Coresets and Approximate Clustering for Bregman Divergences*

Marcel R. Ackermann† and Johannes Blömer†

(revised und updated version of February 4, 2009)

Abstract

We study the generalized k-median problem with respect to a Bregman divergence D_ϕ. Given a finite set $P \subseteq \mathbb{R}^d$ of size n, our goal is to find a set C of size k such that the sum of errors $\mathrm{cost}(P,C) = \sum_{p\in P} \min_{c\in C}\{\mathrm{D}_\phi(p,c)\}$ is minimized. The Bregman k-median problem plays an important role in many applications, e.g. information theory, statistics, text classification, and speech processing. We give the first coreset construction for this problem for a large subclass of Bregman divergences, including important dissimilarity measures such as the Kullback-Leibler divergence and the Itakura-Saito divergence. Using these coresets, we give a $(1+\epsilon)$-approximation algorithm for the Bregman k-median problem with running time $\mathcal{O}\left(dkn + d2^{(\frac{k}{\epsilon})^{\Theta(1)}} \log^{k+2} n\right)$. This result improves over the previousely fastest known $(1+\epsilon)$-approximation algorithm from [1]. Unlike the analysis of most coreset constructions our analysis does not rely on the construction of ϵ-nets. Instead, we prove our results by purely combinatorial means.

1 Introduction

Clustering is the problem of grouping a set of objects into subsets — so-called clusters — such that similar objects are grouped together. Algorithms for clustering objects have numerous applications in various areas of computer science such as machine learning, data compression, speech and image analysis, data mining, or pattern recognition. Obviously, in different applications we need to cluster different objects, such as text documents, gene sequences, probability distributions or speech signals. Therefore, the notion of (dis-)similarity varies from application to application. The so-called Bregman divergences play an important role in many of these applications.

The quality of a clustering is measured using a well-defined cost function involving the proper dissimilarity measure for a given application. A cost function that has been proved useful in the past decades is the k-median cost function. Here the objective is to partition a set of objects into k clusters, each with a given cluster representative, such that the sum of errors from each object to their representative is minimized. Many approximation algorithms and techniques for this minimization problem have been developed when the dissimilarity measure used is a metric such as the Euclidean distance (known as the Euclidean k-median problem), or when the squared Euclidean distance is used (known as the Euclidean k-means problem). However, until recently for non-metric dissimilarity measures almost no approximation algorithms were known.

This stands in sharp contrast to the fact that many non-metric dissimilarity measures are used in various applications. To name just a few examples, the Mahalanobis divergence is used in statistics, the Pearson correlation is used in genetics, the cosine similarity is used in data mining, the Itakura-Saito divergence is used in speech processing, and the Kullback-Leibler divergence is used in machine learning, data mining, and information theory. In fact, our original motivation for clustering with general divergence measures came from an industrial project on lossless compression of Java and C++ executable code, where we used clustering with respect to the Kullback-Leibler divergence. For an overview of applications of various dissimilarity measures and clustering algorithms we refer to [17, 23].

The Kullback-Leibler divergence is an instance of a broader class of dissimilarity measures that has attained considerable attention in the past few years: the class of Bregman divergences. Bregman divergences include useful dissimilarity measures such as the squared Euclidean distance as well as the above-mentioned Mahalanobis distance and Itakura-Saito divergence. While research on the combinatorial and algorithmic properties of the Bregman k-median problem is relatively new, some interesting results have been obtained recently. In [4], Lloyd's famous k-means heuristic has been generalized to all Bregman divergences. In [22], the notion of Bregman Voronoi diagrams has been studied. A first $(1+\epsilon)$-approximation algorithm applicable to many

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

†Department of Computer Science, University of Paderborn, Germany, {mra,bloemer}@uni-paderborn.de

Bregman divergences has been proposed in [1].

In this paper we take another step in applying methods from computational geometry to clustering with respect to Bregman divergences. Building heavily on Chen's work [8, 9], we describe the first coreset construction for a large subclass of Bregman divergences, including important dissimilarity measures such as the Kullback-Leibler divergence. Unlike the analysis of most coreset constructions our analysis does not rely on the construction of ϵ-nets. Instead, we give a purely combinatorial proof. Using these coresets we describe a $(1+\epsilon)$-approximation algorithm solving the k-median problem for large class of Bregman divergences. The algorithm has running time $\mathcal{O}\left(dkn + d2^{(\frac{k}{\epsilon})^{\Theta(1)}} \log^{k+2} n\right)$. Assuming that d and k are much smaller than n, this result improves significantly over the previously fastest known $(1+\epsilon)$-approximation algorithm from [1] which runs in time $\mathcal{O}(d2^{(\frac{k}{\epsilon})^{\Theta(1)}} n)$.

1.1 Related work. Relatively little is known about the complexity and geometry of the general Bregman k-median problem. In [4], Banerjee et al. have generalized Lloyd's famous k-means heuristic to all Bregman divergences. Thereby, they give a unified explanation for earlier observations that the k-means heuristic is applicable to individual measures such as the Itakura-Satito divergence [6] or the Kullback-Leibler divergence [11]. In [22], Nielsen et al. study the notion of Bregman Voronoi diagrams and show how to compute them efficiently. A first $(1+\epsilon)$-approximation algorithm applicable to the k-median problem for many Bregman divergences has been proposed by Ackermann et al. in [1]. This result generalizes an earlier algorithm from [18] for the squared Euclidean distances to a variety of Bregman divergences and some other dissimilarity measures.

Coresets in the context of the Euclidean k-median and the Euclidean k-means problem have been known for some time (see [21, 3, 15, 10, 13]). A (k,ϵ)-coreset for a set P is a small (weighted) set such that for any set C of k cluster centers the (weighted) clustering cost of the coreset is an approximation for the clustering cost of the original set P with relative error at most ϵ. From this definition it follows that a good approximate clustering for the set P is given by a good approximate clustering for the coreset. Many coreset constructions have been given, most notably by Har-Peled and Kushal in [14] where the size of the coreset is independent of the size of the input set (but still exponential in dimension d). Of the recent results, the coreset construction of Chen in [8, 9] is of particular interest to our contribution. The size of Chen's coresets is only linear in d. Thus, these coresets are well suited for high-dimensional settings.

Furthermore, a relaxed notion of coresets has been introduced by Feldman et al. in [12]. The definition of *weak coresets* differs slightly from the original definition of a coreset. For a finite set Γ a set S is called a Γ-weak coreset for P if for all sets of k cluster centers from Γ the clustering cost of the coreset is a good approximation to the clustering cost of the whole set P. This notion of weak coreset plays a crucial role in our work.

