Improved Smoothed Analysis of the k-Means Method[*]

Bodo Manthey[†] Heiko Röglin[‡]

Abstract

The k-means method is a widely used clustering algorithm. One of its distinguished features is its speed in practice. Its worst-case running-time, however, is exponential, leaving a gap between practical and theoretical performance. Arthur and Vassilvitskii [3] aimed at closing this gap, and they proved a bound of $\mathrm{poly}(n^k, \sigma^{-1})$ on the smoothed running-time of the k-means method, where n is the number of data points and σ is the standard deviation of the Gaussian perturbation. This bound, though better than the worst-case bound, is still much larger than the running-time observed in practice.

We improve the smoothed analysis of the k-means method by showing two upper bounds on the expected running-time of k-means. First, we prove that the expected running-time is bounded by a polynomial in $n^{\sqrt{k}}$ and σ^{-1}. Second, we prove an upper bound of $k^{kd} \cdot \mathrm{poly}(n, \sigma^{-1})$, where d is the dimension of the data space. The polynomial is independent of k and d, and we obtain a polynomial bound for the expected running-time for $k, d \in O(\sqrt{\log n / \log \log n})$.

Finally, we show that k-means runs in smoothed polynomial time for one-dimensional instances.

1 Introduction

The k-means method is a very popular algorithm for clustering high-dimensional data. It is a local search algorithm based on ideas by Lloyd [10]: Initiated with k arbitrary cluster centers, it assigns every data point to its nearest center, and then readjusts the centers, reassigns the data points, ... until it stabilizes. (In Section 1.1, we describe the algorithm formally.) The k-means method terminates in a local optimum, which might be far worse than the global optimum. However, in practice it works very well. It is particularly popular because of its simplicity and its speed: "In practice, the number of iterations is much less than the number of samples", as Duda et al. [6, Section 10.4.3] put it. According to Berkhin [5], the k-means method "is by far the most popular clustering tool used in scientific and industrial applications."

The practical performance and popularity of the k-means method is at stark contrast to its performance in theory. The only upper bounds for its running-time are based on the observation that no clustering appears twice in a run of k-means: Obviously, n points can be distributed among k clusters in only k^n ways. Furthermore, the number of Voronoi partitions of n points in $\mathbb{R}^d$ into k classes is bounded by a polynomial in n^{kd} [8], which yields an upper bound of $\mathrm{poly}(n^{kd})$. On the other hand, Arthur and Vassilvitskii [2] showed that k-means can run for $2^{\Omega(\sqrt{n})}$ iterations in the worst case.

To close the gap between good practical and poor theoretical performance of algorithms, Spielman and Teng introduced the notion of smoothed analysis [12]: An adversary specifies an instance, and this instance is then subject to slight random perturbations. The smoothed running-time is the maximum over the adversarial choices of the expected running-time. On the one hand, this rules out pathological, isolated worst-case instances. On the other hand, smoothed analysis, unlike average-case analysis, is not dominated by random instances since the instances are not completely random; random instances are usually not typical instances and have special properties with high probability. Thus, smoothed analysis also circumvents the drawbacks of average-case analysis. For a survey of smoothed analysis, we refer to Spielman and Teng [13].

The goal of this paper is to bound the smoothed running-time of the k-means method. There are basically two reasons why the smoothed running-time of the k-means method is a more realistic measure than its worst-case running-time: First, data obtained from measurements is inherently noisy. So even if the original data were a bad instance for k-means, the data measured is most likely a slight perturbation of it. Second, if the data possesses a meaningful k-clustering, then slightly perturbing the data should preserve this clustering. Thus, smoothed analysis might help to obtain a faster k-means method: We take the data measured, perturb it slightly, and then run k-means on the per-

turbed instance. The bounds for the smoothed running-time carry over to this variant of the k-means method.

1.1 k-Means Method.

An instance of the k-means clustering problem is a point set $\mathcal{X} \subseteq \mathbb{R}^d$ 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

$$\sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \|x - c_i\|^2$$

is minimized. Given the cluster centers, every data point should obviously be assigned to the cluster whose center is closest to it. The name k-means stems from the fact that, given the clusters, the centers $c_1, \ldots, c_k$ should be 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 proceeds now as follows:

1. Select cluster centers $c_1, \ldots, c_k$.

2. Assign every $x \in \mathcal{X}$ to the cluster $\mathcal{C}_i$ whose cluster center c_i is closest to it.

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.

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

1.2 Related Work.

The problem of finding a good clustering can be approximated arbitrarily well: Bădoiu et al. [4], Matoušek [11], and Kumar et al. [9] devised polynomial time approximation schemes with different dependencies on the approximation ratio $(1 + \varepsilon)$ as well as n, k, and d: $O(2^{O(k\varepsilon^{-2} \log k)} \cdot nd)$, $O(n\varepsilon^{-2k^2 d} \log^k n)$, and $O(\exp(k/\varepsilon) \cdot nd)$, respectively.

While the polynomial time approximation schemes show that k-means clustering can be approximated arbitrarily well, the method of choice for finding a k-clustering is the k-means method due to its performance in practice. However, the only polynomial bound for k-means holds for $d = 1$, and only for instances with polynomial spread [7], which is the maximum distance of points divided by the minimum distance.

