CGL

Computational Geometric Learning

Probabilistic k-Median Clustering in Data Streams

Christiane Lammersen Melanie Schmidt Christian Sohler

CGL Technical Report No.: 02

Part of deliverable: WP-1/RTD
Site: TUDO
Month: 12

Project co-funded by the European Commission within FP7 (2010–2013)
under contract nr. IST-25582, an Ebco Eppich Fellowship,
a PIMS Fellowship and by DFG grant SO 514/4-3.

1 Introduction

Most of the real world datasets contain uncertain, imprecise, or incomplete data. Typical examples are measurements of sensor networks or datasets arising from record linkage across multiple data sources. Imagine, for example, that we maintain a dataset representing our knowledge on a certain set of entities. Now, we want to add information from an additional dataset containing data records of some, but not all of our entities. The new dataset might also contain data records of entities not present in our knowledge base. In our scenario, we might not be sure whether two data records from the two sets belong to the same entity or not, but we might have a good idea of how likely this is. Thus, if we view our knowledge base as a set X of multi-dimensional points, each data record from our new set can be seen as a discrete probability distribution over X with a total probability between 0 and 1. If we are now interested in analyzing our new data, we have to deal with uncertainty, i. e., we must be able to cope with uncertain points.

One important tool for data analysis that is needed also in the case of uncertain data is *clustering*. Clustering is the problem to partition a given set of objects into subsets called clusters such that objects in the same cluster are similar and objects in different clusters are dissimilar. Therefore, clustering is a useful tool for data compression, object classification, or pattern recognition. Often, in addition to containing uncertain data, real world data sets can be very large and are given as a data stream or stored on hard disks. In these cases, random access to the data would be very time consuming or is even not possible. Thus, one needs streaming algorithms for clustering large sets of uncertain data.

In this paper, we study the development of clustering algorithms for uncertain data based on *coresets*. Intuitively, a coreset is a small set of weighted points that approximates the input points with respect to the studied clustering problem. Instead of clustering the original dataset, a clustering algorithm can then be run on the small coreset to obtain a $(1+\varepsilon)$-approximation in shorter running time. We develop two different coreset constructions, one for the *metric* and one for the *Euclidean* uncertain k-median-clustering problem.

1.1 Related Work

Clustering Uncertain Data. In recent publications, some traditional clustering heuristics have been extended so that they can handle uncertain data. For instance, Chau et al. [10] and Ngai et al. [33] extended Lloyd's k-means algorithm [16, 31], and Kriegel and Pfeifle [28, 29] and Xu and Li [35] extended the density-based clustering algorithm DBSCAN [14] for handling uncertainty. Furthermore, Günnemann et al. [20] developed a subspace clustering heuristic for uncertain data. A survey of uncertain data mining and management applications can be found in [2].

Surprisingly, only a few theoretical results on clustering uncertain data have been obtained so far [12, 19]. Cormode and McGregor [12] introduced the study of *probabilistic clustering problems*, where probabilistic means that the input is a set of probabilistic points, each formalized as a probability distribution function

which describes the possible locations of the points. There are two possible variations of probabilistic clustering. In the first variation, called *unassigned clustering*, each point is assigned to the closest cluster center. In the second variation, called *assigned clustering*, each point is assigned to a fixed cluster center, no matter where it is actually located. In this paper, we focus on the assigned version of the *probabilistic k-median problem*. The k-median objective is one of the most frequently used clustering objectives. In our probabilistic version of the k-median problem, we allow that the total realization probability of a point can be smaller than 1, i. e., it is possible that a point is not realized at all. The goal is now to find k cluster centers, which are deterministic points in the considered metric space, and an assignment of the points to these cluster centers such that the sum of the expected distances of the points to the cluster centers is minimized. Note that we have to assign each point to a cluster center before we know where it is eventually realized.

Cormode and McGregor [12] achieved a $\mathcal{O}(1)$-approximation for the assigned metric k-median problem by first computing the 1-median of each probabilistic input point and then clustering the 1-medians. They also considered other clustering problems. For the unassigned Euclidean k-median problem and both the unassigned and assigned Euclidean k-means problem, they obtained a $(1+\varepsilon)$-approximation by using a simple reduction to weighted deterministic clustering problems. For the unassigned metric k-center problem, they proposed a bicriteria approximation algorithm which results in a constant factor approximation but uses $2k$ instead of k cluster centers. In a follow-up work, Guha and Munagala [19] improved the last-mentioned result. More precisely, by using a reduction to a "truncated" version of a deterministic metric k-median problem, they obtained a constant factor approximation for both the unassigned and assigned metric k-center problem that preserves the number of allowed cluster centers k.

Clustering Certain Data. The classical k-median-clustering problem, which does not consider uncertainty and which we refer to as *the deterministic k-median-clustering problem*, is well-studied. We start with an overview for the Euclidean k-median problem. Arora et al. [4] developed the first $(1+\varepsilon)$-approximation algorithm for the Euclidean k-median problem in the plane by extending the technique developed by Arora [3] for the Euclidean TSP. The running time of their algorithm is $\mathcal{O}(n^{\mathcal{O}(1/\varepsilon)+1})$. Their work was improved by Kolliopoulos and Rao [27], who solved the d-dimensional case with $d \geq 2$ and reduced the running time to $\mathcal{O}(2^{\tilde{\mathcal{O}}(1/\varepsilon)^{d-1}} n \log^{d+6} n)$.

Further improvements have been obtained on the base of coresets. The first coreset construction for deterministic clustering problems was given by Bădoiu et al. [6]. Their coreset leads to a $(1+\varepsilon)$-approximation with an expected running time of $\mathcal{O}(2^{\text{poly}(k,1/\varepsilon)} d^{\mathcal{O}(1)} n \log^{\mathcal{O}(k)} n)$, which is polynomial in the dimension. Later, Agarwal et al. [1] proposed a definition of coresets as it is nowadays used for clustering problems. Har-Peled and Mazumdar [22] used coresets to achieve a running time of $\mathcal{O}(n + \text{poly}(1/\varepsilon)^{d-1} k^{\mathcal{O}(1)} \log^{\mathcal{O}(1)} n)$, which is linear for fixed ε, d, and k. They showed how to maintain their coreset in insertion-only data streams. Har-Peled and Kushal [21] constructed a coreset

of size $\mathcal{O}(k^2\varepsilon^{-d})$, which is independent of the number of input points. Another result was given by Frahling and Sohler [17], who proposed a coreset of size $\mathcal{O}(k\varepsilon^{-d}\log(n))$ that can be maintained in insertion-deletion data streams. Furthermore, Kumar, Sabharwal and Sen [30] provided a $\mathcal{O}(2^{(k/\varepsilon)^{\mathcal{O}(1)}}dn)$ time $(1+\varepsilon)$-approximation for the k-median problem. Chen [11] developed a coreset of size $\mathcal{O}(dk^2\varepsilon^{-2}\log(n)\log(k/\varepsilon))$ and used it to achieved a $(1+\varepsilon)$-approximation in time $\mathcal{O}(ndk + 2^{(k/\varepsilon)^{\mathcal{O}(1)}d^2\log^{k+2}n})$. Finally, Feldman et al. [15] constructed a weak coreset of size $\text{poly}(k, \varepsilon^{-1})$, which is independent of the number of input points n and the dimension d, leading to a PTAS with running time $\mathcal{O}(ndk + d(k/\varepsilon)^{\mathcal{O}(1)} + 2^{\mathcal{O}(k/\varepsilon)})$.

For general metric spaces, a $(1+\varepsilon)$-approximation with running time polynomial in k is not possible for $\varepsilon < 1.73$ unless NP $\subseteq$ DTIME$(n^{\mathcal{O}(\log\log n)})$ [25]. Also, for fixed k and a constant approximation factor, the deterministic k-median problem proved demanding. Charikar et al. [9] gave the first constant-factor approximation. Indyk [24] developed a randomized bicriteria approximation with constant approximation factor and $\mathcal{O}(k)$ centers. Based on that, Guha et al. [18] developed a constant-factor approximation with running time $\tilde{\mathcal{O}}(nk)$ and showed how to maintain it in insertion-only data streams. Mettu and Plaxton [32] proved a $\Omega(nk)$ lower bound for any constant-factor approximation, even for randomized algorithms, and also achieved a running time of $\tilde{\mathcal{O}}(nk)$. Chen [11] used coresets and developed a $(10+\varepsilon)$-approximation algorithm with running time $\mathcal{O}(nk+k^7\varepsilon^{-5}\log^5 n)$. The coreset has size $\mathcal{O}(dk^2\varepsilon^{-2}\log n\log(k/\varepsilon))$. The approximation factor was then consecutively improved to 6 by Jain and Vazirani [26], to 4 by Charikar and Guha [8], and finally to 3 by Arya et al. [5]. The latter algorithm uses local search and analyzes the locality gap.

To the best of our knowledge, there does not exist any coreset construction for probabilistic clustering problems.

1.2 Our Contribution

The focus of our work is defining and constructing probabilistic coresets. How should such a coreset look like? It can certainly contain probabilistic points. However, in addition to the number of such points, the overall storage size is also influenced by the representation size of the probability distributions of the points. Thus, we define a coreset by restricting both the number and the representation size of the probabilistic points.

Our first coreset construction deals with the metric k-median problem. Here, a probabilistic coreset can be constructed by a reduction to the deterministic metric k-median problem. Note that the expected distances between points do not satisfy the properties of a metric space because two copies of the same probabilistic point do not have expected distance zero as it should be for identical elements of a metric space. Thus, the intuitive reduction does not work. However, the expected *earth mover distance* can be used to define a metric space and hence makes the use of existing coreset constructions for the deterministic metric k-median problem possible.

In the Euclidean case, one can also transform a problem instance into an input for the deterministic metric k-median problem, because a Euclidean space

is also a metric space. However, the resulting instance is indeed metric but not Euclidean anymore. Thus, a coreset construction based on the transformed input can only use metric coreset constructions and not constructions specifically for Euclidean inputs. Now, the problem is that while metric algorithms provide a coreset which grants an approximation for all centers picked from a finite subset of the metric space, coresets for the Euclidean clustering problem have to approximate center sets from the infinite Euclidean space. Due to this fact, one cannot use a metric coreset construction to compute a coreset for the Euclidean k-median problem.

Thus, we develop a coreset construction for the probabilistic Euclidean k-median problem by extending and further developing a technique of Chen [11] used for the deterministic k-median problem.The main idea of this construction is to use a bicritera approximation to partition the input into subsets of points which are close to each other (in terms of their optimal clustering cost) and then sampling representatives from each such subset.

We use approximate 1-medians for the probability distributions of all points to compute a center set $\mathcal{A}$ of a bicriteria approximation in the probabilistic setting. When partitioning the points according to $\mathcal{A}$, we do not only partition the points according to the distances of their associated 1-medians to the centers in $\mathcal{A}$, but also according to the width of their probability distributions. This is because we do not only have to ensure that the (1-medians of the) points in the same subset are close to each other (in terms of their optimal clustering cost), but also that they behave similar according to the cost induced by the width of their probability distributions. After partitioning, we sample points from all subsets. The sampling becomes weighted due to the different total realization probabilities. In order to obtain representatives of the probabilistic input points that are not influenced by their former total realization probabilities, we weight each sample point individually (additionally to distributing the total weight of the subset among all sample points). Finally, we approximate the probability distributions of all sample points.

Both coreset constructions also work for the *weighted* probabilistic k-median problem. Furthermore, we show how to maintain the Euclidean coreset in data streams.

2 Preliminaries

First, we give some notations concerning dealing with metric spaces. Let $M = (X, \mathrm{D})$ be any metric space, where X is a set of points and D is a distance function defined on X. For any finite subset $Y \subseteq X$ and any point $x \in X$, let $\mathrm{D}(x, Y) := \min_{y \in Y}\{\mathrm{D}(x, y)\}$ denote the minimal distance between x and points in Y. For any finite subsets $Y, Z \subseteq X$, let $\mathrm{D}(Y, Z) := \min_{y \in Y}\{\mathrm{D}(y, Z)\}$ denote the minimal distance between points in Y and Z. For any finite set $Y \subseteq X$, let $\mathrm{diam}(Y) := \max_{y,z \in Y} \mathrm{D}(y, z)$ denote the *diameter* of Y. For any point $x \in X$ and a non-negative value R, we define $\mathcal{B}(x, R)$ to be the ball with center x and radius R.

Next, we define the *probabilistic k-median-clustering problem.* Let $\mathcal{X} :=$

$\{x_1, \ldots, x_m\} \subseteq X$ be a finite set of m points from the metric space M, and let $V := \{v_1, \ldots, v_n\}$ be a set of n *nodes*, where each node v_i follows an independent probability distribution $\mathcal{D}_i$ over $\mathcal{X}$. For any $i \in [n]$ and any $j \in [m]$, we denote the probability that the node v_i is realized at x_j by p_{ij}. We denote the total probability that v_i is realized by $p_i := \sum_{j=1}^{m} p_{ij}$. We assume that $p_i \leq 1$, which means that with probability $1 - p_i$ the node v_i is not realized. Let $p_{\min}$ be the smallest realization probability, i. e., $p_{\min} \leq p_{ij}$ for each $i \in [n]$ and each $j \in [m]$.

Definition 1 (Probabilistic k-Median)**.** *Given $k \in \mathbb{N}$ and a finite set of κ possible center locations $\mathcal{C} \subseteq X$, the* probabilistic metric k-median-clustering problem *for the set of nodes V is to find a set $C := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ of k cluster centers and an assignment $\rho : V \to C$ such that the expected k-median-clustering cost*

$$\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}(V, C, \rho)\right] := \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \cdot \mathrm{D}(x_j, \rho(v_i))$$

is minimized.

Analogously, given $k \in \mathbb{N}$, a finite set of κ possible center locations $\mathcal{C} \subseteq X$, and a positive weight function $w : V \to \mathbb{R}_{\geq 0}$ on the set of nodes V, the weighted probabilistic metric k-median-clustering problem *for V is to find a set $C := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ of k cluster centers and an assignment $\rho : V \to C$ such that the expected k-median-clustering cost*

$$\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, C, \rho)\right] := \sum_{i=1}^{n} w(v_i) \sum_{j=1}^{m} p_{ij} \cdot \mathrm{D}(x_j, \rho(v_i))$$

is minimized. We denote the cost of an optimal clustering by $\mathrm{cost}_k^*(V)$ and the total weight by $W := \sum_{v_i \in V} w(v_i) p_i$.

In case that the metric space is a Euclidean space $\mathbb{R}^d$, the definition of the unweighted (resp. weighted) probabilistic Euclidean k-median-clustering problem is verbatim, except that the set of possible center locations is $\mathcal{C} = \mathbb{R}^d$.

The unweighted (resp. weighted) deterministic k-median-clustering problem is a special case of the unweighted (resp. weighted) probabilistic k-median-clustering problem, where $m = n$ and, for each node v_i, we have $p_{ii} = 1$ and $p_{ij} = 0$ for all $j \neq i$. Note that, in the literature, it is typically assumed that $\mathcal{C} = \mathcal{X}$, i. e., the set of possible center locations is equal to the set of clustered points. Thus, our definition is a generalization of the typical definition of the deterministic k-median-clustering problem.

Additionally note that if, for each node v_i, we have one $j \in [m]$ with $p_{ij} = p_i$ and hence $p_{ij'} = 0$ for all $j' \neq j$, then the unweighted (resp. weighted) probabilistic k-median clustering can be immediately reduced to a weighted deterministic k-median clustering. Note that we use the notation cost_k^* in this case as well, then referring to the optimal weighted deterministic k-median-clustering cost.

Finally, we give a definition of a *coreset* for the probabilistic k-median-clustering problem. Our definition restricts both the number of probabilistic points in the

coreset and the size of the probability distributions describing the points. Let $U := \{u_1, \ldots, u_s\}$ be a set of s nodes where each $u_o \in U$ follows an independent probability distribution $\mathcal{D}'_o$ over $\mathcal{X}$. For any $o \in [s]$ and any $j \in [m]$, we denote the probability that u_o is realized at x_j by p'_{oj}. We denote the total probability that u_o is realized by $p'_o := \sum_{j=1}^{m} p'_{oj}$.

Definition 2 (Coreset for Probabilistic k-Median)**.** *Given the set of nodes V, let $U := \{u_1, \ldots, u_s\} \subseteq V$ be a weighted set of nodes with positive weight function $w : U \to \mathbb{R}_{\geq 0}$, and let $\mathcal{D}' := \{\mathcal{D}'_1, \ldots, \mathcal{D}'_s\}$ be a set of s probability distributions over $\mathcal{X}$ defining the distribution of nodes in U. Given $k \in \mathbb{N}$, a finite set of κ possible center locations $\mathcal{C} \subseteq X$, and a precision parameter ε, $0 < \varepsilon \leq 1$, the set U is called (k, ε)-coreset of V for the probabilistic metric k-median-clustering problem if, for each $C \subseteq \mathcal{C}$ of size $|C| = k$, we have*

$$\left| \min_{\rho: U \to C} \mathbf{E}_{\mathcal{D}'} \left[\mathrm{cost}_w(U, C, \rho) \right] - \min_{\rho: V \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}(V, C, \rho) \right] \right| \leq \varepsilon \cdot \min_{\rho: V \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}(V, C, \rho) \right] \quad .$$

In case that the metric space is a Euclidean space $\mathbb{R}^d$, the definition of a (k, ε)-coreset is verbatim, except that the set of possible center locations is $\mathcal{C} = \mathbb{R}^d$. If the nodes in V are already weighted by a positive weight function $w' : V \to \mathbb{R}_{\geq 0}$, then the definition changes by replacing $\mathrm{cost}(V, C, \rho)$ *with* $\mathrm{cost}_{w'}(V, C, \rho)$.

