

# Linear Time Algorithms for Clustering Problems in any dimensions

Amit Kumar<sup>1</sup>, Yogish Sabharwal<sup>2</sup>, and Sandeep Sen<sup>3</sup>

<sup>1</sup> Dept of Comp Sc & Engg, Indian Institute of Technology, New Delhi-110016, India  
amitk@cse.iitd.ernet.in

<sup>2</sup> IBM India Research Lab, Block-I, IIT Delhi, Hauz Khas, New Delhi-110016, India  
ysabharwal@in.ibm.com

<sup>3</sup> 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  $\varepsilon$  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, 16, 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. However, 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  $\varepsilon$  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 running 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 $d = 2$ $O(n + \rho k^{O(1)} \log^{O(1)} n)$ (discrete also) where $\rho = \exp[O((1 + \log 1/\varepsilon)/\varepsilon)^{d-1}]$	Arora [1] Har-Peled et al. [10]
discrete $k$ -median	$O(\rho n \log n \log k)$	Kolliopoulos et al. [13]
$k$ -means	$O(n/\varepsilon^d)$ for $k = 2$ $O(n\varepsilon^{-2k^2 d} \log^k n)$ $O(g(k, \varepsilon) n \log^k n)$ $g(k, \varepsilon) = \exp[(k^3/\varepsilon^8)(\ln(k/\varepsilon) \ln k)]$ $O(n + k^{k+2} \varepsilon^{-(2d+1)k} \log^{k+1} n \log^k \frac{1}{\varepsilon})$ $O(2^{(k/\varepsilon)^{O(1)}} dn)$	Inaba et al. [11] Matousek [15] de la Vega et al. [6] Har-Peled et al. [10] 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 satisfied, 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  $\mathbb{R}^d$ , 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. Further the running time of this algorithm should be independent of  $n$ . 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, there 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  $\varepsilon$ .

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  $C_1, \dots, C_k$ . A solution also finds  $k$  points, which we call *centers*,  $c_1, \dots, c_k$ . 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  $\text{OPT}_k(P)$  be the cost of the optimal solution to the clustering problem  $\mathcal{C}(f, k)$  on input  $P$ .*

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

Reducibility captures the fact that if  $P$  is  $(k, \alpha)$ -reducible for a small constant  $\alpha$ , 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  $(k, \alpha)$ -irreducible, where  $\alpha = \epsilon/1200k$ . Indeed, suppose this is not the case. Let  $i$  be the highest integer such that  $P$  is  $(i, \alpha)$ -irreducible. Then,  $\text{OPT}_k(P) \leq (1 + 150k\alpha)^{k-i} \text{OPT}_i(P) \leq (1 + \epsilon/4) \text{OPT}_i(P)$ . Therefore, if we can get a  $(1 + \epsilon/4)$ -approximation algorithm for  $\mathcal{C}(f, i)$  on input  $P$ , then we have a  $(1 + \epsilon)$ -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  $\text{OPT}_k(P, C)$  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 the 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  $Q \in \mathbb{R}^d$  and a parameter  $\alpha$  as input.  $\mathcal{A}$  first randomly samples a set  $R$  of size  $(\frac{1}{\alpha})^{O(1)}$  points from  $Q$ . Starting from  $R$ ,  $\mathcal{A}$  produces a set of points, which we call  $\text{core}(R)$ , of size at most  $2^{(\frac{1}{\alpha})^{O(1)}}$ .  $\mathcal{A}$  satisfies the condition that with constant probability there is at least one point  $c \in \text{core}(R)$  such that  $\text{OPT}_1(Q, \{c\}) \leq (1 + \alpha) \text{OPT}_1(Q)$ . Further the time taken by  $\mathcal{A}$  to produce  $\text{core}(R)$  from  $R$  is at most  $O(2^{(\frac{1}{\alpha})^{O(1)}} \cdot dn)$ .
- **Tightness Property :** Let  $P$  be a set of points which is  $(k, \alpha)$ -irreducible for some constant  $\alpha$ . Consider an optimal solution to  $\mathcal{C}(f, k)$  on  $P$  – let  $C = \{c_1, \dots, c_k\}$  be the centers in this solution. Suppose we have a set of  $i$  points  $C'_i = \{c'_1, \dots, c'_i\}$ , such that  $\text{OPT}_k(P, C') \leq (1 + \alpha/k)^i \text{OPT}_k(P)$ , where  $C' = \{c'_1, \dots, c'_i, c_{i+1}, \dots, c_k\}$ . Let  $P'_1, \dots, P'_k$  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  $P'_{i+1}$  be the largest cluster amongst  $P'_{i+1}, \dots, P'_k$ . Then there exists a set of points  $S$  such that the following conditions hold :