Arthur and Vassilvitskii [3] have analyzed the running-time of the k-means method subject to Gaussian perturbation: The points are drawn according to independent d-dimensional Gaussian distributions with standard deviation σ. Arthur and Vassilvitskii proved that the expected running-time after perturbing the input with Gaussians with standard deviation σ is poly-

nomial in n^k, d, the diameter of the perturbed point set, and $1/\sigma$.

Recently, Arthur [1] showed that the probability that the running-time of k-means subject to Gaussian perturbations exceeds a polynomial in n, d, the diameter of the instance, and $1/\sigma$ is bounded by $O(1/n)$. However, his argument does not yield any significant bound on the expected running-time of k-means: The probability of $O(1/n)$ that the running-time exceeds a polynomial bound is too large to yield an upper bound for the expected running-time, except for the trivial upper bound of $\text{poly}(n^{kd})$.

1.3 New Results.

We improve the smoothed analysis of the k-means method by proving two upper bounds on its running-time. First, we show that the smoothed running-time of k-means is bounded by a polynomial in $n^{\sqrt{k}}$ and $1/\sigma$.

THEOREM 1.1. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a set of n points drawn according to independent Gaussian distributions whose means are in $[0, 1]^d$. Then the expected running-time of the k-means method on the instance $\mathcal{X}$ is bounded from above by a polynomial in $n^{\sqrt{k}}$ and $1/\sigma$.*

Thus, compared to the previously known bound, we decrease the exponent by a factor of $\sqrt{k}$. Second, we show that the smoothed running-time of k-means is bounded by $k^{kd} \cdot \text{poly}(n, 1/\sigma)$. In particular, this decouples the exponential part of the bound from the number n of points.

THEOREM 1.2. *Let $\mathcal{X}$ be drawn as described in Theorem 1.1. Then the expected running-time of the k-means method on the instance $\mathcal{X}$ is bounded from above by $k^{kd} \cdot \text{poly}(n, 1/\sigma)$.*

An immediate consequence of Theorem 1.2 is the following corollary, which proves that the expected running-time is polynomial in n and $1/\sigma$ if k and d are small compared to n. This result is of particular interest since d and k are usually much smaller than n.

COROLLARY 1.1. *Let $k, d \in O(\sqrt{\log n / \log \log n})$. Let $\mathcal{X}$ be drawn as described in Theorem 1.1. Then the expected running-time of k-means on the instance $\mathcal{X}$ is bounded by a polynomial in n and $1/\sigma$.*

David Arthur [1] presented an insightful proof that k-means runs in time polynomial in n, $1/\sigma$, and the diameter of the instance with a probability of at least $1 - O(1/n)$. It is worth pointing out that his result is orthogonal to our results: neither do our results imply polynomial running time with probability $1 - O(1/n)$,

nor does Arthur's result yield any non-trivial bound on the expected running-time (not even poly$(n^k, 1/\sigma)$) since the success probability of $1 - O(1/n)$ is way too small. The exception is our result for $d = 1$, which yields not only a bound on the expectation, but also a bound that holds with high probability. However, the original definition of smoothed analysis [12] is in terms of expectation, not in terms of bounds that hold with a probability of $1 - o(1)$.

To prove our bounds, we prove a lemma about perturbed point sets (Lemma 2.1). The lemma bounds the number of points close to the boundaries of Voronoi partitions that arise during the execution of k-means. It might be of independent interest, in particular for smoothed analyses of geometric algorithms and problems.

Finally, we prove a polynomial bound for the running-time of k-means in one dimension.

THEOREM 1.3. *Let $\mathcal{X} \subseteq \mathbb{R}$ be drawn according to 1-dimensional Gaussian distributions as described in Theorem 1.1. Then the expected running-time of k-means on $\mathcal{X}$ is polynomial in n and $1/\sigma$. Furthermore, the probability that the running-time exceeds a polynomial in n and $1/\sigma$ is bounded by $1/\operatorname{poly}(n)$.*

We remark that this result for $d = 1$ is not implied by the result of Har-Peled and Sadri [7] that the running-time of one-dimensional k-means is polynomial in n and the spread of the instance. The reason is that the expected value of the square of the spread is unbounded.

The restriction of the adversarial points to be in $[0, 1]^d$ is necessary: Without any bound, the adversary can place the points arbitrarily far away, thus diminishing the effect of the perturbation. We can get rid of this restriction and obtain the same results by allowing the bounds to be polynomial in the diameter of the adversarial instance. However, for the sake of clarity and to avoid another parameter, we have chosen the former model.

1.4 Outline. To prove our two main theorems, we first prove a property of perturbed point sets (Section 2): In any step of the k-means algorithm, there are not too many points close to any of the at most $\binom{k}{2}$ hyperplanes that bisect the centers and that form the Voronoi regions. To put it another way: No matter how k-means partitions the point set $\mathcal{X}$ into k Voronoi regions, the number of points close to any boundary is rather small with overwhelming probability.

We use this lemma in Section 3: First, we use it to prove Lemma 3.1, which bounds the expected number of iterations in terms of the smallest possible distance of two clusters. Using this bound, we derive a first upper bound for the expected number of iterations (Lemma 3.2), which will result in Theorem 1.2 later on.

In Sections 4 and 5, we distinguish between iterations in which at most $\sqrt{k}$ or at least $\sqrt{k}$ clusters gain or lose points. This will result in Theorem 1.1.

We consider the special case of $d = 1$ in Section 6. For this case, we prove an upper bound polynomial in n and $1/\sigma$ until the potential has dropped by at least 1.