3 Coreset for Metric k-Median

3.1 Morphing Probability Distributions

We can compute a coreset for the probabilistic metric k-median-clustering problem by reducing the problem to the deterministic metric k-median-clustering problem. Imagine for the moment that the total realization probabilities p_i are uniform for all input nodes $v_i \in V$, i. e., we have $p_i = p$ for all $i \in [n]$ and some fixed probability p, $0 < p \leq 1$. Then, two probability distributions can be 'morphed' into one another. This can be done by using the *earth mover distance* (EMD), which gives the cost needed to morph one probability distribution into another. It is defined as follows:

Definition 3. *Let $v_{i'}$ and $v_{i''}$ be two nodes in V with $p_{i'} = p_{i''}$. We say a mapping $\varrho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ morphs $v_{i'}$ into $v_{i''}$ if it satisfies, for all $x_{j'}, x_{j''} \in \mathcal{X}$,*

$$\sum_{x_j \in \mathcal{X}} \varrho(x_{j'}, x_j) = p_{i'j'} \quad \text{and} \quad \sum_{x_j \in \mathcal{X}} \varrho(x_j, x_{j''}) = p_{i''j''} \quad .$$

The cost of the mapping is defined as

$$\mathrm{morph}(v_{i'}, v_{i''}, \varrho) := \sum_{x_{j'} \in \mathcal{X}} \sum_{x_{j''} \in \mathcal{X}} \varrho(x_{j'}, x_{j''}) \cdot \mathrm{D}(x_{j'}, x_{j''}) \quad .$$

The expected earth mover distance $\mathrm{EMD}(v_{i'}, v_{i''})$ *between $v_{i'}$ and $v_{i''}$ is the minimum cost of a mapping that morphs $v_{i'}$ into $v_{i''}$.*

The EMD forms a metric space [34], and thus provides a way to reduce the probabilistic k-median problem to its deterministic version. In the following lemma, we will show that this reduction indeed works and can be extended to the case of non-uniform realization probabilities.

Lemma 4. *Given $k \in \mathbb{N}$ and a finite set of κ possible center locations $\mathcal{C} \subseteq X$, the computation of a (k,ε)-coreset of V for the probabilistic metric k-median-clustering problem can be reduced to the computation of a (k,ε)-coreset for the deterministic metric k-median-clustering problem.*

Proof. Note that the cost to assign a node to a center in a probabilistic k-median clustering coincides with the EMD between the node and a node which has all its realization probability concentrated in the center. Furthermore, the EMD defined on a set of nodes with uniform total realization probabilities is a metric [34]. Let us assume for the moment that the realization probability of each node is a fixed value p. Now, for each possible center location $x \in \mathcal{C}$, we consider an artificial node denoted by $\text{node}(x)$ that is located at x and has total realization probability p. We define a new metric space that has a point $\sigma(v_i)$ for each node $v_i \in V$ and a point $\sigma(\text{node}(x))$ for each artificial node $\text{node}(x)$. The distance between two points in the new metric space is given by the EMD between the corresponding original nodes. Let $\sigma(V)$ be the points resulting from the nodes in V, and let $\sigma(\text{node}(\mathcal{C}))$ be the points resulting from the artificial nodes. The computation of a (k,ε)-coreset of V for the probabilistic metric k-median problem is equivalent to the computation of a (k,ε)-coreset of $\sigma(V)$ for the deterministic metric k-median problem in which we use the new metric defined above and in which the k centers are points from $\sigma(\text{node}(\mathcal{C}))$.

To get rid of the assumption that the nodes have uniform total realization probabilities, we use a grouping technique. More precisely, let V be a set of nodes with non-uniform total realization probabilities. Recall that we denote the smallest occurring realization probability by $p_{\min}$. We round the total realization probability of each node $v_i \in V$ down to a value $\widetilde{p_i} := p_{\min} \cdot (1+\varepsilon)^\ell$ where ℓ is the largest natural number including 0 such that the new realization probability $\widetilde{p_i}$ is smaller than or equal to the original realization probability p_i. Obviously, we have $\widetilde{p_i} \geq p_i/(1+\varepsilon) \geq (1-\varepsilon)p_i$. We can achieve the rounding by changing successively the realization probability p_{ij} at an arbitrary point $x_j \in \mathcal{X}$ to a value $\widetilde{p_{ij}}$ with $(1-\varepsilon)p_{ij} \leq \widetilde{p_{ij}} < p_{ij}$ until we have reduced the realization probability of v_i by the desired amount. Since each realization probability is reduced by at most an ε-fraction, the expected clustering cost for V is decreased by at most an ε-fraction.

Now, we have at most $\lceil \log_{1+\varepsilon}(1/p_{\min}) \rceil + 1 = \mathcal{O}(\varepsilon^{-1}\log(1/p_{\min}))$ groups of nodes with the same total realization probability. For each group, we find a (k,ε)-coreset using the EMD-approach as explained above. Since the union of (k,ε)-coresets for disjoint sets is a (k,ε)-coreset for the union of the original sets [22], the union of the (k,ε)-coresets for the groups is a (k,ε)-coreset for the set nodes in V with rounded realization probabilities. Hence, we have computed a $(k,2\varepsilon)$-coreset for the nodes in V with unrounded realization probabilities. Thus, running our algorithm with a precision parameter $\varepsilon' \leq \varepsilon/2$ leads to the desired result. □

3.2 Approximating Probability Distributions

Using an algorithm of Edmonds and Karp [13], the EMD between any two nodes $v_{i'}$ and $v_{i''}$ with $p_{i'} = p_{i''}$ can be computed in $\mathcal{O}(m^3)$ time. In order to improve the running time needed to compute the EMD between nodes and to have a small representation of a coreset node, we approximate the probability distribution $\mathcal{D}_i$ of each node $v_i \in V$ by computing a $(1, \varepsilon)$-coreset of v_i for the probabilistic 1-median problem. This means, for any center $c \in \mathcal{C}$, the expected cost to assign the node v_i to the center c is $(1 \pm \varepsilon)$-approximated by the coreset. The $(1, \varepsilon)$-coreset construction can directly be transferred from weighted deterministic 1-median clustering as given in the proof of the following lemma.

Lemma 5. *Let $v_i \in V$ be any node and $\mathcal{C} \subseteq X$ be the set of possible center locations. The computations of a subset $Z_i \subseteq \mathcal{X}$ and an approximated probability distribution $\mathcal{D}'_i : Z_i \to [0, 1]$ that satisfies*

$$\left| \sum_{x_j \in Z_i} \mathcal{D}'_i(x_j) \cdot \mathrm{D}(x_j, c) - \sum_{x_j \in \mathcal{X}} p_{ij} \cdot \mathrm{D}(x_j, c) \right| \leq \varepsilon \sum_{x_j \in \mathcal{X}} p_{ij} \cdot \mathrm{D}(x_j, c)$$

for each center $c \in \mathcal{C}$ can be reduced to the computation of a $(1, \varepsilon)$-coreset for the deterministic 1-median problem.

Proof. We use a similar grouping technique as in the proof of Lemma 4. More precisely, we round each realization probability p_{ij} down to a value $\widetilde{p_{ij}} := p_{\min} \cdot (1 + \varepsilon)^\ell$ where ℓ is the largest natural number including 0 such that the new realization probability $\widetilde{p_{ij}}$ is smaller than or equal to the original realization probability p_{ij}. This decreases the clustering cost of v_i by at most an ε-fraction. For each exponent ℓ that occurs, we build one corresponding group $\mathcal{X}_{i,\ell} \subseteq \mathcal{X}$ of points with the same realization probability. There are at most $\mathcal{O}(\varepsilon^{-1} \log(1/p_{\min}))$ groups. We proceed with each group $\mathcal{X}_{i,\ell}$ in the same way. We call a $(1, \varepsilon)$-coreset algorithm for the deterministic 1-median clustering to compute a weighted subset $Z_{i,\ell} \subseteq \mathcal{X}_{i,\ell}$ with positive weight function $w' : Z_{i,\ell} \to \mathbb{R}_{\geq 0}$ such that, for each center $c \in \mathcal{C}$, we have

$$\left| \sum_{x_j \in Z_{i,\ell}} w'(x_j) \cdot \mathrm{D}(x_j, c) - \sum_{x_j \in \mathcal{X}_{i,\ell}} \mathrm{D}(x_j, c) \right| \leq \varepsilon \sum_{x_j \in \mathcal{X}_{i\ell}} \mathrm{D}(x_j, c) \ .$$

Summing up over all groups and setting $p'_{ij} := w'(x_j) \cdot \widetilde{p_{ij}}$, we obtain

$$\left| \sum_{\ell} \sum_{x_j \in Z_{i,\ell}} p'_{ij} \cdot \mathrm{D}(x_j, c) - \sum_{x_j \in \mathcal{X}} \widetilde{p_{ij}} \cdot \mathrm{D}(x_j, c) \right| \leq \varepsilon \sum_{x_j \in \mathcal{X}} \widetilde{p_{ij}} \cdot \mathrm{D}(x_j, c) \ .$$

Thus, we obtained a $(1, \varepsilon)$-coreset for the probabilistic k-median clustering of v_i with rounded realization probabilities. Hence, we have computed a $(k, 2\varepsilon)$-coreset for v_i with unrounded realization probabilities. Thus, running our algorithm with a precision parameter $\varepsilon' \leq \varepsilon/2$ leads to the desired result. □

3.3 Analysis of the Construction

In our coreset construction, we apply the algorithm for the deterministic metric k-median clustering developed by Chen [11]. Note that, although it is not explicitly mentioned by the author, this algorithm works for our more general version of the deterministic metric k-median clustering (where the set of clustered points and the set of possible center locations can be arbitrary finite subsets of points from the underlying metric space). The number of possible center locations has only a logarithmic influence on the coreset size.

Lemma 6 ([11])**.** *Given a set Z of z points to be clustered and a set $\mathcal{C}$ of κ possible center locations in a metric space, $k \in \mathbb{N}$, a precision parameter ε, $0 < \varepsilon < 1$, and an error probability parameter δ, $0 < \delta < 1$, one can compute a weighted set $W \subseteq Z$ in $\mathcal{O}(zk \log(1/\delta))$ time such that $|W| = \mathcal{O}(k\varepsilon^{-2}(k \log(\kappa) + \log(1/\delta)) \log(z))$ and W is a (k, ε)-coreset of Z for the deterministic metric k-median-clustering problem with probability $1 - \delta$.*

Theorem 7. *Given a node set $V := \{v_1, \ldots, v_n\}$ defined on $M = (X, D)$, a set of associated probability distributions $\mathcal{D}_1, \ldots, \mathcal{D}_n$ over $\mathcal{X} := \{x_1, \ldots, x_m\} \subset X$, a set $\mathcal{C} \subseteq X$ of κ possible center locations, $k \in \mathbb{N}$, a precision parameter ε, $0 < \varepsilon < 1$, and an error probability parameter δ, $0 < \delta < 1$, one can compute a weighted subset $U := \{u_1, \ldots, u_s\} \subseteq V$ and a set of probability distributions $\mathcal{D}' := \{\mathcal{D}'_1, \ldots, \mathcal{D}'_s\}$ such that U and $\mathcal{D}'$ build a (k, ε)-coreset of V for the probabilistic metric k-median problem with probability $1 - \delta$. The size of U is $\mathcal{O}(k\varepsilon^{-3}(k \log(\kappa) + \log(n/\delta)) \log(n) \log(1/p_{\min}))$, and each probability distribution in $\mathcal{D}'$ assigns $\mathcal{O}(\varepsilon^{-3} \log^2(\kappa nm/(\delta p_{\min})))$ points from $\mathcal{X}$ a positive probability. The coreset construction requires $\mathcal{O}\left(\varepsilon^{-6}k^2 \cdot \log^6\left(\kappa nm/(p_{\min} \cdot \varepsilon\delta)\right)\right)$ bits of space and has a running time of $\mathcal{O}(nm + \varepsilon^{-10}kn \log^7(\kappa nm/(\delta p_{\min})) + \varepsilon^{-1}m \log(nm/(\delta p_{\min})))$.*

If V is weighted, then the size of U is $\mathcal{O}\left(\varepsilon^{-3}k^2 \cdot \log^3\left(\kappa W/(w_{\min} \cdot p_{\min} \cdot \varepsilon\delta)\right)\right)$, and each probability distribution in $\mathcal{D}'$ assigns $\mathcal{O}(\varepsilon^{-3} \log^2(\kappa W/(w_{\min} \cdot p_{\min} \cdot \varepsilon\delta)))$ points from $\mathcal{X}$ a positive probability. The coreset construction requires $\mathcal{O}\left(\varepsilon^{-6}k^2 \cdot \log^6\left(\kappa W/(w_{\min} \cdot p_{\min} \cdot \varepsilon\delta)\right)\right)$ bits of space and has a running time of $\mathcal{O}(\varepsilon^{-10}knm \log^7(\kappa W/(w_{\min} \cdot p_{\min} \cdot \varepsilon\delta)))$.

Proof. For each node $v_i \in V$, we round the realization probability p_{ij} at each point $x_j \in \mathcal{X}$ down to a value $p_{\min}(1 + \varepsilon)^\ell$ and build the groups $\mathcal{X}_{i,\ell}$. This can be done in $\mathcal{O}(nm)$ time. Then, we approximate the probability distribution $\mathcal{D}_i$. Due to Lemma 6, with probability $1 - \delta/(nm)$, a $(1, \varepsilon)$-coreset of $\mathcal{X}_{i,\ell}$ can be computed in $\mathcal{O}(m \log(nm/\delta))$ time. Since there are $\mathcal{O}(\varepsilon^{-1} \log(1/p_{\min}))$ groups, the coresets for all groups and hence the approximated probability distribution $\mathcal{D}'_i$ for the rounded realization probabilities of v_i can be computed in $\mathcal{O}(\varepsilon^{-1}m \log(nm/\delta) \log(1/p_{\min}))$ time. Due to Union Bound and the fact that there are less than m groups, $\mathcal{D}'_i$ has the desired property with probability $1 - \delta/n$. Due to Lemma 6 and the fact that there are $\mathcal{O}(\varepsilon^{-1} \log(1/p_{\min}))$ groups $\mathcal{X}_{i,\ell}$, the probability distribution $\mathcal{D}'_i$ assigns $\mathcal{O}(\varepsilon^{-3} \log(\kappa nm/\delta) \log(m) \log(1/p_{\min}))$ points from $\mathcal{X}$ a positive probability.

Due to Union Bound and the fact that there are n nodes, $\mathcal{D}'$ has the desired

property with probability $1-\delta$. The running time to compute $\mathcal{D}'$ is $\mathcal{O}(nm + \varepsilon^{-1}m\log(nm/\delta)\log(1/p_{\min}))$.

Then, we round the total realization probabilities of all the nodes in V to a value $p_{\min}(1+\varepsilon)^{\ell}$ and build $\mathcal{O}(\varepsilon^{-1}\log(1/p_{\min}))$ groups of nodes. This can be done in $\mathcal{O}(nm)$ time. Obviously, each group contains at most n nodes. For each group, we compute a (k,ε)-coreset using the EMD-approach. The EMD between two nodes can be computed in $\mathcal{O}(\varepsilon^{-9}\log^3(\kappa nm/\delta)\log^3(m)\log^3(1/p_{\min}))$ time since each probability distribution assigns $\mathcal{O}(\varepsilon^{-3}\log(\kappa nm/\delta)\log(m)\log(1/p_{\min}))$ points from $\mathcal{X}$ a positive probability and the running time for computing the EMD is at most cubic in the number of points. Thus, due to Lemma 6, the construction of a coreset for a group needs $\mathcal{O}(nk\log(n/\delta)\cdot\varepsilon^{-9}\log^3(\kappa nm/\delta)\log^3(m)\log^3(1/p_{\min}))$ time and has the desired property with probability $1-\delta/n$. Since there are $\mathcal{O}(\varepsilon^{-1}\log(1/p_{\min}))$ groups, the coreset construction for all groups has a running time of $\mathcal{O}(\varepsilon^{-10}nk\log(n/\delta)\log^3(\kappa nm/\delta)\log^3(m)\log^4(1/p_{\min}))$ and works with probability $1-\delta$.

Due to Lemma 6 and since there are $\mathcal{O}(\varepsilon^{-1}\log(1/p_{\min}))$ groups each consisting of at most n nodes, the size of the resulting coreset U is $\mathcal{O}(k\varepsilon^{-3}(k\log(\kappa)+\log(n/\delta))\log(n)\log(1/p_{\min}))$.

Overall, our algorithm has a running time of

$$\mathcal{O}(nm+\varepsilon^{-1}m\log(nm/\delta)\log(1/p_{\min})+\varepsilon^{-10}nk\log(n/\delta)\log^3(\kappa nm/\delta)\log^3(m)\log^4(1/p_{\min}))\ .$$

Since our construction used twice the rounding technique and twice a coreset construction each having error probability δ, it computes a $(k,(1+\varepsilon)^4-1)$-coreset with probability $1-2\delta$. Thus, running our algorithm with precision parameter $\varepsilon' \leq \varepsilon/15$ and with an error probability parameter of $\delta' \leq \delta/2$ leads to the desired result.

We described the proof in the unweighted case because of readability and since in contrast to the Euclidean case we do not need the weighted case for maintaining the coreset in data streams. The construction works weighted as well, and the weighted results are stated in the theorem. □

4 Coreset for Euclidean k-Median

4.1 Coreset Construction

Our coreset construction for the Euclidean probabilistic k-median problem consists of three steps: (i) partitioning the nodes into disjoint subsets, (ii) drawing random sample nodes from each such subset, and (iii) approximating the probability distribution of each sample node. The union of all sample nodes together with their approximated probability distributions forms the desired coreset. Next, we give a detailed description of our construction.

Step (i) partitions the nodes into groups that are similar in terms of their locations and their contributions to the total clustering cost. It is based on two computations:

- For every node $v_i \in V$, compute a point y_i that satisfies $\sum_{j=1}^{m} p_{ij} \cdot \mathrm{D}(x_j, y_i) \leq 2 \cdot \min_{x_{j'} \in X} \sum_{j=1}^{m} p_{ij} \cdot \mathrm{D}(x_j, x_{j'})$, i. e., y_i is a 2-approximation of the probabilistic 1-median for v_i. Let $Y := \{y_1, y_2, \ldots, y_n\}$ be the weighted set of the resulting n points where the weight of y_i is set to $w(y_i) := w(v_i)$.

