Clustering with Faulty Centers

Emily Fox
University of Texas at Dallas, USA

Hongyao Huang
University of Texas at Dallas, USA

Benjamin Raichel
University of Texas at Dallas, USA

Abstract

In this paper we introduce and formally study the problem of k-clustering with faulty centers. Specifically, we study the faulty versions of k-center, k-median, and k-means clustering, where centers have some probability of not existing, as opposed to prior work where clients had some probability of not existing. For all three problems we provide fixed parameter tractable algorithms, in the parameters k, d, and ε, that $(1+\varepsilon)$-approximate the minimum expected cost solutions for points in d dimensional Euclidean space. For Faulty k-center we additionally provide a 5-approximation for general metrics. Significantly, all of our algorithms have only a linear dependence on n.

2012 ACM Subject Classification Theory of computation → Facility location and clustering; Theory of computation → Computational geometry

Keywords and phrases clustering, approximation, probabilistic input, uncertain input

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2022.40

Funding *Emily Fox*: Research partially supported by NSF grant CCF-1942597.
Hongyao Huang: Research partially supported by NSF grant CCF-1750780.
Benjamin Raichel: Research partially supported by NSF grant CCF-1750780.

1 Introduction

There is a vast body of computational geometry literature which considers input points that are certain, that is they always exist and their location is known. However, uncertainty naturally arises when we are dealing with real world inputs. To model uncertain inputs, several works have considered the notion of probabilistic points. Two models for probabilistic points are commonly used: (i) the *existential* model [18, 20, 21], and (ii) the *locational* model [7, 11]. In the *existential* model, each probabilistic point has a certain fixed location if it exists, but it has a given probability of not existing. In the *locational* model, each probabilistic point always exists but its location is uncertain, and is instead specified by a probability density function over some region.

In this paper, we consider variants of the k-clustering problem under the *existential* model for the cluster centers. Specifically, we consider the k-center, k-median, and k-means problems, where the input points that must be covered are certain to exist, but each one of the k selected centers has an independent probability of existing, i.e. of being open to cover points. Our goal is then to select centers so as to minimize the expected furthest distance, sum of distances, or sum of squared distances, that points must travel to their nearest open center. We denote this as the *Faulty k-Clustering* problem. Prior papers have considered k-clustering in probabilistic input models, but where the centers are certain and the points needing to be covered are probabilistic (see for example [7]). To the best of our knowledge we are the first to consider probabilistic k-clustering, where the uncertainty is on the cluster centers. Our variant of k-clustering is quite natural, as real world facilities can often have some probability of failure. Indeed, this real world motivation of faulty centers

has inspired other previous work, though not in our probabilistic setting. Specifically, in the Fault Tolerant Clustering Problem (see for example [23]), the centers are certain to exist, though each point must travel to its l-th closest center. This objective attempts to provide robustness (i.e. failure tolerance) in the chosen centers. However, it less faithfully models the case where individual centers fail with some probability, since for example while one point's closest center may be closed, a different point's closest center may be open.

Related Work

The k-clustering problem with certain (i.e. non-probabilistic) input points is a classic and fundamental topic in computational geometry. The three most common variants are k-center, k-median, and k-means clustering. All three problems are known to be NP-hard. k-center is NP-hard to approximate within any factor less than 2 in general metric spaces [17] and hard to approximate within a factor of roughly 1.82 in the plane [8]. k-means is known to be NP-hard even when $k = 2$ [2], while k-median is known to be hard to approximate within a factor of $(1 + 2/e)$ [19]. Despite the hardness of these problems, there are many well known approximation algorithms. The standard greedy algorithm for k-center by Gonzalez [10] achieves an optimal 2-approximation to the optimal radius r_{opt}. By an alternative method, Hochbaum and Shmoys [16] also achieved a 2-approximation for k-center. For k-median and k-means it is known that local search achieves a constant factor approximation in polynomial time. (See discussion in [12] and references therein.) In Euclidean space, PTAS's exists for these problems when k, d, and ε are bounded. Specifically, Agarwal and Procopiuc [1] achieve a $(1 + \varepsilon)$-approximation for k-center in $O(n \log k) + (k/\varepsilon)^{O(k^{1-1/d})}$ time. For k-median and k-means a number of corset based $(1 + \varepsilon)$-approximation algorithms have been given which run in linear time in n, including Har-Peled and Mazumdar [14], and subsequent papers improving the time dependency on k, d, and ε [5, 9].

A number of prior works have considered variants of k-clustering where the client points that need to be covered are probabilistic, as opposed to our model where the centers are probabilistic. Perhaps most notably is the work of Cormode and McGregor [7]. They consider client points under the locational model, though as they also allow clients to have non-zero probability to not exist, their model also captures the existential model. For k-median and k-means they achieve $(1 + \varepsilon)$-approximations to the minimum expected cost solution in Euclidean space, and a constant factor approximation for k-median in general metrics. Their main focus, however, is on the more challenging case of k-center, for which they provide bi-criteria approximations for general metrics. That is, in addition to approximating the radius, they are allowed to exceed the requested number centers, and they provide different tradeoffs between the two kinds of approximation. Guha and Munagala [11] subsequently provided a non-bi-criteria approximation for k-center, obtaining an $O(1)$-approximation on only the expected radius. For points in $\mathbb{R}^d$, Huang and Li [18] later achieved the first PTAS for k-center, when k, d are fixed constants.

There have been a number of other follow up results to [7], again for the case of probabilistic clients not centers. For k-center on the real line, $\mathbb{R}^1$, Wang and Zhang [27] showed that the problem can be solved exactly in polynomial time. Munteanu *et al.* [25] considered the special case where $k = 1$ for both the k-center and the k-median objectives, and achieved polynomial time $(1 + \varepsilon)$-approximations. Moreover, in the data mining community, previous works have considered variations of probabilistic k-median [24] and k-means clustering [3, 26].

The idea of modeling faulty centers in clustering problems has also been considered in previous works under the setting of fault tolerant k-clustering [4, 22, 23]. In fault tolerant clustering centers are certain rather than probabilistic, and one assigns each point to its l

nearest center for some integer $l \geq 1$. This is unlike our probabilistic model where each point is assigned to its nearest center that happens to be open.

1.1 Preliminaries

In the standard k-clustering problem, we are given a set P of n points in a metric space, called *clients*, and an integer parameter $k > 0$. The task is to select k points from P, called *centers*, so as to minimize some cost function of the distances of the client points in P to their nearest centers. In *k-center clustering* one minimizes the maximum distance from a client to its nearest center. In *k-median clustering* one minimizes the sum of distances from each client to their nearest center. Finally, in *k-means clustering*, one minimizes the sum of squared distances from each client to their nearest center. Note that for a given set of centers, the distances from the clients to their nearest centers, defines a vector of length n. The goal of k-center, k-median, or k-means, is then to minimize the ℓ_∞, ℓ_1, or ℓ_2 norm of this vector, respectively.

