

Worst-Case and Smoothed Analysis of k -Means Clustering with Bregman Divergences

Bodo Manthey¹ and Heiko Röglin^{2*}

¹ Department of Applied Mathematics, University of Twente
b.manthey@utwente.nl

² Department of Quantitative Economics, Maastricht University
heiko@roeglin.org

Abstract. The k -means algorithm is the method of choice for clustering large-scale data sets and it performs exceedingly well in practice. Most of the theoretical work is restricted to the case that squared Euclidean distances are used as similarity measure. In many applications, however, data is to be clustered with respect to other measures like, e.g., relative entropy, which is commonly used to cluster web pages. In this paper, we analyze the running-time of the k -means method for Bregman divergences, a very general class of similarity measures including squared Euclidean distances and relative entropy. We show that the exponential lower bound known for the Euclidean case carries over to almost every Bregman divergence. To narrow the gap between theory and practice, we also study k -means in the semi-random input model of smoothed analysis. For the case that n data points in \mathbb{R}^d are perturbed by noise with standard deviation σ , we show that for almost arbitrary Bregman divergences the expected running-time is bounded by $\text{poly}(n^{\sqrt{k}}, 1/\sigma)$ and $k^{kd} \cdot \text{poly}(n, 1/\sigma)$.

1 Introduction

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

* Supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD).

squared Euclidean distances and KLD are special cases of *Bregman divergences*, a very general class that contains most practically important distance measures.

Even though a lot of theoretical research has been conducted on clustering algorithms, the by far most successful algorithm in industrial and scientific applications is the seemingly ad hoc *k-means method* [6], a local search algorithm due to Lloyd [12]: Start with an arbitrary set of k centers and repeat the following two steps until the process stabilizes: 1) Assign every data point to its closest center. 2) Readjust the positions of the centers such that they are optimal for the current assignment. The k -means method works very well in practice. One of its distinguished features is its speed: It has been observed that the number of iterations it needs to find a local optimum is much smaller than the number of objects to be clustered [8, Section 10.4.3]. This is in stark contrast to its worst-case running-time: The only upper bound is $n^{O(kd)}$ [11], which is based on the observation that no clustering appears twice in a run of k -means. On the other hand, Vattani [15] showed that k -means can run for $2^{\Omega(n)}$ iterations in the worst case. This lower bound 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 [14] 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 [?] showed that for squared Euclidean distances 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 improved this bound to $\text{poly}(n^{\sqrt{k}}, 1/\sigma)$ and we additionally obtained a bound of $k^{kd} \cdot \text{poly}(n, 1/\sigma)$ [13]. Recently, Arthur et al. [3] showed that the smoothed running-time of k -means is polynomial in n and $1/\sigma$.

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 $\text{poly}(n^{\sqrt{k}}, 1/\sigma)$ and $k^{kd} \cdot \text{poly}(n, 1/\sigma)$.

1.1 k -Means Method

An instance for k -means clustering is a set $\mathcal{X} \subseteq \mathbb{R}^d$ consisting of n points. The aim is to find a clustering $\mathcal{C}_1, \dots, \mathcal{C}_k$ of \mathcal{X} , i.e., a partition of \mathcal{X} , as well as cluster centers $c_1, \dots, c_k \in \mathbb{R}^d$ such that the potential $\sum_{i=1}^k \sum_{x \in \mathcal{C}_i} d_\Phi(x, c_i)$ is minimized, where d_Φ denotes some distance measure on \mathbb{R}^d . Given the cluster centers, every data point should be assigned to the cluster whose center is closest to it. The other way round, given the clusters, the centers c_1, \dots, c_k should be chosen so as to minimize the potential. In the next section, we will see that for Bregman divergences this is the case if the centers are chosen as the centers of mass, i.e., $c_i = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} x$. The k -means method for Bregman divergences proceeds now as follows (observe that since the potential decreases in every step, no clustering occurs twice, and the algorithm eventually terminates):

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

1.2 Bregman Divergences

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

Definition 1. Let $X \subseteq \mathbb{R}^d$, and let $\Phi : X \rightarrow \mathbb{R}$ be a strictly convex function such that Φ is differentiable on the relative interior $\text{ri}(X)$ of X . The Bregman divergence $d_\Phi : X \times \text{ri}(X) \rightarrow [0, \infty)$ is defined as

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

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

For a finite set of points $C \subseteq X$, we denote the center of mass of C by $\text{cm}(C) = \frac{1}{|C|} \sum_{x \in C} x$. For Bregman divergences the potential can be expressed in terms of the center of mass in the following way [5, Proposition 1]: For every c ,

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

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

Another important property of Bregman divergences is that the bisector of two centers c and c' , i.e., the set $\{x \in X \mid d_{\Phi}(x, c) = d_{\Phi}(x, c')\}$, is a hyperplane, which follows immediately from the definition of d_{Φ} . 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, 11]. The proof of this bound relies only on the fact that the bisectors are hyperplanes. Hence, also for general Bregman divergences, the worst-case number of iterations cannot exceed W .

