

# Worst-Case and Smoothed Analysis of $k$ -Means Clustering with Bregman Divergences\*

Bodo Manthey<sup>†</sup>

University of Twente  
Department of Applied Mathematics  
P. O. Box 217, 7500 AE Enschede  
The Netherlands  
b.manthey@utwente.nl

Heiko Röglin<sup>‡</sup>

Maastricht University  
Department of Quantitative Economics  
P. O. Box 616, 6200 MD Maastricht  
The Netherlands  
heiko@roeglin.org

April 7, 2010

The  $k$ -means algorithm is the method of choice for clustering large-scale data sets and it performs exceedingly well in practice despite its exponential worst-case running-time. To narrow the gap between theory and practice,  $k$ -means has been studied in the semi-random input model of smoothed analysis, which often leads to more realistic conclusions than mere worst-case analysis. For the case that  $n$  data points in  $\mathbb{R}^d$  are perturbed by Gaussian noise with standard deviation  $\sigma$ , it has been shown that the expected running-time is bounded by a polynomial in  $n$  and  $1/\sigma$ . This result assumes that squared Euclidean distances are used as distance measure.

In many applications, however, data is to be clustered with respect to Bregman divergences rather than squared Euclidean distances. A prominent example is the Kullback-Leibler divergence (a.k.a. relative entropy) that is commonly used to cluster web pages. To broaden the knowledge about this important class of distance measures, we analyze the running-time of the  $k$ -means method for Bregman divergences. We first give a smoothed analysis of  $k$ -means with (almost) arbitrary Bregman divergences, and we show bounds of  $\text{poly}(n^{\sqrt{k}}, 1/\sigma)$  and  $k^{kd} \cdot \text{poly}(n, 1/\sigma)$ . The latter yields a polynomial bound if  $k$  and  $d$  are small compared to  $n$ . On the other hand, we show that the exponential lower bound carries over to a huge class of Bregman divergences.

---

\*A preliminary version of this paper appeared in *Proc. 20th Int. Symp. on Algorithms and Computation (ISAAC 2009)*, vol. 5878 of *Lecture Notes in Computer Science*, pp. 1024–1033, Springer 2009. This paper also includes parts of “Improved Smoothed Analysis of the  $k$ -Means Method” [15].

<sup>†</sup>Work done in part at Saarland University, Department of Computer Science, Postfach 151150, 66041 Saarbrücken, Germany.

<sup>‡</sup>Supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD).

# 1 Introduction

Clustering a set of objects into a certain number of classes so as to maximize the similarity of objects in the same class is a fundamental problem with applications in various areas like information retrieval, bioinformatics, and data compression. Usually the objects are represented by points in  $\mathbb{R}^d$ , and they are to be clustered into  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$  that can be represented by centers  $c_1, \dots, c_k \in \mathbb{R}^d$  such that the sum  $\sum_{i=1}^k \sum_{x \in \mathcal{C}_i} d(x, c_i)$  becomes minimal for some distance measure  $d$ . A common distance function  $d$  are squared Euclidean distances but in many practical applications other distance measures are required. For instance, when clustering text documents like web pages often the *bag-of-words model* [7] is applied, in which the objects to be clustered are probability distributions over the set of all words. A popular distance measure for probability distributions is the *Kullback-Leibler divergence* (KLD, also known as relative entropy). Both squared Euclidean distances and KLD are special cases of *Bregman divergences*, a very general class that contains most practically important distance measures.

Even though a lot of theoretical research has been conducted on clustering algorithms, the by far most successful algorithm in industrial and scientific applications is the seemingly ad hoc *k-means method* [6], a local search algorithm due to Lloyd [14]: Start with an arbitrary set of  $k$  centers and repeat the following two steps until the process stabilizes: 1) Assign every data point to its closest center. 2) Readjust the positions of the centers such that they are optimal for the current assignment. The *k-means method* works very well in practice. One of its distinguished features is its speed: It has been observed that the number of iterations it needs to find a local optimum is much smaller than the number of objects to be clustered [8, Section 10.4.3]. This is in stark contrast to its worst-case running-time: The only upper bound is  $n^{O(kd)}$  [12], which is based on the observation that no clustering appears twice in a run of *k-means*. On the other hand, Vattani [18] showed that *k-means* can run for  $2^{\Omega(n)}$  iterations in the worst case. This lower bounds holds for all  $d \geq 2$ .

To reconcile theory and practice, Arthur and Vassilvitskii considered the *k-means method* for squared Euclidean distances in the framework of *smoothed analysis*. This notion has been introduced by Spielman and Teng [17] and it is based on a two-step input model: An adversary specifies an instance, which is then subject to slight random perturbation. The smoothed running-time is defined to be the worst expected running-time the adversary can achieve. If it is small, then (artificial) worst-case instances might still exist, but they are encountered only with very small probability if inputs are subject to some small amount of random noise. In practice, such noise can come, e.g., from measurement errors or numerical imprecision. Unlike worst-case or average-case analyses, smoothed analyses are neither dominated by single worst-case instances nor by completely random instances, and they lead to more realistic conclusions. Arthur and Vassilvitskii [4] showed that the smoothed running-time of *k-means* is  $\text{poly}(n^k, 1/\sigma)$  if the data points are perturbed by Gaussian random variables with standard deviation  $\sigma$ . We have improved this bound to  $\text{poly}(n^{\sqrt{k}}, 1/\sigma)$  and we have additionally obtained a bound of  $k^{kd} \cdot \text{poly}(n, 1/\sigma)$  [15]. Recently, Arthur et al. [3] showed that the smoothed running-time of *k-means* is polynomial in  $n$  and  $1/\sigma$ .

However, with only a few exceptions [1, 2, 5], the theoretical knowledge about *k-means clustering* is limited to the case of squared Euclidean distances. In this paper, we initiate the theoretical study of the *k-means method* for general Bregman divergences. We show that the lower bound of  $2^{\Omega(n)}$  for the worst-case running-time is valid for almost every Bregman divergence, leading, as for squared Euclidean distances, to a huge discrepancy between the-

ory and practice for many commonly used distance measures like Kullback-Leibler divergence or Itakura-Saito divergence. To obtain more realistic theoretical results, we also analyze the smoothed running-time of  $k$ -means for general Bregman divergences. We show that for almost arbitrary Bregman divergences, the smoothed running-time of  $k$ -means is upper-bounded by  $\text{poly}(n^{\sqrt{k}}, 1/\sigma)$  and  $k^{kd} \cdot \text{poly}(n, 1/\sigma)$ .

In the remainder of this section, we describe the  $k$ -means method (Section 1.1), introduce Bregman divergences (Section 1.2), describe the perturbation model (Section 1.3), and state our results (Section 1.4) and technical contributions (Section 1.5).

## 1.1 $k$ -Means Method

An instance for  $k$ -means clustering is a set  $\mathcal{X} \subseteq \mathbb{R}^d$  consisting of  $n$  points. The aim is to find a clustering  $\mathcal{C}_1, \dots, \mathcal{C}_k$  of  $\mathcal{X}$ , i.e., a partition of  $\mathcal{X}$ , as well as cluster centers  $c_1, \dots, c_k \in \mathbb{R}^d$  such that the potential

$$\sum_{i=1}^k \sum_{x \in \mathcal{C}_i} d_{\Phi}(x, c_i)$$

is minimized, where  $d_{\Phi}$  denotes some distance measure on  $\mathbb{R}^d$ . Given the cluster centers, every data point should be assigned to the cluster whose center is closest to it. The other way round, given the clusters, the centers  $c_1, \dots, c_k$  should be chosen so as to minimize the potential. In the next section, we will see that for Bregman divergences this is the case if the centers are chosen as the centers of mass, i.e.,  $c_i = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} x$ . The  $k$ -means method for Bregman divergences proceeds now as follows:

1. Select cluster centers  $c_1, \dots, c_k$ .
2. Assign every  $x \in \mathcal{X}$  to the cluster  $\mathcal{C}_i$  whose cluster center  $c_i$  is closest to it, i.e.,  $d_{\Phi}(x, c_i) \leq d_{\Phi}(x, c_j)$  for all  $j \neq i$ . (If the closest center is not unique and a point is already assigned to one of the closest clusters, then do not change its assignment.)
3. Set  $c_i = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} x$ .
4. If clusters or centers have changed, goto 2. Otherwise, terminate.

The potential decreases in every step. Thus, no clustering occurs twice, and the algorithm eventually terminates.

## 1.2 Bregman Divergences

One of the most commonly used (and most intuitive) functions  $d_{\Phi}$  is  $d_{\Phi}(x, c) = \|x - c\|^2$ , i.e., squared Euclidean distances. But also other distance measures are common, e.g., Kullback-Leibler divergence [7]. Both are special cases of so-called *Bregman divergences* [5].

**Definition 1.1.** Let  $X \subseteq \mathbb{R}^d$ , and let  $\Phi : X \rightarrow \mathbb{R}$  be a strictly convex function such that  $\Phi$  is differentiable on the relative interior  $\text{ri}(X)$  of  $X$ . The Bregman divergence  $d_{\Phi} : X \times \text{ri}(X) \rightarrow [0, \infty)$  is defined as

$$d_{\Phi}(x, c) = \Phi(x) - \Phi(c) - (x - c)^T \nabla \Phi(c).$$

Here,  $\nabla\Phi(c)$  is the gradient of  $\Phi$  at  $c$ . The basic intuition behind Bregman divergences is the following:  $c$  corresponds to a cluster center and  $x$  to a data point. Let  $\bar{\Phi}(x) = \Phi(c) + (x - c)^T \nabla\Phi(c)$  be the linear interpolation of  $\Phi(x)$  from  $c$ . Then  $d_\Phi(x, c)$  measures how well this interpolation is:  $d_\Phi(x, c) = \Phi(x) - \bar{\Phi}(x)$ . Since  $\Phi$  is strictly convex, we have  $\bar{\Phi}(x) \leq \Phi(x)$  with equality only for  $x = c$ . Thus,  $d_\Phi$  is non-negative and  $d_\Phi(x, c) = 0$  if and only if  $x = c$ .

Some important properties of squared Euclidean distances are also true for general Bregman divergences [16]. For a finite set of points  $C \subseteq X$ , we denote the center of mass of  $C$  by  $\text{cm}(C) = \frac{1}{|C|} \sum_{x \in C} x$ . An important property of Bregman divergences is that the potential can be expressed in terms of the center of mass in the following way [5, Proposition 1]: For every  $c$ ,

$$\sum_{x \in C} d_\Phi(x, c) = \sum_{x \in C} d_\Phi(x, \text{cm}(C)) + |C| \cdot d_\Phi(\text{cm}(C), c).$$

In particular, this means that the center of mass minimizes the potential for a given cluster  $C$ , as it does for squared Euclidean distances.

Another important property of Bregman divergences is that the bisector of two centers  $c$  and  $c'$ , i.e., the set  $\{x \in X \mid d_\Phi(x, c) = d_\Phi(x, c')\}$ , is a hyperplane, which follows immediately from the definition of  $d_\Phi$ . The only known worst-case bound for the running-time of  $k$ -means on squared Euclidean distances comes from the observation that no clustering can repeat during the execution of  $k$ -means. This yields a bound of  $W \leq n^{3kd}$  [3, 12]. The proof of this bound relies only on the fact that the bisectors are hyperplanes. Hence, also for general Bregman divergences, the worst-case number of iterations cannot exceed  $W$ .

In the following, we present some prominent distance measures that are Bregman divergences.

**Mahalanobis Distances.** Let us assume that we want to cluster objects that are each characterized by  $d$  quantities. If these quantities are independent, then clusters should be hyperspherically-shaped and squared Euclidean distances provide a good distance measure. However, if the coordinates are correlated, then clusters are expected to have hyperelliptic shapes and squared Euclidean distances are not the right measure. In that case, let  $B \in \mathbb{R}^{d \times d}$  be the covariance matrix of the components of the data points and assume that it is invertible. This means that the matrix  $B$  is symmetric and positive definite. Let  $A = B^{-1}$ , then the right distance measure taking into account the correlations is the *Mahalanobis distance*  $d_{m_A}$  for  $m_A(x) = x^T A x$ . The gradient of  $m_A$  is  $\nabla m_A(c) = 2Ac$ , which yields  $d_{m_A}(x, c) = (x - c)^T A(x - c)$ . (Letting  $A$  be the identity matrix  $I$  shows that Mahalanobis distances are a generalization of squared Euclidean distances.)

**Kullback-Leibler Divergence and Generalized I-Divergence.** The *Kullback-Leibler divergence* (KLD, relative entropy) is a very popular Bregman divergence. Here,  $X = \{x \in \mathbb{R}^d \mid x \geq 0, \sum_{i=1}^d x_i \leq 1\}$  and an element  $x = (x_1, \dots, x_d) \in X$  represents a probability distribution on a discrete set with  $d + 1$  elements (where  $(x_1, \dots, x_{d+1})$  with  $x_{d+1} = 1 - \sum_{i=1}^d x_i$  is the vector of probabilities). For  $\text{KLD}(x) = \sum_{i=1}^{d+1} x_i \log(x_i)$ , we obtain

$$d_{\text{KLD}}(x, c) = \sum_{i=1}^{d+1} x_i \log\left(\frac{x_i}{c_i}\right),$$

where  $x_{d+1} = 1 - \sum_{i=1}^d x_i$  and  $c_{d+1} = 1 - \sum_{i=1}^d c_i$ . Intuitively, the Kullback-Leibler divergence is a measure for the expected difference in the number of bits that are required to code samples

drawn according to  $x$  when, on the one hand, we use an optimal code based on  $c$  and, on the other hand, we use an optimal code based on  $x$ . KLD plays a crucial role in a variety of applications like clustering text documents and image classification. For instance, for clustering web pages, every data point  $x$  represents a probability distribution on a set of words that appear on the corresponding web page [7].

We will also consider the *generalized I-divergence* (GID), which generalizes KLD to a larger domain: For this, we have  $X = \{x \in \mathbb{R}^d \mid x \geq 0\}$ , the potential function  $\text{GID}(x) = \sum_{i=1}^d x_i \log(x_i)$ , and

$$d_{\text{GID}}(x, c) = \sum_{i=1}^d x_i \log\left(\frac{x_i}{c_i}\right) - \sum_{i=1}^d (x_i - c_i).$$

**Itakura-Saito Divergence.** Another Bregman divergence that is commonly used in signal processing and in particular in speech processing is the *Itakura-Saito divergence* (ISD) [5, 11]. We have again  $X = \{x \in \mathbb{R}^d \mid x \geq 0\}$ , and the potential function is given by the Burg entropy  $\text{ISD}(x) = -\sum_{i=1}^d \log(x_i)$ . From this, we get the Bregman divergence

$$d_{\text{ISD}}(x, c) = \sum_{i=1}^d \frac{x_i}{c_i} - \log\left(\frac{x_i}{c_i}\right) - 1.$$

### 1.3 Perturbation Models for Bregman Divergences

If the Bregman divergence is defined on the whole space  $\mathbb{R}^d$ , i.e., if  $X = \mathbb{R}^d$ , then it is often natural to assume that the points are perturbed by adding Gaussian noise to them. More precisely, we can assume that an adversary is allowed to place initially  $n$  points in  $[0, 1]^d$ , and that each of these points is perturbed by adding a Gaussian random variable with standard deviation  $\sigma$  to each of its coordinates. Equivalently, we can also assume that each point from  $\mathcal{X}$  is a Gaussian random vector with standard deviation  $\sigma$  whose mean can be chosen by the adversary in  $[0, 1]^d$ . This perturbation model has been used for the case of squared Euclidean distances [3, 4, 15].