In Sections 3, 4, 5, and 6 we are only concerned with bounding the number of iterations until the potential has dropped by at least 1. Using these bounds and an upper bound on the potential after the first round, we will derive Theorems 1.1, 1.2, and 1.3 as well as Corollary 1.1 in Section 7.

Due to space limitations, some proofs can only be found in the full version of this paper.

1.5 Preliminaries. In the following, $\mathcal{X}$ is the perturbed instance on which we run k-means, i.e., $\mathcal{X} = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ is a set of n points, where each point x_i is drawn according to a d-dimensional Gaussian distribution with mean $\mu_i \in [0, 1]^d$ and standard deviation σ.

Inaba et al. [8] proved that the number of iterations of k-means is poly(n^{kd}) in the worst case. We abbreviate this bound by $W \leq n^{\kappa kd}$ for some constant κ in the following.

Let $D \geq 1$ be chosen such that, with a probability of at least $1 - W^{-1}$, every data point from $\mathcal{X}$ lies in the hypercube $\mathcal{D} := [-D, 1 + D]^d$ after the perturbation. In Section 7, we prove that D can be bounded by a polynomial in n and σ, and we use this fact in the following sections. We denote by $\mathcal{F}$ the *failure event* that there exists one point in $\mathcal{X}$ that does not lie in the hypercube $\mathcal{D}$ after the perturbation. We say that a cluster is *active* in an iteration if it gains or loses at least one point.

We will always assume in the following that $d \leq n$ and $k \leq n$, and we will frequently bound both d and k by n to simplify calculations. Of course, $k \leq n$ holds for every meaningful instance since it does not make sense to partition n points into more than n clusters. Furthermore, we can assume $d \leq n$ for two reasons: First, the dimension is usually much smaller than the number of points, and, second, if $d > n$, then we can project the points to a lower-dimensional subspace without changing anything.

Let $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ denote the set of clusters. For a natural number k, let $[k] = \{1, \ldots, k\}$. In the following, we will assume that numbers such as $\sqrt{k}$ are integers. For the sake of clarity, we do not write down the tedious floor and ceiling functions that are

actually necessary. Since we are only interested in the asymptotics, this does not affect the validity of the proofs. Furthermore, we assume in the following sections that $\sigma \leq 1$. This assumption is only made to simplify the arguments and we describe in Section 7 how to get rid of it.

2 A Property of Perturbed Point Sets

The following lemma shows that, with high probability, there are not too many points close to the hyperplanes dividing the clusters. It is crucial for our bounds for the smoothed running-time: If not too many points are close to the bisecting hyperplanes, then, eventually, one point that is further away from the bisecting hyperplanes must go from one cluster to another, which causes a significant decrease of the potential.

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

$$\varepsilon := \frac{\sigma^4}{32n^2 dD^2} \cdot \left(\frac{\sigma}{3Dn^{3+2\kappa}}\right)^{4a}$$

from the bisector that it crosses.

Proof. We consider a step of the k-means algorithm, and we refer to the configuration before this step as the *first configuration* and to the configuration after this step 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 we consider is the reassignment of the points according to the Voronoi diagram in the first configuration.

Let $B \subseteq \mathcal{X}$ with $|B| = \ell := kd/a$ be a set of points that change their assignment during the step. There are at most n^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 choices for this partition [8]. 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 $A_i \cup B_i$. Since the positions of the points in $\mathcal{X} \setminus B$ are assumed to be fixed by an adversary, and since we apply a union bound over the partition $A_1, \ldots, A_k$, 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 randomness of the points in B. In particular, the bisecting 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

$$(\mathcal{E}) \qquad \forall i, j \colon \forall b \in B_{i,j} \colon \operatorname{dist}(b, H_{i,j}) \leq \varepsilon \, ,$$

where $\operatorname{dist}(x, H) = \min_{y \in H} \|x - y\|$ denotes the shortest distance of a point x to a hyperplane H. In the following, we denote this event by $\mathcal{E}$. If the hyperplanes $H_{i,j}$ were fixed, the probability of $\mathcal{E}$ could readily be seen to be at most $\left(\frac{2\varepsilon}{\sigma\sqrt{2\pi}}\right)^\ell \leq \left(\frac{\varepsilon}{\sigma}\right)^\ell$. But the hyperplanes are not fixed since their positions and orientations depend on the points in the sets $B_{i,j}$. Therefore, we are only able to prove the following weaker bound in Lemma 2.2:

$$\Pr[\mathcal{E} \wedge \neg\mathcal{F}] \leq \left(\frac{3D}{\sigma}\right)^{kd} \cdot \left(\frac{32n^2 dD^2 \varepsilon}{\sigma^4}\right)^{\ell/4} \, ,$$

where $\neg\mathcal{F}$ denotes the event that, after the perturbation, all points of $\mathcal{X}$ lie in the hypercube $\mathcal{D} = [-D, D + 1]^d$. Now the union bound yields the following upper bound on the probability that a set B with the stated properties exists:

$$\Pr[\mathcal{E}] \leq \Pr[\mathcal{E} \wedge \neg\mathcal{F}] + \Pr[\mathcal{F}]$$
$$\leq n^{3\ell} W \cdot \left(\frac{3D}{\sigma}\right)^{kd} \cdot \left(\frac{32n^2 dD^2 \varepsilon}{\sigma^4}\right)^{\ell/4} + W^{-1}$$
$$= n^{3\ell} W \cdot \left(\frac{1}{n^{3+2\kappa}}\right)^{kd} + W^{-1}$$
$$\leq n^{3\ell + \kappa kd} \cdot \left(\frac{1}{n^{3+2\kappa}}\right)^{kd} + W^{-1}$$
$$\leq n^{-\kappa kd} + W^{-1} \leq 2W^{-1} \, .$$

