Agnostic Clustering*

Maria Florina Balcan[1], Heiko Röglin[2**], and Shang-Hua Teng[3]

[1] College of Computing, Georgia Institute of Technology
ninamf@cc.gatech.edu
[2] Department of Quantitative Economics, Maastricht University
heiko@roeglin.org
[3] Computer Science Department, University of Southern California
shanghua.teng@gmail.com

Abstract. Motivated by the principle of agnostic learning, we present an extension of the model introduced by Balcan, Blum, and Gupta [3] on computing low-error clusterings. The extended model uses a weaker assumption on the target clustering, which captures data clustering in presence of outliers or ill-behaved data points. Unlike the original target clustering property, with our new property it may no longer be the case that all plausible target clusterings are close to each other. Instead, we present algorithms that produce a small list of clusterings with the guarantee that all clusterings satisfying the assumption are close to some clustering in the list, proving both upper and lower bounds on the length of the list needed.

1 Introduction

Problems of clustering data from pairwise distance or similarity information are ubiquitous in science. Typical examples of such problems include clustering proteins by function, images by subject, or documents by topic. In many of these clustering applications there is an unknown target or desired clustering, and while the distance information among data is merely heuristically defined, the real goal in these applications is to minimize the clustering error with respect to the target clustering.

A commonly used approach for data clustering is to first choose a particular distance-based objective function Φ (e.g., k-median or k-means) and then design a clustering algorithm that (approximately) optimizes this objective function [1, 2, 7]. The implicit hope is that approximately optimizing the objective function will in fact produce a clustering of low clustering error, i.e. a clustering that is pointwise close to the target clustering. Mathematically, the implicit assumption is that the clustering error of any c-approximation to Φ on the data set is bounded by some ϵ. We will refer to this assumed property as the (c, ϵ) property for Φ.

* This work was done in part while the authors were at Microsoft Research, New England.

** Supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD).

Balcan, Blum, and Gupta [3] have shown that by making this implicit assumption explicit, one can efficiently compute a low-error clustering even in cases when the approximation problem of the objective function is NP-hard. In particular, they show that for any $c = 1 + \alpha > 1$, if data satisfies the (c, ϵ) property for the k-median or the k-means objective, then one can produce a clustering that is $O(\epsilon)$-close to the target, even for values c for which obtaining a c-approximation is NP-hard.

However, the (c, ϵ) property is a strong assumption. In real data there may well be some data points for which the (heuristic) distance measure does not reflect cluster membership well, causing the (c, ϵ) property to be violated. A more realistic assumption is that the data satisfies the (c, ϵ) property only after some number of outliers or ill-behaved data points, i.e., a ν fraction of the data points, have been removed. We will refer to this property as the (ν, c, ϵ) property.

While the (c, ϵ) property leads to the situation that all plausible clusterings (i.e., all the clusterings satisfying the (c, ϵ) property) are $O(\epsilon)$-close to each other, two different sets of outliers could result in two different clusterings satisfying the (ν, c, ϵ) property. We therefore analyze the clustering complexity of this property [4], i.e, the size of the smallest ensemble of clusterings such that any clustering satisfying the (ν, c, ϵ) property is close to a clustering in the ensemble; we provide tight upper and lower bounds on this quantity for several interesting cases, as well as efficient algorithms for outputting a list such that any clustering satisfying the property is close to one of those in the list.

Perspective: The clustering framework we analyze in this paper is related in spirit to the agnostic learning model in the supervised learning setting [6]. In the Probably Approximately Correct (or PAC) learning model of Valiant [8], also known as the realizable setting, the assumption is that the data distribution over labeled examples is correctly classified by some fixed but unknown concept in some concept class, e.g., by a linear separator. In the agnostic setting [6] however, the assumption is weakened to the hope that most of the data is correctly classified by some fixed but unknown concept in some concept space, and the goal is to compete with the best concept in the class by an efficient algorithm. Similarly, one can view the (ν, c, ϵ) property as an agnostic version of the (c, ϵ) property since we assume that the (ν, c, ϵ) property is satisfied if the (c, ϵ) property is satisfied on most but not all of the points and moreover the points where the property is not satisfied are adversarially chosen.

Our results: We present several algorithmic and information-theoretic results in this new clustering model.

For most of this paper we focus on the k-median objective function. In the case where the target clusters are large (have size $\Omega((\epsilon/\alpha + \nu)n)$) we show that the algorithm in [3] can be used in order to output a single clustering that is $(\nu + \epsilon)$-close to the target clustering. We then show that in the more general case there can be multiple significantly different clusterings that can satisfy the (ν, c, ϵ) property. This is true even in the case where most of the points come from large clusters; in this case, however, we show that we can in polynomial time output a small list of k-clusterings such that any clustering that satisfies

the property is close to one of the clusterings in the list. In the case where most of the points come from small clusters, we provide information-theoretic bounds on the clustering complexity of this property.

We also show how both the analysis in [3] for the (c, ϵ) property and our analysis for the $(\nu, 1+\alpha, \epsilon)$ property can be adapted to the *inductive* case, where we imagine our given data is only a small random sample of the entire data set. Based on the sample, our algorithm outputs a clustering or a list of clusterings of the full domain set that are evaluated with respect to the underlying distribution.

We conclude by discussing how our analysis extends to the k-means objective function as well.

2 The Model

The clustering problems we consider fall into the following general framework: we are given a metric space $\mathcal{M} = (X, d)$ with point set X and a distance function $d : \binom{X}{2} \to \mathbb{R}_{\geq 0}$ satisfying the triangle inequality — this is the ambient space. We are also given the actual point set $S \subseteq X$ we want to cluster; we use n to denote the cardinality of S. A k-clustering $\mathcal{C}$ is a partition of S into k (possibly empty) sets $C_1, C_2, \ldots, C_k$. In this work, we always assume that there is a *true* or *target* k-clustering $\mathcal{C}_T$ for the point set S.

Commonly used clustering algorithms seek to minimize some objective function or "score". For example, the *k-median* clustering objective assigns to each cluster C_i a "median" $c_i \in C_i$ and seeks to minimize $\Phi_1(\mathcal{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, c_i)$. Another example is the *k-means* clustering objective, which assigns to each cluster C_i a "center" $c_i \in X$ and seeks to minimize $\Phi_2(\mathcal{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, c_i)^2$. Given a function Φ and an instance $(\mathcal{M}, S)$, let $\mathrm{OPT}_\Phi = \min_{\mathcal{C}} \Phi(\mathcal{C})$, where the minimum is over all k-clusterings of S.