- Compute a set $\mathcal{A} \subseteq Y$ which is a center set of an $[\alpha, \beta]$-bicriteria approximation to $\mathrm{cost}_k^*(Y)$. That is, $\mathcal{A} := \{a_1, \ldots, a_\tau\}$ satisfies

$$\min_{\rho: Y \to \mathcal{A}} \mathbf{E}_{\mathcal{D}_Y} \left[\mathrm{cost}_w(Y, \mathcal{A}, \rho)\right] \leq \alpha \, \mathrm{cost}_k^*(Y) \ ,$$

  where $\alpha \geq 1$ is some constant, $\tau \leq \beta k$, $\beta = \mathcal{O}(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)))$, $W := \sum_{v_i \in V} w(v_i) p_i$ is the expected total weight of the nodes, $w_{\min} := \min\{\min_{v_i \in V}\{w(v_i)\}, 1\}$ is either the minimum weight of a node in V or 1, and $p_{\min}$ is the smallest realization probability, i. e., $p_{\min} \leq p_{ij}$ for each $i \in [n]$ and each $j \in [m]$. Let $\sigma_Y : Y \to \mathcal{A}$ be an assignment satisfying $\mathbf{E}_{\mathcal{D}_Y} \left[\mathrm{cost}_w(Y, \mathcal{A}, \sigma_Y)\right] = \min_{\rho: Y \to \mathcal{A}} \mathbf{E}_{\mathcal{D}_Y} \left[\mathrm{cost}_w(Y, \mathcal{A}, \rho)\right]$. Note that σ_Y assigns each point in Y to the closest center in $\mathcal{A}$. Let $\sigma_V : V \to \mathcal{A}$ be the corresponding assignment for V such that $\sigma_V(v_i) = \sigma_Y(y_i)$ for all $i \in [n]$.

In Section 4.3, we will see how to efficiently construct the sets $\mathcal{A}$ and Y by using known results. In Section 4.2, we will prove that $\mathcal{A}$ is a $[3\alpha + 2, \beta]$-bicriteria approximation to $\mathrm{cost}_k^*(V)$. This enables us to find a bound R on the average radius of the optimal cost of a probabilistic k-median clustering for V and an upper bound $2^\nu R$ on the distance between an arbitrary point in Y and its closest center in $\mathcal{A}$. The value of ν is $\lceil \log((9\alpha + 6)W/(w_{\min} \cdot p_{\min})) \rceil$. Based on this, we define the following partitioning: For all $\ell \in [\tau]$, define $Y_\ell \subseteq Y$ as the subset of points in Y that σ_Y assigns to a_ℓ, i. e., Y_ℓ is the set of all points whose closest point in $\mathcal{A}$ is a_ℓ (ties broken arbitrarily). Set $V_\ell := \{v_i | y_i \in Y_\ell\}$. Furthermore, for each $\ell \in [\tau]$ and each $h \in \{0, 1, \ldots, \nu\}$, let

$$Y_{\ell,h} := \begin{cases} Y_\ell \cap \mathcal{B}(a_\ell, R) & h = 0 \\ Y_\ell \cap [\mathcal{B}(a_\ell, 2^h R) \backslash \mathcal{B}(a_\ell, 2^{h-1} R)] & h \geq 1 \end{cases}$$

be the *h-th ring set* for the center a_ℓ and the corresponding set Y_ℓ. Set $V_{\ell,h} := \{v_i | y_i \in Y_{\ell,h}\}$. Since $2^\nu R$ is an upper bound on the distance between an arbitrary point in Y and its closest center in $\mathcal{A}$, every point in Y lies in exactly one ring set. Hence, the ring sets for the centers in $\mathcal{A}$ partition Y and thus also V into disjoint sets. This first partitioning gives us subsets of nodes which have relatively close 1-medians. However, as the probability distributions of the nodes can have different variances, they may not behave similarly according to the cost function. Thus, we further subdivide the ring sets according to the width $\sum_{x_j \in \mathcal{X}} (p_{ij}/p_i) D(x_j, y_i)$ of the probability distribution of v_i. Let $2^\mu R$ be an upper bound on the width of the probability distributions. The value for μ is $\lceil \log((6\alpha + 4)W/(w_{\min} \cdot p_{\min})) \rceil$. For each $\ell \in [\tau]$, each $h \in \{0, 1, \ldots, \nu\}$, and each $a \in \{0, 1, \ldots, \mu\}$, let

$$Y_{\ell,h,a} := \begin{cases} \{y_i \in Y_{\ell,h} \mid \sum_{x_j \in \mathcal{X}} (p_{ij}/p_i) \cdot \mathrm{D}(x_j, y_i) \leq R\} & a = 0 \\ \{y_i \in Y_{\ell,h} \mid 2^{a-1} R < \sum_{x_j \in \mathcal{X}} (p_{ij}/p_i) \cdot \mathrm{D}(x_j, y_i) \leq 2^a R & a \geq 1 \end{cases} .$$

Now, $V_{\ell,h,a}$ is the corresponding node set. The sets $V_{\ell,h,a}$ give us the desired partitioning.

Step (*ii*) samples the points in the following way: From each $V_{\ell,h,a}$, we sample a multiset $U_{\ell,h,a}$ with

$$s''' := \left\lceil c \cdot \varepsilon^{-2} \cdot [\log(1/\delta) + k\,(\log(k) + \log(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon))) + d\log(1/\varepsilon))] \right\rceil$$

nodes, where c is a sufficiently large constant. The nodes are sampled with replacement, and, in each sample step, a node $v_i \in V_{\ell,h,a}$ is picked with probability $w(v_i)p_i / \sum_{v_{i'} \in V_{\ell,h,a}} w(v_{i'})p_{i'}$. We set the weight of the sample node v_i to $w'(v_i) = \sum_{v_{i'} \in V_{\ell,h,a}} w(v_{i'})p_{i'}/(p_i s''')$ (see Lemma 9). We store all sampled nodes in the multiset

$$U := \{u_1, \ldots, u_s\} := \bigcup_{\ell=1}^{\tau} \bigcup_{h=0}^{\nu} \bigcup_{a=0}^{\mu} U_{\ell,h,a} \ .$$

Step (*iii*) computes for each coreset node v_i an approximated probability distribution $\mathcal{D}_i$ that assigns positive probability to at most $\mathcal{O}(\varepsilon^{-2}\gamma^2 \log^2(nmW)(d\log(1/\varepsilon) + \log(\gamma)\log(nmW) + \log(n/\delta)))$ points from $\mathcal{X}$. We will show in Section 4.3 how to do this computation using known results.

4.2 Correctness

In this section, we analyze our approach and prove that it indeed constructs a coreset. The analysis starts with the following result of Haussler [23], which enables us to bound the number of sample nodes necessary to reduce the error in the clustering cost below a certain level:

Lemma 8 ([23, 11])**.** *Let $L \geq 0$ and M be fixed constants, and let $f(\cdot)$ be a function defined on a set T such that $M \leq f(v) \leq M + L$ for all $v \in T$. Let S be a set of s' samples drawn independently and uniformly at random from T, and let $N > 0$ be a parameter. If*

$$s' \geq \frac{L^2}{2N^2} \cdot \ln\left(\frac{2}{\delta}\right) \ ,$$

then

$$\mathbf{Pr}\left[\left|\frac{f(T)}{|T|} - \frac{f(S)}{|S|}\right| \geq N\right] \leq \delta \ ,$$

where $f(T) = \sum_{v \in T} f(v)$ and $f(S) = \sum_{u \in S} f(u)$.

To use Lemma 8 for bounding the error, we have to find an upper bound on the contribution of a node to the total clustering cost (the function f). Additionally, we have to overcome the difficulty that our nodes have different realization probabilities while the sampling process in Lemma 8 samples uniformly at random.

Simulating Weighted Sampling

We would like to use weighted sampling to construct our coreset, i. e., to sample a node v_i with probability proportional to $w(v_i)$ and p_i. How can we reduce this to uniform sampling to use the above stated Lemma 8? It is intuitive to construct a set $\widetilde{V}$ which contains multiple copies of all nodes in V according to their weights and realization probabilities. Let S denote the resulting sample set. For each $v_i \in V$, the set $\widetilde{V}$ contains $\lfloor w(v_i)p_i/(w_{\min} \cdot p_{\min} \cdot \varepsilon) \rfloor$ copies of v_i each having a realization probability of $w_{\min} \cdot p_{\min} \cdot \varepsilon$. More precisely, let $v_{i,z}$ be the z-th such copy. Then, for each $x_j \in \mathcal{X}$, we set the probability that $v_{i,z}$ is realized at x_j to $\widetilde{p}_{ij} := p_{ij} \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon / p_i$. Let $\widetilde{\mathcal{D}}$ be the resulting set of probability distributions. Note that the expected weight of the node v_i is $w(v_i)p_i$ and the expected total weight of all the nodes in $\bigcup_z v_{i,z}$ is at least $w(v_i)p_i - w_{\min} \cdot p_{\min} \cdot \varepsilon \geq (1-\varepsilon)w(v_i)p_i$ and at most $w(v_i)p_i$. Thus, our construction changes the clustering-cost contribution of v_i by at most an ε-fraction of its original contribution. For sake of simplicity, we will always assume in the following that $w(v_i)p_i/(w_{\min} \cdot p_{\min} \cdot \varepsilon) \in \mathbb{N}$ for each $v_i \in V$.

Now, let T be any subset of V, and let $\widetilde{T} := \{v_{i,z} \in \widetilde{V} \mid v_i \in T\}$. Given any set $C \subseteq X$ of cluster centers and an assignment $\rho : T \to C$, we define the clustering cost of $\widetilde{T}$ by

$$\mathbf{E}_{\widetilde{\mathcal{D}}}\left[\mathrm{cost}(\widetilde{T}, C, \rho)\right] := \sum_{v_{i,z} \in \widetilde{T}} \sum_{j=1}^{m} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)) \ .$$

Note that $\widetilde{T}$ and T have the same clustering behavior. More precisely, for any set $C \subseteq X$ of cluster centers and any assignment $\rho : T \to C$, we have

$$\begin{aligned}
\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(T, C, \rho)\right] &= \sum_{v_i \in T} w(v_i) \sum_{j=1}^{m} p_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)) \\
&= \sum_{v_i \in T} \frac{w(v_i)p_i}{w_{\min} \cdot p_{\min} \cdot \varepsilon} \sum_{j=1}^{m} \frac{p_{ij} w_{\min} \cdot p_{\min} \cdot \varepsilon}{p_i} \cdot \mathrm{D}(x_j, \rho(v_i)) \\
&= \sum_{v_i \in T} \sum_{z=1}^{w(v_i)p_i/(w_{\min} \cdot p_{\min} \cdot \varepsilon)} 1 \cdot \sum_{j=1}^{m} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)) \\
&= \sum_{v_{i,z} \in \widetilde{T}} \sum_{j=1}^{m} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)) \\
&= \mathbf{E}_{\widetilde{\mathcal{D}}}\left[\mathrm{cost}(\widetilde{T}, C, \rho)\right] \ .
\end{aligned}$$

However, this is only half the way to our goal. We also have to find a transformation between the sample sets in both node descriptions that connects their clustering cost. Thus, let S be a set of sample nodes where each $u = v_i \in S$ is picked independently and non-uniformly at random from T according to the weights and realization probabilities, i. e., u is picked with probability $w(v_i)p_i / \sum_{v_{i'} \in T} w(v_{i'})p_{i'}$. Let $\widetilde{S}$ be a set of nodes that contains for each

$u = v_i \in S$ one of the copies $v_{i,z} \in \widetilde{T}$ chosen uniformly at random from all these copies. Then, $\widetilde{S}$ is a set of sample nodes from $\widetilde{T}$ where each sample node is picked independently and uniformly at random from $\widetilde{T}$ since each copy $v_{i,z}$ is picked with probability

$$\frac{w(v_i)p_i}{\sum_{v_{i'} \in T} w(v_{i'})p_{i'}} \cdot \frac{w_{\min} \cdot p_{\min} \cdot \varepsilon}{w(v_i)p_i} = \frac{w_{\min} \cdot p_{\min} \cdot \varepsilon}{\sum_{v_{i'} \in T} w(v_{i'})p_{i'}} = \frac{1}{|\widetilde{T}|} \ .$$

We set the weight of any sample node $v_i \in S$ to

$$w'(v_i) := \frac{\sum_{v_{i'} \in T} w(v_{i'})p_{i'}}{p_i|S|} = \frac{|\widetilde{T}|}{|\widetilde{S}|} \frac{w_{\min} \cdot p_{\min} \cdot \varepsilon}{p_i} \ .$$

The following relationship between the clustering behavior of S and $\widetilde{S}$ holds:

$$\begin{aligned}
\mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_{w'}(S, C, \rho)\right] &= \sum_{v_i \in S} w'(v_i) \sum_{x_j \in \mathcal{X}} p_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)) \\
&= \sum_{v_{i,z} \in \widetilde{S}} w'(v_i) \sum_{x_j \in \mathcal{X}} p_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)) \\
&= \sum_{v_{i,z} \in \widetilde{S}} \frac{|\widetilde{T}| w_{\min} \cdot p_{\min} \cdot \varepsilon}{p_i|\widetilde{S}|} \sum_{x_j \in \mathcal{X}} p_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)) \\
&= \frac{|\widetilde{T}|}{|\widetilde{S}|} \sum_{v_{i,z} \in \widetilde{S}} \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)) \\
&= \frac{|\widetilde{T}|}{|\widetilde{S}|} \cdot \mathbf{E}_{\widetilde{\mathcal{D}}} \left[\mathrm{cost}(\widetilde{S}, C, \rho)\right] \ . \qquad (1)
\end{aligned}$$

Prepared with this transformations, we can now state the main lemma of this subsection by finding a good upper bound on the clustering cost of any node v_i.

Lemma 9. *Let T be any subset of V, let $Y(T) := \{y_i \in Y \mid v_i \in T\}$ be the subset of the approximated probabilistic 1-medians of all the nodes contained in T, and let $\delta', \xi > 0$ be given parameters. Let S be a sample of $s'' := \lceil \xi^{-2} \ln(2/\delta') \rceil$ nodes picked independently and non-uniformly at random from T according to their weights and realization probabilities, where each node $v_i \in S$ is assigned weight $w'(v_i) := \sum_{v_{i'} \in T} w(v_{i'})p_{i'}/(p_i s'')$. For a fixed set $C \subseteq X$, we have that*

$$\begin{aligned}
&\left| \min_{\rho: T \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(T, C, \rho)\right] - \min_{\rho: S \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_{w'}(S, C, \rho)\right] \right| \\
\leq \ & \xi \left(\sum_{v_{i'} \in T} w(v_{i'})p_{i'} \right) \left(\mathrm{D}(Y(T), C) + \mathrm{diam}(Y(T)) + \max_{y_i \in Y(T)} \sum_{x_j \in \mathcal{X}} \frac{p_{ij}}{p_i} \cdot \mathrm{D}(x_j, y_i) \right)
\end{aligned}$$

with probability at least $1 - \delta'$.

Proof. Consider $\widetilde{T}$ and $\widetilde{V}$ as defined above. We define the function

$$f(v_{i,z}) := \min_{c \in C} \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, c)$$

over the nodes $v_{i,z} \in \widetilde{V}$.

For $c \in C$, set $y^c := \arg\min_{y' \in Y(T)} \mathrm{D}(y', c)$. Then, by the triangle inequality, for every node $v_{i,z} \in \widetilde{T} \subseteq \widetilde{V}$, it holds that

$$\begin{aligned}
0 \;\leq\; & f(v_{i,z}) = \min_{c \in C} \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, c) \leq \min_{c \in C} \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot [\mathrm{D}(x_j, y_i) + \mathrm{D}(y_i, y^c) + \mathrm{D}(y^c, c)] \\
\leq\; & \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, y_i) + \min_{c \in C} \sum_{x_j \in \mathcal{X}} \frac{p_{ij} \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon}{p_i} \cdot \mathrm{D}(y_i, y^c) + \min_{c \in C} \sum_{x_j \in \mathcal{X}} \frac{p_{ij} \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon}{p_i} \cdot \mathrm{D}(y^c, c \\
\leq\; & \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, y_i) + \min_{c \in C} \mathrm{D}(y_i, y^c) \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon + \min_{c \in C} \mathrm{D}(y^c, c) \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon \\
\leq\; & \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, y_i) + \mathrm{diam}(Y(T)) \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon + \mathrm{D}(Y(T), C) \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon \\
\leq\; & \mathrm{D}(Y(T), C) \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon + \mathrm{diam}(Y(T)) \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon + \max_{y_{i'} \in Y(T)} \sum_{x_j \in \mathcal{X}} \widetilde{p}_{i'j} \cdot \mathrm{D}(x_j, y_{i'}) \;\; .
\end{aligned}$$

Due to Lemma 8, setting

$$L := \mathrm{D}(Y(T), C) \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon + \mathrm{diam}(Y(T)) \cdot w_{\min} \cdot p_{\min} \cdot \varepsilon + \max_{y_i \in Y(T)} \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, y_i) \;\; ,$$

$M := 0$, and $N := \xi L$, we have that, for a sample $\widetilde{S}$ of size

$$s'' = \left\lceil \xi^{-2} \cdot \ln\left(\frac{2}{\delta'}\right) \right\rceil \geq \frac{L^2}{2N^2} \cdot \ln\left(\frac{2}{\delta'}\right)$$

from $\widetilde{T}$, it holds that

$$\begin{aligned}
& \mathbf{Pr}\left[\left| \frac{1}{|\widetilde{T}|} \sum_{v_{i,z} \in \widetilde{T}} \min_{c \in C} \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, c) - \frac{1}{|\widetilde{S}|} \sum_{v_{i,z} \in \widetilde{S}} \min_{c \in C} \sum_{x_j \in \mathcal{X}} \widetilde{p}_{ij} \cdot \mathrm{D}(x_j, c) \right| \geq \xi L \right] \\
= \;\; & \mathbf{Pr}\left[\left| \frac{f(\widetilde{T})}{|\widetilde{T}|} - \frac{f(\widetilde{S})}{|\widetilde{S}|} \right| \geq N \right] \leq \delta' \;\; .
\end{aligned}$$

Due to Equality (1), this implies that

$$\begin{aligned}
&\left|\min_{\rho:T\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(T,C,\rho)\right]-\min_{\rho:S\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(S,C,\rho)\right]\right|\\
=&\left|\min_{\rho:T\to C}\mathbf{E}_{\widetilde{\mathcal{D}}}\left[\mathrm{cost}(\widetilde{T},C,\rho)\right]-\min_{\rho:S\to C}\frac{|\widetilde{T}|}{|\widetilde{S}|}\cdot\mathbf{E}_{\widetilde{\mathcal{D}}}\left[\mathrm{cost}(\widetilde{S},C,\rho\right]\right|\\
=&\ |\widetilde{T}|\cdot\left|\frac{1}{|\widetilde{T}|}\min_{\rho:T\to C}\mathbf{E}_{\widetilde{\mathcal{D}}}\left[\mathrm{cost}(\widetilde{T},C,\rho)\right]-\frac{1}{|\widetilde{S}|}\min_{\rho:S\to C}\mathbf{E}_{\widetilde{\mathcal{D}}}\left[\mathrm{cost}(\widetilde{S},C,\rho)\right]\right|\\
\leq&\ \xi|\widetilde{T}|L\\
=&\ \xi|\widetilde{T}|\left(\mathrm{D}(Y(T),C)\cdot w_{\min}\cdot p_{\min}\cdot\varepsilon+\mathrm{diam}(Y(T))\cdot w_{\min}\cdot p_{\min}\cdot\varepsilon+\max_{y_i\in Y(T)}\sum_{x_j\in\mathcal{X}}\widetilde{p}_{ij}\cdot\mathrm{D}(x_j,y_i)\right)\\
=&\ \xi\left(\sum_{v_{i'}\in T}w(v_{i'})p_{i'}\right)\left(\mathrm{D}(Y(T),C)+\mathrm{diam}(Y(T))+\max_{y_i\in Y(T)}\sum_{x_j\in\mathcal{X}}\frac{p_{ij}}{p_i}\cdot\mathrm{D}(x_j,y_i)\right)
\end{aligned}$$

with probability at least $1-\delta'$. □

Now assume we are given a partitioning $V=\bigcup_{r=1}^{\lambda}P_r$ of disjoint subsets of V and sample sufficiently many points from each P_r to apply Lemma 9. Then, with high probability, the total error induced is

$$\begin{aligned}
\xi\Bigg(&\sum_{r=1}^{\lambda}\sum_{v_{i'}\in P_r}w(v_{i'})p_{i'}\mathrm{D}(Y(P_r),C)+\sum_{r=1}^{\lambda}\sum_{v_{i'}\in P_r}w(v_{i'})p_{i'}\,\mathrm{diam}(Y(P_r))\\
&+\sum_{r=1}^{\lambda}\sum_{v_{i'}\in P_r}w(v_{i'})p_{i'}\max_{y_i\in Y(P_r)}\sum_{x_j\in\mathcal{X}}\frac{p_{ij}}{p_i}\cdot\mathrm{D}(x_j,y_i)\Bigg)\ .
\end{aligned}$$

Thus, we search for a partitioning that allows for a good upper bound on these three error terms while keeping λ small. The following claim shows that the first term is independent of the partitioning and is already bounded because of the way we defined the y_i.

