Lloyd's update—the reassignment phase For the clustering assignment step, we show a small distance between the current clustering centroids and those in the ground-truth implies a small number of misclassifications upon reassignment. Specifically, fix any ground-truth cluster T_s, if γ_t is sufficiently small, we show

- Any points added to S_r must not be close to another centroid μ_s in the ground-truth (Lemma 2).
- ρ_{in}^s and ρ_{out}^s are upper bounded by γ_t (Lemma 3).

Lemma 2. *If $\gamma_t < \frac{1}{4}$, then $\forall r \in [k], \forall x \in V(\nu_r)$,*

1. $\|x - \mu_s\| \geq (\frac{1}{2} - 2\gamma_t)\|\mu_r - \mu_s\|, \forall s \neq r$
2. $\|x - \mu_r\| \leq \frac{1}{1-4\gamma_t}\|x - \mu_s\|$

The first statement of Lemma 2 in turn implies there cannot be too many misclassified points of S_s if the ground-truth is well clusterable, since otherwise they would induce a k-means cost larger than ϕ_*.

Lemma 3. *If $\gamma_t < \frac{1}{4}$, and if for some $s \in [k]$, $\forall r \neq s$, $\|\mu_s - \mu_r\| \geq y\frac{\sqrt{\phi_*}}{\sqrt{n_s}}$, then $\rho_{out}^s \leq \frac{4}{(1-4\gamma_t)^2 y^2}$ and $\rho_{in}^s \leq \frac{4}{(1-4\gamma_t)^2 y^2}$.*

Lloyd's update—the mean-adjustment phase Now we show a small number of misclassifications in turn implies a smaller (or at least the same) centroidal distance after the mean-adjustment phase. That is, we upper bound $\|\mu_s - m(S_s)\|$ using ρ_{out} and ρ_{in}. We achieve this through two observations: 1). The number of misclassified points is small. 2). The misclassified points do not incur too much additional cost to the k-means objective.

We first present a well known property of the k-means objective, using which we can measure the distance between any point c and the mean of a cluster Y using $\frac{\sqrt{\phi(c,Y)}}{\sqrt{|Y|}}$ as a unit.

Lemma 4 (Lemma 2.1 of [20])**.** *For any point set Y and any point c in $\mathbb{R}^d$, $\phi(c, Y) = \phi(m(Y), Y) + |Y|\|m(Y) - c\|^2$.*

For $Y \subset T_s$, this further implies

$$\|m(Y) - \mu_s\| \leq \frac{\sqrt{\phi_*^s}}{\sqrt{|Y|}}$$

and that

$$\|m(Y) - \mu_s\| \leq \frac{\sqrt{|T_s \setminus Y|}\sqrt{\phi_*^s}}{|Y|}$$

These two inequalities are used in proving our main lemma.

Lemma 5 (main lemma)**.** *Fix a target clustering T_s and let S_s be a set of points created by removing $\rho_{out}^s n_s$ points from T_s (we denote these points by $T_{s\to r}$) and adding $\rho_{in}^s(r) n_s$ points ($T_{r\to s}$) from each cluster $r \neq s$. If*

- *The added points satisfy $\|m(T_{r\to s}) - \mu_r\| \geq R\|m(T_{r\to s}) - \mu_s\|$*
- *$\rho_{in}^s + \rho_{out}^s < \frac{1}{2}$, where $\rho_{in}^s = \sum_{r\neq s}\rho_{in}^s(r)$*

Then $\|m(S_s) - \mu_s\| \leq (\sqrt{\frac{\rho_{out}^s}{n_s}} + \frac{1}{R}\sqrt{\frac{\rho_{in}^s}{n_s}})2\sqrt{\phi_*}$

