Clustering for Metric and Non-Metric Distance Measures[*]

Marcel R. Ackermann[†] Johannes Blömer[†] Christian Sohler[†]

(revised version of February 4, 2008)

Abstract

We study a generalization of the k-median problem with respect to an arbitrary dissimilarity measure D. Given a finite set P, our goal is to find a set C of size k such that the sum of errors $\mathrm{D}(P, C) = \sum_{p \in P} \min_{c \in C} \{\mathrm{D}(p, c)\}$ is minimized. The main result in this paper can be stated as follows: There exists an $\mathcal{O}\big(n 2^{(\frac{k}{\epsilon})^{\mathcal{O}(1)}}\big)$ time $(1+\epsilon)$-approximation algorithm for the k-median problem with respect to D, if the 1-median problem can be approximated within a factor of $(1+\epsilon)$ by taking a random sample of constant size and solving the 1-median problem on the sample exactly. Using this characterization, we obtain the first linear time $(1+\epsilon)$-approximation algorithms for the k-median problem in an arbitrary metric space with bounded doubling dimension, for the Kullback-Leibler divergence (relative entropy), for Mahalanobis distances, and for some special cases of Bregman divergences. Moreover, we obtain previously known results for the Euclidean k-median problem and the Euclidean k-means problem in a simplified manner. Our results are based on a new analysis of an algorithm from [20].

1 Introduction

Clustering is the (meta-)problem of partitioning a given set of objects into subsets of similar objects. It has application in various areas of computer science such as machine learning, data compression, data mining, or pattern recognition. Depending on the application we want to cluster such diverse objects as text documents, probability distributions, feature vectors, etc. Obviously, different objects and different applications also require different notions of (dis-)similarity of objects. As a consequence, there are numerous different formulations of clustering. In theoretical computer science many approximation algorithms have been developed for variants of clustering, where the points come from a metric space. However, for non-metric dissimilarity measures almost no approximation algorithms are known.

This stands in sharp contrast to the fact that we have non-metric dissimilarity measures in many applications. Text documents are often clustered using the cosine-similarity measure on TFIDF vectors (term frequency inverse document frequency), in bioinformatics Pearson correlation is used frequently, and in statistics Bregman divergences are to be minimized, e.g. the Kullback-Leibler divergence (relative entropy).

In fact, our work started with an industrial project on lossless compression of Java and C++ executable code. For lossless compression achieving a significant compression ratio we had to use statistical models of the executable code based on a large number of probability distributions. The space required to store these distributions threatened to outweigh any savings achieved by compressing the executable code. To solve this problem, rather than storing all distributions of our statistical models we tried to identify a good set of representatives for these distributions. Replacing one distribution in a statistical model by some representative incurs a certain loss in compression that is well approximated by the Kullback-Leibler divergence between the two distributions. Hence computing a set of representatives for distributions in a statistical model for executable code compression immediately leads to a clustering problem involving the Kullback-Leibler divergence.

In this paper we do a first step towards understanding clustering problems with non-metric dissimilarity measures, like Kullback-Leibler divergence. We consider a problem that is relatively well understood in the case of Euclidean and metric distances: k-median clustering. In k-median clustering we have a representative (sometimes called prototype) for each cluster. In the geometric version of the problem this is the cluster center. We are interested in minimizing the sum of error of the clustering, i.e. the error that is made by representing each input object by its corresponding representative. Since we allow non-metric dissimilarity measures, this version of k-median also captures other variants like the well-known Euclidean k-means clustering, where the goal is to minimize the sum of squared errors (with respect to Euclidean distance).

[*] This research was supported by Deutsche Forschungsgemeinschaft (DFG), grant BL 314/6-1.

[†] University of Paderborn, Dept. o. Comp. Science, D-33095 Paderborn, Germany. eMail: {mra,bloemer,csohler}@upb.de

1.1 Related work.

There has been a rich research on clustering during the recent decades. For an overview see [16], [30] or [25].

It is known that in arbitrary metric spaces the k-median clustering problem is $\mathcal{NP}$-hard. It is even $\mathcal{NP}$-hard to approximate within a factor of 1.73 (see [17]). Several constant factor polynomial time approximation algorithms are known. The fastest known algorithm is due to Mettu and Plaxton [26]. In the case of the d-dimensional Euclidean space, Kolliopoulus and Rao [18] stated a $(1+\epsilon)$-approximation algorithm with running time $\mathcal{O}\big(\rho_{\epsilon,d}\, n \log k \log n\big)$ where $\rho_{\epsilon,d} = 2^{\mathcal{O}(1+\frac{1}{\epsilon}\log\frac{1}{\epsilon})^{d-1})}$. This was improved to $\mathcal{O}\big(n + \rho_{\epsilon,d}\, k^{\mathcal{O}(1)} \log^{\mathcal{O}(1)} n\big)$ by Har-Peled and Mazumdar [13] by using coresets. Badiou et al. [1] proposed an algorithm with running time $\mathcal{O}\big(d^{\mathcal{O}(1)} n \log^{\mathcal{O}(k)} n 2^{(\frac{k}{\epsilon})^{\mathcal{O}(1)}}\big)$.

Considering the Euclidean k-means problem, the fastest known algorithm solving the problem exactly in time $\mathcal{O}(n^{dk+1})$ was given by Inaba et al. [14]. A widely used heuristic for the k-means problem is Lloyd's algorithm [22]. Matousek [24] provided a deterministic $(1+\epsilon)$-approximation algorithm with running time $\mathcal{O}\big(n \log^k n\, (\frac{1}{\epsilon})^{2dk^2}\big)$. Har-Peled and Mazumdar [13] improved the running time to $\mathcal{O}\big(n + (\frac{1}{\epsilon})^{2d+1} k^{k+2} \log^{k+1} n \log^k \frac{1}{\epsilon}\big)$. For the high dimensional case, Fernandez de la Vega et al. [11] achieved a $(1+\epsilon)$-approximation in time $\mathcal{O}\big(d^{\mathcal{O}(1)} n \log^{\mathcal{O}(k)} n 2^{(\frac{k}{\epsilon})^{\mathcal{O}(1)}}\big)$. Recently, Feldman et al [10] provided a PTAS with running time $\mathcal{O}\big(ndk + d(\frac{k}{\epsilon})^{\mathcal{O}(1)} + 2^{\tilde{\mathcal{O}}(\frac{k}{\epsilon})}\big)$.

Of the recent work, the following approach is of particular interest to our work. Kumar et al. [20] proposed a simple $\mathcal{O}\big(dn 2^{(\frac{k}{\epsilon})^{\mathcal{O}(1)}}\big)$-time $(1+\epsilon)$-approximation algorithm for the Euclidean k-means problem. This approach was later generalized to certain Euclidean clustering problems such as the Euclidean k-median problem (see [21]). Chen [7] combined the algorithm from [20] with a new coreset construction and improved the running time to $\mathcal{O}\big(ndk + d^2 n^{\sigma} 2^{(\frac{k}{\epsilon})^{\mathcal{O}(1)}}\big)$ for arbitrary $\sigma > 0$.

Methods for k-median clustering by Kullback-Leibler divergence were first suggested by Pereira et al. [27]. Baker and McCallum [3] proposed a simple agglomerative greedy strategy which turns out to perform surprisingly well in empirical tests. Independently, Slonin and Tishby [28] stated a similar algorithm. Dhillon et al. [9] proposed a local improvement heuristic for clustering by Kullback-Leibler divergence which is an adaptation of Lloyd's k-means algorithm. Banerjee et al. [4] generalized this approach to the class of all Bregman divergences. All these recent strategies have in common that they lack any provable approximation ratio and rely on empirical evaluation.

