Active Property Testing

Maria-Florina Balcan[*]   Eric Blais[†]   Avrim Blum[‡]   Liu Yang[§]

Abstract

One of the motivations 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 fictitious examples constructed by the 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 in learning to the domain of *testing*.

We develop both general results for this *active testing* model as well as efficient testing algorithms for a number of 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 [41] 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.

Our results show that testing can be a powerful tool in realistic models for learning, and further that active testing exhibits an interesting and rich structure. Our work in addition brings together tools from a range of areas including U-statistics, noise-sensitivity, self-correction, and spectral analysis of random matrices, and develops new tools that may be of independent interest.

---

1   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 [32, 56]. 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 [49], juntas [28, 11], DNF formulas [25], and decision trees [25]. (See Ron's survey [56] 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 typical problem 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 in cases such as image classification—e.g., classifying handwritten digits into the numerals they represent—where a human labeler would be examining the same representation as the computer, queries can be problematic because the space of reasonable images is a very sparse subset of the entire domain. Indeed, now-classic experiments on membership-query learning algorithms for digit recognition ran into exactly this problem, leading to poor results [7]. 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* [58, 17, 61, 19, 4, 9, 14, 36, 21, 45]. 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 introducing, analyzing, and developing efficient algorithms for 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 2 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 with respect to 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 [32, 41] in which only random sampling of a small number points from the distribution is allowed. 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 an interesting and rich structure and strengthens the connection between testing and learning.

1.1 Our Results

We show that for a number of important properties for learning—including unions of intervals, linear threshold functions, and various assumptions used in semi-supervised learning—one can test in the active testing model with substantially fewer labels than needed to learn. We in addition consider the even more stringent *passive testing* model introduced by Goldreich, Goldwasser, and Ron [32] (where the only operation available to the algorithm is to draw a random labeled sample from D) and give new positive results for that model as well. We further show that for both active and passive testing models, we can *characterize* (up to constant factors) the intrinsic number of label requests needed to test any given property $\mathcal{P}$ with respect to any given distribution D in a new quantity we call the *testing dimension* of $\mathcal{P}$ with respect to D. We then use these dimension notions to prove several near-tight lower bounds. We expand on each of these points below.

Unions of intervals. The function $f : [0,1] \to \{0,1\}$ is a *union of d intervals* if the set $f^{-1}(1)$ consists of at most d intervals in $[0,1]$. It is known that $\Theta(d)$ queries are necessary and sufficient for *learning* functions from this class. Kearns and Ron [41] 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. However, prior to the current work, no non-trivial upper bound was known for the problem of distinguishing unions of d intervals from functions ϵ-far from unions of d intervals (as opposed to far from d/ϵ intervals).

We give an algorithm that tests unions of d intervals with only $O(1)$ queries in the active testing model. This result holds over *any* underlying distribution (known or unknown). Moreover, in the case that the underlying distribution is uniform, we require only $O(\sqrt{d})$ *unlabeled* samples. Thus, as a byproduct we improve over the prior best result in the passive testing model as well. Note that Kearns and Ron [41] show that $\Omega(\sqrt{d})$ examples are required to test unions of intervals over the uniform distribution in the passive testing model, so this result is tight. Moreover, one can show that in the distribution-free testing model of Halevy and Kushilevitz [35] one cannot perform testing of this class from $O(1)$ queries; thus, this class demonstrates a separation between these models (see Appendix A).

At the heart of the analysis of our algorithm is a characterization of functions that are unions of intervals in terms of their *noise sensitivity*, shown via developing a local self-corrector for this class. The noise sensitivity of boolean functions is a powerful tool that has led to recent advances in hardness of approximation [42, 52], learning theory [43, 44, 24], and differential privacy [16]. (See also [53] for more details on the applications of noise sensitivity to the study of boolean functions.) Our work presents a novel application of noise sensitivity in the domain of property testing.

Linear threshold functions. The function $f : \mathbb{R}^n \to \{0,1\}$ is a *linear threshold function* if there are $n+1$ parameters $w_1, \ldots, w_n, \theta \in \mathbb{R}$ such that $f(x) = \mathrm{sgn}(w_1 x_1 + \cdots + w_n x_n - \theta)$ for every $x \in \mathbb{R}^n$. Linear threshold functions are perhaps the most widely-used function class in machine learning. We show that both active and passive testing of testing linear threshold functions in R^n can be done with $O(\sqrt{n})$ labeled examples over the Gaussian distribution. This is substantially less than the $\Omega(n)$ labeled examples needed for learning (even over the Gaussian distribution [47]) and yields a new upper bound for the passive testing model as well. The key challenge here is that

estimating a statistic due to Matulef et al. [49]—which can be done with $O(1)$ queries if arbitrary queries are allowed [49]—would require $\Theta(n)$ samples if done from independent pairs of random examples in the natural way; this is no better than learning. We overcome this obstacle by re-using non-independent pairs of examples in the estimation, together with an analysis and modification of the statistic that allow for use of a theorem of Arcones [3] on the concentration of U-statistics. At a technical level, this result uses the fact that even though typical values of $(x \cdot y)^2$ may be quite large—i.e., $\Theta(n)$—when x and y have every coordinate selected from the standard normal, 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.

Interestingly, we show these bounds are nearly tight, giving lower bounds of $\tilde{\Omega}(n^{1/3})$ and $\tilde{\Omega}(\sqrt{n})$ on the number of labeled examples needed for active and passive testing respectively. The proof of these lower bounds relies on our notion of active and passive *testing dimensions*. More precisely, by using the notion of dimension, we reduce the problem of proving the lower bounds to that of bounding the operator norm of random matrices. This task is then completed by appealing to recent results on the non-asymptotic analysis of random matrices [64]. Our lower bound demonstrates a separation between the active model and the standard (arbitrary-query) testing model.

Disjoint unions of testable properties. We also show that any disjoint union of testable properties remains testable in the active testing model, allowing one to build testable properties out of simpler components; this is then used to provide label-efficient testers for several properties used in semi-supervised learning including the cluster and margin assumptions. See Section 5 for details.

Testing dimension. One of the most powerful notions in learning theory is that of the *dimension* or *intrinsic complexity* of a class of functions. Such notions of dimension (e.g., VC dimension [63], SQ dimension [12], Rademacher complexity [6]) have been exceedingly effective in determining the sample complexity for learning classes of functions in various learning models. Y. Mansour and G. Kalai (personal communication, see also [40]) posed the question of whether comparable notions of dimension might exist for testing. In this work, we answer in the affirmative and introduce the first such notions of dimension for property testing, for both our new model of active testing and the passive testing model.

We show that these notions of testing dimension characterize (up to constant factors) the intrinsic number of labeled examples required to test the given property with respect to a given distribution in the active and passive testing models, respectively. We also introduce a simpler "coarse" notion of testing dimension that characterizes the set of properties testable with $O(1)$ queries in the active testing model.

We use these testing dimensions to obtain lower bounds on the query complexity for testing a number of different properties in both active and passive testing models. Notably, we show that $\Omega(\log n)$ queries are needed to distinguish dictator functions from random functions in both models. This shows that active testing of dictators is as hard as learning dictator functions, and also implies a lower bound of $\Omega(\log n)$ queries for testing a large number of properties—including decision trees, functions of low Fourier degree, juntas, DNFs—in the active testing model.[1]

1.2 Related Work

Active learning. Active learning has become a topic of substantial importance in machine learning due to the rise of applications in which unlabeled data can be sampled much more cheaply than data can be labeled, including text classification [50, 61], medical imaging [39], and image and music retrieval [60, 48] among many others [30, 65, 66]. This has led to significant work in algorithmic development including a yearly active-learning competition, with monetary prizes.[2] There has also been substantial progress in the theoretical understanding of its underlying principles, including both algorithmic guarantees and the design and analysis of appropriate sample complexity measures for this setting [29, 4, 5, 9, 14, 18, 22, 21, 36, 45, 10, 2, 1, 37, 51]. Active learning, unlike passive learning,

---

[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]See http://www.causality.inf.ethz.ch/activelearning.php.

has no known strong Structural Risk Minimization bounds, which further motivates our work. We note that while our model is motivated by active learning, our techniques are very different from those in the active learning literature.

Other Testing Models. In addition to the standard model of property testing [57] and the passive model of property testing [32, 41] discussed above, other models have been introduced to address different testing scenarios. The *tolerant* testing model, introduced by Parnas, Ron, and Rubinfeld [54] was introduced to model situations where the tester must not only accept functions that have a given property but also must accept functions that are close to having the property. The *distribution-free* testing model was introduced by Halevy and Kushilevitz [35] (see also [33, 34, 31, 26]) to explore the setting where the tester does not know the underlying distribution D. Both of these models allow arbitrary queries, however, and so do not address the machine learning settings motivating this work in which one can only query inputs from a large sample of unlabeled points. In Appendix A, we discuss the technical relations between active testing and these other models.

2 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 [57], 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 2.1 (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 2.1 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 [32], 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 2.2 (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 $\text{poly}(n)$-sample, q-query ϵ-tester for $\mathcal{P}$ over D.[3]

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 2.1, 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 Appendix A for a comparison of active testing to other testing models.

3 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 $[\ell_1, u_1], \dots, [\ell_d, u_d]$ 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

---

[3]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.

of d intervals requires $\Omega(d)$ queries. By contrast, we show that active testing of unions of d intervals can be done with a number of label requests that is *independent* of d, 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. Note that previously it was not known how to test this class from $O(1)$ queries even in the (standard) membership query model even over the uniform distribution.[4]

Theorem 3.1. *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 3.1 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 3.2. Fix $\delta > 0$. The *local δ-noise sensitivity* of the function $f : [0, 1] \to \{0, 1\}$ at $x \in [0, 1]$ is $\mathrm{NS}_\delta(f, x) = \mathrm{Pr}_{y \sim_\delta x}[f(x) \neq f(y)]$, 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 3.3. *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. $\square$

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 Appendix C).