The notion of distance between two k-clusterings $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ and $\mathcal{C}' = \{C'_1, C'_2, \ldots, C'_k\}$ that we use throughout the paper, is the fraction of points on which they disagree under the optimal matching of clusters in $\mathcal{C}$ to clusters in $\mathcal{C}'$; we denote that as $\mathrm{dist}(\mathcal{C}, \mathcal{C}')$. Formally,

$$\mathrm{dist}(\mathcal{C}, \mathcal{C}') = \min_{\sigma \in S_k} \frac{1}{n} \sum_{i=1}^{k} |C_i - C'_{\sigma(i)}|,$$

where S_k is the set of bijections $\sigma : \{1, \ldots, k\} \to \{1, \ldots, k\}$. We say that two clusterings $\mathcal{C}$ and $\mathcal{C}'$ are *ϵ-close* if $\mathrm{dist}(\mathcal{C}, \mathcal{C}') \leq \epsilon$ and we say that a clustering has *error* ϵ if it is ϵ-close to the target.

The $(1+\alpha, \epsilon)$-property: The following notion originally introduced in [3] and later studied in [5] is central to our discussion:

Definition 1. *Given an objective function Φ (such as k-median or k-means), we say that instance (S, d)* satisfies the $(1+\alpha, \epsilon)$-property for Φ *with respect to the target clustering $\mathcal{C}_T$ if all clusterings $\mathcal{C}$ with $\Phi(\mathcal{C}) \leq (1+\alpha) \cdot \mathrm{OPT}_\Phi$ are ϵ-close to the target clustering $\mathcal{C}_T$ for (S, d).*

The $(\nu, 1+\alpha, \epsilon)$-property: In this paper, we study the following more robust variation of Definition 1:

Definition 2. *Given an objective function Φ (such as k-median or k-means), we say that instance (S, d)* satisfies the $(\nu, 1+\alpha, \epsilon)$-property for Φ *with respect to the target clustering $\mathcal{C}_T$ if there exists a set of points $S' \subseteq S$ of size at least $(1-\nu)n$ such that (S', d) satisfies the $(1+\alpha, \epsilon)$-property for Φ with respect to the clustering $\mathcal{C}_T \cap S'$ induced by the target clustering on S'.*

In other words our hope is that the $(1+\alpha, \epsilon)$-property for objective Φ is satisfied only after outliers or ill-behaved data points have been removed. Note that unlike the case $\nu = 0$, in general the $(\nu, 1+\alpha, \epsilon)$-property could be satisfied with respect to *multiple significantly different* clusterings, since we allow the set of outliers or ill-behaved data points to be *arbitrary*. As a consequence we will be interested in the size of the smallest list any algorithm could hope to output that guarantees that at least one clustering in the list has small error. Given the instance (S, d), we say that a given clustering $\mathcal{C}$ is *consistent* with the $(\nu, 1+\alpha, \epsilon)$-property for Φ if (S, d) satisfies the $(\nu, 1+\alpha, \epsilon)$-property for Φ with respect to $\mathcal{C}$. The following notion originally introduced in [4] provides a formal measure of the inherent usefulness of a given property.

Definition 3. *Given an instance (S, d) and the $(\nu, 1+\alpha, \epsilon)$-property for Φ, we define the* (γ, k)-clustering complexity *of the instance (S, d) with respect to the $(\nu, 1+\alpha, \epsilon)$-property for Φ to be the length of the shortest list of clusterings $h_1, \ldots, h_t$ such that any consistent k-clustering is γ-close to some clustering in the list. The* (γ, k) clustering complexity *of the $(\nu, 1+\alpha, \epsilon)$-property for Φ is the maximum of this quantity over all instances (S, d).*

Ideally, the $(\nu, 1+\alpha, \epsilon)$ property should have (γ, k) clustering complexity polynomial in $k, 1/\epsilon, 1/\nu, 1/\alpha$, and $1/\gamma$. Sometimes we analyze the clustering complexity of our property restricted to some family of interesting clusterings. We define this analogously:

Definition 4. *Given an instance (S, d) and the $(\nu, 1+\alpha, \epsilon)$-property for Φ, we define the* (γ, k)-restricted clustering complexity *of the instance (S, d) with respect to the $(\nu, 1+\alpha, \epsilon)$-property for Φ and with respect to some family of clusterings $\mathcal{F}$ to be the length of the shortest list of clusterings $h_1, \ldots, h_t$ such that any consistent k-clustering in the family $\mathcal{F}$ is γ-close to some clustering in the list. The* (γ, k) restricted clustering complexity *of the $(\nu, 1+\alpha, \epsilon)$-property for Φ and $\mathcal{F}$ is the maximum of this quantity over all instances (S, d).*

For example, we will analyze the $(\nu, 1+\alpha, \epsilon)$-property restricted to clusterings in which every cluster has size $\Omega((\epsilon/\alpha + \nu)n)$ or to the case where the average cluster size is at least $\Omega((\epsilon/\alpha + \nu)n)$.

Throughout the paper we use the following notations: For $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, \ldots, n\}$. Furthermore, log denotes the logarithm to base 2. We say that a list $C_1, C_2, C_3, \ldots$ of clusterings is *laminar* if C_{i+1} can be obtained from C_i by merging some of the clusters of C_i.

3 k-Median based Clustering: the $(1+\alpha, \epsilon)$-property

We start by summarizing in Section 3.1 consequences of the $(1+\alpha, \epsilon)$-property that are critical for the new results we present in this paper. We also describe the algorithm presented in [3] for the case that all clusters in the target clustering are large. Then in Section 3.2 we show how this algorithm can be extended to and analyzed in the inductive case.

3.1 Key properties of the $(1+\alpha, \epsilon)$-property

Given an instance of k-median specified by a metric space $\mathcal{M} = (X, d)$ and a set of points $S \subseteq X$, fix an optimal k-median clustering $\mathcal{C}^* = \{C_1^*, \ldots, C_k^*\}$, and let c_i^* be the center point for C_i^*. For $x \in S$, let $w(x) = \min_i d(x, c_i^*)$ be the contribution of x to the k-median objective in $\mathcal{C}^*$ (i.e., x's "weight"), and let $w_2(x)$ be x's distance to the second-closest center point among $\{c_1^*, c_2^*, \ldots, c_k^*\}$. Also, let $w = \frac{1}{n}\sum_{x \in S} w(x) = \frac{\mathrm{OPT}}{n}$ be the average weight of the points. Finally, let $\epsilon^* = \mathrm{dist}(\mathcal{C}_T, \mathcal{C}^*)$; so, from the $(1+\alpha, \epsilon)$-property we have $\epsilon^* < \epsilon$.

