*Sublinear-Time Approximation Algorithms for Clustering via Random Sampling**

Artur Czumaj,[1] Christian Sohler[2]

[1] *Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom; e-mail: czumaj@dcs.warwick.ac.uk*

[2] *Heinz Nixdorf Institute and Department of Computer Science, University of Paderborn, D-33102 Paderborn, Germany; e-mail: csohler@uni-paderborn.de*

Received 18 November 2005; accepted 25 May 2006; received in final form 12 October 2006
Published online 15 December 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/rsa.20157

ABSTRACT: We present a novel analysis of a random sampling approach for four clustering problems in metric spaces: *k-median*, *k-means*, *min-sum k-clustering*, and *balanced k-median*. For all these problems, we consider the following simple sampling scheme: select a small sample set of input points uniformly at random and then run some approximation algorithm on this sample set to compute an approximation of the best possible clustering of this set. Our main technical contribution is a significantly strengthened analysis of the approximation guarantee by this scheme for the clustering problems.

The main motivation behind our analyses was to design *sublinear-time* algorithms for clustering problems. Our second contribution is the development of new approximation algorithms for the aforementioned clustering problems. Using our random sampling approach, we obtain for these problems

Correspondence to: Artur Czumaj

*A preliminary extended abstract of this work appeared in *Proceedings of the 31st Annual International Colloquium on Automata, Languages and Programming (ICALP)*, pp. 396–407, 2004.

Contract grant sponsor: NSF ITR
Contract grant number: CCR-0313219
Contract grant sponsor: NSF
Contract grant number: CCR-0105701
Contract grant sponsor: DFG
Contract grant number: Me 872-8/3

the first time approximation algorithms that have running time independent of the input size, and depending on k and the *diameter* of the metric space only. © 2006 Wiley Periodicals, Inc. Random Struct. Alg., 30, 226–256, 2007

Keywords: clustering; k*-median;* k*-means; min-sum clustering; random sampling*

1. INTRODUCTION

The problem of clustering large data sets into subsets (clusters) of similar characteristics has been extensively studied in computer science, operations research, and related fields. Clustering problems arise in various applications, for example, in data mining, data compression, bioinformatics, pattern recognition, and pattern classification. In some of these applications, massive datasets have to be processed, e.g., web pages, network flow statistics, or call-detail records in telecommunication industry. Processing such massive data sets in more than linear time is by far too expensive and often even linear time algorithms may be too slow. One reason for this phenomenon is that massive data sets do not fit into main memory and sometimes even secondary memory capacities are too low. Hence, there is the desire to develop algorithms whose running times are not only polynomial, but in fact are *sublinear* in n (for very recent survey expositions, see, e.g., [8,9,24]). In a typical sublinear-time algorithm, a subset of the input is selected according to some random process and then processed by an algorithm. With high probability, the outcome of this algorithm should be a good approximation of the outcome of an exact algorithm running on the whole input. In many cases, the randomized process that selects the sample is very simple, e.g., it may select a subset uniformly at random.

In this paper, we address the problem of designing *sublinear-time* approximation algorithms using *uniformly random sampling* for clustering problems in metric spaces. We consider four clustering problems: the *k-median problem*, the *k-means problem*, the *min-sum k-clustering problem*, and the *balanced k-median problem*. Given a finite metric space (V, μ), the *k-median problem* is to find a set $C \subseteq V$ of k centers that minimizes $\sum_{p \in V} \mu(p, C)$, where $\mu(p, C)$ denotes the distance from p to the nearest point in C. The *k-means problem* differs from the k-median problem in that we minimize the sum of the squares of the distances of all points to the nearest center, that is, $\sum_{p \in V} (\mu(p, C))^2$. The *min-sum k-clustering problem* for a metric space (V, μ) is to find a partition of V into k subsets $C_1, \dots, C_k$ such that $\sum_{1 \le i \le k} \sum_{p,q \in C_i} \mu(p, q)$ is minimized. The *balanced k-median problem* (which is perhaps less standard than the other three problems) for a metric space (V, μ) is to find a set $\{c_1, \dots, c_k\} \subseteq V$ of k-centers and a partition of V into k subsets $V_1, \dots, V_k$ that minimizes $\sum_{1 \le i \le k} |V_i| \sum_{p \in V_i} \mu(p, c_i)$. Note that in this problem we allow that $c_i \notin V_i$ for an optimal partition $V_1, \dots, V_k$. We also consider the variant of the problem with the constraint $c_i \in V_i$. The results obtained in this paper hold for both variants of the problem.

For all these clustering problems we study the following "simple sampling" algorithm:

- pick a random sample S of points,
- run an approximation algorithm for clustering for the sample, and
- return the clustering induced by the solution for the sample.

The main goal of this article is to design a generic method of analyzing this sampling scheme and to obtain a significantly stronger quantitative bounds for the performance of this method. Using our approach, for a large spectrum of input parameters, we obtain *sublinear-time algorithms* for the four clustering problems mentioned earlier. These are the first approximation algorithms for these problems whose running time is *fully independent of the input size*, $|V|$.

1.1. Previous Research

1.1.1. k-Median. The k-median clustering problem is one of the most studied clustering problem in the literature, both, in theoretical and applied research. It is well-known that the k-median clustering in metric spaces is $\mathcal{NP}$-hard and it is even $\mathcal{NP}$-hard to approximate within a factor of $1 + \frac{2}{e}$ [19]. There exist polynomial time approximation algorithms with constant approximation ratios [2, 5, 6, 16, 20, 26]. When the underlying space is the Euclidean plane, Arora et al. [1] obtained even a PTAS for k-median (extension to higher dimensions and improvements in the running time have been obtained in [3, 14, 21, 23]). The k-median problem has been also extensively investigated in the data stream model, see, e.g., in [7, 12, 14].

A few sublinear-time algorithms for the k-median problem are also known, that is, algorithms with a running time of $o(n^2)$ (if we consider an arbitrary metric space (V, μ) with $|V| = n$, then its description size is $\Theta(n^2)$), see, e.g., [16, 26–28]. The algorithm of Indyk [16] computes in $O(nk)$ time a set of $O(k)$ centers whose cost approximates the value of the k-median by a constant factor. Mettu and Plaxton [26] gave a randomized $O(1)$-approximate k-median algorithm that runs in time $O(n(k + \log n))$ subject to the constraint $R \le 2^{O(n/\log(n/k))}$, where R denotes the ratio between the maximum and the minimum distance between any pair of distinct points in the metric space. Recently, Meyerson et al. [27] presented a sublinear-time for the k-median problem under an assumption that each cluster has size $\Omega(\epsilon n/k)$; their algorithm requires time $O((k^2/\epsilon)\log(k/\epsilon))$ and gives an $O(1)$-approximation guarantee with high probability.

In this article, we consider a model where an upper bound on the diameter of the metric space, Δ is given, that is, where $\mu : V \times V \to [0, \Delta]$. We are interested in the average error of the cost of a sample that is chosen uniformly at random (with repetition) from V. This model has been introduced by Mishra et al. [28] and is motivated by previous works in statistics and learning theory. For example, research on uniform convergence in statistics tries to characterize conditions under which a sample set S is large enough, such that for any function f from a class F, the mean of f on the sample deviates by at most ϵ from the true mean of f.

We remark that an average additive error of ϵ translates to an additive error of ϵn for the total cost. This implies that these are not *multiplicative* approximation algorithms since the cost for an optimal clustering can be much smaller than ϵn. However, for most instances arising in practice, the average contribution of a point is some small constant. For these instances, our results translate to approximation algorithms in the classical sense.

Next, let us notice that an additive error cannot be avoided if one wants to design algorithms with $o(n)$ running time. This can be seen as follows. If we take a random sample of size s then it is unlikely that our sample contains points from a cluster with, say, $\frac{n}{100s}$ points. If this cluster has distance to the remaining points roughly Δ, then by not placing a center in this cluster, one can incur an (average) error of $\frac{\Delta}{s}$. Since the cost of clustering the remaining points may be arbitrarily small, this error is of additive nature. Furthermore, it is easy to see that any algorithm that has an additive average error of at most ϵ requires $s \ge \Delta/\epsilon$, i.e., the sample size must depend on Δ.

We also remark that scaling the instance cannot help to overcome these problems. If we scale the metric such that the new metric has maximum distance 1, then apply the algorithm and rescale back to our original instance, the additive error will become $\epsilon\Delta$ rather than ϵ.

Mishra et al. [28] studied the quality of k-median clusterings obtained by random sampling in this model. Let $\mathbb{A}_\alpha$ be an arbitrary α-approximation algorithm for k-median. Using techniques from statistics and computational learning theory, Mishra et al. [28] proved that if we sample a set S of $s = \widetilde{O}\big(\big(\frac{\alpha\Delta}{\epsilon}\big)^2(k \ln n + \ln(1/\delta))\big)$ points from V i.u.r. (independently and uniformly at random) and run algorithm $\mathbb{A}_\alpha$ to approximate the k-median solution for S, then with probability at least $1 - \delta$ the average distance of each point to the nearest center in the set of centers output by $\mathbb{A}_\alpha$ is at most $2\alpha\,\mathrm{med}_{\mathrm{avg}}(V, k) + \epsilon$, where $\mathrm{med}_{\mathrm{avg}}(V, k)$ denotes the *average distance* to the optimal k-median solution C, that is, $\mathrm{med}_{\mathrm{avg}}(V, k) = \frac{\sum_{v\in V}\mu(v,C)}{n}$. Using this result, Mishra et al. [28] developed a generic sublinear-time approximation algorithm for k-median. If the algorithm $\mathbb{A}_\alpha$ has the running time of $T(s)$, then the resulting algorithm runs in $T(s)$ time for $s = \widetilde{O}\big(\big(\frac{\alpha\Delta}{\epsilon}\big)^2(k \ln n + \ln(1/\delta))\big)$ and computes with probability at least $1 - \delta$ a set of k centers such that the average distance to the nearest center is at most $2\alpha\,\mathrm{med}_{\mathrm{avg}}(V, k) + \epsilon$. Notice that since there exist $O(1)$-approximation algorithms for k-median with $T(s) = O(s^2)$, this approach leads to an approximation algorithm for the k-median problem whose dependency on n is only $\widetilde{O}(\log^2 n)$, rather than $\Omega(n^2)$ or $\Omega(nk)$ as in the algorithms discussed earlier. On the other hand, the running time of this algorithm depends on Δ; however, as discussed earlier (see also [26–28]), such a dependency is necessary to obtain this kind of approximation guarantee.

1.1.2. k-Means Clustering. The k-means problem is another standard clustering problem that has been widely studied in the literature. As in the case of the k-median problem, the k-means problem in metric spaces is $\mathcal{NP}$-hard to approximate within a factor of $1 + c$, for some positive constant c. Notice that the k-means problem in a metric space is identical to the k-median problem in a non-metric space where all the distances are squared. The squaring of the distances does not maintain the triangle inequality, but it is well-known that the resulting distance function is almost a metric: all of the properties of a metric space are satisfied except that the triangle inequality only holds to within a factor of 2. Therefore, in particular, many algorithms for the k-median problem can be easily transformed to work for the k-means problem as well.

The k-means clustering problem has been studied especially extensively in geometric setting, where the input point set is in a Euclidean space $\mathbb{R}^d$ and the centers are not restricted to be input points but may be arbitrary points in $\mathbb{R}^d$. Inaba et al. [15] described an exact algorithm for this problem that requires time $O(n^{kd+1})$. There are numerous $(1 + \varepsilon)$-approximation algorithms for k-means clustering in $\mathbb{R}^d$, see, e.g., [3, 11, 14, 15, 22, 25]. The most recent algorithm, by Kumar et al. [22], is a sampling-based algorithm that finds a $(1+\varepsilon)$-approximation of k-means clustering for a set of n points in $\mathbb{R}^d$ in time $O(2^{(k/\varepsilon)^{O(1)}}dn)$, which is linear for fixed k and ε.

1.1.3. Min-Sum k-Clustering. The min-sum k-clustering problem was first formulated (for general graphs) by Sahni and Gonzales [30]. There is a 2-approximation algorithm by Guttman-Beck and Hassin [13], with running time $n^{O(k)}$. Recently, Bartal et al. [4] presented an $O\big(\frac{1}{\epsilon}\log^{1+\epsilon} n\big)$-approximation algorithm with $n^{O(1/\epsilon)}$ running time, and then Fernandez de la Vega et al. [11] gave a $(1 + \epsilon)$-approximation algorithm with running time $O(n^{3k}2^{O((1/\epsilon)^{k^2})})$. For points in $\mathbb{R}^d$, Schulman [29] introduced an algorithm for distance functions ℓ_2^2, ℓ_1, and ℓ_2 that computes a solution that is either within a $(1+\epsilon)$-factor of the optimum or that disagrees with the optimal clustering in at most an ϵ fraction of the points. For the basic case of $k = 2$ (which is the complement of the Max-Cut problem), Indyk [18, Theorem 38] gave a $(1+\epsilon)$-approximation algorithm that runs in time $O(2^{1/\epsilon^{O(1)}} n(\log n)^{O(1)})$, which is sublinear in the full input description size, but superlinear in n.

1.1.4. Balanced k-Median. It is known that in metric spaces the solution to balanced k-median is to within a factor of 2 of that of min-sum k-clustering, e.g. [4, Claim 1]. Therefore, balanced k-median has been usually considered in connection with the min-sum k-clustering problem discussed earlier. The problem was first studied by Guttman-Beck and Hassin [13] who gave an exact $O(n^{O(k)})$-time algorithm, and Bartal et al. [4] obtained an $O\big(\frac{1}{\epsilon}\log^{1+\epsilon} n\big)$-approximation in time $n^{O(1/\epsilon)}$ based on metric embeddings into HSTs [10]. We are not aware of any sublinear-time algorithm for balanced k-median.

1.2. New Contribution

In this article, we investigate the quality of a simple *uniform sampling* approach to clustering problems and apply novel analyzes to obtain improved bounds for the running time of clustering algorithms.