Claim 10. *Let $C\subseteq X$ be any set of k cluster centers, and let $V=\bigcup_{r=1}^{\lambda}P_r$ be any partitioning of V into disjoint subsets. If each $y_i\in Y$ is a 2-approximation of the probabilistic 1-median of v_i, then we have*

$$\sum_{r=1}^{\lambda}\sum_{v_{i'}\in P_r}w(v_{i'})p_{i'}\cdot\mathrm{D}(Y(P_r),C)\leq\sum_{y_i\in Y}w(y_i)p_i\cdot\mathrm{D}(y_i,C)$$

and

$$\sum_{y_i\in Y}w(y_i)p_i\cdot\mathrm{D}(y_i,C)\leq 3\cdot\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]\ .$$

Proof. Let $C \subseteq X$ be any set of k centers for the probabilistic k-median clustering problem for V, and let $\rho : V \to C$ be an assignment from nodes in V to centers in C that produces minimal cost using centers in C. Note that

$$\begin{aligned}\sum_{r=1}^{\lambda} \sum_{v_{i'} \in P_r} w(v_{i'}) p_{i'} \cdot \mathrm{D}(Y(P_r), C) &= \sum_{r=1}^{\lambda} \sum_{v_{i'} \in P_r} w(v_{i'}) p_{i'} \min_{y_i \in Y(P_r)} \mathrm{D}(y_i, C) \\ &\leq \sum_{r=1}^{\lambda} \sum_{v_{i'} \in P_r} w(v_{i'}) p_{i'} \cdot \mathrm{D}(y_{i'}, C) \\ &= \sum_{y_i \in Y} w(y_i) p_i \mathrm{D}(y_i, C)\end{aligned}$$

and thus the first part of the claim holds.

Now note that $\sum_{i=1}^{n} \sum_{j=1}^{m} w(v_i) p_{ij} \cdot \mathrm{D}(x_j, y_i) \leq 2 \cdot \mathrm{cost}_n^*(V) \leq 2 \cdot \mathrm{cost}_k^*(V)$ because in an optimal solution with n centers, the cost induced by v_i cannot be lower than the cost induced by v_i if it is assigned to its 1-median. Thus, because all y_i are 2-approximations of the corresponding v_i, Y provides a 2-approximation of $\mathrm{cost}_n^*(V)$. Furthermore, $2 \cdot \mathrm{cost}_n^*(V) \leq 2 \cdot \mathrm{cost}_k^*(V)$ because $k \leq n$. Combining this with the triangle inequality, we gain

$$\begin{aligned}\sum_{y_i \in Y} w(y_i) p_i \cdot \mathrm{D}(y_i, C) &\leq \sum_{i=1}^{n} w(v_i) p_i \cdot \mathrm{D}(y_i, \rho(v_i)) = \sum_{i=1}^{n} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \cdot \mathrm{D}(y_i, \rho(v_i)) \\ &\leq \sum_{i=1}^{n} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \cdot (\mathrm{D}(y_i, x_j) + \mathrm{D}(x_j, \rho(v_i))) \\ &\leq 2 \cdot \mathrm{cost}_k^*(V) + \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, C, \rho)\right] \\ &\leq 3 \cdot \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, C, \rho)\right] \ .\end{aligned}$$

□

Note that we proved the following claim during the proof of Claim 10. We state it explicitly since we will need it later on. Interestingly, it provides a bound on the last term of our error term for the case that we partition V into $|V|$ partitions each containing one node (which we will not do because $|V|$ is too large).

Claim 11. *If each $y_i \in Y$ is a 2-approximation of the probabilistic 1-median of v_i, then*

$$\sum_{i=1}^{n} \sum_{j=1}^{m} w(v_i) p_{ij} \cdot \mathrm{D}(x_j, y_i) \leq 2 \cdot \mathrm{cost}_n^*(V) \leq 2 \cdot \mathrm{cost}_k^*(V) \ .$$

Bicriteria Approximation

Let $\mathcal{A} \subseteq Y$ be the center set of the computed $[\alpha, \beta]$-bicriteria approximation to $\mathrm{cost}_k^*(Y)$. Recall that $\alpha \geq 1$ is some constant, $\tau \leq \beta k$, $\beta = \mathcal{O}(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)))$. We show that $\mathcal{A}$ is not only a $[\alpha, \beta]$-bicriteria approximation to

$\text{cost}_k^*(Y)$, but also a $[3\alpha+2, \beta]$-bicriteria approximation to $\text{cost}_k^*(V)$. This is true because $\mathbf{E}_{\mathcal{D}}[\text{cost}_w(V, \mathcal{A}, \sigma_V)]$ can be bounded by using Claim 11 to bound the probabilistic clustering cost when using the y_i as centers and by bounding the weighted distance from each y_i to its closest a_ℓ in $\mathcal{A}$. The latter fact is proven by the following claim.

Claim 12. *If each $y_i \in Y$ is a 2-approximation of the probabilistic 1-median of v_i and $\mathcal{A}$ is an $[\alpha, \mathcal{O}(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)))]$-bicriteria approximation to $\text{cost}_k^*(Y)$, then*

$$\sum_{\ell=1}^{\tau} \sum_{y_i \in Y_\ell} w(y_i) p_i \cdot \mathrm{D}(y_i, a_\ell) \leq 3\alpha \cdot \text{cost}_k^*(V) \ .$$

Proof. Let $C \subseteq X$ be any set of k optimal centers for the probabilistic k-median clustering problem for V, and let $\rho : V \to C$ be an optimal assignment from nodes in V to centers in C. Note that ρ provides an assignment for Y as well (assign y_i to $\rho(v_i)$), but for Y this is not necessarily optimal. By triangle inequality, we have

$$\begin{aligned}
\text{cost}_k^*(Y) &\leq \sum_{i=1}^{n} w(y_i) p_i \cdot \mathrm{D}(y_i, \rho(v_i)) = \sum_{i=1}^{n} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \cdot \mathrm{D}(y_i, \rho(v_i)) \\
&\leq \sum_{i=1}^{n} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \cdot (\mathrm{D}(y_i, x_j) + \mathrm{D}(x_j, \rho(v_i))) \\
&\leq 3 \cdot \text{cost}_k^*(V) \ ,
\end{aligned}$$

where the last inequality follows from Claim 11 and the definition of ρ. Since

$$\sum_{\ell=1}^{\tau} \sum_{y_i \in Y_\ell} w(y_i) p_i \cdot \mathrm{D}(y_i, a_\ell) = \min_{\rho: Y \to \mathcal{A}} \mathbf{E}_{\mathcal{D}_Y}[\text{cost}_w(Y, \mathcal{A}, \rho)] \leq \alpha \, \text{cost}_k^*(Y) \ ,$$

the claim follows. □

Now, we are prepared to bound $\mathbf{E}_{\mathcal{D}}[\text{cost}_w(V, \mathcal{A}, \sigma_V)]$ and thus prove that $\mathcal{A}$ is a bicriteria approximation to $\text{cost}_k^*(V)$.

Lemma 13. *If each $y_i \in Y$ is a 2-approximation of the probabilistic 1-median of v_i and $\mathcal{A}$ is an $[\alpha, \mathcal{O}(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)))]$-bicriteria approximation to $\text{cost}_k^*(Y)$, then*

$$\mathbf{E}_{\mathcal{D}}[\text{cost}_w(V, \mathcal{A}, \sigma_V)] \leq (3\alpha + 2) \cdot \text{cost}_k^*(V) \ .$$

Proof. It follows from triangle inequality and Claims 11 and 12 that

$$\begin{aligned}
\mathbf{E}_{\mathcal{D}}[\text{cost}_w(V, \mathcal{A}, \sigma_V)] &\leq \sum_{\ell=1}^{\tau} \sum_{y_i \in Y_\ell} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \cdot (\mathrm{D}(x_j, y_i) + \mathrm{D}(y_i, a_\ell)) \\
&= \sum_{y_i \in Y} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \cdot \mathrm{D}(x_j, y_i) + \sum_{\ell=1}^{\tau} \sum_{y_i \in Y_\ell} w(y_i) p_i \cdot \mathrm{D}(y_i, a_\ell) \\
&\leq (3\alpha + 2) \cdot \text{cost}_k^*(V) \ .
\end{aligned}$$

□

Lemma 13 also implies that $\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]/(3\alpha+2)$ is a lower bound on the optimal cost of a probabilistic k-median clustering for V. Thus,

$$R := \frac{\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]}{(3\alpha+2)W}$$

is a lower bound on the expected average optimal cost per unit weight. We want to find a bound (depending on R) for the distance between an arbitrary point in Y and its closest center in $\mathcal{A}$. Recall that $w_{\min} := \min\{\min_{v_i\in V}\{w(v_i)\},1\}$ is the minimum weight of a node and $p_{\min}$ is the smallest realization probability. It follows that no point can be in distance greater than $\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(Y,\mathcal{A},\sigma_Y)\right]/(w_{\min}\cdot p_{\min})$ from the closest center in $\mathcal{A}$. Thus, by triangle inequality, Claim 11, and the definition of R, we get

$$\begin{aligned}
\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(Y,\mathcal{A},\sigma_Y)\right] &= \sum_{i=1}^{n} w(y_i)p_i\cdot \mathrm{D}(y_i,\sigma_Y(y_i)) = \sum_{i=1}^{n} w(v_i)p_i\cdot \mathrm{D}(y_i,\sigma_V(v_i)) \\
&\le \sum_{i=1}^{n}\sum_{j=1}^{m} w(v_i)p_{ij}\cdot(\mathrm{D}(y_i,x_j)+\mathrm{D}(x_j,\sigma_V(v_i))) \\
&\le 2\cdot \mathrm{cost}_n^*(V) + \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right] \\
&\le 3\cdot \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right] = (9\alpha+6)RW \ .
\end{aligned}$$

Set $\nu := \lceil \log((9\alpha+6)W/(w_{\min}\cdot p_{\min}))\rceil$. Then,

$$\begin{aligned}
2^{\nu}R &= 2^{\lceil \log((9\alpha+6)W/(w_{\min}\cdot p_{\min}))\rceil}R \\
&\ge (9\alpha+6)RW/(w_{\min}\cdot p_{\min}) \\
&\ge \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(Y,\mathcal{A},\sigma_Y)\right]/(w_{\min}\cdot p_{\min}) \ ,
\end{aligned}$$

so $2^{\nu}R$ provides the desired bound. Due to the definition of $Y_{\ell,h}$, the diameter $\mathrm{diam}(Y_{\ell,h})$ is bounded by $2(2^hR)$ for each $\ell\in[\tau]$ and each $h\in\{0,1,\ldots,\nu\}$. As we see in the following lemma, we can use this bound to show an upper bound on the sum of the weighted diameters of all subsets $\mathrm{diam}(Y_{\ell,h})$ that is within a constant factor of the optimal clustering cost. Hence, we can bound the second term of the error term for the partitioning $V=\bigcup_{\ell=1}^{\tau}\bigcup_{h=0}^{\nu}V_{\ell,h}$.

Claim 14. *If each $y_i\in Y$ is a 2-approximation of the probabilistic 1-median of v_i and $\mathcal{A}$ is an $[\alpha,\mathcal{O}(\log(W/(w_{\min}\cdot p_{\min}\cdot\varepsilon)))]$-bicriteria approximation to $\mathrm{cost}_k^*(Y)$, then*

$$\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{y_i\in Y_{\ell,h}} w(y_i)p_i\cdot 2^hR \le (6\alpha+1)\cdot\mathrm{cost}_k^*(V)$$

and

$$\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{y_i\in Y_{\ell,h}} w(y_i)p_i\cdot \mathrm{diam}(Y_{\ell,h}) \le (12\alpha+2)\cdot\mathrm{cost}_k^*(V) \ .$$

Proof. Let $v_i \in V$ be an arbitrary node, and let $y_i \in Y_{\ell,h}$ for some $\ell \in [t]$ and $h \in \{0, \dots, \nu\}$. Note that $2^h R = R$ if $h = 0$, and $\mathrm{D}(y_i, a_\ell) \geq 2^{h-1}R$ if $h \geq 1$. Hence, we have $2^h R \leq 2\mathrm{D}(y_i, a_\ell) + R$. It follows that

$$\begin{aligned}
\sum_{\ell=1}^{t}\sum_{h=0}^{\nu}\sum_{y_i \in Y_{\ell,h}} w(y_i)p_i \cdot 2^h R &\leq \sum_{\ell=1}^{t}\sum_{h=0}^{\nu}\sum_{y_i \in Y_{\ell,h}} w(y_i)p_i \left(2\mathrm{D}(y_i, a_\ell) + R\right) \\
&\leq \sum_{\ell=1}^{t}\sum_{y_i \in Y_\ell} w(y_i)p_i \left(2\mathrm{D}(y_i, a_\ell) + R\right) \\
&= 6\alpha \,\mathrm{cost}_k^*(V) + \frac{\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}(V, \mathcal{A}, \sigma_V)\right]}{3\alpha + 2} \\
&\leq (6\alpha + 1)\,\mathrm{cost}_k^*(V) \ ,
\end{aligned}$$

where the second inequality follows from Claim 12 and the last inequality follows from Lemma 13. Now, since $\mathrm{diam}(Y_{\ell,h}) \leq 2(2^h R)$, the above inequality also implies the second part of the claim. □

Up to now, we found a partitioning of our nodes that consists of subsets with small diameter and small distance to an arbitrary C. It remains to bound the maximal value of $\sum_{x_j \in \mathcal{X}} (p_{ij}/p_i) D(x_j, y_i)$ for each partition. The problem is that the different realizations of v_i do not necessarily lie near y_i but can be arbitrarily far away. However, if they are, then *every* assignment of v_i to a cluster center will have to pay this distance (weighted with the realization probability). Thus, we further subdivide the ring sets into subsets with the same behavior regarding $\sum_{x_j \in \mathcal{X}} (p_{ij}/p_i) D(x_j, y_i)$.

Set $\mu := \lceil \log((6\alpha + 4)W/(w_{\min} \cdot p_{\min})) \rceil$. Note that, for any node $v_i \in V$, we have

$$\begin{aligned}
2^\mu R w(v_i) p_i &\geq 2^{\lceil \log((6\alpha+4)W/(w_{\min} \cdot p_{\min})) \rceil} R \cdot w_{\min} \cdot p_{\min} \\
&\geq 2(3\alpha + 2)RW = 2 \cdot \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, \mathcal{A}, \sigma_V)\right] \\
&\geq 2 \cdot \mathrm{cost}_n^*(V) \\
&\geq \min_{\rho: V \to Y} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, Y, \rho)\right]
\end{aligned}$$

and the expected assignment cost of v_i to its approximated probabilistic 1-median y_i cannot be larger than $\min_{\rho: V \to Y} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, Y, \rho)\right]$. Thus, we have $Y_{\ell,h} = \bigcup_{a \in \{0,\dots,\mu\}} Y_{\ell,h,a}$ and this partitioning finally leads to the desired bound on the last term of our error term.

