WORST-CASE AND SMOOTHED ANALYSIS OF K-MEANS CLUSTERING WITH BREGMAN DIVERGENCES*

Bodo Manthey[†] *and Heiko Röglin*[‡]

Abstract. The k-means method 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 σ, 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 the 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.

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 a wide range of areas. Usually the objects are represented by points in $\mathbb{R}^d$, and they are to be clustered into k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ that can be represented by centers $c_1, \ldots, c_k \in \mathbb{R}^d$ 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

[*] A preliminary version of this paper has appeared in *Proc. 20th Int. Symp. on Algorithms and Computation (ISAAC 2009)* [18]. This paper also includes parts of "Improved Smoothed Analysis of the k-Means Method" [17].

[†] *University of Twente, Department of Applied Mathematics, P. O. Box 217, 7500 AE Enschede, Netherlands*, b.manthey@utwente.nl. Work done in part at Saarland University, Department of Computer Science, Saarbrücken, Germany.

[‡] *University of Bonn, Department of Computer Science, Friedrich-Ebert-Allee 144, 53113 Bonn, Germany*, roeglin@cs.uni-bonn.de. Supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD) and by a Veni grant from the Netherlands Organisation for Scientific Research (NWO).

distance measures are required. For instance, when clustering text documents like web pages often the *bag-of-words model* [8] 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 many 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* [7], a local search algorithm due to Lloyd [16]: 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 [9, Section 10.4.3]. This is in stark contrast to its worst-case running-time: The only upper bound is $n^{O(kd)}$ [14], which is based on the observation that no clustering appears twice in a run of k-means. On the other hand, Vattani [24] 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 [22]. We refer to two surveys [19, 23] for an overview over smoothed analysis. 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 noise with standard deviation σ. We have improved this bound to $\text{poly}(n^{\sqrt{k}}, 1/\sigma)$, and we have additionally obtained a bound of $k^{kd} \cdot \text{poly}(n, 1/\sigma)$ [17]. Recently, Arthur et al. [3] have shown 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 theory 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 $\mathrm{poly}(n^{\sqrt{k}}, 1/\sigma)$ and $k^{kd} \cdot \mathrm{poly}(n, 1/\sigma)$.

In the next section, we define Bregman divergences (Section 2.1), describe the k-means method (Section 2.2), and discuss perturbation models for Bregman divergences (Section 2.3). We summarize our results in Section 3. After that, we present the generic smoothed analysis (Section 4) and apply it to some prominent Bregman divergences (Section 5). Finally, we prove a lower bound on the running-time of k-means with Bregman divergences (Section 6) and conclude with some open problems (Section 7).

2 Preliminaries

2.1 Bregman Divergences

The idea behind Bregman divergences and their use as distance measures is quite simple: Assume that we have a strictly convex function Φ, and assume that we have two points x and c whose distance we want to measure. We take the linear interpolation $\overline{\Phi}$ of Φ from c. We say that the distance from x to c is the amount by which we underestimate $\Phi(x)$, i.e., $\Phi(x) - \overline{\Phi}(x)$. Since Φ is strictly convex, $\overline{\Phi}(x)$ underestimates $\Phi(x)$. Thus, we have $\overline{\Phi}(x) \leq \Phi(x)$ with equality only for $x = c$. The following definition makes this rigorous.

Definition 2.1. *Let* $\mathbb{D} \subseteq \mathbb{R}^d$*, and let* $\Phi : \mathbb{D} \to \mathbb{R}$ *be a strictly convex function such that* Φ *is differentiable on the relative interior* $\mathrm{ri}(\mathbb{D})$ *of* $\mathbb{D}$*. The* Bregman divergence $d_\Phi : \mathbb{D} \times \mathrm{ri}(\mathbb{D}) \to [0, \infty)$ *is defined as*

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

where $\nabla\Phi(c)$ *is the gradient of* Φ *at* c*.*

Note that d_Φ is not a metric (even squared Euclidean distances, which are a Bregman divergence, are not a metric): First, it does not satisfy the triangle inequality. Second, it is often not even symmetric.

Some important properties of squared Euclidean distances are true for Bregman divergences [20]. For a finite set of points $C \subseteq \mathbb{D}$, we denote the center of mass of C by $\mathrm{cm}(C) = \frac{1}{|C|} \sum_{x \in C} x$. 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,

$$\sum_{x \in C} d_\Phi(x, c) = \sum_{x \in C} d_\Phi(x, \mathrm{cm}(C)) + |C| \cdot d_\Phi(\mathrm{cm}(C), c). \tag{1}$$

Another important property of Bregman divergences is that the bisector of two centers c and c', i.e., the set $\{x \in \mathbb{D} \mid d_\Phi(x, c) = d_\Phi(x, c')\}$, is a hyperplane. This follows immediately from the definition of d_Φ, and it is important both for our analysis and to obtain a worst-case upper bound for k-means.

In the following, we present some prominent Bregman divergences.

Mahalanobis Distances. Assume that we want to cluster objects that are each characterized by d quantities. If these quantities are independent, then clusters should be axis-aligned hyper-ellipses. If the d quantities share the same scale, then the clusters are even hyperspheres and squared Euclidean distances provide a good distance measure.

However, if the coordinates are correlated or scaled differently, then clusters cease to be axis-aligned or hyperspheres, but are some hyper-ellipse. In that case, let $C \in \mathbb{R}^{d \times d}$ be the covariance matrix of the components of the data points and assume that it is invertible. This means that the matrix C is symmetric and positive definite. Let $A = C^{-1}$, then the right distance measure taking into account the correlations is the *Mahalanobis distance* d_{m_A} for $m_A(x) = x^T Ax$. 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. This means that $d_{m_I}(x,c) = \|x-c\|^2$.)

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

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

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 [8].

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

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

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) [5,13]. We have again $\mathbb{D} = \{x \in \mathbb{R}^d \mid x \geq 0\}$, and the potential function is given by the Burg entropy $\mathrm{ISD}(x) = -\sum_{i=1}^d \log(x_i)$. From this, we get the Bregman divergence

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

2.2 k-Means Method

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

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

is minimized, where Φ is a strictly convex function and, thus, d_Φ is a Bregman divergence (see Definition 2.1).

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 $c_1, \ldots, c_k$ should be chosen so as to minimize the potential. According to (1), we should choose

$$c_i = \frac{1}{|\mathcal{C}_i|} \cdot \sum_{x \in \mathcal{C}_i} x,$$

i.e., we should choose c_i as the center of mass of its cluster $\mathcal{C}_i$, in order to minimize the potential.

The k-means method proceeds as follows:

1. Select cluster centers $c_1, \ldots, c_k \in \mathbb{D} \subseteq \mathbb{R}^d$ arbitrarily.

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. Otherwise, we break ties arbitrarily.)

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 in a local optimum.

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, 14]. The proof of this bound relies only on the fact that the bisectors are hyperplanes. This is true not only for squared Euclidean distances, but for any Bregman divergence. Hence, also for Bregman divergences, the worst-case number of iterations cannot exceed W.

