ON THE NUMBER OF STEPS OF LLOYD'S K-MEANS METHOD

BY

BARDIA SADRI

B.S., Sharif University of Technology, 1999

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2004

Urbana, Illinois

Abstract

This thesis presents polynomial upper and lower bounds on the number of iterations performed by Lloyd's method for k-means clustering. The upper bounds are polynomial in the number of points, number of clusters, and the spread of the point set. The presented lower bound shows that in the worst case the k-means heuristic needs to perform $\Omega(n)$ iterations, for n points on the real line and two centers. Surprisingly, the spread of this lower bound construction is *polynomial.* This is the first construction showing that the k-means heuristic requires more than a polylogarithmic number of iterations. Furthermore, two alternative algorithms with guaranteed performances are presented, which are simple variants of Lloyd's method. Results of experimental studies on these algorithms are also presented.

To Azadeh,

Acknowledgments

I owe many thanks to Sariel Har-Peled for doing far more than only supervising this research. I would also like to thank Edgar Ramos for useful discussions as well as for the support of this research. I am also in debt to Boris Aronov for his careful and thorough examining of the manuscript and his many constructive suggestions.

Finally, I must thank Pankaj K. Agarwal and David Mount for inspiring comments on the problems studied in this thesis and other related problems. In particular, David Mount provided us with the test data sets used in [KMN$^+$02].

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Clustering

In a general and informal sense, the purpose of *clustering* is to group a given set of data elements into *clusters* in such a way that elements in each cluster more or less resemble each other while elements from different clusters are substantially dissimilar. Resemblance and difference are defined to suit the nature of data and the sought for application of clustering. By strictly specifying the type of the considered elements, the way in which similarity and difference are measured, and the desired form of optimality in classification, numerous members of this loosely defined family of clustering problems are composed. Many of these problems are of interest in various scientific and computational disciplines and are widely studied from both theoretical and empirical perspectives.

Consider for example the following real-world clustering problem. Suppose we need to assign telephone area codes to towns and cities in a country. For each area code a central switching facility has to be built and be connected to the switching centers of all towns with that area code. Suppose our budget allows us to build only 20 such switching centers. To minimize the amount of wiring job needed and perhaps many other reasons, we like the total distance of the central switching facility in each area code to the switching centers of the town with that area code to be as small as possible. In fact, we wish to place central facilities and assign towns to them in such a way to do as little total wiring job as we can. We can formulate all this into a the following clustering problem.

Given a map showing the location of switching centers of the towns, we wish to group the towns into 20 *clusters, assigning each group a distinct area code and a location for a central switching facility, in such a way that the the total distance between the town switching centers to their corresponding central switching facility is minimized.*

As is the case in this example, clustering is often an *optimization* problems: there is a way to quantify the quality of any given clustering and the goal is to find the best, minimizing or maximizing the given quality function. It is almost always the case that there is an underlying notion of distance between pairs of individual data elements based on which the clustering quality function is defined. In *geometric* clustering problems, the data elements are points in a space endowed with a notion of distance, i.e. a metric space.

The desired number of clusters may or may not be known in advance and thus, accordingly, may or may not be considered as part of the input to the clustering problem. In certain applications, the number of clusters is to be determined by the clustering algorithm and the final number of clusters returned can be a parameter in measuring the quality of the clustering. However, the clustering problems studied in this thesis does not belong to this group and the number of desired clusters is always an input parameter.

1.2 Geometric Clustering

In this section, we formalize the notion of *geometric clustering*. It is important to notice that the definition of "geometric clustering" given below does not fully encompass the family of problems suggested by this term.

Definition 1.2.1 In a geometric *clustering* problem, we are given a finite set $X \subset \Re^d$ of n points and an integer $k \geq 2$, and we seek a partition (clustering) $\mathcal{S} = (S_1, \ldots, S_k)$ of

X into k disjoint nonempty subsets along with a set $C = \{c_1, \ldots, c_k\}$ of k corresponding centers, that minimizes a suitable cost function among all such k-clusterings of X. The cost function typically represents how tightly each cluster is packed and how separated different clusters are. A center c_i is said to *serve* the points in its cluster S_i.

Below we mention three of the most widely investigated classical geometric clustering problems, characterized and distinguished from each other only by their cost functions. The notation $\|\cdot\|$ denotes the Euclidean (ℓ_2) distance.

k-center. The goal in this clustering problem is to minimize, among all clusters, the maximum distance of a point in the cluster to the center of the cluster. More formally, the clustering cost function $\phi_\infty(\mathcal{S}, C)$ of k-center is

$$\phi_\infty(\mathcal{S}, C) = \max_i \max_{x \in S_i} \|x - c_i\|.$$

k-median. The sum of distances between the points and their corresponding centers is to be minimized. Thus, the clustering cost function is the following.

$$\phi_1(\mathcal{S}, C) = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - c_i\| .$$

k-means. Similar to k-median with the difference that that squares of distances are considered. Thus we seek to minimize the sum of the *squares* of distances between the points and their corresponding centers.

$$\phi_2(\mathcal{S}, C) = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - c_i\|^2 .$$

Throughout this thesis, we only consider the *k-means* clustering. Thus our clustering cost function will be

$$\phi_2(\mathcal{S}, C) = \sum_{i=1}^{k} \psi(S_i, c_i),$$

where $\psi(S,c) = \sum_{x\in S} \|x-c\|^2$. It can be easily observed that for any cluster S_i, the point c that minimizes the sum $\sum_{x\in S_i} \|x-c\|^2$, is the centroid of S_i, which we shall denote by $c(S_i)$, and therefore in an optimal clustering, $c_i = c(S_i)$. Thus the above cost function can be written as

$$\phi_2(\mathcal{S}) = \sum_{i=1}^{k} \sum_{x\in S_i} \|x - c(S_i)\|^2 .$$

It can also be observed that in an optimal k-clustering, each point of S_i is closer to the center corresponding to S_i than to any other center. Thus, an optimal k-clustering is imposed by a Voronoi diagram whose sites are the centroids of the clusters. Such partitions are related to centroidal Voronoi tessellations, (see [DFG99]).

1.3 Lloyd's Method for k-means

A k-means clustering algorithm that is used widely because of its simplicity and ease of implementation is the *k-means heuristic*, also called *Lloyd's method*. This algorithm starts with an arbitrary k-clustering $\mathcal{S}_0$ of X with the initial k centers chosen to be the centroids of the clusters of $\mathcal{S}_0$. Then it repeatedly performs local improvements by applying the following "hill-climbing" step.

Definition 1.3.1 Given a clustering $\mathcal{S} = (S_1, \ldots, S_k)$ of X, a k-MEANS *step* returns a clustering $\mathcal{S}' = (S'_1, \ldots, S'_k)$ by letting S'_i equal to the intersection of X with the Voronoi cell of $c(S_i)$ in the Voronoi partitioning imposed by centers $c(S_1), \ldots, c(S_k)$. The (new) center of S'_i will be $c(S'_i)$. In a clustering $\mathcal{S} = (S_1, \ldots, S_k)$ of X, a point $x \in X$ is *misclassified* if there exists $1 \le i \ne j \le k$, such that $x \in S_i$ but $\|x - c(S_j)\| < \|x - c(S_i)\|$. Thus a k-MEANS step can be broken into two stages: (i) every misclassified point is assigned to its closest center, and (ii) Centers are moved to the centroids of their newly formed clusters (see Figure 1.1).

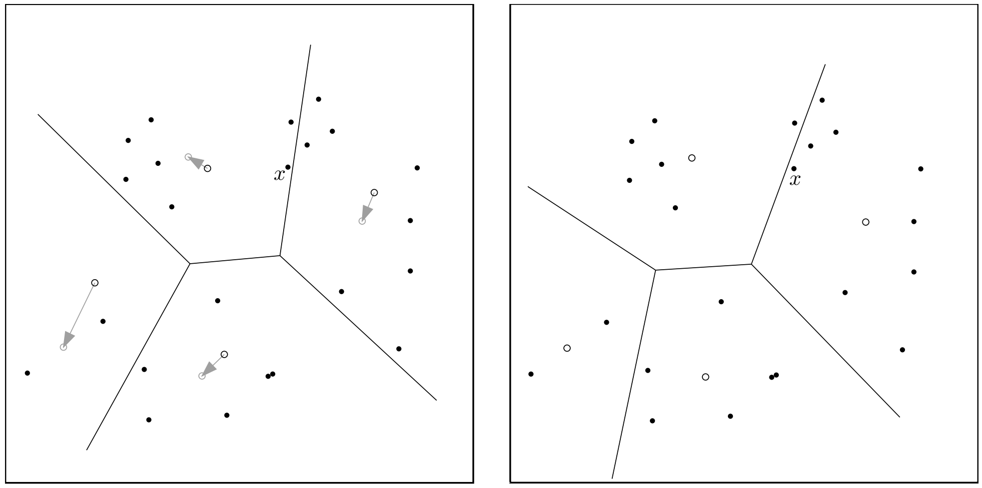

Figure 1.1: A k-Means step.
Small circles represent the centers and clusters are determined by their Voronoi diagram. On the left, the arrows point from the current centers to to the centroids of the clusters. The point indicated as x becomes misclassified when centers move to centroid of their clusters and is thus reassigned on the right.

