Active Property Testing

Maria-Florina Balcan
School of Computer Science
Georgia Institute of Technology
Atlanta, GA, USA.
Email: ninamf@cc.gatech.edu

Eric Blais Avrim Blum Liu Yang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA.
Email: [eblais, avrim, liuy]@cs.cmu.edu

***Abstract*—One motivation for property testing of boolean functions is the idea that testing can provide a fast preprocessing step before learning. However, in most machine learning applications, it is not possible to request for labels of arbitrary examples constructed by an algorithm. Instead, the dominant query paradigm in applied machine learning, called *active learning*, is one where the algorithm may query for labels, but *only on points in a given (polynomial-sized) unlabeled sample*, drawn from some underlying distribution D. In this work, we bring this well-studied model to the domain of *testing*.**

We develop both general results for this *active testing* model as well as efficient testing algorithms for several important properties for learning, demonstrating that testing can still yield substantial benefits in this restricted setting. For example, we show that testing unions of d intervals can be done with $O(1)$ label requests in our setting, whereas it is known to require $\Omega(d)$ labeled examples for learning (and $\Omega(\sqrt{d})$ for passive testing [22] where the algorithm must pay for *every* example drawn from D). In fact, our results for testing unions of intervals also yield improvements on prior work in both the classic query model (where any point in the domain can be queried) and the passive testing model as well. For the problem of testing linear separators in R^n over the Gaussian distribution, we show that both active and passive testing can be done with $O(\sqrt{n})$ queries, substantially less than the $\Omega(n)$ needed for learning, with near-matching lower bounds. We also present a general combination result in this model for building testable properties out of others, which we then use to provide testers for a number of assumptions used in semi-supervised learning.

In addition to the above results, we also develop a general notion of the *testing dimension* of a given property with respect to a given distribution, that we show characterizes (up to constant factors) the intrinsic number of label requests needed to test that property. We develop such notions for both the active and passive testing models. We then use these dimensions to prove a number of lower bounds, including for linear separators and the class of dictator functions.

***Keywords*-Property testing, Active learning, Boolean functions, Linear threshold functions, Unions of intervals**

I. Introduction

Property testing and machine learning have many natural connections. In property testing, given black-box access to an unknown boolean function f, one would like with few queries to distinguish the case that f has some given property $\mathcal{P}$ (belongs to the class of functions $\mathcal{P}$) from the case that f is far from any function having that property. In machine learning one would like to find a good approximation g of f, typically under the assumption that f belongs to a given class $\mathcal{P}$. This connection is in fact a natural motivation for property testing: to cheaply determine whether learning with a given hypothesis class is worthwhile [19], [26]. If the labeling of examples is expensive, or if a learning algorithm is computationally expensive to run, or if one is deciding from what source to purchase one's data, performing a cheap test in advance could be a substantial savings. Indeed, query-efficient testers have been designed for many common function classes considered in machine learning including linear threshold functions [25], juntas [18], [7], DNF formulas [16], and decision trees [16]. (See Ron's survey [26] for much more on the connection between learning and property testing.)

However, there is a disconnect between the most commonly used property-testing and machine learning models. Most property-testing algorithms rely on the ability to query functions on arbitrary points of their choosing. On the other hand, most machine learning problems unfortunately do not allow one to perform queries on fictitious examples constructed by an algorithm. Consider, for instance, a scenario such as machine learning for medical diagnosis. Given a large database of patients with each patient described by various features (height, age, family history, smoker or not, etc.), one would like to learn a function that predicts from these features whether or not a patient has a given medical condition (diabetes, for example). To perform this learning task, the researchers can run a (typically expensive) medical test on any of the patients to determine if the patient has the medical condition. However, researchers *cannot* ask whether the patient would still have the disease were the values of some of his features changed! Moreover, researchers cannot make up a feature vector out of whole cloth and ask if that feature vector has the disease. As another example, in classifying documents by topic, selecting an existing document on the web and asking a labeler "Is this about sports or business?" may be perfectly reasonable. However, the typical representation of a document in a machine learning sytem is as a vector of word-counts in R^n (a "bag of words", without any information about the order in which they appear in the document). Thus, modifying

some existing vector, or creating a new one from scratch, would not produce an object that we could expect a human labeler to easily classify. The key issue is that for most problems in machine learning, the example and the label are in fact *both* functions of some underlying more complex object. Even for problems where this issue does not occur, such as classifying handwritten digits into the numerals they represent, queries can be problematic because the space of reasonable pixel images is a very sparse subset of the entire domain. Indeed, now-classic experiments on membership-query learning for digit recognition ran into exactly this problem, leading to poor results [5]. In this case, the problem is that the distribution one cares about (the distribution of natural handwritten digits) is not one that the algorithm can easily construct new examples from.

As a result of these issues, the dominant query paradigm in machine learning in recent years is not one where the algorithm can make arbitrary queries, but instead is a weaker model known as *active learning* [28], [12], [29], [13], [2], [6], [10], [20], [15], [23]. In active learning, there is an underlying distribution D over unlabeled examples (say the distribution of documents on the web, represented as vectors over word-counts) that we assume can be sampled from cheaply: we assume the algorithm may obtain a polynomial number of samples from D. Then, the algorithm may ask an oracle for labels (these oracle calls are viewed as expensive), *but only on points in its sample*. The goal of the active learning algorithm is to produce an accurate hypothesis while requesting as few labels as possible, ideally substantially fewer than in passive learning where *every* example drawn from D is labeled by the oracle.

In this work, we bridge this gap between testing and learning by developing and analyzing a model of testing that parallels active learning, which we call *active testing*. As in active learning, we assume that our algorithm is given a polynomial number of unlabeled examples from the underlying distribution D and can then make label queries, but *only* over the points in its sample. From a small number of such queries, the algorithm must then answer whether the function has the given property, or is far, with respect to D, from any function having that property (see Section II for formal definitions). We show that even with this restriction, we can still efficiently test important properties for machine learning including unions of intervals, linear separators, and a number of properties considered in semi-supervised learning. Moreover, these testers reveal important structural characteristics of these classes. We additionally develop a notion of *testing dimension* that characterizes the number of examples needed to test a given property over a given distribution, much like notions of dimension in machine learning. We do this for both the active testing model and the weaker *passive testing* model [19], [22] in which the algorithm must query every example it draws from D. In fact, as part of our analysis, we also develop improved algorithms for several important classes for the passive testing model as well. Overall, our results demonstrate that active testing exhibits a rich structure and strengthens the connection between testing and learning.

A. Our Results and Structure of this Paper

After formally defining the active testing model in Section II, we begin in Section III by presenting an active tester for the class of *unions of d intervals* on the line. This tester requires only a constant number of label requests (independent of d) and applies for any (even unknown) distribution D. Additionally, when D is the uniform distribution, our tester requires only $O(\sqrt{d})$ unlabeled samples, immediately yielding an $O(\sqrt{d})$-sample passive tester in this case. The best prior result for this class applied only to the relaxed problem of distinguishing a union of d intervals from functions ϵ-far from a union of d/ϵ intervals, over the uniform distribution (using $O(1)$ queries in the arbitrary-query model and $O(\sqrt{d})$ samples in the passive-testing model) [22].