Proof. $\|m(S_s) - \mu_s\| = \|\frac{m(S_s\cap T_s)|S_s\cap T_s| + \sum_{r\neq s} m(S_s\cap T_r)|S_s\cap T_r|}{|S_s|} - \mu_s\| = \|\frac{m(S_s\cap T_s)n_s(1-\rho_{out}^s) + \sum_{r\neq s} m(S_s\cap T_r)\rho_{in}^s(r)n_s}{|S_s|} - \mu_s\| \leq \frac{n_s(1-\rho_{out}^s)\|m(S_s\cap T_s)-\mu_s\|}{|S_s|} + \sum_{r\neq s}\frac{\rho_{in}^s(r)n_s}{|S_s|}\|m(S_s \cap T_r) - \mu_s\| \leq \frac{2n_s(1-\rho_{out}^s)\|m(S_s\cap T_s)-\mu_s\|}{n_s} + \sum_{r\neq s}\frac{2\rho_{in}^s(r)n_s}{n_s}\|m(S_s\cap T_r) - \mu_s\| \leq 2(1-\rho_{out}^s)\|m(S_s\cap T_s) - \mu_s\| + \sum_{r\neq s}\frac{2\rho_{in}^s(r)}{R}\|m(S_s \cap T_r) - \mu_r\|$ The second inequality uses the assumption $\rho_{in}^s + \rho_{out}^s < \frac{1}{2}$ and the last inequality uses the first assumption. We have $\|m(S_s \cap T_s) - \mu_s\| \leq \frac{\sqrt{\rho_{out}^s n_s}\sqrt{\phi_*^s}}{n_s(1-\rho_{out}^s)}$ and $\|m(S_s \cap T_r) - \mu_r\| \leq \frac{\sqrt{\phi_*^r}}{\sqrt{n_s\rho_{in}^s(r)}}$. So $\sum_{r\neq s}\frac{2\rho_{in}^s(r)n_s}{Rn_s}\|m(S_s \cap T_r) - \mu_r\| \leq \frac{2}{R\sqrt{n_s}}\sum_{r\neq s}\sqrt{\rho_{in}^s(r)}\sqrt{\phi_*^r}$. Applying Cauchy Schwarz inequality, we get $\sum_{r\in[k]}\sqrt{\rho_{in}^s(r)}\sqrt{\phi_*^r} \leq \sqrt{\sum_{r\in[k]}\rho_{in}^s(r)}\sqrt{\sum_{r\in[k]}\phi_*^r} = \sqrt{\rho_{in}^s\phi_*}$ Our statement thus follows through. □

Remark: Note this result alone does not depend on our clusterability assumption. However, in order to translate this bound into an upper bound on the ratio $\frac{\Delta_s^{t+1}}{\|\mu_s - \mu_r\|}$, we would need the center separation to be of the same order, i.e., lower bounded by $\Omega(\frac{\sqrt{\phi_*}}{\sqrt{n_s}})$.

Applying the reassignment and mean-adjustment phases recursively, our first conclusion is when the current solution is close to a well-clusterable solution Lloyd's algorithm converges rapidly.

Theorem 1. *Assume there is a dataset-solution pair (X, T_*) satisfying $d_{rs}^*(f)$-center separability, with $f > 32$. If at iteration t, $\forall r \in [k], \Delta_r^t < \beta_t\frac{\sqrt{\phi_*}}{\sqrt{n_r}}$ with $\beta_t < \max\{\gamma\frac{f}{8}, \frac{128}{9f}\}$ with $\gamma < 1$, then $\forall r \in [k], \Delta_r^{t+1} < \beta_{t+1}\frac{\sqrt{\phi_*}}{\sqrt{n_r}}$, with $\beta_{t+1} < \max\{\frac{\gamma}{2}\frac{f}{8}, \frac{128}{9f}\}$.*

Theorem 1 suggests when the $\max_r \|\nu_r - \mu_r\|$ is sufficiently small, Lloyd's update converges linearly to the ground-truth centroids until it reaches a plateau-like phase. Combining it with Lemma 1, we reach our main conclusion.

Theorem 2. *Assume (X, T_*) satisfies $d^*_{rs}(f)$-center separability with $f > 32$. If we cluster X using Algorithm 1, where we choose a g-approximate k-means algorithm with $g < \frac{f^2}{128} - 1$ for the seeding, and execute Lloyd's update until convergence, then all but $\frac{81}{8f^2}$ fraction of the points will be correctly classified with respect to T_*.*

Remark: By Theorem 2, if $f = \Omega(\sqrt{k})$, then using a $O(k)$-approximate k-means algorithm for seeding in Lloyd's algorithm suffices to correctly cluster all but $O(\frac{1}{k})$-fraction of points.

4 A simple and fast heuristic seeding

Since the goal of analyzing Lloyd's algorithm is to justify the practical success of popular heuristics, requiring it to be initialized by an approximation algorithm seems unreasonable in this regard; most approximation algorithms to our knowledge are computationally expensive and complicated to implement. After all, in Section 3 as well as in [21, 5], the seeding algorithm has been treated as a blackbox.

