The Effectiveness of Lloyd-Type Methods for the *k*-Means Problem

RAFAIL OSTROVSKY, University of California, Los Angeles
YUVAL RABANI, The Hebrew University of Jerusalem
LEONARD J. SCHULMAN, California Institute of Technology
CHAITANYA SWAMY, University of Waterloo

We investigate variants of Lloyd's heuristic for clustering high-dimensional data in an attempt to explain its popularity (a half century after its introduction) among practitioners, and in order to suggest improvements in its application. We propose and justify a *clusterability* criterion for data sets. We present variants of Lloyd's heuristic that quickly lead to provably near-optimal clustering solutions when applied to well-clusterable instances. This is the first performance guarantee for a variant of Lloyd's heuristic. The provision of a guarantee on output quality does not come at the expense of speed: some of our algorithms are candidates for being *faster in practice* than currently used variants of Lloyd's method. In addition, our other algorithms are faster on well-clusterable instances than recently proposed approximation algorithms, while maintaining similar guarantees on clustering quality. Our main algorithmic contribution is a novel probabilistic seeding process for the starting configuration of a Lloyd-type iteration.

Categories and Subject Descriptors: F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems—*computations on discrete structures*; G.3 [**Probability and Statistics**]: *Probabilistic algorithms (including Monte Carlo)*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Randomized algorithms, approximation algorithms

A preliminary version of this article appeared in *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006.
R. Ostrovsky was supported in part by NSF grants 0830803, 09165174, 1065276, 1118126, and 1136174; US-Israel BSF grant 2008411; OKAWA Foundation Research Award; IBM Faculty Research Award; Xerox Faculty Research Award; B. John Garrick Foundation Research Award, Teradata Research Award; and Lockheed-Martin Corporation Research Award; and the Defense Advanced Research Projects Agency through the U.S. Office of Naval Research under Contract N00014-11-1-0392. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Y. Rabani was supported by ISF grant 52/03 and BSF grant 2002282. L. J. Schulman was supported in part by NSF CCF-0515342, NSF CCF-1038578, NSA H98230-06-1-0074, and NSF ITR CCR-0326554. The research of C. Swamy was supported partially by NSERC grant 327620-09 and an Ontario Early Researcher Award.
Part of this work was done while Y. Rabani was visiting UCLA and Caltech.
The work of C. Swamy was done while the author was postdoctoral scholar at Caltech.
Authors' addresses: R. Ostrovsky, Department of Computer Science and Department of Mathematics, UCLA, Box 951596 3732D BH, Los Angeles, CA 90095-1596; email: rafail@cs.ucla.edu; Y. Rabani, The Rachel and Selim Benin School of Computer Science and Engineering and the Center of Excellence on Algorithms, The Hebrew University of Jerusalem, Jerusalem 91904, Israel; email: yrabani@cs.huji.ac.il; L. J. Schulman, Caltech, 1200 East California Blvd., Pasadena, CA 91125; email: schulman@caltech.edu; C. Swamy, Department of Combinatorics and Optimization, Faculty of Mathematics, MC 4033, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada; email: cswamy@math.uwaterloo.ca.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2012 ACM 0004-5411/2012/12-ART28 $15.00
DOI 10.1145/2395116.2395117 http://doi.acm.org/10.1145/2395116.2395117

ACM Reference Format:
Ostrovsky, R., Rabani, Y., Schulman, L., and Swamy, C. 2012. The effectiveness of Lloyd-type methods for the k-means problem. J. ACM 59, 6, Article 28 (December 2012), 22 pages.
DOI = 10.1145/2395116.2395117 http://doi.acm.org/10.1145/2395116.2395117

1. INTRODUCTION

Overview. There is presently a wide and unsatisfactory gap between the practical and theoretical clustering literatures. For decades, practitioners have been using heuristics of great speed but uncertain merit; the latter should not be surprising since the problem is *NP*-hard in almost any formulation. However, in the last few years, algorithms researchers have made considerable innovations, and even obtained polynomial-time approximation schemes (PTAS's) for some of the most popular clustering formulations. Yet these contributions have not had a noticeable impact on practice. Practitioners instead continue to use a variety of heuristics (Lloyd, EM, agglomerative methods, etc.) that have no known performance guarantees.

There are two ways to approach this disjuncture. The most obvious is to continue developing new techniques until they are so good—down to the implementations—that they displace entrenched methods. The other is to look toward popular heuristics and ask whether there are reasons that justify their extensive use, but elude the standard theoretical criteria; and in addition, whether theoretical scrutiny suggests improvements in their application. This is the approach we take in this article.

As in other prominent cases [Spielman and Teng 2001; Papadimitriou et al. 2000], such an analysis typically involves some abandonment of the worst-case inputs criterion. (In fact, part of the challenge is to identify simple conditions on the input, that allow one to prove a performance guarantee of wide applicability.) Our starting point is the notion that (as discussed in Schulman [2000]) one should be concerned with k-clustering data that possesses a meaningful k-clustering. What does it mean for the data to have a meaningful k-clustering? Here are two examples of settings where one would intuitively not consider the data to possess a meaningful k-clustering. If nearly optimum cost can be achieved by two very different k-way partitions of the data then the identity of the optimal partition carries little meaning (e.g., if the data was generated by random sampling from a source, then the optimal cluster regions might shift drastically upon resampling). Alternatively, if a near-optimal k-clustering can be achieved by a partition into fewer than k clusters, then that smaller value of k should be used to cluster the data. If near-optimal k-clusterings are hard to find only when they provide ambiguous classification or marginal benefit (i.e., in the absence of a meaningful k-clustering), then such hardness should not be viewed as an acceptable obstacle to algorithm development. Instead, the performance criteria should be revised.

Specifically, we consider the *k-means* formulation of clustering: given a finite set $X \subseteq \mathbb{R}^d$, find k points ("centers") to minimize the sum over all points $x \in X$ of the squared distance between x and the center to which it is assigned. In an optimal solution, each center is assigned the data in its Voronoi region and is located at the center of mass of this data. Perhaps the most popular heuristic used for this problem is Lloyd's method, which consists of the following two phases: (a) "Seed" the process with some initial centers (the literature contains many competing suggestions of how to do this); (b) Iterate the following *Lloyd step* until the clustering is "good enough": cluster all the data in the Voronoi region of a center together, and then move the center to the centroid of its cluster.

Although Lloyd-style methods are widely used, to our knowledge, there is no known mathematical analysis that attempts to explain or predict the performance of these heuristics. In this paper, we take the first step in this direction. We show that if the data is well clusterable according to a certain "clusterability" or "separation" condition (that we introduce and discuss later), then various Lloyd-style methods do indeed perform well and return a provably near-optimal clustering. Our contributions are threefold:

(a) We introduce a separation condition and justify it as a reasonable abstraction of well-clusterability for the analysis of k-means clustering algorithms. Our condition is simple, and abstracts a notion of well clusterability alluded to earlier: letting $\Delta_k^2(X)$ denote the cost of an optimal k-means solution of input X, we say that X is *ϵ-separated for k-means* if $\Delta_k^2(X)/\Delta_{k-1}^2(X) \leq \epsilon^2$. (A similar condition for $k = 2$ was used for ℓ_2^2 edge-cost clustering in Schulman [2000].)

    Our motivation for proposing this condition is that a significant drop in the k-clustering cost is already used by practitioners as a diagnostic for choosing the value of k [Duda et al. 2000; Section 10.10]. Furthermore, we show that: (i) The data satisfies our separation condition if and only if it satisfies the other intuitive notion of well clusterability suggested earlier, namely that any two low-cost k-clusterings disagree on only a small fraction of the data; and (ii) The condition is robust under noisy (even adversarial) perturbation of the data. In Section 5, we prove rigorous versions of (i) and (ii).

(b) We present a novel and efficient sampling process for seeding Lloyd's method with initial centers, which allows us to prove the effectiveness of these methods.

(c) We demonstrate the effectiveness of (our variants of) the Lloyd heuristic under the separation condition. Specifically: (i) Our simplest variant uses only the new seeding procedure, requires a single Lloyd-type descent step, and achieves a constant-factor approximation in time linear in $|X|$. This algorithm has success probability exponentially small in k, but we show that (ii) a slightly more complicated seeding process based on our sampling procedure yields a constant-factor approximation guarantee with constant probability, again in linear time. Since only one run of seeding+descent is required in both algorithms, these are candidates for being faster in practice than currently used Lloyd variants, which are used with multiple re-seedings and many Lloyd steps per re-seeding. (iii) We also give a PTAS by combining our seeding process with a sampling procedure of Kumar et al. [2004], whose running time is linear in $|X|$ and exponential in k. This PTAS is significantly faster, and also simpler, than the PTAS of Kumar et al. [2004] (applying the separation condition to both algorithms; the latter does not run faster under the condition).

Literature and Problem Formulation. Let $X \subseteq \mathbb{R}^d$ be the given point set and $n = |X|$. In the k-means problem, the objective is to partition X into k clusters $\bar{X}_1, \ldots, \bar{X}_k$ and assign each point in every cluster $\bar{X}_i$ to a common *center* $\bar{c}_i \in \mathbb{R}^d$, so as to minimize the "k-means cost" $\sum_{i=1}^k \sum_{x \in \bar{X}_i} \|x - \bar{c}_i\|^2$, where $\|.\|$ denotes the ℓ_2 norm. We let $\Delta_k^2(X)$ denote the optimum k-means cost. Observe that given the centers $\bar{c}_1, \ldots, \bar{c}_k$, it is easy to determine the best clustering corresponding to these centers: cluster $\bar{X}_i$ simply consists of all points $x \in X$ for which $\bar{c}_i$ is the nearest center (breaking ties arbitrarily). Conversely given a clustering $\bar{X}_i, \ldots, \bar{X}_k$, the best centers corresponding to this clustering are obtained by setting $\bar{c}_i$ to be the center of mass (centroid) of cluster X_i, that is, setting $\bar{c}_i = \frac{1}{|\bar{X}_i|} \cdot \sum_{x \in \bar{X}_i} x$. It follows that both of these properties simultaneously hold in an optimal solution, that is, $\bar{c}_i$ is the centroid of cluster $\bar{X}_i$, and each point in $\bar{X}_i$ has $\bar{c}_i$ as its nearest center.

The problem of minimizing the k-means cost is one of the earliest and most intensively studied formulations of the clustering problem, both because of its mathematical elegance and because it bears closely on statistical estimation of mixture models of k point sources under spherically symmetric Gaussian noise. We briefly survey the most relevant literature here. The k-means problem seems to have been first considered by Steinhaus [1956]. A simple greedy iteration to minimize cost was suggested by Lloyd in 1957 [Lloyd 1982] (and less methodically in the same year by Cox [1957]; also apparently by psychologists between 1959–67 [Tryon and Bailey 1970]). This and similar iterative descent methods soon became the dominant approaches to the problem [Max 1960; MacQueen 1967; Dempster et al. 1977; Linde et al. 1980] (see also Gersho and Gray [1992], Gray and Neuhoff [1998], and Jain et al. [1999], and the references therein); they remain so today, and are still being improved [Alsabti et al. 1998; Pelleg and Moore 1999; Phillips 2002; Kanungo et al. 2002]. Lloyd's method (in any variant) converges only to local optima however, and is sensitive to the choice of the initial centers [Milligan 1980]. Consequently, a lot of research has been directed toward seeding methods that try to start off Lloyd's method with a good initial configuration [Forgey 1965; Kaufman and Rousseeuw 1990; Fisher 1996; Higgs et al. 1997; Snarey et al. 1997; Bradley and Fayyad 1998; Meila and Heckerman 1998; Pena et al. 1999].

Very few theoretical guarantees are known about Lloyd's method or its variants. The convergence rate of Lloyd's method has recently been investigated in Dasgupta [2003], Har-Peled and Sadri [2005], and Arthur and Vassilvitskii [2006], and in particular, Arthur and Vassilvitskii [2006] shows that Lloyd's method can require a superpolynomial number of iterations to converge. More recently, Vattani [2011] shows that Lloyd's method can take exponential time even in two dimensions. In contrast, Arthur et al. [2011] proved that Lloyd's algorithm has polynomial-smoothed complexity.

The only work (besides ours) on the performance guarantee of Lloyd's method is Arthur and Vassilvitskii [2007], which appeared after the publication of a conference version of this article. [Ostrovsky et al. 2006] but was performed independently of our work. They present a randomized seeding procedure that is quite similar to our seeding process (see Section 4.1.1) and show that Lloyd's method initialized with these seed centers is an $O(\log k)$-approximation algorithm for every data set. Observe that, in contrast, we obtain a much stronger guarantee, namely a constant approximation ratio, however under the assumption that the data is well separated. Aggarwal et al. [2009] and Ailon et al. [2009] independently refined and extended the analysis of Arthur and Vassilvitskii [2007] to obtain bicriteria and streaming algorithms for k-means respectively. Aggarwal et al. showed that sampling $O(k)$ centers using the seeding method in Arthur and Vassilvitskii [2007] yields a constant-factor approximation, while Ailon et al. [2009] showed that sampling $O(k \log k)$ centers via a slightly different randomized seeding method yields an $O(1)$-approximation and leverage this to obtain a streaming approximation algorithm.