In Section IV we consider the problem of testing *linear threshold functions* in R^n. For this problem we show that both active and passive testing can be done with $O(\sqrt{n})$ labeled examples over the Gaussian distribution, substantially less than the $\Omega(n)$ labeled examples needed for learning (even over the Gaussian distribution [24]). The key challenge here is that estimating a statistic due to Matulef et al. [25]—which can be done with $O(1)$ queries if arbitrary queries are allowed [25]—would require $\Theta(n)$ samples if done from independent pairs of random examples in the natural way. We overcome this obstacle by re-using non-independent pairs of examples in the estimation. At a technical level, this result uses the fact that even though typical values of $(x \cdot y)^2$ are fairly large, for any boolean function f it will be the case that for "most" values y, the quantity $(\mathbb{E}_x[f(x)x \cdot y])^2$ is quite small—which can be shown via a Fourier decomposition of f. This in turn allows one to show strong concentration via a theorem of Arcones [1] on the concentration of U-statistics.

In Section V we show that any *disjoint union* of a polynomial number of testable properties remains testable in the active testing model, allowing one to build testable properties out of simpler components; this can then be used to provide label-efficient testers for several properties used in semi-supervised learning.

Finally, in Section VI we consider notions of *intrinsic dimension* for testing. Such notions of dimension (e.g., VC dimension [30], SQ dimension [8], Rademacher complexity [4]) have been particularly effective in determining sample complexity for learning. Y. Mansour and G. Kalai (personal communication, see also [21]) posed the question of whether comparable notions of dimension might exist for testing. In this work we answer in the affirmative, for both active and passive testing models, and use these notions of dimension to obtain a number of lower bounds. Notably, we show that

$\Omega(\log n)$ queries are needed to distinguish dictator functions from random functions in both models[1] as well as lower bounds of $\tilde{\Omega}(n^{1/3})$ and $\tilde{\Omega}(\sqrt{n})$ for active and passive testing (respectively) of linear threshold functions.

Due to space limitations, a number of proofs have been omitted from this extended abstract. See the full version of this paper [3] for omitted proofs and a detailed discussion of the relation of active testing to other testing models.

II. The Active Property Testing Model

A *property* $\mathcal{P}$ of boolean functions is simply a subset of all boolean functions. We will also refer to properties as *classes* of functions. The *distance* of a function f to the property $\mathcal{P}$ with respect to a distribution D over the domain of the function is $dist_D(f, \mathcal{P}) := \min_{g \in \mathcal{P}} \Pr_{x \sim D}[f(x) \neq g(x)]$. A *tester* for $\mathcal{P}$ is a randomized algorithm that must distinguish (with high probability) between functions in $\mathcal{P}$ and functions that are far from $\mathcal{P}$. In the standard property testing model introduced by Rubinfeld and Sudan [27], a tester is allowed to query the value of the function on any input in order to make this decision. We consider instead a model in which we add restrictions to the possible queries:

Definition II.1 (Property tester)**.** An *s-sample, q-query ϵ-tester* for $\mathcal{P}$ over the distribution D is a randomized algorithm A that draws a sample S of size s from D, queries for the value of f on q points of S, and then

1) Accepts w.p. at least $\frac{2}{3}$ when $f \in \mathcal{P}$, and
2) Rejects w.p. at least $\frac{2}{3}$ when $dist_D(f, \mathcal{P}) \geq \epsilon$.

We will use the terms "label request" and "query" interchangeably. Definition II.1 coincides with the standard definition of property testing when the number of samples is unlimited and the distribution's support covers the entire domain. In the other extreme case where we fix $q = s$, our definition then corresponds to the *passive testing* model of Goldreich, Goldwasser, and Ron [19], where the inputs queried by the tester are sampled from the distribution. Finally, by setting s to be polynomial in an appropriate measure of the input domain or property $\mathcal{P}$, we obtain the *active testing* model that is the focus of this paper:

Definition II.2 (Active tester)**.** A randomized algorithm is a *q-query active ϵ-tester* for $\mathcal{P} \subseteq \{0,1\}^n \to \{0,1\}$ over D if it is a $\mathrm{poly}(n)$-sample, q-query ϵ-tester for $\mathcal{P}$ over D.[2]

In some cases, the domain of our functions is not $\{0,1\}^n$. In those cases, we require s to be polynomial in some other appropriate measure of complexity of the domain or property $\mathcal{P}$ that we specify explicitly. Note that in Definition II.1, since we do not have direct membership query access (at arbitrary points), our tester must accept w.p. at least $\frac{2}{3}$ when f is such that $dist_D(f, \mathcal{P}) = 0$, even if f does not satisfy $\mathcal{P}$ over the entire input space. See the full version of this paper [3] for a comparison of active testing to other testing models.

[1] Building on this analysis, Noga Alon (personal communication) has recently developed a stronger $\Omega(k \log n)$ lower bound for the active testing dimension of juntas via use of the Kim-Vu polynomial method.

[2] We emphasize that the name *active tester* is chosen to reflect the connection with active learning. It is *not* meant to imply that this model of testing is somehow "more active" than the standard property testing model.

III. Testing Unions of Intervals

The function $f : [0,1] \to \{0,1\}$ is a *union of d intervals* if there are at most d non-overlapping intervals $[\ell_1, u_1], \ldots, [\ell_d, u_d]$ such that $f(x) = 1$ iff $\ell_i \leq x \leq u_i$ for some $i \in [d]$. The VC dimension of this class is $2d$, so learning a union of d intervals requires $\Omega(d)$ queries. Kearns and Ron [22] showed that under the uniform distribution, the relaxed problem of distinguishing unions of d intervals from functions that are ϵ-far from unions of d/ϵ intervals can be done with a constant number of queries in the standard arbitrary-query testing model, and with $O(\sqrt{d})$ samples in the passive testing model (suppressing dependence on ϵ).

Here, we show that the *non-relaxed* problem of distinguishing unions of d intervals from functions ϵ-far from unions of d intervals can be done with a constant number of label requests in the active testing model, for *any* (even unknown) distribution D. Specifically, we prove that we can test unions of d intervals in the active testing model using only $O(1/\epsilon^4)$ label requests from a set of $poly(d, 1/\epsilon)$ unlabeled examples. Furthermore, over the uniform distribution, we need a total of only $O(\sqrt{d}/\epsilon^5)$ unlabeled examples, which immediately implies an $O(\sqrt{d}/\epsilon^5)$ sample bound for passive testing. Note that Kearns and Ron [22] show that $\Omega(\sqrt{d})$ examples are required to test unions of intervals over the uniform distribution in the passive testing model, so this latter result, as a function of d, is tight.

Theorem III.1. *For any (known or unknown) distribution D, testing unions of d intervals in the active testing model can be done using only $O(1/\epsilon^4)$ queries. In the case of the uniform distribution, we further need only $O(\sqrt{d}/\epsilon^5)$ unlabeled examples.*

We prove Theorem III.1 by beginning with the case that the underlying distribution is uniform over $[0,1]$, and afterwards show how to generalize to arbitrary distributions. Our tester is based on showing that unions of intervals have a *noise sensitivity* characterization.