In this section, we leverage the same clusterability assumption we made in analyzing Lloyd's algorithm to devise a simple and fast seeding algorithm. Algorithm 2, similar to the *buckshot algorithm* [12] which is used in practice for text clustering, achieves a constant k-means approximation under our clusterability assumption as a standalone algorithm, and serves as a seeding algorithm for Lloyd's algorithm that satisfies the requirement in Theorem 2. Moreover, the time complexity of this algorithm is independent of the data size, making it highly scalable to massive datasets.

The algorithm is based on uniform random sampling of the dataset, a common seeding strategy for Lloyd's algorithm. However, its obvious drawback is that small clusters may not be seeded while large clusters may contain more than one seed. To ensure that each cluster is seeded, it is natural to consider over-seeding, i.e., sampling $m > k$ points from X. Then the challenge becomes selecting k seeds from the sampled points. We show Single-Linkage [15], a commonly used heuristic (usually for hierarchical clustering [6]), can be used to merge points that belong to the same ground-truth cluster. Our main result for this section is that for a well-clusterable dataset, the heuristic seeding procedure presented in Algorithm 2 followed by Lloyd's algorithm correctly classifies most of the dataset with significant probability.

Algorithm 2 Heuristic seeding

1: $\{\nu_i, i \in [m]\} \leftarrow$ sample m points from X uniformly at random with replacement
2: $\{S_1, \dots, S_k\} \leftarrow$run Single-Linkage on $\{\nu_i, i \in [m]\}$ until there are only k connected components left
3: $C_0 = \{\nu^*_r, r \in [k]\} \leftarrow$ take the mean of the points in each connected component $S_r, r \in [k]$

Theorem 3. *Assume (X, T_*) satisfies $d^*_{rs}(f)$-center separability with $f > \max\{\frac{1}{\alpha}, 32\}$. If we cluster X using Algorithm 1, where we use Algorithm 2 for the seeding step, and execute Lloyd's update until convergence to refine the solution, then with probability at least $1 - m\exp(-2(\frac{f}{4} - 1)^2 w^2_{\min}) - k\exp(-mp_{\min})$ all but $\frac{81}{8f^2}$ fraction of the points will be correctly classified with respect to T_*.*

Remark: The success probability here doesn't approach 1 as $m \to \infty$. Instead, m should be carefully chosen to be neither too large nor too small. For example, when $w_{\min}$ and $p_{\min}$ are bounded away from zero, and $f = \Omega(\sqrt{k})$, then choosing m to be $\Theta(k)$ will ensure a significant success probability (in this case, if $k, n \to \infty$, the success probability does approach 1 as $m \to \infty$). Theorem 3 follows directly from Theorem 1 and Theorem 4.

Theorem 4. *Assume (X, T_*) satisfies $d^*_{rs}(f)$-center separability with $f > \frac{1}{\alpha}$. If we obtain seeds $\{\nu^*_r, r \in [k]\}$ by applying Algorithm 2 to X. Then $\forall \mu_r, \exists \nu^*_r$ s.t. $\|\mu_r - \nu^*_r\| \le \frac{\sqrt{f}}{2}\sqrt{\frac{\phi^r_*}{n_r}}$ with probability at least $1 - m\exp(-2(\frac{f}{4} - 1)^2 w^2_{\min}) - k\exp(-mp_{\min})$.*

Proof idea: we first show that Single-Linkage has the property of correctly identifying k connected components of a graph G, provided for all edges of G, all intra-cluster edges are shorter than any inter-cluster edges. Then we show that the edge set E induced by sample $\{\nu_i\}$ satisfies the condition with significant probability, where each connected component $\{\nu_{r(j)}\}$ corresponds to samples from the ground-truth cluster T_r. Finally, taking the mean of points in each connected component gives the desired result (the proofs of Theorem 4 and its lemmas can be found in the Appendix).

