Sublinear-Time Approximation for Clustering via Random Sampling [⋆]

Artur Czumaj[1] and Christian Sohler[2]

[1] Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA. czumaj@cis.njit.edu

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

Abstract. In this paper we present a novel analysis of a random sampling approach for three clustering problems in metric spaces: *k-median*, *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 points uniformly at random from V 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 the first time approximation algorithms that have the running time independent of the input size, and depending on k and the diameter of the metric space only.

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., [7, 16]). 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 some approximation of the outcome of

[⋆] Research partly supported by NSF ITR grant CCR-0313219, NSF grant CCR-0105701, and DFG grant Me 872-8/2.

an exact algorithm running on the whole input. In many cases the randomized process that selects the sample is very simple, e.g., a uniformly random subset is selected.

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 three clustering problems: the *k-median 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 *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 two 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 $C_1, \dots, C_k$ that minimizes $\sum_{1 \le i \le k} |C_i| \cdot \sum_{p \in C_i} \mu(p, c_i)$.

For all these three clustering problems we study the following "simple sampling" algorithm: pick a random sample S of points, run an approximation algorithm for the sample, and return the clustering induced by the solution for the sample. The main goal of this paper 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 three clustering problems above. These are the first approximation algorithms for these problems whose running time is *independent of the input size*, $|V|$.

1.1 Previous research

***k*-median.** The k-median clustering problem is perhaps the most studied clustering problem in the literature, both, in theoretical research and in applications. 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 $1 + \frac{2}{e}$ [13]. There exist polynomial time approximation algorithms with constant approximation ratios [2, 4, 5, 11, 14, 17]. 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 [15], and more recently in [10]). The k-median problem has been also extensively investigated in the data stream model, see e.g., recent works in [6, 10].

There exist a few sublinear-time algorithms for the k-median problem, that is algorithms with the 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., [11, 17–19]. The algorithm of Indyk [11] computes in $O(n\,k)$ time a set of $O(k)$ centers whose cost approximates the value of the k-median by a constant factor. Mettu and Plaxton [17] gave a randomized $O(1)$-approximate k-median algorithm that runs in time $O(n(k + \log n))$ subject to the constraint $R = 2^{O(n/\log(n/k))}$, where R denotes the ratio between the maximum and the minimum distance between any pair of distinct points in the metric space. Very recently, Meyerson et al. [18] presented a sublinear-time for the k-median problem under an assumption that each cluster has size $\Omega(n\,k/\epsilon)$; their algorithm requires time $O((k^2/\epsilon)\, \log(k/\epsilon))$ and gives a $O(1)$-approximation guarantee with high probability.

Notice that all the sublinear-time ($o(n^2)$-time) algorithms mentioned above made some assumptions about the input. We follow this approach and in this paper we consider a model with the diameter of the metric space $\varDelta$ given, that is, with $\mu : V \times V \to [0, \varDelta]$. Such a model has been investigated before by Mishra et al. [19], who studied the quality of k-median clusterings obtained by random sampling. Let $\mathbb{A}_\alpha$ be an arbitrary α-approximation algorithm for k-median. Using techniques from statistics and computational learning theory, Mishra et al. [19] proved that if we sample a set S of $s = \widetilde{O}\big(\big(\frac{\alpha \varDelta}{\epsilon}\big)^2 \; (k \, \ln n + \ln(1/\delta))\big)$ points from V i.u.r. (*independently and uniformly at random*) and run algorithm $\mathbb{A}_\alpha$ to find an approximation of k-median for S, then with probability at least $1 - \delta$ the outputted set of k centers has the *average distance* to the nearest center of at most $2\,\alpha\, med_{avg}(V, k) + \epsilon$, where $med_{avg}(V, k)$ denotes the *average distance* to the k-median C, that is, $med_{avg}(V, k) = \frac{\sum_{v \in V} \mu(v,C)}{n}$. Using this result, Mishra et al. [19] 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 $s = \widetilde{O}\left(\big(\frac{\alpha \varDelta}{\epsilon}\big)^2 \cdot (k \, \ln n + \ln(1/\delta))\right)$ 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 $2\,\alpha\, med_{avg}(V, k) + \epsilon$. 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 $\varOmega(n^2)$ or $\varOmega(n\,k)$ as in the algorithms discussed above. On the other hand, the running time of this algorithm depends on $\varDelta$, and as discussed in [19] (see also [17, 18]), such a dependency is necessary to obtain this kind of approximation guarantee.

