Scalable constant k-means approximation via heuristics on well-clusterable data

Cheng Tang
tangch@gwu.edu
Department of Computer Science
George Washington University
Washington, DC, 22202

Claire Monteleoni
cmontel@gwu.edu
Department of Computer Science
George Washington University
Washington, DC, 22202

Abstract

We present a simple heuristic clustering procedure, with running time independent of the data size, that combines random sampling with Single-Linkage (Kruskal's algorithm), and show that with sufficient probability, it has a constant approximation guarantee with respect to the optimal k-means cost, provided an optimal solution satisfies a center-separability assumption. As the separation increases, it has better performance: fix any $\epsilon, \delta > 0$, if the center separation is sufficiently large, it has a $(1+\epsilon)$-approximation guarantee with probability at least $1-\delta$.

1 Introduction

While there is a rich body of literature on approximation algorithms for the k-means clustering problem [16, 10, 12, 8], less work has focused on proving guarantees for practically used schemes, e.g., Lloyd's algorithm [15] and linkage-based algorithms [7]. Ostrovsky et al. [17] first showed that when seeded with k-means++ [1], a Lloyd-like algorithm efficiently finds a $(1+\epsilon)$-approximation to the k-means objective (i.e., a Polynomial Time Approximation Scheme, PTAS) with high probability on well-clusterable instances. With a weaker clusterability assumption, Kumar and Kannan [11] showed that the *k-SVD + constant k-means approximation + Lloyd's update* scheme is a PTAS for the k-means clustering problem. Subsequent analysis [4] proposed a center-separability assumption as a simplification of [11], under which they showed that after projecting data to the subspace obtained by k-SVD, any constant k-means approximation is a PTAS, provided the center separation is sufficiently large (Sec. 3, [4]). A drawback of [11, 4] is that the required k-SVD step limits the applicability of their clustering scheme to $d > k$. The performance of linkage-based algorithms for center-based clustering, including k-means, on well-clusterable data were investigated by [3, 5], where the linkage algorithms are used to find a hierarchical clustering and some smart pruning is needed for finding the final k-clustering.

We show that a simple heuristic, one that combines random sampling with Single-Linkage (the latter terminates when k-components are left, eliminating the need for pruning), is a PTAS for the k-means problem with high probability when the underlying data satisfies a clusterability assumption that is comparable to those in [17, 11, 4, 2]. Yet, its running time is independent of the data size while, to our knowledge, this is not the case for most algorithms with such strong approximation guarantees. We thus demonstrate a positive case of computational gain by exploiting the structure of easy data.

1.1 Preliminaries

The input of our clustering problem is 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) non-degenerate[1] optimal k-means clustering

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

$T_* = \{T_s, s \in [k]\}$, which in addition satisfies $d^*_{rs}(f)$-weak center separability, defined below. Let $n_s := |T_s|, \forall s \in [k]$, and let $n_{\min} := \min_{s\in[k]} n_s$ and $n_{\max} := \max_{s\in[k]} n_s$.

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 $\{x \in \mathbb{R}^d, \|x - c_r\| < \|x - c_s\|, \forall s \neq r\}$.

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 optimal k-means cost.

Characterization of (X, T_*) Three properties of (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_* ($0 < \alpha \leq 1$ always holds; $\alpha = 1$ when the ground-truth is perfectly balanced). We let $w_r := \frac{(\phi^r_*/n_r)}{\max_{x\in T_r} \|x-\mu_r\|^2}$ characterize the ratio between average and maximal "spread" of cluster T_r, and we let $w_{\min} := \min_{r\in[k]} w_r$. Note $p_{\min} \leq \frac{1}{k}$, so it should not be treated as a constant as k increases; α and $w_{\min}$, on the other hand, do not necessarily grow with k (nor n, d), and we treat them as constants.

Our clusterability assumption We present two assumptions. The second is stronger (but within a factor of $\sqrt{k}$) than the first.

Definition 1 ($d^*_{rs}(f)$-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., $\phi_1 := \max_r \phi^r_*, \phi_2 := \max_{s,s\neq 1} \phi^s_*$.*

This clusterability assumption is reminiscent of the mean separation assumption in the earlier work on learning mixtures of Gaussians [9], where the means of different components are required to be at least $\Omega(\sigma_{max})$ apart, with σ_{max} being the largest deviation of a single component. Since most of the mass of a Gaussian component is within one standard deviation of their mean, σ_{max} provides a rough bound of "cluster width" of each component. Thus, mean separation implies that the within-cluster distance is on average smaller than the between-cluster distance. Here, we do not have any probabilistic assumptions, however, μ_r, μ_s are the empirical mean of their respective clusters. Also note that $\sqrt{\frac{\phi^r_*}{n_r}}$ is the empirical deviation for cluster T_r. However, instead of requiring the centers to be at least $\Omega(\sqrt{\frac{\phi_1}{n_1}})$ apart, we need a more strict condition $\Omega(\sqrt{\frac{\phi_1}{n_r}}), \forall r$, due to the technical difficulties that arise by not having measure concentration. When analyzing the performance of Algorithm 1 together with Lloyd's algorithm [15], we need a stronger assumption as below, which depends on the global k-means cost.