Lloyd's algorithm, to which we shall refer as "k-MeansMtd" throughout this thesis, performs the k-Means step repeatedly and stops when the assignment of the points to the centers does not change from that of the previous step. This happens when there remains no misclassified points and consequently in the last k-Means step $\mathcal{S}' = \mathcal{S}$. Clearly the clustering cost is reduced when each point is mapped to the closest center and also when each center moves to the centroid of the points it serves. Thus, the clustering cost is strictly reduced in each of the two stages of a k-Means step. This in particular implies that no clustering can be seen twice during the course of execution of k-MeansMtd. Since there are only finitely many k-clusterings, the algorithm terminates in finite time.

A slight technical detail involves the event of a center losing all the points it serves. The original k-means heuristic does not specify a particular solution to this problem. Candidate strategies used in practice include: placing the lonely center somewhere else arbitrarily or randomly, leaving it where it is to perhaps acquire some points in futures steps, or completely removing it. For the sake of convenience in our analysis, we adopt the last strategy, namely, whenever a center is left serving no points, we remove that center permanently and continue with the remaining centers.

k-MEANSMTD and its variants are widely used in practice [DHS01]. It is known that the output of k-MEANSMTD is not necessarily a global minimum, and it can be arbitrarily bad compared to the optimal and clustering and it is well known that the answer returned by the algorithm and the number of steps depends on the initial choice of the centers, i.e. the initial clustering [KMN$^+$02]. These shortcomings of k-MEANSMTD has lead to development of efficient polynomial approximation schemes for the k-means clustering problem both in low [Mat00, ES03, HPM] and high dimensions [dKKR03]. Unfortunately, those algorithms have had little impact in practice, as they are complicated and probably impractical because of large constants. A local search algorithm based on k-MEANS steps was suggested by Kanungo *et al.* [KMN$^+$02], which yields a constant-factor approximation, and it seems to perform reasonably well in practice.

Up to this point, no meaningful theoretical bound was known for the number of steps k-MEANSMTD can take to terminate in the worst case. Inaba *et al.* [IKI94] observe that the number of distinct Voronoi partitions of a given n-point set $X \subset \Re^d$ induced by k sites is at most $O(n^{kd})$ which gives a trivial similar upper bound on the number of steps of k-MEANSMTD considering the fact that clustering cost monotonically decreases and as a result no k-clustering can be seen twice. However, the fact that k in typical application can be in the hundreds together with the relatively fast convergence of k-MEANSMTD observed in practice, make this bound somewhat meaningless. The difficulty of proving any super-linear lower bound further suggests the looseness of this bound.

It thus appears that the combinatorial behavior of k-MEANSMTD is far from being well understood. This thesis provides a lower bound and several upper bounds on the number of iterations performed by k-MEANSMTD and some close variants. To our knowledge, our lower bound is the first that is super-polylogarithmic. Our upper bounds are *polynomial* in the spread Δ of the input point set, k, and n (the *spread* of a point set is the ratio between its diameter and the distance between its closest pair). The bounds are meaningful for most inputs.

In Chapter 2, we present an $\Omega(n)$ lower bound on the number of iterations performed by k-MeansMtd. More precisely, we show that for an adversarially chosen initial two centers and a set of n points on the line, k-MeansMtd takes $\Omega(n)$ steps. Note, that this matches the trivial upper bound in $1d$, as the number of Voronoi partitions in one dimension with two centers is linear.

In Chapter 3, we provide a polynomial upper bound for the one-dimensional case. Chapter 4 presents an an upper bound for the case where the points lie on a grid. In Chapter 5, we investigate two alternative related algorithms, and provide polynomial upper bounds on the number of iterations they perform. Some experimental results are presented in Chapter 6. In Chapter 7, we conclude by mentioning a few open problems.

2 A Linear Lower Bound

2.1 A One-Dimensional Construction for Two Clusters

In this section, we describe a set of $2n$ points, along with an initial pair of centers, on which k-MeansMtd takes $\Omega(n)$ steps to terminate for $n \geq 2$.

Fix $n \geq 2$. Our set X will consist of $2n$ numbers

$$y_1 < \cdots < y_n < x_n < \cdots < x_1$$

with $y_i = -x_i$, for $i = 1, \ldots, n$.

At the ith iteration, we denote by l_i and r_i the current left and right centers, respectively, and by L_i and R_i the new sets of points assigned to l_i and r_i, respectively. Furthermore, for each $i \geq 0$, we denote by α_i the Voronoi boundary $\frac{1}{2}(l_i + r_i)$ between the centers l_i and r_i. Thus

$$L_i = \{x \in X \mid x < \alpha_i\} \qquad \text{and} \qquad R_i = \{x \in X \mid x \geq \alpha_i\}.$$

Let x_1 be an arbitrary positive real number and let $x_2 < x_1$ be a positive real number to be specified shortly. Initially, we let $l_1 = x_2$ and $r_1 = x_1$ and consequently $\alpha_1 = \frac{1}{2}(x_1 + x_2)$. Thus in the first iteration, $L_1 = \{y_1, \ldots, y_n, x_n, \ldots, x_2\}$ and $R_1 = \{x_1\}$. We will choose $x_2, \ldots, x_n$ such that at the end of the ith step we have $L_i = \{y_1, \ldots, y_n, x_n, \ldots, x_{i+1}\}$ and $R_i = \{x_i, \ldots, x_1\}$ (see Figure 2.1). Suppose for the inductive hypothesis that at the $(i-1)$th step we have

$$L_{i-1} = \{y_1, \ldots, y_n, x_n, \ldots, x_{i+1}, x_i\} \qquad \text{and} \qquad R_{i-1} = \{x_{i-1}, \ldots, x_1\}\,.$$

Thus we can compute l_i and r_i as follows

$$l_i = \frac{y_1 + \cdots + y_n + x_n + \cdots + x_i}{2n - i + 1} \qquad \text{and} \qquad r_i = \frac{x_{i-1} + \cdots + x_1}{i-1}.$$

Since $y_1 + \cdots + y_n + x_n + \cdots + x_i = -(x_{i-1} + \cdots + x_1)$, we get for α_i:

$$\begin{aligned} \alpha_i = \frac{1}{2}(l_i + r_i) &= \frac{1}{2}\left(\frac{x_{i-1} + \cdots + x_1}{i-1} - \frac{x_{i-1} + \cdots + x_1}{2n-i+1}\right) \\ &= \frac{n-i+1}{(i-1)(2n-i+1)}(x_{i-1} + \cdots + x_1) \\ &= \frac{n-i+1}{(i-1)(2n-i+1)} \cdot s_{i-1}, \end{aligned}$$

where $s_{i-1} = \sum_{j=1}^{i-1} x_j$.