Definition III.2. Fix $\delta > 0$. The *local δ-noise sensitivity* of the function $f : [0,1] \to \{0,1\}$ at $x \in [0,1]$ is $\mathrm{NS}_\delta(f, x) = \Pr_{y \sim_\delta x}[f(x) \neq f(y)]$, where $y \sim_\delta x$ represents a draw of y uniform in $(x - \delta, x + \delta) \cap [0,1]$. The *noise sensitivity* of f is

$$\mathbb{NS}_\delta(f) = \Pr_{x, y \sim_\delta x}[f(x) \neq f(y)]$$

or, equivalently, $\mathbb{NS}_\delta(f) = \mathbb{E}_x \mathrm{NS}_\delta(f, x)$.

A simple argument shows that unions of d intervals have (relatively) low noise sensitivity:

Proposition III.3. *Fix $\delta > 0$ and let $f : [0,1] \to \{0,1\}$ be a union of d intervals. Then $\mathbb{NS}_\delta(f) \leq d\delta$.*

Proof sketch: Draw $x \in [0,1]$ uniformly at random and $y \sim_\delta x$. The inequality $f(x) \neq f(y)$ can only hold when a boundary $b \in [0,1]$ of one of the d intervals in f lies in between x and y. For any point $b \in [0,1]$, the probability that $x < b < y$ or $y < b < x$ is at most $\frac{\delta}{2}$, and there are at most $2d$ boundaries of intervals in f, so the proposition follows from the union bound. ∎

The key to the tester is showing that the converse of the above statement is approximately true as well: for δ small enough, every function that has noise sensitivity not much larger than $d\delta$ is close to being a union of d intervals. (Full proof in the full version [3]).

Lemma III.4. *Fix $\delta = \frac{\epsilon^2}{32d}$. Let $f : [0,1] \to \{0,1\}$ be a function with noise sensitivity bounded by $\mathbb{NS}_\delta(f) \leq d\delta(1 + \frac{\epsilon}{4})$. Then f is ϵ-close to a union of d intervals.*

Proof outline: The proof proceeds in two steps. First, we show that so long as f has low noise-sensitivity, it can be "locally self-corrected" to a function $g : [0,1] \to \{0,1\}$ that is $\frac{\epsilon}{2}$-close to f and is a union of at most $d(1 + \frac{\epsilon}{4})$ intervals. We then show that g – and every other function that is a union of at most $d(1 + \frac{\epsilon}{4})$ intervals – is $\frac{\epsilon}{2}$-close to a union of d intervals.

To construct the function g, we consider a smoothed function $f_\delta : [0,1] \to [0,1]$ obtained by taking the convolution of f and a uniform kernel of width 2δ. We define τ to be some appropriately small parameter. When $f_\delta(x) \leq \tau$, then this means that nearly all the points in the δ-neighborhood of x have the value 0 in f, so we set $g(x) = 0$. Similarly, when $f_\delta(x) \geq 1 - \tau$, then we set $g(x) = 1$. (This procedure removes any "local noise" that might be present in f.) This leaves all the points x where $\tau < f_\delta(x) < 1 - \tau$. Let us call these points *undefined*. For each such point x we take the largest value $y \leq x$ that is defined and set $g(x) = g(y)$. The key technical part of the proof involves showing that the construction described above yields a function g that is $\frac{\epsilon}{2}$-close to f and that is a union of $d(1 + \frac{\epsilon}{4})$ intervals. Due to space constraints, we defer the argument to the full version of this paper [3]. ∎

The noise sensitivity characterization of unions of intervals obtained by Proposition III.3 and Lemma III.4 suggest a natural approach for building a tester: design an algorithm that estimates the noise sensitivity of the input function and accepts iff this noise sensitivity is small enough. This is indeed what we do:

UNION OF INTERVALS TESTER(f, d, ϵ)
Parameters: $\delta = \frac{\epsilon^2}{32d}, r = O(\epsilon^{-4})$.
1) For rounds $i = 1, \ldots, r$,
   1.1 Draw $x \in [0,1]$ uniformly at random.
   1.2 Draw samples until we obtain $y \in (x - \delta, x + \delta)$.
   1.3 Set $Z_i = \mathbf{1}[f(x) \neq f(y)]$.
2) **Accept** iff $\frac{1}{r}\sum Z_i \leq d\delta(1 + \frac{\epsilon}{8})$.

The algorithm makes $2r = O(\epsilon^{-4})$ queries to the function. Since a draw in Step 1.2 is in the desired range with probability 2δ, the number of samples drawn by the algorithm is a random variable with very tight concentration around $r(1 + \frac{1}{2\delta}) = O(d/\epsilon^6)$. The draw in Step 1.2 also corresponds to choosing $y \sim_\delta x$. As a result, the probability that $f(x) \neq f(y)$ in a given round is exactly $\mathbb{NS}_\delta(f)$, and the average $\frac{1}{r}\sum Z_i$ is an unbiased estimate of the noise sensitivity of f. By Proposition III.3, Lemma III.4, and Chernoff bounds, the algorithm therefore errs with probability less than $\frac{1}{3}$ provided that $r > c \cdot 1/(d\delta\epsilon^2) = c \cdot 32/\epsilon^4$ for some suitably large constant c.

Improved unlabeled sample complexity: Notice that by changing Steps 1.1-1.2 slightly to pick the first pair (x, y) such that $|x - y| < \delta$, we immediately improve the unlabeled sample complexity to $O(\sqrt{d}/\epsilon^5)$ without affecting the analysis. In particular, this procedure is equivalent to picking $x \in [0,1]$ then $y \sim_\delta x$.[3] As a result, up to $poly(1/\epsilon)$ terms, we also strengthen the *passive testing* bounds of Kearns and Ron [22] which only distinguish the case that f is a union of d intervals from the case that f is ϵ-far from the class of unions of d/ϵ intervals. (Their results use $O(\sqrt{d}/\epsilon^{1.5})$ examples.) Kearns and Ron [22] show that $\Omega(\sqrt{d})$ examples are necessary for passive testing, so in terms of d this is optimal.

Active testing over arbitrary distributions: We now consider the case that examples are drawn from some arbitrary distribution D. First, let us consider the easier case that D is known. In that case, we can reduce the problem of testing over general distributions to that of testing over the uniform distribution on $[0,1]$ by using the CDF of D. In particular, given point x, define $p_x = \Pr_{y \sim D}[y \leq x]$. So, for x drawn from D, p_x is uniform in $[0,1]$.[4] As a result we can just replace Step 1.2 in the tester with sampling until we obtain y such that $p_y \in (p_x - \delta, p_x + \delta)$. Now, suppose D is not known. In that case, we do not know the p_x and p_y values exactly. However, we can use the fact that the VC-dimension of the class of initial intervals on the line equals 1 to uniformly estimate all such values from a polynomial-sized unlabeled sample. In particular, $O(1/\gamma^2)$ unlabeled examples are sufficient so that with high probability, *every* point x has property that the estimate $\hat{p}_x$ of p_x computed with respect to the sample (the fraction of

[3] Except for events of $O(\delta)$ probability mass at the boundary.

