Universal Weak Coreset

Ragesh Jaiswal*, **Amit Kumar***

Department of Computer Science and Engineering
Indian Institute of Technology Delhi
{rjaiswal, amitk}@cse.iitd.ac.in

Abstract

Coresets for k-means and k-median problems yield a small summary of the data, which preserves the clustering cost with respect to any set of k centers. Recently coresets have also been constructed for constrained k-means and k-median problems. However, the notion of coresets has the drawback that (i) they can only be applied in settings where the input points are allowed to have weights, and (ii) in general metric spaces, the size of the coresets can depend logarithmically on the number of points. The notion of *weak coresets*, which has less stringent requirements than coresets, has been studied in the context of classical k-means and k-median problems. A weak coreset is a pair (J, S) of subsets of points, where S acts as a summary of the point set and J as a set of potential centers. This pair satisfies the properties that (i) S is a good summary of the data as long as the k centers are chosen from J only, and (ii) there is a good choice of k centers in J with a cost close to the optimal cost. We develop this framework, which we call *universal weak coresets*, for constrained clustering settings. In conjunction with recent coreset constructions for constrained settings, our designs give greater data compression, are conceptually simpler, and apply to a wide range of constrained k-median and k-means problems.

Introduction

Center-based clustering problems such as k-MEDIAN and the k-MEANS are important data processing tasks. Given a set of center locations $F \subset \mathcal{X}$, a set $X \subset \mathcal{X}$ of n points in a metric $(\mathcal{X}, D)$, and a parameter k, the goal here is to partition the set of points into k *clusters*, say $X_1, \ldots, X_k$, and assign the points in each cluster to a corresponding *cluster center*, say $c_1, \ldots, c_k \in F$ respectively, such that the objective $\sum_{i=1}^{k} \sum_{x \in X_i} D(x, c_i)^z$ is minimized. Here, z is a parameter which is 1 for k-MEDIAN and 2 for k-MEANS.[1] In the past decade, there has been significant effort in designing coresets for such settings. Given a k-MEDIAN or k-MEANS clustering instance as above, a coreset with parameter ε is a weighted subset S of points in the metric space with the following property: for every set C of k points in the metric space, the assignment cost of X to C is within $(1 \pm \varepsilon)$ of that of S. More formally, let $w(x)$ denote the weight of a point $x \in S$, and for a point $x \in X$, let $D(x, C)$ be the distance between x and the closest point in C. Then the following condition is satisfied for every subset C of k points (where $z = 1$ or $z = 2$ depending on the clustering problem being considered):

$$(1-\varepsilon) \sum_{x \in S} w(x) D(x, C)^z \leq \sum_{x \in X} D(x, C)^z \leq (1+\varepsilon) \sum_{x \in S} w(x) D(x, C)^z \quad (1)$$

Coresets are useful for several reasons: (i) There are efficient algorithms for constructing small-sized coresets. Hence, some of the fastest known algorithms for k-MEANS and k-MEDIAN problems proceed in a two step fashion: first, find a succinct coreset, and then run a less efficient algorithm on the coreset; (ii) in streaming settings, where one cannot afford to store the entire dataset, a coreset provides a summary of the data without compromising on the quality of clustering. Further, it is well known that coresets from two distinct data sets can be composed to yield a new coreset for the union of these two datasets. Hence, coresets are amenable to settings where data arrives over time; (iii) in scenarios where the set of k centers may change over time, a coreset represents an efficient way of computing the clustering cost.

For most applications, the requirements of a coreset may seem too strong. Indeed, a less stringent notion of *weak coreset* was defined by (Feldman, Monemizadeh, and Sohler 2007). A weak coreset, with a parameter ε, for a point set X as above is a pair (J, S) of subsets of points in the metric space, with S being a weighted subset of points, such that (i) the condition (1) is satisfied for all subsets C, where $|C| = k$ and $C \subseteq J$; and (ii) there is a subset C of k centers in J such that the assignment cost of X to C is within $(1 + \varepsilon)$ of the optimal clustering cost of X. The motivation for defining a weak coreset is that one could obtain weak coresets with better guarantees than a coreset. Indeed, this shall be the case for the problems considered in this work.

To understand why weak coresets may have better guarantees than coresets, we briefly discuss coreset construction techniques. Typical constructions use random sampling-based ideas. One starts with an initial set of $O(k)$ centers ob-

*These authors contributed equally.

[1] Note that even though our results are stated for $z = 1$ and $z = 2$, they also hold for a general value of z.

tained by a fast approximation algorithm. For each of these centers $c \in C$, we partition the data into "rings" of geometrically increasing size around c. From each of these rings, one samples $\mathrm{poly}(\frac{k}{\varepsilon})$ points and appropriately assigns them weights – these weighted sampled points "represent" the points in the ring as far as c is concerned, i.e., their assignment cost to c is very close to that of the original set of points in the ring with high probability. These sampled points form the desired coreset. However, for the coreset property to hold, these sampled points must have near-optimal assignment cost for *every* set of k centers. Since there are about n^k possibilities for the choice of k centers, we need to sample $\big(\mathrm{poly}(\frac{k}{\varepsilon}) \cdot \log n\big)$ points from each ring to ensure the coreset property. In geometric settings, concepts such as an ε-net and ε-centroid set have been used to reduce the coreset size. However, in general metric spaces, there are lower bounds (see (Baker et al. 2020a; Cohen-Addad et al. 2022)), suggesting that the size of the coreset will have a dependency on $\log n$.

Weak coresets allow us to remove the dependency on $\log n$ even in general metric spaces. Since the near-optimal clustering guarantees need to hold with respect to k centers chosen from J only, the set of such possibilities reduces to $|J|^k$. Thus, a small-sized J would typically imply a small-sized sample S as well. Further, weak coresets allow us to maintain a near-optimal clustering in streaming setting. Indeed, the sets J and S can be constructed in a streaming setting. Since the set of k centers needs to be selected from J only, and each can be tested with respect to S, we can also maintain a set of near-optimal k centers in a streaming setting.