Consider a complete graph $G = (V, E)$. Any k-clustering $\{V_1, \dots, V_k\}$ of the vertex set induces a bipartition of the edge set $E = E_{in} \cup E_{out}$ s.t. $e = (v_i, v_j) \in E_{in}$ if $v_i, v_j \in V_r$ for some $r \in [k]$, and $e = (v_i, v_j) \in E_{out}$ if $v_i \in V_r, v_j \in V_s, r \neq s$. Let $w(e) := \|v_i - v_j\|$, the correctness of Single-Linkage on instances described above is formally stated below.

Lemma 6. *Assume a complete graph $G = (V, E)$ admits a k-clustering $\{V_1^*, \ldots, V_k^*\}$ of V with the induced edge bi-partition E_{in}^*, E_{out}^* such that $\forall e_1 \in E_{in}^*, \forall e_2 \in E_{out}^*$, we have $w(e_1) < w(e_2)$. Then running Single-Linkage on $G_0 := (V, \emptyset)$ until k-components left, results in a graph G_{SL} such that for each connected component, r, of G_{SL} the vertex set, V_{SL}^r, corresponds to exactly one cluster V_r^* of V.*

Then Lemma 7 and 8 together imply that with significant probability, the ground-truth clustering induces a non-degenerate k-clustering of $\{\nu_i, i \in [m]\}$, represented as $\{\{\nu_i\} \cap T_r, r \in [k]\}$, which satisfies the property required by Lemma 6.

Lemma 7. *Let $T_{\pi(i)}$ denote the grounth-truth cluster a sample ν_i belongs to. Define two events: $A := \{\forall \nu_i, i \in [m], \|\nu_i - \mu_{\pi(i)}\| \leq \frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^{\pi(i)}}{n_{\pi(i)}}}\}$, and $B := \{\forall T_r, r \in [k], T_r \cap \{\nu_i, i \in [m]\} \neq \emptyset\}$. Then $Pr(A \cap B) \geq 1 - m\exp(-2(\frac{f}{4} - 1)^2 w_{\min}^2) - k\exp(-mp_{\min})$.*

Lemma 8. *For any $\nu_i \in \{\nu_i, i \in [m]\}$, let $T_{\pi(i)}$ denote the ground-truth cluster it belongs to. If $\forall \nu_i \in \{\nu_i, i \in [m]\}$, $\|\nu_i - \mu_{\pi(i)}\|^2 \leq \frac{f}{4}\frac{\phi_*^{\pi(i)}}{n_{\pi(i)}}$ and $f > \frac{1}{\alpha}$. Then for any $i, j \in [m]$ s.t. $\pi(i) = \pi(j)$, and for any $p, q \in [m]$ s.t. $\pi(p) \neq \pi(q)$, $\|\nu_i - \nu_j\| < \|\nu_p - \nu_q\|$.*

4.1 Approximation guarantee for the k-means problem

Additionally, we show that Algorithm 2 achieves constant k-means approximation under an assumption weaker than Definition 1.

Definition 3 ($d_{rs}^*(f)$-weak center separability). *A dataset-solution pair (X, T_*) satisfies $d_{rs}^*(f)$-weak center separability if $\forall r \in [k], s \neq r$, $\|\mu_r - \mu_s\| \geq d_{rs}^*$, where $d_{rs}^* = f(\sqrt{\phi_1 + \phi_2})(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}})$, where ϕ_1 and ϕ_2 are the k-means cost of the largest and second largest (w.r.t. k-means cost) clusters in an optimal k-means solution, i.e., $\phi_1 := \max_r \phi_*^r, \phi_2 := \max_{s, s\neq 1} \phi_*^s$.*

Theorem 5. *Assume T_* is an optimal k-means solution with respect to X, which satisfies $d_{rs}^*(f)$-weak center separability with $f > \max\{\frac{1}{\alpha}, 16\}$. If we cluster X using Algorithm 2, then with probability at least $1 - m\exp(-2(\frac{f}{4} - 1)^2 w_{\min}^2) - k\exp(-mp_{\min})$, the final solution is a 4-approximation to the k-means objective.*

The proof, similar to Theorem 3.2 of [5], utilizes Lemma 2 and Theorem 4, and is included in the Appendix. In Theorem 5 we have fixed f, m as constants to get a constant approximation guarantee with probability depending on f, m. If we instead fix any approximation factor $1 + \epsilon > 1$, and failure probability $\delta > 0$, then by allowing f, m to depend on these two parameters, we can achieve $1 + \epsilon$-approximation guarantee with probability at least $1 - \delta$, as shown in the corollary below.