Lemma 3.4. *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 Appendix C. $\square$

---

[4]The best prior result achieved a relaxed guarantee of distinguishing the case that f is a union of d intervals from the case that f is ϵ-far from a union of d/ϵ intervals [41].

The noise sensitivity characterization of unions of intervals obtained by Proposition 3.3 and Lemma 3.4 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 3.3, Lemma 3.4, 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 $y \sim_\delta x$.[5] As a result, up to $poly(1/\epsilon)$ terms, we also improve over the *passive testing* bounds of Kearns and Ron [41] which are able only to distinguish the case that f is a union of d intervals from the case that f is ϵ-far from being a union of d/ϵ intervals. (Their results use $O(\sqrt{d}/\epsilon^{1.5})$ examples.) Kearns and Ron [41] 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 $[0, 1]$.[6] 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 points in the *sample* that are $\leq x$) will be within γ of the correct p_x value [13]. 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 3.1.

---

[5]Except for events of $O(\delta)$ probability mass at the boundary.
[6]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 $[p_x^L, p_x^U]$.

4 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 $f(x) = \text{sgn}(w_1 x_1 + \cdots w_n x_n - \theta)$.[7] The main result of this section is that 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 4.1. *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. Furthermore, no (even computationally inefficient) algorithm can test 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 [47]. Linear threshold functions can be tested with a constant number of queries in the standard (arbitrary query) property testing model [49].

The starting point for the upper bound in Theorem 4.1 is a characterization lemma of linear threshold functions in terms of the following self-correlation statistic. To be precise, 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 4.2. 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 4.3 (Matulef et al. [49])**.** *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.*

The proof of Lemma 4.3 relies on the Hermite decomposition of functions. In fact, the original characterization of Matulef et al. [49] is stated in terms of the level-1 Hermite weight of functions. The above characterization follows easily from their result. For completeness, we include the details in Appendix D.

Lemma 4.3 suggests an obvious approach to testing for linear threshold functions from random examples: simply estimate the self-correlation coefficient of Definition 4.2 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 [3] 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.

---

[7]Here, $\text{sgn}(z) = \mathbf{1}[z \geq 0]$ is the standard sign function.

LTF TESTER(f, ϵ)
Parameters: $\tau = \sqrt{4n \log(4n/\epsilon^3)}$, $m = 800\tau/\epsilon^3 + 32/\epsilon^6$.
 1. Draw $x^1, x^2, \dots, x^m$ independently at random from $\mathbb{R}^n$.
 2. Query $f(x^1), f(x^2), \dots, f(x^m)$.
 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 4.3, 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

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

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 $g(x, y) = \begin{cases} f(x)f(y) \langle x, y \rangle & |\langle x, y \rangle| \leq \tau \\ 0 & \text{otherwise} \end{cases}$ allows us to apply Arcones' theorem.

Lemma 4.4 (Arcones [3]). *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 $x^1, \dots, x^q$ independently at random and setting $U_q(h) = \binom{q}{2}^{-1} \sum_{i<j} h(x^i, x^j)$. Then for every $t > 0$,*

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

An argument combining Lemma 4.4 with a separate argument showing that g is "close" to an unbiased estimator for $\rho(f)$ provides the desired guarantee for the LTF TESTER. The complete proof is presented in Appendix D.

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 4.1 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 6, where we will use the notion of testing dimension to establish the bound.

5 Testing Disjoint Unions of Testable Properties

We now show that active testing has the feature that a disjoint union 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 Appendix G.

Specifically, given properties $\mathcal{P}_1, \dots, \mathcal{P}_N$ over domains $X_1, \dots, X_N$, 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 5.1. *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. See Appendix E. □

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 [15, 20], 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 5.1 implies the cluster assumption is testable with $poly(1/\epsilon)$ queries.[8] 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.[9]

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 Appendix G for definitions and analysis).

Theorem 5.2. *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.*

6   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 [12] or Rademacher complexity [6] 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 $\mathcal{P}$. We are now ready to formulate our notions of dimension.

---

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

[9]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.

6.1 Passive Testing Dimension

Definition 6.1. 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 6.2. *The sample complexity of passive testing property $\mathcal{P}$ over distribution D is $\Theta(d_{passive}(\mathcal{P}, D))$.*

Proof. See Appendix F. $\qquad\square$

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 in Sections 6.3 and 6.4). 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 $\mathcal{P}$.

6.2 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 6.3. 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 6.4. *If $d_{coarse}(\mathcal{P}, D) = O(1)$ the 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.*

Proof. See Appendix F. $\qquad\square$

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 6.5. 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

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

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 6.6. *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.*

Proof. See Appendix F. □

We now use these notions of dimension to prove lower bounds for testing several properties.

6.3 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 6.7. *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 B.1, 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

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

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. □

11

6.4 Application: LTFs

The testing dimension also lets us prove the lower bounds in Theorem 4.1 regarding the query complexity for testing linear threshold functions. Specifically, those bounds follow directly from the following result.

Theorem 6.8. *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 Appendix F.4.

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 $f(x) = \text{sgn}(w \cdot x)$. 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 $\text{sgn}(Xw)$ 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 B.1. 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 [64] to obtain the desired inequality.

7 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. These testing algorithms bring together tools from a range of areas including U-statistics, noise-sensitivity, and self-correction, and develop characterizations of certain function classes that may be of independent interest. 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 near-matching lower bounds. In the context of 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.

References

[1] Nir Ailon. Active learning ranking from pairwise preferences with almost optimal query complexity. *Journal of Machine Learning Research*, 13:137–164, 2012.

[2] Nir Ailon, Ron Begleiter, and Esther Ezra. A new active learning scheme with applications to learning to rank from pairwise preferences. *CoRR*, abs/1110.2136, 2011.

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

[4] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[5] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*, 2007.

[6] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[7] Eric Baum and Kenneth Lang. Query learning can work poorly when a human oracle is used. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1993.

[8] Mihir Bellare, Oded Goldreich, and Madhu Sudan. Free bits, PCPs and non-approximability – towards tight results. *SIAM J. Comput.*, 27(3):804–915, 1998.

[9] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.

[10] Alina Beygelzimer, Daniel Hsu, John Langford, and Tong Zhang. Agnostic active learning without constraints. In *NIPS*, pages 199–207, 2010.

[11] Eric Blais. Testing juntas nearly optimally. In *Proc. 41st Annual ACM Symposium on the Theory of Computing*, pages 151–158, 2009.

[12] 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 *STOC*, pages 253–262, 1994.

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

[14] Rui M. Castro and Robert D. Nowak. Minimax bounds for active learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*, 2007.

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

[16] Mahdi Cheraghchi, Adam Klivans, Pravesh Kothari, and Homin K. Lee. Submodular functions are noise stable. In *Proc. 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1586–1592, 2012.

[17] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 201–221, 1994.

[18] Sanjoy Dasgupta. Analysis of a greedy active learning strategy. *Advances in Neural Information Processing Systems*, 17, 2004.

[19] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

[20] Sanjoy Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 2011. To appear.

[21] Sanjoy Dasgupta, Daniel J. Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. *Advances in Neural Information Processing Systems*, 20, 2007.

[22] Sanjoy Dasgupta, Adam Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the 18th Annual Conference on Computational Learning Theory (COLT)*, 2005.

[23] Jason V. Davis and Inderjit Dhillon. Differential entropic clustering of multivariate gaussians. In *Advances in Neural Information Processing Systems 19*, 2006.

[24] Ilias Diakonikolas, Prahladh Harsha, Adam Klivans, Raghu Meka, Prasad Raghavendra, Rocco A. Servedio, and Li-Yang Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *Proc. 42nd Annual ACM Symposium on the Theory of Computing*, pages 533–542, 2010.

[25] Ilias Diakonikolas, Homin Lee, Kevin Matulef, Krzysztof Onak, Ronitt Rubinfeld, Rocco Servedio, and Andrew Wan. Testing for concise representations. In *Proc. 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 549–558, 2007.

[26] Elya Dolev and Dana Ron. Distribution-free testing algorithms for monomials with a sublinear number of queries. In *Proceedings of the 13th international conference on Approximation, and 14 the International conference on Randomization, and combinatorial optimization: algorithms and techniques*, APPROX/RANDOM'10, pages 531–544. Springer-Verlag, 2010.

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

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

[29] Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[30] T. Gao and D. Koller. Active classification based on value of classifier. In *NIPS*, 2011.

[31] Dana Glasner and Rocco A. Servedio. Distribution-free testing lower bound for basic boolean functions. *Theory of Computing*, 5(1):191–216, 2009.

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

[33] Shirley Halevy and Eyal Kushilevitz. Distribution-free connectivity testing. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 3122 of *Lecture Notes in Computer Science*, pages 393–404. Springer Berlin / Heidelberg, 2004.

[34] Shirley Halevy and Eyal Kushilevitz. A lower bound for distribution-free monotonicity testing. In *Approximation, Randomization and Combinatorial Optimization*, volume 3624 of *Lecture Notes in Computer Science*, pages 612–612. Springer Berlin / Heidelberg, 2005.

[35] Shirley Halevy and Eyal Kushilevitz. Distribution-free property-testing. *SIAM Journal on Computing*, 37(4):1107–1138, 2007.

[36] Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th Annual International Conference on Machine Learning (ICML)*, 2007.

[37] Steve Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

[38] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[39] Juan Eugenio Iglesias, Ender Konukoglu, Albert Montillo, Zhuowen Tu, and Antonio Criminisi. Combining generative and discriminative models for semantic segmentation of ct scans via active learning. In *Information Processing in Medical Imaging (IPMI)*, 2011.

14

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

[41] Michael Kearns and Dana Ron. Testing problems with sublearning sample complexity. *Journal of Computer and System Sciences*, 61(3):428 – 456, 2000.

[42] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for MAX-CUT and other two-variable CSPs? *SIAM J. Comput.*, 37(1):319–357, 2007.

[43] Adam Klivans, Ryan O'Donnell, and Rocco Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.*, 68(4):808–840, 2004.

[44] Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via gaussian surface area. In *Proc. 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 541–550, 2008.