1.2.1. k-Median Problem. We first study the *k-median* problem. Our sampling is identical to the one by Mishra et al. [28]; however, our analysis is stronger and leads to significantly better bounds. Let $\alpha \geq 1$, $0 < \delta < 1$, $0 < \beta \leq 1$, and $\epsilon > 0$ be arbitrary parameters. We prove that if we pick a sample set of size $\widetilde{O}\big(\frac{\Delta\alpha}{\epsilon\beta^2}(k + \alpha\ln(1/\delta))\big)$ i.u.r., then an α-approximation of the optimal solution for the sample set yields an approximation of the average distance to the nearest median to within $2(\alpha+\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)+\epsilon$ with probability at least $1-\delta$; notice, in particular, that this gives the sample size *independent of n*, which we consider the main contribution of our result. As noted before (see also [28]), it is impossible to obtain a sample complexity independent of both Δ and n.

Comparing our result with the one from [28], we improve the sample complexity by a factor of $\frac{\Delta\log n}{\epsilon}$ while obtaining a slightly worse approximation ratio of $2(\alpha + \beta)\,\mathrm{med}_{\mathrm{avg}}(V,k) + \epsilon$, instead of $2\alpha\,\mathrm{med}_{\mathrm{avg}}(V,k) + \epsilon$ as in [28]. As a highlight, we obtain an algorithm that in time $\widetilde{O}\big(\big(\frac{\Delta}{\epsilon}(k + \log(1/\delta))\big)^2\big)$—*fully independent of n*—has an average distance to the nearest median of at most $O(\mathrm{med}_{\mathrm{avg}}(V,k)) + \epsilon$ with probability at least $1-\delta$.

Furthermore, our analysis can be improved if we assume the input points are in Euclidean space $\mathbb{R}^d$. In this case, we improve the approximation guarantee to $(\alpha+\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k) + \epsilon$ at the cost of increasing the sample size to $\widetilde{O}\big(\frac{\Delta\alpha}{\epsilon\beta^2}(kd + \log(1/\delta))\big)$. This bound also significantly improves the analysis of Mishra et al. [28].

1.2.2. k-Means Clustering. Our analysis of the sampling algorithm for k-median mentioned earlier can be easily modified to handle the k-means clustering problem. The only real difference is the loss of one Δ factor in the analysis.

Let $\mathrm{mean}_{\mathrm{avg}}(V,k)$ denote the *average distance* to the optimal k-means solution C, that is, $\mathrm{mean}_{\mathrm{avg}}(V,k) = \frac{\sum_{v\in V}(\mu(v,C))^2}{n}$. Then, we can prove that if we pick a sample set of size $\widetilde{O}\big(\frac{\Delta^2\alpha}{\epsilon\beta^2}(k + \alpha\ln(1/\delta))\big)$ i.u.r., then an α-approximation of the optimal solution for the sample set yields an approximation of the average distance to the nearest k-means to within $4(\alpha+\beta)\,\mathrm{mean}_{\mathrm{avg}}(V,k)+\epsilon$ with probability at least $1-\delta$. Using this result, we can obtain an algorithm that in time $\widetilde{O}\big(\big(\frac{\Delta^2}{\epsilon}(k+\log(1/\delta))\big)^2\big)$ has the average distance to the nearest k-means at most $O(\mathrm{mean}_{\mathrm{avg}}(V,k))+\epsilon$ with probability at least $1-\delta$. We can also extend our analysis to the Euclidean version of the problem. We prove that a set of points in Euclidean space $\mathbb{R}^d$, a sample set of size $\widetilde{O}\big(\frac{\Delta^2\alpha}{\epsilon\beta^2}(kd+\log(1/\delta))\big)$ can be used to obtain an algorithm that returns a k-means solution of average cost at most $(\alpha+\beta)\,\mathrm{mean}_{\mathrm{avg}}(V,k)+\epsilon^2$ with probability at least $1-\delta$.

1.2.3. Min-Sum k-Clustering and k-Median Problems. The *min-sum k-clustering* and the *balanced k-median* problems are combinatorially more complex than the k-median problem. For these two problems, we give the *first* sublinear-time algorithms. Since in metric spaces the solution to the balanced k-median problem is within a factor of 2 of that of the min-sum k-clustering problem, we will consider the balanced k-median problem only.

We consider the problem of minimizing the average balanced k-median cost, that is, the cost of the balanced k-median normalized by the square of the number of input elements. We use the same approach as for the k-median problem. Let $\epsilon > 0$, $\alpha \geq 1$, $\beta > 0$, and $0 < \delta < 1$ be arbitrary parameters. We prove that if we pick a sample set of size $\widetilde{O}\big(\frac{\Delta(k+\ln(1/\rho))}{\epsilon}((\alpha/\beta)^2+\Delta k^2/\epsilon)\big)$ i.u.r., then an α-approximation of the optimal solution for the sample set approximates the average balanced k-median cost to within $(2\alpha+\beta)\,\mathrm{med}^b_{\mathrm{avg}}(V,k)+\epsilon$ with probability at least $1-\delta$, where $\mathrm{med}^b_{\mathrm{avg}}(V,k)$ denotes the average cost of the optimal solution for balanced k-median. Notice that similarly as for the k-median problem, the sample size is independent of n.

Unlike the k-median problem, the output of balanced k-median is supposed to consist of a set of k centers $c_1,\dots,c_k$ and a partition (clustering) of the input V into $V_1\cup\cdots\cup V_k$ that minimizes (or approximately minimizes) $\sum_{i=1}^k |V_i| \sum_{v\in V_i}\mu(v,c_i)$. Our sampling algorithm, when combined with the algorithm due to Bartal et al. [4], leads to a randomized algorithm that in time independent of n returns the set of k centers $c_1,\dots,c_k$ for which the value of $\frac{\sum_{i=1}^k |V_i|\sum_{v\in V_i}\mu(v,c_i)}{n^2}$ is at most $\log^{O(1)} n\,\mathrm{med}^b_{\mathrm{avg}}(V,k)+\epsilon$ with probability at least $1-\delta$. If one also knows the number of elements that are assigned to each cluster in an approximate solution, then one can compute in $O(nk)+\widetilde{O}(k^{2.5}\sqrt{n})$ time an optimal clustering [31]. Since our algorithm can be modified to provide the cluster sizes, we can use this approach to compute a good solution quickly from the implicit representation as a balanced k-median.

1.3. High Level Description of Our Approach

Before we begin to analyze specific problems, we first discuss our high level approach. We study the approximation guarantee of the following natural sampling scheme. Choose a multiset S of s elements i.u.r. from V, for some suitable chosen s. Then run an α-approximation algorithm $\mathbb{A}$ for the problem of interest on S. Our goal is to study the quality of the solution computed by $\mathbb{A}$ on S.

```
Generic sampling scheme $(V, \mathbb{A}, s)$

choose a multiset $S \subseteq V$ of size $s$ i.u.r. (with repetitions)
run $\alpha$-approximation algorithm $\mathbb{A}$ on input $S$ to compute a solution $C^*$ (set of $k$ centers)
return set $C^*$
```

Let us denote by $\mathrm{cost}(X, C)$ the cost of the clustering (for the problem under consideration) of set X with the center set C and let C_{opt} denote an optimal solution for V.

To analyze the approximation guarantee of this approach, we proceed in three steps.

(i) We show that w.h.p. and after normalization $\mathrm{cost}(S, C_{\mathrm{opt}})$ is an approximation of $\mathrm{cost}(V, C_{\mathrm{opt}})$.

(ii) Since C_{opt} may not be a feasible solution for S (e.g., in the k-median problem C_{opt} may not be contained in S), we show that there is a *feasible* solution in S, which has cost at most $\frac{c}{\alpha}\,\mathrm{cost}(S, C_{\mathrm{opt}})$ for some constant $c \geq \alpha$.

(iii) We show that w.h.p. every possible solution for V with cost more than $c\,\mathrm{cost}(V, C_{\mathrm{opt}})$ is either not a feasible solution for S or has cost more than $c\,\mathrm{cost}(S, C_{\mathrm{opt}})$ for S.

These three claims will allow us to conclude the analysis as follows. Since S contains a solution with cost at most $\frac{c}{\alpha}\,\mathrm{cost}(S, C_{\mathrm{opt}})$, $\mathbb{A}$ will compute a solution C^* with cost at most $c\,\mathrm{cost}(S, C_{\mathrm{opt}})$. Since every solution for V with cost more than $c\,\mathrm{cost}(V, C_{\mathrm{opt}})$ has cost more than $c\,\mathrm{cost}(S, C_{\mathrm{opt}})$ for S, we know that $\mathbb{A}$ computes a solution C^* with cost at most $c\,\mathrm{cost}(V, C_{\mathrm{opt}})$ for V. Hence, our sampling is a c-approximation algorithm.

We apply this approach to study sampling algorithms for four problems: the k-median problem, the k-means problem, the balanced k-median problem, and the min-sum k-clustering problem.

2. ANALYSIS OF THE *k*-MEDIAN PROBLEM

We first consider the k-median problem. A *k-median of V* is a set C of k points (*centers*) in V that minimizes the value of

$$\sum_{v \in V} \min_{1 \leq i \leq k} \mu(v, c_i) \equiv \sum_{v \in V} \mu(v, C).$$

The *k-median problem* is to compute a k-median for a given metric space (V, μ).

Let

$$\mathrm{med}_{\mathrm{opt}}(V, k) = \min_{C \subseteq V, |C| = k} \sum_{v \in V} \mu(v, C)$$

denote the cost of a k-median of V and let

$$\mathrm{med}_{\mathrm{avg}}(V, k) = \frac{1}{|V|}\,\mathrm{med}_{\mathrm{opt}}(V, k)$$

denote the average cost of a k-median of V. In a similar manner, for a given $U \subseteq V$ and $C \subseteq V$, we define the average cost of solution C to be

$$\mathrm{cost}_{\mathrm{avg}}(U, C) = \frac{1}{|U|} \sum_{v \in U} \mu(v, C).$$

The following theorem summarizes our analysis and it is the main result of this section.

Theorem 1. *Let (V, μ) be a metric space. Let $0 < \delta < 1$, $\alpha \geq 1$, $0 < \beta \leq 1$ and $\epsilon > 0$ be approximation parameters. Let $\mathbb{A}$ be an α-approximation algorithm for the k-median problem in metric spaces. If we choose a sample set $S \subseteq V$ of size s i.u.r., with*

$$s \geq \frac{c\alpha}{\beta}\left(k + \frac{\Delta}{\epsilon\beta}\left(\alpha\ln(1/\delta) + k\ln\left(\frac{k\Delta\alpha}{\epsilon\beta^2}\right)\right)\right),$$

for an appropriate constant c and we run algorithm $\mathbb{A}$ with input S, then for the solution C^* obtained by $\mathbb{A}$, with probability at least $1 - \delta$ it holds the following

$$\mathrm{cost}_{\mathrm{avg}}(V, C^*) \leq 2(\alpha + \beta)\,\mathrm{med}_{\mathrm{avg}}(V, k) + \epsilon.$$

To begin our analysis of the quality of the approximation of C^* and the proof of Theorem 1, let us introduce some basic notation.

Definition 2.1 (φ-good/bad solutions). A set of k centers C is a *φ-bad solution* of the k-median of V if $\mathrm{cost}_{\mathrm{avg}}(V, C) > \varphi\,\mathrm{med}_{\mathrm{avg}}(V, k)$. If C is not a φ-bad solution, then it is a *φ-good solution.*

For the k-median problem, we want to prove that for a certain sample size s our algorithm is a $(2(\alpha + \beta))$-approximation algorithm. Following the approach described in the previous section, we have to show that our sample set S contains w.h.p. a solution with cost at most $2(1 + \beta/\alpha)\,\mathrm{med}_{\mathrm{avg}}(V, k)$, and hence, any α-approximation for S returns a $2(\alpha + \beta)$-approximation for V w.h.p. We prove the following lemma.

Lemma 2.2. *Let S be a multiset of size $s \geq \frac{3\Delta\alpha(1+\alpha/\beta)\ln(1/\delta)}{\beta\,\mathrm{med}_{\mathrm{avg}}(V,k)}$ chosen from V i.u.r. If an α-approximation algorithm for k-median $\mathbb{A}$ is run on input S, then the following holds for the solution C^* returned by $\mathbb{A}$:* $\mathbf{Pr}\big[\mathrm{cost}_{\mathrm{avg}}(S, C^*) \leq 2(\alpha + \beta)\,\mathrm{med}_{\mathrm{avg}}(V, k)\big] \geq 1 - \delta$.

Proof. Let C_{opt} denote a k-median of V and let X_i denote the random variable for the distance of the i-th point in S to the nearest center of C_{opt}. Then, $\mathrm{cost}_{\mathrm{avg}}(S, C_{\mathrm{opt}}) = \frac{1}{s}\sum_{1\leq i\leq s} X_i$. Furthermore, since $\mathbf{E}[X_i] = \mathrm{med}_{\mathrm{avg}}(V, k)$, we also have $\mathrm{med}_{\mathrm{avg}}(V, k) = \frac{1}{s}\mathbf{E}\big[\sum X_i\big]$. Therefore,

$$\mathbf{Pr}\left[\mathrm{cost}_{\mathrm{avg}}(S, C_{\mathrm{opt}}) > \left(1 + \tfrac{\beta}{\alpha}\right)\mathrm{med}_{\mathrm{avg}}(V, k)\right] = \mathbf{Pr}\left[\sum_{1\leq i\leq s} X_i > \left(1 + \tfrac{\beta}{\alpha}\right)\mathbf{E}\left[\sum_{1\leq i\leq s} X_i\right]\right].$$

Observe that each X_i satisfies $0 \leq X_i \leq \Delta$. Therefore, we apply a Hoeffding bound (Lemma A.2) to obtain:

$$\mathbf{Pr}\left[\sum_{1\leq i\leq s} X_i > (1 + \beta/\alpha)\mathbf{E}\left[\sum_{1\leq i\leq s} X_i\right]\right] \leq e^{-\frac{s\,\mathrm{med}_{\mathrm{avg}}(V,k)\min\{(\beta/\alpha),(\beta/\alpha)^2\}}{3\Delta}} \leq \delta. \tag{1}$$

Let C be the set of k centers in S obtained by replacing each $c \in C_{\mathrm{opt}}$ by its nearest neighbor in S. By the triangle inequality, we get $\mathrm{cost}_{\mathrm{avg}}(S, C) \leq 2\,\mathrm{cost}_{\mathrm{avg}}(S, C_{\mathrm{opt}})$. Hence, multiset S contains a set of k centers whose cost is at most $2(1 + \beta/\alpha)\,\mathrm{med}_{\mathrm{avg}}(V, k)$ with probability at least $1 - \delta$. Therefore, the lemma follows because $\mathbb{A}$ returns an α-approximation C^* of the k-median for S. ■

Next, we show that any solution $C_b \subseteq S$ that is a $(2\alpha + 6\beta)$-bad solution of a k-median of V satisfies $\mathrm{cost}_{\mathrm{avg}}(S, C_b) > 2(\alpha + \beta)\,\mathrm{med}_{\mathrm{avg}}(V, k)$ with high probability.