1.2 Our results. We will use a generalized formulation of the k-median problem. Let $\mathrm{D}(\cdot,\cdot)$ be a *dissimilarity measure* (e.g., a Bregman divergence) that specifies the dissimilarity between two objects from domain $\mathcal{X} \subseteq \mathbb{R}^d$. For $p \in \mathcal{X}$ and $C \subseteq \mathcal{X}$ we also use $\mathrm{D}(p,C) = \min_{c\in C} \mathrm{D}(p,c)$ for simplicity of notation.

The k-median problem with respect to D is defined as follows. We are given a finite set $P \subseteq \mathcal{X}$ of objects from the domain. Our goal is to find a set $C \subseteq \mathcal{X}$ of size k such that the sum of errors

$$\mathrm{cost}(P,C) = \sum_{p\in P} \mathrm{D}(p,C)$$

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.

Throughout this paper, we consider the k-median clustering problem for a large class of Bregman divergences we call *μ-similar Bregman divergences*. To the best of our knowledge, this class includes most of the Bregman divergences that are used in practice, such as the squared Euclidean distance, Mahalanobis distance, Kullback-Leibler divergence, and Itakura-Saito divergence. Our main results can be summarized as follows.

- Given a μ-similar Bregman divergence on domain $\mathcal{X} \subseteq \mathbb{R}^d$. We show that for any $P \subseteq \mathcal{X}$ of size n and for any finite $\Gamma \subseteq \mathcal{X}$ there exists a weak coreset of size $\mathcal{O}\left(\frac{1}{\epsilon^2} k \log(n) \log(k|\Gamma|^k \log n)\right)$. Furthermore, given a set of medians of a constant factor bicriteria approximation of the k-median problem, we show that with high probability such a coreset can be constructed in time $\mathcal{O}\left(dkn + \frac{1}{\epsilon^2} k \log(n) \log(k|\Gamma|^k \log n)\right)$, using the coreset construction from [8, 9].
- Using weak coresets and an adaptation of an algorithm from [1] we give a $(1+\epsilon)$-approximation algorithm with running time $\mathcal{O}\left(dkn + d2^{(\frac{k}{\epsilon})^{\Theta(1)}} \log^{k+2} n\right)$. To our knowledge, this new algorithm is the asymptotically fastest algorithm known for the k-median problem with respect to a number of Bregman divergences such as the Kullback-Leibler divergence and the Itakura-Saito divergence.

We point out that due to the low dependency on d our results are particularly relevant for high-dimensional settings. Additionally, using the merge-and-reduce technique from [15] our weak coresets can also be applied to the data streaming model.

1.3 Our techniques. Unlike the analysis of most "strong" coreset constructions our analysis does not rely on the construction of ϵ-nets (e.g., compare to [15, 8, 9]). Usually, the analysis of coresets via ϵ-nets requires the underlying space to be a metric space. However, most Bregman divergences are asymmetric, do not obey the triangle inequality, and may even possess singularities, i.e., there are $p, q \in \mathcal{X}$ such that $\mathrm{D}(p, q) = \infty$. In the close vicinity of such singularities no meaningful clustering can take place, and also the use of ϵ-nets is infeasible. Therefore, we restrict ourselves to the subclass of so called μ-similar Bregman divergences that avoid these singularities.

μ-similar Bregman divergences feature some quasi-metric properties. That is, they are approximately symmetric and they obey the triangle inequality within a constant factor. Using these properties and a straightforward adaptation of the proofs from [8, 9] based on ϵ-nets, one finds that Chen's coreset construction leads to a $(k, \Theta(1))$-coreset S. However, subtle technical difficulties arise from the asymmetry if we have to show that S is a (k, ϵ)-coreset for arbitrarily small ϵ. Hence, even in the case of μ-singular Bregman divergences, the use of ϵ-nets to prove the existence of (k, ϵ)-coresets for the Bregman k-median problem seems infeasible.

Therefore, we concentrate on the construction of Γ-weak coresets. Our main contribution is the construction of an explicit, small Γ that includes all relevant medians. By bounding the combinatorial complexity of the approximate solutions computed by the algorithm from [1], as well as bounding the combinatorial complexity of Bregman k-median clusterings in general, we are able to prove the existence of such a small Γ by purely combinatorial means. In fact, our construction of weak corsets can be applied to any approximation algorithm if the combinatorial complexity of the algorithm's possible outputs is small.

Our approach is similar in spirit to the approach from [8, 9] for the metric k-median problem and the metric k-means problem, as well as to the approach from [12] for the Euclidean k-means problem.

2 Preliminaries

In this section we give a short introduction to Bregman divergences. We introduce our notion of μ-similarity and prove some important properties of μ-similar Bregman divergences. Furthermore, we derive a bound on the combinatorial complexity of (general) Bregman k-median clusterings.

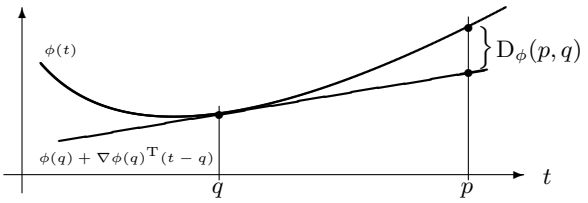

Figure 1: Geometric interpretation of a Bregman divergence.

2.1 Bregman divergences. The dissimilarity measures known as Bregman divergences were first proposed in 1967 by Lev M. Bregman [5]. For a study of Bregman divergences and their optimization problems see [7].

Intuitively, a Bregman divergence can be seen as the error when approximating a convex function by a tangent hyperplane (see Figure 1). We will use the following formal definition.

DEFINITION 2.1. *Let $\mathcal{X} \subseteq \mathbb{R}^d$. For any strictly convex, differentiable function $\phi : \mathcal{X} \to \mathbb{R}$ we define the* Bregman divergence *with respect to ϕ as*

$$\mathrm{D}_\phi(p, q) = \phi(p) - \phi(q) - \nabla\phi(q)^{\mathrm{T}}(p - q)$$

for $p, q \in \mathcal{X}$. Here $\nabla\phi(q)$ denotes the gradient of ϕ at point q.

Note that $\mathrm{D}_\phi(p, q)$ equals the tail of the first-order Taylor expansion of $\phi(p)$ at q. Bregman divergences include many prominent dissimilarity measures like the square of the ℓ_2-norm $\mathrm{D}_{\ell_2^2}(p, q) = \|p - q\|^2$ (with $\phi_{\ell_2^2}(t) = \|t\|^2$), the generalized Kullback-Leibler divergence $\mathrm{D}_{\mathrm{KL}}(p, q) = \sum p_i \ln \frac{p_i}{q_i} - \sum(p_i - q_i)$ (with $\phi_{\mathrm{KL}}(t) = \sum t_i \ln t_i - t_i$), and the Itakura-Saito divergence $\mathrm{D}_{\mathrm{IS}}(p, q) = \sum \frac{p_i}{q_i} - \ln \frac{p_i}{q_i} - 1$ (with $\phi_{\mathrm{IS}}(t) = -\sum \ln t_i$). We point out that, in general, Bregman divergences are asymmetric and do not satisfy the triangle inequality.

