Smoothed Analysis of the k-Means Method

DAVID ARTHUR, Stanford University, Department of Computer Science
BODO MANTHEY, University of Twente, Department of Applied Mathematics
HEIKO RÖGLIN, University of Bonn, Department of Computer Science

The k-means method is one of the most widely used clustering algorithms, drawing its popularity from its speed in practice. Recently, however, it was shown to have exponential worst-case running time. In order to close the gap between practical performance and theoretical analysis, the k-means method has been studied in the model of smoothed analysis. But even the smoothed analyses so far are unsatisfactory as the bounds are still super-polynomial in the number n of data points.

In this paper, we settle the smoothed running time of the k-means method. We show that the smoothed number of iterations is bounded by a polynomial in n and $1/\sigma$, where σ is the standard deviation of the Gaussian perturbations. This means that if an arbitrary input data set is randomly perturbed, then the k-means method will run in expected polynomial time on that input set.

Categories and Subject Descriptors: F.2.0 [**Analysis of Algorithms and Problem Complexity**]: General

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Data Clustering, k-Means Method, Smoothed Analysis

1. INTRODUCTION

Clustering is a fundamental problem in computer science with applications ranging from biology to information retrieval and data compression. In a clustering problem, a set of objects, usually represented as points in a high-dimensional space $\mathbb{R}^d$, is to be partitioned such that objects in the same group share similar properties. The k-means method is a traditional clustering algorithm, originally conceived by Lloyd [1982]. It begins with an arbitrary clustering based on k centers in $\mathbb{R}^d$, and then repeatedly makes local improvements until the clustering stabilizes. The algorithm is greedy and as such, it offers virtually no accuracy guarantees. However, it is both very simple and very fast, which makes it appealing in practice. Indeed, one recent survey of data mining techniques states that the k-means method "is by far the most popular clustering algorithm used in scientific and industrial applications" [Berkhin 2002].

However, theoretical analysis has long been at stark contrast with what is observed in practice. In particular, it was recently shown that the worst-case running time of the k-means method is $2^{\Omega(n)}$ even on two-dimensional instances [Vattani pear]. Conversely, the only upper bounds known for the general case are k^n and $n^{O(kd)}$. Both upper bounds

Authors' e-mail addresses: darthur@cs.stanford.edu, b.manthey@utwente.nl, and heiko@roeglin.org.
A preliminary version of this paper appeared as "k-Means has Polynomial Smoothed Complexity" [Arthur et al. 2009]. This paper also includes parts of "Improved Smoothed Analysis of the k-Means Method" [Manthey and Röglin 2009a].
The work was done in part while Bodo Manthey was at Saarland University, Department of Computer Science, Saarbrücken, Germany.
Heiko Röglin was supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD).

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

are based entirely on the trivial fact that the k-means method never encounters the same clustering twice [Inaba et al. 2000]. In contrast, Duda et al. state that the number of iterations until the clustering stabilizes is often linear or even sublinear in n on practical data sets [Duda et al. 2000, Section 10.4.3]. The only known polynomial upper bound, however, applies only in one dimension and only for certain inputs [Har-Peled and Sadri 2005].

So what does one do when worst-case analysis is at odds with what is observed in practice? We turn to the smoothed analysis of Spielman and Teng [2004], which considers the running time after first randomly perturbing the input. Intuitively, this models how fragile worst-case instances are and if they could reasonably arise in practice. In addition to the original work on the simplex algorithm, smoothed analysis has been applied successfully in other contexts, e.g., for the ICP algorithm [Arthur and Vassilvitskii 2009], online algorithms [Becchetti et al. 2006], the knapsack problem [Beier and Vöcking 2004], and the 2-opt heuristic for the TSP [Englert et al. 2007].

The k-means method is in fact a perfect candidate for smoothed analysis: it is extremely widely used, it runs very fast in practice, and yet the worst-case running time is exponential. Performing this analysis has proven very challenging however. It has been initiated by Arthur and Vassilvitskii who showed that the smoothed running time of the k-means method is polynomially bounded in n^k and $1/\sigma$, where σ is the standard deviation of the Gaussian perturbations [Arthur and Vassilvitskii 2009]. The term n^k has been improved to $\min(n^{\sqrt{k}}, k^{kd} \cdot n)$ by Manthey and Röglin [2009a]. Unfortunately, this bound remains super-polynomial even for relatively small values of k, e.g., $k = \log n$. In this paper we settle the smoothed running time of the k-means method: We prove that it is polynomial in n and $1/\sigma$. The exponents in the polynomial are unfortunately too large to match the practical observations, but this is in line with other works in smoothed analysis, including Spielman and Teng's original analysis of the simplex method [Spielman and Teng 2004]. The arguments presented here, which reduce the smoothed upper bound from exponential to polynomial, are intricate enough without trying to optimize constants, even in the exponent. However, we hope and believe that our work can be used as a basis for proving tighter results in the future.

Note that we only analyze the running time in this paper. We do not analyze the quality of the local optimum found by the k-means method or whether it is a global optimum. In fact, it is not the case that the k-means method usually finds the global optimum in practice. But it usually seems to be fast. Thus, our analysis of the running time matches the observed performance of the k-means method.

1.1. k-Means Method

An input for the k-means method is a set $\mathcal{X} \subseteq \mathbb{R}^d$ of n data points. The algorithm outputs k centers $c_1, \ldots, c_k \in \mathbb{R}^d$ and a partition of $\mathcal{X}$ into k clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$. The k-means method proceeds as follows:

(1) Select cluster centers $c_1, \ldots, c_k \in \mathbb{R}^d$ arbitrarily.
(2) Assign every $x \in \mathcal{X}$ to the cluster $\mathcal{C}_i$ whose cluster center c_i is closest to it, i.e., $\|x - c_i\| \leq \|x - c_j\|$ for all $j \neq i$.
(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.

In the following, an *iteration* of k-means refers to one execution of step 2 followed by step 3. A slight technical subtlety in the implementation of the algorithm is the possible event that a cluster loses all its points in Step 2. There exist some strategies

to deal with this case [Har-Peled and Sadri 2005]. For simplicity, we use the strategy of removing clusters that serve no points and continuing with the remaining clusters.

If we define $c(x)$ to be the center of the cluster that data point x is assigned to, then one can check that each step of the algorithm decreases the following potential function:

$$\Psi = \sum_{x \in \mathcal{X}} \|x - c(x)\|^2 . \tag{1}$$

The essential observation for this is the following: If we already have cluster centers $c_1, \ldots, c_k \in \mathbb{R}^d$ representing clusters, then every data point should be assigned to the cluster whose center is nearest to it to minimize Ψ. On the other hand, given clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$, the centers $c_1, \ldots, c_k$ should be chosen as the centers of mass of their respective clusters in order to minimize the potential, which follows from Lemma 2.3.

In the following, we will speak of k-means rather than of the k-means method for short. The worst-case running time of k-means is bounded from above by $(k^2 n)^{kd} \leq n^{3kd}$, which follows from Inaba et al. [2000] and Warren [1968]. (The bound of $O(n^{kd})$ frequently stated in the literature holds only for constant values for k and d, but in this paper k and d are allowed to depend on n.) This upper bound is based solely on the observation that no clustering occurs twice during an execution of k-means since the potential decreases in every iteration. On the other hand, the worst-case number of iterations has been proved to be $\exp(\sqrt{n})$ for $d \in \Omega(\sqrt{n})$ [Arthur and Vassilvitskii 2006]. This has been improved recently to $\exp(n)$ for $d \geq 2$ [Vattani pear].

1.2. Related Work

The problem of finding optimal k-means clusterings is NP-hard even in the restricted cases where $d = 2$ [Mahajan et al. 2009] or $k = 2$ [Aloise et al. 2009], where a clustering is optimal if it minimizes the potential function (1). On the other hand, the problem allows for polynomial-time approximation schemes [Bădoiu et al. 2002; Matoušek 2000; Kumar et al. 2004] with various dependencies of the running time on n, k, d, and the approximation ratio $1 + \varepsilon$. The running times of these approximation schemes depend exponentially on k. Recent research on this subject also includes the work by Gaddam et al. [2007] and Wagstaff et al. [2001]. However, the most widely used algorithm for k-means clustering is still the k-means method due to its simplicity and speed.

Despite its simplicity, the k-means method itself and variants thereof are still the subject of research [Kanungo et al. 2002; Arthur and Vassilvitskii 2007; Ostrovsky et al. 2006]. Let us mention in particular the work by Har-Peled and Sadri [2005] who have shown that a certain variant of the k-means method runs in polynomial time on certain instances. In their variant, a data point is said to be $(1 + \varepsilon)$-misclassified if the distance to its current cluster center is larger by a factor of more than $(1 + \varepsilon)$ than the distance to its closest center. Their *lazy k-means method* only reassigns points that are $(1 + \varepsilon)$-misclassified. In particular, for $\varepsilon = 0$, lazy k-means and k-means coincide. They show that the number of steps of the lazy k-means method is polynomially bounded in the number of data points, $1/\varepsilon$, and the spread of the point set (the spread of a point set is the ratio between its diameter and the distance between its closest pair).

In an attempt to reconcile theory and practice, Arthur and Vassilvitskii [2009] performed the first smoothed analysis of the k-means method: If the data points are perturbed by Gaussian perturbations of standard deviation σ, then the smoothed number of iterations is polynomial in n^k, d, the diameter of the point set, and $1/\sigma$. However, this bound is still super-polynomial in the number n of data points. They conjectured that k-means has indeed polynomial smoothed running time, i.e., that the smoothed number of iterations is bounded by some polynomial in n and $1/\sigma$.

Since then, there has been only partial success in proving the conjecture. Manthey and Röglin [2009a] improved the smoothed running time bound by devising two bounds: The first is polynomial in $n^{\sqrt{k}}$ and $1/\sigma$. The second is $k^{kd} \cdot \mathrm{poly}(n, 1/\sigma)$, where the degree of the polynomial is independent of k and d. Additionally, they proved a polynomial bound for the smoothed running time of k-means on one-dimensional instances.

1.3. Our Contribution

We prove that the k-means method has polynomial smoothed running time. This finally proves Arthur and Vassilvitskii's conjecture [Arthur and Vassilvitskii 2009].

THEOREM 1.1. *Fix an arbitrary set $\mathcal{X}' \subseteq [0,1]^d$ of n points and assume that each point in $\mathcal{X}'$ is independently perturbed by a normal distribution with mean 0 and standard deviation σ, yielding a new set $\mathcal{X}$ of points. Then the expected running time of k-means on $\mathcal{X}$ is bounded by a polynomial in n and $1/\sigma$.*

We did not optimize the exponents in the polynomial as the arguments presented here, which reduce the smoothed upper bound from exponential to polynomial, are already intricate enough and would not yield exponents matching the experimental observations even when optimized. We hope that similar to the smoothed analysis of the simplex algorithm, where the first polynomial bound [Spielman and Teng 2004] stimulated further research culminating in Vershynin's improved bound [Vershynin 2009], our result here will also be the first step towards a small polynomial bound for the smoothed running time of k-means. As a reference, let us mention that the upper bound on the expected number of iterations following from our proof is

$$O\left(\frac{n^{34}\log^4(n)k^{34}d^8}{\sigma^6}\right).$$

The idea is to prove, first, that the potential after one iteration is bounded by some polynomial and, second, that the potential decreases by some inverse polynomial amount in every iteration (or, more precisely, in every sequence of a few consecutive iterations). To do this, we prove upper bounds on the probability that the minimal improvement is small. The main challenge is the huge number of up to $(k^2n)^{kd}$ possible clusterings. Each of these clusterings yields a potential iteration of k-means, and a simple union bound over all of them is too weak to yield a polynomial bound.

To prove the bound of $\mathrm{poly}(n^{\sqrt{k}}, 1/\sigma)$ [Manthey and Röglin 2009a], a union bound was taken over the $(k^2n)^{kd}$ clusterings. This is already a technical challenge as the set of possible Voronoi-based clusterings is fixed only after the points are fixed. To show a polynomial bound, we reduce the number of cases in the union bound by introducing the notion of *transition blueprints*. Basically, every iteration of k-means can be described by a transition blueprint. The blueprint describes the iteration only roughly, so that several iterations are described by the same blueprint. Intuitively, iterations with the same transition blueprint are correlated in the sense that either all of them make a small improvement or none of them do. This dramatically reduces the number of cases that have to be considered in the union bound. On the other hand, the description conveyed by a blueprint is still precise enough to allow us to bound the probability that any iteration described by it makes a small improvement.

We distinguish between several types of iterations, based on which clusters exchange how many points. Sections 4.1 to 4.5 deal with some special cases of iterations that need separate analyses. After that, we analyze the general case (Section 4.6). The difficulty in this analysis is to show that every transition blueprint contains "enough randomness". We need to show that this randomness allows for sufficiently tight upper

bounds on the probability that the improvement obtained from any iteration corresponding to the blueprint is small.

Finally, we put the six sections together to prove that k-means has polynomial smoothed running time (Section 4.7) for $d \geq 2$. For completeness, we include a couple of technical lemmas and proofs from an earlier smoothed analysis [Manthey and Röglin 2009a]. This includes a proof that k-means has polynomial smoothed running time also for $d = 1$ in Section 5, which completes the proof of Theorem 1.1.

2. PRELIMINARIES

For a finite set $X \subseteq \mathbb{R}^d$, let $\mathrm{cm}(X) = \frac{1}{|X|} \sum_{x \in X} x$ be the center of mass of the set X. If $H \subseteq \mathbb{R}^d$ is a hyperplane and $x \in \mathbb{R}^d$ is a single point, then $\mathrm{dist}(x, H) = \min\{\|x - y\| \mid y \in H\}$ denotes the distance of the point x to the hyperplane H. Furthermore, for $a \in \mathbb{N}$, we denote by $[a]$ the set $\{1, 2, \ldots, a\}$.

For our smoothed analysis, an adversary specifies an instance $\mathcal{X}' \subseteq [0,1]^d$ of n points. Then each point $x' \in \mathcal{X}'$ is perturbed by adding an independent d-dimensional Gaussian random vector with standard deviation σ to x' to obtain the data point x. These perturbed points form the input set $\mathcal{X}$. For convenience we assume that $\sigma \leq 1$. This assumption is without loss of generality: The number of iterations that k-means needs is invariant under scaling of the point set $\mathcal{X}$. If $\sigma > 1$, 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 $\mathrm{poly}(n) \leq \mathrm{poly}(n, 1/\sigma)$. Additionally we assume $k \leq n$ and $d \leq n$: First, $k \leq n$ is satisfied after the first iteration since at most n clusters can contain a point. Second, k-means is known to have polynomial smoothed complexity for $d \in \Omega(n/\log n)$ [Arthur and Vassilvitskii 2006]. The restriction of the adversarial points to be in $[0,1]^d$ is necessary as, otherwise, the adversary can diminish the effect of the perturbation by placing all points far apart from each other. Another way to cope with this problem is to state the bounds in terms of the diameter of the adversarial instance [Arthur and Vassilvitskii 2009]. However, to avoid having another parameter, we have chosen the former model.