[45] Vladimir Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

[46] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

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

[48] Michael I. Mandel, Graham E. Poliner, and Daniel P. W. Ellis. Support vector machine active learning for music retrieval. *Multimedia Syst.*, 12(1):3–13, 2006.

[49] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco A. Servedio. Testing halfspaces. In *Proc. 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 256–264, 2009.

[50] Andrew McCallum and Kamal Nigam. Employing EM in pool-based active learning for text classification. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 350–358, 1998.

[51] Stanislav Minsker. Plug-in approach to active learning. *JMLR*, 13:67–90, 2012.

[52] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. *Ann. Math.*, 171(1):295–341, 2010.

[53] Ryan O'Donnell. *Computational applications of noise sensitivity*. PhD thesis, Massachusetts Institute of Technology, 2003. AAI0805205.

[54] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, 72(6):1012–1042, 2006.

[55] Michal Parnas, Dana Ron, and Alex Samorodnitsky. Testing basic boolean formulae. *SIAM J. Discret. Math.*, 16(1):20–46, 2003.

[56] Dana Ron. Property testing: A learning theory perspective. *Foundations and Trends in Machine Learning*, 1(3):307–402, 2008.

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

[58] H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the 5th Annual ACM workshop on Computational learning theory*, pages 287–294, 1992.

[59] Georgi E. Shilov. *Linear Algebra*. Dover, 1977.

[60] Simon Tong and Edward Y. Chang. Support vector machine active learning for image retrieval. In *ACM Multimedia*, pages 107–118, 2001.

[61] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 4:45–66, 2001.

[62] Vladimir Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

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

[64] 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`.

[65] C. Vondrick and D. Ramanan. Video annotation and tracking with active learning. In *NIPS*, 2011.

[66] Jenna Wiens and John V. Guttag. Active learning applied to patient-adaptive heartbeat classification. In *NIPS*, pages 2442–2450, 2010.

A Comparison of Active Testing and Other Property Testing Models

In this section, we compare the active testing model with four existing models of property testing: the standard property testing model as introduced by Rubinfeld and Sudan [57], the passive testing model first studied by Goldreich, Goldwasser, and Ron [32], the tolerant property testing model introduced by Parnas, Ron, and Rubinfeld [54], and the distribution-free property testing model of Halevy and Kushilevitz [35].

A.1 Standard and Passive Property Testing

Fix some sets X, Y and let $\mathcal{P}$ be some property of functions $f : X \to Y$. Let D be some distribution over X. Recall that the standard model of property testing is defined as follows.

Definition A.1 (Standard Property Tester [57])**.** A *q-query (standard) ϵ-tester* for $\mathcal{P}$ over the distribution D is a randomized algorithm A that queries the value of a function f on q of its inputs and then

1. Accepts with probability at least $\frac{2}{3}$ when $f \in \mathcal{P}$, and
2. Rejects with probability at least $\frac{2}{3}$ when $\mathrm{d}_D(f, \mathcal{P}) \geq \epsilon$.

The most commonly-studied case is where the distribution D is uniform over the domain of the function. When that is not the case, note that we can assume that the tester *knows* the distribution D. For the alternate model where the tester does not know D, see Section A.3.

The *passive* property testing model is similar to the standard property testing model, except that the queries made by the tester in this model are drawn at random from D.

Definition A.2 (Passive Property Tester [32])**.** A *q-query passive ϵ-tester* for $\mathcal{P}$ over the distribution D is a randomized algorithm A that draws q samples independently at random from D, queries the value of a function f on each of these q samples, and then

1. Accepts with probability at least $\frac{2}{3}$ when $f \in \mathcal{P}$, and
2. Rejects with probability at least $\frac{2}{3}$ when $\mathrm{d}_D(f, \mathcal{P}) \geq \epsilon$.

The query complexity of a property under a given testing model is the minimum query complexity of any tester for the property in this model. We denote the query complexity of properties in the standard, passive, and active testing models with the following notation.

Definition A.3 (Query complexity)**.** The *query complexity* of $\mathcal{P}$ over D in the standard property testing model is

$$Q_{D,\epsilon}(\mathcal{P}) := \min\{q > 0 : \text{there exists a } q\text{-query } \epsilon\text{-tester for } \mathcal{P}\}.$$

Similarly, the query complexity of $\mathcal{P}$ over D in the active and passive testing models is

$$Q_{D,\epsilon}^{a}(\mathcal{P}) := \min\{q > 0 : \text{there exists a } q\text{-query active } \epsilon\text{-tester for } \mathcal{P}\}$$
$$Q_{D,\epsilon}^{p}(\mathcal{P}) := \min\{q > 0 : \text{there exists a } q\text{-query passive } \epsilon\text{-tester for } \mathcal{P}\}.$$

With this notation in place, we can now formally establish the relationship between the standard, active, and passive models of property testing.

Theorem A.4. *For every property $\mathcal{P}$, every distribution D, and every $\epsilon > 0$,*

$$Q_{D,\epsilon}(\mathcal{P}) \leq Q_{D,\epsilon}^{a}(\mathcal{P}) \leq Q_{D,\epsilon}^{p}(\mathcal{P}). \tag{1}$$

Furthermore, the three testing models are distinct: there exist properties $\mathcal{P}$, distributions D, and constants $\epsilon > 0$ such that $Q_{D,\epsilon}(\mathcal{P}) < Q_{D,\epsilon}^{a}(\mathcal{P})$ and there also exist $\mathcal{P}, D, \epsilon$ such that $Q_{D,\epsilon}^{a}(\mathcal{P}) < Q_{D,\epsilon}^{p}(\mathcal{P})$.

Proof. Both inequalities in (1) are obtained with simple arguments. For the first inequality, note that we can always simulate an active tester in the standard property testing model by internally sampling[10] a random subset of the inputs in the domain of the function f and having the active tester select from those inputs. The second inequality follows from the fact that we can simulate a passive tester in the active testing model by querying the function on the first $Q_{D,\epsilon}^{p}(\mathcal{P})$ samples drawn at random from D.

The distinctness of the three models of property testing is not as immediate, but it follows from the main results in our paper. Theorem 6.7 shows that testing dictatorship in the active testing model requires $\Omega(\log n)$ queries. The same property can be tested with $O(1/\epsilon)$ queries in the standard testing model [8, 55], so this establishes the first strict inequality. For the second strict inequality, consider the property of being a union of d intervals. Theorem 3.1 shows that we can test this property with $O(1/\epsilon^4)$ queries in the active testing model but $\Omega(\sqrt{d})$ queries are required to test the same property in the passive model [41]. □

A.2 Tolerant Testing

The tolerant property testing model is an extension of the standard model of property testing with one extra requirement: the tester must accept functions with a given property $\mathcal{P}$ *as well as* functions that are (very) close to $\mathcal{P}$. Formally, the model is defined as follows.

Definition A.5 (Tolerant Property Tester [54])**.** Fix $0 \leq \epsilon_1 < \epsilon_2 \leq 1$. A *$q$-query tolerant (ϵ_1, ϵ_2)-tester* for $\mathcal{P}$ over the distribution D is a randomized algorithm A that queries the value of a funciton f on q of its inputs and then

1. Accepts with probability at least $\frac{2}{3}$ when $d_D(f, \mathcal{P}) \leq \epsilon_1$, and
2. Rejects with probability at least $\frac{2}{3}$ when $d_D(f, \mathcal{P}) \geq \epsilon_2$.

Definition A.6. The *query complexity* of $\mathcal{P}$ over D in the tolerant property testing model is

$$Q_{D,\epsilon_1,\epsilon_2}^{\text{tol}}(\mathcal{P}) := \min\{q > 0 : \text{there exists a } q\text{-query tolerant } (\epsilon_1, \epsilon_2)\text{-tester for } \mathcal{P}\}.$$

---

[10]Note that here we use the fact that a standard property tester knows the underlying distribution D and can therefore generate samples from this distribution "for free".

One may ask whether every property that has a query-efficient tolerant tester also has a query-efficient tester in the active model. Our lower bound on the query complexity for testing dictator functions in the active model gives a negative answer to this question: there are properties that require significantly more queries to test in the active model than in the tolerant testing model.

Theorem A.7. *There exist* $\mathcal{P}$*,* D*, and* $0 \leq \epsilon_1 < \epsilon_2 \leq 1$ *for which* $Q^{\mathrm{tol}}_{D,\epsilon_1,\epsilon_2}(\mathcal{P}) < Q^a_{D,\epsilon_2}(\mathcal{P})$.

Proof. Consider the property $\mathcal{P}$ of being a dictator function and let D be the uniform distribution over the hypercube. Theorem 6.7 shows that $Q^a_{D,\epsilon_2}(\mathcal{P}) = \Omega(\log n)$. By contrast, standard testers for dictator functions [8, 55] are tolerant (ϵ_1, ϵ_2)-testers with query complexity $O(1/(\epsilon_2 - \epsilon_1)^2)$ so the inequality in the theorem statement holds when $\epsilon_2 - \epsilon_1 = \Theta(1)$. □

We believe that the tolerant and active property testing models are incomparable—i.e., that there exist properties $\mathcal{P}$ (along with distributions D and parameters $\epsilon_1 < \epsilon_2$) for which the inequality in Theorem A.7 is reversed and $Q^{\mathrm{tol}}_{D,\epsilon_1,\epsilon_2}(\mathcal{P}) > Q^a_{D,\epsilon_2}(\mathcal{P})$. We leave the proof (or disproof) of this assertion as an open problem.

A.3 Distribution-free testing

In the above property testing models, the tester knows the underlying distribution D. To model the scenario where the tester does not know D, Halevy and Kushilevitz [35] introduced the distribution-free testing model. (See also [33, 34, 31, 26].) The model is defined formally as follows.