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

### 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  $O(2^{\frac{k}{\varepsilon}})^{O(1)} \cdot dn$ . The techniques are very similar to the ones used by the authors in [14]. We defer the proof to the appendix. 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 used to get a close approximation to the optimal 1-median solution. Given a set of points  $P$ , let  $\text{AvgMed}(P)$  denote  $\frac{\text{OPT}_1(P)}{|P|}$ , 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  $\mathbb{R}^d$ , and  $\varepsilon$  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  $\text{span}(X)$  contains a point  $x$  such that  $\text{OPT}_1(P, \{x\}) \leq (1 + \varepsilon)\text{OPT}_1(P)$ . and (ii)  $X$  contains a point  $y$  at distance at most  $2\text{AvgMed}(P)$  from  $x$ .*

We now show that if we can upper and lower bound  $\text{AvgMed}(P)$  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  $\mathbb{R}^d$  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  $a \leq \text{AvgMed}(P) \leq b$ . 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  $\text{OPT}_1(P, \{z\}) \leq (1 + 2\varepsilon)\text{OPT}_1(P)$ . 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  $\text{span}(X)$  in an effort to get within close distance of  $x$ .

For each point  $p \in X$ , and each integer  $i$  in the range  $[\lceil \log_{\frac{\varepsilon}{4}} a \rceil, \lceil \log b \rceil]$  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  $\text{span}(X)$  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.  $d(y, x) \leq \varepsilon \text{AvgMed}(P)$  : Using triangle inequality, we see that

$$f(P, y) \leq f(P, x) + |P|d(y, x) \leq (1 + 2\varepsilon)\text{OPT}_1(P).$$

Therefore  $y$  itself is the required point.

2.  $d(y, x) > \varepsilon \text{AvgMed}(P)$  : Consider the value of  $i$  such that  $2^{i-1} \leq \text{AvgMed}(P, 1) \leq 2^i$  – while constructing  $Y$ , we must have considered this value of  $i$  for all points in  $X$ . Let  $t = 2^i$ . Clearly,  $t/2 \leq \text{AvgMed}(P) \leq t$ .

Observe that  $d(y, x) \leq 2\text{AvgMed}(P) \leq 2t$ . Therefore, by the manner in which we have constructed  $G_y(t)$ , there must be a point  $p \in G_y(t)$  for which  $d(p, x) \leq \varepsilon t / 2 \leq \varepsilon \text{AvgMed}(P)$ . This implies that

$$f(P, p) \leq f(P, x) + |P|d(x, p) \leq (1 + 2\varepsilon)\text{OPT}_1(P).$$

Therefore  $p$  is the required point.

This completes the proof of the lemma.

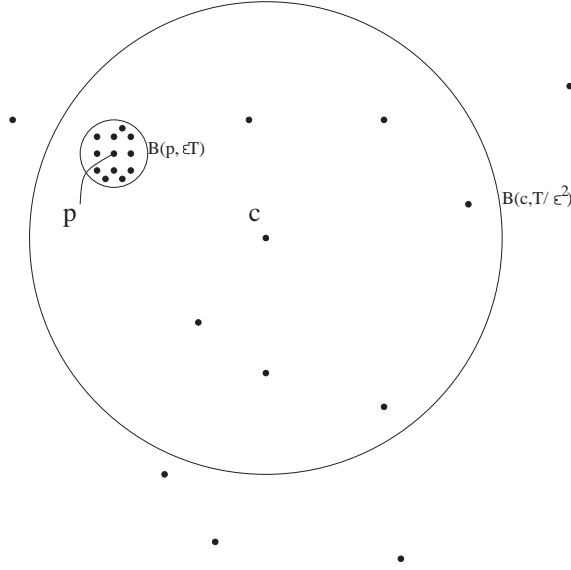
We now show the existence of the random sampling procedure.

**Theorem 1.** *Let  $P$  be a set of  $n$  points in  $\mathbb{R}^d$ , and let  $\varepsilon$  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  $\text{core}(R)$  such that with constant probability there is a point  $x \in \text{core}(R)$  satisfying  $f(P, x) \leq (1 + O(\varepsilon))\text{OPT}_1(P)$ . Further, the time taken to construct  $\text{core}(R)$  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  $\text{AvgMed}(P)$ . Consider the ball  $B_1$  of radius  $T/\varepsilon^2$  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  $\text{AvgMed}(P)$ .

Let  $B_2$  denote the ball of radius  $\varepsilon T$  centered at  $p$ . One of the following two cases must happen :



**Fig. 1.** Points distribution around the Median

- 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  $a = v/\varepsilon$  and  $b = \frac{v\varepsilon^2}{2|Q|}$ . Notice that  $b/a$  is  $O(\varepsilon^{O(1)})$ . We can now use the Lemma 2 to construct the desired core set.
- There are at most  $2\varepsilon|P'|$  points of  $P'$  outside  $B_2$  : Suppose  $d(p, c) \leq 4\varepsilon T$ . In this case  $f(P, p) \leq (1 + O(\varepsilon))\text{OPT}_1(P)$  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  $d(c, p) - 2\varepsilon R$ . 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

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

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

This proves the theorem.

Thus we have shown the existence of the random sampling procedure.