Definition 2 ($(d^*_{rs}(f)$-center separability)**.** *A dataset-solution pair (X, T_*) satisfies $d^*_{rs}(f)$-center separability if we redefine $d^*_{rs}(f)$ above as $d^*_{rs}(f) := f\sqrt{\phi_*}(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}})$.*

Although stronger than weak center separability, $(d^*_{rs}(f)$-center separability is implied by the assumption in [17]. Furthermore, in the case $d < k$ and $f = O(\sqrt{k})$, it is implied by the assumption in [11]; when $f = O(1)$, it is similar to the assumption in [4].

2 Main results

In large-scale applications, such as computer vision, clustering algorithms are often run on a random sample of the entire data (i.e., a subset of data sampled uniformly at random) [6, 13, 14]. Our

Algorithm 1 Heuristic clustering

Input: X, m, k
Output: $\{S_1, \ldots, S_k\}$
1: $\{\nu_i, i \in [m]\} \leftarrow$ sample m points from X (i.i.d.) uniformly at random with replacement
2: $\{\tilde{S}_1, \ldots, \tilde{S}_k\} \leftarrow$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 $\tilde{S}_r, r \in [k]$
4: $X = S_1 \cup \cdots \cup S_k \leftarrow k$-partition X according to the Voronoi region induced by C_0

main results provide an example where such an heuristic, as described in Algorithm 1, has provable guarantee. In the context of k-means clustering, this leads us to the conclusion that Algorithm 1 is a constant approximation k-means algorithm with high probability, whose performance can be further improved by Lloyd's algorithm. It also suggests that if the dataset has a clusterable structure, the sample size could be independent of the data size, a desirable property for dealing with massive datasets.

Theorem 1. *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 1, then with probability at least $1-m\exp(-2(\frac{f}{4}-1)^2 w^2_{\min})-k\exp(-mp_{\min})$, the final solution is a 4-approximation to the k-means objective.*

The proof, similar to Theorem 3.2 of [4], follows directly from Theorem 3 and Lemma 1.

Proof. Consider each cluster S_r in the final solution. Its k-means cost, by definition, is $\phi(\{m(S_r)\}, S_r) \le \phi(\{\mu_r\}, S_r) = \phi(\{\mu_r\}, S_r \cap T_r) + \phi(\{\mu_r\}, \cup_{s\neq r} S_r \cap T_s)$. By Theorem 3 and our assumption on center separation, $\gamma \le \frac{\sqrt{f}}{2f} < \frac{1}{4}$, we can apply Lemma 1 to get $\phi(\{\mu_r\}, \cup_{s\neq r} S_r \cap T_s) = \sum_{s\neq r}\sum_{x\in S_r\cap T_s}\|x-\mu_r\|^2 \le \sum_{s\neq r}\sum_{x\in S_r\cap T_s}\frac{1}{(1-4\gamma)^2}\|x-\mu_s\|^2$, by Lemma 1. Since $f > 16$, we get $\frac{1}{(1-4\gamma)^2} \le 4$. Summing over all $r \in [k]$, $\phi(\{S_r, r \in [k]\}) \le \sum_r \phi(\{\mu_r\}, S_r \cap T_r) + \sum_r \frac{1}{(1-4\gamma)^2}\sum_{s\neq r}\sum_{x\in S_r\cap T_s}\|x-\mu_s\|^2 \le 4(\sum_r \phi(\{\mu_r\}, S_r \cap T_r) + \sum_r\sum_{s\neq r}\sum_{x\in S_r\cap T_s}\|x-\mu_s\|^2) = 4\{\sum_r(\sum_{x\in S_r\cap T_r}\|x-\mu_r\|^2 + \sum_{s\neq r}\sum_{x\in S_r\cap T_s}\|x-\mu_s\|^2)\} = 4\{\sum_r\sum_{x\in S_r}\|x-C_*(x)\|^2\} = 4\phi_*$ (C_* is the set of optimal centroids). □

Intuitively, we want neither under-sampling, which may fail to cover some optimal clusters, nor over-sampling, which may include outliers. The intuition translates into the success probability of Algorithm 1: m should be carefully chosen to be neither too large nor too small.

In Theorem 1 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 1 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 $\frac{\log\frac{2k}{\delta}}{p_{\min}} < m < \frac{\delta}{2}\exp\{2(\frac{f}{4}-1)^2 w^2_{\min}\}$, then Algorithm 1 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 1 has polynomial dependence on k. The quadratic dependence of our assumption on $\frac{1}{\epsilon}$ can be relaxed to $\frac{1}{\sqrt{\epsilon}}$, if we run Lloyd's algorithm to refine the clustering and use $d^*_{rs}(f)$-center separability instead.