Lemma 5 ([3]). *If the k-median instance $(\mathcal{M}, S)$ satisfies the $(1+\alpha, \epsilon)$-property with respect to $\mathcal{C}_T$, then*

(a) less than $6\epsilon n$ points $x \in S$ have $w_2(x) - w(x) < \frac{\alpha w}{2\epsilon}$,
(b) if each cluster in $\mathcal{C}_T$ has size at least $2\epsilon n$, less than $(\epsilon - \epsilon^*)n$ points $x \in S$ on which $\mathcal{C}_T$ and $\mathcal{C}^*$ agree have $w_2(x) - w(x) < \frac{\alpha w}{\epsilon}$, and
(c) for every $z \geq 1$, at most $z\epsilon n/\alpha$ points $x \in S$ have $w(x) \geq \frac{\alpha w}{z\epsilon}$.

Algorithm 1 k-median, the case of large target clusters

Input: τ, b.

Step 1 Construct the graph $G_\tau = (S, E_\tau)$ by connecting all pairs $\{x, y\} \in \binom{S}{2}$ with $d(x, y) \leq \tau$.

Step 2 Create a new graph $H_{\tau,\mathrm{b}}$ where we connect two points by an edge if they share more than bn neighbors in common in G_τ.

Step 3 Let $\mathcal{C}'$ be any clustering obtained by taking the largest k components in $H_{\tau,\mathrm{b}}$, adding the vertices of all other smaller components to any of these.

Step 4 For each point $x \in S$ and each cluster C_j', compute the median distance $d_{\mathrm{med}}(x, j)$ between x and all points in C_j'.
Insert x into the cluster C_i'' for $i = \mathrm{argmin}_j d_{\mathrm{med}}(x, j)$.

Output: Clustering $\mathcal{C}''$

Theorem 6 ([3]). *Assume that the k-median instance satisfies the $(1+\alpha, \epsilon)$-property. If each cluster in $\mathcal{C}_T$ has size at least $(3 + 10/\alpha)\epsilon n + 2$, then given w we can efficiently find a clustering that is ϵ-close to $\mathcal{C}_T$. If each cluster in $\mathcal{C}_T$ has size at least $(4 + 15/\alpha)\epsilon n + 2$, then we can efficiently find a clustering that is ϵ-close to $\mathcal{C}_T$ even without being given w.*

Since some of the elements of this construction are essential in our subsequent proofs, we summarize in the following the main ideas of this proof.

Main Ideas of the Construction: Assume first that we are given w. We use Algorithm 1 with $\tau = \frac{2\alpha w}{5\epsilon}$ and $\mathrm{b} = (1 + 5/\alpha)\epsilon$. For the analysis, let us define $d_{crit} = \frac{\alpha w}{5\epsilon}$. We call point x *good* if both $w(x) < d_{crit}$ and $w_2(x) - w(x) \geq 5d_{crit}$, else x is called *bad*; by Lemma 5 and the fact that $\epsilon^* \leq \epsilon$, if all clusters in the target have size greater than $2\epsilon n$, then at most a $(1 + 5/\alpha)\epsilon$-fraction of points is bad. Let X_i be the *good* points in the optimal cluster C_i^*, and let $B = S \setminus \cup X_i$ be the bad points. For instances satisfying the $(1 + \alpha, \epsilon)$-property, the threshold graph G_τ defined in Algorithm 1 has the following properties: (i) For all x, y in the same X_i, the edge $\{x, y\} \in E(G_\tau)$. (ii) For $x \in X_i$ and $y \in X_{j \neq i}$, $\{x, y\} \notin E(G_\tau)$. Moreover, such points x, y do not share any neighbors in G_τ (by the triangle inequality). This implies that each X_i is contained in a distinct component of the graph $H_{\tau,\mathrm{b}}$; the remaining components of $H_{\tau,\mathrm{b}}$ contain vertices from the "bad bucket" B. Since the X_i's are larger than B, we get that the clustering C' obtained in Step 3 by taking the largest k components in H and adding the vertices of all other smaller components to one of them differs from the optimal clustering $\mathcal{C}^*$ only in the bad points which constitute an $O(\epsilon/\alpha)$ fraction of the total.

To argue that the clustering $\mathcal{C}''$ is ϵ-close to $\mathcal{C}_T$, we call a point x "red" if it satisfies $w_2(x) - w(x) < 5d_{crit}$, "yellow" if it is not red but $w(x) \geq d_{crit}$, and "green" otherwise. So, the green points are those in the sets X_i, and we have partitioned the bad set B into red points and yellow points. The clustering $\mathcal{C}'$ agrees with $\mathcal{C}^*$ on the green points, so without loss of generality we may assume $X_i \subseteq C'_i$. Since each cluster in $\mathcal{C}'_i$ has a strict majority of green points all of which are clustered as in $\mathcal{C}^*$, this means that for a non-red point x, the median distance to points in its correct cluster with respect to $\mathcal{C}^*$ is less than the median distance to points in any incorrect cluster. Thus, $\mathcal{C}''$ agrees with $\mathcal{C}^*$ on all non-red points. Since there are at most $(\epsilon - \epsilon^*)n$ red points on which $\mathcal{C}_T$ and $\mathcal{C}^*$ agree by Lemma 5 — and $\mathcal{C}''$ and $\mathcal{C}_T$ might disagree on all these points — this implies $\mathrm{dist}(\mathcal{C}'', \mathcal{C}_T) \leq (\epsilon - \epsilon^*) + \epsilon^* = \epsilon$, as desired.

The "unknown w" Case: If we are not given the value w, and every target cluster has size at least $(4 + 15/\alpha)\epsilon n + 2$, we instead run Algorithm 1 (with $\tau = \frac{2\alpha w}{5\epsilon}$ and $\mathrm{b} = (1 + 5/\alpha)\epsilon$ repeatedly for different values of w, starting with $w = 0$ (so the graph G_τ is empty) and at each step increasing w to the next value such that G_τ contains at least one new edge. We say that a point is missed if it does not belong to the k largest components of $H_{\tau,\mathrm{b}}$. The number of missed points decreases with increasing w, and we stop with the smallest w, for which we miss at most $bn = (1 + 5/\alpha)\epsilon n$ points and each of the k largest components contains more than $2bn$ points. Clearly, for the correct value of w, we miss at most bn points because we miss only bad points. Additionally, every X_i contains more than $2bn$ points. This implies that our guess for w can only be smaller than the correct w and the resulting graphs G_τ and $H_{\tau,\mathrm{b}}$ can only have fewer edges than the corresponding graphs for the correct w. However, since we miss at most bn points and every set X_i contains more than

bn points, there must be good points from every good set X_i that are not missed. Hence, each of the k largest components corresponds to a distinct cluster $\mathcal{C}_i^*$. We might misclassify all bad points and at most bn good points (those not in the k largest components), but this nonetheless guarantees that each $\mathcal{C}_i'$ contains at least $|X_i| - bn \geq bn + 2$ correctly clustered green points (with respect to $\mathcal{C}^*$) and at most bn misclassified points. Therefore, as shown above for the case of known w, the resulting clustering $\mathcal{C}''$ will correctly cluster all non-red points as in $\mathcal{C}^*$ and so is at distance at most ϵ from $\mathcal{C}_T$.

