Helly-Type Theorems in Property Testing

Sourav Chakraborty[1], Rameshwar Pratap[1], Sasanka Roy[1], and Shubhangi Saraf[2]

[1] Chennai Mathematical Institute,
Chennai, India.
e-mail:{sourav,rameshwar,sasanka}@cmi.ac.in
[2] Department of Mathematics and Department of Computer Science,
Rutgers University.
e-mail:shubhangi.saraf@rutgers.edu

Abstract. Helly's theorem is a fundamental result in discrete geometry, describing the ways in which convex sets intersect with each other. If S is a set of n points in $\mathbb{R}^d$, we say that S is (k, G)-clusterable if it can be partitioned into k clusters (subsets) such that each cluster can be contained in a translated copy of a geometric object G. In this paper, as an application of Helly's theorem, by taking a constant size sample from S, we present a testing algorithm for (k, G)-clustering, *i.e.*, to distinguish between two cases: when S is (k, G)-clusterable, and when it is ϵ-far from being (k, G)-clusterable. A set S is ϵ-far $(0 < \epsilon \leq 1)$ from being (k, G)-clusterable if at least ϵn points need to be removed from S to make it (k, G)-clusterable. We solve this problem for $k = 1$ and when G is a symmetric convex object. For $k > 1$, we solve a *weaker* version of this problem. Finally, as an application of our testing result, in clustering with outliers, we show that one can find the *approximate* clusters by querying a constant size sample, with high probability.

1 Introduction

Given a set of n points in $\mathbb{R}^d$, deciding whether all the points can be contained in a unit radius ball is a well known problem in Computational Geometry. Of course, the goal is to solve this problem as quickly as possible. In order to solve this problem exactly, one has to look at all the n points in the worst case scenario. But if n is too large, an algorithm with linear running time may not be fast enough. Thus, one may be interested in "solving" the above problem by taking a very small size sample and outputting the "right answer" with high probability. In this paper, we consider the *promise* version of this problem. More precisely, for the given *proximity parameter* ϵ (where $0 < \epsilon \leq 1$), our goal is to distinguish between the following two cases:

- all the points can be contained in a unit radius ball,
- no unit radius ball can contain more than $(1 - \epsilon)$ fraction of points.

The above *promise* problem falls in the realm of property testing (see [10], [9] and [17]). In property testing, the goal is to look at a very small fraction of the input and decide whether the input satisfies the property or is "far" from satisfying it. Property testing algorithms for computational geometric problems have been studied earlier in

[6], [5] and [1]. In this paper, we study the above problem in property testing setting and give a simple algorithm to solve it. The algorithm queries only a constant number of points (where the constant depends on the dimension d and ϵ, but is independent of n) and correctly distinguishes between the two cases mentioned above with probability at least $2/3$. While the algorithm is very simple, the proof of correctness is a little involved, for which we use Helly's theorem. Helly's theorem ([11]) states that if a family of convex sets in $\mathbb{R}^d$ has a non-empty intersection for every $d+1$ sets, then the whole family has a non-empty intersection. In fact, since Helly's theorem also works for symmetric convex bodies, we can solve the above problem for any symmetric convex body instead of just a unit radius ball. Thus, we have

Theorem 1. *Let A be a symmetric convex body. If S is a set of n points in $\mathbb{R}^d$ as input with the proximity parameter ϵ (where $0 < \epsilon \leq 1$), then there is an algorithm $\mathcal{A}$ that randomly samples $O(\frac{d}{\epsilon^{d+1}})$ many points and*

- *$\mathcal{A}$ accepts, if all the points in S can be contained in a translated copy of A,*
- *$\mathcal{A}$ rejects with probability $\geq 2/3$, if any translated copy of A can contain at most $(1-\epsilon)n$ points.*

The running time of $\mathcal{A}$ is $O(\frac{d}{\epsilon^{d+1}})$.

One would like to generalize the above problem for more than one object, *i.e.*, given k translated copies of object B, the goal is to distinguish between the following two cases with high probability:

- all n points can be contained in k translated copies of B,
- at least ϵ fraction of points cannot be contained in any k translated copies of B.

We would like to conjecture that a similar algorithm, as stated in Theorem 1, would also work for the generalized k object problem. Unfortunately, Helly's theorem does not hold for the k object setting, but we would like to conjecture that a version of the Helly-type theorem does hold for this setting. Assuming the above conjecture, we can obtain a similar algorithm for the k object setting. We can also unconditionally solve a *weaker* version of the k object problem.