The equation is by our choice of ε, the inequalities are due to some simplifications and $W \leq n^{\kappa kd}$. $\square$

LEMMA 2.2. *The probability of the event $\mathcal{E} \wedge \neg\mathcal{F}$ is bounded from above by*

$$\left(\frac{3D}{\sigma}\right)^{kd} \cdot \left(\frac{32n^2 dD^2 \varepsilon}{\sigma^4}\right)^{\ell/4} \quad .$$

3 An Upper Bound

Lemma 2.1 yields an upper bound on the number of iterations that k-means needs: Since there are only few points close to hyperplanes, eventually a point switches from one cluster to another that initially was not close to a hyperplane. The results of this section lead to the proof of Theorem 1.2.

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

LEMMA 3.1. *For every $a \in [k]$, with a probability of at least $1 - 3W^{-1}$, every sequence of $k^{kd/a} + 1$ consecutive steps of the k-means algorithm (not including the first one) reduces the potential by at least*

$$\frac{\varepsilon^2 \cdot \min\{\Delta^2, 1\}}{36dD^2 k^{kd/a}} ,$$

where Δ denotes the smallest distance of two cluster centers that occurs during the sequence and ε is defined as in Lemma 2.1.

In order to obtain a bound on the number of iterations that k-means needs, we need to bound the distance Δ of the closest cluster centers. This is done in the following lemma, which exploits Lemma 3.1. The following lemma is the crucial ingredient of the proof of Theorem 1.2.

LEMMA 3.2. *Let $a \in [k]$ be arbitrary. Then the expected number of steps until the potential drops by at least 1 is bounded from above by*

$$\gamma \cdot k^{2kd/a} \cdot nkd \left(\frac{d^2 n^4 D}{\sigma \varepsilon}\right)^2$$

for a sufficiently large absolute constant γ.

Proof. With a probability of at least $1 - 3W^{-1}$, the number of iterations until the potential drops by at least

$$\frac{\varepsilon^2 \cdot \min\{\Delta^2, 1\}}{36dD^2 k^{kd/a}}$$

is at most $k^{kd/a} + 1$ due to Lemma 3.1. We estimate the contribution of the failure event, which occurs only with probability $3W^{-1}$, to the expected running time by 3 and ignore it in the following. Let T denote the random variable that equals the number of sequences of length $k^{kd/a} + 1$ until the potential has dropped by one.

The random variable T can only exceed t if

$$\min\{\Delta^2, 1\} \leq \frac{36dD^2 k^{kd/a}}{\varepsilon^2 \cdot t} ,$$

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

$$\begin{aligned}
\mathrm{E}\,[T] &= \textstyle\sum_{t=1}^{W} \Pr[T \geq t] \\
&\leq \int_0^W \Pr\left[\min\{\Delta^2, 1\} \leq \frac{36dD^2 k^{kd/a}}{\varepsilon^2 \cdot t}\right] \mathrm{d}t \\
&\leq t' + \int_{t'}^W \Pr\left[\Delta \leq \frac{6\sqrt{d}Dk^{kd/(2a)}}{\varepsilon \cdot \sqrt{t}}\right] \mathrm{d}t ,
\end{aligned}$$

for $t' = \left(\frac{(24d+96)n^4\sqrt{d}Dk^{kd/(2a)}}{\sigma\varepsilon}\right)^2$.

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 between 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. From this, we can conclude the following: for every point that is assigned to $\mathcal{C}_1$ or $\mathcal{C}_2$ and that has a distance of at least δ from the bisector H, as compensation another point must be assigned to $\mathcal{C}_1$ or $\mathcal{C}_2$ that has a distance of at most $\delta/2$ from H. Hence, the total number of points assigned to $\mathcal{C}_1$ or $\mathcal{C}_2$ can be at most twice as large as the total number of points assigned to $\mathcal{C}_1$ or $\mathcal{C}_2$ that are at a distance of at most δ 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 δ.

2. There exist two subsets of points whose union has cardinality 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 as follows using a union bound and Lemma 4.4 (see also Arthur and Vassilvitskii [3, Proposition 5.6]):

$$n^{2d}\left(\frac{4d\delta}{\sigma}\right)^{2d-d} + (2n)^{4d}\cdot\left(\frac{\delta}{\sigma}\right)^d \leq \left(\frac{(4d+16)n^4\delta}{\sigma}\right)^d \quad .$$

Hence, $\Pr\left[\Delta \leq \frac{6\sqrt{d}Dk^{kd/(2a)}}{\varepsilon \cdot \sqrt{t}}\right] \leq \left(\frac{\sqrt{t'}}{\sqrt{t}}\right)^d$ and, for $d \geq 3$, we obtain

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

For $d = 2$, we obtain

$$\mathrm{E}\left[T\right] \leq t' + \int_{t'}^{W} \left(\frac{\sqrt{t'}}{\sqrt{t}}\right)^d \mathrm{d}t \leq t' + t' \cdot \left[\ln(t)\right]_1^W$$
$$= t' \cdot (1 + \ln(W)) \leq 2\kappa nkd \cdot t' \ .$$