Claim 15. *If each $y_i \in Y$ is a 2-approximation of the probabilistic 1-median of v_i and $\mathcal{A}$ is an $[\alpha, \mathcal{O}(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)))]$-bicriteria approximation to $\mathrm{cost}_k^*(Y)$, then it holds that*

$$\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i \in V_{\ell,h,a}} w(v_i)p_i \max_{y_{i'} \in Y_{\ell,h,a}} \sum_{x_j \in \mathcal{X}} \frac{p_{i'j}}{p_{i'}} \cdot \mathrm{D}(x_j, y_{i'}) \leq 5\,\mathrm{cost}_k^*(V) \ .$$

Proof. It holds that

$$\begin{aligned}
&\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i \max_{y_{i'}\in Y_{\ell,h,a}} \sum_{x_j\in\mathcal{X}} \frac{p_{i'j}}{p_{i'}}\cdot \mathrm{D}(x_j,y_{i'}) \\
\leq\ &\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i \left(R + 2\sum_{x_j\in\mathcal{X}} \frac{p_{ij}}{p_i}\cdot \mathrm{D}(x_j,y_i)\right) \\
\leq\ &\left(\sum_{v_i\in V} w(v_i)p_i\right)\cdot R \ +\ 2\sum_{y_i\in Y} w(y_i)\sum_{x_j\in\mathcal{X}} p_{ij}\cdot \mathrm{D}(x_j,y_i) \\
\leq\ &\mathrm{cost}_k^*(V) \ +\ 4\,\mathrm{cost}_k^*(V)\ ,
\end{aligned}$$

where the first inequality follows from the definition of $Y_{\ell,h,a}$ and the last inequality follows from the definition of R and Claim 11. □

Bounding the Number of Possible Center Locations

Definition 16. *Let* $\Phi := \lceil \log(18(3\alpha+2)W/(w_{\min}\cdot p_{\min}\cdot\varepsilon))\rceil$*, and let* $\mathcal{U} := \bigcup_{\ell=1}^{\tau}\mathcal{B}(a_\ell, 2^\Phi R)$ *be the union of 'huge' balls around all points in* $\mathcal{A}$*. For* $\ell\in[\tau]$ *and* $h\in\{0,\dots,\Phi\}$*, let*

$$L_{\ell,h} := \begin{cases} \mathcal{B}(a_\ell, R) & h=0 \\ \mathcal{B}(a_\ell, 2^h R)\backslash\mathcal{B}(a_\ell, 2^{h-1}R) & h\geq 1 \end{cases}.$$

For $h\in\{0,\dots,\Phi\}$*, set* $r_h := 2^h\varepsilon R/(195\alpha\sqrt{d})$*. For* $\ell\in[\tau]$ *and* $h\in\{0,\dots,\Phi\}$*, consider an axis-parallel grid with side length* r_h *to partition* $L_{\ell,h}$ *into cells. Pick an arbitrary point from each grid cell in* $L_{\ell,h}$ *and store these representative points in* $\mathcal{G}_{\ell,h}$*. Finally, set* $\mathcal{G} := \bigcup_{\ell=1}^{\tau}\bigcup_{h=0}^{\Phi}\mathcal{G}_{\ell,h}$*.*

Claim 17. *We have* $\ln(|\mathcal{G}|) = \mathcal{O}(\log(k)+\log(\log(W/(w_{\min}\cdot p_{\min}\cdot\varepsilon)))+d\log(1/\varepsilon))$*.*

Proof. By a volume argument, Chen [11] showed that

$$|\mathcal{G}| \leq \tau(\Phi+1)\left(\frac{b'\alpha}{\varepsilon}\right)^d \text{ with } b' := 4\cdot 195\sqrt{\pi e}\ .$$

By the definition of τ and Φ, we get

$$\begin{aligned}
\ln(|\mathcal{G}|) &\leq \ln\left(\beta k\left(\lceil\log(18(3\alpha+2)W/(w_{\min}\cdot p_{\min}\cdot\varepsilon))\rceil+1\right)\left(\frac{b'\alpha}{\varepsilon}\right)^d\right) \\
&= \ln\left(k\cdot\mathcal{O}(\log(W/(w_{\min}\cdot p_{\min}\cdot\varepsilon)))\left(\lceil\log(18(3\alpha+2)W/(w_{\min}\cdot p_{\min}\cdot\varepsilon))\rceil+1\right)\left(\frac{b'\alpha}{\varepsilon}\right)^d\right) \\
&= \mathcal{O}(\log(k)+\log(\log(W/(w_{\min}\cdot p_{\min}\cdot\varepsilon)))+d\log(1/\varepsilon))\ .
\end{aligned}$$

□

Lemma 18. *For all sets C of at most k centers chosen from $\mathcal{G}$, it holds that*

$$\left|\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}(V,C,\rho)\right]-\min_{\rho:U\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(U,C,\rho)\right]\right|\leq\varepsilon/5\cdot\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}(V,C,\rho)\right]$$

with probability at least $1-\delta/2$.

Proof. Fix an arbitrary set $C\subseteq X$ of size at most k. Due to Lemma 9, setting $\xi:=\varepsilon/(5(12\alpha+10))$ and $\delta':=|\mathcal{G}|^{-k}\delta/(2\tau(\nu+1)(\mu+1))$, it holds that

$$\begin{aligned}&\left|\min_{\rho:V_{\ell,h,a}\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V_{\ell,h,a},C,\rho)\right]-\min_{\rho:U_{\ell,h,a}\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U_{\ell,h,a},C,\rho)\right]\right|\\ \leq\;&\frac{\varepsilon}{5(12\alpha+10)}\left(\sum_{v_i\in V_{\ell,h,a}}w(v_i)p_i\right)\left(\mathrm{D}(Y_{\ell,h,a},C)+\mathrm{diam}(Y_{\ell,h,a})+\max_{y_i\in Y_{\ell,h,a}}\sum_{x_j\in\mathcal{X}}\frac{p_{ij}}{p_i}\cdot\mathrm{D}(x_j,y_i)\right)\end{aligned}$$

with probability at least $1-\delta'$ for $\ell\in[\tau]$, $h\in\{0,\ldots,\nu\}$, and $a\in\{0,\ldots,\mu\}$. Due to the Union Bound, the above inequality is true with probability $1-\tau(\nu+1)(\mu+1)\delta'=1-|\mathcal{G}|^{-k}\delta/2$ for all $V_{\ell,h,a}$ simultaneously. Since there are at most $|\mathcal{G}|^k$ different ways to select a set C of size at most k from $\mathcal{G}$, the above inequality holds for every such set C with probability at least $1-\delta/2$. Recall that $p_{\min}$ is the smallest realization probability, $w_{\min}$ is the minimum weight of a node in V, and $W=\sum_{v_i\in V}w(v_i)p_i$ is the expected total weight of the nodes in V. The required sample size is

$$\begin{aligned}s''&=\left\lceil\xi^{-2}\ln(2/\delta')\right\rceil\\&=\left\lceil\left(\frac{12\alpha+10}{\varepsilon}\right)^2\ln\left(2\frac{\tau(\nu+1)(\mu+1)|\mathcal{G}|^k}{\delta}\right)\right\rceil\\&\leq\left\lceil c'\cdot\varepsilon^{-2}\cdot[\log(1/\delta)+k\log(|\mathcal{G}|)]\right\rceil\\&\leq\left\lceil c\cdot\varepsilon^{-2}\cdot[\log(1/\delta)+k\left(\log(k)+\log(\log(W/(w_{\min}\cdot p_{\min}\cdot\varepsilon)))+d\log(1/\varepsilon)\right)]\right\rceil=s'''\end{aligned}$$

for some sufficiently large constants c' and c. Summing the error up over all $\ell\in[\tau]$, $h\in\{0,\ldots,\nu\}$, and $a\in\{0,\ldots,\mu\}$, we get

$$\begin{aligned}&\left|\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]-\min_{\rho:U\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]\right|\\=\;&\left|\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\min_{\rho:V_{\ell,h,a}\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V_{\ell,h,a},C,\rho)\right]-\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\min_{\rho:U_{\ell,h,a}\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U_{\ell,h,a},C,\rho)\right]\right|\\\leq\;&\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\left|\min_{\rho:V_{\ell,h,a}\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V_{\ell,h,a},C,\rho)\right]-\min_{\rho:U_{\ell,h,a}\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U_{\ell,h,a},C,\rho)\right]\right|\\\leq\;&\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\frac{\varepsilon}{5(12\alpha+10)}\left(\sum_{v_i\in V_{\ell,h,a}}w(v_i)p_i\right)\left(\mathrm{D}(Y_{\ell,h,a},C)+\mathrm{diam}(Y_{\ell,h,a})+\max_{y_{i'}\in Y_{\ell,h,a}}\sum_{x_j\in\mathcal{X}}\frac{p_{i'j}}{p_{i'}}\cdot\mathrm{D}(x_j\right.\\\leq\;&\frac{\varepsilon}{5(12\alpha+10)}\left(3\cdot\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]+\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{y_i\in Y_{\ell,h}}w(y_i)p_i\,\mathrm{diam}(Y_{\ell,h})+5\,\mathrm{cost}_k^*(V)\right)\end{aligned}$$

$$\leq \frac{\varepsilon}{5(12\alpha+10)}\left(3\cdot\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]+(12\alpha+2)\,\mathrm{cost}_k^*(V)+5\,\mathrm{cost}_k^*(V)\right)$$
$$\leq \frac{\varepsilon}{5}\cdot\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right] \ ,$$

where the second last and the third last inequality follows from Claims 10, 14, and 15. □

Lemma 19. *Let $C\subseteq\mathbb{R}^d$ with a center $c\in C$ that is outside $\mathcal{U}$. Then*

$$\left|\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]-\min_{\rho:U\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]\right|\leq\varepsilon\cdot\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right] \ .$$

Proof. If there is no $v_i\in V$ which is assigned to c in the optimal clustering of V, then c can be ignored. Thus, assume that there exists a node v_i that is assigned to c, i. e., $c\in\arg\min_{c'\in C}\sum_{j=1}^m w(v_i)p_{ij}\mathrm{D}(x_j,c')$. Let a_ℓ be the center in $\mathcal{A}$ to which v_i is assigned by σ_V. Since $c\notin\mathcal{U}$, we have $\mathrm{D}(a_\ell,c)>2^\Phi R$. Hence, we get the following lower bound on the optimal clustering cost:

$$\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]\geq\sum_{j=1}^m w(v_i)p_{ij}\mathrm{D}(x_j,c)\geq\sum_{j=1}^m w(v_i)p_{ij}\left(\mathrm{D}(a_\ell,c)-\mathrm{D}(x_j,a_\ell)\right)$$
$$>w(v_i)p_i\cdot 2^\Phi R-\sum_{j=1}^m w(v_i)p_{ij}\mathrm{D}(x_j,a_\ell)\geq(18(3\alpha+2)W/\varepsilon)\cdot R-\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]$$
$$\geq\frac{18}{\varepsilon}\cdot\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]-\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]\geq\frac{17}{\varepsilon}\cdot\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right] \ .$$

We get

$$\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]<\frac{\varepsilon}{17}\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right] \ . \tag{2}$$

On the other hand, we can bound the error of our coreset as we will see in the following. First, note that for all nodes $v_{i'}$ and $v_{i''}$ from the same subset $V_{\ell,h,a}$, we have that

$$-R+\frac{1}{2}\sum_{x_j\in\mathcal{X}}\frac{p_{i''j}}{p_i''}\mathrm{D}(x_j,y_{i''})<\sum_{x_j\in\mathcal{X}}\frac{p_{i'j}}{p_i'}\mathrm{D}(x_j,y_{i'})<R+2\sum_{x_j\in\mathcal{X}}\frac{p_{i''j}}{p_i''}\mathrm{D}(x_j,y_{i''}) \ . \tag{3}$$

For $v_{i'}$ and $v_{i''}$ from the same $V_{\ell,h,a}$, note that either $h=0$ and thus $\max\{\mathrm{D}(y_{i'},a_\ell),\mathrm{D}(y_{i''},a_\ell)\}\leq R$ or $h>1$ and then $\max\{\mathrm{D}(y_{i'},a_\ell),\mathrm{D}(y_{i''},a_\ell)\}\leq 2\cdot 2^{h-1}R\leq 2\cdot\min\{\mathrm{D}(y_{i'},a_\ell),\mathrm{D}(y_{i''},a_\ell)\}$. Thus, it holds that

$$\mathrm{D}(y_{i'},y_{i''})\leq\mathrm{D}(y_{i'},a_\ell)+\mathrm{D}(y_{i''},a_\ell)\leq 3\min\{\mathrm{D}(y_{i'},a_\ell),\mathrm{D}(y_{i''},a_\ell)\}+R \ . \tag{4}$$

Next, note that

$$\begin{aligned}\min_{\rho:U\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]&=\min_{\rho:U\to C}\sum_{\ell=1}^\tau\sum_{h=0}^\nu\sum_{a=0}^\mu\sum_{v_i\in U_{\ell,h,a}}\sum_{x_j\in\mathcal{X}}w'(v_i)p_{ij}\mathrm{D}(x_j,\rho(v_i))\\&=\min_{\rho:U\to C}\sum_{\ell=1}^\tau\sum_{h=0}^\nu\sum_{a=0}^\mu\sum_{v_i\in U_{\ell,h,a}}\sum_{x_j\in\mathcal{X}}\frac{\sum_{v_{i'}\in V_{\ell,h,a}}w(v_{i'})p_{i'}}{s'''p_i}p_{ij}\mathrm{D}(x_j,\rho(v_i))\\&=\min_{\rho:U\to C}\sum_{\ell=1}^\tau\sum_{h=0}^\nu\sum_{a=0}^\mu\frac{\sum_{v_{i'}\in V_{\ell,h,a}}w(v_{i'})p_{i'}}{s'''}\sum_{v_i\in U_{\ell,h,a}}\sum_{x_j\in\mathcal{X}}\frac{p_{ij}}{p_i}\mathrm{D}(x_j,\rho(v_i)) \ .\end{aligned}$$

Thus, we can also represent our coreset in a way that each node $v_i \in U_{\ell,h,a}$ has weight $w''(v_i) = \sum_{v_{i'} \in V_{\ell,h,a}} w(v_{i'})p_{i'}/s'''$ and normalized realization probabilities p_{ij}/p_i. In this representation, the weights sum up to $\sum_{v_i \in U_{\ell,h,a}} w''(v_i) = \sum_{v_{i'} \in V_{\ell,h,a}} w(v_{i'})p_{i'}$. We want to use this in order to split the node set $V_{\ell,h,a}$ into s''' disjoint subsets of nodes $V_{\ell,h,a,1}, \dots, V_{\ell,h,a,s'''}$ such that each $v_{i'} \in U_{\ell,h,a}$ covers a subset $V_{\ell,h,a,z}$ with $z \in [s''']$, i. e., $w''(v_{i'}) = \sum_{v_i \in V_{\ell,h,a,z}} w(v_i)p_i$. The only obstacle could be that a node v_i has to be covered by two or more coreset nodes. However, then we can split v_i into two or more copies of v_i each having an adequate fraction of $w(v_i)p_i$. Let $\pi : [n] \to [n]$ be the mapping such that the corset node $v_{\pi(i)}$ covers v_i. Using this notation, we can rewrite the coreset clustering cost as

$$
\begin{aligned}
& \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\sigma)\right] \\
= \; & \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_{i'}\in U_{\ell,h,a}} w'(v_{i'})\left[\sum_{x_j\in\mathcal{X}} p_{i'j}\mathrm{D}(x_j,\sigma(v_{i'}))\right] \\
= \; & \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_{i'}\in U_{\ell,h,a}} \frac{\sum_{v_{i''}\in V_{\ell,h,a}} w(v_{i''})p_{i''}}{s'''p_{i'}}\left[\sum_{x_j\in\mathcal{X}} p_{i'j}\mathrm{D}(x_j,\sigma(v_{i'}))\right] \\
= \; & \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_{i'}\in U_{\ell,h,a}} w''(v_{i'})\left[\sum_{x_j\in\mathcal{X}} \frac{p_{i'j}}{p_{i'}}\mathrm{D}(x_j,\sigma(v_{i'}))\right] \\
= \; & \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{z}\sum_{v_i\in V_{\ell,h,a,z}} w(v_i)p_i\left[\sum_{x_j\in\mathcal{X}} \frac{p_{\pi(i)j}}{p_{\pi(i)}}\mathrm{D}(x_j,\sigma(v_{\pi(i)}))\right] \\
= \; & \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i\left[\sum_{x_j\in\mathcal{X}} \frac{p_{\pi(i)j}}{p_{\pi(i)}}\mathrm{D}(x_j,\sigma(v_{\pi(i)}))\right] \quad . \qquad (5)
\end{aligned}
$$