1.2 The generalized k-median problem.

In this section we introduce our formulation of the k-median problem, which captures a number of well-known clustering problems. We are given an arbitrary ground set Δ of possible objects. On this ground set a *dissimilarity measure* $\mathrm{D} : \Delta \times \Delta \to \mathbb{R}_+$ is defined, that specifies the dissimilarity between two objects from the ground set. We make no assumption on the dissimilarity measure other than $\mathrm{D}(x,y) = 0$ if and only if $x = y$. For $P \subseteq \Delta$ and $c \in \Delta$ we also use $\mathrm{D}(P,c) = \sum_{p \in P} \mathrm{D}(p,c)$ for simplicity of notation.

The k-median problem is defined as follows. We are given a finite set $P \subseteq \Delta$ of objects from the ground set. Our goal is to find a set $C \subseteq \Delta$ of size k such that the sum of errors

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

is minimized. We denote the cost of such an optimal solution by $opt_k(P)$. The elements of C are called k-medians of P.

1.3 Our results.

The main result in this paper can be roughly stated as follows: Every dissimilarity measure has a linear time $(1+\epsilon)$-approximation algorithm for the k-median problem, if the 1-median problem can be approximated within a factor of $(1+\epsilon)$ by taking a random sample of constant size and solving the 1-median problem on the sample exactly. Later, we apply this result to obtain a number of new clustering algorithms for metric and non-metric distance measures. Stated in detail, the result is as follows.

Theorem 1.1. *Given $k \in \mathbb{N}$ and $\epsilon < 1$. Assume that for $\delta < 1$ and $\gamma = \frac{\epsilon}{3}$ dissimilarity measure D satisfies:*

a) For every finite subset $S \subseteq \Delta$ an optimal 1-median $\Gamma(S) \in \Delta$, i.e. $\mathrm{D}\big(S, \Gamma(S)\big) = opt_1(S)$, can be computed in time depending only on $|S|$.

b) There exists a constant $m_{\gamma,\delta} \in \mathbb{N}$ such that for every subset $P \subseteq \Delta$ of size n and for every uniform sample multiset $S \subseteq P$ of size $m_{\gamma,\delta}$ an optimal 1-median $\Gamma(S) \in \Delta$ satisfies

$$\Pr\big[\mathrm{D}\big(P, \Gamma(S)\big) \le (1+\gamma)\, opt_1(P)\big] \ge 1 - \delta.$$

Then there exists an algorithm that with constant probability returns a $(1+\epsilon)$-approximation of the k-median problem with respect to D for input instance P of size n in time $\mathcal{O}\big(n 2^{(\frac{k}{\epsilon})^{\mathcal{O}(1)}}\big)$.

We call conditions (a) and (b) the $[\gamma, \delta]$-*sampling property.* Property (a) captures mainly the fact that the problem is well-posed, i.e. a k-median is computable, etc. Property (b) requires that a solution for a constant size uniform sample is a $(1+\epsilon)$-approximation for the 1-median problem. However, for some dissimilarity measures like the ℓ_2-norm on $\mathbb{R}^d$, property (a) is not satisfied, because one cannot compute an exact solution to the Euclidean 1-median problem. To deal with these problems we generalize our property in Section 4. Even under this weaker restriction we obtain a $(1+\epsilon)$-approximation algorithm of the k-median problem.

1.4 New clustering algorithms.

Using our main result, we obtain the first linear time $(1+\epsilon)$-approximation algorithms for the following problems:

- k-median in an arbitrary metric space (X, D) with bounded doubling dimension that satisfies condition (a).
- k-median of probability distributions with respect to the Kullback-Leibler divergence (relative entropy), under the assumption that for all distributions every element has at least some (small constant) probability. This is the first approximation algorithm for k-median under the Kullback-Leibler divergence that provides *any* non-trivial approximation ratio.
- k-median in $\mathbb{R}^d$ with respect to Mahalanobis distances and some special cases of Bregman divergences that can be reduced to a Mahalanobis distance.

1.5 Our techniques.

To obtain our main result we describe a generalized and improved analysis of algorithm Irred-k-means from [20]. This algorithm has already been generalized to other clustering problems. In [21], sufficient conditions for dissimilarity measures in the $\mathbb{R}^d$ that allow for the application of the algorithm from [20] have been given. However, symmetry or the triangle inequality are always assumed. Our generalization does not require these assumptions. Therefore, we can obtain results for non-metric dissimilarity measures like the Kullback-Leibler divergence, which does not seem to be possible using previous results.

Our new approach does not only generalize to non-metric dissimilarity measures, it can also be used to obtain the results for the Euclidean k-median and the Euclidean k-means problem from [20, 21]. Moreover, these results can be obtained by a significantly simplified analysis.

To obtain our results for specific dissimilarity measures like the Kullback-Leibler divergence, Mahalanobis distances, etc. we show that the 1-median of a constant sized uniform sample set $S \subseteq P$ is an approximate 1-median of P. These sampling results are the second main contribution of this paper.

1.6 Organization.

In Section 2 we present and analyze a $(1+\epsilon)$-approximation algorithm for the k-median problem with respect to D satisfying the $[\gamma, \delta]$-sampling property. In Section 3 we give several examples of dissimilarity measures satisfying this property. We discuss a generalization of the $[\gamma, \delta]$-sampling property in Section 4.

2 Algorithm for $[\gamma, \delta]$-sampleable dissimilarity measures

In this section we describe and analyze our main algorithm.

2.1 Superset sampling.

The algorithm uses the superset sampling technique due to [20]. The technique is an immediate consequence of probabilistic concentration bounds. In the sequel, $m_{\gamma,\delta}$ and Γ are as described in Theorem 1.1.

Lemma 2.1. (superset sampling technique) *Let* D *satisfy the* $[\gamma, \delta]$*-sampling property. Let be* $P \subseteq \Delta$ *of size* n *and* $P' \subseteq P$ *with* $|P'| \geq \alpha n$*. Let be* $S \subseteq P$ *an uniform sample multiset of size at least* $\frac{2}{\alpha} m_{\gamma,\delta}$*. With probability at least* $\frac{1-\delta}{5}$ *there exists a subset* $S' \subseteq S$ *with* $|S'| = m_{\gamma,\delta}$ *and optimal 1-median* $\Gamma(S')$ *satisfying*

$$\mathrm{D}\big(P', \Gamma(S')\big) \leq (1+\gamma)\, opt_1(P') \ .$$

For a sample set $S \subseteq P$ of size $\frac{2}{\alpha} m_{\gamma,\delta}$ we set $T = \big\{\Gamma(S') \mid S' \subset S,\ |S'| = m_{\gamma,\delta}\big\}$. As an immediate consequence of the lemma, for any fixed subset $P' \subseteq P$ with $|P'| \geq \alpha|P|$ with constant probability T contains an approximate 1-median of P'. We call the elements of T candidates for approximate medians.

2.2 The algorithm.

First, we explain an idealized version of algorithm Irred-k-means for the case of $k = 2$. Let P_1 and P_2, denote the clusters of an optimal 2-median clustering of input set P with $|P_1| \geq \alpha|P|$. Here $\alpha < \frac{1}{4}$ is a parameter to be specified later. Our strategy can be stated as follows.

1. Use the superset sampling technique to obtain $\tilde{c}_1$ from P with $\mathrm{D}(P_1, \tilde{c}_1) \leq (1+\gamma)\, opt_1(P_1)$.
2. Let N be the smallest subset of the closest points from P towards $\tilde{c}_1$ such that for the remaining points $R = P \setminus N$ holds $|P_2 \cap R| \geq \alpha|R|$. Assign N to $\tilde{c}_1$.