Definition A.8 (Distribution-free Tester [35]). An *s-sample, q-query distribution-free ϵ-tester* for $\mathcal{P}$ is a randomized algorithm A that draws s independent samples from the (unknown) distribution D, queries the value of the (unknown) function f on those s samples and $q - s$ additional inputs of its choosing, and then

1. Accepts with probability at least $\frac{2}{3}$ when $f \in \mathcal{P}$, and
2. Rejects with probability at least $\frac{2}{3}$ when $\mathrm{d}_D(f, \mathcal{P}) \geq \epsilon$.

Definition A.9. The *query complexity* of the property $\mathcal{P}$ in the distribution-free model is

$$Q^{\mathrm{df}}_\epsilon(\mathcal{P}) := \min\{q > 0 : \text{for some } 0 < s \leq q, \text{there exists an } s\text{-sample, } q\text{-query distribution-free } \epsilon\text{-tester for } \mathcal{P}\}.$$

Superficially, the distribution-free and active testing models appear to be similar: in both models, the tester first samples the underlying distribution D and then queries the value of the function on some inputs. The challenges in the two models, however, are mostly orthogonal and, as a result, the two models of property testing are incomparable. This statement is made precise by the following two results.

Theorem A.10. *There exist properties* $\mathcal{P}$ *such that for every distribution* D *and every large enough constant* $\epsilon > 0$*,* $Q^a_{D,\epsilon}(\mathcal{P}) < Q^{\mathrm{df}}_\epsilon(\mathcal{P})$.

Proof. Fix a large enough $d > 0$ and let $\mathcal{P}$ be the property consisting of the set of unions of d intervals. Theorem 3.1 shows that for every distribution D, we have $Q^a_{D,\epsilon}(\mathcal{P}) = O(1/\epsilon^4)$. To complete the proof of the theorem, we now show that $Q^{\mathrm{df}}_\epsilon(\mathcal{P}) = \Omega(\sqrt{d})$.

Consider the following two distributions on pairs of functions $f : [0, 1] \to \{0, 1\}$ and distributions D on $[0, 1]$. For the distribution $\mathcal{F}_{\mathrm{yes}}$, choose a set S of d points sampled independently and uniformly at random from $[0, 1]$. Define D to be the uniform distribution over S, and let $f : [0, 1] \to \{0, 1\}$ be a random function defined by choosing $f(x)$ uniformly at random for every $x \in S$ and setting $f(x) = 0$ for all $x \in [0, 1] \setminus S$. Clearly, every such function f is a union of d intervals.

The distribution $\mathcal{F}_{\mathrm{no}}$ is defined similarly except that in this case we let S be a set of $10d$ points. We define D to be uniform over S and again define $f : [0, 1] \to \{0, 1\}$ by choosing $f(x)$ uniformly at random for all $x \in S$ and

setting $f(x) = 0$ for all remaining points. In this case, whp the resulting functions are far from unions of d intervals over D_2.

Let A be a distribution-free tester for unions of d intervals. The tester A must accept with high probability when we draw a function f and distribution D from $\mathcal{F}_{\text{yes}}$ and it must reject with high probability when instead we draw a function and distribution from $\mathcal{F}_{\text{no}}$. Clearly, querying the functions on points that were not drawn from the distribution D will not help A since with probability 1 it will observe $f(x) = 0$ on those points. Assume now that A makes $s = o(\sqrt{d})$ draws to the distribution D. By the birthday paradox, with probability $1 - o(1)$, the s samples drawn from the distribution are distinct. In this case, the distributions on the values of the function on those s inputs are uniformly random so it has no way to distinguish whether the input was drawn from $\mathcal{F}_{\text{yes}}$ or from $\mathcal{F}_{\text{no}}$. This contradicts the assumption that A is a valid distribution-free tester for unions of d intervals and completes the proof of the lower bound on $Q_\epsilon^{\text{df}}(\mathcal{P})$. $\qquad\square$

Theorem A.11. *There exist properties $\mathcal{P}$, distributions D, and parameters $\epsilon > 0$ such that $Q_\epsilon^{\text{df}}(\mathcal{P}) < Q_{D,\epsilon}^a(\mathcal{P})$.*

Proof. Let $\mathcal{P}$ be the property of being a dictator function, let D be the uniform distribution over the hypercube, and let $\epsilon > 0$ be some constant. Theorem 6.7 shows that $Q_{D,\epsilon}^a(\mathcal{P}) = \Omega(\log n)$. By contrast, Halevy and Kushilevitz [35] showed that it is possible to test dictator functions in the distribution-free model with a constant number of queries when ϵ is constant and so $Q_\epsilon^{\text{df}}(\mathcal{P}) = O(1)$. $\qquad\square$

B Proof of a Property Testing Lemma

The following lemma is a generalization of a lemma that is widely used for proving lower bounds in property testing [27, Lem. 8.3]. We use this lemma to prove the lower bounds on the query complexity for testing dictator functions and testing linear threshold functions.

Lemma B.1. *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 $\pi_S(E_S) \le \frac{1}{5} 2^{-q}$ such that*

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

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

Proof. Consider any decision tree $\mathcal{A}$ of depth q. Each internal node of the tree consists of a query $y \in U$ and a subset $T \subseteq \mathbb{R}$ such that its children are labeled by T and $\mathbb{R} \setminus T$, respectively. The leaves of the tree are labeled with either "accept" or "reject", and let L be the set of leaves labeled as accept. Each leaf $\ell \in L$ corresponds to a set $S_\ell \subseteq U^q$ of queries and a subset $T_\ell \subseteq \mathbb{R}^\ell$, where $f : X \to \mathbb{R}$ leads to the leaf ℓ iff $f(S_\ell) \in T_\ell$. The probability that $\mathcal{A}$ (correctly) accepts an input drawn from π is

$$a_1 = \sum_{\ell \in L} \int_{T_\ell} \pi_{S_\ell}(y)\,dy.$$

Similarly, the probability that $\mathcal{A}$ (incorrectly) accepts an input drawn from π' is

$$a_2 = \sum_{\ell \in L} \int_{T_\ell} \pi'_{S_\ell}(y)\,dy.$$

The difference between the two rejection probabilities is bounded above by

$$a_1 - a_2 \le \sum_{\ell \in L} \int_{T_\ell \setminus E_{S_\ell}} \pi_{S_\ell}(y) - \pi'_{S_\ell}(y)\,dy + \sum_{\ell \in L} \int_{T_\ell \cap E_{S_\ell}} \pi_{S_\ell}(y)\,dy.$$

The conditions in the statement of the lemma then imply that

$$a_1 - a_2 < \sum_{\ell \in L} \int_{T_\ell} \tfrac{1}{6} \pi_{S_\ell}(y) dy + \tfrac{5}{6} \sum_\ell \int_{E_{S_\ell}} \pi_{S_\ell}(y) dy \leq \tfrac{1}{3}.$$

To complete the proof, we note that $\mathcal{A}$ errs on an input drawn from $\mathrm{Fair}(\pi, \pi', U)$ with probability

$$\tfrac{1}{2}(1 - a_1) + \tfrac{1}{2}a_2 = \tfrac{1}{2} - \tfrac{1}{2}(a_1 - a_2) > \tfrac{1}{3}. \qquad \square$$

C Proofs for Testing Unions of Intervals

In this section we complete the proofs of the technical results in Section 3.

Proposition 3.3 (Restated). *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. For any fixed $b \in [0, 1]$, the probability that $x < b < y$ when $x \sim U(0, 1)$ and $y \sim U(x - \delta, x + \delta)$ is

$$\Pr_{x,y}[x < b < y] = \int_0^\delta \Pr_{y \sim U(b-t-\delta, b-t+\delta)}[y \geq b] \mathrm{d}t = \int_0^\delta \frac{\delta - t}{2\delta} \mathrm{d}t = \frac{\delta}{4}.$$

Similarly, $\Pr_{x,y}[y < b < x] = \frac{\delta}{4}$. So the probability that b lies between x and y is at most $\frac{\delta}{2}$.

When f is the union of d intervals, $f(x) \neq f(y)$ only if at least one of the boundaries $b_1, \ldots, b_{2d}$ of the intervals of f lies in between x and y. So by the union bound, $\Pr[f(x) \neq f(y)] \leq 2d(\delta/2) = d\delta$. Note that if b is within distance δ of 0 or 1, the probability is only lower. $\qquad \square$

Lemma 3.4 (Restated). *Fix $\delta = \frac{\epsilon^2}{32d}$. Let $f : [0, 1] \to \{0, 1\}$ be any function with noise sensitivity $\mathbb{NS}_\delta(f) \leq d\delta(1 + \frac{\epsilon}{4})$. Then f is ϵ-close to a union of d intervals.*

Proof. The proof proceeds in two steps: We first show that f is $\frac{\epsilon}{2}$-close to a union of $d(1 + \frac{\epsilon}{2})$ intervals, then we show that every union of $d(1 + \frac{\epsilon}{2})$ intervals is $\frac{\epsilon}{2}$-close to a union of d intervals.

Consider the "smoothed" function $f_\delta : [0, 1] \to [0, 1]$ defined by

$$f_\delta(x) = \mathbb{E}_{y \sim_\delta x} f(y) = \frac{1}{2\delta} \int_{x-\delta}^{x+\delta} f(y) \mathrm{d}y.$$

The function f_δ is the convolution of f and the uniform kernel $\phi : \mathbb{R} \to [0, 1]$ defined by $\phi(x) = \frac{1}{2\delta} \mathbf{1}[|x| \leq \delta]$.

Fix $\tau = \frac{4}{\epsilon} \mathbb{NS}_\delta(f)$. We introduce the function $g^* : [0, 1] \to \{0, 1, *\}$ by setting

$$g^*(x) = \begin{cases} 1 & \text{when } f_\delta(x) \geq 1 - \tau, \\ 0 & \text{when } f_\delta(x) \leq \tau, \text{ and} \\ * & \text{otherwise} \end{cases}$$

for all $x \in [0, 1]$. Finally, we define $g : [0, 1] \to \{0, 1\}$ by setting $g(x) = g^*(y)$ where $y \leq x$ is the largest value for which $g(y) \neq *$. (If no such y exists, we fix $g(x) = 0$.)