Throughout the following, we assume that the perturbed point set $\mathcal{X}$ is contained in some hypercube of side-length D, i.e., $\mathcal{X} \subseteq [-D/2, D/2]^d = \mathcal{D}$. We choose D such that the probability of $\mathcal{X} \not\subseteq \mathcal{D}$ is bounded from above by n^{-3kd}. Then, as the worst-case number of iterations is bounded by n^{3kd} [Inaba et al. 2000], the event $\mathcal{X} \not\subseteq \mathcal{D}$ contributes only an insignificant additive term of $+1$ to the expected number of iterations, which we ignore in the following.

Since Gaussian random vectors are heavily concentrated around their mean and all means are in $[0,1]^d$, we can choose $D = \sqrt{90kd\ln(n)}$ to obtain the desired failure probability for $\mathcal{X} \not\subseteq \mathcal{D}$, as shown by the following calculation, in which Z denotes a one-dimensional Gaussian random variable with mean 0 and standard deviation 1:

$$\begin{aligned}\Pr\big[\mathcal{X} \not\subseteq \mathcal{D}\big] &\leq nd \cdot \Pr\big[|Z| \geq D/2 - 1\big] \leq 2nd \cdot \Pr\big[Z \geq D/3\big] \\ &\leq \frac{2nd}{\sqrt{2\pi}} \cdot \exp(-D^2/18) \leq n^2 \cdot \exp(-D^2/18) \leq n^{-3kd}\,,\end{aligned}$$

where we used the tail bound $\Pr[Z \geq z] \leq \frac{\exp(-z^2/2)}{z\sqrt{2\pi}}$ for Gaussians [Durrett 1991].

For our smoothed analysis, we use essentially three properties of Gaussian random variables. Let X be a d-dimensional Gaussian random variable with standard deviation σ. First, the probability that X assumes a value in any fixed ball of radius ε is

at most $(\varepsilon/\sigma)^d$. (This is a very rough estimate, obtained from multiplying the maximum density of a Gaussian with an upper bound for the volume of a d-dimensional ball. A sharper bound, however, would not improve the running time bound significantly.) Second, let $b_1, \ldots, b_{d'} \in \mathbb{R}^d$ be orthonormal vectors for some $d' \leq d$. Then the vector $(b_1 \cdot X, \ldots, b_{d'} \cdot X) \in \mathbb{R}^{d'}$ is a d'-dimensional Gaussian random variable with the same standard deviation σ. Third, let H be any hyperplane. Then the probability that a Gaussian random variable assumes a value that is within a distance of at most ε from H is bounded by ε/σ. This follows also from the first two properties if we choose $d' = 1$ and b_1 to be the normal vector of H.

We will often upper-bound various probabilities, and it will be convenient to reduce the exponents in these bounds. Under certain conditions, this can be done safely regardless of whether the base is smaller or larger than 1.

FACT 2.1. *Let $p \in [0,1]$ be a probability, and let $A, c, b, e,$ and e' be non-negative real numbers satisfying $c \geq 1$ and $e \geq e'$. If $p \leq A + c \cdot b^e$, then it is also true that $p \leq A + c \cdot b^{e'}$.*

PROOF. If b is at least 1, then $A + c \cdot b^{e'} \geq 1$ and it is trivially true that $p \leq A + c \cdot b^{e'}$. Otherwise, $b^e \leq b^{e'}$, and the result follows. □

2.1. Potential Drop in an Iteration of k-Means

During an iteration of the k-means method there are two possible events that can lead to a significant potential drop: either one cluster center moves significantly, or a data point is reassigned from one cluster to another and this point has a significant distance from the bisector of the clusters (the bisector is the hyperplane that bisects the two cluster centers). In the following we quantify the potential drops caused by these events.

The potential drop caused by reassigning a data point x from one cluster to another can be expressed in terms of the distance of x from the bisector of the two cluster centers and the distance between these two centers.

LEMMA 2.2. *Assume that, in an iteration of k-means, a point $x \in \mathcal{X}$ switches from $\mathcal{C}_i$ to $\mathcal{C}_j$. Let c_i and c_j be the centers of these clusters, and let H be their bisector. Then reassigning x decreases the potential by $2 \cdot \|c_i - c_j\| \cdot \operatorname{dist}(x, H)$.*

PROOF. The potential decreases by $\|c_i - x\|^2 - \|c_j - x\|^2 = (2x - c_i - c_j) \cdot (c_j - c_i)$. Let v be the unit vector in the $c_j - c_i$ direction. Then $(2x - c_i - c_j) \cdot v = 2 \operatorname{dist}(x, H)$ because v is orthogonal to H. The observation $c_j - c_i = \|c_i - c_j\| \cdot v$ completes the proof. □

The following lemma, which also follows from basic linear algebra, reveals how moving a cluster center to the center of mass decreases the potential.

LEMMA 2.3 ((KANUNGO ET AL. [2004])). *Assume that the center of a cluster $\mathcal{C}$ moves from c to $\operatorname{cm}(\mathcal{C})$ during an iteration of k-means, and let $|\mathcal{C}|$ denote the number of points in $\mathcal{C}$ when the movement occurs. Then the potential decreases by $|\mathcal{C}| \cdot \|c - \operatorname{cm}(C)\|^2$.*

2.2. The Distance between Centers

As the distance between two cluster centers plays an important role in Lemma 2.2, we analyze how close together two simultaneous centers can be during the execution of k-means. This has already been analyzed implicitly [Manthey and Röglin 2009a, Proof of Lemma 3.2], but the variant below gives stronger bounds. From now on, when we refer to a k-means iteration, we will always mean an iteration *after the first one*. By restricting ourselves to this case, we ensure that the centers at the beginning of the iteration are the centers of mass of actual clusters, as opposed to the arbitrary choices that were used to seed k-means.

Definition 2.4. Let δ_ε denote the minimum distance between two cluster centers at the beginning of a k-means iteration in which (1) the potential Ψ drops by at most ε, and (2) at least one data point switches between the clusters corresponding to these centers.

LEMMA 2.5. *Fix real numbers $Y \geq 1$ and $e \geq 2$. Then, for any $\varepsilon \in [0, 1]$,*

$$\Pr\left[\delta_\varepsilon \leq Y\varepsilon^{1/e}\right] \leq \varepsilon \cdot \left(\frac{O(1) \cdot n^5 Y}{\sigma}\right)^e.$$

PROOF. We first define two events $\mathcal{E}_1$ and $\mathcal{E}_2$, and we argue that $\delta_\varepsilon \leq Y\varepsilon^{1/e}$ can only occur if one of these events occurs. Then we bound the probability of $\mathcal{E}_1 \cup \mathcal{E}_2$. Let $t = 3d + \lfloor e \rfloor$. The events are defined as follows:

$\mathcal{E}_1$. There exist a subset $T \subseteq \mathcal{X}$ with $|T| = t + 1$ and hyperplanes $H \subseteq \mathbb{R}^d$ and $H_0 \subseteq \mathbb{R}^d$ such that, for every $x \in T$, $\operatorname{dist}(x, H) \leq 3nY\varepsilon^{1/e}$ or $\operatorname{dist}(x, H_0) \leq 3nY\varepsilon^{1/e}$.
$\mathcal{E}_2$. There exist subsets $A \subseteq \mathcal{X}$ and $A' \subseteq \mathcal{X}$ with $A \neq A'$ and $|A \cup A'| \leq t$ such that $\|\operatorname{cm}(A) - \operatorname{cm}(A')\| \leq \sqrt{\varepsilon}$.

Consider an arbitrary k-means iteration I that results in a potential drop of at most ε, and let I_0 denote the previous iteration. Also consider an arbitrary pair of clusters that exchange at least one data point during I. We define the following:

— Let a_0 and b_0 denote the centers of the two clusters at the beginning of iteration I_0 and let H_0 denote the hyperplane bisecting a_0 and b_0.
— Let A and B denote the set of data points in the two clusters at the beginning of iteration I. Note that H_0 splits A and B.
— Let a and b denote the centers of the two clusters at the beginning of iteration I, and let H denote the hyperplane bisecting a and b. Note that $a = \operatorname{cm}(A)$ and $b = \operatorname{cm}(B)$.
— Let A' and B' denote the set of data points in the two clusters at the end of iteration I. Note that H splits A' and B'.
— Let a' and b' denote the centers of the two clusters at the end of iteration I. Note that $a' = \operatorname{cm}(A')$ and $b' = \operatorname{cm}(B')$.

Now suppose we have $\|a - b\| \leq Y\varepsilon^{1/e}$.

First we consider the case $|A' \cup A| \geq t + 1$. We claim that every point in A must be within a distance of $nY\varepsilon^{1/e}$ of H_0. Indeed, if this were not true, then since H_0 splits A and B, and since $a = \operatorname{cm}(A)$ and $b = \operatorname{cm}(B)$, we would have $\|a - b\| \geq \operatorname{dist}(a, H_0) > \frac{nY\varepsilon^{1/e}}{|A|} \geq Y\varepsilon^{1/e}$, giving a contradiction. Furthermore, as I results in a potential drop of at most ε, Lemma 2.3 implies that $\|a' - a\|, \|b' - b\| \leq \sqrt{\varepsilon}$, and therefore,

$$\|a' - b'\| \leq \|a' - a\| + \|a - b\| + \|b - b'\| \leq Y\varepsilon^{1/e} + 2\sqrt{\varepsilon} \leq 3Y\varepsilon^{1/e}.$$

In particular, we can repeat the above argument to see that every point in A' must be within a distance of $3nY\varepsilon^{1/e}$ of H. This means that there are two hyperplanes such that every point in $A \cup A'$ is within a distance of $3nY\varepsilon^{1/e}$ of at least one of these hyperplanes. Hence, this case can only occur if $\mathcal{E}_1$ occurs.

Next we consider the case $|A' \cup A| \leq t$. We must have $A' \neq A$ since some point is exchanged between clusters A and B during iteration I. Lemma 2.3 implies that $\|\operatorname{cm}(A') - \operatorname{cm}(A)\| \leq \sqrt{\varepsilon}$ must hold for iteration I. Otherwise, I would result in a potential drop of more than ε. Hence, this case can only occur if $\mathcal{E}_2$ occurs.

It remains to analyze the probability of $\mathcal{E}_1 \cup \mathcal{E}_2$. Following the arguments by Arthur and Vassilvitskii [2009, Proposition 5.9], we obtain that the probability of $\mathcal{E}_1$ is at most

$$n^{t+1} \cdot \left(\frac{12dnY\varepsilon^{1/e}}{\sigma}\right)^{t+1-2d} = n^{3d+\lfloor e\rfloor+1} \cdot \left(\frac{12dnY\varepsilon^{1/e}}{\sigma}\right)^{d+\lfloor e\rfloor+1} \leq \left(\frac{12dn^4Y\varepsilon^{1/e}}{\sigma}\right)^{d+\lfloor e\rfloor+1}. \quad (2)$$

This bound can be proven as follows: Arthur and Vassilvitskii [2009, Lemma 5.8] have shown that we can approximate arbitrary hyperplanes H and H_0 by hyperplanes $\tilde{H}$ and $\tilde{H}_0$ that pass through d points of $\mathcal{X}$ exactly such that any point $x \in \mathcal{X}$ within a distance of L of H or H_0 has a distance of at most $2dL$ from $\tilde{H}$ or $\tilde{H}_0$, respectively. A union bound over all choices for these $2d$ points and the remaining $t+1-2d$ points yields the term n^{t+1}. Once $\tilde{H}$ and $\tilde{H}_0$ are fixed, the probability that a random point is within distance $2dL$ of at least one of the hyperplanes is bounded from above by $4dL/\sigma$. Taking into account that the remaining $t+1-2d$ points are independent Gaussians yields a final bound of $n^{t+1}(4dL/\sigma)^{t+1-2d}$ with $L = 3nY\varepsilon^{1/e}$.

Next we analyze the probability of $\mathcal{E}_2$. Consider some fixed A and A', and let x_0 be a data point in the symmetric difference of A and A'. Then $\mathrm{cm}(A') - \mathrm{cm}(A)$ can be written as $\sum_{x\in X} c_x \cdot x$ for constants c_x with $|c_{x_0}| \geq \frac{1}{n}$. We consider only the randomness in the perturbed position of x_0 and allow all other points in X to be fixed adversarially. Then $\mathrm{cm}(A') - \mathrm{cm}(A)$ follows a normal distribution with standard deviation at least $\frac{\sigma}{n}$, and hence $\|\mathrm{cm}(A') - \mathrm{cm}(A)\| \leq \sqrt{\varepsilon}$ with a probability of at most $(n\sqrt{\varepsilon}/\sigma)^d$. The total number of possible sets A and A' is bounded by $(4n)^t$: we choose t candidate points to be in $A \cup A'$ and then for each point, we choose which set(s) it belongs to. Taking a union bound over all possible choices, we see that $\mathcal{E}_2$ can occur with a probability of at most

$$(4n)^t \cdot \left(\frac{n\sqrt{\varepsilon}}{\sigma}\right)^d = \left(\frac{4^{3+\lfloor e\rfloor/d} n^{4+\lfloor e\rfloor/d}\sqrt{\varepsilon}}{\sigma}\right)^d. \quad (3)$$

Combining equations (2) and (3), we have

$$\Pr[\delta_\varepsilon \leq Y\varepsilon^{1/e}] \leq \Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2] \leq \left(\frac{12dn^4Y\varepsilon^{1/e}}{\sigma}\right)^{d+\lfloor e\rfloor+1} + \left(\frac{4^{3+\lfloor e\rfloor/d} n^{4+\lfloor e\rfloor/d}\sqrt{\varepsilon}}{\sigma}\right)^d.$$

Note that $d + \lfloor e\rfloor + 1 \geq e$ and $d \geq 2$, so we can reduce exponents according to Fact 2.1:

$$\begin{aligned}
\Pr[\delta_\varepsilon \leq Y\varepsilon^{1/e}] &\leq \left(\frac{12dn^4Y\varepsilon^{1/e}}{\sigma}\right)^e + \left(\frac{4^{3+\lfloor e\rfloor/d} n^{4+\lfloor e\rfloor/d}\sqrt{\varepsilon}}{\sigma}\right)^2 \\
&\leq \varepsilon \cdot \left(\frac{12dn^4Y}{\sigma}\right)^e + \varepsilon \cdot \left(\frac{4^{6+e}n^{8+e}}{\sigma^2}\right) && \text{(since } d \geq 2\text{)} \\
&\leq \varepsilon \cdot \left(\frac{12n^5Y}{\sigma}\right)^e + \varepsilon \cdot \left(\frac{4^4n^5}{\sigma}\right)^e && \text{(since } d \leq n,\ e \geq 2 \text{ and } \sigma \leq 1\text{)} \\
&\leq \varepsilon \cdot \left(\frac{O(1)\cdot n^5Y}{\sigma}\right)^e.
\end{aligned}$$

□

3. TRANSITION BLUEPRINTS

Our smoothed analysis of k-means is based on the potential function Ψ (see (1)). If $\mathcal{X} \subseteq \mathcal{D}$, then after the first iteration, Ψ will always be bounded from above by a polynomial in n and $1/\sigma$. (The reason for this is simple: After the first iteration, all cluster centers, which are convex combinations of points, are also in $\mathcal{D} = [-D/2, D/2]^d$. Thus, the maximum distance from a point to its closest center is at most $D\sqrt{d}$. Since $D = \sqrt{90kd\ln(n)}$, the sum of squared distances is bounded by a polynomial.) Therefore, k-means terminates quickly if we can lower-bound the drop in Ψ during each iteration. So what must happen for a k-means iteration to result in a small potential drop? Recall that any iteration consists of two distinct phases: assigning points to clusters, and then recomputing center positions. Furthermore, each phase can only decrease the potential. According to Lemmas 2.2 and 2.3, an iteration can only result in a small potential drop if none of the centers move significantly and no point is reassigned that has a significant distance to the corresponding bisector. The previous analyses [Arthur and Vassilvitskii 2009; Manthey and Röglin 2009a] essentially use a union bound over all possible iterations to show that it is unlikely that there is an iteration in which none of these events happens. Thus, with high probability, we get a significant potential drop in every iteration. As the number of possible iterations can only be bounded by $(k^2n)^{kd}$, these union bounds are quite wasteful and yield only super-polynomial bounds.