2.3 Perturbation Models for Bregman Divergences

2.3.1 Natural Perturbations and Exponential Families

If the Bregman divergence is defined on the whole space $\mathbb{R}^d$, i.e., if $\mathbb{D} = \mathbb{R}^d$, then it is often considered 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 an independent d-dimensional Gaussian random variable with standard deviation σ to it. Equivalently, we can also assume that each point from $\mathcal{X}$ is a Gaussian random vector with standard deviation σ 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, 17]. But if $\mathbb{D}$ is a proper subset of $\mathbb{R}^d$, as it is the case for KLD or GID, then Gaussians cannot be used as it might yield points outside the feasible region $\mathbb{D}$. For this reason, special care is needed when defining perturbation models for Bregman divergences.

However, it is not surprising that Gaussian noise is not suitable for all Bregman divergences. In fact, other probability distributions are often more natural: Banerjee et al. [5], making observations by Forster and Warmuth rigorous [11], show a nice correspondence between Bregman divergences and exponential families of probability distributions [6]. What they basically show is that the Bregman divergence to some center c equals the negative log-likelihood of a corresponding parametric exponential family with expectation parameter c (up to some fixed function that is independent of the parameter). Let us briefly explain exponential families and what "corresponding" means in this context.

A parametric exponential family of probability distributions is given by its density

$$x \mapsto \exp\bigl(\theta^T \tilde{x} - \Phi^*(\theta)\bigr) \cdot p_0(\tilde{x}).$$

Here, $\tilde{x}$ is a sufficient statistic for the family and θ is the parameter. For instance, if we choose $\tilde{x}^T = x$, $\theta = \mu$, $\Phi^*(\theta) = \theta^2/2 = \|\mu\|^2/2$, and $p_0(x) = (2\pi)^{d/2} \cdot \exp(-\|x\|^2/2)$, we obtain that the d-dimensional Gaussian distribution with uniform variance is an exponential family, parameterized by its its mean [5, Example 9]. We refer to Barndorff-Nielsen [6] for a thorough introduction to exponential families.

Assume that we have a Bregman divergence d_Φ generated by some strictly convex function Φ. Let Φ^* be the Legendré-Fenchel dual (or convex conjugation) of Φ [21], i.e., we have

$$\Phi^*(\theta) = \sup\bigl\{x^T\theta - \Phi(x) \mid x \in \mathbb{D}\bigr\}.$$

The dual Φ^* is also strictly convex. From the dual Φ^*, we obtain a parametric exponential family of probability distributions as described above, and this parametric exponential family is the one that corresponds to d_Φ.

Squared Euclidean distances are a Bregman divergence for $\Phi(x) = \|x\|^2$. Replacing this by $\Phi(x) = \|x\|^2/2$ only scales the distances. Now we have $\Phi^*(\theta) = \|\theta\|^2/2$, which shows that squared Euclidean distances correspond to the exponential family of Gaussian distributions.

The exponential families corresponding to KLD are multinomial distributions. For Itakura-Saito divergence, we have exponential distributions [5].

However, while Gaussian perturbations are perfect for a smoothed analysis, multinomial distributions cannot be used easily. Multinomial distributions are discrete, and we need continuous probability distributions for our analysis to work. If we make them continuous, we obtain multivariate Gaussian distributions, which can yield negative values and, thus, are useless for KLD.

For these reasons, we decided to use exponential distributions for KLD, Itakura-Saito and generalized I-divergence in the following way: Given the adversarial points, we add independent exponentially distributed random variables to each coordinate, each with the same parameter. In case of KLD, the point thus obtained is not in the domain. Thus, we scale it afterwards to make sure that the some of the coordinates is 1.

Let us stress that our smoothed analysis is not restricted to these perturbation models: In the following section, we describe which properties we demand from the perturbation models and which parameters we extract from the Bregman divergences. Then we can apply our smoothed analysis to any combination of Bregman divergence and perturbation model that satisfies these properties.

2.3.2 Assumptions and Parameterization

In order to make our smoothed analysis applicable to a large class of Bregman divergences and perturbation models, we decided to consider very general perturbation models that need to satisfy only a couple of properties. In this section, we list these properties and explain how the perturbation model is parameterized. (The parameters will influence the smoothed running-time bounds.)

Assumptions. First of all, 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.

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

We assume that the following properties are satisfied for $\sigma \in (0, 1]$:

(1) For any $\varepsilon \geq 0$, any hyperplane H, and any point in $x \in \mathbb{D} \cap [0, 1]^d$, the probability that the perturbed version of x has a distance of at most ε from H is bounded from above by $\sqrt{\varepsilon}/\sigma$.

(2) For any $x \in \mathbb{D} \cap [0, 1]^d$, the perturbed version of x can be described by a probability density function that is bounded from above by $(1/\sigma)^d$ on $\mathbb{R}^d$.

(3) Perturbed points cannot be too far away from the hypercube $[0, 1]^d$: Let D be chosen such that with a probability of 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-case number of steps of k-means.

Let us make a few remarks about these assumptions:

- For Gaussian noise, the probability of being close to a hyperplane is even ε/σ. However, to gain some flexibility for choosing other perturbation models, we use the weaker bound of $\sqrt{\varepsilon}/\sigma$.
- The bound on the density immediately implies that for any $\varepsilon \geq 0$, any $c \in \mathbb{R}^d$, and any $x \in \mathbb{D} \cap [0, 1]^d$, the perturbed version of x lies in some hyperball with radius ε and center c with a probability of at most $(2\varepsilon/\sigma)^d$.
- The bounds on the smoothed running-time that we obtain depend polynomially on D. For Gaussian random vectors with mean in $[0, 1]^d$ and standard deviation $\sigma \leq 1$, we can choose D polynomially in n.
- The parameter D, where $2D + 1$ is the side length of the cube $\mathcal{D}$, of course depends on σ. However, as we have $\sigma \leq 1$ and D should increase with σ, we can simply use always the value for D that we obtain from $\sigma = 1$.
- For Bregman divergences with a bounded domain, such as KLD, we can choose D simply sufficiently large such that the whole domain is contained in $\mathcal{D}$.

Parameterization. For our analysis to work, we have to define a few parameters that basically measure how close a Bregman divergence is to the squared Euclidean distance.

Recall that $\mathbb{D}$ is the domain of the Bregman divergence and $\mathcal{D} = [-D, D + 1]^d$ is a hypercube that contains all points (after perturbation) with a probability of at least $1 - W^{-1}$. For $\varepsilon \geq 0$, let $\mathcal{I}(\varepsilon)$ be the interior of $\mathbb{D} \cap \mathcal{D}$ that has a distance of at least ε to the boundary:

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

Note that, on the one hand, for squared Euclidean distances, we have $\mathbb{D} = \mathbb{R}^d$ and thus $\mathbb{D} \cap \mathcal{D} = \mathcal{D}$. On the other hand, for KLD, we have $\mathbb{D} \subseteq [0, 1]^d \subseteq \mathcal{D}$ and thus $\mathbb{D} \cap \mathcal{D} = \mathbb{D}$.

