Linear Time Algorithms for Clustering Problems in any dimensions

Amit Kumar[1], Yogish Sabharwal[2], and Sandeep Sen[3]

[1] Dept of Comp Sc & Engg, Indian Institute of Technology, New Delhi-110016, India
amitk@cse.iitd.ernet.in
[2] IBM India Research Lab, Block-I, IIT Delhi, Hauz Khas, New Delhi-110016, India
ysabharwal@in.ibm.com
[3] Dept of Comp Sc & Engg, Indian Institute of Technology, Kharagpur, India
ssen@cse.iitkgp.ernet.in

Abstract. We generalize the k-means algorithm presented by the authors [14] and show that the resulting algorithm can solve a larger class of clustering problems that satisfy certain properties (*existence of a random sampling procedure* and *tightness*). We prove these properties for the k-median and the discrete k-means clustering problems, resulting in $O(2^{(k/\varepsilon)^{O(1)}}dn)$ time $(1+\varepsilon)$-approximation algorithms for these problems. These are the first algorithms for these problems linear in the size of the input (nd for n points in d dimensions), independent of dimensions in the exponent, assuming k and ε to be fixed. A key ingredient of the k-median result is a $(1+\varepsilon)$-approximation algorithm for the 1-median problem which has running time $O(2^{(1/\varepsilon)^{O(1)}}d)$. The previous best known algorithm for this problem had linear running time.

1 Introduction

The problem of clustering a group of data items into similar groups is one of the most widely studied problems in computer science. Clustering has applications in a variety of areas, for example, data mining, information retrieval, image processing, and web search ([5, 7, 9]). Given the wide range of applications, many different definitions of clustering exist in the literature ([8, 4]). Most of these definitions begin by defining a notion of distance (similarity) between two data items and then try to form clusters so that data items with small distance between them get clustered together.

Often, clustering problems arise in a geometric setting, i.e., the data items are points in a high dimensional Euclidean space. In such settings, it is natural to define the distance between two points as the Euclidean distance between them. Two of the most popular definitions of clustering are the *k-means clustering problem* and the *k-median clustering problem.* Given a set of points P, the k-means clustering problems seeks to find a set K of k centers, such that $\sum_{p\in P} d(p,K)^2$ is minimized, whereas the k-median clustering problems seeks to find a set K of k centers, such that $\sum_{p\in P} d(p,K)$ is minimized. Note that the points in K can be arbitrary points in the Euclidean space. Here $d(p,K)$ refers

to the distance between p and the closest center in K. We can think of this as each point in P gets assigned to the closest center. The points that get assigned to the same center form a cluster. These problems are NP-hard for even $k = 2$ (when dimension is not fixed). Interestingly, the center in the optimal solution to the 1-mean problem is the same as the center of mass of the points. Howvever, in the case of the 1-median problem, also known as the Fermat-Weber problem, no such closed form is known. We show that despite the lack of such a closed form, we can obtain an approximation to the optimal 1-median in $O(1)$ time (independent of the number of points). There exist variations to these clustering problems, for example, the discrete versions of these problems, where the centers that we seek are constrained to lie on the input set of points.

1.1 Related work

A lot of research has been devoted to solving these problems exactly (see [11] and the references therein). Even the best known algorithms for the k-median and the k-means problem take at least $\Omega(n^d)$ time. Recently, some work has been devoted to finding $(1+\varepsilon)$-approximation algorithm for these problems, where ε can be an arbitrarily small constant. This has led to algorithms with much improved running times. Further, if we look at the applications of these problems, they often involve mapping subjective features to points in the Euclidean space. Since there is an error inherent in this mapping, finding a $(1+\varepsilon)$-approximate solution does not lead to a deterioration in the solution for the actual application.

The following table summarizes the recent results for the problems, in the context of $(1+\varepsilon)$-approximation algorithms. Some of these algorithms are randomized with the expected runing time holding good for any input.

Problem	Result	Reference
1-median	$O(n/\varepsilon^2)$	Indyk [12]
k-median	$O(n^{O(1/\varepsilon)+1})$ for $d = 2$	Arora [1]
	$O(n + \varrho k^{O(1)}\log^{O(1)} n)$ (discrete also)	Har-Peled et al. [10]
	where $\varrho = exp[O((1+\log 1/\varepsilon)/\varepsilon)^{d-1}]$	
	$O(2^{(k/\varepsilon)^{O(1)}} d^{O(1)} n \log^{O(k)} n)$	Badoiu et al. [3]
discrete k-median	$O(\varrho n \log n \log k)$	Kolliopoulos et al. [13]
k-means	$O(n/\varepsilon^d)$ for $k = 2$	Inaba et al. [11]
	$O(n\varepsilon^{-2k^2 d}\log^k n)$	Matousek [15]
	$O(g(k,\varepsilon) n \log^k n)$	de la Vega et al. [6]
	$g(k,\varepsilon) = exp[(k^3/\varepsilon^8)(\ln(k/\varepsilon)\ln k]$	
	$O(n + k^{k+2}\varepsilon^{-(2d+1)k}\log^{k+1} n \log^k \frac{1}{\varepsilon})$	Har-Peled et al. [10]
	$O(2^{(k/\varepsilon)^{O(1)}} dn)$	Kumar et al. [14]