The k-means problem is NP-hard even for $k = 2$ [Drineas et al. 2004]. Recently there has been substantial progress in developing approximation algorithms for this problem. Matoušek [2000] gave the first PTAS for this problem, with running time polynomial in n, for a fixed k and dimension. Subsequently a succession of algorithms have appeared [Ostrovsky and Rabani 2002; Badoiu et al. 2002; de la Vega et al. 2003; Effros and Schulman 2004a, 2004b; Har-Peled and Mazumdar 2004; Kumar et al. 2004] with varying runtime dependency on n, k and the dimension. The most recent of these is the algorithm of Kumar et al. [2004], which presents a linear-time PTAS for a fixed k. There are also various constant-factor approximation algorithms for the related *k-median* problem [Jain and Vazirani 2001; Charikar et al. 2002; Charikar and Guha 1999; Jain et al. 2003; Mettu and Plaxton 2004], which also yield approximation algorithms for k-means, and have running time polynomial in n, k and the dimension; recently Kanungo et al. [2004] adapted the k-median algorithm of [2004] to obtain a $(9 + \epsilon)$-approximation algorithm for k-means.

However, none of these methods match the simplicity and speed of the popular Lloyd's method. Researchers concerned with the runtime of Lloyd's method bemoan the need for n nearest-neighbor computations in each descent step [Kanungo et al. 2002] ! Interestingly, the last reference provides a data structure that provably speeds up the nearest-neighbor calculations of Lloyd descent steps, under the condition that the optimal clusters are well separated. (This is unrelated to providing performance guarantees for the outcome.) Their data structure may be used in any Lloyd-variant, including ours, and is well suited to the conditions under which we prove performance of our method; however, ironically, it may not be worthwhile to precompute their data structure since our method requires so few descent steps.

2. PRELIMINARIES

We use the following notation throughout. For a point set S, we use $\mathrm{ctr}(S)$ to denote the center of mass of S. Let partition $X_1 \cup \cdots \cup X_k = X$ be an optimal k-means clustering of the input X, and let $c_i = \mathrm{ctr}(X_i)$ and $c = \mathrm{ctr}(X)$. So $\Delta_k^2(X) = \sum_{i=1}^k \sum_{x \in X_i} \|x - c_i\|^2 = \sum_{i=1}^k \Delta_1^2(X_i)$. Let $n_i = |X_i|$, $n = |X|$, and $r_i^2 = \frac{\Delta_1^2(X_i)}{n_i}$, that is, r_i^2 is the "mean squared error" in cluster X_i. We sometimes refer to r_i as the radius of cluster X_i. Define $D_i = \min_{j \neq i} \|c_j - c_i\|$.

Definition 2.1. We say that X is *ϵ-separated* for k-means if $\Delta_k^2(X) \leq \epsilon^2 \Delta_{k-1}^2(X)$.

We assume throughout that X is ϵ-separated for k-means for a suitably small constant ϵ. The upper bound required on ϵ depends on which seeding procedure we use (and we mention these bounds as we go along), but certainly $\epsilon \leq 6 \cdot 10^{-7}$ suffices for all our analyses to go through. (This upper bound on ϵ is rather small, but we emphasize that we have not attempted to optimize any constants and have focused instead on simplicity of presentation.) We use the following basic lemmas quite frequently.

LEMMA 2.2. *For every x, $\sum_{y \in X} \|x - y\|^2 = \Delta_1^2(X) + n\|x - c\|^2$. Hence $\sum_{\{x,y\} \subseteq X} \|x - y\|^2 = n\Delta_1^2(X)$.*

LEMMA 2.3. *Consider any set $S \subseteq \mathbb{R}^d$ and any partition $S_1 \cup S_2$ of S with $S_1 \neq \emptyset$. Let s, s_1, s_2 denote respectively $\mathrm{ctr}(S), \mathrm{ctr}(S_1), \mathrm{ctr}(S_2)$. Then, (i) $\Delta_1^2(S) = \Delta_1^2(S_1) + \Delta_1^2(S_2) + \frac{|S_1||S_2|}{|S|}\|s_1 - s_2\|^2$, and (ii) $\|s_1 - s\|^2 \leq \frac{\Delta_1^2(S)}{|S|} \cdot \frac{|S_2|}{|S_1|}$.*

PROOF. Let $a = |S_1|$ and $b = |S_2| = |S| - |S_1|$. We have

$$\begin{aligned}
\Delta_1^2(S) &= \sum_{x \in S_1} \|x - s\|^2 + \sum_{x \in S_2} \|x - s\|^2 \\
&= \left(\Delta_1^2(S_1) + a\|s_1 - s\|^2\right) + \left(\Delta_1^2(S_2) + b\|s_2 - s\|^2\right) \qquad \text{(by Lemma 2.2)} \\
&= \Delta_1^2(S_1) + \Delta_1^2(S_2) + \frac{ab}{a+b} \cdot \|s_1 - s_2\|^2.
\end{aligned}$$

The second equality follows from Lemma 2.2 by noting that s is also the center of mass of the point set where a points are located at s_1 and b points are located at s_2. So the optimal 1-means cost of this point set is given by $a\|s_1 - s\|^2 + b\|s_2 - s\|^2$, which (by Lemma 2.2) is also equal to $\frac{ab}{a+b} \cdot \|s_1 - s_2\|^2$. This proves part (i). Part (ii) follows by substituting $\|s_1 - s\| = \|s_1 - s_2\| \cdot b/(a+b)$ in part (i) and dropping the $\Delta_1^2(S_1)$ and $\Delta_1^2(S_2)$ terms. $\square$

3. THE 2-MEANS PROBLEM

We first consider the 2-means case. We assume that the input X is ϵ-separated for 2-means. We present an algorithm that returns a solution of cost at most $\left(1 + O(\epsilon^2)\right)\Delta_2^2(X)$ in linear time. An appealing feature of our algorithm is its simplicity, both in description and analysis. In Section 4, where we consider the k-means case, we will build upon this algorithm to obtain both a linear-time constant-factor (of the form $1 + O(\epsilon^2)$) approximation algorithm and a PTAS with running-time exponential in k, but linear in n, d.

The chief algorithmic novelty in our 2-means algorithm is a *nonuniform* sampling process to pick two seed centers. Our sampling process is very simple: we pick the pair $x, y \in X$ with probability proportional to $\|x - y\|^2$. This biases the distribution towards pairs that contribute a large amount to $\Delta_1^2(X)$ (noting that $n\Delta_1^2(X) = \sum_{\{x,y\}\subseteq X} \|x - y\|^2$). We emphasize that, as improving the seeding is the only way to get Lloyd's method to find a high-quality clustering, the topic of picking the initial seed centers has received much attention in the experimental literature (see, e.g., Pena et al. [1999] and references therein). However, to the best of our knowledge, this simple and intuitive seeding method is new to the vast literature on the k-means problem. By putting more weight on pairs that contribute a lot to $\Delta_1^2(X)$, the sampling process aims to pick the initial centers from the cores of the two optimal clusters. We define the core of a cluster precisely later, but loosely speaking, it consists of points in the cluster that are significantly closer to this cluster-center than to any other center. Lemmas 3.1 and 3.2 make the benefits of this approach precise. Thus, in essence, we are able to leverage the separation condition to nearly isolate the optimal centers. Once we have the initial centers within the cores of the two optimal clusters, we show that a simple Lloyd-like step, which is also simple to analyze, yields a good performance guarantee: we consider a suitable ball around each center and move the center to the centroid of this ball to obtain the final centers. This "ball-k-means" step is adopted from Effros and Schulman [2004b], where it is shown that if the k-means cost of the current solution is small compared to $\Delta_{k-1}^2(X)$ (which holds for us since the initial centers lie in the cluster-cores) then a Lloyd step followed by a ball-k-means step yields a clustering of cost close to $\Delta_k^2(X)$. In our case, we are able to eliminate the Lloyd step, and show that the ball-k-means step alone guarantees a good clustering.

(1) *Sampling.* Randomly select a pair of points from the set X to serve as the initial centers, picking the pair $x, y \in X$ with probability proportional to $\|x - y\|^2$. Let $\hat{c}_1, \hat{c}_2$ denote the two picked centers.
(2) *"Ball-**k**-Means" Step.* For each $\hat{c}_i$, consider the ball of radius $\|\hat{c}_1 - \hat{c}_2\|/3$ around $\hat{c}_i$ and compute the centroid $\bar{c}_i$ of the portion of X in this ball. Return $\bar{c}_1, \bar{c}_2$ as the final centers.

Running Time. The entire algorithm runs in time $O(nd)$. Step (2) clearly takes only $O(nd)$ time. We show that the sampling step can be implemented to run in $O(nd)$ time. Consider the following two-step sampling procedure: (a) first pick center $\hat{c}_1$ by choosing a point $x \in X$ with probability equal to $\frac{\sum_{y\in X}\|x-y\|^2}{\sum_{x,y\in X}\|x-y\|^2} = (\Delta_1^2(X) + n\|x - c\|^2)/2n\Delta_1^2(X)$ (using Lemma 2.2); (b) pick the second center by choosing point $y \in X$ with probability equal to $\|y - \hat{c}_1\|^2/(\Delta_1^2(X) + n\|c - \hat{c}_1\|^2)$. This two-step sampling procedure is equivalent to the sampling process in Step (1), that is, it picks pair $x_1, x_2 \in X$ with probability $\boxed{\frac{\|x_1-x_2\|^2}{\sum_{\{x,y\}\subseteq X}\|x-y\|^2}}$. Each step takes only $O(nd)$ time since $\Delta_1^2(X)$ can be precomputed in $O(nd)$ time.

Analysis. The analysis hinges on the important fact that under the separation condition, the radius r_i of each optimal cluster is substantially smaller than the intercluster separation $\|c_1 - c_2\|$ (Lemma 3.1). This allows us to show in Lemma 3.2 that, with high probability, each initial center $\hat{c}_i$ lies in the core (suitably defined) of a distinct optimal cluster, say X_i, and hence $\|c_1 - c_2\|$ is much larger than the distances $\|\hat{c}_i - c_i\|$ for $i = 1, 2$. Assuming that $\hat{c}_1, \hat{c}_2$ lie in the cores of the clusters, we prove in Lemma 3.3

that the ball around $\hat{c}_i$ contains only, and most of the mass of cluster X_i, and therefore the centroid $\bar{c}_i$ of this ball is very "close" to c_i. This, in turn, implies that the cost of the clustering around $\bar{c}_1, \bar{c}_2$ is small.

LEMMA 3.1. $\max(r_1^2, r_2^2) \le \frac{\epsilon^2}{1-\epsilon^2}\|c_1 - c_2\|^2 = O(\epsilon^2)\|c_1 - c_2\|^2$.

PROOF. By part (i) of Lemma 2.3 we have $\Delta_1^2(X) = \Delta_2^2(X) + \frac{n_1 n_2}{n} \cdot \|c_1 - c_2\|^2$ which is equivalent to $\frac{n}{n_1 n_2} \cdot \Delta_2^2(X) = \|c_1 - c_2\|^2 \frac{\Delta_2^2(X)}{\Delta_1^2(X) - \Delta_2^2(X)}$. This implies that $r_1^2 \cdot \frac{n}{n_2} + r_2^2 \cdot \frac{n}{n_1} \le \frac{\epsilon^2}{1-\epsilon^2}\|c_1 - c_2\|^2$. □

Let $\rho = \frac{100\epsilon^2}{1-\epsilon^2}$. We require that $\rho < \frac{1}{4}$ (so we need $\epsilon^2 < \frac{1}{401}$). We define the core of cluster X_i as the set $X_i^{\text{cor}} = \{x \in X_i : \|x - c_i\|^2 \le \frac{r_i^2}{\rho}\}$. By Markov's inequality, $|X_i^{\text{cor}}| \ge (1-\rho)n_i$ for $i = 1, 2$.

LEMMA 3.2. $\Pr\left[\{\hat{c}_1, \hat{c}_2\} \cap X_1^{\text{cor}} \neq \emptyset \text{ and } \{\hat{c}_1, \hat{c}_2\} \cap X_2^{\text{cor}} \neq \emptyset\right] \ge 1 - 4\rho$.

PROOF. To simplify our expressions, we assume that all the points are scaled by $\frac{1}{\|c_1 - c_2\|}$ (so $\|c_1 - c_2\| = 1$). By part (i) of Lemma 2.3, we have $\Delta_1^2(X) = \Delta_2^2(X) + \frac{n_1 n_2}{n} \cdot \|c_1 - c_2\|^2$ which implies that $\Delta_1^2(X) \le \frac{n_1 n_2}{n(1-\epsilon^2)}$ (since $\Delta_2^2(X) \le \epsilon^2 \Delta_1^2(X)$ due to the separation condition). Let c_i' denote the center of mass of X_i^{cor}. Applying part (ii) of Lemma 2.3 (taking $S = X_i$ and $S_1 = X_i^{\text{cor}}$) we get that $\|c_i' - c_i\|^2 \le \frac{\rho}{1-\rho} \cdot r_i^2$. The probability of the event in the lemma is A/B where

$$
\begin{aligned}
A &= \sum_{x \in X_1^{\text{cor}}} \sum_{y \in X_2^{\text{cor}}} \|x - y\|^2 = \sum_{x \in X_1^{\text{cor}}} \left(\Delta_1^2(X_2^{\text{cor}}) + |X_2^{\text{cor}}| \|x - c_2'\|^2\right) \\
&= |X_1^{\text{cor}}| \Delta_1^2(X_2^{\text{cor}}) + |X_2^{\text{cor}}| \Delta_1^2(X_1^{\text{cor}}) + |X_1^{\text{cor}}||X_2^{\text{cor}}| \|c_1' - c_2'\|^2 \ \ge\ (1-\rho)^2 n_1 n_2 \|c_1' - c_2'\|^2,
\end{aligned}
$$

and $B = \sum_{\{x,y\} \subseteq X} \|x - y\|^2 = n\Delta_1^2(X) \le \frac{n_1 n_2}{1-\epsilon^2}$. By these bounds on $\|c_i' - c_i\|$ and Lemma 3.1, we get $\|c_1' - c_2'\| \ge \|c_1 - c_2\| - 2\sqrt{\frac{\rho}{1-\rho}} \max(r_1, r_2) \ge 1 - 2\epsilon\sqrt{\frac{\rho}{(1-\rho)(1-\epsilon^2)}} \ge 1 - \frac{\rho}{5\sqrt{1-\rho}}$. So $A \ge \left(1 - 2\rho - \frac{2\rho}{5\sqrt{1-\rho}}\right) n_1 n_2$, and $A/B \ge \left(1 - \epsilon^2 - 2\rho - \frac{2\rho}{5\sqrt{1-\rho}}\right) \ge 1 - 4\rho$. □

So we may assume that each initial center $\hat{c}_i$ lies in X_i^{cor}. Let $\hat{d} = \|\hat{c}_1 - \hat{c}_2\|$ and $B_i = \{x \in X : \|x - \hat{c}_i\| \le \hat{d}/3\}$. Recall that $\bar{c}_i$ is the centroid of B_i, and we return $\bar{c}_1, \bar{c}_2$ as our final solution.

