The Effectiveness of Lloyd-type Methods for the k-Means Problem

Rafail Ostrovsky* rafail@cs.ucla.edu
Yuval Rabani† rabani@cs.technion.ac.il
Leonard J. Schulman‡ schulman@caltech.edu
Chaitanya Swamy§ cswamy@ist.caltech.edu

Abstract

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.

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 paper.

As in other prominent cases [**?**] 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 [**?**]) 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 (for example, 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

*Computer Science Department and Department of Mathematics, UCLA. Supported in part by IBM Faculty Award, Xerox Innovation Group Award, a gift from Teradata, Intel equipment grant, and NSF Cybertrust grant No. 0430254.

†Computer Science Department, Technion — Israel Institute of Technology, Haifa 32000, Israel. Part of this work was done while visiting UCLA and Caltech. Supported in part by ISF 52/03, BSF 2002282, and the Fund for the Promotion of Research at the Technion.

‡Caltech, Pasadena, CA 91125. Supported in part by NSF CCF-0515342, NSA H98230-06-1-0074, and NSF ITR CCR-0326554.

§Center for the Math. of Information, Caltech, Pasadena, CA 91125.

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 on high-dimensional data. In this paper, we take the first such step. We show that *if the data is well-clusterable* according to a certain "clusterability" or "separation" condition (that we introduce and justify), *then various Lloyd-style methods do indeed perform well and return a provably near-optimal clustering*. Our contributions are threefold:

(1) 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 [**?**].)
One motivation for proposing this condition is that a significant drop in the k-clustering cost is already used in practice as a diagnostic for choosing the value of k (see [**?**] §10.10). Furthermore, we show (in Section 5) 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.

(2) 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.

(3) 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, Sabharwal and Sen [**?**], whose whose running time is linear in $|X|$ and the dimension, and exponential in k. This PTAS is significantly faster, and also simpler, than the PTAS of Kumar et al. [**?**] (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 in 1956 [**?**]. A simple greedy iteration to minimize cost was suggested in 1957 by Lloyd [**?**] (and less methodically in the same year by Cox [**?**]; also apparently by psychologists between 1959-67 [**?**]). This and similar iterative descent methods soon became the dominant approaches to the problem [**?**, **?**, **?**, **?**] (see also [**?**, **?**, **?**] and the references therein); they remain so today, and are still being improved [**?**, **?**, **?**, **?**]. Lloyd's method (in any variant) converges only to local optima however, and is sensitive to the choice of the initial centers [**?**]. Consequently, a lot of research has been directed toward seeding methods that try to start off Lloyd's method with a good initial configuration [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. 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 [**?**, **?**, **?**] and in particular, [**?**] shows that Lloyd's method can require a superpolynomial number of iterations to converge.

The k-means problem is *NP*-hard even for $k = 2$ [**?**]. Recently there has been substantial progress in developing approximation algorithms for this problem. Matoušek [**?**] 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 [**?**, **?**, **?**, **?**, **?**, **?**, **?**] with varying runtime dependency on n, k and the dimension. The most recent of these is the algorithm of Kumar, Sabharwal and Sen [**?**], which presents a linear time

PTAS for a fixed k. There are also various constant-factor approximation algorithms for the related *k-median* problem [**?**, **?**, **?**, **?**, **?**], which also yield approximation algorithms for k-means, and have running time polynomial in n, k and the dimension; recently Kanungo et al. [**?**] adapted the k-median algorithm of [**?**] 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 [**?**] ! 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. Define $D_i = \min_{j \neq i} \|c_j - c_i\|$. We assume throughout that X is ϵ-separated for k-means, that is, $\Delta_k^2(X) \leq \epsilon^2 \Delta_{k-1}^2(X)$, where $0 < \epsilon \leq \epsilon_0$ with ϵ_0 being a suitably small constant. We use the following basic lemmas quite frequently.