Connection to Clustering: We can also view this problem in the context of clustering. Clustering ([15],[12], [2]) is a common problem that arises in the analysis of large data sets. In a typical clustering problem, we have a set of n input points in d dimensional space and our goal is to partition the points into k clusters. There are two ways to define the cluster size (cost):

- the maximum pairwise distance between an arbitrary pair of points in the cluster,
- twice the maximum distance between a point and a chosen centroid.

The first one is called as k-center clustering for diameter cost and the second one is called as k-center clustering for radius cost. In the k-center problem, our goal is to minimize the maximum of these distances. Computing k-center clustering is NP-hard: even for 2 clusters in general Euclidean space (of dimension d); and also for general number of k clusters even on a plane.

In this paper, we assume that the cluster can be of symmetric convex shape also. Given a set S of n points and a symmetric convex body A in $\mathbb{R}^d$, we say that the set of points is (k, A)-clusterable if all the points can be contained in k translated copies of A. In the *promise* version of the problem, for a given proximity parameter ϵ (where $0 < \epsilon \leq 1$), our goal is to distinguish between the cases when S is (k, A)-clusterable and when it is ϵ-far from being (k, A)-clusterable. We say that S is ϵ-far from being (k, A)-clusterable if at least ϵn points need to be removed from S in order to make it (k, A)-clusterable.

We solve the above problem for $k = 1$ with constant number of queries. For $k > 1$, we solve a *weaker* version of the problem. In order to solve the *promise* version of the problem, we have designed a randomized algorithm which is generally called as *tester*.

Our algorithms can also be used to find an *approximately good* clustering. In clustering with outliers (anomalies), when we have the ability to ignore some points as outliers, we present a randomized algorithm that takes a constant size sample from input and outputs radii and centers of the clusters. The benefit of our algorithm is that we construct an *approximate* representation of such clustering in time which is independent of the input size.

The most interesting part of our result is that we initiate application of Helly-type theorem in property testing in order to solve the clustering problem.

1.1 Other related work

Alon *et al.* [1] presented testing algorithm for (k, b)-*clustering*. A set of points is said to be (k, b)-*clusterable* if it can be partitioned into k *clusters*, where radius (or diameter) of every cluster is at most b. Section 5 of [1] presents a testing algorithm for radius cost under the L_2 metric. The analysis of this algorithm can be easily generalized to any metric under which each cluster is determined by a *simple* convex set (a convex set in $\mathbb{R}^d$ is called *simple* if its VC-dimension is $O(d)$).

For testing 1-center clustering, our result and the result from [1] give constant query testing algorithms. Although the two results have incomparable query complexity (in terms of number of queries depending on ϵ), for testing k-center clustering, we give a *weaker* query complexity algorithm which works for fixed k and d, and for $\epsilon \in (\epsilon', 1]$ where $\epsilon' = \epsilon'(k, t)$ (where t is a constant which depends on the shape of the geometric object). Alon *et al.* used the sophisticated VC-dimension technique while we have used Helly-type results.

1.2 Organization of the paper

In Section 2, we introduce the notations, definitions and state Helly and *Helly-type* theorems that are used in this paper. In Section 3, we design the *tester* for $(1, A)$-cluster testing for a given symmetric convex body A. In Section 4, we design the *tester* for (k, G)-cluster testing for a given geometric object G. In Section 5, as an application of results from Sections 3 and 4, we present an algorithm to find *approximate* clusters with outliers.

2 Preliminaries

2.1 Definitions

n-piercing: *A family of sets is called n-pierceable if there exists a set S of n points such that each member of the family has a non-empty intersection with S.*
Homotheticity: *Let A and B be two geometric bodies in $\mathbb{R}^d$. A is homothetic to B if there exist $v \in \mathbb{R}^d$ and $\lambda > 0$ such that $A = v + \lambda B$ (where λ is called scaling factor of B). In particular, when $\lambda = 1$, A is said to be a **translated copy** of B.*
Symmetric convex body: *A convex body A is called symmetric if it is centrally symmetric with respect to the origin, i.e., a point $v \in \mathbb{R}^d$ lies in A if and only if its reflection through the origin $-v$ also lies in A. In other words, for every pair of points $v_1, v_2 \in \mathbb{R}^d$, if $v_1 \in v_2 + A$, then $v_2 \in v_1 + A$ and vice versa. Circles, ellipses, n-gons (for even n) with parallel opposite sides are examples of symmetric convex bodies.*

2.2 Property Testing

In property testing, the goal is to query a very small fraction of the input and decide whether the input satisfies a certain predetermined property or is "far" from satisfying it. Let $x = \{0,1\}^n$ be a given input string. Then, a property testing algorithm, with query complexity $q(|x|)$ and proximity parameter ϵ for a decision problem L, is a randomized algorithm that makes at most $q(|x|)$ queries to x and distinguishes between the following two cases:

- if x is in L, then the algorithm *Accepts* x with probability at least $\frac{2}{3}$,
- if x is ϵ-far from L, then the algorithm *Rejects* x with probability at least $\frac{2}{3}$.

