A PTAS for k-Means Clustering Based on Weak Coresets

Dan Feldman[*]
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
dannyf@post.tau.ac.il

Morteza Monemizadeh[†]
Heinz Nixdorf Institute and
Computer Science
Department
University of Paderborn
D-33102 Paderborn, Germany
monemi@hni.uni-paderborn.de

Christian Sohler[‡]
Heinz Nixdorf Institute and
Computer Science
Department
University of Paderborn
D-33102 Paderborn, Germany
csohler@uni-paderborn.de

ABSTRACT

Given a point set $\mathcal{P} \subseteq \mathbb{R}^d$ the k-means clustering problem is to find a set $C = \{c_1, \ldots, c_k\}$ of k points and a partition of $\mathcal{P}$ into k clusters $C_1, \ldots, C_k$ such that the sum of squared errors $\sum_{i=1}^{k} \sum_{p \in C_i} \|p - c_i\|_2^2$ is minimized. For given centers this cost function is minimized by assigning points to the nearest center. The k-means cost function is probably the most widely used cost function in the area of clustering.

In this paper we show that every unweighted point set $\mathcal{P}$ has a weak (ϵ, k)-coreset of size $\text{poly}(k, 1/\epsilon)$ for the k-means clustering problem, i.e. its size is *independent* of the cardinality $|\mathcal{P}|$ of the point set and the dimension d of the Euclidean space $\mathbb{R}^d$. A weak coreset is a weighted set $\mathcal{S} \subseteq \mathcal{P}$ together with a set $\mathcal{T}$ such that $\mathcal{T}$ contains a $(1 + \epsilon)$-approximation for the optimal cluster centers from P and for every set of k centers from $\mathcal{T}$ the cost of the centers for $\mathcal{S}$ is a $(1 \pm \epsilon)$-approximation of the cost for P.

We apply our weak coreset to obtain a PTAS for the k-means clustering problem with running time $O(nkd + d \cdot \text{poly}(k/\epsilon) + 2^{\tilde{O}(k/\epsilon)})$.

Categories and Subject Descriptors

F.2.2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity—*Nonnumerical Algorithms and Problems*

General Terms

Algorithms, Theory

Keywords

Geometric Optimization, k-means, approximation, coresets

1. INTRODUCTION

Clustering is the process to partition a given set of objects into sets called clusters such that objects in the same cluster are similar and objects in different sets are dissimilar. Clustering has many applications in different areas including bioinformatics, pattern recognition, data compression, and information retrieval. Because of the wide variety of applications, there is no general formulation of clustering. However, some formulations have been very successful for a variety of applications. One of these is the k-means clustering problem. In this problem we are given a set $\mathcal{P}$ of points in $\mathbb{R}^d$ and we try to find a set $C \subseteq \mathbb{R}^d$ of k cluster centers $\{c_1, \ldots, c_k\}$ and a corresponding partition $C_1, \ldots, C_k$ of $\mathcal{P}$ into k clusters such that the sum of squared errors $\sum_{i=1}^{k} \sum_{p \in C_i} \|p - c_i\|_2^2$ is minimized. The k-means clustering problem has been studied intesively both in theory and practice. One of the most widely used clustering algorithm is Lloyd's algorithm [15]. Although this algorithm is only a heuristic, i.e. it does not guarantee a certain approximation guarantee, it has proved to be very useful in many applications. Trying to explain the popularity of Lloyd's algorithm, Ostrovsky et al. [19] showed that for well-separated instances, a variant of this algorithm is a (1+ϵ)-approximation algorithm for k-means clustering with running time $O(2^{(k/\epsilon)}dn)$. However, their separation criterion depends on ϵ, i.e. the smaller ϵ becomes the stronger separation is required.

Early work PTAS for k-means clustering started with the work of Inaba et al. [12] who observed that the number of Voronoi partitions of k points in $\mathbb{R}^d$ is n^{dk} and thus an *optimal* clustering be computedin time $O(n^{dk+1})$. Matousek [17] presented a $(1 + \epsilon)$-approximation algorithm for k-means clustering, with running time $O(n\epsilon^{-2k^2 d} \log^k n)$. Har-Peled and Mazumdar [10] used coresets to improve the running time to $O(n + k^{k+2}\epsilon^{-(2d+1)} \log^{k+1} n \log^{k+1} 1/\epsilon)$. Fernandez de la Vega et al. [4] proposed a $(1 + \epsilon)$-approximation algorithm, for high dimensions (they refer to it as l_2^2 k-median clustering), with running time $O(g(k, \epsilon)d^{O(1)}n \log^{O(k)} n)$, where $g(k, \epsilon) = \exp\left\{(k^3/\epsilon^8)(\ln(k/\epsilon)) \ln k\right\}$. Kumar et al. [13, 14] showed a $(1 + \epsilon)$-approximation algorithm for k-means clustering running in $O(2^{(k/\epsilon)^{O(1)}} dn)$ time. Chen [3] gave a new coreset construction that can be combined with the previously mentioned algorithm[13] to improve the running time to $O(ndk + 2^{(k/\epsilon)^{O(1)}} d^2 n^{\sigma})$.

Our results.

In this paper we develop a $(1+\epsilon)$-approximation for the k-means clustering problem with running time $O(nkd + d \cdot \text{poly}(k/\epsilon) + 2^{\tilde{O}(k/\epsilon)})$. This significantly improves the previous best PTAS with running time $O(ndk + 2^{(k/\epsilon)^{O(1)}} d^2 n^{\sigma})$ [3].

[*] Supported by the German-Israel Foundation for Scientific Research and Development

[†] Supported by the DFG Research Training Group GK-693 of the Paderborn Institute for Scientific Computation (PaSCo)

[‡] Research supported in part by DFG grant So 514/1-1.

SCG'07, June 6–8, 2007, Gyeongju, South Korea.
Copyright 2007 ACM 978-1-59593-705-6/07/0006 ...$5.00.

The main ingredient of our algorithm is a procedure to compute in $O(nkd)$ time a weak (k,ϵ)-coreset of size $\mathrm{poly}(k/\epsilon)$, which is independent of $|\mathcal{P}|$ and d. This weak coreset is a weighted set $\mathcal{S}$ of points together with another set of points $\mathcal{T}$ such that the cost any set of k centers from $\mathcal{T}$ is a $(1 \pm \epsilon)$-approximstion of the cost of $\mathcal{P}$ and $\mathcal{T}$ contains a $(1+\epsilon)$-approximation of the optimal solution. Such a set is called (k,ϵ)-approximate centroid set. Our coreset implies that one can construct in $O(nkd)$ time a (k,ϵ)-approximate centroid set of size independent of n and d.

Another interesting application of our coreset is the kernel k-means algorithm. In this algorithm points are implicitly mapped to a high dimensional space. The dimension of this space may be very large (possibly infinite) and so coreset constructions that depend on d are of little use. Also we cannot apply dimensionality reduction techniques since the point coordinates in the high dimensional space are not known. If this space has finite dimension we can apply our coreset construction and obtain a PTAS for the kernel k-means algorithm.

Finally, we remark that our construction can be carried over to the k-median problem. Further, we can also apply a variant of our coreset to obtain data streaming algorithms, but the size will be depend on $\log n$.

Our techniques.

The main new technique in this paper is a non-uniform sampling scheme to construct the coreset. Points are sampled based on their distance from a constant factor approximation. In contrast to previous approaches we also use non-uniformly distributed weights for the points.

Other related work.

Har-Peled and Mazumdar [10] used (strong) coresets to obtain faster algorithms for clustering problems. Roughly speaking, a strong coreset for k-means is a weighted subset $\mathcal{S}$ of $\mathcal{P}$, so that for any set of k points in $\mathbb{R}^d$, the weighted sum of squared distances from points in $\mathcal{S}$ to the nearest centers is approximately the same as (differs by a factor of $1 \pm \epsilon$ from) the sum of squared distances from points in $\mathcal{P}$ to the nearest centers. Their coreset was of size $O(k\epsilon^{-d}\log n)$. They also used coresets to compute an $(1+\epsilon)$-approximation k-median and k-means clustering in the streaming model of computation using $O(k\epsilon^{-d}\log^{2d+2} n)$ space. Their algorithms handle streams with insertions only. Then Frahling and Sohler [7] showed that they can maintain a coreset of size $O(k\epsilon^{-d-2}$ $\log n)$ for the same problem using a different coreset construction, which also works for data streams with insertions and deletions. Har-Peled and Kushal [9] recently showed that one can construct coresets for k-means with size independent of n, namely of size $O(k^3\epsilon^{-d-1})$. Very recently Feldman, Fiat, and Sharir [6] extended this type of coresets for linear centers or facilites where facilities can be lines, flats. For high dimensional spaces Chen [3] proposed a coreset of size $O(k^2 d\epsilon^{-2}\log^2 n)$.

Matousek's result [17] was based on the idea of centroid set. A centroid set is a set that contains at least one k-tuple, which forms (approximately) optimal centers for k-means clustering. In particular, he showed that there exists an ϵ-approximate centroid set of size $n/\epsilon^d \log(1/\epsilon)$. Effros and Shulman [5] showed that there exists a centroid set of size $\epsilon^{-d-1}(k^4 + k^2\epsilon^{-2})$. This result showed that it might be possible to have a centroid set and coreset independent of input set. Then Har-Peled and Mazumdar [10] and Har-Peled and Kushal [9] used from this fact that Matousek's construction is weight sensitive and they used their coreset as input set for Matousek's construction to obtain an ϵ-approximate centroid set of size $k/\epsilon^{2d}\log n \log(1/\epsilon)$ and $k^3/\epsilon^{2d+1}\log(1/\epsilon)$ respectively. We should mention that an implicit result of Kumar et al. [12, 13, 14] is an ϵ-approximate centroid set of size $n^{1/\epsilon}$.