For these three problems we now formally define their respective cost functions for any given subset $C \subseteq P$. Specifically, for k-center, k-median, and k-means we respectively define the functions f, g, and h, as follows.

$$f_P(C) = \max_{p \in P} ||p - C|| = \max_{p \in P} \min_{c \in C} ||p - c||, \qquad g_P(C) = \sum_{p \in P} ||p - C|| = \sum_{p \in P} \min_{c \in C} ||p - c||,$$

$$h_P(C) = \sum_{p \in P} ||p - C||^2 = \sum_{p \in P} \min_{c \in C} ||p - c||^2.$$

The goal of k-center, k-median, or k-means is then to select the subset $C \subseteq P$ of size k that minimizes f_P, g_P, or h_P respectively. Note that for k-center clustering specifically, we often refer to $r = f_P(C)$ as the radius of the solution C, as r can be viewed as the radius of k equal radius balls covering the points in P.

Here we consider a variation of the standard k-clustering problem where there is uncertainty on whether any one of the chosen centers will be open, i.e. uncertainty on whether points in P can be covered by that center. Specifically, in addition to the point set P, as part of the input we are also given a vector V whose ith entry v_i is the probability that p_i will be open if it is chosen as a center. For any point p_i, when convenient we use $prob(p_i) = v_i$ to denote its associated probability.

▶ **Definition 1.** *Let P be a set of n points in a metric space, $V \in [0,1]^n$ be a corresponding vector of probabilities, and $C \subseteq P$ be any subset. Any subset $R \subseteq C$ is called a* realization *of C, and let $Real(C)$ denote the set of all realizations (i.e. the power set of C). For a given realization R, a center $p \in C$ is said to be* open *(resp.* closed*) is $p \in R$ (resp. $p \notin R$). Each center $p \in C$ is open independently with probability $prob(p)$, thus the probability $R \in Real(C)$ occurs (i.e. is the set of open centers) is $Prob(R) = (\Pi_{p \in R}\ prob(p))(\Pi_{p \in C \setminus R}\ (1 - prob(p)))$.*

For this probabilistic version of k-clustering, our cost functions $f_{P,V}(C)$, $g_{P,V}(C)$, and $h_{P,V}(C)$ are now random variables, which for a given realization $R \subseteq C$ are equal to $f_P(R)$, $g_P(R)$, and $h_P(R)$, respectively. (Note that throughout we use the single subscript f_P to denote the non-probabilistic cost function, and the double subscript $f_{P,V}$ to denote the corresponding random variable version.) Our goal is now to find the subset C minimizing the expected value $E[f_{P,V}(C)]$, $E[g_{P,V}(C)]$, or $E[h_{P,V}(C)]$, where the expectation is taken over the distribution of $Real(C)$ determined by V.

▶ **Problem 2** (Faulty k-Center Clustering). *As input you are given a set P of n points in a metric space, a corresponding vector $V \in [0,1]^n$ of independent probabilities, and a positive integer parameter k. Find the subset C_{opt} of k centers which minimizes $E[f_{P,V}(C)]$. That is, $C_{opt} = \arg\min_{C \subseteq P, |C|=k} E[f_{P,V}(C)]$.*

▶ **Problem 3** (Faulty k-Median Clustering). *As input you are given a set P of n points in a metric space, a corresponding vector $V \in [0,1]^n$ of independent probabilities, and a positive integer parameter k. Find the subset C_{opt} of k centers which minimizes $E[g_{P,V}(C)]$. That is, $C_{opt} = \arg\min_{C \subseteq P, |C|=k} E[g_{P,V}(C)]$.*

▶ **Problem 4** (Faulty k-Means Clustering). *As input you are given a set P of n points in a metric space, a corresponding vector $V \in [0,1]^n$ of independent probabilities, and a positive integer parameter k. Find the subset C_{opt} of k centers which minimizes $E[h_{P,V}(C)]$. That is, $C_{opt} = \arg\min_{C \subseteq P, |C|=k} E[h_{P,V}(C)]$.*

▶ Remark 5. It is possible that all selected centers are closed, i.e. $R = \emptyset$. Thus to make sure the problem is well defined, we set $f_P(\emptyset)$, $g_P(\emptyset)$, and $h_P(\emptyset)$ equal to different specified values. Natural choices for these would depend on the input set P. For example, setting them equal to the optimal (non-probabilistic) 1-center, 1-median, or 1-mean solution, respectively. An alternative, though related approach, is to fix some center which is open with probability 1 and always included in the solution (i.e. not a part of the k selected centers).

Note that if all entries in V are the same then the probability that $R = \emptyset$ is the same, regardless of which subset C of size k is chosen, and thus the choice of $f_P(\emptyset)$ (resp. $g_P(\emptyset)$ and $h_P(\emptyset)$) does not affect the relative ordering of $E[f_{P,V}(C)]$ for different C. Furthermore, even in the case when the entries in V differ, for the solution our algorithm returns (as described below) the probability that $R = \emptyset$ is less than or equal to that for the optimal solution.

1.2 Our Contribution

To the best of our knowledge, we are the first to formally study the faulty k-clustering problem, where the probabilities are on the centers rather than the clients. As stated above, this is a natural setting, as centers may have some probability of failure.

For the three most common k-clustering variants, k-center, k-median, and k-means, we provide fixed parameter tractable approximation algorithms for their faulty versions. Specifically, for Faulty k-Center we provide an $O(8^k kn)$ time 5-approximation in general metrics. All remaining results are $(1+\varepsilon)$-approximations for the Euclidean metric $\mathbb{R}^d$, where we always assume d is a constant. For Faulty k-Center we provide an $O(n \log k) + 2^{O(k)}/\varepsilon^{d(k+1)}$ time $(1+\varepsilon)$-approximation in the Euclidean case. For Faulty k-Median we provide an $O(kn) + (2^{O(k \log k)}/\varepsilon^{d(k+1)})n^{o(1)}$ time $(1+\varepsilon)$-approximation in the Euclidean case. Finally, for Faulty k-Means we provide an $O(kn) + (2^{O(k \log k)}/\varepsilon^{(2d+1)k})n^{o(1)}$ time $(1+\varepsilon)$-approximation in the Euclidean case.

It is important to note that all of our algorithms have only a linear dependence on n. Moreover, for all three problems, in the Euclidean case we are providing a $(1+\varepsilon)$-approximation, that is an EPTAS for fixed k and d.

Recall that the standard (non-faulty) versions of these problems are NP-hard, even in Euclidean settings, with additional results on hardness of approximation or for special cases, depending on which one of the three problems is considered. As the non-faulty versions are a special case of the faulty versions (where probabilities are all 1), these hardness results immediately apply to our problems. Moreover, our problems have the additional challenge that each choice of centers has an exponential number of possible realizations.