Here, "x is ϵ-far from L" means that the Hamming distance between x and any string in L is at least $\epsilon|x|$. A property testing algorithm is said to have *one-sided error* if it satisfies the stronger condition that the accepting probability for instances $x \in L$ is 1 instead of $\frac{2}{3}$.

2.3 Helly's and Fractional Helly's Theorem

In 1913, Eduard Helly proved the following theorem:

Theorem 2. *(Helly's Theorem [11]) Given a finite family of convex sets $C_1, C_2, ..., C_n$ in $\mathbb{R}^d$ (where $n \geq d+1$) such that if intersection of every $d+1$ of these sets is non-empty, then the whole collection has a non-empty intersection.*

Katchalski and Liu proved the following result which can be viewed as a fractional version of the Helly's Theorem.

Theorem 3. *(Fractional Helly's Theorem [16]) For every α (where $0 < \alpha \leq 1$), there exists $\beta = \beta(d, \alpha)$ with the following property. Let $C_1, C_2, ..., C_n$ be convex sets in $\mathbb{R}^d$ (where $n \geq d+1$) and if at least $\alpha\binom{n}{d+1}$ of the collection of subfamilies of size $d+1$ has a non-empty intersection, then there exists a point contained in at least βn sets.*

Independently, Kalai [8] and Eckhoff [13] proved that $\beta(d, \alpha) = 1 - (1-\alpha)^{\frac{1}{(d+1)}}$.

2.4 *Helly-type* theorem for more than one piercing in convex bodies

Helly's theorem on intersections of convex sets focuses on 1-pierceable families. Danzer *et al.* [7] investigated the following Helly-type problem : If d and m are positive integers, what is the least $h = h(d, m)$ such that a family of boxes (with parallel edges) in $\mathbb{R}^d$ is m-pierceable if each of its h-membered subfamilies is m-pierceable? Following is the main result of their paper:

Theorem 4. *1.* $h(d, 1) = 2$ *for all* d *(where* $d \geq 1$*);*
2. $h(1, m) = m + 1$ *for all* m*;*
3. $h(d, 2) = \begin{cases} 3d \text{ for odd } d; \\ 3d - 1 \text{ for even } d; \end{cases}$
4. $h(2, 3) = 16$*;*
5. $h(d, m) = \infty$ *for* $d \geq 2, n \geq 3$ *and* $(d, m) \neq (2, 3)$.

Katchalski *et al.* proved a result for families of homothetic triangles in a plane ([14]). This result is similar to the intersection property of axis parallel boxes in $\mathbb{R}^d$, studied by Danzer *et al.* This result can also be considered as a Helly-type theorem for more than one piercing of convex bodies. Theorem 5, below, presents the main result of their paper.

Theorem 5. *Let* $\mathcal{T}$ *be a family of homothetic triangles in a plane. If any nine of them can be pierced by two points, then all the members of* $\mathcal{T}$ *can be pierced by two points.*

3 Robust Helly for one piercing of symmetric convex body

Helly's theorem is a fundamental result in discrete geometry, describing the ways in which convex sets intersect with each other. In our case, we will focus on those subset of convex sets whose intersection properties behave *symmetric* in certain ways. Observation 6 explains this in detail. In order to design the *tester* for $(1, A)$-cluster testing problem, we will crucially use this observation, Helly's and fractional Helly's theorem.