We first claim that $dist(f, g) \leq \frac{\epsilon}{2}$. To see this, note that

$$\begin{aligned} dist(f, g) &= \Pr_x[f(x) \neq g(x)] \\ &\leq \Pr_x[g^*(x) = *] + \Pr_x[f(x) = 0 \wedge g^*(x) = 1] + \Pr_x[f(x) = 1 \wedge g^*(x) = 0] \\ &= \Pr_x[\tau < f_\delta(x) < 1 - \tau] + \Pr_x[f(x) = 0 \wedge f_\delta(x) \geq 1 - \tau] + \Pr_x[f(x) = 1 \wedge f_\delta(x) \leq \tau]. \end{aligned}$$

20

We bound the three terms on the RHS individually. For the first term, we observe that $\mathrm{NS}_\delta(f, x) = \min\{f_\delta(x), 1 - f_\delta(x)\}$ and that $\mathbb{E}_x \mathrm{NS}_\delta(f, x) = \mathbb{NS}_\delta(f)$. From these identities and Markov's inequality, we have that

$$\Pr_x[\tau < f_\delta(x) < 1 - \tau] = \Pr_x[\mathrm{NS}_\delta(f, x) > \tau] < \frac{\mathbb{NS}_\delta(f)}{\tau} = \frac{\epsilon}{4}.$$

For the second term, let $S \subseteq [0, 1]$ denote the set of points x where $f(x) = 0$ and $f_\delta(x) \geq 1 - \tau$. Let $\Gamma \subseteq S$ represent a δ-net of S. Clearly, $|\Gamma| \leq \frac{1}{\delta}$. For $x \in \Gamma$, let $B_x = (x - \delta, x + \delta)$ be a ball of radius δ around x. Since $f_\delta(x) \geq 1 - \tau$, the intersection of S and B_x has mass at most $|S \cap B_x| \leq \tau\delta$. Therefore, the total mass of S is at most $|S| \leq |\Gamma|\tau\delta = \tau$. By the bounds on the noise sensitivity of f in the lemma's statement, we therefore have

$$\Pr_x[f(x) = 0 \wedge f_\delta(x) \geq 1 - \tau] \leq \tau \leq \tfrac{\epsilon}{8}.$$

Similarly, we obtain the same bound on the third term. As a result, $dist(f, g) \leq \frac{\epsilon}{4} + \frac{\epsilon}{8} + \frac{\epsilon}{8} = \frac{\epsilon}{2}$, as we wanted to show.

We now want to show that g is a union of $m \leq d\delta(1 + \frac{\epsilon}{2})$ intervals. Each left boundary of an interval in g occurs at a point $x \in [0, 1]$ where $g^*(x) = *$, where the maximum $y \leq x$ such that $g^*(y) \neq *$ takes the value $g^*(y) = 0$, and where the minimum $z \geq x$ such that $g^*(z) \neq *$ has the value $g^*(z) = 1$. In other words, for each left boundary of an interval in g, there exists an interval (y, z) such that $f_\delta(y) \leq \tau$, $f_\delta(z) \geq 1 - \tau$, and for each $y < x < z$, $f_\delta(x) \in (\tau, 1 - \tau)$. Fix any interval (y, z). Since f_δ is the convolution of f with a uniform kernel of width 2δ, it is Lipschitz continuous (with Lipschitz constant $\frac{1}{2\delta}$). So there exists $x \in (y, z)$ such that the conditions $f_\delta(x) = \frac{1}{2}$, $x - y \geq 2\delta(\frac{1}{2} - \tau)$, and $z - x \geq 2\delta(\frac{1}{2} - \tau)$ all hold. As a result,

$$\int_y^z \mathrm{NS}_\delta(f, t) \, dt = \int_y^x \mathrm{NS}_\delta(f, t) \, dt + \int_x^z \mathrm{NS}_\delta(f, t) \, dt \geq 2\delta(\tfrac{1}{2} - \tau)^2.$$

Similarly, for each right boundary of an interval in g, we have an interval (y, z) such that

$$\int_y^z \mathrm{NS}_\delta(f, t) \, dt \geq 2\delta(\tfrac{1}{2} - \tau)^2.$$

The intervals (y, z) for the left and right boundaries are all disjoints, so

$$\mathbb{NS}_\delta(f) \geq \sum_{i=1}^{2m} \int_{y^i}^{z^i} \mathrm{NS}_\delta(f, t) \, dt \geq 2m \frac{\delta}{2}(1 - 2\tau)^2.$$

This means that

$$m \leq \frac{d\delta(1 + \epsilon/4)}{\delta(1 - 2\tau)^2} \leq d(1 + \tfrac{\epsilon}{2})$$

and g is a union of at most $d(1 + \frac{\epsilon}{2})$ intervals, as we wanted to show.

Finally, we want to show that any function that is the union of $m \leq d(1 + \frac{\epsilon}{2})$ intervals is $\frac{\epsilon}{2}$-close to a union of d intervals. Let $\ell_1, \ldots, \ell_m$ represent the lengths of the intervals in g. Clearly, $\ell_1 + \cdots + \ell_m \leq 1$, so there must be a set S of $m - d \leq d\epsilon/2$ intervals in f with total length

$$\sum_{i \in S} \ell_i \leq \frac{m - d}{m} \leq \frac{d\epsilon/2}{d(1 + \frac{\epsilon}{2})} < \frac{\epsilon}{2}.$$

Consider the function $h : [0, 1] \to \{0, 1\}$ obtained by removing the intervals in S from g (i.e., by setting $h(x) = 0$ for the values $x \in [b_{2i-1}, b_{2i}]$ for some $i \in S$). The function h is a union of d intervals and $dist(g, h) \leq \frac{\epsilon}{2}$. This completes the proof, since $dist(f, h) \leq dist(f, g) + dist(g, h) \leq \epsilon$. $\qquad \square$

D   Proofs for Testing LTFs

We complete the proof that LTFs can be tested with $\tilde{O}(\sqrt{n})$ samples in this section.

D.1   Proof of Lemma 4.3

The proof of Lemma 4.3 uses the Hermite decomposition of functions. We begin by introducing this notion and related definitions.

Definition D.1. The *Hermite polynomials* are a set of polynomials $h_0(x) = 1, h_1(x) = x, h_2(x) = \frac{1}{\sqrt{2}}(x^2 - 1), \ldots$ that form a complete orthogonal basis for (square-integrable) functions $f : \mathbb{R} \to \mathbb{R}$ over the inner product space defined by the inner product $\langle f, g \rangle = \mathbb{E}_x[f(x)g(x)]$, where the expectation is over the standard Gaussian distribution $\mathcal{N}(0, 1)$.

Definition D.2. For any $S \in \mathbb{N}^n$, define $H_S = \prod_{i=1}^{n} h_{S_i}(x_i)$. The *Hermite coefficient* of $f : \mathbb{R}^n \to \mathbb{R}$ corresponding to S is $\hat{f}(S) = \langle f, H_S \rangle = \mathbb{E}_x[f(x)H_S(x)]$ and the *Hermite decomposition* of f is $f(x) = \sum_{S \in \mathbb{N}^n} \hat{f}(S)H_S(x)$. The *degree* of the coefficient $\hat{f}(S)$ is $|S| := \sum_{i=1}^{n} S_i$.

The connection between linear threshold functions and the Hermite decomposition of functions is revealed by the following key lemma of Matulef et al. [49].

Lemma D.3 (Matulef et al. [49]). *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 $\sum_{i=1}^{n} \hat{f}(e_i)^2 = W(\mathbb{E}_x f)$. Moreover, every function $g : \mathbb{R}^n \to \{-1, 1\}$ that satisfies $\left| \sum_{i=1}^{n} \hat{g}(e_i)^2 - W(\mathbb{E}_x g) \right| \leq 4\epsilon^3$, is ϵ-close to being a linear threshold function.*

In other words, Lemma D.3 shows that $\sum_i \hat{f}(e_i)^2$ characterizes linear threshold functions. To obtain Lemma 4.3, it suffices to show that this sum is equivalent to $\mathbb{E}_{x,y}[f(x)f(y)\langle x, y \rangle]$. This identity is easily obtained:

Lemma D.4. *For any function $f : \mathbb{R}^n \to \mathbb{R}$, we have $\sum_{i=1}^{n} \hat{f}(e_i)^2 = \mathbb{E}_{x,y}[f(x)f(y)\langle x, y \rangle]$.*

Proof. Applying the Hermite decomposition of f and linearity of expectation,

$$\mathbb{E}_{x,y}[f(x)f(y)\langle x, y \rangle] = \sum_{i=1}^{n} \sum_{S,T \in \mathbb{N}^n} \hat{f}(S)\hat{f}(T)\mathbb{E}_x[H_S(x)x_i]\mathbb{E}_y[H_T(y)y_i].$$

By definition, $x_i = h_1(x_i) = H_{e_i}(x)$. The orthonormality of the Hermite polynomials therefore guarantees that $\mathbb{E}_x[H_S(x)H_{e_i}(x)] = \mathbf{1}[S = e_i]$. Similarly, $\mathbb{E}_y[H_T(y)y_i] = \mathbf{1}[T = e_i]$. $\square$

D.2   Analysis of LTF TESTER

We now complete the analysis of the LTF TESTER algorithm.

For a fixed function $f : \mathbb{R}^n \to \mathbb{R}$, define $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ to be $g(x, y) = f(x)f(y)\langle x, y \rangle$. Let $g^* : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the truncation of g defined by setting

$$g^*(x, y) = \begin{cases} f(x)f(y)\langle x, y \rangle & \text{if } |\langle x, y \rangle| \leq \sqrt{4n \log(4n/\epsilon^3)} \\ 0 & \text{otherwise.} \end{cases}$$

Our goal is to estimate $\mathbb{E}g$. The following lemma shows that $\mathbb{E}g^*$ provides a good estimate of this value.

