Faster Algorithms for the Constrained k-Means Problem

Anup Bhattacharya[1], Ragesh Jaiswal[2], and Amit Kumar[3]

1 **Department of Computer Science and Engineering, Indian Institute of Technology Delhi, New Delhi, India**
anupb@cse.iitd.ernet.in
2 **Department of Computer Science and Engineering, Indian Institute of Technology Delhi, New Delhi, India**
rjaiswal@cse.iitd.ernet.in
3 **Department of Computer Science and Engineering, Indian Institute of Technology Delhi, New Delhi, India**
amitk@cse.iitd.ernet.in

Abstract

The classical center based clustering problems such as k-means/median/center assume that the optimal clusters satisfy the locality property that the points in the same cluster are close to each other. A number of clustering problems arise in machine learning where the optimal clusters do not follow such a locality property. For instance, consider the *r-gather clustering problem* where there is an additional constraint that each of the clusters should have at least r points or the *capacitated clustering problem* where there is an upper bound on the cluster sizes. Consider a variant of the k-means problem that may be regarded as a general version of such problems. Here, the optimal clusters $O_1, ..., O_k$ are an *arbitrary* partition of the dataset and the goal is to output k-centers $c_1, ..., c_k$ such that the objective function $\sum_{i=1}^{k} \sum_{x \in O_i} ||x - c_i||^2$ is minimized. It is not difficult to argue that any algorithm (without knowing the optimal clusters) that outputs a single set of k centers, will not behave well as far as optimizing the above objective function is concerned. However, this does not rule out the existence of algorithms that output a *list* of such k centers such that at least one of these k centers behaves well. Given an error parameter $\varepsilon > 0$, let ℓ denote the size of the smallest list of k-centers such that at least one of the k-centers gives a $(1 + \varepsilon)$ approximation w.r.t. the objective function above. In this paper, we show an upper bound on ℓ by giving a randomized algorithm that outputs a list of $2^{\tilde{O}(k/\varepsilon)}$ k-centers. We also give a closely matching lower bound of $2^{\tilde{\Omega}(k/\sqrt{\varepsilon})}$. Moreover, our algorithm runs in time $O\left(nd \cdot 2^{\tilde{O}(k/\varepsilon)}\right)$. This is a significant improvement over the previous result of Ding and Xu who gave an algorithm with running time $O\left(nd \cdot (\log n)^k \cdot 2^{poly(k/\varepsilon)}\right)$ and output a list of size $O\left((\log n)^k \cdot 2^{poly(k/\varepsilon)}\right)$. Our techniques generalize for the k-median problem and for many other settings where non-Euclidean distance measures are involved.

1998 ACM Subject Classification I.5.3 Clustering

Keywords and phrases k-means, k-median, approximation algorithm, sampling

Digital Object Identifier 10.4230/LIPIcs.STACS.2016.16

1 Introduction

Clustering problems intend to classify high dimensional data based on the proximity of points to each other. There is an inherent assumption that the clusters satisfy *locality* property – points close to each other (in a geometric sense) should belong to the same category. Often,

33rd Symposium on Theoretical Aspects of Computer Science (STACS 2016).
Editors: Nicolas Ollinger and Heribert Vollmer; Article No. 16; pp. 16:1–16:13
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany
SYMPOSIUM ON THEORETICAL ASPECTS OF COMPUTER SCIENCE

we model such problems by the notion of a center based clustering problem. We would like to identify a set of centers, one for each cluster, and then the clustering is obtained by assigning each point to the nearest center. For example, the k-means problem is defined in the following manner: given a dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and an integer k, output a set of k centers $\{c_1, \ldots, c_k\} \subset \mathbb{R}^d$ such that the objective function $\sum_{x \in X} \min_{c \in \{c_1, \ldots, c_k\}} ||x - c||^2$ is minimized. The k-median and the k-center problems are defined in a similar manner by defining a suitable objective function.

However, often such clustering problems entail several *side constraints*. Such constraints limit the set of feasible clusterings. For example, the r-gather k-means clustering problem is defined in the same manner as the k-means problem, but has the additional constraint that each cluster must have at least r points in it. In such settings, it is no longer true that the clustering is obtained from the set of centers by the Voronoi partition. Ding and Xu [5] began a systematic study of such problems, and this is the starting point of our work as well. They defined the so-called *constrained k-means* problem. An instance of such a problem is specified by a set of points X, a parameter k, and a set $\mathbb{C}$, where each element of $\mathbb{C}$ is a partitioning of X into k disjoint subsets (or clusters). Since the set $\mathbb{C}$ may be exponentially large, we will assume that it is specified in a succinct manner by an efficient algorithm which decides membership in this set. A solution needs to output an element $\mathbb{O} = \{O_1, \ldots, O_k\}$ of $\mathbb{C}$, and a set of k centers, $c_1, \ldots, c_k$, one for each cluster in $\mathbb{O}$. The goal is to minimize $\sum_{i=1}^{k} \sum_{x \in O_i} ||x - c_i||^2$. It is easy to check that the center c_i must be the mean of the corresponding cluster O_i. Note that the k-means problem is a special case of this problem where the set $\mathbb{C}$ contains all possible ways of partitioning X into k subsets. The constrained k-median problem can be defined similarly. We will make the natural assumption (which is made by Ding and Xu as well) that it suffices to find a set of k centers. In other words, there is an (efficient) algorithm $A^{\mathbb{C}}$, which given a set of k centers $c_1, \ldots, c_k$, outputs the clustering $\{O_1, \ldots, O_k\} \in \mathbb{C}$ such that $\sum_{i=1}^{k} \sum_{x \in O_i} ||c_i - x||^2$ is minimized. Such an algorithm is called a *partition algorithm* by Ding and Xu [5] [1]. For the case of the k-means problem, this algorithm will just give the Voronoi partition with respect to $c_1, \ldots, c_k$, whereas in the case of the r-gather k-means clustering problem, the algorithm $A^{\mathbb{C}}$ will be given by a suitable min-cost flow computation (see section 4.1 in [5]).

Ding and Xu [5] considered several natural problems arising in diverse areas, e.g. machine learning, which can be stated in this framework. These included the so-called r-gather k-means, r-capacity k-means and l-diversity k-means problems. Their approach for solving such problems was to output a list of candidate sets of centers (of size k) such that at least one of these were close to the optimal centers. We formalize this approach and show that if k is small, then one can obtain a PTAS for the constrained k-means (and the constrained k-median) problems whose running time is linear plus a constant number of calls to $A^{\mathbb{C}}$.