2.2 μ-similarity. Among the Bregman divergences one class of dissimilarity measures plays an important role in our work. For any positive definite matrix $A \in \mathbb{R}^{d \times d}$ the *Mahalanobis distance* with respect to A is defined as

$$\mathrm{D}_A(p, q) = (p - q)^{\mathrm{T}} A\,(p - q)$$

for $p, q \in \mathbb{R}^d$. The Mahalanobis distance was introduced in 1936 by P. C. Mahalanobis [20] based on the inverse of the correlation matrix of two random variables. The Mahalanobis distance is a Bregman divergence given

by the generating function $\phi_A(t) = t^{\mathrm{T}} A\, t$. Unlike most Bregman divergences, Mahalanobis distances are symmetric. Furthermore, they satisfy the following *double triangle inequality*.

LEMMA 2.1. *For a Mahalanobis distance* D_A *and for all* $p, q, r \in \mathbb{R}^d$ *we have* $\mathrm{D}_A(p,q) \leq 2\big(\mathrm{D}_A(p,r) + \mathrm{D}_A(r,q)\big)$.

Proof. Since A is positive definite there exists a non-singular matrix $B \in \mathbb{R}^{d\times d}$ with $A = B^{\mathrm{T}}B$. We obtain

$$\begin{aligned}\mathrm{D}_A(p,q) &= (p-q)^{\mathrm{T}} B^{\mathrm{T}} B\,(p-q)\\ &= (Bp - Bq)^{\mathrm{T}}(Bp - Bq)\\ &= \|Bp - Bq\|^2\\ &\leq 2\big(\|Bp - Br\|^2 + \|Br - Bq\|^2\big)\\ &= 2\big(\mathrm{D}_A(p,r) + \mathrm{D}_A(r,q)\big).\end{aligned}$$

Here the inequality $\|x-y\|^2 \leq 2\left(\|x-z\|^2 + \|z-y\|^2\right)$ holds for all $x, y, z \in \mathbb{R}^d$. □

To some extent, Mahalanobis distances are prototypical for many Bregman divergences that are used in practice. This observation is formalized in the following notion of μ-similarity, that has already been used in [1].

DEFINITION 2.2. *A Bregman divergence* D_ϕ *on domain* $\mathcal{X} \subseteq \mathbb{R}^d$ *is called* μ-similar *for positive real constant* μ *iff there exists a positive definite matrix* A *such that for Mahalanobis distance* D_A *and for each* $p, q \in \mathcal{X}$ *we have*

$$\mu\, \mathrm{D}_A(p,q) \leq \mathrm{D}_\phi(p,q) \leq \mathrm{D}_A(p,q).$$

Example: Let $0 < \lambda < \upsilon$. We consider the generalized Kullback-Leibler divergence $\mathrm{D}_{\mathrm{KL}}(p,q) = \sum p_i \ln \frac{p_i}{q_i} - \sum(p_i - q_i)$ on domain $\mathcal{X} = [\lambda, \upsilon]^d \subseteq \mathbb{R}^d$. We show that D_{KL} is μ-similar with $\mu = \frac{\lambda}{\upsilon}$ and $A = \frac{1}{2\lambda} I_d$.

To see this, note that for the strictly convex function $\phi_{\mathrm{KL}}(t) = \sum_{i=0}^{d} t_i \ln t_i - t_i$, 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

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

Here $\nabla^2\phi_{\mathrm{KL}}(\xi)$ denotes the Hessian matrix of ϕ_{KL} at point ξ. 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 $i \neq j$ the Hessian is a diagonal matrix and we obtain

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

Using $\frac{1}{\upsilon} \leq \frac{1}{\xi_i} \leq \frac{1}{\lambda}$ for each ξ_i, the observation follows. □

To the best of our knowledge, most Bregman divergences D_ϕ that are used in practice are μ-similar when restricted to a domain $\mathcal{X}$ that avoids the singularities of ϕ. A small overview of some μ-similar Bregman divergences is given in Figure 2. It is an important property that μ-similar Bergman divergences are approximately symmetric and satisfy a variant of the double triangle inequality, as is stated in the following lemma.

LEMMA 2.2. *For* μ*-similar Bregman divergence* D_ϕ *and for all* $p, q, r \in \mathcal{X}$ *we have* $\mathrm{D}_\phi(p,q) \leq \frac{1}{\mu}\mathrm{D}_\phi(q,p)$ *and* $\mathrm{D}_\phi(p,q) \leq \frac{2}{\mu}\big(\mathrm{D}_\phi(p,r) + \mathrm{D}_\phi(q,r)\big)$.

Proof. This lemma follows easily from Definition 2.2 and the symmetry and the double triangle inequality of Mahalanobis distances (Lemma 2.1). □

2.3 Combinatorial complexity of Bregman k-median clusterings. It is known that for all Bregman divergences every two clusters of an optimal k-median clustering solution are separated by a hyperplane (see Lemma 3.1 of [22]). Hence, a single cluster of such a solution is separated from the other $k-1$ clusters by at most $k-1$ hyperplanes. This property is formalized in the following definition.

DEFINITION 2.3. *We say set* $L \subseteq P$ *is a* j-linearly separable subset of P *iff there exist at most* j *oriented hyperplanes* $H_1, H_2, \ldots, H_j$ *such that*

i) $L \subseteq H_i^+$ *for all halfspaces* H_i^+ *defined by* H_i*, and*

ii) for $L^{\mathrm{c}} = P \setminus L$ *we have* $L^{\mathrm{c}} \cap \left(\bigcap_{i=1}^{j} H_i^+\right) = \emptyset$.

In terms of this notion, we find that for all Bregman divergences an optimal solution of the k-median clustering problem forms a partition by $(k-1)$-linearly separable subsets of P. We can use this observation to bound the combinatorial complexity of a given k-median clustering problem.

LEMMA 2.3. *Let* $P \subseteq \mathbb{R}^d$, $|P| = n$. *There are at most* n^{dk} $(k-1)$*-linearly separable subsets of* P.