We resolve this problem by introducing the notion of *transition blueprints*. Such a blueprint is a description of an iteration of k-means that *almost* uniquely determines everything that happens during the iteration. In particular, one blueprint can simultaneously cover many similar iterations, which will dramatically reduce the number of cases that have to be considered in the union bound. We begin with the notion of a transition graph, which is part of a transition blueprint.

Definition 3.1. Given a k-means iteration, we define its *transition graph* to be the labeled, directed multigraph with one vertex for each cluster, and with one edge $(\mathcal{C}_i, \mathcal{C}_j)$ with label x for each data point x switching from cluster $\mathcal{C}_i$ to cluster $\mathcal{C}_j$.

We define a vertex in a transition graph to be *balanced* if its in-degree is equal to its out-degree. Similarly, a cluster is balanced during a k-means iteration if the corresponding vertex in the transition graph is balanced.

To make the full blueprint, we also require information on approximate positions of cluster centers. We will see below that for an unbalanced cluster this information can be deduced from the data points that change to or from this cluster. For balanced clusters we turn to brute force: We tile the hypercube $\mathcal{D}$ with a lattice L_ε, where consecutive points are are at a distance of $\sqrt{n\varepsilon/d}$ from each other, and choose one point from L_ε for every balanced cluster.

Definition 3.2. An (m, b, ε)-*transition blueprint* $\mathcal{B}$ consists of a weakly connected transition graph G with m edges and b balanced clusters, and one lattice point in L_ε for each balanced cluster in the graph. A k-means iteration is said to *follow* $\mathcal{B}$ if G is a connected component of the iteration's transition graph and if the lattice point selected for each balanced cluster is within a distance of at most $\sqrt{n\varepsilon}$ of the cluster's actual center position.

If $\mathcal{X} \subseteq \mathcal{D}$, then by the Pythagorean theorem, every cluster center must be within distance $\sqrt{n\varepsilon}$ of some point in L_ε. Therefore, every k-means iteration follows at least one transition blueprint.

As m and b grow, the number of valid (m, b, ε)-transition blueprints grows exponentially, but the probability of failure that we will prove in the following section decreases

equally fast, making the union bound possible. This is what we gain by studying transition blueprints rather than every possible configuration separately.

For an unbalanced cluster $\mathcal{C}$ that gains the points $A \subseteq \mathcal{X}$ and loses the points $B \subseteq \mathcal{X}$ during the considered iteration, the *approximate center of* $\mathcal{C}$ is defined as

$$\frac{|B|\operatorname{cm}(B) - |A|\operatorname{cm}(A)}{|B| - |A|}.$$

If $\mathcal{C}$ is balanced, then the approximate center of $\mathcal{C}$ is the lattice point specified in the transition blueprint. The *approximate bisector of* $\mathcal{C}_i$ *and* $\mathcal{C}_j$ is the bisector of the approximate centers of $\mathcal{C}_i$ and $\mathcal{C}_j$. Now consider a data point x switching from some cluster $\mathcal{C}_i$ to some other cluster $\mathcal{C}_j$. We say the *approximate bisector corresponding to* x is the hyperplane bisecting the approximate centers of $\mathcal{C}_i$ and $\mathcal{C}_j$. Unfortunately, this definition applies only if $\mathcal{C}_i$ and $\mathcal{C}_j$ have distinct approximate centers, which is not necessarily the case (even after the random perturbation and even if both clusters are unbalanced). We will call a blueprint *non-degenerate* if the approximate bisector is in fact well defined for each data point that switches clusters. The intuition is that, if one actual cluster center is far away from its corresponding approximate center, then during the considered iteration the cluster center must move significantly, which causes a potential drop according to Lemma 2.3. Otherwise, the approximate bisectors are close to the actual bisectors and we can show that it is unlikely that all points that change their assignment are close to their corresponding approximate bisectors. This will yield a potential drop according to Lemma 2.2.

The following lemma formalizes what we mentioned above: If the center of an unbalanced cluster is far away from its approximate center, then this causes a potential drop in the corresponding iteration.

LEMMA 3.3. *Consider an iteration of k-means where a cluster $\mathcal{C}$ gains a set A of points and loses a set B of points with $|A| \neq |B|$. If $\left\|\operatorname{cm}(\mathcal{C}) - \frac{|B|\operatorname{cm}(B) - |A|\operatorname{cm}(A)}{|B| - |A|}\right\| \geq \sqrt{n\varepsilon}$, then the potential decreases by at least ε.*

PROOF. Let $\mathcal{C}' = (\mathcal{C} \setminus B) \cup A$ denote the cluster after the iteration. According to Lemma 2.3, the potential drops in the considered iteration by at least

$$\begin{aligned}
&|\mathcal{C}'| \cdot \|\operatorname{cm}(\mathcal{C}') - \operatorname{cm}(\mathcal{C})\|^2 \\
&= (|\mathcal{C}| + |A| - |B|) \left\| \frac{|\mathcal{C}|\operatorname{cm}(\mathcal{C}) + |A|\operatorname{cm}(A) - |B|\operatorname{cm}(B)}{|\mathcal{C}| + |A| - |B|} - \operatorname{cm}(\mathcal{C}) \right\|^2 \\
&= \frac{\big||B| - |A|\big|^2}{|\mathcal{C}| + |A| - |B|} \left\| \operatorname{cm}(\mathcal{C}) - \frac{|B|\operatorname{cm}(B) - |A|\operatorname{cm}(A)}{|B| - |A|} \right\|^2 \geq \frac{(\sqrt{n\varepsilon})^2}{n}.
\end{aligned}$$

□

Now we show that we get a significant potential drop if a point that changes its assignment is far from its corresponding approximate bisector. Formally, we will be studying the following quantity $\Lambda(\mathcal{B})$.

Definition 3.4. Fix a non-degenerate (m, b, ε)-transition blueprint $\mathcal{B}$. Let $\Lambda(\mathcal{B})$ denote the maximum distance between a data point in the transition graph of $\mathcal{B}$ and its corresponding approximate bisector.

LEMMA 3.5. *Fix $\varepsilon \in [0, 1]$ and a non-degenerate (m, b, ε)-transition blueprint $\mathcal{B}$. If there exists an iteration that follows $\mathcal{B}$ and that results in a potential drop of at most ε, then*

$$\delta_\varepsilon \cdot \Lambda(\mathcal{B}) \leq 6D\sqrt{nd\varepsilon}.$$

PROOF. Fix an iteration that follows $\mathcal{B}$ and that results in a potential drop of at most ε. Consider a data point x that switches between clusters $\mathcal{C}_i$ and $\mathcal{C}_j$ during this iteration. Let p and q denote the center positions of these two clusters at the beginning of the iteration, and let p' and q' denote the approximate center positions of the clusters. Also let H denote the hyperplane bisecting p and q, and let H' denote the hyperplane bisecting p' and q'.

We begin by bounding the divergence between the hyperplanes H and H'.

CLAIM 3.6. *Let u and v be arbitrary points on H. Then we have* $\operatorname{dist}(v, H') - \operatorname{dist}(u, H') \leq \frac{4\sqrt{n\varepsilon}}{\delta_\varepsilon} \cdot \|v - u\|$.

PROOF. Let θ denote the angle between the normal vectors of the hyperplanes H and H'. We move the vector $\overrightarrow{p'q'}$ to become $\overrightarrow{pq''}$ for some point $q'' = p + q' - p'$, which ensures $\angle qpq'' = \theta$. Note that $\|q'' - q\| \leq \|q'' - q'\| + \|q' - q\| = \|p - p'\| + \|q' - q\| \leq 2\sqrt{n\varepsilon}$ by Lemma 3.3.

Let r be the point where the bisector of the angle $\angle qpq''$ hits the segment $\overline{qq''}$. By the sine law, we have

$$\begin{aligned} \sin\left(\frac{\theta}{2}\right) &= \sin(\angle prq) \cdot \frac{\|r - q\|}{\|p - q\|} \\ &\leq \frac{\|q'' - q\|}{\|p - q\|} \leq \frac{2\sqrt{n\varepsilon}}{\delta_\varepsilon}. \end{aligned}$$

Let y and y' be unit vectors in the direction $\overrightarrow{pq}$ and $\overrightarrow{p'q'}$, respectively, and let z be an arbitrary point on H'. Then,

$$\begin{aligned} \operatorname{dist}(v, H') - \operatorname{dist}(u, H') &= |(v - z) \cdot y'| - |(u - z) \cdot y'| \\ &\leq |(v - u) \cdot y'| && \text{(by the triangle inequality)} \\ &= |(v - u) \cdot y + (v - u) \cdot (y' - y)| \\ &= |(v - u) \cdot (y' - y)| && \text{(since } u, v \in H\text{)} \\ &\leq \|v - u\| \cdot \|y' - y\|. \end{aligned}$$

Now we consider the isosceles triangle formed by the normal vectors y and y'. The angle between y and y' is θ. Using the sine law again, we get

$$\|y' - y\| = 2 \cdot \sin\left(\frac{\theta}{2}\right) \leq \frac{4\sqrt{n\varepsilon}}{\delta_\varepsilon},$$

and the claim follows. □

We now continue the proof of Lemma 3.5. Let h denote the foot of the perpendicular from x to H, and let $m = \frac{p+q}{2}$. Then,

$$\begin{aligned} \operatorname{dist}(x, H') &\leq \|x - h\| + \operatorname{dist}(h, H') \\ &= \operatorname{dist}(x, H) + \operatorname{dist}(m, H') + \operatorname{dist}(h, H') - \operatorname{dist}(m, H') \\ &\leq \operatorname{dist}(x, H) + \operatorname{dist}(m, H') + \frac{4\sqrt{n\varepsilon}}{\delta_\varepsilon} \cdot \|h - m\|, \end{aligned} \tag{4}$$

where the last inequality follows from Claim 3.6. By Lemma 2.2, we know that the total potential drop during the iteration is at least $2 \cdot \|p - q\| \cdot \operatorname{dist}(x, H)$. However, we assumed that this drop was at most ε, so we therefore have $\operatorname{dist}(x, H) \leq \frac{\varepsilon}{2\delta_\varepsilon}$. Also, by

Lemma 3.3,

$$\operatorname{dist}(m, H') \leq \left\| \frac{p'+q'}{2} - \frac{p+q}{2} \right\| \leq \frac{1}{2} \cdot \|p'-p\| + \frac{1}{2} \cdot \|q'-q\| \leq \sqrt{n\varepsilon}.$$

Furthermore, $\|h-m\| \leq \|m-x\| \leq D\sqrt{d}$ since $h-m$ is perpendicular to $x-h$ and $m-x$ lies in the hypercube $[-D, D]^d$. Plugging these bounds into equation (4), we have

$$\begin{aligned}
\operatorname{dist}(x, H') &\leq \frac{\varepsilon}{2\delta_\varepsilon} + \sqrt{n\varepsilon} + \frac{4D\sqrt{nd\varepsilon}}{\delta_\varepsilon} \\
&\leq \sqrt{n\varepsilon} + \frac{5D\sqrt{nd\varepsilon}}{\delta_\varepsilon} && \text{since } \varepsilon \leq 1 \\
&\leq \frac{6D\sqrt{nd\varepsilon}}{\delta_\varepsilon} && \text{since } \delta_\varepsilon \leq \|p-q\| \leq D\sqrt{d}.
\end{aligned}$$

This bound holds for all data points x that switch clusters, so the lemma follows.

4. ANALYSIS OF TRANSITION BLUEPRINTS

In this section, we analyze transition blueprints for $d \geq 2$. The special case of $d = 1$ is deferred to Section 5. The first five cases deal with special types of blueprints that require separate attention and do not fit into the general framework of case six. The sixth and most involved case deals with general blueprints. Overall, the transition blueprints are divided as follows:

Section 4.1. Transition blueprints in which one balanced cluster gains and loses at least 1 and at most $z_1 d$ points, for a constant z_1.
Section 4.2. Transition blueprints with a node of degree 1, that is, a cluster that either gains exactly one point or loses exactly one point.
Section 4.3. Non-degenerate transition blueprints with at least three disjoint pairs of adjacent unbalanced nodes of degree two.
Section 4.4. Transition blueprints with constant maximum degree. This means that every cluster gains or loses at most a constant number of points.
Section 4.5. Degenerate transition blueprints, that is, blueprints in which the approximate centers of some pair of clusters that exchange at least one point coincide.
Section 4.6. General transition blueprints that fall under none of the above categories.

In the following sections, we define and analyze the random variables $\Delta_1, \ldots, \Delta_6$. The random variable Δ_i is the smallest decrease of the potential caused by any iteration that follows a transition blueprint dealt with in Section 4.i. The only exception is Δ_4, which describes the smallest improvement caused by a sequence of four consecutive iterations dealt with in Section 4.4.

To finally bound the running-time in Section 4.7, let Δ be the smallest decrease of the potential Ψ caused by any sequence of four consecutive iterations. Then we can bound Δ from below by the minimum of the Δ_i.

When analyzing these random variables, we will ignore the case that a cluster can lose all its points in one iteration. If this happens, then k-means continues with one cluster less, which can happen only k times. Since the potential Ψ does not increase even in this case, this gives only an additive term of $4k$ to our analysis. (The four comes from the fact that we consider sequences of four consecutive iterations in the definition of Δ.)

In the lemmas in this section, we do not specify the parameters m and b when talking about transition blueprints. When we say *an iteration follows a blueprint with some*

property P, we mean that there are parameters m and b such that the iteration follows an (m, b, ε)-transition blueprint with property P, where ε will be clear from the context.