Observation 6 *Let* A *be a symmetric convex body in* $\mathbb{R}^d$ *containing* n *points, then* n *translated copies of* A *centered at these* n *points have a common intersection. Moreover, a translated copy of* A *centered at a point in the common intersection contains all these* n *points.*

Lemma 7 *Given a set* S *of* n *points in* $\mathbb{R}^d$*, if every* $d + 1$ *(where* $d + 1 \leq n$*) of them are contained in (a translated copy of) a symmetric convex body* A*, then all the* n *points are contained in (a translated copy of)* A*.*

Proof: Consider a set $\mathcal{B}$ of translated copies of A centered at points in S. Since every $d + 1$ of the given points are contained in (a translated copy of) A, by Observation 6, every $d + 1$ elements in $\mathcal{B}$ has a non-empty intersection. By Helly's theorem, all elements in $\mathcal{B}$ have a non-empty intersection. Let q be a point from this intersection. Then q belongs to every element in $\mathcal{B}$ and hence, by Observation 6, all the centers of the elements in $\mathcal{B}$, *i.e.*, all the n points in S, are contained in (a translated copy of) A centered at q. □

Lemma 8 *Let S be a set of n points in $\mathbb{R}^d$ (where $n \geq d+1$). If at least ϵn (where $0 < \epsilon \leq 1$) points cannot be contained in any translated copy of a symmetric convex body A, then at least ϵ^{d+1} fraction of all the $d+1$ size subsets of S (number of such subsets is $\binom{n}{d+1}$) cannot be contained in any translated copy of A.*

Proof: Consider a set $\mathcal{B}$ of translated copies of A centered at points in S. Now, by fractional Helly's theorem, for every α (where $0 < \alpha \leq 1$), there exists $\beta = \beta(d, \alpha)$ such that if at least an α fraction of $\binom{n}{d+1}$ subsets (of size $d+1$) in $\mathcal{B}$ has a non-empty intersection, then there exists a point (say p) which is contained in at least β fraction of elements of $\mathcal{B}$.

Consider a translated copy of A centered at p. By Observation 6, for every α (where $0 < \alpha \leq 1$), there exists $\beta = \beta(d, \alpha)$ such that if at least an α fraction of $\binom{n}{d+1}$ subsets (of size $d+1$) in S are contained in A, then at least βn points are contained in A.

Thus, if at least $(1-\beta)n$ points cannot be contained in A, then at least $1-\alpha$ fraction of $\binom{n}{d+1}$ subsets (of size $d+1$) in S cannot be contained in A. (Contrapositive of the above statement.)

Since $\beta = 1 - (1-\alpha)^{\frac{1}{(d+1)}}$ ([8], [13]), choosing $1-\beta$ as ϵ makes $1-\alpha$ equal to ϵ^{d+1}, which are the required values of the parameters. □

Theorem 9. *Consider a set of n points in $\mathbb{R}^d$ ($n \geq d+1$) located such that at least ϵn (where $0 < \epsilon \leq 1$) points cannot be contained in any translated copy of a symmetric convex body A. If we randomly sample $\frac{1}{\epsilon^{d+1}} \ln \frac{1}{\delta}$ (where $0 < \delta \leq 1$) many sets of $d+1$ points, then there exists a set in the sample which cannot be contained in any translated copy of A, with probability at least $1-\delta$.*

Proof: By Lemma 8, if at least ϵn points cannot be contained in (any translated copy of) A, then at least ϵ^{d+1} fraction of $\binom{n}{d+1}$ sets (of size $d+1$) cannot be contained in (any translated copy of) A. A set of $d+1$ points cannot be contained in A with probability ϵ^{d+1}. Hence, the probability that it can be contained in A is $1-\epsilon^{d+1}$. Thus, the probability that all the sampled sets are contained in A is $\leq (1-\epsilon^{d+1})^{\frac{1}{\epsilon^{d+1}} \ln \frac{1}{\delta}} \leq e^{-\ln \frac{1}{\delta}} = \delta$. □

Algorithm 1 is a randomized algorithm, *tester*, for $(1, A)$-cluster testing problem.

Data: A set S of n points in $\mathbb{R}^d$ (input is given as black-box), $0 < \delta, \epsilon \leq 1$.
Result: Returns a set of $d+1$ points, if it exists, which cannot be contained in A or accepts (*i.e.*, all the points can be contained in A).

```
repeat
    select a set (say W) of d+1 points uniformly at random from S
    if W cannot be contained in A then
        return W as witness
    end
until (1/ε^{d+1}) ln(1/δ) many times;
if no witness found then
    return /* all the points can be contained in A */
end
```

Algorithm 1: $(1, A)$-cluster testing in a symmetric convex body A

This algorithm has a one sided error, *i.e.*, if all the points can be contained in a symmetric convex body A then it accepts the input, else it outputs a witness with probability at least $1-\delta$. Correctness of the algorithm follows from Theorem 9. Thus, in the problem of testing $(1, A)$-clustering for a symmetric convex body A, the sample size is independent of the input size and hence the property is *testable*. Moreover, the *tester* works for all the possible values of ϵ (for $0 < \epsilon \leq 1$).