3.2 The Inductive Case

In this section we consider an *inductive* model in which the set S is merely a small random subset of points of size n from a much larger abstract instance space X, $|X| = N$, $N \gg n$, and the clustering we output is represented *implicitly* through a hypothesis $h : X \to Y$.

Algorithm 2 Inductive k-median

Input: (S, d), $\epsilon \leq 1$, $\alpha > 0$, k, n.

Training Phase:

Step 1 Set $w = \min\{d(x, y) \mid x, y \in S\}$ and $\tau = \frac{2\alpha w}{5\epsilon}$.

Step 2 Apply Steps 1, 2 and 3 in Algorithm 1 with parameters τ and $b = 2(1 + 5/\alpha)\epsilon$ to generate a clustering $C_1' \ldots C_k'$ of the sample S.

Step 3 If the total number of points in $C_1' \ldots C_k'$ is at least $(1-b)n$ and each $|C_i| \geq 2bn$, then terminate the training phase. Else increase τ to the smallest $\tau' > \tau$ for which $G_\tau \neq G_{\tau'}$ and go to Step 2.

Testing Phase:

When a new point z arrives, compute for every cluster C_i' the median distance of z to all sample points in C_i'. Assign z to the cluster that minimizes this median distance.

Our main result in this section is the following:

Theorem 7. *Assume that the k-median instance (X, d) satisfies the $(1 + \alpha, \epsilon)$-property and that each cluster in $\mathcal{C}_T$ has size at least $(6+30/\alpha)\epsilon N+2$. If we draw a sample S of size $n = \Theta\left(\frac{1}{\epsilon} \ln\left(\frac{k}{\delta}\right)\right)$, then we can use Algorithm 2 to produce a clustering that is ϵ-close to the target with probability at least $1 - \delta$.*

Proof. Let X_i be the *good* points in the optimal cluster C_i^*, and let $B = S \setminus \cup X_i$ be the bad points defined as in Theorem 6 *over the whole instance space X*. In particular, if w is the average weight of the points in the optimal k-median solution over the whole instance space, we call point x *good* if both $w(x) < d_{crit}$ and $w_2(x) - w(x) \geq 5d_{crit}$, else x is called *bad*. Let X_i be the *good* points in the optimal cluster C_i^*, and let $B = S \setminus \cup X_i$ be the bad points. Since each cluster in $\mathcal{C}_T$ has size at least $(6 + 30/\alpha)\epsilon N + 2$ we can show using a similar reasoning as in Theorem 6 that $|X_i| > 5|B|$. Also, since our sample is large

enough, $n = \Theta\left(\frac{1}{\epsilon}\ln\left(\frac{k}{\delta}\right)\right)$, by Chernoff bounds, with probability at least $1-\delta$ over the sample we have $|B \cap S| < 2(1+5/\alpha)\epsilon n$ and $|X_i \cap S| \geq 4(1+5/\alpha)\epsilon n$, and so $|X_i \cap S| > 2|B \cap S|$ for all i. This then ensures that if we apply Steps 1, 2 and 3 in Algorithm 1 with parameters $\tau = \frac{2\alpha w}{5\epsilon}$ and $b = 2(1+5/\alpha)\epsilon$ we generate a clustering $C'_1 \ldots C'_k$ of the sample S that is $O(b)$-close to the target on the sample. In particular, all good points in the sample that are in the same cluster form cliques in the graph $H_{\tau,b}$ and good points from different clusters are in different connected components of this graph. So, taking the largest connected components of this graphs gives us a clustering that is $O(b)$-close to the target clustering restricted to the sample S.

If we do not know w, then we use the same approach as in Theorem 6. That is, we start by setting $w = 0$ and increase it until the k largest components in the corresponding graph $H_{\tau,b}$ cover a large fraction of the points. The key point is that the correctness of this approach followed from the fact that the number of good points in every cluster is more than twice the total number of bad points. As we have argued above, this is satisfied with probability at least $1-\delta$ for the sample S as well, and hence, using arguments similar to the ones in Theorem 6 implies that we cluster the whole space with error at most ϵ. □

Note that one can speed up Algorithm 2 as follows. Instead of repeatedly calling Algorithm 1 from scratch, we can store the graphs G and H and only add new edges to them in every iteration of Algorithm 2. Note also that in the test phase, when a new point z arrives, we compute for every cluster C'_i the median distance of z to all *sample points* in C'_i (and not to all the points added so C''_i), and assign z to the cluster that minimizes this median distance. Note also that a natural approach which will not work (due to the bad points) is to compute a centroid/median for each C'_i and then insert new points based on this Voronoi diagram.

4 k-Median based Clustering: the $(\nu, 1+\alpha, \epsilon)$-property

We now study k-median clustering under the $(\nu, 1+\alpha, \epsilon)$-property. If $\mathcal{C}$ is an arbitrary clustering consistent with the property, and its set of outliers or ill-behaved data points is $S \setminus S'$, we will refer to $w = \frac{\mathrm{OPT}}{n}$ as the *value of* $\mathcal{C}$ or the *value of* S', where OPT is the value of the optimal k-clustering of the set S'. We start with the simple observation that if we are given a value w corresponding to a consistent clustering $\mathcal{C}_T$ on a subset S', then we can efficiently find a clustering that is $(\nu+\epsilon)$-close to $\mathcal{C}_T$ if all clusters in $\mathcal{C}_T$ are large.