2. PRELIMINARIES

A set of points $\mathcal{P}$ in $\mathbb{R}^d$ is weighted, if each point $p \in \mathcal{P}$ is associated with a weight $w_p > 0$. We define $w(\mathcal{P}) = \sum_{p\in\mathcal{P}} w_p$ to be the total weight of $\mathcal{P}$. We consider an (unweighted) set of points $\mathcal{P} \subseteq \mathbb{R}^d$ as a weighted set with $w_p = 1$, for each $p \in \mathcal{P}$.

For two points $p, q \in \mathbb{R}^d$ we use $\mathrm{dist}(p,q)$ to denote the Euclidean distance between p and q. $\Delta(p,q) = (\mathrm{dist}(p,q))^2$ will denote the square of the Euclidean distance. We generalize these definitions to sets: Given a point $p \in \mathbb{R}^d$ and a set of points $Q \subseteq \mathbb{R}^d$ we define $\mathrm{dist}(p,Q) = \min_{q\in Q}\mathrm{dist}(p,q)$ and $\Delta(p,Q) = \min_{q\in Q}\Delta(p,q)$. For a weighted set $\mathcal{P} \subseteq \mathbb{R}^d$ and unweighted set $\mathcal{K} \subseteq \mathbb{R}^d$ we define $\mathbf{cost}(\mathcal{P},\mathcal{K}) = \sum_{p\in P} w(p)\cdot\Delta(p,\mathcal{K})$. Further, we define the distance between two sets $Q, R \subseteq \mathbb{R}^d$ as $\mathrm{dist}(Q,R) = \min_{q\in Q}\mathrm{dist}(q,R)$.

Definition 1 (k-means clustering). *Given a set $\mathcal{P}$ of points in the $\mathbb{R}^d$ the k-means problem is to find a set of k centers $\mathcal{K} \subseteq \mathbb{R}^d$ such that $\mathbf{cost}(\mathcal{P},\mathcal{K})$ is minimized.*

Given an integer $k \geq 1$, we denote by $\mathbf{OPT}(\mathcal{P},k) = \min_{|\mathcal{K}|=k}$ $\mathbf{cost}(\mathcal{P},\mathcal{K})$ the optimal k-mean cost of $\mathcal{P}$. A set $\mathcal{K} \subset \mathbb{R}^d$, $|\mathcal{K}| = k$, is a β-approximation for an optimal k-means solution of P, if $\mathbf{cost}(\mathcal{P},\mathcal{K}) \leq \beta\cdot\mathbf{OPT}(\mathcal{P},k)$. The point $\mu_{\mathcal{P}}(\mathcal{P}) = \frac{\sum_{p\in\mathcal{P}} p}{|\mathcal{P}|}$ is the *centroid* of $\mathcal{P}$. For the 1-means problem the centroid is known to be the optimal cluster center, i.e. $\mathbf{OPT}(\mathcal{P},1) = \sum_{p\in\mathcal{P}}\Delta(p,\mu_{\mathcal{P}}(\mathcal{P}))$. Inaba and *et al.*[12] showed that if we draw a random sample U of size $O(1/\epsilon)$ with constant probability the centroid of U is with constant probability a $(1+\epsilon)$-approximation for the centroid of point set $\mathcal{P}$, that is, $\mathbf{cost}(\mathcal{P},\mu_{\mathcal{P}}(\mathbf{U})) \leq (1+\epsilon)\mathbf{cost}(\mathcal{P},\mu_{\mathcal{P}}(\mathcal{P}))$. This implies (see also [13, 14])

Corollary 2. *Let $\mathcal{P}$ be a set of points in $\mathbb{R}^d$. Then there exists a set $U \subseteq \mathcal{P}$ of size $2/\epsilon$, such that*

$$\mathbf{cost}(\mathcal{P},\mu_{\mathcal{P}}(\mathbf{U})) \leq (1+\epsilon)\mathbf{cost}(\mathcal{P},\mu_{\mathcal{P}}(\mathcal{P})) \ .$$

A set $\mathcal{T} \subseteq \mathbf{R}^d$ is a (k,ϵ)-*approximate centroid set* for $\mathcal{P}$, if there exists a subset $C \subseteq \mathcal{T}$ of size k such that $\mathbf{cost}(\mathcal{P},C) \leq (1+\epsilon)\cdot\mathbf{OPT}(\mathcal{P},k)$. For technical reasons our definition of a weak coreset slightly differs from previous definitions given in [2] and [11].

Definition 3 (Weak (k,ϵ)-coreset). *Let $\mathcal{P}$ be a (possibly) weighted set in $\mathbf{R}^d$. A pair $(\mathcal{S},\mathcal{T})$, $\mathcal{S} \subseteq \mathcal{P}$, is called a weak (k,ϵ)-coreset, if $\mathcal{T}$ is a (k,ϵ)-approximate centroid set for $\mathcal{P}$ and*

$$|\mathbf{cost}(\mathcal{S},\mathcal{K}) - \mathbf{cost}(\mathcal{P},\mathcal{K})| \leq \epsilon\cdot\mathbf{cost}(\mathcal{P},\mathcal{K})$$

for any set $\mathcal{K} \subseteq \mathcal{T}$ with $|\mathcal{K}| = k$.

Notice that our coreset does not imply that an optimal solution for $\mathcal{S}$ is a $(1+\epsilon)$-approximation for $\mathcal{P}$. We can only give guarantees for points from $\mathcal{T}$.

3. THE CORESET CONSTRUCTION

The first step of our algorithm is similar to the coreset constructions in [10] and [3]. We run the constant factor approximation algorithm for k-means clustering due to [18] that in $O(nkd)$ time returns k centers $\mathcal{C} = \{c_1,\cdots,c_k\}$. Let β denote the approximation factor of this algorithm and let C_i denote the set of points from $\mathcal{P}$ that are nearer to c_i than to any other center in $\mathcal{C}$ (ties can be broken arbitrarily). We partition each C_i into two sets C_i^{in} and C_i^{out}. The set C_i^{in} contains all points that are close to c_i, i.e. all points contained in a closed ball $\mathbf{b}(c_i, r_i) = \{p \in \mathbb{R}^d \mid \text{dist}(p, c_i) \leq r_i\}$ with center c_i and radius $r_i = \sqrt{\frac{\mathbf{cost}(C_i, c_i)}{\epsilon \cdot |C_i|}})$. Thus we have $C_i^{\text{in}} = C_i \bigcap \mathbf{b}(c_i, r_i)$. The set C_i^{out} contains the remaining points of C_i, i.e. $C_i^{\text{out}} = C_i \setminus \mathbf{b}(c_i, r_i)$.

To construct the coreset we proceed differently for the points in C_i^{in} and C_i^{out}. From the sets C_i^{in} we draw a set of s_i^{in} points independently and uniformly at random. Then we assign to each point the weight $|C_i^{\text{in}}|/s_i^{\text{in}}$. Let S_i^{in} denote the resulting weighted sample. From each set C_i^{out} we draw a sample set $S_i^{\text{out}} = \{s_1, \cdots, s_{s_i^{\text{out}}}\}$ according to a non-uniform probability distribution. The weights of the points will also be distributed non-uniformly.

We proceed for each cluster separately. The probability of choosing point $q \in C_i^{\text{out}}$ is $p_q = \Pr[q \in S_i^{\text{out}}] = \frac{\Delta(q, c_i)}{\mathbf{cost}(C_i^{\text{out}}, c_i)}$. Each sample point q is assigned a weight $w_q = \frac{\mathbf{cost}(C_i^{\text{out}}, c_i)}{s_i^{\text{out}} \Delta(q, c_i)}$, i.e. the weight of a point depends on its distance to the center c_i. The further a point is away from the center, the smaller its weight.

Finally, we set $\mathcal{S} = \bigcup_{i=1}^{k} \left(S_i^{\text{in}} \cup S_i^{\text{out}}\right)$ and $\mathcal{T}$ will be the set of all centroids of combinations of $2/\epsilon$ points from $\mathcal{S}$ (we allow repetition of points). We will show that for large enough sample sizes, our construction indeed gives a weak coreset. We remark that the set $\mathcal{T}$ depends on the choice of the set $\mathcal{S}$.

4. ANALYSIS

Overview.

We first show that $\mathcal{T}$ is a $(k, 6\epsilon)$-approximate centroid set, if the cost of any subset $\mathcal{K} \subseteq \mathcal{T}$, $|\mathcal{K}| = k$, and the cost of an optimal solution $\mathcal{O}$ are approximated within a factor of $(1 \pm \epsilon)$. Then we show that for an arbitrary set $\mathcal{K}$ of k centers $|\mathbf{cost}(\mathcal{S}, \mathcal{K}) - \mathbf{cost}(\mathcal{P}, \mathcal{K})| \leq \epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K})$ with probability $1 - \lambda$ for large enough s_i^{in} and s_i^{out}. This implies that with this probability $\mathcal{T}$ is a $(k, 6\epsilon)$-approximate centroid set. Finally, we show that for any subset of $\mathcal{T}$ of size k we get $|\mathbf{cost}(\mathcal{S}, \mathcal{K}) - \mathbf{cost}(\mathcal{P}, \mathcal{K})| \leq \epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K})$. Here, the difficulty is that the set $\mathcal{T}$ depends on the random process and hence there are dependencies (this is, why we cannot immediately apply the previous result).

