Testing of Clustering *

Noga Alon†
Department of Mathematics
Tel-Aviv University
Ramat Aviv, ISRAEL
noga@math.tau.ac.il

Seannie Dar
The Academic College
of Tel-Aviv-Yaffo
Tel-Aviv, ISRAEL

Michal Parnas
The Academic College
of Tel-Aviv-Yaffo
Tel-Aviv, ISRAEL
michalp@mta.ac.il

Dana Ron
Department of EE – Systems
Tel-Aviv University
Ramat Aviv, ISRAEL
danar@eng.tau.ac.il

Abstract

A set X of points in $\Re^d$ is (k, b)-*clusterable* if X can be partitioned into k subsets (*clusters*) so that the diameter (alternatively, the radius) of each cluster is at most b. We present algorithms that by sampling from a set X, distinguish between the case that X is (k, b)-clusterable and the case that X is ϵ-far from being (k, b')-clusterable for any given $0 < \epsilon \leq 1$ and for $b' \geq b$. In ϵ-far from being (k, b')-clusterable we mean that more than $\epsilon \cdot |X|$ points should be removed from X so that it becomes (k, b')-clusterable. We give algorithms for a variety of cost measures that use a sample of size independent of $|X|$, and polynomial in k and $1/\epsilon$.

Our algorithms can also be used to find *approximately good* clusterings. Namely, these are clusterings of all but an ϵ-fraction of the points in X that have optimal (or close to optimal) cost. The benefit of our algorithms is that they construct an *implicit representation* of such clusterings in time independent of $|X|$. That is, without actually having to partition all points in X, the implicit representation can be used to answer queries concerning the cluster any given point belongs to.

*An extended abstract of this work will appear in the proceedings of FOCS 2000.

†Research supported in part by a USA Israeli BSF grant, by a grant from the Israel Science Foundation and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

1 Introduction

Clustering problems arise in many areas, and have a variety of applications (*cf.* [6, 24, 34, 26, 11, 39, 28, 25]). In one of its standard forms, the problem is to decide whether a given set X of n points in d-dimensional Euclidean space can be partitioned into k subsets (*clusters*) so that the *cost* of each cluster is at most b. The cost of a cluster can be defined as its *diameter* (i.e, the maximum distance between pairs of points in the cluster), or its *radius* (that is, the minimum radius of a ball containing all the points in the cluster).[1] If such a k-way partition exists, then we say that X is (k, b)-*clusterable* (with respect to the diameter or the radius cost). Unfortunately, both decision problems are NP-Complete for $d \geq 2$ (and variable k) [14, 29], and remain hard even when only a certain constant approximation of the cluster size is sought [13].

In this work we consider the following relaxation of the above decision problems: For a given approximation parameter $\beta \geq 0$, and distance parameter $0 \leq \epsilon \leq 1$, we would like to determine whether the set X is (k, b)-clusterable or ϵ-*far* from being $(k, (1+\beta)b)$-clusterable. In ϵ-*far* from $(k, (1+\beta)b)$-clusterable we mean that more than an ϵ fraction of the points in X should be removed (or moved) so that X becomes $(k, (1+\beta)b)$ clusterable. Given this relaxation of the decision problem, we seek algorithms that will be significantly faster than those required for solving the exact decision problems. In particular, we ask that our algorithms observe as few points as possible from X, and run in time sub-linear in $n = |X|$, or even independent of n.

We refer to algorithms that perform such relaxed (approximate) decision tasks as *testing* algorithms: They are required to output accept if X is (k, b)-clusterable, and to output reject with probability at least $2/3$, if X is ϵ-far from $(k, (1+\beta)b)$-clusterable. (If neither holds, the testing algorithm may output either accept or reject). Such testing algorithms can be useful as an alternative to an exact or even approximate decision procedure when the number of points n is very large. Even if n is not too large and there is time to run a clustering algorithm on all the points, testing can be applied as a preliminary step in order to approximate the quality of the best achievable clustering.

Our Results. We present and analyze testing algorithms both for the diameter cost and the radius cost. All our algorithms run in time *independent* of $n = |X|$, and use a sample from X that has size polynomial in k and ϵ.

We describe algorithms for the L_2 metric (Euclidean), as defined above, which in the case of the radius cost easily extend to other metrics (such as L_∞). We also give algorithms that work under any metric, for $\beta = 1$. With the exception of our algorithms for general metrics, all our algorithms have the following form: They uniformly select a sample of points from X and run an exact decision procedure for verifying whether the sample is (k, b)-clusterable. Specifically, we show:

1. For general metrics we give algorithms that work for $\beta = 1$. For both costs, the sample selected is of size $\tilde{O}(k/\epsilon)$, and the running time is $\tilde{O}(k^2/\epsilon)$. We also observe that any algorithm for testing diameter clustering for $\beta < 1$ (under a general metric), requires a sample of size $\Omega(\sqrt{n/\epsilon})$.

2. For the L_2 metric and the diameter cost, the sample is of size $\tilde{O}\left(\frac{k^2}{\epsilon} \cdot \left(\frac{2}{\beta}\right)^{2d}\right)$. A dependence on $1/\beta$ as well as an exponential dependence on the dimension are unavoidable: We prove

[1] The first problem is also known as *center* clustering (since all points in a given cluster are at distance at most b from the center of the bounding ball), and the second as *pairwise* clustering.

a lower bound of $\Omega(\beta^{-(d-1)/4})$ on the size of the sample required for testing (for $k = 1$ and constant ϵ).

3. For the L_2 metric and the radius cost, the algorithm works for $\beta = 0$ and the sample size is $\tilde{O}\left(\frac{d \cdot k}{\epsilon}\right)$.

In Items 2 and 3 we only stated the size of the sample selected by the algorithms. The running times depend on the exact decision procedures applied, and given the difficulty of the problems is exponential in k and d.

In addition to the above, our algorithms can be used to obtain approximately good clusterings. A k-way partition P is an *ϵ-good* (k, b')-clustering of X if it is a partition having cost at most b' of all but at most an ϵ fraction of the points in X. If X is (k, b)-clusterable, then using our testing algorithms, it is possible to obtain in time independent of n an *implicit representation* of an ϵ-good $(k, (1+\beta)b)$-clustering of X. Namely, given this implicit representation we can determine for any given point $x \in X$, the cluster it belongs to. This can be done in time $O(k)$ per point (or even $O(\log k)$, depending on the cost measure). For example, in the case of radius clustering, the implicit representation is simply a set of k cluster centers. The benefit of such an implicit representation is that it allows to answer queries of the form: "do points $x, y \in X$ belong to the same cluster?" without actually having to partition all points. This approach was previously applied in [17] to graph partitioning problems and a related approach was applied in [15].

Independently from our work, Mishra, Oblinger and Pitt [30] study the problem of approximately good clustering when the cost measure is the sum of distances (or distances squared) to the cluster centers. Their algorithms use a sample of size independent of n and polynomial in $1/\epsilon$, d, k and M, where the points belong to $[0, M]^d$.

Techniques. The following approach is a common thread passing through the analysis of most of our algorithms. Recall that our algorithms work by sampling from X. The sample is viewed as being selected in phases, where we show that with high probability, in each phase certain *progress* is made. In particular, in case X is ϵ-far from being $(k, (1+\beta)b)$-clusterable, this progress leads to rejection after a bounded number of phases. For example, in the case of the diameter cost and a single cluster ($k = 1$), progress is measured in terms of reducing the volume of the region in $\Re^d$ which contains all points having distance at most b from every sample point.

For the radius cost under the L_2 metric, our analysis uses *ϵ-nets* and their relation to the VC-dimension of families of sets. This relation was previously exploited both in the context of learning and in the context of computational geometry.

Perspective. In this paper we approach the problem of clustering from within the framework of Property Testing [37, 17]. In property testing the goal is to decide whether a given object (e.g., graph or function) has a predetermined property (e.g., connectivity or monotonicity), or is *far* from having the property. The notion of being far from having the property depends on the type of object considered. For example, if the object is a graph, then we say that it is far from having a particular property if many edge modifications should be made so that it obtains the property. In "many" we mean at least a certain fraction ϵ of all edges in the graph.

Previous work in property testing has mainly dealt with properties of functions [8, 37, 36, 12, 27, 16, 9] and properties of graphs [17, 18, 19, 3, 32, 4, 7]. Recently, Alon *et. al.* [5] gave a testing algorithm for membership of strings in regular languages, and Ergun *et. al.* [12] study problems related to testing convexity in two dimensions.

Here we further extend the scope of property testing to the domain of clustering problems. Our proof techniques combine geometric analysis with probabilistic analysis that is characteristic of work in property testing. We thus hope to enrich both areas of research.

Other Related Work. Hochbaum and Shmoys [23] were the first to show that it is hard to approximate the cost of an optimal clustering to within a factor of 2 for a general distance function. They also give a 2-approximation algorithm for the problem [22, 23]. As noted above, Feder and Greene [13] show that constant approximation is also hard for L_2 and L_∞ metrics (where the specific constants depend on the metric and cost measure used). An approximation factor of 2 can be achieved efficiently for all geometric variants we consider [20, 13]. For the radius cost, and under both L_∞ and L_2 metrics, Agrawal and Procopiuc [1] give an algorithm for finding a clustering having cost at most $(1+\beta)$ larger than optimal in time $O\left(n \cdot \log k + (k/\beta)^{O(d^2 k^{1-1/d})}\right)$. For more information on clustering, see [10, 2, 33] and references within.

Organization. In Section 2 we introduce the notations and definitions that are used in the paper. In Section 3 we discuss testing when the underlying distance function is a general metric. In Sections 4 and 5, we present the algorithms that work under the L_2 metric, for the diameter cost and the radius cost, respectively. In Section 4 we also give our lower bound for the diameter cost. Results for one dimension are given in Section 6.

2 Preliminaries

We denote by $\mathrm{dist}(x,y)$ the distance between two points x, y. We follow the standard practice of assuming that the distance between a pair of points can be computed in constant time. Since most of this paper deals with the L_2 metric, in what follows we refer to the Euclidean distance. Thus, if $x, y \in \Re^d$, that is $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$, then the Euclidean distance between x and y is $\mathrm{dist}(x,y) \stackrel{\mathrm{def}}{=} \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$.

Given a subset $S \subseteq X$, denote by $d(S)$ the *diameter* of S, that is, the maximum distance between any two points in S. Denote by $r(S)$ the *radius* of the smallest ball containing S. That is, $r(S) = \min_{y \in \Re^d} \max_{x \in S} \mathrm{dist}(x,y)$. The point $y \in \Re^d$ for which the minimum radius is achieved is the *center* of the *minimum bounding ball* of S.

Let $P = (X^1, \ldots, X^k)$ be a partition of X. The *diameter* of the partition P is defined as $D(P) = \max_j d(X^j)$. The *radius* of P is defined as $R(P) = \max_j r(X^j)$.