2 FPT Approximation Algorithms for k-Center

In this section we develop fixed parameter tractable (in the parameter k) approximation algorithms for Faulty k-Center Clustering. Specifically, for general metric spaces we achieve a 5-approximation to the optimal radius and in fixed dimensional Euclidean space we achieve a $(1+\varepsilon)$-approximation, i.e. a PTAS. For comparison, recall that for the standard non-probabilistic version of k-center clustering it is hard to approximate the radius within any constant factor less than 2 in general metric spaces [17], and there is a PTAS in Euclidean space. First, we present definitions and a core lemma, which are common to both algorithms.

Consider any instance P, k of Faulty k-Center, and let $\rho = \{\rho_1, \ldots, \rho_m\}$ be a partition of P into m disjoint subsets. Define the diameter of ρ as,

$$diam(\rho) = \max_i diam(\rho_i) = \max_i \max_{p,q \in \rho_i} ||p - q||.$$

For any subset $Z \subseteq P$, let $\rho(Z) = \{\rho_1(Z), \ldots, \rho_m(Z)\}$, where $\rho_i(Z) = Z \cap \rho_i$, denote the partition of Z induced by ρ. We then define the *characteristic vector* of Z with respect to ρ, denoted $char(Z, \rho)$ as the m dimensional integer vector whose ith entry is $|\rho_i(Z)|$. Define the *canonical subset*, $Canon(char(Z, \rho))$, of a characteristic vector $char(Z, \rho) = (w_1, \ldots, w_m)$, as the subset $S \subseteq P$ consisting of the w_i points with highest probability from ρ_i, for all i.[1]

The following key lemma intuitively states that for any set of centers Q, if we replace each subset $\rho_i(Q)$ with an equal number of higher probability points from ρ_i, then the expected cost of the solution increases by at most the diameter of ρ. Roughly speaking, this holds as each center is moved distance at most ρ, and its probability of being open does not decrease.

▶ **Lemma 6.** *Let $\rho = \{\rho_1, \ldots, \rho_m\}$ be a partition of P into m subsets. Let $Q \subseteq P$ be any subset, and let $S = Canon(char(Q, \rho))$. Then we have,*

$$E[f_{P,V}(S)] \leq diam(\rho) + E[f_{P,V}(Q)].$$

Proof. Observe that for all i, $|\rho_i(S)| = |\rho_i(Q)|$, since S and Q have the same characteristic vector. For any i, label the points in $\rho_i(S) = \{s^i_1, \ldots, s^i_{w_i}\}$ and similarly in $\rho_i(Q) = \{q^i_1, \ldots, q^i_{w_i}\}$ in decreasing order of their probability. We define a bijection $b : S \to Q$, such that $b(s^i_j) = q^i_j$. Observe that the bijection b defines a bijection between $Real(S)$ and $Real(Q)$. Abusing notation slightly, for any realization R_S of S we use $b(R_S)$ to denote the corresponding realization in Q. Observe that by construction, $char(R_S, \rho) = char(b(R_S), \rho)$.

Let $R_S \in Real(S)$ be any realization of S. Consider any point $p \in P$. Let s be the closest point in R_S to p, and let q be the closest point in $b(R_S)$ to p. Let i be the index such that $q \in \rho_i$. Since $char(R_S, \rho) = char(b(R_S), \rho)$, there must be some point $s' \in R_S$ such that $s' \in \rho_i(S) \subseteq \rho_i$. Therefore, by the triangle inequality,

$$||p - s|| \leq ||p - s'|| \leq ||p - q|| + ||q - s'|| \leq ||p - q|| + diam(\rho).$$

This implies $f_P(R_S) = \max_{p \in P} ||p - R_S|| \leq \max_{p \in P} ||p - b(R_S)|| + diam(\rho) = f_P(b(R_S)) + diam(\rho)$.

For two vectors u, v of the same dimension, let $u \leq v$ denote that u is coordinate-wise smaller than v, i.e. $u_i \leq v_i$ for all i. So let V' be a probability vector such that $V' \leq V$. Then observe that $E[f_{P,V}(X)] \leq E[f_{P,V'}(X)]$ for any subset $X \subseteq P$. In other words, if each

[1] For points of equal probability, let there be an arbitrary but fixed ordering.

center in X has a smaller or equal probability to be open under V', then the expected cost cannot decrease when replacing V with V'.

For any point $p \in P$, let $prob_V(p)$ denote the corresponding probability from the vector V. We define a new probability vector V' such that for any $p \in P$, if $p \in S$ then $prob_{V'}(p) = prob_V(b(p))$, and if $p \notin S$ then $prob_{V'}(p) = prob_V(p)$. Observe that since $\rho_i(S)$ consists of the w_i points with highest probability in ρ_i, we have that $prob_V(s_j^i) \geq prob_V(q_j^i)$ for any i, j. Therefore $V' \leq V$, and so by the above discussion, $E[f_{P,V}(S)] \leq E[f_{P,V'}(S)]$.

Let $prob_V(R_S)$ and $prob_{V'}(R_S)$ denote the probability that R_S is realized under V and V', respectively. Observe that $prob_{V'}(R_S) = prob_V(b(R_S))$. Therefore, we have the following,

$$\begin{aligned} E[f_{P,V}(S)] \leq E[f_{P,V'}(S)] &= \sum_{R_S \in Real(S)} prob_{V'}(R_S) \cdot f_P(R_S) \\ &= \sum_{b(R_S) \in Real(Q)} prob_V(b(R_S)) \cdot f_P(R_S) \\ &\leq \sum_{b(R_S) \in Real(Q)} prob_V(b(R_S)) \cdot (f_P(b(R_S)) + diam(\rho)) \\ &= E[f_{P,V}(Q)] + diam(\rho) \end{aligned}$$ ◀

We remark that given an optimal subset of centers C_{opt} and the canonical subset S of any arbitrary characteristic vector, it is not necessarily the case that $E[f_{P,V}(Canon(char(C_{opt}, \rho)))] \leq E[f_{P,V}(S)]$. However, the fact that canonical subsets have approximately the same cost as the best subset of the same characteristic vector is sufficient for proving our algorithms' correctness.

2.1 General Metrics

Here we develop a fixed parameter tractable algorithm for Faulty k-Center Clustering in general metric spaces, which achieves a 5-approximation to the optimal radius.

Consider a subset $C = \{c_1, \ldots, c_k\}$ of k centers from P. For any other subset $Z \subseteq P$, let $Vor_i(Z, C)$ denote the subset of points in Z whose nearest center in C is c_i, e.g. when P is a point set in Euclidean space then this is the subset of points of Z in the Voronoi cell of c_i. Observe that $Vor(Z, C) = \{Vor_1(Z, C), \ldots, Vor_k(Z, C)\}$ defines a partition $\rho(Z) = \{\rho_1(Z), \ldots, \rho_k(Z)\}$ of Z where $\rho_i(Z) = Vor_i(Z, C)$. Thus the definitions of characteristic vectors and canonical subsets from above apply, where to simplify notation we write $char(Z, C) = char(Z, Vor(P, C))$. Moreover, by the triangle inequality

$$\begin{aligned} diam(Vor(P, C)) &= \max_i \max_{p,q \in Vor_i(P,C)} ||p - q|| \leq 2 \max_i \max_{p \in Vor_i(P,C)} ||p - c_i|| \\ &= 2 \max_{p \in P} ||p - C|| = 2 f_P(C), \end{aligned}$$

and thus we immediately have the following corollary of Lemma 6.