$\mathcal{T}$ is a (k, ϵ)-approximate centroid set.

LEMMA 4. *Let $P \subseteq \mathbb{R}^d$ be a point set and let $0 < \epsilon < 1/2$ and $k \geq 1$. Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a weighted point set, and let $\mathcal{T}$ be the set of all centroids of combinations (with repetition) of $2/\epsilon$ points from $\mathcal{S}$. If we have $|\mathbf{cost}(\mathcal{S}, \mathcal{K}) - \mathbf{cost}(\mathcal{P}, \mathcal{K})| \leq \epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K})$ for every set $\mathcal{K} \subseteq \mathcal{T}$ of k points and if $|\mathbf{cost}(\mathcal{S}, \mathcal{O}) - \mathbf{cost}(\mathcal{P}, \mathcal{O})| \leq \epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{O})$ for an optimal set $\mathcal{O} \subseteq \mathbb{R}^d$ of k centers, then $\mathcal{T}$ is a $(k, 6\epsilon)$-approximate centroid set.*

Proof. Let $\mathcal{O}$ denote an optimal set of cluster centers and let $O_1, \ldots, O_k$ be the induced clustering. By Corollary 2 we know that for every cluster O_i the set $\mathcal{T}$ contains a $(1+\epsilon)$-approximation. Since the cost of $\mathcal{O}$ is approximated within a factor of $(1 \pm \epsilon)$ and we lose at most another factor of $(1 + \epsilon)$ when we move each center to the nearest point from $\mathcal{T}$, we have a set $\mathcal{K}^*$ of k centers in $\mathcal{T}$ with $\mathbf{cost}(\mathcal{S}, \mathcal{K}^*) \leq (1 + \epsilon)^2 \cdot \mathbf{cost}(\mathcal{P}, O)$. Since we also know that $\mathbf{cost}(\mathcal{S}, \mathcal{K}^*) \geq (1 - \epsilon)\mathbf{cost}(\mathcal{P}, \mathcal{K}^*)$, we know that $\mathbf{cost}(\mathcal{P}, \mathcal{K}^*) \leq (1 + \epsilon)^2/(1 - \epsilon) \cdot \mathbf{cost}(\mathcal{P}, \mathcal{O})$. For $\epsilon \leq 1/2$ we can obtain that $\mathcal{T}$ is a $(k, 6\epsilon)$-approximate centroid set. □

Arbitrary centers are approximated within a factor $(1 \pm \epsilon)$.

The first step of our analysis will be to show that for an arbitrary fixed set $\mathcal{K} = \{k_1, \cdots, k_l, \cdots, k_k\}$ of k centers, the cost of $\mathcal{S}$ is a $(1 \pm \epsilon)$-approximation of the cost of $\mathcal{P}$. We will prove the following lemma.

LEMMA 5. *Given a point set $\mathcal{P}$ in $\mathbf{R}^d$ and a set $\mathcal{K} \subseteq \mathbb{R}^d$ of k centers. Let $\epsilon, \lambda > 0$ be parameters. Then there is a constant c such that for $s_i^{\text{in}}, s_i^{\text{out}} \geq c \cdot \frac{\ln(k/\lambda)}{\epsilon^4}$, $1 \leq i \leq k$, the sample set $\mathcal{S} = \bigcup_{i=1}^{k} \left(S_i^{\text{in}} \cup S_i^{\text{out}}\right)$ computed by our algorithm satisfies $|\mathbf{cost}(\mathcal{S}, \mathcal{K}) - \mathbf{cost}(\mathcal{P}, \mathcal{K})| \leq \epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K})$ with probability $\geq 1 - \lambda$.*

Proof. Let $S_i = S_i^{\text{in}} \cup S_i^{\text{out}}$and let k_l denote the nearest center from $\mathcal{K}$ to $\mathbf{b}(c_i, r_i)$. The analysis will distinguish between the cases (a) $\text{dist}(k_l, c_i) \geq r_i + \frac{r_i}{\epsilon} = \frac{r_i(1+\epsilon)}{\epsilon}$ and (b) $\text{dist}(k_l, c_i) < r_i + \frac{r_i}{\epsilon} = \frac{r_i(1+\epsilon)}{\epsilon}$. We will assume $\epsilon \leq 1/2$.

Case (a).

Every point $p \in \mathbf{b}(c_i, r_i)$ has distance at least $\text{dist}(\mathbf{b}(c_i, r_i), k_l)$ to the nearest center from $\mathcal{K}$. Since we are in case (a), it has distance at most $\text{dist}(\mathbf{b}(c_i, r_i), k_l) + 2r_i \leq (1+2\epsilon) \cdot \text{dist}(\mathbf{b}(c_i, r_i), k_l)$ to the nearest center from $\mathcal{K}$. By our construction we have that the sum of the weights of the points in S_i^{in} is exactly $|C_i^{\text{in}}|$. Hence, we get

$$\begin{aligned}
& \left|\mathbf{cost}(C_i^{\text{in}}, \mathcal{K}) - \mathbf{cost}(S_i^{\text{in}}, \mathcal{K})\right| \\
\leq \; & |C_i^{\text{in}}| \left(\left((1 + 2\epsilon) \cdot \text{dist}(\mathbf{b}(c_i, r_i), k_l)\right)^2 - \text{dist}(\mathbf{b}(c_i, r_i), k_l)^2\right) \\
\leq \; & 8 \cdot \epsilon \cdot |C_i^{\text{in}}| \cdot \Delta(\mathbf{b}(c_i, r_i), k_l) \\
\leq \; & 8 \cdot \epsilon \cdot \mathbf{cost}(C_i^{\text{in}}, \mathcal{K}) \ .
\end{aligned}$$

Next we consider the points from C_i^{out}. Let $W = \sum_{p \in S_i^{\text{out}}} w_p$ be the random variable for the sum of weights of points in S_i^{out}. In case (a) we will approximate the error for the outer points by the sum of their contributions. We have

$$\begin{aligned}
& |\mathbf{cost}(C_i^{\text{out}}, \mathcal{K}) - \mathbf{cost}(S_i^{\text{out}}, \mathcal{K})| \\
\leq \; & \mathbf{cost}(C_i^{\text{out}}, \mathcal{K}) + \mathbf{cost}(S_i^{\text{out}}, \mathcal{K}) \\
\leq \; & \sum_{q \in C_i^{\text{out}}} \mathbf{cost}(q, k_l) + \sum_{p \in S_i^{\text{out}}} w_p \cdot \mathbf{cost}(p, k_l) \ .
\end{aligned}$$

Now we use the doubled triangle inequality, i.e. $\Delta(p, r) \leq 2(\Delta(p, q)$ $+ \Delta(q, r))$ for all $p, q, r \in \mathbb{R}^d$, to obtain

$$\begin{aligned}
& \sum_{q \in C_i^{\text{out}}} \Delta(q, k_l) + \sum_{p \in S_i^{\text{out}}} w_p \cdot \Delta(p, k_l) \\
\leq \; & \sum_{q \in C_i^{\text{out}}} 2\left(\Delta(p, c_i) + \Delta(c_i, k_l)\right) \\
+ \; & \sum_{p \in S_i^{\text{out}}} 2w_p \left(\Delta(p, c_i) + \Delta(c_i, k_l)\right) \\
\leq \; & \sum_{q \in C_i^{\text{out}}} 2\big(\Delta(q, c_i) + 2(r_i^2 + \Delta(\mathbf{b}(c_i, r_i), k_l))\big) \\
+ \; & \sum_{p \in S_i^{\text{out}}} 2w_p \big(\Delta(p, c_i) + 2(r_i^2 + \Delta(\mathbf{b}(c_i, r_i), k_l))\big) \\
\leq \; & \big(6\mathbf{cost}(C_i, c_i) + 4|C_i^{\text{out}}| \cdot \Delta(\mathbf{b}(c_i, r_i), k_l)\big) \\
+ \; & \Big(\sum_{p \in S_i^{\text{out}}} 6w_p \Delta(p, c_i) + \sum_{p \in S_i^{\text{out}}} 4w_p \Delta(\mathbf{b}(c_i, r_i), k_l)\big)
\end{aligned}$$

$$\begin{aligned}
&\le \big(6r_i^2\epsilon \cdot |C_i| + 4\epsilon \cdot |C_i| \cdot \Delta(\mathbf{b}(c_i, r_i), k_l)\big) \\
&+ \big(6\mathbf{cost}(C_i^{\mathrm{out}}, c_i) + 4\Delta(\mathbf{b}(c_i, r_i), k_l) \sum_{p \in S_i^{\mathrm{out}}} w_p\big) \\
&\le \big(6r_i^2\epsilon \cdot |C_i| + 4\epsilon \cdot |C_i| \cdot \Delta(\mathbf{b}(c_i, r_i), k_l)\big) \\
&+ \big(6 \cdot \mathbf{cost}(C_i, c_i) + 4W \cdot \Delta(\mathbf{b}(c_i, r_i), k_l)\big) \\
&\le \big(6r_i^2\epsilon \cdot |C_i| + 4\epsilon \cdot |C_i| \cdot \Delta(\mathbf{b}(c_i, r_i), k_l)\big) \\
&+ \big(6r_i^2\epsilon \cdot |C_i| + 4W \cdot \Delta(\mathbf{b}(c_i, r_i), k_l)\big)
\end{aligned}$$