In the following, we present some prominent Bregman divergences.

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

Kullback-Leibler Divergence and Generalized I-Divergence. The *Kullback-Leibler divergence* (KLD, relative entropy) is a very popular Bregman divergence. Here, $X = \{x \in \mathbb{R}^d \mid x \geq 0, \sum_{i=1}^d x_i \leq 1\}$ and an element $x = (x_1, \dots, x_d) \in X$ represents a probability distribution on a discrete set with $d + 1$ elements (where (x_1, \dots, x_{d+1}) with $x_{d+1} = 1 - \sum_{i=1}^d x_i$ is the vector of probabilities). For $\text{KLD}(x) = \sum_{i=1}^{d+1} x_i \log(x_i)$, we obtain $d_{\text{KLD}}(x, c) = \sum_{i=1}^{d+1} x_i \log(\frac{x_i}{c_i})$, 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 [7].

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

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, 10]. We have again $X = \{x \in \mathbb{R}^d \mid x \geq 0\}$, and the potential function is given by the Burg entropy $\text{ISD}(x) = -\sum_{i=1}^d \log(x_i)$. From this, we get the Bregman divergence $d_{\text{ISD}}(x, c) = \sum_{i=1}^d \frac{x_i}{c_i} - \log(\frac{x_i}{c_i}) - 1$.

1.3 Perturbation Models for Bregman Divergences

If the Bregman divergence is defined on the whole space \mathbb{R}^d , i.e., if $X = \mathbb{R}^d$, then it is often natural to assume that the points are perturbed by adding Gaussian noise to them. More precisely, we can assume that an adversary is allowed to place initially n points in $[0, 1]^d$, and that each of these points is perturbed by adding a Gaussian with standard deviation σ to each of its coordinates.

On the other hand, if X is a proper subset of \mathbb{R}^d , as it is the case for KLD or GID, then such a perturbation model cannot be applied as it might yield points outside the feasible region X . For this reason, we decided to consider very general perturbation models that need to satisfy only a couple of properties, which we will summarize in the following. In Section 2, we present concrete perturbation models with these properties for some special Bregman divergences.

We assume that the perturbation model is parameterized by some $\sigma \in (0, 1]$ that measures the amount of randomness in the sense that the smaller σ is chosen, the weaker is the perturbation. If every point is perturbed by Gaussian noise, then σ can be chosen as the standard deviation. We assume that the following properties are satisfied for $\sigma \in (0, 1]$:

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

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

Additionally, we need the property that perturbed points cannot be too far away from their initial positions in $X \cap [0, 1]^d$. For this, let D be chosen such that with probability at least $1 - W^{-1}$ every point from the perturbed point set \mathcal{X} is contained in the hypercube $\mathcal{D} = [-D, 1 + D]^d$, where $W \leq n^{3kd}$ denotes the worst number of steps. 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 $\sigma \leq 1$, D can be chosen polynomially in n .

1.4 Parameterization

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

For $\varepsilon \geq 0$, let $\mathcal{I}(\varepsilon)$ be the interior of $X \cap \mathcal{D}$ that has a distance of at least ε to the boundary: $\mathcal{I}(\varepsilon) = \{x \in X \cap \mathcal{D} \mid \text{dist}(x, \partial(X \cap \mathcal{D})) \geq \varepsilon\}$.

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

$$\forall x, y \in X \cap \mathcal{D}: d_\Phi(x, y) \geq \xi \cdot \|x - y\|^2 \quad \text{and} \quad \forall x, y \in \mathcal{I}': d_\Phi(x, y) \leq \xi' \cdot \|x - y\|^2.$$

Observe that for the definition of ξ' , only the interior of $X \cap \mathcal{D}$ is relevant. This is important as otherwise ξ' would be unbounded for many Bregman divergences. The ratio ξ'/ξ is closely related to the μ in the notion of μ -similarity introduced by Ackermann et al. [2]. However, Bregman divergences like KLD, GID, or ISD are not μ -similar for any μ 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. We emphasize that no such restrictions are necessary for our smoothed analysis. There may be points close to the boundary of the domain, but we can take special care of those points. This technical challenge is the reason for the definition of \mathcal{I} and \mathcal{I}' above.