On the other hand, if  $X$  is a proper subset of  $\mathbb{R}^d$ , as it is the case for KLD or GID, then such a perturbation model cannot be applied as it might yield points outside the feasible region  $X$ . For this reason, we decided to consider very general perturbation models that need to satisfy only a couple of properties, which we will summarize in the following. In Section 3, we present concrete perturbation models with these properties for some special Bregman divergences.

Let us assume that the perturbation model is parameterized by some  $\sigma \in (0, 1]$  that essentially measures the amount of randomness. This means that the smaller the parameter  $\sigma$  is chosen, the weaker is the perturbation and the closer is the analysis to a worst-case analysis. If every point is perturbed by Gaussian noise as described above, then the parameter  $\sigma$  can be chosen as the standard deviation. We assume that the following properties are satisfied for  $\sigma \in (0, 1]$ :

- For any  $\varepsilon \geq 0$ , any hyperplane  $H$ , and any point in  $x \in X \cap [0, 1]^d$ , the probability that  $x$  has a distance of at most  $\varepsilon$  from  $H$  after the perturbation is bounded from above by  $\sqrt{\varepsilon}/\sigma$ .
- For any  $x \in X \cap [0, 1]^d$ , the density of the perturbation of  $x$  is bounded from above by  $(1/\sigma)^d$  on  $\mathbb{R}^d$ .

Let us remark two things about our assumptions on the perturbation model: For Gaussian noise, the probability of being close to a hyperplane is even  $\varepsilon/\sigma$ . However, to gain some flexibility for choosing other perturbation models, we decided to use the weaker bound of  $\sqrt{\varepsilon}/\sigma$ . Second, the bound on the density immediately implies that for any  $\varepsilon \geq 0$ , any  $c \in \mathbb{R}^d$ , and any  $x \in X \cap [0, 1]^d$ , the perturbed version of  $x$  lies in the hyperball with radius  $\varepsilon$  and center  $c$  with probability at most  $(2\varepsilon/\sigma)^d$ .

Additionally, we need the property that perturbed points cannot be too far away from their initial positions in  $X \cap [0, 1]^d$ . For this, let  $D$  be chosen such that with probability at least  $1 - W^{-1}$  every point from the perturbed point set  $\mathcal{X}$  is contained in the hypercube  $\mathcal{D} = [-D, 1+D]^d$ , where  $W \leq n^{3kd}$  denotes the worst number of steps of  $k$ -means. The bounds on the smoothed running-time that we obtain depend polynomially on  $D$ . It is easy to see that for Gaussian random vectors with mean in  $[0, 1]^d$  and standard deviation  $\sigma \leq 1$ ,  $D$  can be chosen polynomially in  $n$ .

## 1.4 Parameterization and Our Results

In this section, we make precise what we mean by “almost arbitrary Bregman divergences.” To do this, we define a couple of parameters of Bregman divergences. For the remainder of the paper we assume that  $X$ , the domain of the distance measure, is a convex set.

For  $\varepsilon \geq 0$ , let  $\mathcal{I}(\varepsilon)$  be the interior of  $X \cap \mathcal{D}$  that has a distance of at least  $\varepsilon$  to the boundary:

$$\mathcal{I}(\varepsilon) = \{x \in X \cap \mathcal{D} \mid \text{dist}(x, \partial(X \cap \mathcal{D})) \geq \varepsilon\}.$$

For a given perturbation model, we choose  $\varepsilon^*$  such that  $\Pr[x \notin \mathcal{I}(\varepsilon^*)] \leq n^{-13}$ , where  $x$  denotes the perturbed version of an arbitrary point in  $X \cap [0, 1]^d$ . In the following, we use the notations  $\mathcal{I} = \mathcal{I}(\varepsilon^*)$  and  $\mathcal{I}' = \mathcal{I}(\varepsilon^*/(2n))$ . An important property of this definition is the following: If  $A \subseteq \mathcal{X}$  is a subset of the data points, and  $A$  contains a point from  $\mathcal{I}$ , then  $\text{cm}(A) \in \mathcal{I}(\varepsilon^*/n) \subseteq \mathcal{I}'$ , i.e., the center of mass of  $A$  is also at a distance of at least  $\varepsilon^*/n$  from the boundary.

To relate the Bregman divergence  $d_\Phi$  to squared Euclidean distances, we introduce two parameters  $\xi$  and  $\xi'$  such that, for all  $x, y \in X \cap \mathcal{D}$ ,

$$d_\Phi(x, y) \geq \xi \cdot \|x - y\|^2 \tag{1}$$

and, for all  $x, y \in \mathcal{I}'$ ,

$$d_\Phi(x, y) \leq \xi' \cdot \|x - y\|^2.$$

Observe that for the definition of  $\xi'$ , only the interior of  $X \cap \mathcal{D}$  is relevant. This is important: If we had let  $x, y \in X \cap \mathcal{D}$  instead of  $x, y \in \mathcal{I}'$  for the definition of  $\xi'$ , then  $\xi'$  is unbounded for many Bregman divergences. The ratio  $\xi'/\xi$  is closely related to the  $\mu$  in the notion of  $\mu$ -similarity introduced by Ackermann et al. [2]. However, Bregman divergences like KLD, GID, or ISD are not  $\mu$ -similar for any  $\mu$  on their whole domain. To make them  $\mu$ -similar, their domains have been restricted such that all data points must be sufficiently far away from the singularities. We emphasize that no such restrictions are necessary for our smoothed analysis. There may be points close to the boundary of the domain, but we can take special care of those points. This technical challenge is the reason for the definition of  $\mathcal{I}$  and  $\mathcal{I}'$  above.

We also need a lower bound on the “second derivative” of  $\Phi$ : We have

$$2\xi \leq \frac{\|\nabla\Phi(x) - \nabla\Phi(y)\|}{\|x - y\|} \tag{2}$$

for all  $x, y \in X \cap \mathcal{D}$  with  $x \neq y$ . This can be seen by the following calculation:

$$\begin{aligned} 2\xi\|x - y\|^2 &\leq d_\Phi(x, y) + d_\Phi(y, x) && \text{by (1)} \\ &= (x - y)^T (\nabla\Phi(x) - \nabla\Phi(y)) && \text{by definition of } d_\Phi \\ &\leq \|x - y\| \cdot \|\nabla\Phi(x) - \nabla\Phi(y)\|. \end{aligned}$$

Dividing both sides by  $\|x - y\|^2$  yields the desired bound. Similarly, we need an upper bound, which unfortunately cannot be derived easily from  $\xi'$ :

$$Q' = \sup_{x, y \in \mathcal{I}', x \neq y} \frac{\|\nabla\Phi(x) - \nabla\Phi(y)\|}{\|x - y\|}.$$

Again, we need the upper bound  $Q'$  only for the interior.

In the following, we assume  $d \leq n$  and  $k \leq n$ , which is satisfied in any reasonable instance of the clustering problem. Additionally, we assume that  $d \geq 4$ , which is also no restriction from a practical point of view, as the dimension is usually significantly larger.

Our first result is to show that the expected running-time of the  $k$ -means method is polynomially bounded in  $n^{\sqrt{k}}$  and  $1/\sigma$  for general Bregman divergences (Theorem 3.1). The polynomial, however, depends on the parameters defined above. To be precise, the bound we obtain is  $1/\xi$  times a polynomial in  $n^{\sqrt{k}}$  and  $1/\sigma$ . The polynomial is independent of the Bregman divergence. Hence, the bound grows only linearly in  $1/\xi$  and it is completely independent of  $Q'$  and  $\xi'$ . Our second bound on the expected running-time is  $k^{kd} \text{poly}(n, 1/\sigma)$  (Theorem 3.2). This yields a polynomial smoothed running-time if  $k, d \in O(\sqrt{\log n / \log \log n})$  (Corollary 3.3). Indeed,  $k$  and  $d$  are usually much smaller than  $n$  in practice. This bound depends polynomially on the parameters  $Q'$ ,  $\xi'$ ,  $1/\xi$ , and  $1/\varepsilon^*$ .

Section 2 contains the smoothed analysis, and in Section 3 we present concrete perturbation models for Mahalanobis distances, Kullback-Leibler divergence, generalized I-divergence, and Itakura-Saito divergence and determine the parameters for these concrete examples.

On the negative side, we transfer the lower bound of  $2^{\Omega(n)}$  for squared Euclidean distances to all good-natured Bregman divergences  $d_\Phi$ , where “good-natured” roughly means that all third order derivatives exist and are bounded within a small region (Section 4). This includes Mahalanobis distances, KLD, GID, and ISD.

## 1.5 Technical Contribution

Our smoothed analysis of  $k$ -means with general Bregman divergences uses a novel lemma about perturbed point sets 2.1: Given any Voronoi partition of the point set, it is unlikely that many points are close to the bisectors. However, to analyze general Bregman divergences, we had to tackle several problems. Let us describe the main problem by way of example: For KLD, the parameters  $\xi'$  and  $Q'$  can become arbitrarily large for points close to the boundary of  $X$ . Even after the perturbation, some of the points might still be too close to the boundary to obtain reasonable upper bounds for  $\xi'$  and  $Q'$ . Essentially, we show that the  $kd$  points that are closest to the boundary can be handled separately and that all other points are sufficiently far away from the boundary (i.e., they lie in  $\mathcal{I}$ ) to bound  $\xi'$  and  $Q'$  in a reasonable way.

An obvious question is whether the smoothed polynomial bound [3] carries over to Bregman divergences. The problem with adapting the proof of this bound is that it exploits specific properties of Gaussian perturbations. It uses, in particular, the property that the projection of

a Gaussian random vector onto a lower-dimensional subspace is still a Gaussian random vector with the same standard deviation. It would be very interesting to see if it is possible to relax some of these requirements or if it is possible to design more general perturbation models that still meet the requirements needed for the smoothed polynomial bound. We emphasize that for Bregman divergences whose domain is not the whole  $\mathbb{R}^d$ , Gaussian perturbations are not a suitable perturbation model. This includes, among others, Kullback-Leibler, generalized I- and Itakura-Saito divergence considered in this paper.

In order to prove the lower bound, we first observe that all Mahalanobis distances (in particular squared Euclidean distances) exhibit the same worst-case behavior. Then we show that all “good-natured” Bregman divergences (including all commonly considered examples like KLD, GID, or ISD) behave locally like some Mahalanobis distance, which makes a transfer of the known lower bound for the Euclidean case possible.

## 2 Smoothed Analysis of $k$ -Means with Bregman Divergences

### 2.1 A Property of Perturbed Point Sets

A crucial argument in our smoothed analysis is that, with high probability, there are not too many points close to the hyperplanes dividing the clusters. This means that eventually one point with a relatively large distance from the bisecting hyperplanes must go from one cluster to another, which causes a significant decrease of the potential. In this section, we generalize this lemma to general Bregman divergences. The proof is closely based on the one for squared Euclidean distances, but we introduce here a new idea that shortens the proof significantly and makes the generalization possible.

**Lemma 2.1.** *Let  $a \in [k]$  be arbitrary. With a probability of at least  $1 - 2W^{-1}$ , the following holds: In every iteration of the  $k$ -means algorithm (except for the first one) in which at least  $kd/a$  points change their assignment, at least one of these points has a Euclidean distance larger than*

$$\varepsilon = \left( \frac{\sigma^2}{3Dn^{10}} \right)^{4a}$$

*from the hyperplane that bisects its new and its old cluster center.*

*Proof.* We consider an iteration of the  $k$ -means algorithm, and we refer to the configuration before this iteration as the *first configuration* and to the configuration after this iteration as the *second configuration*. To be precise, we assume that in the first configuration the positions of the centers are the centers of mass of the points assigned to them in this configuration. The step that we consider is the reassignment of the points according to the Voronoi diagram corresponding to the first configuration.

Let  $B \subseteq \mathcal{X}$  with  $|B| = \ell := kd/a$  be a set of points that change their assignment during the step. There are at most  $n^\ell$  choices for the points in  $B$  and at most  $k^{2\ell} \leq n^{2\ell}$  choices for the clusters they are assigned to in the first and the second configuration. We apply a union bound over all these at most  $n^{3\ell}$  choices.

The following sets are defined for all  $i, j \in [k]$  and  $j \neq i$ . Let  $B_i \subseteq B$  be the set of points that leave cluster  $\mathcal{C}_i$ . Let  $B_{i,j} \subseteq B_i$  be the set of points assigned to cluster  $\mathcal{C}_i$  in the first and to cluster  $\mathcal{C}_j$  in the second configuration, i.e., the points in  $B_{i,j}$  leave  $\mathcal{C}_i$  and enter  $\mathcal{C}_j$ . We have  $B = \bigcup_i B_i$  and  $B_i = \bigcup_{j \neq i} B_{i,j}$ .

Let  $A_i$  be the set of points that are in  $\mathcal{C}_i$  in the first configuration except for those in  $B_i$ . We assume that the positions of the points in  $A_i$  are determined by an adversary. Since the sets  $A_1, \dots, A_k$  form a partition of the points in  $\mathcal{X} \setminus B$  that has been obtained in the previous step on the basis of a Voronoi diagram, there are at most  $W \leq n^{3kd}$  choices for this partition [12]. We also apply a union bound over the choices for this partition.

In the first configuration, exactly the points in  $A_i \cup B_i$  are assigned to cluster  $\mathcal{C}_i$ . Let  $c_1, \dots, c_k$  denote the positions of the cluster centers in the first configuration, i.e.,  $c_i$  is the center of mass of  $A_i \cup B_i$ . The positions of the points in  $\mathcal{X} \setminus B$  are assumed to be fixed by an adversary, and we apply a union bound over the partition  $A_1, \dots, A_k$ . Thus, the impact of the set  $A_i$  on the position of  $c_i$  is fixed. However, we want to exploit the randomness of the points in  $B_i$  in the following. Thus, the positions of the centers are not fixed yet but they depend on the random positions of the points in  $B$ . In particular, the separating hyperplane  $H_{i,j}$  of the clusters  $\mathcal{C}_i$  and  $\mathcal{C}_j$  is not fixed but depends on  $B_i$  and  $B_j$ .