Now, we want to use our knowledge to bound the error of clustering with U compared to clustering with V. Let $\sigma : V \to C$ be an assignment such that $\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\sigma)\right] = \min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]$. Note that since $U \subseteq V$, we also have $\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\sigma)\right] = \min_{\rho:U\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]$. We obtain

$$
\begin{aligned}
& \left|\min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right] - \min_{\rho:U\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]\right| \\
= \; & \left|\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\sigma)\right] - \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\sigma)\right]\right| \\
= \; & \left|\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i\left[\sum_{x_j\in\mathcal{X}} \frac{p_{ij}}{p_i}\mathrm{D}(x_j,\sigma(v_i))\right] - \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i\left[\sum_{x_j\in\mathcal{X}} \frac{p_{\pi(i)j}}{p_{\pi(i)}}\mathrm{D}(x_j,\right.\right. \\
= \; & \left|\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i\sum_{x_j\in\mathcal{X}}\left[\frac{p_{ij}}{p_i}\mathrm{D}(x_j,\sigma(v_i)) - \frac{p_{\pi(i)j}}{p_{\pi(i)}}\mathrm{D}(x_j,\sigma(v_{\pi(i)}))\right]\right| \\
\leq \; & \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i\sum_{x_j\in\mathcal{X}}\left[\max\left\{\frac{p_{ij}}{p_i}\mathrm{D}(x_j,\sigma(v_i)), \frac{p_{\pi(i)j}}{p_{\pi(i)}}\mathrm{D}(x_j,\sigma(v_{\pi(i)}))\right\}\right] \quad ,
\end{aligned}
$$

where the inequality holds because both terms are positive. Let $\pi' : [n] \to [n]$ be the assignment that assigns each i to either i or $\pi(i)$ such that $\frac{p_{\pi'(i)j}}{p_{\pi'(i)}}\mathrm{D}(x_j, \sigma(v_{\pi'(i)}) = \max\{\frac{p_{ij}}{p_i}\mathrm{D}(x_j, \sigma(v_i)), \frac{p_{\pi(i)j}}{p_{\pi(i)}}\mathrm{D}(x_j, \sigma(v_{\pi(i)}))\}$ for all $i \in [n]$. Note that assigning $v_{\pi'(i)}$ to a center different from $\sigma(v_{\pi'(i)})$ can only increase the cost. This implies that

$$
\begin{aligned}
& \left| \min_{\rho: V \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(V, C, \rho)\right] - \min_{\rho: U \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_{w'}(U, C, \rho)\right] \right| \\
\leq & \sum_{\ell=1}^{\tau} \sum_{h=0}^{\nu} \sum_{a=0}^{\mu} \sum_{v_i \in V_{\ell,h,a}} w(v_i) p_i \sum_{x_j \in \mathcal{X}} \left[\frac{p_{\pi'(i)j}}{p_{\pi'(i)}} \mathrm{D}(x_j, \sigma(v_{\pi'(i)}) \right] \\
\leq & \sum_{\ell=1}^{\tau} \sum_{h=0}^{\nu} \sum_{a=0}^{\mu} \sum_{v_i \in V_{\ell,h,a}} w(v_i) p_i \sum_{x_j \in \mathcal{X}} \left[\frac{p_{\pi'(i)j}}{p_{\pi'(i)}} \mathrm{D}(x_j, \sigma(v_i)) \right] \\
\leq & \sum_{\ell=1}^{\tau} \sum_{h=0}^{\nu} \sum_{a=0}^{\mu} \sum_{v_i \in V_{\ell,h,a}} w(v_i) p_i \sum_{x_j \in \mathcal{X}} \left[\frac{p_{\pi'(i)j}}{p_{\pi'(i)}} \left(\mathrm{D}(x_j, y_{\pi'(i)}) + \mathrm{D}(y_{\pi'(i)}, y_i) + \mathrm{D}(y_i, \sigma(v_i)) \right) \right] \\
= & \sum_{\ell=1}^{\tau} \sum_{h=0}^{\nu} \sum_{a=0}^{\mu} \sum_{v_i \in V_{\ell,h,a}} w(v_i) p_i \left[\mathrm{D}(y_{\pi'(i)}, y_i) + \mathrm{D}(y_i, \sigma(v_i)) + \sum_{x_j \in \mathcal{X}} \frac{p_{\pi'(i)j}}{p_{\pi'(i)}} \mathrm{D}(x_j, y_{\pi'(i)}) \right] \\
= & \sum_{\ell=1}^{\tau} \sum_{h=0}^{\nu} \sum_{a=0}^{\mu} \sum_{v_i \in V_{\ell,h,a}} w(v_i) p_i \left[\mathrm{D}(y_{\pi'(i)}, y_i) + \sum_{x_j \in \mathcal{X}} \frac{p_{ij}}{p_i} \mathrm{D}(y_i, \sigma(v_i)) + \sum_{x_j \in \mathcal{X}} \frac{p_{\pi'(i)j}}{p_{\pi'(i)}} \mathrm{D}(x_j, y_{\pi'(i)}) \right] \\
\leq & \sum_{\ell=1}^{\tau} \sum_{h=0}^{\nu} \sum_{a=0}^{\mu} \sum_{v_i \in V_{\ell,h,a}} w(v_i) p_i \left[\mathrm{D}(y_{\pi'(i)}, y_i) + \sum_{x_j \in \mathcal{X}} \frac{p_{ij}}{p_i} \left(\mathrm{D}(y_i, x_j) + \mathrm{D}(x_j, \sigma(v_i)) \right) + \sum_{x_j \in \mathcal{X}} \frac{p_{\pi'(i)j}}{p_{\pi'(i)}} \mathrm{D}(x_j, y_\pi \right.
\end{aligned}
$$

where the third and fourth inequality follow by the triangle inequality. By Inequalities (3) and (4), this term is upper bounded by

$$
\begin{aligned}
& \sum_{\ell=1}^{\tau} \sum_{h=0}^{\nu} \sum_{a=0}^{\mu} \sum_{v_i \in V_{\ell,h,a}} w(v_i) p_i \left[\mathrm{D}(y_{\pi'(i)}, y_i) + \sum_{x_j \in \mathcal{X}} \frac{p_{ij}}{p_i} \left(\mathrm{D}(y_i, x_j) + \mathrm{D}(x_j, \sigma(v_i)) \right) + R + 2 \sum_{x_j \in \mathcal{X}} \frac{p_{ij}}{p_i} \mathrm{D}(y_i, \right. \\
= & \sum_{\ell=1}^{\tau} \sum_{h=0}^{\nu} \sum_{a=0}^{\mu} \sum_{v_i \in V_{\ell,h,a}} w(v_i) p_i \left[\mathrm{D}(y_{\pi'(i)}, y_i) + 3 \sum_{x_j \in \mathcal{X}} \frac{p_{ij}}{p_i} \mathrm{D}(y_i, x_j) + \sum_{x_j \in \mathcal{X}} \frac{p_{ij}}{p_i} \mathrm{D}(x_j, \sigma(v_i)) + R \right] \\
\leq & \sum_{\ell=1}^{\tau} \sum_{h=0}^{\nu} \sum_{a=0}^{\mu} \sum_{v_i \in V_{\ell,h,a}} w(v_i) p_i \left[3D(y_i, \sigma_Y(y_i)) + 3 \sum_{x_j \in \mathcal{X}} \frac{p_{ij}}{p_i} \mathrm{D}(y_i, x_j) + \sum_{x_j \in \mathcal{X}} \frac{p_{ij}}{p_i} \mathrm{D}(x_j, \sigma(v_i)) + 2R \right] \quad .
\end{aligned}
$$

Note that, by Claim 11, it holds

$$
\sum_{v_i \in V} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, y_i) \leq 2\, \mathrm{cost}_n^*(V) \leq 2 \cdot \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}(V, \mathcal{A}, \sigma_V)\right] \quad .
$$

By triangle inequality, this implies

$$
\sum_{v_i \in V} w(v_i) p_i \mathrm{D}(y_i, \sigma_Y(y_i)) \leq 2\, \mathrm{cost}_n^*(V) + \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(V, \mathcal{A}, \sigma_V)\right] \leq 3 \cdot \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(V, \mathcal{A}, \sigma_V)\right] \quad .
$$

By the definition of R, we also now that $RW = R\sum_{v_i \in V} w(v_i)p_i < \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, \mathcal{A}, \sigma_V)\right]/2$. Thus, we have

$$\begin{aligned}
&\left|\min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, C, \rho)\right] - \min_{\rho:U\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U, C, \rho)\right]\right| \\
\leq\ & \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i\left[3D(y_i, \sigma_Y(y_i)) + 3\sum_{x_j\in\mathcal{X}}\frac{p_{ij}}{p_i}\mathrm{D}(y_i, x_j) + \sum_{x_j\in\mathcal{X}}\frac{p_{ij}}{p_i}\mathrm{D}(x_j, \sigma(v_i)) + 2R\right] \\
\leq\ & 16\cdot\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, \mathcal{A}, \sigma_V)\right] + \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}}\sum_{x_j\in\mathcal{X}} w(v_i)p_{ij}\mathrm{D}(x_j, \sigma(v_i)) \\
\leq\ & 17\cdot\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, \mathcal{A}, \sigma_V)\right] \ .
\end{aligned}$$

Thus, we have

$$\left|\min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, C, \rho)\right] - \min_{\rho:U\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U, C, \rho)\right]\right| < 17\cdot\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, \mathcal{A}, \sigma_V)\right] \ .$$

By Inequality (2), we obtain

$$\begin{aligned}
&\left|\min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, C, \rho)\right] - \min_{\rho:U\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U, C, \rho)\right]\right| \\
<&17\cdot\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, \mathcal{A}, \sigma_V)\right] \leq \varepsilon\min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V, C, \rho)\right] \ .
\end{aligned}$$

□

Lemma 19 implies that for all C with a center c outside $\mathcal{U}$, our sampling set is already a coreset. Thus, we only have to deal with the case that $C \subseteq \mathcal{U}$. Therefore, suppose that $C = \{c_1, \ldots, c_k\} \subseteq \mathcal{U}$. Let $C' = \{c'_1, \ldots, c'_k\}$, where $c'_t \in \mathcal{G}$ is the representative point of the cell containing c_t for $t \in [k]$. In the following lemma, we bound the error introduced by C' for one specific node v_i.

Lemma 20. *Let $v_i \in V$ be a node. If $C \subseteq \mathcal{U}$ and $|C| \leq k$, then*

$$\begin{aligned}
&\left|\min_{c\in C}\sum_{x_j\in\mathcal{X}} w(v_i)p_{ij}\mathrm{D}(x_j, c) - \min_{c'\in C'}\sum_{x_j\in\mathcal{X}} w(v_i)p_{ij}\mathrm{D}(x_j, c')\right| \\
\leq\ & \frac{\varepsilon}{195\alpha}\left(w(v_i)p_iR + 2\min_{c\in C}\sum_{x_j\in\mathcal{X}} w(v_i)p_{ij}\mathrm{D}(x_j, c) + 2\sum_{x_j\in\mathcal{X}} w(v_i)p_{ij}\mathrm{D}(x_j, \sigma_V(v_i))\right) \ .
\end{aligned}$$

Proof. Let t and z be indices such that

$$c_t \in \arg\min_{c\in C}\sum_{x_j\in\mathcal{X}} w(v_i)p_{ij}\mathrm{D}(x_j, c) \quad \text{and} \quad c'_z \in \arg\min_{c'\in C'}\sum_{x_j\in\mathcal{X}} w(v_i)p_{ij}\mathrm{D}(x_j, c') \ .$$

We consider the case when $\sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, c_t) \leq \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, c'_z)$, as the other case is similar. By the triangle inequality, it holds that

$$\begin{aligned}
& \left| \min_{c \in C} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, c) - \min_{c' \in C'} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, c') \right| = \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, c'_z) - \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D} \\
\leq & \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, c'_t) - \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, c_t) \leq \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \left(\mathrm{D}(x_j, c_t) + \mathrm{D}(c_t, c'_t) - \mathrm{D}(x_j, c_t) \right) \\
= & \; w(v_i) p_i \mathrm{D}(c_t, c'_t) \ .
\end{aligned}$$

The distance $\mathrm{D}(c_t, c'_t)$ is at most the diagonal of the cell which c_t lies in, i. e., $w(v_i) p_i \mathrm{D}(c_t, c'_t) \leq w(v_i) p_i \cdot \sqrt{d} \cdot r_h$ for appropriate $h \geq 0$. If $\mathrm{D}(c_t, \mathcal{A}) \leq R$, i. e., $h = 0$, then

$$w(v_i) p_i \cdot \mathrm{D}(c_t, c'_t) \leq w(v_i) p_i \cdot \sqrt{d} \cdot \frac{1}{195} \cdot \frac{\varepsilon R}{\alpha \sqrt{d}} = \frac{\varepsilon}{195 \alpha} \cdot w(v_i) p_i R \ ,$$

which implies the required bound. Otherwise, we have $2^{h-1} R \leq \mathrm{D}(c_t, \mathcal{A}) \leq 2^h R$ for some $h \geq 1$. By the triangle inequality, we get

$$\begin{aligned}
w(v_i) p_i \cdot \mathrm{D}(c_t, c'_t) & \leq w(v_i) p_i \cdot \sqrt{d} \cdot \frac{1}{195} \cdot \frac{2^h \varepsilon R}{\alpha \sqrt{d}} = w(v_i) p_i \cdot \frac{2\varepsilon}{195\alpha} \cdot 2^{h-1} R \\
& \leq w(v_i) p_i \cdot \frac{2\varepsilon}{195\alpha} \cdot \mathrm{D}(c_t, \mathcal{A}) \leq w(v_i) p_i \cdot \frac{\varepsilon}{195\alpha} \cdot 2 \cdot \mathrm{D}(c_t, \sigma_V(v_i)) \\
& = \frac{\varepsilon}{195\alpha} \cdot 2 \cdot \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(c_t, \sigma_V(v_i)) \\
& \leq \frac{\varepsilon}{195\alpha} \cdot 2 \cdot \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \left(\mathrm{D}(x_j, c_t) + \mathrm{D}(x_j, \sigma_V(v_i)) \right) \\
& \leq \frac{\varepsilon}{195\alpha} \cdot \left(2 \min_{c \in C} \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, c) + 2 \sum_{x_j \in \mathcal{X}} w(v_i) p_{ij} \mathrm{D}(x_j, \sigma_V(v_i)) \right) \ .
\end{aligned}$$

□

Lemma 21. *If $C \subseteq \mathcal{U}$ and $|C| \leq k$, then*

$$\left| \min_{\rho : V \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(V, C, \rho) \right] - \min_{\rho : V \to C'} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(V, C', \rho) \right] \right| \leq \varepsilon \cdot \min_{\rho : V \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(V, C, \rho) \right]$$

and

$$\left| \min_{\rho : U \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(U, C, \rho) \right] - \min_{\rho : U \to C'} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(U, C', \rho) \right] \right| \leq \varepsilon \cdot \min_{\rho : V \to C} \mathbf{E}_{\mathcal{D}} \left[\mathrm{cost}_w(V, C, \rho) \right] \ .$$

Proof. Summing up the inequality of Lemma 20 over all nodes in V, we obtain

$$\begin{aligned}
&\left|\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]-\min_{\rho:V\to C'}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C',\rho)\right]\right|\\
\leq\ &\frac{\varepsilon}{195\alpha}\left(\sum_{v_i\in V}w(v_i)p_iR+2\min_{\rho:V\to C}\sum_{v_i\in V}\sum_{x_j\in\mathcal{X}}w(v_i)p_{ij}\mathrm{D}(x_j,\rho(v_i))+2\sum_{v_i\in V}\sum_{x_j\in\mathcal{X}}w(v_i)p_{ij}\mathrm{D}(x_j,\sigma_V(v_i))\right)\\
\leq\ &\frac{\varepsilon}{195\alpha}\left(WR+2\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]+2\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]\right)\\
\leq\ &\frac{\varepsilon}{195\alpha}\left(2\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]+3\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]\right)\\
\leq\ &\frac{\varepsilon}{195\alpha}\left(2\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]+(9\alpha+6)\,\mathrm{cost}_k^*(V)\right)\\
\leq\ &\frac{\varepsilon}{195\alpha}\left((9\alpha+8)\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]\right)\\
\leq\ &\varepsilon\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}(V,C,\rho)\right]\ ,
\end{aligned}$$

which proves the first part of the lemma. Now, we sum up the inequality of Lemma 20 over all weighted nodes in U. We can apply the lemma since $w'(v_i)$ is independent of j. Note that $w'(v_i)p_i=\sum_{v_{i'}\in V_{\ell,h,a}}w(v_{i'})p_{i'}/s'''$, and thus we have $\sum_{v_i\in U_{\ell,h,a}}w'(v_i)p_i\leq\sum_{v_{i'}\in V_{\ell,h,a}}w(v_{i'})p_{i'}$. Note that, in the proof of Lemma 19, we showed

$$\begin{aligned}
&\left|\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]-\min_{\rho:U\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]\right|\\
\leq\ &\max\left\{\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right],\min_{\rho:U\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]\right\}\\
=\ &\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}}w(v_i)p_i\sum_{x_j\in\mathcal{X}}\left[\max\left\{\frac{p_{ij}}{p_i}\mathrm{D}(x_j,\sigma(v_i)),\frac{p_{\pi(i)j}}{p_{\pi(i)}}\mathrm{D}(x_j,\sigma(v_{\pi(i)}))\right\}\right]\\
\leq\ &17\cdot\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]\ ,
\end{aligned}$$

so $\min_{\rho:U\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]\leq 17\cdot\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,\mathcal{A},\sigma_V)\right]\leq 17(3\alpha+2)\,\mathrm{cost}_k^*(V)$. Due to Lemma 20, we obtain

$$\begin{aligned}
&\left|\min_{\rho:U\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]-\min_{\rho:U\to C'}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C',\rho)\right]\right|\\
=\ &\left|\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in U_{\ell,h,a}}\left(\min_{c\in C}\sum_{x_j\in\mathcal{X}}w'(v_i)p_{ij}\mathrm{D}(x_j,c)-\min_{c'\in C'}\sum_{x_j\in\mathcal{X}}w'(v_i)p_{ij}\mathrm{D}(x_j,c')\right)\right|\\
\leq\ &\frac{\varepsilon}{195\alpha}\left(\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in U_{\ell,h,a}}w'(v_i)p_iR+2\min_{c\in C}\sum_{x_j\in\mathcal{X}}w'(v_i)p_{ij}\mathrm{D}(x_j,c)+2\sum_{x_j\in\mathcal{X}}w'(v_i)p_{ij}\mathrm{D}(x_j,\sigma_V(v_i))\right)\\
\leq\ &\frac{\varepsilon}{195\alpha}\left(WR+2\min_{\rho:U\to C}\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_{w'}(U,C,\rho)\right]+2\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in U_{\ell,h,a}}w'(v_i)\sum_{x_j\in\mathcal{X}}p_{ij}\mathrm{D}(x_j,\sigma_V(v_i))\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\varepsilon}{195\alpha}\left(3WR + 34(3\alpha+2)\operatorname{cost}_k^*(V) + 8\operatorname{cost}_k^*(V)\right) + (12\alpha+2)\operatorname{cost}_k^*(V)\\
&\leq \frac{\varepsilon}{195\alpha}(114\alpha+81)\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\operatorname{cost}(V,C,\rho)\right]\\
&\leq \varepsilon\min_{\rho:V\to C}\mathbf{E}_{\mathcal{D}}\left[\operatorname{cost}(V,C,\rho)\right] \quad ,
\end{aligned}
$$

where the third inequality follows from the above stated inequality and

$$
\begin{aligned}
&\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in U_{\ell,h,a}} w'(v_i)\sum_{x_j\in\mathcal{X}} p_{ij}\mathrm{D}(x_j,\sigma_V(v_i))\\
\leq\ &\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in U_{\ell,h,a}} w'(v_i)\sum_{x_j\in\mathcal{X}} p_{ij}(\mathrm{D}(x_j,y_i)+\mathrm{D}(y_i,\sigma_Y(y_i)))\\
=\ &\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in U_{\ell,h,a}} w'(v_i)\sum_{x_j\in\mathcal{X}} p_{ij}\mathrm{D}(x_j,y_i) + \sum_{v_i\in U_{\ell,h,a}} w'(v_i)p_i\mathrm{D}(y_i,\sigma_Y(y_i)))\\
\overset{\text{(5) on p. 24}}{\leq}\ &\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i\sum_{x_j\in\mathcal{X}}\frac{p_{\pi(i)j}}{p_{\pi(i)}}\mathrm{D}(x_j,y_{\pi(i)}) + \sum_{v_i\in U_{\ell,h,a}} w'(v_i)p_i2^hR\\
\overset{\text{(3) on p. 23}}{\leq}\ &\sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i\left[R+2\sum_{x_j\in\mathcal{X}}\frac{p_{ij}}{p_i}\mathrm{D}(x_j,y_i)\right] + \sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i2^hR\\
\overset{\text{Claim 11}}{\leq}\ &WR + 4\operatorname{cost}_k^*(V) + \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{a=0}^{\mu}\sum_{v_i\in V_{\ell,h,a}} w(v_i)p_i2^hR\\
\leq\ &WR + 4\operatorname{cost}_k^*(V) + \sum_{\ell=1}^{\tau}\sum_{h=0}^{\nu}\sum_{v_i\in V_{\ell,h}} w(v_i)p_i2^hR \quad .\\
\overset{\text{Claim 14}}{\leq}\ &WR + 4\operatorname{cost}_k^*(V) + (6\alpha+1)\operatorname{cost}_k^*(V) \quad .
\end{aligned}
$$

This proves the second part of the lemma. □

4.3 Running Time

We start with describing the constructions of Y, $\mathcal{A}$ and Z_i for any $i \in [n]$. For Y, we need to compute approximated 1-medians. In contrast to the k-median problem for $k > 1$, the 1-median problem does not involve deciding how to assign nodes to centers. This makes the problem much easier and we can reduce the weighted probabilistic 1-median problem to the weighted deterministic 1-median problem where the point $x_j \in \mathcal{X}$ has weight $w(v_i) \cdot p_{ij}$. Similar to Section 4.2, we assume that $w(v_i)p_{ij}/(w_{\min} \cdot p_{\min} \cdot \varepsilon) \in \mathbb{N}$ for each $v_i \in V$ and each $x_j \in \mathcal{X}$. This allows us to transfer the weighted deterministic 1-median problem to an unweighted deterministic 1-median problem by dividing all weights by $w_{\min} \cdot p_{\min} \cdot \varepsilon$ to gain integer weights and then replacing each weighted point by multiple copies of the point each having one unit weight. Now, we use a result by Kumar et al. [30].