We also need the following lower bound on the “second derivative” of Φ , which follows by a simple calculation from the previous definition: $\frac{\|\nabla\Phi(x) - \nabla\Phi(y)\|}{\|x - y\|} \geq 2\xi$ for all $x, y \in X \cap \mathcal{D}$ with $x \neq y$. Similarly, we need an upper bound (only for the interior): $Q' := \sup_{x, y \in \mathcal{I}', x \neq y} \frac{\|\nabla\Phi(x) - \nabla\Phi(y)\|}{\|x - y\|}$.

1.5 Our Results

In the following, we assume $d \leq n$, $k \leq n$, and $d \geq 4$, which are no severe restrictions from a practical point of view. Let P be the maximal potential after the first iteration of k -means, provided that all points of \mathcal{X} lie in \mathcal{D} .

Theorem 2. *Let d_Φ be a Bregman divergence. Then the smoothed running-time of k -means is bounded from above by $\frac{P}{\xi} \cdot \text{poly}(n^{\sqrt{k}}, \frac{1}{\sigma})$ and by $P \cdot k^{kd} \cdot \frac{Q'^2 \xi'^3}{4\xi^5 \varepsilon^{*2}} \cdot \text{poly}(n, \frac{1}{\sigma})$, where the polynomials are independent of d , k , and the parameters.*

The second bound in the theorem yields a polynomial smoothed running-time if $k, d \in O(\sqrt{\log n / \log \log n})$. Indeed, k and d are usually much smaller than n in practice.

On the negative side, in Section 3, we transfer the lower bound of $2^{\Omega(n)}$ for squared Euclidean distances to all good-natured Bregman divergences, where “good-natured” means that all third order derivatives exist and are bounded in a small region, which includes Mahalanobis distances, KLD, GID, and ISD.

1.6 Technical Contribution

In an earlier analysis [13], we presented two different approaches for analyzing the smoothed running-time, leading to upper bounds of $k^{kd} \cdot \text{poly}(n, \sigma^{-1})$ and $\text{poly}(n^{\sqrt{k}}, \sigma^{-1})$ for squared Euclidean distances. Both of these approaches are based on a novel lemma about perturbed point sets, stating that, given any Voronoi partition of the point set, it is unlikely that many points are close to the bisectors [13, Lemma 2.1]. Clearly, the structure of the smoothed analysis presented in this paper is similar to the earlier one [13]. However, we had to tackle several severe problems when transferring the results from squared Euclidean distances to general Bregman divergences. First of all, the proof of the aforementioned lemma about perturbed point sets cannot be generalized directly to Bregman divergences. In the course of finding a generalization, we found a shorter and simpler proof of the lemma. Given this result, the bound of $k^{kd} \cdot \text{poly}(n, \sigma^{-1})$ follows roughly in the same way as in the Euclidean case, but some additional technical problems have to be addressed. Let us describe the main problem by way of example: For KLD, the parameters ξ' and Q' can become arbitrarily large for points close to the boundary of X . Even after the perturbation, some of the points might still be too close to the boundary to obtain reasonable upper bounds for ξ' and Q' . Essentially, we show that the kd points that are closest to the boundary can be handled separately and that all other points are sufficiently far away from the boundary (i.e., they lie in \mathcal{I}) to bound ξ' and Q' in a reasonable way.

An obvious question is whether the smoothed polynomial bound [3] carries over to Bregman divergences. The problem with adapting the proof of this bound is that it exploits specific properties of Gaussian perturbations. It uses, in particular, the property that the projection of a Gaussian random vector onto a lower-dimensional subspace is still a Gaussian with the same standard deviation. It would be very interesting to see if it is possible to relax some of these requirements or if it is possible to design more general perturbation models that still meet the requirements needed for the smoothed polynomial bound.

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

Due to lack of space, all proofs are deferred to the full version of this paper. In the following section, we will only apply Theorem 3 to four common Bregman divergences.

2 Applying the Smoothed Analysis

2.1 Mahalanobis Distances

For Mahalanobis distances, we use the same perturbation model that has been used for squared Euclidean distances [?,13]: 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 $\text{poly}(n)$.