So far, our discussion has focused on the classical k-MEDIAN and k-MEANS settings. However, there has been significant recent activity in the more general class of *constrained* clustering problems. A constrained clustering problem specifies additional conditions on a feasible partitioning of the input points into k clusters. For example, the r-*gathering* problem requires that each cluster in a feasible partitioning must contain at least r data points. Similarly, the well-known *capacitated* clustering problem specifies an upper bound on the size of each cluster. Constrained clustering formulations can also capture various types of *fairness* constraints: each data point has a *label* assigned to it, and we may require upper or lower bounds on the number (or fraction) of points with a certain label in each cluster. Some of these constrained problems are discussed in the Applications section.

Coresets for constrained clustering settings were recently constructed by (Bandyapadhyay, Fomin, and Simonov 2021; Braverman et al. 2022). Note that the standard notion of coreset is meant to preserve the cost of an assignment where points get assigned to the closest center. This prevents using standard coresets in constrained clustering settings where a point may not necessarily get assigned to its closest center. Recent work (Bandyapadhyay, Fomin, and Simonov 2021; Braverman et al. 2022) design "assignment-preserving" coresets that allow their use in constrained settings. In this work, we generalize the notion of weak coresets to *universal weak coresets* for constrained clustering settings. The underlying idea is the same as that of a weak coreset, i.e., we need a weighted subset S of points along with a set J of potential center locations. But now, this pair has the same guarantees as a weak coreset for *any* constrained clustering problem. This universal guarantee has a feature that we need not know in advance the actual constrained clustering problem being solved.

The notion of a universal weak coreset also has the following subtle application. In some specific settings, there is a distinction between known algorithms for weighted and unweighted settings. More specifically, there exist constrained clustering problems, where even if we are given a small-sized set S of points, efficient algorithms for a near-optimal set of k centers with respect to S are known only if the point set S is unweighted. For example, a recent development (Chakraborty, Das, and Krauthgamer 2023) for the k-MEDIAN problem in the Ulam metric has broken the 2-approximation barrier. However, their $(2-\delta)$-approximation algorithm works only on unweighted input permutations. In such settings, we may not be able to efficiently find a good set of centers even if S is a coreset. However, when given a weak coreset (J, S), we know that we need to look for centers that are subsets of J only, and we can use the cost preservation property of the weighted set S to find good centers from J. This allows us to efficiently handle such constrained clustering problems as well.

Breaking the coreset $\log n$ barrier Since it is known (Baker et al. 2020a) that the $\log n$ factor in the size of a coreset is unavoidable in general metric spaces, we must relax the notion of a coreset to break the $\log n$ barrier. Our notion of a universal weak coreset provides a framework for an appropriate relaxation that allows us to break the $\log n$ barrier. More specifically, we relax the condition on the set J to: there exists a subset C of k centers in J such that the assignment cost of X to C is within $(\alpha+\varepsilon)$ of the optimal clustering cost of X, where α is allowed to be > 1. Moreover, the *universal* property on J says that this $(\alpha+\varepsilon)$-approximation holds with respect to *any* target clustering (not only the optimal Voronoi partitioning). The property on the set S remains unchanged. We call this an α-universal weak coreset. Note that a α-universal weak coreset helps to find an α-approximate solution. The relaxation from $(1+\varepsilon)$ to $(\alpha+\varepsilon)$ guarantee is not a significant compromise if α is the best approximation guarantee known for a constrained clustering problem, which is indeed true for several constrained problems we discuss in this paper. On the other hand, this relaxation allows the universal weak coreset size, $(|J| + |S|)$, to be $\mathrm{poly}(\frac{k}{\varepsilon})$, i.e., independent of n. Our main results include constructions of such universal weak coresets:

Informal result: *There is a* 3*-universal weak coreset for the* k-MEDIAN *and a* 9*-universal weak coreset for the* k-MEANS *problem in general metric spaces (the* $3, 9$ *factors improve to* $2, 4$ *for the special case when* $X \subseteq F$*). Further, there is a* 1*-universal weak coreset construction for* k-MEDIAN/k-MEANS *in the Euclidean setting. All these weak coresets have* $\mathrm{poly}(\frac{k}{\varepsilon})$ *size.*

Since the above-mentioned coresets work for constrained settings, we can use an α-universal weak coreset to obtain an

α-approximate solution for an arbitrary version of the constrained clustering problem. These include balanced clustering, fair clustering, l-diversity clustering, and potentially many more.

Related work Two decades ago, coresets were introduced (Har-Peled and Mazumdar 2004) primarily as a tool to design streaming algorithms for the k-MEDIAN/k-MEANS problems. Subsequently, it became an independent computational object of study, and some remarkable techniques and results (Har-Peled and Kushal 2007; Chen 2009; Feldman, Monemizadeh, and Sohler 2007; Langberg and Schulman 2010; Feldman and Langberg 2011; Feldman, Schmidt, and Sohler 2020) have been obtained. More recent developments (Sohler and Woodruff 2018; Huang et al. 2018; Becchetti et al. 2019; Huang and Vishnoi 2020; Baker et al. 2020b; Braverman et al. 2021; Cohen-Addad, Saulpic, and Schwiegelshohn 2021) have focused on improvements on the size of coresets in various metrics. Recent developments have also been on coresets for constrained settings (Huang, Jiang, and Vishnoi 2019; Schmidt, Schwiegelshohn, and Sohler 2020; Bandyapadhyay, Fomin, and Simonov 2021; Braverman et al. 2022), with (Bandyapadhyay, Fomin, and Simonov 2021; Braverman et al. 2022) being most relevant to our work. The idea of designing a coreset that works for more than one problem has also been explored in a recent work (Maalouf et al. 2023).

Organization In the next section, we define the notion of a universal weak coreset. In the subsequent section, we give constructions of such coresets. Finally, in the Applications section, we describe applications of universal weak coresets in finding approximate solutions to several constrained clustering problems.