Since dist$(\mathbf{b}(c_i, r_i), k_l) \ge \frac{r_i}{\epsilon}$ implies $r_i^2 \le \epsilon^2 \cdot \Delta(\mathbf{b}(c_i, r_i), k_l)$ we have for $\epsilon \le 1/2$:

$$\begin{aligned}
&\le (2\epsilon \cdot |C_i| \cdot \Delta(C_i^{\mathrm{in}}, k_l) \cdot (6\epsilon^2 + 2) + 4W \cdot \Delta(\mathbf{b}(c_i, r_i), k_l) \\
&\le 10\epsilon \cdot |C_i^{\mathrm{in}}| \cdot \Delta(\mathbf{b}(c_i, r_i), k_l) + 4W \cdot \Delta(\mathbf{b}(c_i, r_i), k_l)
\end{aligned}$$

Next we show that W is at most $\epsilon \cdot |C_i^{\mathrm{in}}|$ with high probability. Define the random variable $Y_j = w_{s_j}$ to be the weight of the jth sample point in S_i^{out}. Hence $W = \sum_{j=1}^{s_i^{\mathrm{out}}} Y_j$. The expected value $\mathbf{E}[Y_j]$ of Y_j is, by definition, $\mathbf{E}[Y_j] = \sum_{q \in C_i^{\mathrm{out}}} p_q w_q = \frac{|C_i^{\mathrm{out}}|}{s_i^{\mathrm{out}}} \le \frac{\epsilon |C_i|}{s_i^{\mathrm{out}}}$. We also have $|C_i^{\mathrm{in}}| \ge (1 - \epsilon)|C_i|$. Hence, $\mathbf{E}[Y_j] \le \frac{2\epsilon \cdot |C_i^{\mathrm{in}}|}{s_i^{\mathrm{out}}}$ for $\epsilon \le 1/2$. Thus, $\mathbf{E}[W] \le 2\epsilon \cdot |C_i^{\mathrm{in}}|$. An upper bound for the weight of sample point $q \in C_i^{\mathrm{out}}$ is given by $w_q = \frac{\mathbf{cost}(C_i^{\mathrm{out}}, c_i)}{s_i^{\mathrm{out}} \Delta(q, c_i)} \le \frac{\mathbf{cost}(C_i^{\mathrm{out}}, c_i)}{s_i^{\mathrm{out}} \frac{\mathbf{cost}(C_i, c_i)}{\epsilon |C_i|}} \le \frac{\epsilon |C_i|}{s_i^{\mathrm{out}}}$. Define $Z_j = \frac{Y_j}{\frac{\epsilon |C_i|}{s_i^{\mathrm{out}}}} \le 1$. Let $Z = \sum_{j=1}^{s_i^{\mathrm{out}}} Z_j$ then $\mathbf{E}[Z] \le s_i^{\mathrm{out}}$. By Hoeffding bound we obtain:

$$\begin{aligned}
&\mathbf{Pr}\Big[\Big|\sum_{j=1}^{s_i^{\mathrm{out}}} Y_j - \mathbf{E}[Y]\Big| > \epsilon |C_i|\Big] = \mathbf{Pr}\big[|Z - \mathbf{E}[Z]| > s_i^{\mathrm{out}}\big] \\
\le\ & \mathbf{Pr}\Big[|Z - \mathbf{E}[Z]| > \frac{s_i^{\mathrm{out}}}{\mathbf{E}[Z]} \mathbf{E}[Z]\Big] \\
\le\ & 2 \exp\left(- \frac{\mathbf{E}[Z] \cdot \min\left\{ \left(\frac{s_i^{\mathrm{out}}}{\mathbf{E}[Z]}\right), \left(\frac{s_i^{\mathrm{out}}}{\mathbf{E}[Z]}\right)^2 \right\}}{3} \right) \\
=\ & 2 \exp\big(- s_i^{\mathrm{out}}/3\big)\ .
\end{aligned}$$

We choose $s_i^{\mathrm{out}} \ge 3 \ln(2k/\lambda)$. This implies $\mathbf{Pr}[|Y - \mathbf{E}[Y]| > \epsilon |C_i|] \le \lambda/k$. Hence, we get that $\mathbf{Pr}[W > 4\epsilon \cdot |C_i^{\mathrm{in}}|] \le \lambda/k$. It follows that

$$\begin{aligned}
\sum_{q \in C_i^{\mathrm{out}}} \mathbf{cost}(q, k_l) \quad &+ \sum_{p \in S_i^{\mathrm{out}}} w_p \cdot \mathbf{cost}(p, k_l) && (1) \\
&\le 26\epsilon \cdot |C_i^{\mathrm{in}}| \cdot \Delta(\mathbf{b}(c_i, r_i), k_l) && (2)
\end{aligned}$$

with probability at least $1 - \lambda/k$. Since $|\mathbf{cost}(C_i^{\mathrm{out}}, \mathcal{K}) - \mathbf{cost}(S_i^{\mathrm{out}}, \mathcal{K})| \le \mathbf{cost}(C_i^{\mathrm{out}}, \mathcal{K}) + \mathbf{cost}(S_i^{\mathrm{out}}, \mathcal{K})$ this is an upper bound for the error of the outer points. Overall error for the sample set $S_i^{\mathrm{in}} \cup S_i^{\mathrm{out}}$ in case (a) would be

$$\begin{aligned}
&|\mathbf{cost}(S_i^{\mathrm{in}} \cup S_i^{\mathrm{out}}, \mathcal{K}) - \mathbf{cost}(C_i^{\mathrm{in}} \cup C_i^{\mathrm{out}}, \mathcal{K})| \\
\le\ & 8\epsilon \mathbf{cost}(C_i^{\mathrm{in}}, \mathcal{K}) + 26\epsilon \mathbf{cost}(C_i^{\mathrm{in}}, \mathcal{K}) \le 34\epsilon \mathbf{cost}(C_i^{\mathrm{in}}, \mathcal{K}).
\end{aligned}$$

Case (b).

LEMMA 6 (HAUSSLER [8]). *Let $h(\cdot)$ be a function defined on a set $\mathcal{P}$, such that for all $p \in \mathcal{P}$, we have $0 \le h(p) \le M$, where M is a fixed constant. Let $\mathcal{S} = \{p_1, \ldots, p_s\}$ be a multiset of s samples drawn independently and identically from $\mathcal{P}$, and let $\delta > 0$ be a parameter. If $s \ge (M^2/2\delta^2) \cdot \ln(2/\lambda)$, then* $\mathbf{Pr}\left[\left|\frac{h(\mathcal{P})}{|\mathcal{P}|} - \frac{h(\mathcal{S})}{|\mathcal{S}|}\right| \ge \delta\right] \le \lambda$, *where* $h(\mathcal{S}) = \sum_{s \in \mathcal{S}} h(s)$.

We want to apply Lemma 6 to analyze the error of the uniform sampling from C_i^{in}. Therefore, let $h(p) = \Delta(p, \mathcal{K})$ for $p \in C_i^{\mathrm{in}}$. Since C_i^{in} is contained in a ball of radius r_i we get that $\max_{p \in P} h(p) \le (\mathrm{dist}(C_i^{\mathrm{in}}, \mathcal{K}) + 2r_i)^2 \le 2(\Delta(C_i^{\mathrm{in}}, \mathcal{K}) + 4r_i^2)$. Hence, we can use $M = 2(\Delta(C_i^{\mathrm{in}}, \mathcal{K}) + 4r_i^2)$. We define $\mathbf{cost}_{avg}(C_i^{\mathrm{in}}, \mathcal{K}) = \frac{h(C_i^{\mathrm{in}})}{|C_i^{\mathrm{in}}|}$, and $\mathbf{cost}_{avg}(S_i^{\mathrm{in}}, \mathcal{K}) = \frac{h(S_i^{\mathrm{in}})}{|S_i^{\mathrm{in}}|}$. Then we set $\delta = \xi M$. Thus, if $s_i^{\mathrm{in}} \ge \frac{1}{2\xi^2} \cdot \ln(4k/\lambda)$, then

$$\mathbf{Pr}[|\mathbf{cost}_{avg}(C_i^{\mathrm{in}}, \mathcal{K}) - \mathbf{cost}_{avg}(S_i^{\mathrm{in}}, \mathcal{K})| \ge \xi 2(\Delta(C_i^{\mathrm{in}}, \mathcal{K}) + 4r_i^2)]$$

$\le \lambda/2k$.