[4] We are assuming here that D is continuous and has a pdf. If D has point masses, then instead define $p_x^L = \Pr_y[y < x]$ and $p_x^U = \Pr_y[y \leq x]$ and select p_x uniformly in $[p_x^L, p_x^U]$.

points in the *sample* that are $\leq x$) will be within γ of the correct p_x value [9]. If we define $\hat{\mathbb{NS}}_\delta(f)$ to be the noise-sensitivity of f computed using these estimates, then we get $\frac{\delta-\gamma}{\delta+\gamma}\mathbb{NS}_{\delta-\gamma}(f) \leq \hat{\mathbb{NS}}_\delta(f) \leq \frac{\delta+\gamma}{\delta-\gamma}\mathbb{NS}_{\delta+\gamma}(f)$. This implies that $\gamma = O(\epsilon\delta)$ is sufficient so that the noise-sensitivity estimates are sufficiently accurate for the procedure to work as before.

Putting these results together, we have Theorem III.1.

IV. Testing Linear Threshold Functions

A boolean function $f : \mathbb{R}^n \to \{0,1\}$ is a *linear threshold function* (LTF) if there exist $n+1$ real-valued parameters $w_1, \ldots, w_n, \theta$ such that for each $x \in \mathbb{R}^n$, we have $f(x) = \mathrm{sgn}(w_1 x_1 + \cdots w_n x_n - \theta)$. The main result of this section is that under the Gaussian distribution, it is possible to efficiently test whether a function is a linear threshold function in the active and passive testing models with substantially fewer labeled examples than needed for learning, along with near-matching lower bounds.

Theorem IV.1. *We can efficiently test linear threshold functions under the Gaussian distribution with $O(\sqrt{n \log n})$ labeled examples in both active and passive testing models.*

Theorem IV.2. *No (even computationally inefficient) algorithm can test linear threshold functions under the Gaussian distribution with $\tilde{o}(n^{1/3})$ labeled examples for active testing or $\tilde{o}(\sqrt{n})$ labeled examples for passive testing.*

Note that the class of linear threshold functions requires $\Omega(n)$ labeled examples for *learning*, even over the Gaussian distribution [24]. Linear threshold functions can be tested with a constant number of queries in the standard (arbitrary query) property testing model [25].

The starting point for the upper bound in Theorem IV.1 is a characterization lemma of linear threshold functions in terms of a certain self-correlation statistic. To be precise, in Lemma IV.4 we are scaling so that each coordinate is drawn independently from $\mathcal{N}(0,1)$—so a typical example will have length $\Theta(\sqrt{n})$.

Definition IV.3. The *self-correlation coefficient* of the function $f : \mathbb{R}^n \to \mathbb{R}$ is $\rho(f) := \mathbb{E}_{x,y}[f(x)f(y)\langle x, y\rangle]$.

Lemma IV.4 (Matulef et al. [25])**.** *There is an explicit continuous function $W : \mathbb{R} \to \mathbb{R}$ with bounded derivative $\|W'\|_\infty \leq 1$ and peak value $W(0) = \frac{2}{\pi}$ such that every linear threshold function $f : \mathbb{R}^n \to \{-1,1\}$ satisfies $\rho(f) = W(\mathbb{E}_x f)$. Moreover, every function $g : \mathbb{R}^n \to \{-1,1\}$ that satisfies $|\rho(g) - W(\mathbb{E}_x g)| \leq 4\epsilon^3$, is ϵ-close to being a linear threshold function.*

Lemma IV.4 suggests a natural approach to testing for linear threshold functions from random examples: simply estimate the self-correlation coefficient of Definition IV.3 by repeatedly drawing pairs of labeled examples (x_i, y_i) from the Gaussian distribution in R^n and computing the empirical average of the quantities $f(x_i)f(y_i)\langle x_i, y_i\rangle$ observed. The problem with this approach, however, is that the dot-product $\langle x_i, y_i\rangle$ will typically have magnitude $\Theta(\sqrt{n})$ (one can view it as essentially the result of an n step random walk). Therefore to estimate the self-correlation coefficient to accuracy $O(1)$ via independent random samples in this way would require $\Omega(n)$ labeled examples. This is of course not very useful, since it is the same as the number of labeled examples needed to *learn* an LTF.

We will be able to achieve an improved bound, however, using the following idea: rather than averaging over independent pairs (x, y), we will draw a smaller sample and average over all (non-independent) pairs within the sample. That is, we request q random labeled examples $x_1, \ldots, x_q$, and now estimate $\rho(f)$ by computing $\binom{q}{2}^{-1}\sum_{i<j} f(x_i)f(x_j)\langle x_i, x_j\rangle$. Of course, the terms in the summation are no longer independent. However, they satisfy the property that even though the quantity $f(x)f(y)\langle x, y\rangle$ is typically large, for most values y, the quantity $\mathbb{E}_x[f(x)f(y)\langle x, y\rangle]$ is small. (This can be shown via a Fourier decomposition of the function f.) This, together with additional truncation of the quantity in question, will allow us to apply a Bernstein-type inequality for U-statistics due to Arcones [1] in order to achieve the desired concentration.

The resulting LTF Tester is given in Figure 1. This algorithm has two advantages. First, it is a valid tester in both the active and passive property testing models since the q inputs queried by the algorithm are all drawn independently at random from the standard n-dimensional Gaussian distribution. Second, the algorithm itself is very simple. As in many cases with property testing, however, the analysis of this algorithm is more challenging.

LTF Tester(f, ϵ)
Parameters: $\tau = \sqrt{4n\log(4n/\epsilon^3)}$, $m = 800\tau/\epsilon^3 + 32/\epsilon^6$.

1) Draw $x^1, x^2, \ldots, x^m$ independently at random from $\mathbb{R}^n$.
2) Query $f(x^1), f(x^2), \ldots, f(x^m)$.
3) Set $\tilde{\mu} = \frac{1}{m}\sum_{i=1}^m f(x^i)$.
4) Set $\tilde{\rho} = \binom{m}{2}^{-1}\sum_{i\neq j} f(x^i)f(x^j)\langle x^i, x^j\rangle \cdot \mathbf{1}[|\langle x^i, x^j\rangle| \leq \tau]$.
5) **Accept** iff $|\tilde{\rho} - W(\tilde{\mu})| \leq 2\epsilon^3$.

Figure 1. LTF Tester

Given Lemma IV.4, as noted above, the key challenge in the proof of correctness of the LTF Tester is controlling the error of the estimate $\tilde{\rho}$ of $\rho(f)$ in Step 4, which we do with concentration of measure results for U-statistics. The *U-statistic* (of order 2) with symmetric kernel function

$g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is

$$U_g^q(x^1, \ldots, x^q) := \binom{q}{2}^{-1} \sum_{1 \le i < j \le q} g(x^i, x^j).$$

U-statistics are unbiased estimators of the expectation of their kernel function and, even more importantly, when the kernel function is "well-behaved", the tails of their distributions satisfy strong concentration. In our case, the thresholded kernel function

$$g(x,y) = \begin{cases} f(x)f(y)\langle x, y\rangle & |\langle x, y\rangle| \le \tau \\ 0 & \text{otherwise} \end{cases}$$

allows us to apply Arcones' theorem.

