# Differentially-Private Sublinear-Time Clustering

Jeremiah Blocki[*], Elena Grigorescu[†], Tamalika Mukherjee
Department of Computer Science, Purdue University.
{jblocki, elena-g, tmukherj}@purdue.edu

February 13, 2021

**DRAFT**

**Abstract**

Clustering is an essential primitive in unsupervised machine learning. We bring forth the problem of sublinear-time diferentially-private clustering as a natural and well-motivated direction of research. We combine the $k$-means and $k$-median sublinear-time results of Mishra et al. (SODA, 2001) and of Czumaj and Sohler (Rand. Struct. and Algorithms, 2007) with recent results on private clustering of Balcan et al. (ICML 2017), Gupta et al. (SODA, 2010) and Ghazi et al. (NeurIPS, 2020) to obtain sublinear-time private $k$-means and $k$-median algorithms via subsampling. We also investigate the privacy benefits of subsampling for group privacy.

## 1 Introduction

Preserving privacy in data collection and distribution have long been a concern for industrial and governmental agencies, who are now rapidly adopting privacy standards and policies [24, 11, 6, 13]. Differential privacy [12] is the gold standard of privacy protection. A randomized function computed on a database is *differentially private* if the distribution of the function's output does not change by much with the presence or absence of an individual record. While existing research mostly focuses on computing efficient polynomial-time differentially-private algorithms, in dealing with a large amount of data, even linear-time algorithms may be prohibitive in costs. Hence, algorithms that can quickly output approximately accurate solutions while preserving privacy are of great interest in real-world computations on large datasets (e.g., billions of Facebook or Google, or Microsoft users). However, despite the fact that the literature on differentially private algorithms has grown rapidly in recent years, sublinear-time private algorithms for many natural problems are still lacking. In this work we focus on clustering problems and provide some basic sublinear-time private solutions derived from the existing efficient analogues.

Clustering is an essential primitive in unsupervised machine learning. Since many machine learning models deal with sensitive data, private clustering has been studied extensively in the polynomial-time setting [28, 14, 17, 35, 4, 22, 20, 32, 16, 34, 29]. Two of the most widely studied

[*] J. B. was supported in part by NSF CNS-1931443 and NSF CCF-1910659.

[†] E.G was supported in part by NSF CCF-1910659 and NSF CCF-1910411.

variants of clustering are the $k$-median and $k$-means problem. In the $k$-median problem, we are given $n$ data points, and the goal is to find $k$ centers that minimize the sum of distances from the data points to their nearest centers. The setup is the same for $k$-means, except the goal is to find $k$ centers that minimize the sum of the squares of distances from the data points to their nearest centers. Both types of clustering are classical problems, and there is a rich field of research devoted to them in the non-private setting [2, 7, 18, 8, 9, 3, 25, 1, 30].

## 1.1 Contributions

We bring forth the problem of sublinear-time private clustering as a natural and well-motivated direction of research, and show some basic results derived from the non-private analogues on subsampled data. We expect that our results will entice further interest in understanding the best privacy guarantees in sublinear clustering settings.

**Private sublinear clustering.** We combine the techniques of sublinear-time clustering algorithms from Mishra et al. [27] and Czumaj et al. [10] with the private polynomial-time approximation clustering algorithms with a constant multiplicative factor of Balcan et al. [4], Gupta et al. [17] and Ghazi et al. [16] to obtain private sublinear-time clustering algorithms for $k$-median and $k$-means clustering in metric spaces, as well as better approximation guarantees for the particular case of Euclidean space. To the best of our knowledge, these are the first *sublinear-time differentially-private clustering* algorithms formalized in the privacy literature.

We analyze the following sampling algorithm: pick a random sample from the input set; run a private $k$-median (or $k$-means) polynomial-time approximation algorithm on the random sample to obtain a $k$-median (or $k$-means) clustering of the sample; output this clustering. We show that for a small sample size, the average cost of the clustering induced by the random sample is not too far from the average cost of the optimum clustering of the input set. Our analysis closely follows the works of Mishra et al. [27] and Czumaj et al. [10], who gave sublinear time algorithms for clustering in the non-private setting using a constant $\alpha$-approximation polynomial-time algorithm as a black-box. We extend their analysis to handle the case of using an $(\alpha, \gamma)$-approximation polynomial time algorithm as black-box [1]. The approximation guarantee achieved by our algorithm is essentially the same as that of the black-box private algorithms (modulo an extra additive factor of $\epsilon$) [2] For an arbitrary metric space $(V, d)$ consisting of $n$ points and input set $D \subseteq V$,

1. Assuming a private $(\alpha, \gamma)$-factor approximation $k$-median algorithm, that runs in time $T(n)$, we can draw a sample $S \subseteq D$ of size $\textit{poly}\,(\alpha, k \ln n)$ and obtain a $k$-median clustering $\hat{c}_S$ in time $T(s)$ such that with high probability $\textsf{avg-cost}(\hat{c}_S) \leq \alpha \cdot \textsf{avg-cost}(c_D) + \gamma + \epsilon$, where $c_D$ is the optimum $k$-median clustering of $D$.

2. Assuming a private $(\alpha, \gamma)$-factor approximation $k$-means algorithm that runs in time $T(n)$, we can draw a sample $S \subseteq D$ of size $\textit{poly}\,(\alpha, k \ln n)$ and obtain a $k$-means clustering $\hat{c}_S$ such

[1] We note that an additive approximation factor $\gamma > 0$ is unavoidable for any *private* clustering algorithm, thus this extension was necessary. To see why, consider the following two multisets of input data points $D_1 = \{x_1, \ldots, x_1, x_2, \ldots x_{k-1}, x_k\}$ and $D_2 = \{x_1, \ldots, x_1, x_2, \ldots x_{k-1}, x_{k+1}\}$, where $x_1$ occurs in both sets $n - k + 1$ times. Note that the optimal cost for $k$-median in both cases is zero, and in the non-private setting, the algorithm can simply output $\{x_1, \ldots, x_k\}$ as the solution. But a private algorithm must have an additive error since the set of centers computed by our algorithm cannot be affected by the change of replacing $x_k$ in $D_1$ by $x_{k+1}$ in $D_2$, and the input points being private, should not be revealed by our algorithm.

[2] This extra additive factor of $\epsilon$ is unavoidable in order to design clustering algorithms with running time $o(n)$, see [10] for an exposition.

that with high probability $\mathsf{avg\text{-}cost}(\hat{c}_S) \leq \alpha \cdot \mathsf{avg\text{-}cost}(c_D) + \gamma + \epsilon$, where $c_D$ is the optimum $k$-means clustering of $D$.

For the special case of $k$-median in $d$-dimensional Euclidean space, we achieve a sample complexity that is independent of the size of the input set $D \subseteq \mathbb{R}^d$ consisting of $n$ points.

1. Assuming a private $(\alpha, \gamma)$-factor approximation $k$-median algorithm, that runs in time $T(n)$, we can draw a sample $S \subseteq D$ of size $poly(\alpha, d, k)$ and obtain a $k$-median clustering $\hat{c}_S$ in time $T(s)$ such that with high probability $\mathsf{avg\text{-}cost}(\hat{c}_S) \leq \alpha \cdot \mathsf{avg\text{-}cost}(c_D) + \gamma + \epsilon$, where $c_D$ is the optimum $k$-median clustering of $D$.

2. Assuming a private $(\alpha, \gamma)$-factor approximation $k$-means algorithm, that runs in time $T(n)$, we can draw a sample $S \subseteq D$ of size $poly(\alpha, d, k)$ and obtain a $k$-means clustering $\hat{c}_S$ in time $T(s)$ such that with high probability $\mathsf{avg\text{-}cost}(\hat{c}_S) \leq \alpha \cdot \mathsf{avg\text{-}cost}(c_D) + \gamma + \epsilon$, where $c_D$ is the optimum $k$-means clustering of $D$.

**Group privacy for sampling algorithms.** Group privacy ensures that for pairs of inputs that differ on a small number of points, the privacy loss is still bounded. For example, in the setting of a health survey administered to families, a family may wish to preserve all its members' privacy. Any $\xi$-differentially private algorithm, ensures $(g\xi, 0)$-privacy for groups of size $g$. We show that our random sampling algorithm has better group privacy guarantees. In other words, an algorithm that runs an $(\xi, 0)$-differentially private mechanism on a subsample is $(T \cdot \xi, \delta_T)$-differentially private for groups of size $g$, for $0 \leq T \leq g$, where $\delta_T$ is the probability of the number of samples from the $g$ elements is $> T$. We note that $\delta_T$ is often negligible even for $T \ll g$. In such cases, the guarantee of $(T \cdot \xi, \delta_T)$-differential privacy is arguably much stronger than the naive guarantee of $(g\xi, 0)$-group privacy.

## 1.2 Related Work

Sublinear-time approximate $k$-median clustering of a space in which the diameter of points is bounded was introduced by Mishra et al. [27]. They modeled clusterings as functions and studied the quality of $k$-median clusterings obtained by random sampling using computational learning theory techniques. For a metric space, their work shows that if we sample a set of size $poly(\alpha, k \ln n)$ and run an $\alpha$-approximation clustering algorithm on the sample, then with high probability, the set of centers outputted is at most $2\alpha \cdot \mathsf{avg\text{-}cost}(c_{OPT}) + \epsilon$. Their sampling model was adapted by Czumaj et al. [10], who achieved a sample complexity that is independent of $n$ for the $k$-median clustering problem in arbitrary metric spaces. They also extended the random sampling model and their analysis to give sublinear-time results for clustering variants such as $k$-means and min-sum clustering.

Private clustering was first studied by Gupta et al. [17], and Feldman et al. [14]. Gupta et al. [17] modified the local search algorithm for $k$-median by Arya et al. [2] to choose candidate centers in each iteration via the exponential mechanism [26] and produced a polynomial-time algorithm that achieves $(O(1), \tilde{O}(k^2 M))$-approximation ($M$ is the diameter of the space) in discrete spaces. However, their algorithm is highly inefficient in Euclidean space (see [22] for a detailed exposition). A recent line of work has focused on producing an efficient polynomial time algorithm for clustering that achieves a constant (multiplicative) factor approximation in high-dimensional Euclidean space by adopting the techniques of Gupta et al. while maintaining efficiency [4, 22]. A different approach to private clustering was taken by [14]. They gave an efficient algorithm for $k$-median and $k$-means in Euclidean space by introducing the notion of private coresets. A recent line of work has adopted their techniques to give clustering algorithms with better approximation guarantees and efficiency [15, 29, 16].

Privacy amplification by subsampling has been formally studied by Balle et al. [5]. Our result is a simple observation that tailors the privacy amplification achieved with respect to group privacy for a generic sampling algorithm that runs a private algorithm as a black-box in the sampling step.

# 2 Preliminaries

In the following discussion, let $(V, d)$ be an arbitrary metric space.