4.1. Balanced Clusters of Small Degree

LEMMA 4.1. *Fix $\varepsilon \geq 0$ and a constant $z_1 \in \mathbb{N}$. Let Δ_1 denote the smallest improvement made by any iteration that follows a blueprint with a balanced node of in- and outdegree at least 1 and at most $z_1 d$. Then,*

$$\Pr\big[\Delta_1 \leq \varepsilon\big] \leq \varepsilon \cdot \left(\frac{n^{4z_1+1}}{\sigma^2}\right).$$

PROOF. We denote the balanced cluster of in- and outdegree at most $z_1 d$ by $\mathcal{C}$. By Lemma 2.3, if the center of $\mathcal{C}$ moves by at least $\sqrt{\varepsilon/|\mathcal{C}|}$, then the potential drops by at least ε. Let A and B with $|A| = |B| \leq z_1 d$ be the sets of data points corresponding to the incoming and outgoing edges of $\mathcal{C}$, respectively. If $|A|\operatorname{cm}(A)$ and $|B|\operatorname{cm}(B)$ differ by at least $\sqrt{n\varepsilon} \geq \sqrt{|\mathcal{C}|\varepsilon}$, then the cluster center moves by at least $\sqrt{\varepsilon/|\mathcal{C}|}$ as shown by the following reasoning: Let c be the center of mass of the points that belong to $\mathcal{C}$ at the beginning of the iteration and remain in $\mathcal{C}$ during the iteration. Then the center of mass of $\mathcal{C}$ moves from $\frac{(|\mathcal{C}|-|A|)c+|A|\operatorname{cm}(A)}{|\mathcal{C}|}$ to $\frac{(|\mathcal{C}|-|A|)c+|B|\operatorname{cm}(B)}{|\mathcal{C}|}$. Since $|A| = |B|$, these two locations differ by

$$\left\| \frac{|B|\operatorname{cm}(B) - |A|\operatorname{cm}(A)}{|\mathcal{C}|} \right\| \geq \sqrt{\varepsilon/|\mathcal{C}|}\,.$$

We exploit only the randomness of a single point $x \in A \setminus B$. Thus, we let an adversary fix all points in $B \cup A \setminus \{x\}$ arbitrarily. In order for $|A|\operatorname{cm}(A)$ to be $\sqrt{n\varepsilon}$-close to $|B|\operatorname{cm}(B)$, this point x must fall into a hyperball of radius $\sqrt{n\varepsilon}$. This happens with a probability of at most $(\sqrt{n\varepsilon}/\sigma)^d$.

Now we apply a union bound over all possible choices of A and B. We can assume that both A and B contain exactly $z_1 d$ points. Otherwise, we can pad them by adding the same points to both A and B. This does not affect the analysis since we only consider the point x. Hence, the number of choices is bounded by $n^{2z_1 d}$, and we get

$$\begin{aligned}\Pr\big[\Delta_1 \leq \varepsilon\big] &\leq \Pr\big[\exists A, B, |A| = |B| = z_1 d\colon \big\||A|\operatorname{cm}(A) - |B|\operatorname{cm}(B)\big\| \leq \sqrt{n\varepsilon}\big] \\ &\leq n^{2z_1 d}\left(\frac{\sqrt{n\varepsilon}}{\sigma}\right)^d \leq \left(\frac{n^{2z_1+\frac{1}{2}}\sqrt{\varepsilon}}{\sigma}\right)^d.\end{aligned}$$

Using Fact 2.1 and $d \geq 2$ concludes the proof. □

4.2. Nodes of Degree One

LEMMA 4.2. *Fix $\varepsilon \in [0, 1]$. Let Δ_2 denote the smallest improvement made by any iteration that follows a blueprint with a node of degree* 1*. Then,*

$$\Pr\big[\Delta_2 \leq \varepsilon\big] \leq \varepsilon \cdot \left(\frac{O(1) \cdot n^{11}}{\sigma^2}\right).$$

PROOF. Assume that a point x switches from cluster $\mathcal{C}_1$ to cluster $\mathcal{C}_2$, and let c_1 and c_2 denote the positions of the cluster centers at the beginning of the iteration. Let ν be the distance between c_1 and c_2. Then c_2 has a distance of $\nu/2$ from the bisector of c_1 and c_2, and the point x is on the same side of the bisector as c_2.

If $\mathcal{C}_1$ has only one edge, then the center of cluster $\mathcal{C}_1$ moves during this iteration by at least $\frac{\nu}{2(|\mathcal{C}_1|-1)}$, where $|\mathcal{C}_1|$ denotes the number of points belonging to $\mathcal{C}_1$ at the

beginning of the iteration: the point x has a distance of at least $\nu/2$ from c_1, which yields a movement of

$$\left\| c_1 - \frac{c_1|\mathcal{C}_1| - x}{|\mathcal{C}_1| - 1} \right\| = \left\| \frac{c_1 - x}{|\mathcal{C}_1| - 1} \right\| \geq \frac{\nu}{2(|\mathcal{C}_1| - 1)} .$$

Hence, the potential drops by at least $(|\mathcal{C}_1| - 1)\big(\frac{\nu}{2|\mathcal{C}_1|-2}\big)^2 \geq \frac{\nu^2}{4|\mathcal{C}_1|} \geq \frac{\nu^2}{4n}$.

If $\mathcal{C}_2$ has only one edge, then let α be the distance of the point x to the bisector of c_1 and c_2. The potential drop is caused by two effects. First, by reassigning the point, we get a potential drop of $2\alpha\nu$. Second, $\|x - c_2\| \geq |\nu/2 - \alpha|$. Thus, $\mathcal{C}_2$ moves by at least

$$\left\| c_2 - \frac{c_2|\mathcal{C}_2| + x}{|\mathcal{C}_2| + 1} \right\| \geq \left\| \frac{c_2 - x}{|\mathcal{C}_2| + 1|} \right\| \geq \frac{|\nu/2 - \alpha|}{|\mathcal{C}_2| + 1} .$$

This causes a potential drop of at least $(|\mathcal{C}_2|+1)(\nu/2-\alpha)^2/(|\mathcal{C}_2|+1)^2 = (\nu/2-\alpha)^2/(|\mathcal{C}_2|+1) \geq (\nu/2 - \alpha)^2/n$. Combining the two estimates, the potential drops by at least

$$2\alpha\nu + \frac{(\nu/2 - \alpha)^2}{n} \geq \frac{(\nu/2 + \alpha)^2}{n} \geq \frac{\nu^2}{4n} .$$

We can assume $\nu \geq \delta_\varepsilon$ since δ_ε denotes the closest distance between any two simultaneous centers in iterations leading to a potential drop of at most ε. To conclude the proof, we combine the two cases: If either $\mathcal{C}_1$ or $\mathcal{C}_2$ has only one edge, the potential drop can only be bounded from above by ε if $\varepsilon \geq \frac{\nu^2}{4n} \geq \frac{\delta_\varepsilon^2}{4n}$. Hence, Lemma 2.5 yields

$$\Pr\big[\Delta_2 \leq \varepsilon\big] \leq \Pr\big[\delta_\varepsilon^2/(4n) \leq \varepsilon\big] = \Pr\left[\delta_\varepsilon \leq \sqrt{4n\varepsilon}\right] \leq \varepsilon \cdot \left(\frac{O(1) \cdot n^{11}}{\sigma^2}\right).$$

□

4.3. Pairs of Adjacent Nodes of Degree Two

Given a transition blueprint, we now look at pairs of adjacent nodes of degree 2. Since we have already dealt with the case of balanced clusters of small degree (Section 4.1), we can assume that the nodes involved are unbalanced. This means that one cluster of the pair gains two points while the other cluster of the pair loses two points.

LEMMA 4.3. *Fix $\varepsilon \in [0, 1]$. Let Δ_3 denote the smallest improvement made by any iteration that follows a non-degenerate blueprint with at least three disjoint pairs of adjacent unbalanced nodes of degree 2. Then,*

$$\Pr\big[\Delta_3 \leq \varepsilon\big] \leq \varepsilon \cdot \left(\frac{O(1) \cdot n^{30}}{\sigma^6}\right).$$

PROOF. Fix a transition blueprint $\mathcal{B}$ containing at least 3 disjoint pairs of adjacent unbalanced degree-two nodes. We first bound $\Pr[\Lambda(\mathcal{B}) \leq \lambda]$, that is, the probability that all points that switch clusters have a distance of at most λ from their corresponding approximate bisectors. For $i = 1, 2, 3$, let a_i, b_i, and c_i denote the data points corresponding to the edges in the i^{th} pair of adjacent degree-two nodes, and assume without loss of generality that b_i corresponds to the inner edge (the edge that connects the pair of degree-two nodes).

Let $\mathcal{C}_i$ and $\mathcal{C}_i'$ be the clusters corresponding to one such pair of nodes. Since $\mathcal{C}_i$ and $\mathcal{C}_i'$ are unbalanced, we can further assume without loss of generality that $\mathcal{C}_i$ loses both data points a_i and b_i during the iteration, and $\mathcal{C}_i'$ gains both data points b_i and c_i.

Now, $\mathcal{C}_i$ has its approximate center at $p_i = \frac{a_i+b_i}{2}$ and $\mathcal{C}_i'$ has its approximate center at $q_i = \frac{b_i+c_i}{2}$. Since $\mathcal{B}$ is non-degenerate, we know $p_i \neq q_i$ and hence $a_i \neq c_i$. Let H_i denote

the hyperplane bisecting a_i and c_i, and let H_i' denote the hyperplane bisecting p_i and q_i. Since H_i is the image of H_i' under a dilation with center b_i and scale 2, we have

$$\Lambda(\mathcal{B}) \geq \max_i \left(\operatorname{dist}(b_i, H_i') \right) = \frac{\max_i \left(\operatorname{dist}(b_i, H_i) \right)}{2}. \tag{5}$$

All three pairs of adjacent degree-two nodes are disjoint, so we know b_i is distinct from b_j for $j \neq i$ and distinct from a_j and c_j for all j. This implies the position of b_i is independent of b_j for $j \neq i$, and it is also independent of the position and orientation of H_j for all j. In particular, the quantities $\operatorname{dist}(b_i, H_i)$ follow independent one-dimensional normal distributions with standard deviation σ. Therefore, for any $\lambda \geq 0$, we have

$$\Pr\left[\Lambda(\mathcal{B}) \leq \lambda\right] \leq \Pr\left[\max_i \left(\operatorname{dist}(b_i, H_i) \right) \leq 2\lambda\right] \leq \left(\frac{2\lambda}{\sigma}\right)^3.$$

Let $\mathbb{B}$ denote the set of non-degenerate transition blueprints containing at least three disjoint pairs of unbalanced degree-two nodes. The preceding analysis of $\Pr[\Lambda(\mathcal{B}) \leq \lambda]$ depends only on $\{a_i, b_i, c_i\}$ so we can use a union bound over all choices of $\{a_i, b_i, c_i\}$ as follows:

$$\Pr\left[\exists \mathcal{B} \in \mathbb{B} : \Lambda(\mathcal{B}) \leq \lambda\right] \leq n^9 \cdot \left(\frac{2\lambda}{\sigma}\right)^3 = \left(\frac{2n^3\lambda}{\sigma}\right)^3. \tag{6}$$

Now, Lemma 3.5 implies that if an iteration follows a blueprint $\mathcal{B}$ and results in a potential drop of at most ε, then $\delta_\varepsilon \cdot \Lambda(\mathcal{B}) \leq 6D\sqrt{nd\varepsilon}$. We must therefore have either $\delta_\varepsilon \leq \varepsilon^{1/6}$ or $\Lambda(\mathcal{B}) \leq 6D\sqrt{nd} \cdot \varepsilon^{1/3}$. We bound the probability that this happens using Lemma 2.5 and equation (6):

$$\begin{aligned}
\Pr\left[\Delta_3 \leq \varepsilon\right] &\leq \Pr\left[\delta_\varepsilon \leq \varepsilon^{1/6}\right] + \Pr\left[\exists \mathcal{B} \in \mathbb{B} : \Lambda(\mathcal{B}) \leq 6D\sqrt{nd} \cdot \varepsilon^{1/3}\right] \\
&\leq \varepsilon \cdot \left(\frac{O(1) \cdot n^5}{\sigma}\right)^6 + \varepsilon \cdot \left(\frac{12Dn^3\sqrt{nd}}{\sigma}\right)^3 \\
&= \varepsilon \cdot \left(\frac{O(1) \cdot n^{30}}{\sigma^6}\right),
\end{aligned}$$

since $D = \sqrt{90kd \cdot \ln(n)}$, $\sigma \leq 1$, and $d, k \leq n$. □

4.4. Blueprints with Constant Degree

Now we analyze iterations that follow blueprints in which every node has constant degree. It might happen that a single iteration does not yield a significant improvement in this case. But we get a significant improvement after four consecutive iterations of this kind. The reason for this is that during four iterations one cluster must assume three different configurations (by configuration we mean the set of points that are assigned to the cluster). One case in the previous analyses [Arthur and Vassilvitskii 2009; Manthey and Röglin 2009a] is iterations in which every cluster exchanges at most $O(dk)$ data points with other clusters. The case considered in this section is similar, but instead of relying on the somewhat cumbersome notion of *key-values* used in the previous analyses, we present a simplified and more intuitive analysis here, which also sheds more light on the previous analyses.

We define an *epoch* to be a sequence of consecutive iterations in which no cluster center assumes more than two different positions (by position we mean the center of mass of the points assigned to the cluster). Since the centers of mass of any two distinct subsets of $\mathcal{X}$ are different with probability one, an equivalent definition of an

epoch is that during an epoch, there are at most two different sets $\mathcal{C}_i', \mathcal{C}_i''$ of points that any cluster $\mathcal{C}_i$ assumes. Arthur and Vassilvitskii [2009] used the obvious upper bound of 2^k for the length of an epoch (the term *length* refers to the number of iterations in the sequence). We improve this upper bound to three. By the definition of length of an epoch, this means that after at most four iterations, either k-means terminates or one cluster assumes a third configuration.

LEMMA 4.4. *The length of any epoch is at most three.*

PROOF. Let x be any data point that changes clusters during an epoch, and let $\mathcal{C}_i$ be an arbitrary cluster that contains x at some point. Let $\mathcal{C}_i'$ and $\mathcal{C}_i''$ be the two configurations of $\mathcal{C}_i$ during this epoch, and let c_i' and c_i'' be the centers corresponding to $\mathcal{C}_i'$ and $\mathcal{C}_i''$. Without loss of generality, we assume that $x \in \mathcal{C}_i''$ and $x \notin \mathcal{C}_i'$.

First, we show that if x stays fixed in some cluster $\mathcal{C}_a$ during one iteration but then changes to a new cluster $\mathcal{C}_b$ in the next iteration, then the latter iteration already belongs to a new epoch. Since x belonged to $\mathcal{C}_a$ and not to $\mathcal{C}_b$ at first, the former must have been in configuration $\mathcal{C}_a''$ and the latter in configuration $\mathcal{C}_b'$. After not switching clusters once, the only reason that x will want to switch on a subsequent iteration is that either $\mathcal{C}_a$ changes to some other configuration $\tilde{\mathcal{C}}_a$ that contains x or that $\mathcal{C}_b$ changes to some other configuration $\tilde{\mathcal{C}}_b$ that does not contain x. In both cases, after x switches to $\mathcal{C}_b$, we have three configurations either for $\mathcal{C}_a$ or for $\mathcal{C}_b$.

Second, we observe that it is impossible for x to switch from $\mathcal{C}_a$ to $\mathcal{C}_b$ and, later, from $\mathcal{C}_a$ to a third cluster $\mathcal{C}_c$ during an epoch: The former implies $\|x - c_b'\| < \|x - c_c'\|$ while the latter implies $\|x - c_b'\| > \|x - c_c'\|$, a contradiction. This already implies that x visits either only distinct clusters or that at some point, it changes clusters cyclically.