To guarantee that only x_i deserts from L_{i-1} to R_i, in the ith iteration, we need that $x_{i+1} < \alpha_i < x_i$. Thus, it is natural to set $x_i = \tau_i \alpha_i$, where $\tau_i > 1$, for $i = 1, \ldots, n$. Picking the coefficients $\tau_1, \ldots, \tau_n$ is essentially the only part of this construction that is under our control. We set

$$\tau_i = 1 + \frac{1}{n-i+1} = \frac{n-i+2}{n-i+1},$$

for $i = 1, \ldots, n$. Since $\tau_i > 1$, $x_i = \tau_i \alpha_i > \alpha_i$, for $i = 1, \ldots, n$. Next, we verify that

$x_{i+1} < \alpha_i$. By definition,

$$\begin{aligned}
x_{i+1} &= \tau_{i+1}\alpha_{i+1} \\
&= \tau_{i+1} \cdot \frac{n-i}{i(2n-i)} \cdot s_i \\
&= \tau_{i+1} \cdot \frac{n-i}{i(2n-i)} \cdot (x_i + s_{i-1}) \\
&= \tau_{i+1} \cdot \frac{n-i}{i(2n-i)} \left(\tau_i^{-1} + \frac{(i-1)(2n-i+1)}{n-i+1} \right) \alpha_i \\
&= \frac{n-i+1}{i(2n-i)} \left(\frac{n-i+1}{n-i+2} + \frac{(i-1)(2n-i+1)}{n-i+1} \right) \alpha_i.
\end{aligned}$$

It can be verified through elementary simplifications that the coefficient of α_i above is always less than 1 implying that $x_{i+1} < \alpha_i < x_i$, for $i = 1, \ldots, n-1$.

We can compute a recursive formula for x_{i+1} in terms of x_i, as follows

$$\begin{aligned}
x_{i+1} &= \tau_{i+1}\alpha_{i+1} \\
&= \frac{n-i+1}{n-i} \cdot \frac{n-i}{i(2n-i)} \cdot s_i \\
&= \frac{n-i+1}{i(2n-i)} \cdot (x_i + s_{i-1}) \\
&= \frac{n-i+1}{i(2n-i)} \left(x_i + \frac{(i-1)(2n-i+1)}{n-i+1} \cdot \alpha_i \right) \\
&= \frac{n-i+1}{i(2n-i)} \left(x_i + \frac{(i-1)(2n-i+1)}{n-i+1} \left(1 + \frac{1}{n-i+1} \right)^{-1} x_i \right) \\
&= \frac{n-i+1}{i(2n-i)} \left(1 + \frac{(i-1)(2n-i+1)}{n-i+1} \left(\frac{n-i+1}{n-i+2} \right) \right) x_i. \\
&= \frac{n-i+1}{i(2n-i)} \left(1 + \frac{(i-1)(2n-i+1)}{n-i+2} \right) \cdot x_i,
\end{aligned}$$

for $i = 1, \ldots, n-1$. Thus letting

$$c_i = \frac{n-i+1}{i(2n-i)} \left(1 + \frac{(i-1)(2n-i+1)}{n-i+2} \right),$$

we get that

$$x_{i+1} = c_i x_i, \tag{2.1}$$

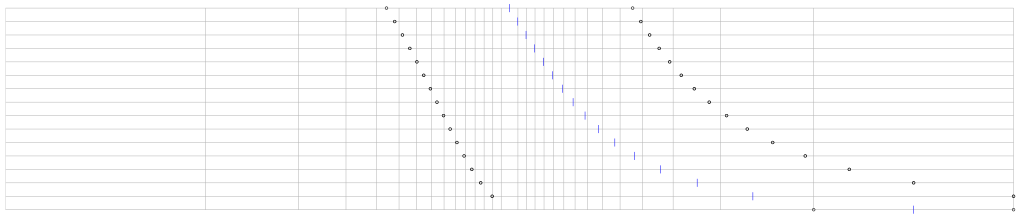

Figure 2.1: The linear lower bound construction.
Vertical lines correspond to points and horizontal lines represent steps (from bottom to top). The small circles stand for centers at each step and the blue small vertical segments shows the Voronoi boundary location.

for $i = 1, \ldots, n-1$.

Theorem 2.1.1 *For each $n \geq 2$, there exists a set of $2n$ points on a line with two initial center positions for which k-MeansMtd takes exactly n steps to terminate.*

2.2 The Spread of the Point Set

It is interesting to examine the spread of the above construction. In particular, somewhat surprisingly, the spread of this construction is polynomial, hinting (at least intuitively) that "bad" inputs for k-MeansMtd are not that contrived.

By Eq. (2.1), we have $x_{i+1} = c_i x_i$. Notice that by the given construction $c_i < 1$ for all $i = 1, \ldots, n-1$ since $x_{i+1} < x_i$. In the sequel we will show that x_n is only polynomially smaller than x_1, namely $x_n = \Omega(x_1/n^4)$. We then derive a bound on the distance between any consecutive pair x_i and x_{i+1}. These two assertions combined, imply that the point set has a spread bounded by $O(n^5)$. The following lemma follows from elementary algebraic simplifications.

Lemma 2.2.1 *For each $1 \leq i \leq n/2$, it holds that*

$$c_i \geq \left(1 - \frac{1}{i}\right)^2,$$

and for each $n/2 < i < n-1$, it holds that

$$c_i \geq \left(1 - \frac{1}{(n-i+1)}\right)^2.$$

Furthermore, for $i \geq 2$, we have

$$c_i \leq \left(1 - \frac{1}{2i}\right).$$

Corollary 2.2.2 *For any $n > 0$ we have $x_n = \Omega(x_1/n^4)$.*

Proof.

$$\begin{aligned}
x_n &= c_1 \cdot \prod_{i=2}^{n-1} c_i \cdot x_1 \\
&\geq c_1 x_1 \cdot \prod_{i=2}^{\lfloor n/2 \rfloor} \left(1 - \frac{1}{i}\right)^2 \cdot \prod_{i=\lfloor n/2 \rfloor+1}^{n-1} \left(1 - \frac{1}{n-i}\right)^2 \\
&= c_1 x_1 \cdot \left(1 - \frac{1}{2}\right)^2 \cdots \left(1 - \frac{1}{\lfloor n/2 \rfloor}\right)^2 \cdot \left(1 - \frac{1}{\lfloor n/2 \rfloor}\right)^2 \cdots \left(1 - \frac{1}{2}\right)^2 \\
&= c_1 x_1 \cdot \left(\prod_{i=2}^{\lfloor n/2 \rfloor} \frac{(i-1)^2}{i^2}\right)^2 \\
&= c_1 x_1 \cdot \left(\frac{1}{\lfloor n/2 \rfloor}\right)^4
\end{aligned}$$

The claim follows as $c_1 = n/(2n-1) = \Theta(1)$. ■

Lemma 2.2.3 *For each $i = 1, \ldots, n-1$, $x_i - x_{i+1} \geq x_i/3i$.*

Proof. Since $x_{i+1} = c_i x_i$, we have $x_i - x_{i+1} = x_i(1 - c_i)$. For $i = 1$, we have $c_1 = n/(2n-1) \leq 2/3$, when $n \geq 2$. Therefore we have $x_1 - x_2 \geq x_1/3$. For $i = 2, \ldots, n-1$, using Lemma 2.2.1 we get $1 - c_i \geq 1/2i$. Thus,

$$x_i - x_{i+1} = x_i(1 - c_i) \geq x_i \cdot \frac{1}{2i} > \frac{x_i}{3i},$$

as claimed. ■

Theorem 2.2.4 *The spread of the point set constructed in Theorem 2.1.1 is $O(n^5)$.*

Proof. By Lemma 2.2.3, for each $i = 1, \ldots, n-1$, $x_i - x_{i+1} \geq x_i/3i$. Since $x_i > x_n$ and by Corollary 2.2.2, $x_n = \Omega(x_1/n^4)$, it follows that $x_i - x_{i+1} = \Omega(x_1/n^5)$. This lower bound for the distance between two consecutive points is also true for y_i's due to the symmetric construction of the point set around 0. On the other hand, since $x_n = \Omega(x_1/n^4)$,

$$x_n - y_n = 2x_n = \Omega(x_1/n^4).$$

Thus every pair of points are at distance at least $\Omega(x_1/n^5)$. Since the diameter of the point set is $2x_1$, we get a bound of $O(n^5)$ for the spread of the point set. ■

3 An Upper Bound for One Dimension

In this section, we prove an upper bound on the number of steps of k-MeansMtd in one dimensional Euclidean space. As we shall see, the bound does not involve k but is instead related to the spread Δ of the point set X. Without loss of generality we can assume that the closest pair of points in X are at distance 1 and thus the diameter of the set X is Δ. Before proving the upper bound, we mention a technical lemma from [KMN$^+$02]. The proof is included for completeness.