The expected (average) value of a function $f$ over uniform elements of a set $X$ is denoted as $\mathbb{E}_X[f]$, i.e. $\mathbb{E}_X[f] = \sum_{x \in X} \Pr_{x \leftarrow X}[x] \cdot f(x)$, where $x$ is picked uniformly from $X$.

**Differential Privacy.** Datasets $D$ and $D'$ are *neighboring* if removing or adding one point in $D$ results in $D'$[3]. A randomized algorithm $\mathcal{M}$ taking as input a dataset $D$ is $(\xi, \delta)$-*differentially private* if for any two neighboring data sets $D$ and $D'$, and for any subset $C$ of outputs of $\mathcal{M}$ it holds that $\Pr[\mathcal{M}(D) \in C] \le e^{\xi} \cdot \Pr[\mathcal{M}(D') \in C] + \delta$. If $\delta = 0$, $\mathcal{M}$ is $\xi$-*differentially private.*

**Clustering.** Given an input set $D \subseteq V$, the goal of the *k-median* clustering problem is to find a set of centers (i.e. a clustering) $\{c_1, \dots, c_k\} \subseteq V$ such that the cost of clustering $\sum_{x \in D} \min_i d(x, c_i)$ is minimized. The goal of the *k-means* clustering problem is to find a clustering $\{c_1, \dots, c_k\} \subseteq V$ such that the cost $\sum_{x \in D} \min_i d^2(x, c_i)$ is minimized. Using the notation of [27], we express clusterings as functions, i.e. for a choice of centers $c_1, \dots, c_k$, let $f_{c_1,\dots,c_k}$ denote a function for this $k$-median clustering $\{c_1, \dots, c_k\}$ such that for each point $x \in D$, we have that $f_{c_1,\dots,c_k}(x)$ is the distance of the closest center $c_i$ to $x$, in other words, $f_{c_1,\dots,c_k}(x) = \min_i d(x, c_i)$. Then the cost of the $k$-median clustering $\{c_1, \dots, c_k\}$ can now be expressed as $\sum_{x \in D} f_{c_1,\dots,c_k}(x)$. Similarly, for a choice of centers $c_1, \dots, c_k$, let $f_{c_1,\dots,c_k}$ denotes a function for this $k$-means clustering $\{c_1, \dots, c_k\}$ such that for each point $x \in D$, $f_{c_1,\dots,c_k}(x) = \min_i d^2(x, c_i)$, then the cost of $k$-means clustering $\{c_1, \dots, c_k\}$ can also be expressed as $\sum_{x \in D} f_{c_1,\dots,c_k}(x)$.

Note that our goal of finding a $k$-median (or $k$-means) clustering $\{c_1, \dots, c_k\}$ of $D$ that minimizes the cost $\sum_{x \in D} f_{c_1,\dots,c_k}(x)$ is equivalent (up to a scaling factor) to finding a $k$-median (or $k$-means) clustering $\{c_1, \dots, c_k\}$ of $D$ such that the average cost $\mathbb{E}_D[f_{c_1,\dots,c_k}]$ is minimized. The class of *k-median* cost functions is naturally defined as $F_V = \{f_{c_1,\dots,c_k} : f_{c_1,\dots,c_k}(x) = \min_i d(x, c_i),\ c_1, \dots, c_k \in V\}$. The optimum clustering for set $D \subseteq V$ can be expressed as a function $c_D := \text{argmin }_{f \in F_V} \mathbb{E}_D[f]$. Similarly, the class of *k-mean* cost functions is naturally defined as $F_V^2 = \{f_{c_1,\dots,c_k} : f_{c_1,\dots,c_k}(x) = \min_i d^2(x, c_i),\ c_1, \dots, c_k \in V\}$. The optimum clustering for set $D \subseteq V$ can be expressed as a function $c_D := \text{argmin }_{f \in F_V^2} \mathbb{E}_D[f]$.

An $(\alpha, \gamma)$-approximation algorithm for $k$-median (or $k$-means) takes as input a set $D$ (say), and outputs clustering $\hat{c_1}, \dots, \hat{c_k}$ such that $\sum_{x \in D} f_{\hat{c_1},\dots,\hat{c_k}}(x) \le \alpha \cdot \sum_{x \in D} f_{c_1,\dots,c_k}(x) + \gamma$, where $c_1, \dots, c_k$ is the optimum clustering for $D$. Equivalently, if $\hat{c}_D$ denotes the approximate clustering of $D$, and $c_D$ denotes the optimum clustering of $D$, then the $(\alpha, \gamma)$-approximation algorithm guarantees that $\mathbb{E}_D[\hat{c}_D] \le \alpha \cdot \mathbb{E}_D[c_D] + \gamma$, up to a scaling factor for $\gamma$.

[3] alternatively, if changing one data point in $D$ results in $D'$.

# 3 Private Sublinear time Approximate Clustering

In this section we describe the generic random sampling algorithm $\mathcal{A}'$ using a private $(\xi, \delta)$-differentially private as a black-box, and in the sequel, we show that $\mathcal{A}'$ is $(\xi', \delta')$-differentially private where $\xi'$ and $\delta'$ are functions of $\xi, \delta$ (see Theorem 1). Additionally, we give the accuracy of $\mathcal{A}'$, i.e., the minimum sample size needed to guarantee that with high probability the approximate clustering cost of the sample $S$ will be close to the true clustering cost of the input set $D$ when $D$ is a subset of an arbitrary metric space (see Theorem 3,Theorem 7) and in the special case of Euclidean space (see Theorem 5, Theorem 8).

**Remark.** For the metric setting, both [10] and [27] consider clusterings where the centers are a subset of the set of input data points (this type of clustering is known as discrete clustering). By carefully conditioning on this requirement, [10] can make the sample complexity independent of $n$. Unfortunately, due to privacy concerns, we must consider the set of chosen $k$ centers to be any subset of the entire metric space, and not restricted to the input set (this type of clustering is known as continuous clustering). Thus we cannot hope to achieve a sample complexity independent of $n$ in the metric setting, using their approach.

We present techniques used by [27] for our $k$-median clustering analysis and describe the techniques used by [10] for our $k$-means clustering analysis.

## 3.1 Generic Algorithm $\mathcal{A}'$

We first present the basic sampling algorithm we employ, this model was first introduced in [27]. Note that the sampling probability $\xi_1$ should be chosen as $o(1)$.

**Algorithm 1** General Sampling Scheme $\mathcal{A}'$

```
On input $D, \xi_1$
Sample each element of $D$ independently w.p. $\xi_1$ and let $S$ be the sample set.
Run $(\xi, \delta)$-DP $(\alpha, \gamma)$-approximation algorithm $\mathcal{A}$ on $S$ to compute a set of private $k$-centers for $S$, denoted by $C^*$.
Output the clustering $C^*$.
```

## 3.2 Privacy of $\mathcal{A}'$

In this section we show that for an algorithm $\mathcal{A}'(D)$ which takes $D$ as input and runs a $(\xi, \delta)$-differentially private algorithm $\mathcal{A}$ on random sample $S \subseteq D$, it is the case that $\mathcal{A}'$ is $(\xi', \delta')$-differentially private. Many works prove something similar to the following, e.g., [23, 5]. We include the proof here for the sake of clarity and completeness [4].

**Theorem 1.** *If $\mathcal{A}$ is an $(\xi, \delta)$-differentially private algorithm, and algorithm $\mathcal{A}'$ is the generic sampling algorithm defined above where each element is sampled independently with probability $\xi_1$, then $\mathcal{A}'$ is $(\xi', \delta')$-differentially private, where $\xi' = \ln \max\left\{\xi_1(e^{\xi} - 1) + 1, (\xi_1(e^{-\xi} - 1) + 1)^{-1}\right\}$, and $\delta' = \max\{\frac{e^{-\xi}\delta\xi_1}{(\xi_1(e^{-\xi}-1)+1)}, \delta\xi\}$.*

[4] Our proof slightly generalizes the analysis given by Adam Smith in his blog post [33].

Observe that if $\mathcal{A}$ is $(\xi, \delta)$-DP, then trivially, $\mathcal{A}'$ is also $(\xi, \delta)$-DP. The privacy bounds achieved in the above theorem are significantly better than these naive bounds. For example, if we consider $\xi = 0.5, \xi_1 = 0.001$, for any $\delta \in [0, 1)$, we achieve $\xi' < 0.00065$, and $\delta' = 0.001\delta$, which is orders of magnitude smaller than $\xi$ and $\delta$.

*Proof.* Let $D$ and $D'$ be neighboring data sets i.e. $D' = D \cup \{x\}$, and let us fix any subset $C$ of all possible outcomes in the output space.

Let $S$ be the set sampled from $D$ and $S'$ be the set sampled from $D'$, where each element from $D$ is independently chosen to belong to $S$ w.p. $\xi_1$, and similarly, each element from $D'$ is selected in $S'$ w.p. $\xi_1$. Since $\mathcal{A}$ is differentially private we have that for any valid subset of outcomes $C$, for all $y \notin S$,

$$\Pr[\mathcal{A}(S \cup \{y\}) \in C] \leq e^{\xi} \Pr[\mathcal{A}(S) \in C] + \delta \ , \tag{1}$$

$$\Pr[\mathcal{A}(S) \in C] \leq e^{\xi} \Pr[\mathcal{A}(S \cup \{y\}) \in C] + \delta \ . \tag{2}$$

We will show that $\Pr_{S' \leftarrow D'}[\mathcal{A}(S') \in C] \leq e^{\xi'} \Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] + \delta'$, and $\Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] \leq e^{\xi'} \Pr_{S' \leftarrow D'}[\mathcal{A}(S') \in C] + \delta'$, which shows that $\Pr[\mathcal{A}'(D') \in C] \leq e^{\xi'} \Pr[\mathcal{A}(D) \in C] + \delta'$, and $\Pr[\mathcal{A}'(D) \in C] \leq e^{\xi'} \Pr[\mathcal{A}'(D') \in C] + \delta'$, and hence $\mathcal{A}'$ is differentially private.