LEMMA 3.3. *For each i, we have $X_i^{\text{cor}} \subseteq B_i \subseteq X_i$. Hence, $\|\bar{c}_i - c_i\|^2 \le \frac{\rho}{1-\rho} \cdot r_i^2$.*

PROOF. Let $\theta = \frac{\epsilon}{\sqrt{\rho(1-\epsilon^2)}} \le \frac{1}{10}$. As argued in Lemma 3.2, we have $\|\hat{c}_i - c_i\| \le \sqrt{\frac{\rho}{1-\rho}} r_i \le \theta\|c_1 - c_2\|$ for $i = 1, 2$ (by Lemma 3.1). So $\frac{4}{5} \le \frac{\hat{d}}{\|c_1 - c_2\|} \le \frac{6}{5}$. For any $x \in B_i$ we have $\|x - c_i\| \le \frac{\hat{d}}{3} + \|\hat{c}_i - c_i\| \le \frac{\|c_1 - c_2\|}{2}$, so $x \in X_i$. Also for any $x \in X_i^{\text{cor}}$, $\|x - \hat{c}_i\| \le 2\theta\|c_1 - c_2\| \le \frac{\hat{d}}{3}$, so $x \in B_i$. Now by part (ii) of Lemma 2.3, with $S = X_i$ and $S_1 = B_i$, we obtain that $\|\bar{c}_i - c_i\|^2 \le \frac{\rho}{1-\rho} \cdot r_i^2$ since $|B_i| \ge |X_i^{\text{cor}}|$ for each i. □

THEOREM 3.4. *This algorithm returns a clustering of cost at most $\frac{\Delta_2^2(X)}{1-\rho}$ with probability at least $1 - O(\rho)$ in time $O(nd)$, where $\rho = \frac{100\epsilon^2}{1-\epsilon^2}$.*

PROOF. The cost of the solution is at most $\sum_{i,x\in X_i} \|x-\bar{c}_i\|^2 = \sum_i \big(\Delta_1^2(X_i)+n_i\|\bar{c}_i-c_i\|^2\big) \le \frac{\Delta_2^2(X)}{1-\rho}$. □

4. THE k-MEANS PROBLEM

We now consider the k-means setting. We assume that $\Delta_k^2(X) \le \epsilon^2\Delta_{k-1}^2(X)$. We describe a linear time constant-factor approximation algorithm, and a PTAS that returns a $(1+\omega)$-optimal solution in time $O\big(2^{O(k/\omega)}nd\big)$. The algorithms consist of various ingredients, which we describe separately first for ease of understanding, before gluing them together to obtain the final algorithm.

Conceptually, both algorithms proceed in two stages. The first stage is a seeding stage, which performs the bulk of the work and guarantees that at the end of this stage there are k seed centers positioned at nearly the right locations. By this, we mean that if we consider distances at the scale of the intercluster separation, then at the end of this stage, each optimal center has a (distinct) initial center located in close proximity — this is precisely the leverage that we obtain from the k-means separation condition (as in the 2-means case). We shall employ three simple seeding procedures with varying time vs. quality guarantees that will exploit this condition and seed the k centers at locations very close to the optimal centers. In Section 4.1.1, we consider a natural generalization of the sampling procedure used for the 2-means case, and show that this picks the k initial centers from the cores of the optimal clusters. This sampling procedure runs in linear time but it succeeds with probability that is exponentially small in k. In Section 4.1.2, we present a very simple deterministic greedy deletion procedure, where we start off with all points in X as the centers and then greedily delete points (and move centers) until there are k centers left. The running time here is $O(n^3d)$. Our deletion procedure is similar to the reverse greedy algorithm proposed by Chrobak et al. [2006] for the k-median problem. Chrobak et al. [2006] show that their reverse greedy algorithm attains an approximation ratio of $O(\log n)$, which is tight up to a factor of $\log\log n$. In contrast, for the k-means problem, if $\Delta_k^2(X) \le \epsilon^2\Delta_{k-1}^2(X)$, we show that our greedy deletion procedure followed by a clean-up step (in the second stage) yields a $\big(1 + O(\epsilon^2)\big)$-approximation algorithm. Finally, in Section 4.1.3, we combine the sampling and deletion procedures to obtain an $O(nkd + k^3d)$-time initialization procedure. We sample $O(k)$ centers, which ensures that every cluster has an initial center in a slightly expanded version of the core, and then run the deletion procedure on an instance of size $O(k)$ derived from the sampled points to obtain the k seed centers.

Once the initial centers have been positioned sufficiently close to the optimal centers, we can proceed in two ways in the *second-stage* (Section 4.2). One option is to use a ball-k-means step, as in 2-means, which yields a clustering of cost $\big(1 + O(\epsilon^2)\big)\Delta_k^2(X)$ due to exactly the same reasons as in the 2-means case. Thus, combined with the initialization procedure of Section 4.1.3, this yields a constant-factor approximation algorithm with running time $O(nkd + k^3d)$. The entire algorithm is summarized in Section 4.3.

The other option, which yields a PTAS, is to use a sampling idea of Kumar et al. [2004]. For each initial center, we compute a list of candidate centers for the corresponding optimal cluster as follows: we sample a small set of points uniformly at random from a slightly expanded Voronoi region of the initial center, and consider the centroid of every subset of the sampled set of a certain size as a candidate. We exhaustively search for the k candidates (picking one candidate per initial center) that yield the least cost solution, and output these as our final centers. The fact that each optimal center c_i has an initial center in close proximity allows us to argue that the entire optimal cluster X_i is contained in the expanded Voronoi region of this initial center, and moreover that $|X_i|$ is a significant fraction of the total mass in this region. Given this property, as argued by Kumar et al. [2004, Lemma 2.3], a random sample from the expanded Voronoi region also (essentially) yields a random sample from X_i, which allows us to compute a good estimate of the centroid of X_i, and hence of $\Delta_1^2(X_i)$. We obtain a $(1+\omega)$-optimal solution in time $O\big(2^{O(k/\omega)}nd\big)$ with constant probability. Since we incur an exponential dependence on k anyway, we just use the simple sampling procedure of Section 4.1.1 in the first-stage to pick the k initial centers. Although the running time is exponential in k, it is significantly better than the running time of $O\big(2^{(k/\omega)^{O(1)}}nd\big)$ incurred by the algorithm of Kumar et al. [2004]; we also obtain a simpler PTAS. Both of these features can be traced to the separation condition, which enables us to nearly isolate the positions of the optimal centers in the first stage. Kumar et al. [2004] do not have any such facility, and therefore need to sequentially "guess" (i.e., exhaustively search) the various centroids, incurring a corresponding increase in the run time. This PTAS is described in Section 4.4.

The following lemma, which is a simple extension of Lemma 3.1 to the k-means case and is proved via an almost identical argument, will be used repeatedly.

LEMMA 4.1. *For every i, we have $r_i^2 \le \frac{\epsilon^2}{1-\epsilon^2} \cdot \min_{j \ne i} \|c_i - c_j\|^2$.*

4.1. Seeding Procedures Used in Stage I

4.1.1. Sampling. We pick k initial centers as follows: first pick two centers $\hat{c}_1, \hat{c}_2$ as in the 2-means case, that is, choose $x, y \in X$ with probability proportional to $\|x-y\|^2$. Suppose we have already picked i centers $\hat{c}_1, \ldots, \hat{c}_i$ where $2 \le i < k$. Now pick a random point $x \in X$ with probability proportional to $\min_{j \in \{1,\ldots,i\}} \|x - \hat{c}_j\|^2$ and set that as center $\hat{c}_{i+1}$. (We remark that the randomized seeding method in Arthur and Vassilvitskii [2007] is very similar; the only difference is that they choose the first center by sampling a point uniformly at random from X, whereas we choose $x \in X$ to be the first center with probability proportional to $\sum_{y \in X} \|x - y\|^2$.)

Running Time. The sampling procedure consists of k iterations, each of which takes $O(nd)$ time. This is because after sampling a new point $\hat{c}_{i+1}$, we can update the quantity $\min_{j \in \{1,\ldots,i+1\}} \|x - \hat{c}_j\|$ for each point x in $O(d)$ time. So the overall running time is $O(nkd)$.

Analysis. Let ρ be a parameter satisfying $\epsilon^2 \ll \rho < \frac{1}{6}$ that we will set later. We need to set ρ to different values (depending on ϵ and satisfying these bounds) depending on whether we are analyzing the case where exactly k points are sampled or $O(k)$ points are sampled (we set $\rho = \epsilon^{2/3}$ for the former, and $\rho = \sqrt{\epsilon}$ for the latter; see Corollary 4.6). As in the 2-means case, we define the core of cluster X_i as $X_i^{\mathrm{cor}} = \{x \in X_i : \|x - c_i\|^2 \le \frac{r_i^2}{\rho}\}$. We show that under our separation assumption, this sampling procedure will pick the k initial centers to lie in the cores of the clusters $X_1, \ldots, X_k$ with probability $\big(1 - O(\rho)\big)^k$. We also show in Lemma 4.5 that if more than k, but still $O(k)$, points are sampled, then with *constant probability*, every cluster will contain a sampled point that lies in a somewhat larger core, that we call the *outer core* of the cluster. This analysis will be useful in Section 4.1.3.

LEMMA 4.2. *With probability $1 - O(\rho)$, the first two centers $\hat{c}_1, \hat{c}_2$ lie in the cores of different clusters, that is,* $\Pr[\bigcup_{i \ne j}(x \in X_i^{\mathrm{cor}}$ *and* $y \in X_j^{\mathrm{cor}})] \ge 1 - 4\rho$.

PROOF. The key observation is that for any pair of distinct clusters X_i, X_j, the 2-means separation condition holds, that is, $\Delta_2^2(X_i \cup X_j) = \Delta_1^2(X_i) + \Delta_1^2(X_j) \le \epsilon^2 \Delta_1^2(X_i \cup X_j)$. This is because

$$\Delta_{k-1}^2(X) \le \sum_{\ell \ne i,j} \Delta_1^2(X_\ell) + \Delta_1^2(X_i \cup X_j) = \Delta_k^2(X) + \big(\Delta_1^2(X_i \cup X_j) - \Delta_2^2(X_i \cup X_j)\big).$$

So $\Delta_1^2(X_i \cup X_j) - \Delta_2^2(X_i \cup X_j) \geq (\frac{1}{\epsilon^2} - 1)\Delta_k^2(X) \geq (\frac{1}{\epsilon^2} - 1)\Delta_2^2(X_i \cup X_j)$. So, using Lemma 3.2, we obtain that $\sum_{x \in X_i^{\mathrm{cor}}, y \in X_j^{\mathrm{cor}}} \|x - y\|^2 \geq (1 - 4\rho) \sum_{\{x,y\} \subseteq X_i \cup X_j} \|x - y\|^2$. Summing over all pairs i, j yields the lemma. □

Now suppose that the first i centers picked, $\hat{c}_1, \ldots, \hat{c}_i$, all lie in cluster cores. We show in Lemma 4.3 that conditioned on this event, center $\hat{c}_{i+1}$ lies in the core of distinct cluster with probability $1 - O(\rho)$. Suppose that $\hat{c}_1, \ldots, \hat{c}_i$ lie in the cores of clusters $X_{j_1}, \ldots, X_{j_i}$. Given a set S of points, we use $d(x, S)$ to denote $\min_{y \in S} \|x - y\|$.

LEMMA 4.3. $\Pr\big[\hat{c}_{i+1} \in \bigcup_{\ell \notin \{j_1, \ldots, j_i\}} X_\ell^{\mathrm{cor}} \,|\, \hat{c}_1, \ldots, \hat{c}_i$ *lie in the cores of* $X_{j_1}, \ldots, X_{j_i}\big] \geq 1 - 6\rho$.