Lemma 2.1 *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 2.2 *Let $S \subseteq \mathbb{R}^d$, and $S_1 \cup S_2$ be a partition of S with $S_1 \neq \emptyset$. Let $s = \mathrm{ctr}(S)$, $s_1 = \mathrm{ctr}(S_1)$, and $s_2 = \mathrm{ctr}(S_2)$. 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

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

The last equality follows from Lemma 2.1 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, and so the optimal 1-means cost of this point set is given by $a\|s_1 - s\|^2 + b\|s_2 - s\|^2$. 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. ■

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 $\big(1 + f(\epsilon)\big)\Delta_2^2(X)$ in linear time, for a suitably defined function f that satisfies $\lim_{\epsilon \to 0} f(\epsilon) = 0$. 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 $1 + f(\epsilon)$) 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 *non-uniform* 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., [**?**] 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 [**?**], 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 $\frac{\sum_{y\in X}\|x-y\|^2}{\sum_{x,y\in X}\|x-y\|^2} = \big(\Delta_1^2(X) + n\|x-c\|^2\big)/2n\Delta_1^2(X)$ (using Lemma 2.1); (b) pick the second center by choosing point $y \in X$ with probability equal to $\|y-\hat{c}_1\|^2/\big(\Delta_1^2(X) + n\|c-\hat{c}_1\|^2\big)$. 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 $\frac{\|x_1-x_2\|^2}{\sum_{\{x,y\}\subseteq X}\|x-y\|^2}$. 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 inter-cluster 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 $\max(r_1^2, r_2^2) \le \frac{\epsilon^2}{1-\epsilon^2}\|c_1 - c_2\|^2$.

Proof: By part (i) of Lemma 2.2 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 $r_1^2\cdot\frac{n}{n_2} + r_2^2\cdot\frac{n}{n_2} \le \frac{\epsilon^2}{1-\epsilon^2}\|c_1-c_2\|^2$. ■

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

Lemma 3.2 $\Pr[\hat{c}_1, \hat{c}_2$ *lie in distinct cores*$] = 1 - O(\rho)$.

Proof: To simplify our expressions, we assume that the points are scaled by $\frac{1}{\|c_1-c_2\|}$. We have $\Delta_1^2(X) = \Delta_2^2(X) + \frac{n_1n_2}{n}\cdot\|c_1-c_2\|^2$ by part (i) of Lemma 2.2, so $\Delta_1^2(X) \le \frac{n_1n_2}{n(1-\epsilon^2)}$. Let $c_i' = \mathrm{ctr}(X_i^{\mathrm{cor}})$. By part (ii) of Lemma 2.2 (with $S = X_i$, $S_1 = X_i^{\mathrm{cor}}$), $\|c_i' - c_i\|^2 \le \frac{\rho}{1-\rho}\cdot r_i^2$. The probability of the stated event is A/B where $A = \sum_{x\in X_1^{\mathrm{cor}}, y\in X_2^{\mathrm{cor}}}\|x-y\|^2 = |X_1^{\mathrm{cor}}|\Delta_1^2(X_2^{\mathrm{cor}}) + |X_2^{\mathrm{cor}}|\Delta_1^2(X_1^{\mathrm{cor}}) + |X_1^{\mathrm{cor}}||X_2^{\mathrm{cor}}|\|c_1'-c_2'\|^2$, and $B = \sum_{\{x,y\}\subseteq X}\|x-y\|^2 = n\Delta_1^2(X) \le \frac{n_1n_2}{1-\epsilon^2}$. By the bounds on $\|c_i'-c_i\|$ and Lemma 3.1, we get $\|c_1'-c_2'\| \ge 1 - 2\epsilon\sqrt{\frac{\rho}{(1-\rho)(1-\epsilon^2)}}$. So $A/B = \big(1-O(\rho)\big)$. ■

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.

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

Proof: By Lemma 3.1 and the definition of X_i^{cor}, we have $\|\hat{c}_i - c_i\| \le \theta\|c_1-c_2\|$ for $i = 1, 2$ where $\theta = \frac{\epsilon}{\sqrt{\rho(1-\epsilon^2)}} \le \frac{1}{10}$. 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 $x \in X_i$. For any $x \in X_i^{\mathrm{cor}}$, $\|x-\hat{c}_i\| \le 2\theta\|c_1-c_2\| \le \frac{\hat{d}}{3}$, so $x \in B_i$. Now by part (ii) of Lemma 2.2, with $S = X_i$ and $S_1 = B_i$, we get $\|\bar{c}_i - c_i\|^2 \le \frac{\rho}{1-\rho}\cdot r_i^2$ since $|B_i| \ge |X_i^{\mathrm{cor}}|$. ■

Theorem 3.4 *The above 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 $\rho = \Theta(\epsilon^2)$.*

