

# Notes on Estimating the Quality of Clustering

## 1 Tolerant Property Testing and Distance Approximation

In this lecture we return to some extensions of property testing and focus on the particular problem of clustering a set of points.

Recall that the goal of a property testing algorithm is to distinguish between the case in which an object has a predetermined property  $\mathcal{P}$  and the case in which it is  $\epsilon$ -far from any object that has the property, for a given distance parameter  $\epsilon$ . In other words, the algorithm should distinguish between having distance 0 to the property and having distance greater than  $\epsilon$ .

One natural extension, which we'll refer to as *tolerant testing*, is to distinguish between having distance at most  $\epsilon_1$ , where  $\epsilon_1 \geq 0$ , and having distance greater than  $\epsilon_2$  (for  $\epsilon_2 > \epsilon_1$ ). Another natural extension is to *distance approximation*. Namely, here the goal is to approximate the distance to having the property. In order to understand the relation between the two variants we need to be a bit more precise in our definitions.

We shall say that an algorithm is a *purely additive* distance approximation algorithm for a property  $\mathcal{P}$  if, for every object  $O$  and for any given  $0 < \delta \leq 1$ , it outputs an estimate  $\hat{\epsilon}$  such that with probability at least  $2/3$  we have that

$$\epsilon_{\mathcal{P}}(O) - \delta \leq \hat{\epsilon} \leq \epsilon_{\mathcal{P}}(O) + \delta$$

where  $\epsilon_{\mathcal{P}}(O)$  is the distance that  $O$  has to (the closest object that has) property  $\mathcal{P}$ . More generally we may consider other forms of approximation, both weaker, where we allow both a multiplicative factor and an additive factor:  $\epsilon_{\mathcal{P}}(O)/C - \delta \leq \hat{\epsilon} \leq C \cdot \epsilon_{\mathcal{P}}(O) + \delta$  for some  $C > 1$  and stronger, where we allow only a multiplicative factor of the form  $(1 \pm \gamma)$  where  $0 < \gamma \leq 1$ , and we ask that the algorithm work for every  $\gamma$ .

We shall say that an algorithm is a *fully-tolerant* testing algorithm for a property  $\mathcal{P}$  if, for every object  $O$  and for every given  $\epsilon_1, \epsilon_2$  such that  $0 \leq \epsilon_1 < \epsilon_2 \leq 1$ , if  $\epsilon_{\mathcal{P}}(O) > \epsilon_2$  then it rejects, and if  $\epsilon_{\mathcal{P}}(O) \leq \epsilon_1$  then it accepts.

ON THE RELATION BETWEEN THE TWO EXTENSIONS. Clearly, a purely additive distance approximation algorithm is stronger (more precisely, not weaker) than a fully tolerant testing algorithm. Given  $\epsilon_1, \epsilon_2$  for the latter problem, we run the algorithm for the former problem with  $\delta = \frac{\epsilon_2 - \epsilon_1}{2}$ . We then consider the output  $\hat{\epsilon}$ . If it is greater than  $\epsilon - \delta/2$  then we reject and if it is at most  $\epsilon - \delta/2$  we accept. A transformation can be done in the other direction at a logarithmic cost in  $1/\epsilon_{\mathcal{P}}(O)$ . The idea is to do a kind of binary search (where there are some small subtleties). There are also analogous relations between the other variants.

IMPLICATIONS OF TESTING ON TOLERANT TESTING The first observation is that if we have a testing algorithm that works by uniformly selecting the points it queries, and it performs  $Q(\epsilon)$  queries, then it can be easily transformed into a tolerant testing algorithm that works for  $\epsilon_2 = \epsilon$  and  $\epsilon_1 = 1/(cQ(\epsilon))$  for some constant  $c$ , and the query complexity of the tolerant testing algorithm is  $c'Q(\epsilon)$ , for a constant  $c'$ . (In fact, if the algorithm has one-sided error, then we use it as is, and get  $c = 3$ ,  $c' = 1$ . If it has two sided error we repeat it several times and take a majority vote.)

For problems in which the query complexity is linear in  $1/\epsilon$  (such as linearity testing) we get tolerant testing for  $\epsilon_1 = \epsilon_2/c$  for a constant  $c$ . However, we can't get full tolerance from this transformation, and furthermore, for many problems  $Q(\epsilon)$  is polynomial in  $1/\epsilon$  so we get something quite weak (e.g.,  $\epsilon_1 = (\epsilon_2)^2$ ).