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

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

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

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

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

Observe that for the definition of ξ' only the interior of $\mathbb{D} \cap D$ is relevant. This is important: If we had let $x, y \in \mathbb{D} \cap \mathcal{D}$ instead of $x, y \in \mathcal{I}'$ for the definition of ξ', then ξ' is unbounded for many Bregman divergences.

We also need a lower bound on the "second derivative" of Φ: We have

$$2\xi \leq \frac{\|\nabla\Phi(x) - \nabla\Phi(y)\|}{\|x - y\|} \tag{3}$$

for all $x, y \in \mathbb{D} \cap \mathcal{D}$ with $x \neq y$. Since $\nabla\Phi^* = (\nabla\Phi)^{-1}$ [21, Theorem 26.5], we can view this as a Lipschitz condition on the gradient of the dual of Φ:

$$\frac{\|\nabla\Phi^*(x) - \nabla\Phi^*(y)\|}{\|x - y\|} \leq \frac{1}{2\xi}.$$

We can derive (3) from (2) as follows:

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

Dividing both sides by $\|x - y\|^2$ yields the desired bound. Similarly, we need an upper bound, which unfortunately cannot be derived easily from ξ'. Instead, we define it in terms of the maximum eigenvalue of the Hessian matrix of Φ, which is denoted by $\nabla^2\Phi$:

$$Q' = \sup_{x \in \mathcal{I}'} \lambda_{\max}\big(\nabla^2\Phi(x)\big).$$

In particular, this implies $\|\nabla\Phi(x) - \nabla\Phi(y)\| \leq Q' \cdot \|x - y\|$ for $x, y \in \mathcal{I}'$. Again, note that we need the upper bound Q' only for the interior.

Let us conclude this part with some remarks:

- The ratio ξ'/ξ is closely related to the μ in the notion of μ-similarity introduced by Ackermann et al. [2]. However, for any μ, Bregman divergences like KLD, GID, or ISD are not μ-similar on their whole domain. To make them μ-similar, their domains have been restricted such that all data points must be sufficiently far away from the singularities [2]. We emphasize that our smoothed analysis does not need such restrictions. There may be points arbitrarily close to the boundary of the domain, but we can take special care of these points. This technical challenge is the reason for the definition of $\mathcal{I}$ and $\mathcal{I}'$ above.

- For our Bregman divergences, the boundary is defined by hyperplanes. Thus, Property (1) yields that it is unlikely for a single point to be very close to the boundary. In particular, the probability that more than $O(kd)$ points are outside of $\mathcal{I}$ is bounded from above by $\text{poly}(n^{-kd})$.

- Our smoothed running-time bounds will depend polynomially on ξ, ξ' and Q'.

3 Our Contributions

3.1 Smoothed Analysis

Results. 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 Bregman divergences (Theorem 5.1). 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', ξ', and ε^*. (Thus, a special consideration of the boundary of the domain, as we did by introducing $\mathcal{I}$ and $\mathcal{I}'$, is not necessary to obtain this bound.)

Our second bound on the smoothed running-time is $k^{kd} \operatorname{poly}(n, 1/\sigma)$ (Theorem 5.2). This yields a polynomial smoothed running-time if $k, d \in O(\sqrt{\log n / \log\log n})$ (Corollary 5.3). Indeed, k and d are usually much smaller than n in practice; in fact, they are often even considered as constants. This second bound depends polynomially on the parameters Q', ξ', $1/\xi$, and $1/\varepsilon^*$.

Section 4 contains the core of the analysis. In Section 5, we state our generic smoothed analysis theorems and apply them to the specific Bregman divergences introduced in Section 2.1.

Note that we only analyze the running-time in this paper. We do not analyze how close the local optimum found by the k-means method is to the global optimum. In fact, k-means method does not usually seem find the global optimum in practice. But it usually seems to be fast. Thus, our smoothed analysis of time but not of accuracy matches the observed performance of the k-means method.

Main Idea and Technical Difficulties. Our smoothed analysis of k-means with Bregman divergences uses a novel lemma about perturbed point sets (Lemma 4.1): 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 still had to tackle several problems. Let us describe the main problem by way of example: For KLD, if we had defined the parameters ξ' and Q' in terms of the whole domain, they would have been unbounded. Even after the perturbation, some of the points might still be too close to the boundary to obtain reasonable upper bounds for ξ' and Q'. Essentially, we show that, first, the kd points that are closest to the boundary can be handled separately and that, second, all other points are sufficiently far away from the boundary (this means that they lie in $\mathcal{I}$) to allow for a reasonable upper bound for both ξ' and Q'.

3.2 Lower Bounds

To complement our smoothed analysis, we transfer the lower bound of $2^{\Omega(n)}$ for squared Euclidean distances to basically all Bregman divergences d_Φ, whose third-order derivatives exist and are bounded within a small region (Section 6). This includes Mahalanobis distances, KLD, GID, and ISD.

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 if a Bregman divergence is sufficiently smooth (this includes all commonly considered examples like KLD, GID, or ISD), then it behaves locally like some Mahalanobis distance. This makes a transfer of the known lower bound for the Euclidean case possible.

4 Smoothed Analysis of k-Means with Bregman Divergences

In this section, we present the generic smoothed analysis of k-means with Bregman divergences. In order to bound the expected running-time of k-means we analyze how long it takes in expectation until the potential decreases by at least 1. For this, it is first of all important to understand which events cause a significant drop of the potential. In Section 4.2 we identify two such events: the potential decreases significantly in an iteration of k-means if either one of the centers moves significantly or if a point is reassigned that has a significant distance from the bisector separating its old and its new cluster center.

To bound the probability of these events, we prove in Section 4.1 a crucial geometric property of perturbed inputs: it is unlikely that many points are close to bisectors in any Voronoi partitioning of the data points. In particular, this property implies that the probability that there exists a Voronoi partitioning in which more than $kd/2$ data points have a distance of at most ε from one of the bisectors is negligible for an appropriately chosen ε. As these $kd/2$ points can be assigned in at most $k^{kd/2}$ different ways to the clusters, in any sequence of $k^{kd/2}+1$ consecutive iterations of k-means one point must change its assignment that has a distance of more than ε at the beginning of the sequence. This implies that either one of the centers must move significantly in the sequence of iterations or that this point has a significant distance to the corresponding bisector when it gets reassigned. This argument is made formal in Section 4.3 and it leads to a bound on the smoothed running-time of $k^{kd}\operatorname{poly}(n, 1/\sigma)$ (Theorem 5.2).

To obtain the second bound that is polynomial in $n^{\sqrt{k}}$ and $1/\sigma$ (Theorem 5.1), we distinguish between iterations of k-means with at most $\sqrt{k}$ active clusters (Section 4.4) and with at least $\sqrt{k}$ active clusters (Section 4.5). (A cluster is *active* in an iteration if it gains or loses points during the iteration.) If a lot of clusters are active, then it is unlikely that for none of them the center moves significantly. To be more precise, we show that if at least $\sqrt{k}$ clusters are active, then with high probability at least one center changes its position by at least $n^{-O(\sqrt{k})}$. If at most $\sqrt{k}$ clusters are active, then either at most $2dk$ points change their assignment or there exists a pair of clusters that exchange more than $2d$ points. In the former case it is unlikely that none of the centers moves by at least $n^{-O(\sqrt{k})}$. In the latter case it is unlikely that all points that switch between the two aforementioned clusters are close to the bisector.