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

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 $O\big(2^{O(k/w)}nd\big)$. 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 inter-cluster 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 employ three simple seeding procedures, with varying time vs. quality guarantees, that exploit this separation condition to 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 $O(n^3 d)$. Our deletion procedure is similar to the *reverse greedy algorithm* proposed by Chrobak, Kenyon and Young [**?**] for the k-median problem. Chrobak et al. 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 $\Delta_k^2(X) \leq \epsilon^2 \Delta_{k-1}^2(X)$, we show that our greedy deletion procedure followed by a clean-up step (in the second stage) yields a $\big(1 + f(\epsilon)\big)$-approximation algorithm.Finally, in Section 4.1.3 we combine the sampling and deletion procedures to obtain an $O(nkd + k^3 d)$-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 $\big(1 + f(\epsilon)\big)\Delta_k^2(X)$ 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 $O(nkd + k^3 d)$. 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. [**?**]. 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. (Lemma 2.3 in [**?**]), 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.; 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. 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.

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 $i \geq 2$ centers $\hat{c}_1, \ldots, \hat{c}_i$. 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 $\hat{c}_{i+1}$.

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 $\epsilon^2 \ll \rho < 1$ be a parameter that we will set later. As in the 2-means case, we define the core of cluster X_i as $X_i^{\mathrm{cor}} = \big\{x \in X_i : \|x - c_i\|^2 \leq \frac{r_i^2}{\rho}\big\}$. We show that under our separation assumption, the above 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 that if we sample more than k, but still $O(k)$, points, 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 4.1 *With probability $1 - O(\rho)$, the first two centers $\hat{c}_1, \hat{c}_2$ lie in the cores of different clusters, that is, $\Pr[\bigcup_{i \neq j}(x \in X_i^{\mathrm{cor}}$ and $y \in X_j^{\mathrm{cor}})] = 1 - O(\rho)$.*

Proof: The key observation is that for any pair of distinct clusters X_i, X_j, the 2-means separation condition holds for $X_i \cup X_j$, that is, $\Delta_2^2(X_i \cup X_j) =$

$\Delta_1^2(X_i) + \Delta_1^2(X_j) \leq \epsilon^2\Delta_1^2(X_i \cup X_j)$. So using Lemma 3.2 we obtain that $\sum_{x\in X_i^{\rm cor}, y\in X_j^{\rm cor}} \|x-y\|^2 = \big(1-O(\rho)\big)\sum_{\{x,y\}\subseteq X_i\cup X_j}\|x-y\|^2$. Summing over all pairs i,j yields the lemma. ■

Now inductively suppose that the first i centers picked $\hat{c}_1,\ldots,\hat{c}_i$ lie in the cores of clusters $X_{j_1},\ldots,X_{j_i}$. We show that conditioned on this event, $\hat{c}_{i+1} \in \bigcup_{\ell\notin\{j_1,\ldots,j_i\}} X_\ell^{\rm cor}$ with probability $1-O(\rho)$. Given a set S of points, we use $d(x,S)$ to denote $\min_{y\in S}\|x-y\|$.

Lemma 4.2 $\Pr\big[\hat{c}_{i+1}$ *lies in* $\bigcup_{\ell\notin\{j_1,\ldots,j_i\}} X_\ell^{\rm cor} \mid \hat{c}_1,\ldots,\hat{c}_i$ *lie in the cores of* $X_{j_1},\ldots,X_{j_i}\big] = 1-O(\rho)$.

Proof: For convenience, we re-index the clusters so that $\{j_1,\ldots,j_i\} = \{1,\ldots,m\}$. Let $\hat{C} = \{\hat{c}_1,\ldots,\hat{c}_i\}$. For any cluster X_j, let $p_j \in \{1,\ldots,i\}$ be such that $d(c_j,\hat{C}) = \|c_j - \hat{c}_{p_j}\|$. Let $A = \sum_{j=m+1}^k \sum_{x\in X_j^{\rm cor}} d(x,\hat{C})^2$, 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 $\max_{x\in X_j^{\rm cor}} \|x-c_j\|/d(c_j,\hat{C})$. For any point $x\in X_j^{\rm cor}$, $j\geq m+1$, we have $d(x,\hat{C}) \geq (1-\alpha)d(c_j,\hat{C})$. By Lemma 3.1, $\alpha \leq \frac{\epsilon/\sqrt{\rho(1-\epsilon^2)}}{1-\epsilon/\sqrt{\rho(1-\epsilon^2)}} \leq \frac{2\epsilon}{\sqrt{\rho(1-\epsilon^2)}} < 1$ for a small enough ρ. Therefore, $A = \sum_{j=m+1}^k \sum_{x\in X_j^{\rm cor}} d(x,\hat{C})^2 \geq \sum_{j=m+1}^k (1-\rho)(1-\alpha)^2 n_j d(c_j,\hat{C})^2$.