In order to complete the proof, we have to estimate the probability of the event

$$\forall i, j \in [k] : \forall b \in B_{i,j} : \text{dist}(b, H_{i,j}) \leq \varepsilon, \quad (\mathcal{E})$$

where  $\text{dist}(x, H) = \min_{y \in H} \|x - y\|$  denotes the shortest Euclidean distance of a point  $x$  to a hyperplane  $H$ . We denote this event by  $\mathcal{E}$ . If the hyperplanes  $H_{i,j}$  were fixed, then, by our assumption on the perturbation model, the probability of  $\mathcal{E}$  could readily be seen to be at most  $(\frac{\sqrt{\varepsilon}}{\sigma})^\ell$ . However, the hyperplanes are not fixed as their positions and orientations depend on the points in the sets  $B_{i,j}$ . Since the union bound also fixes the number of points in  $B_i$  and  $B_j$ , it suffices to know the sums  $\sum_{b \in B_i} b$  and  $\sum_{b \in B_j} b$  to deduce the exact position of the hyperplane  $H_{i,j}$ . Hence, once all sums  $\sum_{b \in B_i} b$  are fixed, all hyperplanes are fixed as well. The drawback is, of course, that fixing the sum  $\sum_{b \in B_i} b$  has an impact on the distribution of the random positions of the points in  $B_i$ . Basically, we show that after fixing the sum  $\sum_{b \in B_i} b$ , we can still exploit the randomness of  $|B_i| - 1$  points. For a set  $B_i$  with at least two points this means that we can exploit the randomness of at least half of its points. Sets  $B_i$  with only one point need a special treatment. In the following, we define for each  $i$  a set  $B'_i \subseteq B_i$  and a point  $b_i \in A_i \cup B_i$  with the intuition that for fixing the sum  $\sum_{b \in B'_i \cup \{b_i\}} b$ , we sacrifice the randomness of  $b_i$ , while we can still exploit the randomness of all the points in  $B'_i$ . For sets  $B_i$  with at least two points, we can choose  $b_i \in B_i$  arbitrarily and  $B'_i = B_i \setminus \{b_i\}$ . If  $|B_i| = 1$ , then only one point leaves  $\mathcal{C}_i$ , and  $B'_i$  would be empty. In this case, however,  $A_i \neq \emptyset$  because otherwise only a single point would belong to  $\mathcal{C}_i$ , whose position would be identical to the cluster center. Thus, this point would not leave cluster  $\mathcal{C}_i$ . This allows us to choose a point  $b_i \in A_i$  and to set  $B'_i = B_i$ . We remove the point  $b_i$  from  $A_i$  and assume that its position is not fixed yet. For this, we have to include the choices for the points  $b_i$  for those sets  $B_i$  with  $|B_i| > 0$  into the union bound, leaving us with an additional factor of  $n^\ell$ .

Let  $Z$  denote a particular choice in the union bound, and let  $\mathcal{E}_Z$  be the respective event. In Lemma 2.2 we prove that, for any choice  $Z$ , we can exploit the randomness of all the points in the sets  $B'_i$  and we obtain the following bound:

$$\Pr[\mathcal{E}_Z \wedge \neg \mathcal{F}] \leq \left(\frac{3D}{\sigma^2}\right)^{dk} \cdot \varepsilon^{\ell/4},$$

where  $\neg \mathcal{F}$  denotes the event that, after the perturbation, all points of  $\mathcal{X}$  lie in  $\mathcal{D}$ . Now the union bound yields the following upper bound on the probability that a set  $B$  with the stated

properties exists:

$$\begin{aligned}
\Pr[\mathcal{E}] &\leq \Pr[\mathcal{F}] + \Pr[\mathcal{E} \wedge \neg\mathcal{F}] \\
&\leq \Pr[\mathcal{F}] + \sum_Z \Pr[\mathcal{E}_Z \wedge \neg\mathcal{F}] \\
&\leq W^{-1} + n^{4\ell} W \cdot \left(\frac{3D}{\sigma^2}\right)^{dk} \cdot \varepsilon^{\ell/4} \\
&\leq W^{-1} + n^{-3kd} \leq 2W^{-1}.
\end{aligned}$$

The inequalities are due to some simplifications,  $W \leq n^{3kd}$ , and our choice of  $\varepsilon$ .  $\square$

**Lemma 2.2.** *For every choice  $Z$  in the union bound in Lemma 2.1, the probability of the event  $\mathcal{E}_Z \wedge \neg\mathcal{F}$  is bounded from above by*

$$\left(\frac{3D}{\sigma^2}\right)^{dk} \cdot \varepsilon^{\ell/4}.$$

*Proof.* For  $y_i, y_j \in \mathbb{R}^d$ , we denote by  $H_{i,j}(y_i, y_j)$  the bisector of the clusters  $\mathcal{C}_i$  and  $\mathcal{C}_j$  that is obtained for  $g_i := \sum_{b \in B'_i \cup \{b_i\}} b = y_i$  and  $g_j = y_j$ . Let  $k^*$  be the number of clusters  $\mathcal{C}_i$  with  $|B_i| > 0$ . Without loss of generality, these are the clusters  $\mathcal{C}_1, \dots, \mathcal{C}_{k^*}$ . This convention allows us to rewrite the probability of  $\mathcal{E}_Z \wedge \neg\mathcal{F}$  as

$$\begin{aligned}
\Pr[\forall i, j: \forall b \in B_{i,j} \setminus \{b_i\}: \text{dist}(b, H_{i,j}) \leq \varepsilon] &\leq \int_{y_1 \in \mathcal{D}} \cdots \int_{y_{k^*} \in \mathcal{D}} \left( \prod_{i=1}^{k^*} f_{g_i}(y_i) \right) \\
&\cdot \Pr[\forall i, j: \forall b \in B_{i,j} \setminus \{b_i\}: \text{dist}(b, H_{i,j}(y_i, y_j)) \leq \varepsilon \mid \forall i: g_i = y_i] dy_{k^*} \dots dy_1, \quad (3)
\end{aligned}$$

where  $f_{g_i}$  is the density of the random vector  $g_i$ . Our notation is a bit sloppy: If  $|B_{i,j}| > 0$  and  $j \notin \{1, \dots, k^*\}$ , then  $H_{i,j}$  depends only on  $y_i$ . In this case, we should actually write  $H_{i,j}(y_i)$  instead of  $H_{i,j}(y_i, y_j)$  in the formula above. In order to keep the notation less cumbersome, we ignore this subtlety and assume that  $H_{i,j}(y_i, y_j)$  is implicitly replaced by  $H_{i,j}(y_i)$  whenever necessary. Points from different sets  $B_i$  and  $B_j$  are independent even under the assumption that the sums  $g_i$  and  $g_j$  are fixed. Hence, we can further rewrite the probability as

$$\begin{aligned}
&\int \cdots \int \left( \prod_{i=1}^{k^*} f_{g_i}(y_i) \right) \\
&\cdot \left( \prod_{i=1}^{k^*} \Pr[\forall j: \forall b \in B_{i,j} \setminus \{b_i\}: \text{dist}(b, H_{i,j}(y_i, y_j)) \leq \varepsilon \mid g_i = y_i] \right) dy_{k^*} \dots dy_1. \quad (4)
\end{aligned}$$

Now let us consider the probability

$$\Pr[\forall j: \forall b \in B_{i,j} \setminus \{b_i\}: \text{dist}(b, H_{i,j}(y_i, y_j)) \leq \varepsilon \mid g_i = y_i]$$

for a fixed  $i$  and for fixed values  $y_i$  and  $y_j$ . To simplify the notation, let  $B'_i = \{q_1, \dots, q_m\}$ , and let the corresponding hyperplanes (which are fixed because  $y_i$  and the  $y_j$ 's are given) be  $H_1, \dots, H_m$ . (A hyperplane may occur several times in this list if more than one point goes from  $\mathcal{C}_i$  to some cluster  $\mathcal{C}_j$ .) Then the probability simplifies to

$$\Pr[\forall j: \text{dist}(q_j, H_j) \leq \varepsilon \mid g_i = y_i]. \quad (5)$$

Let  $H_j(\varepsilon)$  be the slab of width  $2\varepsilon$  around  $H_j$ , i.e.,  $H_j(\varepsilon) = \{x \in \mathbb{R}^d \mid \text{dist}(x, H_j) \leq \varepsilon\}$ . Let  $f$  be the joint density of the random vectors  $q_1, \dots, q_m, g_i$ . Then the probability in (5) can be bounded from above by

$$\int_{z_1 \in H_1(\varepsilon)} \cdots \int_{z_m \in H_m(\varepsilon)} \frac{f(z_1, \dots, z_m, y_i)}{f_{g_i}(y_i)} dz_m \dots dz_1 .$$

Now let  $f_i$  be the density of the random vector  $q_i$ , and let  $f_{m+1}$  be the density of  $b_i$ . This allows us to rewrite the joint density, and we obtain the upper bound

$$\begin{aligned} & \int_{z_1 \in H_1(\varepsilon)} \cdots \int_{z_m \in H_m(\varepsilon)} \frac{f_1(z_1) \cdot \dots \cdot f_m(z_m) \cdot f_{m+1}(y_i - \sum_{j=1}^m z_j)}{f_{g_i}(y_i)} dz_m \dots dz_1 \\ & \leq \frac{1}{\sigma^d f_{g_i}(y_i)} \cdot \int_{z_1 \in H_1(\varepsilon)} \cdots \int_{z_m \in H_m(\varepsilon)} f_1(z_1) \cdot \dots \cdot f_m(z_m) dz_m \dots dz_1 \\ & = \frac{1}{\sigma^d f_{g_i}(y_i)} \left( \prod_{i=1}^m \int_{z_i \in H_i(\varepsilon)} f_i(z_i) dz_i \right) \leq \frac{1}{\sigma^d f_{g_i}(y_i)} \cdot \left( \frac{\sqrt{\varepsilon}}{\sigma} \right)^m . \end{aligned}$$

The first and the last inequality follow from the properties that the perturbation model has to fulfil: The first inequality follows from  $f_m(\cdot) \leq 1/\sigma^d$ , and the last inequality follows because the probability that a random vector assumes a position within distance  $\varepsilon$  of a given hyperplane is at most  $\sqrt{\varepsilon}/\sigma$ .

We plug this bound into (3). The density  $f_{g_i}$  cancels out for every  $i$ , and we obtain

$$\left( \frac{3D}{\sigma} \right)^{dk^*} \cdot \left( \frac{\sqrt{\varepsilon}}{\sigma} \right)^{|B'_1| + \dots + |B'_k|} \leq \left( \frac{3D}{\sigma} \right)^{dk} \cdot \left( \frac{\sqrt{\varepsilon}}{\sigma} \right)^{\ell/2} \leq \left( \frac{3D}{\sigma^2} \right)^{dk} \cdot \varepsilon^{\ell/4}$$

since  $k^* \leq k$ ,  $|B'_1| + \dots + |B'_k| \geq (|B_1| + \dots + |B_k|)/2 = \ell/2$ , and  $\sigma \leq 1$ .  $\square$

## 2.2 Properties of Bregman Divergences and $k$ -Means

In this section, we collect properties of Bregman divergences that we need for the smoothed analysis. In order to relate the movement of a cluster center to the potential drop, we use the following lemma, which immediately follows from Banerjee et al. [5, Proposition 1].

**Lemma 2.3.** *If in an iteration of the  $k$ -means algorithm a cluster center changes its position from  $c$  to  $c'$ , then the potential drops by at least  $d_\Phi(c', c)$ .*

In order to relate a point's change of assignment to the potential drop, we use the following lemma.

**Lemma 2.4.** *Let  $c_1, c_2, x \in \mathbb{R}^d$ , and assume that  $x$  has a Euclidean distance of  $\varepsilon$  from the bisector of  $c_1$  and  $c_2$  and is lying on the same side of the bisector as  $c_1$ . Then*

$$d_\Phi(x, c_2) - d_\Phi(x, c_1) \geq 2\varepsilon \xi \|c_1 - c_2\| .$$

*Proof.* The point  $x$  has a Euclidean distance of at least  $\varepsilon$  from the hyperplane

$$H = \{y \mid d_\Phi(y, c_1) = d_\Phi(y, c_2)\} .$$

Let  $\delta = \|c_2 - c_1\|$ , and let  $x' \in H$  be any point on this hyperplane. Then  $d_\Phi(x', c_1) = d_\Phi(x', c_2)$  and  $\|x - x'\| \geq \varepsilon$ . We obtain

$$\begin{aligned} d_\Phi(x, c_2) - d_\Phi(x, c_1) &= d_\Phi(x, c_2) - d_\Phi(x', c_2) + d_\Phi(x', c_1) - d_\Phi(x, c_1) \\ &= \langle x - x', \nabla\Phi(c_1) - \nabla\Phi(c_2) \rangle \\ &= \|x - x'\| \cdot \|\nabla\Phi(c_1) - \nabla\Phi(c_2)\| \cdot \cos \alpha \\ &\geq 2\varepsilon\xi\|c_1 - c_2\| \cdot \cos \alpha, \end{aligned}$$

where  $\alpha$  is the angle between  $x - x'$  and  $\nabla\Phi(c_1) - \nabla\Phi(c_2)$  and the inequality follows from (2). As  $\nabla\Phi(c_1) - \nabla\Phi(c_2)$  is orthogonal to the hyperplane  $H$ , we can achieve  $\cos(\alpha) = 1$  by choosing  $x'$  to be the orthogonal projection of  $x$  onto  $H$ . This results in  $\cos(\alpha) \in \{-1, 1\}$ . But  $d_\Phi(x, c_2) - d_\Phi(x, c_1) > 0$  rules out  $\cos(\alpha) = -1$ . This concludes the proof.  $\square$

We say that  $\mathcal{X}$  is  $\varepsilon$ -separated if, for every hyperplane  $H$ , there are at most  $2d$  points in  $\mathcal{X}$  that are within a distance of at most  $\varepsilon$  of  $H$ . The following lemma, due to Arthur and Vassilvitskii [4, Proposition 5.6], shows that  $\mathcal{X}$  is likely to be  $\varepsilon$ -separated. As its proof is only based on an upper bound on the probability that a point has a distance of at most  $\varepsilon$  from a fixed hyperplane, a modified version holds also in our more general setting, when taking into account the upper bound of  $\sqrt{\varepsilon}/\sigma$  for the aforementioned probability.

**Lemma 2.5.** *For  $\varepsilon \geq 0$ , the point set  $\mathcal{X}$  is not  $\varepsilon$ -separated with a probability of at most*

$$n^{2d} \cdot \left( \frac{\sqrt{2d\varepsilon}}{\sigma} \right)^d.$$

### 2.3 An Upper Bound

Lemma 2.1 yields an upper bound for the number of iterations that  $k$ -means needs: Since in any configuration there are only few points close to the bisectors, eventually a point switches from one cluster to another that initially was not close to a bisector. The results of this section lead to the proof of Theorem 3.2. First, we bound the number of iterations in terms of the distance  $\Delta$  of the closest cluster centers that occur during the run of  $k$ -means.