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

$$2 + \left(k^{kd/a} + 1\right) \cdot 2\kappa nkd \cdot \left(\frac{(24d+96)n^4\sqrt{d}Dk^{kd/(2a)}}{\sigma\varepsilon}\right)^2 \ ,$$

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

$$\gamma \cdot k^{2kd/a} \cdot nkd \cdot \left(\frac{d^2n^4D}{\sigma\varepsilon}\right)^2 \ . \quad \square$$

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

In this and the following section, we aim at proving the main lemmas that lead to Theorem 1.1. To do this, we distinguish two cases: In this section, we deal with the case that at most $\sqrt{k}$ clusters are active. In this case, either few points change clusters, which yields a potential drop caused by the movement of the centers. Or many points change clusters. Then, in particular, many points switch between two clusters, and not all of them can be close to the hyperplane bisecting the corresponding centers, which yields the potential due to the reassignment.

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. The obvious upper bound for the length of an epoch is 2^k, which is stated also by Arthur and Vassilvitskii [3]: After that many iterations, at least one cluster must have assumed a third position. For our analysis, however, 2^k is too big, and we bring it down to a constant.

LEMMA 4.1. *The length of any epoch is less than four.*

Proof. Let x be any data point that changes from one cluster to another during an epoch, and let $i_1, i_2, \ldots, i_\ell$ be the indices of the different clusters to which x belongs in that order. (We have $i_j \neq i_{j+1}$, but x can change back to a cluster it has already visited. So, e.g., $i_j = i_{j+2}$ is allowed.) For every i_j, we then have two different sets $\mathcal{C}'_{i_j}$ and $\mathcal{C}''_{i_j}$ with centers c'_{i_j} and c''_{i_j} such that $x \in \mathcal{C}''_{i_j} \setminus \mathcal{C}'_{i_j}$. Since x belongs always to exactly one cluster, we have $\mathcal{C}_{i_j} = \mathcal{C}'_{i_j}$ for all except for one j for which $\mathcal{C}_{i_j} = \mathcal{C}''_{i_j}$. Now assume that $\ell \geq 4$. Then, when changing from $\mathcal{C}_{i_1}$ to $\mathcal{C}_{i_2}$, we have $\|x - c'_{i_2}\| < \|x - c'_{i_4}\|$

since x prefers $\mathcal{C}_{i_2}$ over $\mathcal{C}_{i_4}$ and, when changing to $\mathcal{C}_{i_4}$, we have $\|x - c'_{i_4}\| < \|x - c'_{i_2}\|$. This contradicts the assumption that $\ell \geq 4$.

Now assume that x does not change from $\mathcal{C}_{i_j}$ to $\mathcal{C}_{i_{j+1}}$ for a couple of steps, i.e., x waits until it eventually changes clusters. Then the reason for eventually changing to $\mathcal{C}_{i_{j+1}}$ can only be that either $\mathcal{C}_{i_j}$ has changed to some $\tilde{\mathcal{C}}_{i_j}$, which makes x prefer $\mathcal{C}_{i_{j+1}}$. But, since $\tilde{\mathcal{C}}_{i_j} \neq \mathcal{C}''_{i_j}$ and $x \in \tilde{\mathcal{C}}_{i_j}$, we have a third cluster for $\mathcal{C}_{i_j}$. Or $\mathcal{C}_{i_{j+1}}$ has changed to $\tilde{\mathcal{C}}_{i_{j+1}}$, and x prefers $\tilde{\mathcal{C}}_{i_{j+1}}$. But then $\tilde{\mathcal{C}}_{i_{j+1}} \neq \mathcal{C}'_{i_j}$ and $x \notin \tilde{\mathcal{C}}_{i_{j+1}}$, and we have a third cluster for $\mathcal{C}_{i_{j+1}}$.

We can conclude that x visits at most three different clusters, and changes its cluster in every iteration of the epoch. Furthermore, the order in which x visits its clusters is periodic with a period length of at most three. Finally, even a period length of three is impossible: Suppose x visits $\mathcal{C}_{i_1}$, $\mathcal{C}_{i_2}$, and $\mathcal{C}_{i_3}$. Then, to go from $\mathcal{C}_{i_j}$ to $\mathcal{C}_{i_{j+1}}$ (arithmetic is modulo 3), we have $\|x - c'_{i_{j+1}}\| < \|x - c'_{i_{j-1}}\|$. Since this holds for $j = 1, 2, 3$, we have a contradiction.

This holds for every data point. Thus, after at most four iterations either k-means terminates, which is fine, or some cluster assumes a third configuration, which ends the epoch, or some clustering repeats, which is impossible. $\square$

Similar to Arthur and Vassilvitskii [3], we define a *key-value* to be an expression of the form $K = \frac{s}{t} \cdot \mathrm{cm}(S)$, where $s, t \in \mathbb{N}$, $s \leq n^2$, $t < n$, and $S \subseteq \mathcal{X}$ is a set of at most $4d\sqrt{k}$ points. (Arthur and Vassilvitskii allow up to $4dk$ points.) For two key-values K_1, K_2, we write $K_1 \equiv K_2$ if and only if they have identical coefficients for every data point.