4 Robust Helly for more than one piercing of convex bodies

4.1 Helly-type results for more than one piercing of convex bodies

The following lemma says that a *"Helly-type"* result is not true for circles even for 2-piercing. The result can be easily generalized for higher dimensions also. (The proof of the following lemma was suggested by Prof. Jeff Kahn in a private communication.)

Lemma 10 *Consider a set of n circles in a plane. For any constant w (where $w < n$), the condition that every w circles are pierced at two points is not sufficient to ensure that all the circles in the set are pierced at two points.*

We present a proof of the above lemma in the full version of this paper [3].

Using arguments similar to the proof of above lemma, it is easy to prove that a *"Helly-type"* result for more than one piercing is also not true for a set of translated ellipsoids. Katchalski *et al.* [14] and Danzer *et al.* [7] proved a *"Helly-type"* result for more than one piercing of triangles and boxes, respectively. According to [14], a *"Helly-type"* result for more than one piercing is not true for centrally symmetric hexagon (with parallel opposite edges). Similar type of result is true for triangles and pentagons (with pair of parallel edges) which are not symmetric convex bodies. Thus, among symmetric convex bodies (spheres, ellipsoids and n-gons (for $n \leq 6$)), a *"Helly-type"* result for more than one piercing is possible only for parallelograms. We have following observation regarding the same (we present a proof in the full version of this paper [3]):

Observation 11 *Let S be a set of n points in $\mathbb{R}^d$. If every set of h points (for finite possible values of h, see Theorem 4) in S is contained in m (where $m > 0$) translated parallelograms, then all the n points are contained in m translated parallelograms.*

4.2 Fractional Helly for more than one piercing of convex bodies

We now design a *weaker* version of *tester* for (k, G)-clustering (where G is a bounded geometric object and $k > 1$). The tester works for some particular value of $\epsilon \in (\epsilon'(k,t), 1]$, where t is some constant that depends on the shape of geometric object.

We state the following conjecture for more than one piercing of convex bodies.

Conjecture 12 *For every α (where $0 < \alpha \leq 1$), there exists $\beta = \beta(\alpha, k, d)$ with the following property. Let $C_1, C_2, .., C_n$ be convex sets in R^d, $n \geq k(d+1)$, such that at least $\alpha.\binom{n}{k(d+1)}$ of the collection of subfamilies of size $k(d+1)$ are pierced at k points, then at least βn sets are pierced at k points. Also, β approaches 1 as α approaches 1.*

Lemma 13 *If Conjecture 12 is true, then we have the following: Consider a set of n points in $\mathbb{R}^d$ (where $n \geq k(d+1)$). If at least ϵn (where $0 < \epsilon \leq 1$) points cannot be contained in any k translated copies of symmetric convex body A, then at least $\gamma(\beta(\epsilon, k, d))$ fraction of $\binom{n}{k(d+1)}$ sets cannot be contained in any k translated copies of A.*

Proof: Proof of this lemma is similar to the proof of Lemma 8. □

In the above lemma, γ is an appropriately chosen function to compute the value of $1-\alpha$, *i.e.*, the fraction of $\binom{n}{k(d+1)}$ sets which cannot be contained in k translated copies of A.

Now, we prove a *weaker* version of Conjecture 12. We show that for bounded geometric objects, a weaker version of fractional Helly for more than one piercing is true. We use *greedy* approach to prove the same. We prove it for some $\epsilon \in (\epsilon', 1]$, where $\epsilon' = \epsilon'(k, t)$ (where t is a constant that depends on the shape of the geometric object). The result is true only for constant k and d.

Lemma 14 *Consider k translated copies of a geometric object G and a set of n points in $\mathbb{R}^d$ (for constant k and d). Then there exist $\epsilon' = \epsilon'(k, t)$ (where $\epsilon'(k, t) = 1 - \frac{1}{2(t+1)(k+1)}$, t is a constant that depends on the shape of the geometric object) such that for all $\epsilon \in (\epsilon', 1]$, if at least ϵn points cannot be contained in any k translated copies of G, then there exist at least $\Omega(n^{k+1})$ many witnesses of $k+1$ points which cannot be contained in any k translated copies of G.*