As $w(C_i^{\mathrm{in}}) = w(S_i^{\mathrm{in}})$ we have with probability at least $1 - \lambda/2k$

$$|\mathbf{cost}(C_i^{\mathrm{in}}, \mathcal{K}) - \mathbf{cost}(S_i^{\mathrm{in}}, \mathcal{K})| \le 2\xi |C_i^{\mathrm{in}}|(\Delta(C_i^{\mathrm{in}}, \mathcal{K}) + 4r_i^2)$$

It is easy to see that $|C_i^{\mathrm{in}}|\Delta(C_i^{\mathrm{in}}, \mathcal{K}) \le \mathbf{cost}(C_i^{\mathrm{in}}, \mathcal{K})$ and summing this up for all sets C_i^{in}, for $i = 1, \cdots, k$, we have $|\mathbf{cost}(\cup_i C_i^{\mathrm{in}}, \mathcal{K}) - \mathbf{cost}(\cup_i S_i^{\mathrm{in}}, \mathcal{K})| \le 2\xi \left(\sum_i \mathbf{cost}(C_i^{\mathrm{in}}, \mathcal{K}) + \sum_i |C_i^{\mathrm{in}}| 4r_i^2\right)$. As $r_i = \sqrt{\frac{\mathbf{cost}(C_i, c_i)}{\epsilon \cdot |C_i|}}$ and $C_i^{\mathrm{in}} \le C_i$, we have

$$\begin{aligned}
&|\mathbf{cost}(\cup_i C_i^{\mathrm{in}}, \mathcal{K}) - \mathbf{cost}(\cup_i S_i^{\mathrm{in}}, \mathcal{K})| \\
\le\ & 2\xi \left(\sum_i \mathbf{cost}(C_i^{\mathrm{in}}, \mathcal{K}) + \sum_i 4 \frac{\mathbf{cost}(C_i, c_i)}{\epsilon} \right) \\
\le\ & 2\xi \left(\sum_i \mathbf{cost}(C_i^{\mathrm{in}}, \mathcal{K}) + 4 \frac{\mathbf{OPT}(\mathcal{P}, k)}{\beta \epsilon} \right).
\end{aligned}$$

Recall that β is the approximation factor of the solution $\{c_1, \ldots, c_k\}$. We set $\xi = \beta \epsilon^2/10$. Then we get $|\mathbf{cost}(\cup_i C_i^{\mathrm{in}}, \mathcal{K}) - \mathbf{cost}(\cup_i S_i^{\mathrm{in}}, \mathcal{K})| \le \epsilon \left(\sum_i \mathbf{cost}(C_i^{\mathrm{in}}, \mathcal{K})\right) \le \epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K})$ which holds with probability at least $1 - \lambda/2$ and for $s_i^{\mathrm{in}} \ge \frac{50}{\beta^2 \epsilon^4} \ln(4k/\lambda)$.

Let $\mathbf{cost}_{avg}(C_i^{\mathrm{out}}, \mathcal{K}) = \frac{\mathbf{cost}(C_i^{\mathrm{out}}, \mathcal{K})}{|C_i^{\mathrm{out}}|}$. Define the random variable $X_j = \frac{w_{s_j}}{|C_i^{\mathrm{out}}|} \cdot \Delta(s_j, \mathcal{K})$ for the average contribution of the jth sample point in S_i^{out} to the nearest center of $\mathcal{K}$. The expected value of $\mathbf{E}[X_j]$ is

$$\begin{aligned}
\mathbf{E}[X_j] &= \sum_{q \in C_i^{\mathrm{out}}} p_q \cdot \frac{w_q}{|C_i^{\mathrm{out}}|} \cdot \Delta(q, \mathcal{K}) \\
&= \frac{1}{s_i^{\mathrm{out}} \cdot |C_i^{\mathrm{out}}|} \sum_{q \in C_i^{\mathrm{out}}} \Delta(q, \mathcal{K}) = \frac{\mathbf{cost}_{avg}(C_i^{\mathrm{out}}, \mathcal{K})}{s_i^{\mathrm{out}}}.
\end{aligned}$$

We define $\mathbf{cost}_{avg}(S_i^{\mathrm{out}}, \mathcal{K}) = \frac{1}{|C_i^{\mathrm{out}}|} \cdot \sum_{p \in S_i^{\mathrm{out}}} w_p \cdot \Delta(p, \mathcal{K}) = \sum_{j=1}^{s_i^{\mathrm{out}}} X_j$ (notice that the averaging is done by dividing by $|C_i^{\mathrm{out}}|$ and not by the sum of the weights of the points in S_i^{out}). Hence, $\mathbf{E}[\mathbf{cost}_{avg}(S_i^{\mathrm{out}}, \mathcal{K})] = \sum_{j=1}^{s_i^{\mathrm{out}}} \mathbf{E}[X_j] = \mathbf{cost}_{avg}(C_i^{\mathrm{out}}, \mathcal{K})$. For each

$q \in C_i^{out}$ we have

$$\begin{aligned}\Delta(q, k_l) &\leq 2\left(\Delta(q, c_i) + \Delta(c_i, k_l)\right) \\ &\leq 2\left(\Delta(q, c_i) + (\frac{r_i(1+\epsilon)}{\epsilon})^2\right) \\ &\leq 4\Delta(q, c_i)\left[\frac{(1+\epsilon)^2}{\epsilon^2}\right].\end{aligned}$$

Observe that each X_j satisfies

$$\begin{aligned}X_j &= \frac{w_{s_j}}{|C_i^{out}|} \cdot \Delta(s_j, \mathcal{K}) \\ &\leq \frac{w_{s_j}}{|C_i^{out}|} \cdot \Delta(s_j, k_l) \leq \frac{w_{s_j}}{|C_i^{out}|} \cdot 4\Delta(s_j, c_i)\left[\frac{(1+\epsilon)^2}{\epsilon^2}\right] \\ &\leq \frac{\mathbf{cost}(C_i^{out}, c_i)}{\Delta(s_j, c_i) \cdot s_i^{out} \cdot |C_i^{out}|} \cdot 4\Delta(s_j, c_i)\left[\frac{(1+\epsilon)^2}{\epsilon^2}\right] \\ &\leq 4\left[\frac{(1+\epsilon)^2}{\epsilon^2}\right] \cdot \frac{\mathbf{cost}_{avg}(C_i^{out}, c_i)}{s_i^{out}}.\end{aligned}$$

Define random variable $Z_j = \frac{X_j}{4\left[\frac{(1+\epsilon)^2}{\epsilon^2}\right] \cdot \frac{\mathbf{cost}_{avg}(C_i^{out}, c_i)}{s_i^{out}}} \leq 1$ and let $Z = \sum_{j=1}^{s_i^{out}} Z_j$. Applying Hoeffding bound we have

$$\begin{aligned}&\mathbf{Pr}[|\mathbf{cost}_{avg}(S_i^{out}, \mathcal{K}) - \mathbf{cost}_{avg}(C_i^{out}, \mathcal{K})| \geq \epsilon \cdot \\ &\qquad\qquad \mathbf{cost}_{avg}(C_i^{out}, c_i)] \\ &= \mathbf{Pr}[|\sum_{j=1}^{s_i^{out}} X_j - \sum_{j=1}^{s_i^{out}} \mathbf{E}[X_j]| \geq \epsilon \cdot \mathbf{cost}_{avg}(C_i^{out}, c_i)] \\ &= \mathbf{Pr}[|Z - \mathbf{E}[Z]| \geq \frac{\epsilon^3 s_i^{out}}{4(1+\epsilon^2)\mathbf{E}[Z]}\mathbf{E}[Z]] \\ &\leq 2\exp\left(-\frac{\mathbf{E}[Z] \cdot \min(\frac{\epsilon^3 s_i^{out}}{4((1+\epsilon)^2)\mathbf{E}[Z]}, \left(\frac{\epsilon^3 s_i^{out}}{4((1+\epsilon)^2)\mathbf{E}[Z]}\right)^2)}{3}\right)\end{aligned}$$

Choosing $s_i^{out} \geq \frac{12((1+\epsilon)^2)}{\epsilon^3}\ln(4k/\lambda)$ gives $\mathbf{Pr}[|\mathbf{cost}_{avg}(S_i^{out}, \mathcal{K}) - \mathbf{cost}_{avg}(C_i^{out}, \mathcal{K})| \geq \epsilon \cdot \mathbf{cost}_{avg}(C_i^{out}, c_i)] \leq \lambda/(2k)$. Multiplying by $|C_i^{out}|$ gives

$$|\mathbf{cost}(S_i^{out}, \mathcal{K}) - \mathbf{cost}(C_i^{out}, \mathcal{K})| \leq \epsilon \cdot \mathbf{cost}(C_i^{out}, c_i)$$

with probability at least $1 - \lambda/(2k)$.

Now we can combine cases (a) and (b). Summing up over all sets C_i^{in} and C_i^{out}, for $i = 1, \cdots, k$, there exists a constant c' such that for $s_i^{in}, s_i^{out} \geq c' \cdot \frac{\ln(k/\lambda)}{\epsilon^4}$:

$$\begin{aligned}|\mathbf{cost}(\mathcal{S}, \mathcal{K}) - \mathbf{cost}(\mathcal{P}, \mathcal{K})| &\leq 34\epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{C}) &(3)\\ &\leq 34\epsilon\beta \cdot \mathbf{cost}(\mathcal{P}, \mathcal{O}) &(4)\\ &\leq 34\epsilon\beta \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K}) &(5)\end{aligned}$$

with probability at least $1 - \lambda$ and where $\mathcal{S} = \bigcup_{i=1}^k \left\{S_i^{in} \cup S_i^{out}\right\}$ and $\mathcal{O}$ is an optimal solution for $\mathcal{P}$. Replacing ϵ by $\epsilon/(34\beta)$ gives Lemma 5. □

Centers from $\mathcal{T}$ *are approximated within a factor* $(1 \pm \epsilon)$.

Now we want to prove the following lemma.