Proposition 8. *Assume that the target $\mathcal{C}_T$ is consistent with the $(\nu, 1+\alpha, \epsilon)$-property for k-median. Assume that each target cluster has size at least $(3 + 10/\alpha)\epsilon n + 2 + 2\nu n$. Let $S' \subseteq S$ with $|S'| \geq (1-\nu)n$ be its corresponding set of non-outliers. If we are given the value of S', then we can efficiently find a clustering that is $(\nu+\epsilon)$-close to $\mathcal{C}_T$.*

Proof. We can use the same argument as in Theorem 6 with the modification that we treat the outliers or ill-behaved data points $S \setminus S'$ as additional red bad

points. To prove correctness, observe that the only property we used about red bad points is that in the graph G_τ none of them connects to points from two different sets X_i and X_j. Due to the triangle inequality, this is also satisfied for the "outliers". The proof then proceeds as in Theorem 6 above. □

4.1 Large Target Clusters

We now show that the $(\nu+\epsilon, k)$-clustering complexity of the $(\nu, 1+\alpha, \epsilon)$-property is 1 in the "large clusters" case. Specifically:

Theorem 9. *Let $\mathcal{F}$ be the family of clusterings with the property that every cluster has size at least $(4 + 15/\alpha)\epsilon n + 2 + 3\nu n$. Then the $(\nu + \epsilon, k)$ restricted clustering complexity of the $(\nu, 1 + \alpha, \epsilon)$-property with respect to $\mathcal{F}$ is 1, and we can efficiently find a clustering that is $(\nu + \epsilon)$-close to any clustering in $\mathcal{F}$ that is consistent with the $(\nu, 1 + \alpha, \epsilon)$-property; in particular this clustering is $(\nu + \epsilon)$-close to the target $\mathcal{C}_T$.*

Proof. Let $\mathcal{C}_1$ be an arbitrary clustering consistent with the $(\nu, 1+\alpha, \epsilon)$-property of minimal value of w. Let $\mathcal{C}_2$ be any other consistent clustering. By definition we know that there exist sets of points S_1 and S_2 of size at least $(1 - \nu)n$ such that (S_i, d) satisfies the $(1+\alpha, \epsilon)$-property with respect to the induced clustering $\mathcal{C}_i \cap S_i$ on S_i, for $i = 1, 2$. Let w and w' denote the values of the clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$ on the sets S_1 and S_2, respectively; and by assumption we have $w \leq w'$. Furthermore, let $\mathcal{C}_1^*$ and $\mathcal{C}_2^*$ denote the optimal k-clusterings on the sets S_1 and S_2, respectively. We set $\tau = \frac{2\alpha w}{5\epsilon}$ and $\tau' = \frac{2\alpha w'}{5\epsilon}$, and b $= (1 + 5/\alpha)\epsilon + \nu$ and consider the graphs $H_{\tau,b}$ and $H_{\tau',b}$. Let $K_1, \ldots, K_k$ be the k largest connected components in the graph $H_{\tau,b}$, and let $K'_1, \ldots, K'_k$ be the k largest connected components in the graph $H_{\tau',b}$. For $j \in [2]$, let $B_j = (S_j \setminus \cup_i X_i^j) \cup (S \setminus S_j)$ denote the bad set of clustering $\mathcal{C}_j^*$. As in Theorem 6, we can show that $|B_j| \leq ((1 + 5/\alpha)\epsilon + \nu)n$. For $i \in [k]$, we denote by X_i^1 the intersection of K_i with the good set of clustering $\mathcal{C}_1^*$ and we denote by X_i^2 the intersection of K'_i with the good set of clustering $\mathcal{C}_2^*$. By the assumption that the size of the target clusters is more than three times the size of the bad set, we have $X_i^j \geq 2|B_j|$ for all $i \in [k]$ and $j \in [2]$.

As $H_{\tau,b} \subseteq H_{\tau',b}$, this implies that (up to reordering) $K_i \subseteq K'_i$ for every i. This is because otherwise, if we end up merging two components K_i and K_j before reaching w', then one of the clusters K'_l must be a subset of B_1 and so it must be strictly smaller than $(4 + 15/\alpha)\epsilon n + 2 + 3\nu n$. This implies that the clusterings $\mathcal{C}_1^*$ and $\mathcal{C}_2^*$ are $O(\epsilon/\alpha+\nu)$-close to each other since they can only differ on the bad set $B_1 \cup B_2$. By Proposition 8, this implies that also the clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$ are $O(\epsilon/\alpha + \nu)$-close to each other.

Moreover, since $X_i^j \geq 2|B_j|$ for all $i \in [k]$ and $j \in [2]$, using an argument similar to the one in Theorem 6 yields that the clusterings $\mathcal{C}'_w$ and $\mathcal{C}'_{w'}$ obtained by running Algorithm 1 with w and w', respectively, are *identical*; moreover this clustering is $(\nu + \epsilon)$-close to both $\mathcal{C}_1$ and $\mathcal{C}_2$. This follows as the outliers in the sets $S \setminus S_1$ and $S \setminus S_2$ can be treated as additional red bad points as described

in Proposition 8 above. Since $\mathcal{C}_1$ is an arbitrary clustering consistent with the $(\nu, 1+\alpha, \epsilon)$-property with a minimal value of w and $\mathcal{C}_2$ is any other consistent clustering, we obtain that the $(\nu+\epsilon, k)$-clustering complexity is 1.

By the same arguments, we can also use the algorithm for unknown w, described after Theorem 6, to get $(\nu+\epsilon)$-close to any consistent clustering when we do not know the value of w beforehand. □

4.2 Target Clusters that are Large on Average

We show here that if we allow some of the target clusters to be small, then the (γ, k) clustering complexity of the $(\nu, 1+\alpha, \epsilon)$-property is larger than one — it can be as large as k even for $\gamma = 1/k$. Specifically:

Theorem 10. *For $k \leq \nu n$ and $\gamma \leq (1-\nu)/k$ the (γ, k)-clustering complexity of the $(\nu, 1+\alpha, \epsilon)$-property is $\Omega(k)$.*

Proof Sketch. Let $A_1, \ldots, A_k$ be sets of size $n(1-\nu)/k$ and let $x_1, \ldots, x_k$ be additional points not belonging to any of the sets $A_1, \ldots, A_k$ such that the optimal k-median solution on the set $A_1 \cup \ldots \cup A_k$ is the clustering $\mathcal{C} = \{A_1, \ldots, A_k\}$ and the instance $(A_1 \cup \ldots \cup A_k, d)$ satisfies the $(1+\alpha, \epsilon)$-property. We assume that $S \subseteq \mathbb{N}$ and that every set A_i consists of $n(1-\nu)/k$ points at exactly the same position $a_i \in \mathbb{N}$. In our construction, we will have $a_1 < \ldots < a_k$.