Min-sum *k*-clustering. The min-sum k-clustering problem was first formulated (for general graphs) by Sahni and Gonzales [21]. There is a 2-approximation algorithm by Guttman-Beck and Hassin [9] with running time $n^{O(k)}$. Recently, Bartal et al. [3] presented an $O(\frac{1}{\epsilon} \log^{1+\epsilon} n)$-approximation algorithm with $O(n^{1/\epsilon})$ running time and then Fernandez de la Vega et al. [8] gave an $(1 + \epsilon)$-approximation algorithm with the running time of $O(n^{3k} \cdot 2^{O((1/\epsilon)^{k^2})})$. For point sets in the $\mathbb{R}^d$, Schulman [20] introduced an algorithm for distance functions ℓ_2^2, ℓ_1 and ℓ_2 that computes a solution that is either within $(1+\epsilon)$ of the optimum or that disagrees with the optimum in at most an ϵ fraction of points. For the basic case of $k = 2$ (which is complement to the Max-Cut), Indyk [12] gave an $(1 + \epsilon)$-approximation algorithm that runs in $O(n^{1+\gamma} \cdot (\log n)^{(1/\epsilon)^{O(1)}})$ time for any $\gamma > 0$, which is sublinear in the full input description size but superlinear in n.

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, see, e.g. [3, Claim 1]. Therefore, balanced k-median has been usually considered in connection with the min-sum k-clustering problem discussed above. The problem was first studied by Guttman-Beck and Hassin [9] who gave an exact $O(n^{k+1})$-time algorithm and Bartal et al. [3] obtained an $O(\frac{1}{\epsilon} \log^{1+\epsilon} n)$-approximation in time $n^{O(1/\epsilon)}$ based on metric embeddings into HSTs. We are not aware of any sublinear-time algorithm for balanced k-median.

1.2 New contribution

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

We first study the k-**median** problem. Our sampling is identical to the one by Mishra et al. [19], however our analysis is stronger and leads to significantly better bounds. Let $\alpha \geq 1$, $0 < \delta < 1$, and $\epsilon > 0$ be arbitrary parameters. We prove that if we pick a sample set of size $\widetilde{O}(\frac{\Delta \cdot \alpha}{\epsilon^2} \cdot (k + \alpha \ln(1/\delta)))$ 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 $2\,(\alpha + \epsilon)\, med_{avg}(V,k) + \epsilon$ with probability at least $1 - \delta$; notice in particular, that this gives the sample size *independent of* n. As noted in [19], it is impossible to obtain a sample complexity independent of both Δ and n.

Comparing our result to the one from [19], we improve the sample complexity by a factor of $\Delta \cdot \log n$ while obtaining a slightly worse approximation ratio of $2\,(\alpha + \epsilon)\, med_{avg}(V,k) + \epsilon$, instead of $2\,\alpha\, med_{avg}(V,k) + \epsilon$ as in [19]. However, since the algorithm with the best known approximation guarantee has $\alpha = 3 + \frac{1}{c}$ for the running time of $O(n^c)$ time [2], we significantly improve the running time of [19] for all realistic choices of the input parameters while achieving the same approximation guarantee. As a highlight, we obtain an algorithm that in time $\widetilde{O}((\frac{\Delta\, k}{\epsilon^2} \cdot (k + \log(1/\delta)))^2)$ — *fully independent of* n — has the average distance to the nearest median at most $O(med_{avg}(V,k)) + \epsilon$ with probability at least $1 - \delta$.

Furthermore, our analysis can be significantly improved if we assume the input points are in Euclidean space $\mathbb{R}^d$. In this case we improve the approximation guarantee to $(\alpha + \epsilon)\, med_{avg}(V,k) + \epsilon$ in the cost of increasing the sample size to $\widetilde{O}(\frac{\Delta \cdot \alpha}{\epsilon^2} \cdot (k\,d + \log(1/\delta)))$. This bound also significantly improves an analysis from [19]. Due to space limitations we omit the corresponding proof in this extended abstract.