Universal Weak Coreset

We define the notion of universal weak coreset formally in this section. We shall use $[k]$ to denote the set $\{1, ..., k\}$. In the discussion, 'with high probability' should be interpreted as 'with probability at least 0.99'. Let $\mathcal{X}$ denote a metric space with metric D defined on it. We now formally define a constrained clustering problem. While describing an instance $\mathcal{I}$, we would like to separate the actual constraints on feasible clusterings and the underlying clustering instance. A clustering instance $\mathcal{I}'$ is given by a tuple (X, F, w, k), where X is the set of all input points with a corresponding weight function $w : X \to \mathbb{R}^+$, a set F of potential center locations and a value k, which denotes the number of clusters.

A constrained clustering instance consists of a tuple (X, F, w, k) as above and a k-tuple $\Gamma = (t_1, \ldots, t_k)$ of non-negative real values such that $\sum_{i\in[k]} t_i = \sum_{x\in X} w(x)$. Intuitively, the value t_i denotes the total weight of the points assigned to the i^{th} cluster. However, a point in X can be partially assigned to several clusters, but the sum of these partial weight assignments should equal $w(x)$. In other words, an assignment is given by a mapping $\sigma : X \times [k] \to \mathbb{R}^+$, such that $\sum_{i\in[k]} \sigma(x, i) = w(x)$ for each $x \in X$. An assignment σ is said to be *consistent* with $\Gamma = (t_1, \ldots, t_k)$, denoted $\sigma \sim \Gamma$, if $\sum_{x\in X} \sigma(x, i) = t_i$ for all $i \in [k]$. Thus, the k-tuple Γ denotes how the weights of the points in X get partitioned into the k clusters. Given an instance $\mathcal{I} = ((X, F, w, k), \Gamma)$ of constrained clustering, and an ordered set $C \subseteq F$ of k centers, the clustering cost, denoted $\mathsf{cost}_z(X, w, C, \Gamma)$, where $z = 1$ or 2, is defined as follows (here $C = (c_1, \ldots, c_k)$):

$$\mathsf{cost}_z(X, w, C, \Gamma) \equiv \min_{\sigma\sim\Gamma} \left\{ \sum_{i=1}^{k} \sum_{x\in X} \sigma(x, i) \cdot D(x, c_i)^z \right\}.$$

Now, the optimal cost of clustering over the choice of centers C is denoted as follows:

$$\mathsf{opt}_z(X, w, \Gamma) \equiv \min_{C:C\subseteq F,|C|=k} \{\mathsf{cost}_z(X, w, C, \Gamma)\}.$$

We are now ready to define the notion of weak coresets. The parameter z shall be either 1 or 2 in the following. We shall also fix a parameter $\varepsilon > 0$ for the rest of the discussion. This should be treated as an arbitrarily small but positive constant.

Definition 1 (α-Universal Weak Coreset)**.** *Given a clustering instance $\mathcal{I} = (X, F, w, k)$, an α-universal weak coreset is a tuple (J, S, v), where $J \subset F$ is a subset of potential center locations, and $S \subset X$ is a weighted subset of points with weight function $v : S \to \mathbb{R}^+$ such that for any assignment $\sigma : X \times [k] \to \mathbb{R}^+$: the following conditions hold with high probability:*

(A) J contains a subset $(c_1, ..., c_k)$ with

$$\sum_{i=1}^{k} \sum_{x\in X} \sigma(x, i) \cdot D(x, c_i)^z \leq (\alpha + \varepsilon) \cdot \sum_{i=1}^{k} \sum_{x\in X} \sigma(x, i) \cdot D(x, c_i^*)^z.$$

where $(c_1^*, ..., c_k^*)$ is the optimal center set that respects σ, i.e., $(c_1^*, ..., c_k^*) = \arg\min_{(s_1,...,s_k)} \left\{ \sum_{i=1}^{k} \sum_{x\in X} \sigma(x, i) \cdot D(x, s_i)^z \right\}$

(B) For every subset $C \subseteq J$, $|C| = k$ and every Γ:

$$\mathsf{cost}_z(X, w, C, \Gamma) \in (1 \pm \varepsilon) \cdot \mathsf{cost}_z(S, v, C, \Gamma).$$

The size of a weak coreset (J, S, v) is defined as $(|J| + |S|)$.

An α-universal weak coreset allows us to summarise the dataset so that this summary is sufficient to obtain an $(\alpha+\varepsilon)$-approximate solution to any constrained version of the clustering problem in time that is dependent only on the size $(|J| + |S|)$ of the coreset. This could lead to fast approximation algorithms if the universal coreset construction is efficient and its size is independent of the data size, namely $n = |X| + |F|$.[2] In the next section, we shall see that this is indeed possible. Let us see a canonical approximation algorithm that finds an $(\alpha + \varepsilon)$-approximate solution from an α-universal weak coreset.

[2]It is sufficient to consider the k nearest facility locations for every point in X. This takes care of scenarios where F is infinite.

Theorem 1. *Consider a clustering instance (X, F, w, k) and let (J, S, v) be an α-universal weak coreset for it. Given a constrained clustering instance $((X, F, w, k), \Gamma)$, there is an algorithm $\mathcal{A}$ that, with high probability, outputs a set of k centers $C \subseteq F$ such that:*

$$\mathsf{cost}_z(X, w, C, \Gamma) \leq (\alpha + \varepsilon) \cdot \mathsf{opt}_z(X, w, \Gamma).$$

Moreover, the running time of $\mathcal{A}$ is $\tilde{O}(|J|^k \cdot |S|)$.