We say that $\mathcal{X}$ is δ-*sparse* if, for every key-values K_1, K_2, K_3, K_4 with $\|K_1 + K_2 - K_3 - K_4\| \leq \delta$, we have $K_1 + K_2 \equiv K_3 + K_4$.

LEMMA 4.2. *The probability that the point set $\mathcal{X}$ is not δ-sparse is at most*

$$n^{16d\sqrt{k}+12} \cdot \left(\frac{n^4\delta}{\sigma}\right)^d .$$

After four iterations, one cluster has assumed a third center or k-means terminates. This yields the following lemma (see also Arthur and Vassilvitskii [3, Corollary 5.2]).

LEMMA 4.3. *Assume that $\mathcal{X}$ is δ-sparse. Then, in every sequence of four consecutive iterations that do not lead to termination and such that in every of these iterations*

- *at most $\sqrt{k}$ clusters are active and*
- *each cluster gains or loses at most $2d\sqrt{k}$ points,*

the potential decreases by at least $\frac{\delta^2}{4n^4}$.

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 distance ε of H. The following lemma, due to Arthur and Vassilvitskii [3, Proposition 5.6], shows that $\mathcal{X}$ is likely to be ε-separated.

LEMMA 4.4. (ARTHUR, VASSILVITSKII [3]) $\mathcal{X}$ *is not ε-separated with a probability of at most*

$$n^{2d} \cdot \left(\frac{4d\varepsilon}{\sigma}\right)^d.$$

Given that $\mathcal{X}$ is ε-separated, every iteration with at most $\sqrt{k}$ active clusters in which one cluster gains or loses at least $2d\sqrt{k}$ points yields a significant decrease of the potential.

LEMMA 4.5. *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 $2\varepsilon^2/n$.*

This lemma is similar to Proposition 5.4 of Arthur and Vassilvitskii [3]. We present here a corrected proof based on private communication with Sergei Vassilvitskii.

Proof. If 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 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 hyperplane bisecting the cluster centers c_i and c_j. Assume that x switches from $\mathcal{C}_i$ to $\mathcal{C}_j$.

Then the potential decreases by at least $\|c_i - x\|^2 - \|c_j - x\|^2 = (2x - c_i - c_j) \cdot (c_j - c_i)$. Let v be the unit vector in $c_j - c_i$ direction. Then $(2x - c_i - c_j) \cdot v \geq 2\varepsilon$. We have $c_j - c_i = \alpha v$ for $\alpha = \|c_j - c_i\|$, and hence, it remains to bound $\|c_j - c_i\|$ from below. If we can prove $\alpha \geq \varepsilon/n$, then we have a potential drop of at least $(2x - c_i - c_j) \cdot \alpha v \geq \alpha 2\varepsilon \geq 2\varepsilon^2/n$ as claimed.

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. In particular, this implies that $\|c_i - c_j\| \geq \text{dist}(c_i, H) + \text{dist}(c_j, H)$.

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

Now set $\delta_i = n^{-16-(16+i)\cdot\sqrt{k}} \cdot \sigma$ and $\varepsilon_i = \sigma \cdot n^{-4-i\sqrt{k}}$. Then the probability that the instance is not δ_i-sparse is bounded from above by

$$n^{16d\sqrt{k}+12+4d-16d-(16+i)d\cdot\sqrt{k}} \leq n^{-id\sqrt{k}}.$$

The probability that the instance is not ε_i-separated is bounded from above by (we use $d \leq n$ and $4 \leq n$)

$$n^{4d-4d-id\sqrt{k}} = n^{-id\sqrt{k}}.$$

We abbreviate the fact that an instance is δ_i-sparse and ε_i-separated by *i-nice*. Now Lemmas 4.3 and 4.5 immediately yield the following lemma.

LEMMA 4.6. *Assume that $\mathcal{X}$ is i-nice. Then the number of sequences of at most four consecutive iterations, each of which with at most $\sqrt{k}$ active clusters, until the potential has dropped by at least 1 is bounded from above by*

$$\left(\min\left\{\frac{1}{4} \cdot n^{-36-(32+2i)\sqrt{k}} \cdot \sigma^2, 2\sigma^2 \cdot n^{-9-i2\sqrt{k}}\right\}\right)^{-1}$$
$$\leq \frac{n^{(c+2i)\cdot\sqrt{k}}}{\sigma^2} =: S_i$$

for a suitable constant c.

The first term comes from δ_i, which yields a potential drop of at least $\delta_i^2/(4n^4)$. The second term comes from ε_i, which yields a drop of at least $2\varepsilon_i^2/n$.

Putting the pieces together yields the main lemma of this section.

LEMMA 4.7. *The expected number of sequences of at most four consecutive iterations, each of which with at most $\sqrt{k}$ active clusters, until the potential has dropped by at least 1 is bounded from above by*

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

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

In this section, we consider steps of the k-means algorithm in which at least $\sqrt{k}$ different clusters gain or lose points. The improvement yielded by such a step can only be small if none of the cluster centers changes its position significantly due to the reassignment of points, which, intuitively, becomes increasingly unlikely the more clusters are active. We show that, indeed, if at least $\sqrt{k}$ clusters are active, then with high probability one of them changes its position by $n^{-O(\sqrt{k})}$, yielding a potential drop in the same order of magnitude.

The following observation, which has also been used by Arthur and Vassilvitskii [3], relates the movement of a cluster center to the potential drop.