Lemma 3.0.5 ([KMN$^+$02]) *Let S be a set of points in $\Re^d$ with centroid $c = c(S)$ and let z be an arbitrary point in $\Re^d$. Then*

$$\psi(S,z) - \psi(S,c) = \sum_{x \in S} \Big(\|x-z\|^2 - \|x-c\|^2 \Big) = |S| \cdot \|c-z\|^2 .$$

Proof. Using $\langle u, v \rangle$ to denote inner product of vectors u and v, we have

$$\begin{aligned}
\psi(S,z) - \psi(S,c) &= \sum_{x \in S} \Big(\|x-z\|^2 - \|x-c\|^2 \Big) \\
&= \sum_{x \in S} (\langle x-z, x-z \rangle - \langle x-c, x-c \rangle) \\
&= \sum_{x \in S} (\langle x, x \rangle - 2\langle x, z \rangle + \langle z, z \rangle - (\langle x, x \rangle - 2\langle x, c \rangle + \langle c, c \rangle)) \\
&= |S|\langle z, z \rangle - |S|\langle c, c \rangle + 2\sum_{x \in S} (\langle x, c \rangle - \langle x, z \rangle) \\
&= |S|\langle z, z \rangle - |S|\langle c, c \rangle + 2\sum_{x \in S} \langle x, c-z \rangle .
\end{aligned}$$

Since

$$\sum_{x \in S} \langle x, c - z \rangle = \left\langle \sum_{x \in S} x, c - z \right\rangle = \langle |S| \, c, c - z \rangle = |S| \langle c, c - z \rangle \, ,$$

we get

$$\begin{aligned} |S| \langle z, z \rangle - |S| \langle c, c \rangle + 2 \sum_{x \in S} \langle x, c - z \rangle &= |S| \langle z, z \rangle - |S| \langle c, c \rangle + 2 |S| \langle c, c - z \rangle \\ &= |S| \langle z, z \rangle - \langle c, c \rangle - 2 \langle c, z \rangle + 2 \langle c, c \rangle \\ &= |S| \langle c - z, c - z \rangle \\ &= |S| \cdot \|c - z\|^2 \, . \end{aligned}$$

■

The above lemma quantifies the contribution of a center c_i to the cost improvement in a k-Means step as a function of the distance it moves. More formally, if in a k-Means step a k-clustering $\mathcal{S} = (S_1, \ldots, S_k)$ is changed to the other k-clustering $\mathcal{S}' = (S'_1, \ldots, S'_k)$, then

$$\phi(\mathcal{S}') - \phi(\mathcal{S}) \geq \sum_{i=1}^{k} |S'_i| \cdot \left\| c(S'_i) - c(S_i) \right\|^2 .$$

Note that in the above analysis we only consider the improvement resulting from the second stage of k-Means step in which the centers are moved to the centroids of their clusters. There is an additional gain from reassigning the points in the first stage of a k-Means step that we currently ignore.

In all our upper bound arguments we use the fact that if the initial set of centers is chosen from inside the convex hull of the input point set X (even if this is not the case, all centers move inside the convex hull of X after one step), the initial clustering cost is no more than $n\Delta^2$. This simply follows from the fact that each of the n points in X is at distance no more than Δ from its assigned center.

Theorem 3.0.6 *The number of steps of k-MeansMtd on a set $X \subset \Re$ of n points with spread Δ is at most $O(n\Delta^2)$.*

Proof. Consider a k-Means step that changes a k-clustering $\mathcal{S}$ into another k-clustering $\mathcal{S}'$. The crucial observation is that in this step, there exists a cluster that is only extended or shrunk from its right end. To see this consider the leftmost cluster S_1. Either S_1 is modified in this step, in which case this modification can only happen in form of extension or shrinking at its right end, or it remains the same. In the latter case, the same argument can be made about S_2, and so on.

Thus assume that S_1 is extended on right by receiving a set T from the cluster directly to its right, namely S_2 (S_2 cannot lose all its points to S_1 as it has at least one point to the right of c_2 and this point is closer to c_2 than to c_1 and cannot go to S_1). Notice that $c(T)$ is to the right of the leftmost point in T and at distance at least $(|T|-1)/2$ from this leftmost point (because every pair of points are at distance one or more in T and $c(T)$ gets closest to its leftmost point when every pair of consecutive points in T are placed at the minimum distance of 1 from each other). Similarly, the centroid of S_1 is to the left of the rightmost point of S_1 and at distance at least $(|S_1|-1)/2$ from it. Thus,

$$\|c(S_1) - c(T)\| \geq \frac{|T|-1}{2} + \frac{|S_1|-1}{2} + 1 = \frac{|T|+|S_1|}{2},$$

where the extra 1 is added because the distance between the leftmost point in T and the rightmost point in S_1 is at least 1. The centroid of S_1' will therefore be at distance

$$\frac{|T|}{|S_1|+|T|}\|c(S_1) - c(T)\| \geq \frac{|T|}{|S_1|+|T|} \cdot \frac{|T|+|S_1|}{2} = \frac{|T|}{2} \geq \frac{1}{2}$$

from $c(S_1)$ and to its right. Consequently, by Lemma 3.0.5, the improvement in clustering cost is at least $1/4$.

Similar analysis implies a similar improvement in the clustering cost for the case where we remove points from S_1. Since the initial clustering cost is at most $n\Delta^2$, the number of steps is no more than $n\Delta^2/(1/4) = 4n\Delta^2$. ∎

Remark 3.0.7 The upper bound of Theorem 3.0.6 as well as all other upper bounds proved later in this thesis can be slightly improved by observing that at the end of any

k-MEANS step (or a substitute step used in the alternate algorithms considered later), we have a clustering $\mathcal{S} = (S_1, \ldots, S_k)$ of the input point set X with centers $c_1, \ldots, c_k$, respectively, where for each $i = 1, \ldots, k$, $c_i = c(S_i)$. Let $\hat{c} = c(X)$. By Lemma 3.0.5 we can write for each $i = 1, \ldots, k$

$$\psi(S_i, c_i) = \psi(S_i, \hat{c}) - |S_i| \cdot \|\hat{c} - c_i\|^2 .$$

Summing this equation up for every $i = 1, \ldots, k$,

$$\phi(\mathcal{S}') = \sum_{x \in X} \|x - \hat{c}\|^2 - \sum_{i=1}^{k} |S_i| \,.\, \|\hat{c} - c_i\|^2 < \sum_{x \in X} \|\hat{c} - x\|^2 = \frac{1}{n} \sum_{x,y \in X} \|x - y\|^2 ,$$

we get the better upper bound of $1/n \sum_{x,y \in X} \|x - y\|^2$ that can replace the trivial bound of $n\Delta^2$. Notice that depending on the point set X, this improved upper bound can be by a factor of $O(n)$ smaller than $n\Delta^2$. Nevertheless, in all our upper bound results we employ the weaker bound for the purpose of readability, while all those bounds can be made more precise by applying the above-mentioned improvement.

4 Upper Bound for Points on a d-Dimensional Grid

In this section, we prove an upper bound on the number of steps of k-MeansMtd when the input points belong to the integer grid $\{1, \ldots, M\}^d$. This is the case in many practical applications where every data point has a large number of fields with each field having values in a small discrete range. For example, this includes clustering of pictures, where every pixel forms a single coordinate (or three coordinates, corresponding to the RGB values) and the value of every coordinate is restricted to be an integer in the range 0–255.

The main observation is that the centroids of any two subsets of $\{1, \ldots, M\}^d$ are either equal or are suitably far away. Since each step of k-MeansMtd moves at least one center or else stops, this observation guarantees a certain amount of improvement to the clustering cost in each step.

Lemma 4.0.8 *Let S_1 and S_2 be two nonempty subsets of $\{1, \ldots, M\}^d$ with $|S_1|+|S_2| \leq n$. Then, either $c(S_1) = c(S_2)$ or*

$$\|c(S_1) - c(S_2)\| \geq \frac{1}{n^2}.$$

Proof. If $c(S_1) \neq c(S_2)$ then they differ in at least one coordinate. Let u_1 and u_2 be the values of $c(S_1)$ and $c(S_2)$ in one such coordinate, respectively. By definition, $u_1 = s_1/|S_1|$ and $u_2 = s_2/|S_2|$ where s_1 and s_2 are integers in the range $\{1, \ldots, nM\}$. In other words $|u_1 - u_2|$ is the difference of two distinct fractions, both with denominators less than n. It follows that $|u_1 - u_2| \geq 1/n^2$ and consequently

$$\|c(S_1) - c(S_2)\| \geq |u_1 - u_2| \geq 1/n^2.$$

■

Theorem 4.0.9 *The number of steps of* k-MEANSMTD *when executed on a point set* X *taken from the grid* $\{1, \ldots, M\}^d$ *is at most* dn^5M^2.