Corollary 1. *Assume the conditions in Theorem 5 hold. For any $\delta > 0, \epsilon > 0$, if $f = \Omega(\sqrt{\log(\frac{\frac{1}{\delta}\log\frac{k}{\delta}}{p_{\min}})} + \frac{1}{\epsilon^2})$, and choosing $\frac{\log\frac{2k}{\delta}}{p_{\min}} < m < \frac{\delta}{2}\exp\{2(\frac{f}{4} - 1)^2 w_{\min}^2\}$, then Algorithm 2 has $(1 + \epsilon)$-approximation guarantee with respect to the optimal k-means objective with probability at least $1 - \delta$.*

Therefore, it suffices to have $m = \Omega(\frac{\log\frac{k}{\delta}}{p_{\min}})$ (this is at least $\Omega(k\log\frac{k}{\delta})$). Since the algorithm is only run on a sample of size m, as long as $p_{\min} = \Omega(\exp(-k))$, the runtime of Algorithm 2 has polynomial dependence on k.

5 Local convergence and stopping criteria

Assuming Lloyd's algorithm is executed until convergence, we analyzed its global convergence on well-clusterable datasets. In this section, we turn to study its local convergence and stopping criteria.

In practice, early stopping is commonly used to prevent Lloyd's algorithm from running too long. Four criteria are frequently used [27]: maximal number of iterations, between-iteration centroid movement, between-iteration cluster re-assignment, and change of between-iteration k-means cost. The first criterion is usually set arbitrarily by the user according to the upper limit of time she is willing to spend, and used as a backup for other criteria. Of the remaining three, the centroid-movement based criterion has an advantage in the large-scale setting, where the computation of cluster reassignment or k-means cost (or its change) is impractical since they rely on the property of the entire dataset. To our knowledge, no theoretical analysis exists for their performance.

Adapting Theorem 1 to local convergence, we give justification of a criterion that is a modification of the centroid movement criterion. Consider an intermediate solution $C_{t_0-1} := \{c_r, r \in [k]\}$ of Algorithm 1. By Lloyd's update rule, they are the means of clusters in $S_{t_0-1} := \{S_{t_0-1}^r, r \in [k]\}$, which is not necessarily a local or global optimum, i.e., $S_{t_0} \neq S_{t_0-1}$. Let $C_t := \{\nu_r, r \in [k]\}, \forall t \geq t_0$, let $\delta_r^t := \|\nu_r - c_r\|$, and $\delta^t = \max_r \delta_r^t$ (δ_t is the between-iteration centroid movement), the following holds.

Corollary 2. *Assume a dataset-solution pair (X, S_{t_0-1}) satisfies $d_{rs}^*(f)$-center separability with $f > 32$. If $\delta^{t_0} < \frac{1}{8}\min_{r\neq s, s\in[k]}\|c_r - c_s\|$, then $\forall T \geq t_0$, $\forall r, \delta_r^T < \frac{128}{9f}\sqrt{\frac{\phi_{t_0-1}}{n_{t_0-1}^r}}$, where, $\phi_{t_0-1} := \phi(C_{t_0-1}, S_{t_0-1})$,*

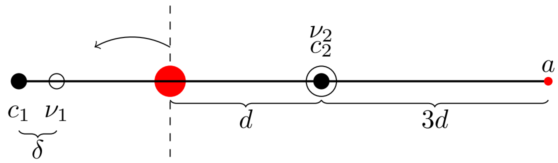

Figure 1: Two clusters in a bad instance when center separation is small

and $n^r_{t_0-1} := |S^r_{t_0-1}|$.

Corollary 2 also suggests we can use $\frac{\delta^t}{\max_{r\neq s}\|c_r-c_s\|} < \frac{1}{8}$ as a certificate of local convergence, given (X, S_{t_0-1}) is d^*_{rs}-center separable.

This leads us to ask: do the utility of stopping criteria really depend on clusterability of the dataset? Our next result provides evidence that clusterability does matter. We construct a bad instance, one that fails our clusterability assumption, such that for arbitrarily small δ^{t_0}, after one Lloyd's iteration, the updated centroids can have a large "jump" in the solution space.