**Lemma 2.6.** *With a probability of at least  $1 - 4W^{-1}$ , every sequence of  $k^{kd/2} + 1$  consecutive iterations of the  $k$ -means algorithm (not including the first one) reduces the potential by at least*

$$\frac{1}{k^{kd/2}} \left( \frac{\xi^{5/2} \varepsilon \varepsilon^* \Delta}{6D\sqrt{d}Q'\xi^{3/2}} \right)^2,$$

where  $\Delta$  denotes the smallest distance of two cluster centers that occurs during the sequence and  $\varepsilon$  is defined as in Lemma 2.1 for  $a = 4$ .

*Proof.* Consider the configuration directly before the sequence of steps is performed. Due to Lemma 2.1, the probability that more than  $kd/4$  points are within distance  $\varepsilon$  of one of the bisectors is at most  $2W^{-1}$ . Additionally, only with a probability of at most  $W^{-1}$ , there exists a point from  $\mathcal{X}$  that does not lie in the hypercube  $\mathcal{D}$ , and only with a probability of at most  $W^{-1}$ , there are more than  $kd/4$  points outside of  $\mathcal{I}$ . Let us assume in the following that none of these failure events occurs.

The at most  $kd/2$  points that are either close to a bisector or not contained in  $\mathcal{I}'$  can assume at most  $k^{kd/2}$  different configurations. Thus, during the considered sequence, at least one point in  $\mathcal{I}'$  that is initially not within distance  $\varepsilon$  of one of the bisectors must change its assignment. Let us call this point  $x$ , and let us assume that it changes from cluster  $\mathcal{C}_1$  to cluster  $\mathcal{C}_2$ . Furthermore, let  $c_1$  and  $c_2$  be the positions of the centers before the sequence. We distinguish two cases. First, if  $x$  is closer to  $c_2$  than to  $c_1$  with respect to  $d_\Phi$  already in the beginning of the sequence, then  $x$  will change its assignment in the first step. According to Lemma 2.4, the potential decreases by at least  $2\varepsilon\xi\Delta$ .

The second case is that  $x$  is closer to  $c_1$  than to  $c_2$  with respect to  $d_\Phi$ . Then, according to Lemma 2.4,

$$d_\Phi(x, c_2) - d_\Phi(x, c_1) \geq 2\xi\varepsilon\Delta.$$

In this case,  $x$  can only change to cluster  $\mathcal{C}_2$  after at least one of the centers of  $\mathcal{C}_1$  or  $\mathcal{C}_2$  has moved. Let  $c'_1$  and  $c'_2$  denote the centers of  $\mathcal{C}_1$  and  $\mathcal{C}_2$  immediately before the reassignment of  $x$ . Then

$$d_\Phi(x, c'_1) - d_\Phi(x, c'_2) \geq 0.$$

Together, this implies

$$d_\Phi(x, c_2) - d_\Phi(x, c'_2) + d_\Phi(x, c'_1) - d_\Phi(x, c_1) \geq 2\xi\varepsilon\Delta. \quad (6)$$

Since the point  $x$  lies in  $\mathcal{I}$  and it belongs to cluster  $\mathcal{C}_1$  when  $c_1$  and  $c'_1$  are computed,  $c_1$  and  $c'_1$  must both belong to  $\mathcal{I}'$ . Let us first consider the case that also  $c_2$  and  $c'_2$  both belong to  $\mathcal{I}'$ . In this case, we will derive a lower bound on  $d_\Phi(c'_1, c_1) + d_\Phi(c'_2, c_2)$ . For  $i \in \{1, 2\}$ , we can rewrite  $d_\Phi(c'_i, c_i)$  as follows:

$$d_\Phi(c'_i, c_i) = d_\Phi(x, c_i) - d_\Phi(x, c'_i) - \langle x - c'_i, \nabla\Phi(c'_i) - \nabla\Phi(c_i) \rangle. \quad (7)$$

Together with Equations (6) and (7),  $c_1, c'_1, c_2, c'_2 \in \mathcal{I}'$  implies

$$\begin{aligned} & d_\Phi(c'_1, c_1) + d_\Phi(c'_2, c_2) \\ & \geq \frac{\xi}{\xi'} \cdot (d_\Phi(c_1, c'_1) + d_\Phi(c_2, c'_2)) \\ & \geq \frac{\xi}{\xi'} \cdot (2\xi\varepsilon\Delta - |\langle x - c_1, \nabla\Phi(c_1) - \nabla\Phi(c'_1) \rangle| - |\langle x - c'_2, \nabla\Phi(c'_2) - \nabla\Phi(c_2) \rangle|) \\ & \geq \frac{\xi}{\xi'} \cdot (2\xi\varepsilon\Delta - \|x - c_1\| \cdot \|\nabla\Phi(c_1) - \nabla\Phi(c'_1)\| - \|x - c'_2\| \cdot \|\nabla\Phi(c'_2) - \nabla\Phi(c_2)\|) \\ & \geq \frac{\xi}{\xi'} \cdot (2\xi\varepsilon\Delta - Q' \cdot \|x - c_1\| \cdot \|c_1 - c'_1\| - Q' \cdot \|x - c'_2\| \cdot \|c'_2 - c_2\|) \\ & \geq \frac{\xi}{\xi'} \cdot (2\xi\varepsilon\Delta - 3D\sqrt{d}Q' \cdot \|c'_1 - c_1\| - 3D\sqrt{d}Q' \cdot \|c'_2 - c_2\|) \\ & \geq \frac{\xi 2\xi\varepsilon\Delta}{\xi'} - \frac{3\sqrt{\xi}D\sqrt{d}Q'}{\xi'} \cdot \left( \sqrt{d_\Phi(c'_1, c_1)} + \sqrt{d_\Phi(c'_2, c_2)} \right). \end{aligned}$$

For  $i = \operatorname{argmax}_{j \in \{1, 2\}} d_\Phi(c'_j, c_j)$ , this yields

$$2d_\Phi(c'_i, c_i) + \frac{6\sqrt{\xi}D\sqrt{d}Q'}{\xi'} \cdot \sqrt{d_\Phi(c'_i, c_i)} \geq \frac{\xi 2\xi\varepsilon\Delta}{\xi'}, \quad (8)$$

which in turn implies

$$d_{\Phi}(c'_i, c_i) + \sqrt{d_{\Phi}(c'_i, c_i)} \geq \frac{2\xi^2\varepsilon\Delta}{6D\sqrt{d}Q'\xi'}.$$

As the right side of the inequality is at most 1, this implies

$$\sqrt{d_{\Phi}(c'_i, c_i)} \geq \frac{2\xi^2\varepsilon\Delta}{12D\sqrt{d}Q'\xi'}.$$

Since  $c_1, c'_1 \in \mathcal{I}'$ , and since we consider the case that also  $c_2, c'_2 \in \mathcal{I}'$ , we obtain

$$\|c'_i - c_i\| \geq \sqrt{d_{\Phi}(c'_1, c_1)}/\xi' \geq \frac{2\xi^2\varepsilon\Delta}{12D\sqrt{d}Q'\xi'^{3/2}} =: Z.$$

Each time the center of  $\mathcal{C}_i$  moves by some amount  $\delta$  with respect to the Euclidean distance, the potential drops by at least  $\xi\delta^2$  (see Lemma 2.3). Since this function is convex, the smallest potential drop is obtained if the center moves by  $Z/k^{kd/2}$  in each iteration. Thus, the decrease of the potential due to the movement of the center is at least

$$k^{kd/2} \cdot \frac{\xi Z^2}{k^{kd}} = \frac{1}{k^{kd/2}} \left( \frac{2\xi^{5/2}\varepsilon\Delta}{12D\sqrt{d}Q'\xi'^{3/2}} \right)^2,$$

which concludes this case.

To finish the proof, we have to consider the case that  $c_2 \notin \mathcal{I}'$  or  $c'_2 \notin \mathcal{I}'$ . In this case, we also consider the position  $c''_2$  of the center of cluster  $\mathcal{C}_2$  after this iteration, that is, after  $x$  is reassigned and the center of  $\mathcal{C}_2$  is recomputed. Since  $x \in \mathcal{I}$ , we know that  $c''_2 \in \mathcal{I}(\varepsilon^*/n)$ . As  $c_2$  or  $c'_2$  does not lie in  $\mathcal{I}' = \mathcal{I}(\varepsilon^*/(2n))$ , this implies  $\|c_2 - c''_2\| \geq \varepsilon^*/(2n)$  or  $\|c'_2 - c''_2\| \geq \varepsilon^*/(2n)$ , respectively. Hence, in  $k^{kd/2} + 1$  steps the center of  $\mathcal{C}_2$  must have moved by at least  $\varepsilon^*/(2n)$ . By the same arguments as above, this yields a lower bound for the potential drop of

$$(k^{kd/2} + 1) \cdot \left( \frac{\sqrt{\xi}\varepsilon^*}{2n(k^{kd/2} + 1)} \right)^2 \geq \frac{1}{k^{kd/2}} \left( \frac{\sqrt{\xi}\varepsilon^*}{4n} \right)^2,$$

which concludes the proof as the bound claimed in the lemma is smaller than the bounds obtained in the two cases.  $\square$

Next we need to analyze the random variable  $\Delta$ , the smallest possible distance of two centers that can occur during the execution of  $k$ -means.

**Lemma 2.7.** *For  $\delta \geq 0$ , we have*

$$\Pr[\Delta \leq \delta] \leq \left( \frac{1028n^8\delta}{\sigma^2} \right)^{d/2}.$$

*Proof.* Let us consider a situation reached by  $k$ -means in which there are two clusters  $\mathcal{C}_1$  and  $\mathcal{C}_2$  whose centers are at a distance of  $\delta$  from each other. We denote the positions of these centers by  $c_1$  and  $c_2$ . Let  $H$  be the bisector of  $c_1$  and  $c_2$ . The points  $c_1$  and  $c_2$  are the centers of mass of the points assigned to  $\mathcal{C}_1$  and  $\mathcal{C}_2$ , respectively, and they have a Euclidean distance of at most  $\delta$  from  $H$ .

From this, we can conclude the following: for every point that is assigned to  $\mathcal{C}_1$  or  $\mathcal{C}_2$  and that has a distance of at least  $2\delta$  from the bisector  $H$ , as compensation another point must be

assigned to  $\mathcal{C}_1$  or  $\mathcal{C}_2$  that has a distance of at most  $\delta$  from  $H$ . Hence, the total number of points assigned to  $\mathcal{C}_1$  or  $\mathcal{C}_2$  can be at most twice as large as the total number of points assigned to  $\mathcal{C}_1$  or  $\mathcal{C}_2$  that are at a distance of at most  $2\delta$  from  $H$ . Hence, there can only exist two centers at a distance of at most  $\delta$  if one of the following two properties is met:

1. There exists a hyperplane from which more than  $2d$  points have a distance of at most  $2\delta$ .
2. There exist two subsets of points whose union has a cardinality of at most  $4d$  and whose centers of mass are at a distance of at most  $\delta$ .

The probability that one of these events occurs can be bounded from above using a union bound and Lemma 2.5:

$$\Pr[\Delta \leq \delta] \leq n^{2d} \left( \frac{\sqrt{4d\delta}}{\sigma} \right)^d + (2n)^{4d} \cdot \left( \frac{2\delta}{\sigma} \right)^d \leq \left( \frac{1028n^8\delta}{\sigma^2} \right)^{d/2}. \quad \square$$

The following lemma is the crucial ingredient of the proof of Theorem 3.2.

**Lemma 2.8.** *Let  $a \in [k]$  be arbitrary, and let  $\varepsilon$  (which depends on  $a$ ) be chosen as in Lemma 2.1. Then the expected number of steps until the potential drops by at least 1 is bounded from above by*

$$\gamma k^{kd} \cdot \left( \frac{n^{11} D Q' \xi'^{3/2}}{2\sigma \xi^{5/2} \varepsilon \varepsilon^*} \right)^2$$

for a sufficiently large absolute constant  $\gamma$ .

*Proof.* With a probability of at least  $1 - 4W^{-1}$ , the number of iterations until the potential drops by at least

$$\frac{1}{k^{kd/2}} \left( \frac{2\xi^{5/2} \varepsilon \varepsilon^* \Delta}{12nD\sqrt{d}Q'\xi'^{3/2}} \right)^2$$

is at most  $k^{kd/2} + 1 \leq 2k^{kd/2}$  due to Lemma 2.6. We estimate the contribution of the failure event, which occurs only with probability  $4W^{-1}$ , to the expected running-time by 4 and ignore it in the following. Let  $T$  denote the random variable that equals the maximum number of sequences of length  $2k^{kd/2}$  until the potential has dropped by at least one. The random variable  $T$  can only exceed  $t$  if

$$\Delta^2 \leq \frac{k^{kd/2}}{t} \left( \frac{12nD\sqrt{d}Q'\xi'^{3/2}}{2\xi^{5/2} \varepsilon \varepsilon^*} \right)^2,$$

leading to the following bound on the expected value of  $T$ :

$$\begin{aligned} \mathbb{E}[T] &= \sum_{t=1}^W \Pr[T \geq t] \leq 4 + \int_0^W \Pr \left[ \Delta^2 \leq \frac{k^{kd/2}}{t} \left( \frac{12nD\sqrt{d}Q'\xi'^{3/2}}{2\xi^{5/2} \varepsilon \varepsilon^*} \right)^2 \right] dt \\ &\leq 4 + t' + \int_{t'}^W \Pr \left[ \Delta \leq \frac{12k^{kd/4} n D \sqrt{d} Q' \xi'^{3/2}}{2\sqrt{t} \xi^{5/2} \varepsilon \varepsilon^*} \right] dt \end{aligned}$$

with

$$t' = \left( \frac{12336k^{kd/4} n^9 D \sqrt{d} Q' \xi'^{3/2}}{2\sigma^2 \xi^{5/2} \varepsilon \varepsilon^*} \right)^2.$$

According to Lemma 2.7,

$$\Pr \left[ \Delta \leq \frac{12k^{kd/4}nD\sqrt{d}Q'\xi^{3/2}}{2\sqrt{t}\xi^{5/2}\varepsilon\varepsilon^*} \right] \leq \left( \frac{\sqrt{t'}}{\sqrt{t}} \right)^{d/2}.$$

For  $d \geq 4$ , this yields

$$\begin{aligned} \mathbb{E}[T] &\leq 4 + t' + \int_{t'}^W \left( \frac{\sqrt{t'}}{\sqrt{t}} \right)^{d/2} dt \leq 4 + t' + \int_{t'}^W \frac{t'}{t} dt \\ &\leq 4 + t' + t' \cdot [\ln(t)]_1^W = 4 + t' \cdot (1 + \ln(W)) \leq 12nk d \cdot t'. \end{aligned}$$

Altogether, this shows that the expected number of steps until the potential drops by at least 1 can be bounded from above by

$$2k^{kd/2} \cdot 12nk d \cdot \left( \frac{12336k^{kd/4}n^9D\sqrt{d}Q'\xi^{3/2}}{2\sigma^2\xi^{5/2}\varepsilon\varepsilon^*} \right)^2,$$

which can, for a sufficiently large absolute constant  $\gamma$ , be bounded from above by

$$\gamma k^{kd} \cdot \left( \frac{n^{11}DQ'\xi^{3/2}}{2\sigma\xi^{5/2}\varepsilon\varepsilon^*} \right)^2. \quad \square$$

## 2.4 Iterations with at most $\sqrt{k}$ Active Clusters

In this section, we analyze steps with at most  $\sqrt{k}$  active clusters. In such a step, either every cluster exchanges altogether at most  $2d\sqrt{k}$  points with other clusters or there are two clusters that exchange at least  $2d + 1$  points with each other. In the former case, the potential will drop due to a significant movement of the centers. In the latter case, the potential drops due to the reassignment.