On the other hand, for any point $x\in X_j$, we have $d(x,\hat{C}) \leq \|x-\hat{c}_{p_j}\|$. For $j=1,\ldots,m$, $\hat{c}_{p_j}$ lies in $X_j^{\rm cor}$, so $\|c_j-\hat{c}_{p_j}\| \leq \frac{r_j}{\sqrt{\rho}}$. Therefore, $B \leq \sum_{j=1,x\in X_j}^k \|x-\hat{c}_{p_j}\|^2 \leq \sum_{j=1}^k \big(\Delta_1^2(X_j) + n_j\|c_j-\hat{c}_{p_j}\|^2\big) \leq \big(1+\frac{1}{\rho}\big)\Delta_k^2(X) + \sum_{j=m+1}^k n_j d(c_j,\hat{C})^2$. 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 $\big(\frac{1}{\epsilon^2}-1\big)\Delta_k^2(X) \leq n_j d(c_j,\hat{C})^2$ 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 $A = \big(1-O(\rho+\frac{\epsilon}{\sqrt{\rho}})\big)B$. If we set $\rho = \Omega(\epsilon^{2/3})$, we obtain $A/B = 1-O(\rho)$. ■

Next, we analyze the case when more than k points are sampled. Let $\rho_1 = \rho^3$. Define the *outer core* of X_i to be $X_i^{\rm out} = \{x\in X_i : \|x-c_i\|^2 \leq \frac{r_i^2}{\rho_1}\}$. Note that $X_i^{\rm cor} \subseteq X_i^{\rm out}$. 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. By mimicking the proof of Lemma 4.2, one can show (Lemma 4.3) 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. Observe that while Lemma 4.2 only shows that the "good" event happens *conditioned* on the fact that previous samples were also "good", Lemma 4.3 gives an *unconditional* bound. This crucial difference allows us to argue, via a straightforward martingale analysis, that if we sample N points from X, then with some constant probability, each outer core $X_i^{\rm out}$ will contain a sampled point. Corollary 4.4 summarizes the results.

Lemma 4.3 *Suppose that we have sampled i points $\{\hat{x}_1,\ldots,\hat{x}_i\}$ from X. Let $X_1,\ldots,X_m$ be all the clusters whose outer cores contain some sampled point $\hat{x}_j$. Then $\Pr[\hat{x}_{i+1}\in\bigcup_{j=m+1}^k X_j^{\rm cor}] \geq 1-5\rho$.*

Corollary 4.4 *(i) If we sample k points $\hat{c}_1,\ldots,\hat{c}_k$, then with probability $\big(1-O(\rho)\big)^k$, $\rho=\Omega(\epsilon^{2/3})$, each point $\hat{c}_i$ lies in a distinct core $X_i^{\rm cor}$, so $\|\hat{c}_i-c_i\| \leq r_i/\sqrt{\rho}$.*
(ii) If we sample $N = \frac{2k}{1-5\rho} + \frac{2\ln(2/\rho)}{(1-5\rho)^2}$ points $\hat{c}_1,\ldots,\hat{c}_N$, $\rho=\sqrt{\epsilon}$, then with probability $1-O(\rho)$, each outer core $X_i^{\rm out}$ contains some $\hat{c}_i$, so $\|\hat{c}_i-c_i\| \leq r_i/\sqrt{\rho^3}$.

4.1.2 Greedy deletion procedure.

We maintain a set of centers $\hat{C}$ that are used to cluster X. We initialize $\hat{C} \leftarrow X$ and delete centers till k centers remain. 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 call $R(x)$ the Voronoi set of x. Repeat the following steps until $|\hat{C}| = k$.