Proof. The algorithm tries out all ordered subsets $C := (c_1, \ldots, c_k)$ of size k of J. For each such subset, one can find an assignment $\sigma : S \times [k] \to \mathbb{R}^+$ that is consistent with Γ and minimizes $\mathsf{cost}_z(S, v, C, \Gamma)$. This can be done by setting up a suitable min-cost flow network. Thus, we can efficiently compute $\mathsf{cost}_z(X, w, C, \Gamma)$. Finally, we output the subset C for which $\mathsf{cost}_z(X, w, C, \Gamma)$ is minimized. The desired result follows easily from the properties of a universal weak coreset. □

Note that if the coreset construction is efficient (*i.e., polynomial in* $n, k, 1/\varepsilon$) and the coreset size is $f(k, \varepsilon)$, for some function f, then the above theorem gives an FPT (*Fixed-Parameter Tractable*) approximation algorithm with parameter k. This means that as long as k is a fixed constant, the algorithm runs in polynomial time. We now give efficient constructions of universal weak coresets.

Universal Weak Coreset Construction

As described in the previous section, a universal weak coreset is a pair (J, S) of sets. The set J comes with the guarantee that for any constrained clustering instance, J contains a near-optimal center set for that clustering with high probability. Our construction of the set J uses the results of (Goyal, Jaiswal, and Kumar 2020). The main idea is to use D^2-sampling (with respect to a center set that gives constant factor approximation for the unconstrained clustering problem on the dataset) to sample sufficiently many points from the dataset. This ensures that each cluster has a good representation in the sample. This is sufficient to guarantee an approximation guarantee since a uniformly sampled point from every cluster will be a good center in expectation. The set S, on the other hand, is constructed using the algorithm of (Braverman et al. 2022). The construction works by partitioning the points into $\mathrm{poly}(k/\varepsilon)$ "rings" based on the distance of the points from the nearest center in a good k-center set for the unconstrained problem. It uniformly samples from rings with abundant points and readjusts the weight of the chosen points. The sparse rings are combined into groups, and two weighted points representing the entire group are given for each such group. More details about the construction of these sets are given in the subsequent discussions.

We now give an algorithm for constructing coresets. Recall that there are two sets in the definition of a universal weak coreset: J and S. The set S represents the input points that need to be clustered, whereas the set J represents the potential center locations. We will construct these two sets independently using two known lines of results.

Constructing J: Let us first see the construction of the set J that follows from developments on D^z-sampling based algorithms for the *list-k-median/means* problems (Goyal, Jaiswal, and Kumar 2020; Bhattacharya et al. 2020) – the idea of list k-MEDIAN or k-MEANS gave a unified way of handling a large class of constrained clustering problems.

The following problem is addressed by (Goyal, Jaiswal, and Kumar 2020; Bhattacharya et al. 2020). Given a clustering instance (X, F, w, k), and a parameter $\varepsilon > 0$, output a list $\mathcal{L} = \{C_1, \ldots, C_\ell\}$, where each $C_i \subseteq F$ is an ordered set of k centers such that the following property is satisfied: for any partition $P_1, \ldots, P_k$ of the point set, there exists a set of k centers $C = (c'_1, \ldots, c'_k) \in \mathcal{L}$ such that

$$\sum_{i \in [k]} \sum_{x \in P_i} D(x, c'_i)^z \leq (\alpha + \varepsilon) \sum_{i \in [k]} \sum_{x \in P_i} D(x, c^*_i)^z,$$

where c^*_i is the optimal center for P_i. The goal is to minimize the size ℓ of $\mathcal{L}$ (the above property needs to hold with high probability). To solve this problem, (Goyal, Jaiswal, and Kumar 2020; Bhattacharya et al. 2020) find a suitable set $M \subseteq F$ (using a D^z-sampling technique) and then iterate over all subsets of size k of M to generate the list $\mathcal{L}$. We state the relevant result from (Goyal, Jaiswal, and Kumar 2020) that we shall use to construct the set J.[3]

Theorem 2 ((Goyal, Jaiswal, and Kumar 2020))**.** *There is a randomised algorithm, that outputs a set $M \subset F$ of size $O\left(\mathrm{poly}(\frac{k}{\varepsilon})\right)$ with the following property: For any assignment $\sigma : X \times [k] \to \mathbb{R}^+$, with high probability, there is a set of k centers $C := \{c_1, \ldots, c_k\} \in M$ such that:*

$$\sum_{i=1}^{k} \sum_{x \in X} \sigma(x, i) \cdot D(x, c_i)^z \leq (3^z + \varepsilon) \cdot \sum_{i=1}^{k} \sum_{x \in X} \sigma(x, i) \cdot D(x, c^*_i)^z,$$

where $(c^*_1, ..., c^*_k)$ is the optimal set of centers that respect σ, i.e., $(c^*_1, ..., c^*_k) = \arg\min_{(s_1, \ldots, s_k)} \sum_{i=1}^{k} \sum_{x \in X} \sigma(x, i) \cdot D(x, s_i)^z$. The running time of this algorithm is $O(n|M|)$.

It is not difficult to see that the set M in the above theorem is precisely the set J that we need for a 3^z-universal weak coreset. For the special case of $X \subseteq F$ (i.e., a center can be located at any of the input points), (Goyal, Jaiswal, and Kumar 2020) gave an improved guarantee of $(2^z + \varepsilon)$ instead of $(3^z + \varepsilon)$. So, the same improvement transfers to the universal weak coreset. For the Euclidean metric, (Bhattacharya et al. 2020) used sampling ideas similar to (Goyal, Jaiswal, and Kumar 2020) to give a result similar to Theorem 2. However, the approximation guarantee here is $(1+\varepsilon)$ and the size of M is $(\frac{k}{\varepsilon})^{O(\frac{1}{\varepsilon})}$. This gives a 1-universal weak coreset property for the set J in the Euclidean setting.

[3]Note that the result in this particular form is not explicitly stated in (Goyal, Jaiswal, and Kumar 2020) since this was not the primary goal of that work. In particular, the result stated here is a weighted version of the results in (Goyal, Jaiswal, and Kumar 2020). However, it follows from their analysis.