We start by analyzing the former case. As was done for squared Euclidean distances [4], we define an *epoch* to be a sequence of consecutive iterations in which no cluster center assumes more than two different positions. Equivalently, there are at most two different sets  $\mathcal{C}'_i, \mathcal{C}''_i$  that every cluster  $\mathcal{C}_i$  assumes. It has been shown that the length of any epoch is at most three [3], where length refers to the number of iterations of the epoch. The proof of this does not use any specific properties of squared Euclidean distances and holds for general Bregman divergences as well.

We use the notion of  $(\eta, c)$ -coarseness used by Arthur et al. [3]:  $\mathcal{X}$  is  $(\eta, c)$ -coarse if for any pairwise different subsets  $\mathcal{C}_1, \mathcal{C}_2$ , and  $\mathcal{C}_3$  of  $\mathcal{X}$  with  $|\mathcal{C}_1 \triangle \mathcal{C}_2| \leq c$  and  $|\mathcal{C}_2 \triangle \mathcal{C}_3| \leq c$  either  $\|\text{cm}(\mathcal{C}_1) - \text{cm}(\mathcal{C}_2)\| > \eta$  or  $\|\text{cm}(\mathcal{C}_2) - \text{cm}(\mathcal{C}_3)\| > \eta$ . Since the length of any epoch is at most three, after at most four iterations, one cluster assumes a third positions. Assume that  $\mathcal{X}$  is  $(\eta, c)$ -coarse and that in four consecutive iterations, no cluster gains or loses more than  $c$  points. Then one cluster center moves by at least  $\eta$  during one of these iterations. Combining this with Lemma 2.3 and (1), we get a potential drop of at least  $\xi\eta^2$ .

**Lemma 2.9.** *Assume that  $\mathcal{X}$  is  $(\eta, c)$ -coarse and consider a sequence of four consecutive iterations. If in each of these iterations every cluster exchanges at most  $c$  points, then the potential decreases by at least  $\xi\eta^2$ .*

It remains to prove an upper bound for the probability that  $\mathcal{X}$  is not  $(\eta, c)$ -coarse. This bound needs only the probability for the event that a single point falls into a hyperball of a specific radius  $\varepsilon$  [3, Lemma 4.8]. For Gaussian noise, this probability is at most  $(\varepsilon/\sigma)^d$ . Here we only have an upper bound of  $(2\varepsilon/\sigma)^d$ , which yields the following, slightly weaker bound.

**Lemma 2.10.** *For  $\eta \geq 0$ , the probability that  $\mathcal{X}$  is not  $(\eta, c)$ -coarse is at most  $(6n)^{2c} \cdot (4nc\eta/\sigma)^d$ .*

Now we turn to the case that one cluster gains or loses many points. Given that  $\mathcal{X}$  is  $\varepsilon$ -separated, every iteration with at most  $\sqrt{k}$  active clusters in which one cluster gains or loses more than  $2d\sqrt{k}$  points yields a significant decrease of the potential.

**Lemma 2.11.** *Assume that  $\mathcal{X}$  is  $\varepsilon$ -separated. For every iteration with at most  $\sqrt{k}$  active clusters, the following holds: If a cluster gains or loses more than  $2d\sqrt{k}$  points, then the potential drops by at least  $2\xi\varepsilon^2/n$ .*

*Proof.* Assume that a cluster  $\mathcal{C}_i$  gains or loses more than  $2d\sqrt{k}$  points in a single iteration with at most  $\sqrt{k}$  active clusters. Then there exists another cluster  $\mathcal{C}_j$  with which  $\mathcal{C}_i$  exchanges at least  $2d+1$  points. Since  $\mathcal{X}$  is  $\varepsilon$ -separated, one of these points, say,  $x$ , must be at a distance of at least  $\varepsilon$  from the bisector of  $c_i$  and  $c_j$ . According to Lemma 2.4,  $d_{\Phi}(x, c_i) - d_{\Phi}(x, c_j) \geq \varepsilon 2\xi \cdot \|c_j - c_i\|$ .

It remains to be proved that  $\|c_j - c_i\| \geq \frac{\varepsilon}{n}$ . Let  $H'$  be the hyperplane bisecting the centers of  $\mathcal{C}_i$  and  $\mathcal{C}_j$  in the previous iteration. While  $H'$  does not necessarily bisect  $c_i$  and  $c_j$ , it divides the data points belonging to  $\mathcal{C}_i$  and  $\mathcal{C}_j$  correctly. This implies  $\|c_i - c_j\| \geq \text{dist}(c_i, H') + \text{dist}(c_j, H')$ .

Consider the at least  $2d+1$  data points switching between  $\mathcal{C}_i$  and  $\mathcal{C}_j$ . One of them must be at a distance of at least  $\varepsilon$  from  $H'$  because  $\mathcal{X}$  is  $\varepsilon$ -separated. Let us assume that this point belongs to  $\mathcal{C}_i$ . Then  $\text{dist}(c_i, H') \geq \varepsilon/n$  as  $\mathcal{C}_i$  contains at most  $n$  points. Thus,  $\|c_i - c_j\| \geq \varepsilon/n$ .  $\square$

We consider  $(\eta, c)$ -coarseness for  $c = 2d\sqrt{k}$ . For a set of points that is  $(\eta, 2d\sqrt{k})$ -coarse and  $\varepsilon$ -separated, any sequence of four consecutive steps with at most  $\sqrt{k}$  active clusters yields an improvement of at least  $\min\{\xi\eta^2, 2\xi\varepsilon^2/n\}$ : either Lemma 2.9 or Lemma 2.11 applies. This yields the main lemma of this section, and it will lead to Theorem 3.1.

**Lemma 2.12.** *The expected number of sequences of at most four consecutive iterations, each with at most  $\sqrt{k}$  active clusters, until the potential has dropped by at least 1 is bounded from above by*

$$\frac{1}{\xi} \cdot \text{poly}\left(n^{\sqrt{k}}, \frac{1}{\sigma}\right),$$

where the polynomial is independent of the parameters of the Bregman divergence and  $d \geq 4$ .

*Proof.* Let  $\Delta$  be the smallest improvement made by any sequence of four consecutive iterations with at most  $\sqrt{k}$  active clusters. The random variable  $\Delta$  can only be smaller than some value  $x \geq 0$  if either the instance is not  $\varepsilon(x)$ -separated for  $\varepsilon(x) = \sqrt{nx/(2\xi)}$  or not  $\eta(x)$ -coarse for  $\eta(x) = \sqrt{x/\xi}$ . Hence, for  $x \leq 1$ ,

$$\begin{aligned} \Pr[\Delta \leq x] &\leq \Pr[\mathcal{X} \text{ is not } \eta(x)\text{-coarse}] + \Pr[\mathcal{X} \text{ is not } \varepsilon(x)\text{-separated}] \\ &\leq \left(\frac{2n^{12\sqrt{k}+2} \cdot \sqrt{x/\xi}}{\sigma}\right)^d + n^{2d} \cdot \left(\frac{\sqrt{2d \cdot \sqrt{nx/(2\xi)}}}{\sigma}\right)^d \\ &\leq \left(\frac{2n^{14\sqrt{k}}}{\sigma}\right)^d \left(\left(\frac{x}{\xi}\right)^{\frac{d}{2}} + \left(\frac{x}{2\xi}\right)^{\frac{d}{4}}\right). \end{aligned}$$

Let  $T$  be the random variable of the maximal number of sequences of four consecutive iterations with at most  $\sqrt{k}$  active clusters until the potential has dropped by one. We obtain the following estimate for the expected value of  $T$ :

$$\begin{aligned} \mathbb{E}[T] &= \sum_{t=1}^W \Pr[T \geq t] \leq \sum_{t=1}^W \Pr\left[\Delta \leq \frac{1}{t}\right] \leq 1 + \int_{t=1}^W \Pr\left[\Delta \leq \frac{1}{t}\right] dt \\ &\leq 1 + \int_1^W \min\left\{1, \left(\frac{2n^{14\sqrt{k}}}{\sigma}\right)^d \left(\frac{1}{t\xi}\right)^{\frac{d}{2}}\right\} dt + \int_1^W \min\left\{1, \left(\frac{2n^{14\sqrt{k}}}{\sigma}\right)^d \left(\frac{1}{2t\xi}\right)^{\frac{d}{4}}\right\} dt \\ &\leq \frac{1}{\xi} \cdot \text{poly}\left(n^{\sqrt{k}}, \frac{1}{\sigma}\right) \cdot \log(W) \leq \frac{1}{\xi} \cdot \text{poly}\left(n^{\sqrt{k}}, \frac{1}{\sigma}\right), \end{aligned}$$

where the second-to-last inequality uses the assumption  $d \geq 4$ , and the last inequality uses  $\log W \leq 3kd \log n$ .  $\square$

## 2.5 Iterations with at least $\sqrt{k}$ Active Clusters

In this section, we consider steps of the  $k$ -means algorithm in which at least  $\sqrt{k}$  different clusters gain or lose points. The improvement that such an iteration yields can only be small if none of the cluster centers changes its position significantly due to the reassignment of points. Intuitively, this becomes increasingly unlikely as the number of active clusters increases. For squared Euclidean distances, we showed that, indeed, if at least  $\sqrt{k}$  clusters are active, then with high probability one of them changes its position by  $n^{-O(\sqrt{k})}$ . This yields a potential drop in the same order of magnitude.

**Lemma 2.13.** *The expected number of steps with at least  $\sqrt{k}$  active clusters until the potential drops by at least 1 is bounded from above by*

$$\frac{1}{\xi} \cdot \text{poly}\left(n^{\sqrt{k}}, \frac{1}{\sigma}\right),$$

where the polynomial is independent of the parameters of the Bregman divergence.

*Proof.* We consider one step of the  $k$ -means algorithm with at least  $\sqrt{k}$  active clusters. Let  $\varepsilon$  be defined as in Lemma 2.1 for  $a = 1$ . We distinguish two cases: Either one point that is reassigned during the considered iteration has a distance of at least  $\varepsilon$  from the bisector that it crosses, or all points are at a distance of at most  $\varepsilon$  from their respective bisectors. In the former case, we immediately get a potential drop of at least  $2\xi\varepsilon\Delta$ , where  $\Delta$  denotes the minimal distance of two cluster centers. In the latter case, Lemma 2.1 implies that with high probability less than  $kd$  points are reassigned during the considered step. We apply a union bound over the choices for these points. In the union bound, we fix not only these points but also the clusters they are assigned to before and after the step. We denote by  $A_i$  the set of points that are assigned to cluster  $\mathcal{C}_i$  in both configurations and we denote by  $B_i$  and  $B'_i$  the sets of points assigned to cluster  $\mathcal{C}_i$  before and after the step, respectively, except for the points in  $A_i$ . Analogously to Lemma 2.1, we assume that the positions of the points in  $A_1 \cup \dots \cup A_k$  are fixed adversarially, and we apply a union bound on the different partitions  $A_1, \dots, A_k$  that are realizable. Altogether, we have a union bound over less than  $n^{3kd} \cdot n^{3kd} = n^{6kd}$  events. Let

$c_i$  be the position of the cluster center of  $C_i$  before the reassignment, and let  $c'_i$  be the position after the reassignment. Then

$$c_i = \frac{|A_i| \cdot \text{cm}(A_i) + |B_i| \cdot \text{cm}(B_i)}{|A_i| + |B_i|} ,$$

where  $\text{cm}(\cdot)$  denotes the center of mass of a point set. Since  $c'_i$  can be expressed analogously, we can write the change of position of the cluster center of  $C_i$  as

$$c_i - c'_i = |A_i| \cdot \text{cm}(A_i) \left( \frac{1}{|A_i| + |B_i|} - \frac{1}{|A_i| + |B'_i|} \right) + \frac{|B_i| \cdot \text{cm}(B_i)}{|A_i| + |B_i|} - \frac{|B'_i| \cdot \text{cm}(B'_i)}{|A_i| + |B'_i|} .$$

Due to the union bound,  $\text{cm}(A_i)$  and  $|A_i|$  are fixed. Additionally, also the sets  $B_i$  and  $B'_i$  are fixed but not the positions of the points in these two sets. If we considered only a single center, then we could easily estimate the probability that  $\|c_i - c'_i\| \leq \beta$ . For this, we additionally fix all positions of the points in  $B_i \cup B'_i$  except for one of them, say  $b_i$ . Given this, we can express the event  $\|c_i - c'_i\| \leq \beta$  as the event that  $b_i$  assumes a position in a ball whose position depends on the fixed values and whose radius, which depends on the number of points in  $|A_i|$ ,  $|B_i|$ , and  $|B'_i|$ , is not larger than  $n\beta$ . Hence, the probability is bounded from above by

$$\left( \frac{2n\beta}{\sigma} \right)^d .$$

However, we are interested in the probability that this is true for all centers simultaneously. Unfortunately, the events are not independent for different clusters. We estimate this probability by identifying a set of  $\ell/2$  clusters whose randomness is independent enough, where  $\ell \geq \sqrt{k}$  is the number of active clusters. More precisely, we do the following: Consider a graph whose nodes are the active clusters and that contains an edge between two nodes if and only if the corresponding clusters exchange at least one point. We identify a dominating set in this graph, i.e., a subset of nodes that covers the graph in the sense that every node not belonging to this subset has at least one edge into the subset. We can assume that the dominating set, which we identify, contains at most half of the active clusters. (In order to find such a dominating set, start with the graph and throw out edges until the remaining graph is a tree. Then put the nodes on odd layers to the left side and the nodes on even layers to the right side, and take the smaller side as the dominating set.)

