VC Dimension and Distribution-Free Sample-Based Testing*

Eric Blais†
University of Waterloo
Waterloo, Canada
eric.blais@uwaterloo.ca

Renato Ferreira Pinto Jr.
Google Canada
Waterloo, Canada
r4ferrei@uwaterloo.ca

Nathaniel Harms‡
University of Waterloo
Waterloo, Canada
nharms@uwaterloo.ca

ABSTRACT

We consider the problem of determining which classes of functions can be tested more efficiently than they can be learned, in the distribution-free sample-based model that corresponds to the standard PAC learning setting. Our main result shows that while VC dimension by itself does not always provide tight bounds on the number of samples required to test a class of functions in this model, it can be combined with a closely-related variant that we call "lower VC" (or LVC) dimension to obtain strong lower bounds on this sample complexity.

We use this result to obtain strong and in many cases nearly optimal bounds on the sample complexity for testing unions of intervals, halfspaces, intersections of halfspaces, polynomial threshold functions, and decision trees. Conversely, we show that two natural classes of functions, juntas and monotone functions, can be tested with a number of samples that is polynomially smaller than the number of samples required for PAC learning.

Finally, we also use the connection between VC dimension and property testing to establish new lower bounds for testing radius clusterability and testing feasibility of linear constraint systems.

CCS CONCEPTS

• **Theory of computation** → **Randomness, geometry and discrete structures**; **Streaming, sublinear and near linear time algorithms**; **Lower bounds and information complexity**; *Probabilistic computation*; • **Mathematics of computing** → *Probability and statistics.*

KEYWORDS

property testing, vc dimension, distribution-free, halfspaces, decision trees, clustering, juntas, monotonicity

ACM Reference Format:
Eric Blais, Renato Ferreira Pinto Jr., and Nathaniel Harms. 2021. VC Dimension and Distribution-Free Sample-Based Testing. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21), June 21–25, 2021, Virtual, Italy*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3406325.3451104

*The full version of this paper is [10].
†Funded by an NSERC Discovery grant.
‡Funded by an NSERC Canada Graduate Scholarship.

STOC '21, June 21–25, 2021, Virtual, Italy

ACM ISBN 978-1-4503-8053-9/21/06…$15.00
https://doi.org/10.1145/3406325.3451104

1 INTRODUCTION

For which classes of functions can we test membership in the class more efficiently than we can learn a good approximation of a function in the class? This is a central question in property testing that was initially posed in the seminal work of Goldreich, Goldwasser, & Ron [27] and Kearns & Ron [33], and has since received a considerable amount of attention in different models of property testing and learning. In the standard PAC learning model of Valiant [52], the learner is *sample-based* (using only random examples of the function) and *distribution-free* (it must work for any distribution, unknown to the learner). Goldreich, Goldwasser, & Ron [27] introduce distribution-free sample-based testers and "stress that [this model] is essential for some of the potential applications" listed in that paper, but despite much recent interest in both distribution-free and sample-based testing (e.g. [4, 6, 8, 11, 17, 22, 26, 28, 30, 44]), even basic questions for this model remain unanswered. For example, are halfspaces more efficiently testable than learnable?

More precisely, fix a set $\mathcal{H}$ of Boolean-valued functions over some domain $\mathcal{X}$. There is an unknown distribution $\mathcal{D}$ over $\mathcal{X}$, the distance $\mathrm{dist}_{\mathcal{D}}(f,g)$ between two functions $f,g : \mathcal{X} \to \{0,1\}$ is $\mathbb{P}_{x\sim\mathcal{D}}[f(x) \neq g(x)]$, and the distance between f and the set $\mathcal{H}$ is $\inf_{h\in\mathcal{H}} \mathrm{dist}_{\mathcal{D}}(f,h)$. In both the learning and testing problems, the algorithm is given a set of m labelled examples $(x, f(x))$ with each x drawn independently from $\mathcal{D}$. For some fixed $\epsilon > 0$, the goals of the algorithms are:

Learning: When $f \in \mathcal{H}$, output a function h that satisfies $\mathrm{dist}_{\mathcal{D}}(f,h) \leq \epsilon$;
Testing: Accept if $f \in \mathcal{H}$ and reject if $\mathrm{dist}_{\mathcal{D}}(f,\mathcal{H}) \geq \epsilon$.

In both cases, the algorithms are required to satisfy the condition with probability at least $\frac{2}{3}$ (in this paper, we study testing algorithms with two-sided error). Let $m_\epsilon^{\mathrm{learn}}(\mathcal{H})$ and $m_\epsilon^{\mathrm{test}}(\mathcal{H})$ denote the minimum sample complexity of a learning and testing algorithm for $\mathcal{H}$, respectively. Except in pathological cases (such as $\mathcal{H}$ being a singleton), $m_\epsilon^{\mathrm{test}}(\mathcal{H}) = O(m_\epsilon^{\mathrm{learn}}(\mathcal{H}))$ [27].[1] The main question can now be phrased as:

> *For which classes $\mathcal{H}$ of Boolean-valued functions is $m_\epsilon^{\mathrm{test}}(\mathcal{H}) \ll m_\epsilon^{\mathrm{learn}}(\mathcal{H})$?*

The fundamental result of PAC learning (see e.g. [46]) is that the VC dimension of $\mathcal{H}$ determines $m_\epsilon^{\mathrm{learn}}(\mathcal{H})$. Recall that a set $T \subseteq \mathcal{X}$ is *shattered* by $\mathcal{H}$ if for every $\ell : T \to \{0,1\}$ there is a function

[1] The definitions above correspond to the standard (ϵ,δ)-PAC learning definition with $\delta = \frac{1}{3}$ and to distribution-free sample-based property testing, respectively. Note that for property testing over fixed (and known) distributions, the upper bound on sample complexity holds for *proper* (not general) learning sample complexity [27].

$f \in \mathcal{H}$ that agrees with ℓ on all points in T. The *VC dimension of $\mathcal{H}$ with respect to $S \subseteq \mathcal{X}$* is

$$\mathrm{VC}_S(\mathcal{H}) := \max\{k : \exists T \subseteq S \text{ of size } |T| = k \text{ shattered by } \mathcal{H}\}.$$

(When $S = \mathcal{X}$ we will often omit the subscript and write simply $\mathrm{VC}(\mathcal{H})$.) When $\epsilon > 0$ is constant, $m_\epsilon^{\text{learn}}(\mathcal{H}) = \Theta(\mathrm{VC}(\mathcal{H}))$, so to understand the relationship between $m_\epsilon^{\text{test}}(\mathcal{H})$ and $m_\epsilon^{\text{learn}}(\mathcal{H})$, it is necessary to understand the relationship between $m_\epsilon^{\text{test}}(\mathcal{H})$ and the VC dimension.

The VC dimension has appeared in the property testing literature, but its use has been limited to upper bounds (e.g. [2, 3, 6, 27], see however the recent work of Livni & Mansour [36] which shows some lower bounds for *graph-based discrimination* between distributions, a problem related to testing properties of distributions, in terms of a VC-like notion of dimension); the relationship mentioned above implies $m_\epsilon^{\text{test}}(\mathcal{H}) = O(\mathrm{VC}(\mathcal{H}))$ for constant ϵ. This paper is concerned with finding *lower bounds* in terms of the VC dimension. Such lower bounds would be desirable not only for understanding the relationship between testing and learning, but also because they would be *combinatorial* in nature, obtained via an analysis of the structure of the function class, whereas nearly all known lower bounds in sample-based property testing (e.g., [6, 11, 27, 33, 44]) are *distributional*: a probability distribution specific to the problem is constructed and shown to be hard to test.

It is clear that VC dimension cannot, in general, be a lower bound on the sample complexity of testing: consider the following example from [27]. Let $\mathcal{H}$ be the set of all Boolean functions $f : \{0,1\}^n \to \{0,1\}$ that satisfy $f(x) = 1$ for all $x \in \{0,1\}^n$ with $x_1 = 1$. $\mathrm{VC}(\mathcal{H}) = \Theta(2^n)$ since the 2^{n-1} points x with $x_1 = 0$ are shattered, while $m_\epsilon^{\text{test}}(\mathcal{H}) = O(1/\epsilon)$. Therefore, the relationship of VC dimension to (distribution-free sample-based) property testing is more complicated than to (PAC) learning, and we must introduce some new ideas.

1.1 Our Results

The central message of the current work is that the VC dimension *can* give lower bounds on $m_\epsilon^{\text{test}}(\mathcal{H})$ when it is combined with a closely-related combinatorial measure. For any class $\mathcal{H}$ of Boolean-valued functions over $\mathcal{X}$ and any subset $S \subseteq \mathcal{X}$, define the *LVC dimension* (or *Lower Vapnik-Chervonenkis dimension*) of $\mathcal{H}$ with respect to S to be

$$\mathrm{LVC}_S(\mathcal{H}) := \max\{k : \forall T \subseteq S \text{ of size } |T| = k, T \text{ is shattered by } \mathcal{H}\}.$$

The definition of LVC dimension differs from that of the VC dimension only by the replacement of the existential quantifier with a universal one. This immediately implies that $\mathrm{LVC}(\mathcal{H}) \le \mathrm{VC}(\mathcal{H})$ for every class $\mathcal{H}$ and motivates our choice to call this measure "lower" VC dimension. And in some cases, the LVC dimension of a class can be much smaller than its VC dimension. (See Section 5.1 for a discussion of some concepts in learning theory related to LVC dimension.) Our main theorem gives a general lower bound on $m_\epsilon^{\text{test}}(\mathcal{H})$ in terms of the VC and LVC dimensions of $\mathcal{H}$. In this theorem, the set S may be interpreted as a choice of subdomain on which $\mathrm{LVC}_S(\mathcal{H})$ is large relative to $\mathrm{VC}_{\mathcal{X}}(\mathcal{H})$.

Theorem 1.1. *There are constants $C, \epsilon_0 > 0$ such that for any class $\mathcal{H}$ of Boolean-valued functions over $\mathcal{X}$ and any $S \subseteq \mathcal{X}$, if $|S| \ge 5 \cdot \mathrm{VC}_S(\mathcal{H})$ and $\mathrm{LVC}_S(\mathcal{H}) \ge C \cdot \mathrm{VC}_S(\mathcal{H})^{3/4}\sqrt{\log \mathrm{VC}_S(\mathcal{H})}$, then for all $\epsilon \le \epsilon_0$,*

$$m_\epsilon^{\text{test}}(\mathcal{H}) = \Omega\left(\frac{\mathrm{LVC}_S(\mathcal{H})^2}{\mathrm{VC}_S(\mathcal{H}) \log \mathrm{VC}_S(\mathcal{H})}\right).$$

This bound is tight in the sense that there are classes $\mathcal{H}$ for which $m_\epsilon^{\text{test}}(\mathcal{H}) = \Theta\left(\frac{\mathrm{VC}(\mathcal{H})}{\log \mathrm{VC}(\mathcal{H})}\right)$ (for any constant ϵ), while $\mathrm{LVC}_S(\mathcal{H}) = \mathrm{VC}(\mathcal{H})$ for some S with $|S| \ge 5 \cdot \mathrm{VC}(\mathcal{H})$.

For many natural classes $\mathcal{H}$ of functions, there is a set S where $\mathrm{LVC}_S(\mathcal{H}) = \mathrm{VC}_S(\mathcal{H}) = \tilde{\Omega}(\mathrm{VC}(\mathcal{H}))$. For these classes, the following direct consequence of Theorem 1.1 is most convenient.

Corollary 1.2. *There is a constant $\epsilon_0 > 0$ such that the following holds. For every class $\mathcal{H}$ of Boolean-valued functions over $\mathcal{X}$, if $\exists S \subseteq \mathcal{X}$ for which $\mathrm{LVC}_S(\mathcal{H}) = \mathrm{VC}_S(\mathcal{H})$ and $|S| \ge 5 \cdot \mathrm{VC}_S(\mathcal{H})$, then for all $\epsilon \le \epsilon_0$,*

$$m_\epsilon^{\text{test}}(\mathcal{H}) = \Omega\left(\frac{\mathrm{VC}_S(\mathcal{H})}{\log \mathrm{VC}_S(\mathcal{H})}\right).$$

We use Theorem 1.1 and Corollary 1.2 to establish sample complexity lower bounds for distribution-free sample-based testing of many natural classes of functions, which shows, essentially, that two-sided testing is not significantly more efficient than learning (and, in the full version [10], that two-sided testing is not significantly more efficient than one-sided testing). The main new lower bounds we obtain are summarized in Table 1, with the VC dimensions of each class included for comparison. (For the formal definitions of each class, see the section devoted to that class.)

Note that many of these lower bounds are tight up to a logarithmic factor. While a factor of $\log(\mathrm{VC})$ is sometimes necessary (as for "symmetric" classes, see the full version [10]), it is not clear whether it is necessary for the classes in the table. We now discuss these lower bounds in more detail. For standard definitions in property testing and learning, see the full version of this paper [10].

Unions of k intervals. In [6] (see also [33, 41]) it was shown that there is an algorithm that can test unions of k intervals over *any* distribution on $[0,1]$ with only $O(\sqrt{k})$ samples—as long as the distribution is known to the algorithm. Our lower bound for this class shows that the sample must be quadratically larger if the distribution is not known to the algorithm.
Our bound also has implications for the *active testing* model [6], where a tester can draw some unlabelled samples from the unknown distribution $\mathcal{D}$ and then query the value of the target function on any of the sampled points. Blum and Hu [12] showed that it is possible to tolerantly test unions of k intervals in this model with $O(k)$ samples and $O(1)$ queries. Theorem 3.2 implies that $\widetilde{\Omega}(k)$ samples are necessary, even for intolerant active testers (regardless of how many samples are queried), so their result is essentially optimal.

Halfspaces. When testing over the Gaussian distribution on $\mathbb{R}^n$, only $O(\sqrt{n})$ samples suffice to test halfspaces [6]; in fact, $\widetilde{O}(\sqrt{n})$ samples suffice in the "partially distribution-free" setting where the distribution is unknown but promised to be rotation-invariant [30]. (With query access to the function, only a *constant* number of queries are required to test

Table 1: Summary of results.