Proof. Fix a single hyperplane H_i. There exists another hyperplane G_i defined by d points from P such that $P \cap H_i^+ = P \cap G_i^+$. This argument can be repeated for each of $k-1$ hyperplanes $H_1, H_2, \ldots, H_{k-1}$. Therefore, $L = P \cap \bigcap_{i=1}^{k-1} H_i^+$ is properly defined by selecting $d(k-1)$ points from P. It follows that the number of $(k-1)$-linearly separable subsets L of P is bounded by $\binom{n}{d(k-1)} \leq n^{dk}$. □

domain $\mathcal{X}$	$\phi(t)$	μ	A	$\mathrm{D}_\phi(p,q)$
$\mathbb{R}^d$	squared ℓ_2-norm $\|t\|_2^2$	1	I_d	squared Euclidean distance $\|p-q\|_2^2$
$\mathbb{R}^d$	generalized norm $t^{\mathrm{T}}At$	1	A	Mahalanobis distance $(p-q)^{\mathrm{T}}A(p-q)$
$[\lambda,\nu]^d \subseteq \mathbb{R}_+^d$	neg. Shannon entropy $\sum t_i \ln(t_i) - t_i$	$\frac{\lambda}{\nu}$	$\frac{1}{2\lambda} I_d$	Kullback-Leibler divergence $\sum p_i \ln(\frac{p_i}{q_i}) - \sum(p_i - q_i)$
$[\lambda,\nu]^d \subseteq \mathbb{R}_+^d$	Burg entropy $-\sum \ln(t_i)$	$\frac{\lambda^2}{\nu^2}$	$\frac{1}{2\lambda^2} I_d$	Itakura-Saito divergence $\sum \frac{p_i}{q_i} - \ln(\frac{p_i}{q_i}) - 1$
$[\lambda,\nu]^d \subseteq \mathbb{R}_+^d$	harmonic ($\alpha > 0$) $\sum \frac{1}{t_i^\alpha}$	$\frac{\lambda^{\alpha+2}}{\nu^{\alpha+2}}$	$\frac{\alpha(\alpha-1)}{2\lambda^{\alpha+2}} I_d$	harmonic divergence ($\alpha > 0$) $\sum \frac{1}{p_i^\alpha} - \frac{\alpha+1}{q_i^\alpha} + \frac{\alpha p_i}{q_i^{\alpha+1}}$
$[\lambda,\nu]^d \subseteq \mathbb{R}_+^d$	norm-like ($\alpha \geq 2$) $\sum t_i^\alpha$	$\frac{\lambda^{\alpha-2}}{\nu^{\alpha-2}}$	$\frac{\alpha(\alpha-1)}{2} \nu^{\alpha-2} I_d$	norm-like divergence ($\alpha \geq 2$) $\sum p_i^\alpha + (\alpha-1) q_i^\alpha - \alpha p_i q_i^{\alpha-1}$
$[\lambda,\nu]^d \subseteq \mathbb{R}^d$	exponential $\sum e^{t_i}$	$e^{-(\nu-\lambda)}$	$\frac{e^\nu}{2} I_d$	Exponential loss $\sum e^{p_i} - (p_i - q_i + 1) e^{q_i}$
$[-\nu,\nu]^d \subseteq (-1,1)^d$	Hellinger-like $-\sum \sqrt{1-t_i^2}$	$2(1-\nu^2)^{\frac{3}{2}}$	$\frac{1}{2(1-\nu^2)^{\frac{3}{2}}} I_d$	Hellinger-like divergence $\sum \frac{1-p_i q_i}{\sqrt{1-q_i^2}} - \sqrt{1-p_i^2}$

Figure 2: Some μ-similar Bregman divergences.

In [4], Banerjee et al. show a remarkable connection between Bregman divergences and Lloyd's well known k-means heuristic [19]. Among other results, they showed that for each Bregman divergence the optimal 1-median of any given set S is uniquely defined by the centroid $c_S = \frac{1}{|S|}\sum_{p\in S} p$ of set S (also known as the center of gravity of S). Using Lemma 2.3, we conclude that for any given input set $P \subseteq \mathbb{R}^d$ at most n^{dk} points from $\mathbb{R}^d$ have to be considered as one of the optimal k-medians of P.

COROLLARY 2.1. *Let Γ_P be the set of all centroids of all $(k-1)$-linearly separable subsets of P. Then $|\Gamma_P| \leq n^{dk}$.*

3 Coreset construction

In this section we present a weak coreset construction applicable to all μ-similar Bregman divergences. In particular, we show how to construct a Γ-weak coreset for an arbitrary but fixed and finite Γ.

3.1 Γ-weak coresets. Recently, coresets have emerged as a standard technique in computational geometry. Generally speaking, a coreset is a small weighted set S that maintains the same clustering behavior as the large input set P. Such a coreset can be used as a smaller input set for an approximation algorithm.

Let D_ϕ be a μ-similar Bregman divergence on domain $\mathcal{X} \subseteq \mathbb{R}^d$. Coresets are usually defined such that the weighted sum of errors $\mathrm{cost}_w(S,C) = \sum_{s\in S} w(s)\,\mathrm{D}_\phi(s,C)$ is a $(1\pm\epsilon)$-approximation of $\mathrm{cost}(P,C)$ *for any set* C of size $|C| = k$. However, if we recall that the combinatorial complexity of Bregman clusterings is bounded this seems to be an unnecessarily strict demand.

Therefore, we will use a relaxed notion of coresets where only center points from a finite but significant subset $\Gamma \subseteq \mathcal{X}$ are considered. We call this a Γ-weak coreset. Weak coresets have already been used in [12]. However, our notion differs slightly from the previous definition.

DEFINITION 3.1. *Let $P \subseteq \mathcal{X}$ and $\Gamma \subseteq \mathcal{X}$. A weighted multiset $S \subseteq \mathcal{X}$ with weight function $w : S \to \mathbb{R}_{\geq 0}$ such that $\sum_{s\in S} w(s) = |P|$ is called a* Γ-weak (k,ϵ)-coreset *of P iff for all $C \subseteq \Gamma$ of size $|C| = k$ we have*

$$|\mathrm{cost}(P,C) - \mathrm{cost}_w(S,C)| \leq \epsilon\,\mathrm{cost}(P,C)\ . \tag{3.1}$$

3.2 Chens's coreset construction. We make use of Chen's coreset construction from [8, 9]. Let $A = \{a_1, a_2, \ldots, a_\kappa\}$ be the medians of an $[\alpha,\beta]$-bicriteria approximation of a k-median D_ϕ-clustering of P, i.e., $\mathrm{cost}(P,A) \leq \alpha\, opt_k(P)$ and $|A| = \kappa \leq \beta k$. A simple algorithm to obtain such a bicriteria approximation for μ-similar Bregman divergences is given in the appendix.

Let $P_1, P_2, \ldots, P_\kappa$ be the partition of P induced by assigning each $p \in P$ to their closest $a_i \in A$, i.e. $p \in P_i$ iff $a_i = \arg\min_{a\in A} \mathrm{D}_\phi(p,a)$, breaking ties arbitrarily. Furthermore, let $R = \frac{1}{\alpha n}\mathrm{cost}(P,A)$. Note that $R \leq \frac{1}{n}\, opt_k(P)$. Let $B_r(a_i) = \{x \mid \mathrm{D}_\phi(x,a_i) \leq r\}$ denote the D_ϕ-ball of radius r centered at a_i.