Lemma 22 ([30]). *Let $\mathcal{X}$ be a set of points in $\mathbb{R}^d$, and let ε, $0 < \varepsilon < 1$, be a precision parameter. There exists an algorithm with running time $\mathcal{O}(2^{(1/\varepsilon)^{\mathcal{O}(1)}})$ that finds a point that is a $(1+\varepsilon)$-approximation to the deterministic Euclidean 1-median of $\mathcal{X}$ with constant probability.*

Thus, computing one 2-approximated 1-median with constant probability takes constant running time. Let $1 - \delta'$ be the failure probability of the algorithm cited in Lemma 22. If we run $(1/\delta')\log(n/\delta)$ copies of this algorithm, each with precision parameter $\varepsilon = 1$, then the probability that it simultaneously fails in all tries is bounded by $(1-\delta')^{(1/\delta')\log(n/\delta)} \leq 1 - \delta/n$. Thus, we find a 2-approximation of the probabilistic 1-median for one v_i with this success probability in time $\mathcal{O}(\log(n/\delta))$. In this manner, we can completely compute Y in $\mathcal{O}(n\log(n/\delta))$ time and gain by the Union Bound that each $y_i \in Y$ is a 2-approximation of the weighted probabilistic 1-median of v_i with probability $1 - \delta$.

Computing $\mathcal{A}$ means that we have to compute a bicritera approximation to the weighted deterministic k-median-clustering problem for the points in Y. We use a bicriteria approximation by Indyk [24] which originally deals with the metric case. Note that choosing the best centers from the input point set is a 2-approximation to the optimal solution for a deterministic Euclidean k-median problem.

Lemma 23 ([24]). *Given a weighted set Z of n points from a metric space, where the minimum weight of a point is w_Z and the total weight of the points is W_Z, and a natural number $k = \mathcal{O}(\sqrt{n})$, one can compute $\mathcal{O}(k\log(W_Z/w_Z))$ cluster centers in time $\mathcal{O}(nk\log(1/\delta)\log(\log(W_Z/w_Z)))$ such that, with probability $1-\delta$, the cost of the weighted deterministic k-median clustering of Z using these centers is within a constant factor of the optimal weighted deterministic k-median-clustering cost.*

We apply Lemma 23 using the Euclidean space as the metric space and defining Z to be the weighted set of points in Y where the weight of any $y_i \in Y$ is set to $w(v_i)p_i/(w_{\min} \cdot p_{\min} \cdot \varepsilon)$. Recall that $W = \sum_{v_i \in V} w(v_i)p_i$, $w_{\min} = \min\{\min_{v_i \in V}\{w(v_i)\}, 1\}$, and $p_{\min} = \min_{v_i \in V}\{p_i\}$. Then, the minimum weight w_Z is at least 1 and the total weight W_Z is at most $W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)$. Thus, the set $\mathcal{A}$ consists of $\mathcal{O}(k\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon))$ centers and can be computed in $\mathcal{O}(kn\log(1/\delta)\log(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon))))$ time. We can compute the assignment $\sigma_Y : Y \to \mathcal{A}$ and hence to which set $Y_{\ell,h}$ each point in Y belongs in $\mathcal{O}(kn\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)))$ time.

The computation of the Z_i is similar to the metric case (confer Section 3.2), but we have to use an algorithm for the Euclidean version of the problem. We use the following result by Chen [11].

Lemma 24 ([11]). *Given a set $\mathcal{X}$ of m integer-weighted points in $\mathbb{R}^d$ with total weight $W_{\mathcal{X}}$, a precision parameter ε, $0 < \varepsilon < 1$, and an error probability parameter δ, $0 < \delta < 1$, one can compute a weighted set Z in $\mathcal{O}(md\log(1/\delta)\log(\log(W_{\mathcal{X}})))$ time such that $|Z| = \mathcal{O}(\varepsilon^{-2}\log^2(W_{\mathcal{X}})(d\log(1/\varepsilon)+\log(\log(W_{\mathcal{X}}))+\log(1/\delta)))$ and Z is a $(1,\varepsilon)$-coreset of $\mathcal{X}$ for the weighted Euclidean deterministic 1-median-clustering problem with probability $1-\delta$.*

Let $\mathcal{X}_i$ be the subset of points $x_j \in \mathcal{X}$ with $p_{ij} > 0$. We set the weight of each point $x_j \in \mathcal{X}_i$ to $w(v_i) \cdot p_{ij}/(w_{\min} \cdot p_{\min} \cdot \varepsilon)$. Then, every point in $\mathcal{X}_i$ has an integer weight and the total weight of $\mathcal{X}_i$ is at most $W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)$. Thus, by running the algorithm from Lemma 24 with error probability parameter $\delta' := \delta/n$ on $\mathcal{X}_i$, we obtain in $\mathcal{O}(md \log(n/\delta) \log(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon))))$ time and with probability $1-\delta/n$ a $(1,\varepsilon)$-coreset of $\mathcal{X}_i$ for the weighted deterministic 1-median problem, which is a $(1,\varepsilon)$-coreset of v_i for the probabilistic 1-median problem. We can compute all coresets in this manner which gives a success probability of more than $1-\delta$ by the Union Bound. The size of each coreset is $\mathcal{O}(\varepsilon^{-2} \log^2(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon))(d \log(1/\varepsilon) + \log(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon))) + \log(n/\delta)))$.

Theorem 25. *Given a node set $V := \{v_1, \ldots, n_n\} \subset \mathbb{R}^d$, a set of associated probability distributions $\mathcal{D}_1, \ldots, \mathcal{D}_n$ over $\mathcal{X} := \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, a weight function $w : V \to \mathbb{R}_{\geq 0}$, a number $k \in \mathbb{N}$, a precision parameter ε, $0 < \varepsilon < 1$, and an error probability parameter δ, $0 < \delta < 1$, one can compute a weighted subset $U := \{u_1, \ldots, u_s\} \subset V$ and set of probability distributions $\mathcal{D}' := \{\mathcal{D}'_1, \ldots, \mathcal{D}'_s\}$ such that U and $\mathcal{D}'$ build a (k,ε)-coreset of V for the Euclidean weighted probabilistic k-median problem with probability $1-\delta$. The size of U is*

$$\mathcal{O}\left(\varepsilon^{-2} k^2 d \cdot \log^4\left(\frac{kW}{w_{\min} \cdot p_{\min} \cdot \varepsilon\delta}\right)\right),$$

and each probability distribution in $\mathcal{D}'$ assigns

$$\mathcal{O}\left(\varepsilon^{-2} d \cdot \log^3\left(\frac{nW}{w_{\min} \cdot p_{\min} \cdot \varepsilon\delta}\right)\right)$$

points from $\mathcal{X}$ a positive probability. The coreset construction has a running time of

$$\mathcal{O}\left(knm \log\left(\frac{n}{\delta}\right) \log\left(\log\left(\frac{W}{w_{\min} \cdot p_{\min} \cdot \varepsilon}\right)\right)\right)$$

and a space requirement of

$$\mathcal{O}\left(\varepsilon^{-4} k^2 d^2 \cdot \log^8\left(\frac{knW}{w_{\min} \cdot p_{\min} \cdot \varepsilon\delta}\right)\right) \quad .$$

Proof. Due to Lemma 19 and 21, using the original probability distributions in $\mathcal{D}$, U is a (k,ε)-coreset for V with probability $1-2\delta$ since Y and $\mathcal{A}$ have the desired properties with probability $1-2\delta$. Using the approximated probability distributions in $\mathcal{D}'$, it follows that for all sets $C \subseteq \mathbb{R}^d$ of size at most k, we have

$$\begin{aligned}(1-\varepsilon)^2 \cdot \min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right] &\leq \min_{\rho:U\to C} \mathbf{E}_{\mathcal{D}'}\left[\mathrm{cost}_w(U,C,\rho)\right] \\ &\leq (1+\varepsilon)^2 \cdot \min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]\end{aligned}$$

with probability $1-3\delta$. Hence, for all sets $C \subseteq \mathbb{R}^d$ of size at most k, we have

$$\left|\min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right] - \min_{\rho:U\to C} \mathbf{E}_{\mathcal{D}'}\left[\mathrm{cost}_w(U,C,\rho)\right]\right| \leq 3\varepsilon \cdot \min_{\rho:V\to C} \mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}_w(V,C,\rho)\right]$$

with probability $1 - 3\delta$. Thus, by running our coreset construction with a precision parameter $\varepsilon' \leq \varepsilon/3$ and an error probability parameter $\delta' \leq \delta/3$, U together with the probability distributions $\mathcal{D}'$ builds a (k, ε)-coreset of V for the probabilistic k-median-clustering problem with probability $1 - \delta$.

Since we obtain the set of coreset nodes U by drawing a sample of s''' nodes from each set $V_{\ell,h,a}$, where $\ell \in [\tau]$, $h \in \{0, 1, \ldots, \nu\}$, and $a \in \{0, 1, \ldots, \mu\}$, the size of U is

$$\begin{aligned} & s''' \cdot \tau(\nu+1)(\mu+1) \\ = \; & \mathcal{O}\left(\varepsilon^{-2} k \cdot \log^3\left(\frac{W}{w_{\min} \cdot p_{\min} \cdot \varepsilon}\right) \cdot \left[\log\left(\frac{1}{\delta}\right) + k\left(\log(k) + \log\left(\log\left(\frac{W}{w_{\min} \cdot p_{\min} \cdot \varepsilon}\right)\right)\right) + d \log\left(\frac{1}{\varepsilon}\right.\right.\right. \end{aligned}$$

As described above, we can compute the set of approximated 1-medians Y in time $\mathcal{O}(n \log(n/\delta))$ and $\mathcal{A}$ in time $\mathcal{O}(kn \log(1/\delta) \log(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon))))$, each with success probability $1 - \delta$. Given Y and $\mathcal{A}$, we can compute the expected assignment cost of each node $v_i \in V$ to its approximated 1-median y_i and to which set $Y_{\ell,h,a}$ each point in Y belongs in time $\mathcal{O}(nm)$. Sampling $s'''\tau(\nu+1)(\mu+1) \leq n$ nodes takes $\mathcal{O}(n)$ time. Finally, computing all Z_i with success probability $1-\delta$ takes time $\mathcal{O}(nm \log(n/\delta) \log(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon))))$. Thus, the overall running time to compute the coreset is $\mathcal{O}(n \log(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)))(m \log(n/\delta) + k \log(1/\delta)))$.

Since each realization probability can be represented by using $\mathcal{O}(\log(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)))$ bits of space, the description length of our coreset and the total space requirement of our streaming algorithm is

$$|U| \cdot \max_{i \in [n]} |\mathcal{D}'_i| \cdot \mathcal{O}\left(\log\left(\frac{W}{w_{\min} \cdot p_{\min} \cdot \varepsilon}\right)\right) = \mathcal{O}\left(\varepsilon^{-4} k^2 d^2 \cdot \log^8\left(\frac{knW}{w_{\min} \cdot p_{\min} \cdot \varepsilon\delta}\right)\right) \; .$$

□

5 The Streaming Algorithm

We show how to maintain a small (k, ε)-coreset for the probabilistic Euclidean k-median problem in the data stream model. More precisely, the set V is given as a stream of n weighted nodes in worst-case order. Each node v_i is given as a consecutive chunk in the data stream that is a sequence of up to m point-probability pairs in worst case order representing the discrete probability distribution $\mathcal{D}_i$ of the node v_i.

Typically, streaming algorithms are only allowed to perform one sequential scan over the data and to require an update time and local memory that is polylogarithmic in the size of the input stream. Let $W := \sum_{v_i \in V} w(v_i) p_i$ be the expected total weight of the nodes in V, $w_{\min} := \min\{\min_{v_i \in V}\{w(v_i)\}, 1\}$ be either the minimum weight of a node in V or 1, and $p_{\min}$ be the smallest realization probability, i. e., $p_{\min} \leq p_{ij}$ for each $i \in [n]$ and each $j \in [m]$. Then, the space requirement of our streaming algorithm is polylogarithmic in n, m, and $W/(w_{\min} \cdot p_{\min})$ and the update time per node is polylogarithmic in n, and $W/(w_{\min} \cdot p_{\min})$. Note that the update time per node might be linear in m since

the probability distribution of a node can be represented by m point-probability pairs.

Our streaming algorithm computes for each node v_i, given as a stream of up to m point-probability pairs, its approximated probability distribution $\mathcal{D}'_i$ and its approximated 1-median y_i using the streaming variant of the coreset construction for the deterministic k-median problem by Chen [11]. We will later explain how this computation is done. Let us call $(\mathcal{D}'_i, y_i)$ a *light node*. Thus, from now on, we can assume that the input is a stream of n light nodes $(\mathcal{D}'_1, y_1), \ldots, (\mathcal{D}'_n, y_n)$. To maintain a coreset of this input stream, we use the merge-and-reduce technique [7, 22]. More precisely, the light nodes are organized in a small number of coresets, each representing $2^\ell N$ light nodes (for some integer ℓ and a fixed constant N). Every time when two coresets representing the same number of light nodes exist, we take the union (merge) and create a new coreset (reduce).

The construction is based on the following observation:

Observation 26. *(i) If U_1 and U_2 are (k, ε)-coresets for disjoint sets V_1 and V_2, then $U_1 \cup U_2$ is a (k, ε)-coresets for $V_1 \cup V_2$.*

(ii) If U_1 is a (k, ε_1)-coresets for U_2 and U_2 is a (k, ε_2)-coresets for U_3, then U_1 is a $(k, (1+\varepsilon_1)(1+\varepsilon_2) - 1)$-coreset for U_3.

The idea is as follows. We maintain buckets $B_0, B_1, \ldots, B_{\lceil \log(n/N) \rceil}$. Bucket B_0 can store between 0 and N nodes. For $\ell \geq 1$, bucket B_ℓ is either empty or stores a coreset U_ℓ of N coreset nodes representing $2^{\ell-1}N$ nodes from the data stream. Next, we explain the method in detail.

All nodes in the data stream are processed in the same way. Let v_i be the i-th node read from the data stream. Then, v_i is inserted into bucket B_0. If bucket B_0 is full, then all nodes from B_0 are moved to bucket B_1. If bucket B_1 is empty, we are done. Otherwise, we compute a coreset U_2 of size N from the union of the $2N$ nodes stored in B_0 and B_1 using our coreset construction with precision parameter $\varepsilon' := \varepsilon/(2\lceil \log(n) \rceil)$ and error probability parameter $\delta_i := \delta/(\lceil \log(n) \rceil i^2)$. Afterwards, both buckets B_0 and B_1 are emptied and the N coreset nodes from U_2 are moved into bucket B_2. If B_2 is empty, we are done. Otherwise, we compute a coreset U_3 of size N from the union of the $2N$ coreset nodes stored in B_1 and B_2 using our coreset construction with precision parameter ε' and error probability parameter δ_i. Note that U_3 is a $(k, (1 + \varepsilon/(2\lceil \log(n) \rceil))^2 - 1)$-coreset for $2^2 N$ nodes from the data stream. Then, buckets B_1 and B_2 are emptied and the N coreset nodes from U_3 are moved into bucket B_3. If bucket B_3 is empty, we are done. Otherwise, we repeat this process until we reach an empty bucket. Observe that we always compute a $(k, \varepsilon/(2\lceil \log(n) \rceil))$-coreset of two merged coresets so that U_ℓ is a $(k, (1 + \varepsilon/(2\lceil \log(n) \rceil))^\ell - 1)$-coreset of size N representing $2^{\ell-1}N$ nodes from the data stream.