Domain	Class $\mathcal{H}$	$m_\epsilon^{\text{test}}(\mathcal{H})$		VC$(\mathcal{H})$
$[n]$ or $\mathbb{R}$	Unions of k intervals	$\Omega\left(\frac{k}{\log k}\right)$	Theorem 3.2	$\Theta(k)$
$\mathbb{R}^n$	Halfspaces	$\Omega\left(\frac{n}{\log n}\right)$	Theorem 3.4	$\Theta(n)$
	Intersections of k halfspaces	$\Omega\left(\frac{nk}{\log(nk)}\right)$	Theorem 3.9	$\Theta(nk\log k)$
	Degree-k PTFs over $\mathbb{R}^n$	$\Omega\left(\frac{\binom{n+k}{k}}{\log\binom{n+k}{k}}\right)$	Theorem 5.10	$\Theta\left(\binom{n+k}{k}\right)$
	Size-k decision trees	$\Omega\left(\frac{k}{\log k}\right)$	Theorem 3.11	$\Omega(k)$
$\{0,1\}^n$	Halfspaces	$\Omega\left(\frac{n}{\log n}\right)$	Theorem 4.2	$\Theta(n)$
	Degree-k PTFs	$\Omega\left(\frac{(n/4ek)^k}{k\log(n/k)}\right)$	Theorem 4.4	$\Theta\left(\binom{n}{\le k}\right)$
	Size-k decision trees	$\Omega\left(\frac{k}{\log k\cdot\log\log k}\right)$	Theorem 4.7	$\Omega(k), O(k\log n)$

halfspaces over the Gaussian distribution or the uniform distribution on the hypercube [38].) Epstein & Silwal [21, 47] show that $\Omega(n/\epsilon)$ samples are required for testing the class of halfspaces over $\mathbb{R}^n$ with one-sided error. Our lower bound establishes a quadratic gap in sample complexity between rotation-invariant and general distribution-free testing, and the lower bound holds even for the hypercube $\{0,1\}^n$.

Intersections of halfspaces, and PTFs. Intersections of halfspaces [13, 18] and polynomial threshold functions [20, 31, 34, 42] have received much attention in the learning theory literature, but very few bounds are known on the sample or query complexity for testing these classes. As far as we know, the only bound known for testing intersections of k halfspaces is an upper bound of $\exp(k\log k)$ queries for testing the class over the Gaussian distribution [19] and no bound is known for testing polynomial threshold functions of degree greater than 1. So our results appear to establish the first non-trivial lower bounds specific for either of these classes in any model of property testing.

Decision trees. Kearns and Ron [33] first studied the problem of testing size-k decision trees, showing that $\Omega(\sqrt{k})$ samples are necessary to test the class over the uniform distribution and that this bound can be matched in the parameterized property testing model where the algorithm must only distinguish size-k decision trees from functions that are far from size-k' decision trees over the uniform distribution for some $k' > k$. The sample complexity of the (non-parameterized) size-k decision tree testing problem over the uniform distribution is not known. (The query complexity for testing size-k decision trees is also far from settled: despite recent notes to the contrary in [14, 45], the best current lower bound for the query complexity of testing size-k decision trees is $\Omega(\log k)$ [15, 49]; see also [9] for a stronger lower bound for testers with one-sided error.)

Our techniques can also be used to establish lower bounds for other models of testing. First, we show an application to testing properties of sets of points—properties that correspond to unsupervised learning problems. Namely, the *radius clustering* problem can be represented by the class C_k that consists of all sets of points $X \subseteq \mathbb{R}^n$ that can be covered by the union of at most k unit-radius balls. A distribution $\mathcal{D}$ on $\mathbb{R}^n$ is *k-clusterable* if its support is in C_k, and it is *ϵ-far from k-clusterable* if the total variation distance between $\mathcal{D}$ and any k-clusterable distribution is at least ϵ. Alon et al. [2] showed that $O(\frac{1}{\epsilon}nk\log(nk))$ samples from $\mathcal{D}$ suffice to ϵ-test k-clusterability—that is, to distinguish k-clusterable distributions from those that are ϵ-far from k-clusterable. (This can be improved when ϵ is a constant to $O(\frac{1}{\epsilon}nk\log(k)\log\frac{1}{\epsilon})$; see [29] and Section 6.1.) Prior to this work, the only lower bound for the sample complexity of this problem was Epstein and Silwal's recent lower bound of $\Omega(n/\epsilon)$ samples for ϵ-testing 1-clusterability with one-sided error [21, 47]. We give a lower bound for two-sided error testers that is tight up to poly-log factors for all values of k up to $2^{n/6}$.

Theorem 1.3. *For sufficiently small constant $\epsilon > 0$, any two-sided ϵ-tester for k-clusterability in $\mathbb{R}^n$ must have sample complexity $\Omega\left(\frac{nk}{\log(nk)}\right)$.*

A variant of Theorem 1.1 can also be used to prove strong lower bounds for some testing problems even when the underlying distribution is guaranteed to be uniform over an unknown subset of the domain. Such a situation occurs in the recent model of testing LP-type problems introduced by Epstein and Silwal [21]; see Section 6.2 for the details. We show that in this model, testing feasibility of a linear program with n variables and two-sided error requires $n^{1-o(1)}$ queries, almost matching the $O(n/\epsilon)$ upper bound of [21].

The connection between LVC dimension and distribution-free property testing also extends to the tolerant testing model even when the algorithm has *query* access to the function and can adaptively select the queries during its execution. The bound in the theorem applies even to the tolerant testing model where $\epsilon_0 = 0$ (i.e., the algorithm must accept functions that are 0-close to the class) but it does *not* apply to the non-tolerant testing model. This is a subtle point: in the distribution-free setting, a function f may be 0-close to $\mathcal{H}$ (so a $(0,\epsilon_1)$-tolerant tester should accept it) while also satisfying $f \notin \mathcal{H}$ (so a non-tolerant tester is not required to accept it). See the full version [10] for details.

Finally, we examine the necessity of the conditions in Theorem 1.1. We have proved that many commonly-studied classes of functions meet the conditions of the theorem and therefore are impossible to test much more efficiently than learn; but are there commonly-studied classes of functions that fail the condition in the theorem and have efficient distribution-free sample-based testers? [27] gave an example a class where distribution-free sample-based testing is much more efficient than learning, which we repeated above, but this is not a commonly-studied, natural class. In the full version of this paper [10], we prove that two foundational properties in the property testing literature, k-juntas and monotone Boolean functions, have distribution-free sample-based testers with complexity $O(\mathrm{VC}^c)$ for constants $c < 1$.

1.2 Our Techniques

Our main tool is a reduction from testing properties of *distributions* to testing properties of *functions*. Some relationships between these two types of problems have been observed before, e.g. by Sudan [48] and Goldreich & Ron [28], who note that any distribution testing problem can be reduced to a testing problem for a specially-constructed symmetric property of functions with non-Boolean range (symmetric properties are invariant under permutations of the variables). Goldreich & Ron [28] also observed that testing symmetric properties of functions can be reduced to *support-size estimation*, a fundamental problem in distribution testing. We extend this connection between distribution testing and property testing to properties that are not symmetric: we show, in the opposite direction of [28], that support-size testing can be reduced to property testing when the LVC dimension is large.

The generality of our lower bound comes from the Sauer–Shelah–Perles lemma, which is usually used to prove *upper bounds* in terms of the VC dimension. However, we use it to show that a random function is far from the property $\mathcal{H}$ when the underlying distribution has support size larger than the VC dimension. On the other hand, random functions must be accepted by the tester when the underlying distribution is supported on a set smaller than the LVC dimension. In this way, we can show that any testing algorithm must implicitly solve the support-size testing problem by distinguishing between distributions of small vs. large supports. Tight bounds on the support-size estimation problem were attained by Valiant & Valiant [50, 51]; we use a version of the bound due to Wu & Yang [55] which applies to a wider range of parameters that are necessary for our reduction.

With this technique, the problem of attaining lower bounds is transformed into the combinatorial problem of constructing appropriate sets S with large $\mathrm{LVC}_S(\mathcal{H})$ and $\mathrm{VC}_S(\mathcal{H})$. We study this problem in the second half of the paper. For some properties, like halfspaces, this is easy, but other properties require more effort:

Intersections of halfspaces. The VC dimension of intersections of k halfspaces in $\mathbb{R}^n$ is $\Theta(nk \log k)$ in general [18]. We construct a subset $S \subset \mathbb{R}^n$ in which $\mathrm{LVC}_S = \mathrm{VC}_S = \Theta(nk)$. We accomplish this by constructing S such that intersections of k halfspaces on S are equivalent to $\Theta(nk)$-alternating functions on $\mathbb{R}$. This reduction yields a lower bound of $\Omega(nk/\log(nk))$ for testing intersections of k halfspaces.

Polynomial threshold functions on $\mathbb{R}^n$. Although PTFs can be transformed into halfspaces in a higher dimension, we opt to treat PTFs as a special case of a *Dudley class* [7] and explore the connection between LVC dimension, Dudley classes, and *maximum classes*. Dudley classes are those obtained by taking the sign of a function in a fixed vector space $\mathcal{F}$ of real-valued functions, and these classes have VC dimension equal to the dimension of that vector space [54]. Maximum classes are those for which the Sauer–Shelah–Perles lemma is tight (see Section 5.1), and in particular, those classes satisfy LVC = VC (Proposition 5.2). Johnson [32] showed that Dudley classes with domain $\mathbb{R}^n$ where the functions $\mathcal{F}$ are analytic are maximum on an arbitrarily large subset $S \subseteq \mathbb{R}^n$, and therefore $\mathrm{LVC}_S = \mathrm{VC}_S$, so our main result applies. Other examples of analytic Dudley classes include balls in $\mathbb{R}^n$ and trigonometric polynomial threshold functions in $\mathbb{R}^2$, so we automatically obtain lower bounds for these classes.

Halfspaces and PTFs on the Boolean hypercube. Our constructions of the set S for halfspaces and PTFs on domain $\mathbb{R}^n$ fail on the more restrictive domain $\{0,1\}^n$, and indeed the deterministic reduction underlying Theorem 2.9 seems to fail as well, because it is hard or impossible to construct large sets $S \subset \{0,1\}^n$ with high LVC_S (observe that $\{0,1\}^n$ is far from being in general position). Therefore we use a *randomized* reduction that requires some results on the non-singularity of random matrices. In particular, we rely on a theorem of Abbe, Shpilka, & Wigderson [1] to construct a random set $S \subseteq \{\pm 1\}^n$ on which the condition $\mathrm{LVC} = \Omega(\mathrm{VC})$ holds "with high probability", i.e. a *random* set of size $\Omega(\mathrm{VC})$ is shattered.

k-Clusterability. k-Clusterability is a property of *distributions*, not functions, so our main theorem does not apply; however, we can adapt the argument to this setting. In the case $k = 1$, we use concentration results for random points on the n-sphere to show that a random set of n points (with $\|x\|_2 > 1$) is 1-clusterable while a random set of $2n$ points is far from 1-clusterable, which we can then use in a randomized reduction from support-size testing to 1-clusterability. We extend these concentration results to k disjoint spheres to attain the randomized reduction to k-clusterability.

Uniform distributions and testing LP feasibility. In some cases it is desirable to show lower bounds for testing problems where the underlying distribution is promised to be uniform over some unknown subset of the domain. An example is the recent model of Epstein & Silwal [21] for LP-type testing. In some cases, by replacing the support-size testing lower bounds of Wu & Yang [55] with the lower bounds for the *distinct elements* problem [43], we can reproduce the main theorem with a slightly weaker bound, but with the guarantee that the distributions are uniform (over an unknown support).

Upper Bounds. There are three upper bounds in the full version of this paper [10]: for symmetric classes of functions (which establishes the optimality of Theorem 1.1), for juntas, and for monotone functions. The result for symmetric classes follows from a result of Goldreich & Ron [28]. For juntas and monotonicity, see the full version of the paper.

2 GENERAL LOWER BOUND

We prove our main result, Theorem 1.1, in this section. Before we begin, we discuss some examples that illuminate why the conditions in the theorem are important, i.e. the choice of a subset $S \subseteq \mathcal{X}$ with large $\mathsf{LVC}_S(\mathcal{H})$ and the condition $|S| \geq 5 \cdot \mathsf{VC}_S(\mathcal{H})$. Unlike a learning algorithm, a property tester can halt and reject as soon as it sees proof that the unknown function $f : \mathcal{X} \to \{0,1\}$ does not belong to the class. Therefore, we aim to find subsets $S \subseteq \mathcal{X}$ where small "certificates" of non-membership cannot exist. This motivates the definition of $\mathsf{LVC}_S(\mathcal{H})$: any subset $T \subseteq S$ of size $|T| \leq \mathsf{LVC}_S(\mathcal{H})$ cannot contain any certificates of non-membership, for any function $f \notin \mathcal{H}$. So we want to find sets where $\mathsf{LVC}_S(\mathcal{H})$ is as large as possible relative to $\mathsf{VC}(\mathcal{H})$. On the other hand, if $\mathsf{LVC}_S(\mathcal{H}) = \mathsf{VC}(\mathcal{H})$ but $|S| = \mathsf{VC}(\mathcal{H})$, then the class $\mathcal{H}$ restricted to S is trivial: it contains all possible functions on S, so testing is still easy. $|S|$ must be large enough so that most functions are far from $\mathcal{H}$, and this will be guaranteed in general when $|S| > 5 \cdot \mathsf{VC}_S(\mathcal{H})$ (the constant 5 is somewhat arbitrary). The following examples illustrate these phenomena. In the first example, $\mathsf{LVC}_{\mathcal{X}}(\mathcal{H})$ is constant, but a careful choice of large S allows $\mathsf{LVC}_S(\mathcal{H}) = \mathsf{VC}(\mathcal{H})$, and we will obtain lower bounds for this class:

Example 2.1. Let $\mathcal{L}_n$ be the set of halfspaces $\mathbb{R}^n \to \{\pm 1\}$. As is well-known, $\mathsf{VC}_{\mathbb{R}^n}(\mathcal{L}_n) = n+1$. But $\mathsf{LVC}_{\mathbb{R}^n}(\mathcal{L}_n) = 2$, since any 3 colinear points cannot be shattered. On the other hand, if $S \subseteq \mathbb{R}^n$ is a set of points in general position and $|S| > n+1$, then $\mathsf{LVC}_S(\mathcal{L}_n) = \mathsf{VC}_S(\mathcal{L}_n) = n+1$.