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 assume that A is scaled such that the smallest eigenvalue, which is positive, equals 1. Let λ_{\max} be the largest eigenvalue of A . A short calculation shows that the parameters can be chosen such that Theorem 2 yields the following bound:

Theorem 3. *The smoothed running-time of k -means using m_A is bounded from above by $\lambda_{\max} \cdot \text{poly}(n^{\sqrt{k}}, \frac{1}{\sigma})$ and $k^{kd} \cdot \lambda_{\max}^6 \cdot \text{poly}(n, \frac{1}{\sigma})$.*

2.2 Kullback-Leibler Divergence

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

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

Let us make this more precise. For a point $x \in X$, we obtain $x' \in \mathbb{R}^{d+1}$ by adding the component $x_{d+1} = 1 - \sum_{i=1}^d x_i$. Then we draw random numbers y_1, \dots, y_{d+1} independently according to some probability distribution to be specified in a moment. Let $S = \sum_{i=1}^{d+1} x_i + y_i = 1 + \sum_{i=1}^{d+1} y_i$. Then we obtain the perturbed point $z \in \mathbb{R}^d$ by setting $z_i = \frac{x_i + y_i}{S}$. By construction, $z \geq 0$ and $\sum_{i=1}^d z_i \leq 1$. We use the exponential distribution [9], whose density is $\frac{1}{\theta} \cdot \exp(-\frac{x}{\theta})$ for a positive parameter θ .

In the full version of this paper, we show that this perturbation model satisfies the requirements of Section 1.3 for $\theta = 8d\sigma^{d/(d+1)}$. Analyzing the parameters ξ , ξ' , Q' , and ε^* as well as the potential P after the first iteration yields the following theorem.

Theorem 4. *The smoothed running-time of k -means using KLD is bounded from above by $\text{poly}(n^{\sqrt{k}}, \frac{1}{\sigma})$. and $k^{kd} \cdot \text{poly}(n, \frac{1}{\sigma})$.*

2.3 Generalized I-Divergence

For generalized I-divergence and Itakura-Saito divergence, we use the same perturbation model, except for rescaling. Since we do not have to rescale, this allows

us to let the adversary choose any density function f bounded by $\frac{1}{2\sqrt{d}\sigma}$ whose tail bounds are sufficiently small: The probability of a number greater than $\text{poly}(n)$ must be bounded by $\frac{1}{ndW}$. Then we perturb a point by adding independent random numbers drawn according to f . For this perturbation model, Theorem 4 carries over to GID and ISD. Details can be found in the full version.

3 Lower Bound

In this section, we transfer the exponential lower bound proved by Vattani [15] to almost arbitrary Bregman divergences.

Theorem 5 (Vattani [15]). *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.*

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

Lemma 6. *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 $n, k, d \in \mathbb{N}$.*

Now we transfer worst-case instances for Mahalanobis distances to instances for arbitrary good-natured Bregman divergences. For this, we use the observation that any good-natured Bregman divergence d_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 . Hence, essentially we only need to scale down the worst-case instance for d_{m_H} and embed it locally into a small space around z_0 .

Lemma 7. *Let $\Phi : X \rightarrow \mathbb{R}$ be a strictly convex function with $X \subseteq \mathbb{R}^d$ and the following properties: There exist a $z_0 \in X$ and a $\zeta > 0$ such that*

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

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

Combining Vattani’s lower bound with Lemma 6 and Lemma 7, we obtain the main result of this section.

Theorem 8. *The worst-case number of iterations of k -means for the following Bregman divergences is at least $\exp(\Omega(n))$ for n points and $d \geq 2$: Mahalanobis distances for any symmetric positive definite matrix A , Kullback-Leibler divergence (KLD), generalized I-divergence (GID), Itakura-Saito divergence (ISD).*

References

1. Marcel R. Ackermann and Johannes Blömer. Coresets and approximate clustering for Bregman divergences. In *Proc. of the 20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 1088–1097, 2009.
2. Marcel R. Ackermann, Johannes Blömer, and Christian Sohler. Clustering for metric and non-metric distance measures. In *Proc. of the 19th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 799–808, 2008.
3. David Arthur, Bodo Manthey, and Heiko Röglin. k -means has polynomial smoothed complexity. In *Proc. of the 50th Ann. IEEE Symp. on Found. of Computer Science (FOCS)*, 2009. To appear.
4. David Arthur and Sergei Vassilvitskii. Worst-case and smoothed analysis of the ICP algorithm, with an application to the k -means method. In *Proc. of the 47th Ann. IEEE Symp. on Found. of Computer Science (FOCS)*, pages 153–164, 2006.
5. Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.
6. Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, USA, 2002.
7. Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.
8. Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, 2000.
9. William Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley & Sons, 1971.
10. 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.
11. 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.
12. Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
13. 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.
14. 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.
15. Andrea Vattani. k -means requires exponentially many iterations even in the plane. In *Proc. of the 25th ACM Symp. on Computational Geometry (SoCG)*, pages 324–332, 2009.