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 $\widetilde{O}\left(\frac{\Delta}{\epsilon} \cdot \left(\frac{\sqrt{k}\,\alpha^2\,\ln(1/\delta)}{\beta} + \frac{k + \ln(1/\delta)}{\epsilon}\right)\right)$ i.u.r., then an α-approximation of the optimal solution for the sample set approximates the average balanced k-median cost to within $(2\,\alpha + \beta)\, med^b_{avg}(V,k) + \epsilon$ with probability at least $1 - \delta$, where $med^b_{avg}(V,k)$ 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 in 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 $V_1 \cup \dots \cup V_k$ that minimizes (or approximates the minimum) of $\sum_{i=1}^{k} |V_i| \sum_{v \in V_i} \mu(v, c_i)$. Our sampling algorithm leads to a randomized algorithm that in time independent of n

returns the set of k centers $c_1, \ldots, c_k$ for which the value of $\frac{\sum_{i=1}^{k} |V_i| \sum_{v \in V_i} \mu(v,c_i)}{|V|^2}$ is at most $O(med^b_{avg}(V,k)) + \epsilon$ 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 $O(n\,k) + \widetilde{O}(k^{2.5}\,\sqrt{n})$ time an optimal clustering [22]. 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. What is 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.
run $\alpha$-approximation algorithm $\mathbb{A}$ on input $S$ to compute a solution $C^*$ (set of $k$ centers)
return set $C^*$
```

To analyze the approximation guarantee of this approach we proceed in two steps. First, we show that w.h.p. and after normalization $cost(S, C_{opt})$ is an approximation of $cost(V, C_{opt})$, where C_{opt} denotes an optimal solution for V. 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 $\frac{c}{\alpha} \cdot cost(S, C_{opt})$ for some constant $c \geq \alpha$. Then we show that w.h.p. every possible solution for V with cost more than $c \cdot cost(V, C_{opt})$ is either not a feasible solution for S or has cost more than $c \cdot cost(S, C_{opt})$ for S. Since S contains a solution with cost at most $\frac{c}{\alpha} \cdot cost(S, C_{opt})$, $\mathbb{A}$ will compute a solution C^* with cost at most $c \cdot cost(S, C_{opt})$. Since every solution for V with cost more than $c \cdot cost(V, C_{opt})$ has cost more than $c \cdot cost(S, C_{opt})$ for S, we know that $\mathbb{A}$ computes a solution C^* with cost at most $c \cdot cost(V, C_{opt})$ for V. Hence, our sampling is a c-approximation algorithm.

We apply this approach to study sampling algorithms for three problems: the k-median 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 $med_{opt}(V,k) = \min_{C \subseteq V,\, |C|=k} \sum_{v \in V} \mu(v, C)$ denote the *cost of a k-median* of V. Let $med_{avg}(V,k) = \frac{1}{|V|} \cdot med_{opt}(V,k)$ 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 $cost_{avg}(U,C) = \frac{1}{|U|} \sum_{v \in U} \mu(v, C)$. 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 $0 < \delta < 1$, $\alpha \geq 1$, 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*

$$s \ \geq \ c \cdot (1 + \alpha/\epsilon) \cdot \left(k + \frac{\varDelta}{\epsilon} \cdot \left(\alpha \cdot \ln(1/\delta) + k \cdot \ln\left(\tfrac{k \, \varDelta \, (1+\alpha/\epsilon)}{\epsilon} \right) \right) \right) \ ,$$

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

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

To begin our analysis of the quality of the approximation of C^* and the proof of Theorem 1, let us introduce some basic notation. Let $\beta > 0$, $\alpha \geq 1$. A set of k centers C is a *β-bad α-approximation* of k-median of V if $\mathit{cost}_{avg}(V, C) > (\alpha + \beta) \cdot \mathit{med}_{avg}(V, k)$. If C is not a β-bad α-approximation then it is a *β-good α-approximation*.