PROOF. For notational convenience, re-index the clusters so that $\{j_1, \ldots, j_i\} = \{1, \ldots, m\}$. (Note that m could be smaller than i as multiple centers could lie in the core of the same cluster.) Let $\hat{C} = \{\hat{c}_1, \ldots, \hat{c}_i\}$. For any cluster X_j, let $p_j \in \{1, \ldots, i\}$ be the index such that $\hat{c}_{p_j}$ is the center in $\hat{C}$ nearest to c_j, that is, $d(c_j, \hat{C}) = \|c_j - \hat{c}_{p_j}\|$. Let $A = \sum_{j=m+1}^{k} \sum_{x \in X_j^{\mathrm{cor}}} d(x, \hat{C})^2$, and $B = \sum_{j=1}^{k} \sum_{x \in X_j} d(x, \hat{C})^2$. Observe that the probability of the event stated in the lemma is exactly A/B. Let α denote the maximum over all $j \geq m+1$ of the quantity $\max_{x \in X_j^{\mathrm{cor}}} \|x - c_j\| / d(c_j, \hat{C})$. Consider any $j \geq m+1$ and $x \in X_j^{\mathrm{cor}}$. We have $\|x - c_j\| \leq \frac{r_j}{\sqrt{\rho}} \leq \frac{\epsilon}{\sqrt{\rho(1-\epsilon^2)}} \cdot \|c_j - c_{p_j}\|$, where the last inequality follows from Lemma 4.1. Also, $d(c_j, \hat{C}) = \|c_j - \hat{c}_{p_j}\| \geq \|c_j - c_{p_j}\| - \|c_{p_j} - \hat{c}_{p_j}\| \geq \|c_j - c_{p_j}| - \frac{r_{p_j}}{\sqrt{\rho}} \geq \big(1 - \frac{\epsilon}{\sqrt{\rho(1-\epsilon^2)}}\big)\|c_j - c_{p_j}\|$, where the last inequality uses Lemma 4.1. Thus, $\alpha \leq \frac{\epsilon/\sqrt{\rho(1-\epsilon^2)}}{1-\epsilon/\sqrt{\rho(1-\epsilon^2)}} \leq \frac{2\epsilon}{\sqrt{\rho(1-\epsilon^2)}} < 1$ since $\epsilon^2 \ll \rho$.

For any point $x \in X_j^{\mathrm{cor}}$, $j \geq m+1$, we have $d(x, \hat{C}) \geq (1-\alpha) d(c_j, \hat{C})$. Therefore,

$$A = \sum_{j=m+1}^{k} \sum_{x \in X_j^{\mathrm{cor}}} d(x, \hat{C})^2 \geq \sum_{j=m+1}^{k} (1-\rho)(1-\alpha)^2 n_j d(c_j, \hat{C})^2 \geq (1 - \rho - 2\alpha) \sum_{j=m+1}^{k} n_j d(c_j, \hat{C})^2.$$

On the other hand, for any point $x \in X_j$, $j = 1, \ldots, k$, we have $d(x, \hat{C}) \leq \|x - \hat{c}_{p_j}\|$. Also note that for $j = 1, \ldots, m$, $\hat{c}_{p_j}$ lies in X_j^{cor}, so $\|c_j - \hat{c}_{p_j}\| \leq \frac{r_j}{\sqrt{\rho}}$. Therefore,

$$B \leq \sum_{j=1}^{k} \sum_{x \in X_j} \|x - \hat{c}_{p_j}\|^2 \leq \sum_{j=1}^{k} \big(\Delta_1^2(X_j) + n_j \|c_j - \hat{c}_{p_j}\|^2\big) \leq \left(1 + \frac{1}{\rho}\right) \Delta_k^2(X) + \sum_{j=m+1}^{k} n_j d(c_j, \hat{C})^2.$$

Finally, for any $j = m+1, \ldots, k$, if we assign all the points in cluster X_j to the point $\hat{c}_{p_j}$, then the increase in cost is exactly $n_j \|c_j - \hat{c}_{p_j}\|^2$ and at least $\Delta_{k-1}^2(X) - \Delta_k^2(X)$. Therefore $(\frac{1}{\epsilon^2} - 1)\Delta_k^2(X) \leq n_j d(c_j, \hat{C})^2$ for any $j = m+1, \ldots, k$, and $B \leq \frac{1+\epsilon^2/\rho}{1-\epsilon^2} \sum_{j=m+1}^{k} n_j d(c_j, \hat{C})^2$. Comparing with A and plugging in the value of α, we get that $A/B \geq 1 - (\rho + \frac{4\epsilon}{\sqrt{\rho(1-\epsilon^2)}} + \epsilon^2 + \epsilon^2/\rho)$. So if we set $\rho = \epsilon^{2/3}$, we obtain $A/B \geq 1 - 6\rho$. □

Next, we analyze the case when more than k points are sampled. Define the outer core of X_i to be $X_i^{\mathrm{out}} = \{x \in X_i : \|x - c_i\|^2 \leq \frac{r_i^2}{\rho^3}\}$. Note that $X_i^{\mathrm{cor}} \subseteq X_i^{\mathrm{out}}$. Let $N = \frac{2k}{1-5\rho} + \frac{2\ln(2/\delta)}{(1-5\rho)^2}$ where $0 < \delta < 1$ is a desired error tolerance. We prove in Lemma 4.4 that at every sampling step, there is a constant probability that the sampled point lies in the core of some cluster whose outer core does not contain a previously sampled point. The crucial difference between this lemma and Lemma 4.3 is that Lemma 4.3 only shows that the "good" event happens *conditioned* on the fact that previous samples were also "good", whereas here we give an *unconditional* bound. Using this, Lemma 4.5 shows that if we sample N points from X, then with some constant probability, each outer core X_i^{out} will contain a sampled point. The proof is based on a straightforward martingale analysis.

LEMMA 4.4. *Suppose that we have sampled i points* $\{\hat{x}_1, \dots, \hat{x}_i\}$ *from X. Let* $X_1, \dots, X_m$ *be all the clusters whose outer cores contain some sampled point* $\hat{x}_j$. *Then* $\Pr[\hat{x}_{i+1} \in \bigcup_{j=m+1}^{k} X_j^{\mathrm{cor}}] \geq 1 - 5\rho$.

PROOF. For $i = 0, 1$, this follows from Lemma 4.2. We mimic the proof of Lemma 4.3. Let $\hat{C} = \{\hat{x}_1, \dots, \hat{x}_i\}$. We have $X_j^{\mathrm{out}} \cap \hat{C} \neq \emptyset$ for $j = 1, \dots, m$ and $X_j^{\mathrm{out}} \cap \hat{C} = \emptyset$ for $j = m+1, \dots, k$. Let α denote the maximum over all $j \geq m+1$ of the quantity $(\max_{x \in X_j^{\mathrm{cor}}} \|x - c_j\|)/d(c_j, \hat{C})$. Here we have $\alpha \leq \sqrt{\rho^3/\rho} < 1$. Then, for any point $x \in X_j^{\mathrm{cor}}$, $j \geq m+1$, we have $d(x, \hat{C}) \geq (1-\alpha)d(c_j, \hat{C})$ and as in Lemma 4.3, $A = \sum_{j=m+1}^{k} \sum_{x \in X_j^{\mathrm{cor}}} d(x, \hat{C})^2 \geq (1 - \rho - 2\alpha) \sum_{j=m+1}^{k} n_j d(c_j, \hat{C})^2$. On the other hand, again arguing as in Lemma 4.3, we have $B = \sum_{j=1}^{k} \sum_{x \in X_j} d(x, \hat{C})^2 \leq \frac{1+\epsilon^2/\rho^3}{1-\epsilon^2} \sum_{j=m+1}^{k} n_j d(c_j, \hat{C})^2$. Therefore $A/B \geq 1 - \big(\rho + 2\sqrt{\frac{\rho^3}{\rho}} + \frac{\epsilon^2}{\rho^3} + \epsilon^2\big)$. So taking $\rho = \sqrt{\epsilon}$ gives $A/B \geq 1 - 5\rho$. □

LEMMA 4.5. *Suppose we sample* $N = \frac{2k}{1-5\rho} + \frac{2\ln(2/\delta)}{(1-5\rho)^2}$ *points* $\hat{x}_1, \dots, \hat{x}_N$ *from X using the above sampling procedure. Then,* $\Pr[\forall j = 1, \dots, k,$ *there exists some* $\hat{x}_i \in X_j^{\mathrm{out}}] \geq 1 - \delta$.

PROOF. Let Y_t be a random variable that denotes the number of clusters that do not contain a sampled point in their outer cores, after t points have been sampled. We want to bound $\Pr[Y_N > 0]$. Consider the following random walk on the line with W_t denoting the (random) position after t time steps: $W_0 = k$, and $W_{t+1} = W_t$ with probability 5ρ and $W_t - 1$ with probability $1 - 5\rho$. Notice that $\Pr[Y_N > 0] \leq \Pr[W_N > 0]$, because as long as $W_t > 0$, any outcome that leads to a left move in the random walk can be mapped to an outcome (in the probability space corresponding to the sampling process) where the outer core of a new cluster is hit by the currently sampled point. So we bound $\Pr[W_N > 0]$. Define $Z_t = W_t + t(1 - 5\rho)$. Then, $\mathrm{E}\big[Z_{t+1} | Z_1, \dots, Z_t\big] \leq Z_t$, so $Z_0, Z_1, \dots$ forms a supermartingale. Clearly $|Z_{t+1} - Z_t| \leq 1$ for all t. So by Azuma's inequality (see, e.g., Motwani and Raghavan [1995]), $\Pr[Z_N - Z_0 > \sqrt{2N\ln(2/\delta)}] \leq \delta$ which implies that $W_N \leq k + \sqrt{2N\ln(2/\delta)} - N(1 - 5\rho)$ with probability at least $1 - \delta$. Plugging the value of N shows that $N(1 - 5\rho) - \sqrt{2N\ln(2/\delta)} \geq k$. □

COROLLARY 4.6. *(i) If we sample k points* $\hat{c}_1, \dots, \hat{c}_k$, *then with probability* $\big(1 - O(\rho)\big)^k$, *where* $\rho = \epsilon^{2/3}$, *for each i there is a distinct center* $\hat{c}_i \in X_i^{\mathrm{cor}}$, *that is,* $\|\hat{c}_i - c_i\| \leq r_i/\sqrt{\rho}$. *Hence, if* ϵ *is such that* $\frac{\epsilon}{\sqrt{\rho(1-\epsilon^2)}} \leq \frac{1}{10}$, $\rho < \frac{1}{6}$ *(so* $\epsilon \leq \frac{1}{32}$ *suffices), we have that* $\|\hat{c}_i - c_i\| \leq D_i/10$.
(ii) If we sample N points $\hat{x}_1, \dots, \hat{x}_N$, *where* $N = \frac{2k}{1-5\rho} + \frac{2\ln(2/\rho)}{(1-5\rho)^2}$ *and* $\rho = \sqrt{\epsilon}$, *then with probability* $1 - O(\rho)$, *for each i there is a distinct point* $\hat{x}_i \in X_i^{\mathrm{out}}$, *that is,* $\|\hat{x}_i - c_i\| \leq r_i/\sqrt{\rho^3}$.

Notice that part (i) of Corollary 4.6 implies that, as in the 2-means setting, if we perform a ball-k-means step (where we "move" each $\hat{c}_i$ to the centroid of $\{x \in X : \|x - \hat{c}_i\| \leq \min_{j \neq i} \|\hat{c}_j - \hat{c}_i\|/3\}$), we can argue by mimicking the proof of Lemma 3.3 and Theorem 3.4 that the cost of the resulting clustering is at most $\frac{\Delta_k^2(X)}{1-\rho}$. As mentioned at the beginning of Section 4, this algorithm is subsumed by an algorithm that uses oversampling and greedy deletion followed by a ball-k-means step to guarantee an $O(1)$-approximation with constant success probability (see Section 4.3).

4.1.2. Greedy Deletion Procedure. We maintain a set of centers $\hat{C}$ that are currently used to cluster X. For any point $x \in \mathbb{R}^d$, let $R(x) \subseteq X$ denote the points of X in the Voronoi region of x (given the set of centers $\hat{C}$). We refer to $R(x)$ as the Voronoi set of x. Initialize $\hat{C} \leftarrow X$. Repeat the following steps until $|\hat{C}| = k$.

A1. Compute T = cost of clustering X around centers in $\hat{C} = \sum_{x\in\hat{C}} \sum_{y\in R(x)} \|y - x\|^2$. Also for every $x \in \hat{C}$, compute T_x = cost of clustering X around $\hat{C} \setminus \{x\} = \sum_{z\in\hat{C}\setminus\{x\}} \sum_{y\in R_{-x}(z)} \|y - z\|^2$, where $R_{-x}(z)$ denotes the Voronoi set of z given the center set $\hat{C} \setminus \{x\}$.

A2. Pick the center $y \in \hat{C}$ for which $T_x - T$ is minimum and set $\hat{C} \leftarrow \hat{C} \setminus \{y\}$.

A3. Recompute the Voronoi sets $R(x) = R_{-y}(x) \subseteq X$ for each (remaining) center $x \in \hat{C}$. Now we "move" the centers to the centroids of their respective (new) Voronoi sets, that is, for every set $R(x)$, we update $\hat{C} \leftarrow \hat{C} \setminus \{x\} \cup \{\mathrm{ctr}(R(x))\}$.

Running Time. There are $n - k$ iterations of the A1-A3 loop. Each iteration takes $O(n^2 d)$ time: computing T and the sets $R(x)$ for each x takes $O(n^2 d)$ time and we can then compute each T_x in $O(|R(x)|d)$ time (since while computing T, we can also compute for each point its second-nearest center in $\hat{C}$). Therefore, the overall running time is $O(n^3 d)$.