Third, we show that x belongs to at most three different clusters during an epoch. Assume to the contrary that x belongs to at least four clusters: $\mathcal{C}_a, \mathcal{C}_b, \mathcal{C}_c, \mathcal{C}_d$. By the above argument, we can assume x switches from $\mathcal{C}_a$ to $\mathcal{C}_b$ to $\mathcal{C}_c$ to $\mathcal{C}_d$. The change from $\mathcal{C}_a$ to $\mathcal{C}_b$ yields $\|x - c_b'\| < \|x - c_d'\|$. But, for x to switch from $\mathcal{C}_c$ to $\mathcal{C}_d$, we need $\|x - c_d'\| < \|x - c_b'\|$, a contradiction.

We have now shown that x switches between distinct clusters during every iteration before stabilizing at either a single cluster or a cycle of clusters, and that during this time, x can visit at most three different clusters altogether.

Fourth, let us rule out a cycle of length three as well: Suppose x cycles between $\mathcal{C}_a$, $\mathcal{C}_b$, and $\mathcal{C}_c$. Then, to go from $\mathcal{C}_a$ to $\mathcal{C}_b$, we have $\|x - c_b'\| < \|x - c_c'\|$. Similarly, $\|x - c_c'\| < \|x - c_a'\|$ and $\|x - c_a'\| < \|x - c_b'\|$, which is impossible.

We now know that the sequence of clusters to which x belongs is of one of the following forms: (1) $\mathcal{C}_a$, $\mathcal{C}_a$, ..., (2) $\mathcal{C}_a$, $\mathcal{C}_b$, $\mathcal{C}_b$, ..., (3) $\mathcal{C}_a$, $\mathcal{C}_b$, $\mathcal{C}_c$, $\mathcal{C}_c$, ..., (4) $\mathcal{C}_a$, $\mathcal{C}_b$, $\mathcal{C}_a$, $\mathcal{C}_b$, ... (x changes back and forth between $\mathcal{C}_a$ and $\mathcal{C}_b$), or (5) $\mathcal{C}_a$, $\mathcal{C}_b$, $\mathcal{C}_c$, $\mathcal{C}_b$, $\mathcal{C}_c$, ... (x changes back and forth between $\mathcal{C}_b$ and $\mathcal{C}_c$). The list above tell us that after the fourth iteration of an epoch, we get the same clustering as after the second iteration. Since no clustering can repeat during an execution of k-means, this concludes the proof. □

For our analysis, we introduce the notion of (η, c)-*coarseness*. In the following, $\triangle$ denotes the symmetric difference of two sets.

Definition 4.5. We say that $\mathcal{X}$ is (η, c)-*coarse* if for every triple $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ of pairwise distinct subsets 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 $\|\mathrm{cm}(\mathcal{C}_2) - \mathrm{cm}(\mathcal{C}_3)\| > \eta$.

According to Lemma 4.4, in every sequence of four consecutive iterations, one cluster assumes three different configurations. This yields the following lemma.

LEMMA 4.6. *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 η^2.*

PROOF. According to Lemma 4.4, there is one cluster that assumes three different configurations $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ in this sequence. Due to the assumption in the lemma, we have $|\mathcal{C}_1 \triangle \mathcal{C}_2| \leq c$ and $|\mathcal{C}_2 \triangle \mathcal{C}_3| \leq c$. Hence, due to the definition of (η, c)-coarseness, we have $\|\mathrm{cm}(\mathcal{C}_i) - \mathrm{cm}(\mathcal{C}_{i+1})\| > \eta$ for one $i \in \{1, 2\}$. Combining this with Lemma 2.3 concludes the proof. □

LEMMA 4.7. *For $\eta \geq 0$, the probability that $\mathcal{X}$ is not (η, c)-coarse is at most $(7n)^{2c} \cdot (2nc\eta/\sigma)^d$.*

PROOF. Given any sets $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ with $|\mathcal{C}_1 \triangle \mathcal{C}_2| \leq c$ and $|\mathcal{C}_2 \triangle \mathcal{C}_3| \leq c$, we can write $\mathcal{C}_i$, for $i \in \{1, 2, 3\}$, uniquely as the disjoint union of a common ground set $A \subseteq \mathcal{X}$ with a set $B_i \subseteq \mathcal{X}$ with $B_1 \cap B_2 \cap B_3 = \emptyset$. Furthermore,

$$B_1 \cup B_2 \cup B_3 = (C_1 \cup C_2 \cup C_3) \setminus A = (C_1 \triangle C_2) \cup (C_2 \triangle C_3),$$

so $|B_1 \cup B_2 \cup B_3| = |(C_1 \triangle C_2) \cup (C_2 \triangle C_3)| \leq 2c$.

We perform a union bound over all choices for the sets B_1, B_2, and B_3. The number of choices for these sets is bounded from above by $7^{2c}\binom{n}{2c} \leq (7n)^{2c}$: We choose $2c$ candidate points for $B_1 \cup B_2 \cup B_3$, and then for each point, we choose which set(s) it belongs to (it does not belong to all of them, but we allow that it belongs to none of them because otherwise we would not cover the case that $B_1 \cup B_2 \cup B_3$ contains fewer than $2c$ points). We assume in the following that the sets B_1, B_2, and B_3 are fixed. For $i \in \{1, 2\}$, we can write $\mathrm{cm}(\mathcal{C}_i) - \mathrm{cm}(\mathcal{C}_{i+1})$ as

$$\left(\frac{|A|}{|A| + |B_i|} - \frac{|A|}{|A| + |B_{i+1}|}\right) \cdot \mathrm{cm}(A) + \frac{|B_i|}{|A| + |B_i|} \cdot \mathrm{cm}(B_i) - \frac{|B_{i+1}|}{|A| + |B_{i+1}|} \cdot \mathrm{cm}(B_{i+1})\,. \quad (7)$$

Let us first consider the case that we have $|B_i| = |B_{i+1}|$ for one $i \in \{1, 2\}$. Then $\mathrm{cm}(\mathcal{C}_i) - \mathrm{cm}(\mathcal{C}_{i+1})$ simplifies to

$$\frac{|B_i|}{|A| + |B_i|} \cdot (\mathrm{cm}(B_i) - \mathrm{cm}(B_{i+1})) = \frac{1}{|A| + |B_i|} \cdot \left(\sum_{x \in B_i \setminus B_{i+1}} x - \sum_{x \in B_{i+1} \setminus B_i} x\right).$$

Since $B_i \neq B_{i+1}$, there exists a point $x \in B_i \triangle B_{i+1}$. Let us assume without loss of generality that $x \in B_i \setminus B_{i+1}$ and that the positions of all points in $(B_i \cup B_{i+1}) \setminus \{x\}$ are fixed arbitrarily. Then the event that $\|\mathrm{cm}(\mathcal{C}_i) - \mathrm{cm}(\mathcal{C}_{i+1})\| \leq \eta$ is equivalent to the event that x lies in a fixed hyperball of radius $(|A| + |B_i|)\eta \leq n\eta$. Hence, the probability is bounded from above by $(n\eta/\sigma)^d \leq (2nc\eta/\sigma)^d$.

Now assume that $|B_1| \neq |B_2| \neq |B_3|$. For $i \in \{1, 2\}$, we set

$$r_i = \left(\frac{|A|}{|A| + |B_i|} - \frac{|A|}{|A| + |B_{i+1}|}\right)^{-1} = \frac{(|A| + |B_i|) \cdot (|A| + |B_{i+1}|)}{|A| \cdot (|B_{i+1}| - |B_i|)}$$

and

$$Z_i = \frac{|B_{i+1}|}{|A| + |B_{i+1}|} \cdot \mathrm{cm}(B_{i+1}) - \frac{|B_i|}{|A| + |B_i|} \cdot \mathrm{cm}(B_i)\,.$$

According to (7), the event $\|\mathrm{cm}(\mathcal{C}_i) - \mathrm{cm}(\mathcal{C}_{i+1})\| \leq \eta$ is equivalent to the event that $\mathrm{cm}(A)$ falls into the hyperball with radius $|r_i|\eta$ and center $r_i Z_i$. Hence, the event that both $\|\mathrm{cm}(\mathcal{C}_1) - \mathrm{cm}(\mathcal{C}_2)\| \leq \eta$ and $\|\mathrm{cm}(\mathcal{C}_2) - \mathrm{cm}(\mathcal{C}_3)\| \leq \eta$ can only occur if the hyperballs

$\mathcal{B}(r_1 Z_1, |r_1|\eta)$ and $\mathcal{B}(r_2 Z_2, |r_2|\eta)$ intersect. This event occurs if and only if the centers $r_1 Z_1$ and $r_2 Z_2$ have a distance of at most $(|r_1| + |r_2|)\eta$ from each other. Hence,

$$\begin{aligned}&\Pr\big[(\|\mathrm{cm}(\mathcal{C}_1) - \mathrm{cm}(\mathcal{C}_2)\| \le \eta) \wedge (\|\mathrm{cm}(\mathcal{C}_2) - \mathrm{cm}(\mathcal{C}_3)\| \le \eta)\big]\\ &\le \Pr\left[\|r_1 Z_1 - r_2 Z_2\| \le (|r_1| + |r_2|)\eta\right] .\end{aligned}$$

After some algebraic manipulations, we can write the vector $r_1 Z_1 - r_2 Z_2$ as

$$\begin{aligned}&- \frac{|A| + |B_2|}{|A| \cdot (|B_2| - |B_1|)} \cdot \sum_{x \in B_1} x - \frac{|A| + |B_2|}{|A| \cdot (|B_3| - |B_2|)} \cdot \sum_{x \in B_3} x\\ &+ \left(\frac{|A| + |B_1|}{|A| \cdot (|B_2| - |B_1|)} + \frac{|A| + |B_3|}{|A| \cdot (|B_3| - |B_2|)}\right) \cdot \sum_{x \in B_2} x .\end{aligned}$$

Since $B_1 \neq B_3$, there must be an $x \in B_1 \triangle B_3$. We can assume that $x \in B_1 \setminus B_3$. If $x \notin B_2$, we let an adversary choose all positions of the points in $B_1 \cup B_2 \cup B_3 \setminus \{x\}$. Then the event $\|r_1 Z_1 - r_2 Z_2\| \le (|r_1| + |r_2|)\eta$ is equivalent to x falling into a fixed hyperball of radius

$$\begin{aligned}&\left|\frac{|A| \cdot (|B_2| - |B_1|)}{|A| + |B_2|}(|r_1| + |r_2|)\right| \eta\\ &= \left|(|B_2| - |B_1|) \cdot \left(\left|\frac{|A| + |B_1|}{|B_2| - |B_1|}\right| + \left|\frac{|A| + |B_3|}{|B_3| - |B_2|}\right|\right)\right| \eta \le 2nc\eta .\end{aligned}$$

The probability of this event is thus bounded from above by $(2nc\eta/\sigma)^d$.

It remains to consider the case that $x \in (B_1 \cap B_2) \setminus B_3$. Also in this case we let an adversary choose the positions of the points in $B_1 \cup B_2 \cup B_3 \setminus \{x\}$. Now the event $\|r_1 Z_1 - r_2 Z_2\| \le (|r_1| + |r_2|)\eta$ is equivalent to x falling into a fixed hyperball of radius

$$\begin{aligned}&\left|\frac{|A| \cdot (|B_3| - |B_2|)}{|A| + |B_2|}(|r_1| + |r_2|)\right| \eta\\ &= \left|(|B_3| - |B_2|) \cdot \left(\left|\frac{|A| + |B_1|}{|B_2| - |B_1|}\right| + \left|\frac{|A| + |B_3|}{|B_3| - |B_2|}\right|\right)\right| \eta \le 2nc\eta .\end{aligned}$$

Hence, the probability is bounded from above by $(2nc\eta/\sigma)^d$ also in this case.

This concludes the proof because there are at most $(7n)^{2c}$ choices for B_1, B_2, and B_3 and, for every choice, the probability that both $\|\mathrm{cm}(\mathcal{C}_1) - \mathrm{cm}(\mathcal{C}_2)\| \le \eta$ and $\|\mathrm{cm}(\mathcal{C}_2) - \mathrm{cm}(\mathcal{C}_3)\| \le \eta$ is at most $(2nc\eta/\sigma)^d$. □

Combining Lemmas 4.6 and 4.7 immediately yields the following result.

LEMMA 4.8. *Fix $\varepsilon \ge 0$ and a constant $z_2 \in \mathbb{N}$. Let Δ_4 denote the smallest improvement made by any sequence of four consecutive iterations that follow blueprints whose nodes all have degree at most z_2. Then,*

$$\Pr\big[\Delta_4 \le \varepsilon\big] \le \varepsilon \cdot \left(\frac{O(1) \cdot n^{2(z_2+1)}}{\sigma^2}\right) .$$

PROOF. Taking $\eta = \sqrt{\varepsilon}$ and $c = z_2$, Lemmas 4.6 and 4.7 immediately give

$$\Pr\big[\Delta_4 \le \varepsilon\big] \le (7n)^{2z_2} \cdot \left(\frac{2nz_2\sqrt{\varepsilon}}{\sigma}\right)^d .$$

Since $d \ge 2$, the lemma follows from Fact 2.1 and the fact that z_2 is a constant. □

4.5. Degenerate blueprints

LEMMA 4.9. *Fix $\varepsilon \in [0,1]$. Let Δ_5 denote the smallest improvement made by any iteration that follows a degenerate blueprint. Then,*

$$\Pr\left[\Delta_5 \leq \varepsilon\right] \leq \varepsilon \cdot \left(\frac{O(1) \cdot n^{11}}{\sigma^2}\right).$$

PROOF. We first argue that $\Delta_5 \leq \varepsilon$ implies $\delta_\varepsilon \leq 2\sqrt{n\varepsilon}$. For this, consider an arbitrary iteration that follows a degenerate blueprint and that makes an improvement of at most ε. Let $\mathcal{C}_i$ and $\mathcal{C}_j$ be two clusters that have the same approximate center c and that exchange a data point during the iteration, and let c_i and c_j denote the actual centers of these clusters at the beginning of the iteration.

If $\mathcal{C}_i$ is unbalanced, then by Lemma 3.3 it follows that $\|c-c_i\| \leq \sqrt{n\varepsilon}$. If $\mathcal{C}_i$ is balanced, then $\|c-c_i\| \leq \sqrt{n\varepsilon}$ because the iteration follows the blueprint. Similarly one can argue that in any case $\|c-c_j\| \leq \sqrt{n\varepsilon}$, and hence $\|c_i-c_j\| \leq 2\sqrt{n\varepsilon}$. This implies that $\delta_\varepsilon \leq 2\sqrt{n\varepsilon}$.

However, we know from Lemma 2.5 that this occurs with probability at most $\varepsilon \cdot (O(1) \cdot n^{5.5}/\sigma)^2$. □

4.6. Other Blueprints

Now, after having ruled out five special cases, we can analyze the case of a general blueprint.

LEMMA 4.10. *Fix $\varepsilon \in [0,1]$. Let Δ_6 be the smallest improvement made by any iteration whose blueprint does not fall into any of the previous five categories with $z_1 = 8$ and $z_2 = 7$. This means that we consider only non-degenerate blueprints whose balanced nodes have in- and out-degree at least $8d+1$, that do not have nodes of degree one, that have at most two disjoint pairs of adjacent unbalanced node of degree 2, and that have a node with degree at least 8. Then,*

$$\Pr\left[\Delta_6 \leq \varepsilon\right] \leq \varepsilon \cdot \left(\frac{O(1) \cdot n^{33} k^{30} d^3 D^3}{\sigma^6}\right).$$

Proving this lemma requires some preparation. Assume that the iteration follows a blueprint $\mathcal{B}$ with m edges and b balanced nodes. We distinguish two cases: either the center of one unbalanced cluster assumes a position that is $\sqrt{n\varepsilon}$ away from its approximate position or all centers are at most $\sqrt{n\varepsilon}$ far away from their approximate positions. In the former case the potential drops by at least ε according to Lemma 3.3. If this is not the case, the potential drops if one of the points is far away from its corresponding approximate bisector according to Lemma 3.5.

