A simple D^2-sampling based PTAS for k-means and other Clustering problems

Ragesh Jaiswal[1], Amit Kumar[1], and Sandeep Sen[1]

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

Abstract. Given a set of points $P \subset \mathbb{R}^d$, the k-means clustering problem is to find a set of k *centers* $C = \{c_1, ..., c_k\}, c_i \in \mathbb{R}^d$, such that the objective function $\sum_{x \in P} d(x, C)^2$, where $d(x, C)$ denotes the distance between x and the closest center in C, is minimized. This is one of the most prominent objective functions that have been studied with respect to clustering.

D^2-sampling [7] is a simple non-uniform sampling technique for choosing points from a set of points. It works as follows: given a set of points $P \subseteq \mathbb{R}^d$, the first point is chosen uniformly at random from P. Subsequently, a point from P is chosen as the next sample with probability proportional to the square of the distance of this point to the nearest previously sampled points.

D^2-sampling has been shown to have nice properties with respect to the k-means clustering problem. Arthur and Vassilvitskii [7] show that k points chosen as centers from P using D^2-sampling gives an $O(\log k)$ approximation in expectation. Ailon et. al. [5] and Aggarwal et. al. [4] extended results of [7] to show that $O(k)$ points chosen as centers using D^2-sampling give $O(1)$ approximation to the k-means objective function with high probability. In this paper, we further demonstrate the power of D^2-sampling by giving a simple randomized $(1 + \epsilon)$-approximation algorithm that uses the D^2-sampling in its core.

1 Introduction

Clustering problems arise in diverse areas including machine learning, data mining, image processing and web-search [11, 16, 15, 25]. One of the most commonly used clustering problems is the k-means problem. Here, we are given a set of points P in a d-dimensional Euclidean space, and a parameter k. The goal is to find a set C of k *centers* such that the objective function

$$\Delta(P, C) = \sum_{p \in P} d(p, C)^2$$

is minimized, where $d(p, C)$ denotes the distance from p to the closest center in C. This naturally partitions P into k clusters, where each cluster corresponds to the set of points of P which are closer to a particular center than other centers.

It is also easy to show that the center of any cluster must be the mean of the points in it. In most applications, the parameter k is a small constant. However, this problem turns out to be NP-hard even for $k = 2$ [13].

One very popular heuristic for solving the k-means problem is the Lloyd's algorithm [22]. The heuristic is as follows : start with an arbitrary set of k centers as seeds. Based on these k centers, partition the set of points into k clusters, where each point gets assigned to the closest center. Now, we update the set of centers as the means of each of these clusters. This process is repeated till we get convergence. Although, this heuristic often performs well in practice, it is known that it can get stuck in local minima [6]. There has been lot of recent research in understanding why this heuristic works fast in practice, and how it can be modified such that we can guarantee that the solution produced by this heuristic is always close to the optimal solution.

One such modification is to carefully choose the set of initial k centers. Ideally, we would like to pick these centers such that we have a center close to each of the optimal clusters. Since we do not know the optimal clustering, we would like to make sure that these centers are well separated from each other and yet, are representatives of the set of points. A recently proposed idea [24, 7] is to pick the initial centers using D^2-sampling which can be described as follows. The first center is picked uniformly at random from the set of points P. Suppose we have picked a set of $k' < k$ centers – call this set C'. Then a point $p \in P$ is chosen as the next center with probability proportional to $d(p, C')^2$. This process is repeated till we have a set of k centers.

There has been lot of recent activity in understanding how good a set of centers picked by D^2-sampling are (even if we do not run the Lloyd's algorithm on these seed centers). Arthur and Vassilvitskii [7] showed that if we pick k centers with D^2-sampling, then the expected cost of the corresponding solution to the k-means instance is within $O(\log k)$-factor of the optimal value. Ostrovsky et. al. [24] showed that if the set of points satisfied a separation condition (named (ϵ^2, k)-irreducible as defined in Section 2), then these k centers give a constant factor approximation for the k-means problem. Ailon et. al. [5] proved a bi-criteria approximation property – if we pick $O(k \log k)$ centers by D^2-sampling, then it is a constant approximation, where we compare with the optimal solution that is allowed to pick k centers only. Aggarwal et. al. [4] give an improved result and show that it is enough to pick $O(k)$ centers by D^2-sampling to get a constant factor bi-criteria approximation algorithm.

In this paper, we give yet another illustration of the power of the D^2-sampling idea. We give a simple randomized $(1 + \epsilon)$-approximation algorithm for the k-means algorithm, where $\epsilon > 0$ is an arbitrarily small constant. At the heart of our algorithm is the idea of D^2-sampling – given a set of already selected centers, we pick a small set of points by D^2-sampling with respect to these selected centers. Then, we pick the next center as the centroid of a subset of these small set of points. By repeating this process of picking k centers sufficiently many times, we can guarantee that with high probability, we will get a set of k centers whose objective value is close to the optimal value. Further, the running time

of our algorithm is $O(nd \cdot 2^{\tilde{O}(k^2/\epsilon)})$ [1]– for constant value of k, this is a linear time algorithm. It is important to note that PTAS with better running time are known for this problem. Chen [12] give an $O\left(nkd + d^2 n^\sigma \cdot 2^{(k/\epsilon)^{O(1)}}\right)$ algorithm for any $\sigma > 0$ and Feldman et al. [17] give an $O\left(nkd + d \cdot poly(k/\epsilon) + 2^{\tilde{O}(k/\epsilon)}\right)$ algorithm. However, these results often are quite involved, and use the notion of coresets. Our algorithm is simple, and only uses the concept of D^2-sampling.

1.1 Other Related Work

There has been significant research on exactly solving the k-means algorithm (see e.g., [20]), but all of these algorithms take $\Omega(n^{kd})$ time. Hence, recent research on this problem has focused on obtaining fast $(1+\epsilon)$-approximation algorithms for any $\epsilon > 0$. Matousek [23] gave a PTAS with running time $O(n\epsilon^{-2k^2 d} \log^k n)$. Badoiu et al. [9] gave an improved PTAS with running time $O(2^{(k/\epsilon)^{O(1)}} d^{O(1)} n \log^{O(k)} n)$. de la Vega et al. [14] gave a PTAS which works well for points in high dimensions. The running time of this algorithm is $O(g(k,\epsilon) n \log^k n)$ where $g(k,\epsilon) = \exp[(k^3/\epsilon^8)(\ln(k/\epsilon)\ln k]$. Har-Peled et al. [18] proposed a PTAS whose running time is $O(n + k^{k+2}\epsilon^{-(2d+1)k} \log^{k+1} n \log^k \frac{1}{\epsilon})$. Kumar et al. [21] gave the first linear time PTAS for fixed k – the running time of their algorithm is $O(2^{(k/\epsilon)^{O(1)}} dn)$. Chen [12] used the a new coreset construction to give a PTAS with improved running time of $O(ndk + 2^{(k/\epsilon)^{O(1)}} d^2 n^\sigma)$. Recently, Feldman et al. [17] gave a PTAS with running time $O(nkd + d \cdot poly(k/\epsilon) + 2^{\tilde{O}(k/\epsilon)})$ – this is the fastest known PTAS (for fixed k) for this problem.

There has also been work on obtaining fast constant factor approximation algorithms for the k-means problem based on some properties of the input points (see e.g. [24, 8]).

1.2 Our Contributions

In this paper, we give a simple PTAS for the k-means problem based on the idea of D^2-sampling. Our work builds on and simplifies the result of Kumar et al. [21]. We briefly describe their algorithm first. It is well known that for the 1-mean problem, if we sample a set of $O(1/\epsilon)$ points uniformly at random, then the mean of this set of sampled points is close to the overall mean of the set of all points. Their algorithm begins by sampling $O(k/\epsilon)$ points uniformly at random. With reasonable probability, we would sample $O(1/\epsilon)$ points from the largest cluster, and hence we could get a good approximation to the center corresponding to this cluster (their algorithm tries all subsets of size $O(1/\epsilon)$ from the randomly sampled points). However, the other clusters may be much smaller, and we may not have sampled enough points from them. So, they need to prune a lot of points from the largest cluster so that in the next iteration a random sample of $O(k/\epsilon)$ points will contain $O(1/\epsilon)$ points from the second largest cluster, and so on. This

[1] $\tilde{O}$ notation hides a $O(\log k/\epsilon)$ factor which simplifies the expression.

requires a non-trivial idea termed as *tightness* condition by the authors. In this paper, we show that the pruning is not necessary if instead of using uniform random sampling, one uses D^2-sampling.