We define the *list k-means* problem. Given a set of points X and parameters k and ε, we want to output a list $\mathcal{L}$ of sets of k points (or centers). The list $\mathcal{L}$ should have the following property: for *any* partitioning $\mathbb{O} = \{O_1, \ldots, O_k\}$ of X into k clusters, there exists a set $c_1, \ldots, c_k$ in the list $\mathcal{L}$ such that (up-to reordering of these centers)

$$\sum_{i=1}^{k} \sum_{x \in O_i} ||c_i - x||^2 \leq (1 + \varepsilon) \sum_{i=1}^{k} \sum_{x \in O_i} ||x - m_i||^2, \tag{1}$$

[1] [5] also gave a discussion on such partition algorithms for a number of clustering problems with side constraints.

where $m_i = \frac{\sum_{x \in O_i} x}{|O_i|}$ denotes the mean of O_i. Note that the latter quantity is the k-means cost of the clustering $\mathbb{O}$, and so we require $c_1, \ldots, c_k$ to be such that the cost of assigning to these centers is close to the optimal k-means cost of this clustering. We shall use $\mathtt{opt}_k(\mathbb{O})$ to denote the optimal k-means cost of $\mathbb{O}$.

Although such an oblivious approach to clustering may appear too optimistic, we show that it is possible to obtain such a list $\mathcal{L}$ of size $2^{\tilde{O}(k/\varepsilon)}$ in $O\left(nd \cdot 2^{\tilde{O}(k/\varepsilon)}\right)$ time [2]. This improves the result of Ding and Xu [5], where they gave an algorithm which outputs a list of size $O\left((\log n)^k \cdot 2^{\text{poly}(k/\varepsilon)}\right)$. Observe that we address a question which is both algorithmic and existential : how small can the size of $\mathcal{L}$ be, and how efficiently can we find it ? We also give almost matching lower bounds on the size of such a list $\mathcal{L}$. Our algorithm for finding $\mathcal{L}$ relies on the D^2-sampling idea – iteratively find the centers by picking the next one to be *far* from the current set of centers. Although these ideas have been used for the k-means problems (see e.g. [9]), they rely heavily on the fact that given a set of centers, the corresponding clustering is obtained by the corresponding Voronoi partition. Our approach relies in showing that there is a small sized list $\mathcal{L}$ which works well for all possible clusterings.

It is not hard to show that a result for the list k-means problem implies a corresponding result for the constrained k-means problem with the number of calls to $A^{\mathbb{C}}$ being equal to the size of the list $\mathcal{L}$. Therefore, we obtain as corollary of our main result efficient algorithms for the constrained k-means (and the constrained k-median) problems.

1.1 Related Work

The classical k-means problem is one of the most well-studied clustering problems. There is a long sequence of work on obtaining fast PTAS for the k-means and the k-median problems (see e.g., [12, 2, 4, 7, 11, 1, 3, 9, 6] and references therein). Some of these works implicitly maintain a list of centers of size k such that the condition (1) is satisfied for all clusterings $\mathbb{O}$ which correspond to a Voronoi partition (with respect to a set of k centers) of the input set of points, and one picks the best possible set of centers from this list (see e.g., [11, 1, 9]). The list has at most $2^{\text{poly}(k/\varepsilon)}$ elements, and from this, one can recover a $(1+\varepsilon)$-approximation algorithm for the k-means problem with running time $O\left(nd \cdot 2^{\text{poly}(k/\varepsilon)}\right)$.

The more general case of the constrained k-means problem was studied by Ding and Xu [5] who also gave an algorithm that outputs a list of size $O\left((\log n)^k \cdot 2^{\text{poly}(k/\varepsilon)}\right)$. Our work improves upon this result. Moreover, we consider the formulation of the list k-means problem as an important contribution, and feel that similar formulations in other classification settings would be useful.

1.2 Preliminaries

We formally define the problems considered in this paper. The centroid or mean of a finite set of points $X \subset \mathbb{R}^d$ is denoted by $\Gamma(X) = \frac{\sum_{x \in X} x}{|X|}$. Let $\Delta(X)$ denote the 1-means cost of these set of points, i.e., $\sum_{x \in X} ||x - \Gamma(X)||^2$.

An input instance $\mathcal{I}$ for the list k-means (or the list k-median) problem consists of a set of points X, a positive integer k and a positive parameter ε. A partition of X into disjoint subsets $O_1, \ldots, O_k$ will be called a *clustering* of X. Given a clustering $\mathbb{O}^\star = \{O_1^\star, \ldots, O_k^\star\}$

[2] $\tilde{O}$ notation hides a $O(\log \frac{k}{\varepsilon})$ factor.

of X and a set of k centers $C = \{c_1, \ldots, c_k\}$, define $\texttt{cost}_C(\mathbb{O}^\star)$ as the minimum, over all permutations π of C, of $\sum_{i=1}^{k} \sum_{x \in O_i^\star} ||x - c_{\pi(i)}||^2$. Recall that $\texttt{opt}_k(\mathbb{O}^\star)$ denotes the optimal k-means cost of $\mathbb{O}^\star$, i.e., $\sum_{i=1}^{k} \sum_{x \in O_i^\star} ||x - \Gamma(O_i^\star)||^2$.

For a set of points X and a set of points C (of size at most k), define $\Phi_C(X)$ as $\sum_{x \in X} \min_{c \in C} ||x - c||^2$, i.e., we consider the Voronoi partition of X induced by C, and consider the k-means cost of X with respect to this partition. When considering the list k-median problem, we will use the same notation, except that we will consider the Euclidean norm instead of the square of the Euclidean norm. When C is a singleton set $\{c\}$, we shall abuse notation by using $\Phi_c(X)$ instead of $\Phi_{\{c\}}(X)$.

As mentioned in the introduction, the constrained k-means problem is specified by a set of points X, a positive integer k, and a set $\mathbb{C}$ of feasible clusterings of X. Further, we are given an algorithm $A^{\mathbb{C}}$, which given a set of k centers C, outputs the clustering $\mathbb{O}$ in $\mathbb{C}$ which minimizes $\texttt{cost}_C(\mathbb{O})$. The goal is to find a clustering $\mathbb{O} \in \mathbb{C}$ and a set C of size k which minimizes $\texttt{cost}_C(\mathbb{O})$. Note that the centers in C should just be the mean of each cluster in $\mathbb{O}$. On the other hand, if we know C, then we can find the best clustering in $\mathbb{C}$ by calling $A^{\mathbb{C}}$. We use the same notation for the constrained k-median problem.

We now mention a few results which will be used in our analysis. The following fact is well known.

▶ **Fact 1.** *For any $X \subset \mathbb{R}^d$ and $c \in \mathbb{R}^d$ we have $\sum_{x \in X} ||x - c||^2 = \sum_{x \in X} ||x - \Gamma(X)||^2 + |X| \cdot ||c - \Gamma(X)||^2$.*

We next define the notion of D^2-sampling.