Constructing the set S: Now we show how to construct the desired set S. Here, we build on the recent work of (Braverman et al. 2022) in designing "assignment-preserving coresets" for k-MEDIAN and k-MEANS. Their construction works by partitioning the points into $\tilde{O}(k^2\varepsilon^{-z})$ "rings" and then finding suitable (weighted) representatives from each of these rings. The latter procedure requires clever random sampling techniques. The selected representatives S_R from a particular ring R satisfy the following condition: for any set of k centers C, the total assignment cost to C of all the points in the ring R is close to that of S_R. But one would like this property to hold for all n^k possible ways of choosing C. Thus, one needs to apply union bound over all such possibilities, which results in a multiplicative factor of $k \log n$ in the representative size from each ring.[4] As mentioned earlier, one hopes to avoid this barrier by constructing weak coresets. Here, the number of possible choices for the set C reduces to $|J|^k$ instead of n^k. So, the $k \log n$ factor needed in the size of the sampled set S_R from each ring R gets replaced by $k \log |J|$. The trade-off is that instead of the classical coreset allowing a $(1+\varepsilon)$-approximate solution, the α-universal weak coreset only allows a $(\alpha+\varepsilon)$-approximate solution. This is not a significant compromise if α is the best approximation guarantee known for a constrained clustering problem, which is true for several cases. We formally state the result from (Braverman et al. 2022) that we shall use to construct the set S for our universal weak coreset.

Theorem 3 ((Braverman et al. 2022))**.** *Consider a clustering instance (X, F, w, k) and a parameter $\delta \in (0, 1)$. There is a randomised algorithm to construct a weighted set $T \subset X$ of size $O\left(\text{poly}(\frac{k}{\varepsilon}) \cdot \log \frac{1}{\delta}\right)$ with weight function $v : T \to \mathbb{R}^+$ that satisfies the following property: given a set C of k centers[5],*

$$\forall \Gamma, \mathsf{cost}_z(X, w, C, \Gamma) \in (1 \pm \varepsilon) \cdot \mathsf{cost}_z(T, v, C, \Gamma),$$

holds with probability at least $(1-\delta)$. Moreover, the running time of the algorithm is $O(n|T|)$.

The construction of the desired set S using the above result follows from a direct application of union bound over the choice of k center sets in the set J.

Theorem 4. *Consider a clustering instance (X, F, w, k). There is a randomized algorithm for constructing a weighted set $S \subset X$ of size $\left(\text{poly}(\frac{k}{\varepsilon}) \cdot \log |J|\right)$ with weight function $v : S \to \mathbb{R}^+$ such that the following event happens with high probability: for every choice of C centers from J, $|C| = k$, and every Γ:*

$$\mathsf{cost}_z(X, w, C, \Gamma) \in (1 \pm \varepsilon) \cdot \mathsf{cost}_z(S, v, C, \Gamma).$$

Moreover, the running time of the algorithm is $O(n|S|)$.

The following results now follow from Theorem 4 and the discussion after Theorem 2:

Theorem 5 (Main theorem: Metric k-MEDIAN)**.** *There is a 3-universal weak coreset (J, S, v) of size $\left(\text{poly}(\frac{k}{\varepsilon})\right)$ for k-MEDIAN (i.e., $z = 1$) objective in general metric spaces. The time to construct such a coreset is $O(n \cdot (|J| + |S|))$.*

For the special case $X \subseteq F$, the guarantee in the above theorem improves from 3 to 2.

Theorem 6 (Main theorem: Metric k-MEANS)**.** *There is a 9-universal weak coreset (J, S, v) of size $\left(\text{poly}(\frac{k}{\varepsilon})\right)$ for k-MEANS (i.e., $z = 2$) objective in general metric spaces. The time to construct such a coreset is $O(n \cdot (|J| + |S|))$.*

For the special case $X \subseteq F$, the guarantee in the above theorem improves from 9 to 4.

Theorem 7 (Main theorem: Euclidean k-MEDIAN/k-MEANS)**.** *There is a 1-universal weak coreset of size $\left(\text{poly}(\frac{k}{\varepsilon})\right)$ for k-MEDIAN and k-MEANS objectives in the Euclidean metric. The time to construct such a coreset is $O\left(n(k/\varepsilon)^{O(\frac{1}{\varepsilon})}\right)$.*

In the following section, we see applications of the results above.

Applications

In this section, we apply the universal weak coreset constructions to solve constrained versions of the k-MEDIAN and k-MEANS problems. As mentioned earlier, we can view a universal weak coreset as a compression of the original dataset. There are two ways of applying universal weak coresets : (i) Execute a known algorithm for the specific constrained problem on the compressed instance (J, S, v, k) instead of (F, X, w, k), and (ii) Use the meta-algorithm defined in Theorem 1 with appropriate modifications. We now discuss some specific examples.

Clustering With Size-Based Constraints

We consider constrained clustering problems where, besides optimizing the objective function, there are constraints on the size of the clusters. For example, the r-gathering problem requires a lower bound of r on the size of every cluster. Similarly, the capacitated clustering problem has an upper bound on cluster size. These constraints try to capture a "balance" property that limits the variance in the cluster sizes. We can model such size-constrained problems using the balanced k-MEDIAN or k-MEANS problem. Here, in addition to (F, X, w, k), an instance also specifies tuples $(l_1, ..., l_k)$ and $(u_1, ..., u_k)$; where l_i and u_i are the lower and upper bound on the total weight of the i^{th} cluster, respectively. For example, the r-gathering problem is obtained by setting $l_i = r, u_i = \infty$ for all $i \in [k]$. Let us see how the 3-universal weak coreset for k-MEDIAN objective from Theorem 5 implies a 3-approximation algorithm for any instance of the balanced k-MEDIAN problem (the extension to balanced k-MEANS is analogous).

Theorem 8. *Let (J, S, v) be a 3-universal weak coreset for an input instance (F, X, w, k). Let $\mathcal{I} = (F, X, w, k, (l_1, ..., l_k), (u_1, ..., u_k))$ be an instance of the balanced k-MEDIAN problem. Then there is a randomized algorithm $\mathcal{A}$, that with high probability, outputs a k center*

[4] Coreset constructions prior to (Braverman et al. 2022) (e.g., (Bandyapadhyay, Fomin, and Simonov 2021)) had another $\log n$ factor coming from the number of rings. This bottleneck was removed in (Braverman et al. 2022).