Lemma IV.5 (Arcones [1])**.** *For a symmetric function $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, let $\Sigma^2 = \mathbb{E}_x[\mathbb{E}_y[h(x,y)]^2] - \mathbb{E}_{x,y}[h(x,y)]^2$, let $b = \|h - \mathbb{E}h\|_\infty$, and let $U_q(h)$ be a random variable obtained by drawing $x^1, \ldots, x^q$ independently at random and setting $U_q(h) = \binom{q}{2}^{-1} \sum_{i<j} h(x^i, x^j)$. Then for every $t > 0$,*

$$\Pr[|U_q(h) - \mathbb{E}h| > t] \le 4 \exp\left(\frac{qt^2}{8\Sigma^2 + 100bt}\right).$$

One final task before proving Theorem IV.1 is to relate the self-correlation coefficient $\rho(f)$ with the function $g(x,y)$ (to which we will apply Arcones' theorem). This relation is given in Lemma IV.6 below, whose proof is in the full version of this paper [3].

Lemma IV.6. *Let g be defined as above with $\tau = \sqrt{4n\log(4n/\epsilon^3)}$. Then $|\mathbb{E}g - \rho| \le \frac{1}{2}\epsilon^3$.*

Using the above, we can now give the desired guarantee for LTF TESTER.

Proof of Theorem IV.1: Consider the LTF-TESTER algorithm. When the estimates $\tilde{\mu}$ and $\tilde{\rho}$ satisfy

$$|\tilde{\mu} - \mathbb{E}f| \le \epsilon^3 \quad \text{and} \quad |\tilde{\rho} - \mathbb{E}[f(x)f(y)\langle x, y\rangle]| \le \epsilon^3,$$

Lemma IV.4 guarantees that the algorithm correctly distinguishes LTFs from functions that are far from LTFs. To complete the proof, we must therefore show that the estimates are within the specified error bounds with probability at least $2/3$.

The values $f(x^1), \ldots, f(x^m)$ are independent $\{-1, 1\}$-valued random variables. By Hoeffding's inequality,

$$\Pr[|\tilde{\mu} - \mathbb{E}f| \le \epsilon^3] \ge 1 - 2e^{-\epsilon^6 m/2} = 1 - 2e^{-O(\sqrt{n})}.$$

The estimate $\tilde{\rho}$ is a U-statistic with kernel g^* as defined above. This kernel satisfies

$$\|g^* - \mathbb{E}g^*\|_\infty \le 2\|g^*\|_\infty = 2\sqrt{4n\log(4n/\epsilon^3)}$$

and

$$\begin{aligned}\Sigma^2 &\le \mathbb{E}_y\big[\mathbb{E}_x[g^*(x,y)]^2\big] \\ &= \mathbb{E}_y\big[\mathbb{E}_x[f(x)f(y)\langle x, y\rangle \mathbf{1}[|\langle x, y\rangle| \le \tau]]^2\big].\end{aligned}$$

For any two functions $\phi, \psi : \mathbb{R}^n \to \mathbb{R}$, when ψ is $\{0,1\}$-valued the Cauchy-Schwarz inequality implies that $\mathbb{E}_x[\phi(x)\psi(x)]^2 \le \mathbb{E}_x[\phi(x)]\mathbb{E}_x[\phi(x)\psi(x)^2] = \mathbb{E}_x[\phi(x)]\mathbb{E}_x[\phi(x)\psi(x)]$ and so $\mathbb{E}_x[\phi(x)\psi(x)] \le \mathbb{E}_x[\phi(x)]$. Applying this inequality to the expression for Σ^2 gives

$$\begin{aligned}\Sigma^2 &\le \mathbb{E}_y\big[\mathbb{E}_x[f(x)f(y)\langle x, y\rangle]^2\big] \\ &= \mathbb{E}_y\big[\big(\sum_{i=1}^n f(y) y_i \mathbb{E}_x[f(x)x_i]\big)^2\big] \\ &= \sum_{i,j} \hat{f}(e_i)\hat{f}(e_j)\mathbb{E}_y[y_i y_j] = \sum_{i=1}^n \hat{f}(e_i)^2.\end{aligned}$$

By Parseval's identity, we have $\sum_i \hat{f}(e_i)^2 \le \|\hat{f}\|_2^2 = \|f\|_2^2 = 1$. Lemmas IV.6 and IV.5 imply that

$$\begin{aligned}\Pr[|\tilde{\rho} - \mathbb{E}g| \le \epsilon^3] &= \Pr[|\tilde{\rho} - \mathbb{E}g^*| \le \tfrac{1}{2}\epsilon^3] \\ &\ge 1 - 4e^{-\frac{mt^2}{8 + 200\sqrt{n\log(4n/\epsilon^3)}t}} \ge \tfrac{11}{12}.\end{aligned}$$

The union bound completes the proof of correctness. ∎

It is natural to ask whether we can further improve the query complexity of the tester for linear threshold functions by using U-statistics of higher order. The lower bound in Theorem IV.2 shows that this—or any other possible active or passive testing approach—cannot yield a query complexity sub-polynomial in n. We defer the discussion of this lower bound to Section VI, where we will use the notion of testing dimension to establish the bound.

V. TESTING DISJOINT UNIONS OF TESTABLE PROPERTIES

We now show that active testing has the feature that a disjoint union of a polynomial number of testable properties is testable, with a number of queries that is independent of the size of the union; this feature does not hold for passive testing. In addition to providing insight into the distinction between the two models, this fact will be useful in our analysis of semi-supervised learning-based properties mentioned below and discussed more fully in [3].

Specifically, given properties $\mathcal{P}_1, \ldots, \mathcal{P}_N$ over domains $X_1, \ldots, X_N$, define their disjoint union $\mathcal{P}$ over domain $X = \{(i, x) : i \in [N], x \in X_i\}$ to be the set of functions f such that $f(i,x) = f_i(x)$ for some $f_i \in \mathcal{P}_i$. In addition, for any distribution D over X, define D_i to be the conditional distribution over X_i when the first component is i. If each $\mathcal{P}_i$ is testable over D_i then $\mathcal{P}$ is testable over D with only small overhead in the number of queries:

Theorem V.1. *Given properties $\mathcal{P}_1, \ldots, \mathcal{P}_N$, if each $\mathcal{P}_i$ is testable over D_i with $q(\epsilon)$ queries and $U(\epsilon)$ unlabeled samples, then their disjoint union $\mathcal{P}$ is testable over the combined distribution D with $O(q(\epsilon/2) \cdot (\log^3 \frac{1}{\epsilon}))$ queries and $O(U(\epsilon/2) \cdot (\frac{N}{\epsilon}\log^3 \frac{1}{\epsilon}))$ unlabeled samples.*

Proof: Let $p = (p_1, \ldots, p_N)$ denote the mixing weights for distribution D; that is, a random draw from D can be

viewed as selecting i from distribution p and then selecting x from D_i. We are given that each $\mathcal{P}_i$ is testable with failure probability $1/3$ using using $q(\epsilon)$ queries and $U(\epsilon)$ unlabeled samples. By repetition, this implies that each is testable with failure probability δ using $q_\delta(\epsilon) = O(q(\epsilon)\log(1/\delta))$ queries and $U_\delta(\epsilon) = O(U(\epsilon)\log(1/\delta))$ unlabeled samples, where we will set $\delta = \epsilon^2$. We now test property $\mathcal{P}$ as follows:

For $\epsilon' = 1/2, 1/4, 1/8, \ldots, \epsilon/2$ do:
Repeat $O(\frac{\epsilon'}{\epsilon}\log(1/\epsilon))$ times:
1) Choose a random (i, x) from D.
2) Sample until either $U_\delta(\epsilon')$ samples have been drawn from D_i or $(8N/\epsilon)U_\delta(\epsilon')$ samples total have been drawn from D, whichever comes first.
3) In the former case, run the tester for property $\mathcal{P}_i$ with parameter ϵ', making $q_\delta(\epsilon')$ queries. If the tester rejects, then reject.