▶ **Definition 2** (D^2-sampling)**.** Given a set of points $X \subset \mathbb{R}^d$ and another set of points $C \subset \mathbb{R}^d$, D^2-sampling from X w.r.t. C samples a point $x \in X$ with probability $\frac{\Phi_C(\{x\})}{\Phi_C(X)}$. For the case $C = \emptyset$, D^2-sampling is the same as uniform sampling from X.

The following result of Inaba et al. [8] shows that a constant size random sample is a good enough approximation of a set of points X as far as the 1-means objective is concerned.

▶ **Lemma 3** ([8])**.** *Let S be a set of points obtained by independently sampling M points with replacement uniformly at random from a point set $X \subset \mathbb{R}^d$. Then for any $\delta > 0$,*

$$\mathbf{Pr}\left[\Phi_{\Gamma(S)}(X) \leq \left(1 + \frac{1}{\delta M}\right) \cdot \Delta(X)\right] \geq (1 - \delta).$$

We will also use the following simple fact that may be interpreted as approximate version of the triangle inequality for squared Euclidean distance.

▶ **Fact 4** (Approximate triangle inequality)**.** *For any $x, y, z \in \mathbb{R}^d$, we have $||x - z||^2 \leq 2 \cdot ||x - y||^2 + 2 \cdot ||y - z||^2$.*

1.3 Our Results

We now state our results for the list k-means and the list k-median problems.

▶ **Theorem 5.** *Given a set of n points $X \subset \mathbb{R}^d$, parameters k and ε, there is a randomized algorithm which outputs a list $\mathcal{L}$ of $2^{\tilde{O}(k/\varepsilon)}$ sets of centers of size k such that for any clustering $\mathbb{O}^\star = \{O_1^\star, ..., O_k^\star\}$ of X, the following event happens with probability at least $1/2$: there is a set $C \in \mathcal{L}$ such that*

$$\texttt{cost}_C(\mathbb{O}^\star) \leq (1 + \varepsilon) \cdot \texttt{opt}_k(\mathbb{O}^\star).$$

Moreover, the running time of our algorithm is $O\left(nd \cdot 2^{\tilde{O}(k/\varepsilon)}\right)$. *The same statement holds for the list k-median problem as well, except that the size of the list* $\mathcal{L}$ *becomes* $2^{\tilde{O}(k/\varepsilon^{O(1)})}$ *and the running time of our algorithm becomes* $O\left(nd \cdot 2^{\tilde{O}(k/\varepsilon^{O(1)})}\right)$.

As a corollary of this result we get PTAS for the constrained k-means problem (and similarly for the constrained k-median problem). The proof may be found in the full version of this paper. [3]

▶ **Corollary 6.** *There is a randomized algorithm which given an instance of the constrained k-means problem and parameter* $\varepsilon > 0$, *outputs a solution of cost at most* $(1+\varepsilon)$*-times the optimal cost with probability at least* $1/2$. *Further, the time taken by this algorithm is* $O\left(nd \cdot 2^{\tilde{O}(k/\varepsilon)}\right) + 2^{\tilde{O}(k/\varepsilon)} \cdot T$, *where* T *denotes the time taken by* $A^{\mathbb{C}}$ *on this instance.*

Proof. We use the algorithm in Theorem 5 to get a list $\mathcal{L}$ for this data-set. For each set $C \in \mathcal{L}$, we invoke $A^{\mathbb{C}}$ with C as the set of centers – let $\mathbb{O}(C)$ denote the clustering produced by $A^{\mathbb{C}}$. We output the clustering for which $\texttt{cost}_C(\mathbb{O}(C))$ is minimum. Let $\mathbb{O}^\star$ be the optimal clustering, i.e., the clustering in $\mathbb{C}$ for which $\texttt{opt}_k(\mathbb{O}^\star)$ is minimum. We know that with probability at least $1/2$, there is set $C \in \mathcal{L}$ for which $\texttt{cost}_C(\mathbb{O}^\star) \leq (1+\varepsilon)\texttt{opt}_k(\mathbb{O}^\star)$. Now, the solution produced by our algorithm has cost at most $\texttt{cost}_C(\mathbb{O}(C))$, which by definition of $A^{\mathbb{C}}$, is at most $\texttt{cost}_C(\mathbb{O}^\star)$. ◀

We also give a nearly matching lower bound on the size of $\mathcal{L}$. The following result along with Yao's Lemma shows that one cannot reduce the size of $\mathcal{L}$ to less than $2^{\tilde{\Omega}\left(\frac{k}{\sqrt{\varepsilon}}\right)}$.

▶ **Theorem 7.** *Given a parameter* k *and a small enough positive constant* ε, *there exists a set* X *of points in* $\mathbb{R}^d$ *and a set* $\mathbb{C}$ *of clusterings of* X *such that any list* $\mathcal{L}$ *of* k*-centers with the following property must have size at least* $2^{\tilde{\Omega}\left(\frac{k}{\sqrt{\varepsilon}}\right)}$ *: for at least half of the clusterings* $\mathbb{O} \in \mathbb{C}$, *there exists a set* C *in* $\mathcal{L}$ *such that* $\texttt{cost}_C(\mathbb{O}) \leq (1+\varepsilon)\texttt{opt}_k(\mathbb{O})$.

Our techniques also extend to settings involving many other "approximate" metric spaces (see the discussion in the full version of this paper). Another important observation is that in the lower bound result above, the clusterings in $\mathbb{C}$ correspond to Voronoi partitions of X. This throws light on the previous works [11, 1, 6, 9, 10] as to why the running time of all the algorithms was proportional to $2^{\text{poly}(k/\varepsilon)}$: they were implicitly maintaining a list which satisfied (1) for all Voronoi partitions of X, and therefore, our lower bound result applies to their algorithms as well.

1.4 Our Techniques

Our techniques are based on the idea of D^2-sampling that was used by Jaiswal et al. [9] to give a $(1+\varepsilon)$-approximation algorithm for the k-means problem. Our ideas also have similarities to the ideas of Ding and Xu [5]. We discuss these similarities towards the end of this subsection.

One of the crucial ingredients that is used in most of the $(1+\varepsilon)$-approximation algorithms for k-means is Lemma 3. This result essentially states that given a set of points P, if we are able to uniformly sample $O(1/\varepsilon)$ points from it, then the mean of these sampled points

[3] The full version of this paper may be found on Arxiv. Here is the link: `http://arxiv.org/abs/1504.02564`.

will be a good substitute for the mean of P. Consider an optimal clustering $O_1^\star, \ldots, O_k^\star$ for a set of points X. If we could uniformly sample from each of the clusters $O_i^\star$, then by the argument above, we will be done. The first problem one encounters is that one can only sample from the input set of points, and so, if we sample sufficiently many points from X, we need to somehow distinguish the points which belong to $O_i^\star$ in this sample. This can be dealt with using the following argument: suppose we manage to get a small sample S of points (say of size $O(\text{poly}(k/\varepsilon))$) that contain at least $\Omega(1/\varepsilon)$ points uniformly distributed in $O_i^\star$, then we can try all possible subsets of S of size $O(1/\varepsilon)$ and ensure that at least one of the subsets is a uniform sample of appropriate size from $O_i^\star$. Another issue is – how do we ensure that the sample S has sufficient representation from $O_i^\star$? Uniform sampling from the input X will not work since $|O_i^\star|$ might be really small compared to the size of $|X|$. This is where D^2-sampling plays a crucial role and we discuss this next.