▶ **Corollary 7.** *Let $C = \{c_1, \ldots, c_k\}$ be a subset of k centers from P. Let $Q \subseteq P$ be any subset, and let $S = Canon(char(Q, C))$. Then we have,*

$$E[f_{P,V}(S)] \leq 2 f_P(C) + E[f_{P,V}(Q)].$$

Observe that for two different subsets $Z, Z' \subseteq P$ such that $|Z| = |Z'|$ it is possible to have $char(Z, C) = char(Z', C)$, however, we have the following bound on the total number of characteristic vectors for subsets of size k.

▶ **Observation 8.** *Let $C \subseteq P$ be a subset of k points. Then there are $O(4^k)$ possible characteristic vectors for all subsets of size k with respect to C. That is,*

$$\left| \bigcup_{C' \subseteq P, |C'|=k} char(C', C) \right| = O(4^k).$$

Proof. Recall $char(C', C)$ is a vector of length k whose (non-negative) entries sum to k. Any such vector can be represented as a binary vector of length $2k-1$ by writing each entry from the original vector in unary, and then separating entries with a single zero. The number of binary vectors of length $2k-1$ is $2^{2k-1} = O(4^k)$.[2] ◀

Before we present our algorithm, we make one more simple observation.

▶ **Observation 9.** *Given a point set P, probability vector V, and a subset C of k centers, $E[f_{P,V}(C)]$ can be computed in $O(2^k kn)$ time. Specifically, enumerate all possible $O(2^k)$ realizations in $Real(C)$. Then for a given realization R, the probability of R occurring is $(\Pi_{p_i \in R}\ v_i)(\Pi_{p_i \in C \setminus R}\ (1-v_i))$, and thus is computable in $O(k)$ time. Moreover, $f_P(R)$ can be computed in $O(kn)$ time, by checking for each point of P what is the closest point in R.*

▶ **Theorem 10.** *Let C_{opt} denote an optimal solution to Faulty k-Center. Then in $O(8^k kn)$ time one can compute a set $C \subseteq P$ of k centers such that $E[f_{P,V}(C)] \leq 5E[f_{P,V}(C_{opt})]$.*

Proof. The first step of our algorithm is to compute a set of k centers $D \subseteq P$, which is a 2-approximation to the optimal solution to the standard k-center instance on P, that is $f_P(D) \leq 2 \min_{C' \subseteq P, |C'|=k} f_P(C')$. Note that D can be computed in $O(kn)$ time using the standard k-center algorithm of Gonzalez [10]. Next we guess the characteristic vector of C_{opt} with respect to D, $char(C_{opt}, D)$. This is done by enumerating all binary vectors of length $2k-1$ which have k 1's, as discussed in Observation 8. Let $W = (w_1, \ldots, w_k)$ denote the current guess for $char(C_{opt}, D)$. We construct the canonical subset $Canon(W)$, by taking the w_i points with highest probability from $Vor_i(P, D)$ for all i. Next we compute the expected cost of this subset $E[f_{P,V}(Canon(W))]$. After computing this value for all possible guesses of W, we then return as our solution $C = Canon(W)$ with minimum expected cost. (Note if W is not realizable, i.e. if there are fewer than w_i points in $Vor_i(P, D)$, then we simply record $E[f_{P,V}(Canon(W))] = \infty$.)

For the running time, computing D takes $O(kn)$ time. Next, for each guess $W = (w_1, \ldots, w_k)$ of $char(C_{opt}, D)$, computing $Canon(W)$ takes $O(kn)$ time, since finding the w_i points with highest probability from $Vor_i(P, D)$ can easily be done in $O(w_i n)$ time, and hence $O(kn)$ time over all i since $\sum w_i = k$. (Note this step can be performed faster by preprocessing the points, though ultimately it does not affect the asymptotic running time.) Next we must compute $E[f_{P,V}(Canon(W))]$, which can be done in $O(2^k kn)$ time by Observation 9. Thus since there are $O(4^k)$ possible guesses by Observation 8, the total time is $O(4^k(2^k kn + kn) + kn) = O(8^k kn)$

As for correctness, first observe that

$$f_P(D) \leq 2 \min_{C' \subseteq P, |C'|=k} f_P(C') \leq 2 \min_{C' \subseteq P, |C'|=k} E[f_{P,V}(C')] = 2E[f_{P,V}(C_{opt})].$$

[2] Further using the fact that there are exactly k entries which are 1 in this vector of length $2k-1$, gives the more precise bound on the number of such vectors, $\binom{2k-1}{k} = O(4^k/\sqrt{k})$.

Next, for $C = Canon(char(C_{opt}, D))$, by Corollary 7 we have $E[f_{P,V}(C)] \leq 2f_P(D) + E[f_{P,V}(C_{opt})]$. Combining these inequalities thus gives,

$$E[f_{P,V}(C)] \leq 2f_P(D)+E[f_{P,V}(C_{opt})] \leq 4E[f_{P,V}(C_{opt})]+E[f_{P,V}(C_{opt})] = 5E[f_{P,V}(C_{opt})]. \blacktriangleleft$$

2.2 Euclidean PTAS

In this section we provide a fixed parameter tractable $(1+\varepsilon)$-approximation to the optimal radius for instances of Faulty k-Center Clustering where $P \subseteq \mathbb{R}^d$, and where for simplicity we assume d is a constant. To achieve this we consider the axis aligned regular grid of cell side length Δ, which is a value to be determined shortly. Specifically, for any point $p \in P$, where $p = (p^1, \ldots, p^d)$, its cell is given by $cell_\Delta(p) = (\lfloor p^1/\Delta \rfloor, \ldots, \lfloor p^d/\Delta \rfloor)$. Assuming this limited use of the floor function takes $O(1)$ time, in $O(n)$ time we can compute the cell of every point in P. Moreover, as these cells are given by integer vectors, using hashing in the same time we can also compute the set of non-empty grid cells and the corresponding points in each cell. Let $Grid_\Delta(P)$ denote this partition of P into the non-empty grid cells. We have the following standard observation (see for example [15]).

▶ **Observation 11.** *Let $B(c,r)$ denote the ball of radius r and center c, for any point c and radius $r > 0$. Consider the regular grid of cell side length Δ. Then the number of grid cells intersecting $B(c,r)$ is at most $(2 + \lceil 2r/\Delta \rceil)^d$.*

The following theorem uses similar observations about grids and k-center as [1]. In particular, as a starting point, we similarly make use of the algorithm of Feder and Greene [8], which achieves a 2-approximation for k-center clustering in $O(n \log k)$ time.