SOME KNOWN RESULTS RELATED TO TESTING PROBLEMS WE HAVE SEEN. The strongest general result is that in the dense graphs model every property that has a testing algorithm whose query complexity is only a function of  $\epsilon$ , has a fully tolerant testing whose query complexity is only a function of  $\delta$ . Since this is a general transformation, which uses the regularity lemma, the dependence on  $1/\delta$  may be much higher in the tolerant case as compared to the dependence on  $1/\epsilon$  in the standard case. For the class of properties that are defined by partitions, such as bipartiteness,  $k$ -colorability, having a clique of at least a given size  $\rho n$  and more, the general-partition (standard) testing algorithm actually works by estimating the distance to having the property. Thus, for these properties we have purely additive distance approximation algorithms whose query complexity is polynomial in  $1/\delta$ . (Note the difference in the case of clique between estimating the distance to having a clique of size at least  $\rho n$  and approximating the size of the maximum clique.)

There are also results for distance approximation in sparse graphs (some are purely additive and some are not), though there is no known general transformation, and there are results for monotonicity and properties of codes. On the other hand there are some negative results showing that there are properties that have an efficient standard testing algorithm but no efficient tolerant testing algorithm.

## 2 Tolerant Testing of Clustering and Finding Approximately Good Clusterings

Today we shall discuss a property that we haven't talked about before: "Clusterability". Thus we shall "capture several birds": (1) We'll see a property we haven't seen (over objects of a type we haven't seen before); (2) We'll see an example of a tolerant testing algorithm; (3) We'll see how this algorithm can be transformed into a more standard approximation algorithm; (4) We'll see a new proof technique. Our goal is to get a high-level idea of the analysis, and hence we'll skip quite a few proofs.

PROBLEM DEFINITION. Let  $X$  be a set of  $n$  points, and let  $\text{dist} : X \times X \rightarrow \mathfrak{R}$  be a distance function defined over pairs of points in  $X$ . Let  $C : 2^X \rightarrow \mathfrak{R}$  be a *cost measure* on sets of points, defined based on the underlying distance function  $\text{dist}(\cdot, \cdot)$ . For a given  $k$ -way partition  $P = \{X_i\}_{i=1}^k$  of  $X$ , the cost of the partition  $P$  is defined as  $\max_i \{C(X_i)\}$ , and with a slight abuse of notation is denoted by  $C(P)$ .

In particular, we shall be interested in the case where  $C(\cdot)$  is the *diameter* cost measure. Namely,

for a given subset  $S \subseteq X$ ,  $C(S) = \max_{x,y \in S} \{\text{dist}(x,y)\}$ . In all that follows, for simplicity we refer to subsets. However, all definitions extend to multisets.

**Definition 1 (( $k, b$ )-clusterable)** *Let  $k$  be an integer and let  $b$  be a real value. We say that a set  $X$  is  $(k, b)$ -clusterable with respect to the cost function  $C(\cdot)$  (and the underlying distance function  $\text{dist}(\cdot, \cdot)$ ), if there exists a  $k$ -way partition  $P = \{X_i\}_{i=1}^k$  of  $X$  such that  $C(P) \leq b$ .*

**Definition 2 ( $\epsilon$ -far)** *A set  $X$  is said to be  $\epsilon$ -far from being  $(k, b)$ -clusterable with respect to  $C(\cdot)$  for a given  $0 \leq \epsilon \leq 1$ , if for every subset  $Y \subseteq X$  of size at least  $(1-\epsilon)|X|$ , and for every  $k$ -way partition  $P = \{Y_i\}_{i=1}^k$  of  $Y$ , we have  $C(P) > b$ . Otherwise the set  $X$  is  $\epsilon$ -close to being  $(k, b)$ -clusterable with respect to  $C(\cdot)$ .*

**Definition 3 (Hereditary Cost Measures)** *We say that the cost measure  $C(\cdot)$  is a hereditary clustering cost if for every  $k$  and  $b$ , whenever  $X$  is  $(k, b)$ -clusterable with respect to  $C(\cdot)$ , then so is every subset  $Y$  of  $X$ .*

Note that the diameter cost is hereditary. We consider the following “natural” algorithm for tolerant testing of clustering.

**Algorithm 1** (Tolerant Testing Algorithm for  $(k, b)$ -Clustering with respect to cost  $C(\cdot)$ )

1. Uniformly and independently select  $m$  points from  $X$  (where  $m$  will be specified later). Denote the (multi)set of points selected by  $U$ .
2. Let  $\delta = \epsilon_2 - \epsilon_1$ . If  $U$  is  $(\epsilon_1 + \delta/2)$ -close to being  $(k, b)$ -clusterable with respect to  $C(\cdot)$  then accept, otherwise reject.