If all runs have accepted, then accept.

A straightforward calculation shows that the number of queries and samples is as desired. Furthermore, if indeed $f \in \mathcal{P}$ then each call to a tester rejects with probability at most δ so the overall failure probability is at most $(\delta/\epsilon)\log^2(1/\epsilon) < 1/3$. Thus, it suffices to analyze the case that $dist_D(f, \mathcal{P}) \geq \epsilon$.

If $dist_D(f, \mathcal{P}) \geq \epsilon$ then $\sum_{i:p_i \geq \epsilon/(4N)} p_i \cdot dist_{D_i}(f_i, \mathcal{P}_i) \geq 3\epsilon/4$. Moreover, for indices i such that $p_i \geq \epsilon/(4N)$, with high probability Step 2 draws $U_\delta(\epsilon')$ samples, so we may assume for such indices the tester for $\mathcal{P}_i$ is indeed run in Step 3. Let $I = \{i : p_i \geq \epsilon/(4N) \text{ and } dist_{D_i}(f_i, \mathcal{P}_i) \geq \epsilon/2\}$. Thus, we have

$$\sum_{i \in I} p_i \cdot dist_{D_i}(f_i, \mathcal{P}_i) \geq \epsilon/4.$$

Let $I_{\epsilon'} = \{i \in I : dist_{D_i}(f_i, \mathcal{P}_i) \in [\epsilon', 2\epsilon']\}$. Bucketing the above summation by values ϵ' in this way implies that for some value $\epsilon' \in \{\epsilon/2, \epsilon, 2\epsilon, \ldots, 1/2\}$, we have:

$$\sum_{i \in I_{\epsilon'}} p_i \geq \epsilon/(8\epsilon'\log(1/\epsilon)).$$

This in turn implies that with probability at least $2/3$, the run of the algorithm for this value of ϵ' will find such an i and reject, as desired. ∎

As a simple example, consider $\mathcal{P}_i$ to contain just the constant functions **1** and **0**. In this case, $\mathcal{P}$ is equivalent to what is often called the "cluster assumption," used in semi-supervised and active learning [11], [14], that if data lies in some number of clearly identifiable clusters, then all points in the same cluster should have the same label. Here, each $\mathcal{P}_i$ individually is easily testable (even passively) with $O(1/\epsilon)$ labeled samples, so Theorem V.1 implies the cluster assumption is testable with $poly(1/\epsilon)$ queries.[5] However, it is not hard to see that passive testing with $poly(1/\epsilon)$ samples is not possible and in fact requires $\Omega(\sqrt{N}/\epsilon)$ labeled examples.[6]

We build on this to produce testers for other properties often used in semi-supervised learning. In particular, one common assumption used (often called the *margin* or *low-density* assumption) is that there should be some large margin γ of separation between the positive and negative regions (but without assuming the target is necessarily a linear threshold function). Here, we give a tester for this property, which uses a tester for the cluster property as a subroutine, along with analysis of an appropriate weighted graph defined over the data. Specifically, we prove the following result (See the full paper [3] for definitions and analysis).

Theorem V.2. *For any γ, $\gamma' = \gamma(1 - 1/c)$ for constant $c > 1$, for data in the unit ball in R^d for constant d, we can distinguish the case that D_f has margin γ from the case that D_f is ϵ-far from margin γ' using Active Testing with $O(1/(\gamma^{2d}\epsilon^2))$ unlabeled examples and $O(1/\epsilon)$ label requests.*

VI. General Testing Dimensions

The previous sections have discussed upper and lower bounds for a variety of classes. Here, we define notions of *testing dimension* for passive and active testing that characterize (up to constant factors) the number of labels needed for testing to succeed, in the corresponding testing protocols. These will be distribution-specific notions (like SQ dimension [8] or Rademacher complexity [4] in learning), so let us fix some distribution D over the instance space X, and furthermore fix some value ϵ defining our goal. I.e., our goal is to distinguish the case that $dist_D(f, \mathcal{P}) = 0$ from the case $dist_D(f, \mathcal{P}) \geq \epsilon$.

For a given set S of unlabeled points, and a distribution π over boolean functions, define π_S to be the distribution over labelings of S induced by π. That is, for $y \in \{0,1\}^{|S|}$ let $\pi_S(y) = \Pr_{f\sim\pi}[f(S) = y]$. We now use this to define a distance between distributions. Specifically, given a set of unlabeled points S and two distributions π and π' over boolean functions, define

$$\mathrm{d}_S(\pi, \pi') = (1/2) \sum_{y \in \{0,1\}^{|S|}} |\pi_S(y) - \pi'_S(y)|,$$

to be the variation distance between π and π' induced by S. Finally, let Π_0 be the set of all distributions π over functions in $\mathcal{P}$, and let set Π_ϵ be the set of all distributions π' in which a $1 - o(1)$ probability mass is over functions at least ϵ-far from $\mathcal{P}$.

[5] Since the $\mathcal{P}_i$ are so simple in this case, one can actually test with only $O(1/\epsilon)$ queries.

[6] Specifically, suppose region 1 has $1-2\epsilon$ probability mass with $f_1 \in \mathcal{P}_1$, and suppose the other regions equally share the remaining 2ϵ probability mass and either (a) are each pure but random (so $f \in \mathcal{P}$) or (b) are each 50/50 (so f is ϵ-far from $\mathcal{P}$). Distinguishing these cases requires seeing at least two points with the same index $i \neq 1$, yielding the $\Omega(\sqrt{N}/\epsilon)$ bound.

We are now ready to formulate our notions of dimension. Each is based on defining an associated testing game and showing that the minimax value can be characterized in a natural way. There is also an interesting connection between the passive testing dimension and VC-dimension, which we present below. All proofs not presented here appear in the full version [3].

A. Passive Testing Dimension

Definition VI.1. Define the passive testing dimension, $d_{passive} = d_{passive}(\mathcal{P}, D)$, as the largest $q \in \mathbb{N}$ such that,

$$\sup_{\pi \in \Pi_0} \sup_{\pi' \in \Pi_\epsilon} \Pr_{S \sim D^q}(\mathrm{d}_S(\pi, \pi') > 1/4) \leq 1/4.$$

That is, there exist distributions $\pi \in \Pi_0$ and $\pi' \in \Pi_\epsilon$ such that a random set S of $d_{passive}$ examples has a reasonable probability (at least $3/4$) of having the property that one cannot reliably distinguish a random function from π versus a random function from π' from just the labels of S. From the definition it is fairly immediate that $\Omega(d_{passive})$ examples are *necessary* for passive testing; in fact, one can show that $O(d_{passive})$ are sufficient as well.