Indeed, using eq. (1), we have

$$\begin{aligned}
\Pr_{S' \leftarrow D'}[\mathcal{A}(S') \in C] &= \Pr[x \notin S'] \Pr[\mathcal{A}(S') \in C \ | x \notin S'] + \Pr[x \in S'] \Pr[\mathcal{A}(S') \in C \ | x \in S'] \\
&= (1 - \xi_1) \cdot \Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] + \xi_1 \cdot \Pr[\mathcal{A}(S') \in C \ | x \in S'] \\
&\leq (1 - \xi_1) \cdot \Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] + \xi_1 \cdot (e^{\xi} \Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] + \delta) \\
&= (\xi_1(e^{\xi} - 1) + 1) \cdot \Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] + \delta\xi_1
\end{aligned}$$

Now we want to lower bound $\Pr_{S' \leftarrow D'}[\mathcal{A}(S') \in C]$ using eq. (2),

$$\begin{aligned}
\Pr_{S' \leftarrow D'}[\mathcal{A}(S') \in C] &= \Pr[x \notin S'] \Pr[\mathcal{A}(S') \in C \ | x \notin S'] + \\
&+ \Pr[x \in S'] \Pr[\mathcal{A}(S') \in C \ | x \in S'] \\
&= (1 - \xi_1) \cdot \Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] + \\
&+ \xi_1 \cdot \Pr[\mathcal{A}(S') \in C \ | x \in S'] \\
&\geq (1 - \xi_1) \cdot \Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] + \\
&+ \xi_1 \cdot e^{-\xi} \cdot (\Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] - \delta) \\
&= (\xi_1(e^{-\xi} - 1) + 1) \cdot \Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] - e^{-\xi}\delta\xi_1
\end{aligned}$$

It follows that

$$\Pr_{S \leftarrow D}[\mathcal{A}(S) \in C] \leq \frac{\Pr_{S' \leftarrow D'}[\mathcal{A}(S') \in C]}{(\xi_1(e^{-\xi} - 1) + 1)} + \frac{e^{-\xi}\delta\xi_1}{(\xi_1(e^{-\xi} - 1) + 1)}$$

We can set

$$\delta' = \max\{\frac{e^{-\xi}\delta\xi_1}{(\xi_1(e^{-\xi} - 1) + 1)}, \delta\xi\}$$

and

$$\xi' = \ln \max\left\{\xi_1(e^{\xi} - 1) + 1, (\xi_1(e^{-\xi} - 1) + 1)^{-1}\right\} .$$

□

### 3.3 Private $k$-median in Metric Space

We consider an input set $D \subseteq V$, $|V| = n$, and let $M$ be the diameter of $V$.

**Lemma 1** (Uniform Convergence Lemma [19, 31]). *Let $F$ be a finite set of functions on $X$ with $0 \leq f(x) \leq M$ for all $f \in F$ and $x \in X$. Let $S = x_1, \ldots, x_m$ be a sequence of $m$ examples drawn independently and identically from $X$ and let $\epsilon > 0$. Then*

$$\Pr[\exists f \in F : |\mathop{\mathbb{E}}_{X}[f] - \mathop{\mathbb{E}}_{S}[f]| \geq \epsilon] \leq \delta \ ,$$

*when $m \geq \frac{M^2}{2\epsilon^2}(\ln |F| + \ln \frac{2}{\delta})$.*

As noted by [27], the above lemma implies fast uniform convergence of the family of $k$-median cost functions $F_V$. Since $|F_V|$ denotes the number of different $k$-median clusterings, it is given by $|F_V| = \binom{n}{k} = O(n^k)$.

We extend the results of Mishra et al. [27] to show that the sampling algorithm $\mathcal{A}'$ gives a reasonable $k$-median clustering in metric space using an approximation algorithm $\mathcal{A}$ that gives a constant multiplicative error $\alpha$, along with an additive error $\gamma$ as a black-box[5]. Note that this extension (formally defined in Lemma 2) applies to any approximation algorithm (non-private or private) that has a constant multiplicative error, along with an additive error and that is used as the black-box clustering algorithm within our sampling algorithm.

**Lemma 2.** *For an arbitrary metric space $(V, d)$ of $n$ points, and a set of points $D \subseteq V$ assuming an $(\alpha, \gamma)$-approximation $k$-median algorithm, where $\alpha \geq 0.33$ is a constant, that runs in time $T(n)$, we can draw a sample $S \subseteq D$ of size $s$,*

$$s = \Omega\Big(\Big(\frac{\alpha M}{\epsilon}\Big)^2 \Big(k \ln n + \ln \frac{1}{\delta}\Big)\Big)$$

*and obtain a $k$-median clustering $\hat{c}_S$ in time $T(|S|)$ such that with probability at least $1-\delta$, $\mathbb{E}_D[\hat{c}_S] \leq \alpha\, \mathbb{E}_D[c_D] + \gamma + \epsilon$, where $c_D$ represents the optimum clustering of $D$ and $\epsilon > 0$.*

*Proof.* The main idea is to show that the sample approximate cost of $\hat{c}_S$ and the true cost of $c_D$ are close. The proof proceeds by first outlining the different relations between $c_D, \hat{c}_S$ and $c_S$, and then combining them appropriately.

Recall $\hat{c}_S$ is the approximate clustering obtained by running $\mathcal{A}(S)$, $c_D = \text{argmin}_{\ f \in F_V} \mathbb{E}_D[f]$ is the optimum clustering of $D$, and $c_S = \text{argmin}_{\ f \in F_V} \mathbb{E}_S[f]$ is the optimum clustering of sample $S$. We first apply Lemma 1 to the family of $k$-median cost functions $F_V$ and obtain that for $s \geq 8\left(\frac{\alpha M}{\epsilon}\right)^2\left(k \ln n + \ln \frac{4}{\delta}\right)$ with probability at least $1 - \delta/2$, $|\mathbb{E}_S[c_D] - \mathbb{E}_D[c_D]| < \frac{\epsilon}{4\alpha}$ (3), and by the same Lemma 1 and size of sample set $s$ we have that with probability at least $1 - \delta/2$,

[5] Mishra et al. [27] showed similar results when using an $\alpha$-approximation algorithm $\mathcal{A}$ as a black-box, where $\alpha$ is a constant multiplicative error.

$|\mathbb{E}_D[\hat{c}_S] - \mathbb{E}_S[\hat{c}_S]| < \frac{\epsilon}{4\alpha}$ (4). Since $c_S$ is the optimum clustering of $S$, we have that $\mathbb{E}_S[c_S] \leq \mathbb{E}_S[c_D]$. By combining this with (3), we obtain that with probability at least $1-\delta/2$, $\mathbb{E}_S[c_S] - \mathbb{E}_D[c_D] < \frac{\epsilon}{4\alpha}$ (5).

Now, by virtue of running $\mathcal{A}$ on sample $S$, we have the guarantee that $\mathbb{E}_S[\hat{c}_S] \leq \alpha\, \mathbb{E}_S[c_S] + \gamma$. By adding this to (5) and rearranging, we obtain that with probability at least $1 - \delta/2$, $\mathbb{E}_S[\hat{c}_S] - \gamma - \alpha\, \mathbb{E}_D[c_D] \leq \frac{\epsilon}{4}$ (6). Lastly, by combining (4) and (6), we obtain that with probability at least $1 - \delta$, $\mathbb{E}_D[\hat{c}_S] - \alpha\, \mathbb{E}_D[c_D] - \gamma \leq \frac{\epsilon}{4} + \frac{\epsilon}{4\alpha} \leq \epsilon$ (7), where the last step is true for $\alpha \geq 0.33$. □

Given a metric space $(V, d)$ of $n$ points with diameter $M$, a private set $D \subseteq V$, Gupta et al. [17] modify a non-private local clustering algorithm [2] for solving $k$-median to make it differentially-private. Their algorithm starts off with an arbitrary set of $k$-centers and in each iteration, it swaps out an existing center in the set with a better center using the exponential mechanism, and after a sufficient number of steps, the algorithm chooses a good solution from amongst the ones seen so far. They obtain the following accuracy guarantee for their private algorithm.

**Theorem 2.** *[17] Given a metric space $(V, d)$ of $n$ points with diameter $M$, a set $D \subseteq V$, there exists a $\xi$-differentially private $k$-median algorithm that except with probability $O(1/poly(n))$ outputs a $(6, O(Mk^2 \log^2(n/\xi)))$-approximation of a $k$-median clustering of $D$.*

We will use the algorithm in [17] as our black-box algorithm $\mathcal{A}$. By plugging in the approximation guarantees for $\mathcal{A}$ into our Lemma 2, we get the following accuracy guarantee for our algorithm $\mathcal{A}'$.

**Theorem 3** (Accuracy of $\mathcal{A}'$). *For an arbitrary metric space $(V, d)$ of $n$ points, and a private set of points $D \subseteq V$, given the $\xi$-DP $(6, O(Mk^2 \log^2(n/\xi)))$-approximation $k$-median algorithm (from [17]), we have a $\xi'$-DP algorithm $\mathcal{A}'$ (as defined in Theorem 1) that can draw a sample $S \subseteq D$ of size $s$,*

$$s = \Omega\Big( \left(\frac{M}{\epsilon}\right)^2 \left(k \ln n + \ln \frac{1}{\hat{\delta}}\right) \Big)$$

*and obtain a $k$-median clustering $\hat{c}_S$ such that with probability at least $1 - \hat{\delta}$, $\mathbb{E}_D[\hat{c}_S] \leq 6\, \mathbb{E}_D[c_D] + O(Mk^2 \log^2(n/\xi)) + \epsilon$.*

## 3.4 Private $k$-median in Euclidean Space

In this setting, we consider input set $D \subseteq \mathbb{R}^d$ and $|D| = n$. The family of $k$-median cost functions is now $F = \{f_{c_1,\ldots,c_k} : c_1, \ldots, c_k \in \mathbb{R}^d\}$. Note that the number of possible clusterings of $D$ is now uncountably infinite. We want to apply techniques similar to the metric space setting, but in order to estimate $|F|$, we adapt the approach of [27] and use $\epsilon$-nets.

An $\epsilon$-*net* $F_\epsilon$ for $F$ is a family of functions such that for each $f \in F$ and for any set $D$ of points, there exists an $f_\epsilon \in F_\epsilon$ such that $|\mathbb{E}_D[f - f_\epsilon]| \leq \epsilon$. In order to obtain an $\epsilon$-net for $F$, consider the subset of $k$-median cost functions that correspond to centers at evenly-spaced gridpoints. Formally, $\langle x_1, \ldots, x_d \rangle$ is a $d$-dimensional $\eta$-gridpoint if for each $x_i$, $x_i$ is a multiple of $\eta$. [27] showed that if the gridpoints are spaced close enough, then we obtain an $\epsilon$-net for the family of $k$-median cost functions.

**Lemma 3** ($k$-median has a small $\epsilon$-net [27]). *Let $X$ be a set of $n$ points in $\mathbb{R}^d$ of bounded diameter $M$. For $x \in X$, let $F = \{f_{c_1,\ldots,c_k} : f_{c_1,\ldots,c_k}(x) = \min_i d(x, c_i)\}$, and let $F_\epsilon = \{f_{c_1,\ldots,c_k} : f_{c_1,\ldots,c_k}(x) = \min_i d(x, c_i)$ where $c_i$ is a $d$-dimensional*

*$\frac{\epsilon}{2d}$-gridpoint in the bounded diameter space$\}$. For each $f \in F$, there exists $f_\epsilon \in F_\epsilon$ such that $|\mathbb{E}_X[f - f_\epsilon]| \leq \epsilon$.*

Note that $|F_\epsilon| = O\left(\left(\frac{dM}{\epsilon}\right)^{dk}\right)$, thus we can apply Lemma 1 to $F_\epsilon$ and set $D$ to show that the sample cost for each function in $F_\epsilon$ approaches the true cost provided the size of the sample is chosen appropriately.