Proof. Note, that $U = n \cdot (\sqrt{d}M)^2 = ndM^2$ is an upper bound of for the clustering cost of any k-clustering of a point set in $\{1, \ldots, M\}^d$ and that at each step at least one center moves by at least $1/n^2$. Therefore, by Lemma 3.0.5, at every step the cost function decreases by at least $1/n^4$ and the overall number of steps can be no more than $U/(1/n^4) = dn^5M^2$. ■

5 Arbitrary Point Sets and Alternative Algorithms

Unfortunately proving any meaningful bounds for the general case of k-MeansMtd, namely with points in $\Re^d$ with $d > 1$ and no further restrictions, remains elusive. However, in this section, we present two close relatives of k-MeansMtd for which we can prove polynomial bounds on the number of steps. The first algorithm differs from k-MeansMtd in that it moves a misclassified point to its correct cluster, as soon as the misclassified point is discovered (rather than first finding all misclassified points and then reassigning them to their closest centers as is the case in k-MeansMtd). The second algorithm is basically the same as k-MeansMtd with a naturally generalized notion of misclassified points. Our experimental results (Chapter 6) further support the kinship of these two algorithms with k-MeansMtd.

As was the case with our previous upper bounds, our main approach in bounding the number of steps in both these algorithms is through showing substantial improvements in the clustering cost at each step.

5.1 The SinglePnt Algorithm

We introduce an alternative to the k-Means step which we shall call a SinglePnt step.

Definition 5.1.1 In a SinglePnt *step* on a k-clustering $\mathcal{S} = (S_1, \ldots, S_k)$, a misclassified point x is chosen, such that $x \in S_i$ and $\|x - c(S_j)\| < \|x - c(S_i)\|$, for some $1 \le i \ne j \le k$, and a new clustering $\mathcal{S}' = (S'_1, \ldots, S'_k)$ is formed by removing x from S_i

and adding it to S_j. Formally, for each $1 \le l \le k$,

$$S'_l = \begin{cases} S_l & l \neq i, j, \\ S_l \setminus \{x\} & l = i, \\ S_l \cup \{x\} & l = j. \end{cases}$$

The centers are updated to the centroids of the clusters, and therefore only the centers of S_i and S_j change. Note that updating the centers takes constant time.

In a SINGLEPNT step, if the misclassified point is far away from at least one of $c(S_i)$ and $c(S_j)$, then the improvement in clustering cost made in the SINGLEPNT step cannot be too small.

Lemma 5.1.2 *Let S and T be two point sets of sizes n and m, respectively, and let $s = c(S)$ and $t = c(T)$. Suppose that x is a point in T with distances d_S and d_T from s and t, respectively, and such that $d_S < d_T$. Let $S' = S \cup \{x\}$ and $T' = T \setminus \{x\}$ and let $s' = c(S')$ and $t' = c(T')$.*

$$\psi(S, s) + \psi(T, t) - \psi(S', s') - \psi(T', t') \ge \frac{(d_S + d_T)^2}{2(n+m)}.$$

Proof. Indeed,

$$c(S') = \frac{n}{n+1} c(S) + \frac{1}{n+1} x.$$

Thus

$$\begin{aligned} \left\| s - s' \right\| &= \left\| c(S) - c(S') \right\| \\ &= \left\| \frac{1}{n+1} c(S) - \frac{1}{n+1} x \right\| \\ &= \frac{1}{n+1} \left\| c(S) - x \right\| \\ &= \frac{d_S}{n+1}. \end{aligned}$$

Similarly,

$$\left\|t - t'\right\| = \frac{d_T}{m-1}.$$

Thus using Lemma 3.0.5 we get

$$\psi(S', s) - \psi(S', s') = (n+1)\left(\frac{d_S}{n+1}\right)^2 = \frac{d_S^2}{n+1},$$

and similarly

$$\psi(T', t) - \psi(T', t') = \frac{d_T^2}{(m-1)}.$$

As such, since $d_S < d_T$, we have that

$$\psi(S, s) + \psi(T, t) \geq \psi(S', s) + \psi(T', t),$$

and

$$\begin{aligned}
&\psi(S, s) + \psi(T, t) - \psi(S', s') - \psi(T', t') \\
&\qquad \geq \psi(S', s) + \psi(T', t) - \psi(S', s') - \psi(T', t') \\
&\qquad \geq \frac{d_S^2}{n+1} + \frac{d_T^2}{m-1} \\
&\qquad \geq \frac{d_S^2}{n+m} + \frac{d_T^2}{n+m} \\
&\qquad = \frac{d_S^2 + d_T^2}{n+m} \\
&\qquad \geq \frac{(d_S + d_T)^2}{2(n+m)}.
\end{aligned}$$

■

Our modified version of k-MEANSMTD, to which we shall refer as "SINGLEPNT", replaces k-MEANS steps with SINGLEPNT steps. Starting from an arbitrary clustering of the input point set, SINGLEPNT repeatedly performs SINGLEPNT steps until no misclassified points remain. Notice that unlike the k-MEANS step, the SINGLEPNT step does not maintain the property that the clustering achieved at the end of the step is imposed by some Voronoi diagram. However, when the algorithm stops no misclassified points are left, and this property must hold since otherwise further steps would be possible.

Theorem 5.1.3 *On any input $X \subset \Re^d$,* SinglePnt *makes at most $O(kn^2\Delta^2)$ steps before termination.*

Proof. Once again, we assume that no two points in X are less than unit distance apart. Call a SinglePnt step *weak*, if the misclassified point it considers is at distance less than $1/8$ from both *involved centers*, i.e., its current center and the center closest to it. We call a SinglePnt step *strong* if it is not weak. Lemma 5.1.2 shows that in a strong SinglePnt step the clustering cost improves by at least $1/(128n)$. In the sequel we shall show that the algorithm cannot take more than k *consecutive* weak steps, and thus at least one out of every $k+1$ consecutive steps must be strong and thus result an improvement of $1/(128n)$ to the clustering cost; hence the upper bound of $O(kn^2\Delta^2)$.

For a fixed point in time, let $c_1, \ldots, c_k$ denote the current centers, and let $S_1, \ldots, S_k$ denote the corresponding clusters; namely, S_i is the set of points served by c_i, for $i = 1, \ldots, k$. Consider the balls $B_1, \ldots, B_k$ of radius $1/8$ centered at $c_1, \ldots, c_k$, respectively. Observe that since every pair of points in X are at distance at least 1 from each other, each ball B_i can contain at most one point of X. Moreover, the intersection of any subset of the balls $B_1, \ldots, B_k$ can contain at most one point of X. For a point $x \in X$, let $\mathcal{B}(x)$ denote the set of balls among $B_1, \ldots, B_k$ that contain the point x. We refer to $\mathcal{B}(x)$ as the *batch* of x.

By the above observation, the balls (and the corresponding centers) are classified according to the point of X they contain (if they contain such a point at all). Let $\mathcal{B}_X$ be the set of batches of balls that are induced by X and contain more than one ball. Formally,

$$\mathcal{B}_X = \{\mathcal{B}(x) : x \in X, |\mathcal{B}(x)| > 1\}.$$

The set of balls $\bigcup \mathcal{B}_X$ is the set of *active* balls.

A misclassified point x can participate in a weak SinglePnt step only if it belongs to more than one ball; i.e., when $|\mathcal{B}(x)| > 1$. Observe that, if we perform a weak step, and one of the centers move such that the corresponding ball B_i no longer contains any point of X in its interior, then for B_i to contain a point again, the algorithm must perform a strong step. To see this, observe that (weakly) losing a point x may cause a center move a distance of at most $1/8$. Therefore, once a center c_i loses a point x, and thus moves away from x, it does not move far enough for the ball B_i to contain a different point of X.