3. Use superset sampling technique again to obtain $\tilde{c}_2$ from R with $\mathrm{D}(P_2 \cap R, \tilde{c}_2) \leq (1+\gamma)\, opt_1(P_2 \cap R)$.
4. Assign all remaining points to their closest approximate median and return $\{\tilde{c}_1, \tilde{c}_2\}$ as $(1+\gamma)$-approximate solution.

This idealized strategy faces two problems. First, using the superset sampling technique we do not get a single approximate median $\tilde{c}_1$. Instead we get set T of candidates for approximate medians. To solve this problem we simply try all possible candidates as approximate median $\tilde{c}_1$ and choose the candidate which leads to minimal cost. Note that for constant α and $m_{\gamma,\delta}$ the candidate set is also of constant size $2^{\frac{2}{\alpha} m_{\gamma,\delta}}$. The same procedure is used for obtaining $\tilde{c}_2$ in step 3.

Second, it is obvious that we do not know the optimal clusters P_1 and P_2. Thus, we do not know how to choose N from step 2 explicitly. To cope with this problem we approximate N by partitioning P into subsets $N^{(1)}, N^{(2)}, \ldots, N^{(\lceil \log n \rceil)}$. Here, $N^{(1)}$ denotes the $\frac{n}{2}$ closest points towards $\tilde{c}_1$, $N^{(2)}$ the next $\frac{n}{4}$ closest points, $N^{(3)}$ the next $\frac{n}{8}$ closest points, and so on. Let be $R^{(j)} = P \setminus \bigcup_{i=1}^{j} N^{(i)}$ and let ν be the minimal value such that $|P_2 \cap R^{(\nu)}| \geq \alpha |R^{(\nu)}|$. Instead of N we will assign the points from $N^{(1)}, N^{(2)}, \ldots, N^{(\nu)}$ to $\tilde{c}_1$.

We still do not know the value ν. However, we can guess ν by trying all possible values and choosing the value that leads to minimal cost.

Algorithm Cluster in Figure 1 gives a precise recursive definition of the aforementioned strategy for arbitrary k. The algorithm alternates between two different types of phases: A *sampling phase* and a *pruning phase*. In the sampling phase new candidates for approximate medians are computed (according to Lemma 2.1) and tried in a recursive manner. In the pruning phase new values for ν are tried: The next $\frac{1}{2}|R|$ closest points to already computed approximate medians are assigned and discarded from future consideration. Hence, the algorithm computes a set of possible solutions to the k-median problem for all candidates of $\tilde{c}_i$ and all values of ν. In the final step the algorithm chooses the best solution found this way.

2.3 Comparison of our contribution and [20]. Algorithm Cluster is identical to algorithm Irred-k-means from [20]. However, our interpretation and analysis of the algorithm differ significantly from [20].

As above, we cosider the case $k = 2$ and denote by P_1 and P_2 the optimal clusters. Given an approximate median $\tilde{c}_1$ of cluster P_1, the goal of the analysis in [20] is to show that in the pruning phase only points from P_1 will be assigned to $\tilde{c}_1$. In our analysis, the goal is to show that most points assigned to $\tilde{c}_1$ come from P_1.

```
CLUSTER($R, l, \tilde{C}$):
  $R$   set of remaining input points
  $l$   number of medians yet to be found
  $\tilde{C}$   set of medians already found

if $l = 0$ then return $\tilde{C}$
else
    if $l \geq |R|$ then return $\tilde{C} \cup R$
    else
        /* sampling phase */
        sample a multiset $S$ of size $\frac{2}{\alpha} m_{\gamma,\delta}$ from $R$
        $T \leftarrow \{\Gamma(S') \mid S' \subset S,\ |S'| = m_{\gamma,\delta}\}$
        for all $\tilde{c} \in T$ do
            $C^{(\tilde{c})} \leftarrow$ CLUSTER$(R, l-1, \tilde{C} \cup \{\tilde{c}\})$
        end for
        /* pruning phase */
        $N \leftarrow \{\frac{1}{2}|R|$ minimal $p \in R$ w.r.t. $\mathrm{D}(p, \tilde{C})\}$
        $C^* \leftarrow$ CLUSTER$(R \setminus N, l, \tilde{C})$
        return $C^{(\tilde{c})}$ or $C^*$ with minimum cost
    end if
end if
```

Figure 1: Clustering algorithm for arbitrary k and fixed positive real constants α, γ, δ

To achieve their goal, the analysis in [20] relies heavily on the symmetry and the triangle inequality for the Euclidean distance. Furthermore, the notion of irreducibility is fundamental for the analysis of [20].

On the other hand, in the analyis given below we explicitly allow that also points from P_2 are assigned to $\tilde{c}_1$. We use the parameter α in algorithm Cluster to control the number of points that are incorrectly assigned. By choosing α small enough and the sample size large enough we are able to bound the approximation factor by $1 + \epsilon$. This approach is purely combinatorial and, hence, requires only minimal assumptions on the dissimilarity meassure. In particular, symmetry, triangle inequality, and the notion of irreducible clusterings are no longer needed.

Note that in [21] the methods from [20] have been applied in a more general setting. However, the results in [21] come not close to our main result as stated in Theorem 1.1.

2.4 Analysis for $k = 2$. To simplify notation, first we analyze algorithm Cluster for the case of $k = 2$. In the sequel, let D satisfy the $[\gamma, \delta]$-sampling property.

Theorem 2.1. *Let $\alpha < \frac{1}{4}$ be an arbitrary positive constant. The algorithm* Cluster *started with parameters $(P, 2, \emptyset)$ computes a solution $\tilde{C}$ of the 2-median problem* *for input instance P of size n satisfying*

$$\Pr\left[\mathrm{D}(P,\tilde{C}) \leq (1+8\alpha)(1+\gamma)\, opt_2(P)\right] \geq \left(\frac{1-\delta}{5}\right)^2,$$

using $\mathcal{O}\left(n2^{(\frac{2}{\alpha}m_{\gamma,\delta})^{\mathcal{O}(1)}}\right)$ *arithmetic operations.*

Proof of Theorem 2.1. Assume for simplicity of notation that n is a power of 2. Further, let P_1 and P_2 denote the partition of P of the optimal 2-clustering with the optimal set of medians $C = \{c_1, c_2\}$, i.e. $D(P,C) = opt_k(P)$ and $D(P_i, c_i) = opt_1(P_i)$. Assume $|P_1| \geq \frac{1}{2}|P| > \alpha|P|$.

Denote by T_1 the candidate set from step 7 during the initial call of the algorithm. By Lemma 2.1 with probability at least $\frac{1-\delta}{5}$ set T_1 contains a $\tilde{c}_1$ with $\mathrm{D}(P_1,\tilde{c}_1) \leq (1+\gamma)\,\mathrm{D}(P_1,c_1)$.

We consider two cases. First, we assume there exists a recursive call with parameters $\big(R, 1, \{\tilde{c}_1\}\big)$ such that $|P_2 \cap R| \geq \alpha|R|$. Later we consider the case when there is no such recursive call.

So let us assume there exists a recursive call with $|P_2 \cap R| \geq \alpha|R|$. Let R be the largest input set with that property. Let T_2 be the candidate set from step 7 of this call. Again by Lemma 2.1 with probability $\frac{1-\delta}{5}$ set T_2 contains a $\tilde{c}_2$ satisfying $\mathrm{D}(P_2 \cap R, \tilde{c}_2) \leq (1+\gamma)\,\mathrm{D}(P_2 \cap R, c_2')$. Here c_2' is the optimal 1-median of $P_2 \cap R$, i.e. $D(P_2 \cap R, c_2') = opt_1(P_2 \cap R)$. Thus, $\tilde{C} = \{\tilde{c}_1, \tilde{c}_2\}$ yields an upper bound $D(P,\tilde{C})$ on the cost of the solution returned by algorithm Cluster.