4.1 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 [17], but we introduce here a new idea that shortens the proof significantly and makes the generalization possible.

Lemma 4.1. *Let $a \in [k] := \{1, \ldots, k\}$ be arbitrary. With a probability of at least $1 - 2W^{-1}$, the following holds: In every iteration of the k-means method (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*

$$\varepsilon = \left(\frac{\sigma^2}{3Dn^{11}}\right)^{4a}$$

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

Proof. We consider an iteration of the k-means method, 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. If more than ℓ points change their assignment during the step, then B can be an arbitrary subset of these points with $|B| = \ell$. There are at most n^ℓ 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 $A_1, \ldots, A_k$ 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 [14]. 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 $c_1, \ldots, c_k$ 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 $A_1, \ldots, A_k$. 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

$$\forall i,j \in [k] : \forall b \in B_{i,j} \colon \; \mathrm{dist}(b, H_{i,j}) \leq \varepsilon \,, \tag{$\mathcal{E}$}$$

where $\mathrm{dist}(x,H) = \min_{y \in H} \|x - y\|$ 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 $\left(\frac{\sqrt{\varepsilon}}{\sigma}\right)^{\ell}$. 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 $b_i \in (A_i \cup B_i) \setminus B'_i$ 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. Since $|B| = \ell$, there are at most ℓ sets B_i with $|B_i| > 0$. Hence, we choose for at most ℓ indices i a point b_i, leaving us with an additional factor of at most n^{ℓ} in the union bound. Altogether, there are at most $n^{4\ell} W$ choices in the union bound.

Let Z denote a particular choice in the union bound, and let $\mathcal{E}_Z$ be the respective event. In Lemma 4.2 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:

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

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:

$$\begin{aligned}
\Pr\bigl[\mathcal{E}\bigr] &\leq \Pr\bigl[\mathcal{F}\bigr] + \Pr\bigl[\mathcal{E} \wedge \neg\mathcal{F}\bigr] \;\leq\; \Pr\bigl[\mathcal{F}\bigr] + \sum_{Z} \Pr\bigl[\mathcal{E}_Z \wedge \neg\mathcal{F}\bigr] \\
&\leq W^{-1} + n^{4\ell} W \cdot \left(\frac{3nD}{\sigma^2}\right)^{dk} \cdot \varepsilon^{\ell/4} \\
&\leq W^{-1} + n^{4kd} W \cdot \left(\frac{3nD}{\sigma^2}\right)^{dk} \cdot \left(\frac{\sigma^2}{3Dn^{11}}\right)^{a\ell} \;\leq\; W^{-1} + n^{-3kd} \;\leq\; 2W^{-1} \,.
\end{aligned}$$

The inequalities are due to some simplifications, $W \le n^{3kd}$, and our choice of ε. □

Lemma 4.2. *For every choice Z in the union bound in the proof of Lemma 4.1, the probability of the event $\mathcal{E}_Z \wedge \neg\mathcal{F}$ is bounded from above by*

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

Proof. The union bound fixes the sets A_i, B_i, and B_i'. Additionally for every i the point b_i (its identity not its position) is fixed by the union bound. The positions of all points in A are chosen arbitrarily by an adversary while all other points are left random. The center of cluster i in the first configuration is the center of mass of the points currently assigned to that cluster. If $|B_i| > 1$, then $b_i \in B_i$, $B_i = B_i' \cup \{b_i\}$ and exactly the points $A_i \cup B_i$ are assigned to cluster i in the first configuration. If $|B_i| = 1$, then $b_i \notin B_i$, $b_i \notin A_i$ (remember that in this case we have removed b_i from A_i), and exactly the points $A_i \cup B_i \cup \{b_i\}$ are assigned to cluster i in the first configuration. Hence, if $|B_i| > 1$ the center of cluster i is at the position

$$\begin{aligned}\mathrm{cm}(A_i \cup B_i) &= \frac{|A_i|}{|A_i| + |B_i|} \cdot \mathrm{cm}(A_i) + \frac{|B_i|}{|A_i| + |B_i|} \cdot \mathrm{cm}(B_i)\\ &= \frac{|A_i|}{|A_i| + |B_i|} \cdot \mathrm{cm}(A_i) + \frac{1}{|A_i| + |B_i|} \cdot \sum_{b \in B_i' \cup \{b_i\}} b.\end{aligned}$$

Since the positions of all points in A_i are specified by an adversary and the cardinalities of the sets A_i and B_i are fixed in the union bound, the center of cluster i is completely determined if additionally the sum $\sum_{b \in B_i' \cup \{b_i\}} b$ is given. Similarly one can argue that also in the case $|B_i| = 1$ it suffices to know the sum $\sum_{b \in B_i' \cup \{b_i\}} b$ in order to determine the center of cluster i.

We define for each i a random vector g_i as $g_i := \sum_{b \in B_i' \cup \{b_i\}} b$. 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 $g_i = y_i$ and $g_j = y_j$. Let $k^\star$ be the number of clusters $\mathcal{C}_i$ with $|B_i| > 0$. Without loss of generality, these are the clusters $\mathcal{C}_1, \ldots, \mathcal{C}_{k^\star}$. This convention allows us to rewrite the probability of $\mathcal{E}_Z \wedge \neg\mathcal{F}$. When the event $\neg\mathcal{F}$ occurs, then all input points are contained in $\mathcal{D}$. Hence every random vector g_i is contained in $n\mathcal{D} := [-nD, 1 + nD]^d$. Hence,

$$\begin{aligned}\Pr\bigl[\forall i, j \colon \forall b \in B_{i,j} \setminus \{b_i\} \colon\ \mathrm{dist}(b, H_{i,j}) \le \varepsilon\bigr] \le \int_{y_1 \in n\mathcal{D}} \cdots \int_{y_{k^\star} \in n\mathcal{D}} \left(\prod_{i=1}^{k^\star} f_{g_i}(y_i)\right)\\ \cdot \Pr\bigl[\forall i, j \colon \forall b \in B_{i,j} \setminus \{b_i\} \colon\ \mathrm{dist}(b, H_{i,j}(y_i, y_j)) \le \varepsilon \bigm| \forall i \colon g_i = y_i\bigr]\, \mathrm{d}y_{k^\star} \ldots \mathrm{d}y_1 ,\end{aligned}$$

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 $j \notin \{1, \ldots, k^\star\}$, 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 $H_{i,j}(y_i, y_{j_i})$ 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

$$\int \cdots \int \left(\prod_{i=1}^{k^\star} f_{g_i}(y_i) \right) \\ \cdot \left(\prod_{i=1}^{k^\star} \Pr\big[\forall j \colon \forall b \in B_{i,j} \setminus \{b_i\} \colon \operatorname{dist}(b, H_{i,j}(y_i, y_j)) \le \varepsilon \,\big|\, g_i = y_i\big] \right) \mathrm{d}y_{k^\star} \ldots \mathrm{d}y_1 \,. \quad (4)$$