For such a k-way partition P of X, we consider the following two cost measures:

1. Diameter Cost: $Cost(P) = D(P)$.
2. Radius Cost: $Cost(P) = R(P)$.

Hence, a set X is (k,b)-clusterable according to one of the above cost measures if there *exists* a k-way partition $P = (X^1, \ldots, X^k)$ of X such that $Cost(P) \leq b$. The set is ϵ-*far* from being $(k, (1+\beta)b)$-clusterable for a given $0 \leq \epsilon \leq 1$ and $\beta \geq 0$, if for *every* subset $Y \subseteq X$ of size at most $(1-\epsilon)|X|$, and for every k-way partition $P_Y = (Y^1, \ldots, Y^k)$ of Y, we have $Cost(P_Y) > (1+\beta)b$.

Since all our algorithms have one-sided error, we shall use the following definition of testing algorithms for clustering. We say that an algorithm is a *Diameter-Clustering* (*Radius-Clustering*)

Tester if given access to points in a set $X \subset \Re^d$ and parameters k, b, ϵ, and β, the algorithm accepts X if it is (k, b)-clusterable with respect to the diameter cost (radius cost), and rejects X with probability at least 2/3 if it is ϵ-far from being $(k, (1+\beta)b)$-clusterable.

3 Testing of Clustering Under General Metrics

We begin by describing an algorithm for testing diameter clustering when the underlying distance function is any metric and $\beta = 1$. The algorithm distinguishes between the case in which X is (k, b)-clusterable, and the case in which X is ϵ-far from $(k, 2b)$-clusterable, under the assumption that the distances between points in X obey the triangle inequality. The basic idea of the algorithm is to try and find points in X that are *representatives* of different clusters. That is, their pairwise distances are greater than the allowed diameter b. In case X is (k, b)-clusterable then there can be at most k such representatives. On the other hand, as we show in the analysis of the algorithm, if X is ϵ-far from (k, b)-clusterable, then with probability at least 2/3 the algorithm will find $k+1$ such representatives. The algorithm for the radius cost as well as its analysis, are very similar. Furthermore, a certain refinement of this idea serves as a basis for the analysis of some of our other algorithms.

Algorithm 1 (general metric, diameter cost, $k \geq 1$, $\beta = 1$)

1. *Let* rep_1 *be an arbitrary point in* X *(a representative for the first cluster).*
2. $i \leftarrow 1$*;* find-new-rep $\leftarrow$ TRUE*.*
3. *while* $i < k+1$ *and* find-new-rep = TRUE *do*
   (a) *Uniformly and independently select a sample of size* $\ln(3k)/\epsilon$*.*
   (b) *If there exists a point* x *in the sample, such that* $\text{dist}(x, \text{rep}_j) > b$ *for every* $j \leq i$*, then* $i \leftarrow i+1$ *and* $\text{rep}_i = x$*.*
   (c) *Else (all points in the sample are at distance at most* b *from some* rep_j*),* find-new-rep $\leftarrow$ FALSE*.*
4. *If* $i \leq k$ *then accept, otherwise, reject.*

Since there are at most k iterations of the while loop, and in each iteration the algorithm computes at most $k \cdot \ln(3k)/\epsilon$ distances, the running time of the algorithm is $O(k^2 \log k/\epsilon)$

Theorem 1 *Algorithm 1 is a diameter-clustering tester for* $\beta = 1$ *under any metric.*

Proof: We first observe that the algorithm rejects only if it finds $k+1$ points whose pairwise distances are all greater than b. Therefore, if X is (k, b)-clusterable, then the algorithm never rejects. Hence, from now on assume that X is ϵ-far from $(k, 2b)$-clusterable, and we show that the algorithm rejects with probability at least 2/3. That is, we show that with probability at least 2/3, in every iteration a new representative is found, resulting in $k+1$ representatives.

Consider any particular iteration. We say that a point $x \in X$ is a *candidate representative* with respect to $\text{rep}_1, \ldots, \text{rep}_i$ if it has distance greater than b from each of these points. We claim that as long as $i \leq k$, there must be more than ϵn such candidate representatives. Assume in contradiction

that there are at most ϵn such points. Then we could remove these points from X, and for every other point $y \in X$, assign y to a cluster j such that $\mathrm{dist}(y, \mathrm{rep}_j) \leq b$. By the triangle inequality, the diameter of each resulting cluster is at most $2b$, which contradicts our assumption concerning X. Now, if we uniformly select $\ln(3k)/\epsilon$ points from X, then the probability that no candidate representative is selected is less than $(1-\epsilon)^{\ln(3k)/\epsilon} < \exp(-\epsilon(\ln(3k)/\epsilon) = \frac{1}{3k}$. The probability that this occurs in *any* iteration is at most $1/3$, and the theorem follows. ■

Finding an approximately good clustering. Suppose X is (k, b)-clusterable and so the algorithm terminates with at most k representatives $\mathrm{rep}_1, \ldots, \mathrm{rep}_i$. By the above analysis, with probability at least $2/3$, the algorithm does not terminate as long as the number of candidate representatives with respect to $\mathrm{rep}_1, \ldots, \mathrm{rep}_i$ is greater than ϵn. This implies that with probability at least $2/3$, the final representatives have the following property: They define an implicit representation of a partition having diameter at most $2b$ of all but at most an ϵ-fraction of the points in X. That is, excluding the at most ϵn points that are candidate representatives (i.e., that are at distance greater than b from the at most k representative $\mathrm{rep}_1, \ldots, \mathrm{rep}_i$), every other point $x \in X$ can be assigned to some cluster j for which $\mathrm{dist}(x, \mathrm{rep}_j) \leq b$. The time required to find the cluster a given point belongs to is $O(k)$.

A Lower Bound for $\beta < 1$. If all that is known about the distance function between points in X is that it obeys the triangle inequality, then the above result is tight in the following sense. It is not possible to test for diameter clustering for $\beta < 1$ using a sample of size independent of n or even of size $o(\sqrt{n})$. To see why this is true consider a metric that is defined by a complete graph on $N = 2n$ vertices with the following weights (distances) on the edges. There exists a perfect matching between the vertices such that each edge in the matching has weight 2, and every other edge has weight 1. If X corresponds to any subset of size n of the vertices such that no two vertices in X are matched, then X is $(1, 1)$-clusterable. On the other hand, if X contains more than ϵn pairs of matched vertices, then it is ϵ-far from $(1, 2-\delta)$-clusterable for any $\delta > 0$. However, in order to distinguish between the two cases with non-negligible probability, the algorithm has to sample $\Omega(\sqrt{n/\epsilon})$ vertices.

Testing Radius Clustering Under General Metrics. The algorithm for radius clustering is the same as Algorithm 1 except that a point is selected as a new representative only if it is at distance greater than $2b$ from each representative selected so far. By the triangle inequality, if X is (k, b)-clusterable then there can be at most k representatives. On the other hand, if X is ϵ-far from $(k, 2b)$-clusterable, then as long as $i \leq k$ there must be more than ϵn candidate representatives (as the previous representatives can serve as cluster centers). Hence the analysis of the radius-clustering algorithm follows the same lines as that of the diameter-clustering algorithm. Furthermore, as in the case of diameter clustering, here too we can use the representatives found by the algorithm to induce an approximately good clustering of X.

4 Testing of Diameter Clustering Under the L_2 Metric

4.1 The case $k = 1$

We start by studying the problem of testing for a single cluster. In the next subsection we generalize the algorithm presented here and its analysis to any number of clusters k.

Algorithm 2 (L_2 metric, diameter cost, $k = 1$, $d \geq 1$, $0 < \beta \leq 1$)

1. *Uniformly and independently select* $m = \Theta\left(\frac{1}{\epsilon} \cdot d^{3/2} \log(1/\beta) \left(\frac{2}{\beta}\right)^d\right)$ *points in* X.
2. *If the distance between every pair of points in the sample is at most* b *then accept, otherwise reject.*

Step 2 of the algorithm can clearly be done in time $O(m^2)$. If $d \leq 3$ the problem can be solved in time $O(m \log m)$ [35].

Theorem 2 *Algorithm 2 is a diameter-clustering tester for* $k = 1$ *under the* L_2 *metric.*

We shall use the following lemma. For an illustration see Figure 4.1.

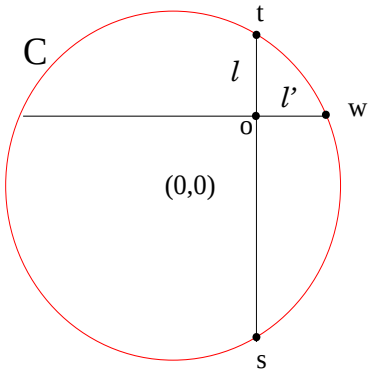

Figure 1: An illustration for the proof of Lemma 1.

Lemma 1 *Let* C *be a circle of radius at most* b. *Let* s *and* t *be any two points on* C, *and let* o *be a point on the segment connecting* s *and* t, *such that* $\mathrm{dist}(s, o) \geq b$. *Consider the line perpendicular to the line through* s *and* t *at* o *and let* w *be its (closer) meeting point with the circle* C. *Then* $\mathrm{dist}(w, o) \geq \mathrm{dist}(o, t)/2$.

Proof: Denote by $\ell = \mathrm{dist}(o, t)$ and $\ell' = \mathrm{dist}(w, o)$. We place the center of the circle C at the origin $(0, 0)$, and set the y axis to be parallel to the line connecting s and t. Let $r \leq b$ be the radius of C. If we denote the y coordinate of t by η, then its x coordinate is $\alpha = \sqrt{r^2 - \eta^2}$. Given the distances between the points (and the orientation of the coordinate system), the point o is at coordinates $(\alpha, \eta - \ell)$, and the point w at $(\alpha + \ell', \eta - \ell)$. Since w is on the circle, we must have

$$(\alpha + \ell')^2 + (\eta - \ell)^2 = r^2 \tag{1}$$

which implies that

$$\ell' = \sqrt{r^2 - (\eta - \ell)^2} - \alpha \tag{2}$$

If we now substitute $\alpha = \sqrt{r^2 - \eta^2}$ we get

$$\ell' \quad = \quad \sqrt{r^2 - (\eta - \ell)^2} - \sqrt{r^2 - \eta^2}$$

$$
\begin{aligned}
&= \frac{(r^2-(\eta-\ell)^2)-(r^2-\eta^2)}{\sqrt{r^2-(\eta-\ell)^2}+\sqrt{r^2-\eta^2}} \\
&= \frac{\eta^2-(\eta-\ell)^2}{\sqrt{r^2-(\eta-\ell)^2}+\sqrt{r^2-\eta^2}} \\
&= \frac{\ell(2\eta-\ell)}{\sqrt{r^2-(\eta-\ell)^2}+\sqrt{r^2-\eta^2}} \\
&> \frac{\ell(2\eta-\ell)}{2r}
\end{aligned}
$$

Since s and t are both on the circle and $\mathrm{dist}(s,t) = \mathrm{dist}(s,o) + \mathrm{dist}(o,t) \geq b + \ell$, we have that $2\eta \geq b + \ell$. Hence, since $r \leq b$, we obtain that $\ell' \geq \ell/2$ as claimed. ■

Proof of Theorem 2 for $d = 2$. We start by proving the theorem for two dimensions, and then show how it generalizes to any d dimensions. Clearly if X is $(1,b)$-clusterable then the algorithm accepts. We thus focus on proving that if X is ϵ-far from being $(1,(1+\beta)b)$-clusterable then the algorithm rejects with probability at least $2/3$.