Analysis. Let ρ be a parameter such that $\rho \leq \frac{1}{10}$, $\epsilon/\sqrt{\rho(1-\epsilon^2)} \leq \frac{1}{14}$ (so we need $\epsilon^2 \leq \frac{1}{1961}$). Recall that $D_i = \min_{j\neq i} \|c_j - c_i\|$. Define $d_i^2 = \Delta_k^2(X)/n_i$. We will use a different notion of a cluster-core here, but the notion will still capture the fact that the core consists of points that are quite close to the cluster-center compared to the intercluster distance, and contains most of the mass of the cluster. Let $\mathcal{B}(x, r) = \{y \in \mathbb{R}^d : \|x - y\| \leq r\}$ denote the ball of radius r centered at x. Define the *kernel* of X_i to be the ball $Z_i = \mathcal{B}(c_i, d_i/\sqrt{\rho})$ and the core of X_i as $X_i^{\mathrm{cor}} = X_i \cap Z_i$. Observe that $r_i \leq d_i$, so by Markov's inequality $|X_i^{\mathrm{cor}}| \geq (1-\rho)n_i$. Also, since $\Delta_{k-1}^2(X) - \Delta_k^2(X) \leq n_i D_i^2$ we have that $d_i^2 \leq D_i^2 \cdot \frac{\epsilon^2}{1-\epsilon^2}$. Therefore, $X_i^{\mathrm{cor}} = X \cap Z_i$. We prove that,

> at the start of every iteration, for every i, there is a (distinct) center $x \in \hat{C}$ that lies in Z_i. (*)

Clearly (*) holds at the beginning, since $\hat{C} = X$ and $X_i^{\mathrm{cor}} \neq \emptyset$ for every cluster X_i. First we show (Lemma 4.7) that if $x \in \hat{C}$ is the only center that lies in a slightly enlarged version of the ball Z_i for some i, then x is not deleted. Lemma 4.8 then makes the crucial observation that even after a center y is deleted, if the new Voronoi set $R_{-y}(x)$ of a center $x \in \hat{C}$ captures points from X_i^{cor}, then $R_{-y}(x)$ cannot "extend" too far into some other cluster X_i, that is, for $x' \in R_{-y}(x) \cap X_j$ where $j \neq i$, $\|y - c_i\|$ is not much larger than $\|y - c_j\|$. It will then follow that invariant (*) is maintained.

LEMMA 4.7. *Suppose (*) holds at the start of an iteration, and $x \in \hat{C}$ is the only center in $\mathcal{B}\big(c_i, \frac{4d_i}{\sqrt{\rho}}\big)$ for some cluster X_i, then $x \in \hat{C}$ after Step A2.*

PROOF. Since property (*) holds, we also know that $x \in Z_i$ and so $X_i^{\mathrm{cor}} \subseteq R(x) \cap X_i$ (see Figure 1). If x is deleted in Step A2, then all points in X_i^{cor} will be reassigned to a center at least $\frac{4d_i}{\sqrt{\rho}}$ away from c_i. So the cost-increase $T_x - T$ is at least $A =$

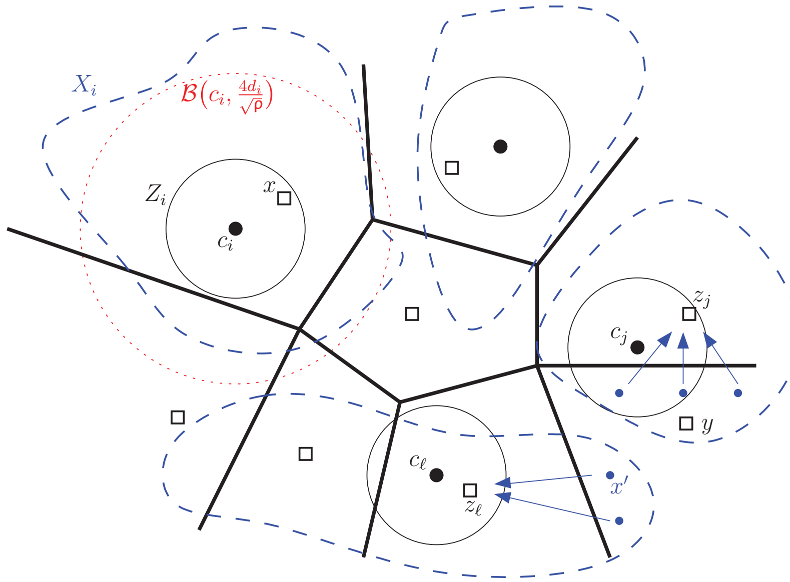

Fig. 1. A snapshot at the beginning of the A1-A3 loop. The squares are points in $\hat{C}$. The solid lines show the Voronoi regions with respect to $\hat{C}$; the dashed lines show the optimal clusters. x is the only point in $\hat{C}$ in $\mathcal{B}(c_i, \frac{4d_i}{\sqrt{\rho}})$. The arrows show the reasignment used to upper bound the cost T_y in Lemma 4.7.

$\frac{5(1-\rho)}{\rho} \cdot n_i d_i^2 = \frac{5(1-\rho)}{\rho} \cdot \Delta_k^2(X)$. Now since $|\hat{C}| > k$, there is some j (j could be i) such that the Voronoi region of c_j (with respect to the optimal center-set) contains at least two centers from $\hat{C}$. We will show that deleting one of these centers will be less expensive than deleting x. Let $z_\ell \in \hat{C}$ be the center closest to c_ℓ for $\ell = 1, \ldots, k$. Note that $z_\ell \in Z_\ell$. Let $y \in \hat{C}$, $y \neq z_j$ be another center in the Voronoi region of c_j. Suppose we delete y. We can upper bound the cost-increase $T_y - T$ by the cost-increase due to the reassignment where we assign all points in $R_y \cap X_\ell$ to z_ℓ for $\ell = 1, \ldots, k$. For any $x' \in R(y) \cap X_\ell$, we have $\|x' - z_\ell\| \leq \|x' - c_\ell\| + \|c_\ell - z_\ell\| \leq \|x' - c_\ell\| + \frac{d_\ell}{\sqrt{\rho}}$. For $\ell \neq j$, we also have $D_\ell \leq \|c_j - y\| + \|y - c_\ell\| \leq 2\|y - c_\ell\|$ since c_j is closer to y than c_ℓ, and

$$\|y-c_\ell\| \leq \|x'-c_\ell\|+\|x'-y\| \leq \|x'-c_\ell\|+\|x'-z_\ell\| \leq 2\|x'-c_\ell\|+\|c_\ell-z_\ell\| \leq 2\|x'-c_\ell\|+\frac{d_\ell}{\sqrt{\rho}}.$$

Therefore, $D_\ell \leq 4\|x'-c_\ell\|+\frac{2d_\ell}{\sqrt{\rho}}$ which implies that $(\frac{\sqrt{1-\epsilon^2}}{\epsilon} - \frac{2}{\sqrt{\rho}})d_\ell \leq 4\|x'-c_\ell\|$. Combining this with the bound on $\|x' - z_\ell\|$, we get that for $\ell \neq j$, $\|x' - z_\ell\| \leq \beta\|x' - c_\ell\|$ where $\beta = 1 + \frac{4\epsilon}{\sqrt{\rho(1-\epsilon^2)}-2\epsilon}$. Hence, the cost-increase of the reassignment is at most

$$\begin{aligned} B &= \sum_{\ell=1}^{k} \sum_{x' \in R(y) \cap X_\ell} \|x' - z_\ell\|^2 \\ &\leq \sum_{x' \in R(y) \cap X_j} 2\left(\|x' - c_j\|^2 + \frac{d_j^2}{\rho}\right) + \sum_{\ell \neq j} \sum_{x' \in R(y) \cap X_\ell} \beta^2 \|x' - c_\ell\|^2 \\ &\leq \max(2, \beta^2)\Delta_k^2(X) + \tfrac{2}{\rho} \cdot n_j d_j^2 = \left(\max(2, \beta^2) + \tfrac{2}{\rho}\right)\Delta_k^2(X). \end{aligned}$$

Any ρ satisfying the bounds stated in Section 4.1.2 ensures that $A > B$ (since $\beta < \frac{4}{3}$ and $\rho < \frac{3}{7}$). Thus, x is not the cheapest center to delete, which completes the proof. $\square$

LEMMA 4.8. *Suppose center $y \in \hat{C}$ is deleted in Step A2. Let $x \in \hat{C} \setminus \{y\}$ be such that $R_{-y}(x) \cap X_j^{\text{cor}} \neq \emptyset$ for some j. Then, for any $x' \in R_{-y}(x) \cap X_\ell$, $\ell \neq j$, we have $\|x' - c_j\| \leq \|x' - c_\ell\| + \frac{\max(d_\ell + 6d_j, 4d_\ell + 3d_j)}{\sqrt{\rho}}$.*

PROOF. Suppose that y lies in the Voronoi region of center c_i (with respect to the optimal centers). Let $\hat{C}' = \hat{C} \setminus \{y\}$. There must be a center $z_i \in \hat{C}'$ such that $\|z_i - c_i\| \leq \frac{4d_i}{\sqrt{\rho}}$. If $y \notin Z_i$, this follows from property (*) otherwise this follows from Lemma 4.7. For any $\ell \neq i$, we know by property (*) that there is some center $z_\ell \in \hat{C}'$ that lies in Z_ℓ. Let x'' be a point in $R_{-y}(x) \cap X_j^{\text{cor}}$. Then,

$$\|x - c_j\| \leq \|x - x''\| + \|x'' - c_j\| \leq \|x'' - z_j\| + \|x'' - c_j\| \leq \|z_j - c_j\| + 2\|x'' - c_j\| \leq \|z_j - c_j\| + \frac{2d_j}{\sqrt{\rho}}.$$

Now considering the point x', we have

$$\begin{aligned} \|x' - c_j\| &\leq \|x' - x\| + \|x - c_j\| \leq \|x' - z_\ell\| + \|x - c_j\| \\ &\leq \|x' - c_\ell\| + \|z_\ell - c_\ell\| + \|x - c_j\| \leq \|x' - c_\ell\| + \|z_\ell - c_\ell\| + \|z_j - c_j\| + \frac{2d_j}{\sqrt{\rho}}. \end{aligned}$$

If $j = i$, then we get that, $\|x' - c_i\| \leq \|x' - c_\ell\| + \frac{d_\ell + 6d_i}{\sqrt{\rho}}$. For any other j, we have that $\|x' - c_j\| \leq \|x' - c_\ell\| + \frac{4d_\ell + 3d_j}{\sqrt{\rho}}$ (since it could be that $\ell = i$). $\square$

LEMMA 4.9. *Suppose that property (*) holds at the beginning of some iteration in the deletion phase. Then (*) also holds at the end of the iteration, that is, after Step A3.*

PROOF. Suppose that we delete center $y \in \hat{C}$ that lies in the Voronoi region of center c_i (with respect to the optimal centers) in Step A2. Let $\hat{C}' = \hat{C} \setminus \{y\}$ and $R'(x) = R_{-y}(x)$ for any $x \in \hat{C}'$. Fix a cluster X_j. Let $S = \{x \in \hat{C}' : R'(x) \cap X_j^{\text{cor}} \neq \emptyset\}$ and $Y = \bigcup_{x \in S} R'(x)$. We show that there is some set $R'(x), x \in \hat{C}'$ whose centroid $\text{ctr}(R'(x))$ lies in the ball Z_j, which will prove the lemma. By Lemma 4.8 and noting that $d_\ell^2 \leq \frac{\epsilon^2}{1-\epsilon^2} \cdot D_\ell^2$ for every ℓ, for any $x' \in Y \cap X_\ell$ where $\ell \neq j$, we have $\|x' - c_j\| \leq \|x' - c_\ell\| + \frac{\epsilon}{\sqrt{\rho(1-\epsilon^2)}} \cdot \max(D_\ell + 6D_j, 4D_\ell + 3D_j)$. Also $D_j, D_\ell \leq \|c_j - c_\ell\| \leq \|x' - c_j\| + \|x' - c_\ell\|$. Substituting for D_j, D_ℓ we get that $\|x' - c_j\| \leq \beta\|x' - c_\ell\|$, where $\beta = \frac{1 + 7\epsilon/\sqrt{\rho(1-\epsilon^2)}}{1 - 7\epsilon/\sqrt{\rho(1-\epsilon^2)}}$. Using this, we obtain that $A = \sum_{x' \in Y} \|x' - c_j\|^2 \leq \beta^2 \sum_{\ell=1}^{k} \sum_{x' \in Y \cap X_\ell} \|x' - c_\ell\|^2 \leq \beta^2 \Delta_k^2(X)$. We also have

$$\begin{aligned} A = \sum_{x \in S} \sum_{x' \in R'(x)} \|x' - c_j\|^2 &= \sum_{x \in S} \Big(\Delta_1^2(R'(x)) + |R'(x)| \|\text{ctr}(R'(x)) - c_j\|^2 \Big) \\ &\geq |Y| \min_{x \in S} \|\text{ctr}(R'(x)) - c_j\|^2. \end{aligned}$$

Since $X_j^{\text{cor}} \subseteq Y$ we have $|Y| \geq (1 - \rho)n_j$, so we obtain that $\min_{x \in S} \|\text{ctr}(R'(x)) - c_j\| \leq \frac{\beta}{\sqrt{1-\rho}} \cdot d_i$. The bounds on ρ ensure that $\frac{\rho\beta^2}{1-\rho} \leq 1$, so that $\min_{x \in S} \|\text{ctr}(R'(x)) - c_j\| \leq \frac{d_j}{\sqrt{\rho}}$. $\square$

COROLLARY 4.10. *After the deletion phase, for every i, there is a center $\hat{c}_i \in \hat{C}$ with $\|\hat{c}_i - c_i\| \leq \frac{\epsilon}{\sqrt{\rho(1-\epsilon^2)}} \cdot D_i$.*

4.1.3. A linear time seeding procedure. We now combine the sampling idea with the deletion procedure to obtain an initialization procedure that runs in time $O(nkd + k^3d)$ and succeeds with high probability. We first sample $O(k)$ points from X using the sampling procedure. Then we run the deletion procedure on an $O(k)$-size instance consisting of the centroids of the Voronoi regions of the sampled, points, with each centroid having a weight equal to the mass of X in its corresponding Voronoi region. The sampling process will ensure that with high probability, every cluster X_i contains a point $\hat{c}_i$ that is close to its center c_i. This will allow us to argue that the $\Delta_k^2(.)$ cost of the sampled instance is much smaller than its $\Delta_{k-1}^2(.)$ cost, and that the optimal centers for the sampled instance lie near the optimal centers for X. We can then use the analysis of the previous section to argue that after the deletion procedure the k centers are still quite close to the optimal centers for the sampled instance, and hence also close to the optimal centers for X. Fix $\rho_1 = \sqrt{\epsilon}$.