Using A we define a partition $\{P_{ij}\}_{i,j}$ of P by

$$P_{i0} = P_i \cap B_R(a_i)$$

for $i = 1, 2, \ldots, \kappa$ and

$$P_{ij} = P_i \cap \big(B_{2^j R}(a_i) \setminus B_{2^{j-1} R}(a_i)\big)$$

for $i = 1, 2, \ldots, \kappa$ and $j = 1, 2, \ldots, \nu$ where $\nu = \lceil \log(\alpha n) \rceil$. Note that $\{P_{ij}\}_{i,j}$ is indeed a partition of P since the existence of a $p \in P$ with $\mathrm{D}_\phi(p, A) > 2^\nu R$ leads to $\mathrm{cost}(P, A) > 2^\nu R \geq \alpha n R = \mathrm{cost}(P, A)$ which is a contradiction.

For i, j let S_{ij} be a uniform sample multiset from P_{ij} of size $|S_{ij}| = m$. Let $w(s) = \frac{1}{m}|P_{ij}|$ be the weight associated with $s \in S_{ij}$. We define $S = \bigcup_{i,j} S_{ij}$ of size $|S| = m\kappa\nu = m\beta k \lceil \log(\alpha n) \rceil$ as our weak coreset. We will show the following theorem.

THEOREM 3.1. *Let $\Gamma \subseteq \mathcal{X}$ be an arbitrary but fixed and finite set. If $m = \Omega\left(\frac{\alpha^2}{\epsilon^2} \log\big(\frac{\beta}{\delta} k |\Gamma|^k \log(\alpha n)\big)\right)$ then with probability $1 - \delta$ the weighted multiset S is a Γ-weak (k, ϵ)-coreset of P.*

COROLLARY 3.1. *Given a set of medians of an $[\alpha, \beta]$-approximate k-median clustering of P for constants α, β, with high probability a Γ-weak (k, ϵ)-coreset of P of size $\mathcal{O}\left(\frac{1}{\epsilon^2} k \log(n) \log(k|\Gamma|^k \log n)\right)$ can be constructed in time $\mathcal{O}\left(dkn + \frac{1}{\epsilon^2} k \log(n) \log(k|\Gamma|^k \log n)\right)$.*

3.3 Proof of Theorem 3.1. We will make use of the following probabilistic concentration bound, given by Haussler [16].

LEMMA 3.1. ([16]) *Let be $f : P \to \mathbb{R}$ and $F \in \mathbb{R}$ such that we have $0 \leq f(p) \leq F$ for all $p \in P$. Let $S \subseteq P$ be a uniform sample multiset of size $|S| \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$ for constant positive reals ϵ, δ. Then we have*

$$\Pr\left[\left|\frac{1}{|P|}\sum_{p \in P} f(p) - \frac{1}{|S|}\sum_{s \in S} f(s)\right| \leq \epsilon F\right] \geq 1 - \delta \ .$$

Our strategy to prove Theorem 3.1 is as follows. First, we prove inequality (3.1) with high probability for an arbitrary but fixed set C of size k. Subsequently, we use the union bound to show that with probability at least $1 - \delta$ inequality (3.1) is satisfied for all $C \subseteq \Gamma$ of size k.

LEMMA 3.2. *Let $C \subseteq \mathcal{X}$ be a fixed set of size $|C| = k$. If $m \geq \frac{392\alpha^2}{\epsilon^2\mu^2} \ln\left(\frac{2\kappa\nu|\Gamma|^k}{\delta}\right)$ then with probability $1 - \frac{\delta}{|\Gamma|^k}$ we have*

$$|\mathrm{cost}(P, C) - \mathrm{cost}_w(S, C)| \leq \epsilon\, \mathrm{cost}(P, C) \ .$$

Proof. Fix i, j. For all $p \in P_{ij}$ define the function f_{ij} by $f_{ij}(p) = \mathrm{D}_\phi(p, C)$. Let $q^* \in P_{ij}$ denote an input that minimizes f_{ij}. Analogously to Lemma 2.2, we have

$$\begin{aligned} 0 \leq f_{ij}(p) &\leq \frac{4}{\mu}\big(\mathrm{D}_\phi(q^*, C) + \mathrm{D}_\phi(q^*, a_i) + \mathrm{D}_\phi(p, a_i)\big) \\ &\leq \frac{4}{\mu}\big(\mathrm{D}_\phi(q^*, C) + 2^{j+1} R\big) \end{aligned}$$

for all $p \in P_{ij}$. Hence, by Lemma 3.1 with probability $1 - \frac{\delta}{\kappa\nu|\Gamma|^k}$ we have

$$\begin{aligned} &\left|\frac{1}{|P_{ij}|}\,\mathrm{cost}(P_{ij}, C) - \frac{1}{|P_{ij}|}\,\mathrm{cost}_w(S_{ij}, C)\right| \\ &\qquad \leq \frac{\epsilon}{7\alpha}\big(\mathrm{D}_\phi(q^*, C) + 2^{j+1} R\big) \ . \end{aligned}$$

By construction of P_{ij}, for $j \geq 1$ we have $|P_{ij}| 2^{j+1} R \leq 4\,\mathrm{cost}(P_{ij}, A)$. For $j = 0$ we have $|P_{ij}| 2^{j+1} R = 2\,|P_{i0}| R$. We obtain $|P_{ij}| 2^{j+1} R \leq 4\,\mathrm{cost}(P_{ij}, A) + 2\,|P_{ij}| R$ for all $j \geq 0$. Therefore,

$$\begin{aligned} &|\mathrm{cost}(P_{ij}, C) - \mathrm{cost}_w(S_{ij}, C)| \\ &\quad \leq \frac{\epsilon}{7\alpha}\big(|P_{ij}|\,\mathrm{D}_\phi(q^*, C) + |P_{ij}| 2^{j+1} R\big) \\ &\quad \leq \frac{\epsilon}{7\alpha}\big(\mathrm{cost}(P_{ij}, C) + 4\,\mathrm{cost}(P_{ij}, A) + 2\,|P_{ij}| R\big) \ . \end{aligned}$$