We can informally describe our algorithm as follows. We maintain a set of candidate centers C, which is initially empty. Given a set C, $|C| < k$, we add a new center to C as follows. We sample a set S of $O(k/\epsilon^3)$ points using D^2-sampling with respect to C. From this set of sampled points, we pick a subset T and the new center is the mean of this set T. We add this to C and continue.

From the property of D^2-sampling ([4, 5]), with some constant, albeit small probability p', we pick up a point from a hitherto untouched cluster C' of the *optimal clustering*. Therefore by sampling about α/p' points using D^2-sampling, we expect to hit approximately α points from C'. If α is large enough, (c.f. Lemma 1), then the centroid of these α points gives a $(1+\epsilon)$ approximation of the cluster C'. Therefore, with reasonable probability, there will be a choice of a subset T in each iteration such that the set of centers chosen are from C'. Since we do not know T, our algorithm will try out all subsets of size $|T|$ from the sample S. Note that our algorithm is very simple, and can be easily parallelized. Our algorithm has running time $O(dn \cdot 2^{\tilde{O}(k^2/\epsilon)})$ which is an improvement over that of Kumar et al. [21] who gave a PTAS with running time $O\left(nd \cdot 2^{(k/\epsilon)^{O(1)}}\right)$. [2]

Because of the relative simplicity, our algorithm generalizes to measures like Mahalanobis distance and μ-similar Bregman divergence. Note that these do not satisfy triangle inequality and therefore not strict metrics. Ackermann et al. [2] have generalized the framework of Kumar et al. [21] to Bregman divergences but we feel that the D^2-sampling based algorithms are simpler.

We formally define the problem and give some preliminary results in Section 2. In Section 3, we describe our algorithm, and then analyze it subsequently. In Section A, we discuss PTAS for other distance measures.

2 Preliminaries

An instance of the k-means problem consists of a set $P \subseteq \mathbb{R}^d$ of n points in d-dimensional space and a parameter k. For a set of points (called centers) $C \subseteq \mathbb{R}^d$, let $\Delta(P, C)$ denote $\sum_{p \in P} d(p, C)^2$, i.e., the cost of the solution which picks C as the set of centers. For a singleton $C = \{c\}$, we shall often abuse notation, and use $\Delta(P, c)$ to denote $\Delta(P, C)$. Let $\Delta_k(P)$ denote the cost of the optimal k-means solution for P.

Definition 1. *Given a set of points P and a set of centers C, a point $p \in P$ is said to be sampled using* D^2-sampling *with respect to C if the probability of it being sampled, $\rho(p)$, is given by*

$$\rho(p) = \frac{d(p, C)^2}{\sum_{x \in P} d(x, C)^2} = \frac{\Delta(\{p\}, C)}{\Delta(P, C)}.$$

[2] It can be used in conjunction with Chen [12] to obtain a superior running time but at the cost of the simplicity of our approach

We will also need the following definition from [21].

Definition 2 (Irreducibility or separation condition). *Given k and ϵ, a set of points P is said to be (k, γ)-irreducible if*

$$\Delta_{k-1}(P) \geq (1 + \gamma) \cdot \Delta_k(P).$$

We will often appeal to the following result [20] which shows that uniform random sampling works well for 1-means[3].

Lemma 1 (Inaba et al. [20]). *Let S be a set of points obtained by independently sampling M points with replacement uniformly at random from a point set P. Then, for any $\delta > 0$,*

$$\Delta(P, \{m(S)\}) \leq \left(1 + \frac{1}{\delta M}\right) \cdot \Delta(P, \{m(P)\}),$$

holds with probability at least $(1 - \delta)$. Here $m(X) = \left(\frac{\sum_{x \in X} x}{|X|}\right)$ denotes the centroid of a point set X.

Finally, we will use the following property of the squared Euclidean metric. This is a standard result from linear algebra [19].

Lemma 2. *Let $P \subseteq \mathbb{R}^d$ be any point set and let $c \in \mathbb{R}^d$ be any point. Then we have the following:*

$$\sum_{p \in P} d(p, c)^2 = \sum_{p \in P} d(p, m(P))^2 + |P| \cdot d(c, m(P))^2,$$

where $m(P) = \left(\frac{\sum_{p \in P} p}{|P|}\right)$ denotes the centroid of the point set.

Finally, we mention the simple approximate triangle inequality with respect to the squared Euclidean distance measure.

Lemma 3 (Approximate triangle inequality). *For any three points $p, q, r \in \mathbb{R}^d$ we have:*

$$d(p, q)^2 \leq 2 \cdot (d(p, r)^2 + d(r, q)^2).$$

3 PTAS for *k*-means

We first give a high level description behind the algorithm. We will also assume that the instance is (k, ϵ)-irreducible for a suitably small parameter ϵ. We shall then get rid of this assumption later. The algorithm is described in Figure 1. Essentially, the algorithm maintains a set C of centers, where $|C| \leq k$. Initially C is empty, and in each iteration of Step 2(b), it adds one center to C till its size

[3] It turns out that even minor perturbations from uniform distribution can be catastrophic and indeed in this paper we had to work around this.

reaches k. Given a set C, it samples a set of S points from P using D^2-sampling with respect to C (in Step 2(b)). Then it picks a subset T of S of size $M = O(1/\epsilon)$, and adds the centroid of T to C. The algorithm cycles through all possible subsets of size M of S as choices for T, and for each such choice, repeats the above steps to find the next center, and so on. To make the presentation clearer, we pick a k-tuple of M-size subsets $(s_1, \ldots, s_k)$ in advance, and when $|C| = i$, we pick T as the s_i^{th} subset of S. In Step 2(i), we cycle through all such k-tuples $(s_1, \ldots, s_k)$. In the analysis, we just need to show that *one* such k-tuple works with reasonable probability.

We develop some notation first. For the rest of the analysis, we will fix a tuple $(s_1, \ldots, s_k)$ – this will be the "desired tuple", i.e., the one for which we can show that the set C gives a good solution. As our analysis proceeds, we will argue what properties this tuple should have. Let $C^{(i)}$ be the set C at the beginning of the i^{th} iteration of Step 2(b). To begin with $C^{(0)}$ is empty. Let $S^{(i)}$ be the set S sampled during the i^{th} iteration of Step 2(b), and $T^{(i)}$ be the corresponding set T (which is the s_i^{th} subset of $S^{(i)}$).

Let $O_1, \ldots, O_k$ be the optimal clusters, and c_i denote the centroid of points in O_i. Further, let m_i denote $|O_i|$, and wlog assume that $m_1 \geq \ldots \geq m_k$. Note that $\Delta_1(O_i)$ is same as $\Delta(O_i, \{c_i\})$. Let r_i denote the average cost paid by a point in O_i, i.e.,

$$r_i = \frac{\sum_{p \in O_i} d(p, c_i)^2}{m_i}.$$

We will assume that the input set of points P are (k, ϵ)-irreducible. We shall remove this assumption later. Now we show that any two optimal centers are far enough.

Find-k-means(P)
Let $N = (51200 \cdot k/\epsilon^3)$, $M = 100/\epsilon$, and $P = \binom{N}{M}$
1. **Repeat** 2^k times and output the the set of centers C that give least cost
2. **Repeat** for all k-tuples $(s_1, ..., s_k) \in [P] \times [P] \times .... \times [P]$ and pick the set of centers C that gives least cost
(a) $C \leftarrow \{\}$
(b) For $i \leftarrow 1$ to k
Sample a set S of N points with D^2-sampling (w.r.t. centers C)
Let T be the s_i^{th} subset of S. [a]
$C \leftarrow C \cup \{m(T)\}$. [b]

[a] For a set of size N we consider an arbitrary ordering of the subsets of size M of this set.
[b] $m(T)$ denote the centroid of the points in T.

Fig. 1. The k-means algorithm that gives $(1+\epsilon)$-approximation for any (k, ϵ)-irreducible data set. Note that the inner loop is executed at most $2^k \cdot \left(\binom{N}{M}\right)^k \sim 2^k \cdot 2^{\tilde{O}(k/\epsilon)}$ times.