Theorem 27. *Given a natural number k, a precision parameter ε, $0 < \varepsilon \leq 1$, an error probability parameter δ, $0 < \delta \leq 1$, and given the nodes in V and the discrete probability distribution $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ over $\mathcal{X}$ with $|\mathcal{X}| = m$ in form*

of a data stream, there is an algorithm that computes a weighted subset $U := \{u_1, \ldots, u_s\} \subseteq V$ *and a set of probability distributions* $\mathcal{D}' := \{\mathcal{D}'_1, \ldots, \mathcal{D}'_s\}$ *such that the nodes in* U *together with the probability distributions* $\mathcal{D}'$ *build a* (k, ε)*-coreset of* V *for the probabilistic* k*-median-clustering problem with probability* $1 - \delta$*. The size of* U *is*

$$\mathcal{O}\left(\varepsilon^{-2}k^2 d \cdot \log^7\left(\frac{knW}{w_{\min} \cdot p_{\min} \cdot \varepsilon\delta}\right)\right) \ ,$$

where $W := \sum_{v_i \in V} w(v_i)p_i$*,* $w_{\min} := \min\{\min_{v_i \in V}\{w(v_i)\}, 1\}$*, and* $p_{\min} := \min_{i \in [n], j \in [m]}\{p_{ij}\}$*. Each probability distribution in* $\mathcal{D}'$ *assigns*

$$\mathcal{O}\left(\varepsilon^{-2}d \cdot \log^6\left(\frac{nmW}{w_{\min} \cdot p_{\min} \cdot \varepsilon\delta}\right)\right)$$

points from $\mathcal{X}$ *a positive probability. The streaming algorithm requires*

$$\mathcal{O}\left(\varepsilon^{-4}k^2d^2 \cdot \log^{14}\left(\frac{knmW}{w_{\min} \cdot p_{\min} \cdot \varepsilon\delta}\right)\right)$$

space and has an update time per node of

$$\mathcal{O}\left(\varepsilon^{-4}k^3d^2m \cdot \log^{14}\left(\frac{dknmW}{w_{\min} \cdot p_{\min} \cdot \varepsilon\delta}\right)\right) \ .$$

Proof. We use the merge-and-reduce technique on the stream of light nodes as described above. For each merge step, we use the coreset construction of Theorem 25. For any ℓ, U_ℓ is a (k, ε_ℓ)-coreset with

$$\begin{aligned}
\varepsilon_\ell &= (1 + \varepsilon/(2 \cdot \lceil\log(n)\rceil))^\ell - 1 \\
&\leq (1 + \varepsilon/(2 \cdot \lceil\log(n)\rceil))^{\lceil\log(n)\rceil} - 1 \\
&= 1 + \lceil\log(n)\rceil \cdot \frac{\varepsilon}{2 \cdot \lceil\log(n)\rceil} + \frac{\lceil\log(n)\rceil^2}{2} \cdot \frac{\varepsilon^2}{2^2 \cdot \lceil\log(n)\rceil^2} + \frac{\lceil\log(n)\rceil^3}{6} \cdot \frac{\varepsilon^3}{2^3 \cdot \lceil\log(n)\rceil^3} + \cdots - 1 \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{12} + \frac{\varepsilon}{48} + \cdots \leq \varepsilon \ .
\end{aligned}$$

Since the coresets U_ℓ represent disjoint sets of light nodes from the stream, the union of the coresets U_ℓ is always a (k, ε)-coreset for all the light nodes from the input stream read so far.

The error probability is given as follows: Let v_i be the i-th node read from the data stream. Then, each coreset construction that is triggered by v_i is done with an error probability parameter of $\delta_i = \delta/(\lceil\log(n)\rceil i^2)$. It follows from the Union Bound and the fact that v_i triggers at most $\lceil\log(n)\rceil$ coreset constructions that the total error probability for coreset constructions triggered by v_i is δ/i^2. Since the first node v_1 does not trigger a coreset construction, the total error probability of our streaming algorithm is at most $\sum_{i=2}^{n} \delta_i \lceil\log(n)\rceil = \sum_{i=2}^{n} \delta/i^2 \leq \delta$.

Due to Theorem 25 and since we maintain $\mathcal{O}(\log(n))$ coresets each computed with a precision parameter $\varepsilon' = \varepsilon/(2\lceil\log(n)\rceil)$ and an error probability parameter of at least $\delta' := \delta/(\lceil\log(n)\rceil n^2)$, the size of U is

$$\mathcal{O}\left(\log(n) \cdot \varepsilon'^{-2}k^2 d \cdot \log^4\left(\frac{kW}{w_{\min} \cdot p_{\min} \cdot \varepsilon'\delta'}\right)\right) = \mathcal{O}\left(\varepsilon^{-2}k^2 d \cdot \log^7\left(\frac{knW}{w_{\min} \cdot p_{\min} \cdot \varepsilon\delta}\right)\right) \ .$$

Due to Theorem 25, the update time per node is

$$\begin{aligned} & \mathcal{O}\left(T + k|U| \max_{i\in[n]} |\mathcal{D}'_i| \log\left(\frac{|U|}{\delta'}\right) \log\left(\log\left(\frac{W}{w_{\min}\cdot p_{\min}\cdot \varepsilon'}\right)\right)\right) \\ = \; & \mathcal{O}\left(T + \max_{i\in[n]} |\mathcal{D}'_i| \cdot \varepsilon^{-2}k^3 d \cdot \log^8\left(\frac{dknW}{w_{\min}\cdot p_{\min}\cdot \varepsilon\delta}\right)\right) \;\; , \end{aligned}$$

where T is the maximum time to obtain any $\mathcal{D}'_i$ from $\mathcal{D}_i$.

Next, we upper bound $\max_{i\in[n]} |\mathcal{D}'_i|$ and T. Since our streaming algorithm is allowed to use space only polylogarithmic in m, we have to use a streaming variant of the algorithm given in Lemma 24 to compute any $\mathcal{D}'_i$ from $\mathcal{D}_i$. As explained in [11], this can be done by applying the merge-and-reduce technique. The construction follows the one for computing the coreset of light nodes U from the input stream of light nodes V. This time, we maintain buckets $B_0, B_1, \ldots, B_{\lceil \log(m/M)\rceil}$. Bucket B_0 can store between 0 and M points. For $\ell \geq 1$, bucket B_ℓ is either empty or stores a coreset $\mathcal{D}'_{i,\ell}$ of M coreset points representing $2^{\ell-1}M$ points from $\mathcal{D}_i$. Let x_j be the j-th point read from the data stream representing $\mathcal{D}_i$. Each coreset construction triggered by x_j is run with precision parameter $\varepsilon' := \varepsilon/(2\lceil\log(m)\rceil)$ and error probability parameter $\delta' := \delta/(n\lceil\log(m)\rceil j^2)$. Then, $\mathcal{D}'_i$ is a $(1,\varepsilon)$-coreset of $\mathcal{D}_i$ with probability at least $1-\delta/n$. It follows by Union Bound that, for each $v_i \in V$, $\mathcal{D}'_i$ is a $(1,\varepsilon)$-coreset of $\mathcal{D}_i$ with probability at least $1-\delta$.

Due to Theorem 25 and since we maintain $\mathcal{O}(\log(m))$ coresets each computed with a precision parameter $\varepsilon' = \varepsilon/(2\lceil\log(m)\rceil)$ and an error probability parameter of at least $\delta' := \delta/(n\lceil\log(m)\rceil m^2)$, we have

$$\max_{i\in[n]} |\mathcal{D}'_i| = \mathcal{O}\left(\log(m)\cdot \varepsilon'^{-2} d \cdot \log^3\left(\frac{nW}{w_{\min}\cdot p_{\min}\cdot \varepsilon'\delta'}\right)\right) = \mathcal{O}\left(\varepsilon^{-2} d \cdot \log^6\left(\frac{nmW}{w_{\min}\cdot p_{\min}\cdot \varepsilon\delta}\right)\right) \;\; .$$

Due to Lemma 24 and since we maintain $\mathcal{O}(\log(m))$ coresets each computed with a precision parameter $\varepsilon' = \varepsilon/(2\lceil\log(m)\rceil)$ and an error probability parameter of at least $\delta' := \delta/(n\lceil\log(m)\rceil m^2)$, the update time for m point-probability pairs is

$$\begin{aligned} T &= \mathcal{O}\left(m\cdot\log(m)\cdot \max_{i\in[n]} |\mathcal{D}'_i| \cdot d \log\left(\frac{1}{\delta'}\right) \log\left(\log\left(\frac{W}{w_{\min}\cdot p_{\min}\cdot \varepsilon'}\right)\right)\right) \\ &= \mathcal{O}\left(\varepsilon^{-2} d^2 m \cdot \log^9\left(\frac{nmW}{w_{\min}\cdot p_{\min}\cdot \varepsilon\delta}\right)\right) \;\; . \end{aligned}$$

Thus, the update time per node is

$$\begin{aligned} & \mathcal{O}\left(T + \max_{i\in[n]} |\mathcal{D}'_i| \cdot \varepsilon^{-2}k^3 d \cdot \log^8\left(\frac{dknW}{w_{\min}\cdot p_{\min}\cdot \varepsilon\delta}\right)\right) \\ = \; & \mathcal{O}\left(\varepsilon^{-4}k^3 d^2 m \cdot \log^{14}\left(\frac{dknmW}{w_{\min}\cdot p_{\min}\cdot \varepsilon\delta}\right)\right) \;\; . \end{aligned}$$

Since each realization probability can be represented by using $\mathcal{O}(\log(W/(w_{\min}\cdot p_{\min}\cdot\varepsilon)))$ bits of space, the description length of our coreset and the total space

requirement of our streaming algorithm is

$$|U|\cdot\max_{i\in[n]}|\mathcal{D}_i'|\cdot\mathcal{O}\left(\log\left(\frac{W}{w_{\min}\cdot p_{\min}\cdot\varepsilon}\right)\right)=\mathcal{O}\left(\varepsilon^{-4}k^2d^2\cdot\log^{14}\left(\frac{knmW}{w_{\min}\cdot p_{\min}\cdot\varepsilon\delta}\right)\right)\ .$$

□

6 Applications

Coresets can be used to speed up approximation algorithms by computing a solution on the small coreset instead of the possibly huge original input set. For an α-approximation algorithm, this leads to an $\alpha(1+\varepsilon)$ approximation as the weighted cost of the coreset deviates from the true cost by an additive error of at most an ε-fraction of the true cost. In the metric case, the only approximation algorithm for probabilistic k-median problem known so far is the reduction to the deterministic k-median problem by Cormode and McGregor [19]. Using their algorithm, we get a constant-factor approximation. However, the computation of our coreset has roughly the same running time as the algorithm in [19].

In the Euclidean case, no approximation algorithm is known so far. By using our coreset, we get a randomized $(1+\varepsilon)$-approximation in superpolynomial time.

Lemma 28. *There exists a randomized $(1+\varepsilon)$-approximation algorithm for the Euclidean probabilistic k-median problem that works with high constant probability and has a running time of $\mathcal{O}(nm\log(n)\log(\log(W/(w_{\min}\cdot p_{\min})))+(W/(w_{\min}\cdot p_{\min}))^{\log^5(W/(w_{\min}\cdot p_{\min}))})$ for constant k, d, and ε.*

Proof. First, we call our coreset construction for the Euclidean probabilistic k-median problem. Note that computing the best k centers for the weighted coreset are a $(1+\varepsilon)$-approximation for the original problem. Thus, we have to solve the weighted probabilistic k-median problem on our coreset. The k centers can come from the infinite set $\mathbb{R}^d$. Hence, we cannot iterate through all possible center locations. Note that we cannot use the set $\mathcal{G}$ from Definition 16 for this purpose neither because we only showed that for centers outside $\mathcal{U}$ our coreset *is a coreset* but did *not* show that centers outside $\mathcal{U}$ can be ignored.

Thus, we approach the problem from a different angle. In an optimal solution, every node is assigned to one of the k optimal centers. This induces a partition. In each partition, the center to which the nodes are assigned, is the 1-median of the assigned nodes (otherweise using the 1-median would improve the solution). We use this in the following way: We iterate over all $\mathcal{O}(|U|^k)$ possibilities to subdivide our coreset into k partitions. For each partition, we compute a $(1+\varepsilon)$-approximation. Finally, we output the approximated 1-medians of the partition with lowest clustering cost.

This algorithm computes a $(1+\varepsilon)$-approximation of the optimal solution for our weighted coreset: As we test all partitions, we in particular test the optimal partition. By computing $(1+\varepsilon)$-approximations of the 1-medians, we compute a $(1+\varepsilon)$-approximation for this partition and thus for the optimal solution. Note that this provides a $(1+\varepsilon)^2$-approximation for the original input.

Hence, running the algorithm with parameter $\varepsilon' = \varepsilon/3$ leads to the desired approximation guarantee.

To implement the approach, we use the algorithm by Kumar et al. [30], similarly as described below Lemma 22. As we want a $(1+\varepsilon)$-approximation, computing one approximated 1-median with constant error probability takes time $\mathcal{O}(2^{(1/\varepsilon)^{\mathcal{O}(1)}})$, and computing all k 1-medians with error probability δ takes time $\mathcal{O}(k \log(k/\delta) \cdot 2^{(1/\varepsilon)^{\mathcal{O}(1)}})$. The running time for this approximation algorithm is

$$\begin{aligned}
&\mathcal{O}\left(knm \log\left(n/\delta\right) \log\left(\log\left(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon)\right)\right)\right) && \text{coreset construction}\\
&+\mathcal{O}(k^{\left(\varepsilon^{-2} d \cdot \log^6(W/(w_{\min} \cdot p_{\min} \cdot \varepsilon \delta))\right)}) && \text{number of partitionings}\\
&\cdot \mathcal{O}(k \cdot k \log(k/\delta) \cdot 2^{(1/\varepsilon)^{\mathcal{O}(1)}}) && \text{approximated 1-median computation}
\end{aligned}$$

which is $\mathcal{O}(nm \log(n) \log(\log(W/(w_{\min} \cdot p_{\min}))) + (W/(w_{\min} \cdot p_{\min}))^{\log^5(W/(w_{\min} \cdot p_{\min}))})$ for constant k, ε, δ, and d. □

Both coreset constructions can also be used to speed up a slow approximation algorithm, e. g., an algorithm with a running time exponential in the expected total weight W or in the term $W/(w_{\min} p_{\min})$, as soon as such an algorithm exists. This leads to a polynomial-time approximation algorithm since the size of our coreset is only polylogarithmic in $W/(w_{\min} p_{\min})$ and it is beneficial if the algorithm has a better approximation guarantee (in the metric case) or lesser dependence on W, ε, k or d (in the Euclidean case).

Finally, the coreset construction can be beneficial in itself as it enables queries to the clustering cost in a setting where data comes in a stream and can only be stored in sketches.

Acknowledgments

The authors thank the anonymous referees for their detailed and useful comments, especially for suggesting to try to extend Lemma 4 to the non-uniform case.

References

[1] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Approximating extent measures of points. *Journal of the ACM*, 51(4):606–635, 2004.

[2] C. C. Aggarwal and P. S. Yu. A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 21(5):609–623, 2009.

[3] S. Arora. Polynomial time approximation schemes for euclidean traveling salesman and other geometric problems. *Journal of the ACM*, 45(5):753–782, 1998.

[4] S. Arora, P. Raghavan, and S. Rao. Approximation schemes for euclidean -medians and related problems. In *STOC*, pages 106–113, 1998.

[5] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004.

[6] M. Bădoiu, S. Har-Peled, and P. Indyk. Approximate clustering via coresets. In *Proceedings of the 34th STOC*, pages 250–257, 2002.

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

[8] M. Charikar and S. Guha. Improved combinatorial algorithms for facility location problems. *SIAM Journal on Computing*, 34(4):803–824, 2005.

[9] M. Charikar, S. Guha, É. Tardos, and D. B. Shmoys. A constant-factor approximation algorithm for the k-median problem. *J. Comput. Syst. Sci.*, 65(1):129–149, 2002.

[10] M. Chau, R. Cheng, B. Kao, and J. Ng. Uncertain data mining: An example in clustering location data. In *Proceedings of the 10th Pacific-Asia Conference (PAKDD)*, pages 199–204, 2006.

[11] K. Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

[12] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In *Proceedings of the 27th PODS*, pages 191–200, 2008.

[13] J. Edmonds and R. M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM*, 19(2):248–264, 1972.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996.

[15] D. Feldman, M. Monemizadeh, and C. Sohler. A ptas for k-means clustering based on weak coresets. In *Proceedings of the 23rd Symposium on Computational Geometry (SoCG)*, pages 11–18, 2007.

[16] E. Forgey. Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*, 768(21), 1965.

[17] G. Frahling and C. Sohler. Coresets in dynamic geometric data streams. In *Proceedings of the 37th STOC*, pages 209–217, 2005.

[18] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):515–528, 2003.

[19] S. Guha and K. Munagala. Exceeding expectations and clustering uncertain data. In *Proceedings of the 28th PODS*, pages 269–278, 2009.

[20] S. Günnemann, H. Kremer, and T. Seidl. Subspace clustering for uncertain data. In *Proceedings of the SIAM International Conference on Data Mining*, pages 385–396, 2010.

[21] S. Har-Peled and A. Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

[22] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th STOC*, pages 291–300, 2004.

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

[24] P. Indyk. Sublinear time algorithms for metric space problems. In *Proceedings of the 31st STOC*, pages 428–434, 1999.

[25] K. Jain, M. Mahdian, and A. Saberi. A new greedy approach for facility location problems. In *Proceedings of the 34th STOC*, pages 731–740, 2002.

[26] K. Jain and V. V. Vazirani. Primal-dual approximation algorithms for metric facility location and k-median problems. In *Proceedings of the 40th FOCS*, pages 2–13, 1999.

[27] S. G. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. *SIAM Journal on Computing*, 37(3):757–782, 2007.

[28] H.-P. Kriegel and M. Pfeifle. Density-based clustering of uncertain data. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 672–677, 2005.

[29] H.-P. Kriegel and M. Pfeifle. Hierarchical density-based clustering of uncertain data. In *IEEE International Conference on Data Mining (ICDM)*, pages 689–692, 2005.

[30] A. Kumar, Y. Sabharwal, and S. Sen. Linear-time approximation schemes for clustering problems in any dimensions. *Journal of the ACM*, 57(2), 2010.

[31] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–136, 1982.

[32] R. R. Mettu and C. G. Plaxton. Optimal time bounds for approximate clustering. *Machine Learning*, 56(1-3):35–60, 2004.

[33] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip. Efficient clustering of uncertain data. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)*, pages 436–445, 2006.

[34] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the 6th International Conference on Computer Vision (ICCV)*, pages 59–66, 1998.

[35] H. Xu and G. Li. Density-based probabilistic clustering of uncertain data. In *International Conference on Computer Science and Software Engineering, CSSE (4)*, pages 474–477, 2008.