1.2 Our contributions

In this paper, we generalize the algorithm of authors [14] to a wide range of clustering problems. We define a general class of clustering problems and show that if certain conditions are satsified, we can get linear time $(1+\varepsilon)$-approximation algorithms for these problems. We then use our general framework to get the following results. Given a set of n points P in $\Re^d$, we present

1. a randomized algorithm that generates a candidate center set of size $O(2^{1/\varepsilon^{O(1)}})$, such that at least one of the points in this set is a $(1+\varepsilon)$-approximate 1-median of P with constant probability. The running time of the algorithm is $O(2^{1/\varepsilon^{O(1)}}d)$, assuming that the points are stored in a suitable data structure such that a point can be randomly sampled in constant time. This improves on the algorithm of Badoiu et al. [3] which generates a candidate center set of size $O(2^{1/\varepsilon^4}\log n)$ in time $O(d2^{1/\varepsilon^4}\log n)$.
2. a randomized $(1+\varepsilon)$-approximation algorithm for the 1-median problem which runs in time $O(2^{1/\varepsilon^{O(1)}}d)$, assuming that the points are stored in a suitable data structure such that a point can be randomly sampled in constant time.
3. a randomized $(1+\varepsilon)$-approximation algorithm for the k-median problem which runs in $O(2^{(k/\varepsilon)^{O(1)}}nd)$ time.
4. a randomized $(1+\varepsilon)$-approximation algorithm for the discrete k-means clustering which runs in $O(2^{(k/\varepsilon)^{O(1)}}nd)$ time.

All our algorithms yield the desired result with constant probability (which can be made as close to 1 as we wish by a constant number of repetitions). As mentioned earlier, we generalize the result of the authors in [14] to solve a larger class of clustering problems satisfying a set of conditions (c.f. section 2). We then show that the k-median problem and the discrete k-means problem fall in this class of clustering problems. One important condition that the clustering problems must satisfy is that there should be an algorithm to generate a candidate set of points of size independent of n, such that at least one of these points is a close approximation to the optimal center when we desire only one cluster. Armed with such a subroutine, we show how to approximate all the centers in the optimal solution in an iterative manner.

It is easy to see that our algorithms for the k-median and the discrete k-means problems have better running time than the previously known algorithms for these problems, specially when d is very large. In fact, these are the first algorithms for the k-median and the discrete k-means clustering that have running time linear in the size of the input for fixed k and ε.

For the 1-median problem, the candidate center set generation and the actual approximation algorithm have better running time than all previously known algorithms. The algorithms in this paper have the additional advantage of simplicity inherited from generalizing the approach of Kumar et al. [14].

The remaining paper is organized as follows. In Section 2, we describe a general approach for solving clustering problems efficiently. In the subsequent

sections we give applications of the general method by showing that this class of problems includes the k-median, the k-means and the discrete k-means problems. In section 4.3, we also describe an efficient approximation algorithm for the 1-median problem.

2 Clustering Problems

In this section, we give a general definition of clustering problems. Our algorithms will work on any of these problems provided certain conditions are satisfied. We will state these conditions later in the section.

We shall define a clustering problem by two parameters – an integer k and a real-valued cost function $f(Q, x)$, where Q is a set of points, and x is a point in an Euclidean space. We shall denote this clustering problem as $\mathcal{C}(f, k)$. The input to $\mathcal{C}(f, k)$ is a set of points in a Euclidean space.

Given an instance P of n points, $\mathcal{C}(f, k)$ seeks to partition them into k sets, which we shall denote as *clusters*. Let these clusters be $C_1, \ldots, C_k$. A solution also finds k points, which we call *centers*, $c_1, \ldots, c_k$. We shall say that c_i is the center of cluster C_i (or the points in C_i are assigned to c_i). The objective of the problem is to minimize the quantity $\sum_{i=1}^{k} f(C_i, c_i)$.

This is a fairly general definition. Let us see some important special cases.

- k-median : $f(Q, x) = \sum_{q \in Q} d(q, x)$.
- k-means : $f(Q, x) = \sum_{q \in Q} d(q, x)^2$.

We can also encompass the discrete versions of these problems, i.e., cases where the centers have to be one of the points in P. In such problems, we can make $f(Q, x)$ unbounded if $x \notin Q$.