**Lemma 4.** *For $D \subseteq \mathbb{R}^d$, assuming an $(\alpha, \gamma)$-approximation $k$-median algorithm that runs in time $T(n)$, where $\alpha \geq 0.75$ is a constant, we can draw a sample $S$ of size $s$ where*

$$s = \Omega\Big( \left(\frac{M\alpha}{\epsilon}\right)^2 \left(dk \ln \frac{\alpha d M}{\epsilon} + \ln \frac{1}{\delta}\right) \Big) \ ,$$

*and obtain a $k$-median clustering $\hat{c}_S$ in time $T(|S|)$ such that with probability at least $1-\delta$, $\mathbb{E}_D(\hat{c}_S) \leq \alpha\, \mathbb{E}_D(c_D) + \gamma + \epsilon$, where $c_D$ is the optimum $k$-median clustering of $D$ in $\mathbb{R}^d$.*

*Proof.* The main idea is to show that the sample approximate cost of $\hat{c}_S$ and the true cost of $c_D$ are close. First, recall that $c_S, c_D$ are the optimum clusterings of sample $S$ and set $D$ respectively, and $\hat{c}_S$ is the approximate clustering that is obtained by running $\mathcal{A}(S)$. We consider the class $F_{\epsilon/6\alpha}$ as an $\epsilon$-net for the family of $k$-median cost functions $F$. We define $\hat{c}_{S,\epsilon} :=$ closest gridpoint function to $\hat{c}_S$, and $c_{D,\epsilon} :=$ closest gridpoint function to $c_D$. Then by Lemma 3, $|\mathbb{E}_D[c_{D,\epsilon}] - \mathbb{E}_D[c_D]| \leq \epsilon/6\alpha$ (8). Since $|F_{\epsilon/6\alpha}| = \left(\frac{6dM\alpha}{\epsilon}\right)^{dk}$, and the sample size $s$ is appropriately defined in the lemma statement, we can apply Lemma 1 on the class of functions $F_{\epsilon/6\alpha}$ to obtain that with probability at least $1-\delta/2$, $|\mathbb{E}_S[c_{D,\epsilon}] - \mathbb{E}_D[c_{D,\epsilon}]| \leq \epsilon/6\alpha$ (9). Adding (8) and (9), by triangle inequality, we get that with probability at least $1-\delta/2$, $|\mathbb{E}_S[c_{D,\epsilon}] - \mathbb{E}_D[c_D]| \leq \epsilon/3\alpha$ (10).

Since $c_S$ is the optimum clustering of $S$, its average cost over $S$ is better than the average cost given by the clustering $c_{D,\epsilon}$, hence $\mathbb{E}_S[c_S] \leq \mathbb{E}_S[c_{D,\epsilon}]$. Adding (10) to this, with probability at least $1-\delta/2$, $\mathbb{E}_S[c_S] - \mathbb{E}_D[c_D] \leq \epsilon/3\alpha$ (11).

Since we ran an $(\alpha, \gamma)$-approximation on the sample $S$, we have $\mathbb{E}_S[\hat{c}_S] \leq \alpha\, \mathbb{E}_S[c_S] + \gamma$. Combining this with (11), we have that with probability at least $1-\delta/2$, $\mathbb{E}_S[\hat{c}_S] - \gamma - \alpha\, \mathbb{E}_D[c_D] \leq \epsilon/3$ (12). By Lemma 3, the cost of a clustering and its closest gridpoint function is no more than $\epsilon/6\alpha$, thus $|\mathbb{E}_S[\hat{c}_S] - \mathbb{E}_S[\hat{c}_{S,\epsilon}]| \leq \epsilon/6\alpha$. Combining this with (12), we have that with probability at least $1-\delta/2$, $\mathbb{E}_S[\hat{c}_{S,\epsilon}] - \gamma - \alpha\, \mathbb{E}_D[c_D] \leq \epsilon/3 + \epsilon/6\alpha$ (13).

By Lemma 1, with probability at least $1-\delta/2$, $|\mathbb{E}_S[\hat{c}_{S,\epsilon}] - \mathbb{E}_D[\hat{c}_{S,\epsilon}]| \leq \epsilon/6\alpha$. Combining this fact with (13), we get with probability at least $1-\delta$, $\mathbb{E}_D[\hat{c}_{S,\epsilon}] - \alpha\, \mathbb{E}_D[c_D] - \gamma \leq \epsilon/3 + \epsilon/3\alpha$ (14). Lastly, by Lemma 3, $|\mathbb{E}_D[\hat{c}_S] - \mathbb{E}_D[\hat{c}_{S,\epsilon}]| \leq \epsilon/6\alpha$, and by adding this to (14) and rearranging, we get that with probability at least $1-\delta$, $\mathbb{E}_D[\hat{c}_S] - \alpha\, \mathbb{E}_D[c_D] - \gamma \leq \epsilon/3 + \epsilon/2\alpha \leq \epsilon$, where the last step is true for $\alpha \geq 0.75$. □

We will now state the private $(\alpha, \gamma)$-approximation clustering algorithm that we will use as a black-box with the lemma we proved above. Our sublinear-time private $k$-median result is subsequently obtained as a corollary.

Given any $w$-approximation algorithm for $k$-median (respectively $k$-means), Ghazi et al. [16] use differentially-private coresets to give pure and approximate differentially-private algorithms that run in polynomial time and achieve approximation guarantees very close to that of the original algorithm.

**Theorem 4.** *[16] Assume there is a polynomial-time (not necessarily DP) algorithm for k-median (respectively k-means) in $\mathbb{R}^d$ with approximation ratio $w$. Then there is an $\xi$-DP algorithm that runs in time $k^{O_\alpha(1)}$poly($nd$) and with probability 0.99, produces a $\Big(w(1+\alpha), O_{w,\alpha}\Big(\Big(\frac{kd+k^{O_\alpha(1)}}{\xi}\Big)$ poly log $n\Big)\Big)$-approximation for k-median (respectively k-means).*

*Moreover, there is an $(\xi, \delta)$-DP algorithm with the same runtime and approximation ratio but with additive error $O_{w,\alpha}\Big(\Big(\frac{k\sqrt{d}}{\xi} \cdot$ poly log $\big(\frac{k}{\delta}\big)\Big) + \Big(\frac{k^{O_\alpha(1)}}{\xi} \cdot$ poly logn$\Big)\Big)$.*

Note that the state-of-the-art non-private algorithm for $k$-median achieves an approximation ratio of $w = 2.633$ [1]. We use the algorithm from [16] as our black-box algorithm $\mathcal{A}$, and by plugging in the approximation guarantees of $\mathcal{A}$ as stated in Theorem 4 with Lemma 4, we obtain the following accuracy guarantees for our sampling algorithm $\mathcal{A}'$ in the pure differential privacy as well as the approximate differential privacy settings.

**Theorem 5** (Accuracy of $\mathcal{A}'$ for pure and approximate DP)**.** *For private set $D \subseteq \mathbb{R}^d$, and an $\xi$-DP $\Big(w(1+\alpha), O\Big(\Big(\frac{kd+k^{O(1)}}{\xi}\Big)$poly log $n\Big)\Big)$-approximation k-median algorithm (from [16]), that runs in time $k^{O(1)}$poly($nd$), for $w(1+\alpha) \geq 0.75$, we have a $\xi'$-DP algorithm $\mathcal{A}'$ that can draw a sample $S \subseteq D$ of size s,*

$$s = \Omega\Big(\Big(\frac{Mw(1+\alpha)}{\epsilon}\Big)^2\Big(dk \ln \frac{w(1+\alpha)dM}{\epsilon} + \ln \frac{1}{\hat{\delta}}\Big)\Big)$$

*and obtain a k-median clustering $\hat{c}_S$ in time $k^{O(1)}$poly($sd$) such that with probability at least $1 - \hat{\delta}$, $\mathbb{E}_D[\hat{c}_S] \leq w(1+\alpha)\,\mathbb{E}_D[c_D] + O\Big(\Big(\frac{kd+k^{O(1)}}{\xi}\Big)$poly log $n\Big) + \epsilon$.*

*Moreover, by using the $(\xi, \delta)$-DP algorithm from [16] with the same runtime and approximation ratio but with additive error $\gamma' := O\Big(\Big(\frac{k\sqrt{d}}{\xi} \cdot$ poly log$\Big(\frac{k}{\delta}\Big)\Big) + \Big(\frac{k^{O(1)}}{\xi} \cdot$ poly logn$\Big)\Big)$, we obtain a $(\xi', \delta')$-DP algorithm $\mathcal{A}'$ that draws a sample of the same size, and obtains a k-median clustering such that with probability at least $1 - \hat{\delta}$, $\mathbb{E}_D[\hat{c}_S] \leq w(1+\alpha)\,\mathbb{E}_D[c_D] + \gamma' + \epsilon$. Privacy parameters $\xi', \delta'$ are as defined in Theorem 1.*

Note that the state-of-the-art non-private algorithm for $k$-median achieves an approximation ratio of $w = 2.633$ [1].

## 3.5 Private $k$-means clustering in metric space

We follow the techniques of Czumaj et al. [10] and extend their sublinear $k$-means clustering analysis to work for black-box polynomial-time $k$-means algorithms that have an additive factor of $\gamma > 0$, then we combine this extension with the existing private $k$-means clustering algorithm [16] to obtain a private sublinear-time $k$-means clustering algorithm.

Let $(V, d)$ be a metric space consisting of $n$ points with diameter $M$, and let $D \subseteq V$ be an input set. Suppose $S \subseteq D$ is randomly sampled and let $\hat{c}_S$ be the clustering obtained by running the $(\alpha, \gamma)$-approximation $k$-means algorithm $\mathcal{A}(S)$. We extend the Czumaj et al. analysis in the following manner. First, we show that with high probability and chosen sample size, the average cost of approximate clustering $\hat{c}_S$ of the sample $S$ is close to the average cost of the optimum clustering $c_D$ of the input set $D$, i.e. $\alpha \cdot \mathbb{E}_S[\hat{c}_S] + \gamma \leq (\alpha + \beta)\,\mathbb{E}_D[c_D] + \gamma$ . Next, we show that if clustering $c_b$ is a $(\alpha + 3\beta, \gamma)$-"bad" solution of input set $D$, i.e., $\mathbb{E}_D[c_b] > (\alpha + 3\beta)\,\mathbb{E}_D[c_D] + \gamma$, then, for an appropriate sample size, with high probability the clustering $c_b$ is also a "bad" solution for the sample set $S$, in other words, $\mathbb{E}_S[c_b] > (\alpha+2\beta)\,\mathbb{E}_D[c_D]+\gamma$, where $c_D$ is the optimum clustering of $D$. From these two statements, we can conclude that with high probability the approximate clustering $\hat{c}_S$ must be a "good"-solution for the input set $D$, in other words, $\mathbb{E}_D[\hat{c}_S] \leq (\alpha+2\beta)\,\mathbb{E}_D[c_D]+\gamma$. By carefully choosing the parameter $\beta$ as a function of $\epsilon > 0$ and $\mathbb{E}_D[c_D]$, we can remove it from the sample complexity and approximation guarantees. We omit the formal proof here, as it just follows the Czumaj et al. analysis but also keeps track of the additive factor $\gamma$, and ensures that we are considering a feasible clustering to be any subset of the metric space, see Appendix A for a full proof. We present the final lemma below.