We shall view the sample as being selected in $p = 2\pi/\beta^2$ phases, where in each phase $\Theta(\log(p)/\epsilon)$ points are selected (uniformly and independently). We shall show that with high probability over the choice of the sample, in each phase certain *progress* is made. The progress is such that after at most p phases, the diameter of all sample points is greater than b (causing the algorithm to reject). For each $1 \leq j \leq p$, let U_j denote the union of all points selected in the first j phases. We shall need the following definitions.

Definition 1

- *For any given point $x \in \Re^2$, let C_x denote the circle of radius b centered at x.*
- *For a given (finite) set $T \subseteq \Re^2$, let $I(T)$ denote the* intersection *of all circles C_x of points $x \in T$.*
- *For any region R in $\Re^2$, let $A(R)$ denote the size (area) of R.*

By the above definition, for each phase j, every point $y \in I(U_j)$ is at distance at most b from *every* point in the sample selected so far. If in phase $j+1$ a new sample point x falls outside $I(U_j)$, then the algorithm rejects, as this means that the new point is at distance greater than b from some sample point. Otherwise, $x \in I(U_j)$, and we consider the decrease in the area of the intersection caused by the addition of x. That is, $A(I(U_j)) - A(I(U_j \cup \{x\})) = A(I(U_j) \setminus C_x)$.

Definition 2 *We say that a point $x \in X$ is* influential *with respect to $I(U_j)$ if $x \notin I(U_j)$ or if x causes a significant decrease in the area of $I(U_j)$. Namely, if the area $A(I(U_j) \setminus C_x)$ that is removed from the intersection, is greater than $(\beta b)^2/2$.*

We claim that if X is ϵ-far from being $(1,(1+\beta)b)$-clusterable, then for every $1 \leq j \leq p-1$, in phase $j+1$ there are at least ϵn points in X that are influential with respect to $I(U_j)$. Subject to this claim, if the sample in each phase is of size at least $\ln(3p)/\epsilon$, then the probability that an influential point is not selected in a *fixed* phase, is at most

$$(1-\epsilon)^{\ln(3p)/\epsilon} < \exp(-\ln(3p)) = 1/(3p) \ .$$

Hence, the probability that for *some* phase no influential point is selected is less than $1/3$.

Thus, assume from now on that for every $1 \leq j \leq p-1$, the sample selected in phase $j+1$ contains an influential point x with respect to $I(U_j)$. As stated above, if $x \notin I(U_j)$ then the algorithm rejects. Otherwise, x decreases the area of $I(U_j)$ by at least $(\beta b)^2/2$. However, since the area of the initial circle (defined by the first sample point) is πb^2, then the number of phases in which such a decrease can occur is at most $p = 2\pi/\beta^2$.

In order to complete the proof Theorem 2 (for $d = 2$), we must show that for every $1 \leq j \leq p-1$, there are at least ϵn points in X that are influential with respect to $I(U_j)$. Assume contrary to the claim that there are at most ϵn influential points with respect to some $I(U_j)$. Then we can remove these (at most) ϵn influential points from X. The points that remain in X all belong to $I(U_j)$, and as the following lemma shows, they form a cluster of diameter at most $(1+\beta)b$, in contradiction to our assumption on X.

Lemma 2 *Let T be any finite subset of $\Re^2$. Then for every $x, y \in I(T)$ such that x is non-influential with respect to T,* $\mathrm{dist}(x,y) \leq (1+\beta)b$.

Proof: It is clear of course that if $y \in C_x$ then $\mathrm{dist}(x,y) \leq b$. Therefore, let $y \in I_j \setminus C_x$. Consider the line through x and y, and let o be the point where it intersects with C_x. Then,

$$\mathrm{dist}(x,y) = \mathrm{dist}(x,o) + \mathrm{dist}(o,y).$$

Clearly $\mathrm{dist}(x,o) = b$. Thus, we want to show that $\mathrm{dist}(o,y) \leq \beta b$.

Let us draw the tangent to C_x at o and let z and w be the first two points it meets on the boundary of I_j. The points y, w, z define a triangle T, whose height is $h = \mathrm{dist}(o,y)$. Let $\ell_1 = \mathrm{dist}(w,o)$, and $\ell_2 = \mathrm{dist}(z,o)$. Thus, the length of the base of the triangle S is $\ell_1 + \ell_2$. Let $A(T)$ denote the area of T. Since $T \subseteq I_j \setminus C_x$, and x is non-influential, then

$$A(T) = \frac{h(\ell_1+\ell_2)}{2} \leq \frac{(\beta b)^2}{2}\,. \tag{3}$$

We will now show that $h \leq \ell_1 + \ell_2$, and from this conclude that $h \leq \beta b$ as required.

We prove that $\ell_1 \geq h/2$. Let C_1 be the circle on which w sits, and let s and t be the intersection points of the line connecting x and y with the circle C_1 (see Figure 2).

We have, $\mathrm{dist}(o,t) \geq h$ and $\mathrm{dist}(s,o) \geq b$. We can thus apply Lemma 1, and get that

$$\ell_1 = \mathrm{dist}(w,o) \geq \mathrm{dist}(o,t)/2 \geq h/2.$$

In an analogous way we can show that $\ell_2 \geq h/2$. This implies that $h \leq 2\min(\ell_1, \ell_2) \leq \ell_1 + \ell_2$. By Equation (3) we can conclude that:

$$\frac{h^2}{2} \leq \frac{h(\ell_1+\ell_2)}{2} \leq \frac{(\beta b)^2}{2}.$$

■

Extending the Proof to Higher Dimensions. For each sample point x let B_x denote the d-dimensional ball of radius b centered at x. Let $I(U_j)$ be the intersection of all balls centered at points selected in phases 1 to j, and let $V(I(U_j))$ denote the volume of the intersection. Let V_d

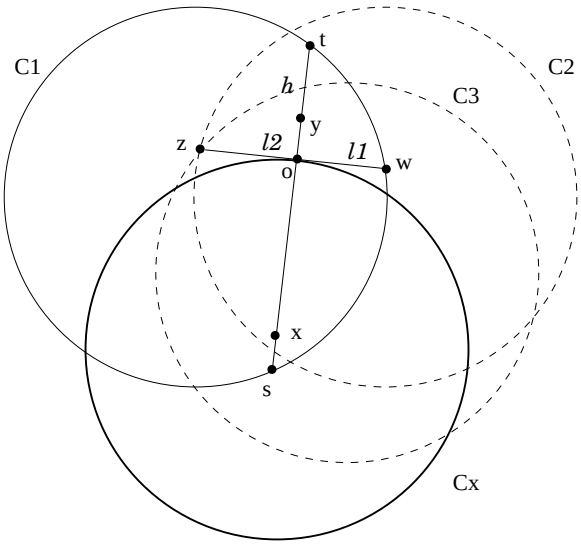

Figure 2: An illustration for the proof of Lemma 2. The circles C_x and C_1 are as defined in the proof. The circles C_2 and C_3 denote additional circles defined by points in the sample.

denote the volume of the d-dimensional unit ball. Here we shall say that a point x is *influential* with respect to $I(U_j)$ if $x \notin I(U_j)$, or if the volume removed by x is at least

$$\begin{aligned} V(I(U_j)) - V(I(U_j \cup \{x\})) &= V(I(U_j) \setminus B_x) \\ &> \frac{(\beta b)^d \cdot V_{d-1}}{d \cdot 2^{d-1}} \ . \end{aligned}$$

Since the volume of the initial ball of radius b (defined by the first sample point) is $V_d \cdot b^d$, then the number p of phases required is at most

$$\frac{d \cdot V_d}{2 \cdot V_{d-1}} \cdot \left(\frac{2}{\beta}\right)^d = O\left(\sqrt{d}\left(\frac{2}{\beta}\right)^d\right).$$

Once again, the following lemma, completes the proof of Theorem 2 for any $d > 2$.

Lemma 3 *Let T be any finite subset of $\Re^d$. Then for every $x, y \in I(T)$ such that x is non-influential with respect to T,* $\mathrm{dist}(x, y) \leq (1 + \beta)b$.

Proof: Let $y \in I_j \setminus B_x$. Consider the line through x and y, and let o be the point where it intersects with B_x. Then,

$$\mathrm{dist}(x, y) = \mathrm{dist}(x, o) + \mathrm{dist}(o, y),$$

where $\mathrm{dist}(x, o) = b$. Again we show that $h = \mathrm{dist}(o, y) \leq \beta b$.

Consider some plane that passes through the line defined by x and y. Draw in this plane the line tangent to B_x at o. Let z and w be the first two points that this line meets on the boundary of I_j. Notice, that any such plane intersects each of the d-dimensional balls defining I_j in a circle of radius at most b. Thus, we can again use Lemma 1, and prove (as in Lemma 2), that $\mathrm{dist}(z, o) \geq h/2$ and $\mathrm{dist}(w, o) \geq h/2$. This will be true for any plane passing through the line defined by x and y.

Therefore, a $(d-1)$-dimensional ball of radius $h/2$ is contained in the intersection of $I_j \setminus B_x$ with the $(d-1)$-dimensional hyperplane tangent to B_x at o. Thus, the cone of height h whose base is this $(d-1)$-dimensional ball of radius $h/2$ is contained in $I_j \setminus B_x$. The volume of this cone is

$$\frac{h(h/2)^{d-1}V_{d-1}}{d}$$

and since x is non-influential we have

$$\frac{h(h/2)^{d-1}V_{d-1}}{d} \leq \frac{(\beta b)^d \cdot V_{d-1}}{d \cdot 2^{d-1}} \quad .$$

Thus, $h \leq \beta b$ as required. ■

4.2 General k

The algorithm for $k > 1$ is a generalization of the algorithm for $k = 1$.

Algorithm 3 (L_2 metric, diameter cost, $k \geq 1$, $d \geq 1$, $0 < \beta \leq 1$)

1. *Uniformly and independently select* $m = \Theta\left(\frac{k^2 \log k}{\epsilon} \cdot d \cdot \left(\frac{2}{\beta}\right)^{2d}\right)$ *points in* X.
2. *If there exists a* k*-way partition* P *of the sample for which* $D(P) \leq b$*, then accept. Otherwise, reject.*

Verifying whether there exists a k-way partition of m points having diameter at most b can be done in time $(O(m))^{d \cdot k^2}$ [38]. The basic observation is that we may consider only partitions for which the convex hulls of the different clusters are disjoint. This is true since given a minimum diameter partition for which some point in cluster i belongs to the convex hull of cluster i', we can move this point from cluster i to cluster i' without increasing the diameter. Thus, in Step 2 the algorithm enumerates all such partitions of the sample and computes their diameter. This is done by considering all choices of $\binom{k}{2}$ hyperplanes among the $O(m^{d+1})$ hyperplanes that separate the m sample points, and then merging subsets of points that fall in the resulting regions into k clusters.