Proof: We say a geometric object G is *best* if it encloses the maximum number of points from the given set of n points. Now, we start with such a best object. Let us say the best object contains at least $c_0(1-\epsilon)n$ points (where $0 < c_0 \leq 1$). Now draw an object, L_G, concentric and homothetic with respect to G, having a scaling factor of $2+\varepsilon$ (for $0 < \varepsilon \ll 1$, see the definition of Homotheticity in Subsection 2.1 where $v = 0$ and $\lambda = 2+\varepsilon$). The annulus obtained by two concentric objects G and L_G can be filled with constant many (say $t\ (= \kappa^d - 1)$, [3] we present a proof of this in the full version of this paper [3]) translated copies of G. Since we started with the best object, the annulus contains at most $tc_0(1-\epsilon)n$ points. Hence, the number of points which are outside L_G is at least $\epsilon n - tc_0(1-\epsilon)n = \epsilon_1 n$, where $\epsilon_1 = \epsilon - tc_0(1-\epsilon)$. We throw away all the points in the annulus. Now, we are left with best object that containing at least $c_0(1-\epsilon)n$ points and the remaining space containing at least $\epsilon_1 n$ points.

Now, we repeat the above process on $\epsilon_1 n$ points and would keep on repeating it until every point is either deleted or contained in some translated copies of G. Thus, total number of points that we have deleted from annuli is at most $t\Sigma_{i\geq 0}c_i(1-\epsilon_i)n$ and total number of points that are inside translated copies of G is at least $\Sigma_{i\geq 0}c_i(1-\epsilon_i)n$ (where $\epsilon_0 = \epsilon$).

By construction, the total number of points inside translated copies of G and the points that have been deleted from annuli is at least n. Thus,

$$\Sigma_{i\geq 0}c_i(1-\epsilon_i)n + t\Sigma_{i\geq 0}c_i(1-\epsilon_i)n \geq n \quad (\text{where } \epsilon_0 = \epsilon).$$

[3] κ is (ceiling of) the ratio of side length of the smallest d-cube circumscribing L_G to that of the largest d-cube (homothetic *w.r.t.* smallest d-cube circumscribing L_G) inscribing G.

$$\Sigma_{i\geq 0} c_i(1-\epsilon_i)n \geq \frac{n}{t+1}.$$

Let G_i denotes the i-th geometric object and $|G_i|$ denotes the number of points contained in it. Thus,

$$\Sigma_{i\geq 0} |G_i| \geq \frac{n}{t+1}.$$

By assumption, k translated copies of G can contain at most $(1-\epsilon)n$ points.
Thus, $|G_i| \leq (1-\epsilon)n$. Since $\epsilon > 1 - \frac{1}{2(t+1)(k+1)}$,

$$|G_i| < \frac{n}{2(t+1)(k+1)}.$$

Now, our goal is to make $k+1$ buckets, $S_1, S_2, .., S_{k+1}$, from G_i's such that each bucket contains at least $\frac{n}{2(t+1)(k+1)}$ points and at most $\frac{n}{(t+1)(k+1)}$ points. We construct these buckets by adding points from G_i's until its size become at least $\frac{n}{2(t+1)(k+1)}$. Since each $|G_i| < \frac{n}{2(t+1)(k+1)}$ and $\Sigma_{i\geq 0}|G_i| \geq \frac{n}{t+1}$, this construction is possible. Thus, for a particular bucket S_i,

$$\frac{n}{2(t+1)(k+1)} \leq |S_i| \leq \frac{n}{(t+1)(k+1)}.$$

Now, choosing one point from each of the $(k+1)$ buckets gives a set of $k+1$ points as a witness, which cannot be contained in k translated copies of G. Thus, there are at least $\left(\frac{1}{2(t+1)(k+1)}\right)^{k+1} n^{k+1}$ $(= \Omega(n^{k+1}))$ many witnesses. □

Theorem 15. *Consider k translated copies of a geometric object G and a set of n points in $\mathbb{R}^d$ (for constant k and d). Then there exist $\epsilon' = \epsilon'(k,t)$ (where t is a constant that depends on the shape of the geometric object) such that for all $\epsilon \in (\epsilon', 1]$, at least ϵn points cannot be contained in any k translated copies of G. Now, if we randomly sample $\frac{1}{c}\ln\frac{1}{\delta}$ (where $0 < \delta \leq 1$ and cn^{k+1} is the number of witnesses, see Lemma 14) many sets of size $k+1$, then there exists a set in the sample which cannot be contained in any k translated copies of G, with probability at least $1-\delta$.*

We present a proof of the above theorem in the full version of this paper [3].