B1. Compute $T = \sum_{x\in\hat{C}}\sum_{y\in R(x)}\|y-x\|^2$, the cost of clustering X around $\hat{C}$. For every $x\in\hat{C}$, compute $T_x = \sum_{z\in\hat{C}\setminus\{x\}}\sum_{y\in R_{-x}(z)}\|y-z\|^2$, where $R_{-x}(z)$ is the Voronoi set of z given the center set $\hat{C}\setminus\{x\}$.

B2. Pick the center $y\in\hat{C}$ for which $T_x - T$ is minimum and set $\hat{C}\leftarrow\hat{C}\setminus\{y\}$.

B3. 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 $\hat{C}\leftarrow\hat{C}\setminus\{x\}\cup\{{\rm ctr}(R(x))\}$.

Running time There are $n-k$ iterations of the B1-B3 loop. Each iteration takes $O(n^2d)$ time: computing T and the sets $R(x)$ for each x takes $O(n^2d)$ 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 $O(n^3d)$.

Analysis Let ρ be a parameter such that $\rho \leq \frac{1}{10}$, $\epsilon/\sqrt{\rho(1-\epsilon^2)} \leq \frac{1}{14}$. Recall that $D_i = \min_{j\neq i}\|c_j - c_i\|$. Define $d_i^2 = \Delta_k^2(X)/n_i$. We use a different notion of a cluster-core here, but one that still captures the fact that

the core consists of points that are quite close to the cluster-center compared to the inter-cluster distance, and contains most of the mass of the cluster. Let $\mathcal{B}(x,r) = \{y \in \mathbb{R}^d : \|x-y\| \leq r\}$. Define $Z_i = \mathcal{B}(c_i, d_i/\sqrt{\rho})$ and the core of X_i as $X_i^{\text{cor}} = X_i \cap Z_i$. Observe that $r_i \leq d_i$, so by Markov's inequality $|X_i^{\text{cor}}| \geq (1-\rho)n_i$. 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, $X_i^{\text{cor}} = X \cap Z_i$. 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. Call this invariant (*). Clearly (*) holds at the beginning, since $\hat{C} = X$ and $X_i^{\text{cor}} \neq \emptyset$ for every cluster X_i. First, we argue (Lemma 4.5) 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. The intuition is that there must be some other center c_j whose Voronoi region (wrt. optimal center-set) contains at least two centers from $\hat{C}$, and deleting one of these centers, the one further away from c_j, should be less expensive. Lemma 4.6 then shows that even after a center y is deleted, if the new Voronoi region $R_{-y}(x)$ of a center $x \in \hat{C}$ captures points from some X_j^{cor}, then $R_{-y}(x)$ cannot "extend" too far into some other cluster X_ℓ, that is, for any $x' \in R_{-y}(x) \cap X_\ell$ where $\ell \neq j$, $\|x'-c_j\|$ is not much larger than $\|x'-c_\ell\|$. This is due to the fact that for both clusters X_j and X_ℓ, there are undeleted centers (due to the invariant and Lemma 4.5) $z_j \in Z_j$ and $z_\ell \in Z_\ell$. Thus, since $R_{-y}(x) \cap X_j^{\text{cor}} \neq \emptyset$, it must be that x is relatively close to c_j, Also, since x' is captured by x and not z_ℓ, one can lower bound $\|x'-z_\ell\|$, and hence $\|x'-c_\ell\|$, in terms of $\|x'-x\|$ (and $\frac{d_\ell}{\sqrt{\rho}}$), which in turn can be lower bounded in terms of $\|x'-c_j\|$ (and $\frac{d_j}{\sqrt{\rho}}$). Lemma 4.7 puts everything together to show that invariant (*) is maintained.

Lemma 4.5 *Suppose (*) holds at the start of an iteration, and $x \in \hat{C}$ is the only center in $\mathcal{B}(c_i, 4d_i/\sqrt{\rho})$ for some cluster X_i, then $x \in \hat{C}$ after step B2.*

Lemma 4.6 *Suppose center $y \in \hat{C}$ is deleted in step B2. 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 $\|x'-c_j\| \leq \|x'-c_\ell\| + \frac{\max(d_\ell+6d_j, 4d_\ell+3d_j)}{\sqrt{\rho}}$.*

Lemma 4.7 *Suppose that property (*) holds at the beginning of some iteration in the deletion phase. Then (*) also holds at the end of the iteration, i.e., after step B3.*