Lemma 2.3. *Let S be a multiset of s points chosen i.u.r. from V with s such that*

$$s \ge c\left((1+\alpha/\beta)k + \frac{(\alpha+\beta)\Delta\left(\ln(1/\delta) + k\ln\left(\frac{k(\alpha+\beta)\Delta}{\beta^2\,\mathrm{med}_{\mathrm{avg}}(V,k)}\right)\right)}{\beta^2\,\mathrm{med}_{\mathrm{avg}}(V,k)}\right),$$

where c is a certain positive constant. Let $\mathbb{C}$ be the set of $(2\alpha + 6\beta)$-bad solutions C of a k-median of V. Then,

$$\mathbf{Pr}\left[\exists C_b \in \mathbb{C} : C_b \subseteq S \textit{ and } \mathrm{cost}_{\mathrm{avg}}(S, C_b) \le 2(\alpha+\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right] \le \delta.$$

Proof. We choose c so that $s \ge \frac{2\alpha+3\beta}{\beta}k$. Let us consider an arbitrary solution C_b that is a $(2\alpha+6\beta)$-bad solution of a k-median of V and let S^* be a multiset of $s-k$ points chosen i.u.r from V. Then,

$$\begin{aligned}
&\mathbf{Pr}\left[C_b \subseteq S \text{ and } \mathrm{cost}_{\mathrm{avg}}(S, C_b) \le 2(\alpha+\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right] \\
&\quad = \mathbf{Pr}\left[\mathrm{cost}_{\mathrm{avg}}(S, C_b) \le 2(\alpha+\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k) \;\middle|\; C_b \subseteq S\right] \mathbf{Pr}\left[C_b \subseteq S\right] \\
&\quad = \mathbf{Pr}\left[\mathrm{cost}_{\mathrm{avg}}(S^*, C_b) \le 2\frac{s}{s-k}((\alpha+\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k))\right] \mathbf{Pr}\left[C_b \subseteq S\right] && (2) \\
&\quad \le \mathbf{Pr}\left[\mathrm{cost}_{\mathrm{avg}}(S^*, C_b) \le 2((\alpha+1.5\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k))\right] \mathbf{Pr}\left[C_b \subseteq S\right], && (3)
\end{aligned}$$

where (2) holds because $(s-k)\,\mathrm{cost}_{\mathrm{avg}}(S^*, C_b) = \mathrm{cost}(S^*, C_b) = \mathrm{cost}(S^* \cup C_b, C_b) = s\,\mathrm{cost}_{\mathrm{avg}}(S^* \cup C_b, C_b)$ and the elements are chosen with repetition, and (3) follows from $s \ge \frac{2\alpha+3\beta}{\beta}k$.

Next, similar to the proof of Lemma 2.2, we prove the following inequality

$$\mathbf{Pr}\left[\mathrm{cost}_{\mathrm{avg}}(S^*, C_b) \le 2(\alpha+1.5\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right] \le e^{\frac{-s\beta^2\,\mathrm{med}_{\mathrm{avg}}(V,k)}{2(\alpha+\beta)\Delta}}. \qquad (4)$$

To prove this, let us denote by X_i the random variable for the distance of the i-th point in S^* to the nearest center of C_b. Since C_b is a $(2\alpha+6\beta)$-bad solution of a k-median of V, we have $\mathbf{E}[X_i] > (2\alpha+6\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)$. Therefore, we have

$$\begin{aligned}
&\mathbf{Pr}\left[\mathrm{cost}_{\mathrm{avg}}(S^*, C_b) \le 2(\alpha+1.5\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right] \\
&\quad = \mathbf{Pr}\left[\sum_{i=1}^{s-k} X_i \le (s-k)2(\alpha+1.5\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right] \\
&\quad \le \mathbf{Pr}\left[\sum_{i=1}^{s-k} X_i \le \left(1 - \frac{3\beta}{2\alpha+6\beta}\right)\mathbf{E}\left[\sum_{i=1}^{s-k} X_i\right]\right].
\end{aligned}$$

Next, since $0 \le X_i \le \Delta$, we apply Hoeffding bound (Lemma A.2) to the above to obtain

$$\begin{aligned}
\mathbf{Pr}\left[\mathrm{cost}_{\mathrm{avg}}(S^*, C_b) \le 2(\alpha+1.5\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right] &\le \exp\left(-\frac{\left(\frac{3\beta}{2\alpha+6\beta}\right)^2 \mathbf{E}\left[\sum_{i=1}^{s-k} X_i\right]}{2\Delta}\right) \\
&\le \exp\left(-\frac{9\beta^2(s-k)\,\mathrm{med}_{\mathrm{avg}}(V,k)}{4\Delta(\alpha+3\beta)}\right),
\end{aligned}$$

where the last inequality follows from $\mathbf{E}[X_i] > (2\alpha + 6\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)$. Since $s \geq \frac{2\alpha+3\beta}{\beta}k$ implies that $s - k \geq \frac{2}{3}s$, the inequality above yields (4).

Once we have inequality (4), we can combine it with the inequality $\mathbf{Pr}[C_b \subseteq S] \leq \binom{s}{k}/\binom{n}{k}$ into (3), and then apply there the upper bound $|\mathbb{C}| \leq \binom{n}{k}$ to conclude:

$$\begin{aligned}
&\mathbf{Pr}\left[\exists C_b \in \mathbb{C} : C_b \subseteq S \text{ and } \mathrm{cost}_{\mathrm{avg}}(S, C_b) \leq 2(\alpha+\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right] \\
&\quad\leq \sum_{C_b \in \mathbb{C}} \mathbf{Pr}\left[C_b \subseteq S \text{ and } \mathrm{cost}_{\mathrm{avg}}(S, C_b) \leq 2(\alpha+\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right] \\
&\quad\leq \sum_{C_b \in \mathbb{C}} \mathbf{Pr}\left[\mathrm{cost}_{\mathrm{avg}}(S^*, C_b) \leq 2\left(\alpha + 1.5\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right)\right] \mathbf{Pr}\left[C_b \subseteq S\right] \\
&\quad\leq \sum_{C_b \in \mathbb{C}} \exp\left(-\frac{\beta^2 s\,\mathrm{med}_{\mathrm{avg}}(V,k)}{2\Delta(\alpha+\beta)}\right) \frac{\binom{s}{k}}{\binom{n}{k}} \\
&\quad\leq \binom{n}{k} \exp\left(-\frac{\beta^2 s\,\mathrm{med}_{\mathrm{avg}}(V,k)}{2\Delta(\alpha+\beta)}\right) \frac{\binom{s}{k}}{\binom{n}{k}} \\
&\quad\leq s^k \exp\left(-\frac{\beta^2 s\,\mathrm{med}_{\mathrm{avg}}(V,k)}{2\Delta(\alpha+\beta)}\right),
\end{aligned}$$

which is smaller than δ for

$$s \geq \max\left\{\frac{4k\Delta(\alpha+\beta)}{\beta^2\,\mathrm{med}_{\mathrm{avg}}(V,k)} \ln\left(\frac{4k\Delta(\alpha+\beta)}{\beta^2\,\mathrm{med}_{\mathrm{avg}}(V,k)}\right), \frac{2\Delta(\alpha+\beta)\ln(1/\delta)}{\beta^2\,\mathrm{med}_{\mathrm{avg}}(V,k)}\right\}.$$

This implies Lemma 2.3. ▪

Proof of Theorem 1. Let β^* be a positive parameter that will be set later in the proof. Let s be chosen such that the prerequisites of Lemmas 2.2 and 2.3 hold with β replaced by β^*, that is,

$$s \geq c(1+\alpha/\beta^*)\left(k + \frac{\Delta}{\beta^*\,\mathrm{med}_{\mathrm{avg}}(V,k)}\left(\alpha\ln(1/\delta) + k\ln\left(\frac{k(\alpha+\beta^*)\Delta}{(\beta^*)^2\,\mathrm{med}_{\mathrm{avg}}(V,k)}\right)\right)\right) \tag{5}$$

for certain constant c. Let S be a multiset of s points chosen i.u.r. from V. Then, by Lemma 2.3 with probability at least $1-\delta$, no set $C \subseteq S$ that is a $(2\alpha + 6\beta^*)$-bad solution of a k-median of V satisfies the inequality

$$\mathrm{cost}_{\mathrm{avg}}(S, C) \leq 2(\alpha+\beta^*)\,\mathrm{med}_{\mathrm{avg}}(V,k).$$

On the other hand, if we run algorithm $\mathbb{A}$ for set S, then by Lemma 2.2, the resulting set C^* of k centers with probability at least $1-\delta$ satisfies

$$\mathrm{cost}_{\mathrm{avg}}(S, C^*) \leq 2(\alpha+\beta^*)\,\mathrm{med}_{\mathrm{avg}}(V,k).$$

This, together with the claim above implies that with probability at least $1-2\delta$ the set C^* is a $(6\beta^*, 2\alpha)-$ good solution of a k-median of V, that is,

$$\mathbf{Pr}\left[\mathrm{cost}_{\mathrm{avg}}(V, C^*) \leq (2\alpha + 6\beta^*)\,\mathrm{med}_{\mathrm{avg}}(V,k)\right] \geq 1 - 2\delta. \tag{6}$$

To complete the proof, we must include the only parameters β and ϵ and remove the dependence of $\mathrm{med}_{\mathrm{avg}}(V,k)$ in the bound of s in (5).

Let us first consider the case when $\mathrm{med}_{\mathrm{avg}}(V,k) \le \epsilon$. We use (5) and (6) with $\beta^* = \frac{1}{6}\epsilon/\mathrm{med}_{\mathrm{avg}}(V,k)$, and since $\beta^* \ge 1/6$, we will obtain that if

$$s \ge c(1+\alpha)\left(k + \frac{\Delta}{\epsilon}\left(\alpha\ln(1/\delta) + k\ln\left(\frac{k(\alpha+1)\Delta}{\epsilon}\right)\right)\right),$$

for certain positive constant c, then with probability at least $1-2\delta$ we have

$$\mathrm{cost}_{\mathrm{avg}}(V,C^*) \le (2\alpha + 6\beta^*)\,\mathrm{med}_{\mathrm{avg}}(V,k) = 2\alpha\,\mathrm{med}_{\mathrm{avg}}(V,k) + \epsilon.$$

Notice that this bound is independent of β.

Next, we consider the case when $\mathrm{med}_{\mathrm{avg}}(V,k) > \epsilon$. Then, by (5) and (6), we have that for a certain constant $c > 0$ and for $\beta = 3\beta^*$, if

$$s \ge c(1+\alpha/\beta)\left(k + \frac{\Delta}{\beta\epsilon}\left(\alpha\ln(1/\delta) + k\ln\left(\frac{k\Delta(1+\alpha/\beta)}{\beta^2\epsilon}\right)\right)\right),$$

then with probability at least $1-2\delta$ we have

$$\mathrm{cost}_{\mathrm{avg}}(V,C^*) \le 2(\alpha+\beta)\,\mathrm{med}_{\mathrm{avg}}(V,k).$$

Theorem 1 follows by combining these two bounds. ∎

3. *k*-MEDIAN APPROXIMATION IN EUCLIDEAN SPACES

Our result from the previous section can be improved if we consider the k-median problem in Euclidean spaces $\mathbb{R}^d$. Let us remind that the Euclidean k-median problem is for an input set V of n points in $\mathbb{R}^d$ to find k centers $C = \{c_1,\dots,c_k\} \subseteq \mathbb{R}^d$ that minimize $\mathrm{med}^{E^d}_{\mathrm{opt}}(V,k) = \sum_{v\in V}\min_{1\le i\le k}\|v-c_i\|_2$, where $\|v-c_i\|_2$ denote the Euclidean distance between v and c_i. (Notice that in the definition of the k-median problem used in Section 2 we required that $C \subseteq V$; now, we require only $C \subseteq \mathbb{R}^d$.)

Let $\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k) = \frac{1}{|V|}\mathrm{med}^{E^d}_{\mathrm{opt}}(V,k)$ and $\mathrm{cost}^{E^d}_{\mathrm{avg}}(S,C) = \frac{1}{|S|}\sum_{u\in S}\min_{1\le i\le k}\|u-c_i\|_2$. Then, we can prove the following analogue of Theorem 1.

Theorem 2. *Let V be a subset of $\mathbb{R}^d$ of size n. Let $0 < \delta < 1$, $\alpha \ge 1$, $\beta \le 1$, and $\epsilon > 0$ be approximation parameters. Let $\mathbb{A}$ be an α-approximation algorithm for the Euclidean k-median problem in $\mathbb{R}^d$. If we choose a sample set $S \subseteq V$ of size s i.u.r., where*

$$s \ge \frac{c\alpha}{\beta}\left(k + \frac{\Delta}{\beta\epsilon}\left(kd\ln(\sqrt{d}\Delta/\epsilon) + \ln(1/\delta)\right)\right),$$

and we run algorithm $\mathbb{A}$ with input S, then for the solution C^* obtained by $\mathbb{A}$, with probability at least $1-\delta$ it holds the following,

$$\mathrm{cost}^{E^d}_{\mathrm{avg}}(V,C^*) \le (\alpha+\beta)\,\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k) + \epsilon.$$

Proof. The proof is almost identical to the proof of Theorem 1 with the exception of a few minor modifications. First of all, since we do not require $C \subseteq V$, in Lemma 2.2 we do not need factor 2 in the approximation bound in that lemma. Therefore, in particular, the bound corresponding to Lemma 2.2 is now as follows: if $s \ge \frac{3\Delta\alpha(1+\alpha/\beta)\ln(1/\delta)}{\beta\,\mathrm{med}_{\mathrm{avg}}(V,k)^{E^d}}$, then $\mathbf{Pr}\big[\mathrm{cost}^{E^d}_{\mathrm{avg}}(S,C^*) \le (\alpha+\beta)\,\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k)\big] \ge 1-\delta$.