▶ **Theorem 12.** *Let $P \subset \mathbb{R}^d$ be an instance of Faulty k-Center Clustering in d-dimensional Euclidean space, and let C_{opt} denote an optimal solution to this instance. Then for any $\varepsilon > 0$, in $O(n \log k) + 2^{O(k)}/\varepsilon^{d(k+1)}$ time[3] one can compute a set $C \subseteq P$ of k centers such that $E[f_{P,V}(C)] \leq (1+\varepsilon)E[f_{P,V}(C_{opt})]$.*

Proof. First, use the algorithm of [8] to compute a set of k centers C' which covers all of P within radius $r = f_P(C') \leq 2 \min_{Z \subseteq P, |Z|=k} f_P(Z) \leq 2E[f_{P,V}(C_{opt})]$. Now set $\Delta = \varepsilon r/(4\sqrt{d})$, and compute $Grid_\Delta(P)$. Let x denote the number of entries in $Grid_\Delta(P)$, where by Observation 11,

$$x \leq k(2 + \lceil 2r/\Delta \rceil)^d = k(2 + \lceil 8\sqrt{d}/\varepsilon \rceil)^d = O(k(1/\varepsilon)^d).$$

Observe that $Grid_\Delta(P)$ is a partition of P, with diameter $diam(Grid_\Delta(P)) \leq \Delta\sqrt{d} = \varepsilon r/4$. Let $S = Canon(char(C_{opt}, Grid_\Delta(P)))$. By Lemma 6 we have,

$$E[f_{P,V}(S)] \leq \frac{\varepsilon r}{4} + E[f_{P,V}(C_{opt})] \leq \frac{\varepsilon}{2}E[f_{P,V}(C_{opt})] + E[f_{P,V}(C_{opt})] = (1+\varepsilon/2)E[f_{P,V}(C_{opt})].$$

Therefore in order to compute a $(1+\varepsilon/2)$-approximation, it suffices to compute all possible characteristic vectors for $char(C_{opt}, Grid_\Delta(P))$, evaluate the expected cost of their corresponding canonical subsets, and take the minimum. To speed up the time to evaluate the expected cost of a canonical subset, we instead compute additive $\varepsilon r/4 \leq (\varepsilon/2)E[f_{P,V}(C_{opt})]$ approximations to these values, thus in total achieving a relative $(1+\varepsilon)$-approximation.

[3] Technically the $2^{O(k)}/\varepsilon^{d(k+1)}$ term assumes $\varepsilon < 1$. If $\varepsilon \geq 1$ this term becomes $2^{O(k)}$. That said, it is standard practice to write the bound in this simplified manner.

As for the running time, similar to Observation 8 one can argue that the number of possible characteristic vectors is at most

$$\binom{x+k-1}{k} \leq \left(\frac{(x+k)e}{k}\right)^k = 2^{O(k)}/\varepsilon^{dk}.$$

For each characteristic vector we need to compute the corresponding canonical subset. Observe that within a given cell of $Grid_\Delta(P)$, the canonical subset consists of the m highest probability points for some value $m \leq k$. Therefore, as a preprocessing step, we can throw out all but the k highest probability points in each cell, and then sort the remaining points in each cell by their probability. Throwing out all but the k highest probability points takes $O(n)$ time in total by using linear time median selection in each cell. All the sorting can be done in $O(x \cdot k \log k)$ time total. After preprocessing, for a specific characteristic vector, it takes $O(x+k)$ time to compute the canonical subset. Suppose that for this canonical subset, we can compute an additive $\varepsilon r/4$-approximation to its expected cost in $O(2^k kx)$ time. Then the total time is

$$O(n \log k + n + xk \log k) + (x + k + 2^k kx)2^{O(k)}/\varepsilon^{dk} = O(n \log k) + 2^{O(k)}/\varepsilon^{d(k+1)}.$$

So let S be any given canonical subset. What remains is to show an additive $\varepsilon r/4$-approximation to its expected cost can be computed in $O(2^k kx)$ time. Let R be any realization of S. Consider the set of points X in some cell of the grid partition $Grid_\Delta(P)$. Consider any two points $p, q \in X$, and let r_p, r_q be the nearest center in R to p and q respectively. By the triangle inequality we have

$$||p - r_p|| \leq ||p - r_q|| \leq ||p - q|| + ||q - r_q|| \leq \sqrt{d}\Delta + ||q - r_q|| = \varepsilon r/4 + ||q - r_q||.$$

Therefore, $||q - r_q|| \leq \max_{p \in X} ||p - R|| \leq \varepsilon r/4 + ||q - r_q||$, that is the distance from q to its nearest center in R is an additive $\varepsilon r/4$ approximation to the maximum distance of a point in the cell to its nearest center in R. Thus to get an $O(kx)$ time additive $\varepsilon r/4$-approximation to $f_P(R)$, it suffices to pick an arbitrary representative q in each cell, compute its nearest center in R, and take the maximum. As there are $O(2^k)$ possible realizations R of S, the claim now follows in a similar fashion to Observation 9. ◀

3 k-Median and k-Means

In this section, we develop a fixed parameter tractable approximation scheme for Faulty k-Median Clustering of a collection of points $P \subset \mathbb{R}^d$ in d-dimensional Euclidean space. As in our approximation scheme for Faulty k-Center, we begin by finding a constant-factor approximate solution to k-Median without faulty centers and then partition the points of P into disjoint subsets based on their location in appropriately sized regions of space. Unlike the algorithm for Faulty k-Center, however, these subsets may have different diameters depending on how far away each subset is from the nearest member of the non-faulty k-Median solution. At the end of this section, we describe the minor changes necessary for our algorithm to apply to Faulty k-Means instead of k-Median. For simplicity, throughout this section we assume $0 < \varepsilon \leq 1$.

We now turn to our algorithm description. Let C'_{opt} denote an optimal solution for non-faulty k-Median. We compute a set of centers $D = \{d_1, \ldots, d_k\}$ such that $g_P(C'_{opt}) \leq g_P(D) \leq 2g_P(C'_{opt})$. The set D can be computed in $O(n) + k^{O(1)} \log^{O(1)} n$ time [14].[4]

[4] The algorithms in [14] are randomized, though they achieve this time with high probability. There are

As in Section 2.1, we partition P into a collection of k subsets defined by the distance from each point to its nearest member of D. For any subset $Z \subseteq P$ and any $i \in \{1, \ldots, k\}$, let $Vor_i(Z, D)$ denote the subset of points of Z whose nearest center in D is d_i.