In what follows we will show that for the cost measures we consider, the query complexity,  $m$ , of the above algorithm is independent of  $|X|$  and is polynomial in  $k$  and  $1/\delta$ . As to the running time of the algorithm, it depends on the particular procedures that are applied in order to decide if the  $m$  sampled points are  $(\epsilon_1 + \delta/2)$ -close to being  $(k, b)$ -clusterable. In general, it is always upper bounded by  $O((k+1)^m \cdot k \cdot T_C(m))$  where  $T_C(m)$  is the time sufficient for computing the cost  $C$  of a subset of size at most  $m$ . For the cost measures we consider, we get an exponential dependence on  $k$ . Since the corresponding decision problems are NP-hard, we cannot expect to do much better.

Recall that a tolerant testing algorithm should accept  $X$  with probability at least  $2/3$  if  $X$  is  $\epsilon_1$ -close to being  $(k, b)$ -clusterable with respect to  $C(\cdot)$ . The following lemma, which follows easily by applying an additive Chernoff bound, ensures that this is the case with Algorithm 1 whenever the cost measure  $C(\cdot)$  is a hereditary clustering cost and the sample size  $m$  is sufficiently large (but still sublinear in  $|X|$ , and actually independent of  $|X|$ ).

**Lemma 1** *Let  $C(\cdot)$  be a hereditary clustering cost, and let  $X$  be a set of points that is  $\epsilon_1$ -close to being  $(k, b)$ -clusterable with respect to  $C(\cdot)$ . Consider the choice of  $m = \Omega(1/\delta^2)$  uniformly and independently selected points from  $X$ , and denote the (multi)set of points selected by  $U$ . Then with probability at least  $2/3$  over the choice of  $U$  it is  $(\epsilon_1 + \delta/2)$ -close to being  $(k, b)$ -clusterable with respect to  $C(\cdot)$ .*

The focus of the analysis of Algorithm 1 is hence on proving that, for an appropriate choice of the sample size  $m$ , if  $X$  is  $\epsilon_2$ -far from being  $(k, b(1 + \beta))$ -clusterable with respect to  $C(\cdot)$ , then it is rejected with probability at least  $2/3$ .

In what follows we describe a framework under which such claims can be proved. This framework extends the abstract combinatorial programs presented by Czumaj and Sohler [CS], which was defined for standard testing. We apply this framework to establish the correctness of Algorithm 1 when the underlying distance measure is a general metric and when it is the Euclidean metric over  $\mathbb{R}^d$ .

## 2.1 A Framework for tolerant testing of clustering: Skeletons and Witnesses

We start by discussing the ideas behind the framework. Many of the previous standard testing algorithms applied the “natural” algorithm, which selects a small sample and checks if the sample has property  $\mathcal{P}$ . Although this algorithm is simple, it is usually not an easy task to prove that it works correctly (if at all). The typical proof technique used was to view the sample as two sub-samples, where the first sub-sample is used as a *skeleton* that induces certain constraints on points in the tested object that do not belong to the sub-sample. These constraints are *always* satisfied when the tested object has the property. The heart of the proof is then focused in showing that in case the object is far from having  $\mathcal{P}$ , then necessarily there are many points that violate the constraints induced by the skeleton. The second sub-sample is then used to provide *witnesses* to these violations. In order to circumvent the use of two sub-samples, [CS] defined a few conditions that should hold regarding the skeletons and the witnesses (once these are defined properly), from which the correctness of the natural algorithm follows. This allows to separate between the combinatorial structure of the proof and the error probability analysis.

We next formalize the notions of skeletons and witnesses in the context of clustering.

**Definition 4 (Skeletons)** *A skeleton defined over a set  $X$  of points is a partition  $P = \{S_i\}_{i=1}^t$  of a subset  $S = \bigcup_{i=1}^t S_i$  of  $X$ . We denote this subset  $S$  by  $\text{Set}(P)$ . For a set of skeletons  $\mathbb{K}$  defined over  $X$ , we say that  $\mathbb{K}$  has order  $s$  if  $|\text{Set}(P)| \leq s$  for every skeleton  $P \in \mathbb{K}$ .*

We emphasize that there may be several skeletons (partitions) associated with the same subset  $S$ .