Lemma 4. *For any* $1 \leq i, j \leq k, i \neq j$,

$$d(c_i, c_j)^2 \geq \epsilon \cdot (r_i + r_j).$$

Proof. Suppose $i > j$, and hence $m_i \geq m_j$. For the sake of contradiction assume $d(c_i, c_j)^2 < \epsilon \cdot (r_i + r_j)$. Then we have,

$$\begin{aligned}
\Delta(O_i \cup O_j, \{c_i\}) &= m_i \cdot r_i + m_j \cdot r_j + m_j \cdot d(c_i, c_j)^2 \quad \text{(using Lemma 2)} \\
&\leq m_i \cdot r_i + m_j \cdot r_j + m_j \cdot \epsilon \cdot (r_i + r_j) \\
&\leq (1+\epsilon) \cdot m_i \cdot r_i + (1+\epsilon) \cdot m_j \cdot r_j \quad \text{(since } m_i \geq m_j\text{)} \\
&\leq (1+\epsilon) \cdot \Delta(O_i \cup O_j, \{c_i, c_j\})
\end{aligned}$$

This implies that the centers $\{c_1, ..., c_k\} \backslash \{c_j\}$ give a $(1+\epsilon)$-approximation to the k-means objective. This contradicts the assumption that P is (ϵ, k)-irreducible.

We give an outline of the proof. Suppose in the first $i-1$ iterations, we have found centers which are close to the centers of some $i-1$ clusters in the optimal solution. Conditioned on this fact, we show that in the next iteration, we are likely to sample enough number of points from one of the remaining clusters (c.f. Corollary 1). Further, we show that the samples from this new cluster are close to uniform distribution (c.f. Lemma 6). Since such a sample does not come from exactly uniform distribution, we cannot apply Lemma 1 directly. In fact, dealing with the slight non-uniformity turns out to be non-trivial (c.f. Lemmas 7 and 8).

We now show that the following invariant will hold for all iterations : let $C^{(i-1)}$ consist of centers $c'_1, \ldots, c'_{i-1}$ (added in this order). Then, with probability at least $\frac{1}{2^i}$, there exist distinct indices $j_1, \ldots, j_{i-1}$ such that for all $l = 1, \ldots, i-1$,

$$\Delta(O_{j_l}, c'_l) \leq (1 + \epsilon/20) \cdot \Delta(O_{j_l}, c_{j_l}) \tag{1}$$

Suppose this invariant holds for $C^{(i-1)}$ (the base case is easy since $C^{(0)}$ is empty). We now show that this invariant holds for $C^{(i)}$ as well. In other words, we just show that in the i^{th} iteration, with probability at least $1/2$, the algorithm finds a center c'_i such that

$$\Delta(O_{j_i}, c'_i) \leq (1 + \epsilon/20) \cdot \Delta(O_{j_i}, c_{j_i}),$$

where j_i is an index distinct from $\{j_1, \ldots, j_{i-1}\}$. This will basically show that at the end of the last iteration, we will have k centers that give a $(1+\epsilon)$-approximation with probability at least 2^{-k}.

We now show that the invariant holds for $C^{(i)}$. We use the notation developed above for $C^{(i-1)}$. Let I denote the set of indices $\{j_1, \ldots, j_{i-1}\}$. Now let j_i be the index $j \notin I$ for which $\Delta(O_j, C^{(i-1)})$ is maximum. Intuitively, conditioned on sampling from clusters in $O_i, \cdots, O_k$ using D^2-sampling, it is likely that enough points from O_{j_i} will be sampled. The next lemma shows that there is good chance that elements from the sets O_j for $j \notin I$ will be sampled.

Lemma 5.

$$\frac{\sum_{l \notin I} \Delta(O_l, C^{(i-1)})}{\sum_{l=1}^{k} \Delta(O_l, C^{(i-1)})} \geq \epsilon/2.$$

Proof. Suppose, for the sake of contradiction, the above statement does not hold. Then,

$$\begin{aligned}
\Delta(P, C^{(i-1)}) &= \sum_{l \in I} \Delta(O_l, C^{(i-1)}) + \sum_{l \notin I} \Delta(O_l, C^{(i-1)}) \\
&< \sum_{l \in I} \Delta(O_l, C^{(i-1)}) + \frac{\epsilon/2}{1-\epsilon/2} \cdot \sum_{l \in I} \Delta(O_l, C^{(i-1)}) \quad \text{(by our assumption)} \\
&= \frac{1}{1-\epsilon/2} \cdot \sum_{l \in I} \Delta(O_l, C^{(i-1)}) \\
&\leq \frac{1+\epsilon/20}{1-\epsilon/2} \cdot \sum_{l \in I} \Delta_1(O_l) \quad \text{(using the invariant for } C^{(i-1)}) \\
&\leq (1+\epsilon) \cdot \sum_{l \in I} \Delta_1(O_l) \;\leq\; (1+\epsilon) \cdot \sum_{l \in [k]} \Delta_1(O_l)
\end{aligned}$$

But this contradicts the fact that P is (k, ϵ)-irreducible.

We get the following corollary easily.

Corollary 1.

$$\frac{\Delta(O_{j_i}, C^{(i-1)})}{\sum_{l=1}^{k} \Delta(O_l, C^{(i-1)})} \geq \frac{\epsilon}{2k}.$$

The above Lemma and its Corollary say that with probability at least $\frac{\epsilon}{2k}$, points in the set O_{j_i} will be sampled. However the points within O_{j_i} are not sampled uniformly. Some points in O_{j_i} might be sampled with higher probability than other points. In the next lemma, we show that each point will be sampled with certain minimum probability.

Lemma 6. *For any $l \notin I$ and any point $p \in O_l$,*

$$\frac{d(p, C^{(i-1)})^2}{\Delta(O_l, C^{(i-1)})} \geq \frac{1}{m_l} \cdot \frac{\epsilon}{64}.$$

Proof. Fix a point $p \in O_l$. Let $j_t \in I$ be the index such that p is closest to c'_t among all centers in $C^{(i-1)}$. We have

$$\begin{aligned}
\Delta(O_l, C^{(i-1)}) &\leq m_l \cdot r_l + m_l \cdot d(c_l, c'_t)^2 \quad \text{(using Lemma 2)} \\
&\leq m_l \cdot r_l + 2 \cdot m_l \cdot \left(d(c_l, c_{j_t})^2 + d(c_{j_t}, c'_t)^2 \right) \quad \text{(using Lemma 3)} \\
&\leq m_l \cdot r_l + 2 \cdot m_l \cdot \left(d(c_l, c_{j_t})^2 + \frac{\epsilon r_t}{20} \right), \qquad (2)
\end{aligned}$$

where the second inequality follows from the invariant condition for $C^{(i-1)}$. Also, we know that

$$\begin{aligned}
d(p, c'_t)^2 &\geq \frac{d(c_{j_t}, c_l)^2}{8} - d(c_{j_t}, c'_t)^2 \quad \text{(using Lemma 3)} \\
&\geq \frac{d(c_{j_t}, c_l)^2}{8} - \frac{\epsilon}{20} \cdot r_t \quad \text{(using the invariant for } C^{(i-1)}) \\
&\geq \frac{d(c_{j_t}, c_l)^2}{16} \quad \text{(Using Lemma 4)} \qquad (3)
\end{aligned}$$

So, we get

$$\frac{d(p, C^{(i-1)})^2}{\Delta(O_l, C^{(i-1)})} \geq \frac{d(c_{j_t}, c_l)^2}{16 \cdot m_l \cdot \left(r_l + 2\left(d(c_{j_t}, c_l)^2 + \frac{\epsilon r_t}{20}\right)\right)} \quad \text{(using (2) and (3))}$$

$$\geq \frac{1}{16 \cdot m_l} \cdot \frac{1}{(1/\epsilon) + 2 + 1/20} \geq \frac{\epsilon}{64 \cdot m_l} \quad \text{(using Lemma 4)}$$

Recall that $S^{(i)}$ is the sample of size N in this iteration. We would like to show that that the invariant will hold in this iteration as well. We first prove a simple corollary of Lemma 1.

Lemma 7. *Let Q be a set of n points, and γ be a parameter, $0 < \gamma < 1$. Define a random variable X as follows : with probability γ, it picks an element of Q uniformly at random, and with probability $1 - \gamma$, it does not pick any element (i.e., is null). Let $X_1, \ldots, X_\ell$ be ℓ independent copies of X, where $\ell = \frac{400}{\gamma\epsilon}$. Let T denote the (multi-set) of elements of Q picked by $X_1, \ldots, X_\ell$. Then, with probability at least $3/4$, T contains a subset U of size $\frac{100}{\epsilon}$ which satsifies*

$$\Delta(P, m(U)) \leq \left(1 + \frac{\epsilon}{20}\right) \Delta_1(P) \tag{4}$$

Proof. Define a random variable I, which is a subset of the index set $\{1, \ldots, \ell\}$, as follows $I = \{t : X_t$ picks an element of Q, i.e., it is not null$\}$. Conditioned on $I = \{t_1, \ldots, t_r\}$, note that the random variables $X_{t_1}, \ldots, X_{t_r}$ are independent uniform samples from Q. Thus if $|I| \geq \frac{100}{\epsilon}$, then Lemma 1 implies that with probability at least 0.8, the desired event (4) happens. But the expected value of $|I|$ is $\frac{400}{\epsilon}$, and so, $|I| \geq \frac{100}{\epsilon}$ with high probability, and hence, the statement in the lemma is true.