Proof: Suppose that we delete center $y \in \hat{C}$ that lies in the Voronoi region of center c_i (wrt. optimal centers) in step B2. 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 4.6 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 $\|y-c_j\| \leq \beta\|y-c_\ell\|$ 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 $A = \sum_{x \in S} \sum_{x' \in R'(x)} \|y - c_j\|^2 = \sum_{x \in S} \big(\Delta_1^2(R'(x)) + |R'(x)| \|\text{ctr}(R'(x)) - c_j\|^2\big) \geq |Y| \min_{x \in S} \|\text{ctr}(R'(x)) - c_j\|^2$. Since $X_j^{\text{cor}} \subseteq Y$ we have $|Y| \geq (1-\rho)n_j$, so $\min_{x \in S} \|\text{ctr}(R'(x)) - c_j\| \leq \frac{\beta}{\sqrt{1-\rho}} \cdot d_i$. The bounds on ρ ensure that $\frac{\rho\beta^2}{1-\rho} \leq 1$, so that $\min_{x \in S} \|\text{ctr}(R'(x)) - c_j\| \leq \frac{d_j}{\sqrt{\rho}}$. ■

Corollary 4.8 *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 a seeding procedure that runs in time $O(nkd + k^3 d)$ 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 ensures 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. By the analysis of the previous section, one can then show that after the deletion procedure the k centers are still close to the optimal centers for the sampled instance, and hence also close to the optimal centers for X. Fix $\rho_1 = \sqrt{\epsilon}$.

C1. **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.

C2. **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 (wrt. the set S). We now ignore X, and consider a weighted instance with point-set $\hat{S} = \{\hat{x} = \text{ctr}(R(x)) : x \in S\}$, where each $\hat{x}$ has a *weight* $w(\hat{x}) = |R(x)|$. Run the deletion procedure of Section 4.1.2, on $\hat{S}$ to obtain k centers $\hat{c}_1, \ldots, \hat{c}_k$.

Running time Step C1 takes $O(nNd) = O(nkd)$ time. The run-time analysis of the deletion phase in Section 4.1.2,

shows that step C2 takes $O(N^3 d) = O(k^3 d)$ time. So the overall running time is $O(nkd + k^3 d)$.

Analysis Recall that $\rho_1 = \sqrt{\epsilon}$. Let $\rho_2 = \rho_1^3$. Let $X_i^{\text{cor}} = \{x \in X_i : \|x - c_i\|^2 \leq \frac{r_i^2}{\rho_1}\}$. Let $X_i^{\text{out}} = \{x \in X_i : \|x - c_i\|^2 \leq \frac{r_i^2}{\rho_2}\}$ denote the outer core of cluster X_i. By part (ii) of Corollary 4.4 we know that with probability $1 - O(\rho_1)$, every cluster X_i contains a sampled point in its outer core after step C1. 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 4.9 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 implies that the centers returned by the deletion phase are close to the optimal centers for X.

Lemma 4.9 $\Delta_k^2(\hat{S}) = O(\epsilon^2)\Delta_{k-1}^2(\hat{S})$. *For each optimal center c_i, there is some $\hat{s}_i$ such that $\|\hat{s}_i - c_i\| \leq \frac{D_i}{25} + \frac{r_i}{\sqrt{\rho_1}}$.*

Lemma 4.10 *For each center c_i, there is a center $\hat{c}_i$ such that $\|\hat{c}_i - c_i\| \leq \frac{D_i}{10}$.*

Proof: Let $\hat{D}_i = \min_{j \neq i} \|\hat{s}_j - \hat{s}_i\|$. Then $(1-2\theta) \leq \frac{\hat{D}_i}{D_i} \leq (1+2\theta)$ where $\theta \leq \left(\frac{1}{25} + \frac{\epsilon}{\sqrt{\rho_1(1-\epsilon^2)}}\right)$. Since $\rho_1 = \sqrt{\epsilon}$, we can ensure that $\theta < \frac{1}{22}$. Choosing ρ for the deletion phase suitably, by Corollary 4.8, we can ensure that we obtain $\hat{c}_i$ such that $\|\hat{c}_i - \hat{s}_i\| \leq \frac{\hat{D}_i}{20}$. Thus, $\|\hat{c}_i - c_i\| \leq \frac{D_i}{10}$. ■

4.2. Procedures used in stage II

Given k seed centers $\hat{c}_1, \ldots, \hat{c}_k$ 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 $\big(1 + f(\epsilon)\big)$-approximation algorithm, or the *centroid estimation* step, based on a sampling idea of Kumar et al. [**?**], which yields a PTAS with running time exponential in k. Define $\hat{d}_i = \min_{j \neq i} \|\hat{c}_j - \hat{c}_i\|$.