## 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  $C' \setminus C'_i$  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  $\mathcal{B}(c'_1, t/4) \cup \dots \cup \mathcal{B}(c'_i, t/4)$ , i.e., the points which are distant at most  $t/4$  from  $C'_i = \{c'_1, \dots, c'_i\}$ .

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

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

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

But this contradicts the fact that  $P$  is  $(k, \alpha)$ -irreducible. This proves the tightness property.

### 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  $\mathbb{R}^d$ , 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  $\mathbb{R}^d$ . For a point  $a \in \mathbb{R}^d$  and a subset  $Q \subseteq X$ , define  $S_Q(a) = \sum_{x \in Q} d(a, x)$  and  $S(a) = S_X(a)$ . Let  $\varepsilon$  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  $\mathbb{R}^d$ , and let  $\varepsilon$  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  $f(P, p) \leq (1 + O(\varepsilon))\text{OPT}_1(P)$  with constant probability (independent of  $\varepsilon$ ). 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  $\text{core}(R_1)$  of  $O(2^{(1/\varepsilon)^{O(1)}})$  points such that with constant probability, there is a point  $x \in \text{core}(R_1)$  satisfying  $f(P, x) \leq (1 + O(\varepsilon))\text{OPT}_1(P)$ .

Now we randomly sample a set  $R_2$  of  $O((1/\varepsilon)^{O(1)})$  points and find the point  $p \in \text{core}(R_1)$  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

Recall that 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]. But we show briefly this for the sake of completeness. 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  $\delta > 0$ ,*

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

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

The proof of the tightness property is very similar to the one shown for the  $k$ -median problem.

## 6 Discrete $k$ -means Clustering

This is the same as the  $k$ -means problem with the additional constraint that the centers must be chosen from the input point set only.

We now exhibit the two properties for this problem.

### 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  $\mathbb{R}^d$ . Let  $c$  be the center of the optimal solution to the (continuous) 1-means problem on  $P$ .

**Lemma 5.** *Let  $\alpha$  be a constant,  $0 < \alpha < 1$ , and  $c'$  be a point in  $\mathbb{R}^d$  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  $\text{OPT}_1(P, \{x'\}) \leq (1 + O(\sqrt{\alpha})) \text{OPT}_1(P)$ .*

*Proof.* Let  $x$  be the center of the optimal discrete 1-means solution, i.e.,  $\text{OPT}_1(P, \{x\}) = \text{OPT}_1(P)$ . 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  $\text{OPT}_1(P) = |P|(T + d(c, x)^2)$  and  $\text{OPT}_1(P, \{x'\}) = |P|(T + d(c, x')^2)$ . 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

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

We now show the existence of the random sampling procedure.

**Theorem 3.** *Let  $\alpha$  be a constant,  $0 < \alpha < 1$ . There exists an algorithm which randomly samples a set  $R$  of  $O(\frac{1}{\alpha})$  points from  $P$ . Using this sample, it finds a singleton set  $\text{core}(R)$  such that with constant probability the point  $x \in \text{core}(R)$  satisfies  $f(P, x) \leq (1 + O(\sqrt{\alpha}))\text{OPT}_1(P)$ . Further, the time taken to construct  $\text{core}(R)$  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 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  $\{c'_1, \dots, c'_i\}$  and  $\{c_{i+1}, \dots, c_k\}$  – let these centers be  $c'_r$  and  $c_l$  respectively. Let  $t = d(c_l, c'_r)$ . Let  $S$  be the set of points  $\mathcal{B}(c'_1, t/4) \cup \dots \cup \mathcal{B}(c'_i, t/4)$ , i.e., the points which are distant at most  $t/4$  from  $C'_i = \{c'_1, \dots, c'_i\}$ .

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

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

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

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

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

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

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

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

But this contradicts the fact that  $P$  is  $(k, \alpha)$ -irreducible. This proves the tightness property.

## References

1. S. Arora, *Polynomial time approximation schemes for Euclidean TSP and other geometric problems*, Proceedings of the 37th annual IEEE Symposium on FOCS, 1996, pp. 2-11.
2. S. Arora, P. Raghavan, and S. Rao, *Polynomial time approximation schemes for the Euclidean  $k$ -median problem*, Proceedings of the 30th annual ACM STOC, 1998.
3. M. Badoiu, S. Har-Peled, P. Indyk, *Approximate clustering via core-sets*, STOC 2002, pp. 250-257.
4. M. Bern and D. Eppstein, *Approximation algorithms for geometric problems*, D. S. Hauchbaum, editor, Approximating algorithms for NP-Hard problems. PWS Publishing Company, 1997.
5. A. Broder, S. Glassman, M. Manasse, and G. Zweig, *Syntactic clustering of the Web*, Proceedings of the 6th Int'l World Wide Web Conf (WWW), 1997, pp. 391-404.
6. W. F. de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani, *Approximation schemes for clustering problems*, Proceedings of the 35th Annual Symposium on Theory of Computing, 2003, pp. 50-58.
7. S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas and R. A. Harshman, *Indexing by latent semantic analysis*, Journal of the Society for Information Science, 41(6):391-407, 1990.
8. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, New York, 2nd edition, 2001.