We are now ready to prove the main lemma.

Lemma 8. *With probability at least $1/2$, there exists a subset $T^{(i)}$ of $S^{(i)}$ of size at most $\frac{100}{\epsilon}$ such that*

$$\Delta(O_{j_i}, m(T^{(i)})) \leq (1 + \frac{\epsilon}{20}) \cdot \Delta_1(O_{j_i}).$$

Proof. Recall that $S^{(i)}$ contains $N = \frac{51200k}{\epsilon^3}$ independent samples of P (using D^2-sampling). We are interested in $S^{(i)} \cap O_{j_i}$. Let $Y_1, \ldots, Y_N$ be N independent random variables defined as follows : for any t, $1 \leq t \leq N$, Y_t picks an element of P using D^2-sampling with respect to $C^{(i-1)}$. If this element is not in O_{j_i}, it just discards it (i.e., Y_t is null). Let γ denote $\frac{\epsilon^2}{128k}$. Corollary 1 and Lemma 6 imply that Y_t picks a particular element of O_{j_i} with probability at least $\frac{\gamma}{m_{j_i}}$. We would now like to apply Lemma 7 (observe that $N = \frac{400}{\gamma\epsilon}$). We can do this by a simple coupling argument as follows. For a particular element $p \in O_{j_i}$, suppose Y_t assigns probability $\frac{\gamma(p)}{m_{j_i}}$ to it. One way of sampling a random variable X_t as in Lemma 7 is as follows – first sample using Y_t. If Y_t is null then, X_t is also

null. Otherwise, suppose Y_t picks an element p of O_{j_i}. Then, X_t is equal to p with probability $\frac{\gamma}{\gamma(p)}$, null otherwise. It is easy to check that with probability γ, X_t is a uniform sample from O_{j_i}, and null with probability $1-\gamma$. Now, observe that the set of elements of O_{j_i} sampled by $Y_1, \ldots, Y_N$ is always a superset of $X_1, \ldots, X_N$. We can now use Lemma 7 to finish the proof.

Thus, we will take the index s_i in Step 2(i) as the index of the set $T^{(i)}$ as guaranteed by the Lemma above. Finally, by repeating the entire process 2^k times, we make sure that we get a $(1+\epsilon)$-approximate solution with high probability. Note that the total running time of our algorithm is $\left(nd \cdot 2^k \cdot 2^{\tilde{O}(k/\epsilon)}\right)$.

Removing the (k,ϵ)-irreducibility assumption : We now show how to remove this assumption. First note that we have shown the following result.

Theorem 1. *If a given point set $(k, \frac{\epsilon}{(1+\epsilon/2)\cdot k})$-irreducible, then there is an algorithm that gives a $(1+\frac{\epsilon}{(1+\epsilon/2)\cdot k})$-approximation to the k-means objective and that runs in time $O(nd \cdot 2^{\tilde{O}(k^2/\epsilon)})$.*

Proof. The proof can be obtained by replacing ϵ by $\frac{\epsilon}{(1+\epsilon/2)\cdot k}$ in the above analysis.

Suppose the point set P is not $(k, \frac{\epsilon}{(1+\epsilon/2)\cdot k})$-irreducible. In that case it will be sufficient to find fewer centers that $(1+\epsilon)$-approximate the k-means objective. The next lemma shows this more formally.

Theorem 2. *There is an algorithm that runs in time $O(nd \cdot 2^{\tilde{O}(k^2/\epsilon)})$ and gives a $(1+\epsilon)$-approximation to the k-means objective.*

Proof. Let P denote the set of points. Let $1 < j \leq k$ be the largest index such that P is $(i, \frac{\epsilon}{(1+\epsilon/2)\cdot k})$-irreducible. If no such i exists, then

$$\Delta_1(P) \leq \left(1 + \frac{\epsilon}{(1+\epsilon/2)\cdot k}\right)^k \cdot \Delta_k(P) \leq (1+\epsilon)\cdot \Delta_k(P),$$

and so picking the centroid of P will give a $(1+\epsilon)$-approximation.

Suppose such an i exists. In that case, we consider the i-means problem and from the previous lemma we get that there is an algorithm that runs in time $O(nd \cdot 2^i \cdot 2^{\tilde{O}(i^2/\epsilon)})$ and gives a $(1+\frac{\epsilon}{(1+\epsilon/2)\cdot k})$-approximation to the i-means objective. Now we have that

$$\Delta_i \leq \left(1 + \frac{\epsilon}{(1+\epsilon/2)\cdot k}\right)^{k-i} \cdot \Delta_k \leq (1+\epsilon)\cdot \Delta_k.$$

Thus, we are done.

4 Other Distance Measures

In the previous sections, we looked at the k-means problem where the dissimilarity or distance measure was the square of Euclidean distance. There are numerous practical clustering problem instances where the dissimilarity measure is not a function of the Euclidean distance. In many cases, the points are not generated from a metric space. In these cases, it makes sense to talk about the general k-median problem that can be defined as follows:

Definition 3 (k-median with respect to a dissimilarity measure). *Given a set of n objects $P \subseteq \mathcal{X}$ and a dissimilarity measure $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, find a subset C of k objects (called medians) such that the following objective function is minimized:*

$$\Delta(P, C) = \sum_{p \in P} \min_{c \in C} D(p, c)$$

In this section, we will show that our algorithm and analysis can be easily generalized and extended to dissimilarity measures that satisfy some simple properties. We will look at some interesting examples. Due to lack of space, we just give our main results for this section. The entire discussion could be found in the Appendix.

Theorem 3 (k-median w.r.t. Mahalanobis distance). *Let $0 < \epsilon \leq 1/2$. There is an algorithm that runs in time $O(nd \cdot 2^{\tilde{O}(k^2/\epsilon)})$ and gives a $(1+\epsilon)$-approximation to the k-median objective function w.r.t. Mahalanobis distances for any point set $P \in \mathbb{R}^d, |P| = n$.*

Theorem 4 (k-median w.r.t. μ-similar Bregman divergences). *Let $0 < \mu \leq 1$ and $0 < \epsilon \leq 1/2$. There is an algorithm that runs in time $O\left(nd \cdot 2^{\tilde{O}\left(\frac{k^2}{\mu \cdot \epsilon}\right)}\right)$ and gives a $(1+\epsilon)$-approximation to the k-median objective function w.r.t. μ-similar Bregman divergence for any point set $P \in \mathbb{R}^d, |P| = n$.*

References

1. Marcel R. Ackermann. *Algorithms for the Bregman k-Median Problem.* PhD thesis, 2010.
2. Marcel R. Ackermann and Johannes Blömer. Coresets and approximate clustering for bregman divergences. In *ACM SIAM Symposium on Discrete Algorithms*, pages 1088–1097, 2009.
3. Marcel R. Ackermann, Johannes Blömer, and Christian Sohler. Clustering for metric and nonmetric distance measures. *ACM Trans. Algorithms*, 6:59:1–59:26, September 2010.
4. Ankit Aggarwal, Amit Deshpande, and Ravi Kannan. Adaptive sampling for k-means clustering. In *APPROX-RANDOM*, pages 15–28, 2009.
5. Nir Ailon, Ragesh Jaiswal, and Claire Monteleoni. Streaming k-means approximation. In *Advances in Neural Information Processing Systems 22*, pages 10–18. 2009.

6. David Arthur and Sergei Vassilvitskii. How slow is the k-means method? In *Proc. 22nd Annual Symposium on Computational Geometry*, pages 144–153, 2006.
7. David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.
8. Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. In *FOCS*, pages 309–318, 2010.
9. Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *STOC*, pages 250–257, 2002.
10. Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, December 2005.
11. A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic clustering of the web.
12. Ke Chen. On k-median clustering in high dimensions. In *SODA*, pages 1177–1185, 2006.
13. Sanjoy Dasgupta. The hardness of k-means clustering. Technical Report CS2008-0916, Department of Computer Science and Engineering, University of California San Diego, 2008.
14. Wenceslas Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In *ACM Symposium on Theory of Computing*, pages 50–58, 2003.
15. S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and A.R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990.
16. C. Faloutsos, R. Barber, M. Flickner, and J. Hafner. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 1994.
17. Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A ptas for k-means clustering based on weak coresets. In *Symposium on Computational Geometry*, pages 11–18, 2007.
18. Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *ACM Symposium on Theory of Computing*, pages 291–300, 2004.
19. Sariel Har-Peled and Bardia Sadri. How fast is the k-means method? In *ACM SIAM Symposium on Discrete Algorithms*, pages 877–885, 2005.
20. M. Inaba, N. Katoh, and H. Imai. Applications of weighted voronoi diagrams and randomization to variance based k-clustering. In *Proceedings of the tenth annual symposium on Computational Geometry*, pages 332–339, 1994.
21. Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2), 2010.
22. S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
23. J. Matousek. On approximate geometric k-clustering. *Discrete and Computational Geometry*, 2000.
24. Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. In *Proc. 47th IEEE FOCS*, pages 165–176, 2006.
25. M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 1991.