We proceed to refine the above partition as follows. Fix any $i \in \{1, \ldots, k\}$. Let $t = \left\lceil \log_{(1+\varepsilon)} 2n \right\rceil = O((1/\varepsilon) \log n)$. For each $j \in \{0, \ldots, t\}$, let $r_j = (g_P(D)/(2n)) \cdot (1+\varepsilon)^j$, and let $B_{i,j} = B(d_i, r_j)$, the ball of radius r_j centered at d_i. We partition the points of $Vor_i(P, D) \cap B_{i,0}$ into $O(1/\varepsilon^d)$ *batches* each of diameter $\varepsilon r_0/4$. For each $j \in \{1, \ldots, t\}$, partition the points of $Vor_i(P, D) \cap (B_{i,j} \setminus B_{i,j-1})$ into $O(1/\varepsilon^{d-1})$ batches each of diameter $\varepsilon r_j/8$. These batches can be computed in time linear in their number and the size of $Vor_i(P, D)$ by partitioning points according to their locations in a sufficiently fine grid. (See the discussion before Observation 11 and Observation 11 itself.)

We perform the above assignment to batches for each $i \in \{1, \ldots, k\}$. Observe that no point $p \in P$ can lie further than $g_P(D)$ from its nearest center in D, implying all points are assigned to exactly one batch. Let $\rho = \{\rho_1, \ldots, \rho_m\}$ denote the partition of P into batches.

▶ **Observation 13.** *For the partition $\rho = \{\rho_1, \ldots, \rho_m\}$, we have $m = O((k/\varepsilon^d) \log n)$.*

Recall the definitions of canonical subsets given in Section 2. As in our algorithm for Faulty k-Center, we will enumerate characteristic vectors for subsets of size k with respect to ρ, taking the best canonical subset for these vectors as our solution. We adapt Lemma 6 for the current setting.

▶ **Lemma 14.** *Let $\rho = \{\rho_1, \ldots, \rho_m\}$ denote the partition of P as described above. Let $Q \subseteq P$ be any subset such that $|Q| \leq k$, and let $S = Canon(char(Q, \rho))$. Then we have,*

$$E[g_{P,V}(S)] \leq (1 + 3\varepsilon/4) E[g_{P,V}(Q)].$$

Proof. Our proof follows the one for Lemma 6 except when comparing costs between two realizations $R_S \in Real(S)$ and $R_Q \in Real(Q)$ with the same characteristic vector. Consider any point $p \in P$. Let s be the closest point in R_S to p, and let q be the closest point in R_Q to p. Let i_q be such that $q \in Vor_{i_q}(P, D)$, and let ℓ be the index such that $q \in \rho_\ell$. Given R_S and R_Q have the same characteristic vector, there must be some point $s' \in R_S$ such that $s' \in \rho_\ell$. We have

$$||p - s|| \leq ||p - s'|| \leq ||p - q|| + ||q - s'||.$$

Suppose $q \in B_{i_q,0}$. Recall, $B_{i_q,0}$ is the ball of radius $r_0 = g_P(D)/(2n)$ centered at d_{i_q}. Then,

$$||p - s|| \leq ||p - q|| + ||q - s'|| \leq ||p - q|| + \frac{\varepsilon r_0}{4} = ||p - q|| + \frac{\varepsilon g_P(D)}{8n}.$$

. Now, suppose $q \in (B_{i_q,j} \setminus B_{i_q,j-1})$ for some $j > 0$. Centers d_{i_p} and d_{i_q} are the closest members of D for p and q respectively, and $||q - D|| \geq r_{j-1}$. By triangle inequality,

$$r_{j-1} \leq ||q - D|| \leq ||q - p|| + ||p - D||.$$

Therefore,

$$||p - s|| \leq ||p - q|| + ||q - s'|| \leq ||p - q|| + \frac{\varepsilon(1+\varepsilon) r_{j-1}}{8} \leq ||p - q|| + \frac{\varepsilon r_{j-1}}{4}$$

deterministic constant factor approximations, though with higher polynomial dependence on n.

$$\leq ||p-q|| + \frac{\varepsilon||p-q||}{4} + \frac{\varepsilon||p-D||}{4}.$$

Finally, summing over all p and observing $g_P(D) \leq 2g_P(R_Q)$.

$$\begin{aligned} g_P(R_S) = \sum_{p\in P} ||p-R_S|| &\leq \sum_{p\in P}\left(||p-R_Q|| + \frac{\varepsilon g_P(D)}{8n} + \frac{\varepsilon||p-R_Q||}{4} + \frac{\varepsilon||p-D||}{4}\right) \\ &\leq g_P(R_Q) + \frac{\varepsilon g_P(R_Q)}{4} + \frac{\varepsilon g_P(D)}{8} + \frac{\varepsilon g_P(D)}{4} \\ &\leq (1+3\varepsilon/4)g_P(R_Q). \end{aligned}$$

The rest of the proof proceeds as in the one for Lemma 6, with "$(1+\varepsilon)E[g_{P,V}(Q) \mid e_Q(u)]$" appearing in place of "$E[f_{P,V}(Q) \mid e_Q(u)] + diam(\rho)$" alongside analogous substitutions. ◀

We conclude with our main theorem for Faulty k-Median.

▶ **Theorem 15.** *Let $P \subset \mathbb{R}^d$ be an instance of Faulty k-Median in d-dimensional Euclidean space, and let C_{opt} denote an optimal solution to this instance. Then for any $\varepsilon > 0$, in $O(kn) + (2^{O(k\log k)}/\varepsilon^{d(k+1)})n^{o(1)}$ time, one can compute a set $C \subseteq P$ of k centers such that $E[g_{P,V}(C)] \leq (1+\varepsilon)E[g_{P,V}(C_{opt})]$.*

Proof. As stated above, our algorithm begins by computing the set $D \subseteq P$ of k centers in $O(n) + k^{O(1)}\log^{O(1)} n$ time [14]. We then find the closest center in D for each point in P, and then refine this partition into batches as described above. This produces a partition ρ into $O((k/\varepsilon^d)\log n)$ batches by Observation 13, and thus ρ is computed in $O(kn + (k/\varepsilon^d)\log n)$ time.

We now pick an approximately best characteristic vector. In order to do so quickly, we first compute a ***(k, ε/8)-coreset*** $\mathcal{S}$ of P. Coreset $\mathcal{S}$ is a weighted set of points such that for any set C of centers, $(1-\varepsilon/8)g_P(C) \leq \sum_{s\in\mathcal{S}} \min_{c\in C} w(s)||s-c|| \leq (1+\varepsilon/8)g_P(C)$ where $w(s)$ denotes the weight of point s. A coreset of size $O((k\log n)/\varepsilon^d)$ can be computed in $O(n\log k)$ time [14].[5] Next, we enumerate all $O((k/\varepsilon^d)\log n)^k)$ possible characteristic vectors W for $char(C_{opt}, \rho)$. We estimate the expected cost of each canonical subset $Canon(W)$ within a $(1+\varepsilon/8)$ factor in $O((2^k k^2 \log n)/\varepsilon^d$ time by using computing the cost of individual realizations with regard to $\mathcal{S}$. Finally, we return the solution C with the minimum estimated expected cost. Lemma 14 and the definition of $\mathcal{S}$ implies our solution has the correct expected cost.