9. C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic and W. Equitz, *Efficient and effective querying by image content*, Journal of Intelligent Information Systems, 3(3):231-262, 1994.
10. S. Har-Peled, S. Mazumdar, *Coresets for  $k$ -Means and  $k$ -Median Clustering and their Applications*, Proceedings of the 36th Annual Symposium on Theory of Computing, 2004, pp. 291-300.
11. M. Inaba, N. Katoh and H. Imai, *Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based  $k$ -Clustering*, Proceedings of the 10th Annual ACM Symposium on Computational Geometry, 1994, pp. 332-339.
12. P. Indyk, *High Dimensional Computational Geometry*, Ph.D. Thesis, Department of Computer Science, Stanford University, September 2004.
13. S. Kolliopoulos, and S. Rao, *A nearly linear time approximation scheme for the Euclidean  $k$ -medians problem*, Proceedings of the 7th European Symposium on Algorithms, volume 1643 of Lecture Notes in Computer Science, 1999, pp. 362-371.
14. A. Kumar, Y. Sabharwal, S. Sen, *A simple linear time  $(1 + \epsilon)$ -approximation algorithm for  $k$ -means clustering in any dimensions*, Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science 2004, pp. 454-462.
15. J. Matousek, *On approximate geometric  $k$ -clustering*, Discrete and Computational Geometry, 24, 2000, pp. 61-84.
16. M. J. Swain, and D. H. Ballard, *Color indexing*, International Journal of Computer Vision, 7:11-32, 1991.

## APPENDIX

### 7 Proofs for the General Algorithm for Clustering

In this section, we provide proofs for our general algorithm for the clustering problems we defined.

Fix a clustering problem  $\mathcal{C}(f, k)$ . Fix an instance consisting of a set  $P$  of  $n$  points in  $\mathbb{R}^d$ . Suppose we are given a constant  $\varepsilon > 0$ . We would like to construct a solution for the clustering problem whose cost is within  $(1 + \varepsilon)$  of the optimal cost.

Suppose  $P'$  is a subset of  $P$  and we want to get a good approximation to the optimal 1-center for the point set  $P'$ . Following the *Random Sampling Procedure* for the clustering problem, we would like to sample from  $P'$ . But the problem is that  $P'$  is not explicitly given to us. The following lemma states that if the size of  $P'$  is close to that of  $P$ , then we can sample a slightly larger set of points from  $P$  and hopefully this sample would contain enough random samples from  $P'$ . Let us define things more formally first. Let  $P$  be a set of points and  $P'$  be a subset of  $P$  such that  $|P'| \geq \beta|P|$ , where  $\beta$  is a constant between 0 and 1. Suppose that we require a sample of  $\lambda = O((1/\alpha)^{O(1)})$  points from  $P'$  to generate the candidate center set. Suppose we take a sample  $S$  of size  $O(\frac{4}{\beta\lambda})$  from  $P$ . Now we consider all possible subsets of size  $\lambda$  of  $S$ . For each of these subsets  $S'$ , we can generate a candidate centre set for the 1-center for the clustering problem using the *Random Sampling Procedure* and consider each one of these as a potential 1-center for the clustering problem instance on  $P'$ . The following lemma shows that one of these subsets must give a close enough approximation to the optimal 1-center solution for  $P'$ .

**Lemma 6.** (*Superset Sampling Lemma*) *Let  $\mathcal{CEN}(S')$  be the centre set generated using the Random Sampling Procedure on sampled subset  $S'$ . Then, the following event happens with constant probability*

$$\min_{c' \in \mathcal{CEN}(S'): S' \subset S, |S'| = \lambda} f(P', c') \leq (1 + \varepsilon) \text{OPT}_1(P').$$

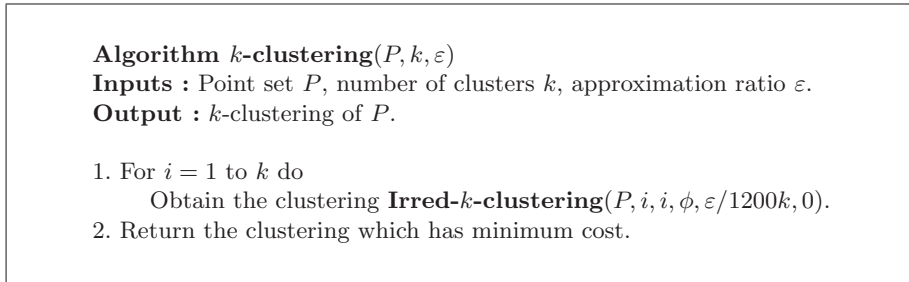
*Proof.* With constant probability,  $S$  contains at least  $\lambda$  points from  $P'$ , the required sample size. The rest follows from the *Random Sampling Procedure* for the clustering problem.

We first present a brief outline of the algorithm.