Proposition 1. *For any* $\delta > 0$*, there exists* (X, S_{t_0-1}) *with* $\delta \ll \min_{r\neq s}\|c_r - c_s\| = \Theta(\sqrt{\frac{\phi^s_{t_0-1}}{n^s_{t_0-1}}} + \sqrt{\frac{\phi^r_{t_0-1}}{n^r_{t_0-1}}})$ *such that* $\delta^{t_0} = \delta$ *but* $\delta^{t_0+1} = \Omega(\sqrt{\frac{\phi^s_{t_0-1}}{n^s_{t_0-1}}} + \sqrt{\frac{\phi^r_{t_0-1}}{n^r_{t_0-1}}}) \gg \delta^{t_0}$.

Proof. Consider a solution S_{t_0-1} that contains two clusters with mean c_1, c_2 s.t. $\delta \ll 2d = \|c_1 - c_2\| = \min_{r\neq s}\|c_r - c_s\|$ (Figure 1). Further assume these two clusters are sufficiently far away from the rest of the clusters. Suppose $S^2_{t_0-1}$ has 4 points; three of them are at distance d to c_2 and one of them (point a) are at distance $3d$ to c_2. Assume $\sqrt{\frac{\phi^1_{t_0-1}}{n^1_{t_0-1}}} < \sqrt{\frac{\phi^2_{t_0-1}}{n^2_{t_0-1}}} = \sqrt{3}d$. Obviously, $\|c_1 - c_2\| = \Theta(\sqrt{\frac{\phi^1_{t_0-1}}{n^1_{t_0-1}}} + \sqrt{\frac{\phi^2_{t_0-1}}{n^2_{t_0-1}}})$. Suppose after one Lloyd's update, for $\nu_1, \nu_2 \in C_{t_0}$, ν_1 moved δ towards c_2 while $\nu_2 = c_2$. Then in S_{t_0} all three points originally assigned to c_2 will be assigned to c_1 and in the updated $\nu_2 \in C_{t_0}$ will move to a, thus $\delta_2^{t_0+1} = 3d$. $\square$

Proposition 1 gives an example where a criterion based only on thresholding centroid movement may stop the algorithm too early and miss the jump, which corresponds to a significant shift in the clustering configuration. Is this just an artificial case that rarely occurs in practice? We turn to empirical study to find out.

5.1 Empirical performance of different stopping criteria

We compare the performance of common stopping criteria introduced previously, and test how they are affected by clusterability.

Experimental setup We generate synthetic datasets to control the degree of clusterability. Starting from random seedings on synthetic datasets, we recorded after how many iterations the following three criteria stop the algorithm [3].

- TH: stops at t when $\delta^t < \frac{1}{8}\min_{r\neq s}\|c_r - c_s\|$ as suggested by Corollary 2.
- $RA(\eta)$: stops when the fraction of points reassigned to another cluster between two consecutive iterations falls below η, where $\eta \in (0, 1)$.
- $KM(\delta)$: stops when the change of k-means cost falls below δ times previous cost, with $\delta \in (0, 1)$.

To understand the utility of stopping criteria, we measured the k-means cost, ϕ_t, its change, $\Delta\phi_t := \phi_t - \phi_{t-1}$, and the between-iteration centroid movement, δ^t, at every iteration. An ideal stopping criterion should stop the algorithm the moment k-means cost enters a stable, plateau-like stage, which should be free from significant centroid movement or change of cost. Lloyd's algorithm is repeated 10 times for each experiment and we report the averages of measured quantities. Lacking guidance in the literature on setting parameters for RA and KM, for each run of the algorithm we randomly draw from $(0, 0.3)$ to set η and δ, separately.

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 (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. More details on the generation of our synthetic data is included in the Appendix.

Results and interpretation Figure 2 shows our measured quantities for the first 20 iterations of

[3] The solution we start from in our experiments should not be interpreted as seeding but as an intermediate solution which Lloyd's algorithm may encounter in practice.