The fact that the blueprint does not belong to any of the previous categories allows us to derive the following upper bound on its number of nodes.

LEMMA 4.11. *Let $\mathcal{B}$ denote an arbitrary transition blueprint with m edges and b balanced nodes in which every node has degree at least 2 and every balanced node has degree at least $2dz_1 + 2$. Furthermore, let there be at most two disjoint pairs of adjacent nodes of degree two in $\mathcal{B}$, and assume that there is one node with degree at least $z_2 + 1 > 2$. Then the number of nodes in $\mathcal{B}$ is bounded from above by*

$$\begin{cases} \frac{5}{6}m - \frac{z_2-4}{3} & \text{if } b = 0, \\ \frac{5}{6}m - \frac{(2z_1d-1)b-2}{3} & \text{if } b \geq 1. \end{cases}$$

PROOF. Let A be the set of nodes of degree two, and let B be the set of nodes of higher degree.

We first bound the number of edges between nodes in A. Towards that end, let G_A denote the subgraph induced by A. We ignore the directions of the edges and treat

G_A as an undirected multigraph. Since every node in G_A has degree at most two, G_A decomposes into a collection of paths, cycles, and isolated vertices. In fact, there cannot be any cycles. If a cycle did exist, then since all vertices in A have degree two, the cycle would form an entire connected component of the transition graph. However, the transition graph is weakly connected and has some vertex of degree at least $z_2 + 1 > 2$, so this is impossible. Therefore, G_A can be decomposed into a collection of paths of length $\ell_1, \ldots, \ell_t$ and possibly some number of isolated vertices. Since the transition graph has at most two pairs of disjoint adjacent degree-two nodes, and since each path of length ℓ_i admits $\left\lceil \frac{\ell_i}{2} \right\rceil$ such pairs, we must have

$$\sum_{i=1}^{t} \ell_i \leq 2 \cdot \sum_{i=1}^{t} \left\lceil \frac{\ell_i}{2} \right\rceil \leq 4 \,.$$

Therefore, there are at most 4 edges connecting nodes in A.

Let $\deg(A)$ and $\deg(B)$ denote the sum of the degrees of the nodes in A and B, respectively. Then $\deg(A) + \deg(B) = 2m$ and the total degree $\deg(A)$ of the vertices in A is $2|A|$. Hence, there are at least $2|A| - 8$ edges between A and B. Therefore,

$$\begin{aligned} & 2|A| - 8 \leq \deg(B) = 2m - 2|A| \\ \Rightarrow\ & |A| \leq \frac{1}{2}m + 2 \,. \end{aligned} \tag{8}$$

Let t denote the number of nodes. The nodes in B have degree at least 3, there is one node in B with degree at least $z_2 + 1$, and balanced nodes have degree at least $2z_1 d + 2$ (and hence, belong to B). Therefore, if $b = 0$,

$$\begin{aligned} & 2m \geq 2|A| + 3(t - |A| - 1) + z_2 + 1 \\ \iff\ & 2m + |A| \geq 3t + z_2 - 2 \\ \Rightarrow\ & \frac{5}{2}m \geq 3t + z_2 - 4 \,. \qquad \text{(due to (8))} \end{aligned}$$

If $b \geq 1$, then the node of degree at least $z_2 + 1$ might be balanced and we obtain

$$\begin{aligned} & 2m \geq 2|A| + (2z_1 d + 2)b + 3(t - |A| - b) \\ \iff\ & 2m + |A| \geq 3t + (2z_1 d - 1)b \\ \Rightarrow\ & \frac{5}{2}m \geq 3t + (2z_1 d - 1)b - 2 \,. \qquad \text{(due to (8))} \end{aligned}$$

The lemma follows by solving these inequalities for t. □

We can now continue to bound $\Pr[\Lambda(\mathcal{B}) \leq \lambda]$ for a fixed blueprint $\mathcal{B}$. The previous lemma implies that a relatively large number of points must switch clusters, and each such point is positioned independently according to a normal distribution. Unfortunately, the approximate bisectors are not independent of these point locations, which adds a technical challenge. We resolve this difficulty by changing variables and then bounding the effect of this change.

LEMMA 4.12. *For a fixed transition blueprint $\mathcal{B}$ with m edges and b balanced clusters that does not belong to any of the previous five categories and for any $\lambda \geq 0$, we have*

$$\Pr\big[\Lambda(\mathcal{B}) \leq \lambda\big] \leq \begin{cases} \left(\frac{\sqrt{d}m^2\lambda}{\sigma}\right)^{\frac{m}{6} + \frac{z_2 - 1}{3}} & \text{if } b = 0, \\ \left(\frac{\sqrt{d}m^2\lambda}{\sigma}\right)^{\frac{m}{6} + \frac{(2z_1 d + 2)b - 2}{3}} & \text{if } b \geq 1. \end{cases}$$

PROOF. We partition the set of edges in the transition graph into *reference edges* and *test edges*. For this, we ignore the directions of the edges in the transition graph and compute a spanning tree in the resulting undirected multi-graph. We let an arbitrary balanced cluster be the root of this spanning tree. If all clusters are unbalanced, then an arbitrary cluster is chosen as the root. We mark every edge whose child is an unbalanced cluster as a reference edge. All other edges are test edges. In this way, every unbalanced cluster $\mathcal{C}_i$ can be incident to several reference edges. But we will refer only to the reference edge between $\mathcal{C}_i$'s parent and $\mathcal{C}_i$ as the reference edge associated with $\mathcal{C}_i$. Possibly except for the root, every unbalanced cluster is associated with exactly one reference edge. Observe that in the transition graph, the reference edge of an unbalanced cluster $\mathcal{C}_i$ can either be directed from $\mathcal{C}_i$ to its parent or vice versa, as we ignored the directions of the edges when we computed the spanning tree. From now on, we will again take into account the directions of the edges.

For every unbalanced cluster i with an associated reference edge, we define the point q_i as

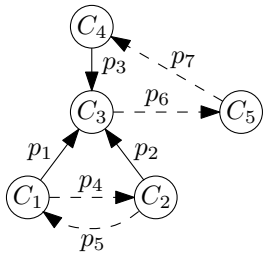
$$q_i = \sum_{x \in A_i} x - \sum_{x \in B_i} x = |A_i| \operatorname{cm}(A_i) - |B_i| \operatorname{cm}(B_i)\,, \tag{9}$$

where A_i and B_i denote the sets of incoming and outgoing edges of $\mathcal{C}_i$, respectively. The intuition behind this definition is as follows: as we consider a fixed blueprint $\mathcal{B}$, once q_i is fixed also the approximate center of cluster i is fixed. If there is at least one balanced cluster, then every unbalanced cluster has an associated reference edge. Hence, once each q_i is fixed, the approximate centers of all unbalanced clusters are also fixed. If all clusters are unbalanced, let q denote the point defined as in (9) but for the root instead of cluster i. If q_i is fixed for every cluster except for the root, then also the value of q is implicitly fixed as $q + \sum q_i = 0$. Hence, once each q_i is fixed, the approximate centers of all unbalanced clusters are also fixed in this case.

Relabeling if necessary, we assume without loss of generality that the clusters with an associated reference edge are the clusters $\mathcal{C}_1, \ldots, \mathcal{C}_r$ and that the corresponding reference edges correspond to the points $p_1, \ldots, p_r$. Furthermore, we can assume that the clusters are topologically sorted: if $\mathcal{C}_i$ is a descendant of $\mathcal{C}_j$ in the spanning tree, then $i < j$.

Let us now assume that an adversary chooses an arbitrary position for q_i for every cluster $\mathcal{C}_i$ with $i \in [r]$. Intuitively, we will show that regardless of how the transition blueprint $\mathcal{B}$ is chosen and regardless of how the adversary fixes the positions of the q_i, there is still enough randomness left to conclude that it is unlikely that all points involved in the iteration are close to their corresponding approximate bisectors. We can alternatively view this as follows: Our random experiment is to choose the md-dimensional Gaussian vector $\bar{p} = (p_1, \ldots, p_m)$, where $p_1, \ldots, p_m \in \mathbb{R}^d$ are the points that correspond to the edges in the blueprint. For each $i \in [r]$ and $j \in [d]$ let $\bar{b}_{ij} \in \{-1, 0, 1\}^{md}$ be the vector so that the j^{th} component of q_i can be written as $\bar{p} \cdot \bar{b}_{ij}$. Then allowing the adversary to fix the positions of the q_i is equivalent to letting him fix the value of every dot product $\bar{p} \cdot \bar{b}_{ij}$.

After the positions of the q_i are chosen, we know the location of the approximate center of every unbalanced cluster. Additionally, the blueprint provides an approximate center for every balanced cluster. Hence, we know the positions of all approximate bisectors. We would like to estimate the probability that all points $p_{r+1}, \ldots, p_m$ have a distance of at most λ from their corresponding approximate bisectors. For this, we further reduce the randomness and project each point p_i with $i \in \{r+1, \ldots, m\}$ onto the normal vector of its corresponding approximate bisector. Formally, for each $i \in \{r+1, \ldots, m\}$, let h_i denote a normal vector to the approximate bisector corresponding to p_i, and let $\bar{b}_{i1} \in [-1, 1]^{md}$ denote the vector such that $\bar{p} \cdot \bar{b}_{i1} \equiv p_i \cdot h_i$. This



$$M = \left(\begin{array}{ccc|cccc} -I_d & 0_d & I_d & 0_d & 0_d & 0_d & 0_d \\ 0_d & -I_d & I_d & 0_d & 0_d & 0_d & 0_d \\ 0_d & 0_d & I_d & 0_d & 0_d & 0_d & 0_d \\ \hline -I_d & I_d & 0_d & \mathcal{B}_4 & 0_d & 0_d & 0_d \\ I_d & -I_d & 0_d & 0_d & \mathcal{B}_5 & 0_d & 0_d \\ 0_d & 0_d & -I_d & 0_d & 0_d & \mathcal{B}_6 & 0_d \\ 0_d & 0_d & 0_d & 0_d & 0_d & 0_d & \mathcal{B}_7 \end{array}\right)$$

Fig. 1. Solid and dashed edges indicate reference and test edges, respectively. When computing the spanning tree, the directions of the edges are ignored. Hence, reference edges can either be directed from parent to child or vice versa. In this example, the spanning tree consists of the edges p_3, p_7, p_1, and p_2, and its root is C_4. We denote by I_d the $d \times d$ identity matrix and by 0_d the $d \times d$ zero matrix. The first three columns of M correspond to q_1, q_2, and q_3. The rows correspond to the points $p_1, \ldots, p_7$. Each block matrix $\mathcal{B}_i$ corresponds to an orthonormal basis of $\mathbb{R}^d$.

means that p_i is at a distance of at most λ from its approximate bisector if and only if $\bar{p} \cdot \bar{b}_{i1}$ lies in some fixed interval $\mathcal{I}_i$ of length 2λ. This event depends only on p_i and is independent of p_j for $j \in \{r+1, \ldots, m\}$ with $j \neq i$, given that the approximate centers are fixed. Thus, the vector $\bar{b}_{i1}$ is a unit vector in the subspace spanned by the vectors $e_{(i-1)d+1}, \ldots, e_{id}$ from the canonical basis. Let $\mathcal{B}_i = \{\bar{b}_{i1}, \ldots, \bar{b}_{id}\}$ be an orthonormal basis of this subspace. Let M denote the $(md) \times (md)$ matrix whose columns are the vectors $\bar{b}_{11}, \ldots, \bar{b}_{1d}, \ldots, \bar{b}_{m1}, \ldots, \bar{b}_{md}$. Figure 1 illustrates these definitions.

For $i \in [r]$ and $j \in [d]$, the values of $\bar{p} \cdot \bar{b}_{ij}$ are fixed by an adversary. Additionally, we allow the adversary to fix the values of $\bar{p} \cdot \bar{b}_{ij}$ for $i \in \{r+1, \ldots, m\}$ and $j \in \{2, \ldots, d\}$. All this together defines an $(m-r)$-dimensional affine subspace U of $\mathbb{R}^{md}$. We stress that the subspace U is chosen by the adversary and no assumptions about U are made. In the following, we will condition on the event that $\bar{p} = (p_1, \ldots, p_m)$ lies in this subspace. We denote by $\mathcal{F}$ the event that $\bar{p} \cdot \bar{b}_{i1} \in \mathcal{I}_i$ for all $i \in \{r+1, \ldots, m\}$. Conditioned on the event that the random vector $\bar{p}$ lies in the subspace U, $\bar{p}$ follows an $(m-r)$-dimensional Gaussian distribution with standard deviation σ. However, we cannot directly estimate the probability of the event $\mathcal{F}$ as the projections of the vectors $\bar{b}_{i1}$ onto the affine subspace U might not be orthogonal. To estimate the probability of $\mathcal{F}$, we perform a change of variables. Let $\bar{a}_1, \ldots, \bar{a}_{m-r}$ be an arbitrary orthonormal basis of the $(m-r)$-dimensional subspace obtained by shifting U so that it contains the origin. Assume for the moment that we had, for each of these vectors $\bar{a}_\ell$, an interval $\mathcal{I}'_\ell$ such that $\mathcal{F}$ can only occur if $\bar{p} \cdot \bar{a}_\ell \in \mathcal{I}'_\ell$ for every ℓ. Then we could bound the probability of $\mathcal{F}$ from above by $\prod \frac{|\mathcal{I}'_\ell|}{\sqrt{2\pi}\sigma}$ as the $\bar{p} \cdot \bar{a}_\ell$ can be treated as independent one-dimensional Gaussian random variables with standard deviation σ after conditioning on U. In the following, we construct such intervals $\mathcal{I}'_\ell$.

It is important that the vectors $\bar{b}_{ij}$ for $i \in [m]$ and $j \in [d]$ form a basis of $\mathbb{R}^{md}$. To see this, let us first have a closer look at the matrix $M \in \mathbb{R}^{md \times md}$ viewed as an $m \times m$ block matrix with blocks of size $d \times d$. From the fact that the reference points are topologically sorted it follows that the upper left part, which consists of the first dr rows and columns, is an upper triangular matrix with non-zero diagonal entries.

As the upper right $(dr) \times d(m-r)$ sub-matrix of M consists solely of zeros, the determinant of M is the product of the determinant of the upper left $(dr) \times (dr)$ sub-matrix and the determinant of the lower right $d(m-r) \times d(m-r)$ sub-matrix. Both of these determinants can easily be seen to be different from zero. Hence, also the determinant of M is not equal to zero, which in turn implies that the vectors $\bar{b}_{ij}$ are linearly independent and form a basis of $\mathbb{R}^{md}$.