Summing up over all i, j, with probability $1 - \frac{\delta}{|\Gamma|^k}$ we have

$$\begin{aligned} &|\mathrm{cost}(P, C) - \mathrm{cost}_w(S, C)| \\ &\leq \sum_{i,j} |\mathrm{cost}(P_{ij}, C) - \mathrm{cost}_w(S_{ij}, C)| \\ &\leq \frac{\epsilon}{7\alpha}\Big(\sum_{i,j} \mathrm{cost}\,(P_{ij}, C) + 4\sum_{i,j} \mathrm{cost}\,(P_{ij}, A) + 2R\sum_{i,j} |P_{ij}|\Big) \\ &= \frac{\epsilon}{7\alpha}\big(\mathrm{cost}(P, C) + 4\,\mathrm{cost}(P, A) + 2nR\big) \\ &\leq \frac{\epsilon}{7\alpha}\big(\mathrm{cost}(P, C) + 4\alpha\, opt_k(P) + 2\, opt_k(P)\big) \\ &\leq \frac{\epsilon}{7\alpha}\big(7\alpha\,\mathrm{cost}(P, C)\big) = \epsilon\,\mathrm{cost}(P, C) \ . \end{aligned}$$

□

By Lemma 3.2, for a fixed choice of $C \subseteq \Gamma$ we have inequality (3.1) with probability $1 - \frac{\delta}{|\Gamma|^k}$. Since there are at most $\binom{|\Gamma|}{k} \leq |\Gamma|^k$ subsets $C \subseteq \Gamma$ of size k we obtain that with probability $1 - \delta$ the weighted multiset S is a Γ-weak (k, ϵ)-coreset, proving Theorem 3.1. □

4 Application to Bregman k-median clustering

In this section we use our Γ-weak coresets to improve the asymptotic running time of an existing $(1 + \epsilon)$-approximation algorithm for the Bregman k-median

clustering problem. In particular, we use a simple adaptation of the $(1+\epsilon)$-approximation algorithm CLUSTER from [1] that works on a weighted input set. To improve the running time, we construct a Γ-weak coreset with respect to a carefully chosen, small Γ that includes all medians relevant to algorithm CLUSTER. The definition of this Γ is the most important part of this section.

Algorithm CLUSTER has already been shown to be applicable for all μ-similar Bregman divergences. The details of this algorithm are not important to us (the reader is directed to [1] for an in-detail description). We only need the following result.

LEMMA 4.1. *Fix input set P and bicriteria approximation A from the coreset construction. Then there exists a set Γ_{CLUSTER} of size $|\Gamma_{\text{CLUSTER}}| \leq n^{(\frac{k}{\epsilon})^{\Theta(1)}}$ such that for every possible (S, w) from Chen's coreset construction applied to P and A, and for every possible output C of algorithm* CLUSTER *started with weighted set S, we have $C \subseteq \Gamma_{\text{CLUSTER}}$.*

The proof of this lemma is somewhat technical and requires some insight into operation of algorithm CLUSTER. However, the main idea of the proof is straightforward. Any cluster center $c \in C$ computed by algorithm CLUSTER is the weighted centroid of a $(\frac{k}{\epsilon})^{\Theta(1)}$-sized subset of S. Since the number of different weights that can appear in applications of algorithm CLUSTER to coresets is small, the number of all possible weighted subsets defining a cluster center c can be bounded by $n^{(\frac{k}{\epsilon})^{\Theta(1)}}$. The full proof of the lemma is given in the appendix.

Now we can give the definition of Γ explicitly. Let C_P denote the optimal k-medians of P and let

$$\Gamma = \Gamma_{\text{CLUSTER}} \cup C_P.$$

By Lemma 4.1 we have the following bound on $|\Gamma|$.

LEMMA 4.2. $|\Gamma| \leq n^{(\frac{k}{\epsilon})^{\Theta(1)}}$

Proof. We know $|C_P| = k$. By Lemma 4.1 we have $|\Gamma_{\text{CLUSTER}}| \leq n^{(\frac{k}{\epsilon})^{\Theta(1)}}$. Therefore, $|\Gamma| \leq k + n^{(\frac{k}{\epsilon})^{\Theta(1)}} \leq n^{(\frac{k}{\epsilon})^{\Theta(1)}}$. □

Note that the definition of Γ depends only on P and A, and is independent of the random choices made during the construction of coreset S. Also note that we do not have to know the exact content of set Γ to construct a Γ-weak coreset using Chen's construction: We only need a size bound on Γ, and this bound is given by Lemma 4.2. Hence, we can state the following approximation algorithm for the Bregman k-median clustering problem.

Algorithm CoreCluster(P, k)**:**

1: *Obtain $\mathcal{O}(\log k)$-approximation A using algorithm* BREGMEANS++(P, k) *(see Appendix).*
2: *Build Γ-weak (k, ϵ)-coreset S of P, using A and Chen's coreset construction.*
3: *Run adaptation of algorithm* CLUSTER *from [1] on weighted input set S to obtain $(1+\epsilon)$-approximate k-median set C.*

THEOREM 4.1. *With constant probability, algorithm* CORECLUSTER *computes a solution C of the k-median problem with respect to μ-similar Bregman divergence D_ϕ for input instance P of size $|P| = n$ satisfying*

$$\text{cost}(P, C) \leq (1 + 7\epsilon)\, opt_k(P)$$

in time $\mathcal{O}\big(dkn + d2^{(\frac{k}{\epsilon})^{\Theta(1)}} \log^{k+2}(n)\big)$.

Proof. Since each step of our algorithm succeeds at least with constant probability, we may assume that with constant probability all three steps yield the desired result.

Let C_P denote the optimal k-medians for P and let C_S denote the optimal k-medians for weighted set S, i.e. $\text{cost}(P, C_P) = opt_k(P)$ and $\text{cost}_w(S, C_S) = opt_k(S)$. Using $C \subseteq \Gamma_{\text{CLUSTER}}$ and the fact that S is a Γ-weak (k, ϵ)-coreset we obtain

$$\text{cost}(P, C) \leq \frac{1}{1-\epsilon} \text{cost}_w(S, C).$$

Since C is a $(1+\epsilon)$-approximation for weighted input set S we get

$$\text{cost}(P, C) \leq \frac{1+\epsilon}{1-\epsilon} \text{cost}_w(S, C_S) \leq \frac{1+\epsilon}{1-\epsilon} \text{cost}_w(S, C_P).$$

Using $C_P \subseteq \Gamma$ and (3.1) we obtain

$$\text{cost}(P, C) \leq \frac{(1+\epsilon)^2}{1-\epsilon} \text{cost}(P, C_P) \leq (1 + 7\epsilon)\, opt_k(P)$$

since $\frac{(1+\epsilon)^2}{1-\epsilon} \leq 1 + 7\epsilon$ for $\epsilon \leq \frac{1}{2}$.