A Other Distance Measures

In the Section 3, we looked at the k-means problem where the dissimilarity or distance measure was the square of Euclidean distance. There are numerous practical clustering problem instances where the dissimilarity measure is not a function of the Euclidean distance. In many cases, the points are not generated from a metric space. In these cases, it makes sense to talk about the general k-median problem that can be defined as follows:

Definition 4 (k-median with respect to a dissimilarity measure). *Given a set of n objects $P \subseteq \mathcal{X}$ and a dissimilarity measure $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, find a subset C of k objects (called medians) such that the following objective function is minimized:*

$$\Delta(P, C) = \sum_{p \in P} \min_{c \in C} D(p, c)$$

In this section, we will show that our algorithm and analysis can be easily generalized and extended to dissimilarity measures that satisfy some simple properties. We will look at some interesting examples. We start by making the observation that in the entire analysis of Section 3 the only properties of the distance measure that we used were given in Lemmas 1, 2, and 3. We also used the symmetry property of the Euclidean metric implicitly. This motivates us to consider dissimilarity measures on spaces where these lemmas (or mild relaxations of these) are true. For such measures, we may replace $d(p, q)^2$ (this is the square of the Euclidean distance) by $D(p, q)$ in all places in Section 3 and obtain a similar result. We will now formalize these ideas.

First, we will describe a property that captures Lemma 1. This is similar to a definition by Ackermann et. al. [3] who discuss PTAS for the k-median problem with respect to metric and non-metric distance measures.

Definition 5 ((f, γ, δ)-Sampling property). *Given $0 < \gamma, \delta \leq 1$ and $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, a distance measure D over space $\mathcal{X}$ is said to have (f, γ, δ)-sampling property if the following holds: for any set $P \subseteq \mathcal{X}$, a uniformly random sample S of $f(\gamma, \delta)$ points from P satisfies*

$$Pr\left[\sum_{p \in P} D(p, m(S)) \leq (1 + \gamma) \cdot \Delta_1(P)\right] \geq (1 - \delta),$$

where $m(S) = \frac{\sum_{s \in S} s}{|S|}$ denotes the mean of points in S.

Definition 6 (Centroid property). *A distance measure D over space $\mathcal{X}$ is said to satisfy the centroid property if for any subset $P \subseteq \mathcal{X}$ and any point $c \in \mathcal{X}$, we have:*

$$\sum_{p \in P} D(p, c) = \Delta_1(P) + |P| \cdot D(m(P), c),$$

where $m(P) = \frac{\sum_{p \in P} p}{|P|}$ denotes the mean of the points in P.

Definition 7 (α-approximate triangle inequality). *Given $\alpha \geq 1$, a distance measure D over space $\mathcal{X}$ is said to satisfy α-approximate triangle inequality if for any three points $p, q, r \in \mathcal{X}, D(p,q) \leq \alpha \cdot (D(p,r) + D(r,q))$*

Definition 8 (β-approximate symmetry). *Given $0 < \beta \leq 1$, a distance measure D over space $\mathcal{X}$ is said to satisfy β-symmetric property if for any pair of points $p, q \in \mathcal{X}$, $\beta \cdot D(q,p) \leq D(p,q) \leq \frac{1}{\beta} \cdot D(q,p)$*

The next theorem gives the generalization of our results for distance measures that satisfy the above basic properties. The proof of this theorem follows easily from the analysis in Section 3. The proof of this theorem is given in Appendix B.

Theorem 5. *Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Let $\alpha \geq 0$, $0 < \beta \leq 1$, and $0 < \delta < 1/2$ be constants and let $0 < \epsilon \leq 1/2$. Let $\eta = \frac{2\alpha^2}{\beta^2}(1+1/\beta)$. Let D be a distance measure over space $\mathcal{X}$ that D follows:*

1. *β-approximate symmetry property,*
2. *α-approximate triangle inequality,*
3. *Centroid property, and*
4. *(f, ϵ, δ)-sampling property.*

Then there is an algorithm that runs in time $O\left(nd \cdot 2^{\tilde{O}(k \cdot f(\epsilon/\eta k, 0.2))}\right)$ and gives a $(1+\epsilon)$-approximation to the k-median objective for any point set $P \subseteq \mathcal{X}, |P| = n$.

The above theorem gives a characterization for when our non-uniform sampling based algorithm can be used to obtain a PTAS for a dissimilarity measure. The important question now is whether there exist interesting distance measures that satisfy the properties in the above Theorem. Next, we look at some distance measures other than squared Euclidean distance, that satisfy such properties.

A.1 Mahalanobis distance

Here the domain is $\mathbb{R}^d$ and the distance is defined with respect to a positive definite matrix $A \in \mathbb{R}^{d \times d}$. The distance between two points $p, q \in \mathbb{R}^d$ is given by $D_A(p,q) = (p-q)^T \cdot A \cdot (p-q)$. Now, we discuss the properties in Theorem 5.

1. (*Symmetry*) For any pair of points $p, q \in \mathbb{R}^d$, we have $D_A(p,q) = D_A(q,p)$. So, the β-approximate symmetry property holds for $\beta = 1$.
2. (*Triangle inequality*) [2] shows that α-approximate triangle inequality holds for $\alpha = 2$.
3. (*Centroid*) The centroid property is shown to hold for Mahalanobis distance in [10].
4. (*Sampling*) [3] (see Corollary 3.7) show that Mahalanobis distance satisfy the (f, γ, δ)-sampling property for $f(\gamma, \delta) = 1/(\gamma\delta)$.

Using the above properties and Theorem 5, we get the following result.

Theorem 6 (k-median w.r.t. Mahalanobis distance). *Let $0 < \epsilon \leq 1/2$. There is an algorithm that runs in time $O(nd \cdot 2^{\tilde{O}(k^2/\epsilon)})$ and gives a $(1+\epsilon)$-approximation to the k-median objective function w.r.t. Mahalanobis distances for any point set $P \in \mathbb{R}^d, |P| = n$.*

A.2 μ-similar Bregman divergence

We start by defining Bregman divergence and then discuss the required properties.

Definition 9 (Bregman Divergence). *Let $\phi : X \to \mathbb{R}^d$ be a continuously-differentiable real-valued and strictly convex function defined on a closed convex set X. The Bregman distance associated with ϕ for points $p, q \in X$ is:*

$$D_\phi(p, q) = \phi(p) - \phi(q) - \Delta\phi(q)^T(p - q)$$

Where $\Delta\phi(q)$ denotes the gradient of ϕ at point q

Intuitively this can be thought of as the difference between the value of ϕ at point p and the value of the first-order Taylor expansion of ϕ around point q evaluated at point p. Bregman divergence includes the following popular distance measures:

- *Euclidean distance.* $D_\phi(p, q) = ||p - q||^2$. Here $\phi(x) = ||x||^2$.
- *Kullback-Leibler divergence.* $D_\phi(p, q) = \sum_i p_i \cdot \ln \frac{p_i}{q_i} - \sum_i (p_i - q_i)$. Here $D_\phi(x) = \sum_i x_i \cdot \ln x_i - x_i$.
- *Itakura-Saito divergence.* $D_\phi(p, q) = \sum_i \left(\ln \frac{p_i}{q_i} - \ln \frac{q_i}{p_i} - 1 \right)$. Here $\phi(x) = -\sum_i \ln x_i$.
- *Mahalanobis distance.* For a symmetric positive definite matrix $U \in \mathbb{R}^{d \times d}$, the Mahalanobis distance is defined as: $D_U(p, q) = (p - q)^T U (p - q)$. Here $\phi_U(x) = x^T U x$.