Next, we proceed to the analysis corresponding to that in Lemma 2.3. All the arguments from that lemma can be used here with the exception of the bound of s^k for the number of locations for k centers contained in S, which is in general not true for the Euclidean k-median problem. However, we can use a standard observation in similar situations, and consider only centers having points on a certain grid in $\mathbb{R}^d$. Indeed, since we know that the input is contained in a d-dimensional cube of side length Δ, we can put a d-dimensional grid with $\big(\frac{\sqrt{d}\Delta}{2\epsilon}\big)^d$ grid points to obtain an additional additive error in k-median of at most ϵn (see also [28, Section 3.2]). Therefore, the number of k centers locations can be upper bounded by $\big(\frac{\sqrt{d}\Delta}{2\epsilon}\big)^{kd}$. With this upper bound, if $\mathbb{C}$ is the set of $(\alpha+3\beta)$-bad solutions of a k-median of V, then

$$\mathbf{Pr}\big[\exists C_b \in \mathbb{C} : \mathrm{cost}^{E^d}_{\mathrm{avg}}(S,C_b) \le (\alpha+\beta)\,\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k)+\epsilon\big]$$
$$\le \left(\frac{\sqrt{d}\Delta}{2\epsilon}\right)^{kd} \exp\left(\frac{-s\beta^2\,\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k)}{(\alpha+3\beta)\Delta}\right).$$

If we use this bound to obtain the result corresponding to Lemma 2.3, then we obtain, that if

$$s \ge c\left((1+\alpha/\beta)k + \frac{kd\Delta(\alpha+\beta)\ln(\sqrt{d}\Delta/\epsilon)}{\beta^2\,\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k)} + \frac{(\alpha+\beta)\Delta}{\beta^2\,\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k)}\ln(1/\delta)\right)$$

then

$$\mathbf{Pr}\big[\exists C_b \in \mathbb{C} : \mathrm{cost}^{E^d}_{\mathrm{avg}}(S,C_b) \le (\alpha+\beta)\,\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k)+\epsilon\big] \le \delta.$$

To complete the proof, we use identical arguments as those used at the end of Section 2. If we consider separately the cases when $\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k) \le \epsilon$ and when $\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k) > \epsilon$, then using arguments from Section 2, we obtain that if

$$s \ge \frac{c\alpha}{\beta}\left(k + \frac{\Delta}{\beta\epsilon}\left(kd\ln(\sqrt{d}\Delta/\epsilon) + \ln(1/\delta)\right)\right)$$

then

$$\mathbf{Pr}\big[\exists C_b \in \mathbb{C} : \mathrm{cost}^{E^d}_{\mathrm{avg}}(S,C_b) \le (\alpha+\beta)\,\mathrm{med}^{E^d}_{\mathrm{avg}}(V,k)+\epsilon\big] \le \delta,$$

what completes the proof of Theorem 2. ▪

4. EXTENSION TO *k*-MEANS CLUSTERING IN METRIC SPACES

The analysis from the previous two sections can be extended to the k-means problem in a straightforward way. Indeed, the k-means problem is identical to the k-median problem with the distance function to be the square of the original "metric distance", i.e., the objective is to find a set C of k centers that minimizes

$$\sum_{v\in V}(\mu(v,C))^2.$$

Similarly to the k-median problem we define the average cost of an optimal k-means solution for the metric space (V, μ) as

$$\text{mean}_{\text{avg}}(V, k) = \frac{1}{|V|} \min_{C \subseteq V, |C|=k} \sum_{v \in V} (\mu(v, C))^2.$$

Next, we define the average cost as

$$(\text{cost}_{\text{avg}})^2(V, C) = \frac{1}{|V|} \sum_{v \in V} (\mu(v, C))^2.$$

It is well-known (and easy to see) that the obtained distance function does not define a metric space, but it is almost a metric: all properties of the metric are satisfied, except that the triangle inequality holds to within a factor 2. Therefore, all our analyses from Sections 2 and 3 holds with a few basic modifications, as described below.

We first prove a modification of Lemma 2.2 that if a multiset S of size $s \geq \frac{3\Delta^2\alpha(1+\alpha/\beta)\ln(1/\delta)}{\beta\,\text{mean}_{\text{avg}}(V,k)}$ is chosen from V i.u.r. then if we run an α-approximation algorithm for k-means $\mathbb{A}$ on input S, then for the solution C^* obtained by $\mathbb{A}$ holds

$$\mathbf{Pr}\left[(\text{cost}_{\text{avg}})^2(S, C^*) \leq 4(\alpha + \beta)\,\text{mean}_{\text{avg}}(V, k)\right] \geq 1 - \delta, \tag{7}$$

where for any $U \subseteq V$, $C' \subseteq V$, $(\text{cost}_{\text{avg}})^2(U, C') = \frac{1}{|U|} \sum_{v \in U} (\mu(u, C'))^2$.

The arguments in the proof are identical to those used in the proof of Lemma 2.2 with two simple differences. First of all, the range of each X_i is now $[0, \Delta^2]$, which is caused by the fact that any square of the distance between a pair of points in V is in that range. This allows us to replace (1) by the following inequality:

$$\mathbf{Pr}\left[\sum_{1 \leq i \leq s} X_i > (1 + \beta/\alpha)\mathbf{E}\left[\sum_{1 \leq i \leq s} X_i\right]\right] \leq e^{-\frac{s\,\text{mean}_{\text{avg}}(V,k)\min\{(\beta/\alpha),(\beta/\alpha)^2\}}{3\Delta^2}} \leq \delta.$$

The second modification is caused by the fact that the triangle inequality does not hold for the k-means problem, but instead we have to use a weakened triangle inequality. And so, following the arguments at the end of the proof of Lemma 2.2, the "weak" triangle inequality implies that $(\text{cost}_{\text{avg}})^2(S, C) \leq 4(\text{cost}_{\text{avg}})^2(S, C_{\text{opt}})$, and hence, the random multiset S contains a set of k centers whose cost is at most $4(1 + \beta/\alpha)\,\text{mean}_{\text{avg}}(V, k)$ with probability at least $1 - \delta$. With these two modifications, we can obtain the variant of Lemma 2.2 with inequality (7) as described earlier.

Next, we use identical changes to prove a modification of Lemma 2.3, that if for a certain constant c, a multiset S of size $s \geq c\Big((1 + \beta/\alpha)k + \frac{(\alpha+\beta)\Delta^2(\ln(1/\delta)+k\ln(k(\alpha+\beta)\Delta/(\beta^2\,\text{mean}_{\text{avg}}(V,k)))}{\beta^2\,\text{mean}_{\text{avg}}(V,k)}\Big)$ is chosen from V i.u.r., then

$$\mathbf{Pr}\left[\exists C_b \in \mathbb{C} : C_b \subseteq S \text{ and } (\text{cost}_{\text{avg}})^2(S, C_b) \leq 4(\alpha + \beta)\,\text{mean}_{\text{avg}}(V, k)\right] \leq \delta,$$

where $\mathbb{C}$ is the set of $(4\alpha + 12\beta)$-bad solutions of a k-mean of V.

Once we have these two modified versions of Lemmas 2.2 and 2.3, we proceed as in the final arguments of the proof of Theorem 1 to conclude with the following result.

Theorem 3. *Let (V, μ) be a metric space. Let $0 < \delta < 1$, $\alpha \ge 1$, $0 < \beta \le 1$ and $\epsilon > 0$ be approximation parameters. Let $\mathbb{A}$ be an α-approximation algorithm for the k-means problem in metric spaces. If we choose a sample set $S \subseteq V$ of size s i.u.r., with*

$$s \ge \frac{c\alpha}{\beta}\left(k + \frac{\Delta^2}{\epsilon\beta}\left(\alpha\ln(1/\delta) + k\ln\left(\frac{k\Delta^2\alpha}{\epsilon\beta^2}\right)\right)\right),$$

for some constant c and we run algorithm $\mathbb{A}$ with input S, then for the solution C^* obtained by $\mathbb{A}$, with probability at least $1-\delta$ it holds the following

$$(\text{cost}_{\text{avg}})^2(V, C^*) \le 4(\alpha+\beta)\,\text{mean}_{\text{avg}}(V,k) + \epsilon.$$

As in the case of the k-median problem, we can extend our analysis to the k-means problem for Euclidean metrics. Like in the Euclidean variant of the k-median problem we allow that the set of centers C is any subset of the Euclidean space $\mathbb{R}^d$. For a point set $V \subseteq \mathbb{R}^d$ and an arbitrary set C of points in the $\mathbb{R}^d$, we define

$$(\text{cost}^{E^d}_{\text{avg}})^2(V, C) = \sum_{v\in V}(\mu(V, C))^2$$

and

$$\text{mean}^{E^d}_{\text{avg}}(V, k) = \min_{C\subseteq\mathbb{R}^d, |C|=k} (\text{cost}^{E^d}_{\text{avg}})^2(V, C).$$

Unlike in the general case, we do not lose an additional factor of 2 compared to the k-median solution. The reason is that like in the Euclidean variant of the k-median problem, we do not have to move each center of the optimal solution to the closest point in the sample. Furthermore, it suffices again to consider only solution on a grid with $\left(\frac{\sqrt{d}\Delta}{2\epsilon}\right)^d$ points. The reason is that for any set of points $V \subseteq \mathbb{R}^d$ and any $y \in \mathbb{R}^d$ we have $\sum_{v\in V}(\mu(v,y))^2 = \sum_{v\in V}(\mu(v,c))^2 + |V|(\mu(c,y))^2$. Thus by considering only solutions on the grid we lose at most an additive term $\epsilon^2 n$. The following extension of Theorem 3 to Euclidean metrics follows by a straightforward application of techniques from Section 3 to the arguments above.

Theorem 4. *Let V be a subset of $\mathbb{R}^d$ of size n. Let $0 < \delta < 1$, $\alpha \ge 1$, $\beta \le 1$, and $\epsilon > 0$ be approximation parameters. Let $\mathbb{A}$ be an α-approximation algorithm for the Euclidean k-means problem in $\mathbb{R}^d$. If we choose a sample set $S \subseteq V$ of size s i.u.r., where*

$$s \ge \frac{c\alpha}{\beta}\left(k + \frac{\Delta^2}{\beta\epsilon}\left(kd\ln(\sqrt{d}\Delta^2/\epsilon) + \ln(1/\delta)\right)\right),$$

and we run algorithm $\mathbb{A}$ with input S, then for the solution C^* obtained by $\mathbb{A}$, with probability at least $1-\delta$ it holds the following,

$$\left(\text{cost}^{E^d}_{\text{avg}}\right)^2(V, C^*) \le (\alpha+\beta)\,\text{mean}^{E^d}_{\text{avg}}(V,k) + \epsilon^2.$$

5. MIN-SUM *k*-CLUSTERING AND BALANCED *k*-MEDIAN IN METRIC SPACES

As we mentioned in Introduction, we follow the approach from [4] and [13] and consider the balanced k-median problem, instead of analyzing min-sum k-clustering.

Let (V, μ) be a metric space. A *balanced k-median of V* is a set $C = \{c_1, \ldots, c_k\}$ of k points (centers) in V that minimizes the value of

$$\min_{\text{partition of } V \text{ into } V_1 \cup \cdots \cup V_k} \sum_{i=1}^{k} |V_i| \sum_{u \in V_i} \mu(u, c_i).$$

The *balanced k-median problem* is for a given (V, μ) to compute a balanced k-median of V and a partition of V into $V_1 \cup \cdots \cup V_k$ that minimizes the sum above. Unless stated otherwise, we will not require $c_i \in V_i$.

Let

$$\text{med}^b_{\text{opt}}(V, k) = \min_{C=\{c_1,\ldots,c_k\} \subseteq V} \min_{\text{partition of } V \text{ into } V_1 \cup \cdots \cup V_k} \sum_{i=1}^{k} |V_i| \sum_{u \in V_i} \mu(u, c_i)$$

denote the *cost of a balanced k-median of V*, and let

$$\text{med}^b_{\text{avg}}(V, k) = \frac{1}{|V|^2} \text{med}^b_{\text{opt}}(V, k)$$

denote the *average cost of a balanced k-median of V*. For a given set $U \subseteq V$ and a set of k centers $C = \{c_1, \ldots, c_k\} \subseteq V$, let us define

$$\text{cost}^b(U, C) = \min_{\substack{\text{partition of } U \\ \text{into } U_1 \cup \cdots \cup U_k}} \sum_{i=1}^{k} |U_i| \sum_{u \in U_i} \mu(u, c_i), \text{ and}$$

$$\text{cost}^b_{\text{avg}}(U, C) = \frac{\text{cost}^b(U, C)}{|U|^2}.$$

We will also use the following notation.

Definition 5.1 ((ϵ, φ)-good/bad solution). A set of k centers C is called a (ϵ, φ)-*bad solution* of balanced k-median of V if $\text{cost}^b_{\text{avg}}(V, C) > \varphi\, \text{med}^b_{\text{avg}}(V, k) + \epsilon$.

If C is not a (ϵ, φ)-bad solution then it is a (ϵ, φ)-*good solution*.

5.1. Sampling Algorithms for the Balanced *k*-Median Problem in Metric Spaces

Our high level approach of analyzing the balanced k-median problem is essentially the same as for the k-median problem. We investigate the generic sampling scheme described in Section 1.3, and in Sections 5.3 and 5.4 we prove the following main theorem.