Theorem VI.2. *The sample complexity of passive testing property $\mathcal{P}$ over distribution D is $\Theta(d_{passive}(\mathcal{P}, D))$.*

Connections to VC dimension: This notion of dimension brings out an interesting connection between learning and testing. In particular, consider the special case that we simply wish to distinguish functions in $\mathcal{P}$ from truly random functions, so π' is the uniform distribution over all functions (this is indeed the form used by our lower bound results and many lower bounds in property testing). In that case, the passive testing dimension becomes the largest q such that for some (multi)set F of functions $f_i \in \mathcal{P}$, a typical sample S of size q would have all 2^q possible labelings occur *approximately the same number of times* over the functions $f_i \in F$. In contrast, the *VC-dimension* of $\mathcal{P}$ is the largest q such that for some sample S of size q, each of the 2^q possible labelings occurs *at least once*. Notice there is a kind of reversal of quantifiers here: in a distributional version of VC-dimension where one would like a "typical" set S to be shattered, the functions that induce the 2^q labelings could be different from sample to sample. However, for the testing dimension, the set F must be fixed in advance. That is the reason that it is possible for a tester to output "no" even though the labels observed are still consistent with some function in $\mathcal{P}$.

B. Active Testing Dimension

For the case of active testing, there are two complications. First, the algorithms can examine their entire $poly(n)$-sized unlabeled sample before deciding which points to query, and secondly they may in principle determine the next query based on the responses to the previous ones (even though all our algorithmic results do not require this feature). If we merely want to distinguish those properties that are actively testable with $O(1)$ queries from those that are not, then the second complication disappears and the first is simplified as well, and the following coarse notion of dimension suffices.

Definition VI.3. Define the coarse active testing dimension, $d_{coarse} = d_{coarse}(\mathcal{P}, D)$, as the largest $q \in \mathbb{N}$ such that,

$$\sup_{\pi \in \Pi_0} \sup_{\pi' \in \Pi_\epsilon} \Pr_{S \sim D^q}(\mathrm{d}_S(\pi, \pi') > 1/4) \leq 1/n^q.$$

Theorem VI.4. *If $d_{coarse}(\mathcal{P}, D) = O(1)$ then active testing of $\mathcal{P}$ over D can be done with $O(1)$ queries, and if $d_{coarse}(\mathcal{P}, D) = \omega(1)$ then it cannot.*

To achieve a more fine-grained characterization of active testing we consider a slightly more involved quantity, as follows. First, recall that given an unlabeled sample U and distribution π over functions, we define π_U as the induced distribution over labelings of U. We can view this as a distribution over *unlabeled* examples in $\{0,1\}^{|U|}$. Now, given two distributions over functions π, π', define $\mathrm{Fair}(\pi, \pi', U)$ to be the distribution over *labeled* examples (y, ℓ) defined as: with probability $1/2$ choose $y \sim \pi_U$, $\ell = 1$ and with probability $1/2$ choose $y \sim \pi'_U$, $\ell = 0$. Thus, for a given unlabeled sample U, the sets Π_0 and Π_ϵ define a *class* of fair distributions over labeled examples. The active testing dimension, roughly, asks how well this class can be approximated by the class of low-depth decision trees. Specifically, let DT_k denote the class of decision trees of depth at most k. The active testing dimension for a given number u of allowed unlabeled examples is as follows:

Definition VI.5. Given a number $u = poly(n)$ of allowed unlabeled examples, we define the active testing dimension, $d_{active}(u) = d_{active}(\mathcal{P}, D, u)$, as the largest $q \in \mathbb{N}$ such that

$$\sup_{\pi \in \Pi_0} \sup_{\pi' \in \Pi_\epsilon} \Pr_{U \sim D^u}(\mathrm{err}^*(\mathrm{DT}_q, \mathrm{Fair}(\pi, \pi', U)) < 1/4) \leq 1/4,$$

where $\mathrm{err}^*(H, P)$ is the error of the optimal function in H with respect to data drawn from distribution P over labeled examples.

Theorem VI.6. *Active testing of property $\mathcal{P}$ over distribution D with failure probability $\frac{1}{8}$ using u unlabeled examples requires $\Omega(d_{active}(\mathcal{P}, D, u))$ label queries, and furthermore can be done with $O(u)$ unlabeled examples and $O(d_{active}(\mathcal{P}, D, u))$ label queries.*

We now use these notions of dimension to prove lower bounds for testing several properties. To do so, we will use the following lemma, a generalization of a lemma widely used for proving lower bounds in property testing [17, Lem. 8.3]. (See [3] for a proof).

Lemma VI.7. *Let π and π' be two distributions on functions $X \to \mathbb{R}$. Fix $U \subseteq X$ to be a set of allowable queries. Suppose that for any $S \subseteq U$, $|S| = q$, there is a set $E_S \subseteq \mathbb{R}^q$*

(possibly empty) satisfying $\pi_S(E_S) \leq \frac{1}{5}2^{-q}$ *such that*

$$\pi_S(y) < \tfrac{6}{5}\pi'_S(y) \quad \textit{for every } y \in \mathbb{R}^q \setminus E_S.$$

Then $\mathrm{err}^*(\mathrm{DT}_q, \mathrm{Fair}(\pi, \pi', U)) > 1/4$.

C. Application: Dictator functions

We prove here that active testing of dictatorships over the uniform distribution requires $\Omega(\log n)$ queries by proving a $\Omega(\log n)$ lower bound on $d_{active}(u)$ for any $u = poly(n)$; in fact, this result holds even for the specific choice of π' as random noise (the uniform distribution over all functions).

Theorem VI.8. *Active testing of dictatorships under the uniform distribution requires* $\Omega(\log n)$ *queries. This holds even for distinguishing dictators from random functions.*

Proof: Define π and π' to be uniform distributions over the dictator functions and over all boolean functions, respectively. In particular, π is the distribution obtained by choosing $i \in [n]$ uniformly at random and returning the function $f : \{0,1\}^n \to \{0,1\}$ defined by $f(x) = x_i$. Fix S to be a set of q vectors in $\{0,1\}^n$. This set can be viewed as a $q \times n$ boolean-valued matrix. We write $c_1(S), \ldots, c_n(S)$ to represent the columns of this matrix. For any $y \in \{0,1\}^q$,

$$\pi_S(y) = \frac{|\{i \in [n] : c_i(S) = y\}|}{n} \qquad \text{and} \qquad \pi'_S(y) = 2^{-q}.$$

By Lemma VI.7, to prove that $d_{active} \geq \frac{1}{2}\log n$, it suffices to show that when $q < \frac{1}{2}\log n$ and U is a set of n^c vectors chosen uniformly and independently at random from $\{0,1\}^n$, then with probability at least $\frac{3}{4}$, every set $S \subseteq U$ of size $|S| = q$ and every $y \in \{0,1\}^q$ satisfy $\pi_S(y) \leq \frac{6}{5}2^{-q}$. (This is like a stronger version of d_{coarse} where $\mathrm{d}_S(\pi, \pi')$ is replaced with an L_∞ distance.)