**Definition 5 (Witnesses)** *For a set of skeletons  $\mathbb{K}$  defined over  $X$ , let  $w : \mathbb{K} \times X \rightarrow \{0, 1\}$  be a witness function. If  $w(P, x) = 1$  then we say that  $x$  is a witness for  $P$ .*

Intuitively, skeletons are representatives of partitions (clusterings) of almost all points in  $X$ , where each part  $S_i$  in the partition is a subset of a different cluster  $X_i$ . For example, we may let a skeleton simply be a partition of at most  $k$  points into singletons, where each point represents a different potential cluster. As to witnesses, intuitively, they are points that provide evidence to imperfections of skeletons. In our example, a witness may simply be a point that is at distance greater than  $b$  from every skeleton point.

**Definition 6 (Good Subsets)** *A subset  $U \subseteq X$  is  $\alpha$ -good with respect to set of skeletons  $\mathbb{K}$ , if there exists a skeleton  $P \in \mathbb{K}$  such that  $\text{Set}(P) \subseteq U$  and there are at most  $\alpha \cdot |U|$  points  $u \in U$  such that  $w(P, u) = 1$  (that is,  $u$  is a witness for  $P$ ). Otherwise the subset  $U$  is  $\alpha$ -bad with respect to  $\mathbb{K}$ .*

The following Lemma will allow us to prove the correctness of Algorithm 1 under the condition that there exists a set of skeletons with certain properties.

**Lemma 2** *Let  $\mathbb{K}$  be a set of skeletons defined over  $X$ , having order  $s$ , and let  $\alpha, \gamma \in (0, 1)$ . Suppose that every skeleton in  $\mathbb{K}$  has at least  $(\alpha + \gamma) \cdot |X|$  points  $x \in X$  that are witnesses for it. Consider selecting, uniformly and independently,  $m = \tilde{\Theta}(s(\log k + 1)/\gamma^2)$  points from  $X$ . Then with probability at least  $2/3$ , the (multi)set of points selected is  $\alpha$ -bad with respect to  $\mathbb{K}$ .*

The next theorem proves that Algorithm 1 is fully tolerant, assuming that there exists a set of skeletons  $\mathbb{K}$  that satisfies the two conditions that are specified in the theorem. The theorem also requires the sample size  $m$  taken by the algorithm to be some large enough function of  $1/(\epsilon_2 - \epsilon_1)$  and of the order  $s$  of  $\mathbb{K}$ .

**Theorem 3** *Suppose that  $\mathbb{K}$  is a set of skeletons with order  $s$  that satisfies the following conditions for any given  $0 \leq \alpha \leq 1$ :*

1. *If  $X$  is  $\alpha$ -far from being  $(k, b(1 + \beta))$ -clusterable with respect to  $C(\cdot)$ , then every skeleton in  $\mathbb{K}$  has at least  $\alpha \cdot |X|$  witnesses for it in  $X$ .*
2. *If a subset  $U \subseteq X$  is  $\alpha$ -bad with respect to  $\mathbb{K}$ , then  $U$  is  $\alpha$ -far from being  $(k, b)$ -clusterable with respect to  $C(\cdot)$ .*

*Let  $m = \tilde{\Omega}(s(\log k + 1)/\delta^2)$  where  $\delta = \epsilon_2 - \epsilon_1$ . If  $X$  is  $\epsilon_2$ -far from being  $(k, b(1 + \beta))$ -clusterable with respect to  $C(\cdot)$ , then Algorithm 1 rejects with probability at least  $2/3$ .*

(Recall that by Lemma 1, if  $C(\cdot)$  is hereditary, then, since  $m = \Omega(1/\delta^2)$ , we know that if  $X$  is  $\epsilon_1$ -close to being  $(k, b)$ -clusterable with respect to  $C(\cdot)$ , then Algorithm 1 accepts with probability at least  $2/3$ .)