It has been observed that $\log^k n = 2^{O(k\log k)}n^{o(1)}$ [6]: If $n \leq 2^{k^2}$, then $\log^k n \leq k^{2k} = 2^{O(k\log k)}$, and if $n > 2^{k^2}$, then $\log^k n \leq 2^{\sqrt{\log n}\log\log n} = n^{o(1)}$. Therefore, for the $(2^{O(k\log k)}/\varepsilon^{dk})n^{o(1)}$ characteristic vectors, in $(2^{O(k\log k)}/\varepsilon^{d(k+1)})n^{o(1)}$ time total we compute the expected costs of their canonical subsets. Adding in other operations yields the promised running time. ◀

3.1 k-Means

We now discuss how to modify our algorithm for Faulty k-Median to instead find a $(1+\varepsilon)$-approximate solution to Faulty k-Means. As before, we start by computing a 2-approximate solution D to non-faulty k-Means in $O(n + k^{k+2}\varepsilon^{-(2d+1)k}\log^{k+1} n \log^k(1/\varepsilon))$ time [14]. Then, we partition the points of P into the k subsets $Vor_i(P, D)$ and refine the partition.

[5] A newer result shows existence of a coreset of size $O(k^2/\varepsilon^d)$ [13]. However, the construction time for this smaller coreset is unclear, and an extra $\log n$ factor does not qualitatively affect our result.

We need to modify the refined partition somewhat to work with k-Means. We set the parameter t that determines the number of concentric balls to $\lceil \log_{1+\varepsilon}(\sqrt{2n}) \rceil$, and let $r_j = \sqrt{h_P(D)/(2n)} \cdot (1+\varepsilon)^j$. We now partition the points of $Vor_i(P,D) \cap B_{i,0}$ into $O(1/\varepsilon^d)$ batches each of diameter $\varepsilon r_0/8$, and for each $j \in \{1,\ldots,t\}$, partition the points of $Vor_i(P,D) \cap (B_{i,j} \setminus B_{i,j-1})$ into $O(1/\varepsilon^{d-1})$ batches each of diameter $\varepsilon r_j/16$.

The rest of the algorithm itself remains the same, but for the analysis we need to revise Lemma 14 to account for the different objective in k-Means.

▶ **Lemma 16.** *Let $\rho = \{\rho_1,\ldots,\rho_m\}$ denote the partition of P for Faulty k-Means. Let $Q \subseteq P$ be any subset such that $|Q| \le k$, and let $S = Canon(char(Q,\rho))$. Then we have,*

$$E[h_{P,V}(S)] \le (1+3\varepsilon/4)E[h_{P,V}(Q)].$$

Proof. We recall the notation from Lemma 6 and Lemma 14 that is used in the novel part of the proof. Let $R_S \in Real(S)$ and $R_Q \in Real(Q)$ have the same characteristic vector. Consider any point $p \in P$. Let s be the closest point in R_S to p, and let q be the closest point in R_Q to p. Let i_q be such that $q \in Vor_{i_q}(P,D)$, and let ℓ be the index such that $q \in \rho_\ell$. Given R_S and R_Q have the same characteristic vector, there must be some point $s' \in R_S$ such that $s' \in \rho_\ell$. We have

$$||p-s|| \le ||p-s'|| \le ||p-q|| + ||q-s'||.$$

The remainder of the proof will rely on the fact that for any $x,y \ge 0$, $xy \le x^2/2 + y^2/2$. This fact is a special case of both Young's [28] inequality for products and the AM-GM inequality.

Suppose $q \in B_{i_q,0}$. Recall, $B_{i_q,0}$ is the ball of radius $r_0 = \sqrt{h_P(D)/(2n)}$ centered at d_{i_q}. Then,

$$\begin{aligned}
||p-s||^2 \le (||p-q|| + ||q-s'||)^2 &\le \left(||p-q|| + \frac{\varepsilon r_0}{8}\right)^2 = \left(||p-q|| + \frac{\varepsilon\sqrt{h_P(D)}}{8\sqrt{2n}}\right)^2 \\
&= ||p-q||^2 + \frac{\varepsilon}{4}||p-q|| \cdot \frac{\sqrt{h_P(D)}}{\sqrt{2n}} + \frac{\varepsilon^2 h_P(D)}{128n} \\
&= ||p-q||^2 + \sqrt{\frac{\varepsilon}{4}}||p-q|| \cdot \sqrt{\frac{\varepsilon h_P(D)}{8n}} + \frac{\varepsilon^2 h_P(D)}{128n} \\
&\le ||p-q||^2 + \frac{\varepsilon||p-q||^2}{8} + \frac{9\varepsilon h_P(D)}{128n}
\end{aligned}$$

Now, suppose $q \in (B_{i_q,j} \setminus B_{i_q,j-1})$ for some $j > 0$. Again, $r_{j-1} \le ||p-q|| + ||p-D||$. Therefore,

$$\begin{aligned}
||p-s||^2 \le (||p-q|| + ||q-s'||)^2 &\le \left(||p-q|| + \frac{\varepsilon(1+\varepsilon)r_{j-1}}{16}\right)^2 \\
&\le \left(||p-q|| + \frac{\varepsilon r_{j-1}}{8}\right)^2 \le \left(\left(1+\frac{\varepsilon}{8}\right)||p-q|| + \frac{\varepsilon||p-D||}{8}\right)^2 \\
&= \left(1 + \frac{\varepsilon}{4} + \frac{\varepsilon^2}{64}\right)||p-q||^2 + 2\left(\frac{\varepsilon}{8} + \frac{\varepsilon^2}{64}\right)||p-q|| \cdot ||p-D|| + \frac{\varepsilon^2||p-D||^2}{64} \\
&\le \left(1 + \frac{17\varepsilon}{64}\right)||p-q||^2 + 2\sqrt{\frac{9\varepsilon}{64}}||p-q|| \cdot \sqrt{\frac{9\varepsilon}{64}}||p-D|| + \frac{\varepsilon||p-D||^2}{64} \\
&\le ||p-q||^2 + \frac{13\varepsilon||p-q||^2}{32} + \frac{5\varepsilon||p-D||^2}{32}
\end{aligned}$$

Finally, summing over all p and observing $h_P(D) \leq 2h_P(R_Q)$.

$$\begin{aligned} h_P(R_S) = \sum_{p\in P} ||p - R_S||^2 &\leq \sum_{p\in P} \left(||p - R_Q||^2 + \frac{13\varepsilon||p - R_Q||^2}{32} + \frac{9\varepsilon h_P(D)}{128n} + \frac{5\varepsilon||p - D||^2}{32} \right) \\ &\leq h_P(R_Q) + \frac{13\varepsilon h_P(R_Q)}{32} + \frac{9\varepsilon h_P(D)}{128} + \frac{5\varepsilon h_P(D)}{32} \\ &\leq (1+\varepsilon)h_P(R_Q). \end{aligned}$$

The rest of the proof proceeds as in the one for Lemma 6, with "$(1+\varepsilon)E[h_{P,V}(Q) \mid e_Q(u)]$" appearing in place of "$E[f_{P,V}(Q) \mid e_Q(u)] + diam(\rho)$" alongside analogous substitutions. ◀

The running time analysis for Faulty k-Means is virtually the same as that for Faulty k-Median, except for the larger time needed to find the initial non-faulty 2-approximate solution. We thus conclude with our final theorem.