Similar to *tester* for $(1, A)$-cluster testing problem, we present a *tester* (Algorithm 2) for problem (k, G)-cluster testing. If all the points can be contained in k translated copies of G then algorithm accepts the input, else it outputs a witness with probability at least $1-\delta$. Correctness of the algorithm follows from Theorem 15. Thus, similar to testing $(1, A)$-clustering, this property is also *testable*. But, the *tester* only works for

constant k and d and for $\epsilon \in (\epsilon', 1]$ (see Lemma 14).

Data: A set S of n points in $\mathbb{R}^d$ (input is given as black-box), $0 < \delta \leq 1$ and $\epsilon \in (\epsilon', 1]$.
Result: Returns a set of $k+1$ points, if it exists, which cannot be contained in k translated copies of G, or accepts (*i.e.*, all the points can be contained in it).

repeat
  select a set (say W) of $k+1$ points uniformly at random from S
  if *W cannot be contained in k translated copies of G* **then**
    return W as witness
  end
until $\frac{1}{c} \ln \frac{1}{\delta}$ *many times*;
if *no witness found* **then**
  return /* *all the points can be contained in k translated copies of G* */
end

Algorithm 2: (k, G)-cluster testing in geometric objects

5 Application in Clustering with Outliers

While considering the clustering problem, we mostly assume that data is perfectly clusterable. But a few random points (outliers, noise) could be added in the data by an adversary. For example, in the k-center clustering, if an adversary adds a point in the data which is very far from the original set of well clustered points, then in the optimum solution that point becomes center of its own cluster and the remaining points are forced to clustered with $(k-1)$ centers only. Also, it is even difficult to locate when a point becomes an outlier. For example: consider a set of points where we need to find its optimal k-center clustering. Take a point from that set and keep moving it far from the remaining set. Now, it is very difficult to locate correctly at which place that point becomes center of its own cluster and the remaining points are left with $(k-1)$-center clusters.

In this work, we consider clustering with outliers by ignoring some fraction of points. Thus, in the case when points are perfectly clusterable, ignoring some fraction of points does not affect the result too much, and the case when outliers are present, the algorithm has the ability to ignore them while computing the final clusters. It may seems that the ability to ignore some fraction of points makes the problem easier, but on the contrary it does not. Because it has not only to decide which point to include in the cluster but also to decide which point to include first. There may be two extreme approaches to solve this problem: 1) Decide which points are outliers and run the clustering algorithm; 2) Do not ignore any points, and after getting final clusters decide which ones are outliers. Unfortunately, neither of these two approaches works well. The first one scales poorly because there are exponentially many choices, and the second one may significantly change the final outcome when outliers are indeed present. This motivates the study of clustering with outliers (see[4]).

Theorem 9 has an application to 1-center clustering with outliers. More precisely, for $0 < \epsilon, \delta \leq 1$, when we have the ability to ignore at least ϵn points as outliers, we present a randomized algorithm which takes a constant size sample from input and correctly output the radius and center of the *approximate* cluster with probability at least $1 - \delta$.

Data: A set S of n points in $\mathbb{R}^d$ (input is given as black-box), $0 < \epsilon, \delta \leq 1$.
Result: Report center and radius of cluster which contain all but at most ϵn points.
1 Uniformly and independently, select $m = \frac{d+1}{\epsilon^{d+1}} \ln \frac{1}{\delta}$ points from S.
2 Compute minimum enclosing ball containing all the sample points and report its center and radius.

Algorithm 3: 1-center clustering with outliers

Theorem 16. *Given a set of n points in $\mathbb{R}^d$ and $0 < \epsilon, \delta \leq 1$, Algorithm 3 correctly outputs, with probability at least $1 - \delta$, a ball containing all but at most ϵn points in constant time by querying a constant size sample (constant depending on d and ϵ). Moreover, if $r_{outlier}$ is the smallest ball containing all but at most ϵn points and r_{min} is the smallest ball containing all the points, then Algorithm 3 outputs the radius r such that $r_{outlier} \leq r \leq r_{min}$.*

We present a proof of the above theorem in the full version of this paper [3].

The problem of clustering with outliers can be generalized for k-center clustering. If Conjecture 12 is true, then it has an application to k-center clustering with outliers. For given $0 < \epsilon, \delta \leq 1$, ignoring at least ϵn points as outliers, we present a randomized algorithm which takes a constant size sample from the input and correctly output the radii and k centers of the *approximate* clusters with probability at least $1 - \delta$.