As stated earlier, we shall assume that we are given a constant $\varepsilon > 0$, and we are interested in finding $(1 + \varepsilon)$-approximation algorithms for these clustering problems.

We now state the conditions the clustering problems should satisfy. We begin with some definitions first. Let us fix a clustering problem $\mathcal{C}(f, k)$. Although we should parameterize all our definitions by f, we avoid this because the clustering problem will be clear from the context.

Definition 1. *Given a point set P, let $\mathtt{OPT}_k(P)$ be the cost of the optimal solution to the clustering problem $\mathcal{C}(f, k)$ on input P.*

Definition 2. *Given a constant α, we say that a point set P is* (k, α)-irreducible *if $\mathtt{OPT}_{k-1}(P) \geq (1 + 150\alpha)\mathtt{OPT}_k(P)$. Otherwise we say that the point set is* (k, α)-reducible.

Reducibility captures the fact that if P is (k, α)-reducible for a small constant α, then the optimal solution for $\mathcal{C}(f, k-1)$ on P is close to that for $\mathcal{C}(f, k)$ on P. So if we are solving the latter problem, it is enough to solve the former one. In fact, when solving the problem $\mathcal{C}(f, k)$ on the point set P, we can assume

that P is (k,α)-irreducible, where $\alpha = \epsilon/1200k$. Indeed, suppose this is not the case. Let i be the highest integer such that P is (i,α)-irreducible. Then, $\texttt{OPT}_k(P) \leq (1+150k\alpha)^{k-i}\texttt{OPT}_i(P) \leq (1+\varepsilon/4)\texttt{OPT}_i(P)$. Therefore, if we can get a $(1+\varepsilon/4)$-approximation algorithm for $\mathcal{C}(f,i)$ on input P, then we have a $(1+\varepsilon)$-approximation algorithm for $\mathcal{C}(f,k)$ on P. Thus it is enough to solve instances which are irreducible.

The first property that we want $\mathcal{C}(f,k)$ to satisfy is a fairly obvious one – it is always better to assign a point in P to the nearest center. We state this more formally as follows :

Closeness Property **:** Let Q and Q' be two disjoint set of points, and let $q \in Q$. Suppose x and x' are two points such that $d(q,x) > d(q,x')$. Then the cost function f satisfies the following property

$$f(Q,x) + f(Q',x') \geq f(Q-\{q\},x) + f(Q' \cup \{q\},x').$$

This is essentially saying that in order to find a solution, it is enough to find the set of k centers. Once we have found the centers, the actual partitioning of P is just the Voronoi partitioning with respect to these centers. It is easy to see that the k-means problem and the k-median problem (both the continuous and the discrete versions) satisfy this property.

Definition 3. *Given a set of points P and a set of k points C, let $\texttt{OPT}_k(P,C)$ be the cost of the optimal solution to $\mathcal{C}(f,k)$ on P when the set of centers is C.*

We desire two more properties from $\mathcal{C}(f,k)$. The first property says that if we are solving $\mathcal{C}(f,1)$, then there should be a simple random sampling algorithm. The second property says that suppose we have approximated the first i centers of the optimal solution closely. Then we should be able to easily extract a large number of points in P which get assigned to these centers. We describe these properties in more detail below :

- ***Random Sampling Procedure*** **:** There exists a procedure $\mathcal{A}$ that takes a set of points $Q \in \Re^d$ and a parameter α as input. $\mathcal{A}$ first randomly samples a set R of size $\left(\frac{1}{\alpha}\right)^{O(1)}$ points from Q. Starting from R, $\mathcal{A}$ produces a set of points, which we call $\texttt{core}(R)$, of size at most $2^{\left(\frac{1}{\alpha}\right)^{O(1)}}$. $\mathcal{A}$ satisfies the condition that with constant probability there is at least one point $c \in \texttt{core}(R)$ such that $\texttt{OPT}_1(Q,\{c\}) \leq (1+\alpha)\texttt{OPT}_1(Q)$. Further the time taken by $\mathcal{A}$ to produce $\texttt{core}(R)$ from R is at most $O(2^{\left(\frac{1}{\alpha}\right)^{O(1)}} \cdot dn)$.
- ***Tightness Property*** **:** Let P be a set of points which is (k,α)-irreducible for some constant α. Consider an optimal solution to $\mathcal{C}(f,k)$ on P – let $C = \{c_1,\ldots,c_k\}$ be the centers in this solution. Suppose we have a set of i points $C'_i = \{c'_1,\ldots,c'_i\}$, such that $\texttt{OPT}_k(P,C') \leq (1+\alpha/k)^i\texttt{OPT}_k(P)$, where $C' = \{c'_1,\ldots,c'_i,c_{i+1},\ldots,c_k\}$. Let $P'_1,\ldots,P'_k$ be the partitioning of P if we choose C' as the set of centers (in other words this is the Voronoi partitioning of P with respect to C'). We assume w.l.o.g. that P'_{i+1} is the largest cluster amongst $P'_{i+1},\ldots,P'_k$. Then there exists a set of points S such that the following conditions hold :