For the k-median problem we want to prove for certain s that 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 $2\,(1 + \beta/\alpha) \cdot \mathit{med}_{avg}(V, k)$, and hence, any α-approximation for S returns a $2\,(\alpha + \beta)$-approximation for V w.h.p. We prove the following lemma.

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

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

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

Observe that each X_i satisfies $0 \leq X_i \leq \varDelta$. Therefore, we can apply a Hoeffding bound to obtain:

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

Let C be the set of k centers in S obtained by replacing each $c \in C_{opt}$ by its nearest neighbor in S. By the triangle inequality, we get $\mathit{cost}_{avg}(S, C) \leq 2 \cdot \mathit{cost}_{avg}(S, C_{opt})$. Hence, multiset S contains a set of k centers whose cost is at most $2 \cdot (1 + \beta/\alpha) \cdot \mathit{med}_{avg}(V, k)$ 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 $(6\,\beta)$-bad $(2\,\alpha)$-approximation of a k-median of V satisfies $\mathit{cost}_{avg}(S, C_b) > 2\,(\alpha + \beta) \cdot \mathit{med}_{avg}(V, k)$ with high probability.

Lemma 2. *Let S be a multiset of s points chosen i.u.r. from V with s such that*

$$s \;\geq\; c \cdot \left((1+\alpha/\beta)\, k \;+\; \frac{(\alpha+\beta)\cdot \varDelta \cdot \left(\ln(1/\delta) + k\,\ln\left(\frac{k\,(\alpha+\beta)\,\varDelta}{\beta^2\, med_{avg}(V,k)}\right)\right)}{\beta^2\, med_{avg}(V,k)} \right) \;,$$

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

$$\mathbf{Pr}\Big[\exists C_b \in \mathbb{C} : C_b \subseteq S \textit{ and } cost_{avg}(S, C_b) \;\leq\; 2\,(\alpha+\beta)\, med_{avg}(V,k)\Big] \;\leq\; \delta \;.$$

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

$$\begin{aligned}
&\mathbf{Pr}\Big[C_b \subseteq S \text{ and } cost_{avg}(S,C_b) \;\leq\; 2\,(\alpha+\beta)\, med_{avg}(V,k)\Big] \\
&= \mathbf{Pr}\Big[cost_{avg}(S,C_b) \;\leq\; 2\,(\alpha+\beta)\, med_{avg}(V,k) \;\Big| C_b \subseteq S\Big] \cdot \mathbf{Pr}\Big[C_b \subseteq S\Big] \\
&= \mathbf{Pr}\Big[cost_{avg}(S^*,C_b) \leq 2 \cdot \frac{s}{s-k}\,((\alpha+\beta)\, med_{avg}(V,k))\Big] \cdot \mathbf{Pr}\Big[C_b \subseteq S\Big] \quad (1) \\
&\leq \mathbf{Pr}\Big[cost_{avg}(S^*,C_b) \leq 2\;(\alpha+1.5\,\beta)\, med_{avg}(V,k))\Big] \cdot \mathbf{Pr}\Big[C_b \subseteq S\Big] \;, \quad (2)
\end{aligned}$$

where (1) holds because the elements are chosen with repetition and (2) follows from $s \;\geq\; \frac{2\,\alpha+3\,\beta}{\beta}\,k$. Furthermore, similarly as in the proof of Lemma 1, we can prove the following inequality

$$\mathbf{Pr}\Big[cost_{avg}(S,C_b) \;\leq\; 2\,(\alpha+1.5\,\beta)\, med_{avg}(V,k) \;\Big|\Big] \;\leq\; e^{\frac{-s\,\beta^2\, med_{avg}(V,k)}{(\alpha+3\,\beta)\,\varDelta}} \;. \quad (3)$$

Therefore, we can plug inequality (3) and the identity $\mathbf{Pr}[C_b \subseteq S] = (s/n)^k$ into (2), and combine this with the upper bound $|\mathbb{C}| \leq n^k$, to conclude the proof. □

Proof of Theorem 1. Let s be chosen such that the prerequisites of Lemmas 1 and 2 hold, that is,

$$s \;\geq\; c\,(1+\alpha/\beta)\, \left(k + \frac{\varDelta}{\beta\, med_{avg}(V,k)}\, \left(\alpha \ln(1/\delta) + k \ln\left(\frac{k(\alpha+\beta)\varDelta}{\beta^2\, med_{avg}(V,k)}\right)\right)\right) \quad (4)$$

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

$$cost_{avg}(S,C) \;\leq\; 2\,(\alpha+\beta)\, med_{avg}(V,k) \;.$$

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

$$cost_{avg}(S,C^*) \;\leq\; 2\,(\alpha+\beta)\, med_{avg}(V,k) \;.$$

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

$$cost_{avg}(V, C^*) \le (2\,\alpha + 6\,\beta) \cdot med_{avg}(V, k) \ \ .$$

This implies immediately the following bound:

$$\mathbf{Pr}\Big[cost_{avg}(V, C^*) \le (2\,\alpha + 6\,\beta) \cdot med_{avg}(V, k)\Big] \ \ge \ 1 - 2\,\delta \ \ .$$