Theorem 3 *Algorithm 3 is a diameter-clustering tester under the* L_2 *metric.*

We start by extending the notion of influential points.

Definition 3 *Let* $P_S = (S^1, \ldots, S^k)$ *be a partition of a subset* $S \subset X$*. We say that a point* x *is* influential *with respect to* P_S *if either* $x \notin \bigcup_{i=1}^{k} I(S^i)$ *(that is,* x *is at distance greater than* b *from some point in every* S^i*), or for every* S^i*:*

$$V(I(S^i) \setminus B_x) > \frac{(\beta b)^d \cdot V_{d-1}}{d \cdot 2^{d-1}}$$

(that is, the volume of $I(S^i)$ *is reduced significantly by* x *for every* S^i*). Let* $Y(P_S) \subset X$ *denote the set of all points that are influential with respect to* P_S*.*

Claim 4 *Suppose* X *is* ϵ*-far from* $(k, (1+\beta)b)$*-clusterable, and let* $P_S = (S^1, \ldots, S^k)$ *be a partition of some* $S \subset X$*. Then for any given* $0 < \delta < 1$*, with probability at least* $1 - \delta$*, a uniformly and independently selected sample of size* $s \geq \frac{\ln(1/\delta)}{\epsilon}$ *contains at least one point* $y \in Y(P_S)$*.*

Proof: By Lemma 3, if X is ϵ-far from $(k,(1+\beta)b)$-clusterable, then necessarily $|Y(P_S)| > \epsilon n$. Otherwise, we could remove all influential points and assign each other point $x \in X$ to a cluster i such that x is non-influential with respect to S^i. This would result in a k-way partition of all but at most an ϵ fraction of the points in X, that has diameter at most $(1+\beta)b$.

Therefore, the probability that a sample of size $s \geq \frac{\ln(1/\delta)}{\epsilon}$ will *not* contain any point in $Y(P_S)$ is at most $(1-\epsilon)^s < \exp(-\epsilon \cdot s) = \delta$, as desired. ■

Proof of Theorem 3: Once again, if X is (k,b)-clusterable, then the algorithm always accepts. We thus focus on the case in which X is ϵ-far from being $(k,(1+\beta)b)$-clusterable.

As in the proof of Theorem 2, we view the sample as being selected in phases. Let $p = \Theta(\sqrt{d} \cdot (2/\beta)^d)$ be the number of phases sufficient for the $k=1$ case, and let $p(k) = k \cdot (p+1)$ be the number of phases used here in the analysis of $k > 1$. Let m_j be the size of the sample selected in the j'th phase, where $\sum_{j=1}^{p(k)} m_j = m$. Let U_j denote the union of all samples selected in the first j phases. Thus, $U = U_{p(k)}$ is the complete sample.

Our goal is to show that with probability at least $2/3$ over the choice of the sample, for *every* partition P of $U_{p(k)}$ we have $D(P) > b$. To this end we define a family of *influential* partitions. For each phase j there is a sub-family of influential partitions that correspond to that phase. These are partitions of subsets of U_j. We show that with probability at least $2/3$, for every phase j and every influential partition $\hat{P}$ corresponding to that phase, the sample selected in the next phase contains an influential point for $\hat{P}$. This will imply that after at most $p(k) = k(p+1)$ phases, the diameter of each influential partition, and consequently of *every partition of the sample*, is greater than b.

We define the influential partitions in an inductive manner. In the initial phase (phase 0) there is a single influential partition of a sample of size 1 (i.e., $m_0 = 1$). Suppose that for each influential partition $\hat{P} = (S^1, \ldots, S^k)$ in phase $j-1$, the j'th sample contains a point from $Y(\hat{P})$, and let us denote this point by $y(\hat{P})$. (If there is more than one such point then $y(\hat{P})$ is defined as the one having the smallest index.) Then in phase j we shall have the k influential partitions

$$(S^1 \cup \{y(\hat{P})\}, S^2, \ldots, S^k) \cdots (S^1, \ldots, S^{k-1}, S^k \cup \{y(\hat{P})\})$$

This implies that the total number of influential partitions in phase j is at most k^j.

We now apply Claim 4 to each one of the k^{j-1} influential partitions in phase $j-1$ (that have diameter at most b). If

$$m_j = \frac{(j-1)\ln k + \ln(3p(k))}{\epsilon}$$

then with probability at least $1 - \frac{1}{3p(k)}$, the j'th sample in facts contains a point $y(\hat{P}) \in Y(\hat{P})$ for every influential partition $\hat{P}$ in phase $j-1$. Setting

$$\begin{aligned} m &= \sum_{j=1}^{p(k)} m_j = \Theta\left(\frac{p(k)^2 \cdot \log k + p(k)\log p(k)}{\epsilon}\right) \\ &= \Theta\left(\frac{k^2 \log k}{\epsilon} \cdot d \cdot (2/\beta)^{2d}\right) \end{aligned}$$

we get that with probability at least $2/3$, the j'th sample contains a point $y(\hat{P}) \in Y(\hat{P})$ for every phase j and every influential partition $\hat{P}$ from phase $j-1$.

Assume the above event in fact holds and so in particular the influential partitions are well defined. We now show that this implies that after at most $p(k)$ phases, the diameter of every partition of the sample must be greater than b.

Consider any partition $P = (U^1_{p(k)}, \ldots, U^k_{p(k)})$ of $U_{p(k)}$, and let $P_j = (U^1_j, \ldots, U^k_j)$ be its *restriction* to U_j. That is, $U^i_j = U^i_{p(k)} \cap U_j$. We claim that there must exist a sequence of influential partitions $\hat{P}_1, \ldots, \hat{P}_{p(k)}$, where $\hat{P}_j = (S^1_j, \ldots, S^k_j)$, so that the following holds. For every i, $S^i_j \subseteq U^i_j$, and for some i, $S^i_j = S^i_{j-1} \cup \{y(\hat{P}_{j-1})\}$. This follows immediately by induction on j: The base of the induction, $j = 0$ is clear. We assume it is true for $j-1$, and prove it for j. Let $1 \leq i \leq k$ be such that $y(\hat{P}_{j-1}) \in U^i_j$. Then we let $\hat{P}_j = (S^1_{j-1}, \ldots, S^i_{j-1} \cup \{y(\hat{P}_{j-1})\}, \ldots, S_k)$, which by definition of the influential partitions is a influential partition.

Let us fix the above sequence of influential partitions. Since there are $p(k) = k \cdot (p+1)$ phases, there must be some $1 \leq i \leq k$ such that in at least $p+1$ phases $j_1, \ldots, j_{p+1}$, $S^i_{j_t} = S^i_{j_t - 1} \cup \{y(\hat{P}_{j_t - 1})\}$ (the first such phase will cause $S^i_{j_1}$ to be non-empty). But by our analysis of the $k = 1$ case, this implies that $d(S^i_{p(k)}) > b$. Since $S^i_{p(k)} \subseteq U^i_{p(k)}$, we have that $d(U^i_{p(k)}) > b$, and so $D(P) > b$. Since the above holds for every partition P of $U = U_{p(k)}$, the theorem follows. ■

4.3 Finding an Approximately Good Clustering

Similarly to what was shown in the previous section, if X is (k, b)-clusterable, then the testing algorithm can be used to find an implicit representation of an approximately good $(k, (1+\beta)b)$-clustering of X. Here the process is slightly more complex.

Recall that for a set T of points in $\Re^d$, $I(T)$ denotes the intersection of all d-dimensional balls B_x having radius b that are centered at points $x \in T$.

Definition 4 *Let $P_S = (S^1, \ldots, S^k)$ be a partition of a subset $S \subseteq X$. A point $x \in X$ is* good *with respect to P_S if there exists an index $1 \leq i \leq k$ such that $x \in I(S^i)$ and $\mathrm{dist}(x, y) \leq (1+\beta)b$ for every $y \in I(S^i)$. Otherwise, x is* bad *with respect to P_S.*

*A partition P_S is α-*bad *for a given $0 \leq \alpha \leq 1$, if the number of bad points with respect to P_S is greater than αn. Otherwise, P_S is α-*good.

Observe that given a subset $S \subseteq X$ and a partition P_S of S that is ϵ-good, P_S can be used to induce an ϵ-good $(k, (1+\beta)b)$ clustering of X. Also note that by Lemma 3 if a point x is bad with respect to a partition P_S, then x must be influential with respect to P_S.

Algorithm 4 (Approximately good clustering, diameter cost)

1. *Call Algorithm 3 with a sample of size $m = O\left(\frac{d^{5/2} \cdot k^3}{\epsilon} \cdot \left(\frac{2}{\beta}\right)^d \log\left(\frac{d \cdot k}{\epsilon \cdot \beta}\right)\right)$.*
2. *Let P be the k-way partition of the sample that is found by Algorithm 3 (if such a partition is found).*
3. *View the sample as being selected in $p(k) = \Theta(k \cdot \sqrt{d} \cdot (2/\beta)^d)$ phases, where U_j denotes the union of all samples selected in the first j phases, and $|U_j| = \Theta\left(\frac{j}{\epsilon} \cdot \left(d \cdot k^2 \cdot \log \frac{d \cdot k \cdot j \cdot p(k)}{\epsilon}\right)\right)$. Let P_j be the restriction of P to U_j. That is, if $P = (U^1, \ldots, U^k)$, then $P_j = (U^1 \cap U_j, \ldots, U^k \cap U_j)$.*
4. *Take an additional sample of size $\Theta(\log(p(k))/\epsilon)$, and count the number of bad points in this additional sample with respect to each partition P_j.*[2]

[2] Checking whether a point is bad with respect to a given partition can be done by linear programming.

5. *Select the restriction P_g that has the smallest number of bad points in the sample and use it to induce the partition of X. That is, if $P_g = (U_g^1, \ldots, U_g^k)$, then for every $x \in X$, if there exists an index i such that $x \in I(U_g^i)$ and $\mathrm{dist}(x,y) \leq (1+\beta)b$ for every $y \in I(U_g^i)$, then assign x to cluster i.*

Notice that the above algorithm calls Algorithm 3 with a sample of size slightly larger than what was needed in the proof of Theorem 3. We shall return to this issue at the end of this subsection.