[5] Note that the statement is for a single set C of centers and not for every set of k centers.

set C that is a $(3+\varepsilon)$-approximate solution for $\mathcal{I}$. The running time of $\mathcal{A}$ is $\tilde{O}(|J|^k \cdot |S|)$.[6]

Proof. Consider an optimal solution to $\mathcal{I}$, and let $\sigma(x,i)$ denote the weight of point x assigned to cluster i. Define $\Gamma := (\sum_x \sigma(x,1), ..., \sum_x \sigma(x,k))$. From the 3-universal weak coreset property, there is a subset C of J, $|C| = k$, such that $\mathsf{cost}_1(X,w,C,\Gamma) \leq (3+\varepsilon)\mathsf{opt}(X,w,\Gamma)$. Moreover, the set S has the property that $\mathsf{cost}_1(X,w,C,\Gamma)$ and $\mathsf{cost}_1(S,v,C,\Gamma)$ are within $(1+\varepsilon)$ factor of each other. This implies that if we try all possible choices of k centers C from J and, for each such C, find $\mathsf{opt}_1(S,v,C,\Gamma)$, then we can compute $\mathsf{opt}_1(X,w,\Gamma)$ within $(3+\varepsilon)$ approximation factor.

The remaining issue is how to compute $\mathsf{opt}_1(S,v,C,\Gamma)$ for a given choice of C. We do not know Γ here, but we can find the tuple Γ' for which $\mathsf{opt}_1(S,v,C,\Gamma')$ is minimized. Indeed, we can set up a minimum cost flow network where we would like to assign the points fractionally to the centers in C, and for each center in C, we can assign lower and upper bounds (i.e., l_i and u_i) for the amount of weight assigned to it. Solving this min-cost flow problem shall yield the optimal choice of Γ'. Minimizing over $C \subseteq J, |C| = k$, we can find $\mathsf{opt}_1(X,w,\Gamma)$.

Combining the above ideas yields a $(3+\varepsilon)$ approximation algorithm. □

The $(9+\varepsilon)$-approximation for arbitrary balanced versions of the k-MEANS problem in general metrics follows on similar lines using the 9-universal weak coreset from Theorem 6. Similarly, a $(1+\varepsilon)$-approximation for arbitrary balanced versions of the k-MEDIAN and k-MEANS problems in Euclidean metrics can be obtained using 1-universal weak coreset from Theorem 7.

Fair Clustering and Other Labeled Versions

We now consider constrained clustering problems where points have labels, i.e., we are given a label set $L := \{1, ..., m\}$, and each point x has a label $\ell(x) \in L$ associated with it. Labels can capture disparate scenarios where every client may be part of multiple (overlapping) groups (*e.g., groups based on gender, ethnicity, age, etc.*). Every unique combination of groups gets assigned a different label, and hence m denotes the number of distinct combinations of groups to which a point can belong. For a label $j \in L$, let X_j denote the set of points that are assigned label j. Consider a clustering instance (X, F, w, k, ℓ), where we have also incorporated the label mapping. The corresponding fair clustering instance $\mathcal{I}$ is specified by an additional list of k pairs, namely, $(\alpha_1, \beta_1), ..., (\alpha_k, \beta_k)$. An optimal solution needs to find a set of k centers, and an assignment $\sigma : X \times [k] \to \mathbb{R}^+$ for all $x \in X$, such that:

(i) For every $j \in [m]$ and $i \in [k]$, $\frac{\sum_{x \in X_i} \sigma(x,i)}{\sum_{x \in X} \sigma(x,i)} \in [\alpha_i, \beta_i]$, i.e., for every group, the fraction of weights assigned to the i^{th} cluster is in the range $[\alpha_i, \beta_i]$. This captures various fairness notions for points that may belong to a particular group.

(ii) The assignment cost, i.e., $\sum_{i=1}^k \sum_{x \in X} \sigma(x,i) \cdot D^z(x, c_i)$, is minimized.

Our definition of universal weak coreset is for the case $m = 1$, i.e., points have only one label, which may be interpreted as the unlabeled case. However, we can extend the notion of universal weak sets to multi-label settings.

Towards this, we recall that the set J constructed in the previous section (Theorems 5 and 6) satisfies the following property: for any assignment $\sigma : X \times [k] \to \mathbb{R}^+$, there is a $(3^z+\varepsilon)$-approximate center set in J with respect to σ. More specifically, let σ^* denote the optimal assignment and let $C^* \equiv (c_1^*, ..., c_k^*)$ denote the optimal k centers that respects σ^*. The property on set J says that there exists k centers $C \equiv (c_1, ..., c_k)$ such that $\sum_i \sum_x \sigma^*(x,i) \cdot D(x,c_i)^z \leq (3^z+\varepsilon) \cdot \sum_i \sum_x \sigma^*(x,i) \cdot D(x,c_i^*)^z$. This means that as long as our set S has the property that for any assignment respecting the group constraint, the corresponding assignment cost to any $C \subseteq J, |C| = k$ is about the same as that of the point set X, we shall have a 3^z-universal weak coreset for the fair clustering problem as well. Here, we note that we can execute the coreset construction from Theorem 4 separately on each group and take a union of the corresponding coresets obtained. This larger set acts as a coreset for the labeled dataset. We now formalize these ideas. First, we extend the notion of a universal weak coreset to the multi-labeled setting. In the unlabeled version, the assignment of weights to centers in a set C was characterized by a tuple Γ of size k. Since we have m labels now, such an assignment needs to be specified for each label. In other words, we now consider tuples Γ of length mk, i.e., $\Gamma := (t_{1,1}, ..., t_{1,m}, t_{2,1}, ..., t_{2,m}, ..., t_{k,1}, ..., t_{k,m})$, where $t_{i,j}$ is meant to denote the total weight of points with label j assigned to the i^{th} cluster. We can define an assignment σ analogously as a map $X \times [k] \to \mathbb{R}^+$. We say that σ is consistent with Γ, i.e., $\sigma \sim \Gamma$ if for every label j and cluster i, $\sum_{x \in X_j} \sigma(x,i) = t_{i,j}$. Similarly, for a set of centers C, define $\mathsf{cost}_z(X,w,C,\Gamma)$ as

$$\mathsf{cost}_z(X,w,C,\Gamma) \equiv \min_{\sigma \sim \Gamma} \left\{ \sum_{i=1}^k \sum_{x \in X} \sigma(x,i) \cdot D(x,c_i)^z \right\}.$$

Again, $\mathsf{opt}_z(X,w,\Gamma)$ can be defined as the optimum cost over all choices of centers C. Now, the definition of a universal coreset (J,S,v) in this setting is analogous to that in Definition 1 – we need to satisfy conditions (A) and (B).