For the analysis of the running time, assume that we can sample points from a given set in time $\mathcal{O}(1)$. Hence, approximation A can be obtained in time $\mathcal{O}(dkn)$. Furthermore, coreset S can be constructed in time $\mathcal{O}(dkn + |S|)$. The adaptation of algorithm CLUSTER for a weighted coreset S with $\sum_{s \in S} w(s) = n$ has a running time $T(n, k)$ given by the recurrence

$$T(n,k) \leq 2^{(\frac{k}{\epsilon})^{\Theta(1)}} T(n, k-1) + T(\frac{n}{2}, k) + d\big(2^{(\frac{k}{\epsilon})^{\Theta(1)}} + |S|\big).$$

This recurrence is of the type

$$T(l, j) \leq r\, T(l, j-1) + T(l-1, j) + c$$

for constants $r = 2^{(\frac{k}{\epsilon})^{\Theta(1)}}$ and $c = d\big(2^{(\frac{k}{\epsilon})^{\Theta(1)}} + |S|\big)$ where the first parameter of T is replaced by its logarithm.

CLAIM 4.1. $T(l,j) \leq c\,r^j l^j$

Proof. We show the claim by induction. For $l = 1$ or $j < 2$ we have constant running time and we get $T(l,j) \leq c$ lor large enough constant c. Hence, let $l, j \geq 2$. By induction hypothesis we have

$$\begin{aligned} T(l,j) &\leq r\big(c\,r^{j-1}l^{j-1}\big) + c\,r^j(l-1)^j + c \\ &\leq c\,r^j\big(l^{j-1} + (l-1)^j + 1\big)\ . \end{aligned}$$

Since $l, j \geq 2$ we have $l^{j-1} + (l-1)^j + 1 \leq l^j$. Thus, we have $T(l,j) \leq c\,r^j l^j$. □

Using Claim 4.1, $l = \log n$, and $j = k$ we obtain

$$\begin{aligned} T(n,k) &\leq d\big(2^{(\frac{k}{\epsilon})^{\Theta(1)}} + |S|\big)\, 2^{(\frac{k}{\epsilon})^{\Theta(1)}} \log^k(n) \\ &= \mathcal{O}\big(|S|\, d\, 2^{(\frac{k}{\epsilon})^{\Theta(1)}} \log^k(n)\big). \end{aligned}$$

Using $|S| = \mathcal{O}\big((\frac{k}{\epsilon})^{\Theta(1)} \log^2(n)\big)$ from Theorem 3.1 and Lemma 4.2 we obtain the desired running time. □

5 Conclusion

We have shown that there exist small weak corsets for the Bregman k-median problem. We have shown how to use such weak coresets to significantly speed-up an existing approximation algorithm. In doing so, we presented the currently asymptotically fastest algorithm known for the k-median problem with respect to a number of Bregman divergences such as the Kullback-Leibler divergence and the Itakura-Saito divergence. Due to the low dependency of the running time on d this algorithm is particularly relevant for high-dimensional settings. It is noteworthy that our weak corsets can be applied to any approximation algorithm if the combinatorial complexity of the algorithm's possible outputs is small. We also point out that using the merge-and-reduce technique from [15] our weak coresets can be applied to the data streaming model.

However, some open problems remain. Considering the results from [14] and [8, 9] the question arises whether we can construct Bregman coresets that are independent of n or d. Also, are there "strong" coresets for the Bregman k-median problem? What further techniques from computational geometry can be applied to Bregman divergences? Furthermore, it remains an open problem whether there exists a $(1+\epsilon)$-approximation algorithm for Bregman divergences with singularities in their domain.

Finally, it is still unclear whether there are $(1+\epsilon)$-approximation algorithms and coreset constructions for many other important non-Bregman, non-metric dissimilarity measures such as Pearson's correlation or cosine similarity.

Acknowledgments. We thank Christian Sohler for helpful discussion.

References

[1] Marcel R. Ackermann, Johannes Blömer, and Christian Sohler. Clustering for metric and non-metric distance measures. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '08)*, pages 799–808. Society for Industrial and Applied Mathematics, 2008.

[2] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[3] 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.

[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] Andrés Buzo, Augustine H. Gray, Jr., Robert M. Gray, and John D. Markel. Speech coding based upon vector quantization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(5):562–574, Oct 1980.

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

[8] 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.

[9] Ke Chen. On k-median and k-means clustering in metric and Euclidean spaces and their applications. Manuscript, available at: `http://ews.uiuc.edu/~kechen/`, July 2007.

[10] Artur Czumaj and Christian Sohler. Sublinear-time approximation for clustering via random sampling. In *Proceedings of the 31st International Colloquium on Automata, Languages and Programming (ICALP'04)*, pages 396–407, 2004.

[11] 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.

[12] 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.

[13] Gereon Frahling and Christian Sohler. Coresets in dynamic geometric data streams. In *Proceedings of the 27th Annual ACM Symposium on Theory of Computing (STOC'05)*, pages 209–217, New York, NY, USA, 2005. ACM.

[14] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. In *Proceedings of the 21st Annual Symposium on Computational Geometry (SCG'05)*, pages 126–134, New York, NY, USA, 2005. ACM.

[15] 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.

[16] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

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

[18] 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.

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

[20] 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.

[21] Nina Mishra, Dan Oblinger, and Leonard Pitt. Sublinear time approximate clustering. In *Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'01)*, pages 439–447, 2001.

[22] Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. On Bregman Voronoi diagrams. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'07)*, pages 746–755, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

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

A Appendix

A.1 A constant factor approximation algorithm. In this section, we show how to construct a factor $\mathcal{O}(\log k)$-approximation for the k-median problem with respect to a μ-similar Bregman divergence D_ϕ. The following algorithm can be used to obtain an initial $[\mathcal{O}(\log k), 1]$-bicriteria approximation necessary for our coreset construction (Section 3).

As approximation algorithm, we use a non-uniform random sampling approach from [2]. This sampling approach has been originally proposed in the context of Euclidean k-means clustering, as well as for the k-median problem using a j-th power of the ℓ_2-norm as distance measure. We can show that the approach is also applicable to μ-similar Bregman k-median clusterings.

Algorithm BregMeans++(P, k):

1: *Choose an initial point a_1 uniformly at random from P.*
2: *Let A be the set of already chosen points from P. Then element $p \in P$ is chosen with probability $\frac{\mathrm{D}_\phi(p,A)}{\mathrm{cost}(P,A)}$ as next element of A.*
3: *Repeat step 2 until A contains k points.*

We say $A = \{a_1, a_2, \ldots, a_k\}$ is chosen at random according to D_ϕ. We can prove the following theorem using the μ-similarity of D_ϕ and a slight modification of the proof of Theorem 3.1 from [2].

THEOREM A.1. *If D_ϕ is a μ-similar Bregman divergence and $A \subseteq \mathcal{X}$ with $|A| = k$ is chosen according to D_ϕ, we have*

$$\mathrm{E}\left[\mathrm{cost}(P, A)\right] \leq \frac{8}{\mu^2}(2 + \ln k)\, opt_k(P) \ .$$

It is an easy application of Markov's inequality to see that, with high probability, algorithm BREGMEANS++ yields a constant factor approximation (regarding k as a constant) of $opt_k(P)$.