**Proof:** Suppose that  $X$  is  $\epsilon_2$ -far from being  $(k, b(1 + \beta))$ -clusterable with respect to  $C(\cdot)$ . By the first item in the premise of the theorem, every skeleton in  $\mathbb{K}$  has at least  $\epsilon_2 \cdot |X|$  witnesses for it in  $X$ . By Lemma 2, if we set  $\alpha = \epsilon_2 - \delta/2$  and  $\gamma = \delta/2$ , then with probability  $2/3$  over the choice of the sample  $U$ , the set  $U$  is  $(\epsilon_2 - \delta/2)$ -bad. But by the second item in the premise of the theorem, for each such  $U$ , the set  $U$  is  $(\epsilon_2 - \delta/2)$ -far from being  $(k, b)$ -clusterable. Since  $\epsilon_2 - \delta/2 = \epsilon_1 + \delta/2$ , each such  $U$  would cause Algorithm 1 to reject, and the second part of the theorem follows. ■

Note that in order to use this theorem to prove the correctness of Algorithm 1 for specific cost measures, such as the diameter cost measure, we have to define a skeleton set  $\mathbb{K}$  and witnesses for the specific cost measure at hand. Furthermore, we must prove that the two conditions in Theorem 3 hold, and that the order  $s$  of  $\mathbb{K}$  is bounded so that the sample size  $m$  will not be too large.

## 2.2 Clustering Under a General Metric

In this subsection we show how to apply Theorem 3 so as to obtain the following theorem for the diameter cost. Recall that  $\delta = \epsilon_2 - \epsilon_1$ .

**Theorem 4** *Let  $C(\cdot)$  be the diameter cost, and let  $\text{dist}(\cdot, \cdot)$  be any underlying distance function that obeys the triangle inequality. Suppose that we run Algorithm 1 with  $m = \tilde{\Theta}(k/\delta^2)$ . If  $X$  is  $\epsilon_1$ -close to being  $(k, b)$ -clusterable, then with probability at least  $2/3$  Algorithm 1 accepts  $X$ , while if  $X$  is  $\epsilon_2$ -far from being  $(k, 2b)$ -clusterable then with probability at least  $2/3$  it rejects  $X$ .*

We note that it was shown in [ADPR] that even for standard testing, it is not possible to go below the factor of  $(1 + \beta) = 2$  without incurring a dependence on  $n$  (specifically,  $\sqrt{n}$ ). To see why this is true, consider the following two metrics: in one, all  $n$  points are at distance 1 from each other. In the second metric, all pairs of points are at distance 1 from each other with the following exception: every point has a unique “partner” that it is at distance 2 from (this still obeys the triangle inequality). If we set  $k = 1$  and  $b = 1$ , then the first set of points is  $(k, b)$  clusterable. On the other hand, for every  $\beta < 1$ , the second set of points is  $(1/2)$ -far from being  $(k, b(1 + \beta))$  clusterable. This is true since for any set of more than  $n/2$  points, there exists a pair in the set that are at distance  $2 > b(1 + \beta)$ .

In order to apply Theorem 3 we define skeletons and witnesses as follows:

**Definition 7 (Skeletons and Witnesses for General Metrics)** *A skeleton  $P$  is a partition of a subset  $S \subseteq X$ ,  $|S| \leq k$  into singletons. A point  $x \in X$  is a witness for a skeleton  $P$ , if it is at distance greater than  $b$  from every point in  $S$ .*

Since a partition  $P$  is uniquely defined by the set  $S = \text{Set}(P)$ , we simply use  $S$  to denote a skeleton. Let  $\mathbb{K}$  denote the set of all skeletons over  $X$  as described in Definition 7 (so that  $\mathbb{K}$  has order  $k$ ).

The next two lemmas establish that the two items in Theorem 3 hold for  $\mathbb{K}$  as defined above. Theorem 4 directly follows using the fact that  $s = k$ .

**Lemma 5** *Let  $0 \leq \alpha \leq 1$ . If  $X$  is  $\alpha$ -far from being  $(k, 2b)$ -clusterable then every skeleton in  $\mathbb{K}$  has at least  $\alpha|X|$  witnesses in  $X$ .*

**Proof:** Assume, contrary to the claim, that there exists a skeleton  $S$  with less than  $\alpha|X|$  witnesses in  $X$ . Consider the subset  $Y \subset X$  that consists of all points in  $X$  that are not witnesses for  $S$ . By the definition of skeletons and witnesses in this case, each point in  $Y$  is at distance at most  $b$  from some point in  $S$ . We can now assign each point in  $Y$  to the closest point in the skeleton  $S$ . All points that are assigned to the same skeleton point must be at distance at most  $2b$  from each other (using the triangle inequality). Thus  $Y$  is  $(k, 2b)$ -clusterable, and has size at least  $(1 - \alpha)|X|$ . But this means that  $X$  is  $\alpha$ -close to being  $(k, 2b)$ -clusterable, and we have reached a contradiction, as desired. ■

**Lemma 6** *Let  $0 \leq \alpha \leq 1$ . If a subset  $U \subseteq X$  is  $\alpha$ -bad with respect to  $\mathbb{K}$ , then  $U$  is  $\alpha$ -far from being  $(k, b)$ -clusterable.*

**Proof:** Assume, contrary to the claim, that  $U$  is  $\alpha$ -close to being  $(k, b)$ -clusterable. We will show that  $U$  is  $\alpha$ -good (thus reaching a contradiction) by presenting a skeleton  $S \subseteq U$  with at most  $\alpha|U|$  witnesses in  $U$  with respect to  $S$ .