In the second example, the conditions of our theorem fail: finding a good set S is impossible, and indeed there is an efficient distribution-free sample-based tester (see the full version [10]).

Example 2.2. Let $\mathcal{M}$ be the set of monotone functions $P \to \{0,1\}$ where P is any partial order ($f : P \to \{0,1\}$ is monotone if $f(x) \leq f(y)$ whenever $x < y$). Recall that an *antichain* is a set of points $x \in P$ that are incomparable. Observe that a set T is shattered by $\mathcal{M}$ if and only if it is an antichain: a monotone function can take arbitrary values on an antichain, whereas if $x, y \in T$ are comparable, say $x < y$, then $f(x) \leq f(y)$ so T cannot be shattered. Therefore $\mathsf{LVC}_S(\mathcal{M}) = \mathsf{VC}_S(\mathcal{M}) = |S|$ if S is an antichain, and if S is not an antichain then $\mathsf{LVC}_S(\mathcal{M}) = 2$ while $\mathsf{VC}_S(\mathcal{M})$ is the size of the largest antichain in S.

We now turn to the proof of Theorem 1.1. The proof uses two main ingredients: lower bounds on the support size estimation problem, and the Sauer–Shelah–Perles theorem.

2.1 Ingredient 1: Support Size Distinction

A fundamental problem in the field of distribution testing is *support size estimation*: Given sample access to an unknown finitely-supported distribution $\mathcal{D}$ where each element occurs with probability at least $1/n$ (for some n), estimate the size of the support up to an additive ϵn error. Valiant & Valiant [50, 51] showed that for constant ϵ, the number of samples required for this problem is $\Theta\left(\frac{n}{\log n}\right)$. We will adapt this lower bound (in fact an improved version of Wu and Yang [55]) to give lower bounds on distribution-free property testing.

Definition 2.3 (Support-Size Distinction Problem). For any $n \in \mathbb{N}$ and $0 < \alpha < \beta \leq 1$, define $\mathrm{SSD}(n, \alpha, \beta)$ as the minimum number $m \in \mathbb{N}$ such that there exists an algorithm that for any input distribution p over $[n]$, takes m samples from p and distinguishes with probability at least $2/3$ between the cases:

(1) $|\operatorname{supp}(p)| \leq \alpha n$ and $\forall i \in \operatorname{supp}(p), p_i \geq 1/n$; and,
(2) $|\operatorname{supp}(p)| \geq \beta n$ and $\forall i \in \operatorname{supp}(p), p_i \geq 1/n$.

Valiant & Valiant [50] and Wu & Yang [55] each prove lower bounds on support-size *estimation* and they do so essentially by proving lower bounds on support-size distinction. We note that the bound of [50] holds for $\mathrm{SSD}(n, \alpha, \beta)$ when $1/2 < \alpha < \beta < 1$, but this gap of at most $1/2$ is not sufficient for our purposes, so we use the improved version of [55]. However, their lower bound on SSD is not stated explicitly, and therefore we state and prove the following bound explicitly in the full version [10].

Theorem 2.4 ([55]). *There exists a constant C such that, for any $\delta \geq C\frac{\sqrt{\log n}}{n^{1/4}}$ and $\delta \leq \alpha < \beta \leq 1 - \delta$,*

$$\mathrm{SSD}(n, \alpha, \beta) = \Omega\left(\frac{n}{\log n} \log^2 \frac{1}{1-\delta}\right).$$

2.2 Ingredient 2: Sauer–Shelah–Perles Lemma

We will need the Sauer–Shelah–Perles lemma (see e.g. [46]), for which we recall the following definitions:

Definition 2.5. Let $\mathcal{H}$ be a set of functions $\mathcal{X} \to \{0,1\}$ and let $S \subseteq \mathcal{X}$. We will define the *shattering number* as

$$\mathrm{sh}(\mathcal{H}, S) := |\{T \subseteq S \mid T \text{ is shattered by } \mathcal{H}\}|.$$

We define the *growth function* as

$$\Phi(\mathcal{H}, S) := |\{\ell : S \to \{0,1\} \mid \exists h \in \mathcal{H}\ \forall x \in S, \ell(x) = h(x)\}|.$$

We state a version of the Sauer–Shelah–Perles lemma that follows from the so-called Sandwich Theorem, rediscovered by numerous authors (see e.g. [39]):

Lemma 2.6 (Sauer–Shelah–Perles). *Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ and let $S \subseteq \mathcal{X}$ with $\mathsf{VC}_S(\mathcal{H}) = d$. Then $\Phi(\mathcal{H}, S) \leq \mathrm{sh}(\mathcal{H}, S) \leq \sum_{i=0}^{d} \binom{|S|}{i}$.*

This lemma gives us a bound on the probability that a random function over a large set is far from the hypothesis class $\mathcal{H}$.

Lemma 2.7. *There is a constant $K > 1$ (in particular, $K = 3.04$ suffices) and constants $L > 0, \epsilon_0 > 0$ (depending on K) such that, if $\mathcal{H}$ is a class of functions $\mathcal{X} \to \{0,1\}$ with $\mathsf{VC}(\mathcal{H}) = d$ and $T \subseteq \mathcal{X}$ has size $|T| \geq Kd$, then a uniformly random labelling $\ell : T \to \{0,1\}$ satisfies, with probability at least $1 - e^{-Ld}$, $\forall h \in \mathcal{H} : \mathbb{P}_{x \sim T}[h(x) \neq \ell(x)] > \epsilon_0$.*

Proof. For any $T \subseteq \mathcal{X}$ of size $|T| = m$, and each $h \in \mathcal{H}$, the number of functions $\ell : T \to \{0,1\}$ that differ from h on at most ϵm points of T is at most $\sum_{i=1}^{\epsilon m} \binom{m}{i}$. Therefore, by the Sauer–Shelah–Perles lemma, the number of labellings $\ell : T \to \{0,1\}$ that differs on at most ϵm points from the closest $h \in \mathcal{H}$ is at most

$$\left(\sum_{i=0}^{d} \binom{m}{i}\right) \cdot \left(\sum_{i=0}^{\epsilon m} \binom{m}{i}\right) \leq \left(\frac{em}{d}\right)^d \cdot \left(\frac{em}{\epsilon m}\right)^{\epsilon m} = \left(\frac{em}{d}\right)^d \cdot \left(\frac{e}{\epsilon}\right)^{\epsilon m}.$$

The probability that a uniformly random $\ell : T \to \{0,1\}$ satisfies this condition is therefore at most

$$\left(\frac{em}{d}\right)^d \cdot \left(\frac{e}{\epsilon}\right)^{\epsilon m} \cdot 2^{-m} = 2^{d(\log(Ke)+K\epsilon\log(e/\epsilon)-K)} = e^{d(\ln(Ke)+K\epsilon\ln(e/\epsilon)-K\ln(2))},$$

For any $K > 1$ satisfying $K\ln(2) > 1 + \ln(K)$, there is $L > 0, \epsilon_0 > 0$ such that the exponent $d\,(\ln(Ke) + K\epsilon\ln(e/\epsilon) - K\ln(2)) < -Ld$ for all $\epsilon < \epsilon_0$. □

2.3 Main Reduction

We now present the main reduction for the proof of Theorem 1.1. This reduction is inspired by a proof in the recent work of Epstein & Silwal [21]. The reduction can be described intuitively as follows. Suppose there is a class $\mathcal{H}$ of functions $\mathcal{X} \to \{0,1\}$ and a set $S \subseteq \mathcal{X}$ such that are two thresholds $t_1 < t_2$ where:

(1) Any set $T \subset S$ of size $|T| \le t_1$ is shattered by $\mathcal{H}$; and,
(2) A random function on any subset $T \subset S$ of size $|T| \ge t_2$ is far from $\mathcal{H}$ with high probability.

Then a distribution-free tester must accept any function (with high probability) when the distribution has support size at most t_1, and reject a random function (with high probability) when the distribution has support size at least t_2. This is made formal in our main lemma:

Lemma 2.8. *Let $\mathcal{H}$ be a set of functions $\mathcal{X} \to \{0,1\}$. Suppose $S \subseteq \mathcal{X}$ has size $|S| = n$ and $0 < \alpha < \beta \le 1$ satisfy the following conditions:*

(1) *$\forall T \subset S$ such that $|T| \le \alpha n$, T is shattered by $\mathcal{H}$; and,*
(2) *$\forall T \subseteq S$ such that $|T| \ge \beta n$, a uniformly random labelling $\ell : T \to \{0,1\}$ satisfies with probability at least $9/10$ the condition*

$$\forall h \in \mathcal{H} : \mathop{\mathbb{P}}_{x\sim T}[\ell(x) \ne h(x)] \ge \epsilon/\beta\,.$$

Then $m_\epsilon^{\text{test}}(\mathcal{H}) = \Omega(\text{SSD}(n,\alpha,\beta))$.

Proof. Let $f : S \to \{0,1\}$ be a uniformly random function, let $\phi : [n] \to S$ be any bijection, and let $\mathcal{D}$ be any distribution over $[n]$ with $\mathcal{D}(x) \ge 1/n$ for all $x \in \text{supp}(\mathcal{D})$. Write $\phi\mathcal{D}$ for the distribution over S of $\phi(x)$ when $x \sim \mathcal{D}$. We make two claims.

First, if $\mathcal{D}$ has support size at most αn then $\text{dist}_{\phi\mathcal{D}}(f,\mathcal{H}) = 0$. Let $T = \text{supp}(\phi\mathcal{D})$. Then since $|T| \le \alpha n$, by the first condition there exists $h \in \mathcal{H}$ such that $h(x) = f(x)$ on all $x \in T$. So $\text{dist}_{\phi\mathcal{D}}(f,h) = 0$.

Second, if $\mathcal{D}$ has support size at least βn then with probability at least $9/10$ over the choice of f, $\text{dist}_{\phi\mathcal{D}}(f,\mathcal{H}) \ge \epsilon$. Let $T = \text{supp}(\phi\mathcal{D})$ and for any $h \in \mathcal{H}$ write $\Delta(f,h) = \{x \in T : f(x) \ne h(x)\}$. Since $|T| \ge \beta n$ we have by assumption that, with probability at least $9/10$ over the choice of f, for uniform $x \sim T$, $\mathbb{P}[x \in \Delta(f,h)] \ge \epsilon/\beta$. Therefore $|\Delta(f,h)| \ge \frac{\epsilon}{\beta}|T| \ge \epsilon n$. Since $\phi\mathcal{D}(x) = \mathcal{D}(\phi^{-1}(x)) \ge 1/n$ for every $x \in T$, this means that for every $h \in \mathcal{H}$, $\mathop{\mathbb{P}}_{x\sim\phi\mathcal{D}}[f(x) \ne h(x)] \ge \frac{1}{n}|\Delta(f,h)| \ge \epsilon$.

We now prove the lower bound on distribution-free sample testing. Assume there is a distribution-free tester A that uses m samples. The algorithm for support-size distinction is as follows. Given input distribution $\mathcal{D}$ over $[n]$, choose a uniformly random $f : S \to \{0,1\}$, draw m samples $Q = (x_1,\dots,x_m)$ from $\phi\mathcal{D}$ and let $Q_f = ((x_1,f(x_1)),\dots,(x_m,f(x_m)))$; run A on the samples Q_f and accept $\mathcal{D}$ iff A outputs 1.

First suppose that $\mathcal{D}$ has support size at most αn. There exists a function $h \in \mathcal{H}$ with $\text{dist}_{\phi\mathcal{D}}(f,h) = 0$, so $f(x) = h(x)$ for all $x \in \text{supp}(\phi\mathcal{D})$. Therefore the samples Q_f and Q_h have the same distribution, and the algorithm must output 1 on Q_h with probability at least $5/6$, so it must output 1 on Q_f, and therefore accept $\mathcal{D}$, with probability at least $5/6$.

Next suppose that $\mathcal{D}$ has support size at least βn. Then the uniformly random function $f : S \to \{0,1\}$ is ϵ-far from $\mathcal{H}$ with respect to $\phi\mathcal{D}$ with probability at least $9/10$. Assuming this occurs, algorithm A must output 0 with probability at least $5/6$, so $\mathcal{D}$ is rejected with probability at least $2/3$. We conclude

$$m_\epsilon^{\text{test}}(\mathcal{H}) = \Omega(\text{SSD}(n,\alpha,\beta))\,.$$ □

2.4 Proof of the Main Lower Bound

Combining Theorem 2.4 with Lemma 2.7, we obtain the most general form of our main theorem:

Theorem 2.9. *Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ and suppose there is a set $S \subseteq \mathcal{X}$ and a value $\delta \in (0,1/2)$ such that, for $n = |S|$, the following hold:*

(1) *$K \cdot \text{VC}_S(\mathcal{H}) \le (1-\delta)n$, where K is the constant from Lemma 2.7; and,*
(2) *$\text{LVC}_S(\mathcal{H}) \ge \delta n$; and,*
(3) *$\delta \ge C\frac{\sqrt{\log n}}{n^{1/4}}$ where C is the constant from Theorem 2.4.*

Let $d = \text{VC}_S(\mathcal{H})$. Then for some constant $\epsilon_0 > 0$ and all $0 < \epsilon < \epsilon_0$,

$$m_\epsilon^{\text{test}}(\mathcal{H}) = \Omega\left(\frac{n}{\log n}\log^2\frac{1}{1-\delta}\right) = \Omega\left(\frac{d}{\log d}\log^2\frac{1}{1-\delta}\right).$$

Proof. Let $\alpha = \frac{1}{n}\text{LVC}_S(\mathcal{H})$, $\beta = \frac{1}{n}K \cdot \text{VC}_S(\mathcal{H})$, so that $\alpha \ge \delta$ and $\beta \le 1-\delta$. Then from Theorem 2.4,

$$\text{SSD}(n,\alpha,\beta) = \Omega\left(\frac{n}{\log n}\log^2\frac{1}{1-\delta}\right).$$

By definition of LVC, any set $T \subseteq S$ with $|T| \le \alpha n$ satisfies condition 1 of Lemma 2.8, and by Lemma 2.7, any set $T \subseteq S$ such that $|T| \ge \beta n = K \cdot \text{VC}_S(\mathcal{H})$ satisfies condition 2 for sufficiently small (constant) $\epsilon > 0$, so by Lemma 2.8,

$$m_\epsilon^{\text{test}}(\mathcal{H}) = \Omega(\text{SSD}(n,\alpha,\beta)) = \Omega\left(\frac{n}{\log n}\log^2\frac{1}{1-\delta}\right).$$

Finally, since $\frac{1}{1-\delta} \ge 1$ and $n = \Omega(d/(1-\delta)) = \Omega(d)$, we have a lower bound of $\Omega\left(\frac{d}{\log d}\log^2\frac{1}{1-\delta}\right)$. □

The following simplified bound proves Theorem 1.1 from the introduction and will also be used in most of our applications.