#### 7.1 Outline

We can assume that the solution is irreducible, i.e., removing one of the centers does not create a solution which has cost within a small factor of the optimal solution.

We start with  $k$  optimal (unknown) centers. In each iteration, we will consider the optimal clustering formed by  $i$  currently known centers and  $k - i$  optimal



**Fig. 2.** The  $k$ -clustering Algorithm

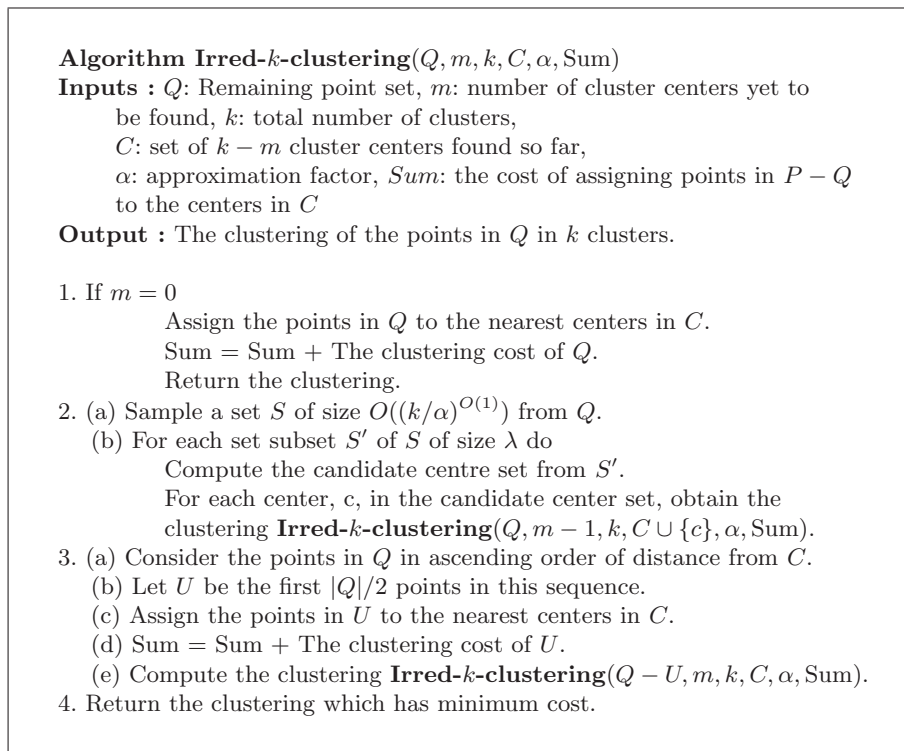
(unknown) centers. Call this the optimal clustering of the current iteration. Our goal will be to approximate the next largest cluster, so that the resulting clustering is an approximation to the optimal clustering of the current iteration. We will bound the overall approximation factor to within a factor of  $(1 + \varepsilon)$  of the optimal clustering.

Suppose we have found centers  $c'_1, \dots, c'_i$ . Suppose  $t$  is the distance between the closest pair of centers  $\{c'_1, \dots, c'_i\}$  and the unknown centers of the optimal clustering  $\{c_{i+1}, \dots, c_k\}$ . We can show that the points at distant at most  $t/4$  from  $\{c'_1, \dots, c'_i\}$  get assigned to  $c_1, \dots, c_i$  by the optimal solution induced by these  $k$  centers. So, we can delete these points. wlog, we assume that the largest cluster from amongst those centered around the unknown clusters  $\{c_{i+1}, \dots, c_k\}$  is centered around  $c_{i+1}$ . Let  $P'_{i+1}$  be this clustering. Now we can show that among the remaining points, the size of  $P'_{i+1}$  is significant. Therefore, we can use random sampling to obtain a center  $c'_{i+1}$  which is a pretty good estimate of  $c_{i+1}$ . Of course we do not know the value of  $t$ , and so a naive implementation of this idea gives an  $O(n(\log n)^k)$  time algorithm.

But now we want to modify it to a linear time algorithm. This is where the algorithm gets more involved. As mentioned above, we can not guess the parameter  $t$ . So we try to guess the size of the point set obtained by removing the balls of radius  $t/4$  around  $\{c_1, \dots, c_i\}$ . So we work with the remaining point set with the hope that the time taken for this remaining point set will also be small and so the overall time will be linear. Now, we describe the actual clustering algorithm.

## 7.2 The Algorithm

The algorithm is described in Figures 2 and 3. Figure 2 is the main algorithm. The inputs are the point set  $P$ ,  $k$  and an approximation factor  $\varepsilon$ . Let  $\alpha$  denote  $\varepsilon/1200k$ . The algorithm  $k$ -clustering( $P, k, \varepsilon$ ) tries to find the highest  $i$  such that  $P$  is  $(i, \alpha)$ -irreducible. Essentially we are saying that it is enough to find  $i$  centers only. Since we do not know this value of  $i$ , the algorithm tries all possible values of  $i$ .