Define $N = P \setminus R$. Using

$$\mathrm{D}(P,\tilde{C}) \leq \mathrm{D}(P_1,\tilde{c}_1) + \mathrm{D}(P_2 \cap N, \tilde{c}_1) + \mathrm{D}(P_2 \setminus N, \tilde{c}_2)$$

we bound each term individually. Using Claim 2.1 and 2.2 we conclude

$$\begin{aligned} &\mathrm{D}(P,\tilde{C}) \leq (1+8\alpha)\,\mathrm{D}(P_1,\tilde{c}_1) + \mathrm{D}(P_2 \setminus N, \tilde{c}_2) \\ &\leq (1+8\alpha)(1+\gamma)\,\mathrm{D}(P_1,c_1) + (1+\gamma)\,\mathrm{D}(P_2 \setminus N, c_2) \\ &\leq (1+8\alpha)(1+\gamma)\, opt_2(P)\ . \end{aligned}$$

Claim 2.1. $\mathrm{D}(P_2 \cap N, \tilde{c}_1) \leq 8\alpha D(P_1,\tilde{c}_1)$.

Proof. Assume $N \neq \emptyset$, otherwise the claim is trivially true. Hence, N is the disjoint union of ν different subsets $N^{(j)}$ of size $\frac{n}{2^j}$ which correspond to the neighborhoods removed in step 12 of the algorithm, i.e. $N = N^{(1)} \cup N^{(2)} \cup \ldots \cup N^{(\nu)}$. We show that for each j the set $N^{(j)}$ contains a large number of points from P_1 and only few points from P_2. Define $R^{(0)} = P$ and $R^{(j)} = R^{(j-1)} \setminus N^{(j)}$. By definition $|R^{(j)}| = |N^{(j)}| = \frac{n}{2^j}$. Note that the $R^{(j)}$ have been input sets of recursive calls prior to the call on $R = R^{(\nu)}$, and hence $|P_2 \cap R^{(j)}| < \alpha|R^{(j)}|$ for all $j < \nu$. We obtain

$$\begin{aligned} (2.1)\quad & \forall j \leq \nu : \\ & |P_2 \cap N^{(j)}| \leq |P_2 \cap R^{(j-1)}| < \alpha|R^{(j-1)}| = 2\alpha\frac{n}{2^j} \end{aligned}$$

where the first inequality holds since $N^{(j)} \subseteq R^{(j-1)}$. We also get

$$\begin{aligned} (2.2)\quad & \forall j \leq \nu : \\ & |P_1 \cap N^{(j)}| = |N^{(j)}| - |P_2 \cap N^{(j)}| \geq (1-2\alpha)\frac{n}{2^j}. \end{aligned}$$

By definition of $N^{(j)}$ we know that for all $j < \nu$ and for $p \in N^{(j)}$ and $p' \in N^{(j+1)}$ we have $\mathrm{D}(p,\tilde{c}_1) \leq \mathrm{D}(p',\tilde{c}_1)$. Using (2.1) and (2.2) we get

$$\begin{aligned} (2.3)\quad & \forall j < \nu : \\ & \mathrm{D}(P_2 \cap N^{(j)}, \tilde{c}_1) \leq \frac{4\alpha}{1-2\alpha}\,\mathrm{D}(P_1 \cap N^{(j+1)}, \tilde{c}_1). \end{aligned}$$

We still need an upper bound on $\mathrm{D}(P_2 \cap N^{(\nu)}, \tilde{c}_1)$. Considering $P_1 \cap R^{(\nu)}$ we obtain

$$\begin{aligned} (2.4)\quad |P_1 \cap R^{(\nu)}| &= |R^{(\nu)}| - |P_2 \cap R^{(\nu)}| \\ &\geq |R^{(\nu)}| - |P_2 \cap R^{(\nu-1)}| > (1-2\alpha)\frac{n}{2^\nu}\ . \end{aligned}$$

By definition of $N^{(\nu)}$ and $R^{(\nu)}$ we also know that for all $p \in N^{(\nu)}$ and $p' \in R^{(\nu)}$ we have $\mathrm{D}(p,\tilde{c}_1) \leq \mathrm{D}(p',\tilde{c}_1)$. Analogously to (2.3), combining (2.1) and (2.4) we conclude

$$(2.5)\qquad \mathrm{D}(P_2 \cap N^{(\nu)}, \tilde{c}_1) \leq \frac{2\alpha}{1-2\alpha}\,\mathrm{D}(P_1 \cap R^{(\nu)}, \tilde{c}_1).$$

Using (2.3) and (2.5) we obtain

$$\begin{aligned} &\mathrm{D}(P_2 \cap N, \tilde{c}_1) = \sum_{j=1}^{\nu} \mathrm{D}(P_2 \cap N^{(j)}, \tilde{c}_1) \\ &\leq 8\alpha \sum_{j=1}^{\nu-1} \mathrm{D}(P_1 \cap N^{(j+1)}, \tilde{c}_1) + 8\alpha\,\mathrm{D}(P_1 \cap R^{(\nu)}, \tilde{c}_1) \\ &\leq 8\alpha\,\mathrm{D}(P_1,\tilde{c}_1) \end{aligned}$$

since $\frac{2\alpha}{1-2\alpha} \leq \frac{4\alpha}{1-2\alpha} \leq 8\alpha$ for $\alpha \leq \frac{1}{4}$. □

Claim 2.2. $\mathrm{D}(P_2 \setminus N, \tilde{c}_2) \leq (1+\gamma)D(P_2 \setminus N, c_2)$.

Proof. By choice of $\tilde{c}_2$, for the optimal 1-median c_2' of $P_2 \setminus N$ we get $\mathrm{D}(P_2 \setminus N, \tilde{c}_2) \leq (1+\gamma)\,\mathrm{D}(P_2 \setminus N, c_2')$. Due to the optimality of c_2' the claim follows. □

Proof of Theorem 2.1 (continued). Finally, we consider the case when there has not been a recursive call on an input set R with $|P_2 \cap R| \geq \alpha|R|$. In this case there is a sequence of recursive calls consecutively using step 13 $\nu = \lceil \log n \rceil$ times. We end up with a single point $q \in R$. This q can be given its own cluster with median $\tilde{c}_2 = q$. This is not contributing any cost to $\mathrm{D}(P, \{\tilde{c}_1, \tilde{c}_2\})$. The cost of $\mathrm{D}(P_2 \cap N, \tilde{c}_1)$ with $N = \bigcup_{j=1}^{\log n} N^{(j)}$ is still bounded as given above. This concludes the proof.

The running time analysis of [20] can easily be adopted to algorithm Cluster. This shows that the running time of the algorithm is linear in n. □

2.5 Analysis for $k > 2$. We can generalize the analysis of algorithm Cluster to the case of $k > 2$, leading to the following theorem for D satisfying the $[\gamma, \delta]$-sampling property.

Theorem 2.2. *Let $\alpha < \frac{1}{4k}$ be an arbitrary positive constant. The algorithm* Cluster *started with parameters $(P, k, \emptyset)$ computes a solution $\tilde{C}$ of the k-median problem for input instance P of size n such that*

$$\Pr\left[\mathrm{D}(P, \tilde{C}) \leq (1 + 8\alpha k^2)(1 + \gamma)\, opt_k(P)\right] \geq \left(\frac{1-\delta}{5}\right)^k,$$

using $\mathcal{O}\left(n 2^{(\frac{k}{\alpha} m_{\gamma,\delta})^{\mathcal{O}(1)}}\right)$ arithmetic operations.