For every active center  $C$  that is not in the dominating set, we do the following: We assume that all the positions of the points in  $B_i \cup B'_i$  are already fixed except for one of them. Given this, we can use the aforementioned estimate for the probability of  $\|c_i - c'_i\| \leq \beta$ . If we iterate this over all points not in the dominating set, we can always use the same estimate; the reason is that the choice of the subset guarantees that, for every node not in the subset, we have a point whose position is not fixed yet. This yields an upper bound of

$$\left( \frac{2n\beta}{\sigma} \right)^{d\ell/2} .$$

Combining this probability with the number of choices in the union bound yields a bound of

$$n^{6kd} \cdot \left( \frac{2n\beta}{\sigma} \right)^{d\ell/2} \leq n^{6kd} \cdot \left( \frac{2n\beta}{\sigma} \right)^{d\sqrt{k}/2} .$$

For

$$\beta = \frac{\sigma}{2n^{18\sqrt{k}+1}}$$

the probability can be bounded from above by  $n^{-3kd} \leq W^{-1}$ .

Now we also take into account the failure probability of  $2W^{-1}$  from Lemma 2.1. This yields that, with a probability of at least  $1 - 3W^{-1}$ , the potential drops in every iteration, in which at least  $\sqrt{k}$  clusters are active, by at least

$$\begin{aligned} \Gamma := \min\{2\xi\varepsilon\Delta, \xi\beta^2\} &\geq \xi \cdot \min\left\{\frac{\sigma^8\Delta}{1296n^{38}D^6d}, \frac{\sigma^2}{n^{36\cdot\sqrt{k}+2}}\right\} \\ &\geq \xi \cdot \min\left\{\Delta \cdot \text{poly}(n^{-1}, \sigma), \text{poly}(n^{-\sqrt{k}}, \sigma)\right\} \end{aligned}$$

since  $d \leq n$  and  $D$  is polynomially bounded in  $\sigma$  and  $n$ . The number  $T$  of steps with at least  $\sqrt{k}$  active clusters until the potential has dropped by one can only exceed  $t$  if  $\Gamma \leq 1/t$ . Hence,

$$\begin{aligned} \mathbb{E}[T] &\leq \sum_{t=1}^{\infty} \Pr[T \geq t] + 3W^{-1} \cdot W \leq 3 + \int_{t=0}^{\infty} \Pr[T \geq t] dt \\ &\leq 3 + \beta^{-2} + \int_{t=\beta^{-2}}^{\infty} \Pr\left[\Gamma \leq \frac{1}{t}\right] dt \\ &\leq 3 + \beta^{-2} + \int_{t=\beta^{-2}}^{\infty} \Pr\left[\Delta\xi \cdot \text{poly}\left(\frac{1}{n}, \sigma\right) \leq \frac{1}{t}\right] dt \\ &\leq 3 + \beta^{-2} + \int_{t=\beta^{-2}}^{\infty} \Pr\left[\Delta \leq \frac{1}{t\xi} \cdot \text{poly}\left(n, \frac{1}{\sigma}\right)\right] dt \\ &\leq 3 + \beta^{-2} + \int_{t=\beta^{-2}}^{\infty} \min\left\{1, \left(\frac{(4d+16) \cdot n^4 \cdot \text{poly}(n, \sigma^{-1})}{t\xi\sigma}\right)^d\right\} dt \\ &= \frac{1}{\xi} \cdot \text{poly}\left(n^{\sqrt{k}}, \frac{1}{\sigma}\right), \end{aligned}$$

where the integral is upper bounded as in the proof of Lemma 2.8.  $\square$

### 3 Applying the Smoothed Analysis

Now we apply the analysis of the previous section to four special Bregman divergences: Mahalanobis distances, Kullback-Leibler divergence, generalized I-divergence, and Itakura-Saito divergence.

In order to get smoothed bounds, we need two ingredients: First, the expected number of steps until the potential drops by at least one. Second, an upper bound for the potential after one iteration. The smoothed bound is the product of both.

We will instantiate the following results with specific Bregman divergences in the remainder of this section. In the remainder of this section, let  $P$  be the maximal potential that we have after the first iteration of  $k$ -means, provided that all points of  $\mathcal{X}$  lie in  $\mathcal{D}$ . First, we exploit Lemmas 2.12 and 2.13.

**Theorem 3.1.** *Let  $d_\Phi$  be a Bregman divergence. Then the smoothed running-time of  $k$ -means is bounded from above by*

$$\frac{P}{\xi} \cdot \text{poly} \left( n^{\sqrt{k}}, \frac{1}{\sigma} \right).$$

Second, we apply Lemma 2.8. Note that the degree of the polynomial is not only independent of  $d$ , but also independent of  $k$ .

**Theorem 3.2.** *Let  $d_\Phi$  be a Bregman divergence. Then the smoothed running-time of  $k$ -means is bounded from above by*

$$P \cdot k^{kd} \cdot \frac{Q'^2 \xi'^3}{4\xi^5 \varepsilon^{*2}} \cdot \text{poly} \left( n, \frac{1}{\sigma} \right).$$

If the parameters  $P$ ,  $1/\xi$ ,  $\xi'$ ,  $Q'$ , and  $1/\varepsilon^*$  are bounded by polynomials, then we get a polynomial smoothed running-time if  $k$  and  $d$  are small compared to  $n$ .

**Corollary 3.3.** *In the setting of Theorem 3.2, if  $P$ ,  $Q'$ , and  $\xi'$  as well as  $1/\xi$ , and  $1/\varepsilon^*$  are bounded from above by  $\text{poly}(n, 1/\sigma)$ , then the smoothed running-time of  $k$ -means is bounded from above by*

$$\text{poly} \left( n, \frac{1}{\sigma} \right)$$

for  $k, d \in O(\sqrt{\log n / \log \log n})$ .

### 3.1 Mahalanobis Distances

For Mahalanobis distances, we use the same perturbation model that has been used for squared Euclidean distances [4, 15]: The adversary chooses  $n$  points in  $[0, 1]^d$ . Then the  $d$  coordinates are perturbed by independent Gaussian perturbations of standard deviation  $\sigma$ . We can choose  $D = \text{poly}(n)$ . Then  $\mathcal{X} \subseteq \mathcal{D} = [-D, D + 1]^d$  with a probability of at least  $1 - W^{-1}$  since Gaussians are concentrated around their mean, which is in  $[0, 1]^d$ . After one iteration of  $k$ -means, every point is assigned to a cluster center within a distance of at most  $\text{poly}(n)$ . Using this, we will bound the potential after one iteration in a moment.

Let  $A \in \mathbb{R}^{d \times d}$  be an arbitrary symmetric positive definite matrix, and consider  $k$ -means using  $m_A$ . Scaling the matrix does not change the behavior of  $k$ -means. Thus, we scale  $A$  such that the smallest eigenvalue, which is positive, becomes 1. Let  $\lambda_{\max}$  be the largest eigenvalue of  $A$ . Then  $\xi = 1$  and  $\xi' = \lambda_{\max}$ . Moreover, we have  $Q' = 2\|A\|$ , where  $\|M\| = \max_{\|x\|=1} \|Mx\|$  is the operator norm of a matrix  $M$  [10, Section 2.3]. The 2-norm of a symmetric matrix equals its largest eigenvalue. Thus,  $Q' = 2\lambda_{\max}$ . As  $\xi'$  and  $Q'$  are bounded on the whole space  $\mathbb{R}^d$ , we can define  $\mathcal{I} = \mathcal{I}' = \mathcal{D}$ . Then, the case yielding to  $\varepsilon^*$  in Lemma 2.8 cannot occur and we can simply remove  $\varepsilon^*$  from the bound in Theorem 3.2.

Now we can also bound the potential after one iteration:  $P$  is bounded by  $\lambda_{\max} \cdot \text{poly}(n)$  if all points lie in  $\mathcal{D}$ . (If not all points assume a value in  $\mathcal{D}$ , then we bound the number of iterations by the worst-case bound of  $W$ , which contributes only a constant to the expected running-time.)

**Theorem 3.4.** *The smoothed running-time of  $k$ -means using  $m_A$  is bounded from above by*

$$\lambda_{\max} \cdot \text{poly} \left( n^{\sqrt{k}}, \frac{1}{\sigma} \right)$$

and

$$k^{kd} \cdot \lambda_{\max}^6 \cdot \text{poly} \left( n, \frac{1}{\sigma} \right).$$

If  $k, d \in O(\sqrt{\log n / \log \log n})$  and the largest eigenvalue of  $A$  is bounded by a polynomial, then, as in Corollary 3.3, we obtain smoothed polynomial running-time.

### 3.2 Kullback-Leibler Divergence

We have to be more careful when choosing a perturbation model for Kullback-Leibler divergence. KLD is defined on a simplex. Thus, we cannot use Gaussian perturbations since these might result in points outside of the domain of KLD.

To get a perturbation model, we take into account that a point represents a probability distribution on a finite set  $\{1, 2, \dots, d+1\}$ . For instance, assume that we want to classify web pages based on a list  $w_1, \dots, w_{d+1}$  of words (the so-called *bag-of-words model* [7]). For a specific web page, let  $n_i$  be the number of occurrences of  $w_i$ . Then  $x_i = \frac{n_i}{\sum_{j=1}^{d+1} n_j}$  is the relative frequency of  $w_i$ . Based on the vectors  $x$ , web pages can be clustered according to their topics since pages about similar topics are likely to contain similar words. To perturb instances, the idea is to add a random number of each word to the web page.

Let us make this more precise. For a point  $x \in X$ , we obtain  $x' \in \mathbb{R}^{d+1}$  by adding the component  $x_{d+1} = 1 - \sum_{i=1}^d x_i$ . Then we draw random numbers  $y_1, \dots, y_{d+1}$  independently according to some probability distribution to be specified in a moment. Let  $S = \sum_{i=1}^{d+1} x_i + y_i = 1 + \sum_{i=1}^{d+1} y_i$ . Then we obtain the perturbed point  $z \in \mathbb{R}^d$  by setting  $z_i = \frac{x_i + y_i}{S}$ . By construction,  $z \geq 0$  and  $\sum_{i=1}^d z_i \leq 1$ .

Now we have to choose a probability distribution. We use the exponential distribution [9], whose density is  $\frac{1}{\theta} \cdot \exp(-\frac{x}{\theta})$  for a positive parameter  $\theta$ . It has mean  $\theta$ , variance  $\theta^2$ , and maximum density  $1/\theta$ .

We choose  $\theta = 8d\sigma^{d/(d+1)}$ . Furthermore, we restrict ourselves to  $\sigma \leq \frac{1}{14d^{4/3}}$ . These choices requires explanation. First,  $\theta = \sigma$  would be the natural choice. To meet the requirements for perturbation models, and to use our framework introduced in Section 2, however, we need this choice of  $\theta$ . We emphasize that choosing  $\theta = \sigma$  would yield exactly the same results since  $\sigma$  and  $\theta = 8d\sigma^{d/(d+1)}$  differ only by a polynomial factor. Second, we need  $\sigma \leq \frac{1}{14d^{4/3}}$  also that our perturbation model meets the requirements. But this does not harm to the result either: On the one hand, it includes the particularly interesting small values of  $\sigma$ . On the other hand, stronger perturbations only decrease the expected running-time, and  $\sigma = 1$  is only polynomially larger than  $\sigma = \frac{1}{14d^{4/3}}$ .

One might argue that Poisson distributions are a more natural model for choosing a random number of words. Poisson distributions are, however, discrete distributions on  $\mathbb{N}$ . A natural way to get a continuous probability distribution would be to add a random number from  $[0, 1)$  to the randomly drawn integer. In this way, the density function becomes a step function that tends exponentially to 0, and the distribution function becomes continuous.

For simplicity, we restrict ourselves to exponential distributions in the following, and we note that the same holds for any distribution with exponentially small tail bounds, like, e.g., the above described variant of a Poisson distribution or Gaussian random variables conditioned on the outcome being non-negative.

Let us now prove that our perturbation model satisfies the requirements of Section 1.3.

**Lemma 3.5.** *Let  $x \in X$  and let  $z \in X$  be the point obtained from  $x$  by perturbation. Let  $H \subseteq \mathbb{R}^d$  be any hyperplane. Then*

$$\Pr[\text{dist}(z, H) \leq \varepsilon] \leq \frac{\sqrt{\varepsilon}}{\sigma},$$

and the density of the random variable  $z$  is bounded from above by  $\sigma^{-d}$ .

*Proof.* Let  $S = \sum_{i=1}^{d+1} x_i + y_i = 1 + \sum_{i=1}^{d+1} y_i$ . Let  $v$  be the normal vector of the hyperplane  $H$ . Without loss of generality, we assume that  $v_d \geq 1/\sqrt{d}$ .

Let  $F$  denote the failure event that  $S \geq 1 + Z$  for some  $Z$  yet to be specified. The event  $F$  occurs only if there is an  $i$  with  $y_i \geq Z/d$ . This happens only with a probability of at most  $(d+1) \exp(-\frac{Z}{d\theta})$ . We choose  $Z$  such that this probability is at most  $\frac{\varepsilon}{2\sigma}$ , which yields  $Z = d\theta \log(\frac{2(d+1)\sigma}{\varepsilon}) \leq d\theta \sqrt{\frac{3d\sigma}{\varepsilon}}$ .

Given  $S$  and  $y_1, \dots, y_{d-1}, y_{d+1}$ , we have  $\text{dist}(z, H) \leq \varepsilon$  only if  $y_d$  assumes a value in an interval of length  $2\varepsilon S/v_d \geq 2\varepsilon S\sqrt{d}$ . This implies that  $\text{dist}(z, H) \leq \varepsilon$  happens only if either  $S \geq 1 + Z$  or if  $y_d$  falls into an interval of length  $2\varepsilon\sqrt{d}(1 + Z)$ . Since the density of  $y_d$  is bounded from above  $1/\theta$ , we obtain

$$\begin{aligned} \Pr[\text{dist}(z, H) \leq \varepsilon] &\leq \frac{\varepsilon}{2\sigma} + \frac{2\varepsilon\sqrt{d}(Z+1)}{\theta} \leq \frac{\varepsilon}{2\sigma} + \frac{2\varepsilon\sqrt{d} \cdot (d\theta\sqrt{\frac{3d\sigma}{\varepsilon}} + 1)}{\theta} \\ &= \frac{\varepsilon}{2\sigma} + \frac{2\varepsilon\sqrt{d}}{\theta} + 2d^2\sqrt{\varepsilon 3\sigma} \leq \frac{3\sqrt{\varepsilon}}{4\sigma} + 2d^2\sqrt{3\sigma\varepsilon} \leq \frac{\sqrt{\varepsilon}}{\sigma}. \end{aligned}$$

The last inequality follows from  $\sigma \leq \frac{1}{14d^{4/3}}$ .