**Fig. 3.** The irreducible  $k$ -clustering algorithm

We now describe the algorithm **Irred- $k$ -clustering**( $Q, m, k, C, \alpha, \text{Sum}$ ). We have found a set  $C$  of  $k - m$  centers already. The points in  $P - Q$  have been assigned to  $C$ . We need to assign the remaining points in  $Q$ . The case  $m = 0$  is clear. In step 2, we try to find a new center by the *Random Sampling Procedure* for the clustering problem. This will work provided a good fraction of the points in  $Q$  do not get assigned to  $C$ . If this is not the case then in step 3, we assign half of the points in  $Q$  to  $C$  and call the algorithm recursively with this reduced point set. For the base case, when  $|C| = 0$ , as  $P_1$  is the largest cluster, we require to sample only  $O(k\lambda)$  points. This is tackled in Step 2. Step 3 is not performed in this case, as there are no centers.

### 7.3 Analysis and Proof of Correctness

It can be shown using techniques similar to those in [14] that if we have the *Random Sampling Procedure* described above, then we can get a  $(1+\varepsilon)$ -approximation algorithm for the clustering problem with constant probability. Further the running time of the algorithm is  $O(2^{(\frac{k}{\varepsilon})^{O(1)}} dn)$ .



**Theorem 4.** *Suppose a point set  $P$  is  $(k, \alpha)$ -irreducible. Then the algorithm **Irred- $k$ -Clustering** $(P, k, k, \emptyset, \alpha, 0)$  returns a solution to the clustering problem  $\mathcal{C}(f, k)$  on input  $P$  of cost at most  $(1 + \alpha)\text{OPT}_k(P)$  with probability  $\gamma^k$ , where  $\gamma$  is a constant.*

*Proof.* Consider an optimal solution to  $\mathcal{C}(f, k)$  on input  $P$ . Let the centers be  $\mathcal{K} = \{c_1, \dots, c_k\}$  and let these partition the point set  $P$  into clusters  $P_1, \dots, P_k$  respectively. The only source of randomization in our algorithm is the invocations to the superset sampling lemma (Lemma 6). Recall that the desired event in the superset sampling lemma happens with constant probability. For ease of exposition, we shall assume that this desired event in fact always happens when we invoke this lemma. At the end of this proof, we will compute the actual probability with which our algorithm succeeds. Thus, unless otherwise stated, we assume that the desired event in the superset sampling lemma always happens.

Observe that when we call **Irred- $k$ -Clustering** with input  $(P, k, k, \emptyset, \alpha, 0)$ , it gets called recursively again several times (although with different parameters). Let  $\mathcal{W}$  be the set of all calls to **Irred- $k$ -Clustering** when we start it with input  $(P, k, k, \emptyset, \alpha, 0)$ . Let  $\mathcal{W}_i$  be those calls in  $\mathcal{W}$  in which the parameter  $C$  (i.e., the set of centers already found) has size  $i$ .

For all values of  $i$ , our algorithm shall maintain the following invariant :

**Invariant :** The set  $\mathcal{W}_i$  contains a call in which the list of parameters  $(Q, m, k, C, \alpha, \text{Sum})$  has the following properties :

- (1) If the optimal solution,  $\mathcal{K}$ , is  $(k, \alpha)$ -irreducible and  $C'_i = \{c'_1, \dots, c'_i\}$  is a set of  $i$  known centers then there exists a set  $C''_i = \{c_{i+1}, \dots, c_k\}$  of  $k - i$  unknown centers,  $C''_i \subseteq \mathcal{K}$ , such that  $\text{OPT}_k(P, C') \leq (1 + \alpha/k)^i \text{OPT}_k(P)$ , where  $C' = C'_i \cup C''_i$ .
- (2) Let  $P'_1, \dots, P'_k$  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  $C'$ ), where  $C'$  is as defined above. Then the set  $P - Q$  is a subset of  $P'_1 \cup \dots \cup P'_i$ .

Clearly, if we show that the invariant holds for  $i = k$ , then we are done. It holds trivially for  $i = 0$ . Suppose the invariant holds for some fixed  $i$ . We shall show that the invariant holds for  $(i + 1)$  as well.

Since the invariant holds for  $i$ , there exist parameter lists in  $\mathcal{W}_i$  which satisfy the invariant mentioned above. Among such parameter lists, choose a list  $(Q, m, k, C, \alpha, \text{Sum})$  for which  $|Q|$  is smallest.

We assume w.l.o.g. that  $P'_{i+1}$  be the largest cluster amongst  $P'_{i+1} \cup \dots \cup P'_k$ . Since the conditions of the *Tightness Property* are met, therefore there exists a set  $S$  contained in  $P'_1 \cup \dots \cup P'_i$  such that  $P - S$  contains at most  $|P'_{i+1}|/\alpha^{O(1)}$  points of  $P'_1 \cup \dots \cup P'_i$ .

Recall that we are looking at the parameter list  $(Q, m, k, C, \alpha, \text{Sum})$  which satisfies the invariant for  $i$ . Let  $P'$  denote  $P - S$ .