B1. *Sampling.* Sample $N = \frac{2k}{1-5\rho_1} + \frac{2\ln(2/\rho_1)}{(1-5\rho_1)^2}$ points from X using the sampling procedure of Section 4.1.1. Let S denote the set of sampled points.

B2. *Deletion Phase.* For each $x \in S$, let $R(x) = \{y \in X : \|y - x\| = \min_{z \in S} \|y - z\|\}$ be its Voronoi set (with respect to the sampled points). We now ignore X, and consider a weighted instance $\hat{S}$ obtained as follows: set $\hat{S} \leftarrow \{\hat{x} = \mathrm{ctr}(R(x)) : x \in S\}$, and assign each $\hat{x}$ a *weight* $w(\hat{x}) = |R(x)|$. Run the deletion procedure of Section 4.1.2, on this new instance to obtain k centers $\hat{c}_1, \ldots, \hat{c}_k$.

Running Time. Step B1 takes $O(nNd) = O(nkd)$ time. The run-time analysis of the deletion phase in Section 4.1.2, shows that Step B2 takes $O(N^3d) = O(k^3d)$ time. So the overall running time is $O(nkd + k^3d)$.

Analysis. Recall that $\rho_1 = \sqrt{\epsilon}$. Let $\rho_2 = \rho_1^3$. For the analysis to go through, we need that $\epsilon \leq 6 \cdot 10^{-7}$. Let $X_i^{\mathrm{cor}} = \{x \in X_i : \|x - c_i\|^2 \leq \frac{r_i^2}{\rho_1}\}$. Let $X_i^{\mathrm{out}} = \{x \in X_i : \|x - c_i\|^2 \leq \frac{r_i^2}{\rho_2}\}$ denote the outer core of cluster X_i. By part (ii) of Corollary 4.6, we know that with probability $1 - O(\rho_1)$, every cluster X_i contains a sampled point in its outer core after Step B1. So assume that this event happens. Let $\hat{s}_1, \ldots, \hat{s}_k$ denote the optimal k centers for $\hat{S}$ and $\hat{c}_1, \ldots, \hat{c}_k$ be the centers returned by the deletion phase. Lemma 4.11 shows that the k-means separation condition also holds for $\hat{S}$, and the optimal centers for $\hat{S}$ are close to the optimal centers for X. This will imply that the centers returned by the deletion phase are close to the optimal centers for X.

LEMMA 4.11.

(i) $\Delta_k^2(\hat{S}) = \hat{\epsilon}^2 \Delta_{k-1}^2(\hat{S})$ *where* $\hat{\epsilon}^2 \leq \frac{1}{4001}$.

(ii) For every optimal center c_i *of* X*, there is a center* $\hat{s}_i$ *such that* $\|\hat{s}_i - c_i\| \leq \frac{D_i}{25} + \frac{r_i}{\sqrt{\rho_1}}$.

PROOF. For each sampled point $x \in S$, recall that $R(x) \subseteq X$ denotes its Voronoi set (with respect to S). For $j = 1, \ldots, k$, let $z_j \in S$ be a sampled point in X_j^{out}, so $\|z_j - c_j\| \leq \frac{r_j}{\sqrt{\rho_2}}$. Consider an optimal $(k-1)$-clustering of $\hat{S}$. We can obtain a $(k-1)$-clustering of X from this by assigning all the points in $R(x)$, where $x \in S$, to the center to which $\mathrm{ctr}(R(x)) \in \hat{S}$ is assigned. The cost-increase in doing so is exactly $A = \sum_{x \in S} \Delta_1^2(R(x))$, so $\Delta_{k-1}^2(X) \leq \Delta_{k-1}^2(\hat{S}) + A$. Since $\|y - x\|^2 \leq \|y - z_j\|^2 \leq 2\big(\|y - c_j\|^2 + \frac{r_j^2}{\rho_2}\big)$ for any $y \in R(x) \cap X_j$, we obtain that $A \leq 2\big(1 + \frac{1}{\rho_2}\big)\Delta_k^2(X)$. To upper bound $\Delta_k^2(\hat{S})$, consider the following k-clustering of $\hat{S}$: for each $\hat{x} = \mathrm{ctr}(R(x)) \in \hat{S}$ where $x \in S \cap X_i$, assign $\hat{x}$ to center c_i. To bound the cost of this assignment, first note that for a point $y \in R(x) \cap X_j$

where $x \in X_i$ and $j \neq i$, we have

$\|y-c_i\| \le \|y-x\|+\|x-c_i\| \le \|y-x\|+\|x-c_j\| \le 2\|y-x\|+\|y-c_j\| \le 3\|y-c_j\|+2\|z_j-c_j\|$.

We also have $\|z_j - c_j\| \le \frac{r_j}{\sqrt{\rho_2}} \le \frac{\epsilon D_j}{\sqrt{\rho_2(1-\epsilon^2)}}$ and $D_j \le \|y - c_i\| + \|y - c_j\|$, which implies that $\|y - c_i\| \le \beta\|y - c_j\|$ where $\beta = \frac{3+2\epsilon/\sqrt{\rho_2(1-\epsilon^2)}}{1-2\epsilon/\sqrt{\rho_2(1-\epsilon^2)}} \le 4$. Thus,

$$\begin{aligned}\Delta_k^2(\hat{S}) &\le \sum_{i=1}^{k}\sum_{x\in S\cap X_i} |R(x)|\|\mathrm{ctr}(R(x)) - c_i\|^2 \le \sum_{i=1}^{k}\sum_{x\in S\cap X_i}\sum_{y\in R(x)} \|y - c_i\|^2 \\ &\le \sum_{i=1}^{k}\sum_{x\in S\cap X_i}\left(\sum_{y\in R(x)\cap X_i}\|y - c_i\|^2 + \sum_{j\neq i, y\in R(x)\cap X_j}\beta^2\|y - c_j\|^2\right). \\ &\le \beta^2\Delta_k^2(X).\end{aligned}$$

Combining the two bounds, we get,

$$\Delta_{k-1}^2(S) \ge \Delta_{k-1}^2(X) - A \ge \left(\frac{1}{\epsilon^2} - 2 - \frac{2}{\rho_2}\right)\Delta_k^2(X) \ge \frac{1/\epsilon^2 - 2 - 2/\rho_2}{\beta^2}\cdot\Delta_k^2(\hat{S}).$$

Since $\rho_1 = \sqrt{\epsilon}$ and $\rho_2 = \rho_1^3$ (and ϵ is tiny), we get that $\Delta_k^2(\hat{S}) = \hat{\epsilon}^2\Delta_{k-1}^2(\hat{S})$ where $4001\hat{\epsilon}^2 \le 1$. This proves part (i).

Consider any center c_i. Suppose $\|\hat{s}_j - c_i\| > D_i/25 + \frac{r_i}{\sqrt{\rho_1}}$ for every point $\hat{s}_j$. Then the cost of clustering X around the centers $\hat{s}_1, \dots, \hat{s}_k$ is at least $\frac{1-\rho_1}{625}\cdot n_i D_i^2 \ge \frac{1-\rho_1}{625}(\frac{1}{\epsilon^2}-1)\Delta_k^2(X)$ (since $n_i D_i^2 \ge \Delta_{k-1}^2(X) - \Delta_k^2(X)$). On the other hand, we also have that the cost of this clustering for X is at most $\Delta_k^2(\hat{S})+A \le \left[2(1+\frac{1}{\rho_2})+\beta^2\right]\Delta_k^2(X)$. The bound on ϵ ensures that this upper bound contradicts the earlier lower bound (on the cost of this clustering). □

LEMMA 4.12. *For each center c_i, there is a center $\hat{c}_i$ such that $\|\hat{c}_i - c_i\| \le \frac{D_i}{10}$.*

PROOF. Notice that mapping each center c_i to its closest center, say $\hat{s}_i$, in $\hat{S}$, gives a one-one map; this follows from part (ii) of Lemma 4.11 and Lemma 4.1 (recall that $D_i = \min_{j\neq i}\|c_i - c_j\|$). Let $\hat{D}_i = \min_{j\neq i}\|\hat{s}_j - \hat{s}_i\|$. Then, $(1 - 2\theta) \le \frac{\hat{D}_i}{D_i} \le (1 + 2\theta)$, where $\theta \le \left(\frac{1}{25} + \frac{\epsilon}{\sqrt{\rho_1(1-\epsilon^2)}}\right)$ (since letting $A = \left(\frac{D_i}{25} + \frac{r_i}{\sqrt{\rho_1}}\right) + \left(\frac{D_j}{25} + \frac{r_j}{\sqrt{\rho_1}}\right)$, we have $\|c_i - c_j\| - A \le \|\hat{s}_i - \hat{s}_j\| \le \|c_i - c_j\| + A$ for all $i, j \neq i$ by Lemma 4.11). Since $\rho_1 = \sqrt{\epsilon}$, we have that $\theta < \frac{1}{22}$. By part (i) of Lemma 4.11, we know that $\hat{S}$ is $\hat{\epsilon}$-separated for k-means where $4001\hat{\epsilon}^2 \le 1$. So we can choose ρ for the deletion phase so that $\rho \le \frac{1}{10}$, $\frac{\hat{\epsilon}}{\sqrt{\rho(1-\hat{\epsilon}^2)}} \le \frac{1}{20}$, and, by Corollary 4.10, ensure that the deletion phase returns a point $\hat{c}_i$ such that $\|\hat{c}_i - \hat{s}_i\| \le \frac{\hat{D}_i}{20}$. Thus, using Lemma 4.11,

$$\|\hat{c}_i - c_i\| \le \frac{D_i}{20}\cdot\frac{12}{11} + \frac{D_i}{25} + \frac{r_i}{\sqrt{\rho_1}} \le \frac{3D_i}{55} + D_i\cdot\theta \le \frac{D_i}{10}. \quad \square$$

4.2. Procedures Used in Stage II

Given k seed centers $\hat{c}_1, \dots, \hat{c}_k$ located sufficiently close to the optimal centers after stage I, we use two procedures in stage II to obtain a near-optimal clustering: the ball-k-means step, which yields a $(1 + O(\epsilon^2))$-approximation algorithm, or the *centroid estimation* step, based on a sampling idea of Kumar et al. [2004], which yields a PTAS with running-time exponential in k. Define $\hat{d}_i = \min_{j\neq i}\|\hat{c}_j - \hat{c}_i\|$. Recall that $D_i = \min_{j\neq i}\|c_j - c_i\|$.

*(A) Ball-**k**-Means Step.* Let B_i be the points of X in a ball of radius $\hat{d}_i/3$ around $\hat{c}_i$, and $\bar{c}_i$ be the centroid of B_i. Return $\bar{c}_1, \dots, \bar{c}_k$ as the final centers.

LEMMA 4.13 (BALL-k-MEANS). *Suppose that, for each i, there is a center $\hat{c}_i$ such that $\|\hat{c}_i - c_i\| \le D_i/10$. Let $\rho = \frac{36\epsilon^2}{1-\epsilon^2}$ and $Y_i = \{x \in X_i : \|x - c_i\|^2 \le \frac{r_i^2}{\rho}\}$. Then, $Y_i \subseteq B_i \subseteq X_i$, and $\|\bar{c}_i - c_i\|^2 \le \frac{\rho}{1-\rho} \cdot r_i^2$.*

The proof of this lemma is essentially identical to that of Lemma 3.3, and hence is omitted.

(B) Centroid Estimation. For each i, we will obtain a set of candidate centers for cluster X_i as follows. Fix $\beta = \frac{1}{1+144\epsilon^2}$. Define the expanded Voronoi region of $\hat{c}_i$ as follows: for any $x \in X$, let $\hat{c}(x)$ denote the center such that $\|x - \hat{c}(x)\| = \min_j \|x - \hat{c}_j\|$. Let $R_i' \subseteq X = \{x \in X : \|x - \hat{c}_i\| \le \|x - \hat{c}(x)\| + \|\hat{c}_i - \hat{c}(x)\|/4\}$. Sample $\frac{4}{\beta\omega}$ points independently and uniformly at random from R_i', where ω is a given input parameter, to obtain a random subset $S_i \subseteq R_i'$. Compute the centroid of every subset of S_i of size $\frac{2}{\omega}$; let T_i be the set consisting of all these centroids. Select the candidates $\bar{c}_1 \in T_1, \dots, \bar{c}_k \in T_k$ that yield the least-cost solution, and return these as the final centers.