Theorem 5. *Let (V, μ) be a metric space. Let $\mathbb{A}$ be an α-approximation algorithm for balanced k-median in metric spaces and let $0 < \delta < 1$, $0 < \epsilon$, and $0 < \beta < \alpha$, be approximation parameters. If we choose a sample set $S \subseteq V$ of size s i.u.r., where*

$$s \geq \frac{c\Delta(k + \ln(1/\delta)) \ln\left(\frac{\alpha k \Delta \ln(1/\delta)}{\beta\epsilon}\right)}{\epsilon} \left(\frac{\alpha^2}{\beta^2} + \frac{\Delta k^2}{\epsilon}\right),$$

for a suitable positive constant c, and we run algorithm $\mathbb{A}$ with input S, then for the solution C^* obtained by $\mathbb{A}$, with probability at least $1 - \delta$ it holds the following

$$\text{cost}^b_{\text{avg}}(V, C^*) \leq 2(\alpha + \beta)\, \text{med}^b_{\text{avg}}(V, k) + \epsilon.$$

The same result (with another constant c) holds for the case when the centers c_i are required to be in the corresponding clusters V_i. Furthermore, in time $O(nk) + \widetilde{O}(k^{2.5}n^{0.5})$ one can find a clustering of V that satisfies the above approximation guarantee.

Since in metric spaces the solution to balanced k-median is within a factor of 2 of that of min-sum k-clustering, Theorem 5 immediately implies the following theorem.

Theorem 6. *Let (V, μ) be a metric space. Let $\mathbb{A}$ be an α-approximation algorithm for min-sum k-clustering in metric spaces and let $0 < \delta < 1$, $\epsilon > 0$, and $0 < \beta < \alpha$, be approximation parameters. If we choose a sample set $S \subseteq V$ of size s i.u.r., where*

$$s \geq \frac{c\Delta(k + \ln(1/\delta)) \ln\left(\frac{\alpha k \Delta \ln(1/\delta)}{\beta\epsilon}\right)}{\epsilon}\left(\frac{\alpha^2}{\beta^2} + \frac{\Delta k^2}{\epsilon}\right),$$

for a suitable positive constant c, and we run algorithm $\mathbb{A}$ with input S, then for the solution C^* obtained by $\mathbb{A}$, with probability at least $1 - \delta$ it holds the following

$$\mathrm{cost}^b_{\mathrm{avg}}(V, C^*) \leq 4(\alpha + \beta)\,\mathrm{med}^b_{\mathrm{avg}}(V, k) + \epsilon.$$

Furthermore, in time $O(nk) + \widetilde{O}(k^{2.5}n^{0.5})$ one can find a clustering of V that satisfies the above approximation guarantee.

5.2. Overview of the Analysis

Our analysis follows the path used in Section 2. However, the analysis for the balanced k-median problem is more difficult. The reason is that we do not know much about the structure of an optimal solution for a given set of centers. The main idea will be to impose constraints on the sizes of the clusters such that on the one hand the number of distinct solutions does not increase too much and on the other hand every potential solution is covered.

In particular, to show that a given bad solution $C_b = \{c_1, \dots, c_k\}$ for the entire point set is also a bad solution for our sample set S (with high probability), we have to consider several potential clusterings of S. To tackle this, we group clusterings of S by their *cardinality vectors* $s^* = (s^*_1, \dots, s^*_k)$, where $\sum_{i=1}^k s^*_i = s = |S|$. We show that any clustering $S_1, \dots, S_k$ of S with centers $c_1, \dots, c_k$ that satisfies $|S_i| = s^*_i$ is not much smaller than the average cost of the best clustering of V with centers C_b, where certain restrictions, induced by the cardinality constraints on the $|S_i|$s, are imposed on the $|V_i|$. Ideally, one would like to set $|V_i| = \frac{s^*_i|V|}{s} = x_i$. The intuition behind this is that, if $V_1, \dots, V_k$ is an optimal such cardinality-constrained clustering, then if we set $S_i = S \cap V_i$ and assume that things behave according to expectation, we would have $|S_i| = s^*_i$, so $S_1, \dots, S_k$ would be an optimal cardinality-constrained clustering of S (by Lemma 5.6); thus, $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*)$ is precisely the cost of this clustering, which is easily bound in terms of the cost of the clustering $V_1, \dots, V_k$ (which is at least $\mathrm{cost}^b_{\mathrm{avg}}(V, C_b)$). However, x_i need not be an integer and random variables need not behave like their expectation. Therefore, one considers the integer vector x' closest to x (such that $\sum_{i=1}^k x'_i = |V|$) and enforces that $|V_i| = x'_i$, and then uses Lemma 5.4 and 5.5 to argue that $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*)$ is close to $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, x', x')$.

5.3. Good Solution for S That Is Also a Good Solution for V

We begin with a result corresponding to Lemma 2.2 for k-median that shows that there exists a clustering for S whose cost is a good approximation of the optimal clustering for V.

Lemma 5.2. *Let C_{opt} be a balanced k-median of V. Let $0 < \gamma \le \frac{1}{2}, 0 < \delta < 1, \epsilon > 0$ be arbitrary parameters. If we choose a multiset $S \subseteq V$ of size $s \ge \frac{6\alpha k \Delta \ln(3k/\delta)}{\gamma\epsilon}$ i.u.r., then*

$$\mathbf{Pr}\left[\mathrm{cost}^b_{\mathrm{avg}}(S, C_{\mathrm{opt}}) \le (1+\gamma)^3 \,\mathrm{med}^b_{\mathrm{avg}}(V,k) + \frac{6k\Delta\ln(3k/\delta)}{\gamma^2 s^2} + \frac{3\epsilon}{2\alpha}\right] \ge 1-\delta.$$

Proof. To simplify the notation, let $\delta_1 = \frac{1}{3}\delta/k$. Let $C_{\mathrm{opt}} = \{c_1, \dots, c_k\}$. Let $V_1^* \cup \dots \cup V_k^*$ be the optimal partition of V, i.e., $\mathrm{med}^b_{\mathrm{opt}}(V,k) = \sum_{i=1}^k |V_i^*| \sum_{u \in V_i^*} \mu(u, c_i)$.

Let us call set V_i^* *dense* if $|V_i^*| \ge \frac{3\ln(1/\delta_1)}{\gamma^2}\frac{|V|}{s}$; V_i^* is *sparse* otherwise. Let S_i be the random variable denoting the multiset $S \cap V_i^*$ (we assume S_i is a multiset, that is, an element can appear multiple times in S_i if it belongs to V_i^* and it appears multiple times in S).

Our first observation is that if V_i^* is dense, then we have $\mathbf{Pr}\big[|S_i| \le (1-\gamma)\frac{s|V_i^*|}{|V|}\big] \le \delta_1$ and $\mathbf{Pr}\big[|S_i| \ge (1+\gamma)\frac{s|V_i^*|}{|V|}\big] \le \delta_1$, and if V_i^* is sparse, then we have $\mathbf{Pr}\big[|S_i| \ge \frac{6\ln(1/\delta_1)}{\gamma^2}\big] \le \delta_1$. To see this, let us first recall that $\mathbf{E}[|S_i|] = s\frac{|V_i^*|}{|V|}$. Hence, by Chernoff bound (Lemma A.1),

$$\mathbf{Pr}\left[|S_i| \le (1-\gamma)\tfrac{s|V_i^*|}{|V|}\right] \le \exp(-\gamma^2 s|V_i^*|/(2|V|)) \le \delta_1,$$

and similarly,

$$\mathbf{Pr}\left[|S_i| \ge (1+\gamma)\tfrac{s|V_i^*|}{|V|}\right] \le \exp(-\gamma^2 s|V_i^*|/(3|V|)) \le \delta_1.$$

Next, let us consider sets V_i^* that are sparse. For any such a set, since $|V_i^*| < \frac{3\ln(1/\delta_1)}{\gamma^2}\frac{|V|}{s}$, we have $\mathbf{E}[|S_i|] < \frac{3\ln(1/\delta_1)}{\gamma^2}$. Therefore, by Chernoff bound (Lemma A.1), we obtain

$$\mathbf{Pr}\left[|S_i| \ge \frac{6\ln(1/\delta_1)}{\gamma^2}\right] \le \exp\left(-\frac{\ln(1/\delta_1)}{\gamma^2}\right) \le \delta_1,$$

where the last inequality holds because $\gamma < 1$.

In view of these bounds, from now on, let us condition on the event that for dense sets V_i^* we have $(1-\gamma)\frac{s|V_i^*|}{|V|} < |S_i| < (1+\gamma)\frac{s|V_i^*|}{|V|}$ and for sparse sets V_i^* we have $|S_i| < \frac{6\ln(1/\delta_1)}{\gamma^2}$. This event holds with probability at least $1 - 2k\delta_1$.

Next, we observe that conditioned on the size r of S_i, the j-th element of S_i is uniformly distributed in V_i^*. To see this, let us consider S to be an ordered multiset. Then, any sequence of s points from V is equally likely to be chosen as S. Now, consider a sequence of points that has exactly r points in V_i^*. We observe that by replacing the j-th point in this sequence with an arbitrary point from V_i we obtain a sequence with r points in V_i^*, which has the same probability to be chosen. Hence, the distribution of the j-th point from S_i (conditioned on the size of S_i) is the uniform distribution in V_i^*.

Thus, for any set V_i^*, let X_i^j be a random variable that denotes the distance between a point selected independently and uniformly at random from V_i and the center c_i. Observe that for any set V_i^*, we have $\mathbf{E}[X_i^j] = \frac{1}{|V_i^*|}\sum_{u\in V_i^*}\mu(u,c_i)$. Let us fix i. We first consider the case when

$$2\frac{|S_i|}{s^2}\gamma\frac{|S_i|}{|V_i^*|}\sum_{u\in V_i^*}\mu(u,c_i) \geq \frac{\epsilon}{\alpha k}. \tag{8}$$

Since $0 \leq X_i^j \leq \Delta$ and $0 < \gamma < 1$, we use Hoeffding bound (Lemma A.2) to obtain

$$\begin{aligned}\mathbf{Pr}\left[\sum_{j=1}^{|S_i|}X_i^j \geq \frac{(1+\gamma)|S_i|\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|}\right] &= \mathbf{Pr}\left[\sum_{j=1}^{|S_i|}X_i^j \geq (1+\gamma)\mathbf{E}\left[\sum_{j=1}^{|S_i|}X_i^j\right]\right]\\ &\leq \exp\left(-\frac{\gamma^2}{3\Delta}\mathbf{E}\left[\sum_{j=1}^{|S_i|}X_i^j\right]\right)\\ &= \exp\left(-\frac{\gamma^2}{3\Delta}\frac{|S_i|\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|}\right)\\ &\leq \exp\left(-\frac{\gamma s\epsilon}{6\Delta\alpha k}\right),\end{aligned} \tag{9}$$

where the last inequality follows from (8) and from the fact that $|S_i| \leq s$.

If (8) does not hold, then let us choose γ^*, $\gamma^* > \gamma$, such that

$$2\frac{|S_i|}{s^2}\gamma^*\frac{|S_i|}{|V_i^*|}\sum_{u\in V_i^*}\mu(u,c_i) = \frac{\epsilon}{\alpha k}.$$

Notice that in that case,

$$\gamma^*\mathbf{E}\left[\sum_{j=1}^{|S_i|}X_i^j\right] = \gamma^*|S_i|\frac{\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|} = \frac{s^2\epsilon}{2\alpha k|S_i|}. \tag{10}$$

Since (8) does not hold and since $\gamma < 1$, we have $\gamma \leq \min\{1,\gamma^*\}$. Therefore, we use Hoeffding bound (Lemma A.2) to obtain

$$\begin{aligned}\mathbf{Pr}\left[\sum_{j=1}^{|S_i|}X_i^j \geq (1+\gamma^*)\mathbf{E}\left[\sum_{j=1}^{|S_i|}X_i^j\right]\right] &\leq \exp\left(-\frac{\min\{\gamma^*,\gamma^{*2}\}}{3\Delta}\frac{|S_i|\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|}\right)\\ &= \exp\left(-\frac{\min\{1,\gamma^*\}}{3\Delta}\frac{\gamma^*|S_i|\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|}\right)\\ &\leq \exp\left(-\frac{\gamma s\epsilon}{6\Delta\alpha k}\right).\end{aligned}$$

Hence, we combine this inequality with (10) to obtain,

$$\begin{aligned}\mathbf{Pr}\left[\sum_{j=1}^{|S_i|}X_i^j \geq (1+\gamma)\mathbf{E}\left[\sum_{j=1}^{|S_i|}X_i^j\right] + \frac{s^2\epsilon}{2\alpha k|S_i|}\right] &= \mathbf{Pr}\left[\sum_{j=1}^{|S_i|}X_i^j \geq (1+\gamma+\gamma^*)\mathbf{E}\left[\sum_{j=1}^{|S_i|}X_i^j\right]\right]\\ &\leq \mathbf{Pr}\left[\sum_{j=1}^{|S_i|}X_i^j \geq (1+\gamma^*)\mathbf{E}\left[\sum_{j=1}^{|S_i|}X_i^j\right]\right]\\ &\leq \exp\left(-\frac{\gamma s\epsilon}{6\Delta\alpha k}\right).\end{aligned}$$

Therefore, if we combine this bound with inequality (9) then we get that if $s \geq \frac{6\Delta\alpha k \ln(1/\delta_1)}{\gamma\epsilon}$, then

$$\mathbf{Pr}\left[\sum_{j=1}^{|S_i|} X_i^j \geq (1+\gamma)\frac{|S_i|\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|} + \frac{s^2\epsilon}{2\alpha k|S_i|}\right] \leq \delta_1.$$

Therefore, from now on, let us condition on the event that for every i, we have that

$$\sum_{u\in S_i}\mu(u,c_i) < (1+\gamma)\frac{|S_i|\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|} + \frac{s^2\epsilon}{2\alpha k|S_i|}.$$