Now let us consider the probability

$$\Pr\big[\forall j \colon \forall b \in B_{i,j} \setminus \{b_i\} \colon \operatorname{dist}(b, H_{i,j}(y_i, y_j)) \le \varepsilon \,\big|\, g_i = y_i\big]$$

for a fixed i and for fixed values y_i and y_j. To simplify the notation, let $\bigcup_j B_{i,j} \setminus \{b_i\} = B'_i = \{q_1, \ldots, q_m\}$, and let the corresponding hyperplanes (which are fixed because y_i and the y_j's are given) be $H_1, \ldots, H_m$. (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

$$\Pr\big[\forall j \colon \operatorname{dist}(q_j, H_j) \le \varepsilon \,\big|\, g_i = y_i\big] \,. \quad (5)$$

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

$$\int_{z_1 \in H_1(\varepsilon)} \cdots \int_{z_m \in H_m(\varepsilon)} \frac{f(z_1, \ldots, z_m, y_i)}{f_{g_i}(y_i)} \,\mathrm{d}z_m \ldots \mathrm{d}z_1 \,.$$

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

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

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

If we plug this bound into (4), the density f_{g_i} cancels out for every i. Hence, the term in the integral does not depend anymore on the values $y_1, \ldots, y_{k^\star}$. We obtain the following upper bound for (4):

$$\begin{aligned} &\left(\frac{\sqrt{\varepsilon}}{\sigma} \right)^{|B'_1| + \ldots + |B'_k|} \frac{1}{\sigma^{dk^\star}} \int_{y_1 \in n\mathcal{D}} \cdots \int_{y_{k^\star} \in n\mathcal{D}} 1 \,\mathrm{d}y_{k^\star} \ldots \mathrm{d}y_1 \\ &\le \left(\frac{3nD}{\sigma} \right)^{dk^\star} \cdot \left(\frac{\sqrt{\varepsilon}}{\sigma} \right)^{|B'_1| + \ldots + |B'_k|} \le \left(\frac{3nD}{\sigma} \right)^{dk} \cdot \left(\frac{\sqrt{\varepsilon}}{\sigma} \right)^{\ell/2} \le \left(\frac{3nD}{\sigma^2} \right)^{dk} \cdot \varepsilon^{\ell/4} \end{aligned}$$

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

4.2 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] (see also (1)).

Lemma 4.3. *If in an iteration of the k-means method 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 4.4. *Let* $c_1, c_2, x \in \mathbb{R}^d$*, and assume that x has a Euclidean distance of* ε *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 ε 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 α is the angle between $x - x'$ and $\nabla\Phi(c_1) - \nabla\Phi(c_2)$ and the inequality follows from Inequality (3). 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. □

We say that $\mathcal{X}$ is *ε-separated* if, for every hyperplane H, there are at most $2d$ points in $\mathcal{X}$ that are within a distance of at most ε of H. The following lemma, due to Arthur and Vassilvitskii [4, Proposition 5.6], shows that $\mathcal{X}$ is likely to be ε-separated. As its proof is only based on an upper bound on the probability that a point has a distance of at most ε 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 4.5. *For $\varepsilon \geq 0$, the point set $\mathcal{X}$ is not ε-separated with a probability of at most*

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

Proof. Arthur and Vassilvitskii's proof is based on the following geometric lemma.

Lemma 4.6. *Let $\mathcal{P}$ be a set of at least d points in $\mathbb{R}^d$, and let H be an arbitrary hyperplane. Then there exists a hyperplane H' passing through d points of $\mathcal{P}$ that satisfies,*

$$\max_{p\in\mathcal{P}} \operatorname{dist}(p, H') \leq 2d \cdot \max_{p\in\mathcal{P}} \operatorname{dist}(p, H).$$

If $\mathcal{X}$ is not ε-separated, then, by the previous lemma, there exists a hyperplane that passes through d points of $\mathcal{X}$ and to which d other points from $\mathcal{X}$ have a distance of at most $2d\varepsilon$.

Let $\mathcal{P}_1 \subseteq \mathcal{X}$ and $\mathcal{P}_2 \subseteq \mathcal{X}$ be two disjoint sets with $|\mathcal{P}_1| = |\mathcal{P}_2| = d$. The probability that for fixed sets all points from $\mathcal{P}_2$ have a distance of at most $2d\varepsilon$ from the hyperplane that passes through all points from $\mathcal{P}_1$ is bounded from above by

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

according to the first property of our perturbation model. Now a union bound over all sets $\mathcal{P}_1$ and $\mathcal{P}_2$ yields the lemma. □

4.3 An Upper Bound

Lemma 4.1 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 5.2. First, we bound the number of iterations in terms of the distance Δ of the closest simultaneous cluster centers that occur during the run of k-means.

Lemma 4.7. *With a probability of at least $1-4W^{-1}$, every sequence of $k^{kd/2}+1$ consecutive iterations of the k-means method (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 Δ denotes the smallest distance of any two simultaneous cluster centers that occur during the sequence and ε is defined as in Lemma 4.1 for $a = 4$.

Proof. Consider the configuration directly before the sequence of steps is performed. Due to Lemma 4.1, the probability that more than $kd/4$ points are within distance ε 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 $\mathcal{I}$. To see the latter recall that $\mathcal{I}$ is chosen such that the probability that any perturbed point lies outside of $\mathcal{I}$ is at most n^{-13}. Hence, the probability that there exists a set $\mathcal{P} \subseteq \mathcal{X}$ with $|\mathcal{P}| = kd/4$ such that all points from $\mathcal{P}$ lie outside of $\mathcal{I}$ can be bounded from above by

$$n^{kd/4} \cdot (n^{-13})^{kd/4} = n^{-3kd} \leq W^{-1}.$$

by a union bound over all choices for $\mathcal{P}$. 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 ε 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_Φ already in the beginning of the sequence, then x will change its assignment in the first step. According to Lemma 4.4, 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_Φ. Then, according to Lemma 4.4,

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

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

$$d_\Phi(x, c_2) - d_\Phi(x, c_2') + d_\Phi(x, c_1') - d_\Phi(x, c_1) \geq 2\xi\varepsilon\Delta\,. \tag{6}$$

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 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:

$$d_\Phi(c_i', c_i) = d_\Phi(x, c_i) - d_\Phi(x, c_i') - \langle x - c_i', \nabla\Phi(c_i') - \nabla\Phi(c_i)\rangle\,. \tag{7}$$

In the following calculation we will use that for any two points $x, y \in \mathcal{D}$ the Euclidean distance $\|x - y\|$ is bounded from above by $3D\sqrt{d}$. Together with Equations (6) and (7),

$c_1, c_1', c_2, c_2' \in \mathcal{I}'$ implies

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

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

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

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 δ with respect to the Euclidean distance, the potential drops by at least $\xi\delta^2$ (see Lemma 4.3). 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. □

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

Lemma 4.8. *For $\delta \geq 0$, we have*

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

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 δ 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 δ from H.