Lemma D.5. *Let $g, g^* : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be defined as above. Then $|\mathbb{E}g - \mathbb{E}g^*| \leq \frac{1}{2}\epsilon^3$.*

Proof. For notational clarity, fix $\tau = \sqrt{4n \log(4n/\epsilon^3)}$. By the definition of g and g^* and with the trivial bound $|f(x)f(y)\langle x,y \rangle| \le n$ we have

$$|\mathbb{E}g - \mathbb{E}g^*| = \left| \Pr_{x,y} \left[|\langle x,y \rangle| > \tau \right] \cdot \mathbb{E}_{x,y} \left[f(x)f(y)\langle x,y \rangle \mid |\langle x,y \rangle| > \tau \right] \right| \le n \cdot \Pr_{x,y} \left[|\langle x,y \rangle| > \tau \right].$$

The right-most term can be bounded with a standard Chernoff argument. By Markov's inequality and the independence of the variables $x_1, \ldots, x_n, y_1, \ldots, y_n$,

$$\Pr_{x,y} \left[\langle x,y \rangle > \tau \right] = \Pr \left[e^{t\langle x,y \rangle} > e^{t\tau} \right] \le \frac{\mathbb{E}e^{t\langle x,y \rangle}}{e^{t\tau}} = \frac{\prod_{i=1}^n \mathbb{E}e^{tx_iy_i}}{e^{t\tau}}.$$

The moment generating function of a standard normal random variable is $\mathbb{E}e^{ty} = e^{t^2/2}$, so

$$\mathbb{E}_{x_i,y_i} \left[e^{tx_iy_i} \right] = \mathbb{E}_{x_i} \left[\mathbb{E}_{y_i} e^{tx_iy_i} \right] = \mathbb{E}_{x_i} e^{(t^2/2)x_i^2}.$$

When $x \sim \mathcal{N}(0,1)$, the random variable x^2 has a χ^2 distribution with 1 degree of freedom. The moment generating function of this variable is $\mathbb{E}e^{tx^2} = \sqrt{\frac{1}{1-2t}} = \sqrt{1 + \frac{2t}{1-2t}}$ for any $t < \frac{1}{2}$. Hence,

$$\mathbb{E}_{x_i} e^{(t^2/2)x_i^2} \le \sqrt{1 + \frac{t^2}{1-t^2}} \le e^{\frac{t^2}{2(1-t^2)}}$$

for any $t < 1$. Combining the above results and setting $t = \frac{\tau}{2n}$ yields

$$\Pr_{x,y} \left[\langle x,y \rangle > \tau \right] \le e^{\frac{nt^2}{2(1-t^2)} - t\tau} \le e^{-\frac{\tau^2}{4n}} = \frac{\epsilon^3}{4n}.$$

The same argument shows that $\Pr[\langle x,y \rangle < -\tau] \le \frac{\epsilon^3}{4n}$ as well. □

The reason we consider the truncation g^* is that its smaller ℓ_∞ norm will enable us to apply a strong Bernstein-type inequality on the concentration of measure of the U-statistic estimate of $\mathbb{E}g^*$.

Lemma D.6 (Arcones [3]). *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 $U_m(h)$ be a random variable obtained by drawing $x^1, \ldots, x^m$ independently at random and setting $U_m(h) = \binom{m}{2}^{-1} \sum_{i<j} h(x^i, x^j)$. Then for every $t > 0$,*

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

We are now ready to complete the proof of the upper bound of Theorem 4.1.

Theorem D.7 (Upper bound in Theorem 4.1, restated). *Linear threshold functions can be tested over the standard n-dimensional Gaussian distribution with $O(\sqrt{n \log n})$ queries in both the active and passive testing models.*

Proof. Consider the LTF-TESTER algorithm. When the estimates $\tilde{\mu}$ and $\tilde{\nu}$ satisfy

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

Lemmas D.3 and D.4 guarantee 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,

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

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

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

and

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

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

$$\Sigma^2 \leq \mathbb{E}_y\big[\mathbb{E}_x[f(x)f(y)\langle x, y\rangle]^2\big] = \mathbb{E}_y\Big[\big(\sum_{i=1}^n f(y)y_i\mathbb{E}_x[f(x)x_i]\big)^2\Big] = \sum_{i,j}\hat{f}(e_i)\hat{f}(e_j)\mathbb{E}_y[y_iy_j] = \sum_{i=1}^n \hat{f}(e_i)^2.$$

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

$$\Pr[|\tilde{\nu} - \mathbb{E}g| \leq \epsilon^3] = \Pr[|\tilde{\nu} - \mathbb{E}g^*| \leq \tfrac{1}{2}\epsilon^3] \geq 1 - 4e^{-\frac{mt^2}{8 + 200\sqrt{n\log(4n/\epsilon^3)}t}} \geq \tfrac{11}{12}.$$

The union bound completes the proof of correctness. $\qquad\square$

E Proofs for Testing Disjoint Unions

Theorem 5.1 (Restated). *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.

First to analyze the total number of queries and samples, since we can assume $q(\epsilon) \geq 1/\epsilon$ and $U(\epsilon) \geq 1/\epsilon$, we have $q_\delta(\epsilon')\epsilon'/\epsilon = O(q_\delta(\epsilon/2))$ and $U_\delta(\epsilon')\epsilon'/\epsilon = O(U_\delta(\epsilon/2))$ for $\epsilon' \geq \epsilon/2$. Thus, the total number of queries made is at most

$$\sum_{\epsilon'} q_\delta(\epsilon/2) \log(1/\epsilon) = O\left(q(\epsilon/2) \cdot \log^3 \frac{1}{\epsilon}\right)$$

and the total number of unlabeled samples is at most

$$\sum_{\epsilon'} \frac{8N}{\epsilon} U_\delta(\epsilon/2) \log(1/\epsilon) = O\left(U(\epsilon/2)\frac{N}{\epsilon} \log^3 \frac{1}{\epsilon}\right).$$

Next, to analyze correctness, 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 $I = \{i : p_i \geq \epsilon/(4N)$ 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. $\square$

F Proofs for Testing Dimensions

F.1 Passive Testing Dimension (proof of Theorem 6.2)

Lower bound: By design, $d_{passive}$ is a lower bound on the number of examples needed for passive testing. In particular, if $d_S(\pi, \pi') \leq 1/4$, and if the target is with probability $1/2$ chosen from π and with probability $1/2$ chosen from π', even the Bayes optimal tester will fail to identify the correct distribution with probability $\frac{1}{2}\sum_{y \in \{0,1\}^{|S|}} \min(\pi_S(y), \pi'_S(y)) = \frac{1}{2}(1 - d_S(\pi, \pi')) \geq 3/8$. The definition of $d_{passive}$ implies that there exist $\pi \in \Pi_0$, $\pi' \in \Pi_\epsilon$ such that $\Pr_S(d_S(\pi, \pi') \leq 1/4) \geq 3/4$. Since π' has a $1 - o(1)$ probability mass on functions that are ϵ-far from $\mathcal{P}$, this implies that over random draws of S and f, the overall failure probability of any tester is at least $(1 - o(1))(3/8)(3/4) > 1/4$. Thus, at least $d_{passive} + 1$ random labeled examples are required if we wish to guarantee error at most $1/4$. This in turn implies $\Omega(d_{passive})$ examples are needed to guarantee error at most $1/3$.

Upper bound: We now argue that $O(d_{passive})$ examples are *sufficient* for testing as well. Toward this end, consider the following natural testing game. The adversary chooses a function f such that either $f \in \mathcal{P}$ or $dist_D(f, \mathcal{P}) \geq \epsilon$. The tester picks a function A that maps labeled samples of size k to accept/reject. That is, A is a deterministic passive testing algorithm. The payoff to the tester is the probability that A is correct when S is chosen iid from D and labeled by f.

If $k > d_{passive}$ then (by definition of $d_{passive}$) we know that for any distribution π over $f \in \mathcal{P}$ and any distribution π' over f that are ϵ-far from $\mathcal{P}$, we have $\Pr_{S \sim D^k}(d_S(\pi, \pi') > 1/4) > 1/4$. We now need to translate this into a statement about the value of the game. Note that any mixed strategy of the adversary can be viewed as

$\alpha\pi + (1 - \alpha)\pi'$ for some distribution π over $f \in \mathcal{P}$, some distribution π' over f that are ϵ-far from $\mathcal{P}$ and some $\alpha \geq 0$. The key fact we can use is that against such a mixed strategy, the Bayes optimal predictor has error exactly

$$\sum_y \min(\alpha\pi_S(y), (1 - \alpha)\pi'_S(y)) \leq \max(\alpha, 1 - \alpha) \sum_y \min(\pi_S(y), \pi'_S(y)),$$

while

$$\sum_y \min(\pi_S(y), \pi'_S(y)) = 1 - (1/2) \sum_y |\pi_S(y) - \pi'_S(y)| = 1 - \mathrm{d}_S(\pi, \pi'),$$

so that the Bayes risk is at most $\max(\alpha, 1 - \alpha)(1 - \mathrm{d}_S(\pi, \pi'))$. Thus, for any $\alpha \in [7/16, 9/16]$, if $\mathrm{d}_S(\pi, \pi') > 1/4$, the Bayes risk is less than $(9/16)(3/4) = 27/64$. Furthermore, any $\alpha \notin [7/16, 9/16]$ has Bayes risk at most $7/16$. Thus, since $\mathrm{d}_S(\pi, \pi') > 1/4$ with probability $> 1/4$ (and if $\mathrm{d}_S(\pi, \pi') \leq 1/4$ then the error probability of the Bayes optimal predictor is at most $1/2$), for any mixed strategy of the adversary, the Bayes optimal predictor has risk less than $(1/4)(7/16) + (3/4)(1/2) = 31/64$.

Now, applying the minimax theorem we get that for $k = d_{passive} + 1$, there exists a mixed strategy A for the tester such that for any function chosen by the adversary, the probability the tester is correct is at least $1/2 + \gamma$ for a constant $\gamma > 0$ (namely, $1/64$). We can now boost the correctness probability using a constant-factor larger sample. Specifically, let $m = c \cdot (d_{passive} + 1)$ for some constant c, and consider a sample S of size m. The tester simply partitions the sample S into c pieces, runs A separatately on each piece, and then takes majority vote. This gives us that $O(d_{passive})$ examples are sufficient for testing with any desired constant success probability in $(1/2, 1)$.