(a) S is contained in $P'_1 \cup \ldots \cup P'_i$.
(b) Let $x \in S, x' \in P - S$. Then, $d(x, \{c'_1, \ldots, c'_i\}) \leq d(x', \{c'_1, \ldots, c'_i\})$.
(c) $P - S$ contains at most $\frac{|P'_{i+1}|}{\alpha^{O(1)}}$ points of $P'_1 \cup \ldots \cup P'_i$.

3 A General Algorithm for Clustering

We can show that if a clustering problem $\mathcal{C}(f, k)$ satisfies the conditions stated in the previous section, then there is an algorithm which with constant probability produces a solution within $(1+\varepsilon)$ factor of the optimal cost. Further the running time of this algorithm is $O(2^{\left(\frac{k}{\varepsilon}\right)^{O(1)}} \cdot dn)$. The techniques are very similar those in [14] and are omitted. We now give applications to various clustering problems. We show that these clustering problems satisfy the tightness property and admit a random sampling procedure as described in the previous section.

4 The *k*-median Problem

As described earlier, the clustering problem $\mathcal{C}(f, k)$ is said to be the k-median problem if $f(Q, x) = \sum_{q \in Q} d(q, x)$. We now exhibit the two properties for this problem.

4.1 Random Sampling Procedure

Badoiu et al. [3] showed that a small random sample can be a used to get a close approximation to the optimal 1-median solution. Given a set of points P, let $\texttt{AvgMed}(P)$ denote $\frac{\texttt{OPT}_1(P)}{|P|}$, i.e., the average cost paid by a point towards the optimal 1-median solution.

Lemma 1. *[3] Let P be a set of points in $\Re^d$, and ε be a constant between 0 and 1. Let X be a random sample of $O(1/\varepsilon^3 \log 1/\varepsilon)$ points from P. Then with constant probability, the following two events happen: (i) The flat span(X) contains a point x such that $\texttt{OPT}_1(P, \{x\}) \leq (1+\varepsilon)\texttt{OPT}_1(P)$. and (ii) X contains a point y at distance at most $2\texttt{AvgMed}(P)$ from x.*

We now show that if we can upper and lower bound $\texttt{AvgMed}(P)$ upto constant factors, then we can construct a small set of points such that at least one of these is a good approximation to the optimal center for the 1-median problem on P.

Lemma 2. *Let P be a set of points in $\Re^d$ and X be a random sample of size $O(1/\varepsilon^3 \log 1/\varepsilon)$ from P. Suppose we happen to know numbers a and b such that $a \leq \texttt{AvgMed}(P) \leq b$. Then, we can construct a set Y of $O(2^{(1/\varepsilon)^{O(1)}} \log(b/\varepsilon a))$ points such that with constant probability there is at least one point $z \in X \cup Y$ satisfying $\texttt{OPT}_1(P, \{z\}) \leq (1 + 2\varepsilon)\texttt{OPT}_1(P)$. Further, the time taken to construct Y from X is $O(2^{(1/\varepsilon)^{O(1)}} d)$.*

Proof. Our construction is similar to that of Badoiu et al. [3]. We can assume that the result stated in Lemma 1 holds (because this happens with constant probability). Let x and y be as in Lemma 1.

We will carefully construct candidate points around the points of X in $span(X)$ in an effort to get within close distance of x.

For each point $p \in X$, and each integer i in the range $[\lfloor \log \frac{\varepsilon}{4} a \rfloor, \lceil \log b \rceil]$ we do the following – let $t = 2^i$. Consider the grid $G_p(t)$ of side length $\varepsilon t/(4|X|) = O(t\varepsilon^4 \log(1/\varepsilon))$ in $span(X)$ centered at p. We add all the vertices of this grid lying within distance at most $2t$ from p to our candidate set Y. This completes the construction of Y. It is easy to see that the time taken to construct Y from X is $O(2^{(1/\varepsilon)^{O(1)}} d)$.