Next, we analyze the maximum density of the random vector  $z$ . For this, we perform a change of variables: Instead of considering the vector  $y = (y_1, \dots, y_{d+1})$ , we consider the vector  $z' = (z_1, \dots, z_d, S)$  and denote by  $h$  its density. The transformation  $\Phi$  with

$$\Phi: (y_1, \dots, y_{d+1}) \mapsto \left( \frac{x_1 + y_1}{1 + \sum_{i=1}^{d+1} y_i}, \dots, \frac{x_d + y_d}{1 + \sum_{i=1}^{d+1} y_i}, 1 + \sum_{i=1}^{d+1} y_i \right)$$

maps  $y$  to  $z'$  and its inverse is

$$\Phi^{-1}: (z'_1, \dots, z'_{d+1}) \mapsto \left( z'_{d+1}z'_1 - x_1, \dots, z'_{d+1}z'_d - x_d, z'_{d+1} - z'_{d+1} \sum_{i=1}^d z'_i \right).$$

A simple calculation shows that the determinant of the Jacobian of  $\Phi^{-1}$  at  $(z'_1, \dots, z'_d, T)$  is  $T^d$ . Let  $f$  denote the density of the exponentially distributed random variables  $y_i$ . Then, the density of  $z$  at  $(z_1, \dots, z_d)$  can be written as

$$\begin{aligned} &\int_0^\infty T^d \cdot \prod_{i=1}^d f(Tz'_i - x_i) \cdot f\left(T - T \sum_{i=1}^d z'_i\right) dT \\ &\leq \int_0^\infty T^d \cdot \frac{\prod_{i=1}^d \exp(-\frac{Tz'_i + x_i}{\theta}) \cdot \exp(-\frac{T - T \sum_{i=1}^d z'_i}{\theta})}{\theta^{d+1}} dT \\ &= \int_0^\infty T^d \cdot \frac{\exp(-\frac{T}{\theta})}{\theta^{d+1}} dT \leq \frac{1}{\theta^{d+1}} \int_0^\infty T^d \exp(-T) dT \\ &= \frac{d!}{\theta^{d+1}} \leq \left(\frac{d}{\theta}\right)^{d+1} \leq \sigma^{-d}. \quad \square \end{aligned}$$

What remains to be done is to analyze the parameters  $\xi$ ,  $\xi'$ ,  $Q'$ , and  $\varepsilon^*$  as well as the potential  $P$  after the first iteration.

All points are contained in  $\mathcal{D}$  because the domain of KLD is a subset of  $[0, 1]^d \subseteq \mathcal{D}$ .

Consider a point  $z$  obtained by perturbing any point  $x$ . The probability of a perturbed point to be  $\varepsilon$ -close to a hyperplane is  $\sqrt{\varepsilon}/\sigma$ . We consider the  $d+1$  hyperplanes  $x_i = 0$  for  $1 \leq i \leq d$  and  $\sum_{i=1}^d x_i = 1$ . The probability that at least one of the  $n$  points comes  $\varepsilon$ -close to one of them is at most  $\frac{n(d+1)\sqrt{\varepsilon}}{\sigma} \leq \frac{2n^2}{\sigma}\sqrt{\varepsilon}$ . We choose  $\varepsilon^* = \frac{1}{4}n^{-29}\sigma^2$ , then the probability that a point comes  $\varepsilon^*$ -close to the boundary of the domain is at most  $n^{-13}$ .

Let us now analyze  $\xi$ . Let  $x, y \in \mathcal{D}$  be arbitrary, and let  $x_{d+1} = 1 - \sum_{i=1}^d x_i$  and  $y_{d+1} = 1 - \sum_{i=1}^d y_i$ , and let  $x', y'$  be the vectors with this additional component. Then

$$d_{\text{KLD}}(x, y) = \sum_{i=1}^{d+1} x_i \log(x_i/y_i) \geq \frac{1}{2} \|x' - y'\| \geq \frac{1}{2} \|x - y\|,$$

where the first inequality follows from Ackermann et al. [2]. This shows  $\xi = \frac{1}{2}$ .

Now we turn to  $\xi'$ . Let  $x, y \in \mathcal{I}'$  be arbitrary. Let  $x'$  and  $y'$  be defined as above. First, we relate  $\|x - y\|$  and  $\|x' - y'\|$ :

$$\begin{aligned} \|x' - y'\|^2 &= \sum_{i=1}^d (x_i - y_i)^2 + \left( \sum_{i=1}^d x_i - y_i \right)^2 \\ &= \|x - y\|^2 + \sum_{1 \leq i, j \leq d} \underbrace{(x_i - y_i)(x_j - y_j)}_{\leq (x_i - y_i)^2 + (x_j - y_j)^2} \\ &\leq (2d + 1) \cdot \|x - y\|^2. \end{aligned}$$

Since  $x, y \in \mathcal{I}'$ , we have  $x_i, y_i \geq \varepsilon^*/2n$  for  $1 \leq i \leq d+1$ . Hence,

$$d_{\text{KLD}}(x, y) \leq \frac{n}{\varepsilon^*} \cdot \|x' - y'\|^2 \leq \frac{n(2d+1)}{\varepsilon^*} \cdot \|x - y\|^2,$$

where the first inequality follows from Ackermann et al. [2]. This shows that  $\xi' \leq \text{poly}(n, \sigma^{-1})$ .

Next comes  $Q'$ . We have  $x, y \in \mathcal{I}'$  and

$$\frac{\|\nabla \text{KLD}(x) - \nabla \text{KLD}(y)\|}{\|x - y\|} \leq d \cdot \max_i \frac{|\log x_i - \log(y_i)|}{|x_i - y_i|}.$$

By the mean value theorem, the latter is  $d$  times the derivative of  $\log$  at some point between  $x_i$  and  $y_i$ . Since  $x, y \in \mathcal{I}'$ , we get  $Q' \leq \frac{2nd}{\varepsilon^*}$ .

Now we bound  $P$ . We note that  $d_{\text{KLD}}(x, c)$  is monotonically increasing in each  $x_i$  and monotonically decreasing in each  $c_i$ . Furthermore, after reassigning the clusters, we have  $c_i \geq x_i/n$ . This yields  $d_{\text{KLD}}(x, c) \leq d \log n$ . Thus, after the first iteration, the potential is bounded from above by  $dn \log n$ .

Putting everything together yields the following theorem.

**Theorem 3.6.** *The smoothed running-time of  $k$ -means using KLD is bounded from above by*

$$\text{poly} \left( n^{\sqrt{k}}, \frac{1}{\sigma} \right).$$

and

$$k^{kd} \cdot \text{poly} \left( n, \frac{1}{\sigma} \right).$$

### 3.3 Generalized I-Divergence

For generalized I-divergence, we use the same perturbation model, except for rescaling. Since we do not have to rescale, this allows us to let the adversary choose any density function  $f$  bounded by  $\frac{1}{2\sqrt{d}\sigma}$  whose tail bounds are sufficiently small: The probability of a number greater than  $\text{poly}(n)$  must be bounded by  $\frac{1}{ndW}$ . Then we perturb a point by adding independent random numbers drawn according to  $f$ . The maximum density is then  $(2\sqrt{d}\sigma)^d$ , which is fine. The probability of coming  $\varepsilon$ -close to a hyperplane  $H$  is also easily analyzed: Let  $v$  be the normal vector with  $v_1 \geq 1/\sqrt{d}$ . We allow the adversary to fix  $z_2, \dots, z_d$ . Then for  $\text{dist}(z, H) \leq \varepsilon$ , the component  $z_1$  must fall into an interval of length at most  $2\varepsilon\sqrt{d}$ , which happens with a probability of at most  $\varepsilon/\sigma \leq \sqrt{\varepsilon}/\sigma$ .

The values for  $\varepsilon^*$ ,  $\xi'$ , and  $Q'$  can be analyzed similarly as for KLD in the previous section. Also  $\xi$  can be analyzed similarly, we only have to use the upper bound of  $\text{poly}(n)$  rather than the upper bound of 1. In the same way, the potential  $P$  after the first iteration can be analyzed.

Overall, we obtain the same results as for KLD.

**Theorem 3.7.** *The smoothed running-time of  $k$ -means using GID is bounded from above by*

$$\text{poly}\left(n^{\sqrt{k}}, \frac{1}{\sigma}\right).$$

and

$$k^{kd} \cdot \text{poly}\left(n, \frac{1}{\sigma}\right).$$

### 3.4 Itakura-Saito Divergence

For the Itakura-Saito divergence, we use the same perturbation model as for generalized I-divergence since their domains are equal and rescaling, as for KLD, is not necessary. We can choose  $D = \text{poly}(n)$  to make sure that  $\mathcal{X} \subseteq [0, D]^d = \mathcal{D}$  with a probability of at least  $1 - W^{-1}$ .

The analysis of  $\varepsilon^* \geq 1/\text{poly}(n)$  is similar to its counterpart for KLD. Let us first analyze  $\xi$  and  $\xi'$ . By definition of a Bregman divergence,  $d_{\text{ISD}}$  is the tail of the first-order Taylor expansion of  $\text{ISD}(x)$  at  $y$ . Thus, there exists a  $\xi \in \mathbb{R}^d$  with  $\xi_i \in [x_i, y_i]$  or  $\xi_i \in [y_i, x_i]$  such that

$$d_{\text{ISD}}(x, y) = \frac{1}{2}(x - y)^T \nabla^2 \text{ISD}(\xi)(x - y),$$

where  $\nabla^2 \text{ISD}(\xi)$  is the Hesse matrix of  $\text{ISD}$  at  $\xi$ . (This exists for all possible  $\xi$ .) The Hesse matrix  $\nabla^2 \text{ISD}(\xi)$  is a diagonal matrix with diagonal entries  $1/\xi_1^2, \dots, 1/\xi_d^2$ . For each such entry, we have  $1/\xi_i^2 \geq \frac{1}{\max(y_i^2, x_i^2)} \geq 1/D^2$ . Thus,  $1/\xi \leq \text{poly}(n)$ , which is fine. On the other hand,  $\frac{1}{\min(y_i^2, x_i^2)} \leq \frac{1}{\varepsilon^*} \leq \text{poly}(n)$ , which shows  $\xi' \leq \text{poly}(n)$ .

Next, we analyze  $Q'$ : For all  $x, y \in \mathcal{I}'$ , we have

$$\frac{\|\nabla \text{ISD}(x) - \nabla \text{ISD}(y)\|}{\|x - y\|} \leq d \cdot \max_i \frac{|1/x_i - 1/y_i|}{|x_i - y_i|}.$$

By the mean value theorem, the latter is  $d$  times the absolute value of the derivative of  $1/z$  at some point between  $x_i$  and  $y_i$ , which is  $1/z^2$ . Since  $x, y \in \mathcal{I}'$ , we get  $Q' \leq \frac{4n^2d}{\varepsilon^{*2}} \leq \text{poly}(n)$ .

If  $\mathcal{X} \subseteq \mathcal{D}$ , then, after the first round, we have  $P \leq \text{poly}(n)$  since  $D = \text{poly}(n)$ . Altogether, we obtain the following result.

**Theorem 3.8.** *The smoothed running-time of  $k$ -means using ISD is bounded from above by*

$$\text{poly}\left(n^{\sqrt{k}}, \frac{1}{\sigma}\right).$$

and

$$k^{kd} \cdot \text{poly}\left(n, \frac{1}{\sigma}\right).$$

## 4 Lower Bound

In this section, we transfer the exponential lower bound proved by Vattani [18] to almost arbitrary Bregman divergences. Our starting point is his lower bound construction.

**Theorem 4.1** (Vattani [18]). *For squared Euclidean distances, there exist sets  $\mathcal{X} \subseteq \mathbb{R}^d$  of  $n$  points on which the  $k$ -means method requires  $2^{\Omega(n)}$  iterations when initialized with a particular set of cluster centers. Here,  $k$  depends on  $n$  and  $d \geq 2$  is arbitrary.*

The general idea to obtain lower bounds for general Bregman divergences is as follows: First, given an arbitrary symmetric positive definite  $A$ , we map the point set  $\mathcal{X}$  in Theorem 4.1 to a point set  $\mathcal{X}'$  such that  $k$ -means behaves on  $\mathcal{X}'$  w.r.t. the Mahalanobis distance  $m_A$  exactly like on  $\mathcal{X}$  w.r.t. squared Euclidean distances. In particular, if the latter requires  $T$  iterations, the former also requires  $T$  iterations. In the second step, we show that every good-natured (which roughly means three times differentiable) Bregman divergence behaves locally like some Mahalanobis distance. Thus, we can transfer the lower bound from squared Euclidean via Mahalanobis to arbitrary distances.

For the second transfer (from Mahalanobis to arbitrary distances), we need a notion of stability of an instance: Let  $d_\Phi$  be a Bregman divergence, let  $\mathcal{X}$  be a point set, and let  $c_1, \dots, c_k$  be initial centers. The instance  $\mathcal{X}, c_1, \dots, c_k$  is called  $d_\Phi$ -stable with slack  $\nu > 0$  if the following holds for all  $x \in \mathcal{X}$  and all iterations: Assume that after reassignment in this iteration,  $x$  belongs to  $\mathcal{C}_i$  with center  $c'_i$ . Then  $d_\Phi(x, c'_i) < d_\Phi(x, c'_j) - \nu$  for all  $j \neq i$ , where  $c'_j$  is the center of cluster  $\mathcal{C}_j$ . We say that an instance is  $d_\Phi$ -stable if there exists a constant  $\nu > 0$  such that it is  $d_\Phi$ -stable with slack  $\nu$ .

If an instance is  $d_\Phi$ -stable, then there is never a point that lies exactly on a bisecting hyperplane. Intuitively, if an instance is  $d_\Phi$ -stable, then (very) slightly perturbing the point set does not change the behavior of  $k$ -means.

First, we show that all Mahalanobis distances are equivalent in terms of the worst-case number of iterations. Vattani's lower bound is for squared Euclidean distances, which are a special case of Mahalanobis distances. Thus, we get an exponential lower bound for all Mahalanobis distances. Let  $W_{d_\Phi}^{k,d}(n)$  be the maximum number of iterations of  $k$ -means on any  $d_\Phi$ -stable instance of  $n$  points in  $\mathbb{R}^d$  using  $d_\Phi$  as the distance measure.

**Lemma 4.2.** *For every symmetric positive definite matrix  $A \in \mathbb{R}^{d \times d}$ , we have  $W_{m_A}^{k,d}(n) = W_{m_I}^{k,d}(n)$  for all  $n, k, d \in \mathbb{N}$ .*

*Proof.* Let  $\mathcal{X} \subseteq \mathbb{R}^d$  be a set of  $n$  points and let  $c_1, \dots, c_k \in \mathbb{R}^d$  be initial cluster centers on which  $k$ -means using squared Euclidean distances needs  $W_{m_I}^{k,d}(n)$  iterations.