Bregman divergences have been shown to satisfy the Centroid property by Banerjee et. al. [10]. All Bregman divergences do not necessarily satisfy the symmetry property or the triangle inequality. So, we cannot hope to use our results for the class of all Bregman divergences. On the other hand, some of the Bregman divergences that are used in practice satisfy a property called *μ-similarity* (see [1] for an overview of such Bregman divergences). Next, we give the definition of μ-similarity.

Definition 10 (μ-similar Bregman divergence). *A Bregman divergence D_ϕ on domain $\mathbb{X} \subseteq \mathbb{R}^d$ is called μ-similar for constant $0 < \mu \leq 1$, if there exists a symmetric positive definite matrix U such that for Mahalanobis distance D_U and for each $p, q \in \mathbb{X}$ we have:*

$$\mu \cdot D_U(p, q) \leq D_\phi(p, q) \leq D_U(p, q). \tag{5}$$

Now, a μ-similar Bregman divergence can easily be shown to satisfy approximate symmetry and triangle inequality properties. This is formalized in the following simple lemma. The proof of this lemma is given in the Appendix C.

Lemma 9. *Let $0 < \mu \leq 1$. Any μ-similar Bregman divergence satisfies the μ-approximate symmetry property and $(2/\mu)$-approximate triangle inequality.*

Finally, we use the sampling property from Ackermann et. al. [3] who show that any μ-similar Bregman divergence satisfy the (f, γ, δ)-sampling property for $f(\gamma, \delta) = \frac{1}{\mu\gamma\delta}$.

Using all the results mentioned above we get the following Theorem for μ-similar Bregman divergences.

Theorem 7 (k-median w.r.t. μ-similar Bregman divergences). *Let $0 < \mu \leq 1$ and $0 < \epsilon \leq 1/2$. There is an algorithm that runs in time $O\left(nd \cdot 2^{\tilde{O}\left(\frac{k^2}{\mu \cdot \epsilon}\right)}\right)$ and gives a $(1+\epsilon)$-approximation to the k-median objective function w.r.t. μ-similar Bregman divergence for any point set $P \in \mathbb{R}^d, |P| = n$.*

B Proof of Theorem 5

Here we give a proof of Theorem 5. For the proof, we repeat the analysis in Section 3 almost word-by-word. One the main things we will be doing here is replacing all instances of $d(p,q)^2$ in Section 3 with $D(p,q)$. So, this section will look very similar to Section 3. First we will restate Theorem 5.

Theorem 8 (Restatement of Theorem 5). *Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Let $\alpha \geq 0$, $0 < \beta \leq 1$, and $0 < \delta < 1/2$ be constants and let $0 < \epsilon \leq 1/2$. Let $\eta = \frac{2\alpha^2}{\beta^2}(1 + 1/\beta)$. Let D be a distance measure over space $\mathcal{X}$ that D follows:*

1. *β-approximate symmetry property,*
2. *α-approximate triangle inequality,*
3. *Centroid property, and*
4. *(f, ϵ, δ)-sampling property.*

Then there is an algorithm that runs in time $O\left(nd \cdot 2^{\tilde{O}(k \cdot f(\epsilon/\eta k, 0.2))}\right)$ and gives a $(1+\epsilon)$-approximation to the k-median objective for any point set $P \subseteq \mathcal{X}, |P| = n$.

We will first assume that the instance is (k, ϵ)-irreducible for a suitably small parameter ϵ. We shall then get rid of this assumption later as we did in Section 3. The algorithm remains the same and is described in Figure 2.

We develop some notation first. For the rest of the analysis, we will fix a tuple $(s_1, \ldots, s_k)$ – this will be the "desired tuple", i.e., the one for which we can show that the set C gives a good solution. As our analysis proceeds, we will argue what properties this tuple should have. Let $C^{(i)}$ be the set C at the beginning of the i^{th} iteration of Step 2(b). To begin with $C^{(0)}$ is empty. Let $S^{(i)}$ be the set S sampled during the i^{th} iteration of Step 2(b), and $T^{(i)}$ be the corresponding set T (which is the s_i^{th} subset of $S^{(i)}$).

Let $O_1, \ldots, O_k$ be the optimal clusters, and $c_1, ..., c_k$ denote the respective optimal cluster centers. Further, let m_i denote $|O_i|$, and wlog assume that $m_1 \geq \ldots \geq m_k$. Let r_i denote the average cost paid by a point in O_i, i.e.,

$$r_i = \frac{\sum_{p \in O_i} D(p, c_i)}{m_i}.$$

Find-k-median(P)

Let $\eta = \frac{2\alpha^2}{\beta^2}(1+1/\beta)$, $N = \frac{(24\eta\alpha\beta k)\cdot f(\epsilon/\eta, 0.2)}{\epsilon^2}$, $M = f(\epsilon/\eta, 0.2)$, and $P = \binom{N}{M}$

1. **Repeat** 2^k times and output the the set of centers C that give least cost
   2. **Repeat** for all k-tuples $(s_1, ..., s_k) \in [P] \times [P] \times .... \times [P]$ and pick the set of centers C that gives least cost
      (a) $C \leftarrow \{\}$
      (b) For $i \leftarrow 1$ to k
      Sample a set S of N points with D^2-sampling (w.r.t. centers C)
      Let T be the s_i^{th} subset of S. [a]
      $C \leftarrow C \cup \{m(T)\}$. [b]

[a] For a set of size N we consider an arbitrary ordering of the subsets of size M of this set.

[b] $m(T)$ denote the centroid of the points in T.

Fig. 2. The algorithm that gives $(1+\epsilon)$-approximation for any (k, ϵ)-irreducible data set. Note that the inner loop is executed at most $2^k \cdot \left(\binom{N}{M}\right)^k \sim 2^k \cdot 2^{\tilde{O}(k \cdot f(\epsilon/\eta, 0.2))}$ times.

First, we show that any two optimal centers are far enough.

Lemma 10. *For any* $1 \leq i < j \leq k$,

$$D(c_j, c_i) \geq \epsilon \cdot (r_i + r_j).$$

Proof. Since $i < j$, we have $m_i \geq m_j$. For the sake of contradiction assume $D(c_j, c_i) < \epsilon \cdot (r_i + r_j)$. Then we have,

$$\begin{aligned}
\Delta(O_i \cup O_j, \{c_i\}) &= m_i \cdot r_i + m_j \cdot r_j + m_j \cdot D(c_j, c_i) \quad \text{(using Centroid property)} \\
&< m_i \cdot r_i + m_j \cdot r_j + m_j \cdot \epsilon \cdot (r_i + r_j) \\
&\leq (1+\epsilon) \cdot m_i \cdot r_i + (1+\epsilon) \cdot m_j \cdot r_j \quad \text{(since } m_i \geq m_j\text{)} \\
&\leq (1+\epsilon) \cdot \Delta(O_i \cup O_j, \{c_i, c_j\})
\end{aligned}$$

This implies that the centers $\{c_1, ..., c_k\} \backslash \{c_j\}$ give a $(1+\epsilon)$-approximation to the k-median objective. This contradicts the assumption that P is (ϵ, k)-irreducible.

The above lemma gives the following Corollary that we will use in the rest of the proof.

Corollary 2. *For any* $i \neq j$, $D(c_i, c_j) \geq (\beta\epsilon) \cdot (r_i + r_j)$.

Proof. If $i > j$, then we have $D(c_i, c_j) \geq \epsilon \cdot (r_i + r_j)$ from the above lemma and hence $D(c_i, c_j) \geq (\beta\epsilon) \cdot (r_i + r_j)$. In case $i < j$, then the above lemma gives $D(c_j, c_i) \geq \epsilon \cdot (r_i + r_j)$. Using β-approximate symmetry property we get the statement of the corollary.

We give an outline of the proof. Suppose in the first $(i-1)$ iterations, we have found centers which are close to the centers of some $(i-1)$ clusters in the optimal solution. Conditioned on this fact, we show that in the next iteration, we are likely to sample enough number of points from one of the remaining clusters (c.f. Corollary 3). Further, we show that the samples from this new cluster are close to uniform distribution (c.f. Lemma 12). Since such a sample does not come from exactly uniform distribution, we cannot use the (f, γ, δ)-sampling property directly. In fact, dealing with the slight non-uniformity turns out to be non-trivial (c.f. Lemmas 13 and 14).