This theorem can be proven analogously to the proof of Theorem 2.1. Generally speaking, during execution of the algorithm we consider the two superclusters P_1' and P_2', with P_1' consisting of the clusters whose medians have already been approximated by $\tilde{C}_i = \{\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\}$ and P_2' consisting of the clusters whose medians have yet to be found. Again, it can be shown that by removing a neighborhood N_i' in step 12 only a small fraction of points from P_2' are removed which leads to $\mathrm{D}(P_2' \cap N_i', \tilde{C}_i) \leq 8\alpha k\, \mathrm{D}(P_1', \tilde{C}_i)$. Summation over all $\tilde{C}_i$ leads to the given bound. We omit the details here for the sake of brevity.

Our main result, Theorem 1.1, is an immediate consequence of Theorem 2.2 by choosing parameters $\alpha = \frac{\epsilon}{16k^2}$ and $\gamma = \frac{\epsilon}{3}$. By running the algorithm multiple times the success probability can be amplified as close to 1 as desired without changing the asymptotic running time.

3 Sampling for large classes of dissimilarity measures

We now show how to apply Theorem 1.1 to various dissimilarity measures.

3.1 Sampling for arbitrary metrics with bounded doubling dimension. Let (X, D) be a metric space and let $\mathrm{diam}(Y) = \max_{x,y \in Y} \mathrm{D}(x, y)$ be the *diameter* of $Y \subseteq X$. A collection $\{Y_1, Y_2, \ldots, Y_\nu\}$ of subsets of Y is called a *β-cover* iff $Y = \bigcup_{i=1}^{\nu} Y_i$ and $\mathrm{diam}(Y_i) \leq \beta$. The *covering number* $\mathrm{C}(Y, \beta)$ is the smallest cardinality of a β-cover of Y, i.e. $\mathrm{C}(Y, \beta) = \min\{\nu \mid \exists Y_1, Y_2, \ldots, Y_\nu \subseteq Y, Y = \bigcup_{i=1}^{\nu} Y_i, \mathrm{diam}(Y_i) \leq \beta\}$. The following definition is taken from [12].

Definition 3.1. (doubling dimension) *For $Y \subseteq X$ let be $\lambda(Y) = \mathrm{C}(Y, \frac{1}{2}\mathrm{diam}(Y))$. Then the* doubling dimension *of (X, D) is defined as* $\mathrm{ddim}(X) = \sup_{Y \subseteq X} \log_2\big(\lambda(Y)\big)$.

In the sequel, let c be the optimal 1-median of input instance $P \subseteq X$ of size n, i.e. $\mathrm{D}(P, c) = opt_1(P)$. We need the following lemma which is an immediate consequence of Markov's inequality and the union bound.

Lemma 3.1. *Let $\delta > 0$. A random sample multiset $S \subseteq P$ of size m satisfies*

$$\Pr\left[\exists q \in S : \mathrm{D}(q, c) \geq \frac{1}{\delta n}\mathrm{D}(P, c)\right] \leq m\delta\ .$$

We also need the following result which is a small modification of a result from [15] (see also [29]).

Lemma 3.2. *Let $\gamma \leq 1$ and $b \in X$ be an arbitrary point with $\mathrm{D}(P, b) > (1 + \frac{4}{5}\gamma)\,\mathrm{D}(P, c)$. A random sample multiset $S \subseteq P$ of size m satisfies*

$$\Pr\left[\mathrm{D}(S, c) \geq \mathrm{D}(S, b) - \frac{\gamma m}{5n}\mathrm{D}(P, c)\right] < \exp\left(-\frac{\gamma^2 m}{144}\right)\ .$$

Proof. This proof is a small modification of the proof of Theorem 34 presented in [29]. For a random sample multiset $S \subseteq P$ we consider the random variable

$$X = \frac{\mathrm{D}(S, b) - \mathrm{D}(S, c) + m\,\mathrm{D}(b, c)}{2\big(\mathrm{D}(b, c) + \frac{\gamma}{5n}\mathrm{D}(P, c)\big)}$$

By the triangle inequality we have $\mathrm{D}(S, b) + m\,\mathrm{D}(b, c) \geq \mathrm{D}(S, c)$ and $\mathrm{D}(S, b) - \mathrm{D}(S, c) \leq m\,\mathrm{D}(b, c)$. Thus, $0 \leq X \leq m$. We are interested in the probability of the event $X \leq \frac{1}{2}m$. This case is equivalent to $\mathrm{D}(S, c) \geq \mathrm{D}(S, b) - \frac{\gamma m}{5n}\mathrm{D}(P, c)$. We use Chernoff bounds to show that this happens only with small probability.

First, we get

$$\mathrm{E}[X] = \frac{m}{n} \cdot \frac{\mathrm{D}(P, b) - \mathrm{D}(P, c) + n\,\mathrm{D}(b, c)}{2\big(\mathrm{D}(b, c) + \frac{\gamma}{5n}\mathrm{D}(P, c)\big)}\ .$$

By choice of b, we have $\mathrm{D}(P, b) - \mathrm{D}(P, c) > \frac{4\gamma}{5}\mathrm{D}(P, c)$ and $\mathrm{D}(P, b) - \mathrm{D}(P, c) > \frac{4\gamma}{5}/(1 + \frac{4\gamma}{5})\,\mathrm{D}(P, b)$. Thus, we get

$$\begin{aligned}
&\mathrm{D}(P,b)-\mathrm{D}(P,c)\\
&=\frac{3-\frac{4\gamma}{5}}{4}(\mathrm{D}(P,b)-\mathrm{D}(P,c))+\frac{1+\frac{4\gamma}{5}}{4}(\mathrm{D}(P,b)-\mathrm{D}(P,c))\\
&>(3-\frac{4\gamma}{5})\frac{\gamma}{5}\mathrm{D}(P,c)+\frac{\gamma}{5}\mathrm{D}(P,b)\\
&=(2-\frac{4\gamma}{5})\frac{\gamma}{5}\mathrm{D}(P,c)+\frac{\gamma}{5}\big(\mathrm{D}(P,c)+\mathrm{D}(P,b)\big)\\
&\geq(2-\frac{4\gamma}{5})\frac{\gamma}{5}\mathrm{D}(P,c)+\frac{\gamma}{5}n\,\mathrm{D}(b,c)\\
&=(2-\gamma)\frac{\gamma}{5}\mathrm{D}(P,c)+\frac{\gamma}{5}\big(n\,\mathrm{D}(b,c)+\frac{\gamma}{5}\mathrm{D}(P,c)\big)\\
&\geq\frac{\gamma}{5}D(P,c)+\frac{\gamma}{5}\big(n\,\mathrm{D}(b,c)+\frac{\gamma}{5}D(P,c)\big)\ .
\end{aligned}$$

Here, the second inequality follows because of the triangle inequality. The last inequality is due to $2-\gamma\geq 1$ for $\gamma\leq 1$. Hence,

$$\begin{aligned}
\frac{2}{m}\mathrm{E}[X]&=\frac{\frac{1}{n}\big(\mathrm{D}(P,b)-\mathrm{D}(P,c)+n\,\mathrm{D}(b,c)\big)}{\mathrm{D}(b,c)+\frac{\gamma}{5n}\mathrm{D}(P,c)}\\
&\geq\frac{(1+\frac{\gamma}{5})\big(\mathrm{D}(b,c)+\frac{\gamma}{5n}D(P,c)\big)}{\mathrm{D}(b,c)+\frac{\gamma}{5n}\mathrm{D}(P,c)}\geq 1+\frac{\gamma}{5}\ .
\end{aligned}$$

So we have $\frac{1}{2}m\leq(1-\mu)\,\mathrm{E}[X]$ for $\mu=\frac{\gamma}{5}/(1+\frac{\gamma}{5})\geq\frac{\gamma}{6}$. Using Chernoff bounds we get

$$\begin{aligned}
\Pr\left[X\leq\frac{1}{2}m\right]&\leq\Pr\left[X\leq\big(1-\frac{\gamma}{6}\big)\,\mathrm{E}[X]\right]\\
&<\exp\left(-\frac{\gamma^2}{72}\mathrm{E}[X]\right)<\exp\left(-\frac{\gamma^2 m}{144}\right)\ .\ \square
\end{aligned}$$

For a random sample multiset S we denote its optimal 1-median by $c_S=\arg\min_{x\in X}\mathrm{D}(S,x)$. We obtain the following result for metrics with bounded doubling dimension.