Theorem 4 *With probability at least $2/3$, the selected partition P_g is ϵ-good.*

Definition 5 *Let $S \subseteq X$ be a set of points. A partition $P_S = (S^1, \ldots, S^k)$ of S is called a* convex-partition *if the convex hulls of the different S^i's are disjoint.*

Lemma 5 *There exists a constant c such that for any fixed set S, with probability at least $1 - \frac{1}{6p(k)}$, a sample of size $s = \frac{d \cdot k^2 \cdot \ln(c|S|) + \ln(6p(k))}{\epsilon}$ contains at least one bad point with respect to each $(\epsilon/2)$-bad convex partition of S.*

Proof: Let Q be any fixed $(\epsilon/2)$-bad convex partition of S. The probability that a sample of size s as stated in the lemma does not contain a bad point with respect to Q is at most $(1-(\epsilon/2))^s < \exp(-(\epsilon/2)s) = \frac{1}{(c|S|)^{dk^2} \cdot 6p(k)}$. It remains to verify that the number of convex partitions of S is at most $(c|S|)^{dk^2}$ for some constant c. Each convex partition of S can be defined by a selection of $\binom{k}{2}$ hyperplanes among the $O(|S|^{d+1})$ hyperplanes that separate $|S|$ points in d dimensions, and then merging subsets of points that fall into the resulting regions into k clusters. The total number of convex partitions is hence $O(|S|)^{dk^2}$. ■

Lemma 6 *With probability at least $5/6$, if we select a sample of size*

$$m = \Theta\left(\frac{d^{\frac{5}{2}} \cdot k^3}{\epsilon} \cdot \left(\frac{2}{\beta}\right)^d \log\left(\frac{d \cdot k}{\epsilon \cdot \beta}\right)\right)$$

then for every phase j and for every convex partition Q of U_j that is $(\epsilon/2)$-bad, the sample selected in phase $j+1$ contains at least one bad point with respect to Q.

Proof: Let m_j be the size of the (additional) sample selected in phase j, so that $|U_j| = |U_{j-1}| + m_j$, and where $|U_0| = 1$. If we apply Lemma 5 with $S = U_{j-1}$ and $m_j = s$, then it is not hard to verify that

$$|U_j| \leq \frac{2j}{\epsilon} \cdot \left(d \cdot k^2 \cdot \log \frac{c \cdot d \cdot k^2 \cdot j \cdot (6p(k))}{\epsilon}\right).$$

Since $p(k) = \Theta(k \cdot \sqrt{d} \cdot (2/\beta)^d)$, we have that $|U_{p(k)}| = O\left(\frac{1}{\epsilon} \cdot d^{5/2} \cdot k^3 \cdot (2/\beta)^d \cdot \log \frac{d \cdot k}{\epsilon \cdot \beta}\right)$. Hence, if we take a sample of size $m = |U_{p(k)}|$ then the probability that for some $1 \leq j \leq p(k)$, and some $(\epsilon/2)$-bad convex partition of U_j, the $(j+1)$-sample does *not* contain a bad point with respect to the partition, is at most $p(k) \cdot \frac{1}{6(p(k)} = \frac{1}{6}$. ■

Corollary 7 *With probability at least $5/6$, there exists an index $1 \leq a \leq p(k)$ such that the restriction P_a is $(\epsilon/2)$-good.*

Proof: Algorithm 3 finds an optimal partition P of the sample by enumerating all convex partitions of the sample. Since the partition P that Algorithm 3 finds is convex, so is each of its restrictions P_j.

By Lemma 6, with probability at least 5/6, the sample selected in each phase contains a bad point with respect to every $(\epsilon/2)$-bad convex partition of the sample selected so far. Suppose that this in fact happens. Then there exists a phase $1 \leq a \leq p(k)$, such that the restriction P_a is $(\epsilon/2)$-good. Otherwise, since every bad point is an influential one, then similarly to what was argued in the proof of Theorem 3, the partition P could not have diameter at most b. ■

Proof of Theorem 4: Let P_a be an $(\epsilon/2)$-good partition guaranteed with probability at least 5/6 by Corollary 7. Then, the probability that the partition P_g selected by Algorithm 4 is ϵ-good is lower bounded by the probability that the following two events both occur: (1) For every ϵ-bad partition P_j, the fraction of bad points in the sample is greater than $\frac{3\epsilon}{4}$; (2) For the $(\epsilon/2)$-good partition P_a, the fraction of bad points in the sample is at most $\frac{3\epsilon}{4}$. Clearly, if both events occur than the selected partition cannot be ϵ-bad. In order to lower bound the probability that both these events occur, we upper bound the probability that either one of them does not occur. By applying a multiplicative Chernoff bound, and using the fact that the number of ϵ-bad partitions is less than $p(k)$, we get that a sample of size $\Theta(\log(p(k))/\epsilon)$ ensures that the probability that one of them does not occur is at most 1/6.

Adding the two sources of failure, that is, the probability that there is no $(\epsilon/2)$-good partition P_a, and the probability that the selected partition P_j is ϵ-bad (given that an $(\epsilon/2)$-good partition P_a exists), we get a total of 1/3 failure probability. ■

As noted previously, the size of the sample used here is larger than that used in Algorithm 3. The reason is that in the analysis of Algorithm 3, we used influential partitions, while here we use convex partitions, whose number is larger. We could not see how to use the former in our (constructive) argument here.

4.4 A Lower Bound for Testing Diameter Clustering

Theorem 5 *For any $\beta > 0$, any algorithm that determines with success probability at least 2/3 whether X is $(1, b)$-clusterable or $\frac{1}{2}$-far from being $(1, (1+\beta)b)$-clusterable with respect to the diameter cost, must sample $\Omega\left(\left(\frac{1}{\beta}\right)^{(d-1)/4}\right)$ points from X.*

In order to prove the theorem we shall need the following lemma.

Lemma 8 *For any dimension d, value $r \in \Re^d$ and $\delta > 0$, it is possible to choose $\Omega\left(\sqrt{d} \cdot \left(\frac{1}{\delta}\right)^{d-1}\right)$ antipodal pairs of points on the surface of the $(d-1)$-dimensional sphere of radius r, where the distance between any two points is larger than $\delta \cdot r$.*

Proof: We choose the pairs one by one in the following way. Choose a pair of antipodal points that are of distance greater than $\delta \cdot r$ from all points chosen so far. Continue to choose antipodal pairs in this way as long as possible.

We claim that the $(d-1)$-dimensional caps of radius $\delta \cdot r$ centered at the points we chose, cover the surface of the $(d-1)$-dimensional sphere of radius r. Otherwise, if there exists a point that is not covered, then it must also be the case that its antipodal point is not covered, and thus we can add an additional pair of antipodal points.

Let θ be the angular diameter of a cap of radius $\delta \cdot r$, and let $\theta_0 = \pi/2 - \theta/2$. Then, $\delta = \sqrt{2 - 2\sin\theta_0}$. Hence, the ratio between the surface area of a $(d-1)$-dimensional sphere of radius r and the surface area of a cap of such a sphere of radius $\delta \cdot r$ is:

$$\frac{\int_{-\pi/2}^{\pi/2} \cos^{d-2} t \, dt}{\int_{\theta_0}^{\pi/2} \cos^{d-2} t \, dt}$$

The numerator is $\Theta(1/\sqrt{d})$ and the denominator is equal to

$$\begin{aligned}
\int_{\theta_0}^{\pi/2} \cos^{d-2} t \, dt &= \int_{\theta_0}^{\pi/2} (1 - \sin^2 t)^{\frac{d-3}{2}} \cos t \, dt \\
&\leq \int_{\theta_0}^{\pi/2} (2(1 - \sin t))^{\frac{d-3}{2}} \cos t \, dt \\
&= -\frac{1}{2} \cdot \frac{2}{d-1} \cdot (2(1 - \sin t))^{\frac{d-1}{2}} \Big|_{\theta_0}^{\pi/2} \\
&= \frac{1}{d-1} \cdot (2(1 - \sin\theta_0))^{\frac{d-1}{2}} \\
&= \frac{\delta^{d-1}}{d-1}
\end{aligned}$$

Hence the number of points we can choose is $\Omega\left(\sqrt{d} \cdot \left(\frac{1}{\delta}\right)^{d-1}\right)$. ■

Proof of Theorem 5: Consider the d-dimensional ball of radius r, where r is slightly greater than $(1+\beta)b/2$. By definition, the distance between any two antipodal points on the surface of this ball is greater than $b(1+\beta)$. By Lemma 8 we can choose $\Omega\left(\sqrt{d} \cdot \left(\frac{1}{\delta}\right)^{d-1}\right)$ antipodal pairs of points on the surface of this ball, such that the distance between any two points is at least $\delta \cdot r$. Thus, by Pythagoras Theorem if we choose $\delta > \frac{2\sqrt{\beta(2+\beta)}}{1+\beta} = \Omega\left(\sqrt{\beta}\right)$ then the distance between any two points that are *not* antipodal is at most b.

Let us fix such a selection of $s = \Omega\left(\sqrt{d} \cdot \left(\frac{1}{\sqrt{\beta}}\right)^{d-1}\right)$ antipodal pairs of positions on the surface of the ball, and suppose X is such that we have $n/(s/2)$ points in each position.[3] Clearly, X is $\frac{1}{2}$-far from being $(1,(1+\beta)b)$-clusterable. However, by the "birthday paradox" (see for example [31]), with high probability, a sample of size $c \cdot \sqrt{s}$ will not contain a pair of points in antipodal positions (for some constant $c < 1$). That is, all points in the sample will be at distance at most b from each other. This implies that our "natural" algorithm (and actually any algorithm having one-sided error) requires $\Omega\left(\left(\frac{1}{\beta}\right)^{(d-1)/4}\right)$ sample points.

To prove the claim for any algorithm, we can apply an argument similar to that used in the lower bound proofs of [18]. Here we sketch the idea. We define two families of sets of n points, where in the first family all sets are far from being $(1,(1+\beta)b)$-clusterable, and in the second family all sets are $(1,b)$-clusterable. The first family is defined by all namings of the n points on the surface of a d-dimensional ball as defined above. In the second family, a set X is defined by

[3] In order that X be an actual set and not a multiset, we can place the points at slightly different but very close positions. Note that our algorithms do not rely on the points in X being different from each other (or at any minimal distance from each other).

selecting for each one of the s pairs of antipodal positions, one of the positions, and putting n/s points in that position. Every such X is $(1, b)$-clusterable.

We now define two processes, one for each family, that constructs a random set X in the family as it answers the algorithm's queries, and completes this construction after the algorithm terminates. Without loss of generality we may assume that the algorithm never queries the same point twice. Then, for each query of the algorithm, the process selects a new point on the sphere in a random fashion which depends on the family to which X belongs.

Assume we are now answering the j'th query, and the processes must decide where to position the new point. Each of the two processes first flips a coin with bias τ, where τ is approximately j/s. According to the outcome, the new point will be placed in the same (or antipodal) position of a previously selected point, or placed in an unoccupied position (whose antipodal position is also unoccupied). In the latter case, an antipodal pair is selected uniformly among all unoccupied pairs, and the new point is placed with equal probability on each position in the pair. In the former case, both processes randomly select an occupied position, where the second process places the new point in the selected position, and the first process places the new point either in this position or in its antipodal position.

Note that as long as the former case does not occur, the distributions on the positions of the points are exactly the same for both processes. However, for a number of queries $j < c \cdot \sqrt{s}$ (for some constant $c < 1$), the probability that an occupied position is selected (in either process) is less than $1/3$. This implies that the statistical differences between the distributions on sequences of queries and answers for the two processes is less than $1/3$, and the theorem follows. ■

Remark: Essentially the same argument as in the above proof gives an $\Omega(\sqrt{n})$ lower bound for $\beta = 0$.