A.2 Details on the adaptation of the algorithm from [1]. Algorithm CLUSTER is a generalization of an earlier algorithm from [18]. It has been shown that with constant probability this algorithm computes a $(1+\epsilon)$-approximation for any μ-similar Bregman divergence.

In a nutshell, algorithm CLUSTER for unweighted point sets works as follows. We leave out most technical details which are explained in [1].

1. Using the superset sampling technique from [18] a first approximate median for a "large" cluster P_1 is found. More precisely, this approximate median is constructed as the centroid of a uniform sample set of size $(\frac{k}{\epsilon})^{\Theta(1)}$ from P.

2. Let $\tilde{C}$ be the set of already found approximate medians. If there is no "large" cluster whose

approximate median has yet to be found, then iteratively the $\frac{n}{2}, \frac{n}{4}, \frac{n}{8}, \ldots$ points closest to C are pruned from the point set. We keep removing points until a new cluster becomes "large" within the remaining point set.

3. Since a new cluster has become "large", we find its approximate median using the superset sampling technique. Again, this approximate median is constructed as the centroid of a uniform sample set of size $(\frac{k}{\epsilon})^{\Theta(1)}$ from P.

4. Steps 2.-3. are repeated until for each cluster an approximate median has been found.

For weighted input sets, only slight modifications to this algorithm are necessary. First, instead of uniform random sampling a point $p \in R$ is sampled with probability $\frac{w(p)}{w(R)}$, where $w(R) = \sum_{p \in R} w(p)$ denotes the total weight of R. Second, when computing the centroid of a set M', the weighted centroid $\sum_{p \in M'} \frac{w(p)}{w(M')}\, p$ has to be computed. Third, recall that $w(S) = n$. Instead of removing the $\frac{n}{2}, \frac{n}{4}, \frac{n}{8}, \ldots$ closest points from the point set, the closest points with a total weight of $\frac{n}{2}, \frac{n}{4}, \frac{n}{8}, \ldots$ are pruned. However, from time to time the weight of the closest points will not add up to exactly $\frac{n}{2^i}$. In this case, a single point p has to be replaced by two copies p_1, p_2 with $w(p) = w(p_1) + w(p_2)$ such that we can find a partition with total weight $\frac{n}{2^i}$.

The pseudocode of the adaptation of algorithm CLUSTER from [1] for weighted input sets is given in Figure 3.

A.3 Proof of Lemma 4.1. Since P and A are fixed so is the partition $\{P_{ij}\}_{i,j}$ of P from the coreset construction. Let S be a weighted multiset with weight function w obtained by the coreset construction from Section 3. Let m be the constant number of elements uniformly sampled from each P_{ij} to obtain S.

First, let us ignore the weight function w. Recall that each approximate median from the output of CLUSTER is obtained as the (weighted) centroid of a subset of size $(\frac{k}{\epsilon})^{\Theta(1)}$ from S. There are at most $\binom{|S|}{(\frac{k}{\epsilon})^{\Theta(1)}} \leq n^{(\frac{k}{\epsilon})^{\Theta(1)}}$ such subsets.

Now let us consider weight function w. Since $|P_{ij}|$ and m are fixed so is the initial weight of each point from input set S. Therefore, the number of weighted centroids formed by the $(\frac{k}{\epsilon})^{\Theta(1)}$-sized subsets of input set S with the initial weight function w is bounded by $n^{(\frac{k}{\epsilon})^{\Theta(1)}}$. However, since sometimes the weight of a point is split in the pruning step of the algorithm the weight of some points will change. So we have to analyze the number of different weights that may be assigned to point $s \in S$ during a run of algorithm CLUSTER.

CLUSTER$(R, w, l, \tilde{C})$:
- R set of remaining input points of total weight $w(R) = n$
- w weight function on R
- j number of medians yet to be found
- 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 M of size $\frac{96k^2}{\epsilon^2 \mu \delta}$ from R
    $T \leftarrow \left\{c \mid c \text{ wghtd. centroid of } M' \subseteq M,\ |M'| = \frac{3}{\epsilon \mu \delta}\right\}$
    for all $\tilde{c} \in T$ **do**
      $C^{(\tilde{c})} \leftarrow$ CLUSTER$(R, w, j-1, \tilde{C} \cup \{\tilde{c}\})$
    end for
    / pruning phase */*
    partition R into set N and $R \setminus N$ such that:
      $\circ\ \forall p \in N, q \in R \setminus N : \mathrm{D}_\phi(p, \tilde{C}) \leq \mathrm{D}_\phi(q, \tilde{C})$ and
      $\circ\ w(N) = w(R \setminus N) = \frac{n}{2}$ (if necessary, split a point)
    let $\tilde{w}$ be the new weight function on $R \setminus N$
    $C^* \leftarrow$ CLUSTER$(R \setminus N, \tilde{w}, j, \tilde{C})$
    return $C^{(\tilde{c})}$ or C^* with minimum cost
  end if
end if

Figure 3: Adaptation of algorithm CLUSTER for weighted sets and μ-similar Dregman divergences D_ϕ.

Observe that the behavior and output of algorithm CLUSTER will not change when the weight function of the input set is scaled by a constant. Therefore, let us consider set S with weight function $\hat{w}$ such that $\hat{w}(s) = m\, w(s)$. Since for $s \in P_{ij}$ we have $w(s) = \frac{1}{m}|P_{ij}|$ it follows that function $\hat{w}$ has only integral weights $\hat{w}(s) = |P_{ij}|$. Hence, there is a one-to-one correspondence to a run of algorithm CLUSTER on unweighted input multiset $\hat{S}$ where each $s \in S$ is replaced by $\hat{w}(s)$ copies of s.

Since splitting of weights for a point from S corresponds to the situation when some points from $\hat{S}$ are pruned and some are not we find that the weights of $\hat{w}$ remain integral during a run of algorithm CLUSTER. Hence there are at most $|P_{ij}| \leq n$ different weights assigned to point $s \in P_{ij}$.

We conclude that there are at most $n^{(\frac{k}{\epsilon})^{\Theta(1)}}$ different weight functions assigned to a $(\frac{k}{\epsilon})^{\Theta(1)}$-sized subsets of input set S. Hence, there are at most $n^{(\frac{k}{\epsilon})^{\Theta(1)}}$ weighted centroids of such a fixed subset of S. It follows that the number of all possible output points from algorithm CLUSTER is bounded by $\left(n^{(\frac{k}{\epsilon})^{\Theta(1)}}\right)^2 = n^{(\frac{k}{\epsilon})^{\Theta(1)}}$. □