Since  $U$  is  $\alpha$ -close to being  $(k, b)$ -clusterable, there exists a subset  $W \subseteq U$ , of size at least  $(1 - \alpha)|U|$  that can be clustered into  $k$  clusters of diameter at most  $b$ . We now choose a point from each cluster in  $W$  to obtain a skeleton  $S \subseteq W \subseteq U$ . The skeleton  $S$  has at most  $\alpha|U|$  witnesses in  $U$  (that is, the points that do not belong to  $W$ ), as claimed. ■

### 2.3 Clustering Under the Euclidean Metric

In this subsection we consider the case that the set of points  $X$  lies in the Euclidean space and the underlying distance function is the Euclidean distance. The cost measure  $C(\cdot)$  is the diameter cost.

**Theorem 7** Let  $C(\cdot)$  be the diameter cost, let  $X \subset \mathbb{R}^d$  for some integer  $d$ , and let the underlying distance  $\text{dist}(\cdot, \cdot)$  be the Euclidean distance between points. Then for any given  $0 < \beta \leq 1$ , if we run Algorithm 1 with  $m = \tilde{\Theta}(k \cdot \delta^{-2} \cdot (1 + (2/\beta))^d)$ , then with probability at least  $2/3$  it accepts  $X$  if  $X$  is  $\epsilon_1$ -close to being  $(k, b)$ -clusterable, and with probability at least  $2/3$  it rejects  $X$  if  $X$  is  $\epsilon_2$ -far from being  $(k, b(1 + \beta))$ -clusterable.

We note that it was shown in [ADPR], that for  $\beta < 1$ , a dependence on  $1/\beta$ , as well as an exponential dependence on the dimension  $d$ , are unavoidable, even for the special case of  $\epsilon_1 = 0$  (i.e., standard testing). In order to prove Theorem 7 we apply Theorem 3, but we shall need slightly more sophisticated notions of skeletons and witnesses than those used in the previous subsection.

**Definition 8 (Intersections of Balls)** For any subset  $Y \subseteq X$  let  $I(Y)$  denote the intersection of all  $d$ -dimensional balls of radius  $b$  centered at the points in  $Y$ . If  $Y = \emptyset$  then  $I(Y) = \mathbb{R}^d$ .

**Definition 9 (Violating and Influential Points)** Let  $Y \subseteq X$ , such that  $I(Y) \neq \emptyset$ . A point  $x \in X$  is violating for  $Y \subseteq X$  if  $x \notin I(Y)$ . The point  $x$  is influential with respect to  $Y$  if  $x \in I(Y)$  and for every  $y \in Y$  it holds that  $\text{dist}(x, y) > \beta b$ .

**Definition 10 (Skeletons and Witnesses for the Euclidean Distance)** Skeletons are defined inductively as follows:

1. The  $k$ -partition  $P = \{\emptyset, \dots, \emptyset\}$  is a skeleton (that is, all  $k$  parts are empty and  $\text{Set}(P) = \emptyset$ ).
2. If  $P = \{S_1, \dots, S_k\}$  is a skeleton and  $x \in X \setminus \text{Set}(P)$  is an influential point with respect to  $S_i$  for some  $1 \leq i \leq k$ , then  $P' = \{S_1, \dots, S_{i-1}, S_i \cup \{x\}, S_{i+1}, \dots, S_k\}$  is a skeleton. (Note that there may be more than one way to add  $x$  to the skeleton  $P$ .)

A point  $x \in X$  is a witness for a skeleton  $P = \{S_1, \dots, S_k\}$ , if for every  $1 \leq i \leq k$  the point  $x$  is either violating or influential with respect to  $S_i$ .

Here too we denote by  $\mathbb{K}$  the set of all skeletons defined over  $X$  as in Definition 10. Clearly, for every set  $S$ , if  $|S| \leq s$ , then the number of partitions  $P \in \mathbb{K}$  such that  $\text{Set}(P) = S$  is bounded by  $k^s$  (where this upper bound will suffice for our purposes). The following lemma, which bounds the size  $s$  of the sets  $S$  that participate in skeletons, is from [CS].