5 Testing of Radius Clustering Under the L_2 Metric

Below is our algorithm for testing with respect to the radius cost under the L_2 metric and for $\beta = 0$. Recall that for this cost and metric, all points in each cluster must be contained in a ball of radius b. The analysis of this algorithm can be easily generalized to any metric under which each cluster is determined by a "simple" convex set (that is, where the family of such sets has VC-dimension $O(d)$). In particular this holds for the L_∞ metric (where these sets are axis aligned cubes). As we see below, the size of the sample is almost linear in d. An alternative analysis of the algorithm, which works for $\beta > 0$ and uses a sample of size independent of d, is given in Subsection 5.1.

Algorithm 5 (L_2 metric, radius cost, $k \geq 1$, $d \geq 1$, $\beta = 0$)

1. *Uniformly and independently select a sample of* $m = \Theta\left(\frac{d \cdot k}{\epsilon} \cdot \log\left(\frac{d \cdot k}{\epsilon}\right)\right)$ *points in* X.
   Let us denote the points selected by U.
2. *If there* exists *a partition* $P = (U^1, ..., U^k)$ *of* U *such that* $R(P) \leq b$ *then accept. Otherwise, reject.*

Finding k balls with minimum radius that contain all m points in the sample (known as the *Euclidean k-Center Problem*) can be done in time $O(m^{kd+2})$ (*cf.* [2, Sec. 7.1]). When d is relatively small it is possible to obtain an improvement on this running time by using the algorithm of

Agrawal and Procopiuc [1], which has running time $m^{O(f(d)\cdot k^{1-1/d})}$ (where $f(d)$ is always bounded by $O(d^{5/2})$).

Remark: Using a result of [17] concerning the relation between learning algorithms and testing algorithms, we could obtain a testing algorithm for radius clustering with the same complexity as Algorithm 5 but with two-sided error. This would be based on the learnability of the concept class defined by unions of k balls. Here we give a direct analysis and obtain one-sided error. Furthermore, the same idea applied here can be used to obtain testing algorithms having one-sided error for any property that can be defined by a family of subsets having bounded VC-dimension.

Theorem 6 *Algorithm 5 is a radius-clustering tester under the L_2 metric for $\beta = 0$.*

We shall need the following definitions (which for sake of the presentation are not given in their full generality). Let $\mathcal{S}$ be a family of subsets of $\Re^d$, let R be a finite subset of $\Re^d$ and let $0 < \epsilon < 1$. We say that $N \subset R$ is an *ϵ-net* of R (with respect to $\mathcal{S}$) if for every $S \in \mathcal{S}$ such that $|S \cap R| > \epsilon \cdot |R|$ there exists at least one point $x \in S \cap N$. In other words, N is an ϵ-net if it "hits" every subset in $\mathcal{S}$ that has a relatively large intersection with R. Our interest in ϵ-nets will soon become clear, but first we need one more definition.

We say that a subset $A \subset \Re^d$ is *shattered* by a family of subsets $\mathcal{S}$, if for every $A' \subseteq A$, there exists $S \in \mathcal{S}$ such that $A' = A \cap S$. The VC-*dimension* of $\mathcal{S}$, denoted by $\mathrm{VCD}(\mathcal{S})$, is the maximum size of a subset $A \subset \Re^d$ that is shattered by $\mathcal{S}$. The VC-dimension of a family of subsets is hence a certain measure of richness (or diversity) of the family.

The following theorem is a special case of a theorem that was proved by Haussler and Welzl [21] based on the work of Vapnik and Chervonenkis [40].

Theorem 7 ([21]) *Let $\mathcal{S}$ be any family of subsets of $\Re^d$, let R be any finite subset of $\Re^d$ and let $0 < \epsilon < 1$. Consider a sample U of size $m \geq \frac{8\mathrm{VCD}(\mathcal{S})}{\epsilon} \cdot \log \frac{8\mathrm{VCD}(\mathcal{S})}{\epsilon}$ selected uniformly and independently from R. Then with probability at least $2/3$, U is an ϵ-net for R with respect to $\mathcal{S}$.*

The proof of Theorem 7 actually gives a bound on the sample size m in terms of a slightly different measure than $\mathrm{VCD}(\mathcal{S})$, which we refer to as the *shatter exponent* (where $\mathrm{VCD}(\mathcal{S})$ is an upper bound on this measure). In our case we can get a slightly better bound on m if we use the shatter exponent directly. We next define it and state a corresponding variant of Theorem 7.

For a subset $A \subset \Re^d$, let $\Phi_{\mathcal{S}}(A) \stackrel{\mathrm{def}}{=} \{A \cap S : \ S \in \mathcal{S}\}$ be the *projection* of $\mathcal{S}$ on A. For any integer m let $\phi_{\mathcal{S}}(m) = \max_{A,\, |A|=m} |\Phi_{\mathcal{S}}(A)|$ be the maximum size of the projection of $\mathcal{S}$ on a set of size m. In particular, by definition of the VC-dimension, for every $m \leq \mathrm{VCD}(\mathcal{S})$, $\phi_{\mathcal{S}}(m) = 2^m$, while for $m > \mathrm{VCD}(\mathcal{S})$, $\phi_{\mathcal{S}}(m) < 2^m$. Let the *shatter exponent*, denoted $\mathrm{SE}(\mathcal{S})$ be the smallest integer such that for every $m \geq 2$, $\phi_{\mathcal{S}}(m) \leq c \cdot m^{\mathrm{SE}(\mathcal{S})}$ for some fixed constant c. It can be shown that for every family of subsets $\mathcal{S}$, $\mathrm{SE}(\mathcal{S}) \leq \mathrm{VCD}(\mathcal{S})$, but as noted above, we can sometimes get a better bound on $\mathrm{SE}(\mathcal{S})$.

Theorem 7' *Let $\mathcal{S}$ be any family of subsets of $\Re^d$, let R be any finite subset of $\Re^d$ and let $0 < \epsilon < 1$. Consider a sample U of size $m \geq \frac{8\mathrm{SE}(\mathcal{S})}{\epsilon} \cdot \log \frac{8\mathrm{SE}(\mathcal{S})}{\epsilon}$ selected uniformly and independently from R. Then with probability at least $2/3$, U is an ϵ-net for R with respect to $\mathcal{S}$.*

Proof of Theorem 6: If X is (k, b)-clusterable, then the algorithm clearly always accepts. Hence, assume from now on that X is ϵ-far from being (k, b)-clusterable. We shall show that the algorithm rejects with probability at least $2/3$.

Let $\mathcal{B}_{k,b}$ be the family of subsets of $\Re^d$ that are defined by unions of k balls each of radius at most b, and let $\overline{\mathcal{B}}_{k,b}$ be the family of complements of subsets in $\mathcal{B}_{k,b}$. By our assumption on X, we have that for every collection of k balls each having radius at most b, there are more than ϵn points in X that do not belong to any of the balls. In other words, for every $S \in \overline{\mathcal{B}}_{k,b}$, we have $|S \cap X| > \epsilon|X|$. This implies that a subset $N \subset X$ is an ϵ-net for X with respect to $\overline{\mathcal{B}}_{k,b}$ if and only if it contains at least one point from every $S \in \overline{\mathcal{B}}_{k,b}$.

Now assume that the sample U selected by Algorithm 5 is an ϵ-net for X. Then, by definition of ϵ-nets and our assumption on X, there is *no* k-way partition P of U such that $R(P) \leq b$. This is true since such a partition corresponds to k balls having radius b, that contain all points in the sample. But this would contradict the assumption that U contains at least one point from every $S \in \overline{\mathcal{B}}_{k,b}$.

In order to bound the size of a sample that is sufficient to ensure that it constitutes an ϵ-net for X with respect to $\overline{\mathcal{B}}_{k,b}$, we bound $\mathrm{SE}(\overline{\mathcal{B}}_{k,b})$. It is easy to verify that $\mathrm{SE}(\overline{\mathcal{B}}_{k,b}) = \mathrm{SE}(\mathcal{B}_{k,b})$, and so it remains to bound $\mathrm{SE}(\mathcal{B}_{k,b})$. Given any set A of m points in $\Re^d$, the number of different subsets $A' = A \cap B$ where $B \in \mathcal{B}_{1,b}$ (i.e., sets defined by single balls), is at most m^{d+1}.[4] This follows from the following fact. For each subset A' such that there exists balls $B \in \mathcal{B}_{1,b}$ for which $A' = A \cap B$, let $B_{A'}$ be such a ball having minimum radius. It is well know that for any such bounding ball there exists a subset $A'' \subseteq A'$ having size at most $d+1$ such that $B_{A'} = B_{A''}$. Hence the number of balls enclosing different subsets of A is at most $\binom{m}{d+1} < m^{d+1}$. Since $\mathcal{B}_{k,b}$ includes unions of k balls, we have that $\mathrm{SE}(\mathcal{B}_{k,b}) \leq k(d+1)$. Hence, Theorem 6 follows by applying Theorem 7'. ■

Finding an approximately good clustering. Suppose X is (k,b)-clusterable and so the algorithm finds a k-way partition P of the sample such that $R(P) \leq b$. That is, the algorithm finds k centers $z^1, \ldots, z^k$ of balls of radius b that contain all sample points. An argument similar to the proof of Theorem 6, shows that with probability at least $2/3$ the centers found by the algorithm actually define an ϵ-good (k,b)-clustering of X. Specifically, as shown in the proof of Theorem 6, with probability at least $2/3$, the sample selected by the algorithm is an ϵ-net for X with respect to $\overline{\mathcal{B}}_{k,b}$. That is, for every $S \in \overline{\mathcal{B}}_{k,b}$ such that $|X \cap S| > \epsilon|X|$, the sample contains at least one point in S. (Note that here it is not true that for every S, $|X \cap S| > \epsilon|X|$, since X is assumed to be (k,b)-clusterable. However, this is immaterial to the claim.) Assume that in fact the sample is an ϵ-net for X. Then by definition of $\overline{\mathcal{B}}_{k,b}$, this means that for every k balls of radius b such that more than $\epsilon|X|$ points of X fall *outside* these balls, the sample contains such a point outside the balls. This in turn implies that for the k balls defined by the centers found by the algorithm, $z^1, \ldots, z^k$, there are at most $\epsilon|X|$ points in X that do not belong to these balls, and so the k centers induce an ϵ-good (k,b)-clustering of X.

Testing and the VC-dimension. The above analysis can be extended to obtain the following relation between the VC-dimension and Testing, very similarly to the way such a relation is obtained between the VC-dimension and PAC Learning.

Consider any property P of boolean functions over some domain Z, and let $\mathcal{F}_P$ be the class of functions having property P. A testing algorithm for property P is given query access to the tested function f (and in particular may ask for the value of f on a uniformly selected sample). If f has property P (that is, f belongs to $\mathcal{F}_P$) then the algorithm should accept. If f is ϵ-far from having

[4] In fact, the bound b on the radius of the balls can be used to obtain a bound of m^d. However, the reasoning is slightly more complicated.

property P (that is, for every function $g \in \mathcal{F}_P$, $\Pr[g(z) \neq f(z)] > \epsilon$, where the probability is over a uniformly selected z), the algorithm should reject with probability at least $2/3$.