Hence, in every weak iteration a point x changes the cluster it belongs to in $\mathcal{B}(x)$. This might result in a shrinking of the active set of balls. On the other hand, while only weak SINGLEPNT steps are being taken, any cluster S_j can change only by winning or losing the point x_i that stabs the corresponding ball B_j. It follows that once a set S_j loses the point x, then it can never get it back since that would correspond to an increase in the clustering cost. Therefore the total number of possible consecutive weak SINGLEPNT steps is bounded by

$$\sum_{x\in X, |\mathcal{B}(x)|>1} |\mathcal{B}(x)| \leq k.$$

■

5.2 The LAZY-k-MEANS Algorithm

Our second variant to k-MEANSMTD, which we name "LAZY-k-MEANS", results from a natural generalization misclassified points (Definition 1.3.1). Intuitively, the difference between the LAZY-k-MEANS and k-MEANSMTD is that LAZY-k-MEANS at each step only reassigns those misclassified points to their closest centers that are *substantially* misclassified, namely the points that benefit from reclassification by at least a constant factor.

Definition 5.2.1 Given a clustering $\mathcal{S} = (S_1, \ldots, S_k)$ of a point set X, if for a point

$x \in S_i$ there exists a $j \neq i$, such that

$$\|x - c(S_i)\| > (1+\varepsilon)\,\|x - c(S_j)\|\,,$$

then x is said to be $(1+\varepsilon)$-*misclassified* for center pair $(c(S_i), c(S_j))$. The centers $c(S_i)$ and $c(S_j)$ are referred to as *switch centers* for x. We also say that $c(S_i)$ is the *losing center* and $c(S_j)$ is the *winning center* for x.

Thus LAZY-k-MEANS with parameter ε starts with an arbitrary k-clustering. In each step, it (i) reassigns every $(1+\varepsilon)$-misclassified point to its closest center and (ii) moves every center to the centroid of its new cluster. Indeed, k-MEANSMTD is simply LAZY-k-MEANS with parameter $\varepsilon = 0$. Naturally, the algorithm stops when no $(1+\varepsilon)$-misclassified points are left.

In the sequel we bound the maximum number of steps taken by LAZY-k-MEANS. We shall use the following fact from elementary Euclidean geometry.

Fact 5.2.2 *Given two points c and c' with $\|c - c'\| = \ell$, the locus of the points x with $\|x - c'\| > (1+\varepsilon)\,\|x - c\|$ is an open ball of radius $R = \ell(1+\varepsilon)/(\varepsilon(2+\varepsilon))$ called the ε-Apollonius ball for c with respect to c'. This ball is centered on the line containing the segment cc' at distance $R + \ell\varepsilon/(2(2+\varepsilon))$ from the bisector of cc', and on the same side of the bisector as c (see Figure 5.1).*

Lemma 5.2.3 *For any three points x, c, and c' in $\Re^d$ with $\|x - c\| \leq \|x - c'\|$ we have*

$$\left\|x - c'\right\|^2 - \|x - c\|^2 = 2h\left\|c - c'\right\|,$$

where h is the distance from x to the bisector of c and c'.

Proof. Let y be the intersection point of the segment cc' with the $(d-1)$-dimensional hyperplane parallel to the bisector of c and c' and containing x. By Pythagorean equality we have $\|x - c\|^2 = \|x - y\|^2 + \|y - c\|^2$ and $\|x - c'\|^2 = \|x - y\|^2 + \|y - c'\|^2$.

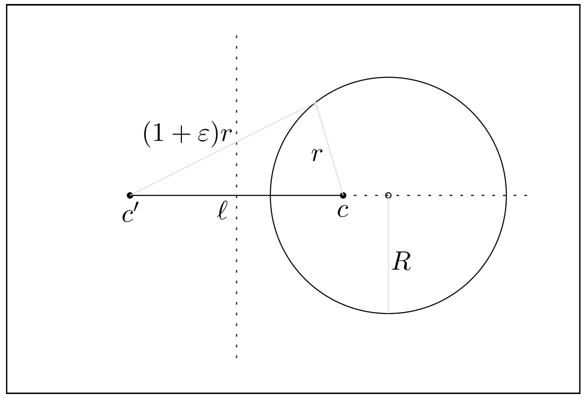

Figure 5.1: The ε-Apollonius ball for c with respect to c'.

Subtracting the first equality from the second, we obtain

$$\begin{aligned} \left\|x - c'\right\|^2 - \|x - c\|^2 &= \left\|y - c'\right\|^2 - \|y - c\|^2 \\ &= (\|y - c\| + \left\|y - c'\right\|)(\|y - c\| - \left\|y - c'\right\|) \\ &= 2h \left\|c - c'\right\|, \end{aligned}$$

since $\|y - c\| - \|y - c'\| = 2h$. ∎

Theorem 5.2.4 *The number of steps of the* Lazy-k-Means *algorithm with parameter* ε *is* $O(n\Delta^2\varepsilon^{-3})$.

Proof. We will show that every two consecutive steps of Lazy-k-Means with parameter ε make an improvement of at least

$$\lambda^* = \frac{\varepsilon^3(2+\varepsilon)}{256(1+\varepsilon)^2} \geq \frac{\varepsilon^3}{512} = \Omega(\varepsilon^3).$$

Let $\ell_0 = \varepsilon(2+\varepsilon)/(16(1+\varepsilon))$. Notice that $\ell_0 < 1/8$ for $0 < \varepsilon \leq 1$. We call a misclassified point x *strongly* misclassified, if its switch centers c and c' are at distance at most ℓ_0 from each other, and *weakly* misclassified otherwise.

If at the beginning of a LAZY-k-MEANS step there exists a strongly misclassified point x for a center pair (c, c'), then since every point in the ε-Apollonius ball for c' with respect to c is at distance at least $\ell_0\varepsilon/(2(2+\varepsilon))$ from the bisector of cc', by Lemma 5.2.3 the reclassification improvement in clustering cost resulting from assigning x to c' is

$$\|x-c\|^2 - \|x-c'\|^2 = \frac{\ell_0^2\varepsilon}{2+\varepsilon} \geq \frac{\varepsilon^3(2+\varepsilon)}{256(1+\varepsilon)^2} = \lambda^*.$$

Thus we assume that all misclassified points are weakly misclassified. Let x be one such point for center pair (c, c'). By our assumption $\|c - c'\| < \ell_0$. Observe that in such a case, the radius of the ε-Apollonius ball for c' with respect to c is $\ell(1+\varepsilon)/(\varepsilon(2+\varepsilon)) < 1/16$. In particular, since there exists a ball of radius $1/16$ containing both x and c', the ball of radius $1/8$ centered at c', which we denote by $B(c', 1/8)$, includes x. Also since $\|c - c'\| < 1/8$ as verified above, we get $c \in B(c', 1/8)$ as well. In other words, both switch centers c and c' are at distance less than $1/4$ from x. Now, since every pair of points in X are at distance 1 or more, any center can be a switch center for at most one weakly misclassified point. This in particular implies that in the considered LAZY-k-MEANS step, no cluster is modified by more than a single point.

When the misclassified points are assigned to their closest centers, the centers that do not lose or win any points stay at their previous locations. A center c' that wins a point x moves closer to x since x is the only point it wins while losing no other points. Similarly, a center c that loses a point x moves away from x since x is the only point it loses without winning any other points. A losing center c moves away from its lost point x by a distance of at most $\|c - x\| < 1/4$ since its previous number of served points was at least 2 (otherwise, we would have $c = x$ and thus x could not be misclassified). Therefore, when c moves to the centroid of its cluster (now missing x), $\|x - c\| < 1/2$ and consequently $\|c - y\| > 1/2$ for any $x \neq y \in X$. As a result, c can not be a switch center for any weakly misclassified point in the subsequent LAZY-k-MEANS step.

On the other hand, the winning center c' to whose cluster x is added, moves closer to x and since no center other than c and c' in $B(x, 1/4)$ moves (as there is no point other than x they can win or lose), x will not be misclassified in the next Lazy-k-Means step.

It follows from the above discussion that the next Lazy-k-Means step cannot have any weakly misclassified points and thus either the algorithm stops or some strongly misclassified point will exist, resulting an improvement of at least λ^*. Thus the total number of steps taken by Lazy-k-Means with parameter ε is at most $2n\Delta^2/\lambda^* = O(n\Delta^2\varepsilon^{-3})$. ■

6 Experimental Results

We introduced both SinglePnt and Lazy-k-Means alternatives to k-MeansMtd as similar, equally easy to implement algorithms that are simpler to analyze than k-MeansMtd itself. However, as mentioned in the introduction, k-MeansMtd is mainly of interest only in practice because of its ease of implementation and its relatively fast termination (small number of steps). It thus raises the question of how our alternative algorithms perform in practice in comparison to k-MeansMtd.