From Markov's inequality, we can conclude that 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 2δ from H. Hence, there can only exist two centers at a distance of at most δ 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 2δ.

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 δ.

The probability that one of these events occurs can be bounded from above using a union bound and Lemma 4.5:

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

The following lemma is the crucial ingredient of the proof of Theorem 5.2. It basically says that the potential decreases significantly every couple of steps.

Lemma 4.9. *Let $a \in [k]$ be arbitrary, and let $\varepsilon = \left(\frac{\sigma^2}{3Dn^{11}}\right)^{4a}$ be as in Lemma 4.1. Then the expected number of steps until the potential drops by at least* 1 *is bounded from above by*

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

for a sufficiently large absolute constant γ.

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 4.7. 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:

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

with

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

According to Lemma 4.8,

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

For $d \geq 4$, this yields

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

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

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

which can, for a sufficiently large absolute constant γ, be bounded from above by

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

This concludes the technical preparation of Theorem 5.2.

4.4 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 [17], where length refers to the number of iterations of the epoch.[1] 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 (η, c)-coarseness used by Arthur et al. [3]: $\mathcal{X}$ is (η, 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 $\|\mathrm{cm}(\mathcal{C}_1) - \mathrm{cm}(\mathcal{C}_2)\| > \eta$ or $\|\mathrm{cm}(\mathcal{C}_2) - \mathrm{cm}(\mathcal{C}_3)\| > \eta$. Since the length of any epoch is at most three, after at most four iterations, one cluster assumes a third position. Assume that $\mathcal{X}$ is (η, 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 η during one of these iterations. Combining this with Lemma 4.3 and (2), we get a potential drop of at least $\xi\eta^2$.

Lemma 4.10. *Assume that $\mathcal{X}$ is (η, 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 (η, c)-coarse. This bound needs only the probability for the event that a single point falls into a hyperball of a specific radius ε [3, Lemma 4.7]. 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.

[1] We claimed earlier [17] that any epoch has a length of at most two. This has been corrected to three in the full version of the paper [3].

Lemma 4.11. *For $\eta \geq 0$, the probability that $\mathcal{X}$ is not (η, 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 ε-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 4.12. *Assume that $\mathcal{X}$ is ε-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 ε-separated, one of these points, say, x, must be at a distance of at least ε from the bisector of c_i and c_j. According to Lemma 4.4, $d_\Phi(x, c_i) - d_\Phi(x, c_j) \geq \varepsilon 2\xi \cdot \|c_j - c_i\|$.

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 $\|c_i - c_j\| \geq \operatorname{dist}(c_i, H') + \operatorname{dist}(c_j, H')$.

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 ε from H' because $\mathcal{X}$ is ε-separated. Let us assume that this point belongs to $\mathcal{C}_i$. Then $\operatorname{dist}(c_i, H') \geq \varepsilon/n$ as $\mathcal{C}_i$ contains at most n points. Thus, $\|c_i - c_j\| \geq \varepsilon/n$. □

We consider (η, c)-coarseness for $c = 2d\sqrt{k}$. For a set of points that is $(\eta, 2d\sqrt{k})$-coarse and ε-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 4.10 or Lemma 4.12 applies. This yields the main lemma of this section. Together with Lemma 4.14 of Section 4.5, it will lead to Theorem 5.1.

Lemma 4.13. *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*

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

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

Proof. Let Δ be the smallest improvement made by any sequence of four consecutive iterations with at most $\sqrt{k}$ active clusters. The random variable Δ 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 $x \le 1$,

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

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:

$$\begin{aligned}\mathrm{E}\left[T\right] = \sum_{t=1}^{W} \Pr\big[T \ge t\big] \;&\le\; \sum_{t=1}^{W} \Pr\left[\Delta \le \frac{1}{t}\right] \;\le\; 1 + \int_{t=1}^{W} \Pr\left[\Delta \le \frac{1}{t}\right] \mathrm{d}t \\ &\le 1 + \int_1^W \min\left\{1, \left(\frac{2n^{14\sqrt{k}}}{\sigma}\right)^d \left(\frac{1}{t\xi}\right)^{\frac{d}{2}}\right\} \mathrm{d}t \\ &\quad + \int_1^W \min\left\{1, \left(\frac{2n^{14\sqrt{k}}}{\sigma}\right)^d \left(\frac{1}{2t\xi}\right)^{\frac{d}{4}}\right\} \mathrm{d}t \\ &\le \frac{1}{\xi} \cdot \mathrm{poly}\left(n^{\sqrt{k}}, \frac{1}{\sigma}\right) \cdot \log(W) \;\le\; \frac{1}{\xi} \cdot \mathrm{poly}\left(n^{\sqrt{k}}, \frac{1}{\sigma}\right) ,\end{aligned}$$

where the second-to-last inequality uses the assumption $d \ge 4$, and the last inequality uses $\log W \le 3kd \log n$. □

4.5 Iterations with at Least $\sqrt{k}$ Active Clusters

In this section, we consider steps of the k-means method 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 4.14. *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*

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

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

Proof. We consider one step of the k-means method with at least $\sqrt{k}$ active clusters. Let ε be defined as in Lemma 4.1 for $a = 1$. We distinguish two cases: Either one point that is reassigned during the considered iteration has a distance of at least ε from the bisector that it crosses, or all points are at a distance of at most ε from their respective bisectors. In the former case, we immediately get a potential drop of at least $2\xi\varepsilon\Delta$, where Δ denotes the minimal distance of two cluster centers. In the latter case, Lemma 4.1 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 4.1, we assume that the positions of the points in $A_1 \cup \ldots \cup A_k$ are fixed by an adversary, and we apply a union bound on the different partitions $A_1, \ldots, A_k$ 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 $\mathcal{C}_i$ before the reassignment, and let c'_i be the position after the reassignment. Then

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

where $\mathrm{cm}(\cdot)$ 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

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

Due to the union bound, $\mathrm{cm}(A_i)$ 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 $\|c_i - c'_i\| \le \beta$. 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 $(2n\beta/\sigma)^{d\ell/2}$.

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

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

For

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

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

Now we also take into account the failure probability of $2W^{-1}$ from Lemma 4.1. 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

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

since $d \le n$ and D is polynomially bounded in σ 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 $\Gamma \le 1/t$. Hence,

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

where the integral is upper bounded as in the proof of Lemma 4.9. □

5 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 4.13 and 4.14.

Theorem 5.1. *Let d_Φ 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 4.9. Note that the degree of the polynomial is not only independent of d, but also independent of k.

Theorem 5.2. *Let d_Φ 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, $1/\xi$, ξ', 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 5.3. *In the setting of Theorem 5.2, if P, Q', and ξ' 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 $k, d \in O\big(\sqrt{\log n / \log\log n}\big)$.