F.2 Coarse Active Testing Dimension (proof of Theorem 6.4)

Lower bound: First, we claim that any nonadaptive active testing algorithm that uses $\leq d_{coarse}/c$ label requests must use more than n^c unlabeled examples (and thus no algorithm can succeed using $o(d_{coarse})$ labels). To see this, suppose algorithm A draws n^c unlabeled examples. The number of subsets of size d_{coarse}/c is at most $n^{d_{coarse}}/6$ (for $d_{coarse}/c \geq 3$). So, by definition of d_{coarse} and the union bound, with probability at least $5/6$, *all* such subsets S satisfy the property that $\mathrm{d}_S(\pi, \pi') < 1/4$. Therefore, for any sequence of such label requests, the labels observed will not be sufficient to reliably distinguish π from π'. Adaptive active testers can potentially choose their next point to query based on labels observed so far, but the above immediately implies that even adaptive active testers cannot use an $o(\log(d_{coarse}))$ queries.

Upper bound: For the upper bound, we modify the argument from the passive testing dimension analysis as follows. We are given that for any distribution π over $f \in \mathcal{P}$ and any distribution π' over f that are ϵ-far from $\mathcal{P}$, for $k = d_{coarse} + 1$, we have $\Pr_{S \sim D^k}(\mathrm{d}_S(\pi, \pi') > 1/4) > n^{-k}$. Thus, we can sample $U \sim D^m$ with $m = \Theta(k \cdot n^k)$, and partition U into subsamples $S_1, S_2, \ldots, S_{cn^k}$ of size k each. With high probability, at least one of these subsamples S_i will have $\mathrm{d}_S(\pi, \pi') > 1/4$. We can thus simply examine each subsample, identify one such that $\mathrm{d}_S(\pi, \pi') > 1/4$, and query the points in that sample. As in the proof for the passive bound, this implies that for any strategy for the adversary in the associated testing game, the best response has probability at least $1/2 + \gamma$ of success for some constant $\gamma > 0$. By the minimax theorem, this implies a testing strategy with success probability $1/2 + \gamma$ which can then be boosted to $2/3$. The total number of label requests used in the process is only $O(d_{coarse})$.

Note, however, that this strategy uses a number of unlabeled examples $\Omega(n^{d_{coarse}+1})$. Thus, this only implies an active tester for $d_{coarse} = O(1)$. Nonetheless, combining the upper and lower bounds yields Theorem 6.4.

F.3 Active Testing Dimension (proof of Theorem 6.6)

Lower bound: for a given sample U, we can think of an adaptive active tester as a decision tree, defined based on which example it would request the label of next given that the previous requests have been answered in any given way. A tester making k queries would yield a decision tree of depth k. By definition of $d_{active}(u)$, with probability

at least $3/4$ (over choice of U), any such tester has error probability at least $(1/4)(1 - o(1))$ over the choice of f. Thus, the overall failure probability is at least $(3/4)(1/4)(1 - o(1) > 1/8$.

Upper bound: We again consider the natural testing game. We are given that for any mixed strategy of the adversary with equal probability mass on functions in $\mathcal{P}$ and functions ϵ-far from $\mathcal{P}$, the best response of the tester has expected payoff at least $(1/4)(3/4) + (3/4)(1/2) = 9/16$. This in turn implies that for any mixed strategy at all, the best response of the tester has expected payoff at least $33/64$ (if the adversary puts more than $17/32$ probability mass on either type of function, the tester can just guess that type with expected payoff at least $17/32$, else it gets payoff at least $(1 - 1/16)(9/16) > 33/64$). By the minimax theorem, this implies existence of a randomized strategy for the tester with at least this payoff. We then boost correctness using $c \cdot u$ samples and $c \cdot d_{active}(u)$ queries, running the tester c times on disjoint samples and taking majority vote.

F.4 Lower Bounds for Testing LTFs (proof of Theorem 6.8)

We complete the proofs for the lower bounds on the query complexity for testing linear threshold functions in the active and passive models. This proof has three parts. First, in Section F.4.1, we introduce some preliminary (technical) results that will be used to prove the lower bounds on the passive and coarse dimensions of testing LTFs. In Section F.4.2, we introduce some more preliminary results regarding random matrices that we will use to bound the active dimension of the class. Finally, in Section F.4.3, we put it all together and complete the proof of Theorem 6.8.

The high level idea of our proof is we will show that for a random LTF given by weight vector $\mathbf{w} \sim N(0, I_{n \times n})$, even if we are given the exact value $\mathbf{w} \cdot \mathbf{x}$ for each example $\mathbf{x}$ (rather than just $sgn(\mathbf{w} \cdot \mathbf{x})$), we still could not distinguish these values from random Gaussian noise. Towards this end, for two distributions P, Q over $\mathcal{R}^K$, we use $||P - Q||$ to denote the total variation distance between them. For instance, given two distributions π, π' over boolean functions, and given a sample S, we have $d_S(\pi, \pi') = ||\pi_S - \pi'_S||$.

F.4.1 Preliminaries for $d_{passive}$ and d_{coarse}

Fix any K. Let the dataset $X = \{x_1, x_2, \cdots, x_K\}$ be sampled iid according to a $N(0, I_{n \times n})$ distribution[11]. Let $\mathbf{X} \in \mathcal{R}^{K \times n}$ be the corresponding data matrix.

Suppose $\mathbf{w} \sim \mathcal{N}(0, I_{n \times n})$. We let

$$\mathbf{z} = \mathbf{Xw},$$

and note that the conditional distribution of $\mathbf{z}$ given X is normal with mean 0 and (X-dependent) covariance matrix, which we denote by Σ. Further applying a threshold function to $\mathbf{z}$ gives $\mathbf{y}$ as the predicted label vector of an LTF.

Lemma F.1. *For any square non-singular matrix B, $\log(det(B)) = Tr(\log(B))$, where $\log(B)$ is the matrix logarithm of B.*

Proof. From [38], we know since every eigenvalue of A corresponds to the eigenvalue of $\exp(A)$, thus

$$det(\exp(A)) = \exp(Tr(A)) \tag{2}$$

where $\exp(A)$ is the matrix exponential of A. Taking logarithm of both sides of (2), we get

$$\log(det(\exp(A))) = Tr(A) \tag{3}$$

Let $B = \exp(A)$ (thus $A = \log(B)$). Then (3) can rewritten as $\log(det(B)) = Tr(\log B)$. $\qquad \square$

Fixing X, let $P_{\mathbf{z}/\sqrt{n}|X}$ denote the conditional distribution over $\mathbf{z}/\sqrt{n}$ given by choosing $\mathbf{w} \sim N(0, I_{n \times n})$ and letting $\mathbf{z} = \mathbf{Xw}$.

---

[11]In fact, essentially the same argument would work for many other product distributions, including uniform on $\{-1, +1\}^n$

Lemma F.2. *For sufficiently large n, and a value $K = \Omega(\sqrt{n/\log(K/\delta)})$, with probability at least $1 - \delta$ (over X),*

$$\|\mathbb{P}_{(\mathbf{z}/\sqrt{n})|X} - \mathcal{N}(0, I)\| \leq 1/4.$$

Proof. For sufficiently large n, for any pair $\mathbf{x}_i$ and $\mathbf{x}_j$, by Bernstein's inequality, with probability $1 - \delta'$,

$$\mathbf{x}_i^T \mathbf{x}_j \in \left[-2\sqrt{n \log \frac{2}{\delta'}}, 2\sqrt{n \log \frac{2}{\delta'}} \right]$$

for $i \neq j$, while concentration inequalities for χ^2 random variables [46] imply that with probability $1 - \delta'$,

$$\mathbf{x}_i^T \mathbf{x}_j \in \left[n - 2\sqrt{n \log \frac{2}{\delta'}}, n + 2\sqrt{n \log \frac{2}{\delta'}} + 2\log \frac{2}{\delta} \right]$$

for $i = j$. By the union bound, setting $\delta' = \delta/K^2$, the above inclusions hold simultaneously for all i, j with probability at least $1 - \delta$. For the remainder of the proof we suppose this (probability $1 - \delta$) event occurs.

For $i \neq j$,

$$
\begin{aligned}
Cov(z_i/\sqrt{n}, z_j/\sqrt{n}|X) &= \frac{\mathbb{E}[z_i z_j | X]}{n} \\
&= \frac{1}{n}\mathbb{E}\left[\left(\sum_{l=1}^{n} w_l \cdot x_{il}\right)\left(\sum_{l=1}^{n} w_l \cdot x_{jl}\right) | X \right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{l,m=1,1}^{n,n} w_l w_m x_{il} x_{jm} | X \right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_l w_l^2 x_{il} x_{jl} | X \right] = \frac{1}{n}\mathbb{E}\left[\sum_l x_{il} x_{jl} | X \right] \\
&= \frac{1}{n}\sum_l x_{il} x_{jl} = \frac{1}{n}\mathbf{x}_i^T \mathbf{x}_j \in \left[-2\sqrt{\frac{\log(2K^2/\delta)}{n}}, 2\sqrt{\frac{\log(2K^2/\delta)}{n}} \right]
\end{aligned}
$$

because $\mathbb{E}[w_l w_m] = 0$ (for $l \neq m$) and $\mathbb{E}[w_l^2] = 1$. Similarly, we have

$$Var(z_i/\sqrt{n}|X) = \frac{1}{n}\mathbf{x}_i^T \mathbf{x}_i \in \left[1 - 2\sqrt{\frac{\log \frac{2K^2}{\delta}}{n}}, 1 + 2\sqrt{\frac{\log \frac{2K^2}{\delta}}{n}} + \frac{2\log \frac{2K^2}{\delta}}{n} \right].$$

Let $\beta = 2\sqrt{\frac{\log(2K^2/\delta)}{n}} + \frac{2\log \frac{2K^2}{\delta}}{n}$. Thus Σ is a $K \times K$ matrix, with $\Sigma_{ii} \in [1 - \beta, 1 + \beta]$ for $i = 1, \cdots, K$ and $\Sigma_{ij} \in [-\beta, \beta]$ for all $i \neq j$.