We now show the existence of the desired point $z \in X \cup Y$. Consider the following cases:

1. $d(y,x) \leq \varepsilon \texttt{AvgMed}(P)$: Using triangle inequality, we see that
$$f(P,y) \leq f(P,x) + |P| d(y,x) \leq (1+2\varepsilon)\texttt{OPT}_1(P).$$
Therefore y itself is the required point.
2. $d(y,x) > \varepsilon \texttt{AvgMed}(P)$: Consider the value of i such that $2^{i-1} \leq AvgMed(P,1) \leq 2^i$ – while constructing Y, we must have considered this value of i for all points in X. Let $t = 2^i$. Clearly, $t/2 \leq \texttt{AvgMed}(P) \leq t$.
Observe that $d(y,x) \leq 2\texttt{AvgMed}(P) \leq 2t$. Therefore, by the manner in which we have constructed $G_y(t)$, there must be a point $p \in G_y(t)$ fow which $d(p,x) \leq \varepsilon t/2 \leq \varepsilon \texttt{AvgMed}(P)$. This implies that
$$f(P,p) \leq f(P,x) + |P| d(x,p) \leq (1+2\varepsilon)\texttt{OPT}_1(P).$$
Therefore p is the required point.

This completes the proof of the lemma.

We now show the existence of the random sampling procedure.

Theorem 1. *Let P be a set of n points in $\Re^d$, and let ε be a constant, $0 < \varepsilon < 1/12$. There exists an algorithm which randomly samples a set R of $O((\frac{1}{\varepsilon})^{O(1)})$ points from P. Using this sample only, it constructs a set of points $\texttt{core}(R)$ such that with constant probability there is a point $x \in \texttt{core}(R)$ satisfying $f(P,x) \leq (1+O(\varepsilon))\texttt{OPT}_1(P)$. Further, the time taken to construct $\texttt{core}(R)$ from R is $O(2^{(1/\varepsilon)^{O(1)}} d)$.*

Proof. Consider the optimal 1-median solution for P – let c be the center in this solution. Let T denote $\texttt{AvgMed}(P)$. Consider the ball B_1 of radius T/ε^2 around c. Let P' be the points of P contained in B_1. It is easy to see that $|P'| \geq (1-\varepsilon^2)n$.

Sample a point p at random from P. With constant probability, it lies in P'. Randomly sample a set Q of $1/\varepsilon$ points from P. Again, with constant probability, these points lie in P'. So we assume that these two events happen. Let $v = \sum_{q \in Q} d(q,p)$. We want to show that v is actually close to $\texttt{AvgMed}(P)$.

Let B_2 denote the ball of radius εT centered at p. One of the following two cases must happen :

- There are at least $2\varepsilon|P'|$ points of P' outside B_2 : In this case, with constant probability, the sample Q contains a point outside B_2. Therefore, $v \geq \varepsilon T$. Also notice that any two points in B_1 are at distance at most $2T/\varepsilon^2$ from each other. So, $v \leq 2T|Q|/\varepsilon^2$. We choose $a = \frac{v\varepsilon^2}{2|Q|}$ and $b = v/\varepsilon$. Notice that b/a is $O(1/\varepsilon^{O(1)})$. We can now use the Lemma 2 to construct the desired core set.
- There are at most $2\varepsilon|P'|$ points of P' outside $\mathcal{B}_2$: Suppose $d(p,c) \leq 4\varepsilon T$. In this case $f(P,p) \leq (1+O(\varepsilon))\mathtt{OPT}_1(P)$ and we are done. So assume this is not the case. Note that the number of points outside B_2 is at most $|P - P'| + 2\varepsilon|P'| \leq \varepsilon^2 n + 2\varepsilon(1-\varepsilon^2)n \leq 3\varepsilon n$. Now suppose we assign all points of P from c to p. Let us see the change in cost. The distance the points in B_2 have to travel decreases by at least $d(c,p) - 2\varepsilon T$. The increase in the distance for points outside B_2 is at most $d(c,p)$. So the overall decrease in cost is at least

$$|B_2|(d(c,p) - 2\varepsilon T) - (n - |B_2|)d(c,p) > 0$$

  if we use $|B_2| \geq n(1-3\varepsilon)$ and $d(c,p) \geq 4\varepsilon T$. This yields a contradiction because c is the optimal center. Thus we are done in this case as well.

4.2 Tightness Property

We now show the existence of tightness property. We will use the same notation as used while defining the tightness property in Section 2. We need to show the existence of the desired set S.

Consider the closest pair of centers between the sets $C' \backslash C'_i$ and C'_i – let these centers be c_l and c'_r respectively. Let $t = d(c_l, c'_r)$. Let S be the set of points $\mathcal{B}(c'_1, t/4) \cup \cdots \cup \mathcal{B}(c'_i, t/4)$, i.e., the points which are distant at most $t/4$ from $C'_i = \{c'_1, \ldots, c'_i\}$.

Clearly, S is contained in $P'_1 \cup \cdots \cup P'_i$. This shows (a). Also, for any $x \in S, x' \in P - S$, $d(x, \{c'_1, \ldots, c'_i\}) \leq d(x', \{c'_1, \ldots, c'_i\})$. This proves (b).