▶ **Theorem 17.** *Let $P \subset \mathbb{R}^d$ be an instance of Faulty k-Means in d-dimensional Euclidean space, and let C_{opt} denote an optimal solution to this instance. Then for any $\varepsilon > 0$, in $O(kn) + (2^{O(k\log k)}/\varepsilon^{(2d+1)k})n^{o(1)}$ time, one can compute a set $C \subseteq P$ of k centers such that $E[h_{P,V}(C)] \leq (1+\varepsilon)E[h_{P,V}(C_{opt})]$.*

References

1. P. K. Agarwal and C. M. Procopiuc. Exact and approximation algorithms for clustering. *Algorithmica*, 33(2):201–226, 2002. doi:10.1007/s00453-001-0110-y.
2. Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Mach. Learn.*, 75(2):245–248, 2009. doi:10.1007/s10994-009-5103-0.
3. Michael Chau, Reynold Cheng, Ben Kao, and Jackey Ng. Uncertain data mining: An example in clustering location data. In *Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conference (PAKDD)*, volume 3918, pages 199–204. Springer, 2006. doi:10.1007/11731139_24.
4. Shiva Chaudhuri, Naveen Garg, and R. Ravi. The p-neighbor k-center problem. *Inf. Process. Lett.*, 65(3):131–134, 1998. doi:10.1016/S0020-0190(97)00224-X.
5. Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009. doi:10.1137/070699007.
6. Vincent Cohen-Addad, Marcin Pilipczuk, and Michal Pilipczuk. Efficient approximation schemes for uniform-cost clustering problems in planar graphs. In *27th Annual European Symposium on Algorithms (ESA)*, pages 33:1–33:14, 2019. doi:10.4230/LIPIcs.ESA.2019.33.
7. Graham Cormode and Andrew McGregor. Approximation algorithms for clustering uncertain data. In *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 191–200. ACM, 2008. doi:10.1145/1376916.1376944.
8. T. Feder and D. H. Greene. Optimal algorithms for approximate clustering. In *20th Annual ACM Symposium on Theory of Computing (STOC)*, pages 434–444. ACM, 1988. doi:10.1145/62212.62255.
9. Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In Jeff Erickson, editor, *Proceedings of the 23rd ACM Symposium on Computational Geometry, Gyeongju, South Korea, June 6-8, 2007*, pages 11–18. ACM, 2007. doi:10.1145/1247069.1247072.
10. T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi:10.1016/0304-3975(85)90224-5.

11 Sudipto Guha and Kamesh Munagala. Exceeding expectations and clustering uncertain data. In *Proceedings of the Twenty-Eigth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS*, pages 269–278. ACM, 2009. doi:10.1145/1559795.1559836.

12 Sariel Har-Peled. *Geometric Approximation Algorithms*, volume 173 of *Mathematical Surveys and Monographs*. AMS, Boston, MA, USA, 2011.

13 Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

14 Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 291–300, 2004. doi:10.1145/1007352.1007400.

15 Sariel Har-Peled and Benjamin Raichel. Net and prune: A linear time algorithm for euclidean distance problems. *J. ACM*, 62(6):44:1–44:35, 2015. doi:10.1145/2831230.

16 D. S. Hochbaum and D. B. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184, 1985. doi:10.1287/moor.10.2.180.

17 W. Hsu and G. L. Nemhauser. Easy and hard bottleneck location problems. *Discrete Applied Mathematics*, 1(3):209–215, 1979. doi:10.1016/0166-218X(79)90044-1.

18 Lingxiao Huang and Jian Li. Stochastic k-center and j-flat-center problems. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 110–129. SIAM, 2017. doi:10.1137/1.9781611974782.8.

19 Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In John H. Reif, editor, *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 731–740. ACM, 2002. doi:10.1145/509907.510012.

20 Pegah Kamousi, Timothy M. Chan, and Subhash Suri. Closest pair and the post office problem for stochastic points. In *Algorithms and Data Structures - 12th International Symposium (WADS)*, volume 6844 of *Lecture Notes in Computer Science*, pages 548–559. Springer, 2011. doi:10.1007/978-3-642-22300-6_46.

21 Pegah Kamousi, Timothy M. Chan, and Subhash Suri. Stochastic minimum spanning trees in euclidean spaces. In *Proceedings of the 27th ACM Symposium on Computational Geometry (SoCG)*, pages 65–74. ACM, 2011. doi:10.1145/1998196.1998206.

22 Samir Khuller, Robert Pless, and Yoram J. Sussmann. Fault tolerant k-center problems. *Theor. Comput. Sci.*, 242(1-2):237–245, 2000. doi:10.1016/S0304-3975(98)00222-9.

23 Nirman Kumar and Benjamin Raichel. Fault tolerant clustering revisited. In *Proceedings of the 25th Canadian Conference on Computational Geometry (CCCG)*. Carleton University, Ottawa, Canada, 2013. URL: http://cccg.ca/proceedings/2013/papers/paper_36.pdf.

24 Christiane Lammersen, Melanie Schmidt, and Christian Sohler. Probabilistic k-median clustering in data streams. *Theory Comput. Syst.*, 56(1):251–290, 2015. doi:10.1007/s00224-014-9539-7.

25 Alexander Munteanu, Christian Sohler, and Dan Feldman. Smallest enclosing ball for probabilistic data. In *30th Annual Symposium on Computational Geometry (SoCG)*, page 214. ACM, 2014. doi:10.1145/2582112.2582114.

26 Wang Kay Ngai, Ben Kao, Chun Kit Chui, Reynold Cheng, Michael Chau, and Kevin Y. Yip. Efficient clustering of uncertain data. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)*, pages 436–445. IEEE Computer Society, 2006. doi:10.1109/ICDM.2006.63.

27 Haitao Wang and Jingru Zhang. One-dimensional k-center on uncertain data. *Theor. Comput. Sci.*, 602:114–124, 2015. doi:10.1016/j.tcs.2015.08.017.

28 William Henry Young. On classes of summable functions and their fourier series. *Proc. R. Soc. Lond. A*, 87:225–229, 1912. doi:10.1098/rspa.1912.0076.