**Lemma 8** Let  $P = \{S_i\}_{i=1}^k$  be a skeleton in  $\mathbb{K}$ . Then  $|\text{Set}(P)| \leq k(1 + (2/\beta))^d$ .

Lemmas 9 and 10, stated and proved below, establish the two items in Theorem 3 for the set of skeletons  $\mathbb{K}$  as defined above. Theorem 7 readily follows (using lemma 8).

**Lemma 9** Let  $0 \leq \alpha \leq 1$ . If  $X$  is  $\alpha$ -far from being  $(k, b(1 + \beta))$ -clusterable, then every skeleton in  $\mathbb{K}$  has at least  $\alpha|X|$  witnesses in  $X$ .

**Proof:** Assume, contrary to the claim, that there exists a skeleton  $P = \{S_i\}_{i=1}^k$  with less than  $\alpha|X|$  witnesses in  $X$ . Let  $Z \subseteq X$  be the subset of all points that are not witnesses with respect

to  $P$ . Hence  $|Z| > (1 - \alpha)|X|$ . We next show that  $Z$  is  $(k, b(1 + \beta))$ -clusterable, and reach a contradiction to the premise of the lemma.

For each  $z \in Z$  there exists an index  $i$ ,  $1 \leq i \leq k$  such that  $z$  is not violating and non-influential with respect to  $S_i$ . We assign such a point  $z$  to the  $i$ 'th cluster. Note that, in particular, all points in  $S_i$  are assigned to the  $i$ 'th cluster. We next show that the distance between any two points in the  $i$ 'th cluster is at most  $(1 + \beta)b$ .

First observe that all points in  $S_i$  are at distance at most  $b$  from each other. This holds by the construction of skeletons in Definition 10, since a point  $x$  can be added to  $S_i$  only if it is influential with respect to  $S_i$ , which in particular requires that  $x \in I(S_i)$ . As to points  $z \in Z \setminus S_i$  that were assigned to the  $i$ 'th cluster: note that any such point  $z$  is non-violating and non-influential with respect to  $S_i$ . Hence  $z \in I(S_i)$  (that is,  $\text{dist}(z, x) \leq b$  for every  $x \in S_i$ ) and there exists a point  $x_z \in S_i$  such that  $\text{dist}(z, x_z) \leq \beta b$ . Consider any other point  $y$  that belongs to the  $i$ 'th cluster. If  $y \in S_i$ , then, since  $z \in I(S_i)$  we have that  $\text{dist}(z, y) \leq b$ . Otherwise, by the triangle inequality we get that  $\text{dist}(z, y) \leq \text{dist}(z, x_z) + \text{dist}(x_z, y) \leq \beta b + b = (1 + \beta)b$ . Thus  $Z$  is  $(k, b(1 + \beta))$ -clusterable, as claimed. ■

**Lemma 10** *Let  $0 \leq \alpha \leq 1$ . If a subset  $U \subseteq X$  is  $\alpha$ -bad with respect to  $\mathbb{K}$ , then  $U$  is  $\alpha$ -far from being  $(k, b)$ -clusterable.*

**Proof:** Assume, contrary to the claim, that  $U$  is  $\alpha$ -close to being  $(k, b)$ -clusterable. That is, there exists a subset  $W \subseteq U$ ,  $|W| \geq (1 - \alpha)|U|$ , and a partition,  $\{W_i\}_{i=1}^k$ , such that the diameter of each  $W_i$  is at most  $b$ .

We next prove that there exists a skeleton  $P = \{S_i\}_{i=1}^k$ ,  $S_i \subseteq W_i \subseteq U$ , with at most  $\alpha|U|$  witnesses in  $U$ . We will build  $P$  in the following iterative manner:

1. We start with the skeleton  $P^0 = \{\emptyset, \dots, \emptyset\}$ .
2. Let  $P^j = \{S_i^j\}_{i=1}^k$  be the skeleton at the beginning of the  $j$ 'th iteration. If there exists an index  $i$  and a point  $x \in W_i$  that is an influential point with respect to  $S_i^j$ , then we let  $P^{j+1} = \{S_1^j, \dots, S_{i-1}^j, S_i^j \cup \{x\}, S_{i+1}^j, \dots, S_k^j\}$ .
3. When for every  $i$ , the subset  $W_i$  does not contain any influential points with respect  $S_i^j$  then we stop.