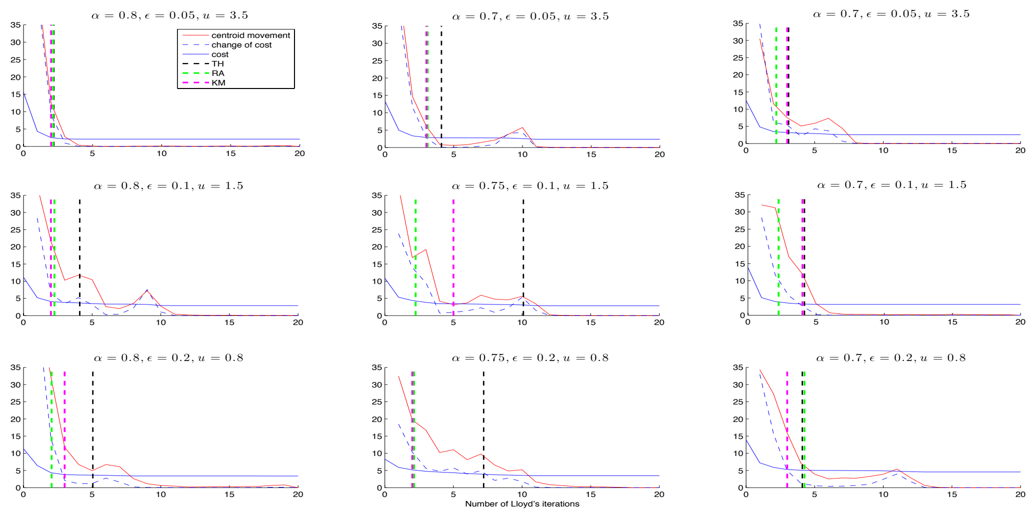

Figure 2: In each subfigure, we plot δ^t, $\Delta\phi_t$, and ϕ_t (scaled differently for convenient display) versus t; the vertical bars marks the stopped iteration according to different stopping criteria. The subfigures vary by clusterability of the dataset, parameterized by α, ϵ, u; clusterability decreases from top to bottom and from left to right.

Lloyd's algorithm in 9 datasets with varying clusterability. To interpret the performance of stopping criteria, let us first understand the plots for ϕ_t (blue), $\Delta\phi_t$ (dashed blue), and δ^t (red). As expected, ϕ_t monotonically decreases with t since Lloyd's algorithm decreases the k-means objective at every iteration. However, the latter two are usually not monotone, and the general trend is that as the dataset becomes more clusterable, they become smoother. We observe that the red plots resemble those of the dashed blue. This means significant centroid movements, which correspond to sudden shifts in the clustering configuration, usually lead to solutions with large drops in the k-means cost. The presence of large spikes in some of the plots suggests that bad cases, where $\delta^t \ll \delta^{t+1}$ such as the one in Proposition 1, do indeed arise in practice (the jumps were even more pronounced in individual runs, before averaging).

Qualitatively, a good criterion should stop the algorithm at the iteration corresponding to the last significant spike in red or dashed blue, or at an iteration t_0 where δ^{t_0} is sufficiently small and $\delta^{t_0} \geq \delta^T, \forall T \geq t_0$ as in Corollary 2. A stop too early will miss the potential drastic shift in the clustering solution while a stop too late wastes computation. From our experiments, we observed that, 1) No stopping criterion consistently satisfies the desired property; clusterability (as parameterized by ϵ, α, u) of the dataset heavily influences the performance of all criteria; other parameters, such as data size and dimension, did not have a significant influence on the performance of criteria, for a fixed level of clusterability. 2) When the dataset is more clusterable, all stopping criteria were able to stop the algorithm at a good point and their choices of stopping point are similar, e.g., the upper-left plot in Figure 2. 3) As the dataset becomes less clusterable, we saw noticeable differences in the stopping criteria; TH (black) seems to stop at a better point more often than RA (green) or KM (magenta), e.g., it catches the last spike in the middle plot in Figure 2.

6 Future work

Future exploration into clusterability assumptions is needed, as current assumptions [5, 4, 21, 28], as well as ours, are still rather strong. Meanwhile, although stochastic Lloyd's algorithm and variants [9, 30] are widely used for large-scale clustering (e.g., it is implemented in popular packages such as scikit-learn [29]), there is little theoretical understanding of it. Building on our understanding of the batch Lloyd's algorithm, it may be promising to combine techniques in stochastic optimization to analyze its stochastic variants. Finally, our empirical findings provide evidence that existing stopping criteria may be insufficient when working with less clusterable data. Given the importance of stopping criteria in stochastic Lloyd's variants, it will be interesting to investigate whether a carefully designed early stopping strategy can work well with all solutions.

Acknowledgements

We thank all our anonymous reviewers for their constructive feedback on improving our initial submission.

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.