Corollary 2.10. *There is a constant $L > 0$ such that the following holds. Let $S \subseteq \mathcal{X}$ satisfy $n := |S| \ge 5 \cdot \text{VC}_S(\mathcal{H})$. If $\text{LVC}_S(\mathcal{H}) > L \cdot \text{VC}_S(\mathcal{H})^{3/4}\sqrt{\log \text{VC}_S(\mathcal{H})}$, then*

$$m_\epsilon^{\text{test}}(\mathcal{H}) = \Omega\left(\frac{\text{LVC}_S(\mathcal{H})^2}{\text{VC}_S(\mathcal{H})\log\text{VC}_S(\mathcal{H})}\right).$$

Proof. We may assume $n = |S| = 5 \cdot \mathrm{VC}_S(\mathcal{H})$ since by taking subsets of a set S of size larger than $\mathrm{VC}_S(\mathcal{H})$, we do not decrease the LVC dimension and do not increase the VC dimension; we can choose a subset that also does not decrease the VC dimension. We may set $K = 4$ in Theorem 2.9. Let $\delta = \frac{\mathrm{LVC}_S(\mathcal{H})}{2K\mathrm{VC}_S(\mathcal{H})}$, so

$$\delta n = \frac{\mathrm{LVC}_S(\mathcal{H})}{2K\mathrm{VC}_S(\mathcal{H})} \cdot 5\mathrm{VC}_S(\mathcal{H}) \le \mathrm{LVC}_S(\mathcal{H}).$$

We also have $(1-\delta)n \ge \left(1 - \frac{1}{8}\right) 5 \cdot \mathrm{VC}_S(\mathcal{H}) \ge 4\mathrm{VC}_S(\mathcal{H}) = K \cdot \mathrm{VC}_S(\mathcal{H})$. Finally,

$$\delta = \frac{\mathrm{LVC}_S(\mathcal{H})}{8\mathrm{VC}_S(\mathcal{H})} \ge \frac{L\sqrt{\log \mathrm{VC}_S(\mathcal{H})}}{8\mathrm{VC}_S(\mathcal{H})^{1/4}} = \frac{5^{1/4}L\sqrt{\log(n/5)}}{8n^{1/4}},$$

so for large enough constant $L > 0$ this is at least $C\frac{\sqrt{\log n}}{n^{1/4}}$ for the constant C in Theorem 2.4, so the conditions for Theorem 2.9 are satisfied, and we obtain a lower bound of

$$\Omega\left(\frac{\mathrm{VC}_S(\mathcal{H})}{\log \mathrm{VC}_S(\mathcal{H})} \log^2 \frac{1}{1-\delta}\right).$$

Finally, using the inequality $\log^2 \frac{1}{1-\delta} \ge \log^2(e^\delta) = \Omega(\delta^2)$ we get the conclusion. □

The proof of the second part of Theorem 1.1, which establishes the tightness of the lower bound, can be found in the full version of this paper [10].

3 GEOMETRIC CLASSES

In this section, we use Corollary 1.2 to prove lower bounds on the number of samples required to test unions of intervals, halfspaces, and intersections of halfspaces.

Technical note: For the domain $\mathbb{R}^n$, the tester may assume that the distribution $\mathcal{D}$ is defined on the same σ-algebra as the Lebesgue measure. The distributions arising from the above reduction are finitely supported but for the functions considered in this paper, one may replace finitely supported distributions with distributions that are absolutely continuous with respect to the Lebesgue measure without changing the results, by replacing each point in the support with an arbitrarily small ball.

3.1 Unions of Intervals

A function $f : \mathbb{R} \to \{0,1\}$ is a *union of k intervals* if there are k intervals $[a_1, b_1], \ldots, [a_k, b_k]$, where we allow $a_i = -\infty$ and $b_i = \infty$, such that $f(x) = 1$ iff x is contained in some interval $[a_i, b_i]$. Let $\mathcal{I}_k$ denote the class of such functions.

The analysis of the LVC dimension of $\mathcal{I}_k$ is a straightforward variant of the standard analysis of the VC dimension of the class and serves as a good introduction to the high-level structure of the arguments that will be used in later proofs as well.

Proposition 3.1. $\mathrm{LVC}_{\mathbb{R}}(\mathcal{I}_k) = \mathrm{VC}_{\mathbb{R}}(\mathcal{I}_k) = 2k$.

Proof. Let $S \subset \mathbb{R}$ have size $2k$ and let $\ell : S \to \{0,1\}$ be arbitrary. Write $S = \{s_1, \ldots, s_{2k}\}$ where $s_1 < \cdots < s_{2k}$ and partition S into k consecutive pairs (s_i, s_{i+1}) for odd i. Then for each pair (s_i, s_{i+1}) we can choose a single interval that contains exactly the points in s_i, s_{i+1} labelled 1 by ℓ. Therefore S is shattered by k intervals.

On the other hand, let $S \subset \mathbb{R}$ have size $|S| = 2k+1$, let $s_1 < \cdots < s_{2k+1}$ be the points in S, and suppose $\ell(i) = 1$ iff i is odd. Then any interval can contain at most 1 point of S labelled 1, unless it also contains a 0-point. Therefore S is not shattered. So a set S is shattered iff $|S| \le 2k$, implying the conclusion. □

Applying Corollary 2.10, we obtain:

Theorem 3.2. *For some constant $\epsilon > 0$, $m_\epsilon^{\mathrm{test}}(\mathcal{I}_k) = \Omega\left(\frac{k}{\log k}\right)$.*

3.2 Halfspaces

A halfspace is a function $f : \mathbb{R}^n \to \{\pm 1\}$ of the form $f(x) = \mathrm{sign}\left(w_0 + \sum_{i=1}^n w_i x_i\right)$ where each $w_i \in \mathbb{R}$. In this subsection, write $\mathcal{L}_n$ for the class of halfspaces (or *Linear threshold functions*) with domain $\mathbb{R}^n$.

The analysis of the LVC dimension follows immediately from the following well-known shattering properties of halfspaces. (See, e.g., [46].)

Proposition 3.3. *Any set $S \subset \mathbb{R}^n$ of size $n+1$ in general position can be shattered by $\mathcal{L}_n$, and any set $T \subset \mathbb{R}^n$ of n linearly independent vectors can be shattered by $\mathcal{L}_n$. No set of size $n+2$ is shattered by $\mathcal{L}_n$.*

Applying Corollary 2.10, we obtain our lower bound for domain $\mathbb{R}^n$:

Theorem 3.4. *For some constant $\epsilon > 0$,*

$$m_\epsilon^{\mathrm{test}}(\mathcal{L}_n) = \Omega\left(\frac{n}{\log n}\right).$$

Proof. This holds by Corollary 2.10, since we may choose any set $S \subset \mathbb{R}^n$ of size $|S| \ge 5(n+1)$ in general position, which by the above proposition satisfies $\mathrm{LVC}_S(\mathcal{L}_n) = \mathrm{VC}_S(\mathcal{L}_n) = n+1$. □

3.3 Intersections of Halfspaces

Let $\mathcal{L}_n^{\cap k}$ denote the class of all Boolean-valued functions obtained by taking the intersections of k halfspaces over $\mathbb{R}^n$. Formally, $\mathcal{L}_n^{\cap k}$ is the set of functions

$$f(x) = h_1(x) \wedge h_2(x) \wedge \cdots \wedge h_k(x)$$

where each h_i is a halfspace. It was recently shown by Csikós, Mustafa, & Kupavskii [18] that the VC dimension of this classes is

$$\mathrm{VC}(\mathcal{L}_n^{\cap k}) = \Theta(nk \log k).$$

Csikós *et al.* remark that it was long assumed (incorrectly) that the VC dimension of the class was $\Theta(nk)$, which is what one might intuitively expect. We exhibit an infinite set S on which $\mathrm{VC}_S(\mathcal{L}_n^{\cap k}) = \mathrm{LVC}_S(\mathcal{L}_n^{\cap k}) = \Theta(nk)$. We do so with an analysis of alternating functions and polynomial threshold functions.

For any n, define the mapping $\psi : \mathbb{R} \to \mathbb{R}^n$ as follows:

$$\psi_n(x) := \begin{cases} (x, x^2, x^3, \ldots, x^n) & \text{if } n \text{ is even} \\ (0, x, x^2, \ldots, x^{n-1}) & \text{if } n \text{ is odd}. \end{cases}$$

Let $\mathcal{A}_m$ be the set of function $\mathbb{R} \to \{0,1\}$ that alternate at most m times.

Proposition 3.5. *The set $\mathcal{P}$ of functions $\mathrm{sign}(p(x))$ on $\mathbb{R}$ where p is a polynomial of degree at most d is equal to the set $\mathcal{A}_d$.*

Proof. This follows from the fact that number of alternations of the function $\text{sign}(p)$ is exactly the number of zeroes of p, which is at most d. On the other hand, any function alternating at most d times may be represented by $\text{sign}(p)$ where p is a polynomial whose zeroes are exactly the points where the function alternates. □

Proposition 3.6. *For any even m and any k, $\mathcal{A}_m^{\cup k} = \mathcal{A}_{mk}$.*

Proof. It is clear that the union of k m-alternating functions will alternate at most mk times, so $\mathcal{A}_m^{\cup k} \subseteq \mathcal{A}_{mk}$, so we must show that $\mathcal{A}_{mk} \subseteq \mathcal{A}_m^{\cup k}$. We will do so by induction on k, where the base case $k = 1$ is trivial. For $k > 1$, let $f \in \mathcal{A}_{mk}$ and let $t_1 < \cdots < t_{mk}$ be the alternations (i.e. f is constant on each interval (t_i, t_{i+1}) and $(-\infty, t_1)$, (t_{mk}, ∞)). There are two cases: First suppose that the first alternation of $f \in \mathcal{A}_{mk}$ alternates from 0 to 1; or, symmetrically, suppose that the last alternation of f alternates from 1 to 0. Then the function g equal to f on $x \le t_m$ and 0 on $x > t_m$ is the union of $m/2$ intervals, and $g \in \mathcal{A}_m$. Let f' be 0 on $x \le t_m$ and equal to f on $x > t_m$, so that f is the union of f' and g, and $f' \in \mathcal{A}_{m(k-1)}$. By induction f' is the union of $k-1$ m-alternating functions, so $f \in \mathcal{A}_m \cup \mathcal{A}_m^{\cup(k-1)} = \mathcal{A}_m^{\cup k}$.

In the second case, the first and last alternations of f alternate from 1 to 0 and 0 to 1, respectively. Let g take value 1 on $(-\infty, t_1]$, $[t_{mk}, \infty)$ as well as on the first $m/2 - 1$ intervals

$$[t_2, t_3], [t_4, t_5], \ldots, [t_{m-2}, t_{m-1}],$$

and 0 otherwise. Then $g \in \mathcal{A}_m$ and the function $f' = f - g$ is in $\mathcal{A}_{m(k-1)}$. So by induction $f' \in \mathcal{A}_m^{\cup(k-1)}$ and $f \in \mathcal{A}_m \cup \mathcal{A}_m^{\cup(k-1)} = \mathcal{A}_m^{\cup k}$. □

Proposition 3.7. *For any even m, any k, and any set $S \subseteq \mathbb{R}$ with $|S| > mk$, $\text{VC}_S(\mathcal{A}_m^{\cap k}) = \text{LVC}_S(\mathcal{A}_m^{\cap k}) = mk$.*

Proof. For a class $\mathcal{H}$, write $\overline{\mathcal{H}}$ of the set of functions $f = -g$ where $g \in \mathcal{H}$ (i.e. the set of complements of functions in $\mathcal{H}$). Note that $\overline{\mathcal{A}}_m = \mathcal{A}_m$ since the complement preserves alternations. By De Morgan's laws, $(\overline{\mathcal{H}_n})^{\cap k} = \overline{\mathcal{H}_n^{\cup k}}$. Then $\mathcal{A}_m^{\cap k} = (\overline{\mathcal{A}_m})^{\cap k} = \overline{\mathcal{A}_m^{\cup k}} = \overline{\mathcal{A}_{mk}} = \mathcal{A}_{mk}$. The conclusion follows since $\text{VC}_S(\mathcal{A}_{mk}) = \text{LVC}_S(\mathcal{A}_{mk}) = mk$ by the same argument as for unions of intervals. □

Lemma 3.8. *For any $k \ge 1$ and $S \subset \mathbb{R}$ with $|S| > nk$, if n is even then $\text{LVC}_{\psi_n(S)}(\mathcal{L}_n^{\cap k}) = \text{VC}_{\psi_n(S)}(\mathcal{L}_n^{\cap k}) = nk$ and if n is odd then $\text{LVC}_{\psi_n(S)}(\mathcal{L}_n^{\cap k}) = \text{VC}_{\psi_n(S)}(\mathcal{L}_n^{\cap k}) = (n-1)k$.*

Proof. First suppose that n is even and consider a halfspace $h(y) = \text{sign}(t + \sum_{i=1}^n w_i y_i)$, where $y = \psi_n(x)$ for some $x \in S$. Then $h(\psi_n(x)) = \text{sign}(t + \sum_{i=1}^n w_i x^i)$, which is the sign of a degree-n polynomial on x. Therefore the set of halfspaces h on the set $\psi(S)$ is equivalent to the set of degree-n polynomials on S, which by Proposition 3.5 is equal to the set of n-alternating functions, so by Proposition 3.7 we have $\text{LVC}_{\psi(S)}(\mathcal{L}_n^{\cap k}) = \text{VC}_{\psi(S)}(\mathcal{L}_n^{\cap k}) = \text{VC}(\mathcal{A}_n^{\cap k}) = nk$. When n is odd, the same argument shows that $\text{LVC}_{\psi_n(S)}(\mathcal{L}_n^{\cap k}) = \text{VC}_{\psi_n(S)}(\mathcal{L}_n^{\cap k}) = (n-1)k$. □

Applying Corollary 2.10 with a sufficiently large set $S \subset \mathbb{R}$, we obtain the theorem:

Theorem 3.9. *For some constant $\epsilon > 0$ and any n, k,*

$$m_\epsilon^{\text{test}}(\mathcal{L}_n^{\cap k}) = \Omega\left(\frac{nk}{\log(nk)}\right).$$

3.4 Decision Trees

For any parameters n and k, let $\mathcal{T}_{n,k}$ denote the set of functions $f : [0,1]^n \to \{0,1\}$ which can be computed by decision trees with at most k nodes, where each node is of the form "$x_i < t$?" for some $t \in \mathbb{R}$. We can bound the LVC dimension of decision trees using the same argument as for unions of intervals.