Given a set of points $X \subseteq \mathbb{R}^d$ and candidate centers $c_1, ..., c_i \in \mathbb{R}^d$, D^2-sampling with respect to the centers $c_1, ..., c_i$ samples a point $x \in X$ with probability proportional to $\min_{c \in \{c_1, \ldots, c_i\}} ||x - c||^2$. Note that this process "boosts" the probability of a cluster $O_j^\star$ that has many points far from the set $\{c_1, \ldots, c_i\}$. Therefore, even if a cluster $O_j^\star$ has a small size, we will have a good chance of sampling points from it (if it is far from the current set of centers). However, this nonuniform sampling technique gives rise to another issue. The points being sampled are no longer uniform samples from the optimal clusters. Depending on the current set of centers, different points in a cluster $O_j^\star$ have different probability of getting sampled. This issue is not that grave for the k-means problem where the optimal clusters are Voronoi regions since we can argue that the probabilities are not very different. However, for the constrained k-means problem where the optimal clusters are allowed to be arbitrary partition of the input points, this problem becomes more serious. This can be illustrated using the following example. Suppose we have managed to pick centers $c_1, \ldots, c_i$ that are good (in terms of cluster cost) for the optimal clusters $O_1^\star, \ldots, O_i^\star$. At this point let $O_j^\star$ denote the cluster other than $O_1^\star, \ldots, O_i^\star$, such that a point sampled using D^2 sampling w.r.t. $c_1, \ldots, c_i$ is most likely to be from $O_j^\star$. Suppose we sample a set S of $O(k/\varepsilon)$ points using D^2-sampling. Are we guaranteed (w.h.p.) to have a subset in S that is a uniform sample from $O_j^\star$? The answer is no (actually quite far from it). This is because the optimal clusters may form an arbitrary partition of the data-set and it is possible that most of the points in $O_j^\star$ might be very close to the centers $c_1, \ldots, c_i$. In this case the probability of sampling such points will be close to 0. The way we deal with this scenario is that we consider a multi-set S' that is the union of the set of samples S and $O(1/\varepsilon)$ copies of each of $c_1, \ldots, c_i$. We then argue that all the points in $O_j^\star$ that are far from $c_1, \ldots, c_i$ will have a good chance of being represented in S (and hence in S'). On the other hand, even though the points that are close to one of $c_1, \ldots, c_i$ will not be represented in S (and hence S'), the center (among $c_1, \ldots, c_i$) that is close to these points have good representation in S' and these centers may be regarded as "proxy" for the points in $O_j^\star$.

Ding and Xu [5], instead of using the idea of D^2-sampling, rely on the ideas of Kumar et al. [11] which involves uniform sampling of points and then pruning the data-set by removing the points that are close to centers that are currently being considered. In their work, they also encounter the problem that points from some optimal cluster might be close to the current set of good centers (and hence will be removed before uniform sampling). Ding and Xu [5] deal with this issue using what they call a "simplex lemma". Consider the same scenario as in the previous paragraph. At a very high level, they consider grids inside several simplices defined by the current centers $c_1, \ldots, c_i$ and the sampled points. Using the simplex lemma, they argue that one of the points inside these grids will be a good center for the cluster $O_j^\star$.

We now give an overview of the paper. In Section 2, we give the algorithm for generating the list of sets of centers for an instance of the list k-means problem. The algorithm is analyzed in Section 3. Details about the lower bound construction (Theorem 7 and extensions to the k-median problem, and other distance metric settings, are discussed in the full version.)

2 The Algorithm

Consider an instance of the list k-means problem. Let X denote the set of points, and ε be a positive parameter. The algorithm **List-k-means** is described in Algorithm 1. It maintains a set C of centers, which is initially empty. Each recursive call to the function **Sample-centers** increases the size of C by one. In Step 2 of this function, the algorithm tries out various candidates which can be added to C (to increase its size by 1). First, it builds a multi-set S as follows: it independently samples (with replacement) $O(k/\varepsilon^3)$ points using D^2-sampling from X w.r.t. the set C. Further, it adds $O(1/\varepsilon)$ copies of each of the centers in C to the set S. Having constructed S, we consider all subsets of size $O(1/\varepsilon)$ of S – for each such subset we try adding the mean of this set to C. Thus, each invocation of **Sample-centers** makes multiple recursive calls to itself ($\binom{|S|}{M}$ to be precise). It will be useful to think of the execution of this algorithm as a tree $\mathcal{T}$ of depth k. Each node in the tree can be labeled with a set C – it corresponds to the invocation of **Sample-centers** with this set as C (and i being the depth of this node). The children of a node denote the recursive function calls by the corresponding invocation of **Sample-centers**. Finally, the leaves denote the set of candidate centers produced by the algorithm.

List-k-means(X, k, ε)
- Let $N = \frac{136448 \cdot k}{\varepsilon^3}$, $M = \frac{100}{\varepsilon}$
- Initialize $\mathcal{L}$ to $\emptyset$.
- Repeat 2^k times:
  - Make a call to **Sample-centers**$(X, k, \varepsilon, 0, \{\})$.
- Return $\mathcal{L}$.

Sample-centers$(X, k, \varepsilon, i, C)$
(1) If $(i = k)$ then add C to the set $\mathcal{L}$.
(2) else
  (a) Sample a multi-set S of N points with D^2-sampling (w.r.t. centers C)
  (b) $S' \leftarrow S$
  (c) For all $c \in C$: $S' \leftarrow S' \cup \{M \text{ copies of } c\}$
  (d) For all subsets $T \subset S'$ of size M:
    (i) $C \leftarrow C \cup \{\Gamma(T)\}$.
    (ii) **Sample-centers**$(X, k, \varepsilon, i+1, C)$

Algorithm 1 Algorithm for list k-means.