*Claim.*  $P'_{i+1} \cup \dots \cup P'_k$  is contained in  $P' \cap Q$ .

*Proof.* We already know that  $S$  is contained in  $P'_1 \cup \dots \cup P'_i$ . Therefore,  $P'_{i+1} \cup \dots \cup P'_k$  is contained in  $P'$ .

Suppose that the claim is not true. Then there exists a point  $x \in P'_{i+1} \cup \dots \cup P'_k$  such that  $x \in P - Q$ . This means that  $x$  was eliminated in a previous iteration. Let that iteration be  $j < i$ . Then there exist centers  $c'_1, \dots, c'_j \in C'_i$ , such that in iteration  $j$ ,  $d(x, \{c'_1, \dots, c'_j\}) \leq d(x, \{c_{i+1}, \dots, c_k\})$ . This implies that  $d(x, C'_i) \leq d(x, C''_i)$ . This contradicts that  $x \in P'_{i+1} \cup \dots \cup P'_k$  by the *closeness property* of the clustering problem.

*Claim.*  $|P'_{i+1}| \geq \frac{\alpha^{O(1)}}{k} |P'|$ .

*Proof.* There are at most  $|P'_{i+1}|/\alpha^{O(1)}$  elements of  $P'_1 \cup \dots \cup P'_i$  in  $P'$ . Therefore, since  $P'_{i+1}, \dots, P'_k$  are the clusters associated with the centers in  $C''_i$  and  $P'_{i+1}$  is the largest of these clusters, we have  $|P'| \leq |P'_{i+1}|/\alpha^{O(1)} + |P'_{i+1}| + \dots + |P'_k| \leq |P'_{i+1}|/\alpha^{O(1)} + k|P'_{i+1}| \leq \frac{k}{\alpha^{O(1)}} |P'_{i+1}|$ .

It follows that  $|P'_{i+1}| \geq \frac{\alpha^{O(1)}}{k} |P' \cap Q|$ . So, if we knew  $P'$ , then using Lemma 6, we can get a point  $c'_{i+1}$  which is a  $(1 + \alpha/k)$  approximation to  $c_{i+1}$  (as the 1-center of the cluster  $P'_{i+1}$ ) by sampling  $O((k/\alpha)^{O(1)})$  points from  $P' \cap Q$ , and generating the candidate center set of size  $O(2^{(k/\alpha)^{O(1)}})$  as described by the *Random Sampling Procedure*. But of course we do not know  $P'$ .

**Lemma 7.**  $|P' \cap Q| \geq |Q|/2$ .

*Proof.* Suppose not, i.e.,  $|P' \cap Q| \leq |Q|/2$ .

*Claim.* Consider the points in  $Q$  sorted in ascending order of the distance from  $C$ . Let  $U$  be the first  $|Q|/2$  points in this order. Then  $U$  does not contain a point of  $P' \cap Q$ .

*Proof.* Follows from condition (b) of the *Tightness Property* for the clustering problem.

So, if  $U$  is as defined in the claim above, then  $P' \cap Q$  is a subset of  $Q - U$ . Since  $P'_{i+1} \cup \dots \cup P'_k$  is contained in  $P' \cap Q$  (because of Claim 7.3 and the fact that  $Q$  is in the parameter list which satisfies the invariant for  $i$ ), it follows that  $P'_{i+1} \cup \dots \cup P'_k$  is a subset of  $Q - U$ . Thus, the parameter list  $(Q - U, C, k, m, \alpha, \text{Sum})$  which is formed in Step(e) of the algorithm satisfies the invariant for  $i$  as well, i.e., it is in  $\mathcal{C}_i$ . But this violates the fact that  $(Q, C, k, m, \alpha, \text{Sum})$  was the parameter list satisfying the invariant for  $i$  in  $\mathcal{C}_i$  for which  $|Q|$  is smallest. This proves the lemma.

The lemma above implies that  $|P' \cap Q| \geq |Q|/2$ . Combined with Claim 7.3, we get  $|P'_{i+1}| \geq \frac{\alpha^{O(1)}|Q|}{4k}$ . The superset sampling lemma combined with the claim above imply that by sampling  $O((k/\alpha)^{O(1)})$  points from  $Q$  and generating the candidate center set as described by the *Random Sampling Procedure* for the clustering problem, we shall get a point  $c'_{i+1}$  such that  $f(P'_{i+1}, c'_{i+1}) \leq (1 +$

$\alpha/k)f(P'_{i+1}, c_{i+1})$ , where  $c_{i+1} \in C''_i$  is the center of  $P'_{i+1}$  in the optimal clustering induced by  $C'$ . This is the case handled by the step 2(b) in the algorithm **Irred- $k$ -Clustering**. In this case the algorithm is called again with parameters  $(Q, m-1, k, C \cup \{c'_{i+1}\}, \alpha, \text{Sum})$ . It is easy to see now that this parameter list satisfies the invariant for  $i+1$ . The set of known centers  $C'_{i+1}$  for the next iteration is  $C'_i \cup \{c'_{i+1}\}$  and the set of unknown centers  $C''_{i+1}$  is  $C''_i \setminus \{c_{i+1}\}$ . Since  $f(P'_{i+1}, c'_{i+1}) \leq (1+\alpha/k)f(P'_{i+1}, c_{i+1})$  and the clustering problem satisfies the closeness property, it follows that  $\mathcal{F}(P, C'_{i+1} \cup C''_{i+1}) \leq (1+\alpha/k)\mathcal{F}(P, C') \leq (1+\alpha/k)^{i+1}\text{OPT}_k(P)$ . Thus we have shown that the invariant holds for all values of  $i$ .