To complete the proof we only must remove the dependence of $med_{avg}(V, k)$ in the bound of s in (4) and relate β to ϵ. For $med_{avg}(V, k) \ge 1$, Theorem 1 follows directly from our discussion above by replacing $6\,\beta$ by ϵ. For $med_{avg}(V, k) < 1$, Theorem 1 follows by replacing β by $\epsilon / med_{avg}(V, k)$. For more details we refer to the full version of the paper. □

3 Min-sum k-clustering and balanced k-median in metric spaces

As we mentioned in Introduction, we follow the approach from [3] and [9] 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 $C = \{c_1, \dots, c_k\}$ of k points (centers) in V that minimizes the value of

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

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 $V_1 \cup \dots \cup V_k$ that minimizes the sum above.

Let

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

denote the *cost of a balanced k-median of* V, and let $med^b_{avg}(V, k) = \frac{1}{|V|^2} med^b_{opt}(V, k)$ denote the *average cost of a balanced k-median of* V. For a given set $U \subseteq V$ and a set of k centers $C = \{c_1, \dots, c_k\} \subseteq V$, let us define

$$cost^b(U, C) = \min_{\substack{\text{partition of } U \\ \text{into } U_1 \cup \dots \cup U_k}} \sum_{i=1}^{k} |U_i| \sum_{u \in U_i} \mu(u, c_i) \ \text{ and } \ cost^b_{avg}(U, C) = \frac{cost^b(U, C)}{|U|^2} \ \ .$$

A set of k centers C is called a (ϵ, β)-*bad α-approximation* of balanced k-median of V if $cost^b_{avg}(V, C) > (\alpha + \beta) \cdot med^b_{avg}(V, k) + \epsilon$. If C is not a (ϵ, β)-bad α-approximation then it is a (ϵ, β)-*good α-approximation*.

3.1 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 Section 3.2 we prove the following main theorem.

Theorem 2. *Let (V, μ) be a metric space. Let $\mathbb{A}$ be an α-approximation algorithm for balanced k-median in metric spaces and let $0 \le \epsilon \le 1/4$, $\beta \ge \frac{4\,\alpha\,\epsilon}{1-2\,\epsilon}$, $0 < \delta < 1$ be approximation parameters. If we choose a sample set $S \subseteq V$ of size s i.u.r., where*

$$s \;\ge\; \frac{c \cdot \varDelta}{\epsilon} \cdot \left(\frac{\sqrt{k}\;\ln(k/\delta)\,\alpha^2}{\beta} \;+\; \frac{\ln(k/\delta) + k \cdot \ln(k\,\varDelta/\epsilon)}{\epsilon} \right) \;,$$

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

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

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

Moreover, the solution C^* approximates an optimal solution for the min-sum k-clustering problem within a factor two times larger than claimed above.

The last claim in Theorem 2 follows from the fact that in metric spaces the solution to balanced k-median is within a factor of 2 of that of min-sum k-clustering.

3.2 Analysis of Generic sampling scheme for balanced k-median

Our analysis follows the path used in Section 2. The main difference is that we must explicitly use "outliers" in our analysis, what makes it significantly more complicated.

We begin with a result corresponding to Lemma 1 for k-median.

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

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

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

Let us call set V_i^* *dense* if $|V_i^*| \ge \frac{3 \cdot \ln(1/\delta_1)}{\gamma^2} \cdot \frac{|V|}{s}$; V_i^* is *sparse* otherwise. Let S_i be the random variable that denotes 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 (that can be easily proven using a Chernoff bound) is that if V_i^* is dense, then we have $\mathbf{Pr}\Big[|S_i| \le (1-\gamma) \cdot \frac{s \cdot |V_i^*|}{|V|}\Big] \le \delta_1$ and $\mathbf{Pr}\Big[|S_i| \ge (1+\gamma) \cdot \frac{s \cdot |V_i^*|}{|V|}\Big] \le \delta_1$, and if V_i^* is sparse, then we have $\mathbf{Pr}\Big[|S_i| \ge \frac{6 \cdot \ln(1/\delta_1)}{\gamma^2}\Big] \le \delta_1$.