Data: A set S of n points in $\mathbb{R}^d$ (input is given as black-box), $0 < \epsilon, \delta \leq 1$.
Result: Reports k centers and radii of clusters which contains all but at most ϵn points.
1 Uniformly and independently, select $m = \frac{k(d+1)}{\gamma(\beta(\epsilon,k,d))} \ln \frac{1}{\delta}$ points from S.
2 Compute k minimum enclosing balls containing all the sample points and report their centers and radii.

Algorithm 4: k-center clustering with outliers

Theorem 17. *Consider a set of n points in $\mathbb{R}^d$. If Conjecture 12 is true and $0 < \epsilon, \delta \leq 1$, then with probability at least $1 - \delta$, Algorithm 4 output k balls containing all but at most ϵn points in constant time by querying a constant size sample (constant depending on k, d and ϵ). Moreover, for $1 \leq i \leq k$, if $r^{(i)}_{outlier}$ is the radius of the optimal i-th cluster by ignoring at most ϵn points as outliers and $r^{(i)}_{min}$ is the radius of the optimal i-th cluster when all points are present, then Algorithm 4 outputs the radius $r^{(i)}$ such that $r^{(i)}_{outlier} \leq r^{(i)} \leq r^{(i)}_{min}$.*

We present a proof of the above theorem in the full version of this paper [3].

6 Conclusion and Open Problems

In this paper, we initiated an application of the Helly (and *Helly-type*) theorem in property testing. For $(1, A)$-cluster testing in a symmetric convex body A, we showed that testing can be done with constant number of queries and hence proved that the property is *testable*. Alon *et al.* [1] also solved a similar problem with constant number of queries, using combination of sophisticated arguments in geometric and probabilistic analysis. For 1-center clustering, our result had an incomparable query complexity in relation (in terms of number of queries depending on ϵ) with the result of Alon *et al*. We stated a conjecture related to fractional *Helly-type* theorem for more than one piercing of convex bodies. Using a greedy approach, we proved a weaker version of the conjecture which we used for testing (k, G)-clustering. We also gave a characterization of the type of symmetric convex body for which Helly-type result for more that one piercing would be true. Finally, as an application of testing result in clustering with outliers, we showed that one can find, with high probability, the *approximate* clusters by querying a constant size sample.

References

1. Noga Alon, Seannie Dar, Michal Parnas, and Dana Ron. Testing of clustering. *SIAM J. Discrete Math.*, 16(3):393–417, 2003.
2. M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press, 1973.
3. Sourav Chakraborty, Rameshwar Pratap, Sasanka Roy, and Shubhangi Saraf. Helly-type theorems in property testing. *CoRR*, abs/1307.8268, 2013.
4. Moses Charikar, Samir Khuller, David M. Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. pages 642–651, 2001.
5. Artur Czumaj and Christian Sohler. Property testing with geometric queries. In *ESA*, pages 266–277, 2001.
6. Artur Czumaj, Christian Sohler, and Martin Ziegler. Property testing in computational geometry. In *ESA*, pages 155–166, 2000.
7. L. Danzer and B. Grünbaum Branko. Intersection properties of boxes in $\mathbb{R}^d$. *Combinatorica*, 2(3):237–246, 1982.
8. J. Eckhoff. An upper bound theorem for families of convex sets. *Geom. Dediata 19*, (75):217–227, 1985.
9. Oded Goldreich. Combinatorial property testing (a survey). *Electronic Colloquium on Computational Complexity (ECCC)*, 4(56), 1997.
10. Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.
11. E. Helly. Über Mengen konvexer Köper mit gemeinschaftlichen Punkten (germen). *Jahresber. Deutsch.Math. Verein.*, (32):175–176, 1923.
12. A. K. Jain and R. C. Dubes. *Algorithms for Clustering*. Prentice-Hall, 1988.
13. G. Kalai. Intersection patterns of convex sets. *Israel J. Math.*, (48):161–174, 1984.
14. M. Katchalski and D. Nashtir. On a conjecture of danzer and grunbaum. *Proc. A.M.S*, (124):3213–3218, 1996.
15. L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.
16. M.Katchalski and A. Liu. A problem of geometry in $\mathbb{R}^n$. *Proc. A.M.S*, (75):284–288, 1979.
17. Dana Ron. Property testing: A learning theory perspective. *Foundations and Trends in Machine Learning*, 1(3):307–402, 2008.