In what follows we shall sometimes view boolean functions as sets. In particular, the VC-dimension of $\mathcal{F}_P$ is defined as the VC-dimension of the family of subsets: $\{S_f\}_{f \in \mathcal{F}_P}$ where $S_f \stackrel{\text{def}}{=} \{z : f(z) = 1\}$. Suppose there is an algorithm $\mathcal{A}$ that given a sample of labeled examples $\{z_i, b_i\}$ where $z_i \in Z$ and $b_i \in \{0, 1\}$, determines whether there exists a function in $\mathcal{F}_P$ that is consistent with the sample. That is, if $\exists g \in \mathcal{F}_P$, such that $g(z_i) = b_i$ for every i, then $\mathcal{A}$ outputs accept, and otherwise it outputs reject. We shall refer to $\mathcal{A}$ as a *consistency checker* for $\mathcal{F}_P$.

Theorem 8 *For any property P, a consistency checker for $\mathcal{F}_P$ can be used for testing P by applying it to a uniformly selected sample of size $m \geq \frac{8\text{VCD}(\mathcal{F}_P)}{\epsilon} \cdot \log \frac{8\text{VCD}(\mathcal{F}_P)}{\epsilon}$.*

Proof: The proof of Theorem 8 is a generalization of the proof of Theorem 6. By definition of a consistency checker, if $f \in \mathcal{F}_P$ (that is, f has property P), then it accepts. Let $\overline{\mathcal{F}_P} \stackrel{\text{def}}{=} \{\neg g : g \in \mathcal{F}_P\}$ (so that in particular $\text{VCD}(\overline{\mathcal{F}_P}) = \text{VCD}(\mathcal{F}_P)$). Then by Theorem 7, for any function f, with probability at least $2/3$, a sample of size m as stated in the theorem is an ϵ-net for f (i.e., S_f) with respect to $\overline{\mathcal{F}_P}$ (i.e. $\{S_g\}_{g \in \overline{\mathcal{F}_P}}$). As argued in the proof of Theorem 8, this implies that if f is ϵ-far from having property P, then it is rejected with probability at least $2/3$. ■

We note that in many cases (e.g. the property of monotonicity), the VC-dimension of the class of functions defined by the property is prohibitively large, and we seek other techniques (that in particular may use adaptive querying).

5.1 An Alternative Analysis for Radius Clustering

Here we present an alternative analysis of the radius clustering algorithm. Recall that Algorithm 5, which worked for $\beta = 0$, selected a sample of size roughly linear in d. Below we show that it is possible to trade the dependence on d (in terms of the sample complexity) with a dependence on $1/\beta$.

We start be analyzing the algorithm for $k = 1$

Algorithm 6 (L_2 metric, radius cost, $k = 1$, $d \geq 1$, $0 < \beta \leq 1$)

1. *Uniformly and independently select $m = \Theta\left(\frac{\log(1/\beta)}{\epsilon \cdot \beta}\right)$ points in X. Denote the set of points selected by U.*
2. *If $r(U) \leq b$ then accept, otherwise reject.*

Theorem 9 *Algorithm 6 is a radius-clustering tester for $k = 1$ under the L_2 metric.*

We shall prove the theorem by appealing to the following lemmas.

Lemma 9 *Let $S \subset \Re^d$, and let $y \in \Re^d$ and $\alpha > 0$ be such that $r(S \cup \{y\}) \leq r(S) \cdot (1 + \alpha)$. Then the distance between y and the center of the minimum sphere bounding S is at most $r(S) \cdot (1 + \alpha + \sqrt{\alpha^2 + 2\alpha})$.*

Proof: Let $r = r(S)$. Without loss of generality we may assume that the center of the minimum sphere bounding S is at the origin of our coordinate system and that the center of a sphere of

radius $r(1+\alpha)$ bounding $S \cup \{y\}$ is at $(\ell, 0, \ldots, 0)$, $\ell > 0$. In order to obtain the stated claim, it suffices, by the triangle inequality to show that $\ell \leq r \cdot \sqrt{\alpha^2 + 2\alpha}$.

Assume to the contrary that $\ell > r \cdot \sqrt{\alpha^2 + 2\alpha}$. For any point $q = (q_1, \ldots, q_d)$ in S, let q' denote the $(d-1)$-dimensional vector $(q_2, \ldots, q_d)$. Since S is bounded by a sphere of radius r centered at the origin, we have that

$$q_1^2 + \|q'\|_2^2 \leq r^2 \tag{4}$$

(where $\|q'\|_2^2 = \sum_{i=2}^{d} (q_i)^2$ denotes the L_2 norm squared of the vector q'). Since S is also bounded by a sphere of radius $r(1+\alpha)$ centered at $(\ell, 0, \ldots, 0)$, we have that

$$(q_1 - \ell)^2 + \|q'\|_2^2 \leq r^2(1+\alpha)^2. \tag{5}$$

Let $a = \frac{\ell^2 - r^2(\alpha^2+2\alpha)}{2\ell}$, so that by our counter assumption $a > 0$. We consider two cases.

Case 1: $q_1 \geq a$. In this case

$$\begin{aligned}
(q_1 - a)^2 + \|q'\|_2^2 &= q_1^2 + \|q'\|_2^2 - 2q_1 \cdot a + a^2 \\
&\leq r^2 - 2 \cdot a^2 + a^2 && (6) \\
&= r^2 - a^2 && (7)
\end{aligned}$$

where Equation (6) follows from Equation (4), our case assumption that $q_1 \geq a$, and from our counter hypothesis by which $a > 0$.

Case 2: $q_1 \leq a$. In this case

$$\begin{aligned}
(q_1 - a)^2 + \|q'\|_2^2 &= q_1^2 + \|q'\|_2^2 - 2q_1 \cdot a + a^2 \\
&= q_1^2 + \ell^2 - 2q_1\ell - \ell^2 + 2q_1\ell + \|q'\|_2^2 - 2q_1 \cdot a + a^2 \\
&= (q_1 - \ell)^2 + \|q'\|_2^2 + 2q_1(\ell - a) + a^2 - \ell^2 \\
&\leq r^2(1+\alpha)^2 + 2 \cdot a \cdot (\ell - a) + a^2 - \ell^2 && (8) \\
&= r^2(1+\alpha)^2 + 2a\ell - \ell^2 - a^2 \\
&= r^2(1+\alpha)^2 + (\ell^2 - r^2(\alpha^2 + 2\alpha)) - \ell^2 - a^2 && (9) \\
&= r^2 - a^2 && (10)
\end{aligned}$$

where Equation (8) follows from Equation (5) and the case assumption that $q_1 \leq a$, and Equation (9) follows from the definition of a.

Hence, in either case we get that S is contained in a sphere centered at $(a, 0, \ldots, 0)$ having radius strictly smaller than r, contradicting the minimality of r. ■

As a corollary to Lemma 9 we get:

Lemma 10 *Let $S \subset \Re^d$, and let $z \in \Re^d$ be the center of the minimum sphere bounding S. Consider any point $y \in \Re^d$ such that* $\mathrm{dist}(y, z) > t$ *for some $t \geq r(S)$. Then*

$$r(S \cup \{y\}) > r(S) \cdot \frac{1}{2} \cdot \left(\frac{t}{r(S)} + \frac{r(S)}{t} \right)$$

Proof: Assume, contrary to the claim that $r(S \cup \{y\}) \leq r(S) \cdot \frac{1}{2} \cdot \left(\frac{t}{r(S)} + \frac{r(S)}{t}\right)$. Let α be such that

$$r(S \cup \{y\}) = r(S) \cdot (1 + \alpha) \; .$$

By Lemma 9, this implies that

$$\mathrm{dist}(y, z) \leq r(S) \cdot (1 + \alpha + \sqrt{\alpha^2 + 2\alpha})$$

where z denotes the center of the minimum sphere bounding S. Let θ be such that $1 + \alpha = \cosh\theta$. Since $\sinh\theta = \sqrt{\cosh^2\theta - 1}$, we have that $\sqrt{\alpha^2 + 2\alpha} = \sinh\theta$. Since $\cosh\theta + \sinh\theta = e^{\theta}$, we get that

$$\mathrm{dist}(y, z) \leq r(S) \cdot e^{\theta} \; . \tag{11}$$

On the other hand, since $\cosh\theta - \sinh\theta = e^{-\theta}$, we have that

$$1 + \alpha = \frac{1}{2}\left(e^{\theta} + e^{-\theta}\right) \; . \tag{12}$$

But by definition of α and our assumption on the relation between $r(S \cup \{y\})$ and $r(S)$,

$$1 + \alpha \leq \frac{1}{2} \cdot \left(\frac{t}{r(S)} + \frac{r(S)}{t}\right) \tag{13}$$

and so $e^{\theta} + e^{-\theta} \leq \frac{t}{r(S)} + \frac{r(S)}{t}$. Since the function $f(x) = (x + 1/x)$ increases with x for $x \geq 1$, it follows that

$$e^{\theta} \leq \frac{t}{r(S)} \quad . \tag{14}$$

Combining Equation (14) with Equation (11) we get

$$\mathrm{dist}(y, z) \leq r(S) \cdot e^{\theta} \leq r(S) \cdot \frac{t}{r(S)} = t \tag{15}$$

contradicting the premise of the lemma that $\mathrm{dist}(y, z) > t$. ■

Proof of Theorem 9: If X is $(1, b)$-clusterable, then the algorithm clearly always accepts. Hence, assume from now on that X is ϵ-far from $(1, (1 + \beta)b)$-clusterable, and we shall show that the algorithm rejects with probability at least $2/3$.

We view the sample of size m as being selected in $p = \Theta(1/\beta)$ *phases*, where in each phase $\Theta(\log(p)/\epsilon)$ points are selected (uniformly and independently). Let U_i be the union of the samples selected in the first i phases, and let $r_i = r(U_i)$.

We show that with probability at least $2/3$ over the choice of the sample, for every phase i, $r_i \geq r_{i-1}(1 + \alpha_i)$ for some sufficiently large α_i. It will follow that after $O(1/\beta)$ phases, we must obtain that $r_i > b$, causing the algorithm to reject.

For each new phase i, let $z_{i-1} \in \Re^d$ be the center of the minimum sphere bounding U_{i-1}. Since X is ϵ-far from $(1, (1+\beta)b)$-clusterable, there are at least ϵn points $y \in X$ such that $\mathrm{dist}(y, z_{i-1}) > (1 + \beta)b$. Let us refer to these points as *influential* (with respect to U_{i-1}). Suppose that in each phase the sample taken is of size at least $\ln(3p)/\epsilon$. Then for any *fixed* phase, the probability that an influential point is not selected is at most $(1 - \epsilon)^{\ln(3p)/\epsilon} < \exp(-\ln(3p)) = 1/(3p)$. The probability that for *some* phase no influential point is selected is therefore less than $1/3$. Hence, assume from now on that for every phase i, the sample selected in this phase contains an influential point y with respect to U_{i-1}.