We now show that the following invariant will hold for all iterations : let $C^{(i-1)}$ consist of centers $c'_1, \ldots, c'_{i-1}$ (added in this order). Then, with probability at least $\frac{1}{2^i}$, there exist distinct indices $j_1, \ldots, j_{i-1}$ such that for all $l = 1, \ldots, i-1$,

$$\Delta(O_{j_l}, c'_l) \leq (1 + \epsilon/\eta) \cdot \Delta(O_{j_l}, c_{j_l}) \tag{6}$$

Where η is a fixed constant that depends on α and β. With foresight, we fix the value of $\eta = \frac{2\alpha^2}{\beta^2} \cdot (1 + 1/\beta)$. Suppose this invariant holds for $C^{(i-1)}$ (the base case is easy since $C^{(0)}$ is empty). We now show that this invariant holds for $C^{(i)}$ as well. In other words, we just show that in the i^{th} iteration, with probability at least $1/2$, the algorithm finds a center c'_i such that

$$\Delta(O_{j_i}, c'_i) \leq (1 + \epsilon/\eta) \cdot \Delta(O_{j_i}, c_{j_i}),$$

where j_i is an index distinct from $\{j_1, \ldots, j_{i-1}\}$. This will basically show that at the end of the last iteration, we will have k centers that give a $(1+\epsilon)$-approximation with probability at least 2^{-k}.

We now show that the invariant holds for $C^{(i)}$. We use the notation developed above for $C^{(i-1)}$. Let I denote the set of indices $\{j_1, \ldots, j_{i-1}\}$. Now let j_i be the index $j \notin I$ for which $\Delta(O_j, C^{(i-1)})$ is maximum. Intuitively, conditioned on sampling from clusters in $O_i, \cdots, O_k$ using D^2-sampling, it is likely that enough points from O_{j_i} will be sampled. The next lemma shows that there is good chance that elements from the sets O_j for $j \notin I$ will be sampled.

Lemma 11.

$$\frac{\sum_{l \notin I} \Delta(O_l, C^{(i-1)})}{\sum_{l=1}^{k} \Delta(O_l, C^{(i-1)})} \geq \epsilon/2.$$

Proof. Suppose, for the sake of contradiction, the above statement does not hold. Then,

$$\begin{aligned}
\Delta(P, C^{(i-1)}) &= \sum_{l \in I} \Delta(O_l, C^{(i-1)}) + \sum_{l \notin I} \Delta(O_l, C^{(i-1)}) \\
&< \sum_{l \in I} \Delta(O_l, C^{(i-1)}) + \frac{\epsilon/2}{1-\epsilon/2} \cdot \sum_{l \in I} \Delta(O_l, C^{(i-1)}) \quad \text{(by our assumption)} \\
&= \frac{1}{1-\epsilon/2} \cdot \sum_{l \in I} \Delta(O_l, C^{(i-1)}) \\
&\leq \frac{1+\epsilon/\eta}{1-\epsilon/2} \cdot \sum_{l \in I} \Delta_1(O_l) \quad \text{(using the invariant for } C^{(i-1)}) \\
&\leq (1+\epsilon) \cdot \sum_{l \in I} \Delta_1(O_l) \quad \text{(using } \eta = (2\alpha^2/\beta^2) \cdot (1+1/\beta) \geq 4) \\
&\leq (1+\epsilon) \cdot \sum_{l \in [k]} \Delta_1(O_l)
\end{aligned}$$

But this contradicts the fact that P is (k, ϵ)-irreducible.

We get the following corollary easily.

Corollary 3.

$$\frac{\Delta(O_{j_i}, C^{(i-1)})}{\sum_{l=1}^{k} \Delta(O_l, C^{(i-1)})} \geq \frac{\epsilon}{2k}.$$

The above Lemma and its Corollary say that with probability at least $\frac{\epsilon}{2k}$, points in the set O_{j_i} will be sampled. However the points within O_{j_i} are not sampled uniformly. Some points in O_{j_i} might be sampled with higher probability than other points. In the next lemma, we show that each point will be sampled with certain minimum probability.

Lemma 12. *For any $l \notin I$ and any point $p \in O_l$,*

$$\frac{D(p, C^{(i-1)})}{\Delta(O_l, C^{(i-1)})} \geq \frac{1}{m_l} \cdot \frac{\epsilon}{3\alpha\beta\eta}.$$

Proof. Fix a point $p \in O_l$. Let $j_t \in I$ be the index such that p is closest to c'_t among all centers in $C^{(i-1)}$. We have

$$\begin{aligned}
\Delta(O_l, C^{(i-1)}) &\leq m_l \cdot r_l + m_l \cdot D(c_l, c'_t) \quad \text{(using Centroid property)} \\
&\leq m_l \cdot r_l + \alpha \cdot m_l \cdot (D(c_l, c_{j_t}) + D(c_{j_t}, c'_t)) \quad \text{(Using triangle inequality)} \\
&\leq m_l \cdot r_l + \alpha \cdot m_l \cdot \left(D(c_l, c_{j_t}) + \frac{\epsilon r_{j_t}}{\eta} \right), \qquad (7)
\end{aligned}$$

where the last inequality follows from the invariant condition for $C^{(i-1)}$. Also, we know that the following inequalities hold:

$$\alpha \cdot (D(p, c'_t) + D(c'_t, c_{j_t})) \geq D(p, c_{j_t}) \quad \text{(from approximate triangle inequality)} \qquad (8)$$

$$\alpha \cdot (D(c_l, p) + D(p, c_{j_t})) \geq D(c_l, c_{j_t}) \quad \text{(from approximate triangle inequality)} \tag{9}$$

$$D(p, c_l) \leq D(p, c_{j_t}) \quad \text{(since } p \in O_l\text{)} \tag{10}$$

$$\beta \cdot D(c_l, p) \leq D(p, c_l) \leq (1/\beta) \cdot D(c_l, p) \quad \text{(from approximate symmetry)} \tag{11}$$

$$D(c_{j_t}, c'_t) \leq (\epsilon/\eta) \cdot r_{j_t} \quad \text{(from invariant condition)} \tag{12}$$

$$\beta \cdot D(c_{j_t}, c'_t) \leq D(c'_t, c_{j_t}) \leq (1/\beta) \cdot D(c_{j_t}, c'_t) \quad \text{(from approximate symmetry)} \tag{13}$$

Inequalities (9), (10), and (11) gives the following:

$$\begin{aligned}
&D(p, c_{j_t}) + D(c_l, p) \geq \frac{D(c_l, c_{j_t})}{\alpha} \\
&\Rightarrow D(p, c_{j_t}) + \frac{D(p, c_l)}{\beta} \geq \frac{D(c_l, c_{j_t})}{\alpha} \quad \text{(using (11))} \\
&\Rightarrow D(p, c_{j_t}) + \frac{D(p, c_{j_t})}{\beta} \geq \frac{D(c_l, c_{j_t})}{\alpha} \quad \text{(using (10))} \\
&\Rightarrow D(p, c_{j_t}) \geq \frac{D(c_l, c_{j_t})}{\alpha(1 + 1/\beta)}
\end{aligned} \tag{14}$$

Using (8) and (14), we get the following:

$$D(p, c'_t) \geq \frac{D(c_l, c_{j_t})}{\alpha^2(1 + 1/\beta)} - D(c'_t, c_{j_t})$$

Using the previous inequality and (13) we get the following:

$$\begin{aligned}
D(p, c'_t) &\geq \frac{D(c_l, c_{j_t})}{\alpha^2(1 + 1/\beta)} - \frac{D(c_{j_t}, c'_t)}{\beta} \\
&\geq \frac{D(c_l, c_{j_t})}{\alpha^2(1 + 1/\beta)} - \frac{\epsilon}{\eta\beta} \cdot r_{j_t} \quad \text{(using the invariant for } C^{(i-1)}\text{)} \\
&\geq \frac{D(c_l, c_{j_t})}{\eta\beta^2} \quad \text{(Using Corollary 2)}
\end{aligned} \tag{15}$$

So, we get

$$\begin{aligned}
\frac{D(p, C^{(i-1)})}{\Delta(O_l, C^{(i-1)})} &\geq \frac{D(c_l, c_{j_t})}{(\eta\beta^2) \cdot m_l \cdot \left(r_l + \alpha\left(D(c_l, c_{j_t}) + \frac{\epsilon r_t}{\eta}\right)\right)} \quad \text{(using (7) and (15))} \\
&\geq \frac{1}{(\eta\beta^2) \cdot m_l} \cdot \frac{1}{1/(\beta\epsilon) + \alpha + 1/(\eta\beta)} \\
&\geq \frac{\epsilon}{(3\eta\alpha\beta)} \cdot \frac{1}{m_l} \quad \text{(using Corollary 2)}
\end{aligned}$$

Recall that $S^{(i)}$ is the sample of size N in this iteration. We would like to show that that the invariant will hold in this iteration as well. We first prove a simple corollary of Lemma 1.