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

For any set V_i^*, let X_i^j be the random variable that denotes the distance between the jth randomly selected element from S_i and the center c_i. Observe that for any set V_i^*, we have $\mathbf{E}[X_i^j] = \frac{1}{|V_i^*|}\cdot\sum_{u\in V_i^*}\mu(u,c_i)$. Let us fix i and let us first assume that

$$2\cdot\frac{|S_i|}{s^2}\cdot\gamma\cdot\frac{|S_i|}{|V_i^*|}\cdot\sum_{u\in V_i^*}\mu(u,c_i) \;\geq\; \epsilon/\alpha \ . \tag{5}$$

Since $0 \leq X_i^j \leq \mathit{\Delta}$, we use Hoeffding bound to prove

$$\mathbf{Pr}\Big[\sum_{j=1}^{|S_i|} X_i^j \;\geq\; (1+\gamma)\cdot|S_i|\cdot\frac{\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|}\Big] \leq \exp\Big(-\frac{\gamma}{3\cdot\mathit{\Delta}}\cdot s\cdot\epsilon/(2\alpha)\Big) \tag{6}$$

where the last inequality follows from (5). If (5) does not hold, then let γ^*, $\gamma^* > \gamma$, be such that

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

Notice that in that case,

$$\gamma^*\cdot\mathbf{E}\Big[\sum_{j=1}^{|S_i|} X_i^j\Big] \;=\; \gamma^*\cdot|S_i|\cdot\frac{\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|} \;=\; \frac{s^2\cdot\epsilon}{2\cdot\alpha\cdot|S_i|} \;\geq\; \frac{s\cdot\epsilon}{2\cdot\alpha} \ . \tag{7}$$

Observe that since (5) does not hold and since $\gamma \leq 1$, we have $\gamma \leq \min\{1,\gamma^*\}$. Therefore, we can use the Hoeffding bound to prove that

$$\begin{aligned}\mathbf{Pr}\Big[\sum_{j=1}^{|S_i|} X_i^j \geq (1+\gamma^*)\cdot\mathbf{E}\Big[\sum_{j=1}^{|S_i|} X_i^j\Big]\Big] &\leq \exp\Big(-\tfrac{\min\{\gamma^*,{\gamma^*}^2\}\cdot|S_i|}{3\cdot\mathit{\Delta}}\cdot\tfrac{\sum_{u\in V_i^*}\mu(u,c_i)}{|V_i^*|}\Big)\\ &\leq \exp\Big(-\frac{\gamma\cdot s\cdot\epsilon}{6\cdot\mathit{\Delta}\cdot\alpha}\Big) \ .\end{aligned} \tag{8}$$

Notice that the inequalities (6) – (8) imply that if $s \geq \frac{6\alpha\cdot\mathit{\Delta}\cdot\ln(1/\delta_1)}{\gamma\cdot\epsilon}$, then

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

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

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

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

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

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

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

what concludes the proof of Lemma 3. □

Lemma 3 (with $\gamma \approx \alpha/\beta$) can be combined with arguments used in Lemma 1 to prove the following.

Corollary 1. *Let $0 < \beta < \alpha$ and $\epsilon > 0$. Let S be a multiset of size $s \geq \frac{c\sqrt{k}\varDelta \ln(3k/\delta)\,\alpha^2}{\beta\,\epsilon}$ chosen from V i.u.r., where c is some constant. If an α-approximation algorithm for balanced k-median $\mathbb{A}$ is run with input S, then for the solution C^* obtained by $\mathbb{A}$ holds*

$$\mathbf{Pr}\Big[cost^b_{avg}(S, C^*) \;\leq\; 2\,(\alpha + \beta) \cdot med^b_{avg}(V,k) + \epsilon \Big] \;\geq\; 1 - \delta \enspace . \qquad \square$$

The next step in our analysis is to consider bad approximations. Our analysis follows the approach used before in the proof of Lemma 2; the main difference is a larger number of parameters used in the analysis. Corollary 1 proves that typically there is a set of k centers in the sample S that has the average cost close to $med^b_{avg}(V,k)$. Now, we show that any $C_b \subseteq S$ that is a $(5\,\epsilon, 2\,\beta)$-bad $(2\,\alpha)$-approximation of a balanced k-median of V satisfies $cost_{avg}(S, C_b) \;>\; 2\,(\alpha + \beta) \cdot med^b_{avg}(V,k) + \epsilon$ with high probability. Details of the proof of the following lemma are deferred to the full version of the paper.