Theorem 2. *Assume T_* is an optimal k-means solution with respect to X, which satisfies $d^*_{rs}(f)$-center separability. And for any $\delta > 0, \epsilon > 0$, if $f = \Omega(\sqrt{\log(\frac{\frac{1}{\delta}\log\frac{1}{\delta}}{p_{\min}})} + \sqrt{\frac{1}{\epsilon}})$, and choosing $\frac{\log\frac{2k}{\delta}}{p_{\min}} < m < \frac{\delta}{2}\exp\{2(\frac{f}{4}-1)^2 w^2_{\min}\}$, then if we run Lloyd's algorithm with seeds $\{\nu_r^*, r \in [k]\}$*

obtained from Algorithm 1, the converged Lloyd's solution has a $(1+\epsilon)$-approximation guarantee with respect to the optimal k-means objective with probability at least $1-\delta$.

Due to space limits we removed some proofs[2].

2.1 Analysis

Lemma 1 shows when the centroids in C_0 is sufficiently close to those in an optimal solution (guaranteed by Theorem 3), the mis-clustered points of each cluster S_r must be "outliers" with respect to its optimal cluster T_s, for some $s \neq r$. Consequently, assigning them to T_r does not increase the cost too much.

Lemma 1. *If $\gamma := \max_{r,s\neq r} \frac{\|\nu_s^* - \mu_s\|}{\|\mu_r - \mu_s\|} < \frac{1}{4}$, then $\forall r \in [k], \forall x \in V(\nu_r^*)$, $\|x - \mu_r\| \leq \frac{1}{1-4\gamma}\|x - \mu_s\|$*

Our main result regarding Algorithm 1 is presented below.

Theorem 3. *Assume T_* is an optimal k-means solution with respect to X, which satisfies $d_{rs}^*(f)$-weak center separability with $f > \max\{\frac{1}{\alpha}, 4\}$. If we cluster X using Algorithm 1, then $\forall \mu_r, \exists \nu_r^*$ s.t. $\|\mu_r - \nu_r^*\| \leq \frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^r}{n_r}}$ with probability at least $1 - m\exp(-2(\frac{f}{4}-1)^2 w_{\min}^2) - k\exp(-mp_{\min})$.*

Proof outline To prove the theorem, we first show that Single-Linkage as used in Algorithm 1 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 (Lemma 2). 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 optimal cluster T_r (Lemma 3 and 4). Finally, taking the mean of points in each connected component gives the desired result.

Consider a complete graph $G = (V, E)$. Any k-clustering $\{V_1, \ldots, V_k\}$ of the vertex set induces a bi-partition 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 2. *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 $w(e_1) < w(e_2)$ (the edge weights are just the Euclidean distances between vertices). 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.*

Now we show that with significant probability, the ground-truth clustering induces a non-degenerate k-clustering of $\{\nu_i, i \in [m]\}$, $\{\{\nu_i\} \cap T_r, r \in [k]\}$, which satisfies the property required by Lemma 2, which follows by combining Lemma 3 and 4.

Lemma 3. *Let $T_{\pi(i)}$ denote the optimal 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 4. *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 $f > \max\{\frac{1}{\alpha}, 4\}$. Then for any $i, j \in [m]$ s.t. $\pi(i) = \pi(j)$, and for any $p, q \in [m]$ s.t. $\pi(p) \neq \pi(q)$, $\|\nu_i - \nu_j\| < \|\nu_p - \nu_q\|$.*

Finally, combining the seeding guarantee from Lemma 3 and 4 with the property of Single-Linkage in Lemma 2 completes the proof of Theorem 3.

References

[1] 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.

[2] They can be found at **http://faculty.cs.gwu.edu/∼cmontel/nips15workshop1_supp.pdf**

[2] 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.

[3] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(1):49–54, 2012.

[4] 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.

[5] Maria Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. In *Automata, Languages, and Programming*, pages 63–74. Springer Berlin Heidelberg, 2012.

[6] 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.

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

[8] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300. ACM, 2004.

[9] Ravindran Kannan and Santosh Vempala. Spectral algorithms. *Found. Trends Theor. Comput. Sci.*, 4:157–288, March 2009.

[10] 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.

[11] 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.

[12] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\varepsilon)$-approximation algorithm for geometric k-means clustering in any dimensions. *Proceedings-Annual Symposium on Foundations of Computer Science*, pages 454–462, 2004.

[13] 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.

[14] 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.

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

[16] Jirí Matousek. On approximate geometric k-clustering. *Discrete & Computational Geometry*, 24(1):61–84, 2000.

[17] 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.