We performed a series of experiments analogous to those done in [KMN$^+$02], as described below, to compare the number of rounds, number of reclassified points, and quality of final clustering produced by these two alternative algorithms with those of k-MeansMtd. We use the same inputs used by Kanungo *et al.* for our experiments. See [KMN$^+$02] for detailed description of those inputs. We have tried to implement each of the algorithms in the simplest possible way and avoided using any advanced point location or nearest neighbor search structure or algorithm. Due to the great similarity between the three algorithms considered here, it is expected that any technique used for improving the performance of any of these algorithms, to be suitable for improving the other two variants in a somewhat similar way.

k-MeansMtd and Lazy-k-Means iterate over points and assign each point to the closest center. While doing this the new set of centers are calculated and existence of a $(1 + \varepsilon)$-misclassified point is checked. SinglePnt examines the points one by one, moving to the first point when reaching the end of the list, checking if they are misclassified or not. When a misclassified point is discovered it is assigned to its closest center and the location of the two switching centers is updated. The algorithm stops when it cannot find a misclassified point for n consecutive steps[1].

Our experimental results are summarized in Table 6.1 and Table 6.2. In conformance to [KMN+02] the costs referred to in these tables is the total final clustering cost, divided by the number of points. In that sense we report the "average" cost per point. Table 6.1 is produced by running, only once, each of the four algorithms with the same set of randomly chosen center for each combination of point set and number of centers considered. By studying several such tables it seems that the total number of reclassified points and the quality of clustering found by SinglePnt tends to be very close to those of k-MeansMtd. Notice that in Table 6.1, the number of steps of SinglePnt are left blank as they are equal to the number of reclassified points and cannot be compared with the number of steps of k-MeansMtd or Lazy-k-Means.

Table 6.2 summarizes the results of running 100 tests similar to the one reported in Table 6.1 each with different initial set of centers picked randomly from the bounding box of the given point set. The best, worst, and average final clustering costs are reported in each case.

We have not discussed the running times as we made no effort in optimizing our implementations. It is however interesting that both of the two alternative algorithms tend to be faster than k-MeansMtd's in a typical implementation such as ours. SinglePnt seems to be typically more than 20% faster than Lloyd. In particular, we emphasize, that our simple implementation is considerably slower than the implementation of Kanungo *et al.* [KMN+02] that uses data structure similar to kd-tree to speed up the computation of the Voronoi partitions. We believe that we would get similar performance gains by using their data structure.

Below are the the descriptions of the test point sets used in these experiments [KMN+02].

ClusGauss: The data consists of 10,000 points in $\Re^3$, generated from a distribution

[1]The input point sets used in these experiments together with the source-code of our implementation is available on the web at `http://www.uiuc.edu/~sariel/papers/03/lloyd_kmeans`.

Data Set	k	Method	Steps	Reclassified	Final Cost
ClusGauss $n = 10,000$ $d = 3$	25	k-MeansMtd	24	4748	0.081615
		SinglePnt	-	4232	0.081622
		Lazy-k-Means, $\epsilon = 0.05$	17	2377	0.082702
		Lazy-k-Means, $\epsilon = 0.20$	18	1554	0.089905
	50	k-MeansMtd	20	4672	0.031969
		SinglePnt	-	4391	0.031728
		Lazy-k-Means, $\epsilon = 0.05$	16	2244	0.032164
		Lazy-k-Means, $\epsilon = 0.20$	22	1974	0.034661
	100	k-MeansMtd	22	5377	0.009639
		SinglePnt	-	4958	0.009706
		Lazy-k-Means, $\epsilon = 0.05$	15	2512	0.010925
		Lazy-k-Means, $\epsilon = 0.20$	19	1748	0.013092
MultiClus $n = 10,000$ $d = 3$	50	k-MeansMtd	21	2544	0.033870
		SinglePnt	-	2419	0.033941
		Lazy-k-Means, $\epsilon = 0.05$	16	1121	0.034622
		Lazy-k-Means, $\epsilon = 0.20$	25	722	0.038042
	100	k-MeansMtd	18	1744	0.009248
		SinglePnt	-	1732	0.008854
		Lazy-k-Means, $\epsilon = 0.05$	11	740	0.009902
		Lazy-k-Means, $\epsilon = 0.20$	15	584	0.010811
	500	k-MeansMtd	12	1768	0.002495
		SinglePnt	-	1694	0.002522
		Lazy-k-Means, $\epsilon = 0.05$	9	528	0.002757
		Lazy-k-Means, $\epsilon = 0.20$	11	444	0.002994
Lena22 $n = 65,536$ $d = 4$	8	k-MeansMtd	36	62130	335.408625
		SinglePnt	-	57357	335.440866
		Lazy-k-Means, $\epsilon = 0.05$	27	50298	338.594668
		Lazy-k-Means, $\epsilon = 0.20$	21	44040	355.715258
	64	k-MeansMtd	211	111844	94.098422
		SinglePnt	-	81505	94.390640
		Lazy-k-Means, $\epsilon = 0.05$	88	55541	97.608823
		Lazy-k-Means, $\epsilon = 0.20$	24	30201	120.274428
	256	k-MeansMtd	167	111110	48.788216
		SinglePnt	-	101522	48.307815
		Lazy-k-Means, $\epsilon = 0.05$	92	57575	51.954810
		Lazy-k-Means, $\epsilon = 0.20$	79	32348	61.331614
Lena44 $n = 16,384$ $d = 16$	8	k-MeansMtd	63	18211	2700.589245
		SinglePnt	-	16467	2700.587691
		Lazy-k-Means, $\epsilon = 0.05$	20	9715	2889.747540
		Lazy-k-Means, $\epsilon = 0.20$	27	9201	3008.783333
	64	k-MeansMtd	61	21292	1525.846646
		SinglePnt	-	16422	1615.667299
		Lazy-k-Means, $\epsilon = 0.05$	45	13092	1555.520952
		Lazy-k-Means, $\epsilon = 0.20$	16	7527	1907.962692
	256	k-MeansMtd	43	21394	1132.746162
		SinglePnt	-	28049	1122.407317
		Lazy-k-Means, $\epsilon = 0.05$	28	12405	1156.884049
		Lazy-k-Means, $\epsilon = 0.20$	27	7993	1320.303278
Kiss $n = 10,000$ $d = 3$	8	k-MeansMtd	18	5982	687.362264
		SinglePnt	-	7026	687.293930
		Lazy-k-Means, $\epsilon = 0.05$	18	3277	690.342895
		Lazy-k-Means, $\epsilon = 0.20$	23	2712	720.891998
	64	k-MeansMtd	202	29288	202.044849
		SinglePnt	-	35228	185.519927
		Lazy-k-Means, $\epsilon = 0.05$	92	12471	221.936175
		Lazy-k-Means, $\epsilon = 0.20$	44	6080	263.497185
	256	k-MeansMtd	144	17896	105.438490
		SinglePnt	-	16992	106.112133
		Lazy-k-Means, $\epsilon = 0.05$	61	7498	120.317362
		Lazy-k-Means, $\epsilon = 0.20$	27	3479	150.156231

Table 6.1: A comparison of all algorithms in a typical run.
Number of steps, number of reclassified points, and final average clustering cost in a typical execution of each of the four algorithms on data sets mentioned in [KMN+02].