By placing the point x_1 very far away from all the sets A_i and by placing A_1 and A_2 much closer together than any other pair of sets, we can achieve that the optimal k-median solution on the set $A_1 \cup \ldots \cup A_k \cup \{x_1\}$ is the clustering $\{A_1 \cup A_2, A_3, \ldots, A_k, \{x_1\}\}$ and that the instance $(A_1 \cup A_k \cup \{x_1\}, d)$ satisfies the $(1+\alpha, \epsilon)$-property. We can continue analogously and place x_2 very far away from all the sets A_i and from x_1. Then the optimal k-median clustering on the set $A_1, \ldots, \cup \ldots \cup A_k \cup \{x_1, x_2\}$ will be $\{A_1 \cup A_2 \cup A_3, A_4, \ldots, A_k, \{x_1, x_2\}\}$ if A_2 and A_3 are much closer together than A_i and A_{i+1} for $i \geq 3$. The instance also satisfies the $(1+\alpha, \epsilon)$-property. This way, each of the clusterings $\{A_1 \cup \ldots \cup A_i, A_{i+1} \ldots A_k, \{x_1\}, \{x_2\}, \ldots, \{x_{i-1}\}\}$ is a consistent target clustering, and the distance between any of them is at least γ. □

Note that in the example in Theorem 10 all the clusterings that satisfy the $(\nu, 1+\alpha, \epsilon)$-property have the feature that the total number of points that come from large clusters (of size at least $n(1-\nu)/k$) is at least $(1-\nu)n$. We show that in such cases we also have an upper bound of k on the clustering complexity.

Theorem 11. *Let $b = (6 + 10/\alpha)\epsilon + \nu$. Let $\mathcal{F}$ be the family of clusterings with the property that the total number of points that come from clusters of size at least $2bn$ is at least $(1-\beta)n$. Then the $(2b+\beta, k)$ restricted clustering complexity of the $(\nu, 1+\alpha, \epsilon)$-property with respect to $\mathcal{F}$ is at most k and we can efficiently construct a list of length at most k such that any clustering in $\mathcal{F}$ that is consistent with the $(\nu, 1+\alpha, \epsilon)$-property is $(2b+\beta)$-close to one of the clusterings in the list.*

Proof. The main idea of the proof is to use the structure of the graphs H to show that the clusterings that are consistent with the $(\nu, 1+\alpha, \epsilon)$-property are almost laminar with respect to each other. Note that for all $w < w'$ we have $G_w \subseteq G_{w'}$ and $H_w \subseteq H_{w'}$. Here we used G_w and H_w as abbreviations for G_τ and H_τ with $\tau = \frac{2\alpha w}{5\epsilon}$. In the following, we say that a cluster is large if it contains at least $2bn$ elements. To find a list of clusterings that "covers" all the relevant clusterings, we use the following algorithm. We keep increasing the value of w until we reach a value w_1 such that the following is satisfied: Let $K_1, \ldots, K_k$ denote the k largest connected components of the graph H_{w_1} and assume $|K_1| \geq |K_2| \geq \ldots \geq |K_k|$. We set $k_1 = \max\{i \in [k] \mid |K_i| \geq bn\}$ and stop for the smallest w_1 for which the clusters $K_1, \ldots, K_{k_1}$ cover together a significant fraction of the space, namely a $1 - (b + \beta)$ fraction. Let $\tilde{S} = K_1 \cup \ldots \cup K_{k_1}$. The first clustering we add to the list contains a cluster for each of the components $K_1, \ldots, K_{k_1}$ and it assigns the points in $S \setminus \tilde{S}$ arbitrarily to those. Now we increase the value of w and each time we add an edge in H_w between two points in different components K_i and K_j, we merge the corresponding clusters to obtain a new clustering with at least one cluster less. We add this clustering to our list and we continue until only one cluster is left. As in every step, the number of clusters decreases by at least one, the list of clusterings produced this way has length at most $k_1 \leq k$. Let $w_1, w_2, \ldots$ denote the values of w for which the clusterings are added to the list. To complete the proof, we show that any clustering $\mathcal{C}$ satisfying the property is $(2b+\beta)$-close to one of the clusterings in the list we constructed. Let $w_{\mathcal{C}}$ denote the value corresponding to $\mathcal{C}$. First we notice that $w_{\mathcal{C}} \geq w_1$. This follows easily from the structure of the graph $H_{w_{\mathcal{C}}}$: it has one connected component for every large cluster in $\mathcal{C}$ and each of these components must contain at least bn points as every large cluster contains at least $2bn$ points and the bad set contains at most bn points. Also by definition and the fact that the size of the bad set is bounded by bn, it follows that these components together cover at least a $1 - (b + \beta)$ fraction of the points. This proves that $w_{\mathcal{C}} \geq w_1$ by the definition of w_1. Now let i be maximal such that $w_i \leq w_{\mathcal{C}}$. We show that the clustering we output at w_i is $(2b+\beta)$-close to the clustering $\mathcal{C}$. Let $K'_1, \ldots, K'_{k'}$ denote the components in H_{w_i} that evolved from the K_i and let $K''_1, \ldots, K''_{k''}$ denote the evolved components in $H_{w_{\mathcal{C}}}$. As $w_{\mathcal{C}} < w_{i+1}$, $k' = k''$ we can assume (up to reordering) that $K'_i = K''_i$ on the set $\tilde{S}$. As all points in $\tilde{S}$ that are not in the bad set for w_i are clustered in $\mathcal{C}$ according to the components $K''_1, \ldots, K''_{k''}$, the clusterings corresponding to w_i and $w_{\mathcal{C}}$ can only differ on $S \setminus \tilde{S}$ and the bad set for w_i. Using the fact $|S \setminus \tilde{S}| \leq (b+\beta)n$ and that the size of the bad set is bounded by bn, we get that the clustering we output at w_i is $(2b+\beta)$-close to the clustering $\mathcal{C}$, as desired. □

Moreover, if every large cluster is at least as large as $(12+20/\alpha)\epsilon n+2\nu n+2\beta$, then, as for w_1 the size of the missed set is at most $(6 + 10/\alpha)\epsilon n + \nu n + \beta$, the intersection of the good set with every large cluster is larger than the missed set for w_i for any i. This then implies that if we apply the median argument from Step 4 of Algorithm 1, the clustering we get for w_i is $(\nu + \epsilon + \beta)$-close to the clustering $\mathcal{C}$ if i is chosen as in the previous proof. Together with Theorem 11 this implies the following corollary.