3 Analysis

In this section we prove Theorem 5 for the list k-means problem. Let $\mathcal{L}$ denote the set of candidate solutions produced by **List-k-means**, where a solution corresponds to a set of centers C of size k. These solutions are output at the leaves of the execution tree $\mathcal{T}$. Fix a

clustering $\mathbb{O}^\star = \{O_1^\star, \ldots, O_k^\star\}$ of X. Recall that a node v at depth i in the execution tree $\mathcal{T}$ corresponds to a set C of size i – call this set C_v. Our proof will argue inductively that for each i, there will be a node v at depth i such that the centers chosen so far in C_v are *good* with respect to a subset of i clusters in $O_1^\star, \ldots, O_k^\star$. We will argue that the following invariant $P(i)$ is maintained during the recursive calls to **Sample-centers**:

***P*(*i*)**: With probability at least $\frac{1}{2^{i-1}}$, there is a node v_i at depth $(i-1)$ in the tree $\mathcal{T}$ and a set of $(i-1)$ distinct clusters $O_{j_1}^\star, O_{j_2}^\star, ..., O_{j_{i-1}}^\star$ such that

$$\forall l \in \{1, ..., i-1\}, \Phi_{c_l}(O_{j_l}^\star) \leq \left(1 + \frac{\varepsilon}{2}\right) \cdot \Delta(O_{j_l}^\star) + \frac{\varepsilon}{2k} \cdot \mathtt{opt}_k(\mathbb{O}^\star), \tag{2}$$

where $c_1, \ldots, c_{i-1}$ are the centers in the set C_{v_i} corresponding to v_i. Recall that $\Delta(O_{j_l}^\star)$ refers to the optimal 1-means cost of $O_{j_l}^\star$.

The proof of the main theorem follows easily from this invariant property – indeed, the statement $P(k)$ holds with probability at least $1/2^k$. Since the algorithm **List-k-means** invokes **Sample-centers** 2^k times, the probability of the statement in $P(k)$ being true in at least one of these invocations is at least a constant. We now prove the invariant by induction on i. The base case for $i = 1$ follows trivially: the vertex v_1 is the root of the tree $\mathcal{T}$ and C_{v_1} is empty. Now assume that $P(i)$ holds for some $i \geq 1$. We will prove that $P(i+1)$ also holds. We first condition on the event in $P(i)$ (which happens with probability at least $\frac{1}{2^{i-1}}$). Let v_i and $O_{j_1}^\star, \ldots, O_{j_{i-1}}^\star$ be as guaranteed by the invariant $P(i)$. Let $C_{v_i} = \{c_1, \ldots, c_{i-1}\}$ (as in the statement $P(i)$). For sake of ease of notation, we assume without loss of generality that the index j_i is i, and we shall use C_i to denote C_{v_i}. Thus, the center c_l corresponds to the cluster $O_l^\star$, $1 \leq l \leq i-1$. Note that for a cluster $O_{i'}^\star, i' \geq i$, $\Phi_{C_i}(O_{i'}^\star)$ is proportional to the probability that a point sampled from X using D^2-sampling w.r.t. C_i comes from the set $O_{i'}^\star$ – let $\bar{i} \in \{i, \ldots, k\}$ be the index i' for which $\Phi_{C_i}(O_{i'}^\star)$ is maximum. We will argue that the invocation of **Sample-centers** corresponding to v_i will try out a point c_i (in Step 2(d)(i)) such that the following property will hold with probability at least 1/2: $\Phi_{c_i}(O_{\bar{i}}^\star) \leq (1 + \varepsilon/2) \cdot \Delta(O_{\bar{i}}^\star) + (\varepsilon/2k) \cdot \mathtt{opt}_k(\mathbb{O}^\star)$. For doing this, we break the analysis into the following two parts. These two parts are discussed in the next two subsections that follow.

Case I. $\left(\frac{\Phi_{C_i}(O_{\bar{i}}^\star)}{\sum_{j=1}^k \Phi_{C_i}(O_j^\star)} < \frac{\varepsilon}{13k}\right)$**:** This captures the scenario where the probability of sampling from any of the uncovered clusters is very small. Note that for the classical k-means problem, this is not an issue because in this case we can argue that the current set of centers C already provides a good approximation for the entire set of data points and we are done. However, for us this is an issue — for example, assuming $i > 2$, it is possible that some of the points in $O_{\bar{i}}^\star$ are close to c_1, whereas the remaining points of this cluster are close to c_2. Still we need to output a center for $O_{\bar{i}}^\star$. In this case we argue that it will be sufficient to output a suitable convex combination of c_1 and c_2.

Case II. $\left(\frac{\Phi_{C_i}(O_{\bar{i}}^\star)}{\sum_{j=1}^k \Phi_{C_i}(O_j^\star)} \geq \frac{\varepsilon}{13k}\right)$**:** In this case, we argue that with good probability we will sample sufficient points from $O_{\bar{i}}^\star$ during Step 2(a) of **Sample-centers**. Further, we will show that a suitable combination of such points along with centers in C_i will be a good center for $O_{\bar{i}}^\star$.

3.1 Case I: $\left(\frac{\Phi_{C_i}(O_i^\star)}{\sum_{j=1}^{k}\Phi_{C_i}(O_j^\star)} < \frac{\varepsilon}{13k}\right)$

In this case we argue that a convex combination of the centers in C_i provides a good approximation to $\Delta(O_i^\star)$. Intuitively, this is because the points in $O_i^\star$ are close to the points in the set C_i. This convex combination is essentially "simulated" by taking $O(1/\varepsilon)$ copies of each of the centers $c_1, ..., c_{i-1}$ in the multi-set S and then trying all possible subsets of size $O(1/\varepsilon)$. The formal analysis follows. First, we note that $\Phi_{C_i}(O_i^\star)$ should be small compared to $\mathtt{opt}_k(\mathbb{O}^\star)$. The proof is deferred to the full version of the paper.

▶ **Lemma 8.** $\Phi_{C_i}(O_i^\star) \leq \frac{\varepsilon}{6k} \cdot \mathtt{opt}_k(\mathbb{O}^\star)$.

For each point $p \in O_i^\star$, let $c(p)$ denote the closest center in C_i. We now define a multi-set $O_i^{'}$ as $\{c(p) : p \in O_i^\star\}$. Note that $O_i^{'}$ is obtained by taking multiple copies of points in C_i. The remaining part of the proof proceeds in two steps. Let $m^\star$ and m' denote the mean of $O_i^\star$ and $O_i^{'}$ respectively. We first show that $m^\star$ and m' are close, and so, assigning all the points of $O_i^\star$ to m' will have cost close to $\Delta(O_i^\star)$. Secondly, we show that if we have a good approximation m'' to m', then assigning all the points of $O_i^\star$ to m'' will also incur small cost (comparable to $\Delta(O_i^\star)$). We now carry out these steps in detail. Observe that

$$\sum_{p \in O_i^\star} ||p - c(p)||^2 = \Phi_{C_i}(O_i^\star). \tag{3}$$

▶ **Lemma 9.** $||m^\star - m'||^2 \leq \frac{\Phi_{C_i}(O_i^\star)}{|O_i^\star|}$.