Proposition 3.10. *Let $S \subset \mathbb{R}^n$ be any subset of the line $\{x \in \mathbb{R}^n : x_2 = \cdots = x_n = 0\}$ with $|S| > k$. Then $\text{LVC}_S(\mathcal{T}_{n,k}) = \text{VC}_S(\mathcal{T}_{n,k}) = k+1$.*

Proof. Observe that on any sequence $s_1 < s_2 < \cdots < s_m$ in S, any function $f \in \mathcal{T}_{n,k}$ can alternate at most k times, since there are at most k nodes in the decision tree labelled "$x_1 < t$" for some values t. Therefore $T \subseteq S$ is shattered iff $|T| \le k+1$. □

Combining this proposition with Corollary 2.10 completes the proof of the lower bound for testing decision trees:

Theorem 3.11. *For some constant $\epsilon > 0$ and any k, n,*

$$m_\epsilon^{\text{test}}(\mathcal{T}_{n,k}) = \Omega\left(\frac{k}{\log k}\right).$$

4 CLASSES OF BOOLEAN FUNCTIONS

The techniques used in the last section do not carry over to classes of functions over the Boolean hypercube. This is because $\{\pm 1\}^n$ is very far from being in general position—indeed, up to 2^{n-1} points can belong to an affine subspace of dimension $n-1$, by, for example, taking the subspace obtained by setting the first coordinate to 1. In this section, we will instead choose the set S uniformly at random from $\{\pm 1\}^n$ and show that the properties we need for the reduction in Lemma 2.8 hold with high probability.

4.1 Halfspaces

We first introduce some notation and a theorem that will be used also for PTFs in the next subsection. For a vector $a \in \{0,1\}^n$ and $x \in \mathbb{R}^n$ we will write $x^a = \prod_{i=1}^n x_i^{a(i)}$. Write $|a| = \sum_i a(i)$. Let $\psi_k : \mathbb{R}^n \to \mathbb{R}^{\binom{n}{\le k}}$ be defined as follows:

$$\psi_k(x) = (x^a)_{a \in \{0,1\}^n : |a| \le k}.$$

We will use the following theorem of Abbe, Shpilka, & Wigderson [1]:

Theorem 4.1 ([1]). *Let n, k, m be positive integers such that*

$$m < \binom{n - \log\binom{n}{\le k} - t}{\le k}.$$

Then for independent, uniformly random vectors $x_1, \ldots, x_m \sim \{\pm 1\}^n$, the vectors $\psi_k(x_1), \ldots, \psi_k(x_m) \in \{\pm 1\}^{\binom{n}{\le k}}$ are linearly independent with probability at least $1 - 2^{-t}$.

Let $\mathcal{L}_n^{\pm}$ denote the set of halfspaces (or linear threshold functions) over $\{\pm 1\}^n$.

Theorem 4.2. *For some constant $\epsilon > 0$ and all n,*

$$m_\epsilon^{\text{test}}(\mathcal{L}_n^{\pm}) = \Omega\left(\frac{n}{\log n}\right).$$

Proof. Set $m = 5(n+1), \alpha = 1/11, \beta = 4/5$. We will repeat the reduction from $\text{SSD}(m, \alpha, \beta)$ to testing $\mathcal{L}_n$ as in Lemma 2.8 and Theorem 2.9 with the fixed set S replaced by a random set S of size m drawn from $\{\pm1\}^n$. First suppose that the input distribution $\mathcal{D}$ over $[m]$ has support size at most $\alpha m < n/2$. Then $T := \text{supp}(\phi\mathcal{D})$ is a uniformly random subset of $\{\pm1\}^n$ of size at most $n/2$, so since $|T| \le n/2 < n - \log(1+n) - C$ for any constant C, by Theorem 4.1 (with $k = 1$), the points in T are linearly independent with probability at least $9/10$. In this case, T is shattered by $\mathcal{L}_n^{\pm}$, so the remainder of the proof goes through as in Lemma 2.8. When $\mathcal{D}$ has support size at least $\beta m = 4(n+1)$, the proof goes through as in Lemma 2.8 and Corollary 2.10 with the constant $K = 4$, and we obtain the lower bound. □

4.2 Polynomial Threshold Functions

Let $\mathcal{P}_{n,k}$ denote the class of polynomial threshold functions with degree k over $\{\pm1\}^n$. The above mapping $\psi_k : \{\pm1\}^n \to \{\pm1\}^d$ with $d = \binom{n}{\le k}$ establishes an equivalence between PTFs and halfspaces in a higher dimension:

Lemma 4.3. *Write $d = \binom{n}{\le k}$. A set $S \subseteq \mathbb{R}^n$ is shattered by $\mathcal{P}_{n,k}$ if and only if $\psi_k(S)$ is shattered by $\mathcal{L}_d^{\pm}$.*

Proof. We shall index the coordinates of $\{\pm1\}^d$ with vectors $a \in \{0,1\}^n$ satisfying $|a| \le k$. Let $\ell : S \to \{\pm1\}$ be any labelling of S. Note that ψ_k is a bijection (which can be seen just from the vectors a with $|a| = 1$. If there is a degree-k polynomial $p(x) = \sum_{a\in\{0,1\}^n, |a|\le k} w_a x^a$ such that $\text{sign}(p(x)) = \ell(x)$ for every $x \in S$, then for every $x \in S$ we have

$$\ell(x) = \text{sign}(p(x)) = \text{sign}\left(w_0 + \sum_{a\in\{0,1\}^n, |a|\le k} w_a x^a\right)$$
$$= \text{sign}\left(w_0 + \sum_{a\in\{0,1\}^n, |a|\le k} w_a \psi_k(x)_a\right).$$

Observe that the function on the right is an LTF in $\mathcal{L}_d^{\pm}$, so there is an LTF consistent with the labelling $\ell \circ \psi_k^{-1}$ on $\psi_k(S)$. So, if S is shattered by $\mathcal{P}_{n,k}$ then $\psi_k(S)$ is shattered by $\mathcal{L}_d^{\pm}$, because ψ_k acts also as a bijection between labellings of S and $\psi_k(S)$. On the other hand, the same equation shows that for any labelling $\ell : \psi_k(S) \to \{\pm1\}$, if there is an LTF $f : \mathbb{R}^d$ such that $f(\psi_k(x)) = \ell(\psi_k(x))$ for each $x \in \psi_k(S)$ then there is a PTF $g : \mathbb{R}^n \to \{\pm1\}$ such that $g(x) = f(\psi(x)) = \ell(\psi(x))$ for each $x \in S$. Therefore S is shattered by $\mathcal{P}_k$ iff $\psi_k(S)$ is shattered by $\mathcal{L}_d^{\pm}$. □

Theorem 4.4. *Write $\mathcal{P}_{n,k}^{\pm}$ for the set of degree-k PTFs with domain $\{\pm1\}^n$. There exists some constants C' and $\epsilon > 0$ such that for all $k < n/C'$,*

$$m_\epsilon^{\text{test}}(\mathcal{P}_{n,k}^{\pm}) = \Omega\left(\frac{\left(\binom{n}{\le k} - \log\binom{n}{k} - O(1)\right)^2}{\binom{n}{\le k}\log\binom{n}{\le k}}\right) = \Omega\left(\frac{(n/4ek)^k}{k\log(n/k)}\right).$$

Proof. Let $d = \binom{n}{\le k}$ and set $m = 5d$. Let $\beta = 4/5$, $t = \log(10)$, and

$$\alpha := \frac{1}{5}\binom{n}{\le k}^{-1}\binom{n - \log\binom{n}{\le k} - t}{\le k}$$

As was the case with halfspaces, we let S be a uniformly random set of m points drawn from $\{\pm1\}^n$, let $\phi : [m] \to S$ be a random mapping obtained by assigning a uniform and independently random $x \in S$ to each $i \in [m]$, and complete the reduction from $\text{SSD}(m, \alpha, \beta)$ to testing $\mathcal{P}_{n,k}$ as in Lemma 2.8 and Theorem 2.9, which we verify below.

We must first verify that $\alpha \ge C\frac{\sqrt{\log m}}{m^{1/4}}$, where C is the constant in Theorem 2.4, for which it suffices to prove that $\alpha \ge \hat{C}\frac{\sqrt{\log d}}{d^{1/4}}$ for a slightly larger $\hat{C} > C$, since $m = 5d$. For an appropriately large choice of constant C', and sufficiently large $n > 2t$,

$$\log\binom{n}{\le k} + t \le \log\binom{n}{\le n/C'} + t \le \log\left(\left(\frac{en}{n/C'}\right)^{n/C'}\right) + t$$
$$= \frac{n}{C'}\log(eC') + t \le n/2,$$

so

$$\alpha \ge \frac{1}{5}\binom{n}{\le k}^{-1}\binom{n/2}{\le k} \ge \frac{1}{5}\left(\frac{n}{2k}\right)^k\left(\frac{k}{en}\right)^k = \left(\frac{1}{2e}\right)^k.$$

For any constant $\eta > 0$, we may assume $C' > (\hat{C}2e)^{\frac{1}{1/4-\eta}}$, so that, using $\frac{k}{n} \le \frac{1}{C'} \le \frac{1}{(C2e)^{\frac{1}{1/4-\eta}}}$, we get

$$\hat{C}\frac{\sqrt{\log d}}{d^{1/4}} \le C\frac{1}{d^{1/4-\eta}} \le \hat{C}\left(\frac{k}{n}\right)^{k(1/4-\eta)} \le \hat{C}\left(\frac{1}{(\hat{C}2e)^{\frac{1}{1/4-\eta}}}\right)^{k(1/4-\eta)}$$
$$\le \frac{1}{5}\left(\frac{1}{2e}\right)^k \le \alpha.$$

Now we verify correctness. Suppose that the input distribution $\mathcal{D}$ over $[m]$ has support size at most αm and let $T := \text{supp}(\phi\mathcal{D})$. T is a (multi)set of at most

$$\alpha m = d\binom{n}{\le k}^{-1}\binom{n - \log\binom{n}{\le k} - t}{\le k} = \binom{n - \log\binom{n}{\le k} - t}{\le k}$$

uniformly random points from $\{\pm1\}^n$, so by Theorem 4.1 the probability that the points $\psi_k(T)$ are linearly independent is at least $9/10$. In that case, $\psi_k(T)$ is shattered by the halfspaces $\mathcal{H}_d$ over $\{\pm1\}^d$ so by Lemma 4.3, T is shattered by $\mathcal{P}_{n,k}$. Therefore, as in Lemma 2.8, the tester for $\mathcal{P}_{n,k}$ will output 1 with probability at least $5/6$, so the distribution $\mathcal{D}$ is accepted with probability at least $2/3$.

Now suppose that the input distribution $\mathcal{D}$ over $[m]$ has support size at least $\beta m = 4d$, and let $T = \text{supp}(\phi\mathcal{D})$. Since ϕ is a random mapping (with replacement), we must first show that, with high probability, $|T| \ge Kd$ for the constant $K > 3.04$ in Lemma 2.7. Since $k \le n/C'$ for a sufficiently large constant C', we have $4d = 4\binom{n}{\le k} \le 4(eC')^{n/C'} \le 2^{cn}$ for constant $c < 1/3$. Therefore the probability that a random point x in T is unique is at least $1 - \frac{4d}{2^n} \ge 1 - 2^{(c-1)n}$. By the union bound, the probability that any point fails to be unique is at most $4d2^{(c-1)n} = 4\binom{n}{\le k}2^{(c-1)n} \le 2^{(2c-1)n} < 2^{-n/3}$. When this occurs, the support of $\phi\mathcal{D}$ has size at least $4d$ so, as in Theorem 2.9, we may apply Lemma 2.7 to conclude that a random labelling $f : \{\pm1\}^n \to \{\pm1\}$ satisfies $\text{dist}_{\phi\mathcal{D}}(f, \mathcal{P}_{n,k}) \ge \epsilon$ with probability

at least 9/10, for some small enough constant $\epsilon > 0$. Then the tester for $\mathcal{P}_{n,k}$ will output 0 with probability at least 5/6, so the distribution $\mathcal{D}$ is rejected with probability at least 2/3.