Lemma 3.1. *Let (X,D) be a metric space with $\mathrm{ddim}(X)\leq B$. For $\delta>0$ exists a constant λ_δ such that every random sample multiset $S\subset P$ of size $m\geq\lambda_\delta B\frac{1}{\gamma^2}\log\frac{1}{\gamma}$ satisfies*

$$\Pr\left[\mathrm{D}(P,c_S)\leq(1+\gamma)\,\mathrm{D}(P,c)\right]\geq 1-\delta\ .$$

Proof. Consider a ball U with radius $r=\frac{6m}{\delta n}\mathrm{D}(P,c)$ and center c. By Lemma 3.1 with probability $1-\frac{\delta}{2}$ all sample points lie within ball U' with radius $r'=\frac{1}{3}r$ and center c. Now consider an arbitrary $q\in X\backslash U$. If $S\subseteq U'$ we get $\mathrm{D}(S,q)\geq 2r'm$. However, $\mathrm{D}(S,c)\leq r'm$. Thus, we can conclude that $c_S\in U$ with probability $1-\frac{\delta}{2}$. Therefore, from now on we will only consider solutions contained in U.

By $\mathrm{ddim}(X)\leq B$ we know that every $Y\subseteq X$ has a $\frac{1}{2}\mathrm{diam}(Y)$-cover of cardinality at most 2^B. Applying this recursively, we obtain that U has an $\frac{r}{2^j}$-cover of cardinality at most 2^{jB} for any $j\in\mathbb{N}$. Thus, for $j=\lceil\log\frac{30m}{\delta\gamma}\rceil$ we obtain a $\frac{\gamma}{5n}\mathrm{D}(P,c)$-cover of cardinality $l\leq\left(\frac{60m}{\delta\gamma}\right)^B$. Let $\{U_1,U_2,\ldots,U_l\}$ be such a cover and let $N=\{x_1,x_2,\ldots,x_l\}$ be an arbitrary set of points with $x_i\in U_i$.

Define $N_b=\{x\in N\mid \mathrm{D}(P,x)>(1+\frac{4\gamma}{5})\,\mathrm{D}(P,c)\}$. We apply Lemma 3.2 to N_b. By the union bound, for each δ there exists a constant λ_δ such that for $m\geq\lambda_\delta B\frac{1}{\gamma^2}\log\frac{1}{\gamma}$ we have

$$\begin{aligned}
\Pr&\left[\exists x\in N_b:\mathrm{D}(S,c)\geq\mathrm{D}(S,x)-\frac{\gamma m}{5n}\mathrm{D}(P,c)\right]\\
&<\left(\frac{60m}{\delta\gamma}\right)^B\cdot\exp\left(-\frac{\gamma^2 m}{144}\right)<\frac{\delta}{2}\ .
\end{aligned}$$

So, again with probability $1-\frac{\delta}{2}$, for all $x\in N_b$ we have $\mathrm{D}(S,x)>\mathrm{D}(S,c)+\frac{\gamma m}{5n}\mathrm{D}(P,c)$.

Now consider an optimal 1-median c_S of S. Let q be the closest point in N to c_S. Since $c_S\in U$ we have $\mathrm{D}(q,c_S)\leq\frac{\gamma}{5n}\mathrm{D}(P,c)$. Furthermore, by the triangle inequality for all $x\in N_b$ we have

$$\begin{aligned}
\mathrm{D}(S,q)&\leq\mathrm{D}(S,c_S)+\frac{\gamma m}{5n}\mathrm{D}(P,c)\\
&\leq\mathrm{D}(S,c)+\frac{\gamma m}{5n}\mathrm{D}(P,c)<\mathrm{D}(S,x).
\end{aligned}$$

Hence, $q\notin N_b$ and $\mathrm{D}(P,q)\leq(1+\frac{4\gamma}{5})\,\mathrm{D}(P,c)$. By the triangle inequality we conclude

$$\mathrm{D}(P,c_S)\leq\mathrm{D}(P,q)+n\,\mathrm{D}(q,c_S)\leq(1+\gamma)\,\mathrm{D}(P,c)\ .$$

This event happens with probability at least $(1-\frac{\delta}{2})^2>1-\delta$. □

Corollary 3.1. *An arbitrary metric space (X,D) with $\mathrm{ddim}(X)\leq B$ satisfies the $[\gamma,\delta]$-sampling property with $m_{\gamma,\delta}=\lambda_\delta B\frac{1}{\gamma^2}\log\frac{1}{\gamma}$, provided that we have access to an algorithm that computes $\Gamma(S)=\arg\min_{x\in X}\mathrm{D}(S,x)$ in time depending only on $|S|$.*

3.2 Sampling for Mahalanobis distances.

For any positive definite matrix $A\in\mathbb{R}^{d\times d}$ we define the *Mahalanobis distance* (see [23]) with respect to A for $x,y\in\mathbb{R}^d$ as

$$\mathrm{D}_A(x,y)=(x-y)^\top\cdot A\cdot(x-y)\ .$$

Note that with $A=I_d$ we get that the square of the ℓ_2-norm on $\mathbb{R}^d$ is a Mahalanobis distance.

In the sequel, for any set P we denote by c_P the *centroid* of P, i.e. $c_P=\frac{1}{|P|}\sum_{x\in P}x$. It is known that c_P is the unique optimal 1-median of P (see [4]).

LEMMA 3.3. *A random sample multiset $S \subseteq P$ of size $m \geq \frac{1}{\gamma\delta}$ satisfies*

$$\Pr\left[\mathrm{D}_A(P, c_S) \leq (1+\gamma)\,\mathrm{D}_A(P, c_P)\right] \geq 1-\delta\ .$$

The proof of this lemma is a straightforward generalization of the proof of Lemma 1 from [14].

COROLLARY 3.2. D_A *on* $\mathbb{R}^d$ *satisfies the* $[\gamma, \delta]$*-sampling property with* $m_{\gamma,\delta} = \frac{1}{\gamma\delta}$ *and* $\Gamma(S) = c_S$.

3.3 Sampling for Bregman divergences.

For any strictly convex, differentiable function $\phi : \Delta \subseteq \mathbb{R}^d \to \mathbb{R}$ we define the *Bregman divergence* (see [5, 6]) with respect to ϕ for $x, y \in \Delta$ as

$$\mathrm{D}_\phi(x, y) = \phi(x) - \phi(y) - \nabla\phi(y)^\top (x-y)\ .$$

Here $\nabla\phi(y)$ denotes the gradient of ϕ at point y. $\mathrm{D}_\phi(x, y)$ can be seen as the tail of the first-order Taylor expansion of $\phi(x)$ at y. Bregman divergences include many prominent dissimilarity measures like the square of the ℓ_2-norm (with $\phi_{\ell_2^2}(t) = \|t\|_2^2$), the Mahalanobis distances ($\phi_A(t) = t^\top A t$) and the Kullback-Leibler divergence ($\phi_{\mathrm{KL}}(t) = \sum t_i \ln t_i$).

We make use of the following fact from [4].