This event holds with probability at least $1 - k\delta_1$. Under the conditioning above, we can proceed to the final conclusion:

$$\begin{aligned}
&\mathrm{cost}^b(S,C)\\
&\leq \sum_{i=1}^{k}|S_i|\sum_{u\in S_i}\mu(u,c_i) \leq \sum_{i:V_i^* \text{ is sparse}}|S_i|\sum_{u\in S_i}\mu(u,c_i) + \sum_{i:V_i^* \text{ is dense}}|S_i|\sum_{u\in S_i}\mu(u,c_i)\\
&\leq \frac{6k\Delta\ln(1/\delta_1)}{\gamma^2} + \sum_{i:V_i^* \text{ is dense}}\frac{(1+\gamma)s|V_i^*|}{|V|}\left(\frac{(1+\gamma)|S_i|\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|} + \frac{s^2\epsilon}{2\alpha k|S_i|}\right)\\
&\leq \frac{6k\Delta\ln(1/\delta_1)}{\gamma^2} + \sum_{i:V_i^* \text{ is dense}}\frac{(1+\gamma)^2 s|V_i^*|\sum_{u\in V_i^*}\mu(u,c_i)}{|V|}\frac{|S_i|}{|V_i^*|}\\
&\quad + \sum_{i:V_i^* \text{ is dense}}\frac{(1+\gamma)s|V_i^*|}{|V|}\frac{s^2\epsilon}{2\alpha k|S_i|}\\
&\leq \frac{6k\Delta\ln(1/\delta_1)}{\gamma^2} + \sum_{i:V_i^* \text{ is dense}}\frac{(1+\gamma)^3 s^2|V_i^*|\sum_{u\in V_i^*}\mu(u,c_i)}{|V|^2} + \frac{(1+\gamma)s^3\epsilon}{2\alpha k|V|}\sum_{i:V_i^* \text{ is dense}}\frac{|V_i^*|}{|S_i|}\\
&\leq \frac{6k\Delta\ln(1/\delta_1)}{\gamma^2} + \frac{(1+\gamma)^3 s^2}{|V|^2}\sum_{i=1}^{k}|V_i^*|\sum_{u\in V_i^*}\mu(u,c_i) + \frac{(1+\gamma)s^3\epsilon}{2\alpha k|V|}\,\frac{k|V|}{s(1-\gamma)}\\
&\leq \frac{6k\Delta\ln(1/\delta_1)}{\gamma^2} + \frac{(1+\gamma)^3 s^2}{|V|^2}\,\mathrm{med}^b_{\mathrm{opt}}(V,k) + \frac{3\epsilon s^2}{2\alpha}.
\end{aligned}$$

This yields the following bound that holds with probability at least $1 - 3k\delta_1 = 1 - \delta$:

$$\mathrm{cost}^b_{\mathrm{avg}}(S,C) \leq \frac{6k\Delta\ln(3k/\delta)}{\gamma^2 s^2} + (1+\gamma)^3\,\mathrm{med}^b_{\mathrm{avg}}(V,k) + \frac{3\epsilon}{2\alpha},$$

what concludes the proof of Lemma 5.2. ■

Lemma 5.2 can be combined with arguments used in Lemma 2.2 to prove the following.

Corollary 5.3. *Let $0 < \beta < \alpha$ and $\epsilon > 0$. Let S be a multiset of size $s \geq \frac{c\alpha^2\Delta k\ln(3k/\delta)}{\beta\epsilon}$ chosen from V i.u.r., where c is an appropriate constant. If an α-approximation algorithm for balanced k-median $\mathbb{A}$ is run with input S, then the solution C^* obtained by $\mathbb{A}$ satisfies*

$$\mathbf{Pr}\left[\mathrm{cost}^b_{\mathrm{avg}}(S,C^*) \leq 2(\alpha+\beta)\,\mathrm{med}^b_{\mathrm{avg}}(V,k) + \epsilon\right] \geq 1 - \delta.$$

Proof. Let us apply Lemma 5.2 with ϵ replaced by $\epsilon/5$, parameter $\gamma = \frac{\beta}{7\alpha}$, and with s chosen to satisfy

$$s \geq \frac{7\alpha}{\beta}\sqrt{\frac{30\alpha k\Delta \ln(3k/\delta)}{\epsilon}}.$$

This inequality implies that

$$\frac{6k\Delta \ln(3k/\delta)}{\gamma^2 s^2} \leq \frac{\epsilon}{5\alpha}.$$

Therefore, by Lemma 5.2, if

$$s \geq \max\left\{\frac{210\alpha^2 k\Delta \ln(3k/\delta)}{\beta\epsilon}, \frac{7\alpha}{\beta}\sqrt{\frac{30\alpha k\Delta \ln(3k/\delta)}{\epsilon}}\right\},$$

and if we choose a multiset $S \subseteq V$ of size s i.u.r., then with probability at least $1 - \delta$ we get

$$\begin{aligned}\mathrm{cost}^b_{\mathrm{avg}}(S, C_{\mathrm{opt}}) &\leq (1+\gamma)^3\, \mathrm{med}^b_{\mathrm{avg}}(V,k) + \frac{6k\Delta \ln(3k/\delta)}{\gamma^2 s^2} + \frac{3\epsilon}{10\alpha}\\ &\leq (1+\beta/\alpha)\, \mathrm{med}^b_{\mathrm{avg}}(V,k) + \frac{\epsilon}{2\alpha}.\end{aligned}$$

To conclude the proof of Corollary 5.3, similarly as in the proof of Lemma 5.2, let us choose C to be the set of k centers in S obtained by replacing each $c \in C_{\mathrm{opt}}$ by its nearest neighbor in the corresponding cluster of an optimal solution for S with centers C_{opt}. By the triangle inequality, $\mathrm{cost}^b_{\mathrm{avg}}(S, C) \leq 2\,\mathrm{cost}^b_{\mathrm{avg}}(S, C_{\mathrm{opt}})$. Hence, multiset S contains a set of k centers whose cost is at most $2(1+\beta/\alpha)\,\mathrm{med}^b_{\mathrm{avg}}(V,k) + \frac{\epsilon}{\alpha}$ with probability at least $1-\delta$. Corollary 5.3 follows since $\mathbb{A}$ returns an α-approximation C^* of the balanced k-median for S. ∎

5.4. Dealing With Bad Approximations

The next step in our analysis is to consider bad approximations. Corollary 5.3 proves that typically there is a set of k centers in the sample S that has the average cost close to $\mathrm{med}^b_{\mathrm{avg}}(V,k)$. Now, we show in Lemma 5.14 that all $C_b \subseteq S$ that are $(12\epsilon, 2\alpha+4\beta)$-bad solutions of a balanced k-median of V satisfy $\mathrm{cost}^b_{\mathrm{avg}}(S, C_b) > (2\alpha+\beta)\,\mathrm{med}^b_{\mathrm{avg}}(V,k) + \epsilon$ with high probability. Our analysis follows the approach used before in the proof of Lemma 2.3, but the technical details are more complex.

In the analysis, we fix a $(6\epsilon, 2\alpha+2\beta)$-bad solution $C_b = \{c_1,\dots,c_k\}$ of a balanced k-median of V. Then we further parameterize the problem. We introduce integer *cardinality constraints* $e = (e_1,\dots,e_k)$ and *weights* $w = (w_1,\dots,w_k)$, which satisfy $\sum_i e_i = |V|$. We consider the following problem:

Find a partition of V into k sets $V_1,\dots,V_k$ that satisfies $|V_i| = e_i$ for every i, and that minimizes

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) = \frac{1}{|V|^2}\sum_i w_i \sum_{v\in V_i} \mu(v, c_i).$$

(Obviously, for the balanced k-median problem, we have to consider only solutions with $e = w$ but the decoupling of cardinality constraints and weights simplifies the analysis.) Again we will not require $c_i \in V_i$.

We start with two simple lemmas that relate the cost of optimal partitions with different cardinality constraints and weight vectors to each other.

Lemma 5.4. *Let $e^{(1)}$ and $e^{(2)}$ be two cardinality constraints such that $\|e^{(1)} - e^{(2)}\|_1 \le \frac{2\epsilon|V|}{\Delta}$. Let w be an arbitrary weight vector with $w_i \le |V|$ for all i. Then*

$$\left|\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e^{(1)}, w) - \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e^{(2)}, w)\right| \le \epsilon.$$

Proof. Assigning a single point from one cluster to another can change the normalized cost of the solution by at most $|V|\Delta/|V|^2 = \Delta/|V|$. We can move from a solution for constraint $e^{(1)}$ to one for $e^{(2)}$ and vice versa by moving at most $\epsilon|V|/\Delta$ points. Therefore, the cost of the solution changes by at most ϵ. ▪

In a similar way we can relate solutions with different weight vectors to each other.

Lemma 5.5. *Let $w^{(1)}$ and $w^{(2)}$ be two weight vectors with $\|w^{(1)} - w^{(2)}\|_1 \le \frac{2\epsilon|V|}{\Delta}$. Let e be an arbitrary cardinality constraint, so $\sum_i e_i = |V|$. Then*

$$\left|\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w^{(1)}) - \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w^{(2)})\right| \le \epsilon.$$

Proof. Changing an entry in weight vector w_i by one can change the cost of the normalized solution by at most $|V|\Delta/|V|^2 = \Delta/|V|$. Since $\|w^{(1)} - w^{(2)}\|_1 \le \frac{2\epsilon|V|}{\Delta}$, the cost of the solution changes by at most ϵ when moving from $w^{(1)}$ to $w^{(2)}$ or vice versa. ▪

5.4.1. Subset Optimality.

Now we can proceed to the main technical tool in our analysis. We want to find a lower bound for the cost of an optimal partition of our sample set for the set of centers C_b. To get this lower bound, we will use the following lemma, which states that certain solutions are optimal solutions for given cardinality constraints and weights.

Lemma 5.6. *Let $V_1, \dots, V_k$ be an optimal partition of V for centers C_b, cardinality constraints e and weights w. Let $w' = w/c$, for an arbitrary positive real c, and let $S_1 \subseteq V_1, \dots, S_k \subseteq V_k$ be arbitrary subsets of $V_1, \dots, V_k$, possibly containing multiple copies of points. Then, $S_1, \dots, S_k$ is an optimal partition for centers C_b, cardinality constraints $e_S = (|S_1|, \dots, |S_k|)$ and weights w', where $|S_i|$ counts every copy of a point in S_i when S_i is a multiset.*

Proof. The proof is by contradiction. We show that nonoptimality of $S_1, \dots, S_k$ for C_b, w', and e implies that $V_1, \dots, V_k$ is not an optimal solution for e and w. To show this, we construct a *cyclic change* of points that improves the cost of the partition $V_1, \dots, V_k$, but preserves the cardinality constraints.

A *cyclic change* is a sequence of points $(p_0, \dots, p_{t-1})$ such that $p_i \ne p_j$ for $i \ne j$. By *applying a cyclic change* $(p_0, \dots, p_{t-1})$ to sets $V_1, \dots, V_k$, we mean the operation of replacing each point p_{j+1} in its cluster V_i by point p_j, where the indices j are taken modulo t.

Observation 5.7. A cyclic change does not affect the cardinality constraints. ▪

The cost of a cyclic change is the change of the cost in the objective function of the clustering problem with centers C_b, cardinality constraints e, and weights w when the cyclic change is applied. Thus, a cyclic change with negative cost for $V_1, \dots, V_k$ would contradict the optimality of $V_1, \dots, V_k$.

For a moment, let us assume that $S = \bigcup_i S_i$ contains no multiple copies of a point. Let $s_i = |S_i|$ for every i. For the purpose of contradiction, let us consider an optimal solution $S'_1, \dots, S'_k$ with $|S'_i| = s_i$ for C_b, w' and e, and suppose that the cost of this solution is strictly smaller than the cost of $S_1, \dots, S_k$. Furthermore, we assume that $S'_1, \dots, S'_k$ has the smallest Hamming distance (i.e., $\sum_{i=1}^{k} |S_i \oplus S'_i|$) among all optimal solutions satisfying the condition above. Then, we construct a cyclic change in the following way. We start with an arbitrary point p_0 such that $\mathrm{Clust}(p_0) \neq \mathrm{Clust}'(p_0)$, where the functions $\mathrm{Clust}(p)$ and $\mathrm{Clust}'(p)$ return the index of the cluster containing p in clustering $S_1, \dots, S_k$ and $S'_1, \dots, S'_k$, respectively. This point is the first point of our cyclic change and we mark it. Then, we choose an unmarked point from cluster $S'_{\mathrm{Clust}(p_0)}$ that is not in $S_{\mathrm{Clust}(p_0)}$. Since the cardinality constraints on S_i and S'_i imply that $|S_i| = |S'_i|$, such a point must exist because $p_0 \in S_{\mathrm{Clust}(p_0)} \backslash S'_{\mathrm{Clust}(p_0)}$. We continue this process until we find a point p_{t-1} with $\mathrm{Clust}(p_{t-1}) = \mathrm{Clust}'(p_0)$. Observe that *after* applying the cyclic change to the sets $S'_1, \dots, S'_k$ to obtain the clustering $T'_1, \dots, T'_k$, we have $p_i \in T'_{\mathrm{Clust}(p_i)}$ for $0 \le i \le t-1$. Furthermore, by the optimality of $S'_1, \dots, S'_k$ and by the fact that the clustering $T'_1, \dots, T'_k$ has smaller Hamming distance to $S_1, \dots, S_k$ than $S'_1, \dots, S'_k$, the cyclic change has positive cost.

We next show that the inverse cyclic change $(p_{t-1}, \dots, p_0)$ has negative cost for $V_1, \dots, V_k$. We know that

$$\sum_{i=0}^{t-1} \left(-\mu(p_i, c_{\mathrm{Clust}'(p_i)}) w'_{\mathrm{Clust}'(p_i)} + \mu(p_i, c_{\mathrm{Clust}(p_i)}) w'_{\mathrm{Clust}(p_i)} \right) > 0,$$

and so

$$\sum_{i=0}^{t-1} \left(-\mu(p_i, c_{\mathrm{Clust}(p_i)}) w_{\mathrm{Clust}(p_i)} + \mu(p_i, c_{\mathrm{Clust}'(p_i)}) w_{\mathrm{Clust}'(p_i)} \right) < 0,$$

which means that the latter cyclic change has negative cost and so it can be used to improve the optimal solution $V_1, \dots, V_k$, which is a contradiction. Hence, $S_1, \dots, S_k$ is an optimal clustering as well.