Consider a set S of q vectors chosen uniformly and independently at random from $\{0,1\}^n$. For any vector $y \in \{0,1\}^q$, the expected number of columns of S that are equal to y is $n2^{-q}$. Since the columns are drawn independently at random, Chernoff bounds imply that

$$\Pr\left[\pi_S(y) > \tfrac{6}{5}2^{-q}\right] \leq e^{-(\frac{1}{5})^2 n 2^{-q}/3} < e^{-\frac{1}{75}n2^{-q}}.$$

By the union bound, the probability that there exists a vector $y \in \{0,1\}^q$ such that more than $\frac{6}{5}n2^{-q}$ columns of S are equal to y is at most $2^q e^{-\frac{1}{75}n2^{-q}}$. Furthermore, when U is defined as above, we can apply the union bound once again over all subsets $S \subseteq U$ of size $|S| = q$ to obtain $\Pr[\exists S, y : \pi_S(y) > \frac{6}{5}2^{-q}] < n^{cq} \cdot 2^q \cdot e^{-\frac{1}{75}n2^{-q}}$. When $q \leq \frac{1}{2}\log n$, this probability is bounded above by $e^{\frac{c}{2}\log^2 n + \frac{1}{2}\log n - \frac{1}{75}\sqrt{n}}$, which is less than $\frac{1}{4}$ when n is large enough, as we wanted to show. ∎

D. Application: LTFs

The testing dimension also lets us prove the lower bounds in Theorem IV.1 regarding the query complexity for testing linear threshold functions. Specifically, those bounds follow directly from the following result.

Theorem VI.9. *For linear threshold functions under the standard* n*-dimensional Gaussian distribution,* $d_{passive} = \Omega(\sqrt{n/\log(n)})$ *and* $d_{active} = \Omega((n/\log(n))^{1/3})$.

Let us give a brief overview of the strategies used to obtain the $d_{passive}$ and d_{active} bounds. The complete proofs for both results, as well as a simpler proof that $d_{coarse} = \Omega((n/\log n)^{1/3})$, can be found in the full version [3].

For both results, we set π to be a distribution over LTFs obtained by choosing $w \sim \mathcal{N}(0, I_{n\times n})$ and outputting $f(x) = \mathrm{sgn}(w \cdot x)$. Set π' to be the uniform distribution over all functions—i.e., for any $x \in \mathbb{R}^n$, the value of $f(x)$ is uniformly drawn from $\{0,1\}$ and is independent of the value of f on other inputs.

To bound $d_{passive}$, we bound the total variation distance between the distribution of $Xw/\sqrt{n}$ given X, and a normal $\mathcal{N}(0, I_{n\times n})$. If this distance is small, then so must be the distance between the distribution of $\mathrm{sgn}(Xw)$ and the uniform distribution over label sequences. In fact, we show this is the case for a broad family of product distributions, characterized by a condition on the moments of the coordinate projections.

Our strategy for bounding d_{active} is very similar to the one we used to prove the lower bound on the query complexity for testing dictator functions in the last section. Again, we want to apply Lemma VI.7. Specifically, we want to show that when $q \leq o((n/\log(n))^{1/3})$ and U is a set of n^c vectors drawn independently from the n-dimensional standard Gaussian distribution, then with probability at least $\frac{3}{4}$, every set $S \subseteq U$ of size $|S| = q$ and almost all $x \in \mathbb{R}^q$, we have $\pi_S(x) \leq \frac{6}{5}2^{-q}$. The difference between this case and the lower bound for dictator functions is that we now rely on strong concentration bounds on the spectrum of random matrices [31] to obtain the desired inequality.

VII. Conclusions

In this work we develop and analyze a model of property testing that parallels the active learning model in machine learning, in which queries are restricted to be selected from a given (polynomially) large unlabeled sample. We demonstrate that a number of important properties for machine learning can be efficiently tested in this setting with substantially fewer queries than needed to learn. We additionally give a combination result allowing one to build testable properties out of others, as well as develop notions of intrinsic *testing dimension* that characterize the number of queries needed to test, and which we then use to prove a number of lower bounds. One open problem is that for testing linear separators, for the active testing model we have an $\tilde{O}(\sqrt{n})$ upper bound and an $\tilde{\Omega}(n^{1/3})$ lower bound. It would be very exciting if the upper bound could be improved, but either way it would be interesting to close

that gap. Additionally, testing of linear separators over more general distributions would be quite interesting.

ACKNOWLEDGMENTS

M.-F. Balcan is supported in part by NSF grant CCF-0953192, AFOSR grant FA9550-09-1-0538, a Microsoft Faculty Fellowship and a Google Research Award. A. Blum is supported in part by the National Science Foundation under grants CCF-0830540, CCF-1116892, and IIS-1065251. L. Yang is supported in part by NSF grant IIS-1065251 and a Google Core AI grant.

REFERENCES

[1] Miguel A. Arcones. A Bernstein-type inequality for U-statistics and U-processes. *Statistics & Probability Letters*, 22(3):239 – 247, 1995.

[2] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proc. 23rd ICML*, 2006.

[3] Maria-Florina Balcan, Eric Blais, Avrim Blum, and Liu Yang. Active property testing. *CoRR*, abs/1111.0897, 2012.

[4] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

[5] Eric Baum and Kenneth Lang. Query learning can work poorly when a human oracle is used. In *Proc. IEEE Inter. Joint Conf. on Neural Networks*, 1993.

[6] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proc. 26th ICML*, 2009.

[7] Eric Blais. Testing juntas nearly optimally. In *Proc. 41st STOC*, pages 151–158, 2009.

[8] Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proc. 26th STOC*, pages 253–262, 1994.

[9] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[10] Rui M. Castro and Robert D. Nowak. Minimax bounds for active learning. In *Proc. 20th COLT*, 2007.

[11] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT press, 2006.

[12] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. In *Proc. 11th ICML*, pages 201–221, 1994.

[13] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Proc. 19th NIPS*, volume 18, 2005.

[14] Sanjoy Dasgupta. Two faces of active learning. *Theor. Comput. Sci.*, 412(19):1767–1781, April 2011.

[15] Sanjoy Dasgupta, Daniel J. Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. *In Proc. 21st NIPS*, 20, 2007.

[16] Ilias Diakonikolas, Homin Lee, Kevin Matulef, Krzysztof Onak, Ronitt Rubinfeld, Rocco Servedio, and Andrew Wan. Testing for concise representations. In *Proc. 48th FOCS*, pages 549–558, 2007.

[17] Eldar Fischer. The art of uninformed decisions. *Bulletin of the EATCS*, 75:97–126, 2001.

[18] Eldar Fischer, Guy Kindler, Dana Ron, Shmuel Safra, and Alex Samorodnitsky. Testing juntas. *J. Comput. Syst. Sci.*, 68:753–787, 2004.

[19] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.

[20] Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proc. 24th ICML*, 2007.

[21] Gil Kalai. Learnability and rationality of choice. *J. Economic Theory*, 113(1):104–117, 2003.

[22] Michael Kearns and Dana Ron. Testing problems with sublearning sample complexity. *J. Comput. and Syst. Sci.*, 61(3):428 – 456, 2000.

[23] Vladimir Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *J. Mach. Learn. Res.*, 11:2457–2485, 2010.

[24] Philip M. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Trans. on Neural Networks*, 6(6):1556–1559, 1995.

[25] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco A. Servedio. Testing halfspaces. In *Proc. 20th SODA*, pages 256–264, 2009.

[26] Dana Ron. Property testing: A learning theory perspective. *Foundations & Trends in Mach. Learn.*, 1(3):307–402, 2008.

[27] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.*, 25:252–271, 1996.

[28] H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proc. 5th COLT*, pages 287–294, 1992.

[29] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 4:45–66, 2001.

[30] Vladimir Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Prob. and its App.*, 16(2):264–280, 1971.

[31] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012. Available at `http://arxiv.org/abs/1011.3027`.