LEMMA 7. *Let $\mathcal{P}$ be a set of points in $\mathbb{R}^d$ and let $0 < \epsilon, \delta < 1/2$ and $k \geq 1$ be parameters. Let $\mathcal{S}$ be a weighted set of points sampled from $\mathcal{P}$ according to our coreset construction using $s_i^{in}, s_i^{out} \geq c \cdot \frac{k\ln(k/\delta)}{\epsilon^5} \cdot \ln(k/\epsilon \cdot \ln(1/\delta))$ for some large enough constant c. Let $\mathcal{T}$ be the set of centroids of subsets from $\mathcal{S}$ (with repetition) of size $2/\epsilon$. Then with probability $1 - \delta$ we get*

$$\forall \mathcal{K} \subseteq \mathcal{T}, |\mathcal{K}| = k: \ |\mathbf{cost}(\mathcal{S}, \mathcal{K}) - \mathbf{cost}(\mathcal{P}, \mathcal{K})| \leq \epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K}) \ .$$

Proof. Let $\mathcal{N}$ denote the set of centroids of all subsets from $\mathcal{P}$ of size $2/\epsilon$. We say that $\mathcal{K} \subseteq \mathcal{N}$ *is well approximated*, if $\left|\mathbf{cost}(\mathcal{S}, \mathcal{K}) - \mathbf{cost}(\mathcal{P}, \mathcal{K})\right| \leq \epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K})$. We want to show that every set $\mathcal{K} \subseteq \mathcal{T}$, $|\mathcal{K}| = k$, is also well approximated. Recall that $\mathcal{T} \subseteq \mathcal{N}$ consists of the centroids of all subsets (with repetition) of size $2/\epsilon$ of $\mathcal{S}$. Wlog. we will assume that for each point $p \in \mathcal{T}$ there is a unique multiset $\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathbf{p})$ of $2/\epsilon$ points from $\mathcal{S}$ that generates p, i.e. $\mu_{\mathcal{P}}(\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathbf{p})) = p$. We cannot directly apply Lemma 5 to show that $\mathcal{K} \subseteq \mathcal{T}$ is well approximated, because $\mathcal{K} \subseteq \mathcal{T}$ imposes the condition that $\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}$, where $\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) = \bigcup_{p \in \mathcal{K}} \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathbf{p})$. Here and in the following we regard both $\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})$ and $\mathcal{S}$ as (unweighted) multisets, i.e. we replace each point p with weight w_p by w_p copies of p. We assume that all relations between multisets take the multiplicity of points into accout. For example, the expression $\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}$ implies that if $\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})$ contains a point multiple times, it appears at least the same number of times in $\mathcal{S}$. Given $\delta > 0$ we want to show that

$$\begin{aligned}&\mathbf{Pr}[\forall \mathcal{K} \subseteq \mathcal{T}, |\mathcal{K}| = k : \mathcal{K} \text{ is well approximated}\,] \\ &= 1 - \mathbf{Pr}[\exists \mathcal{K} \subseteq \mathcal{T}, |\mathcal{K}| = k : \mathcal{K} \text{ is not well approximated}\,] \\ &\geq 1 - \delta \ .\end{aligned}$$

We use the fact that

$$\mathbf{Pr}[\exists \mathcal{K} \subseteq \mathcal{T}, |\mathcal{K}| = k : \mathcal{K} \text{ is not well approximated}\,] \quad (6)$$

$$\begin{aligned}&\leq \sum_{\mathcal{K} \subseteq \mathcal{N}, |\mathcal{K}| = k} \mathbf{Pr}[\mathcal{K} \text{ is not well approximated} \mid \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \\ &\cdot \quad \mathbf{Pr}[\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \ .\end{aligned}$$

We have

$$\begin{aligned}&\mathbf{Pr}[\mathcal{K} \text{ is not } \textbf{well} \text{ approximated} \mid \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \\ &\leq \mathbf{Pr}[|\mathbf{cost}(\cup_{i=1}^k S_i^{out}, \mathcal{K}) - \mathbf{cost}(\cup_{i=1}^k C_i^{out}, \mathcal{K})| \\ &\quad > \epsilon \cdot \mathbf{cost}(\cup_{i=1}^k C_i^{out}, \mathcal{K}) \mid \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \\ &\quad + \mathbf{Pr}[|\mathbf{cost}(\cup_{i=1}^k S_i^{in}, \mathcal{K}) - \mathbf{cost}(\cup_{i=1}^k C_i^{in}, \mathcal{K})| \\ &\quad > \epsilon \cdot \mathbf{cost}(\cup_{i=1}^k C_i^{in}, \mathcal{K}) \mid \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \ .\end{aligned}$$

The condition fixes $2k/\epsilon$ points of the sample set. All remaining points are drawn at random according to the specified distribution. Let us denote by F_i^{in} and F_{out}^{in} these random points, i.e. $S_i^{in} = F_i^{in} \cup \left(C_i^{in} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})\right)$ and $S_i^{out} = F_i^{out} \cup \left(C_i^{out} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})\right)$. We get

$$\begin{aligned}&\mathbf{Pr}[|\mathbf{cost}(S_i^{in}, \mathcal{K}) - \mathbf{cost}(C_i^{in}, \mathcal{K})| > \epsilon \cdot \mathbf{cost}(C_i^{in}, \mathcal{K}) \mid \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) &(7)\\ &\qquad \subseteq \mathcal{S}] &(8)\end{aligned}$$

$$\begin{aligned}&= \mathbf{Pr}[|\mathbf{cost}(F_i^{in}, \mathcal{K}) + \mathbf{cost}(C_i^{in} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), \mathcal{K}) - \mathbf{cost}(C_i^{in}, \mathcal{K})| > \epsilon \cdot \\ &\qquad \mathbf{cost}(C_i^{in}, \mathcal{K})] \\ &\leq \mathbf{Pr}[|\mathbf{cost}(F_i^{in}, \mathcal{K}) - \mathbf{cost}(C_i^{in}, \mathcal{K})| > \epsilon \cdot \mathbf{cost}(C_i^{in}, \mathcal{K}) - \mathbf{cost}(C_i^{in} \cap \\ &\qquad \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), \mathcal{K})]\end{aligned}$$

In a similar way we obtain

$$\begin{aligned}&\mathbf{Pr}[|\mathbf{cost}(S_i^{out}, \mathcal{K}) - \mathbf{cost}(C_i^{out}, \mathcal{K})| > \epsilon \cdot \mathbf{cost}(C_i^{out}, \mathcal{K}) \mid \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) &(9)\\ &\qquad \subseteq \mathcal{S}] &(10)\end{aligned}$$

$$\leq \mathbf{Pr}[|\mathbf{cost}(F_i^{\text{out}}, \mathcal{K}) - \mathbf{cost}(C_i^{\text{out}}, \mathcal{K})| > \epsilon \cdot \mathbf{cost}(C_i^{\text{out}}, \mathcal{K}) \\ -\mathbf{cost}(C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), \mathcal{K})]$$

After rescaling the weights of points in F_i^{in} by $|S_i^{\text{in}}|/|F_i^{\text{in}}|$ we can apply the proof of Lemma 5. Let $\mathbf{Err}_i^{\text{in}}, \mathbf{Err}_i^{\text{out}}$ denote the error bounds derived in the proof of Lemma 5. We distinguish between cases (a) and (b). In case (a) we obtain by Lemma 5 for $\mathbf{Err}_i^{\text{in}} = 8 \cdot \epsilon \cdot \mathbf{cost}(C_i^{\text{in}}, \mathcal{K},)$ and $|F_i^{\text{in}}| \geq c \cdot \frac{\ln(k/\lambda)}{\epsilon^4}$

$$\begin{aligned}
& \lambda/(2k) \\
\geq\ & \mathbf{Pr}\Big[\Big|\frac{|S_i^{\text{in}}|}{|F_i^{\text{in}}|} \cdot \mathbf{cost}(F_i^{\text{in}}, \mathcal{K}) - \mathbf{cost}(C_i^{\text{in}}, \mathcal{K})\Big| > \mathbf{Err}_i^{\text{in}}\Big] \\
=\ & \mathbf{Pr}\Big[\Big|\mathbf{cost}(F_i^{\text{in}}, \mathcal{K}) - \frac{|F_i^{\text{in}}|}{|S_i^{\text{in}}|}\mathbf{cost}(C_i^{\text{in}}, \mathcal{K})\Big| > \frac{|F_i^{\text{in}}|}{|S_i^{\text{in}}|} \cdot \mathbf{Err}_i^{\text{in}}\Big] \\
\geq\ & \mathbf{Pr}\Big[\Big|\mathbf{cost}(F_i^{\text{in}}, \mathcal{K}) - \mathbf{cost}(C_i^{\text{in}}, \mathcal{K})\Big| > \frac{|F_i^{\text{in}}|}{|S_i^{\text{in}}|} \cdot \mathbf{Err}_i^{\text{in}} \\
+\ & (1 - \frac{|F_i^{\text{in}}|}{|S_i^{\text{in}}|}) \cdot \mathbf{cost}(C_i^{\text{in}}, \mathcal{K})\Big] \\
\geq\ & \mathbf{Pr}\Big[\Big|\mathbf{cost}(F_i^{\text{in}}, \mathcal{K}) - \mathbf{cost}(C_i^{\text{in}}, \mathcal{K})\Big| > \mathbf{Err}_i^{\text{in}} \\
+\ & \epsilon \cdot \mathbf{cost}(C_i^{\text{in}}, \mathcal{K})\Big]
\end{aligned}$$

Similarly, we obtain for $\mathbf{Err}_i^{\text{out}} = 26\epsilon \cdot |C_i^{\text{in}}| \cdot \Delta(\mathbf{b}(c_i, r_i), k_l)$

$$\begin{aligned}
& \lambda/(2k) \\
\geq\ & \mathbf{Pr}\Big[\Big|\mathbf{cost}(F_i^{\text{out}}, \mathcal{K}) - \mathbf{cost}(C_i^{\text{out}}, \mathcal{K})\Big| > \mathbf{Err}_i^{\text{out}} \\
+\ & \epsilon \cdot \mathbf{cost}(C_i^{\text{out}}, \mathcal{K})\Big]
\end{aligned}$$