Proof. Let n denote $|O_i^\star|$. Then,

$$||m^\star - m'||^2 = \frac{1}{n^2}\left|\left|\sum_{p \in O_i^\star}(p - c(p))\right|\right|^2 \leq \frac{1}{n}\sum_{p \in O_i^\star}||p - c(p)||^2 = \frac{\Phi_{C_i}(O_i^\star)}{n},$$

where the second last inequality follows from Cauchy-Schwartz [4]. ◀

Now we show that $\Delta(O_i^\star)$ and $\Delta(O_i^{'})$ are close.

▶ **Lemma 10.** $\Delta(O_i^{'}) \leq 2 \cdot \Phi_{C_i}(O_i^\star) + 2 \cdot \Delta(O_i^\star)$.

Proof. The lemma follows by the following inequalities:

$$\begin{aligned}\Delta(O_i^{'}) &= \sum_{p \in O_i^\star}||c(p) - m'||^2 \overset{\text{Fact 1}}{\leq} \sum_{p \in O_i^\star}||c(p) - m^\star||^2 \\ &\overset{\text{Fact 4}}{\leq} 2 \cdot \sum_{p \in O_i^\star}\left(||c(p) - p||^2 + ||p - m^\star||^2\right) = 2 \cdot \Phi_{C_i}(O_i^\star) + 2 \cdot \Delta(O_i^\star).\end{aligned}$$

This completes the proof of the lemma. ◀

Finally, we argue that a good center for $O_i^{'}$ will also serve as a good center for $O_i^\star$.

▶ **Lemma 11.** *Let m'' be a point such that $\Phi_{m''}(O_i^{'}) \leq \left(1 + \frac{\varepsilon}{8}\right) \cdot \Delta(O_i^{'})$. Then $\Phi_{m''}(O_i^\star) \leq \left(1 + \frac{\varepsilon}{2}\right) \cdot \Delta(O_i^\star) + \frac{\varepsilon}{2k} \cdot \mathtt{opt}_k(\mathbb{O}^\star)$.*

[4] For any real numbers $a_1, ..., a_m$, $(\sum_r a_r)^2/m \leq \sum_r a_r^2$.

Proof. Let $n^\star$ denote $|O_{\tilde{i}}^\star|$. Observe that

$$\begin{aligned}
\Phi_{m''}(O_{\tilde{i}}^\star) &= \sum_{p\in O_{\tilde{i}}^\star} ||m''-p||^2 \\
&\overset{\text{Fact 1}}{=} \sum_{p\in O_{\tilde{i}}^\star} ||m^\star-p||^2 + n^\star\cdot||m^\star-m''||^2 \\
&\overset{\text{Fact 4}}{\leq} \Delta(O_{\tilde{i}}^\star) + 2n^\star\left(||m^\star-m'||^2+||m'-m''||^2\right) \\
&\overset{\text{Lemma 9}}{\leq} \Delta(O_{\tilde{i}}^\star) + 2\cdot\Phi_{C_i}(O_{\tilde{i}}^\star) + 2n^\star||m'-m''||^2 \\
&\overset{\text{Fact 1}}{\leq} \Delta(O_{\tilde{i}}^\star) + 2\cdot\Phi_{C_i}(O_{\tilde{i}}^\star) + 2\left(\Phi_{m''}(O_{\tilde{i}}^{'}) - \Delta(O_{\tilde{i}}^{'})\right) \\
&\leq \Delta(O_{\tilde{i}}^\star) + 2\cdot\Phi_{C_i}(O_{\tilde{i}}^\star) + \frac{\varepsilon}{4}\cdot\Delta(O_{\tilde{i}}^{'}) \\
&\overset{\text{Lemma 10}}{\leq} \Delta(O_{\tilde{i}}^\star) + 2\cdot\Phi_{C_i}(O_{\tilde{i}}^\star) + \frac{\varepsilon}{2}\cdot\left(\Phi_{C_i}(O_{\tilde{i}}^\star)+\Delta(O_{\tilde{i}}^\star)\right) \\
&\overset{\text{Lemma 8}}{\leq} \left(1+\frac{\varepsilon}{2}\right)\cdot\Delta(O_{\tilde{i}}^\star) + \frac{\varepsilon}{2k}\cdot\mathtt{opt}_k(\mathbb{O}^\star)
\end{aligned}$$

This completes the proof of the lemma. ◀

The above lemma tells us that it will be sufficient to obtain a $(1+\varepsilon/8)$-approximation to the 1-means problem for the dataset $O_{\tilde{i}}^{'}$. Now, Lemma 3 tells us that there is a subset (again as a multi-set) O'' of size $\frac{16}{\varepsilon}$ of $O_{\tilde{i}}^{'}$ such that the mean m'' of these points satisfies the conditions of Lemma 11. Now, observe that O'' will be a subset of the set S constructed in Step 2 of the algorithm **Sample-center** – indeed, in Step 2(c), we add more than $\frac{16}{\varepsilon}$ copies of *each* point in C_i to S. Now, in Step 2(d), we will try out all subsets of size $\frac{16}{\varepsilon}$ of S and for each such subset, we will try adding its mean to C_i. In particular, there will be a recursive call of this function, where we will have $C_{i+1} = C_i \cup \{m''\}$ as the set of centers. Lemma 11 now implies that C_{i+1} will satisfy the invariant $P(i+1)$. Thus, we are done in this case.

3.2 Case II: $\left(\frac{\Phi_{C_i}(O_{\tilde{i}}^\star)}{\sum_j \Phi_{C_i}(O_j^\star)} \geq \frac{\varepsilon}{13k}\right)$

In this case, we would like to prove that we add a good approximation to the mean of $O_{\tilde{i}}^\star$ to the set C_i. Again, consider the invocation of **Sample-centers** corresponding to C_i. We want the multi-set S to contain a good representation from points in the set $O_{\tilde{i}}^\star$. Secondly, in order to apply Lemma 3, we will need this representation to be a uniform sample from $O_{\tilde{i}}^\star$. Since $\Phi_{C_i}(O_{\tilde{i}}^\star) \geq \frac{\varepsilon}{13k}\cdot\sum_j \Phi_{C_i}(O_j^\star)$, the probability that a point sampled using D^2 sampling w.r.t. C_i is from $O_{\tilde{i}}^\star$ is not too small. So, the multi-set S will have non-negligible representation from the set $O_{\tilde{i}}^\star$. However the points from $O_{\tilde{i}}^\star$ in S may not be a uniform sample from $O_{\tilde{i}}^\star$. Indeed, suppose there is a good fraction of points of $O_{\tilde{i}}^\star$ which are close to C_i, and remaining points of $O_{\tilde{i}}^\star$ are quite far from C_i. Then, D^2-sampling w.r.t. to C_i will not give us a uniform sample from $O_{\tilde{i}}^\star$. To alleviate this problem, we take sufficiently many copies of points in C_i and add them to the multi-set S. In some sense, these copies act as proxy for points in $O_{\tilde{i}}^\star$ that are too close to C_i. Finally, we argue that one of the subsets of S "simulates" a uniform sample from $O_{\tilde{i}}^\star$ and the mean of this subset provides a good approximation for the mean of $O_{\tilde{i}}^\star$. The formal analysis follows.