**Lemma 5.** *For $D \subseteq \mathbb{R}^d$, assuming an $(\alpha,\gamma)$-approximation $k$-means algorithm that runs in time $T(n)$, we can draw a sample $S$ of size $s$,*

$$s \geq c \cdot \max\left\{\frac{M^2\alpha(1+\alpha)\ln(1/\delta)}{\epsilon}, \frac{M^4}{\epsilon^2}\cdot(\ln(1/\delta)+k\ln n)\right\}\ ,$$

*where $c$ is a positive constant, and obtain a $k$-means clustering $\hat{c}_S$ in time $T(|S|)$ such that with probability at least $1-\delta$, $\mathbb{E}_D[\hat{c}_S] \leq \alpha\,\mathbb{E}_D[c_D]+\gamma+\epsilon$, where $c_D$ is the optimum $k$-means clustering of $D$.*

Following the techniques of [17], [4] extended their results to the private $k$-means setting by adapting their analysis to the non-private local search approximation algorithm for $k$-means clustering [21].

**Theorem 6.** *[4] Given a metric space $(V,d)$ of $n$ points with diameter $M$, a set $D \subseteq V$, there exists a $\xi$-differentially private $k$-means algorithm that with probability at least 0.99 produces a $\left(30, O\left((M^2k^4/\xi)\cdot\log^2 n\right)\right)$-approximation for $k$-means clustering.*

We use the algorithm from [4] as our private black-box $k$-means clustering algorithm $\mathcal{A}$. By plugging in the approximation guarantees of $\mathcal{A}$ into our Lemma 5, we obtain the following accuracy guarantee for our sublinear sampling algorithm $\mathcal{A}'$.

**Theorem 7** (Accuracy of $\mathcal{A}'$)**.** *For private set $D \subseteq V$, and an $\xi$-DP $\left(30, O\left((M^2k^4/\xi)\cdot\log^2 n\right)\right)$-approximation $k$-means algorithm (from [4]), that runs in time $T(n)$, we have a $\xi'$-DP algorithm $\mathcal{A}'$ (as defined in Theorem 1) that can draw a sample $S \subseteq D$ of size $s$,*

$$s \geq c \cdot \max\left\{\frac{M^2\alpha(1+\alpha)\ln(1/\hat{\delta})}{\epsilon}, \frac{M^4}{\epsilon^2}\cdot\left(\ln(1/\hat{\delta})+k\ln n\right)\right\}\ ,$$

*where $c$ is a positive constant, and obtain a $k$-means clustering $\hat{c}_S$ in time $T(s)$ such that with probability at least $1-\hat{\delta}$, $\mathbb{E}_D[\hat{c}_S] \leq 30\,\mathbb{E}_D[c_D] + O\left((M^2k^4/\xi)\cdot\log^2 n\right)+\epsilon$.*

### 3.6 Private $k$-means clustering in Euclidean space

The extension of the $k$-means analysis to Euclidean space involves the same steps as outlined in Subsection 3.5, but similar to the analysis for $k$-median in Euclidean space, we need to consider $\epsilon$-nets to estimate the size of possible clusterings (see Subsection 3.4). The full proof of the statement below can be found in Appendix B.

**Lemma 6.** *For $D \subseteq \mathbb{R}^d$, assuming an $(\alpha, \gamma)$-approximation $k$-means algorithm that runs in time $T(n)$, we can draw a sample $S$ of size $s$,*

$$s \geq c \cdot \max \left\{ \frac{M^2 \alpha(1+\alpha)\ln(1/\delta)}{\epsilon}, \frac{M^4}{\epsilon^2} \cdot \left( \ln(1/\delta) + kd \ln\left(\frac{\sqrt{d}M}{2\epsilon}\right)\right)\right\} \ ,$$

*where $c$ is a positive constant, and obtain a $k$-means clustering $\hat{c}_S$ in time $T(|S|)$ such that with probability at least $1-\delta$, $\mathbb{E}_D[\hat{c}_S] \leq \alpha\, \mathbb{E}_D[c_D] + \gamma + \epsilon$, where $c_D$ is the optimum $k$-means clustering of $D$ in $\mathbb{R}^d$.*

Note that the state-of-the-art non-private algorithm for $k$-means achieves an approximation ratio of $w = 6.358$ [1]. We use the private $k$-means algorithm by Ghazi et al. [16] (See Theorem 4) as our black-box private $k$-means clustering algorithm $\mathcal{A}$. By plugging in the approximation guarantees for $\mathcal{A}$ to Lemma 6 we obtain the following accuracy guarantees for the sampling algorithm $\mathcal{A}'$ in both the pure approximate differential privacy setting.

**Theorem 8** (Accuracy of $\mathcal{A}'$ for pure and approximate DP)**.** *For private set $D \subseteq \mathbb{R}^d$, and an $\xi$-DP $\left(w(1+\alpha), O\left(\left(\frac{kd+k^{O(1)}}{\xi}\right) poly\ log\ n\right)\right)$-approximation $k$-means algorithm (from [16]), that runs in time $k^{O(1)}poly(nd)$, for $w(1+\alpha) \geq 0.75$, we have a $\xi'$-DP algorithm $\mathcal{A}'$ that can draw a sample $S \subseteq D$ of size $s$,*

$$s \geq c \cdot \max\left\{ \frac{M^2 \alpha(1+\alpha)\ln(1/\delta')}{\epsilon}, \frac{M^4}{\epsilon^2} \cdot \left( \ln(1/\delta) + kd \ln\left(\frac{\sqrt{d}M}{2\epsilon}\right)\right)\right\} \ ,$$

*and obtain a $k$-means clustering $\hat{c}_S$ in time $k^{O(1)}poly(sd)$ such that with probability at least $1-\hat{\delta}$, $\mathbb{E}_D[\hat{c}_S] \leq w(1+\alpha)\, \mathbb{E}_D[c_D] + O\left(\left(\frac{kd+k^{O(1)}}{\xi}\right) poly\ log\ n\right) + \epsilon$.*

*Moreover, by using the $(\xi, \delta)$-DP algorithm from [16] with the same runtime and approximation ratio but with additive error $\gamma' := O\left(\left(\frac{k\sqrt{d}}{\xi} \cdot poly\ log\left(\frac{k}{\delta}\right)\right) + \left(\frac{k^{O(1)}}{\xi} \cdot poly\ logn\right)\right)$, we obtain a $(\xi', \delta')$-DP algorithm $\mathcal{A}'$ that draws a sample of the same size, and obtains a $k$-means clustering such that with probability at least $1-\hat{\delta}$, $\mathbb{E}_D[\hat{c}_S] \leq w(1+\alpha)\, \mathbb{E}_D[c_D] + \gamma' + \epsilon$. Privacy parameters $\xi', \delta'$ are as defined in Theorem 1.*

# 4 Group Privacy in Sublinear setting

In this section, we give a group privacy result that holds for *any* sampling algorithm $\mathcal{A}'(D)$ that samples a set $S$ from the input set $D$ by independently sampling with probability $\xi_1$ and runs an $\xi$-DP algorithm $\mathcal{A}$ on $S$. Let $D'$ be a set that differs on $g$ elements with respect to $D$, and $0 \leq T \leq g$ be a threshold. Define $\delta_{T,\xi_1,g} := 1 - \sum_{j=0}^{T} \binom{g}{j} \xi_1^j (1-\xi_1)^{g-j}$ , in other words, $\delta_{T,\xi_1,g}$ is the probability of choosing more than $T$ elements that differ from elements in $D'$ in the sample $S$.

Given that $\mathcal{A}$ is $\xi$-DP, we have already shown that $\mathcal{A}'$ is $\xi'$-DP (see Theorem 1). In the following theorem, we show that $\mathcal{A}'$ also gives us better group privacy guarantees.

**Theorem 9.** *If $\mathcal{A}'$ is an $\xi'$-DP sampling algorithm (as described above) then it gives $(T \cdot \xi', \delta_{T,\xi_1,g})$-privacy for groups of size $g$, where $\delta_{T,\xi_1,g} := 1 - \sum_{j=0}^{T} \binom{g}{j} \xi_1^j (1-\xi_1)^{g-j}$.*

*Proof.* Consider two sets $D$ and $D'$ that differ on $g$ elements, i.e., $|D| = |D'| + g$ and set $S \subseteq D$ sampled independently w.p. $\xi_1$. Define the random variable $Y$ to be the number of elements in $S$ sampled from the $g$ differing elements. Fix an output set $C$ in the output space of $\mathcal{A}'$. Then

$$\begin{aligned}
&\Pr[\mathcal{A}'(D) \in C] \\
&= \sum_{i=0}^{g} \Pr[\mathcal{A}'(D) \in C, Y = i] \\
&= \sum_{i=0}^{g} \Pr[\mathcal{A}'(D) \in C | Y = i] \Pr[Y = i] \\
&= \sum_{i=0}^{T} \Pr[\mathcal{A}'(D) \in C | Y = i] \Pr[Y = i] + + \sum_{i=T+1}^{g} \Pr[\mathcal{A}'(D) \in C | Y = i] \Pr[Y = i]
\end{aligned}$$

Applying the naive group privacy bound for each term $\Pr[\mathcal{A}'(D) \in C | Y = i]$ in the first sum,

$$\leq \sum_{i=0}^{T} e^{\xi \cdot i} \Pr[\mathcal{A}'(D') \in C] \Pr[Y = i] + \sum_{i=T+1}^{g} \Pr[\mathcal{A}'(D) \in C | Y = i] \Pr[Y = i]$$

Observe that $\sum_{i=T+1}^{g} \Pr[\mathcal{A}'(D) \in C | Y = i] \Pr[Y = i] \leq \sum_{i=T+1}^{g} \Pr[Y = i] \leq \delta_{T,\xi_1,g}$, therefore,

$$\Pr[\mathcal{A}'(D) \in C] \leq e^{\xi \cdot T} \Pr[\mathcal{A}'(D') \in C] + \delta_{T,\xi_1,g} \ .$$

□

We demonstrate how in many instances, our sampling algorithm $\mathcal{A}'$ achieves better group privacy guarantees for chosen $\xi_1$ and $T$ such that $T \ll g$. (1) If we sample each element of the input set with probability $\xi_1 = 1/\sqrt{g}$, and set threshold $T = 2\sqrt{g}$, then $\mathcal{A}'$ is $(2\sqrt{g}\xi', \delta_{T,\xi_1,g})$ for $\delta_{T,\xi_1,g}$ negligible in $g$. (2) If we sample each element of the input set with probability $\xi_1 = 1/\log g$, and set threshold $T = 2g/\log g$, then $\mathcal{A}'$ is $((2g/\log g)\xi', \delta_{T,\xi_1,g})$ for $\delta_{T,\xi_1,g}$ negligible in $g$.