In a similar way we can obtain bounds for case (b). Summing up over all clusters gives for $\mathcal{F} = \bigcup_{i=1}^{k}(F_i^{\text{in}} \cup F_i^{\text{out}})$:

$$\mathbf{Pr}\Big[\Big|\mathbf{cost}(\mathcal{F}, \mathcal{K}) - \mathbf{cost}(\mathcal{P}, \mathcal{K})\Big| \leq 2\epsilon \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K})\Big] \leq \lambda \ . \quad (11)$$

We will now prove $\mathbf{cost}(\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), \mathcal{K})\mathcal{K} \leq \epsilon/2 \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K})$. Then replacing ϵ by $\epsilon/4$ in equation (11) and combining it with equations (8) and (10) gives

$$\mathbf{Pr}[\mathcal{K} \text{ is not well approximated} \mid \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \leq \lambda \ .$$

LEMMA 8. *For $s_i^{\text{in}}, s_i^{\text{out}} \geq \frac{ck}{\epsilon^5}$, where $c \geq 8$, we have*

$$\mathbf{cost}(\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), \mathcal{K}) \leq \epsilon/2 \cdot \mathbf{cost}(\mathcal{P}, \mathcal{K}) \ .$$

Proof. The analysis will again distinguish between the cases (a) $\text{dist}(k_l, c_i) \geq r_i + \frac{r_i}{\epsilon} = \frac{r_i(1+\epsilon)}{\epsilon}$ and (b) $\text{dist}(k_l, c_i) < r_i + \frac{r_i}{\epsilon} = \frac{r_i(1+\epsilon)}{\epsilon}$. We will assume $\epsilon \leq 1/2$.

Case (a).

First for C_i^{in}

$$\begin{aligned}
& \mathbf{cost}(C_i^{\text{in}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), \mathcal{K}) \\
\leq\ & \frac{2k}{\epsilon}\frac{|C_i^{\text{in}}|}{s_i^{\text{in}}}\Big[2\Big((2r_i)^2 + \Delta(\mathbf{b}(c_i, r_i), k_l)\Big)\Big] \\
\leq\ & \frac{2k}{\epsilon}\frac{|C_i^{\text{in}}|}{\frac{ck}{\epsilon^5}}\Big[2\Big(4\epsilon^2\Delta(\mathbf{b}(c_i, r_i), k_l) + \Delta(\mathbf{b}(c_i, r_i), k_l)\Big)\Big] \\
\leq\ & \epsilon^4|C_i^{\text{in}}| \cdot \Delta(\mathbf{b}(c_i, r_i), k_l) \leq \epsilon^4\mathbf{cost}(C_i^{\text{in}}, \mathcal{K}).
\end{aligned}$$

Then for C_i^{out}

$$\mathbf{cost}(C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), \mathcal{K}) \leq \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} w_p \cdot \mathbf{cost}(p, k_l)$$

$$\begin{aligned}
& \mathbf{cost}(C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), k_l) \\
\leq\ & \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} 2w_p \cdot (\Delta(p, c_i) + \Delta(c_i, k_l)) \\
\leq\ & \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} 2w_p \cdot (\Delta(p, c_i) \\
+\ & 2(r_i^2 + \Delta(\mathbf{b}(c_i, r_i), k_l))) \\
\leq\ & \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} 2w_p \cdot (\Delta(p, c_i) \\
+\ & 2(\Delta(p, c_i) + \Delta(\mathbf{b}(c_i, r_i), k_l))) \\
\leq\ & \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} 2w_p \cdot (3\Delta(p, c_i) \\
+\ & 2\Delta(\mathbf{b}(c_i, r_i), k_l)) \\
\leq\ & 6r_i^2\epsilon|C_i^{\text{in}}| + \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} 4w_p \cdot \Delta(\mathbf{b}(c_i, r_i), k_l) \\
\leq\ & 6\epsilon^3|C_i^{\text{in}}|\Delta(\mathbf{b}(c_i, r_i), k_l) \\
+\ & \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} 4w_p \cdot \Delta(\mathbf{b}(c_i, r_i), k_l) \\
\leq\ & 2\epsilon|C_i^{\text{in}}|\Delta(\mathbf{b}(c_i, r_i), k_l)(3\epsilon^2 + 2) \\
\leq\ & 6\epsilon|C_i^{\text{in}}|\Delta(\mathbf{b}(c_i, r_i), k_l) \leq 6\epsilon\mathbf{cost}(C_i^{\text{in}}, k_l).
\end{aligned}$$

Case (b).

Again first for C_i^{in} and having $r_i = \sqrt{\frac{\mathbf{cost}(C_i, c_i)}{\epsilon.|C_i|}}$

$$\begin{aligned}
\mathbf{cost}(C_i^{\text{in}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), \mathcal{K}) &\leq \frac{2k}{\epsilon}\frac{|C_i^{\text{in}}|}{s_i^{\text{in}}}\left(\frac{r_i(1+\epsilon)}{\epsilon}\right)^2 \\
&\leq \frac{2k}{\epsilon}\frac{|C_i^{\text{in}}|}{\frac{ck}{\epsilon^5}}\left(\frac{r_i(1+\epsilon)}{\epsilon}\right)^2 \\
&\leq \epsilon^2|C_i^{\text{in}}|r_i^2 \\
&\leq \epsilon/6\mathbf{cost}(C_i^{\text{in}}, c_i).
\end{aligned}$$

Then for C_i^{out}

$$\begin{aligned}
& \mathbf{cost}(C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}), \mathcal{K}) \\
\leq\ & \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} w_p \cdot \mathbf{cost}(p, k_l) \\
\leq\ & \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} \frac{\mathbf{cost}(C_i^{\text{out}}, c_i)}{\frac{ck}{\epsilon^5} \cdot \Delta(p, c_i)} \cdot [2(\Delta(p, c_i) + \Delta(c_i, k_l))] \\
\leq\ & \epsilon^4 \cdot \mathbf{cost}(C_i^{\text{out}}, c_i) \\
+\ & \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} \frac{\mathbf{cost}(C_i^{\text{out}}, c_i)}{\frac{ck}{\epsilon^5} \cdot \Delta(p, c_i)} \cdot 2\Delta(c_i, k_l) \\
\leq\ & \epsilon^4 \cdot \mathbf{cost}(C_i^{\text{out}}, c_i) \\
+\ & \sum_{p \in C_i^{\text{out}} \cap \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K})} \frac{\mathbf{cost}(C_i^{\text{out}}, c_i)}{\frac{ck}{\epsilon^5} \cdot r_i^2} \cdot 2(\frac{r_i(1+\epsilon)}{\epsilon})^2 \\
\leq\ & \epsilon^4 \cdot \mathbf{cost}(C_i^{\text{out}}, c_i) \\
+\ & \epsilon^2 \cdot \mathbf{cost}(C_i^{\text{out}}, c_i) \leq \epsilon^2 \cdot \mathbf{cost}(C_i^{\text{out}}, c_i).
\end{aligned}$$

□

Finally, replacing ϵ by $\epsilon/4$ in equation (11) and combining it with equations (8)and (10) gives

$$\mathbf{Pr}[\mathcal{K} \text{ is not well approximated} \mid \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \leq \lambda \ .$$

Plugging this into equation (6) we get

$$\begin{aligned} & \mathbf{Pr}[\exists \mathcal{K} \subseteq \mathcal{T}, |\mathcal{K}| = k : \mathcal{K} \text{ is not well approximated }] \\ \leq & \sum_{\mathcal{K} \subseteq \mathcal{N}, |\mathcal{K}|=k} \mathbf{Pr}[\mathcal{K} \text{ is not well approximated} \mid \mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \\ \cdot & \mathbf{Pr}[\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \\ \leq & \lambda \cdot \sum_{\mathcal{K} \subseteq \mathcal{N}, |\mathcal{K}|=k} \mathbf{Pr}[\mu_{\mathcal{P}}^{-\mathbf{1}}(\mathcal{K}) \subseteq \mathcal{S}] \leq \lambda \cdot |\mathcal{S}|^{2k/\epsilon} \end{aligned}$$

It follows that for $\lambda \leq \delta/|\mathcal{S}|^{2k/\epsilon}$ we obtain the bound stated in the lemma. This is satisfied for $s_i^{\text{in}}, s_i^{\text{out}} \geq c \cdot \frac{k \ln(k/\delta)}{\epsilon^5} \cdot \ln(k/\epsilon \cdot \ln(1/\delta))$ when c is a large enough constant. □

The coreset.

Finally, we put things together.

THEOREM 9. *Given a set $\mathcal{P}$ of n points in $\mathbb{R}^d$ and parameters $\epsilon, \lambda > 0$ and an appropriate constant $c > 0$, if $\mathcal{S}$ is a weighted set of points obtained by our algorithm using $s_i^{\text{in}}, s_i^{\text{out}} \geq c \cdot \frac{k \ln(k/\delta)}{\epsilon^5} \cdot \ln(k/\epsilon \cdot \ln(1/\delta))$ and $\mathcal{T}$ is the set of centroids of subsets of size $2/\epsilon$, then $(\mathcal{S}, \mathcal{T})$ is a weak (k, ϵ)-coreset for point set $\mathcal{P}$ with probability at least $1 - \delta$.*

Proof. We apply Lemma 5 to show that the cost of an optimal set of centers is preserved upto a factor of $(1 \pm \epsilon)$. Then we apply Lemma 7 to show that this is true for all sets of centers from $\mathcal{T}$. From Lemma 4 is follows that $\mathcal{T}$ is a $(k, 6\epsilon)$-approximate centroid set with probability $1 - \delta + \lambda$. Replacing ϵ by $\epsilon/6$, δ by $\delta/2$ and λ by $\delta/2$ we obtain the theorem. □

5. APPLICATIONS

5.1 A k-Means PTAS

We obtain the following PTAS for k-Means clustering. We first compute a weak (k, ϵ)-coreset $\mathcal{S}$. Then we do exhaustive search over all subsets of size k from $\mathcal{T}$. We can slightly improve the running time of this approach using dimensionality reduction. The idea is to use Johnson-Lindenstrauss transform to map $\mathcal{S}$ to a lower dimensional space.