PROPOSITION 3.1. *All Bregman divergences* D_ϕ *with* $P \subseteq \Delta$ *of size* n *and* $q \in \Delta$ *satisfy*

$$\mathrm{D}_\phi(P, q) = \mathrm{D}_\phi(P, c_P) + n\,\mathrm{D}_\phi(c_P, q).$$

As a consequence, we obtain that for all Bregman divergences the centroid c_P is the optimal 1-median of P. In certain cases, we can obtain the $[\gamma, \delta]$-sampling property of Bregman divergences using the $[\gamma, \delta]$-sampling property of a Mahalanobis distance, as is stated in the following lemma.

LEMMA 3.4. *If there exists a positive definite* $A \in \mathbb{R}^{d\times d}$ *and a constant* $0 < \mu \leq 1$ *such that for all* $x, y \in \Delta$

$$\mu \cdot \mathrm{D}_A(x, y) \leq \mathrm{D}_\phi(x, y) \leq \mathrm{D}_A(x, y) \tag{3.6}$$

then a random sample multiset $S \subseteq P$ *of size* $m \geq \frac{1}{\gamma\delta\mu}$ *satisfies*

$$\Pr\left[\mathrm{D}_\phi(P, c_S) \leq (1+\gamma)\,\mathrm{D}_\phi(P, c_P)\right] \geq 1-\delta\ .$$

Proof. From proof of Lemma 3.3 we know that $E[\mathrm{D}_A(c_P, c_S)] = \frac{1}{mn}\mathrm{D}_A(P, c_P)$. Using (3.6) and Markov's inequality with $\gamma \geq \frac{1}{m\delta\mu}$ we get

$$\Pr\left[\mathrm{D}_\phi(c_P, c_S) > \gamma\frac{1}{n}\mathrm{D}_\phi(P, c_P)\right] \leq \Pr\left[\mathrm{D}_A(c_P, c_S) > \gamma\mu\frac{1}{n}\mathrm{D}_A(P, c_P)\right] \leq \delta\ .$$

Since $\mathrm{D}_\phi(P, c_S) = \mathrm{D}_\phi(P, c_P) + n\,\mathrm{D}_\phi(c_P, c_S)$ the lemma follows. □

We can apply this result to the Kullback-Leibler divergence. For any $p, q \in \mathbb{R}_+^d$ the natural[1] *generalized Kullback-Leibler divergence* (see [19, 8]) is defined by

$$\mathrm{D}_{\mathrm{KL}}(p, q) = \sum_{i=1}^d p_i \ln\frac{p_i}{q_i} - \sum_{i=1}^d (p_i - q_i)\ .$$

LEMMA 3.5. *Let be* $\nu > \lambda > 0$ *and define* $A = \frac{1}{2\lambda} I_d$. *For all* $p, q \in [\lambda, \nu]^d$ *we get*

$$\frac{\lambda}{\nu}\mathrm{D}_A(p, q) \leq \mathrm{D}_{\mathrm{KL}}(p, q) \leq \mathrm{D}_A(p, q)\ .$$

Proof. We consider the strictly convex, C^2-function $\phi_{\mathrm{KL}}(t) = \sum_{i=0}^d t_i \ln t_i - t_i$, i.e. all second-order partial derivatives exist and are continuous. The Kullback-Leibler divergence is the tail of the first-order Taylor expansion of $\phi_{\mathrm{KL}}(p)$ at q. Therefore, by the Lagrange form of the remainder term, there exists an ξ with $\xi_i \in [p_i, q_i]$ such that

$$\mathrm{D}_{\mathrm{KL}}(p, q) = \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d \frac{\partial^2}{\partial t_i \partial t_j}\phi_{\mathrm{KL}}(\xi)(p_i - q_i)(p_j - q_j)\ .$$

Since $\frac{\partial^2}{\partial t_i^2}\phi(t) = \frac{1}{t_i}$ and $\frac{\partial^2}{\partial t_i \partial t_j}\phi(t) = 0$ for $i \neq j$ we obtain

$$\mathrm{D}_{\mathrm{KL}}(p, q) = \frac{1}{2}(p-q)^\top \begin{pmatrix} \frac{1}{\xi_1} & & & \\ & \frac{1}{\xi_2} & & \\ & & \ddots & \\ & & & \frac{1}{\xi_d} \end{pmatrix} (p-q)\ .$$

Using $\frac{1}{\nu} \leq \frac{1}{\xi_i} \leq \frac{1}{\lambda}$ the lemma follows. □

COROLLARY 3.3. D_{KL} *on* $[\lambda, \nu]^d$ *satisfies the* $[\gamma, \delta]$*-sampling property with* $m_{\gamma,\delta} = \frac{\nu}{\gamma\delta\lambda}$ *and* $\Gamma(S) = c_S$.

Here the dependency of sample size m on $\frac{1}{\lambda}$ seems awkward. However, we can show that for the case of Kullback-Leibler divergence this dependency can not be improved significantly. To be more specific, consider an input instance $P \subset [p, 1]^2$ of size n consisting of $(1 - \frac{1}{2m})n$ copies of point $(p, 1-p)$ and $\frac{1}{2m}n$ copies of point $(1-p, p)$. By Markov's inequality, with probability at least $\frac{1}{2}$ a sample set S of size m consists solely of copies of $(p, 1-p)$. Therefore, with probability $\frac{1}{2}$ point $(p, 1-p)$ is assigned as approximate 1-median of P. However, we can prove the following lemma.

LEMMA 3.6. *Let be* $p < \frac{1}{4}$ *and* $\gamma \leq \frac{1}{4}(\ln\frac{1}{p})^{-1}$. *Let* P *be defined as above. If* $m < \frac{1}{2p} - 1$, *then point* $(p, 1-p)$ *is not a* $(1+\gamma)$*-approximate 1-median of* P.

[1] For different bases of the logarithm we obtain similar results.

Proof. For simplicity of notation, we identify point $(p, 1-p)$ with the real number p, point $(c, 1-c)$ with c, and so on. Obviously for the optimal 1-median c of P we have $c = p + \frac{1-2p}{2m}$.

Now let us assume for contradiction that p is a $(1+\gamma)$-approximate 1-median of P, or equivalently by Proposition 3.1 $\mathrm{D}_{\mathrm{KL}}(c,p) \leq \frac{\gamma}{n} \mathrm{D}_{\mathrm{KL}}(P,c)$. By definition of set P we get

$$
\begin{aligned}
&\frac{1}{n}\mathrm{D}_{\mathrm{KL}}(P,c) = (1-\frac{1}{2m})\mathrm{D}_{\mathrm{KL}}(p,c) + \frac{1}{2m}\mathrm{D}_{\mathrm{KL}}(1-p,c) \\
&\leq (1-\frac{1}{2m})\mathrm{D}_{\mathrm{KL}}(p,c) + \frac{1}{2m}\mathrm{D}_{\mathrm{KL}}(1-p,c) + \mathrm{D}_{\mathrm{KL}}(c,p) \\
&\leq (1-\frac{1}{2m})\left(p\ln\frac{p}{c} + (1-p)\ln\frac{1-p}{1-c}\right) \\
&\quad + \frac{1}{2m}\left((1-p)\ln\frac{1-p}{c} + p\ln\frac{p}{1-c}\right) \\
&\quad + c\ln\frac{c}{p} + (1-c)\ln\frac{1-c}{1-p} \\
&= (c-p)\ln\frac{1-p}{p} \leq (c-p)\ln\frac{1}{p} \ .
\end{aligned}
$$

Hence, by $\gamma \leq \frac{1}{4}(\ln\frac{1}{p})^{-1}$ we get

$$\mathrm{D}_{\mathrm{KL}}(c,p) \leq \gamma(c-p)\ln\frac{1}{p} \leq \frac{1}{4}(c-p) \ .$$