## 5 Acknowledgements

Elena would like to thank Marek Elias, Michael Kapralov and Aida Mousavifar for initial discussions on this topic while she was visiting EPFL. She also thanks her EPFL hosts for their hospitality.

# References

[1] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *SIAM Journal on Computing*, 49, 2020.

[2] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristic for k-median and facility location problems. STOC, 2001.

[3] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. FOCS, 2010.

[4] Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, and Hongyang Zhang. Differentially private clustering in high-dimensional Euclidean spaces. ICML, 2017.

[5] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. NeurIPS, 2018.

[6] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. SOSP, 2017.

[7] Moses Charikar, Sudipto Guha, Éva Tardos, and David B Shmoys. A constant-factor approximation algorithm for the k-median problem. *Journal of Computer and System Sciences*, 65, 2002.

[8] Ke Chen. On k-median clustering in high dimensions. SODA, 2006.

[9] Ke Chen. A constant factor approximation algorithm for k-median clustering with outliers. SODA, 2008.

[10] Artur Czumaj and Christian Sohler. Sublinear-time approximation algorithms for clustering via random sampling. volume 30, 2007.

[11] Apple Differential Privacy Team. Learning with privacy at scale, 2017.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. volume 7, 2016.

[13] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. CCS, 2014.

[14] Dan Feldman, Amos Fiat, Haim Kaplan, and Kobbi Nissim. Private coresets. STOC, 2009.

[15] Dan Feldman, Chongyuan Xiang, Ruihao Zhu, and Daniela Rus. Coresets for differentially private k-means clustering and applications to privacy in mobile sensor networks. IPSN, 2017.

[16] Badih Ghazi, R. Kumar, and Pasin Manurangsi. Differentially private clustering: Tight approximation ratios. NeurIPS, 2020.

[17] Anupam Gupta, Katrina Ligett, Frank McSherry, Aaron Roth, and Kunal Talwar. Differentially private combinatorial optimization. 2010.

[18] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300, 2004.

[19] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100, 1992.

[20] Zhiyi Huang and Jinyan Liu. Optimal differentially private algorithms for k-means clustering. PODS, 2018.

[21] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 2004.

[22] Haim Kaplan and Uri Stemmer. Differentially private k-means with constant multiplicative error. NeurIPS, 2018.

[23] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? *SIAM Journal on Computing*, 40, 2011.

[24] Daniel Kifer, Solomon Messing, Aaron Roth, Abhradeep Thakurta, and Danfeng Zhang. Guidelines for implementing and auditing differentially private systems. *CoRR*, abs/2002.04049, 2020.

[25] Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. *SIAM Journal on Computing*, 45, 2016.

[26] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. FOCS, 2007.

[27] Nina Mishra, Dan Oblinger, and Leonard Pitt. Sublinear time approximate clustering. SODA, 2001.

[28] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. STOC, 2007.

[29] Kobbi Nissim and Uri Stemmer. Clustering algorithms for the centralized and local models. ALT, 2018.

[30] Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59, 2012.

[31] David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

[32] Moshe Shechner, Or Sheffet, and Uri Stemmer. Private k-means clustering with stability assumptions. AISTATS, 2020.

[33] Adam Smith. Differential privacy and the secrecy of the sample. `https://adamdsmith.wordpress.com/2009/09/02/sample-secrecy/`, Sep 2009.

[34] Uri Stemmer. Locally private k-means clustering. SODA, 2020.

[35] Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, and Hongxia Jin. Differentially private k-means clustering. CODASPY, 2016.

# A Sublinear $k$-means in metric space (with additive error)

In this section we give a detailed proof of Lemma 5. Our proof is nearly identical to that of [10], except that we consider continuous clusterings in metric space. For ease of representation and comparison, we also adopt the notation used in [10], which we recall below.

Let $(V, d)$ be a metric space and $D \subseteq V$ be the input set, and $M$ be the diameter of $V$. Let

$$\mathsf{mean}_{\mathsf{avg}}(D, k) = \frac{1}{|D|} \min_{\substack{C \subseteq V \\ |C|=k}} \sum_{x \in D} d(x, C)^2 ,$$

denote the average cost of an optimum $k$-mean clustering of $D$. Similarly, for any subset $U \subseteq D$ and $C \subseteq V$, define the average cost of a $k$-mean clustering $C$ as

$$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(U, C) = \frac{1}{|U|} \sum_{v \in U} d(v, C)^2 .$$

The analysis from [10] involves two main steps.

1. If $\mathcal{A}(S)$ outputs clustering $C^*$, then we need to show that for a chosen sample size, with high probability ,
$$\alpha \cdot \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C^*) + \gamma \leq (\alpha + \beta)\mathsf{mean}_{\mathsf{avg}}(D, k) + \gamma .$$

2. If clustering $C_b \subseteq \mathbb{R}^d$ is a $(\alpha + 3\beta, \gamma)$-bad solution of input set $D$, i.e., $\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b) > (\alpha + 3\beta)\mathsf{med}_{\mathsf{avg}}(D, k) + \gamma$, then, we need to show that with high probability the clustering $C_b$ is also a bad solution for the sample set $S$,
$$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) > (\alpha + 2\beta)\mathsf{med}_{\mathsf{avg}}(D, k) + \gamma .$$

By combining the above two statements we would get that with high probability the clustering $C^*$ (outputted by $\mathcal{A}(S)$) is an $(\alpha{+}3\beta, \gamma)$-good solution of input set $D$, in other words, $\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C^*) \leq (\alpha + 3\beta)\mathsf{med}_{\mathsf{avg}}(D, k) + \gamma$.

**Lemma 7.** *Let $S$ be a multiset of size $s$ chosen from $D$ i.u.r. For*

$$s \geq \frac{3M^2\alpha(\beta + \alpha)\ln(1/\delta)}{2\beta^2\mathsf{mean}_{\mathsf{avg}}(D, k)} .$$

*If an $(\alpha, \gamma)$-approximation algorithm for k-means $\mathcal{A}$ is run on input $S$, then the following holds for the solution $C^*$ returned by $\mathcal{A}$:*

$$\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C^*) \leq (\alpha + \beta)\mathsf{mean}_{\mathsf{avg}}(D, k) + \gamma] \geq 1 - \delta .$$

*Proof.* Let $C_{OPT}$ denote an optimal $k$-means solution for input set $D$. For $1 \leq i \leq s$, define random variables $X_i$ as the distance of the $i$-th point in $S$ to the nearest center of $C_{OPT}$. Then $\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_{OPT}) = \frac{1}{s}\sum_{1\leq i\leq s} X_i$. Observe that, $\mathbb{E}[X_i] = \sum_{x\in D} \Pr[x \overset{\$}{\leftarrow} D] \cdot d(x, C_{OPT})^2 = \frac{1}{|D|}\sum_{x\in D} d(x, C_{OPT})^2 = \mathsf{mean}_{\mathsf{avg}}(D,k)$ , also $\mathsf{mean}_{\mathsf{avg}}(D,k) = \frac{1}{s}\,\mathbb{E}[\sum_{1\leq i\leq s} X_i]$.

$$\Pr\left[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_{OPT}) > \left(1+\frac{\beta}{\alpha}\right)\mathsf{mean}_{\mathsf{avg}}(D,k)\right] = \Pr\left[\sum_{1\leq i\leq s} X_i > \left(1+\frac{\beta}{\alpha}\right)\mathbb{E}[\sum_{1\leq i\leq s} X_i]\right]$$

Each $0 \leq X_i \leq M^2$. Thus we can apply a Hoeffding bound,

$$\Pr\left[\sum_{1\leq i\leq s} X_i > \left(1+\frac{\beta}{\alpha}\right)\mathbb{E}[\sum_{1\leq i\leq s} X_i]\right] \quad \leq \exp\left(-\frac{s}{3M}\cdot \mathsf{mean}_{\mathsf{avg}}(D,k)\min\{(\beta/\alpha),(\beta/\alpha)^2\}\right)$$

Choosing $s$ as in the lemma statement, the probability above is bounded by $\delta$. Since $\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C^*) \leq \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_{OPT})$, and $\mathcal{A}$ is $(\alpha,\gamma)$-approximation, we have that with probability $1-\delta$,

$$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C^*) \leq (\alpha+\beta)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma$$

□

Next, we need to show that any clustering $C_b$ that is an $(\alpha+3\beta,\gamma)$-bad solution of $k$-means of $D$ satisfies with high probability $\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) > (\alpha+2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k)+\gamma$

**Lemma 8.** *Let $S$ be a multiset of $s$ points chosen i.u.r. from $D$ such that*

$$s \geq \frac{2M^4}{\beta^2(\mathsf{mean}_{\mathsf{avg}}(D,k))^2}\cdot(\ln(1/\delta) + k\ln n)$$

*Let $\mathbb{C}$ be the set of $(\alpha+3\beta,\gamma)$-bad solutions of a $k$-mean clustering of $D$. Then*

$$\Pr[\exists C_b \in \mathbb{C} : \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) \leq (\alpha+2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k)+\gamma] \leq \delta$$

*Proof.* Consider an arbitrary $C_b \in \mathbb{C}$, and define $X_i$ as the distance of the $i$th point in $S$ from the nearest center in $C_b$. Since $C_b$ is a $(\alpha+3\beta,\gamma)$-bad solutions of a $k$-means of $D$, by definition,

$$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b) > (\alpha+3\beta)\mathsf{mean}_{\mathsf{avg}}(D,k)+\gamma \tag{15}$$

Now for $1 \leq i \leq s$, we have that $\mathbb{E}[X_i] = \frac{1}{|D|}\sum_{x\in D}(d(x,C_b))^2 = \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b)$, thus

$$\mathbb{E}[X_i] > (\alpha+3\beta)\mathsf{mean}_{\mathsf{avg}}(D,k)+\gamma \tag{16}$$

Also,

$$\sum_{1\leq i\leq s} X_i = \sum_{x\in S} d(x, C_b) = s\cdot \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b)\ , \tag{17}$$

and $\mathbb{E}[\sum_{1\leq i\leq s} X_i] = s\,\mathbb{E}[X_i]$ for any $i$, recall that $\mathbb{E}[X_i] = \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b)$ and hence technically independent of $i$.