Since  $A$  is symmetric and positive definite, there exists a matrix  $M \in \mathbb{R}^{d \times d}$  such that  $A = M^T M$  (Cholesky factorization [10, Theorem 4.2.5]). Let  $B = M^{-1}$  (since  $A$  is positive definite,  $M$  has full rank), and let  $y = x - x'$  for any  $x, x' \in \mathbb{R}^d$ . Then

$$d_{m_I}(x, x') = \|x - x'\|^2 = y^T y = y^T (B^T M^T) (M B) y = (B y)^T A (B y) = d_{m_A}(B x, B x').$$

Now let  $\mathcal{X}' = \{B x \mid x \in \mathcal{X}\}$  and  $c'_i = B c_i$  for  $i \in [k]$ . Then  $k$ -means behaves w.r.t. squared Euclidean distances on  $\mathcal{X}'$  initialized with centers  $c_1, \dots, c_k$  exactly in the same way as w.r.t. Mahalanobis distance  $d_{m_A}$  on  $\mathcal{X}$  initialized with centers  $c'_1, \dots, c'_k$ . This shows  $W_{m_I}^{k,d}(n) \leq W_{m_A}^{k,d}(n)$ .

To show that  $W_{m_I}^{k,d}(n) \geq W_{m_A}^{k,d}(n)$ , we observe that any worst-case instance for  $d_{m_A}$  can be transformed to an instance for squared Euclidean distances using  $B^{-1}$ .  $\square$

Now we transfer worst-case instances for Mahalanobis distances to instances for arbitrary good-natured Bregman divergences. For this, we use the observation that any good-natured Bregman divergence  $d_\Phi$  behaves locally at some point  $z_0$  like the Mahalanobis distance  $d_{m_H}$ , where  $H$  is the Hessian matrix of  $\Phi$  at  $z_0$ . Hence, essentially we only need to scale down the worst-case instance for  $d_{m_H}$  and embed it locally into a small space around  $z_0$ .

**Lemma 4.3.** *Let  $\Phi : X \rightarrow \mathbb{R}$  be a strictly convex function with  $X \subseteq \mathbb{R}^d$  and the following properties: There exist a  $z_0 \in X$  and a  $\zeta > 0$  such that*

- $Z = \{z \in \mathbb{R}^d \mid \|z - z_0\|_\infty \leq \zeta\} \subseteq X$ ,
- all third-order derivatives of  $\Phi$  exist on  $Z$  and their absolute values are bounded, and
- the Hessian matrix of  $\Phi$  at  $z_0$  is positive definite.

Then  $W_{d_\Phi}^{k,d}(n) \geq W_{m_I}^{k,d}(n)$ .

*Proof.* First we show that  $d_\Phi$  behaves locally around  $z_0$  almost like the Mahalanobis distance  $d_{m_H}$ , where  $H$  denotes the Hessian matrix of  $\Phi$  at  $z_0$ . For this, let  $\tilde{\Phi}(y) = \Phi(z_0 + y)$ , let  $f = \Phi(z_0) = \tilde{\Phi}(0)$ , and let  $g = \nabla \Phi(z_0) = \nabla \tilde{\Phi}(0)$  be the gradient of  $\Phi$  at  $z_0$ . All third-order derivatives of  $\Phi$  on  $Z$  are bounded in absolute value, say, by  $c$ . This implies that all third-order derivatives of  $\tilde{\Phi}$  are bounded by  $c$  in  $\tilde{Z} = \{y \mid \|y\|_\infty \leq \zeta\}$ .

We use the Taylor expansion (cf. Lang [13, §6]) of  $\tilde{\Phi}$ , which yields, for all  $y \in \tilde{Z}$  with  $\|y\|_\infty \leq \varepsilon \leq \zeta$ ,

$$\tilde{\Phi}(y) = f + g^T y + \frac{1}{2} y^T H y + R(y).$$

The remainder term  $R(y)$  is bounded in absolute value by

$$|R(y)| \leq \int_0^1 \frac{(1-t)^2}{2} d^3 c \varepsilon^3 dt \in O(cd^3 \varepsilon^3)$$

since the third-order derivatives are bounded by  $c$ . In the same way, we get

$$\nabla \tilde{\Phi}(y) = g + H y + R'(y)$$

with

$$\|R'(y)\|_\infty \in O(cd^2 \varepsilon^2).$$

Now let  $y, y' \in \tilde{Z}$  with  $\|y\|_\infty, \|y'\|_\infty \leq \varepsilon$ , and let  $z = z_0 + y$  and  $z' = z_0 + y'$ . Then

$$\begin{aligned} d_\Phi(z, z') &= \tilde{\Phi}(y) - \tilde{\Phi}(y') - (y - y')^T \cdot \nabla \tilde{\Phi}(y') \\ &= f + g^T y + \frac{1}{2} y^T H y + R(y) - (f + g^T y' + \frac{1}{2} y'^T H y' + R(y')) \\ &\quad - (y - y')^T \cdot (g + H y' + R'(y')). \end{aligned}$$

We observe that  $|\langle y - y', R'(y') \rangle| \in O(cd^3\varepsilon^3)$  and  $|R(y)|, |R(y')| \in O(cd^3\varepsilon^3)$ . This yields

$$\begin{aligned} d_\Phi(z, z') &= \frac{1}{2} y^T H y - \frac{1}{2} y'^T H y' - (y - y')^T H y' + O(cd^3\varepsilon^3) \\ &= \frac{1}{2} \left( y^T H y + (y')^T H y' - y'^T H y - y^T H y' \right) + O(cd^3\varepsilon^3) \\ &= \frac{1}{2} (y - y')^T H (y - y') + O(cd^3\varepsilon^3) \\ &= \frac{1}{2} d_{m_H}(y, y') + O(cd^3\varepsilon^3), \end{aligned}$$

where the equalities hold due to some rearrangements and since  $H$  is a symmetric matrix.

Due to Lemma 4.2, there exists a set  $\mathcal{X} \subseteq \mathbb{R}^d$  of  $n$  points and centers  $c_1, \dots, c_k \in \mathbb{R}^d$  such that the resulting instance is  $d_{m_H}$ -stable with some slack  $\nu > 0$  and  $k$ -means needs  $W_{m_I}^{k,d}(n)$  iterations using  $m_H$ . We construct an instance  $\tilde{\mathcal{X}} \subseteq Z$  of  $n$  points and initial centers  $\tilde{c}_1, \dots, \tilde{c}_k$  on which  $k$ -means using  $d_\Phi$  also needs  $W_{m_I}^{k,d}(n)$  iterations. We can assume w.l.o.g. that  $\mathcal{X} \subseteq [-1, 1]^d$  and  $c_1, \dots, c_k \in [-1, 1]^d$ . If we use  $\frac{1}{2}d_{m_H}$  instead, then the instance is still stable with slack  $\nu/2$ . If we scale down this instance by a factor of  $\varepsilon > 0$ , then the resulting instance is still  $\frac{1}{2}d_{m_H}$ -stable with slack  $\varepsilon\nu/2$ . Thus, if we distort the distance measure by at most  $\nu\varepsilon/4$ ,  $k$ -means using the scaled down version of  $\mathcal{X}$  and  $\frac{1}{2}d_{m_H}$  behaves exactly the same way as  $k$ -means on  $\mathcal{X}$  using  $d_{m_H}$ .

Let  $\tilde{\mathcal{X}} = \{z_0 + \varepsilon y \mid y \in \mathcal{X}\}$  and  $\tilde{c}_i = z_0 + \varepsilon c_i$ . This yields  $\tilde{\mathcal{X}} \subseteq Z$  and  $\tilde{c}_1, \dots, \tilde{c}_k \in Z$  because  $\varepsilon \leq \zeta$ . The  $k$ -means method behaves on  $\tilde{\mathcal{X}}$  w.r.t.  $d_\Phi$  like on  $\mathcal{X}$  w.r.t.  $d_{m_H}$  if

$$\left| d_\Phi(z, z') - \frac{1}{2} d_{m_H}(y, y') \right| < \frac{\nu\varepsilon}{4}.$$

Since the difference is bounded by  $O(cd^3\varepsilon^3)$ , this can be achieved by making  $\varepsilon > 0$  sufficiently small.  $\square$

Vattani's lower bound construction [18] is  $d_{m_I}$ -stable. Combining this construction with Lemma 4.2 and Lemma 4.3, we obtain the main result of this section.

**Theorem 4.4.** *The worst-case number of iterations of  $k$ -means for the following Bregman divergences is at least  $\exp(\Omega(n))$  for  $n$  points and  $d \geq 2$ :*

1. Mahalanobis distances for any symmetric positive definite matrix  $A$ ,
2. Kullback-Leibler divergence (KLD),
3. generalized I-divergence (GID),
4. Itakura-Saito divergence (ISD).

*Proof.* For Mahalanobis distances, this follows immediately from Vattani’s lower bound [18] and Lemma 4.2.

The domain of the Kullback-Leibler divergence (KLD) is  $X = \{z \in \mathbb{R}^d \mid z \geq 0, \sum_{i=1}^d z_i \leq 1\}$ . We choose  $z_0 = (\frac{1}{d+1}, \dots, \frac{1}{d+1}) \in X$ . Then  $Z = \{z \in Y \mid \|z - z_0\|_\infty \leq \zeta\} \subseteq X$  for  $\zeta = 1/(d+1)^2$ . The convex function corresponding to KLD is  $\text{KLD}(x) = \sum_{j=1}^{d+1} x_j \log x_j$ , where  $x_{d+1} := 1 - \sum_{j=1}^d x_j$ . Simple calculus shows that  $\frac{\partial \text{KLD}(x)}{\partial x_i} = \log x_i - \log x_{d+1}$ ,  $\frac{\partial^2 \text{KLD}(x)}{\partial x_i^2} = \frac{1}{x_i} + \frac{1}{x_{d+1}}$ , and  $\frac{\partial^2 \text{KLD}(x)}{\partial x_i \partial x_j} = \frac{1}{x_{d+1}}$ , for  $i \neq j$ . Hence, the diagonal entries of the Hessian matrix at  $z_0$  are all  $2(d+1)$  while the other entries are all  $(d+1)$ . This matrix is positive definite. It only remains to consider the third-order derivatives, which are of the form

$$\frac{\partial^3 \text{KLD}(x)}{\partial x_i^3} = -\frac{1}{x_i^2} + \frac{1}{x_{d+1}^2} \quad \text{and} \quad \frac{\partial^3 \text{KLD}(x)}{\partial x_i \partial x_j \partial x_\ell} = \frac{1}{x_{d+1}^2}$$

if not  $i = j = \ell$ . For our choice of  $\zeta$  all these derivatives are bounded by  $c = 2(d+1)^2/d$  in  $Z$ , which concludes the proof for KLD.

The lower bound for generalized I-divergence follows analogously by choosing, e.g.,  $z_0 = (1, \dots, 1)$ .

For the Itakura-Saito divergence, we can again choose  $z_0 = (1, \dots, 1)$ . We have  $\frac{\partial \text{ISD}(x)}{\partial x_i} = \frac{-1}{x_i}$  and  $\frac{\partial^2 \text{ISD}(x)}{\partial x_i^2} = \frac{1}{x_i^2}$  and, for  $i \neq j$ ,  $\frac{\partial^2 \text{ISD}(x)}{\partial x_i \partial x_j} = 0$ . Thus, the Hessian matrix at  $z_0$  is the identity matrix, which is of course positive definite. All third-order derivatives are 0 with the exception of  $\frac{\partial^3 \text{ISD}(x)}{\partial x_i^3} = \frac{-2}{x_i^3}$  for  $i \in \{1, \dots, d\}$ . For  $\zeta = 1/2$ , the absolute values of all third-order derivatives around  $z_0$  are bounded by 16, which completes the proof.  $\square$

The results of this section prove that for very general distance measures, the worst-case running-time of  $k$ -means is poor, which complements our smoothed analysis. Furthermore, the two reductions (Lemmas 4.2 and 4.3) indicate that squared Euclidean distances and Mahalanobis distances are in some sense the easiest distances for  $k$ -means, as the lower bound for them carries over to other good-natured Bregman divergences.

## 5 Concluding Remarks

We have shown that the smoothed running-time of  $k$ -means using Bregman divergences is bounded by a polynomial in  $n^{\sqrt{k}}$  and  $1/\sigma$  and by  $k^{kd} \text{poly}(n, 1/\sigma)$ , given that certain parameters that characterize the Bregman divergence are bounded by a polynomial. On the other hand, we proved exponential lower bounds for the worst-case running-time of  $k$ -means using Bregman divergences that are three times differentiable. In particular, these results hold for Mahalanobis distances (the upper bound requires that the largest eigenvalue of the matrix used is bounded by a polynomial), Kullback-Leibler divergence, generalized I-divergence, and Itakura-Saito divergence.

Recently, Arthur et al. [3] proved that the smoothed running-time of  $k$ -means for squared Euclidean distances is bounded by a polynomial in  $n$  and  $1/\sigma$ . An obvious open question is whether this results carries over to Bregman divergences. However, their analysis exploits specific properties of Gaussian noise like, for example, that the projection of a Gaussian onto a lower-dimensional subspace is still a Gaussian with the same standard deviation. There is no straightforward way of adapting this bound to our general perturbation model.

## References

- [1] Marcel R. Ackermann and Johannes Blömer. Coresets and approximate clustering for Bregman divergences. In *Proc. of the 20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 1088–1097, 2009.
- [2] Marcel R. Ackermann, Johannes Blömer, and Christian Sohler. Clustering for metric and non-metric distance measures. In *Proc. of the 19th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 799–808, 2008.
- [3] David Arthur, Bodo Manthey, and Heiko Röglin.  $k$ -means has polynomial smoothed complexity. In *Proc. of the 50th Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 405–414, 2009.
- [4] David Arthur and Sergei Vassilvitskii. Worst-case and smoothed analysis of the ICP algorithm, with an application to the  $k$ -means method. *SIAM Journal on Computing*, 39(2):766–782, 2009.
- [5] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.
- [6] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, USA, 2002.
- [7] Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.
- [8] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, 2000.
- [9] William Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley & Sons, 1971.
- [10] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3rd edition, 1996.
- [11] Robert M. Gray, Andrés Buzo, Augustine H. Gray Jr., and Yasuo Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):367–376, 1980.
- [12] Mary Inaba, Naoki Katoh, and Hiroshi Imai. Variance-based  $k$ -clustering algorithms by Voronoi diagrams and randomization. *IEICE Transactions on Information and Systems*, E83-D(6):1199–1206, 2000.
- [13] Serge Lang. *Real Analysis*. Addison-Wesley, 1969.
- [14] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
- [15] Bodo Manthey and Heiko Röglin. Improved smoothed analysis of the  $k$ -means method. In *Proc. of the 20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 461–470, 2009.

- [16] Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. On Bregman voronoi diagrams. In *Proc. of the 18th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 746–755, 2007.
- [17] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004.
- [18] Andrea Vattani.  $k$ -means requires exponentially many iterations even in the plane. In *Proc. of the 25th ACM Symp. on Computational Geometry (SoCG)*, pages 324–332, 2009.