Corollary 12. *Let $b = (6 + 10/\alpha)\epsilon + \nu$. Let $\mathcal{F}$ be the family of clusterings with the property that the average cluster size n/k is at least $2bn/(1-\beta)$. Then the $(\nu + \epsilon + \beta, k)$ restricted clustering complexity of the $(\nu, 1+\alpha, \epsilon)$-property with respect to $\mathcal{F}$ is at most k and we can efficiently construct a list of length at most k such that any clustering in $\mathcal{F}$ that is consistent with the $(\nu, 1+\alpha, \epsilon)$-property is $(\nu + \epsilon + \beta)$-close to one of the clusterings in the list.*

The Inductive Case We show here how the algorithm in Theorem 11 can be extended to the inductive setting.

Theorem 13. *Let $b = (6 + 10/\alpha)\epsilon + \nu$. Let $\mathcal{F}$ be the family of clusterings with the property that the total number of points that come from clusters of size at least $2bn$ is at least $(1-\beta)n$. If we draw a sample S of size $n = O\big(\frac{1}{\epsilon}\ln\big(\frac{k}{\delta}\big)\big)$, then we can efficiently produce a list of length at most k such that any clustering in the family $\mathcal{F}$ that is consistent with the $(\nu, 1+\alpha, \epsilon)$-property is $3(2b+\beta)$-close to one of the clusterings in the list with probability at least $1-\delta$.*

Proof Sketch. In the training phase, we will run the algorithm in Theorem 11 over the sample to get a list of clusterings $\mathcal{L}$. Then we run an independent "test phase" for each clustering in this list. Let $\mathcal{C}$ be one such clustering in the list $\mathcal{L}$ with clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$, and let $\tilde{S}$ be the set of relevant points defined Theorem 11. In the test phase, when a new point x comes in, then we compute for each cluster $\mathcal{C}'_i$ the medium distance of x to $\mathcal{C}'_i \cap \tilde{S}$, and insert it into the cluster $\mathcal{C}'_i$ to which it has the smallest median distance.

To prove correctness we use the fact that, as shown in Theorem 11, the $(2b+\beta, k)$-clustering complexity of the $(\nu, 1+\alpha, \epsilon)$-property is at most k, when restricted to clusterings in which the total number of points coming from clusters of size at least $2bn$ is at least $(1-\beta)n$. Let $\mathcal{L}$ be a list of $k_1 \leq k$ clusterings such that any consistent clustering is $(2b+\beta)$-close to one of them.

Now the argument is similar to the one in Theorem 7. In the proof of that theorem, we used a Chernoff bound to argue that with probability at least $1-\delta$ the good set of any cluster that is contained in the sample is more than twice as large as the total bad set in the sample. Now we additionally apply a union bound over the at most k clusterings in the list $\mathcal{L}$ to ensure this property for each of the clusterings. From that point on the arguments are analogous to the arguments in Theorem 7. □

4.3 Small Target Clusters

We now consider the general case, where the target clusters can be arbitrarily small. We start with a proposition showing that if we are willing relax the notion of closeness significantly then the clustering complexity is still upper bounded by k even in this general case. With a more careful analysis, we then show a better upper bound on the clustering complexity in this general case.

Proposition 14. *Let $b = (6 + 10/\alpha)\epsilon + \nu$. Then the $((k+4)b, k)$-clustering complexity of the $(\nu, 1+\alpha, \epsilon)$-property is at most k.*

Proof. Let us consider a clustering $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_k)$ and a set $S' \subseteq S$ with $|S'| \geq (1-\nu)n$ such that (S', d) satisfies the $(1+\alpha, \epsilon)$-property with respect to the induced target clustering $\mathcal{C} \cap S'$. Let us first have a look at the graph G_w. There exists a bad set B of size at most bn, and for every cluster i, the points in $X_i = \mathcal{C}_i \setminus B$ form cliques in G_w. There are no edges between X_i and X_j for $i \neq j$ and there is no point $x \in B$ that is simultaneously connected to X_i and X_j for $i \neq j$.

If there are two different consistent clusterings $\mathcal{C}^1$ and $\mathcal{C}^2$ that have the same value w, then, by the properties of G_w, all points in $S \setminus (B_1 \cup B_2)$ are identically clustered. Hence, $\mathrm{dist}(\mathcal{C}^1, \mathcal{C}^2) \leq (|B_1| + |B_2|)/n \leq 2b$. This implies that we do not lose too much by choosing for every value w with multiple consistent clusterings one of them as representative. To be precise, let $w'_1 < w'_2 < \cdots < w'_{s'}$ be a list of all values for which a correct clustering exists and for every w'_i, let $\mathcal{C}'^i$ denote a correct clustering with value w'_i. We construct a sparsified list $\mathcal{L}$ of clusterings as follows: insert $\mathcal{C}'^1$ into $\mathcal{L}$; if the last clustering added to $\mathcal{L}$ is $\mathcal{C}'^i$, add $\mathcal{C}'^j$ for the smallest $j > i$ for which $\mathrm{dist}(\mathcal{C}'^i, \mathcal{C}'^j) \geq (k+2)b$. This way, the list $\mathcal{L}$ will contain clusterings $\mathcal{C}^1, \ldots, \mathcal{C}^s$ with values $w_1, \ldots, w_s$ such that every correct clustering is $(k+4)b$-close to at least one of the clusterings in $\mathcal{L}$.

It remains to bound the length s of the list $\mathcal{L}$. Let us assume for contradiction that $s \geq k+1$. According to the properties of the graphs G_{w_i}, the clusterings that are induced by the clusterings $\mathcal{C}^1, \ldots, \mathcal{C}^{k+1}$ on the set $S \setminus (B_1 \cup \ldots \cup B_{k+1})$ are laminar. Furthermore, as the bad set $B_1 \cup \ldots \cup B_{k+1}$ has size at most $(k+1)bn$, two consecutive clusterings in the list must differ on the set $S \setminus (B_1 \cup \ldots \cup B_{k+1})$, which means together with the laminarity implies that two clusters must have merged. This can happen at most $k-1$ times, contradicting the assumption that $s \geq k+1$. □

We will improve the result in the above proposition by imposing that consecutive clusterings in the list $\mathcal{L}$ in the above proof are significantly different in the laminar part. In particular we will make use of the following lemma which shows that if we have a laminar list of clusterings then the sum of the pairwise distances between consecutive clusterings cannot be too big; this implies that if the pairwise distances between consecutive clusterings are all large, then the list must be short.