Theorem 9. *There is a $(3^z+\varepsilon)$-universal weak coreset of size $\left(m \cdot \mathrm{poly}(\frac{k}{\varepsilon})\right)$ for constrained clustering in the multi-labeled setting.*

Proof. The set J is constructed as in the section on coreset construction. In order to construct the set S, we apply Theorem 4 to each of the sets $X_1, \ldots, X_m$ independently to obtain sets $S_1, \ldots, S_m$. Finally, $S := S_1 \cup \ldots \cup S_m$. The desired result follows from the properties of the universal coreset. □

[6] The overall running time of the approximation algorithm, including the time to construct the universal weak coreset is $n \cdot \mathrm{poly}(\frac{k}{\varepsilon}) + \left(\frac{k}{\varepsilon}\right)^{O(k)}$.

Let us now see why a $(3^z+\varepsilon)$-universal weak coreset can be used for obtaining a $(3^z+\varepsilon)$-approximate solution for multi-labeled constrained clustering problem in FPT time (fixed-parameter tractable time). We state the result for the k-MEDIAN objective in general metric spaces. Similar results will hold for k-MEANS in general metric spaces (i.e., $(9+\varepsilon)$-approximation) and k-MEDIAN or k-MEANS objectives in Euclidean spaces (i.e., $(1+\varepsilon)$-approximation).

Theorem 10. *Let (J, S, v) be a 3-universal weak coreset for a multi-labeled clustering instance (F, X, w, k, ℓ). Consider an instance $\mathcal{I}$ of the constrained clustering problem specified by a set of pairs $\{(\alpha_1, \beta_1), ..., (\alpha_k, \beta_k)\}$. Then there is a randomized algorithm $\mathcal{A}$, which on input $\mathcal{I}$ and (J, S, v) outputs a $(3+\varepsilon)$-approximate solution with high probability. The running time of $\mathcal{A}$ is $|J|^k \cdot (mk)^{O(mk)} \cdot n^{O(1)}$.*

Proof. The proof proceeds along the same line as that for the unlabeled case. We try all $|J|^k$ possible k centers $(c_1, ..., c_k)$ from J and solve the "assignment" problem: find the best fair assignment for the given choice of $(c_1, ..., c_k)$. Our 3-universal weak coreset guarantees the existence of a $(3+\varepsilon)$-approximate solution within J. So, if we can solve the assignment problem optimally, we can find that $(3+\varepsilon)$-approximate solution in J. Such an assignment algorithm was given by (Bandyapadhyay, Fomin, and Simonov 2021) (see Theorem 8.2). The running time of this assignment finding algorithm is $(mk)^{O(mk)} \cdot n^{O(1)}$. □

l-diversity clustering Another well-known constrained clustering problem in the labeled setting is the l-diversity problem. Here, the goal is to cluster the point set X into clusters $(X_1, ..., X_k)$ such that each cluster has at least $1/l$ fraction of the points from each label. Again, the goal is to minimize the k-MEDIAN or k-MEANS assignment cost.

As above, we can use the universal weak coreset construction from Theorem 10 to obtain a $(3^z+\varepsilon)$-approximation algorithm for this problem. Here, we can use the algorithm of (Ding and Xu 2020) to solve the corresponding assignment problem.

Discussion and Open Problems

Classical coresets come with the promise that they help obtain an approximate solution to the k-MEANS or the k-MEDIAN objective in a metric space. This promise holds for most known metric spaces. However, there are certain metrics where a specific approximation guarantee cannot be obtained using a classical coreset. The reason is that the approximation algorithm with such a guarantee does not work on weighted inputs. Note that a classical coreset is a weighted set. For example, a recent development (Chakraborty, Das, and Krauthgamer 2023) in the k-MEDIAN problem in the Ulam metric has broken the 2-approximation barrier. However, their $(2-\delta)$-approximation algorithm works only on unweighted input permutations. So, the classical coreset framework does not help in this setting. On the other hand, the universal weak coreset framework may still be applicable. The reason is that even though we cannot run the approximation algorithm on the set S to find a good center set, we can use S to locate a good center set from J using the cost preservation property of S. So, an interesting open question is whether there is a $(2-\delta)$-universal weak coreset for the Ulam k-median problem. In general, in cases where the guarantee of the set S is limited to cost preservation, i.e., S represents the data only in a limited sense, a universal weak coreset is a more appropriate object to use. It will be interesting to see if there are problems other than the Ulam k-median problem with this property.

Note that there are one pass streaming algorithms for constructing the set S because coresets have composability property (Chen 2009); and there is a constant-pass streaming algorithm for constructing the set J (the algorithm for constructing M in Theorem 2 can be implemented in streaming settings). Thus, both J and S can be constructed in a constant pass streaming setting. We leave it an open problem to design a single-pass streaming algorithm for a universal weak coreset.

Although we give 3-universal weak coreset constructions of size independent of any function of n for k-MEDIAN (and a similar result for k-MEANS), it remains an open problem to construct an α-universal weak coreset for a constant $\alpha < 3$, even for general metric spaces. This will help in obtaining a better than 3 approximation algorithm for several constrained k-MEDIAN problems for which the best-known approximation bound is 3 (similarly a better than 9-approximation for k-MEANS).