LEMMA 5.1. *If in an iteration of the k-means algorithm a cluster center changes its position from c to c', then the potential drops by at least $\|c - c'\|^2$.*

Now we are ready to prove the main lemma of this section.

LEMMA 5.2. *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*

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

Proof. We consider one step of the k-means algorithm with at least $\sqrt{k}$ active clusters. Let ε be defined as in Lemma 2.1 for $a = 1$. We distinguish two cases: Either one point that is reassigned during the considered iteration has a distance of at least 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 $2\varepsilon\Delta$, where Δ denotes the minimal distance of two cluster centers. In the latter case, Lemma 2.1 implies that with high probability less than kd points are reassigned during the considered step. We apply a union bound over the choices for these points. In the union bound, we fix not only these points but also the clusters they are assigned to before and after the step. We denote by A_i the set of points that are assigned to cluster $\mathcal{C}_i$ in both configurations and we denote by B_i and B'_i the sets of points assigned to cluster $\mathcal{C}_i$ before and after the step, respectively, except for the points in A_i. Analogously to Lemma 2.1, we assume that the positions of the points in $A_1 \cup \ldots \cup A_k$ are fixed adversarially, 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

$$n^{\kappa kd} \cdot n^{3kd} \leq n^{(\kappa+3) \cdot kd}$$

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

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

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

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

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

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

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. To be more precise, we do the following: Consider a graph whose nodes are the active clusters and that contains an edge between two nodes if and only if the corresponding clusters exchange at least one point. We identify a dominating set in this graph, i.e., a subset of nodes that covers the graph in the sense that every node not belonging to this subset has at least one edge into the subset. We can assume that the dominating set, which we identify, contains at most half of the active clusters. (In order to find such a dominating set, start with the graph and throw out edges until the remaining graph is a tree. Then put the nodes on odd layers to the left side and the nodes on even layers to the right side, and take the smaller side as the dominating set.)

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

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

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

$$n^{(\kappa+3) \cdot kd} \cdot \left(\frac{n\beta}{\sigma}\right)^{d\ell/2} \leq n^{(\kappa+3) \cdot kd} \cdot \left(\frac{n\beta}{\sigma}\right)^{d\sqrt{k}/2} \ .$$

For
$$\beta = \frac{\sigma}{n^{(4\kappa+6)\cdot\sqrt{k}+1}}$$

the probability can be bounded from above by $n^{-\kappa k d} \leq W^{-1}$.

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

$$\Gamma := \min\{2\varepsilon\Delta, \beta^2\}$$
$$\geq \min\left\{\frac{\sigma^8\Delta}{1296 n^{14+8\kappa}D^6 d}, \frac{\sigma^2}{n^{(8\kappa+12)\cdot\sqrt{k}+2}}\right\}$$
$$\geq \min\left\{\Delta \cdot \text{poly}\left(n^{-1}, \sigma\right), \text{poly}\left(n^{-\sqrt{k}}, \sigma\right)\right\}$$

since $d \leq n$ 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 $\Gamma \leq 1/t$. Hence, $\text{E}[T]$ is bounded from above by

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

where the integral is upper bounded as in the proof of Lemma 3.2. □

6 A Polynomial Bound in One Dimension

In this section, we consider a one-dimensional set $\mathcal{X} \subseteq \mathbb{R}$ of points. The aim of this section is to prove that the expected number of steps until the potential has dropped by at least 1 is bounded by a polynomial in n and $1/\sigma$.

We say that the point set $\mathcal{X}$ is *ε-spreaded* if the following conditions are fulfilled:

- There is no interval of length ε that contains three or more points of $\mathcal{X}$.
- For any four points x_1, x_2, x_3, x_4, where x_2 and x_3 may denote the same point, we have $|x_1 - x_2| > \varepsilon$ or $|x_3 - x_4| > \varepsilon$.

The following lemma justifies the notion of ε-spreadedness.

LEMMA 6.1. *Assume that $\mathcal{X}$ is ε-spreaded. Then the potential drops by at least $\frac{\varepsilon^2}{4n^2}$ in every iteration.*

Assume that $\mathcal{X}$ is ε-spreaded. Then the number of iterations until the potential has dropped by at least 1 is at most $4n^2/\varepsilon^2$ by the lemma above. Let us estimate the probability that $\mathcal{X}$ is ε-spreaded.

LEMMA 6.2. *The probability that $\mathcal{X}$ is not ε-spreaded is bounded from above by $\frac{2n^4\varepsilon^2}{\sigma^2}$.*

Now we have all ingredients for the proof of the main lemma of this section.

LEMMA 6.3. *The number of iterations of k-means until the potential has dropped by at least 1 is bounded by a polynomial in n and $1/\sigma$.*

Proof. Let T be the random variable of the number of iterations until the potential has dropped by at least 1. If $T \geq t$, then $\mathcal{X}$ cannot be ε-spreaded with $4n^2/\varepsilon^2 \leq t$. Thus, in this case, $\mathcal{X}$ is not ε-spreaded with $\varepsilon = \frac{2n}{\sqrt{t}}$. In the worst case, k-means runs for at most $n^{\kappa k}$ iterations. Hence, the expected running time can be bounded by

$$\sum_{t=1}^{n^{\kappa k}}\Pr[T \geq t] \leq \sum_{t=1}^{n^{\kappa k}}\Pr\left[\mathcal{X} \text{ is not } \frac{2n}{\sqrt{t}}\text{-spreaded}\right]$$
$$\leq \sum_{t=1}^{n^{\kappa k}}\frac{8n^4 n^2}{t\sigma^2} \in O\left(\frac{n^6}{\sigma^2} \cdot \log n^{\kappa k}\right) \subseteq O\left(\frac{n^7}{\sigma^2} \cdot \log n\right). \square$$

Finally, we remark that, by choosing $\varepsilon = \frac{\sigma}{n^{2+c}}$, we obtain that the probability that the number of iterations until the potential has dropped by at least exceeds a polynomial in n and $1/\sigma$ is bounded from above by $O(n^{-2c})$. This yields a bound on the running-time of k-means for $d = 1$ that holds with high probability.