LEMMA 10 (JOHNSON-LINDENSTRAUSS LEMMA[16]). *Any set of n points in a Euclidean space can be mapped to $\mathbb{R}^t$ where $t = O(\frac{\log n}{\epsilon^2})$ with distortion $\leq 1 + \epsilon$ in the distances. Such a mapping can be found in $O(nd \log n/\epsilon^2)$.*

We choose the dimension of the target space in such a way that distances between the points in $\mathcal{S} \cup \mathcal{O} \cup \mathcal{T}$ are distorted by at most a factor of $(1 + \epsilon)$. Thus, $t = O(\log |\mathcal{T}|/\epsilon^2)$. Since JL-transform is a linear mapping, we know that the centroid of points is mapped to the centroid of the mapped points. Since we may assume that the centroids of subsets of $\mathcal{T}$ of size k are disjoint in the original space they also will be disjoint in the target space (since their mutual distances are preserved upto a factor of $(1 + \epsilon)$). Thus, a centroid of $2/\epsilon$ points in the target space corresponds to a unique point (the centroid of the points in the original space) and so we can map a solution from the target space back to the original space. Finally, to obtain a solution we do exhaustive search in the set of all subsets of $\mathcal{T}$ of size k and evaluate the cost of each solution in the small target space of the JL-transform.

THEOREM 11. *Given a set $\mathcal{P}$ of n points in $\mathbb{R}^d$ and parameters $\epsilon, \lambda > 0$ and an appropriate constant $c > 0$, there exists a randomized algorithm that computes $(1+\epsilon)$-approximate k-means clustering of $\mathcal{P}$ in time $O(nkd + d \cdot (k/\epsilon)^{O(1)} + 2^{\tilde{O}(k/\epsilon)})$ with probability at least $1 - \lambda$.*

5.2 Streaming

In this section, we adapt the algorithm of Har-Peled and Mazumdar [10] to our randomized coreset. Their algorithm was based on standard dynamization technique of Bentley and Saxe [1] and the following observation about coresets.

OBSERVATION 12. *[10] (i) If C_1 and C_2 are the (k, ϵ)-coresets for disjoint sets P_1 and P_2 respectively, then $C_1 \cup C_2$ is a (k, ϵ)-coreset for $P_1 \cup P_2$.*

(ii) If C_1 is (k, ϵ)-coreset for C_2, and C_2 is a (k, δ)-coreset for C_3, then C_1 is a $(k, (1 + \epsilon)(1 + \delta) - 1)$-coreset for C_3.

Suppose that a sequence of points $p_1, p_2, \ldots$ in $\mathbb{R}^d$ arrive one by one. We want to compute the k-means of the points that arrive so far, and the result should be correct with probability $\geq 1 - \lambda$. The algorithm is quite similar to the ones in [10, 3] but unlike [10, 3], works for weak coresets as well as strong coresets.

Conceptually, we use buckets $B_0, B_1, \ldots$ to store points. The capacity of bucket B_0 is M, where $M = k^2 \epsilon^{-5} \log n$, and the capacity of bucket B_i is $2^{i-1}M$, for $i \geq 1$. We will keep an invariant in the algorithm: B_i is either full or empty, for $i \geq 1$. When p_m arrives, we insert p_m into B_0. If B_0 has less than M points, then we are done. Otherwise, we move all the points of B_0 into a virtual bucket B_1'. If B_1 is empty, move points of B_1' into B_1, and we are done; otherwise we merge the points of B_1' and B_1 into a virtual bucket B_2'. Then we try to move points of B_2' into B_2. We continue the process until we reach a stage r where B_r is empty; and then the points of virtual bucket B_r' are moved into B_r.

We simulate the above tree computation in small space by playing with weights. Let $\mathcal{N}$ denote the set of centroids of all subsets from $\mathcal{P}$ of size $2/\epsilon$. We maintain a weak coreset $(Q_i, \mathcal{N} \cup \mathcal{O})$ (resp. $(Q_i', \mathcal{N} \cup \mathcal{O})$) for each bucket B_i (resp. virtual bucket B_i'), for $i = 0, 1, \ldots$, as follows: Q_0 is B_0 itself; and whenever the points of B_r' and B_r are merged into B_{r+1}', we compute a weak (k, ρ_r)-coreset $(Q_{r+1}', \mathcal{N} \cup \mathcal{O})$ of $(Q_r \cup Q_r', \mathcal{N} \cup \mathcal{O})$ via coreset construction of Lemma 5 with confidence parameter $\lambda_n = \lambda/n^{2k/\epsilon}$, where $r \geq 1$, $\rho_r = \epsilon/cr^2$, n is the number of points received so far, and c is a large positive constant. Simple calculations shows coreset size would be

$$\begin{aligned} 2M + \sum_{i=2}^{\log_2 n} |Q_i| &= O\left(k\epsilon^{-4} \log^9 n \left(\log\left(kn^{2k/\epsilon}/\lambda\right)\right)\right) \\ &= O(k^2 \epsilon^{-5} \log^{10} n) \end{aligned}$$

To analyze the update time for k-means, observe that the amortized time dealing with Q_0 and Q_1 is constant; and for $j = 2, \ldots, \log_2 n$, Q_j is constructed after every $2^{j-1}M$ insertions are made. Therefore the amortized time spent for an update is

$$\begin{aligned} & \left(\sum_{i=2}^{\log_2 n} \frac{1}{2^{i-1}M} \cdot O\left(|Q_{i-1}| dk \cdot \log n^{2k/\epsilon}/\lambda\right) \right) \\ = & \ O\left(dk^2/\epsilon \log^2 n\right) . \end{aligned}$$

We should mention Chen [3] also adapted this standard technique to maintain his coreset in the streaming context obtaining a coreset size of $O(k^2 d\epsilon^{-2} \log^8 n)$, so in comparison the new coreset size is independent of d (which is interesting in the context of kernel k-means), losing two more factors of $\log n$ and three of ϵ.

6. REFERENCES

[1] J.L. Bentley and J.B. Saxe. Decomposable searching problems I: Static-to-dynamic transformation. *J. Algorithms, 1(4):301-358. 1980*, pages 301–358, 1980.

[2] M. Bădoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. *Proc. 34th Annu. ACM Sympos. Theory Comput. (STOC)*, pages 396–407, 2002.

[3] K. Chen. On k-Median clustering in high dimensions. *Proc. 17th Annual ACM-SIAM Symposium of Discrete Algorithms (SODA)*, pages 1177–1185, 2006.

[4] W. Fernandez de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani. Approximation schemes for clustering problems. *Proc. 35th Annu. ACM Sympos. Theory Comput. (STOC)*, pages 50–58, 2003.

[5] M. Effros and L.J. Schulman. Deterministic clustering with data nets. Report TR04-085, Elec. Colloq. Comp. Complexity, http://www.eccc.uni-trier.de/eccc-reports/2004/TR04-085, 2003.

[6] D. Feldman, A. Fiat, and M. Sharir. Coresets for weighted facilities and their applications. *Proc. 47th Annu. IEEE Sympos. Found. Comput. Sci. (FOCS)*, 2006.

[7] G. Frahling and C. Sohler. Coresets in dynamic geometric data streams. *Proc. 37th Annu. ACM Sympos. Theory Comput. (STOC)*, pages 209–217, 2005.

[8] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[9] S. Har-Peled and A. Kushal. Smaller coresets for k-median and k-means clustering. *Proc. 21st Annu. ACM Sympos. Comput. Geom. (SOCG)*, pages 126–134, 2005.

[10] S. Har-Peled and S. Mazumdar. Coresets for k-means and k-median clustering and their applications. *Proc. 36th Annu. ACM Sympos. Theory Comput. (STOC)*, pages 291–300, 2004.

[11] S. Har-Peled and K. R. Varadarajan. Approximation schemes for clustering problems. *Proc. 18th Annu. ACM Sympos. Comput. Geom.(SoCG)*, pages 312–318, 2002.

[12] M. Inaba, N. Katoh, and H. Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. *Proc. 10th Annu. ACM Sympos. Comput. Geom.(SoCG)*, pages 332–339, 1994.

[13] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. *Proc. 45th Annual Symposium on Foundations of Computer Science*, pages 454–462, 2004.

[14] A. Kumar, Y. Sabharwal, and S. Sen. Linear time algorithms for clustering problems in any dimensions. *Proc. 32nd Annual Internat. Colloquium on Automata, Languages, and Programming (ICALP)*, pages 1374–1385, 2005.

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

[16] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[17] J. Matousek. On approximate geometric k-clustering. *Discrete Comput. Geom.*, 24:61–84, 2000.

[18] R.R. Mettu and C.G. Plaxton. Optimal time bounds for approximate clustering. *Machine Learning*, 56:35–60, 2004.

[19] R. Ostrovsky, Y. Rabani, L. Shulman, and C. Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Proc. 47th Annu. IEEE Sympos. Found. Comput. Sci. (FOCS)*, 2006.