We now show that $O(1/\beta)$ phases suffice until $r_i > b$. Let y be an influential point with respect to U_{i-1}. By Lemma 10, $r(U_{i-1} \cup \{y\}) \geq r(U_{i-1}) \cdot \frac{1}{2}\left(\frac{(1+\beta)b}{r_{i-1}} + \frac{r_{i-1}}{(1+\beta)b}\right)$. Therefore, assuming such a point is selected in the i'th sample,

$$r_i \geq r_{i-1} \cdot \frac{1}{2}\left(\frac{(1+\beta)b}{r_{i-1}} + \frac{r_{i-1}}{(1+\beta)b}\right) = \frac{1}{2}\left((1+\beta)b + \frac{r_{i-1}^2}{(1+\beta)b}\right) . \tag{16}$$

We first observe that for every $i > 1$, $r_i \geq \frac{b}{2}$. Thus, we may assume that $r_2 \geq \frac{b}{2}$. On the other hand, since $\frac{(1+\beta)b}{r_{i-1}} + \frac{r_{i-1}}{(1+\beta)b}$ decreases as r_{i-1} increases, then as long as $r_{i-1} \leq b$,

$$r_i \geq r_{i-1} \cdot \frac{1}{2}\left(1+\beta+\frac{1}{1+\beta}\right) = r_{i-1} \cdot \left(1+\frac{\beta^2}{2(1+\beta)}\right) . \tag{17}$$

By applying this lower bound on the rate of increase of r_i (for $i > 2$), we get that after $O(1/\beta^2)$ phases, $r_i > b$. However, we can do a slightly more refined analysis and exploit the fact that for smaller radii, the rate of increase is greater. In particular, let γ be such that $r_{i-1} \leq b(1-\gamma)$. Given Equation (16), it can be shown (using simple manipulations), that for every $\gamma \leq \beta$,

$$r_i \geq r_{i-1} \cdot \frac{1}{2}\left(\frac{(1+\beta)}{(1-\gamma)} + \frac{(1-\gamma)}{(1+\beta)}\right) \geq r_{i-1} \cdot \left(1+\frac{\gamma^2}{2}\right) \tag{18}$$

For each integer $1 \leq a < \log(1/\beta)$, let $s(a)$ be the first phase such that $b \cdot (1-2^{-a}) \leq r_{s(a)} < b \cdot (1-2^{-(a+1)})$ (if there exists such a phase). We would like to upper bound the number of phases t required so that $r_{s(a)+t} \geq b \cdot (1-2^{-(a+1)})$. By Equation (18), as long as $r_i \leq (1-2^{-a+1})$ we have $r_{i+1} \geq r_i \cdot (1+2^{-(2(a+1)+1)})$. Therefore, we need t to be such that $(1-2^{-a}) \cdot (1+2^{-(2(a+1)+1)})^t \geq (1-2^{-(a+1)})$. Since for every $\delta \leq 1/2$, we have the bounds $(1-\delta) \geq \exp(-2\delta)$ and $(1+\delta) \geq \exp(\delta/2)$, it suffices that $t = 2^{a+4}$. It follows that the number of phases required to get from $r_2 \geq b/2$ to $r_i \geq b(1-\beta)$ is at most $16 \cdot \sum_{a=1}^{\log(1/\beta)} 2^a = O(1/\beta)$. Finally, to get from $r_i \geq b(1-\beta)$ to $r_{i+t} > b$, we use the bound from Equation (17), and conclude that it takes an additional $O(1/\beta)$ phases. ■

5.2 $k > 1$

The alternative radius-clustering tester for $k > 1$ is a generalization of the algorithm for the case $k = 1$:

Algorithm 7 (L_2 metric, radius cost, $k > 1$, $d \geq 1$, $0 < \beta \leq 1$)

1. *Uniformly and independently select a sample of* $m = \Theta\left(\frac{k^2 \log k}{\epsilon \cdot \beta^2}\right)$ *points in* X.
   Let us denote the points selected by U.
2. *If there* exists *a partition* $P = (U^1, \ldots, U^k)$ *of* U *such that* $R(P) \leq b$ *then accept. Otherwise, reject.*

The proof of the following theorem is analogous to the proof of Theorem 3, where here we use the arguments from the proof of Theorem 9 as a basis.

Theorem 10 *Algorithm 7 is a radius-clustering tester under the* L_2 *metric.*

6 Clustering in One Dimension

In one dimension the radius and diameter problems are the same, and it is possible to test clustering in an efficient way, for $\beta = 0$, and any L_p metric.

Algorithm 8 (L_p metric, radius and diameter cost, $k \geq 1$, $d = 1$, $\beta = 0$)

1. *Uniformly and independently select* $m = \Theta\left(\frac{k}{\epsilon} \cdot \log \frac{k}{\epsilon}\right)$ *points in* X.
2. *If the sample is* (k, b)*-clusterable then accept, otherwise, reject.*

The problem of clustering in one dimension can be solved by dynamic programming.

Theorem 11 *Algorithm 8 is a radius-clustering tester for* $d = 1$, $\beta = 0$, *and under any* L_p *metric.*

The proof of the Theorem follows directly from the following Lemma, and a standard "balls and bins" analysis.

Lemma 11 *Let* X *be* ϵ*-far from being* (k, b)*-clusterable. Then there exist* k *non-intersecting segments* $[left_i, right_i]$ *each of length* $2b$, *such that there are at least* $(\epsilon n)/(k+1)$ *points from* X *between every two segments, to the left of the leftmost segment and to the right of the rightmost segment.*

Proof: Let us assume for simplicity that X contains distinct points. The first and leftmost segment is placed such that there are $(\epsilon n)/(k+1)$ points from X to the left of it. Since X is ϵ-far from being (k, b)-clusterable, there must exist at least $(\epsilon nk)/(k+1)$ points to the right of this first segment. We thus place the second segment to the right of the first segment, such that there are $(\epsilon n)/(k+1)$ points from X between the two segments. The remaining segments are placed in a similar way. ■

Acknowledgments

We would like to thank Leonard Schulman and Micha Sharir for helpful information on clustering.

References

[1] P. K. Agrawal and C. M. Procopiuc. Exact and apprroximation algorithms for clustering. In *Proceedings of SODA*, pages 658–667, 1998.

[2] P. K. Agrawal and M. Sharir. Algorithms for geometric optimization. *ACM Computing Surveys*, pages 413–458, 1998.

[3] N. Alon, E. Fischer, M. Krivelevich, and M Szegedy. Efficient testing of large graphs. In *Proceedings of FOCS*, pages 645–655, 1999.

[4] N. Alon and M. Krivelevich. Testing k-colorability. Manuscript, 1999.

[5] N. Alon, M. Krivelevich, I. Newman, and M Szegedy. Regular languages are testable with a constant number of queries. In *Proceedings of FOCS*, pages 656–666, 1999.

[6] M. R. Anderberg. *Cluster Analysis for Applications.* Academic Press, 1973.

[7] M. Bender and D. Ron. Testing acyclicity of directed graphs in sublinear time. In *Proceedings of ICALP*, pages 809–820, 2000.

[8] M. Blum, M. Luby, and R. Rubinfeld. Self-testing/correcting with applications to numerical problems. *JACM*, 47:549–595, 1993.

[9] Y. Dodis, O. Goldreich, E. Lehman, S. Raskhodnikova, D. Ron, and A. Samorodnitsky. Improved testing algorithms for monotonocity. In *Proceedings of RANDOM*, pages 97–108, 1999.

[10] D. Hochabaum (Ed.). *Approximation Algorithms for NP-Hard Problems.* PWS Publishing Company, 1997.

[11] Z. Drezner (Ed.). *Facility Location.* Springer Verlag, 1995.

[12] F. Ergun, S. Kannan, S. R. Kumar, R. Rubinfeld, and M. Viswanathan. Spot-checkers. In *Proceedings of STOC*, pages 259–268, 1998.

[13] T. Feder and D. H. Greene. Optimal algorithms for approximate clustering. In *Proceedings of STOC*, pages 434–444, 1988.

[14] R. J. Fowler, M. S. Paterson, and S. L. Tanimoto. Optimal packing and covering in the plane are NP-complete. *IPL*, pages 133–137, 1981.

[15] A. Frieze and R. Kanan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[16] O. Goldreich, S. Goldwasser, E. Lehman, D. Ron, and A. Samordinsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000.

[17] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *JACM*, 45(4):653–750, 1998.

[18] O. Goldreich and D. Ron. Property testing in bounded degree graphs. In *Proceedings of STOC*, pages 406–415, 1997.

[19] O. Goldreich and D. Ron. A sublinear bipartite tester for bounded degree graphs. *Combinatorica*, pages 1–39, 1999.

[20] T. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, pages 1018–1035, 1985.

[21] D. Haussler and E. Welzl. ϵ-nets and simplex range queries. *Descrete and Computational Geometry*, pages 127–151, 1987.

[22] D. Hochbaum and D. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, pages 180–184, 1985.

[23] D. Hochbaum and D. Shmoys. A unified apprroach to approximation algorithms for bottleneck problems. *JACM*, pages 533–550, 1986.

[24] A. K. Jain and R. C. Dubes. *Algorithms for Clustering.* Prentice-Hall, 1988.

[25] J. Jolion, P. Meer, and S. Batauche. Robust clustering with applications in computer vision. *IEEE Trans. Pattern Analysis Mach. Intell.*, pages 791–802, 1991.

[26] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, 1990.

[27] M. Kearns and D. Ron. Testing problems with sub-learning sample complexity. In *Proceedings of COLT*, pages 268–277, 1998.

[28] R. Lupton, F. M. Maley, and N. Young. Data collection for the sloan digital sky survey: A network-flow approach. In *Proceedings of SODA*, pages 296–303, 1996.

[29] N. Megiddo and E. Zemel. A randomized $o(n \log n)$ algorithm for the weighted euclidean 1-center problem. *Journal of Algorithms*, pages 358–368, 1986.

[30] N. Mishra, D. Oblinger, and L. Pitt. Way-sublinear time approximate (pac) clustering. Unpublished manuscript, 2000.

[31] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[32] M. Parnas and D. Ron. Testing the diameter of graphs. In *Proceedings of RANDOM*, pages 85–96, 1999.

[33] C. M. Procopiuc. A survey on clustering. Available from `http://www.cs.duke.edu/~magda`, 2000.

[34] P. Raghavan. Information retrieval algorithms: A survey. In *Proceedings of SODA*, pages 11–18, 1997.

[35] E. A. Ramos. Deterministic algorithms for 3-d diameter and some 2-d lower envelopes. In *16th Symp. on Computational Geometry*, 2000. To appear.

[36] R. Rubinfeld. Robust functional equations and their applications to program testing. *SICOMP*, 28(6):1972–1997, 1999.

[37] R. Rubinfeld and M. Sudan. Robust characterization of polynomials with applications to program testing. *SICOMP*, 25(2):252–271, 1996.

[38] L. Schulman. Private communication, 2000.

[39] J. Shafer, R. Agrawal, and M. Mehta. Sprint: A scalable parallel classifier for data mining. In *Proceedings of the International Conference on Very Large Databases*, 1996.

[40] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its applications*, 17(2):264–280, 1971.