On the other hand, we know that the Kullback-Leibler divergence $\mathrm{D}_{\mathrm{KL}}(c,p)$ is the tail of the first-order Taylor expansion of $\phi_{\mathrm{KL}}(c) = c\ln c - (1-c)\ln(1-c)$ at p. Therefore, by the Lagrange form of the remainder term exist $\xi_1 \in [p,c]$ and $\xi_2 \in [1-c, 1-p]$ satisfying

$$
\begin{aligned}
\mathrm{D}(c,p) &= \frac{1}{2\xi_1}(c-p)^2 + \frac{1}{2\xi_2}(c-p)^2 \\
&\geq \left(\frac{1}{2c} + \frac{1}{2(1-p)}\right)(c-p)^2 \geq \frac{(c-p)^2}{2c(1-p)} \ .
\end{aligned}
$$

This leads to $\frac{c-p}{2c} \leq \frac{c-p}{2c(1-p)} \leq \frac{1}{4}$, or equivalently $c \leq 2p$. Since $c = p + \frac{1-2p}{2m}$ we get $\frac{1-2p}{2m} \leq p$ and finally $m \geq \frac{1}{2p} - 1$, which is a contradiction. □

Now assume with probability $\delta = \frac{1}{2}$ our algorithm could compute a $(1+\gamma)$-approximation by using a sample set S of size $m = \frac{1}{16\gamma p}(\ln\frac{1}{p})^{-1}$. Then for every $p < \frac{1}{4}$ and for $\gamma = \frac{1}{4}(\ln\frac{1}{p})^{-1}$ we have $m < \frac{1}{2p} - 1$, which contradicts the observation above.

Hence, our bound from Corollary 3.3 cannot be improved by a factor of $\frac{1}{32}(\ln\frac{1}{p})^{-1}$.

4 Generalization and Discussion

In this section we generalize our result to an even larger family of dissimilarity measures. For the $[\gamma,\delta]$-sampling property as stated in Theorem 1.1, we require that the optimal 1-median c_S of S can be computed in finite time. However, for the Euclidean k-median problem, no such algorithm is known. Moreover, it has been shown that finding c_S requires finding roots of high-order polynomials, which can not be achieved using only radicals [2].

It turns out that our definition of the $[\gamma,\delta]$-sampling property is far more restrictive than is necessary for our algorithm. All we have to ensure is that given a constant sized sample set $S \subseteq P$, we can find a set of candidates for the approximate 1-median of cluster P. Being a 1-median of S is sufficient for the dissimilarities studied in Section 3, but in general it is not necessary.

Therefore, we can give a more general definition of the $[\gamma,\delta]$-sampling property. It has been shown by Kumar et al. [21] that the Euclidean k-median problem, using the ℓ_2-norm as distance measure, as well as the discrete version of the Euclidean k-means problem satisfy this weaker property.

Property 4.1. *We say a dissimilarity measure* D *satisfies the* $[\gamma,\delta]$-sampling property *iff there exist integer constants $m_{\gamma,\delta}$ and $\tau_{\gamma,\delta}$ such that for each $P \subseteq \Delta$ of size n and for each uniform sample multiset $S \subseteq P$ of size $m_{\gamma,\delta}$ a set $\Gamma(S) \subseteq \Delta$ of size at most $\tau_{\gamma,\delta}$ can be computed which contains a point $\tilde{c} \in \Gamma(S)$ satisfying*

$$\Pr\left[\mathrm{D}(P,\tilde{c}) \leq (1+\gamma)\, opt_1(P)\right] \geq 1-\delta.$$

Furthermore, $\Gamma(S)$ can be computed from S in time depending only on γ, δ, and $|S|$.

Algorithm Cluster can be easily adopted to this new definition. In particular, for each subset $S' \subseteq S$ of size $m_{\gamma,\delta}$ instead of a single point a constant sized set $\Gamma(S')$ is added to the candidate set. Therefore, the asymptotic running time remains the same.

Property 4.1 can be seen as a generalization of the "random sampling procedure property" from [21]. However, our result applies to a larger family of dissimilarity measures. In particular, our result shows that the second requirement from [21] ("tightness property") is not necessary to achieve a $(1+\epsilon)$-approximation for dissimilarity measures satisfying $[\gamma,\delta]$-sampling property.

An interesting direction for future research is to find additional dissimilarity measures that satisfy the $[\gamma,\delta]$-sampling property. Finally, it remains an open problem to give sufficient and necessary conditions for dissimilarity measures such that a $(1+\epsilon)$-approximate solution of the corresponding k-median problem can be found.

References

[1] Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC '02)*, pages 250–257, 2002.

[2] Chandrajit L. Bajaj. The algebraic degree of geometric optimization problems. *Discrete and Computational Geometry*, 3(1):177–191, December 1988.

[3] L. Douglas Baker and Andrew K. McCallum. Distributional clustering of words for text classification. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 96–103, Melbourne, AU, 1998. ACM Press, New York, US.

[4] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research (JMLR)*, 6:1705–1749, October 2005.

[5] Lev M. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[6] Yair Censor and Stavros A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications.* Oxford University Press, 1997.

[7] Ke Chen. On k-median clustering in high dimensions. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '06)*, pages 1177–1185, 2006.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 2nd edition, 2006.

[9] Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classifcation. *Journal of Machine Learning Research (JMLR)*, 3:1265–1287, March 2003.

[10] Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A ptas for k-means clustering based on weak coresets. In *Proceedings of the 23rd ACM Symposium on Computational Geometry (SCG'07)*, pages 11–18, 2007.

[11] Wenceslas Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC '03)*, pages 50–58, 2003.

[12] Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals and low-distortion embeddings. In *Proceedings of the 44th Symposium on Foundations of Computer Science (FOCS'03)*, pages 534–543, 2003.

[13] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC '04)*, pages 291–300, 2004.

[14] Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the 10th annual Symposium on Computational Geometry (SCG '94)*, pages 332–339, New York, NY, USA, 1994. ACM Press.

[15] Piotr Indyk and Mikkel Thorup. Approximate 1-medians. Manuscript, 2000.

[16] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[17] Kamal Jain, Mohammad Mahdian, Evangelos Markakis, Amin Saberi, and Vijay V. Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing lp. *Journal of the ACM (JACM)*, 50(6):795–824, 2003.

[18] Stavros G. Kolliopoulos and Satish Rao. A nearly linear-time approximation scheme for the euclidean κ-median problem. In *Proceedings of the 7th annual European Symposium on Algorithms (ESA '99)*, pages 378–389, 1999.

[19] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[20] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04)*, pages 454–462, Washington, DC, USA, 2004. IEEE Computer Society.

[21] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear time for clustering problems in any dimensions. In *Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP'05)*, pages 1374–1385, Berlin, 2005. Springer Verlag.

[22] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[23] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55, 1936.

[24] Jiri Matousek. On approximate geometric k-clustering. *Discrete & Computational Geometry*, 24(1):61–84, 2000.

[25] Vaniel P. Mercer. Clustering large datasets. Technical report, Linacre College, 2003.

[26] Ramgopal R. Mettu and C. Greg Plaxton. Optimal time bounds for approximate clustering. *Machine Learning*, 56(1–3):35–60, 2004.

[27] Fernando C. Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.

[28] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In *Advances in Neural Information Processing Systems 12 (NIPS 12)*, pages 617–623, 1999.

[29] Mikkel Thorup. Quick k-median, k-center, and facility location for sparse graphs. *SIAM Journal on Computation*, pages 405–432, 2004.

[30] Rui Xu and Donald Wunsch II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16:645 – 678, May 2005.