We want to show that for any $C_b \in \mathbb{C}$, $\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) \leq (\alpha+2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k)+\gamma]$ is low, and then take a union bound over the entire space of $\mathbb{C}$.

$$\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) \leq (\alpha+2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k)+\gamma]$$

Substituting Relation 17 on LHS and Relation 16 on RHS,

$$= \Pr\left[\frac{1}{s}\cdot\sum_{1\leq i\leq s} X_i \leq \frac{(\alpha+2\beta)}{(\alpha+3\beta)}\,\mathbb{E}[X_i] + \gamma\cdot\left(1-\frac{(\alpha+2\beta)}{(\alpha+3\beta)}\right)\right]$$

$$= \Pr\left[\sum_{1\leq i\leq s} X_i \leq \frac{(\alpha+2\beta)}{(\alpha+3\beta)}\cdot s\cdot\mathbb{E}[X_i] + \frac{s\gamma\beta}{\alpha+3\beta}\right]$$

$$= \Pr\left[\sum_{1\leq i\leq s} X_i \leq \frac{(\alpha+2\beta)}{(\alpha+3\beta)}\cdot\mathbb{E}[\sum_{1\leq i\leq s} X_i] + \frac{s\gamma\beta}{\alpha+3\beta}\right]$$

$$= \Pr\left[\sum_{1\leq i\leq s} X_i \leq \left(\frac{(\alpha+2\beta)}{(\alpha+3\beta)} + \frac{s\gamma\beta}{(\alpha+3\beta)\,\mathbb{E}[\sum_{1\leq i\leq s} X_i]}\right)\cdot\mathbb{E}[\sum_{1\leq i\leq s} X_i]\right]$$

$$= \Pr\left[\sum_{1\leq i\leq s} X_i \leq \left(1-\left(\frac{\beta}{(\alpha+3\beta)} - \frac{s\gamma\beta}{(\alpha+3\beta)\cdot s\cdot\mathsf{cost}_{\mathsf{avg}}(D,C_b)}\right)\right)\cdot\mathbb{E}[\sum_{1\leq i\leq s} X_i]\right]$$

Since $0 \leq X_i \leq M^2$, we can apply a Hoeffding bound,

$$\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) \leq (\alpha+2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k)+\gamma]$$

$$\leq \exp\left(-\frac{\mathbb{E}[\sum_{1\leq i\leq s} X_i]}{2M^2}\cdot\left(\frac{\beta}{(\alpha+3\beta)} - \frac{\gamma\beta}{(\alpha+3\beta)\cdot\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C_b)}\right)^2\right)$$

$$= \exp\left(-\frac{s\cdot\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C_b)}{2M^2}\cdot\left(\frac{\beta\cdot\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C_b)-\gamma\beta}{(\alpha+3\beta)\cdot\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C_b)}\right)^2\right)$$

$$= \exp\left(-\frac{s\beta^2}{2M^2\cdot(\alpha+3\beta)^2}\cdot\frac{(\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C_b)-\gamma)^2}{\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C_b)}\right)$$

$$\leq \exp\left(-\frac{s\beta^2}{2M^2\cdot(\alpha+3\beta)^2}\cdot\frac{(\mathsf{mean}_{\mathsf{avg}}(D,k))^2(\alpha+3\beta)^2}{\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C_b)}\right), \qquad \text{Applying relation 15}$$

Now, $\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C_b) \leq M^2$, therefore

$$\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) \leq (\alpha+2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k)+\gamma] \leq \exp\left(-\frac{s\beta^2}{2M^4}\cdot(\mathsf{mean}_{\mathsf{avg}}(D,k))^2\right)$$

By union bound and using the fact that $|\mathbb{C}| \le n^k$,

$$\begin{aligned}&\Pr[\exists C_b \in \mathbb{C} : \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) \le (\alpha + 2\beta)\mathsf{mean}_{\mathsf{avg}}(D, k) + \gamma] \\ &\le n^k \cdot \exp\left(-\frac{s\beta^2}{2M^4} \cdot (\mathsf{mean}_{\mathsf{avg}}(D, k))^2\right)\end{aligned}$$

We choose

$$s \ge \frac{2M^4}{\beta^2(\mathsf{mean}_{\mathsf{avg}}(D, k))^2} \cdot (\ln(1/\delta) + k \ln n)$$

□

Recall Lemma 5, which states that for $D \subseteq \mathbb{R}^d$, assuming an $(\alpha, \gamma)$-approximation $k$-means algorithm that runs in time $T(n)$, we can draw a sample $S$ of size $s$,

$$s \ge c \cdot \max\left\{\frac{M^2\alpha(1 + \alpha)\ln(1/\delta)}{\epsilon}, \frac{M^4}{\epsilon^2} \cdot (\ln(1/\delta) + k \ln n)\right\} ,$$

where $c$ is a positive constant, and obtain a $k$-means clustering $\hat{c}_S$ in time $T(|S|)$ such that with probability at least $1 - \delta$, $\mathbb{E}_D[\hat{c}_S] \le \alpha\, \mathbb{E}_D[c_D] + \gamma + \epsilon$, where $c_D$ is the optimum $k$-means clustering of $D$. The proof of this statement is below.

*Proof.* Let $\beta^*$ be a positive parameter that will be fixed later. Let $s$ be chosen such that sample complexity prerequisites of both Lemma 7 and Lemma 8 are satisfied. Recall from Lemma 9, the sample complexity is as follows,

$$s \ge \frac{3M^2\alpha(\beta^* + \alpha)\ln(1/\delta)}{2{\beta^*}^2\mathsf{mean}_{\mathsf{avg}}(D, k)} , \tag{18}$$

And from Lemma 9 we have

$$s \ge \frac{2M^4}{{\beta^*}^2(\mathsf{mean}_{\mathsf{avg}}(D, k))^2} \cdot (\ln(1/\delta) + k \ln n) , \tag{19}$$

Thus an appropriate sample complexity is

$$s \ge \max\left\{\frac{3M^2\alpha(\beta^* + \alpha)\ln(1/\delta)}{2{\beta^*}^2\mathsf{mean}_{\mathsf{avg}}(D, k)}, \frac{2M^4}{{\beta^*}^2(\mathsf{mean}_{\mathsf{avg}}(D, k))^2} \cdot (\ln(1/\delta) + k \ln n)\right\} \tag{20}$$

For the chosen sample complexity, we have from Lemma 10 that with probability at least $1 - \delta$, no clustering $C \subseteq \mathbb{R}^d$ that is a $(\alpha + 3\beta^*, \gamma)$-bad solution of a $k$-means of $D$ satisfies the inequality

$$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C) \le (\alpha + 2\beta^*)\mathsf{mean}_{\mathsf{avg}}(D, k) + \gamma \tag{21}$$

On the other hand, if we run algorithm $\mathcal{A}(S)$, then by Lemma 9, the resulting clustering $C^*$ with probability at least $1 - \delta$ satisfies

$$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C^*) \le (\alpha + \beta^*)\mathsf{mean}_{\mathsf{avg}}(D, k) + \gamma \tag{22}$$

Thus with probability at least $1 - 2\delta$, the clustering $C^*$ must be a $(\alpha + 3\beta^*, \gamma)$-good solution of a $k$-means of $D$, in other words,

$$\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C^*) \le (\alpha + 3\beta^*)\mathsf{mean}_{\mathsf{avg}}(D, k) + \gamma] \ge 1 - 2\delta . \tag{23}$$

To complete the proof, we must remove the dependency on $\mathsf{mean}_{\mathsf{avg}}(D, k)$ in the sample complexity.

- Case 1: $\mathsf{mean}_{\mathsf{avg}}(D,k) < \epsilon$. Choose $\beta^* = \epsilon/(3 \cdot \mathsf{mean}_{\mathsf{avg}}(D,k) \geq 1/3$, and $\beta = 1/3\beta^* < 1$ then we get that if sample complexity

  $$s \geq c \cdot \max\left\{ \frac{M^2\alpha(1+\alpha\beta)\ln(1/\delta)}{\epsilon}, \frac{M^4}{\epsilon^2} \cdot (\ln(1/\delta) + k \ln n) \right\} ,$$

  where $c$ is a certain positive constant, then with probability $1 - 2\delta$,

  $$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C^*) \leq (\alpha + 3\beta^*)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma \leq \alpha \cdot \mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma + \epsilon ,$$

  and

- Case 2: $\mathsf{mean}_{\mathsf{avg}}(D,k) \geq \epsilon$. Choose $\beta^* = \epsilon/(3 \cdot \mathsf{mean}_{\mathsf{avg}}(D,k)) \leq 1/3$. Then, we get that if sample complexity

  $$s \geq c \cdot \max\left\{ \frac{M^2\alpha(1+\alpha)\ln(1/\delta)}{\epsilon}, \frac{M^4}{\epsilon^2} \cdot (\ln(1/\delta) + k \ln n) \right\} ,$$

  where $c$ is a certain positive constant, then with probability $1 - 2\delta$,

  $$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C^*) \leq (\alpha + 3\beta^*)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma \leq \alpha \cdot \mathsf{mean}_{\mathsf{avg}}(D,k) + 3\epsilon + \gamma ,$$

□

# B Sublinear $k$-means in Euclidean space (with additive error)

In this section, we prove Lemma 6. We follow the same notation as above, and the analysis is very similar to the metric setting. In fact, the statement and proof of the first lemma is identical to the metric setting. Recall we want to show that given the appropriate sample size, with high probability, the cost of the approximate clustering of the sample is close to the cost of the optimum clustering of the input set. Let $D \subseteq \mathbb{R}^d$ be the input set, and $M$ be the diameter of $D$.

**Lemma 9.** *Let $S$ be a multiset of size $s$ chosen from $D$ i.u.r. For*

$$s \geq \frac{3M^2\alpha(\beta+\alpha)\ln(1/\delta)}{2\beta^2\mathsf{mean}_{\mathsf{avg}}(D,k)} .$$

*If an $(\alpha, \gamma)$-approximation algorithm for $k$-means $\mathcal{A}$ is run on input $S$, then the following holds for the solution $C^*$ returned by $\mathcal{A}$:*

$$\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C^*) \leq (\alpha + \beta)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma] \geq 1 - \delta .$$

Next, we need to show that any clustering $C_b$ that is an $(\alpha+3\beta, \gamma)$-bad solution of $k$-means of $D$ satisfies with high probability $\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) > (\alpha+2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma$. Here we use epsilon-nets to approximate the size of the set of bad solutions denoted by $\mathbb{C}$.

**Lemma 10.** *Let $S$ be a multiset of $s$ points chosen i.u.r. from $D$ such that*

$$s \geq \frac{2M^4}{\beta^2(\mathsf{mean}_{\mathsf{avg}}(D,k))^2} \cdot \left( \ln(1/\delta) + kd \ln \left( \frac{\sqrt{d}M}{2\epsilon} \right) \right)$$

*Let $\mathbb{C}$ be the set of $(\alpha + 3\beta, \gamma)$-bad solutions of a $k$-mean clustering of $D$. Then*

$$\Pr[\exists C_b \in \mathbb{C} : \mathsf{cost}_{\mathsf{avg}}^{\mathsf{mean}}(S, C_b) \leq (\alpha + 2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma] \leq \delta$$

*Proof.* Consider an arbitrary $C_b \in \mathbb{C}$, and define $X_i$ as the distance of the $i$th point in $S$ from the nearest center in $C_b$. Since $C_b$ is a $(\alpha + 3\beta, \gamma)$-bad solutions of a $k$-means of $D$, by definition,

$$\mathsf{cost}_{\mathsf{avg}}^{\mathsf{mean}}(D, C_b) > (\alpha + 3\beta)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma \tag{24}$$

Now for $1 \leq i \leq s$, we have that $\mathbb{E}[X_i] = \frac{1}{|D|}\sum_{x \in D}(d(x, C_b))^2 = \mathsf{cost}_{\mathsf{avg}}^{\mathsf{mean}}(D, C_b)$, thus

$$\mathbb{E}[X_i] > (\alpha + 3\beta)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma \tag{25}$$

Also,

$$\sum_{1 \leq i \leq s} X_i = \sum_{x \in S} d(x, C_b) = s \cdot \mathsf{cost}_{\mathsf{avg}}^{\mathsf{mean}}(S, C_b) \ , \tag{26}$$

and $\mathbb{E}[\sum_{1 \leq i \leq s} X_i] = s\,\mathbb{E}[X_i]$ for any $i$, recall that $\mathbb{E}[X_i] = \mathsf{cost}_{\mathsf{avg}}^{\mathsf{mean}}(D, C_b)$ and hence technically independent of $i$.