Lemma 13. *Let Q be a set of n points, and γ be a parameter, $0 < \gamma < 1$. Define a random variable X as follows : with probability γ, it picks an element of Q uniformly at random, and with probability $1-\gamma$, it does not pick any element (i.e., is null). Let $X_1, \ldots, X_\ell$ be ℓ independent copies of X, where $\ell = \frac{4}{\gamma} \cdot f(\epsilon/\eta, 0.2)$. Let T denote the (multi-set) of elements of Q picked by $X_1, \ldots, X_\ell$. Then, with probability at least $3/4$, T contains a subset U of size $f(\epsilon/\eta, 0.2)$ which satsifies*

$$\Delta(P, m(U)) \leq \left(1 + \frac{\epsilon}{\eta}\right) \cdot \Delta_1(P) \tag{16}$$

Proof. Define a random variable I, which is a subset of the index set $\{1, \ldots, \ell\}$, as follows $I = \{t : X_t$ picks an element of Q, i.e., it is not null$\}$. Conditioned on $I = \{t_1, \ldots, t_r\}$, note that the random variables $X_{t_1}, \ldots, X_{t_r}$ are independent uniform samples from Q. Thus if $|I| \geq f(\epsilon/\eta, 0.2)$, then sampling property wrt. D implies that with probability at least 0.8, the desired event (16) happens. But the expected value of $|I|$ is $4 \cdot f(\epsilon/\eta, 0.2)$, and so, $|I| \geq f(\epsilon/\eta, 0.2)$ with high probability, and hence, the statement in the lemma is true.

We are now ready to prove the main lemma.

Lemma 14. *With probability at least $1/2$, there exists a subset $T^{(i)}$ of $S^{(i)}$ of size at most $f(\epsilon/\eta, 0.2)$ such that*

$$\Delta(O_{j_i}, m(T^{(i)})) \leq \left(1 + \frac{\epsilon}{\eta}\right) \cdot \Delta_1(O_{j_i}).$$

Proof. Recall that $S^{(i)}$ contains $N = \frac{(24\eta\alpha\beta k) \cdot f(\epsilon/\eta, 0.2)}{\epsilon^2}$ independent samples of P (using D^2-sampling). We are interested in $S^{(i)} \cap O_{j_i}$. Let $Y_1, \ldots, Y_N$ be N independent random variables defined as follows : for any t, $1 \leq t \leq N$, Y_t picks an element of P using D^2-sampling with respect to $C^{(i-1)}$. If this element is not in O_{j_i}, it just discards it (i.e., Y_t is null). Let γ denote $\frac{\epsilon^2}{6\eta\alpha\beta k}$. Corollary 3 and Lemma 12 imply that Y_t picks a particular element of O_{j_i} with probability at least $\frac{\gamma}{m_{j_i}}$. We would now like to apply Lemma 13 (observe that $N = \frac{4}{\gamma} \cdot f(\epsilon/\eta, 0.2)$). We can do this by a simple coupling argument as follows. For a particular element $p \in O_{j_i}$, suppose Y_t assigns probability $\frac{\gamma(p)}{m_{j_i}}$ to it. One way of sampling a random variable X_t as in Lemma 13 is as follows – first sample using Y_t. If Y_t is null, then X_t is also null. Otherwise, suppose Y_t picks an element p of O_{j_i}. Then X_t is equal to p with probability $\frac{\gamma}{\gamma(p)}$, and null otherwise. It is easy to check that with probability γ, X_t is a uniform sample from O_{j_i}, and null with probability $1 - \gamma$. Now, observe that the set of elements of O_{j_i} sampled by $Y_1, \ldots, Y_N$ is always a superset of $X_1, \ldots, X_N$. We can now use Lemma 13 to finish the proof.

Thus, we will take the index s_i in Step 2(i) as the index of the set $T^{(i)}$ as guaranteed by the Lemma above. Finally, by repeating the entire process 2^k times, we make sure that we get a $(1+\epsilon)$-approximate solution with high probability. Note that the total running time of our algorithm is $\left(nd \cdot 2^k \cdot 2^{\tilde{O}(k \cdot f(\epsilon/\eta, 0.2))}\right)$.

Removing the (k, ϵ)-irreducibility assumption : We now show how to remove this assumption. First note that we have shown the following result.

Theorem 9. *If a given point set $(k, \frac{\epsilon}{(1+\epsilon/2)\cdot k})$-irreducible, then there is an algorithm that gives a $(1+\frac{\epsilon}{(1+\epsilon/2)\cdot k})$-approximation to the k-median objective with respect to distance measure D and that runs in time $O(nd \cdot 2^{\tilde{O}(k\cdot f(\epsilon/k\eta, 0.2))})$.*

Proof. The proof can be obtained by replacing ϵ by $\frac{\epsilon}{(1+\epsilon/2)\cdot k}$ in the above analysis.

Suppose the point set P is not $(k, \frac{\epsilon}{(1+\epsilon/2)\cdot k})$-irreducible. In that case it will be sufficient to find fewer centers that $(1+\epsilon)$-approximate the k-median objective. The next lemma shows this more formally.

Theorem 10. *There is an algorithm that runs in time $O(nd \cdot 2^{\tilde{O}(k\cdot f(\epsilon/\eta k, 0.2))})$ and gives a $(1+\epsilon)$-approximation to the k-median objective with respect to D.*

Proof. Let P denote the set of points. Let $1 < j \leq k$ be the largest index such that P is $(i, \frac{\epsilon}{(1+\epsilon/2)\cdot k})$-irreducible. If no such i exists, then

$$\Delta_1(P) \leq \left(1 + \frac{\epsilon}{(1+\epsilon/2)\cdot k}\right)^k \cdot \Delta_k(P) \leq (1+\epsilon)\cdot \Delta_k(P),$$

and so picking the centroid of P will give a $(1+\epsilon)$-approximation.

Suppose such an i exists. In that case, we consider the i-median problem and from the previous lemma we get that there is an algorithm that runs in time $O(nd \cdot 2^i \cdot 2^{\tilde{O}(i\cdot f(\epsilon/\eta k, 0.2))})$ and gives a $(1+\frac{\epsilon}{(1+\epsilon/2)\cdot k})$-approximation to the i-median objective. Now we have that

$$\Delta_i \leq \left(1 + \frac{\epsilon}{(1+\epsilon/2)\cdot k}\right)^{k-i} \cdot \Delta_k \leq (1+\epsilon)\cdot \Delta_k.$$

Thus, we are done.

C Proof of Lemma 9

Here we give the proof of Lemma 9. For better readability, we first restate the Lemma.

Lemma 15 (Restatement of Lemma 9). *Let $0 < \mu \leq 1$. Any μ-similar Bregman divergence satisfies the μ-approximate symmetry property and $(2/\mu)$-approximate triangle inequality.*

The above lemma follows from the next lwo sub-lemmas.

Lemma 16 (Symmetry for μ-similar Bregman divergence). *Let $0 < \mu \leq 1$. Consider a μ-similar Bregman divergence D_ϕ on domain $\mathbb{X} \subseteq \mathbb{R}^d$. For any two points $p, q \in \mathbb{X}$, we have: $\mu \cdot D_\phi(q,p) \leq D_\phi(p,q) \leq \frac{1}{\mu}\cdot D_\phi(q,p)$*

Proof. Using equation(5) we get the following:

$$\mu \cdot \mathbf{D}_\phi(\mathbf{q},\mathbf{p}) \leq \mu \cdot D_U(q,p) = \mu \cdot D_U(p,q) \leq \mathbf{D}_\phi(\mathbf{p},\mathbf{q}) \leq D_U(p,q) = D_U(q,p) \leq \frac{\mathbf{1}}{\mu}\cdot \mathbf{D}_\phi(\mathbf{q},\mathbf{p}).$$

Lemma 17 (Triangle inequality for μ-similar Bregman divergence). *Let $0 < \mu \leq 1$. Consider a μ-similar Bregman divergence D_ϕ on domain $\mathbb{X} \subseteq \mathbb{R}^d$. For any three points $p, q, r \in \mathbb{X}$, we have: $(\mu/2) \cdot D_\phi(p, r) \leq D_\phi(p, q) + D_\phi(q, r)$*

Proof. We have:

$$\begin{aligned} D_\phi(p, q) + D_\phi(q, r) &\geq \mu \cdot (D_U(p, q) + D_U(q, r)) \\ &\geq (\mu/2) \cdot D_U(p, r) \\ &\geq (\mu/2) \cdot D_\phi(p, r) \end{aligned}$$

The first and third inequality is using equation 5 and the second inequality is using the approximate triangle inequality for Mahalanobis distance.