Lemma 15. *Let $\mathcal{C}^1, \ldots, \mathcal{C}^s$ be a laminar list of clusterings, let $k \geq 2$ denote the number of clusters in $\mathcal{C}^1$, and let $\beta \in (0,1)$. If $\mathrm{dist}(\mathcal{C}^i, \mathcal{C}^{i+1}) \geq \beta$ for every $i \in [s-1]$, then $s \leq \min\{\frac{9\log(k/\beta)}{\beta}, k\}$.*

Proof. When going from $\mathcal{C}^i$ to $\mathcal{C}^{i+1}$, clusters contained in the clustering $\mathcal{C}^i$ merge into bigger clusters contained in $\mathcal{C}^{i+1}$. Merging the clusters $K_1, \ldots, K_\ell \in \mathcal{C}^i$ with $|K_1| \geq |K_2| \geq \cdots \geq |K_\ell|$ into cluster $K \in \mathcal{C}^{i+1}$ contributes $(|K_2| + \cdots + |K_l|)/n$ to the distance between $\mathcal{C}^i$ and $\mathcal{C}^{i+1}$. When going from $\mathcal{C}^i$ to $\mathcal{C}^{i+1}$, multiple such merges can occur and we know that their total contribution to the distance must be at least β. We consider a single merge in which the pieces $K_1, \ldots, K_\ell \in \mathcal{C}^i$ merge into $K \in \mathcal{C}^{i+1}$ virtually as $\ell - 1$ merges and associate with each of them

a type. We say that the merge corresponding to K_i, $i = 2, \ldots, \ell$, has type $j \in \mathbb{N}$ if $|K_i| \in [n/2^{j+1}, n/2^j)$. If K_i has type j, we say that the data points contained in K_i participate in a merge of type j.

For the step from $\mathcal{C}^i$ to $\mathcal{C}^{i+1}$, let x_{ij} denote the total number of virtual merges of type j that occur. The number of merges of type j that can occur during the whole sequence from $\mathcal{C}^1$ to $\mathcal{C}^s$ is bounded from above by 2^{j+1} as each of the n data points can participate at most once in a merge of type j. This follows because once a data point participated in a merge of type j, it is contained in a piece of size at least $n/2^j$.

We are only interested in types $j \leq L = \lfloor \log(k/\beta) \rfloor + 1$. As there can be at most $k-1$ merges from $\mathcal{C}^i$ to $\mathcal{C}^{i+1}$, the total contribution to the distance between $\mathcal{C}^i$ and $\mathcal{C}^{i+1}$ coming from larger types can be at most $k/2^{L+1} \leq \beta/2$. Hence for every $i \in [s-1]$, the total contribution of types $j \leq L$ must be at least $\beta/2$.

In terms of the x_{ij}, these conditions can be expressed as

$$\forall j \in [L]\colon \sum_{i=1}^{s-1} \frac{x_{ij}}{2^{j+1}} \leq 1 \quad \text{and} \quad \forall i \in [s-1]\colon \sum_{j=1}^{L} \frac{x_{ij}}{2^j} \geq \frac{\beta}{2}\,.$$

This yields

$$\frac{(s-1)\beta}{4} \leq \sum_{i=1}^{s-1} \sum_{j=1}^{L} \frac{x_{ij}}{2^{j+1}} \leq L\,,$$

and hence, $s \leq 4L/\beta + 1 \leq \frac{4\lfloor \log(k/\beta) \rfloor + 4}{\beta} + 1 \leq \frac{9 \log(k/\beta)}{\beta}$. As in every step at least two clusters must merge, $s \leq k$ and the lemma follows. □

We can now show the following upper bound on the clustering complexity.

Theorem 16. *Let $b = (6 + 10/\alpha)\epsilon + \nu$. Then the $(9\sqrt{b \log(k/b)}, k)$-clustering complexity of the $(\nu, 1+\alpha, \epsilon)$-property is at most $4\sqrt{\log(k/b)/b}$.*

Proof. We use the same arguments as in Proposition 14. We construct $\mathcal{L}$ in the same way, but with $7\sqrt{b \log(k/b)}$ instead of $(k+2)b$ as bound on the distance of consecutive clusterings. We assume for contradiction that $s \geq t := 4\sqrt{\log(k/b)/b}$ and apply Lemma 15 with $\beta = 7\sqrt{b \log(k/b)} - s'b \geq 3\sqrt{b \log(k/b)}$ to the induced clusterings on $S \setminus (B_1 \cup \ldots \cup B_t)$. This yields $s < t$, contradicting the assumption that $s \geq t$. □

5 Discussion and Open Questions

In this work we extend the results of Balcan, Blum, and Gupta [3] on finding low error clusterings to the *agnostic setting* where we make the weaker assumption that the data satisfies the (c, ϵ) property only after some outliers have been removed.

While we have focused in this paper on the (ν, c, ϵ) property for k-median, most of our results extend directly to the k-means objective as well. In particular,

for the k-means objective one can prove an analog of Lemma 5 with different constants which then can be propagated through the main results of this paper.

It is worth noting that we have assumed implicitly throughout the paper that the fraction of outliers or a good upper bound on it ν is known to the algorithm. In the most general case, where no good upper bound on ν is known, i.e., in the purely agnostic setting, we can run our algorithms $1/\epsilon$ times once for each integral multiplicative of ϵ, thus incurring only a $1/\epsilon$ multiplicative factor increase in the clustering complexity and in the running time.

Open Questions: The main concrete technical questions left open are whether one can show a better upper bound on the clustering complexity in the case of small target clusters and whether in this case there is an efficient algorithm for constructing a short list of clusterings such that every consistent clustering is close to one of the clusterings in the list.

More generally, it would also be interesting to analyze other natural variations of the (c, ϵ) property. For example, a natural direction would be to consider variations that express beliefs that only the c-approximate clusterings that might be returned by natural approximation algorithms are close to the target. In particular, many approximation algorithms for clustering return Voronoi-based clusterings [7]. In this context, a natural relaxation of the (c, ϵ)-property is to assume that only the Voronoi-based clusterings that are c-approximations to the optimal solution are ϵ-close to the target. It would be interesting to analyze whether this is sufficient for efficiently finding low-error clusterings, both in the realizable and in the agnostic setting.

Acknowledgements: We thank Avrim Blum and Mark Braverman for a number of helpful discussions.

References

1. K. Jain, M. Mahdian, and A. Saberi. A new greedy approach for facility location problems. In *STOC*, 2002.
2. M. Charikar, S. Guha, E. Tardos, and D. B. Shmoys. A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, 1999.
3. M. F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2009.
4. M. F. Balcan, A. Blum, and S. Vempala. A discrimantive framework for clustering via similarity functions. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, 2008.
5. M. F. Balcan and M. Braverman. Finding low error clusterings. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
6. M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning Journal*, 1994.
7. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k -means clustering. In *Proceedings of the Eighteenth Annual Symposium on Computational Geometry*, 2002.
8. L. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.