We divide the points in $O_i^\star$ into two parts – points which are close to a center in C_i, and the remaining points. More formally, let the radius R be given by

$$R^2 = \frac{\varepsilon^2}{41} \cdot \frac{\Phi_{C_i}(O_i^\star)}{|O_i^\star|} \tag{4}$$

Define O_i^n as the points in $O_i^\star$ which are within distance R of a center in C_i, and O_i^f be the rest of the points in $O_i^\star$. As in Case I, we define a new set $O_i^{'}$ where each point in O_i^n is replaced by a copy of the corresponding point in C_i. For a point $p \in O_i^n$, define $c(p)$ as the closest center in C_i to p. Now define a multi-set $O_i^{'}$ as $O_i^f \cup \{c(p) : p \in O_i^n\}$. Intuitively, $O_i^{'}$ denotes the set of points that are same as $O_i^\star$ except that points close to centers in C_i have been "collapsed" to these centers by taking appropriate number of copies. Clearly, $|O_i^{'}| = |O_i^\star|$. At a high level, we will argue that any center that provides a good 1-means approximation for $O_i^{'}$ also provides a good approximation for $O_i^\star$. We will then focus on analyzing whether the invocation of **Sample-centers** tries out a good center for $O_i^{'}$.

We give some more notation. Let $m^\star$ and m' denote the mean of $O_i^\star$ and $O_i^{'}$ respectively. Let $n^\star$ and n denote the size of the sets $O_i^\star$ and O_i^n respectively. First, we show that $\Delta(O_i^\star)$ is large with respect to R.

▶ **Lemma 12.** $\Delta(O_i^\star) = \Phi_{m^\star}(O_i^\star) \geq \frac{16n}{\varepsilon^2} R^2$.

Proof. Let c be the center in C_i which is closest to $m^\star$. We divide the proof into two cases:

(i) $||m^\star - c|| \geq \frac{5}{\varepsilon} \cdot R$: For any point $p \in O_i^n$, triangle inequality implies that

$$||p - m^\star|| \geq ||c(p) - m^\star|| - ||c(p) - p|| \geq \frac{5}{\varepsilon} \cdot R - R \geq \frac{4}{\varepsilon} \cdot R.$$

Therefore, $\Delta(O_i^\star) \geq \sum_{p \in O_i^n} ||p - m^\star||^2 \geq \frac{16n}{\varepsilon^2} R^2$.

(ii) $||m^\star - c|| < \frac{5}{\varepsilon} \cdot R$: In this case, we have

$$\begin{aligned} \Phi_{m^\star}(O_i^\star) &\overset{\text{Fact 1}}{=} \Phi_c(O_i^\star) - n^\star \cdot ||m^\star - c||^2 \geq \Phi_{C_i}(O_i^\star) - n^\star \cdot ||m^\star - c||^2 \\ &\overset{(4)}{\geq} \frac{41n^\star}{\varepsilon^2} \cdot R^2 - \frac{25n^\star}{\varepsilon^2} \cdot R^2 \geq \frac{16n}{\varepsilon^2} R^2. \end{aligned}$$

This completes the proof of the lemma. ◀

The proofs of the following two lemmas are similar to those of Lemma 9 and Lemma 10 respectively, and are deferred to the full version of the paper.

▶ **Lemma 13.** $||m^\star - m'||^2 \leq \frac{n}{n^\star} \cdot R^2$

▶ **Lemma 14.** $\Delta(O_i^{'}) \leq 4nR^2 + 2 \cdot \Delta(O_i^\star)$.

We now argue that any center that is good for $O_i^{'}$ is also good for $O_i^\star$.

▶ **Lemma 15.** *Let m'' be such that $\Phi_{m''}(O_i^{'}) \leq \left(1 + \frac{\varepsilon}{16}\right) \cdot \Delta(O_i^{'})$. Then $\Phi_{m''}(O_i^\star) \leq \left(1 + \frac{\varepsilon}{2}\right) \cdot \Delta(O_i^\star)$.*

Proof. The lemma follows from the following inequalities:

$$\begin{aligned} \Phi_{m''}(O_i^\star) &= \sum_{p \in O_i^\star} ||m'' - p||^2 \\ &\overset{\text{Fact 1}}{=} \sum_{p \in O_i^\star} ||m^\star - p||^2 + n^\star \cdot ||m^\star - m''||^2 \end{aligned}$$

$$\begin{aligned}
&\overset{\text{Fact 4}}{\leq} && \Delta(O_{\bar{i}}^{\star}) + 2n^{\star}\left(||m^{\star} - m'||^2 + ||m' - m''||^2\right)\\
&\overset{\text{Lemma 13}}{\leq} && \Delta(O_{\bar{i}}^{\star}) + 2nR^2 + 2n^{\star} \cdot ||m' - m''||^2\\
&\overset{\text{Fact 1}}{\leq} && \Delta(O_{\bar{i}}^{\star}) + 2nR^2 + 2 \cdot \left(\Phi_{m''}(O_{\bar{i}}^{'}) - \Delta(O_{\bar{i}}^{'})\right)\\
&\leq && \Delta(O_{\bar{i}}^{\star}) + 2nR^2 + \frac{\varepsilon}{8} \cdot \Delta(O_{\bar{i}}^{'})\\
&\overset{\text{Lemma 14}}{\leq} && \Delta(O_{\bar{i}}^{\star}) + 2nR^2 + \frac{\varepsilon}{2} \cdot nR^2 + \frac{\varepsilon}{4} \cdot \Delta(O_{\bar{i}}^{\star})\\
&\overset{\text{Lemma 12}}{\leq} && \left(1 + \frac{\varepsilon}{2}\right) \cdot \Delta(O_{\bar{i}}^{\star}).
\end{aligned}$$

This completes the proof of the lemma. ◀

Given the above lemma, all we need to argue is that our algorithm indeed considers a center m'' such that $\Phi_{m''}(O_{\bar{i}}^{'}) \leq (1 + \varepsilon/16) \cdot \Delta(O_{\bar{i}}^{'})$. For this we would need about $O(1/\varepsilon)$ uniform samples from $O_{\bar{i}}^{'}$. However, our algorithm can only sample using D^2-sampling w.r.t. C_i. For ease of notation, let $c(O_{\bar{i}}^{n})$ denote the multi-set $\{c(p) : p \in O_{\bar{i}}^{n}\}$. Recall that $O_{\bar{i}}^{'}$ consists of $O_{\bar{i}}^{f}$ and $c(O_{\bar{i}}^{n})$. The first observation is that the probability of sampling an element from $O_{\bar{i}}^{f}$ is reasonably large (proportional to ε/k). Using this fact, we show how to sample from $O_{\bar{i}}^{'}$ (almost uniformly). Finally, we show how to convert this almost uniform sampling to uniform sampling (at the cost of increasing the size of sample). We defer the proof of the following lemma to the full version of the paper.