It remains to show how to deal with the case where points occur more than once in the sample set S. Assume we have a cyclic change $(p_0, \dots, p_{t-1})$ where point p_i and p_j are copies of the same point of V. Then, either $(p_0, \dots, p_i, p_{j+1}, \dots, p_{t-1})$ or $(p_i, \dots, p_j)$ is a cyclic change of positive cost and we can consider this particular cyclic change. Applying this argument several times, we arrive at a cyclic change without duplicates of points and can apply our arguments from above. ■

5.4.2. Introduction to the Main Construction: Lower Bounding $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*)$.

To find a lower bound for our sample set S with respect to the centers C_b, we show that for a fixed cardinality constraint $s^* = (s^*_1, \dots, s^*_k)$ with $\sum_i s^*_i = s$, and for weights $(s^*_1, \dots, s^*_k)$ (thus identical to the cardinality constraints), the cost of an optimal partition of S (which is $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*)$) will be large with high probability.

This is done in the following way. We first construct a vector $x = \left(\frac{s^*_1|V|}{s}, \dots, \frac{s^*_k|V|}{s}\right)$, and define x' to be the closest integer vector to x in the ℓ_1 norm that satisfies $\sum_i x'_i = |V|$. We consider an optimal partition of V into set $V_1, \dots, V_k$ for centers C_b, cardinality constraints

$e = x'$, and weights $w = x'$. We study the following experiment. We choose s points $p_1, \ldots, p_s$ from V uniformly at random with repetition and let $S = \{p_1, \ldots, p_s\}$. If p_j is in V_i, we assign it to cluster S_i. By Lemma 5.6, we know that the clustering $S_1, \ldots, S_k$ is optimal for the cardinality constraints $|S \cap V_i|$ and weights w. We will argue that *(a)* the normalized cost of clustering $S_1, \ldots, S_k$ with weights s^* approximates well the normalized cost of $V_1, \ldots, V_k$ with weights w, and that *(b)* the cost of clustering $S_1, \ldots, S_k$ with weights $ws/|V| = (w_1 \frac{s}{|V|}, \ldots, w_k \frac{s}{|V|})$ is close to the cost of the optimal clustering with cardinality constraints and weights s^*. We let s'_i be a random variable with value $|S \cap V_i|$ and let us set $s' = (s'_1, \ldots, s'_k)$.

We will prove that for sufficiently large s, with high probability the following claims hold:

- (Lemma 5.8) $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|)$ is larger than $(1-\lambda)\,\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - \epsilon$,
- (Lemma 5.9) $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', s^*)$ is larger than $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|) - \epsilon$,
- (Corollary 5.11) $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*)$ is larger than $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', s^*) - \epsilon$.

These three bounds will yield (see Corollary 5.12) a good lower bound for $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*)$, that is, for the cost of an optimal partition of S with respect to centers C_b and for cardinality constraint s^* with $\sum_i s^*_i = s$, and for weights $(s^*_1, \ldots, s^*_k)$.

Lower Bounding $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|)$. To get a lower bound for $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*)$, we begin with the following result that lower bounds $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|)$ in terms of $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w)$ and ϵ.

Lemma 5.8. *For $s \geq \frac{2\Delta \ln(1/\delta)}{\epsilon\lambda^2}$, we get with probability at least $1-\delta$ that*

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|) > (1-\lambda)\,\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - \epsilon.$$

Proof. Since the weights of the clusters are fixed and rescaled by factor $s/|V|$, the expected contribution of a random point from V is exactly $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w)/s$. Therefore,

$$\mathbf{E}[\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|)] = \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w),$$

and we have to only show the sharp concentration. By Hoeffding bound (Lemma A.2) and the fact that a random point from V contributes with at most Δ/s, we obtain

$$\mathbf{Pr}\big[\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}\big(S, C_b, s', \tfrac{ws}{|V|}\big) \leq (1-\zeta)\mathbf{E}\big[\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}\big(S, C_b, s', \tfrac{ws}{|V|}\big)\big]\big] \leq e^{-\frac{\zeta^2 \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) s}{2\Delta}}.$$

Now, we make distinction between the cases $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) \geq \epsilon$ and $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) < \epsilon$. For $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) \geq \epsilon$, the lemma follows with $s \geq \frac{2\Delta \ln(1/\rho)}{\epsilon\zeta^2}$ and $\zeta = \lambda$. Otherwise, if $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) < \epsilon$, then we choose $\zeta = \epsilon/\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w)$. In this case, we get:

$$\mathbf{Pr}\left[\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|) \leq \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - \epsilon\right] \leq e^{-\frac{\epsilon^2 s}{2\Delta\, \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w)}}.$$

Observe that for any λ, $0 < \lambda \leq 1$, if we choose s such that $s \geq \frac{2\Delta \ln(1/\rho)}{\epsilon\lambda^2}$, then we obtain:

$$\exp\left(-\frac{\epsilon^2 s}{2\Delta\, \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w)}\right) \leq \exp\left(-\frac{\epsilon \ln(1/\rho)}{\lambda^2\, \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w)}\right) \leq \rho,$$

where the last inequality follows from the fact that $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) < \epsilon$ and that $\lambda \leq 1$.

This bound finally implies that in either case, for any λ, $0 < \lambda \leq 1$, if we choose s such that $s \geq \frac{2\Delta \ln(1/\rho)}{\epsilon\lambda^2}$, then we will have

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|) \leq (1-\lambda)\,\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - \epsilon$$

with probability at most ρ. This yields the lemma. ▪

Lower Bounding $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', s^*)$. With a lower bound for $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|)$, we continue our analysis by proving a lower bound for $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', s^*)$ by $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|) - \epsilon$. We apply Lemma 5.5 to get the following.

Lemma 5.9. *If $s \geq 2k\Delta/\epsilon$, then we have*

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', s^*) \geq \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|) - \epsilon.$$

Proof. Recall that w is the nearest (in the ℓ_1 norm) integer vector to $x = \frac{|V|}{s} s^*$ that satisfies $\sum w_i = |V|$. It is easy to verify that $\|w - x\|_1 \leq 2k$. Hence $\|s^* - ws/|V|\|_1 \leq 2k$ as well. The lemma follows now from Lemma 5.5 with parameters $V = S$ and $|S| = s \geq 2k\Delta/\epsilon$. ▪

Lower Bounding $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*)$. Our final step is a lower bound for $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*)$. We first show that typically, the ℓ_1 distance between s' and its expectation is small.

Lemma 5.10. *If $s \geq \frac{3\Delta^2 k^2 \ln(2k/\rho)}{\epsilon^2}$, then with probability at least $1 - \delta$ the following holds:*

$$\|s' - \mathbf{E}[s']\|_1 < \epsilon s/\Delta.$$

Proof. Let $X_{i,j}$ denote the indicator random variable for the event that the j-th point from S is in V_i. Let us observe that if we set

$$\zeta_i = \frac{\epsilon s}{k\Delta \mathbf{E}\left[\sum_{j=1}^{s} X_{i,j}\right]},$$

then we have

$$\begin{aligned}
&\mathbf{Pr}\left[\left|\sum_{j=1}^{s} X_{i,j} - \mathbf{E}\left[\sum_{j=1}^{s} X_{i,j}\right]\right| \geq \epsilon s/(k\Delta)\right] \\
&\quad \leq \mathbf{Pr}\left[\sum_{j=1}^{s} X_{i,j} \geq (1+\zeta)\mathbf{E}\left[\sum_{j=1}^{s} X_{i,j}\right]\right] + \mathbf{Pr}\left[\sum_{j=1}^{s} X_{i,j} \leq (1-\zeta)\mathbf{E}\left[\sum_{j=1}^{s} X_{i,j}\right]\right].
\end{aligned}$$

Let us begin with the analysis of the first term. We assume first that $\zeta \leq 1$. Then, by the Hoeffding bound (Lemma A.2), we get

$$\mathbf{Pr}\left[\sum_j X_{i,j} \geq (1+\zeta)\mathbf{E}\left[\sum_j X_{i,j}\right]\right] \leq e^{-\frac{\zeta^2 \mathbf{E}[\sum_j X_{i,j}]}{3}} = e^{-\left(\frac{\epsilon s}{k\Delta \mathbf{E}[\sum_j X_{i,j}]}\right)^2 \frac{\mathbf{E}[\sum_j X_{i,j}]}{3}} \leq e^{-\frac{\epsilon^2 s}{3\Delta^2 k^2}},$$

where the last inequality uses the fact that $s \geq \mathbf{E}\left[\sum_j X_{i,j}\right]$. Otherwise, if $\zeta > 1$, then we have

$$\mathbf{Pr}\left[\sum_j X_{i,j} \geq (1+\zeta)\mathbf{E}\left[\sum_j X_{i,j}\right]\right] \leq e^{-\frac{\zeta \mathbf{E}[\sum_j X_{i,j}]}{3}} = e^{-\frac{\epsilon s}{3k\Delta}}.$$

Similarly, we can use the Hoeffding bound (Lemma A.2) to obtain the following:

$$\mathbf{Pr}\left[\sum_j X_{i,j} \le (1-\zeta)\mathbf{E}\left[\sum_j X_{i,j}\right]\right] \le e^{-\frac{\zeta^2 \mathbf{E}[\sum_j X_{i,j}]}{2}} \le e^{-\frac{\epsilon^2 s}{2\Delta^2 k^2}}.$$

Hence, by combining the three inequalities above, if we choose

$$s \ge \frac{3\Delta^2 k^2 \ln(2k/\delta)}{\epsilon^2},$$

then we get

$$\mathbf{Pr}\left[\left|\sum_j X_{i,j} - \mathbf{E}\left[\sum_j X_{i,j}\right]\right| \ge \epsilon s/(k\Delta)\right] \le \delta/k.$$

Now, the lemma follows immediately by applying the union bound to the inequality above. ∎

Our next result follows from Lemma 5.4.

Corollary 5.11. *For $s \ge \frac{12\Delta^2 k^2 \ln(2k/\delta)}{\epsilon^2}$, the following bound holds with probability at least $1-\delta$:*

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*) \ge \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', s^*) - \epsilon.$$

Proof. We have

$$\|s' - s^*\|_1 \le \|s' - \mathbf{E}[s']\|_1 + \|\mathbf{E}[s'] - s^*\|_1 \le \frac{\epsilon s}{\Delta} + \frac{s}{|V|}\|x' - x\|_1 \le \frac{\epsilon s}{\Delta} + \frac{\epsilon^2 s^2}{\Delta^2 |V| k} \le 2\epsilon s/\Delta.$$

The corollary follows by substituting ϵ with $\epsilon/2$. ∎

Finally, we can combine Lemmas 5.8 and 5.9 with Corollary 5.11 to obtain the following.

Corollary 5.12. *For every λ and ϵ, $0 < \lambda, \epsilon \le 1$, if*

$$s \ge \frac{12\Delta \ln(4k/\delta)}{\epsilon}\left(\frac{\Delta k^2}{\epsilon} + \frac{1}{\lambda^2}\right),$$

then with probability at least $1-\delta$ we get

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*) > (1-\lambda)\,\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - 3\epsilon.$$

Proof. By Corollary 5.11 we have

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*) \ge \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', s^*) - \epsilon$$

with probability $1-\delta/2$ for our choice of s. Using Lemma 5.9, we get

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*) \ge \mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|) - 2\epsilon.$$

By Lemma 5.8, we get with probability $1 - \delta/2$

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s', ws/|V|) > (1-\lambda)\,\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - \epsilon.$$

Combining the last two inequalities, we get with probability $1 - \delta$

$$\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*) > (1-\lambda)\,\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - 3\epsilon.$$

This proves Corollary 5.12. ■

5.4.3. Quality of Bad Solution: Bad Solutions Are Bad. After the analysis in Section 5.4.2, we are now ready to proceed with the analysis of the quality of bad approximation for a sample set. We begin with a lemma that considers a set of bad centers.

Lemma 5.13. *Let S be a multiset of points chosen i.u.r. from V with s such that*

$$s \ge \frac{c\Delta(k+\ln(1/\delta))\ln\left(\frac{\alpha k\Delta\ln(1/\delta)}{\beta\epsilon}\right)}{\epsilon}\left(\frac{\alpha^2}{\beta^2}+\frac{\Delta k^2}{\epsilon}\right),$$

where c is a suitable constant and δ is an arbitrary confidence parameter. Let C_b be a $(6\epsilon, 2\alpha+2\beta)$-bad solution of a balanced k-median of V. Suppose that $\beta \le \alpha$. Then

$$\mathbf{Pr}[\mathrm{cost}^{b}_{\mathrm{avg}}(S, C_b) > (2\alpha+\beta)\,\mathrm{med}^{b}_{\mathrm{avg}}(V,k) + \epsilon] \ge 1-\delta.$$

Proof. Let us first fix vector s^*. Then, we apply Corollary 5.12 with $\lambda = \beta/(2(\alpha+\beta))$ and $\delta' = \delta/s^k$ to get that with probability $1 - \delta/s^k$:

$$\begin{aligned}\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*) &> (1-\lambda)\,\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - 3\epsilon\\ &= \left(1-\frac{\beta}{2(\alpha+\beta)}\right)\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - 3\epsilon.\end{aligned}$$

Next, we observe that since C_b is a $(6\epsilon, 2\alpha+2\beta)$-bad solution of a balanced k-median of V, we have $\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) > 2(\alpha+\beta)\,\mathrm{med}^{b}_{\mathrm{avg}}(V,k) + 6\epsilon$. Thus, with probability $1-\delta/s^k$:

$$\begin{aligned}\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(S, C_b, s^*, s^*) &> \left(1-\frac{\beta}{2(\alpha+\beta)}\right)\mathrm{cost}^{\mathrm{con}}_{\mathrm{avg}}(V, C_b, e, w) - 3\epsilon\\ &\ge \left(1-\frac{\beta}{2(\alpha+\beta)}\right)\left(2(\alpha+\beta)\,\mathrm{med}^{b}_{\mathrm{avg}}(V,k) + 6\epsilon\right) - 3\epsilon\\ &\ge (2\alpha+\beta)\,\mathrm{med}^{b}_{\mathrm{avg}}(V,k) + \epsilon,\end{aligned}$$

where the last inequality follows from $\beta \le \alpha$.

Since there are at most s^k choices for s^*, the lemma follows from the bound above and the union bound. ■

Once we have the bound for a single set of bad centers, we can proceed with the analysis of all bad sets of centers.