We want to show that for any $C_b \in \mathbb{C}$, $\Pr[\mathsf{cost}_{\mathsf{avg}}^{\mathsf{mean}}(S, C_b) \leq (\alpha + 2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma]$ is low, and then take a union bound over the entire space of $\mathbb{C}$.

$\Pr[\mathsf{cost}_{\mathsf{avg}}^{\mathsf{mean}}(S, C_b) \leq (\alpha + 2\beta)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma]$

Substituting Relation 26 on LHS and Relation 25 on RHS,

$$= \Pr\left[ \frac{1}{s} \cdot \sum_{1 \leq i \leq s} X_i \leq \frac{(\alpha + 2\beta)}{(\alpha + 3\beta)}\,\mathbb{E}[X_i] + \gamma \cdot \left( 1 - \frac{(\alpha + 2\beta)}{(\alpha + 3\beta)} \right) \right]$$

$$= \Pr\left[ \sum_{1 \leq i \leq s} X_i \leq \frac{(\alpha + 2\beta)}{(\alpha + 3\beta)} \cdot s \cdot \mathbb{E}[X_i] + \frac{s\gamma\beta}{\alpha + 3\beta} \right]$$

$$= \Pr\left[ \sum_{1 \leq i \leq s} X_i \leq \frac{(\alpha + 2\beta)}{(\alpha + 3\beta)} \cdot \mathbb{E}[\sum_{1 \leq i \leq s} X_i] + \frac{s\gamma\beta}{\alpha + 3\beta} \right]$$

$$= \Pr\left[ \sum_{1 \leq i \leq s} X_i \leq \left( \frac{(\alpha + 2\beta)}{(\alpha + 3\beta)} + \frac{s\gamma\beta}{(\alpha + 3\beta)\,\mathbb{E}[\sum_{1 \leq i \leq s} X_i]} \right) \cdot \mathbb{E}[\sum_{1 \leq i \leq s} X_i] \right]$$

$$= \Pr\left[ \sum_{1 \leq i \leq s} X_i \leq \left( 1 - \left( \frac{\beta}{(\alpha + 3\beta)} - \frac{s\gamma\beta}{(\alpha + 3\beta) \cdot s \cdot \mathsf{cost}_{\mathsf{avg}}(D, C_b)} \right) \right) \cdot \mathbb{E}[\sum_{1 \leq i \leq s} X_i] \right]$$

Since $0 \le X_i \le M^2$, we can apply a Hoeffding bound,

$$\begin{aligned}
&\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) \le (\alpha + 2\beta)\mathsf{mean}_{\mathsf{avg}}(D, k) + \gamma] \\
&\le \exp\left(-\frac{\mathbb{E}[\sum_{1\le i\le s} X_i]}{2M^2} \cdot \left(\frac{\beta}{(\alpha+3\beta)} - \frac{\gamma\beta}{(\alpha+3\beta)\cdot \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b)}\right)^2\right) \\
&= \exp\left(-\frac{s \cdot \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b)}{2M^2} \cdot \left(\frac{\beta \cdot \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b) - \gamma\beta}{(\alpha+3\beta)\cdot \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b)}\right)^2\right) \\
&= \exp\left(-\frac{s\beta^2}{2M^2\cdot(\alpha+3\beta)^2} \cdot \frac{(\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b) - \gamma)^2}{\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b)}\right) \\
&\le \exp\left(-\frac{s\beta^2}{2M^2\cdot(\alpha+3\beta)^2} \cdot \frac{(\mathsf{mean}_{\mathsf{avg}}(D, k))^2(\alpha+3\beta)^2}{\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b)}\right), \qquad \text{Applying relation 24}
\end{aligned}$$

Now, $\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D, C_b) \le M^2$, therefore

$$\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) \le (\alpha + 2\beta)\mathsf{mean}_{\mathsf{avg}}(D, k) + \gamma] \le \exp\left(-\frac{s\beta^2}{2M^4} \cdot (\mathsf{mean}_{\mathsf{avg}}(D, k))^2\right)$$

By union bound and using the fact that $|\mathbb{C}| \le \left(\frac{\sqrt{d}M}{2\epsilon}\right)^{kd}$,

$$\begin{aligned}
&\Pr[\exists C_b \in \mathbb{C} : \mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S, C_b) \le (\alpha + 2\beta)\mathsf{mean}_{\mathsf{avg}}(D, k) + \gamma] \\
&\le \left(\frac{\sqrt{d}M}{2\epsilon}\right)^{kd} \cdot \exp\left(-\frac{s\beta^2}{2M^4} \cdot (\mathsf{mean}_{\mathsf{avg}}(D, k))^2\right)
\end{aligned}$$

We choose

$$s \ge \frac{2M^4}{\beta^2(\mathsf{mean}_{\mathsf{avg}}(D, k))^2} \cdot \left(\ln(1/\delta) + kd\ln\left(\frac{\sqrt{d}M}{2\epsilon}\right)\right)$$

□

Recall Lemma 6, which states that for $D \subseteq \mathbb{R}^d$, assuming an $(\alpha, \gamma)$-approximation $k$-means algorithm that runs in time $T(n)$, we can draw a sample $S$ of size $s$,

$$s \ge c \cdot \max\left\{\frac{M^2\alpha(1+\alpha)\ln(1/\delta)}{\epsilon}, \frac{M^4}{\epsilon^2} \cdot \left(\ln(1/\delta) + kd\ln\left(\frac{\sqrt{d}M}{2\epsilon}\right)\right)\right\},$$

where $c$ is a positive constant, and obtain a $k$-means clustering $\hat{c}_S$ in time $T(|S|)$ such that with probability at least $1-\delta$, $\mathbb{E}_D[\hat{c}_S] \le \alpha\,\mathbb{E}_D[c_D] + \gamma + \epsilon$, where $c_D$ is the optimum $k$-means clustering of $D$ in $\mathbb{R}^d$. The proof of this statement is below.

*Proof.* Let $\beta^*$ be a positive parameter that will be fixed later. Let $s$ be chosen such that sample complexity prerequisites of both Lemma 9 and Lemma 10 are satisfied. Recall from Lemma 9, the sample complexity is as follows,

$$s \ge \frac{3M^2\alpha(\beta^* + \alpha)\ln(1/\delta)}{2{\beta^*}^2\mathsf{mean}_{\mathsf{avg}}(D, k)}, \qquad (27)$$

And from Lemma 9 we have

$$s \geq \frac{2M^4}{{\beta^*}^2(\mathsf{mean}_{\mathsf{avg}}(D,k))^2} \cdot \left(\ln(1/\delta) + kd\ln\left(\frac{\sqrt{d}M}{2\epsilon}\right)\right) \ , \tag{28}$$

Thus an appropriate sample complexity is

$$s \geq \max\left\{\frac{3M^2\alpha(\beta^* + \alpha)\ln(1/\delta)}{2{\beta^*}^2\mathsf{mean}_{\mathsf{avg}}(D,k)}, \frac{2M^4}{{\beta^*}^2(\mathsf{mean}_{\mathsf{avg}}(D,k))^2} \cdot \left(\ln(1/\delta) + kd\ln\left(\frac{\sqrt{d}M}{2\epsilon}\right)\right)\right\} \tag{29}$$

For the chosen sample complexity, we have from Lemma 10 that with probability at least $1-\delta$, no clustering $C \subseteq \mathbb{R}^d$ that is a $(\alpha + 3\beta^*, \gamma)$-bad solution of a $k$-means of $D$ satisfies the inequality

$$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S,C) \leq (\alpha + 2\beta^*)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma \tag{30}$$

On the other hand, if we run algorithm $\mathcal{A}(S)$, then by Lemma 9, the resulting clustering $C^*$ with probability at least $1-\delta$ satisfies

$$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(S,C^*) \leq (\alpha + \beta^*)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma \tag{31}$$

Thus with probability at least $1-2\delta$, the clustering $C^*$ must be a $(\alpha + 3\beta^*, \gamma)$-good solution of a $k$-means of $D$, in other words,

$$\Pr[\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C^*) \leq (\alpha + 3\beta^*)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma] \geq 1 - 2\delta \ . \tag{32}$$

To complete the proof, we must remove the dependency on $\mathsf{mean}_{\mathsf{avg}}(D,k)$ in the sample complexity.

- Case 1: $\mathsf{mean}_{\mathsf{avg}}(D,k) < \epsilon$. Choose $\beta^* = \epsilon/(3 \cdot \mathsf{mean}_{\mathsf{avg}}(D,k) \geq 1/3$, and $\beta = 1/3\beta^* < 1$ then we get that if sample complexity

  $$s \geq c \cdot \max\left\{\frac{M^2\alpha(1+\alpha\beta)\ln(1/\delta)}{\epsilon}, \frac{M^4}{\epsilon^2} \cdot \left(\ln(1/\delta) + kd\ln\left(\frac{\sqrt{d}M}{2\epsilon}\right)\right)\right\} \ ,$$

  where $c$ is a certain positive constant, then with probability $1-2\delta$,

  $$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C^*) \leq (\alpha + 3\beta^*)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma \leq \alpha \cdot \mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma + \epsilon \ ,$$

  and

- Case 2: $\mathsf{mean}_{\mathsf{avg}}(D,k) \geq \epsilon$. Choose $\beta^* = \epsilon/(3 \cdot \mathsf{mean}_{\mathsf{avg}}(D,k)) \leq 1/3$. Then, we get that if sample complexity

  $$s \geq c \cdot \max\left\{\frac{M^2\alpha(1+\alpha)\ln(1/\delta)}{\epsilon}, \frac{M^4}{\epsilon^2} \cdot \left(\ln(1/\delta) + kd\ln\left(\frac{\sqrt{d}M}{2\epsilon}\right)\right)\right\} \ ,$$

  where $c$ is a certain positive constant, then with probability $1-2\delta$,

  $$\mathsf{cost}^{\mathsf{mean}}_{\mathsf{avg}}(D,C^*) \leq (\alpha + 3\beta^*)\mathsf{mean}_{\mathsf{avg}}(D,k) + \gamma \leq \alpha \cdot \mathsf{mean}_{\mathsf{avg}}(D,k) + 3\epsilon + \gamma \ ,$$

□