LEMMA 4.14 (CENTROID-ESTIMATION). *Suppose that for each i, there is a center $\hat{c}_i$ such that $\|\hat{c}_i - c_i\| \le D_i/10$. Then $X_i \subseteq R_i'$, where R_i' is as defined in the centroid-estimation procedure, and $|X_i| \ge \beta |R_i'|$.*

PROOF. For any $j \ne i$, we have $\frac{4}{5} \cdot \|c_i - c_j\| \le \|\hat{c}_i - \hat{c}_j\| \le \frac{6}{5} \cdot \|c_i - c_j\|$. Hence, $\frac{4D_i}{5} \le \hat{d}_i \le \frac{6D_i}{5}$. Consider any $x \in X_i$ that lies in the Voronoi region of $\hat{c}_j$ (so $\hat{c}(x) = \hat{c}_j$). We have $\|x - c_i\| \le \|x - c_j\|$, therefore $\|x - \hat{c}_i\| \le \|x - \hat{c}_j\| + \frac{D_i + D_j}{10} \le \|x - \hat{c}_j\| + \|c_i - c_j\|/5 \le \|x - \hat{c}_j\| + \|\hat{c}_i - \hat{c}_j\|/4$; so $x \in R_i'$. Suppose $|X_i| \le \beta |R_i'|$. Let $a_j = \frac{|R_i' \cap X_j|}{|R_i'|}$. So $\frac{a_i}{1-a_i} \le \frac{\beta}{1-\beta}$. Consider the clustering where we arbitrarily assign some $n_i \cdot \frac{a_j}{1-a_i}$ points of X_i to center c_j for each $j \ne i$. For any $x \in X_i$ and $j \ne i$, we have $\|x - c_j\|^2 \le 2(\|x - c_i\|^2 + \|c_i - c_j\|^2)$. So the cost of reassigning points in X_i is at most $2\Delta_1^2(X_i) + \frac{2n_i}{1-a_i} \cdot \sum_{j \ne i} a_j \|c_i - c_j\|^2 \le 2\Delta_1^2(X_i) + \frac{2\beta}{1-\beta} \cdot \sum_{j \ne i} a_j |R_i'| \|c_i - c_j\|^2$. We also know that for any $y \in R_i' \cap X_j$,

$$\begin{aligned} \|y - c_i\| &\le \|y - \hat{c}(y)\| + \frac{\|\hat{c}_i - \hat{c}(y)\|}{4} + \frac{D_i}{10} \le \|y - \hat{c}_j\| + \frac{\|y - \hat{c}_i\|}{2} + \frac{D_i}{10} \\ \implies \frac{\|y - c_i\|}{2} &\le \|y - c_j\| + \frac{1.5D_i + D_j}{10}. \end{aligned}$$

Since $D_i, D_j \le \|c_i - c_j\|$, this in turn implies that $\|y - c_i\| \le 2\|y - c_j\| + \|c_i - c_j\|/2$, which implies that $\|c_i - c_j\| \le 6 \cdot \|y - c_j\|$. Therefore, we can bound $a_j |R_i'| \|c_i - c_j\|^2$ by $36 \cdot \sum_{y \in R_i' \cap X_j} \|y - c_j\|^2$. Hence, the cost of this clustering is at most $\max(2, 1 + \frac{72\beta}{1-\beta}) \Delta_k^2(X) \le (1 + \frac{1}{2\epsilon^2}) \Delta_k^2(X)$. The cost of this clustering is also at least $\Delta_{k-1}^2(X)$. This is a contradiction to the assumption that $\Delta_k^2(X) \le \epsilon^2 \Delta_{k-1}^2(X)$. $\square$

4.3. A Linear Time Constant-Factor Approximation Algorithm

This algorithm uses the initialization procedure of Section 4.1.3 followed by a ball-k-means step, and hence runs in time $O(nkd + k^3 d)$.

(1) Execute the seeding procedure of Section 4.1.3 to obtain k initial centers $\hat{c}_1, \dots, \hat{c}_k$.
(2) Run the *ball-k-means step* of Section 4.2 to obtain the final centers.

By Lemma 4.12, we know that with probability $1 - O(\sqrt{\epsilon})$, for each c_i, there is a distinct center $\hat{c}_i$ such that $\|\hat{c}_i - c_i\| \le D_i/10$. Therefore, by Lemma 4.13, for each c_i, we have $\|\bar{c}_i - c_i\|^2 \le \frac{\rho}{1-\rho} \cdot r_i^2$. Hence, by mimicking the proof of Theorem 3.4, we obtain the following theorem.

THEOREM 4.15. *Assuming that $\Delta_k^2(X) \le \epsilon^2 \Delta_{k-1}^2(X)$ for a small enough ϵ, this algorithm returns a solution of cost at most $\frac{1-\epsilon^2}{1-37\epsilon^2} \cdot \Delta_k^2(X)$ with probability $1 - O(\sqrt{\epsilon})$ in time $O(nkd + k^3 d)$.*

4.4. A PTAS for Any Fixed k

The PTAS combines the sampling procedure of Section 4.1.1 (we could also use the seeding procedure of Section 4.1.3) with the centroid estimation step described in Section 4.2.

(1) Use the procedure in Section 4.1.1 to pick k initial centers $\hat{c}_1, \ldots, \hat{c}_k$.
(2) Run the *centroid estimation* procedure of Section 4.2 to obtain the final centers.

The running time is dominated by the exhaustive search in the centroid estimation procedure, which takes time $O\big(2^{(4k/\beta\omega)} nd\big)$. We show that the cost of the final solution is at most $(1+\omega)\Delta_k^2(X)$, with probability γ^k for some constant γ. By repeating the procedure $O(\gamma^{-k})$ times, we can boost this to a constant.

THEOREM 4.16. *Assuming that $\Delta_k^2(X) \le \epsilon^2 \Delta_{k-1}^2(X)$ for a small enough ϵ, there is a PTAS for the k-means problem that returns a $(1+\omega)$-optimal solution with constant probability in time $O(2^{O(k(1+\epsilon^2)/\omega)} nd)$.*

PROOF. By part (i) of Corollary 4.6, we know that with probability $\Theta(1)^k$, the sampling process returns centers $\hat{c}_1, \ldots, \hat{c}_k$ such that for each i, $\|\hat{c}_i - c_i\| \le D_i/10$ (assuming ϵ is suitably small). So, by Lemma 4.14, we know that $|X_i| \ge \beta |R_i'|$ for every i, where recall that $\beta = \frac{1}{1+144\epsilon^2}$. Now Kumar et al. [2004, Lemma 2.3] shows that for every i, with constant probability, there is some candidate point $c_i' \in T_i$ such that $\sum_{x \in X_i} \|x - c_i'\|^2 \le (1+\omega)\Delta_1^2(X_i)$. The cost of the best-candidate solution is at most the cost of the solution due to the points $c_1' \in T_1, \ldots, c_k' \in T_k$, which is at most $(1+\omega)\Delta_k^2(X)$. The overall success probability for one call of the procedure is γ^k for some constant $\gamma < 1$, so by repeating the procedure $O(\gamma^{-k})$ times we can obtain constant success probability. □

5. THE SEPARATION CONDITION

We now show that our separation condition implies, and is implied by, the condition that any two near-optimal k-clusterings disagree on only a small fraction of the data. Let $cost(x_1, \ldots, x_k)$ denote the cost of clustering X around the centers $x_1, \ldots, x_k \in \mathbb{R}^d$. We use $R(x)$ to denote the Voronoi set of point x, that is, the points of X in the Voronoi region of x (with respect to centers that will be clear from the context). Let $S_1 \ominus S_2 = (S_1 \setminus S_2) \cup (S_2 \setminus S_1)$ denote the symmetric difference of S_1 and S_2.

THEOREM 5.1. *Suppose that $X \subseteq \mathbb{R}^d$ is ϵ-separated for k-means for a small enough ϵ. The following hold:*

(i) If there are centers $\hat{c}_1, \ldots, \hat{c}_k$ such that $cost(\hat{c}_1, \ldots, \hat{c}_k) \le \alpha \Delta_{k-1}^2(X)$, where $0 < \alpha \le \frac{1-401\epsilon^2}{400}$, then for each $\hat{c}_i$ there is a distinct optimal center $c_{\sigma(i)}$ such that $|R(\hat{c}_i) \ominus X_{\sigma(i)}| \le 161\epsilon^2 |X_{\sigma(i)}|$;

(ii) If $\hat{X}$ is a point set obtained by perturbing each $x \in X_i$ by a distance of (at most) $\frac{\epsilon \Delta_{k-1}(X)}{\sqrt{n}}$ (in any direction) then $\Delta_k^2(\hat{X}) = O(\epsilon^2)\Delta_{k-1}^2(\hat{X})$.

PROOF. Let $\rho = \left(\frac{\alpha}{\epsilon^2}+1\right)^{-1}$. Note that $\rho \geq \frac{400\epsilon^2}{1-\epsilon^2}$. Define $X_i^{\text{cor}} = \{x \in X_i : \|x - c_i\| \leq \frac{r_i}{\sqrt{\rho}}\}$, and let $d_i'^2 = \epsilon^2 \Delta_{k-1}^2(X)/n_i$. Note that $r_i^2 \leq d_i'^2 \leq \frac{\epsilon^2}{1-\epsilon^2} \cdot D_i^2$. We claim that, for every $\hat{c}_i$, there must be a distinct optimal center, call it c_i, such that $\|\hat{c}_i - c_i\| \leq \frac{2d_i'}{\sqrt{\rho}} \leq \frac{D_i}{10}$. Suppose not. Then, in the clustering around $\hat{c}_1, \ldots, \hat{c}_k$, all the points in X_i^{cor} are assigned to a center that is at least $\frac{2d_i'}{\sqrt{\rho}}$ away from c_i. Therefore, $cost(\hat{c}_1, \ldots, \hat{c}_k) > (\frac{1}{\rho} - 1)n_i d_i'^2 = \alpha \Delta_{k-1}^2(X)$ giving a contradiction. Re-indexing the clusters this way, we show that $\sigma(i) = i$ yields the desired mapping. This is because $R(\hat{c}_i)$ contains each point $x \in X_i$ such that $\|x - c_i\| \leq 2D_i/5$, and therefore $|R(\hat{c}_i) \cap X_i| \geq (1-\rho_1)|X_i|$ where $\rho_1 = \frac{25\epsilon^2}{4(1-\epsilon^2)}$. Also by Lemma 4.14, we have $|X_i| \geq \beta |R(\hat{c}_i)|$ where $\beta = \frac{1}{1+144\epsilon^2}$. Therefore, we get that $|R(\hat{c}_i) \ominus X_i| \leq (2\rho_1 + \frac{1}{\beta} - 1)|X_i| \leq 161\epsilon^2 |X_i|$ for $\epsilon \leq \frac{1}{2}$.

For a point $x \in X$, we use $\hat{x}$ to denote its perturbed image in $\hat{X}$. Note that, for any $y \in \mathbb{R}^d$, we have $\|\hat{x} - y\|^2 \leq 2\left(\|x - y\|^2 + \frac{\epsilon^2 \Delta_{k-1}^2(X)}{n}\right)$. Consider the k-clustering of $\hat{X}$ where we assign all the perturbed points of X_i to c_i. The cost of this clustering for $\hat{X}$ is at most $2\Delta_k^2(X) + 2\epsilon^2 \Delta_{k-1}^2(X)$. Conversely, one can obtain a $(k-1)$-clustering of X from an optimal $(k-1)$-clustering of $\hat{X}$ by assigning each $x \in X$ to the center to which $\hat{x}$ is assigned. Thus, we get that $\Delta_{k-1}^2(X) \leq 2\Delta_{k-1}^2(\hat{X}) + 2\epsilon^2 \Delta_{k-1}^2(X)$. Combining the two bounds, we get that $\Delta_k^2(\hat{X}) \leq \gamma \Delta_{k-1}^2(\hat{X})$ where $\gamma = \frac{8\epsilon^2}{1-2\epsilon^2} = O(\epsilon^2)$. □

THEOREM 5.2. *Let $\epsilon \leq \frac{1}{3}$. Suppose that for every k-clustering $\hat{X}_1, \ldots, \hat{X}_k$ of X of cost at most $\alpha^2 \Delta_k^2(X)$,*

(i) there exists a bijection σ such that $\forall i,\ |\hat{X}_i \ominus X_{\sigma(i)}| \leq \epsilon |X_{\sigma(i)}|$; AND/OR
(ii) there is a bijection σ such that $\sum_{i=1}^k |\hat{X}_i \ominus X_{\sigma(i)}| \leq \frac{\epsilon}{k-1}|X|$.

Then, X is α-separated for k-means.

PROOF. Let $R_1, \ldots, R_{k-1}$ be an optimal $(k-1)$-means solution. We will construct a refinement of $R_1, \ldots, R_{k-1}$ and argue that this has large Hamming distance to $X_1, \ldots, X_k$, and hence has cost at least $\alpha^2 \Delta_k^2(X)$. Since the cost of $R_1, \ldots, R_{k-1}$ is at least the cost of any refinement of it, this will imply that $\Delta_{k-1}^2(X) \geq \alpha^2 \Delta_k^2(X)$. Let R_{k-1} be the largest cluster. We start with an arbitrary refinement $R_1, \ldots, R_{k-2}, \hat{X}_{k-1}, \hat{X}_k$ where $\hat{X}_{k-1} \cup \hat{X}_k = R_{k-1},\ \hat{X}_{k-1}, \hat{X}_k \neq \emptyset$. If the cost of this k-clustering is at least $\alpha^2 \Delta_k^2(X)$ then we are done. So assume that this is not the case, and let σ be the claimed bijection.

For part (i), we introduce a large disagreement by splitting $\hat{X}_{k-1} \cap X_{\sigma(k-1)}$ and $\hat{X}_k \cap X_{\sigma(k)}$ into two equal-sized halves, $A_{k-1} \cup B_{k-1}$ and $A_k \cup B_k$ respectively, and "mismatching" them. More precisely, we claim that the clustering $R_1, \ldots, R_{k-2}, X_{k-1}' = (\hat{X}_{k-1} \setminus A_{k-1}) \cup A_k, X_k' = (\hat{X}_k \setminus A_k) \cup A_{k-1}$ has large Hamming distance. For any bijection σ', if $\sigma'(i) \neq \sigma(i)$ for $i \leq k-2$, then $|R_i \ominus X_{\sigma'(i)}| \geq |R_i \cap X_{\sigma(i)}| \geq (1-\epsilon)|R_i|$; otherwise, $\sigma'(k) \in \{\sigma(k-1), \sigma(k)\}$, so $|X_k' \ominus X_{\sigma'(k)}| \geq \frac{1-\epsilon}{2}|X_k'| \geq \epsilon |X_k'|$ since $X_k' \setminus X_{\sigma(k-1)} \supseteq B_k,\ X_k' \setminus X_{\sigma(k)} \supseteq A_{k-1}$.