Let  $P = \{S_i\}_{i=1}^k$  be the final resulting skeleton. Notice that for every  $i$ , the subset  $W_i$  does not contain a violating point with respect to  $S_i \subseteq W_i$ , because all points in  $W_i$  are at distance at most  $b$  from each other. Also when we finish building  $P = \{S_i\}_{i=1}^k$ , the subset  $W_i$  contains no influential points with respect to  $S_i$ . Thus all points in  $W_i$  are not witnesses with respect to  $P$ . Since there are at most  $\alpha|U|$  points in  $U$  that do not belong to any cluster  $W_i$ , then there are at most  $\alpha|U|$  witnesses with respect to  $P$ . ■

## 2.4 Finding Approximately Good Clusterings

As stated earlier, our tolerant testing algorithm for clustering can be modified to obtain an algorithm that outputs an approximately good clustering of most input points. More precisely, for any given parameters  $\epsilon_1 < \epsilon_2$ , if the set of points  $X$  is  $\epsilon_1$ -close to being  $(k, b)$ -clusterable, then with high

probability the modified algorithm outputs an implicit representation of a  $(k, b(1 + \beta))$ -clustering of all but an  $\epsilon_2$ -fraction of the points of  $X$ . In an “implicit representation” we mean a partition of a small subset of the points that can be used to determine the cluster to which each point in  $X$  belongs to (but at most  $\epsilon_2|X|$  of the points). Specifically, we prove:

**Theorem 11** *Suppose that  $\mathbb{K}$  is a set of skeletons with order  $s$  that satisfies the following conditions:*

1. *For any given skeleton  $P \in \mathbb{K}$ , the subset of all points in  $X$  that are not witnesses for  $P$  is  $(k, b(1 + \beta))$ -clusterable. Furthermore, given a skeleton  $P$  and a point  $x$  that is not a witness for  $P$ , it is possible to determine to which cluster  $x$  belongs without looking at any points outside of  $\text{Set}(P) \cup \{x\}$ .*
2. *For any given  $0 \leq \alpha \leq 1$ , if a subset  $U \subset X$  is  $\alpha$ -close to being  $(k, b)$ -clusterable with respect to  $C(\cdot)$ , then  $U$  is  $\alpha$ -good with respect to  $\mathbb{K}$ .*

*Then the following holds for Algorithm 2 specified below: for any given parameters  $\epsilon_1 < \epsilon_2$ , if  $X$  is  $\epsilon_1$ -close to being  $(k, b)$ -clusterable with respect to  $C(\cdot)$ , then with high constant probability Algorithm 2 outputs an implicit representation of a  $(k, b(1 + \beta))$ -clustering of all but at most an  $\epsilon_2$ -fraction of the points in  $X$ . By an “implicit representation” we mean a skeleton in  $\mathbb{K}$  that can be used to determine the cluster to which each point in  $X$  belongs to (but at most  $\epsilon_2|X|$  of the points).*

The running time of Algorithm 2 depends on the time required to determine if the subset  $U$  sampled by the algorithm is  $(\epsilon_1 + \delta/2)$ -close to being  $(k, b)$ -clusterable, and if so to find a skeleton  $P \in \mathbb{K}$  such that  $\text{Set}(P) \subseteq U$  and there are at most  $(\epsilon_1 + \delta/2) \cdot |U|$  witnesses to  $P$  in  $U$ . Hence the running time depends on the definitions of skeletons and witnesses for the specific cost measure. Similarly, the time required to determine if a point  $x$  is a witness with respect to the skeleton  $P$  (implicit representation) that is output by the algorithm, and if not, to which cluster it belongs, is also problem-dependent.

**Algorithm 2** (Approximate Clustering Algorithm given a set of skeletons  $\mathbb{K}$ )

1. Uniformly and independently select  $m = \tilde{\Theta}(s(\log k + 1)/\delta^2)$  points from  $X$ , where  $\delta = \epsilon_2 - \epsilon_1$ . Let  $U$  denote the (multi)set of points selected.
2. If  $U$  is  $(\epsilon_1 + \delta/2)$ -close to being  $(k, b)$ -clusterable with respect to  $C(\cdot)$  then do:
  - Find a skeleton  $P \in \mathbb{K}$  such that  $\text{Set}(P) \subseteq U$  and such that there exist at most  $(\epsilon_1 + \delta/2) \cdot |U|$  witnesses to  $P$  in  $U$ .
  - Output  $P$ .
3. Else output *Fail*.