We obtain a lower bound of $\Omega\left(\frac{d}{\log d}\log^2\frac{1}{1-\alpha}\right)$, since $1-\beta \geq \alpha$. Using the inequality $\log^2\frac{1}{1-x} \geq \log^2\frac{1}{e^{-x}} = \log^2(e^x) = \Omega(x^2)$, we get

$$\frac{d}{\log d}\log^2\frac{1}{1-\alpha} = \Omega\left(\frac{d}{\log d}\alpha^2\right) = \Omega\left(\frac{\binom{n-\log(d)-t}{\leq k}^2}{d\log d}\right).$$

To obtain the simplified bound, use $n-\log(d)-t \leq n/2$ from above, and $\binom{n/2}{\leq k} \geq (n/2k)^k$ to get

$$\Omega\left(\frac{(n/2k)^{2k}}{d\log d}\right) = \Omega\left(\frac{(n/2k)^{2k}}{(en/k)^k k\log(en/k)}\right) = \Omega\left(\frac{(n/4ek)^k}{k\log(n/k)}\right).$$

□

4.3 Decision Trees

Let $\mathcal{B}_{n,k}$ be the set of functions $f : \{0,1\}^n \to \{0,1\}$ defined by decision trees with k nodes of the form "$x_i = 1$?". When $k \gg \log n$, fairly tight bounds on the VC dimension of $\mathcal{B}_{n,k}$ are known, due to Mansour [37]:

Lemma 4.5 ([37]). *$\mathrm{VC}(\mathcal{B}_{n,k})$ is between $\Omega(k)$ and $O(k\log n)$.*

A lower bound on the LVC dimension of $\mathcal{B}_{n,k}$ is also easily established.

Proposition 4.6. *Every subset $T \subseteq \{0,1\}^n$ of size at most k is shattered by $\mathcal{B}_{n,k}$.*

Proof. We prove by induction on k that any set $T \subseteq S$ of size k is shattered by a decision tree with at most k leaves. Clearly when $k = 1$, for any subset $T \subseteq S$ of size $|T| = 1$, decision trees with 0 nodes and 1 leaf shatter T. For $k > 1$, there exists a coordinate $i \in [n]$ such that $T_0 := \{x \in T : x_i = 0\} \neq \emptyset$ and $T_1 := \{x \in T : x_i = 1\} \neq \emptyset$. Now T_0 is a subset of size $k - |T_1| < k$ so by induction it is shattered by subtrees with at most $k - |T_1|$ leaves, while T_1 is shattered by subtrees with at most $|T_1|$ leaves. Therefore T is shattered by a tree with at most k leaves. Since the number of nodes is at most the number of leaves, T is shattered by $\mathcal{B}_{n,k}$. □

We are now ready to bound the sample and tolerant-query complexities for testing decision trees.

Theorem 4.7. *For any k, $n \geq \log k + \log\log k + \Omega(1)$, and sufficiently small constant $\epsilon > 0$,*

$$m_\epsilon^{\mathrm{test}}(\mathcal{B}_{n,k}) = \Omega\left(\frac{k}{\log k \cdot \log\log k}\right).$$

Proof. Let $S \subset \{0,1\}^n$ be a subcube with dimension $m = \log(6C) + \log k + \log\log\log k$ and let $d = \mathrm{VC}_S(\mathcal{B}_{n,k})$. Then by Lemma 4.5, for some constant C and sufficiently large k,

$$\begin{aligned} d &\leq Ck\log(m) = Ck\log\log(6Ck\log\log k) \\ &\leq Ck\log\log(k^2) = Ck(\log\log k + 1), \end{aligned}$$

so that

$$\begin{aligned} (1-\delta)|S| = (1-\delta)2^m &= 6Ck\log\log k - k \\ &= 5Ck\log\log k + Ck(\log\log k - 1/C) \\ &\geq 5Ck(\log\log k + 1) \geq 5d. \end{aligned}$$

By Proposition 4.6, $\mathrm{LVC}_S(\mathcal{B}_{n,k}) \geq k$, so for $\delta = \frac{1}{6C\log\log k}$,

$$\mathrm{LVC}_S(\mathcal{B}_{n,k}) \geq k = \delta 6Ck\log\log k = \delta|S|,$$

therefore the conditions for Theorem 2.9 are satisfied. We obtain a lower bound of

$$\Omega\left(\frac{k\log\log k}{\log k}\log^2\frac{1}{1-\frac{1}{\log\log k}}\right).$$

Using the inequality $\log^2\frac{1}{1-1/x} \geq \log^2\frac{1}{e^{-1/x}} = \Omega(1/x^2)$, we get the lower bound of

$$\Omega\left(\frac{k\log\log k}{(\log k)(\log\log(k))^2}\right) = \Omega\left(\frac{k}{\log k\cdot\log\log k}\right).$$

□

5 MAXIMUM CLASSES AND ANALYTIC DUDLEY CLASSES

A number of sample complexity lower bounds for testing natural classes of functions can be obtained by considering maximum and analytic Dudley classes, as we describe in this section.

5.1 LVC and the Sauer–Shelah–Perles Lemma

Recall the Sauer–Shelah–Perles lemma and the associated definitions:

Let $\mathcal{H}$ be a set of functions $\mathcal{X} \to \{0,1\}$ and let $S \subseteq \mathcal{X}$. The *shattering number* is

$$\mathrm{sh}(\mathcal{H}, S) := |\{T \subseteq S \mid T \text{ is shattered by } \mathcal{H}\}|,$$

and the *growth function* is

$$\Phi(\mathcal{H}, S) := |\{\ell : S \to \{0,1\} \mid \exists h \in \mathcal{H}\ \forall x \in S, \ell(x) = h(x)\}|.$$

Sauer–Shelah–Perles lemma. *Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ and let $S \subseteq \mathcal{X}$ with $\mathrm{VC}_S(\mathcal{H}) = d$. Then $\Phi(\mathcal{H}, S) \leq \mathrm{sh}(\mathcal{H}, S) \leq \sum_{i=0}^d \binom{|S|}{i}$.*

Much research has studied the cases where this inequality is tight in various ways: A class is called *maximum* on S ([5, 16, 24, 25, 32, 35, 40]) if the sequence of inequalities is tight, i.e. $\mathcal{H}$ is maximum on S if

$$\Phi(\mathcal{H}, S) = \mathrm{sh}(\mathcal{H}, S) = \sum_{i=0}^d \binom{|S|}{i}.$$

A class is called *shatter-extremal* on S (see e.g. [16, 39, 40]) if the first inequality is tight, i.e.

$$\Phi(\mathcal{H}, S) = \mathrm{sh}(\mathcal{H}, S).$$

We are not aware of any studies of the case where the second inequality $\mathrm{sh}(\mathcal{H}, S) \leq \sum_{i=0}^d \binom{|S|}{i}$ is tight; our requirement $\mathrm{LVC}_S(\mathcal{H}) = \mathrm{VC}_S(\mathcal{H})$ fills in the gap:

Proposition 5.1. *A set $\mathcal{H}$ of functions $\mathcal{X} \to \{0,1\}$ satisfies $\mathrm{LVC}_S(\mathcal{H}) = \mathrm{VC}_S(\mathcal{H})$ on a set $S \subseteq \mathcal{X}$ if and only if $\mathrm{sh}(\mathcal{H}, S) = \sum_{i=0}^d \binom{|S|}{i}$, for $d = \mathrm{VC}_S(\mathcal{H})$.*

Proof. This follows from the fact that $\sum_{i=0}^{d}\binom{|S|}{i}$ is exactly the number of sets of size at most d; if the equality holds, all such sets are shattered, so $\mathrm{LVC}_S(\mathcal{H}) = d$. On the other hand if $\mathrm{LVC}_S(\mathcal{H}) = d$ then all sets of size at most d are shattered, so the equality holds. □

We can therefore conclude:

Proposition 5.2. *A set $\mathcal{H}$ of functions $\mathcal{X} \to \{0,1\}$ is maximum on $S \subseteq \mathcal{X}$ if and only if it is both shatter-extremal on S and $\mathrm{LVC}_S(\mathcal{H}) = \mathrm{VC}_S(\mathcal{H})$.*

Then we easily obtain lower bounds for maximum classes using Corollary 2.10.

Theorem 5.3. *Let $\mathcal{H}$ be a set of functions $\mathcal{X} \to \{0,1\}$. Suppose there is $S \subseteq \mathcal{X}$ such that $\mathcal{H}$ is maximum on S and $d := \mathrm{VC}_S(\mathcal{H})$ satisfies $|S| \geq 5d$. Then for sufficiently small constant $\epsilon > 0$,*

$$m_\epsilon^{\mathrm{test}}(\mathcal{H}) = \Omega\left(\frac{d}{\log d}\right).$$

Examples of maximum classes include the set of functions $f : [n] \to \{0,1\}$ with at most $n/5$ 1-valued points [40] (studied in detail in the full version [10]), unions of k intervals [23], and positive halfspaces (halfspaces with normal vectors $w \in \mathbb{R}^n$ satisfying $x_i \geq 0$) [24]. Another standard example is the set of sign vectors arising from an arrangement of hyperplanes:

Example 5.4 ([25]). Let H be a set of $n > d$ hyperplanes in $\mathbb{R}^d$ and write $H = \{h_1, \ldots, h_n\}$ where each $h_i : \mathbb{R}^d \to \{\pm 1\}$ is of the form $h_i(x) = \mathrm{sign}(t + \sum_{j=1}^{d} w_j x_j)$ for some $t, w_j \in \mathbb{R}$. Assume that the hyperplanes are in general position. Let $\mathcal{H}$ be the set of functions $f_x : [n] \to \{\pm 1\}$ obtained by choosing $x \in \mathbb{R}^d$ obtained by setting $f_x(i) = h_i(x)$. Then $\mathrm{VC}_{[n]}(\mathcal{H}) = d$ and $\mathcal{H}$ is maximum on $[n]$, as proved by Gartner & Welzl [25]. Therefore, for any such set $\mathcal{H}$ where $n \geq 5d$ we obtain via Theorem 5.3 that $m_\epsilon^{\mathrm{test}}(\mathcal{H}) = \Omega(d/\log d)$.

5.2 Analytic Dudley Classes

Some examples of maximum classes and classes with $\mathrm{LVC}_S(\mathcal{H}) = \mathrm{VC}_S(\mathcal{H})$ that are arguably more pertinent to property testing can be obtained from a family of classes called *Dudley classes* [7].

Definition 5.5 (Dudley Class). A class $\mathcal{H}$ of functions $\mathcal{X} \to \{\pm 1\}$ is a *Dudley class* if there exists a set $\mathcal{F}$ of functions $\mathcal{X} \to \mathbb{R}$ and a function $h : \mathcal{X} \to \mathbb{R}$ such that:

- $\mathcal{F}$ is a vector space, i.e. $\forall f, g \in \mathcal{F}, \lambda \in \mathbb{R}$, $f + g \in \mathcal{F}$ and $\lambda f \in \mathcal{F}$;
- Every $g \in \mathcal{H}$ can be written as $g(x) = \mathrm{sign}(f(x) + h(x))$.

We will refer to $\mathcal{F}$ as the vector space of $\mathcal{H}$ and h as the threshold of $\mathcal{H}$.

The VC dimension of Dudley classes is equal to the dimension of the vector space $\mathcal{F}$:

Theorem 5.6 ([54] Theorem 3.1). *Let $\mathcal{H}$ be any Dudley class with vector space $\mathcal{F}$. Then $\mathrm{VC}(\mathcal{H}) = \dim(\mathcal{F})$.*

This theorem implies that $\mathrm{LVC}_S(\mathcal{H}) = \mathrm{VC}_S(\mathcal{H})$ on a set $S \subseteq \mathcal{X}$ if and only if the dimension of the vector space remains the same when restricted to any subset of S:

Corollary 5.7. *Let $\mathcal{H}$ be a Dudley class of functions $\mathcal{X} \to \{\pm 1\}$ with vector space $\mathcal{F}$ of functions $\mathcal{X} \to \mathbb{R}$ and threshold h. Then for any set $S \subseteq \mathcal{X}$, $\mathrm{VC}_S(\mathcal{H}) = \mathrm{LVC}_S(\mathcal{H})$ if and only if the vector space $\mathcal{F}$ restricted to any $T \subseteq S$ of size $|T| = d = \mathrm{VC}_S(\mathcal{H})$ has dimension d.*

Proof. This follows from the above theorem, since for any $T \subseteq S$ of size $|T| = d$ on which $\mathcal{F}$ has dimension d, $\mathrm{VC}_T(\mathcal{H}) = d$, so T is shattered. □

A useful condition on Dudley classes that guarantees the above condition was described by Johnson [32]. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *analytic* if it is infinitely differentiable and for every x in the domain, there is an open set $U \ni x$ such that f is equal to its Taylor series expansion on U. We will call a Dudley class *analytic* if its threshold h and each f in the basis of $\mathcal{F}$ is analytic. Johnson proves the following (rewritten in our terminology):

Theorem 5.8 ([32]). *Let $\mathcal{H}$ be any analytic Dudley class on domain $[0,1]^n$ with $\mathrm{VC}(\mathcal{H}) = d$. Then for any $N > n$ there exists a set $S \subset [0,1]^n$ of size $|S| = N$ such that $\mathcal{H}$ is maximum on S with $\mathrm{VC}_S(\mathcal{H}) = d$.*

Then by taking $N \geq 5d$ in the above theorem and applying Theorem 5.3, we obtain:

Corollary 5.9. *Let $\mathcal{H}$ be any analytic Dudley class and suppose $\mathrm{VC}(\mathcal{H}) = d$. Then for some constant $\epsilon > 0$,*

$$m_\epsilon^{\mathrm{test}}(\mathcal{H}) = \Omega\left(\frac{d}{\log d}\right).$$

Examples of analytic Dudley classes include halfspaces (for which we have already proved the lower bound) and PTFs. Other examples due to [32] are balls in $\mathbb{R}^n$ and trigonometric polynomial threshold functions in $\mathbb{R}^d$:

Theorem 5.10. *For some constant $\epsilon > 0$, the following classes $\mathcal{H}$ satisfy the following bounds:*

(1) *Degree-k PTFs on domain $\mathbb{R}^n$ satisfy*

$$m_\epsilon^{\mathrm{test}}(\mathcal{H}) = \Omega\left(\frac{\binom{n+k}{k}}{\log\binom{n+k}{k}}\right).$$

(2) *Balls in $\mathbb{R}^n$, i.e. functions $f : \mathbb{R}^n \to \{\pm 1\}$ of the form $f(x) = \mathrm{sign}(t - \|x - z\|_2)$, satisfy $m_\epsilon^{\mathrm{test}}(\mathcal{H}) = \Omega\left(\frac{n}{\log n}\right)$.*

(3) *Signs of trigonometric polynomials, i.e. functions $\mathbb{R}^2 \to \{\pm 1\}$ of the form:*

$$f(x,y) = \mathrm{sign}\left(t + \sum_{k=1}^{d} a_k \cos(kx) + \sum_{k=1}^{d} b_k \sin(kx) - y\right),$$

satisfy $m_\epsilon^{\mathrm{test}}(\mathcal{H}) = \Omega\left(\frac{d}{\log d}\right)$.