In particular, we can write every $\bar{a}_\ell$ as a linear combination of the vectors $\bar{b}_{ij}$. Let

$$\bar{a}_\ell = \sum_{i,j} c_{ij}^\ell \bar{b}_{ij}$$

for some coefficients $c_{ij}^\ell \in \mathbb{R}$. Since the values of $\bar{p} \cdot \bar{b}_{ij}$ are fixed for $i \in [r]$ and $j \in [d]$ as well as for $i \in \{r+1, \dots, m\}$ and $j \in \{2, \dots, d\}$, we can write

$$\bar{p} \cdot \bar{a}_\ell = \kappa_\ell + \sum_{i=r+1}^{m} c_{i1}^\ell (\bar{p} \cdot \bar{b}_{i1})$$

for some constant κ_ℓ that depends on the fixed values chosen by the adversary. Let $c_{\max} = \max\{|c_{i1}^\ell| \mid i > r\}$. The event $\mathcal{F}$ happens only if, for every $i > r$, the value of $\bar{p} \cdot \bar{b}_{i1}$ lies in some fixed interval of length 2λ. Thus, we conclude that $\mathcal{F}$ can happen only if for every $\ell \in [m-r]$ the value of $\bar{p} \cdot \bar{a}_\ell$ lies in some fixed interval $\mathcal{I}'_\ell$ of length at most $2c_{\max}(m-r)\lambda$. It only remains to bound $c_{\max}$ from above. For $\ell \in [m-r]$, the vector c^ℓ of the coefficients c_{ij}^ℓ is obtained as the solution of the linear system $Mc^\ell = \bar{a}_\ell$. The fact that the upper right $(dr) \times d(m-r)$ sub-matrix of M consists only of zeros implies that the first dr entries of $\bar{a}_\ell$ uniquely determine the first dr entries of the vector c^ℓ. As $\bar{a}_\ell$ is a unit vector, the absolute values of all its entries are bounded by 1. Now we observe that each row of the matrix M contains at most two non-zero entries in the first dr columns because every edge in the transition blueprint belongs to only two clusters. This and a short calculation show that the absolute values of the first dr entries of c are bounded by r: The absolute values of the entries $d(r-1)+1, \dots, dr$ coincide with the absolute values of the corresponding entries in $\bar{a}_\ell$ and are thus bounded by 1. Given this, the rows $d(r-2)+1, \dots, d(r-1)$ imply that the corresponding values in $\bar{a}_\ell$ are bounded by 2 and so on.

Assume that the first dr coefficients of c^ℓ are fixed to values whose absolute values are bounded by r. This leaves us with a system $M'(c^\ell)' = \bar{a}'_\ell$, where M' is the lower right $\big((m-r)d\big) \times \big((m-r)d\big)$ sub-matrix of M, $(c^\ell)'$ are the remaining $(m-r)d$ entries of c^ℓ, and $\bar{a}'_\ell$ is a vector obtained from $\bar{a}_\ell$ by taking into account the first dr fixed values of c^ℓ. All absolute values of the entries of $\bar{a}'_\ell$ are bounded by $2r+1$. As M' is a diagonal block matrix, we can decompose this into $m-r$ systems with d variables and equations each. As every $d \times d$-block on the diagonal of the matrix M' is an orthonormal basis of the corresponding d-dimensional subspace, the matrices in the subsystems are orthonormal. Furthermore, the right-hand sides have a norm of at most $(2r+1)\sqrt{d} \le 3\sqrt{d}r$. Hence, we can conclude that $c_{\max}$ is bounded from above by $3\sqrt{d}r$.

Thus, the probability of the event $\mathcal{F}$ can be bounded from above by

$$\prod_{i=r+1}^{m} \frac{|\mathcal{I}'_i|}{\sqrt{2\pi}\sigma} \le \left(\frac{6\sqrt{d}r(m-r)\lambda}{\sqrt{2\pi}\sigma}\right)^{m-r} \le \left(\frac{\sqrt{d}m^2\lambda}{\sigma}\right)^{m-r},$$

where we used that $r(m-r) \le m^2/4$. Using Fact 2.1, we can replace the exponent $m-r$ by a lower bound. If all nodes are unbalanced, then r equals the number of nodes minus one. Otherwise, if $b \ge 1$, then r equals the number of nodes minus b. Hence, Lemma 4.11 yields

$$\Pr\big[\Lambda(\mathcal{B}) \le \lambda\big] \le \begin{cases} \left(\frac{\sqrt{d}m^2\lambda}{\sigma}\right)^{\frac{m}{6} + \frac{z_2 - 4}{3} + 1} & \text{if } b = 0, \\ \left(\frac{\sqrt{d}m^2\lambda}{\sigma}\right)^{\frac{m}{6} + \frac{(2z_1 d - 1)b - 2}{3} + b} & \text{if } b \ge 1, \end{cases}$$

which completes the proof. □

With the previous lemma, we can bound the probability that there exists an iteration whose transition blueprint does not fall into any of the previous categories and that makes a small improvement.

PROOF OF LEMMA 4.10. Let $\mathbb{B}$ denote the set of (m, b, ε)-blueprints that do not fall into the previous five categories. Here, ε is fixed but there are nk possible choices for m and b. As in the proof of Lemma 4.3, we will use a union bound to estimate the probability that there exists a blueprint $\mathcal{B} \in \mathbb{B}$ with $\Lambda(\mathcal{B}) \le \lambda$. Note that once m and b are fixed, there are at most $(nk^2)^m$ possible choices for the edges in a blueprint, and for every balanced cluster, there are at most $\left(\frac{D\sqrt{d}}{\sqrt{n\varepsilon}}\right)^d$ choices for its approximate center. Since $\mathcal{B}$ does not belong to any of the previous five categories, $m \ge \max(z_2+1, b(dz_1+1)) = \max(8, 8bd+b)$ because there is one vertex with degree at least z_2+1, and there are b vertices with degree at least $2dz_1+2$.

Now we set $Y = k^5 \cdot \sqrt{ndD}$. Lemma 4.12 yields the following bound:

$$\begin{aligned}
&\Pr\left[\exists \mathcal{B} \in \mathbb{B} : \Lambda(\mathcal{B}) \le \frac{6D\sqrt{nd}}{Y} \cdot \varepsilon^{1/3}\right] \\
&\le \sum_{m=8}^{n} (nk^2)^m \cdot \left(\frac{6m^2 dD\sqrt{n}}{Y\sigma} \cdot \varepsilon^{1/3}\right)^{\frac{m}{6}+\frac{z_2-1}{3}} \\
&\quad + \sum_{b=1}^{k} \sum_{m=8bd+b}^{n} \left(\frac{D\sqrt{d}}{\sqrt{n\varepsilon}}\right)^{bd} \cdot (nk^2)^m \cdot \left(\frac{6m^2 dD\sqrt{n}}{Y\sigma} \cdot \varepsilon^{1/3}\right)^{\frac{m}{6}+\frac{(2z_1d+2)b-2}{3}}. \qquad (10)
\end{aligned}$$

Each term in the first sum simplifies as follows:

$$\begin{aligned}
(nk^2)^m \cdot \left(\frac{6m^2 dD\sqrt{n}}{Y\sigma} \cdot \varepsilon^{1/3}\right)^{\frac{m}{6}+\frac{z_2-1}{3}} &\le \left(\frac{6n^{17/2}k^{12}dD}{Y\sigma} \cdot \varepsilon^{1/3}\right)^{\frac{m}{6}+\frac{z_2-1}{3}} \\
&= \left(\frac{6n^8k^7d^{1/2}D^{1/2}}{\sigma} \cdot \varepsilon^{1/3}\right)^{\frac{m}{6}+\frac{z_2-1}{3}}.
\end{aligned}$$

Furthermore, $\frac{m}{6}+\frac{z_2-1}{3} \ge \frac{8}{6}+\frac{6}{3} > 3$, so we can use Fact 2.1 to decrease the exponent here, which gives us

$$\left(\frac{6n^8k^7d^{1/2}D^{1/2}}{\sigma} \cdot \varepsilon^{1/3}\right)^3 = \varepsilon \cdot \left(\frac{O(1) \cdot n^{24}k^{21}d^{3/2}D^{3/2}}{\sigma^3}\right).$$

Similarly, each term in the second sum simplifies as follows:

$$\begin{aligned}
&\left(\frac{D\sqrt{d}}{\sqrt{n\varepsilon}}\right)^{bd} \cdot (nk^2)^m \cdot \left(\frac{6m^2 dD\sqrt{n}}{Y\sigma} \cdot \varepsilon^{1/3}\right)^{\frac{m}{6}+\frac{(2z_1d+2)b-2}{3}} \\
&\le \left(\frac{D\sqrt{d}}{\sqrt{n\varepsilon}}\right)^{bd} \cdot \left(\frac{6n^8k^7d^{1/2}D^{1/2}}{\sigma} \cdot \varepsilon^{1/3}\right)^{\frac{m}{6}+\frac{(2z_1d+2)b-2}{3}}.
\end{aligned}$$

Furthermore,

$$\frac{m}{6}+\frac{(2z_1d+2)b-2}{3} \ge \frac{8bd+b}{6}+\frac{16bd+2b-2}{3} \ge \frac{20bd}{3}.$$

Therefore, we can further bound the terms in the second sum by

$$\left(\left(\frac{D\sqrt{d}}{\sqrt{n\varepsilon}}\right)^{3/20} \cdot \frac{6n^8k^7d^{1/2}D^{1/2}}{\sigma} \cdot \varepsilon^{1/3}\right)^{\frac{m}{6}+\frac{(2z_1d+2)b-2}{3}}$$
$$= \left(\frac{6n^{317/40}k^7d^{23/40}D^{13/20}}{\sigma} \cdot \varepsilon^{31/120}\right)^{\frac{m}{6}+\frac{(2z_1d+2)b-2}{3}}.$$

As noted above,

$$\frac{m}{6} + \frac{(2z_1d+2)b-2}{3} \geq \frac{20bd}{3} > \frac{120}{31},$$

so we can use Fact 2.1 to decrease the exponent, which gives us

$$\varepsilon \cdot \left(\frac{6n^{317/40}k^7d^{23/40}D^{13/20}}{\sigma}\right)^{120/31} < \varepsilon \cdot \left(\frac{O(1) \cdot n^{317/10}k^{28}d^{23/10}D^{13/5}}{\sigma^4}\right).$$

Using these bounds, we can simplify inequality (10):

$$\begin{aligned}
&\Pr\left[\exists \mathcal{B} \in \mathbb{B} : \Lambda(\mathcal{B}) \leq \frac{6D\sqrt{nd}}{Y} \cdot \varepsilon^{1/3}\right] \\
&\leq \varepsilon \cdot n \cdot \left(\frac{O(1) \cdot n^{24}k^{21}d^{3/2}D^{3/2}}{\sigma^3}\right) + \varepsilon \cdot nk \cdot \left(\frac{O(1) \cdot n^{317/10}k^{28}d^{23/10}D^{13/5}}{\sigma^4}\right) \\
&\leq \varepsilon \cdot \left(\frac{O(1) \cdot n^{327/10}k^{29}d^{23/10}D^{13/5}}{\sigma^4}\right).
\end{aligned}$$

On the other hand $Y = k^5 \cdot \sqrt{ndD} \geq 1$, so Lemma 2.5 guarantees

$$\begin{aligned}
\Pr\left[\delta_\varepsilon \leq Y\varepsilon^{1/6}\right] &\leq \varepsilon \cdot \left(\frac{O(1) \cdot n^5 Y}{\sigma}\right)^6 \\
&= \varepsilon \cdot \left(\frac{O(1) \cdot n^{11/2}k^5d^{1/2}D^{1/2}}{\sigma}\right)^6 \\
&= \varepsilon \cdot \left(\frac{O(1) \cdot n^{33}k^{30}d^3D^3}{\sigma^6}\right).
\end{aligned}$$

Finally, we know from Lemma 3.5 that if a non-degenerate blueprint $\mathcal{B}$ results in a potential drop of at most ε, then $\delta_\varepsilon \cdot \Lambda(\mathcal{B}) \leq 6D\sqrt{nd\varepsilon}$. We must therefore have either $\delta_\varepsilon \leq Y\varepsilon^{1/6}$ or $\Lambda(\mathcal{B}) \leq \frac{6D\sqrt{nd}}{Y} \cdot \varepsilon^{1/3}$. Therefore,

$$\begin{aligned}
\Pr\big[\Delta_6 \leq \varepsilon\big] &\leq \Pr\left[\exists \mathcal{B} \in \mathbb{B} : \Lambda(\mathcal{B}) \leq \frac{6D\sqrt{nd}}{Y} \cdot \varepsilon^{1/3}\right] + \Pr\left[\delta_\varepsilon \leq Y\varepsilon^{1/6}\right] \\
&\leq \varepsilon \cdot \left(\frac{O(1) \cdot n^{327/10}k^{29}d^{23/10}D^{13/5}}{\sigma^4}\right) + \varepsilon \cdot \left(\frac{O(1) \cdot n^{33}k^{30}d^3D^3}{\sigma^6}\right) \\
&\leq \varepsilon \cdot \left(\frac{O(1) \cdot n^{33}k^{30}d^3D^3}{\sigma^6}\right),
\end{aligned}$$

which concludes the proof. □

4.7. Proof of the Main Theorem

Given the analysis of the different types of iterations, we can complete the proof that k-means has polynomial smoothed running time.

PROOF OF THEOREM 1.1. Throughout this section, we assume $d \geq 2$. The proof of Theorem 1.1 for $d = 1$ is deferred to Lemma 5.3.

Let T denote the maximum number of iterations that k-means can need on the perturbed data set $\mathcal{X}$, and let Δ denote the minimum possible potential drop over a period of four consecutive iterations. As remarked in Section 2, we can assume that all the data points lie in the hypercube $[-D/2, D/2]^d$ for $D = \sqrt{90kd \cdot \ln(n)}$, because the alternative contributes only an additive term of +1 to $\mathrm{E}[T]$.

After the first iteration, we know $\Psi \leq ndD^2$. This implies that if $T \geq 4t + 1$, then $\Delta \leq ndD^2/t$. However, in the previous sections, we proved that for $\varepsilon \in (0, 1]$,

$$\Pr[\Delta \leq \varepsilon] \leq \sum_{i=1}^{6} \Pr\big[\Delta_i \leq \varepsilon\big] \leq \varepsilon \cdot \frac{O(1) \cdot n^{33} k^{30} d^3 D^3}{\sigma^6}.$$

Recall from Section 2 that $T \leq n^{3kd}$ regardless of the perturbation. Therefore,

$$\begin{aligned}
\mathrm{E}[T] &\leq O(ndD^2) + \sum_{t=ndD^2}^{n^{3kd}} 4 \cdot \Pr\big[T \geq 4t + 1\big] \\
&\leq O(ndD^2) + \sum_{t=ndD^2}^{n^{3kd}} 4 \cdot \Pr\left[\Delta \leq \frac{ndD^2}{t}\right] \\
&\leq O(ndD^2) + \sum_{t=ndD^2}^{n^{3kd}} \frac{4ndD^2}{t} \cdot \left(\frac{O(1) \cdot n^{33} k^{30} d^3 D^3}{\sigma^6}\right) \\
&= O(ndD^2) + \left(\frac{O(1) \cdot n^{34} k^{30} d^4 D^5}{\sigma^6}\right) \cdot \left(\sum_{t=ndD^2}^{n^{3kd}} \frac{1}{t}\right) \\
&= O(ndD^2) + \left(\frac{O(1) \cdot n^{34} k^{30} d^4 D^5}{\sigma^6}\right) \cdot O(kd \cdot \ln(n)) \\
&= \frac{O(1) \cdot n^{34} k^{34} d^8 \cdot \ln^4(n)}{\sigma^6},
\end{aligned}$$

which completes the proof. □

5. A POLYNOMIAL BOUND IN ONE DIMENSION

In this section, we consider a one-dimensional set $\mathcal{X} \subseteq \mathbb{R}$ of points. The aim 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-spread* if the second-smallest distance between any two points is at least ε: For every $x_1, x_2, x_3, x_4 \in \mathcal{X}$, where $x_1 \neq x_4$ but x_2 and x_3 are possibly equal, we have $|x_1 - x_2| \geq \varepsilon$ or $|x_3 - x_4| \geq \varepsilon$. The following lemma justifies the notion of ε-spread.