5.1 Mahalanobis Distances

For Mahalanobis distances, we use the same perturbation model that has been used for squared Euclidean distances [4, 17]: The adversary chooses n points in $[0,1]^d$. Then the d coordinates are perturbed by independent Gaussian perturbations of standard deviation σ. 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 [12, Section 2.3]. The 2-norm of a symmetric matrix equals its largest eigenvalue. Thus, $Q' = 2\lambda_{\max}$. As ξ' 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 ε^* in Lemma 4.9 cannot occur and we can simply remove ε^* from the bound in Theorem 5.2.

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 5.4. *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 $k, d \in O\big(\sqrt{\log n / \log\log n}\big)$ and the largest eigenvalue of A is bounded by a polynomial, then, as in Corollary 5.3, we obtain smoothed polynomial running-time.

Remark 5.5. *The most natural perturbation model for Mahalanobis distances d_{m_A} would be Gaussian perturbations with covariance matrix A^{-1}. Analogously to Lemma 6.2 below, this is almost the same as using squared Euclidean distances and the identity matrix as covariance matrix (thus, using independent Gaussians with the same standard deviation).*

This (almost) makes the polynomial bound [3] applicable. The only issue that remains is the initial potential: We have assumed that all eigenvalues of A are between 1 and $\lambda_{\max}$. If $\lambda_{\max}$ is very large, then it is not possible to bound the potential after the first iteration by a polynomial (the initial points are from $[0,1]^d$, and $d_{m_A}(x, y)$ for $x, y \in [0,1]^d$ can be about $\lambda_{\max}$).

However, if we assume that $\lambda_{\max}$ is bounded by a polynomial in n, then we obtain smoothed polynomial running-time for k-means with Bregman divergences.

5.2 Kullback-Leibler Divergence

Before stating our perturbation model for KLD precisely, let us motivate our choice: A point represents a probability distribution on a finite set $\{1, 2, \ldots, d+1\}$. For instance, assume that we want to classify web pages based on a list $w_1, \ldots, w_{d+1}$ of words (the so-called *bag-of-words model* [8]). 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 copies of each word to the web page.

Now let us describe our perturbation model precisely. For a point $x \in \mathbb{D}$, 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 $y_1, \ldots, y_{d+1}$ 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 [10], whose density is $\frac{1}{\theta} \cdot \exp\bigl(-\frac{x}{\theta}\bigr)$ for a positive parameter θ. It has mean θ, variance θ^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 require explanation. First, $\theta = \sigma$ would be the natural choice. However, to meet the requirements for perturbation model, and to use our framework introduced in Section 4, we need to choose θ slightly larger than σ. Let us emphasize that σ and $\theta = 8d\sigma^{d/(d+1)}$ differ only by a polynomial factor. Second, due to the requirements for the perturbation model, we also need $\sigma \leq \frac{1}{14d^{4/3}}$. But this does not harm the result either: On the one hand, it includes the particularly interesting small values of σ. 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 2.3.

Lemma 5.6. *Let $x \in \mathbb{D}$ and let $z \in \mathbb{D}$ be the point obtained from x by perturbation. Let $H \subseteq \mathbb{R}^d$ be any hyperplane. Then*

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

and the density of the random variable z is bounded from above by σ^{-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 $(d+1)\exp\bigl(-\frac{Z}{d\theta}\bigr)$. We choose Z such that this probability is at most $\frac{\varepsilon}{2\sigma}$, which yields $Z = d\theta \log\bigl(\frac{2(d+1)\sigma}{\varepsilon}\bigr) \leq d\theta\sqrt{\frac{3d\sigma}{\varepsilon}}$.

Given S and $y_1, \ldots, y_{d-1}, y_{d+1}$, we have $\operatorname{dist}(z, H) \leq \varepsilon$ 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 $\operatorname{dist}(z, H) \leq \varepsilon$ 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

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

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 $y = (y_1, \ldots, y_{d+1})$, we consider the vector $z' = (z_1, \ldots, z_d, S)$ and denote by h its density. The transformation Φ with

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

maps y to z' and its inverse is

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

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

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

□

What remains to be done is to choose ε^* and to analyze the parameters ξ, ξ', and Q' 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 ε-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 ε-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 ε^*-close to the boundary of the domain is at most n^{-13}.

Let us now analyze ξ. 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 ξ'. 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 we have $\xi' \leq \text{poly}(n, \sigma^{-1})$.

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

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

and

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

5.3 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 $\operatorname{poly}(n)$ 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 ε-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 $z_2, \ldots, z_d$. Then for $\operatorname{dist}(z, H) \leq \varepsilon$, 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 ε^*, ξ', and Q' can be analyzed similarly as for KLD in the previous section. Also ξ can be analyzed similarly, we only have to use the upper bound of $\operatorname{poly}(n)$ 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 5.8. *The smoothed running-time of k-means using* GID *is bounded from above by*

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

and

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

5.4 Itakura-Saito Divergence

For the Itakura-Saito divergence, we use again exponentially distributed perturbations. This is the natural choice for this distance measure (see Section 2.3.1). Rescaling as we did for KLD is not necessary. We can choose $D = \operatorname{poly}(n)$ to make sure that $\mathcal{X} \subseteq [0, D]^d = \mathcal{D}$ with a probability of at least $1 - W^{-1}$.

The analysis of $\varepsilon^* \geq 1/\operatorname{poly}(n)$ is similar to its counterpart for KLD. Let us first analyze ξ and ξ'. By definition of a Bregman divergence, d_{ISD} is the tail of the first-order Taylor expansion of $\mathrm{ISD}(x)$ 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

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

where $\nabla^2 \operatorname{ISD}(\xi)$ is the Hesse matrix of ISD at ξ. (This exists for all possible ξ.) The Hesse matrix $\nabla^2 \operatorname{ISD}(\xi)$ is a diagonal matrix with diagonal entries $1/\xi_1^2, \ldots, 1/\xi_d^2$. 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, $1/\xi \leq \operatorname{poly}(n)$, which is fine. On the other hand, $\frac{1}{\min(y_1^2, x_i^2)} \leq \frac{1}{\varepsilon^*} \leq \operatorname{poly}(n)$, which shows $\xi' \leq \operatorname{poly}(n)$.

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

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

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 $Q' \leq \frac{4n^2 d}{\varepsilon^{*2}} \leq \operatorname{poly}(n)$.

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

Theorem 5.9. *The smoothed running-time of k-means using* ISD *is bounded from above by*

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

and

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

6 Lower Bound

In this section, we transfer the exponential lower bound proved by Vattani [24] to almost arbitrary Bregman divergences. Our starting point is his lower bound construction.

Theorem 6.1 (Vattani [24]). *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 6.1 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 Bregman divergence behaves locally like some Mahalanobis distance, if it is three times differentiable. 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_Φ be a Bregman divergence, let $\mathcal{X}$ be a point set, and let $c_1, \ldots, c_k$ be initial centers. The instance $\mathcal{X}$, $c_1, \ldots, c_k$ is called *d_Φ-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_Φ-stable* if there exists a constant $\nu > 0$ such that it is d_Φ-stable with slack ν.

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

6.1 Lower Bound for Mahalanobis Distances

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 $W_\Phi^{k,d}(n)$ be the maximum number of iterations of k-means on any d_Φ-stable instance of n points in $\mathbb{R}^d$ using d_Φ as the distance measure.