6 OTHER MODELS OF TESTING

In this section, we mention two additional results that we obtain using the same framework as the results above: lower bounds for testing clusterability and for testing feasibility of LP-type problems. We will prove the first result, and leave the proof of the second result for the full version of this paper [10].

6.1 Testing Clusterability

For a point $x \in \mathbb{R}^n$ and radius $r > 0$, define $B_r(x) = \{y \in \mathbb{R}^n : \|x-y\|_2 \le r\}$. Alon, Dar, Parnas, & Ron [2] introduced the problem of testing clusterability with radius cost:

Definition 6.1 (Radius Clustering). Say that a probability distribution $\mathcal{D}$ over $\mathbb{R}^n$ is *k-clusterable* if there exist k centers $c_1, \dots, c_k \in \mathbb{R}^n$ such that $\mathrm{supp}(\mathcal{D}) \subseteq \cup_{i=1}^k B_1(c_i)$. An ϵ-tester for k-clusterability is a randomized algorithm A that is given sample access to $\mathcal{D}$ and must satisfy the following:

(1) If $\mathcal{D}$ is k-clusterable then $\mathbb{P}[A(\mathcal{D}) = 1] \ge 2/3$; and,
(2) If $\mathcal{D}$ is ϵ-far from being k-clusterable in total variation distance, then $\mathbb{P}[A(\mathcal{D}) = 0] \ge 2/3$.

Alon *et al.* [2] prove an upper bound of $O\left(\frac{nk\log(nk)}{\epsilon}\right)$ samples for one-sided testing of k-clusterability when the distribution is uniform over an unknown set of points. Their proof is by VC dimension arguments. The following theorem updates the upper bound of [2] using modern VC dimension results (see the full version of this paper [10], and also [29]).

Theorem 6.2 (Improved version of [2]). *There is a one-sided, distribution-free ϵ-tester for k-clusterability in $\mathbb{R}^n$ with sample complexity $O\left(\frac{nk\log k}{\epsilon}\log\frac{1}{\epsilon}\right)$.*

In this section, we prove a nearly-optimal $\Omega(nk/\log(nk))$ lower bound on distribution-free testers for k-clusterability with two-sided error.

Let $S_r^n = \{x \in \mathbb{R}^n : \|x\|_2 = r\}$ be the points on the hypersphere of radius r.

Proposition 6.3. *For every $\delta > 0$ there is $\eta > 0$ such that a uniformly random set of n points P drawn from $S_{1+\eta}^n$ is contained within some ball $B_1(x)$ with probability at least $1-\delta$.*

Proof. Unless all n points in P lie on a hyperplane through the origin (which occurs with probability 0), there is a hyperplane through the origin such that all points in P lie on one side. Consider the distribution of P conditional on this event, and without loss of generality assume that the hyperplane is $\{x : x_1 = 0\}$ so that all points $x \in P$ satisfy $x_1 > 0$. Let $\eta > 0$ and consider the ball B of radius 1 centered at $z = (\sqrt{(1+\eta)^2 - 1}, 0, \dots, 0)$. Let $x \in S_{1+\eta}^n$ satisfy $x_1 \ge z_1 = \sqrt{(1+\eta)^2-1} = \sqrt{\eta(2-\eta)}$. Then since $\|x\|_2^2 = (1+\eta)^2$,

$$\|x-z\|_2^2 = (x_1-z_1)^2 + \sum_{i=2}^n x_i^2 = (x_1-z_1)^2 + (1+\eta)^2 - x_1^2$$
$$= z_1^2 - 2x_1z_1 + (1+\eta)^2 \le (1+\eta)^2 - z_1^2 = 1,$$

so all points x with $x_1 \ge z_1$ are contained within the ball B. Conditioned on $x_1 > 0$, the probability that $x_1^2 \ge \eta(2-\eta)$ is at least the probability that $y_1^2 \ge \eta(2-\eta)$ for y drawn uniformly randomly from S_1^n. This probability goes to 1 as $\eta \to 0$, so the probability that $x_1^2 \ge \eta(2-\eta)$ also approaches 1 as $\eta \to 0$. The conclusion follows. □

Proposition 6.4. *For every constant $\delta, \eta > 0$, there is a constant $\epsilon_0 > 0$ such that, for all $\epsilon < \epsilon_0$ and for a uniformly random set P of $m = 2n$ points drawn from $S_{1+\eta}^n$, with probability at least $1 - e^{-\delta n}$, no subset $T \subset P$ of size $(1-\epsilon)m$ is contained within a ball of radius 1.*

Proof. Let $t = (1-\epsilon)m > n$ and let $T \subset P$ have size $|T| = t$. If the points T are contained within a ball of radius 1 then they are contained within a centered halfspace, because the intersection of the ball with $S_{1+\eta}^n$ is equal to the intersection of some halfspace with $S_{1+\eta}^n$. The probability that t uniformly random points on the surface of the sphere lie within some hemisphere is $2^{1-t}\sum_{k=0}^{n-1}\binom{t-1}{k}$ [53]. There are at most $\binom{m}{t}$ subsets of size t, so the probability that any of these subsets lie within a hemisphere is at most

$$\binom{m}{m-t}2^{1-t}\left(\frac{et}{n}\right)^n \le 2^{1-(1-\epsilon)m}\left(\frac{e}{\epsilon}\right)^{\epsilon m}\left(\frac{et}{n}\right)^n$$
$$= 2^{1+\epsilon m\log(e/\epsilon)-(1-\epsilon)m+n\log\left(\frac{e(1-\epsilon)m}{n}\right)}$$
$$= 2^{1+\epsilon 2n\log(e/\epsilon)-(1-\epsilon)2n+n\log(e(1-\epsilon)2)}$$
$$\le 2^{1-2n(1-\epsilon\log(4e^2/\epsilon))}.$$

The conclusion holds since $\epsilon\log(4e^2/\epsilon) \to 0$ as $\epsilon \to 0$. □

Proposition 6.5 (Balls and bins). *Fix $C > 0$, $0 < \delta \le 1$, and let n, k be positive integers with $k \le \frac{1}{10}e^{\delta^2 Cn/3}$. Then if Cnk balls are deposited into k bins uniformly at random, the following hold:*

(1) *With probability at least $9/10$, every bin receives at most $(1+\delta)Cn$ balls;*
(2) *With probability at least $9/10$, every bin receives at least $(1-\delta)Cn$ balls.*

Proof. Let X_{ij} be the indicator variable for the event that the i-th ball goes into the j-th bin, and let the random variable $L_j = \sum_{i=1}^{Cnk} X_{ij}$ denote the final load on the j-th bin. Note that $\mathbb{E}[L_j] = Cn$. By the multiplicative Chernoff bound, we have:

(1) $\mathbb{P}[L_j \ge (1+\delta)Cn] \le e^{-\delta^2 Cn/3}$; and
(2) $\mathbb{P}[L_j \le (1-\delta)Cn] \le e^{-\delta^2 Cn/3}$.

In both cases, by the union bound, the probability that the respective event occurs for any L_j $(1 \le j \le k)$ is at most $k \cdot e^{-\delta^2 Cn/3} \le 1/10$, as desired. □

Lemma 6.6. *For $k < \frac{1}{10}e^{n/6}$, let $A_1, \dots, A_k$ be spheres in $\mathbb{R}^n$ of radius $1+\eta$ for sufficiently small $\eta > 0$, such that the minimum distance between any two spheres is 3. Define the following distribution $\mathcal{S}$ over $\bigcup_{i=1}^n A_i$: Draw $i \in [k]$ uniformly at random and then draw $x \sim A_i$ uniformly at random. Then:*

(1) *If S is a set of $m \le nk/2$ independent points drawn from $\mathcal{S}$, then with probability at least $9/10$, there are k balls of radius 1 whose union contains S;*
(2) *If S is a set of $4nk \le m \le 8nk$ independent points drawn from $\mathcal{S}$ and $\epsilon > 0$ is a sufficiently small constant, then with probability at least $81/100$, no union of k balls of radius 1 contains more than $(1-\epsilon)m$ points of S.*

Proof. First suppose that $m \le nk/2$. If each sphere A_i receives at most n sample points then by Proposition 6.3, setting $\delta, \eta > 0$ arbitrarily small in the statement of that proposition, for each sphere A_i there is a ball B_i of radius 1 containing all points $S \cap A_i$ with

probability arbitrarily close to 1, so there are k balls containing all points of S. Proposition 6.5 (with $C = 1/2$ and $\delta = 1$) shows that the maximum load of any sphere is at most n with probability at least 9/10, so the first conclusion holds.

Now suppose that $4nk \le m \le 8nk$. Note that no ball of radius 1 can contain points from more than 1 sphere A_i. Proposition 6.5 (with $C = 4$ and $\delta = 1/2$) shows that the minimum load of any sphere is at least $2n$ with probability at least 9/10. Assume that this occurs for the rest of this argument.

Let $S_i = S \cap A_i$ for $i = 1, \ldots, k$, and say that S_i is *difficult* if no ball of radius 1 contains at least $(1 - \epsilon')|S_i|$ points in S_i, for constant ϵ' to be defined. Since $|S_i| \ge 2n$, Proposition 6.4 gives that $\mathbb{P}[S_i \text{ is difficult}] \ge 1 - e^{-\delta n}$. Setting $\delta = 1/6$ and by the union bound, the probability that every S_i is difficult is at least $1 - k \cdot e^{-\delta n} \ge 1 - \frac{1}{10} e^{n/6} e^{-\delta n} = 9/10$. Fix ϵ' corresponding to $\delta = 1/6$ in Proposition 6.4.

Assume that every S_i is difficult, and consider any set of k balls $B_1, \ldots, B_k$. Denote their union by $B = \bigcup_i B_i$. Then for each S_i, we have that $|B \cap S_i| \ge (1 - \epsilon')|S_i|$ only if at least two balls B_{j_1}, B_{j_2} intersect S_i. Thus, this can only happen for at most $k/2$ such S_i's. Assume without loss of generality that $S_1, \ldots, S_\ell$ have at least $(1 - \epsilon')$-fraction of their points covered by B, so that $\ell \le k/2$. It follows that

$$|S \setminus B| \ge \sum_{i=\ell+1}^{k} \epsilon' |S_i| \ge \frac{k}{2} \cdot \epsilon' \cdot 2n \ge \frac{\epsilon' m}{8}.$$

Which satisfies the second claim for $\epsilon = \epsilon'/8$, and this happens with probability at least $9/10 \cdot 9/10 = 81/100$ over the choice of S. □

Theorem 6.7 (Restatement of Theorem 1.3). *For sufficiently small constant $\epsilon > 0$, any ϵ-tester for k-clusterability in $\mathbb{R}^n$ requires at least $\Omega\left(\frac{nk}{\log(nk)}\right)$ samples.*

Proof. Let $N = 8nk$ and let $\alpha = 1/16, \beta = 1/2$. We will prove a reduction from support-size distinction to k-clusterability; we may assume that the tester for k-clusterability has success probability at least 5/6 due to standard boosting techniques. For an input distribution $\mathcal{D}$ over $[N]$ with densities at least $1/N$, construct spheres $A_1, \ldots, A_k$ as in Lemma 6.6. Construct the map $\phi : [N] \to \bigcup_{i=1}^{k} A_i$ by sampling $s_1, \ldots, s_N \sim \mathcal{S}$, where $\mathcal{S}$ is the distribution from Lemma 6.6, and setting $\phi(i) = s_i$. Then simulate the tester for k-clusterability by giving the tester samples $\phi(i)$ for $i \sim \mathcal{D}$. We will write $\phi\mathcal{D}$ for the distribution over $\bigcup_{i=1}^{k} A_i$ obtained by sampling $i \sim \mathcal{D}$ and returning $\phi(i)$.

First suppose that $|\operatorname{supp}(\mathcal{D})| \le \alpha N$. Then $\operatorname{supp}(\phi\mathcal{D})$ is a set of at most $\alpha N = nk/2$ points sampled from $\mathcal{S}$, so by Lemma 6.6, with probability at least 9/10 over the choice of ϕ the distribution $\phi\mathcal{D}$ is k-clusterable, so the tester will output 1 with probability at least 5/6, so the total probability of success is at least 2/3.