7 Putting the Pieces Together

In the previous sections, we have only analyzed the expected number of iterations until the potential drops by at least 1. To bound the expected number of iterations that k-means needs to terminate, we need an upper on the potential in the beginning. To get this, we use the following lemma.

LEMMA 7.1. *Let x be a one-dimensional Gaussian random variable with standard deviation σ and mean $\mu \in [0, 1]$. Then, for all $t \geq 1$,*

$$\Pr[x \notin [-t, 1+t]] < \sigma \cdot \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

For $D = \sqrt{2\sigma^2 \ln(n^{1+\kappa k d}d\sigma)} \le \text{poly}(n,\sigma)$, the probability that any component of any of the n data points is not contained in the hypercube $\mathcal{D} = [-D, 1+D]^d$ is bounded from above by $n^{-\kappa k d} \le W^{-1}$. This implies that $\mathcal{X} \subseteq \mathcal{D}$ with a probability of at least $1 - W^{-1}$. If $\mathcal{X} \subseteq \mathcal{D}$, then, after the first iteration, the potential is bounded from above by $nd \cdot (2D+1)^2 = \text{poly}(n)$.

In the beginning, we made the assumption that $\sigma \le 1$. While this covers the small values of σ, which we consider as more relevant, the assumption is only a technical requirement, and we can get rid of it: The number of iterations that k-means needs is invariant under scaling of the point set $\mathcal{X}$. Now assume that $\sigma > 1$. Then we consider $\mathcal{X}$ scaled down by $1/\sigma$, which corresponds to the following model: The adversary chooses points from the hypercube $[0, 1/\sigma]^d \subseteq [0,1]^d$, and then we add d-dimensional Gaussian vectors with standard deviation 1 to every data point. The expected running-time that k-means needs on this instance is bounded from above by the running-time needed for adversarial points chosen from $[0,1]^d$ and $\sigma = 1$, which is $\text{poly}(n) \le \text{poly}(n, 1/\sigma)$.

The remaining parts of the proofs of the theorems and the corollary, which are based on straightforward arguments, can be found in the full version of this paper.

8 Conclusions

We have proved two upper bounds for the smoothed running-time of the k-means method: The first bound is $\text{poly}(n^{\sqrt{k}}, 1/\sigma)$. The second bound is $k^{kd} \cdot \text{poly}(n, 1/\sigma)$, which decouples the exponential growth in k and d from the number of points and the standard deviation. In particular, this yields a smoothed running-time that is polynomial in n and $1/\sigma$ for $k, d \in O(\sqrt{\log n / \log \log n})$.

The obvious question now is whether a bound exists that is polynomial in n and $1/\sigma$, without exponential dependence on k or d. We believe that such a bound exists. However, we suspect that new techniques are required to prove it; bounding the smallest possible improvement from below might not be sufficient. The reason for this is that the number of possible partitions, and thus the number of possible k-means steps, grows exponentially in k, which makes it more likely for small improvements to exist as k grows.

Finally, we are curious if our techniques carry over to other heuristics. In particular Lemma 2.1 is quite general, as it bounds the number of points from above that are close to the boundaries of the Voronoi partitions that arise during the execution of k-means. In fact, we believe that a slightly weaker version of Lemma 2.1 is also true for arbitrary Voronoi partitions and not only for those arising during the execution of k-means. This

insight might turn out to be helpful in other contexts as well.

Acknowledgement

References

[1] David Arthur. Smoothed analysis of the k-means method. Manuscript, 2008.

[2] David Arthur and Sergei Vassilvitskii. How slow is the k-means method? In Nina Amenta and Otfried Cheong, editors, *Proc. of the 22nd ACM Symposium on Computational Geometry (SOCG)*, pages 144–153. ACM Press, 2006.

[3] 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 Foundations of Comp. Science (FOCS)*, pages 153–164. IEEE Computer Society, 2006.

[4] Mihai Bădoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proc. of the 34th Ann. ACM Symposium on Theory of Computing (STOC)*, pages 250–257. ACM Press, 2002.

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

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

[7] Sariel Har-Peled and Bardia Sadri. How fast is the k-means method? *Algorithmica*, 41(3):185–202, 2005.

[8] 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.

[9] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Proc. of the 45th Ann. IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 454–462, 2004.

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

[11] Jiří Matoušek. On approximate geometric k-clustering. *Discrete and Computational Geometry*, 24(1):61–84, 2000.

[12] 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.

[13] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms and heuristics: Progress and open questions. In Luis M. Pardo, Allan Pinkus, Endre Süli, and Michael J. Todd, editors, *Foundations of Computational Mathematics, Santander 2005*, pages 274–342. Cambridge University Press, 2006.