Acknowledgments

Ragesh Jaiswal acknowledges the support from the SERB, MATRICS grant.

References

Baker, D.; Braverman, V.; Huang, L.; Jiang, S. H.-C.; Krauthgamer, R.; and Wu, X. 2020a. Coresets for Clustering in Graphs of Bounded Treewidth. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 569–579. PMLR.

Baker, D.; Braverman, V.; Huang, L.; Jiang, S. H.-C.; Krauthgamer, R.; and Wu, X. 2020b. Coresets for Clustering in Graphs of Bounded Treewidth. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 569–579. PMLR.

Bandyapadhyay, S.; Fomin, F. V.; and Simonov, K. 2021. On Coresets for Fair Clustering in Metric and Euclidean Spaces and Their Applications. In Bansal, N.; Merelli, E.; and Worrell, J., eds., *48th International Colloquium on Automata, Languages, and Programming (ICALP 2021)*, volume 198 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 23:1–23:15. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-195-5.

Becchetti, L.; Bury, M.; Cohen-Addad, V.; Grandoni, F.; and Schwiegelshohn, C. 2019. Oblivious Dimension Reduction for K-Means: Beyond Subspaces and the Johnson-

Lindenstrauss Lemma. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, 1039–1050. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367059.

Bhattacharya, A.; Goyal, D.; Jaiswal, R.; and Kumar, A. 2020. On Sampling Based Algorithms for k-Means. In Saxena, N.; and Simon, S., eds., *40th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2020)*, volume 182 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 13:1–13:17. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. ISBN 978-3-95977-174-0.

Braverman, V.; Cohen-Addad, V.; Jiang, H.; Krauthgamer, R.; Schwiegelshohn, C.; Toftrup, M.; and Wu, X. 2022. The Power of Uniform Sampling for Coresets. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, 462–473. Los Alamitos, CA, USA: IEEE Computer Society.

Braverman, V.; Jiang, S.; Krauthgamer, R.; and Wu, X. 2021. Coresets for Clustering in Excluded-minor Graphs and Beyond. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms*. ACM-SIAM Symposium on Discrete Algorithms, SODA ; Conference date: 10-01-2021 Through 13-01-2021.

Chakraborty, D.; Das, D.; and Krauthgamer, R. 2023. Clustering Permutations: New Techniques with Streaming Applications. In Kalai, Y. T., ed., *14th Innovations in Theoretical Computer Science Conference, ITCS 2023, January 10-13, 2023, MIT, Cambridge, Massachusetts, USA*, volume 251 of *LIPIcs*, 31:1–31:24. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Chen, K. 2009. 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.

Cohen-Addad, V.; Larsen, K. G.; Saulpic, D.; and Schwiegelshohn, C. 2022. Towards Optimal Lower Bounds for K-Median and k-Means Coresets. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, 10381051. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392648.

Cohen-Addad, V.; Saulpic, D.; and Schwiegelshohn, C. 2021. A New Coreset Framework for Clustering. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, 169–182. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380539.

Ding, H.; and Xu, J. 2020. A Unified Framework for Clustering Constrained Data Without Locality Property. *Algorithmica*, 82: 808–852.

Feldman, D.; and Langberg, M. 2011. A Unified Framework for Approximating and Clustering Data. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11, 569–578. New York, NY, USA: Association for Computing Machinery. ISBN 9781450306911.

Feldman, D.; Monemizadeh, M.; and Sohler, C. 2007. A PTAS for k-means clustering based on weak coresets. In *Proceedings of the twenty-third annual symposium on Computational geometry*, SCG '07, 11–18. New York, NY, USA: ACM. ISBN 978-1-59593-705-6.

Feldman, D.; Schmidt, M.; and Sohler, C. 2020. Turning Big Data Into Tiny Data: Constant-Size Coresets for k-Means, PCA, and Projective Clustering. *SIAM Journal on Computing*, 49(3): 601–657.

Goyal, D.; Jaiswal, R.; and Kumar, A. 2020. FPT Approximation for Constrained Metric k-Median/Means. In Cao, Y.; and Pilipczuk, M., eds., *15th International Symposium on Parameterized and Exact Computation (IPEC 2020)*, volume 180 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 14:1–14:19. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. ISBN 978-3-95977-172-6.

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

Har-Peled, S.; and Mazumdar, S. 2004. On Coresets for K-Means and k-Median Clustering. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '04, 291–300. New York, NY, USA: Association for Computing Machinery. ISBN 1581138520.

Huang, L.; Jiang, S. H.-C.; Li, J.; and Wu, X. 2018. Epsilon-Coresets for Clustering (with Outliers) in Doubling Metrics. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 814–825.

Huang, L.; Jiang, S. H.-C.; and Vishnoi, N. K. 2019. *Coresets for Clustering with Fairness Constraints*. Red Hook, NY, USA: Curran Associates Inc.

Huang, L.; and Vishnoi, N. K. 2020. Coresets for Clustering in Euclidean Spaces: Importance Sampling is Nearly Optimal. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, 1416–1429. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369794.

Langberg, M.; and Schulman, L. J. 2010. Universal ε-Approximators for Integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, 598–607. USA: Society for Industrial and Applied Mathematics. ISBN 9780898716986.

Maalouf, A.; Tukan, M.; Braverman, V.; and Rus, D. 2023. AutoCoreset: An Automatic Practical Coreset Construction Framework. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 23451–23466. PMLR.

Schmidt, M.; Schwiegelshohn, C.; and Sohler, C. 2020. Fair Coresets and Streaming Algorithms for Fair k-means. In Bampis, E.; and Megow, N., eds., *Approximation and Online Algorithms*, 232–251. Cham: Springer International Publishing. ISBN 978-3-030-39479-0.

Sohler, C.; and Woodruff, D. P. 2018. Strong Coresets for k-Median and Subspace Approximation: Goodbye Dimension. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 802–813.