Suppose $P - S$ contains more than $|P_l|/\alpha$ points of $P'_1 \cup \cdots \cup P'_i$. In that case, these points are assigned to centers at distance at least $t/4$. It follows that $\mathtt{OPT}_k(P, C')$ is at least $\frac{t|P_l|}{4\alpha}$. This implies that $t|P_l| \leq 4\alpha\mathtt{OPT}_k(P, C')$. But then if we assign all the points in P_l to c'_r, the cost increases by at most

$$|P_l|t \leq 4\alpha\mathtt{OPT}_k(P,C') \leq 4\alpha(1+\alpha/k)^i\mathtt{OPT}_k(P) \leq 4\alpha(1+\alpha/k)^k\mathtt{OPT}_k(P) \leq 12\alpha\mathtt{OPT}_k(P).$$

But this contradicts the fact that P is (k,α)-irreducible.

4.3 Applications to the 1-median Problem

In this section, we present an algorithm for the 1-median problem. Given a set of n points in $\Re^d$, the algorithm with constant probability produces a solution of cost at most $(1+\varepsilon)$ of the optimal cost for any constant $\varepsilon > 0$. The running time of the algorithm is $O(2^{1/\varepsilon^{O(1)}}d)$, assuming that it is possible to randomly sample a point in constant time.

Our algorithm is based on the following idea presented by Indyk [12].

Lemma 3. *[12] Let X be a set of n points in $\Re^d$. For a point $a \in \Re^d$ and a subset $Q \subseteq X$, define $S_Q(a) = \sum_{x \in Q} d(a,x)$ and $S(a) = S_X(a)$. Let ε be a constant, $0 \leq \varepsilon \leq 1$. Suppose a and b are two points such that $S(b) > (1+\varepsilon)S(a)$. Then,*

$$Pr\left(\sum_{x \in Q} d(a,x) \geq \sum_{x \in Q} d(b,x)\right) < e^{-\varepsilon^2 |Q|/64}.$$

We now show the existence of a fast algorithm for approximating the optimal 1-median solution.

Theorem 2. *Let P be a set of n points in $\Re^d$, and let ε be a constant, $0 < \varepsilon < 1$. There exists an algorithm which randomly samples a set R of $O((\frac{1}{\varepsilon})^{O(1)})$ points from P. Using this sample only, it finds a point p such that $f(P,x) \leq (1 + O(\varepsilon))\mathtt{OPT}_1(P)$ with constant probability (independent of ε). The time taken by the algorithm to find such a point p from R is $O(2^{(1/\varepsilon)^{O(1)}} d)$.*

Proof. We first randomly sample a set R_1 of $O((\frac{1}{\varepsilon})^{O(1)})$ points from P and using Theorem 1, construct a set $\mathtt{core}(R_1)$ of $O(2^{(1/\varepsilon)^{O(1)}})$ points such that with constant probability, there is a point $x \in \mathtt{core}(R_1)$ satisfying $f(P,x) \leq (1 + O(\varepsilon))\mathtt{OPT}_1(P)$.

Now we randomly sample a set R_2 of $O((1/\varepsilon)^{O(1)})$ points and find the point $p \in \mathtt{core}(R_1)$ for which $S_{R_2}(p) = f(R_2, p)$ is minimum. By Lemma 3, p is with constant probability a $(1 + O(\varepsilon))$-approximate median of P.

Clearly, the time taken by the algorithm is $O(2^{(1/\varepsilon)^{O(1)}} d)$.

Also note that we can boost the success probability to an arbitrarily small constant by selecting a large enough (yet constant) sample R.

5 k-means clustering

In this problem, $f(Q,x) = \sum_{q \in Q} d(q,x)^2$. The two properties for the k-means problem were shown by the authors in [14]. For a set of points T, let $c(T)$ denote their centroid. The random sampling property follows from the following fact showed by Inaba et al. [11].

Lemma 4. **[11]** *Let T be a set of m points obtained by independently sampling m points uniformly at random from a point set P. Then, for any $\delta > 0$,*

$$f(S, c(T)) < \left(1 + \frac{1}{\delta m}\right)\mathtt{OPT}_1(P)$$

holds with probability at least $1 - \delta$.

The proof of tightness property is similar to that for the k-median problem.

6 Discrete k-means Clustering

This is same as k-means problem with the extra constraint that the centers must be from the input point set only. We now show the two properties here.

6.1 Random Sampling Procedure

We first show that given a good approximation to the center of the optimal (continuous) 1-means problem, we can get a good approximation to the center of the optimal discrete 1-means problem. Let us have some notation first. Let P be a set of n points in $\Re^d$. Let c be the center of the optimal solution to the (continuous) 1-means problem on P.