For part (ii), since $|R_{k-1}| \geq \frac{|X|}{k-1}$, we have $|\hat{X}_{k-1} \cap X_{\sigma(k-1)}| + |\hat{X}_k \cap X_{\sigma(k)}| \geq \frac{1-\epsilon}{k-1}|X|$. After this mismatch operation, for any bijection σ', the total disagreement is at least $|X_{k-1}' \ominus X_{\sigma'(k-1)}| + |X_k' \ominus X_{\sigma'(k)}| \geq \frac{1}{2}\left(|\hat{X}_{k-1} \cap X_{\sigma(k-1)}| + |\hat{X}_k \cap X_{\sigma(k)}|\right) \geq \frac{1-\epsilon}{2(k-1)}|X| \geq \frac{\epsilon}{k-1}|X|$. □

6. CONCLUSIONS AND DISCUSSION

We initiate a mathematical analysis of Lloyd-style methods that attempts to explain the performance of these heuristics. We show that if the data satisfies a natural "clusterability" or "separation" condition, then various Lloyd-style methods perform well and return a near-optimal clustering. Our chief algorithmic contribution is a novel and efficient sampling procedure for seeding Lloyd's method with initial centers, such that if the data satisfies our separation condition then (even) a single Lloyd-type descent step suffices to yield a constant-factor approximation.

It may have struck the reader that there is something too good about our performance guarantees. Since we (and also Arthur and Vassilvitskii [2007][1]) need to use only one round of Lloyd's method, we cannot possibly be taking full advantage of the algorithm, in particular, its capacity to start out with a seeding that is unbalanced across clusters and correct it by shifting centers from one cluster to another. The extent to which Lloyd's method is successful at doing so is, in fact, unclear, and, for this reason, there is much literature exploring the merits of different seeding procedures. Nevertheless, we feel that Lloyd's method is better than we have accounted for, and that our results fall short of explaining (or predicting) the performance of Lloyd-style methods; instead, they suggest that our separation condition is perhaps too stringent (and too restrictive as a measure of data-clusterability). If so, then the main open question that emerges from our work is to demonstrate a condition weaker than ours, for which the initial seeding is not necessarily close to an optimal solution, but yet Lloyd's algorithm can be shown to converge in a small number of rounds to a near-optimal solution.

An orthogonal research direction is to explore further implications of our seeding method and separation condition (or similar ones) for the k-means and possibly other clustering problems. For instance, it might be possible to obtain stronger, or more general, algorithmic results. Subsequent to the publication of the conference version of this article, some results [Bunn and Ostrovsky 2007; Nissim et al. 2007; Balcan et al. 2009; Awasthi et al. 2010] have been obtained in this vein. Nissim et al. [2007] and Bunn and Ostrovsky [2007] use our results to design secure, privacy-preserving ways of computing a near-optimal k-means solution when the data satisfies our separation condition; the former exploits the robustness of our separation condition, and the latter uses our seeding method. Balcan et al. [2009] propose a different separation condition and show that if the data satisfies this condition then one can efficiently compute a solution that is close to a desired target clustering. Building upon this work, Awasthi et al. [2010] propose yet another separation condition and show that their condition is weaker than both our separation condition and the separation condition of Balcan et al. [2009] (the latter implication requires that the target clusters be "large"). They design a PTAS for the k-means problem on instances satisfying their separation condition. In particular, this implies that if the data is ϵ-separated (under our notion) for k-means, then, for any $\alpha > 0$, one can compute a $(1 + \alpha)$-optimal k-means solution in polynomial time. Furthermore, they show that it is *NP*-hard to compute an optimal solution even when the instance satisfies their separation condition (or our ϵ-separation condition).

ACKNOWLEDGMENTS

We thank the referees for their feedback and various helpful suggestions.

REFERENCES

AGGARWAL, A., DESHPANDE, A., AND KANNAN, R. 2009. Adaptive sampling for k-means clustering. In *Proceedings of the 12th International Workshop on Approximation Algorithms for Combinatorial Optimazation Problems (APPROX 2009)*. 15–28.

[1] In fact, Arthur and Vassilvitskii [2007] require no Lloyd steps at all!

AILON, N., JAISWAL, R., AND MONTELEONI, C. 2009. Streaming k-means approximation. In *Proceedings of the 23rd Annual Conference on Neural Information Processing System (NIPS)*. 10–18.

ALSABTI, K., RANKA, S., AND SINGH, V. 1998. An efficient *k*-means clustering algorithm. In *Proceedings of the 1st Workshop on High Performance Data Mining*.

ARTHUR, D., MANTHEY, B., AND RÖGLIN, H. 2011. Smoothed analysis of the k-means method. *J. ACM 58,* 5, 19.

ARTHUR, D., AND VASSILVITSKII, S. 2006. How slow is the *k*-means method? In *Proceedings of the 22nd Annual Symposium on Computational Geometry (SoCG)*. 144–153.

ARTHUR, D. AND VASSILVITSKII, S. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 1027–1035.

ARYA, V., GARG, N., KHANDEKAR, R., MEYERSON, A., MUNAGALA, K., AND PANDIT, V. 2004. Local search heuristics for *k*-median and facility location problems. *SIAM J. Comput. 33*, 544–562.

AWASTHI, P., BLUM, A., AND SHEFFET, O. 2010. Stability yields a PTAS for k-median and *k*-means clustering. In *Proceedings of the 51st Annual Symposium on Foundations of Computer Science (FOCS)*. 309–318.

BADOIU, M., HAR-PELED, S., AND INDYK, P. 2002. Approximate clustering via core-sets. In *Proceedings of the 34th Annual Symposium on Theory of Computing (STOC)*. 250–257.

BALCAN, M., BLUM, A., AND GUPTA, A. 2009. Approximate clustering without the approximation. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 1068–1077.

BRADLEY, P. S. AND FAYYAD, U. 1998. Refining initial points for *K*-means clustering. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*. 91–99.

BUNN, P. AND OSTROVSKY, R. 2007. Secure two-party *k*-means clustering. In *Proceedings of the ACM Conference on Computer and Communications Security*. 486–497.

CHARIKAR, M. AND GUHA, S. 1999. Improved combinatorial algorithms for the facility location and *k*-median problems. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*. 378–388.

CHARIKAR, M., GUHA, S., TARDOS, É., AND SHMOYS, D. B. 2002. A constant-factor approximation algorithm for the *k*-median problem. *J. Comput. Syst. Sci. 65*, 129–149.

CHROBAK, M., KENYON, C., AND YOUNG, N. 2006. The reverse greedy algorithm for the metric *k*-median problem. *Info. Proc. Lett. 97*, 68–72.

COX, D. R. 1957. Note on grouping. *J. ASA 52*, 543–547.

DASGUPTA, S. 2003. How fast is *k*-means? In *Proceedings of the 16th Annual Conference on Computational Learning Theory (COLT)*. 735.

DE LA VEGA, W. F., KARPINSKI, M., KENYON, C., AND RABANI, Y. 2003. Approximation schemes for clustering problems. In *Proceedings of the 35th ACM Symposium on Theory of Computing*. 50–58.

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Y. Stat. Soc. B 39*, 1–38.

DRINEAS, P., FRIEZE, A., KANNAN, R., VEMPALA, S., AND VINAY, V. 2004. Clustering large graphs via the Singular Value Decomposition. *Mach. Learn. 56*, 9–33.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2000. *Pattern Classification*. Wiley-Interscience.

EFFROS, M. AND SCHULMAN, L. J. 2004a. Deterministic clustering with data nets. Electronic Tech rep. ECCC TR04-050.

EFFROS, M. AND SCHULMAN, L. J. 2004b. Deterministic clustering with data nets. In *Proceedings of the International Symposium on Information Theory (ISIT)*.

FISHER, D. 1996. Iterative optimization and simplification of hierarchical clusterings. *J. Artif. Intell. Res. 4*, 147–178.

FORGEY, E. 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics 21*, 768.

GERSHO, A. AND GRAY, R. M. 1992. *Vector quantization and signal compression*. Kluwer.

GRAY, R. M. AND NEUHOFF, D. L. 1998. Quantization. *IEEE Trans. Inform. Theory 44,* 6, 2325–2384.

HAR-PELED, S. AND MAZUMDAR, S. 2004. On coresets for *k*-means and *k*-median clustering. In *Proceedings of the 36th Annual Symposium on Theory of Computing (STOC)*. 291–300.

HAR-PELED, S., AND SADRI, B. 2005. How fast is the *k*-means method? *Algorithmica 41*, 185–202.

HIGGS, R. E., BEMIS, K. G., WATSON, I. A., AND WIKEL, J. H. 1997. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comp. Sci. 37*, 861–870.

JAIN, A. K., MURTY, M. N., AND FLYNN, P. J. 1999. Data clustering: A review. *ACM Comput. Surv. 31,* 3.

JAIN, K., MAHDIAN, M., MARKAKIS, E., SABERI, A., AND VAZIRANI, V. 2003. Greedy facility location algorithms analyzed using dual-fitting with factor-revealing LP. *J. ACM 50*, 795–824.

JAIN, K. AND VAZIRANI, V. 2001. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. *J. ACM 48*, 274–296.

KANUNGO, T., MOUNT, D., NETANYAHU, N., PIATKO, C., SILVERMAN, R., AND WU, A. 2004. A local search approximation algorithm for k-means clustering. *Comput. Geom. 28*, 89–112.

KANUNGO, T., MOUNT, D. M., NETANYAHU, N. S., PIATKO, C. D., SILVERMAN, R., AND WU, A. Y. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 881–892.

KAUFMAN, L. AND ROUSSEEUW, P. J. 1990. *Finding Groups in Data. An Introduction to Cluster Analysis*. Wiley.

KUMAR, A., SABHARWAL, Y., AND SEN, S. 2004. A simple linear time $(1 + \varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Proceedings of the 45th Annual Symposium on Foundations of Computer Science (FOCS)*. 454–462.

LINDE, Y., BUZO, A., AND GRAY, R. M. 1980. An algorithm for vector quantization design. *IEEE Trans. Commun. COM-28*, 84–95.

LLOYD, S. P. 1982. Least squares quantization in PCM. *Special issue on quantization, IEEE Trans. Inform. Theory 28*, 129–137.

MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. 281–297.

MATOUSEK, J. 2000. On approximate geometric k-clustering. *Disc. Computat. Geom. 24*, 61–84.

MAX, J. 1960. Quantizing for minimum distortion. *IEEE Trans. Inform. Theory IT-6,* 1, 7–12.

MEILA, M. AND HECKERMAN, D. 1998. An experimental comparison of several clustering and initialization methods. In *Proceedings of the 14th Conference on Uncertainty in Arificial Intelligent (UAI)*. 386–395.

METTU, R. R. AND PLAXTON, C. G. 2004. Optimal time bounds for approximate clustering. *Mach. Learn. 56*, 35–60.

MILLIGAN, G. 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika 45*, 325–342.

MOTWANI, R. AND RAGHAVAN, P. 1995. *Randomized Algorithms*. Cambridge Univ. Press.

NISSIM, K., RASHKHODNIKOVA, S., AND SMITH, A. 2007. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual Symposium on Theory of Computing (STOC)*. 75–84.

OSTROVSKY, R. AND RABANI, Y. 2002. Polynomial time approximation schemes for geometric clustering problems. *J. ACM 49,* 2, 139–156.

OSTROVSKY, R., RABANI, Y., SCHULMAN, L., AND SWAMY, C. 2006. The effectiveness of Lloyd-type methods for the k-means problem. In *Proceedings of the 47th Annual Symposium on Foundation of Computer Science (FOCS)*. 165–174.

PAPADIMITRIOU, C., RAGHAVAN, P., TAMAKI, H., AND VEMPALA, S. 2000. Latent semantic indexing: A probabilistic analysis. *J. Comput. Syst. Sci. 61*, 217–235.

PELLEG, D. AND MOORE, A. 1999. Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the 5th Annual ACM Conference on Knowledge Discovery and Data Mining (KDD)*. 277–281.

PENA, J. M., LOZANO, J. A., AND LARRANAGA, P. 1999. An empirical comparison of four initialization methods for the k-means algorithm. *Patt. Rec. Lett. 20*, 1027–1040.

PHILLIPS, S. J. 2002. Acceleration of k-means and related clustering problems. In *Proceedings of the 4th Workshop on Algorithm Engineering and Experiments (ALENEX)*.

SCHULMAN, L. J. 2000. Clustering for edge-cost minimization. In *Proceedings of the 32nd Annual Symposium on Theory of Computing (STOC)*. 547–555.

SNAREY, M., TERRETT, N. K., WILLET, P., AND WILTON, D. J. 1997. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model. 15*, 372–385.

SPIELMAN, D. AND TENG, S. 2001. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In *Proceedings of 33rd Annual Symposium on Theory of Computing (STOC)*. 296–305.

STEINHAUS, H. 1956. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci. C1. III vol IV*, 801–804.

TRYON, R. C. AND BAILEY, D. E. 1970. *Cluster Analysis*. McGraw-Hill. 147–150.

VATTANI, A. 2011. k-means requires exponentially many iterations even in the plane. *Disc. Comput. Geom. 45,* 4, 596–616.

Received March 2010; revised April 2012 and September 2012; accepted September 2012