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 $\bar{c}_1, \ldots, \bar{c}_k$ as the final centers.

Centroid estimation. For each i, we will obtain a set of candidate centers for cluster X_i. Fix $\beta = \frac{25}{25+256\epsilon^2}$. Let $R_i' \subseteq X = \{x \in X : \|x - \hat{c}_i\| \leq \min_j \|x - \hat{c}_j\| + \hat{d}_i/4\}$ be the expanded Voronoi region of $\hat{c}_i$. 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 $\bar{c}_1 \in T_1, \ldots, \bar{c}_k \in T_k$ that yield the least-cost solution, and return these as the final centers.

Analysis Recall that $D_i = \min_{j \neq i} \|c_j - c_i\|$. Let $\rho = \frac{36\epsilon^2}{1-\epsilon^2}$. The proof of Lemma 4.11, which analyzes the ball-k-means step, is essentially identical to that of Lemma 3.3.

Lemma 4.11 *Let $Y_i = \big\{x \in X_i : \|x - c_i\|^2 \leq \frac{r_i^2}{\rho}\big\}$. If $\|\hat{c}_i - c_i\| \leq D_i/10$ for each i, then $Y_i \subseteq B_i \subseteq X_i$, and $\|\bar{c}_i - c_i\| \leq \frac{\rho}{1-\rho} \cdot r_i^2$.*

Lemma 4.12 *Suppose $\|\hat{c}_i - c_i\| \leq D_i/10$ for each i. Then $X_i \subseteq R_i'$, where R_i' is as defined above, and $|X_i| \geq \beta |R_i'|$.*

Proof: We have $\frac{4D_i}{5} \leq \hat{d}_i \leq \frac{6D_i}{5}$. Consider any $x \in X_i$ that lies in the Voronoi region of $\hat{c}_j$. We have $\|x - c_i\| \leq \|x - c_j\|$, therefore $\|x - \hat{c}_i\| \leq \|x - \hat{c}_j\| + D_i/5 \leq \|x - \hat{c}_j\| + \hat{d}_i/4$; so $x \in R_i'$. Suppose $|X_i| \leq \beta |R_i'|$. Let $a_j = \frac{|R_i' \cap X_j|}{|R_i'|}$. So $\frac{a_i}{1-a_i} \leq \frac{\beta}{1-\beta}$. Consider the clustering where we arbitrarily assign some $\frac{a_j}{1-a_i}$ points of X_i to center c_j for each $j \neq i$. For any $x \in X_i$ and $j \neq i$, we have $\|x - c_j\|^2 \leq 2(\|x - c_i\|^2 + \|c_i - c_j\|^2)$. So the cost of reassigning points in X_i is at most $2\Delta_1^2(X_i) + \frac{2n_i}{1-a_i} \cdot \sum_{j \neq i} a_j \|c_i - c_j\|^2 \leq 2\Delta_1^2(X_i) + \frac{2\beta}{1-\beta} \cdot \sum_{j \neq i} a_j |R_i'| \|c_i - c_j\|^2$. We also know that for any $y \in R_i' \cap X_j$, $\|y - c_i\| \leq \|y - c_j\| + \frac{D_i + D_j}{10}$, which implies that $\|c_i - c_j\| \leq \frac{8}{5} \cdot \|y - c_j\|$. Therefore, we can bound $a_j |R_i'| \|c_i - c_j\|^2$ by $\frac{64}{25} \sum_{y \in R_i' \cap X_j} \|y - c_j\|^2$. Hence, the cost of this clustering is at most $\max\big(2, 1 + \frac{128\beta}{25(1-\beta)}\big)\Delta_k^2(X) \leq \big(1 + \frac{1}{2\epsilon^2}\big)\Delta_k^2(X)$. The cost of this clustering is also at least $\Delta_{k-1}^2(X)$. This is a contradiction to the assumption that $\Delta_k^2(X) \leq \epsilon^2 \Delta_{k-1}^2(X)$. ■

4.3. A linear time constant-factor approximation algorithm

D1. Execute the seeding procedure of Section 4.1.3 to obtain k initial centers $\hat{c}_1, \ldots, \hat{c}_k$.