As we mentioned earlier, a parameter list  $(Q, m, k, C, \alpha, \text{Sum})$  which satisfies the invariant for  $i=k$  has the desired centers in  $C$ .

It is easy to verify that the cost reported by the algorithm  $\mathcal{F}$  satisfies

$$\text{OPT}_k(P) \leq \mathcal{F} \leq (1+\alpha/k)^k \text{OPT}_k(P) \leq (1+2\alpha)\text{OPT}_k(P) \leq (1+\varepsilon/4)\text{OPT}_k(P).$$

This proves the correctness of our algorithm. We just need to calculate the probability with which the algorithm is called with such a parameter list.

Note that the only source of randomness in **Irred- $k$ -Clustering** is in the Step 2(a). The sampling gives the desired result with constant probability (according to Lemma 6). Further each time we execute Step 2, we decrease  $m$  by 1. So, in any sequence of successive recursive calls, there can be at most  $k$  invocations of Step 2. Now, we have just shown that there is a parameter list in  $\mathcal{W}_k$  for which  $C$  contains a set of centers close to the optimal clusters. Let us look at the sequence of recursive calls which have resulted in this parameter list. In these sequence of calls, as we mentioned above, there are  $k$  invocations of the random sampling. Each of these work correctly with constant probability. Therefore, the probability that we actually see this parameter list during the execution of this algorithm is  $\gamma^k$  for some constant  $\gamma$ .

Now we establish the running time of our algorithm.

**Theorem 5.** *The algorithm **Irred- $k$ -Clustering** when called with parameters  $(P, k, k, \emptyset, \alpha, 0)$  runs in time  $O(2^{(k/\alpha)^{O(1)}} dn)$ , where  $n = |P|$ .*

*Proof.* Let  $T(n, m)$  be the running time of our algorithm on input  $(Q, m, k, C, \alpha, \text{Sum})$  where  $n = |Q|$ . Then in Step 2(b), we have  $u(k, \alpha)$  subsets of the sample, where  $u(k, \alpha) = O(2^{(k/\alpha)^{O(1)}})$ . Computation of the candidate center set from any set  $S'$  in Step 2(b) takes  $O(2^{(k/\alpha)^{O(1)}} \cdot nd)$  time. Steps 3(a)-(d) take  $O(nd)$  time. Therefore we get the recurrence

$$T(n, m) = O(u(k, \alpha))T(n, m-1) + T(n/2, m) + O(u(k, \alpha) \cdot nd).$$

It is not difficult to show from this that  $T(n, k)$  is  $O(2^{(k/\alpha)^{O(1)}} dn)$ .

We can now state our main theorem.

**Theorem 6.** *A  $(1+\varepsilon)$ -approximate solution to a clustering problem satisfying the conditions specified in the previous section for a point set  $P$  in  $\mathbb{R}^d$  can be found in time  $O(2^{(k/\varepsilon)^{O(1)}} dn)$ , with constant probability.*

*Proof.* We can run the algorithm **Irred- $k$ -Clustering**  $c^k$  times for some constant  $c$  to ensure that it yields the desired result with constant probability. This still keeps the running time  $O(2^{(k/\alpha)^{O(1)}} dn)$ . So let us assume this algorithm gives the desired solution with constant probability.

Notice that the running time of our main algorithm in Figure 2 is also  $O(2^{(k/\alpha)^{O(1)}} dn)$ . We just have to show that it is correct.

Let  $i$  be the highest index for which  $P$  is  $(i, \alpha)$ -irreducible. So, it follows that

$$\text{OPT}_i(P) \leq (1+150k\alpha)\text{OPT}_{i+1}(P) \leq \dots \leq (1+150k\alpha)^{k-i}\text{OPT}_k(P) \leq (1+\varepsilon/4)\text{OPT}_k(P).$$

Further, we know that the algorithm **Irred- $k$ -Clustering** on input  $(P, i, i, \emptyset, \alpha, 0)$  yields a set of  $i$  centers  $C$  for which  $\mathcal{F}(P, C) \leq (1 + \varepsilon/4)\text{OPT}_i(P)$ . Therefore, we get a solution of cost at most  $(1 + \varepsilon/4)(1 + \varepsilon/4)\text{OPT}_k(P) \leq (1 + \varepsilon)\text{OPT}_k(P)$ . This proves the theorem.