Lemma 5. *Let α be a constant, $0 < \alpha < 1$, and c' be a point in $\Re^d$ such that $\sum_{p\in P} d(p,c')^2 \leq (1+\alpha)\sum_{p\in P} d(p,c)^2$. Let x' be the point of P closest to c' Then $\texttt{OPT}_1(P,\{x'\}) \leq (1+O(\sqrt{\alpha}))\texttt{OPT}_1(P)$.*

Proof. Let x be the center of the optimal discrete 1-means solution, i.e., $\texttt{OPT}_1(P,\{x\}) = \texttt{OPT}_1(P)$. Let T be the average cost paid by the points of P in the optimal 1-means solution, i.e., $T = \frac{\sum_{p\in P} d(p,c)^2}{|P|}$.

Then $\texttt{OPT}_1(P) = |P|(T + d(c,x)^2)$ and $\texttt{OPT}_1(P,\{x'\}) = |P|(T + d(c,x')^2)$. From the definition of c', we know that $d(c,c')^2 \leq \alpha T$. Notice that

$$d(c,x') \leq d(c,c') + d(c',x') \leq d(c,c') + d(c',x) \leq 2d(c,c') + d(c,x).$$

We know that $f(P,x) = |P|(T + d(c,x)^2)$ and $f(P,x') = |P|(T + d(c,x')^2)$. So

$$\begin{aligned} f(P,x') - f(P,x) = |P|(d(c,x')^2 - d(c,x)^2) &\leq |P|\left((2d(c,c') + d(c,x))^2 - d(c,x)^2\right) \\ &\leq 4|P|\left(d(c,c)^2 + d(c,c')d(c,x)\right) \leq 4|P|\left(\alpha T + \sqrt{\alpha T}d(c,x)\right) \\ &\leq 4|P|\left(\alpha T + \sqrt{\alpha}(T + d(c,x)^2)\right) \leq O(\sqrt{\alpha})\texttt{OPT}_1(P). \end{aligned}$$

We now show the existence of the random sampling procedure.

Theorem 3. *Let α be a constant, $0 < \alpha < 1$. There exists an algorithm which randomly samples a set R of $O\left(\frac{1}{\alpha}\right)$ points from P. Using this sample, it finds a singleton set $\texttt{core}(R)$ such that with constant probability the point $x \in \texttt{core}(R)$ satisfies $f(P,x) \leq (1+O(\sqrt{\alpha}))\texttt{OPT}_1(P)$. Further, the time taken to construct $\texttt{core}(R)$ from R is $O((\frac{1}{\alpha} + n)d)$.*

Proof. Using Lemma 4, we can get a point c' such that $\sum_{p\in P} d(p,c')^2 \leq (1+\alpha)\sum_{p\in P} d(p,c)^2$. As mentioned in the lemma, we do this by by taking the centroid of a random sample of $O(1/\alpha)$ points of P. This takes time $O(\frac{1}{\alpha}\cdot d)$.

The rest follows from the previous lemma.

6.2 Tightness Property

We now show the existence of tightness property. We will use the same notation as used while defining the tightness property in Section 2. We need to show the existence of the desired set S.

Consider the closest pair of centers between the sets $\{c'_1, \ldots, c'_i\}$ and $\{c_{i+1}, \ldots, c_k\}$ – let these centers be c'_r and c_l respectively. Let $t = d(c_l, c'_r)$. Let S be the set of points $\mathcal{B}(c'_1, t/4) \cup \cdots \cup \mathcal{B}(c'_i, t/4)$, i.e., the points which are distant at most $t/4$ from $C'_i = \{c'_1, \ldots, c'_i\}$.

Clearly, S is contained in $P'_1 \cup \cdots \cup P'_i$. This shows (a). Also, for any $x \in S, x' \in P - S$, $d(x, \{c'_1, \ldots, c'_i\}) \leq d(x', \{c'_1, \ldots, c'_i\})$. This proves (b).

Suppose $P - S$ contains more than $|P_l|/\alpha^2$ points of $P'_1 \cup \cdots \cup P'_i$. In that case, these points are assigned to centers at distance at least $t/4$. It follows that $\mathtt{OPT}_k(P, C')$ is at least $\frac{t^2|P_l|}{16\alpha^2}$. This implies that $t^2|P_l| \leq 16\alpha^2 \mathtt{OPT}_k(P, C')$.