Next suppose that $|\operatorname{supp}(\mathcal{D})| \ge \beta N$ so $\operatorname{supp}(\phi\mathcal{D})$ is a set of between $\beta N = 4nk$ and $N = 8nk$ points sampled from $\mathcal{S}$. Then by Lemma 6.6, for sufficiently small constant $\epsilon > 0$, with probability at least 81/100 over the choice of ϕ, $X := \operatorname{supp}(\phi\mathcal{D})$ is at least ϵ/β-far from k-clusterable according to the uniform distribution over X. Since $\mathcal{D}$ (and therefore $\phi\mathcal{D}$) has densities at least $1/N$ on X, any k-clusterable distribution $\phi\mathcal{D}$ must be at least $\frac{(\epsilon/\beta)|X|}{N} \ge \epsilon$-far from $\phi\mathcal{D}$. Therefore the ϵ-tester will output 0 with probability at least 5/6, so the total probability to output 0 is at least 2/3. So the algorithm solves support-size distinction with parameters $N = 8nk, \alpha = 1/16, \beta = 1/2$. Finally, by Theorem 2.4, the number of samples required is at least $\Omega\left(\frac{N}{\log N}\right) = \Omega\left(\frac{nk}{\log(nk)}\right)$. □

6.2 Testing LP-Type Problems

Epstein & Silwal [21] recently introduced property testing for LP-Type problems, which are problems that generalize linear programming. The algorithm has query access to a set S of constraints and must determine with high probability whether an objective function ϕ satisfies $\phi(S) \le k$ or if at least an ϵ-fraction of constraints must be removed in order to satisfy $\phi(S) \le k$. We refer the reader to their paper for the definition of their model. One of their applications is the following upper bound for testing feasibility of an LP, where the algorithm is allowed to query the linear constraints:

Theorem 6.8 ([21]). *There is a tester for feasibility in $\mathbb{R}^n$ with two-sided error and sample complexity $O(n/\epsilon)$.*

Testing if a set $X \subseteq \mathbb{R}^n$ with labels $\ell : X \to \{\pm 1\}$ is realizable by a halfspace can be solved by their algorithm, since for each $x \in X$ one can add the constraint $\ell(x) \cdot (w_0 + \sum_{i=1}^{n} w_i x_i) \ge 1$ to S, with variables $w_0, w_1, \ldots, w_n$. We prove that this is nearly optimal:

Theorem 6.9. *Testing with two-sided error whether a set $X \subseteq \mathbb{R}^n$ with labels $\ell : X \to \{\pm 1\}$ is realizable by a halfspace or whether at least $\epsilon|X|$ labels must be changed to become realizable by a halfspace requires at least $n^{1-o(1)}$ samples.*

The technique for this proof differs slightly from the other results in this paper. For this result, we must use lower bounds for Support-Size Distinction that use probability distributions with integral densities. Such lower bounds were proved in [43]. See the full version [10] for details.

ACKNOWLEDGEMENTS

We thank Cameron Seth for many interesting discussions during the course of this research and the anonymous referees for valuable feedback on the manuscript.

REFERENCES

[1] Emmanuel Abbe, Amir Shpilka, and Avi Wigderson. 2015. Reed–Muller codes for random erasures and errors. *IEEE Transactions on Information Theory* 61, 10 (2015), 5229–5252. https://doi.org/10.1109/TIT.2015.2462817

[2] Noga Alon, Seannie Dar, Michal Parnas, and Dana Ron. 2003. Testing of clustering. *SIAM Journal on Discrete Mathematics* 16, 3 (2003), 393–417. https://doi.org/10.1137/S0895480102410973

[3] Noga Alon, Jacob Fox, and Yufei Zhao. 2019. Efficient arithmetic regularity and removal lemmas for induced bipartite patterns. *Discrete Analysis* 2019 (2019), 14 pp. Issue 3. https://doi.org/10.19086/da.7757

[4] Noga Alon, Rani Hod, and Amit Weinstein. 2016. On Active and Passive Testing. *Combinatorics, Probability and Computing* 25 (2016), 1–20. https://doi.org/10.1017/S0963548315000292

[5] Noga Alon, Shay Moran, and Amir Yehudayoff. 2016. Sign rank versus VC dimension. In *Conference on Learning Theory*. 47–80. http://proceedings.mlr.press/v49/alon16.html

[6] Maria-Florina Balcan, Eric Blais, Avrim Blum, and Liu Yang. 2012. Active property testing. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. IEEE, 21–30. https://doi.org/10.1109/FOCS.2012.64

[7] Shai Ben-David and Ami Litman. 1998. Combinatorial variability of Vapnik-Chervonenkis classes with applications to sample compression schemes. *Discrete*

Applied Mathematics 86, 1 (1998), 3–25. https://doi.org/10.1016/S0166-218X(98)00000-6

[8] Piotr Berman, Meiram Murzabulatov, and Sofya Raskhodnikova. 2019. The power and limitations of uniform samples in testing properties of figures. *Algorithmica* 81, 3 (2019), 1247–1266. https://doi.org/10.1007/s00453-018-0467-9

[9] Eric Blais, Joshua Brody, and Kevin Matulef. 2012. Property Testing Lower Bounds via Communication Complexity. *Computational Complexity* 21, 2 (2012), 311–358. https://doi.org/10.1007/s00037-012-0040-x

[10] Eric Blais, Renato Ferreira Pinto Jr., and Nathaniel Harms. 2020. VC Dimension and Distribution-Free Sample-Based Testing. arXiv:2012.03923 [cs.LG] https://arxiv.org/abs/2012.03923

[11] Eric Blais and Yuichi Yoshida. 2019. A characterization of constant-sample testable properties. *Random Structures & Algorithms* 55, 1 (2019), 73–88. https://doi.org/10.1002/rsa.20807

[12] Avrim Blum and Lunjia Hu. 2018. Active tolerant testing. In *Proceedings of the 31st Conference On Learning Theory*. http://proceedings.mlr.press/v75/blum18a.html

[13] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)* 36, 4 (1989), 929–965. https://doi.org/10.1145/76359.76371

[14] Nader H. Bshouty. 2020. Almost Optimal Testers for Concise Representations. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2020)*, Jarosław Byrka and Raghu Meka (Eds.), Vol. 176. 5:1–5:20. https://doi.org/10.4230/LIPIcs.APPROX/RANDOM.2020.5

[15] Sourav Chakraborty, David García-Soriano, and Arie Matsliah. 2011. Efficient Sample Extractors for Juntas with Applications. In *Automata, Languages and Programming - 38th International Colloquium (ICALP 2011) (Lecture Notes in Computer Science, Vol. 6755)*. Springer, 545–556. https://doi.org/10.1007/978-3-642-22006-7_46

[16] Jérémie Chalopin, Victor Chepoi, Shay Moran, and Manfred K Warmuth. 2019. Unlabeled Sample Compression Schemes and Corner Peelings for Ample and Maximum Classes. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, Vol. 132. 34. https://doi.org/10.4230/LIPIcs.ICALP.2019.34

[17] Xi Chen, Adam Freilich, Rocco A Servedio, and Timothy Sun. 2017. Sample-Based High-Dimensional Convexity Testing. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*. https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2017.37

[18] Mónika Csikós, Nabil H Mustafa, and Andrey Kupavskii. 2019. Tight Lower Bounds on the VC-dimension of Geometric Set Systems. *J. Mach. Learn. Res.* 20 (2019), 81–1. https://www.jmlr.org/papers/volume20/18-719/18-719.pdf

[19] Anindya De, Elchanan Mossel, and Joe Neeman. 2019. Is your function low dimensional?. In *Conference on Learning Theory (COLT 2019) (Proceedings of Machine Learning Research, Vol. 99)*. PMLR, 979–993. http://proceedings.mlr.press/v99/de19a.html

[20] Ilias Diakonikolas, Prahladh Harsha, Adam Klivans, Raghu Meka, Prasad Raghavendra, Rocco A Servedio, and Li-Yang Tan. 2010. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *Proceedings of the forty-second ACM symposium on Theory of computing*. 533–542. https://doi.org/10.1145/1806689.1806763

[21] Rogers Epstein and Sandeep Silwal. 2020. Property Testing of LP-Type Problems. In *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*, Artur Czumaj, Anuj Dawar, and Emanuela Merelli (Eds.), Vol. 168. 98:1–98:18. https://doi.org/10.4230/LIPIcs.ICALP.2020.98

[22] Noah Fleming and Yuichi Yoshida. 2020. Distribution-Free Testing of Linear Functions on $\mathbb{R}^n$. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. https://doi.org/10.4230/LIPIcs.ITCS.2020.22

[23] Sally Floyd. 1989. *Space-bounded learning and the Vapnik-Chervonenkis dimension*. Ph.D. Dissertation. University of California, Berkeley. https://www.icsi.berkeley.edu/pubs/techreports/tr-89-61.pdf

[24] Sally Floyd and Manfred Warmuth. 1995. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine learning* 21, 3 (1995), 269–304. https://doi.org/10.1023/A:1022660318680

[25] Bernd Gärtner and Emo Welzl. 1994. Vapnik-Chervonenkis dimension and (pseudo-) hyperplane arrangements. *Discrete & Computational Geometry* 12, 4 (1994), 399–432. https://doi.org/10.1007/BF02574389

[26] Dana Glasner and Rocco A Servedio. 2009. Distribution-free testing lower bound for basic boolean functions. *Theory of Computing* 5, 1 (2009), 191–216. https://doi.org/10.4086/toc.2009.v005a010

[27] Oded Goldreich, Shari Goldwasser, and Dana Ron. 1998. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)* 45, 4 (1998), 653–750. https://doi.org/10.1145/285055.285060

[28] Oded Goldreich and Dana Ron. 2016. On sample-based testers. *ACM Transactions on Computation Theory* 8, 2 (2016), 1–54. https://doi.org/10.1145/2898355

[29] Sariel Har-Peled. 2014. Determining the number of clusters using property testing algorithm. Theoretical Computer Science Stack Exchange. https://cstheory.stackexchange.com/q/25655

[30] Nathaniel Harms. 2019. Testing halfspaces over rotation-invariant distributions. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 694–713. https://doi.org/10.1137/1.9781611975482.44

[31] Lisa Hellerstein and Rocco A. Servedio. 2007. On PAC learning algorithms for rich Boolean function classes. *Theor. Comput. Sci.* 384, 1 (2007), 66–76. https://doi.org/10.1016/j.tcs.2007.05.018

[32] Hunter R Johnson. 2014. Some new maximum VC classes. *Inform. Process. Lett.* 114, 6 (2014), 294–298. https://doi.org/10.1016/j.ipl.2014.01.006

[33] Michael Kearns and Dana Ron. 2000. Testing problems with sublearning sample complexity. *Journal of Computer and System Science* 61, 3 (2000), 428–456. https://doi.org/10.1006/jcss.1999.1656

[34] Adam R Klivans and Rocco A Servedio. 2004. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. *J. Comput. System Sci.* 68, 2 (2004), 303–318. https://doi.org/10.1016/j.jcss.2003.07.007

[35] Dima Kuzmin and Manfred K Warmuth. 2007. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research* 8, Sep (2007), 2047–2081. http://jmlr.org/papers/v8/kuzmin07a.html

[36] Roi Livni and Yishay Mansour. 2019. Graph-based Discriminators: Sample Complexity and Expressiveness. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019)*. 6696–6705. http://papers.nips.cc/paper/8895-graph-based-discriminators-sample-complexity-and-expressiveness

[37] Yishay Mansour. 1997. Pessimistic decision tree pruning based on tree size. In *Machine Learning-International Workshop then Conference*. Citeseer, 195–201. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.3146

[38] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco A Servedio. 2010. Testing halfspaces. *SIAM J. Comput.* 39, 5 (2010), 2004–2047. https://doi.org/10.1137/070707890

[39] Shay Moran. 2012. *Shattering Extremal Systems*. Ph.D. Dissertation. Universität des Saarlandes Saarbrücken. https://arxiv.org/abs/1211.2980

[40] Shay Moran and Manfred K Warmuth. 2016. Labeled compression schemes for extremal classes. In *International Conference on Algorithmic Learning Theory*. Springer, 34–49. https://doi.org/10.1007/978-3-319-46379-7_3

[41] Joe Neeman. 2014. Testing surface area with arbitrary accuracy. In *Symposium on Theory of Computing, STOC 2014*. ACM, 393–397. https://doi.org/10.1145/2591796.2591807

[42] Ryan O'Donnell and Rocco A Servedio. 2010. New degree bounds for polynomial threshold functions. *Combinatorica* 30, 3 (2010), 327–358. https://doi.org/10.1007/s00493-010-2173-3

[43] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. 2009. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM J. Comput.* 39, 3 (2009), 813–842. https://doi.org/10.1137/070701649

[44] Dana Ron and Asaf Rosin. 2020. Almost Optimal Distribution-Free Sample-Based Testing of k-Modality. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2020)*, Jarosław Byrka and Raghu Meka (Eds.), Vol. 176. 27:1–27:19. https://doi.org/10.4230/LIPIcs.APPROX/RANDOM.2020.27

[45] Mert Saglam. 2018. Near Log-Convexity of Measured Heat in (Discrete) Time and Consequences. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018*. 967–978. https://doi.org/10.1109/FOCS.2018.00095

[46] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[47] Sandeep Silwal. 2020. Personal communication.

[48] Madhu Sudan. 2010. Invariance in property testing. In *Property testing*. Springer, 211–227. https://doi.org/10.1007/978-3-642-16367-8_12

[49] Li-Yang Tan. 2020. Personal communication.

[50] Gregory Valiant and Paul Valiant. 2011. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the forty-third annual ACM Symposium on Theory of Computing*. 685–694. https://doi.org/10.1145/1993636.1993727

[51] Gregory Valiant and Paul Valiant. 2011. The power of linear estimators. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*. IEEE, 403–412. https://doi.org/10.1109/FOCS.2011.81

[52] Leslie G. Valiant. 1984. A Theory of the Learnable. *Commun. ACM* 27, 11 (1984), 1134–1142. https://doi.org/10.1145/1968.1972

[53] James G. Wendel. 1962. A Problem in Geometric Probability. *Math. Scand.* 11 (1962), 109–112. https://eudml.org/doc/165817

[54] Roberta S Wenocur and Richard M Dudley. 1981. Some special Vapnik-Chervonenkis classes. *Discrete Mathematics* 33, 3 (1981), 313–318. https://doi.org/10.1016/0012-365X(81)90274-0

[55] Yihong Wu and Pengkun Yang. 2019. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics* 47, 2 (2019), 857–883. https://doi.org/10.1214/17-AOS1665