LEMMA 5.1. *Assume that $\mathcal{X}$ is ε-spread. Then the potential drops by at least $\frac{\varepsilon^2}{4n^2}$ in every iteration.*

PROOF. Since $d = 1$, we can order the clusters from left to right. A cluster is called active during an iteration if it gains or loses points during this iteration. Let $\mathcal{C}_i$ be the left-most active cluster, and let $\mathcal{C}_j$ be the right-most active cluster.

We consider $\mathcal{C}_i$ first. $\mathcal{C}_i$ exchanges only points with the cluster to its right, for otherwise it would not be the leftmost active cluster. Thus, it cannot gain and lose points simultaneously. Assume that it gains points. Let A_i be the set of points of $\mathcal{C}_i$ before the iteration, and let B_i be the set of points that it gains. Obviously, $\min_{x\in B_i} x > \max_{x\in A_i} x$. If $B_i \cup A_i$ contains at least three points, then we are done: If $|A_i| \geq 2$, then we consider the two rightmost points $x_1 \leq x_2$ of A_i and the leftmost point x_3 of B_i. We have $|x_1 - x_2| \geq \varepsilon$ or $|x_2 - x_3| \geq \varepsilon$ since $\mathcal{X}$ is ε-spread. Thus, x_3 has a distance of at least $\varepsilon/2$ from the center of mass $\mathrm{cm}(A_i)$ because $\mathrm{dist}(x_1, x_3) \geq \varepsilon$, $x_1 \leq x_2 \leq x_3$, and $\mathrm{cm}(A_i) \leq (x_1 + x_2)/2$. Hence,

$$\mathrm{cm}(B_i) \geq \mathrm{cm}(A_i) + \frac{\varepsilon}{2}.$$

Thus, the cluster center moves to the right from $\mathrm{cm}(A_i)$ to

$$\begin{aligned}\mathrm{cm}(A_i \cup B_i) &= \frac{|A_i| \cdot \mathrm{cm}(A_i) + |B_i| \cdot \mathrm{cm}(B_i)}{|A_i \cup B_i|} \\ &\geq \frac{|A_i \cup B_i| \cdot \mathrm{cm}(A_i) + |B_i| \cdot \frac{\varepsilon}{2}}{|A_i \cup B_i|} \geq \mathrm{cm}(A_i) + \frac{\varepsilon}{2n}.\end{aligned}$$

The case $|A_i| = 1$ and $|B_i| \geq 2$ is analogous. The same holds if cluster $\mathcal{C}_j$ switches from A_j to $A_j \cup B_j$ with $|A_j \cup B_j| \geq 3$, or if $\mathcal{C}_i$ or $\mathcal{C}_j$ lose points but initially have at least three points. Thus, in these cases, a cluster moves by at least $\varepsilon/(2n)$, which causes a potential drop by at least $\varepsilon^2/(4n^2)$.

It remains to consider the case that $|A_i \cup B_i| = 2 = |A_j \cup B_j|$. Thus, $A_i = \{a_i\}$, $B_i = \{b_i\}$, and also $A_j = \{a_j\}$, $B_j = \{b_j\}$. We restrict ourselves to the case that $\mathcal{C}_i$ consists only of a_i and gains b_i and that $\mathcal{C}_j$ has a_j and b_j and loses b_j because all other cases can be handled analogously. If only two clusters are active, we have $b_i = b_j$, and we have only three different points. Otherwise, all four points are distinct. In both cases we have either $|a_i - b_i| \geq \varepsilon$ or $|a_j - b_j| \geq \varepsilon$ since $\mathcal{X}$ is ε-spread. But then either the center of $\mathcal{C}_i$ or the center of $\mathcal{C}_j$ moves by at least $\varepsilon/2$, which implies that the potential decreases by at least $\varepsilon^2/4 \geq \varepsilon^2/(4n^2)$. □

If $\mathcal{X}$ is ε-spread, then the number of iterations until the potential drops by at least 1 is at most $4n^2/\varepsilon^2$ by the lemma above. Let us prove that $\mathcal{X}$ is likely to be ε-spread.

LEMMA 5.2. *The probability that $\mathcal{X}$ is not ε-spread is bounded from above by $\frac{n^4\varepsilon^2}{\sigma^2}$.*

PROOF. The point set $\mathcal{X}$ is not ε-spread if there exist points $x_1, x_2, x_3, x_4 \in \mathcal{X}$, where $x_1 \neq x_4$ but $x_2 = x_3$ is allowed, with $|x_1 - x_2| < \varepsilon$ and $|x_3 - x_4| < \varepsilon$. Let an adversary fix points x_1, x_2, x_3, x_4. To avoid dependencies, let the adversary also fix the positions of x_2 and x_3. The probability that x_1 is within a distance of less than ε of x_2 is at most $\frac{2\varepsilon}{\sqrt{2\pi}\sigma} \leq \frac{\varepsilon}{\sigma}$. The probability that x_4 is within a distance of less than ε of x_3 is bounded analogously. Since the positions of x_2 and x_3 are fixed, the two events are independent. Thus, the probability that both distances are smaller than ε is at most $\left(\frac{\varepsilon}{\sigma}\right)^2$. The lemma follows now from a union bound over the at most n^4 possible choices for $x_1, \ldots, x_4$. □

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

LEMMA 5.3. *Fix an arbitrary set $\mathcal{X}' \subseteq [0,1]$ of n points and assume that each point in $\mathcal{X}'$ is independently perturbed by a normal distribution with mean 0 and standard*

deviation σ, yielding a new set $\mathcal{X}$ of points. Then the expected number of iterations of k-means on $\mathcal{X}$ is bounded by $O\big(\frac{n^7k^2\log^2 n}{\sigma^2}\big)$.

PROOF. We choose $D = \sqrt{90k\ln(n)}$. As remarked in Section 2, we can assume that the perturbed point set is a subset of the interval $[-D/2, D/2]$. The alternative contributes an additive term of 1 to the expected number of iterations.

If all points are in the interval $[-D/2, D/2]$, then the potential after the first iteration is bounded from above by nD^2. Let T be the random variable of the number of iterations. If $T \geq t+2$, then $\mathcal{X}$ cannot be ε-spread with $4n^3D^2/\varepsilon^2 \leq t$. Thus, $\mathcal{X}$ cannot be ε-spread with $\varepsilon = 2Dn^{3/2}/\sqrt{t}$. As k-means runs for at most n^{3k} iterations,

$$\begin{aligned}
\mathrm{E}\left[T\right] &= 2 + \sum_{t=1}^{n^{3k}} \Pr\big[T \geq t+2\big] \leq 2 + \sum_{t=1}^{n^{3k}} \Pr\left[\mathcal{X} \text{ is not } \frac{2Dn^{3/2}}{\sqrt{t}}\text{-spread}\right] \\
&\leq 2 + \sum_{t=1}^{n^{3k}} \frac{4D^2n^7}{t\sigma^2} = \sum_{t=1}^{n^{3k}} O\left(\frac{kn^7\log n}{t\sigma^2}\right) \\
&= O\left(\frac{kn^7\log n}{\sigma^2}\cdot\log n^{3k}\right) = O\left(\frac{n^7k^2\log^2 n}{\sigma^2}\right) .
\end{aligned}$$

□

6. CONCLUDING REMARKS

In this paper, we settled the smoothed running time of the k-means method for arbitrary k and d. The exponents in our smoothed analysis are constant but large. We did not make a huge effort to optimize the exponents as the arguments are intricate enough even without trying to optimize constants. Furthermore, we believe that our approach, which is essentially based on bounding the smallest possible improvement in a single step, is too pessimistic to yield a bound that matches experimental observations. A similar phenomenon occurred already in the smoothed analysis of the 2-opt heuristic for the TSP [Englert et al. 2007]. There it was possible to improve the bound for the number of iterations by analyzing sequences of consecutive steps rather than single steps. It is an interesting question if this approach also leads to an improved smoothed analysis of k-means.

Squared Euclidean distances, while most natural, are not the only distance measure used for k-means clustering. The k-means method can be generalized to arbitrary Bregman divergences [Banerjee et al. 2005]. Bregman divergences include the Kullback-Leibler divergence, which is used, e.g., in text classification, or Mahalanobis distances. Due to its role in applications, k-means clustering with Bregman divergences has attracted a lot of attention recently [Ackermann and Blömer 2009; Ackermann et al. 2010]. Recently, upper bounds of $\mathrm{poly}(n^{\sqrt{k}}, 1/\sigma)$ as well as $k^{kd}\cdot\mathrm{poly}(n, 1/\sigma)$ have been shown for the smoothed running time of k-means with (almost arbitrary) Bregman divergences [Manthey and Röglin 2009b]. However, it remains open if the polynomial bound can be transferred to general Bregman divergences as well.

REFERENCES

ACKERMANN, M. R. AND BLÖMER, J. 2009. Coresets and approximate clustering for Bregman divergences. In *Proc. of the 20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*. 1088–1097.

ACKERMANN, M. R., BLÖMER, J., AND SOHLER, C. 2010. Clustering for metric and nonmetric distance measures. *ACM Transactions on Algorithms 6,* 4.

ALOISE, D., DESHPANDE, A., HANSEN, P., AND POPAT, P. 2009. NP-hardness of euclidean sum-of-squares clustering. *Machine Learning 75,* 2, 245–248.

ARTHUR, D., MANTHEY, B., AND RÖGLIN, H. 2009. k-means has polynomial smoothed complexity. In *Proc. of the 50th Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*. 405–414.

ARTHUR, D. AND VASSILVITSKII, S. 2006. How slow is the k-means method? In *Proc. of the 22nd ACM Symp. on Computational Geometry (SoCG)*. 144–153.

ARTHUR, D. AND VASSILVITSKII, S. 2007. k-means++: The advantages of careful seeding. In *Proc. of the 18th ACM-SIAM Symp. on Discrete Algorithms (SODA)*. 1027–1035.

ARTHUR, D. AND VASSILVITSKII, S. 2009. 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.

BĂDOIU, M., HAR-PELED, S., AND INDYK, P. 2002. Approximate clustering via core-sets. In *Proc. of the 34th Ann. ACM Symp. on Theory of Computing (STOC)*. 250–257.

BANERJEE, A., MERUGU, S., DHILLON, I. S., AND GHOSH, J. 2005. Clustering with Bregman divergences. *Journal of Machine Learning Research 6*, 1705–1749.

BECCHETTI, L., LEONARDI, S., MARCHETTI-SPACCAMELA, A., SCHÄFER, G., AND VREDEVELD, T. 2006. Average case and smoothed competitive analysis of the multilevel feedback algorithm. *Mathematics of Operations Research 31,* 1, 85–108.

BEIER, R. AND VÖCKING, B. 2004. Random knapsack in expected polynomial time. *Journal of Computer and System Sciences 69,* 3, 306–329.

BERKHIN, P. 2002. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, USA.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2000. *Pattern Classification*. John Wiley & Sons.

DURRETT, R. 1991. *Probability: Theory and Examples*. Duxburry Press.

ENGLERT, M., RÖGLIN, H., AND VÖCKING, B. 2007. Worst case and probabilistic analysis of the 2-Opt algorithm for the TSP. In *Proc. of the 18th ACM-SIAM Symp. on Discrete Algorithms (SODA)*. 1295–1304.

GADDAM, S. R., PHOHA, V. V., AND BALAGANI, K. S. 2007. K-Means+ID3: A novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods. *IEEE Transactions on Knowledge and Data Engineering 19,* 3, 345–354.

HAR-PELED, S. AND SADRI, B. 2005. How fast is the k-means method? *Algorithmica 41,* 3, 185–202.

INABA, M., KATOH, N., AND IMAI, H. 2000. Variance-based k-clustering algorithms by Voronoi diagrams and randomization. *IEICE Transactions on Information and Systems E83-D,* 6, 1199–1206.

KANUNGO, T., MOUNT, D. M., NETANYAHU, N. S., PIATKO, C. D., SILVERMAN, R., AND WU, A. Y. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24,* 7, 881–892.

KANUNGO, T., MOUNT, D. M., NETANYAHU, N. S., PIATKO, C. D., SILVERMAN, R., AND WU, A. Y. 2004. A local search approximation algorithm for k-means clustering. *Computational Geometry: Theory and Applications 28,* 2-3, 89–112.

KUMAR, A., SABHARWAL, Y., AND SEN, S. 2004. 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 Computer Science (FOCS)*. 454–462.

LLOYD, S. P. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory 28,* 2, 129–137.

MAHAJAN, M., NIMBHORKAR, P., AND VARADARAJAN, K. 2009. The planar k-means problem is NP-hard. In *Proc. of the 3rd Int. Workshop on Algorithms and Computation (WALCOM)*. Lecture Notes in Computer Science Series, vol. 5431. Springer, 274–285.

MANTHEY, B. AND RÖGLIN, H. 2009a. Improved smoothed analysis of the k-means method. In *Proc. of the 20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*. 461–470.

MANTHEY, B. AND RÖGLIN, H. 2009b. Worst-case and smoothed analysis of k-means clustering with Bregman divergences. In *Proc. of the 20th Int. Symp. on Algorithms and Computation (ISAAC)*. Lecture Notes in Computer Science Series, vol. 5878. Springer, 1024–1033.

MATOUŠEK, J. 2000. On approximate geometric k-clustering. *Discrete and Computational Geometry 24,* 1, 61–84.

OSTROVSKY, R., RABANI, Y., SCHULMAN, L., AND SWAMY, C. 2006. The effectiveness of Lloyd-type methods for the k-means problem. In *Proc. of the 47th Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*. 165–176.

SPIELMAN, D. A. AND TENG, S.-H. 2004. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM 51,* 3, 385–463.

VATTANI, A. to appear. k-means requires exponentially many iterations even in the plane. *Discrete & Computational Geometry*.

VERSHYNIN, R. 2009. Beyond Hirsch conjecture: Walks on random polytopes and smoothed complexity of the simplex method. *SIAM Journal on Computing 39,* 2, 646–678.

WAGSTAFF, K. L., CARDIE, C., ROGERS, S., AND SCHRÖDL, S. 2001. Constrained k-means clustering with background knowledge. In *Proc. of the 18th International Conference on Machine Learning (ICML)*. Morgan Kaufmann, 577–584.

WARREN, H. E. 1968. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society 133,* 1, 167–178.

Received ; revised ; accepted