▶ **Lemma 16.** *Let x be a sample from D^2-sampling w.r.t. C_i. Then, $\mathbf{Pr}[x \in O_{\bar{i}}^{f}] \geq \frac{\varepsilon}{15k}$. Further, for any point $p \in O_{\bar{i}}^{f}$, $\mathbf{Pr}[x = p] \geq \frac{\gamma}{|O_{\bar{i}}^{\star}|}$, where γ denotes $\frac{\varepsilon^2}{533k}$.*

Let $X_1, \ldots X_l$ be l points sampled independently using D^2-sampling w.r.t. C_i. We construct a new set of random variables $Y_1, \ldots, Y_l$. Each variable Y_u will depend on X_u only, and will take values either in $O_{\bar{i}}^{'}$ or will be $\perp$. These variables are defined as follows: if $X_u \notin O_{\bar{i}}^{f}$, we set Y_u to $\perp$. Otherwise, we assign Y_u to one of the following random variables with equal probability: (i) X_u or (ii) a random element of the multi-set $c(O_{\bar{i}}^{n})$. The following observation follows from Lemma 16, and its proof is deferred to the full version of the paper.

▶ **Corollary 17.** *For a fixed index u, and an element $x \in O_{\bar{i}}^{'}$, $\mathbf{Pr}[Y_u = x] \geq \frac{\gamma'}{|O_{\bar{i}}^{'}|}$, where $\gamma' = \gamma/2$.*

Corollary 17 shows that we can obtain samples from $O_{\bar{i}}^{'}$ which are nearly uniform (up to a constant factor). To convert this to a set of uniform samples, we use the idea of [9]. For an element $x \in O_{\bar{i}}^{'}$, let γ_x be such that $\frac{\gamma_x}{|O_{\bar{i}}^{'}|}$ denotes the probability that the random variable Y_u is equal to x (note that this is independent of u). Corollary 17 implies that $\gamma_x \geq \gamma'$. We define a new set of independent random variables $Z_1, \ldots, Z_l$. The random variable Z_u will depend on Y_u only. If Y_u is $\perp$, Z_u is also $\perp$. If Y_u is equal to $x \in O_{\bar{i}}^{'}$, then Z_u takes the value x with probability $\frac{\gamma'}{\gamma_x}$, and $\perp$ with the remaining probability. Note that Z_u is either $\perp$ or one of the elements of $O_{\bar{i}}^{'}$. Further, conditioned on the latter event, it is a uniform sample from $O_{\bar{i}}^{'}$. We can now give the key lemma (proof is deferred to the full version).

▶ **Lemma 18.** *Let l be $\frac{128}{\gamma' \cdot \varepsilon}$, and m'' denote the mean of the non-null samples from $Z_1, \ldots, Z_l$. Then, with probability at least $1/2$, $\Phi_{m''}(O_{\bar{i}}^{'}) \leq (1 + \varepsilon/16) \cdot \Delta(O_{\bar{i}}^{'})$.*

Let $C_i^{(l)}$ denote the multi-set obtained by taking l copies of each of the centers in C_i. Now observe that all the non-$\perp$ elements among $Y_1, \ldots, Y_l$ are elements of $\{X_1, \ldots, X_l\} \cup C_i^{(l)}$, and so the same must hold for $Z_1, \ldots, Z_l$. This implies that in Step 2(d) of the algorithm **Sample-centers**, we would have tried adding the point m'' as described in Lemma 18. Therefore, the induction hypothesis continues to hold with probability at least $1/2$. This concludes the proof of Theorem 5.

References

1. Marcel R. Ackermann, Johannes Blömer, and Christian Sohler. Clustering for metric and nonmetric distance measures. *ACM Trans. Algorithms*, 6:59:1–59:26, September 2010. URL: http://doi.acm.org/10.1145/1824777.1824779.
2. Mihai Bādoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proc. of the 34th Annual ACM Symp. on Theory of Computing*, STOC'02, pages 250–257, New York, NY, USA, 2002. ACM. doi:10.1145/509907.509947.
3. Ke Chen. On k-median clustering in high dimensions. In *Proc. of the 17th Annual ACM-SIAM Symp. on Discrete Algorithm*, SODA'06, pages 1177–1185, New York, NY, USA, 2006. ACM. doi:10.1145/1109557.1109687.
4. W. Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In *Proc. of the 35th Annual ACM Symp. on Theory of Computing*, STOC'03, pages 50–58, New York, NY, USA, 2003. ACM. doi:10.1145/780542.780550.
5. Hu Ding and Jinhui Xu. A unified framework for clustering constrained data without locality property. In *Proc. of the 26th Annual ACM-SIAM Symp. on Discrete Algorithms*, SODA'15, pages 1471–1490, 2015. doi:10.1137/1.9781611973730.97.
6. Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proc. of the 23rd Annual Symp. on Computational Geometry*, SoCG'07, pages 11–18, New York, NY, USA, 2007. ACM. doi:10.1145/1247069.1247072.
7. Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proc. of the 36th Annual ACM Symp. on Theory of Computing*, STOC'04, pages 291–300, New York, NY, USA, 2004. ACM. doi:10.1145/1007352.1007400.
8. Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering: (extended abstract). In *Proc. of the 10th Annual Symp. on Computational Geometry*, SoCG'94, pages 332–339, New York, NY, USA, 1994. ACM. doi:10.1145/177424.178042.
9. Ragesh Jaiswal, Amit Kumar, and Sandeep Sen. A simple D^2-sampling based PTAS for k-means and other clustering problems. *Algorithmica*, 70(1):22–46, 2014. doi:10.1007/s00453-013-9833-9.
10. Ragesh Jaiswal, Mehul Kumar, and Pulkit Yadav. Improved analysis of D^2-sampling based PTAS for k-means and other clustering problems. *Information Processing Letters*, 115(2):100–103, 2015. doi:http://dx.doi.org/10.1016/j.ipl.2014.07.009.
11. Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2):5:1–5:32, February 2010. doi:10.1145/1667053.1667054.
12. J. Matoušek. On approximate geometric k-clustering. *Discrete and Computational Geometry*, 24(1):61–84, 2000. doi:10.1007/s004540010019.