Let m_l and m'_r be the centers of the optimal (continuous) 1-means solution of P_l and P'_r respectively. Further, let T_l and T'_r be the average cost paid by P_l and P'_r in this optimal solution respectively, i.e., $T_l = \frac{\sum_{p \in P_l} d(p, m_l)^2}{|P_l|}$ and $T'_r = \frac{\sum_{p \in P'_r} d(p, m'_r)^2}{|P'_r|}$. Observe that $f(P_l, c_l) = |P_l|(T_l + d(c_l, m_l)^2)$ and $f(P_l, c'_r) = |P_l|(T_l + d(c'_r, m_l)^2)$. Therefore, if we assign the points in P_l from c_l to c'_r, the increase in cost is

$$\begin{aligned} |P_l| \left(d(c'_r, m_l)^2 - d(c_l, m_l)^2 \right) &\leq |P_l| \left((d(c'_r, c_l) + d(c_l, m_l))^2 - d(c_l, m_l)^2 \right) \\ &\leq |P_l| \left(t^2 + 2t d(c_l, m_l) \right) \end{aligned}$$

We know that the first term above, i.e., $|P_l|t^2$ is at most $16\alpha^2 \mathtt{OPT}_k(P, C')$. We now need to bound the second term only. We consider two cases

- $t \leq \alpha d(c_l, c_m)$: In this case, $|P_l| \cdot 2td(c_l, m_l) \leq 2\alpha d(c_l, m_l)^2 |P_l| \leq 2\alpha f(P_l, c_l) \leq 2\alpha \mathtt{OPT}_k(P, C')$.
- $t > \alpha d(c_l, c_m)$: In this case, $|P_l| \cdot 2td(c_l, m_l) \leq \frac{2t^2|P_l|}{\alpha} \leq 32\alpha \mathtt{OPT}_k(P, C')$.

Thus, in either case, the cost increases by at most

$$48\alpha \mathtt{OPT}_k(P, C') \leq 48\alpha(1+\alpha/k)^i \mathtt{OPT}_k(P) \leq 48\alpha(1+\alpha/k)^k \mathtt{OPT}_k(P) \leq 144\alpha \mathtt{OPT}_k(P).$$

But this contradicts the fact that P is (k, α)-irreducible.

7 Concluding Remarks

The algorithm can also be extended to solve the above clustering problems when each of the points have an associated (integral) weight with total weight W. The solution to the above clustering problems for the weighted version is the same as the solution to the unweighted version where a point p with weight w is replaced by w points of unit weight. It can be verified that for handling the weighted case: the closeness property remains unchanged; in condition (c) for

the tightness property, the size of the set gets replaced by the weight of the set; The random sampling procedure requires time at most linear in n (number of remaining distinct points) in order to perform the required weighted sampling. The running time thus obtained for the algorithm in the weighted case is $O(2^{(k/\varepsilon)^{O(1)}} n \cdot d\log^k W)$.

References

1. Arora, S.: Polynomial time approximation schemes for Euclidean TSP and other geometric problems. Proceedings of the 37th Annual Symposium on Foundations of Computer Science (1996) 2–11
2. Arora, S., Raghavan, P., Rao, S.: Approximation schemes for Euclidean k-medians and related problems. Proceedings of the thirtieth annual ACM symposium on Theory of computing (1998) 106–113
3. Badoiu, M., Har-Peled, S., Indyk, P.: Approximate clustering via core-sets. Proceedings of the thiry-fourth annual ACM symposium on Theory of computing (2002) 250–257
4. Bern, M., Eppstein, D.: Approximation algorithms for geometric problems. Approximating algorithms for NP-Hard problems. PWS Publishing Company (1997) 296–345
5. Broder, A., Glassman, S., Manasse, M., Zweig, G.: Syntactic clustering of the Web. Proc. of 6th International World Wide Web Conference (1997) 391–404
6. de la Vega, W. F., Karpinski, M., Kenyon, C., Rabani, Y.: Approximation schemes for clustering problems. Proceedings of the thirty-fifth annual ACM symposium on Theory of computing (2003) 50–58
7. Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A.: Indexing by latent semantic analysis. Journal of the American Society for Information Science **41(6)** (1990) 391–407
8. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification. Wiley-Interscience, New York, 2nd edition (2001)
9. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and effective querying by image content. Journal of Intelligent Information Systems **3(3)** (1994) 231–262
10. Har-Peled, S., Mazumdar, S.: On coresets for k-means and k-median clustering. Proceedings of the thirty-sixth annual ACM symposium on Theory of computing (2004) 291–300
11. Inaba, M., Katoh, N., Imai, H.: Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. Proceedings of the tenth annual symposium on Computational Geometry (1994) 332–339
12. Indyk, P.: High Dimensional Computational Geometry. Ph.D. Thesis. Department of Computer Science, Stanford University (2004)
13. Kolliopoulos, S., Rao, S.: A nearly linear time approximation scheme for the Euclidean k-medians problem. Proceedings of the 7th European Symposium on Algorithms (1999) 362-371
14. Kumar, A., Sabharwal, Y., Sen, S.: A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. Proceedings of the 45th Annual Symposium on Foundations of Computer Science (2004) 454–462
15. Matousek, J.: On approximate geometric k-clustering Discrete and Computational Geometry **24** (2000) 61–84