Lemma 5.14. *Let S be a multiset of s points chosen i.u.r. from V with s such that:*

$$s \geq \frac{c\Delta(k+\ln(1/\delta))\ln\left(\frac{\alpha k\Delta\ln(1/\delta)}{\beta\epsilon}\right)}{\epsilon}\left(\frac{\alpha^2}{\beta^2}+\frac{\Delta k^2}{\epsilon}\right),$$

where c is a suitable positive constant and δ is an arbitrary confidence parameter. Let $\mathbb{C}$ be the set of $(12\epsilon, 2\alpha+4\beta)$-bad solutions C of a balanced k-median of V. Then,

$$\mathbf{Pr}\left[\exists C_b \in \mathbb{C} : C_b \subseteq S \text{ and } \operatorname{cost}_{\text{avg}}(S, C_b) \leq (2\alpha+\beta)\operatorname{med}^b_{\text{avg}}(V,k)+\epsilon\right] \leq \delta.$$

Proof. We proceed as in the proof of Lemma 2.3. We choose c so that $s \geq \frac{8(\alpha+\beta)}{\beta}k$ and we let S^* be a multiset of $s-k$ points chosen i.u.r from V. Then,

$$\begin{aligned}
&\mathbf{Pr}\left[\exists C_b \in \mathbb{C} : C_b \subseteq S \text{ and } \operatorname{cost}^b_{\text{avg}}(S, C_b) \leq (2\alpha+\beta)\operatorname{med}^b_{\text{avg}}(V,k)+\epsilon\right]\\
&\quad\leq \sum_{C_b\in\mathbb{C}} \mathbf{Pr}\left[C_b \subseteq S \text{ and } \operatorname{cost}^b_{\text{avg}}(S, C_b) \leq (2\alpha+\beta)\operatorname{med}^b_{\text{avg}}(V,k)+\epsilon\right]\\
&\quad\leq \sum_{C_b\in\mathbb{C}} \mathbf{Pr}\left[\operatorname{cost}^b_{\text{avg}}(S^*, C_b) \leq \frac{s^2}{(s-k)^2}\left((2\alpha+\beta)\operatorname{med}^b_{\text{avg}}(V,k)+\epsilon\right)\right]\mathbf{Pr}\left[C_b \subseteq S\right]\\
&\quad\leq \sum_{C_b\in\mathbb{C}} \mathbf{Pr}\left[\operatorname{cost}^b_{\text{avg}}(S^*, C_b) \leq (2\alpha+2\beta)\operatorname{med}^b_{\text{avg}}(V,k)+2\epsilon\right]\mathbf{Pr}\left[C_b \subseteq S\right].
\end{aligned}$$

Now, we apply Lemma 5.13 with values of s, β and ϵ replaced by s^*, 2β and 2ϵ, respectively. This will imply that if

$$s-k \geq \frac{c\Delta(k+\ln(1/\delta))\ln\left(\frac{\alpha k\Delta\ln(1/\delta)}{\beta\epsilon}\right)}{\epsilon}\left(\frac{(\alpha+\beta)^2}{\beta^2}+\frac{\Delta k^2}{\epsilon}\right),$$

then for any $C_b \in \mathbb{C}$ we have

$$\mathbf{Pr}\left[\operatorname{cost}^b_{\text{avg}}(S^*, C_b) \leq (2\alpha+2\beta)\operatorname{med}^b_{\text{avg}}(V,k)+2\epsilon\right] \leq \delta.$$

Thus, if we plug this bound in the inequality above, we get

$$\begin{aligned}
&\mathbf{Pr}\left[\exists C_b \in \mathbb{C} : C_b \subseteq S \text{ and } \operatorname{cost}^b_{\text{avg}}(S, C_b) \leq (2\alpha+\beta)\operatorname{med}^b_{\text{avg}}(V,k)+\epsilon\right]\\
&\quad\leq \sum_{C_b\in\mathbb{C}} \mathbf{Pr}\left[\operatorname{cost}^b_{\text{avg}}(S^*, C_b) \leq (2\alpha+2\beta)\operatorname{med}^b_{\text{avg}}(V,k)+2\epsilon\right]\mathbf{Pr}\left[C_b \subseteq S\right]\\
&\quad\leq |\mathbb{C}|\delta\binom{s}{k}\Big/\binom{n}{k} \leq s^k\delta.
\end{aligned}$$

This allows us to conclude that if we set

$$s \geq \frac{c\Delta(k+\ln(1/\delta))\ln\left(\frac{\alpha k\Delta\ln(1/\delta)}{\beta\epsilon}\right)}{\epsilon}\left(\frac{\alpha^2}{\beta^2}+\frac{\Delta k^2}{\epsilon}\right),$$

for a suitable constant c, then Lemma 5.14 follows. ∎

5.5. Concluding the Proof of Theorem 5

We conclude the proof of Theorem 5 that will now follow from Corollary 5.3 and Lemma 5.14.

Proof of Theorem 5. By Lemma 5.14, with probability at least $1 - \delta$ no set $C \subseteq S$ that is $(12\epsilon, 2\alpha + 4\beta)$-bad solution of a balanced k-median of V satisfies the inequality

$$\mathrm{cost}^b_{\mathrm{avg}}(S, C) > 2(\alpha + \beta)\, \mathrm{med}^b_{\mathrm{avg}}(V, k) + \epsilon.$$

On the other hand, if we run algorithm $\mathbb{A}$ for set S, then by Corollary 5.3 the resulting set C^* of k-centers with probability at least $1 - \delta$ satisfies

$$\mathrm{cost}^b_{\mathrm{avg}}(S, C^*) \le 2(\alpha + \beta)\, \mathrm{med}^b_{\mathrm{avg}}(V, k) + \epsilon.$$

These two claims imply that with probability at least $1 - 2\delta$ the set C^* is a $(12\epsilon, 2\alpha + 4\beta)$-good solution of a balanced k-median of V, that is,

$$\mathbf{Pr}\left[\mathrm{cost}^b_{\mathrm{avg}}(S, C^*) \le (2\alpha + 4\beta)\, \mathrm{med}^b_{\mathrm{avg}}(V, k) + 12\epsilon\right] \ge 1 - 2\delta.$$

This implies the first part of the claim for the variant of the problem without the constraint $c_i \in V_i$.

Next, we consider the version of the problem with constraint $c_i \in V_i$. We first observe that Corollary 5.3 holds also when instead of allowing C^* to be an arbitrary set of points, we add the constraint that $c_i \in V_i$ for all i. To see this, let us observe that adding the constraint $c_i \in V_i$ can only increase the cost of a bad solution, and at the same time, the solution returned by our algorithm will still be a good solution for the problem *without* this constraint. This implies that the bound in Corollary 5.3 will hold.

We next show that we can enforce the constraint while increasing the average cost of our solution by at most ϵ, which will imply our claim. Let us consider a solution $V_1, \dots, V_k$. If $c_i \in V_j$, $i \neq j$, then we swap c_i with an arbitrary point v in V_i. This can increase the cost of the clustering by at most $\mu(v, c_j)|V_j| \le \Delta n$. We do this for all k clusters. This increases the cost of our clustering by at most Δkn. Thus, if $\epsilon \ge \Delta k/n$ we have $\Delta kn \le \epsilon n^2$ and we increase the average cost by at most ϵ. Otherwise, $s \ge n$ and we can run $\mathbb{A}$ on the original input set. This proves the first part of the claim.

To see the second part of the claim, we consider the partition $S_1, \dots, S_k$ of our sample set computed by algorithm $\mathbb{A}$. Then, we consider the integer vector $x' = (x'_1, \dots, x'_k)$ that is closest to $\left(\frac{|S_1||V|}{s}, \dots, \frac{|S_k||V|}{s}\right)$. We use x'_i as cluster sizes and run the algorithm from [31] to obtain an optimal clustering for these sizes. Using similar arguments as before, we obtain that the cost of this clustering is approximated by the cost of the optimal clustering of our sample. ▪

APPENDIX A: CONCENTRATION BOUNDS

In this section, we present concentration bounds used in the paper. We begin with Chernoff bound and then present Hoeffding bound.

Lemma A.1 (Chernoff bound). *Let $X_1, \ldots, X_N$ be independent random variables, with* $\mathbf{Pr}[X_i = 1] = p$ *and* $\mathbf{Pr}[X_i = 0] = 1 - p$ *for each i and for certain p,* $0 \le p \le 1$. *Let* $X = \sum_{i=1}^{N} X_i$. *Then*

- *for any* ζ, $0 \le \zeta \le 1$,

$$\mathbf{Pr}[X \ge (1+\zeta)\mathbf{E}[X]] \le \exp(-\zeta^2 \mathbf{E}[X]/3),$$
$$\mathbf{Pr}[X \le (1-\zeta)\mathbf{E}[X]] \le \exp(-\zeta^2 \mathbf{E}[X]/2);$$

- *for any* $t \ge 6\mathbf{E}[X]$,

$$\mathbf{Pr}[X \ge t] \le 2^{-t}.$$

Lemma A.2 (Hoeffding bound). *Let $X_1, \ldots, X_N$ be independent random variables, with* $0 \le X_i \le M$ *for each i,* $0 \le i \le N$. *Let* $X = \sum_{i=1}^{N} X_i$. *Then*

- *for any* $\zeta \ge 0$,

$$\mathbf{Pr}[X \ge (1+\zeta)\mathbf{E}[X]] \le \exp\left(-\frac{\mathbf{E}[X]\min\{\zeta,\zeta^2\}}{3M}\right),$$
$$\mathbf{Pr}[X \le (1-\zeta)\mathbf{E}[X]] \le \exp\left(-\frac{\mathbf{E}[X]\zeta^2}{2M}\right).$$

ACKNOWLEDGMENTS

We thank an anonymous referee for making many good suggestions that improved the quality of the paper.

REFERENCES

[1] S. Arora, P. Raghavan, and S. Rao, Approximation schemes for Euclidean k-medians and related problems. In Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC), 1998, pp. 106–113.

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

[3] M. Bădoiu, S. Har-Peled, and P. Indyk, Approximate clustering via core-sets. In Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC), 2002, pp. 250–257.

[4] Y. Bartal, M. Charikar, and D. Raz, Approximating min-sum k-clustering in metric spaces. In Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC), 2001, pp. 11–20.

[5] M. Charikar and S. Guha, Improved combinatorial algorithms for the facility location and k-median problems. In Proceedings of the 40th IEEE Symposium on Foundations of Computer Science (FOCS), 1999, pp. 378–388.

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

[7] M. Charikar, L. O'Callaghan, and R. Panigrahy, Better streaming algorithms for clustering problems. In Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC), 2003, pp. 30–39.

[8] B. Chazelle, Who says you have to look at the input? The brave new world of sublinear computing? In Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2004, p. 134.

[9] A. Czumaj and C. Sohler, Abstract combinatorial programs and efficient property testers, SIAM J Computing 34 (2005), 580–615.

[10] J. Fakcharoenphol, S. Rao, and K. Talwar, A tight bound on approximating arbitrary metrics by tree metrics, J Comput System Sci 69 (2004), 485–497.

[11] W. Fernandez de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani, Approximation schemes for clustering problems. In Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC), 2003, pp. 50–58.

[12] G. Frahling and C. Sohler, Coresets in dynamic geometric data streams. In Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC), 2005, pp. 209–217.

[13] N. Gutmann-Beck and R. Hassin, Approximation algorithms for min-sum p-clustering, Discrete Appl Math 89 (1998), 125–142.

[14] S. Har-Peled and S. Mazumdar, On coresets for k-means and k-median clustering, In Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC), 2004, pp. 291–300.

[15] M. Inaba, N. Katoh, and H. Imai, Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. In Proceedings of the 10th Annual ACM Symposium on Computational Geometry, 1994, pp. 332–339.

[16] P. Indyk, Sublinear time algorithms for metric space problems. In Proceedings of the 31st Annual ACM Symposium on Theory of Computing (STOC), 1999, pp. 428–434.

[17] P. Indyk, A sublinear time approximation scheme for clustering in metric spaces. In Proceedings of the 40th IEEE Symposium on Foundations of Computer Science (FOCS), 1999, pp. 154–159.

[18] P. Indyk, High-Dimensional Computational Geometry, PhD Dissertation, Stanford University, 2000.

[19] K. Jain, M. Mahdian, E. Markakis, A. Saberi, and V. Vazirani, Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP, J ACM 50(6) (2003), 795–824.

[20] K. Jain and V. V. Vazirani, Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation, J ACM 48 (2001), 274–296.

[21] S. G. Kolliopoulos and S. Rao, A nearly linear-time approximation scheme for the Euclidean k-median problems. In Proceedings of the 7th Annual European Symposium on Algorithms (ESA), 1999, pp. 378–389.

[22] A. Kumar, Y. Sabharwal, and S. Sen, A simple linear time $(1 + \varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In Proceedings of the 45th IEEE Symposium on Foundations of Computer Science (FOCS), 2004, pp. 454–462.

[23] A. Kumar, Y. Sabharwal, and S. Sen, Linear time algorithms for clustering problems in any dimensions. In Proceedings of the 32nd Annual International Colloquium on Automata, Languages and Programming (ICALP), 2005, pp. 1374–1385.

[24] R. Kumar and R. Rubinfeld, Sublinear time algorithms, SIGACT News 34(4) (2003), 57–67.

[25] J. Matoušek, On approximate geometric k-clustering, Discrete Computational Geometry 24 (2000), 61–84.

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

[27] A. Meyerson, L. O'Callaghan, and S. Plotkin, A k-median algorithm with running time independent of data size, Machine Learn 56(1-3) (2004), 61–87.

[28] N. Mishra, D. Oblinger, and L. Pitt, Sublinear time approximate clustering. In Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2001, pp. 439–447.

[29] L. J. Schulman, Clustering for edge-cost minimization. In Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC), 2000, pp. 547–555.

[30] S. Sahni and T. Gonzalez, $\mathcal{P}$-complete approximation problems, J ACM 23 (1976), 555–566.

[31] T. Tokuyama and J. Nakano, Geometric algorithms for the minimum cost assignment problem, Random Struct Algorithms 6 (1995), 393–406.