Lemma 6.2. *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 $c_1, \ldots, c_k \in \mathbb{R}^d$ 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 $A = M^TM$ (Cholesky factorization [12, Theorem 4.2.5]). Let $B = M^{-1}$ (since A is positive definite, A and thus also M have full rank), and let $y = x - x'$ for any $x, x' \in \mathbb{R}^d$. Then

$$d_{m_I}(x, x') = \|x - x'\|^2 = y^Ty = y^T(B^TM^T)(MB)y = (By)^TA(By) = d_{m_A}(Bx, Bx').$$

Now let $\mathcal{X}' = \{Bx \mid x \in \mathcal{X}\}$ and $c'_i = Bc_i$ for $i \in [k]$. Then k-means behaves w.r.t. squared Euclidean distances on $\mathcal{X}$ initialized with centers $c_1, \ldots, c_k$ exactly in the same way as w.r.t. Mahalanobis distance d_{m_A} on $\mathcal{X}'$ initialized with centers $c'_1, \ldots, c'_k$. 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}. □

6.2 Lower Bound for Bregman Divergences

Now we transfer worst-case instances for Mahalanobis distances to general Bregman divergences. For this, we use the observation that any Bregman divergence d_Φ behaves locally at some point z_0 like the Mahalanobis distance d_{m_H}, where H is the Hessian matrix of Φ at z_0. For this, we need the assumption that the third-order derivatives of Φ exist. 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 6.3. *Let $\Phi : \mathbb{D} \to \mathbb{R}$ be a strictly convex function with $\mathbb{D} \subseteq \mathbb{R}^d$ and the following properties: There exist a $z_0 \in \mathbb{D}$ and a $\zeta > 0$ such that*

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

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

Proof. First we show that d_Φ behaves locally around z_0 almost like the Mahalanobis distance d_{m_H}, where H denotes the Hessian matrix of Φ 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 Φ at z_0. All third-order derivatives of Φ 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 [15, §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

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

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

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

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

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

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 6.2, there exists a set $\mathcal{X} \subseteq \mathbb{R}^d$ of n points and centers $c_1, \ldots, c_k \in \mathbb{R}^d$ 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 $\tilde{c}_1, \ldots, \tilde{c}_k$ on which k-means using d_Φ also needs $W_{m_I}^{k,d}(n)$ iterations. We can assume w.l.o.g. that $\mathcal{X} \subseteq [-1,1]^d$ and $c_1, \ldots, c_k \in [-1,1]^d$. 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 $\tilde{c}_1, \ldots, \tilde{c}_k \in Z$ because $\varepsilon \leq \zeta$. The k-means method behaves on $\tilde{\mathcal{X}}$ w.r.t. d_Φ like on $\mathcal{X}$ w.r.t. d_{m_H} if

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

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

6.3 Applying the Lower Bound

Vattani's lower bound construction [24] is d_{m_I}-stable. Combining this construction with Lemma 6.2 and Lemma 6.3, we obtain the main result of this section.

Theorem 6.4. *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 [24] and Lemma 6.2.

The domain of the Kullback-Leibler divergence (KLD) is $\mathbb{D} = \{z \in \mathbb{R}^d \mid z \geq 0, \sum_{i=1}^d z_i \leq 1\}$. We choose $z_0 = (\frac{1}{d+1}, \ldots, \frac{1}{d+1}) \in \mathbb{D}$. Then $Z = \{z \in Y \mid \|z - z_0\|_\infty \leq \zeta\} \subseteq \mathbb{D}$ for $\zeta = 1/(d+1)^2$. The convex function corresponding to KLD is $\mathrm{KLD}(x) = \sum_{j=1}^{d+1} x_j \log x_j$, where $x_{d+1} := 1 - \sum_{j=1}^d x_j$. Simple calculus shows that $\frac{\partial \, \mathrm{KLD}(x)}{\partial x_i} = \log x_i - \log x_{d+1}$, $\frac{\partial^2 \, \mathrm{KLD}(x)}{\partial x_i^2} = \frac{1}{x_i} + \frac{1}{x_{d+1}}$, and $\frac{\partial^2 \, \mathrm{KLD}(x)}{\partial x_i \partial x_j} = \frac{1}{x_{d+1}}$, 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

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

if not $i = j = \ell$. For our choice of ζ 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., $z_0 = (1, \ldots, 1)$.

For the Itakura-Saito divergence, we can again choose $z_0 = (1, \ldots, 1)$. We have $\frac{\partial \operatorname{ISD}(x)}{\partial x_i} = \frac{-1}{x_i}$ and $\frac{\partial^2 \operatorname{ISD}(x)}{\partial x_i^2} = \frac{1}{x_i^2}$ and, for $i \neq j$, $\frac{\partial^2 \operatorname{ISD}(x)}{\partial x_i \partial x_j} = 0$. 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 $\frac{\partial^3 \operatorname{ISD}(x)}{\partial x_i^3} = \frac{-2}{x^3}$ for $i \in \{1, \ldots, d\}$. For $\zeta = 1/2$, the absolute values of all third-order derivatives around z_0 are bounded by 16, which completes the proof. □

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 6.2 and 6.3) 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.

7 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 $k^{kd} \operatorname{poly}(n, 1/\sigma)$, 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] have 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.

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.

Acknowledgement

We thank Georg Still and the anonymous referees for their help to improve the presentation.

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 nonmetric distance measures. *ACM Transactions on Algorithms*, 6(4), 2010.

[3] David Arthur, Bodo Manthey, and Heiko Röglin. Smoothed analysis of the k-means method. *Journal of the ACM*, 58(5), 2011.

[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] Ole Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* John Wiley & Sons, 1978.

[7] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, USA, 2002.

[8] 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.

[9] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification.* John Wiley & Sons, 2000.

[10] William Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley & Sons, 1971.

[11] Jürgen Forster and Manfred K. Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.

[12] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 3rd edition, 1996.

[13] 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.

[14] 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.

[15] Serge Lang. *Real Analysis.* Addison-Wesley, 1969.

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

[17] 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.

[18] Bodo Manthey and Heiko Röglin. Worst-case and smoothed analysis of k-means clustering with Bregman divergences. In *Proc. of the 20th Int. Symp. on Algorithms and Computation (ISAAC)*, volume 5878 of *Lecture Notes in Computer Science*, pages 1024–1033. Springer, 2009.

[19] Bodo Manthey and Heiko Röglin. Smoothed analysis: Analysis of algorithms beyond worst case. *it – Information Technology*, 53(6):280–286, 2011.

[20] Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. Bregman Voronoi diagrams. *Discrete & Computational Geometry*, 44(2):281–307, 2010.

[21] R. Tyrrell Rockafeller. *Convex Analysis.* Princeton University Press, 1970.

[22] 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.

[23] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis: An attempt to explain the behavior of algorithms in practice. *Communications of the ACM*, 52(10):76–84, 2009.

[24] Andrea Vattani. k-means requires exponentially many iterations even in the plane. *Discrete and Computational Geometry*, 45(4):596–616, 2011.