Lemma 4. *Let S be a multiset of s points chosen i.u.r. from V with s such that:*

$$s \;\geq\; c \cdot \left(\frac{\varDelta}{\epsilon^2} \cdot (\ln(k/\delta) + k \cdot \ln(k\,\varDelta/\epsilon)) + \frac{1}{\beta} \right) \enspace ,$$

where c is a suitable positive constant. Let $\mathbb{C}$ be the set of $(5\epsilon, 2\,\beta)$-bad $(2\,\alpha)$-approximations C of a balanced k-median of V. Then,

$$\mathbf{Pr}\Big[\exists C_b \in \mathbb{C} : C_b \subseteq S \text{ and } cost_{avg}(S, C_b) \leq (1-\epsilon)^2\,(2\,\alpha + \beta)\, med^b_{avg}(V,k) + \epsilon \Big] \leq \delta.$$

Now Theorem 2 follows from Corollary 1 and Lemma 4. To expand our implicit representation of the clustering, we can use the values v_i^* obtained from the optimum partition of our sample set S as cluster sizes and then use the algorithm from [22]. □

References

1. S. Arora, P. Raghavan, and S. Rao. Approximation schemes for Euclidean k-medians and related problems. *30th STOC*, pp. 106–113, 1998.
2. V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *33rd STOC*, pp. 21–30, 2001.
3. Y. Bartal, M. Charikar, and D. Raz. Approximating min-sum k-clustering in metric spaces. *33rd STOC*, pp. 11–20, 2001.
4. M. Charikar and S. Guha. Improved combinatorial algorithms for the facility location and k-median problems. *40th FOCS*, pp. 378–388, 1999.
5. M. Charikar, S. Guha, É. Tardos, and D. B. Shmoys. A constant-factor approximation algorithm for the k-median problem. *31st STOC*, pp. 1–10, 1999.

   M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. *12th SODA*, pp. 642–651, 2001.
6. M. Charikar, L. O'Callaghan, and R. Panigrahy. Better streaming algorithms for clustering problems. *35th STOC*, pp. 30–39, 2003.
7. B. Chazelle. Who says you have to look at the input? The brave new world of sublinear computing? *15th SODA*, p. 134, 2004.
8. W. Fernandez de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani. Polynomial time approximation schemes for metric min-sum clustering. *35th STOC*, pp. 50–58, 2003.
9. N. Gutmann-Beck and R. Hassin. Approximation algorithms for min-sum p-clustering. *Discrete Applied Mathematics*, 89: 125–142, 1998.
10. S. Har-Peled and S. Mazumdar. Coresets for k-means and k-median clustering and their applications. *36th STOC*, 2004.
11. P. Indyk. Sublinear time algorithms for metric space problems. *31st STOC*, pp. 428–434, 1999.
12. P. Indyk. A sublinear time approximation scheme for clustering in metric spaces. *40th FOCS*, pp. 154–159, 1999.
13. K. Jain, M. Mahdian, and A. Saberi. A new greedy approach for facility location problems. *34th STOC*, pp. 731–740, 2002.
14. K. Jain and V. V. Vazirani. Primal-dual approximation algorithms for metric facility location and k-median problems. *40th FOCS*, pp. 2–13, 1999.
15. S. G. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the Euclidean k-median problems. *7th ESA*, pp. 378–389, 1999.
16. R. Kumar and R. Rubinfeld. Sublinear time algorithms. *SIGACT News*, 34(4):57–67, 2003.
17. R. R. Mettu and C. G. Plaxton. Optimal time bounds for approximate clustering. *18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 344–351, August 2002.
18. A. Meyerson, L. O'Callaghan, and S. Plotkin. A k-median algorithm with running time independent of data size. *Journal of Machine Learning*, 2004.
19. N. Mishra, D. Oblinger, and L. Pitt. Sublinear time approximate clustering. *12th SODA*, pp. 439–447, 2001.
20. L. J. Schulman. Clustering for edge-cost minimization. *32nd STOC*, pp. 547–555, 2000.
21. S. Sahni and T. Gonzalez. $\mathcal{P}$-complete approximation problems. *JACM*, 23: 555-566, 1976.
22. T. Tokuyama, and J. Nakano. Geometric algorithms for the minimum cost assignment problem. *Random Structures and Algorithms*, 6(4): 393-406, 1995.