Data Set	k	Method	Minimum Cost	Maximum Cost	Average Cost
ClusGauss $n = 10,000$ $d = 3$	25	k-MeansMtd	0.068462	0.087951	0.07501276
		SinglePnt	0.067450	0.083194	0.07486010
		Lazy-k-Means, $\varepsilon = 0.20$	0.074667	0.100035	0.08510598
		Lazy-k-Means, $\varepsilon = 0.05$	0.070011	0.092658	0.07803375
	50	k-MeansMtd	0.028841	0.040087	0.03335312
		SinglePnt	0.028376	0.040623	0.03308624
		Lazy-k-Means, $\varepsilon = 0.20$	0.031175	0.046528	0.03719264
		Lazy-k-Means, $\varepsilon = 0.05$	0.029626	0.040811	0.03384180
	100	k-MeansMtd	0.011425	0.016722	0.01401549
		SinglePnt	0.010106	0.017986	0.01365492
		Lazy-k-Means, $\varepsilon = 0.20$	0.011928	0.022015	0.01565268
		Lazy-k-Means, $\varepsilon = 0.05$	0.011730	0.020600	0.01442575
MultiClus $n = 10,000$ $d = 3$	50	k-MeansMtd	0.027563	0.034995	0.03051698
		SinglePnt	0.027412	0.034167	0.03083110
		Lazy-k-Means, $\varepsilon = 0.20$	0.029507	0.055160	0.03620397
		Lazy-k-Means, $\varepsilon = 0.05$	0.028457	0.046314	0.03260643
	100	k-MeansMtd	0.002477	0.004324	0.00308144
		SinglePnt	0.002390	0.004179	0.00303798
		Lazy-k-Means, $\varepsilon = 0.20$	0.002758	0.005175	0.00356282
		Lazy-k-Means, $\varepsilon = 0.05$	0.002331	0.004789	0.00322593
	500	k-MeansMtd	0.002142	0.002731	0.00240768
		SinglePnt	0.002136	0.002805	0.00244548
		Lazy-k-Means, $\varepsilon = 0.20$	0.002539	0.003567	0.00292354
		Lazy-k-Means, $\varepsilon = 0.05$	0.002206	0.002890	0.00254321
Lena22 $n = 65,536$ $d = 4$	8	k-MeansMtd	263.644420	348.604787	299.78905632
		SinglePnt	263.659829	348.527023	307.12394164
		Lazy-k-Means, $\varepsilon = 0.20$	278.337133	414.679356	345.07986265
		Lazy-k-Means, $\varepsilon = 0.05$	271.041374	409.802396	322.99259307
	64	k-MeansMtd	82.074376	102.327255	88.53558757
		SinglePnt	82.190945	104.574941	89.24323986
		Lazy-k-Means, $\varepsilon = 0.20$	100.601485	147.170657	111.93562151
		Lazy-k-Means, $\varepsilon = 0.05$	82.798308	106.231864	94.20319250
	256	k-MeansMtd	44.637740	51.482531	47.66542537
		SinglePnt	44.699224	51.685618	47.81799127
		Lazy-k-Means, $\varepsilon = 0.20$	56.906620	71.491475	62.00216985
		Lazy-k-Means, $\varepsilon = 0.05$	47.178425	54.946136	50.82872342
Lena44 $n = 16,384$ $d = 16$	8	k-MeansMtd	2699.721266	3617.282065	2903.30164756
		SinglePnt	2699.663310	3216.854024	2894.42713876
		Lazy-k-Means, $\varepsilon = 0.20$	2834.438965	4452.875383	3293.73084140
		Lazy-k-Means, $\varepsilon = 0.05$	2725.907276	3649.518829	2977.33094524
	64	k-MeansMtd	1305.357406	1694.965827	1503.17431782
		SinglePnt	1345.821487	1811.663769	1515.08195678
		Lazy-k-Means, $\varepsilon = 0.20$	1564.252624	2385.794013	1785.93841955
		Lazy-k-Means, $\varepsilon = 0.05$	1410.883673	1793.704755	1565.18092988
	256	k-MeansMtd	1044.017122	1311.942456	1151.64441691
		SinglePnt	1055.788028	1308.459754	1168.30843808
		Lazy-k-Means, $\varepsilon = 0.20$	1262.487865	1653.820840	1400.49905496
		Lazy-k-Means, $\varepsilon = 0.05$	1094.884884	1385.345314	1219.27000492
Kiss $n = 10,000$ $d = 3$	8	k-MeansMtd	687.278119	714.789442	700.352315760
		SinglePnt	687.279479	714.731416	697.292832560
		Lazy-k-Means, $\varepsilon = 0.20$	727.017538	947.779405	802.256735040
		Lazy-k-Means, $\varepsilon = 0.05$	689.779010	861.853344	719.140385820
	64	k-MeansMtd	158.607749	208.946701	178.21703676
		SinglePnt	151.642447	203.102940	177.17793706
		Lazy-k-Means, $\varepsilon = 0.20$	222.646398	324.435479	259.62118455
		Lazy-k-Means, $\varepsilon = 0.05$	170.571861	248.648363	208.64482062
	256	k-MeansMtd	96.272602	115.294309	105.30212380
		SinglePnt	97.141907	125.009357	107.08187899
		Lazy-k-Means, $\varepsilon = 0.20$	124.378185	158.922757	140.72908431
		Lazy-k-Means, $\varepsilon = 0.05$	103.672482	129.685819	116.73971102

Table 6.2: Summary of 100 tests.

Minimum, maximum, and average clustering cost on 100 executions of each of the algorithms on each of the data sets with initial centers picked randomly.

consisting of 100 clusters of almost equal size, with centers uniformly distributed in $[-1, 1]^3$. The points in each cluster are drawn from a multivariate Gaussian distribution centered at the cluster center, where each coordinate has a standard deviation of 0.05.

MultiClus: The data consists of 10,000 points in $\Re^3$ which were generated from a distribution consisting of a number of multivariate Gaussian clusters of various sizes and standard deviations. Again cluster centers where sampled uniformly from a the cube $[-1, 1]^3$. The cluster sizes are powers of 2. The probability of generating a cluster of size 2^i is $1/2^i$, and hence there are many small clusters. The standard deviation of a cluster of size m is $0.05/\sqrt{m}$, and hence each cluster is expected to have roughly equal distortion of 0.025.

Lena22 and Lena44: These were taken from an application in image compression through vector quantization. The data were generated by partitioning a 512×512 gray-scale Lena image into $65,536$ 2×2 tiles. Each tile is treated as a point in a 4-dimensional space. Lena44 was generated using 4×4 tiles, thus generating 16,384 points in 16-dimensional space.

Kiss: This is from a color quantization application. 10,000 RGB pixel values were sampled at random from a color image of a painting "The Kiss" by Gustav Klimt. This resulted in 10,000 points in 3-space.

7 Conclusions

We presented several results on the number of iterations performed by the k-MeansMtd clustering algorithm. To our knowledge, our results are the first to provide combinatorial bounds on the performance of k-MeansMtd. We consider this thesis to be a first step in understanding the Lloyd's method. It is our belief that both our lower and upper bounds are loose, and one might need to use other techniques to improve them. In particular, we mention some open problems:

1. There is still a large gap between our lower and upper bounds. In particular, a super-linear lower bound would be interesting even in high-dimensional space.
2. Our current upper bounds include the spread as a parameter. It would be interesting to prove (or disprove) that this is indeed necessary.
3. We have introduced alternative, easy to analyze algorithms, that are comparable to k-MeansMtd both in their description and their behavior in practice. It would be interesting to show provable connections between these algorithms and compare the bounds on the number of steps they require to terminate.

Recently, independently of our results, Sanjoy Dasgupta [Das03] announced results which are similar to a *subset* of our results. In particular, he mentions the one-dimensional lower bound, and a better upper bound for $k < 5$ but only in one dimension. This work of Sanjoy Dasgupta and Howard Karloff seems to be using similar arguments to ours (personal communication) although to our knowledge it has not been written or published yet.

References

[Das03] Sanjoy Dasgupta. Open problems: How fast is k-means. In *COLT*, volume 2777 of *Lecture Notes in Computer Science*. Springer, 2003.

[DFG99] Qiang Du, Vance Faber, and Max Gunzburger. Centroidal voronoi tessellations: Applications and algorithms. *SIAM Rev.*, 41(4):637–676, 1999. CL 008.

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley Interscience, New York, 2001.

[dKKR03] W. Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In *Proceedings of the thirty-fifth ACM symposium on Theory of computing*, pages 50–58. ACM Press, 2003. CL 003.

[ES03] Michelle Effros and Leonard J. Schulman. Rapid clustering with a deterministic data net. manuscript, 2003.

[HPM] Sariel Har-Peled and Soham Mazumdar. Coresets for k-means and k-median clustering and their applications. to appear in SODA 2004.

[IKI94] Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering: (extended abstract). In *Proceedings of the tenth annual symposium on Computational geometry*, pages 332–339. ACM Press, 1994.

[KMN+02] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for k-means clustering. In *Proceedings of the eighteenth annual symposium on Computational geometry*, pages 10–18. ACM Press, 2002. CL 002.

[Mat00] Jiri Matoušek. On approximate geometric k-clustering. *Discrete and Computational Geometry*, 24(1), 2000.