D2. Run the **ball-k-means step** of Section 4.2 to obtain the final centers.

The running time is $O(nkd + k^3 d)$. By Lemma 4.10, we know that with probability $1 - O(\sqrt{\epsilon})$, for each c_i, there is a distinct center $\hat{c}_i$ such that $\|\hat{c}_i - c_i\| \leq D_i/10$. Combined with Lemma 4.11, this yields the following theorem.

Theorem 4.13 *If $\frac{\Delta_k^2(X)}{\Delta_{k-1}^2(X)} \leq \epsilon^2$ for a small enough ϵ, the above 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 with the centroid estimation step.

E1. Use the procedure in Section 4.1.1 to pick k initial centers $\hat{c}_1, \ldots, \hat{c}_k$.

E2. Run the **centroid estimation** procedure of Section 4.2 to obtain the final centers.

The running time is dominated by the exhaustive search in step E2 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 4.14 *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 appropriately setting ρ in the sampling procedure, we can ensure that with probability $\Theta(1)^k$, it returns centers $\hat{c}_1, \ldots, \hat{c}_k$ such that for each i, $\|\hat{c}_i - c_i\| \le D_i/10$ (part (i) of Corollary 4.4). So by Lemma 4.12 we know that $|X_i| \ge \beta|R_i'|$ for every i. Now Lemma 2.3 in [**?**] 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 $c_1' \in T_1, \ldots, c_k' \in T_k$. 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 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 region of point x (the centers 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 *Let $\alpha^2 = \frac{1-401\epsilon^2}{400}$. 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 $cost(\hat{c}_1, \ldots, \hat{c}_k) \le \alpha^2\Delta_{k-1}^2(X)$, then for each $\hat{c}_i$ there is a distinct optimal center $c_{\sigma(i)}$ such that $|R(\hat{c}_i) \ominus X_{\sigma(i)}| \le 28\epsilon^2|X_{\sigma(i)}|$;

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

Notice that part (i) also yields that if X is ϵ-separated for k-means, then any k-clustering of cost at most $\frac{\alpha^2}{\epsilon^2}\Delta_k^2(X)$ has small Hamming distance to the optimal k-clustering (more strongly, each cluster has small Hamming distance to a distinct optimal cluster). We now show the converse.

Theorem 5.2 *Let $\epsilon \le \frac{1}{3}$. 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 $|\hat{X}_i \ominus X_{\sigma(i)}| \le \epsilon|X_{\sigma(i)}|$. Then, X is α-separated for k-means;

(ii) there is a bijection σ such that $\sum_{i=1}^k |\hat{X}_i \ominus X_{\sigma(i)}| \le \frac{2\epsilon}{k-1}|X|$. 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 high cost, implying that $\Delta_{k-1}^2(X)/\Delta_k^2(X)$ is large. 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 $\hat{X}_{k-1} \cup \hat{X}_k = R_{k-1},\ \hat{X}_{k-1}, \hat{X}_k \ne \emptyset$. If this has high cost, then we are done, otherwise 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 $\sigma'(i) \ne \sigma(i)$ for $i \le k-2$, then $|R_i \ominus X_{\sigma'(i)}| \ge |R_i \cap X_{\sigma(i)}| \ge (1-\epsilon)|R_i|$; otherwise, $\sigma'(k) \in \{\sigma(k-1), \sigma(k)\}$, so $|X_k' \ominus X_{\sigma'(k)}| \ge \frac{1-\epsilon}{2}|X_k'|$ 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 $|R_{k-1}| \ge \frac{|X|}{k-1}$, we have $|\hat{X}_{k-1} \cap X_{\sigma(k-1)}| + |\hat{X}_k \cap X_{\sigma(k)}| \ge \frac{1-\epsilon}{k-1}|X|$. After the above mismatch operation, for any bijection σ', the total disagreement is at least $|X_{k-1}' \ominus X_{\sigma'(k-1)}| + |X_k' \ominus X_{\sigma'(k)}| \ge \frac{1}{2}\big(|\hat{X}_{k-1} \cap X_{\sigma(k-1)}| + |\hat{X}_k \cap X_{\sigma(k)}|\big) \ge \frac{1-\epsilon}{2(k-1)}|X|$. ∎