Let $P_1 = \mathcal{N}(0, \Sigma^{K \times K})$ and $P_2 = \mathcal{N}(0, I^{K \times K})$. As the density

$$p_1(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^K \det(\Sigma)}} \exp\left(-\frac{1}{2}\mathbf{z}^T \Sigma^{-1} \mathbf{z}\right)$$

and the density

$$p_2(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^K}} \exp\left(-\frac{1}{2}\mathbf{z}^T \mathbf{z}\right)$$

28

Then L_1 distance between the two distributions P_1 and P_2

$$\int |dP_2 - dP_1| \leq 2\sqrt{K(P_1, P_2)} = 2\sqrt{(1/2)\log\det(\Sigma)},$$

where this last equality is by [23]. By Lemma F.1, $\log(\det(\Sigma)) = Tr(\log(\Sigma))$. Write $A = \Sigma - I$. By the Taylor series

$$\log(I + A) = -\sum_{i=1}^{\infty} \frac{1}{i}(I - (I + A))^i = -\sum_{i=1}^{\infty} \frac{1}{i}(-A)^i$$

Thus,

$$Tr(\log(I + A)) \;=\; \sum_{i=1}^{\infty} \frac{1}{i} Tr((-A)^i). \tag{4}$$

Every entry in A^i can be expressed as a sum of at most K^{i-1} terms, each of which can be expressed as a product of exactly i entries from A. Thus, every entry in A^i is in the range $[-K^{i-1}\beta^i, K^{i-1}\beta^i]$. This means $Tr(A^i) \leq K^i\beta^i$. Therefore, if $K\beta < 1/2$, since $Tr(A) = 0$, the expansion of $Tr(\log(I + A)) \leq \sum_{i=1}^{\infty} K^i\beta^i = O\left(K\sqrt{\frac{\log(K/\delta)}{n}}\right)$.

In particular, for some $K = \Omega(\sqrt{n/\log(K/\delta)})$, $Tr(\log(I + A))$ is bounded by the appropriate constant to obtain the stated result. $\qquad\square$

F.4.2 Preliminaries for d_{active}

Given an $n \times m$ matrix A with real entries $\{a_{i,j}\}_{i\in[n], j\in[m]}$, the *adjoint* (or *transpose* – the two are equivalent since A contains only real values) of A is the $m \times n$ matrix A^* whose (i, j)-th entry equals $a_{j,i}$. Let us write $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ to denote the eigenvalues of $\sqrt{A^*A}$. These values are the *singular values* of A. The matrix A^*A is positive semidefinite, so the singular values of A are all non-negative. We write $\lambda_{\max}(A) = \lambda_1$ and $\lambda_{\min}(A) = \lambda_m$ to represent its largest and smallest singular values. Finally, the *induced norm* (or *operator norm*) of A is

$$\|A\| = \max_{x\in\mathbb{R}^m\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x\in\mathbb{R}^m:\|x\|_2^2=1} \|Ax\|_2.$$

For more details on these definitions, see any standard linear algebra text (e.g., [59]). We will also use the following strong concentration bounds on the singular values of random matrices.

Lemma F.3 (See [64, Cor. 5.35]). *Let A be an $n \times m$ matrix whose entries are independent standard normal random variables. Then for any $t > 0$, the singular values of A satisfy*

$$\sqrt{n} - \sqrt{m} - t \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq \sqrt{n} + \sqrt{m} + t \tag{5}$$

with probability at least $1 - 2e^{-t^2/2}$.

The proof of this lemma follows from Talagrand's inequality and Gordon's Theorem for Gaussian matrices. See [64] for the details. The lemma implies the following corollary which we will use in the proof of our theorem.

Corollary F.4. *Let A be an $n \times m$ matrix whose entries are independent standard normal random variables. For any $0 < t < \sqrt{n} - \sqrt{m}$, the $m \times m$ matrix $\frac{1}{n}A^*A$ satisfies both inequalities*

$$\left\|\frac{1}{n}A^*A - I\right\| \leq 3\frac{\sqrt{m} + t}{\sqrt{n}} \qquad and \qquad \det\left(\frac{1}{n}A^*A\right) \geq e^{-m\left(\frac{(\sqrt{m}+t)^2}{n} + 2\frac{\sqrt{m}+t}{\sqrt{n}}\right)} \tag{6}$$

with probability at least $1 - 2e^{-t^2/2}$.

Proof. When there exists $0 < z < 1$ such that $1 - z \leq \frac{1}{\sqrt{n}}\lambda_{\max}(A) \leq 1 + z$, the identity $\frac{1}{\sqrt{n}}\lambda_{\max}(A) = \|\frac{1}{\sqrt{n}}A\| = \max_{\|x\|_2^2 = 1}\|\frac{1}{\sqrt{n}}Ax\|_2$ implies that

$$1 - 2z \leq (1 - z)^2 \leq \max_{\|x\|_2^2 = 1}\left\|\frac{1}{\sqrt{n}}Ax\right\|_2^2 \leq (1 + z)^2 \leq 1 + 3z.$$

These inequalities and the identity $\|\frac{1}{n}A^*A - I\| = \max_{\|x\|_2^2 = 1}\|\frac{1}{\sqrt{n}}Ax\|_2^2 - 1$ imply that $-2z \leq \|\frac{1}{n}A^*A - I\| \leq 3z$. Fixing $z = \frac{\sqrt{m}+t}{\sqrt{n}}$ and applying Lemma F.3 completes the proof of the first inequality.

Recall that $\lambda_1 \leq \cdots \leq \lambda_m$ are the eigenvalues of $\sqrt{A^*A}$. Then

$$\det(\tfrac{1}{n}A^*A) = \frac{\det(\sqrt{A^*A})^2}{n} = \frac{(\lambda_1 \cdots \lambda_m)^2}{n} \geq \left(\frac{\lambda_1^2}{n}\right)^m = \left(\frac{\lambda_{\min}(A)^2}{n}\right)^m.$$

Lemma F.3 and the elementary inequality $1 + x \leq e^x$ complete the proof of the second inequality. $\square$

F.4.3 Proof of Theorem 6.8

Theorem 6.8 (Restated). *For linear threshold functions under the standard Gaussian distribution in $\mathbb{R}^n$, $d_{passive} = \Omega(\sqrt{n/\log(n)})$ and $d_{active} = \Omega((n/\log(n))^{1/3})$.*

Proof. Let K be as in Lemma F.2 for $\delta = 1/4$. Let $D = \{(x_1, y_1), \ldots, (x_K, y_K)\}$ denote the sequence of labeled data points under the random LTF based on $\mathbf{w}$. Furthermore, let $D' = \{(x_1, y_1'), \ldots, (x_K, y_K')\}$ denote the sequence of labeled data points under a target function that assigns an independent random label to each data point. Also let $\mathbf{z}_i = (1/\sqrt{n})\mathbf{w}^T x_i$, and let $\mathbf{z}' \sim N(0, I_{K \times K})$. Let $E = \{(x_1, \mathbf{z}_1), \ldots, (x_K, \mathbf{z}_K)\}$ and $E' = \{(x_1, \mathbf{z}_1'), \ldots, (x_K, \mathbf{z}_K')\}$. Note that we can think of y_i and y_i' as being functions of $\mathbf{z}_i$ and $\mathbf{z}_i'$, respectively. Thus, letting $X = \{x_1, \ldots, x_K\}$, by Lemma F.2, with probability at least $3/4$,

$$\|\mathbb{P}_{D|X} - \mathbb{P}_{D'|X}\| \leq \|\mathbb{P}_{E|X} - \mathbb{P}_{E'|X}\| \leq 1/4.$$

This suffices for the claim that $d_{passive} = \Omega(K) = \Omega(\sqrt{n/\log(n)})$.

Next we turn to the lower bound on d_{active}. Let us now introduce two distributions $\mathcal{D}_{\text{yes}}$ and $\mathcal{D}_{\text{no}}$ over linear threshold functions and functions that (with high probability) are far from linear threshold functions, respectively. We draw a function f from $\mathcal{D}_{\text{yes}}$ by first drawing a vector $\mathbf{w} \sim \mathcal{N}(0, I_{n \times n})$ from the n-dimensional standard normal distribution. We then define $f : x \mapsto \text{sgn}(\frac{1}{\sqrt{n}}x \cdot \mathbf{w})$. To draw a function g from $\mathcal{D}_{\text{no}}$, we define $g(x) = \text{sgn}(\mathbf{y}_x)$ where each $\mathbf{y}_x$ variable is drawn independently from the standard normal distribution $\mathcal{N}(0, 1)$.

Let $\mathbf{X} \in \mathbb{R}^{n \times q}$ be a random matrix obtained by drawing q vectors from the n-dimensional normal distribution $\mathcal{N}(0, I_{n \times n})$ and setting these vectors to be the columns of $\mathbf{X}$. Equivalently, $\mathbf{X}$ is the random matrix whose entries are independent standard normal variables. When we view $\mathbf{X}$ as a set of q queries to a function $f \sim \mathcal{D}_{\text{yes}}$ or a function $g \sim \mathcal{D}_{\text{no}}$, we get $f(\mathbf{X}) = \text{sgn}(\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{w})$ and $g(\mathbf{X}) = \text{sgn}(\mathbf{y}_{\mathbf{X}})$. Note that $\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{w} \sim \mathcal{N}(0, \frac{1}{n}\mathbf{X}^*\mathbf{X})$ and $\mathbf{y}_{\mathbf{X}} \sim \mathcal{N}(0, I_{q \times q})$. To apply Lemma B.1 it suffices to show that the ratio of the pdfs for both these random variables is bounded by $\frac{6}{5}$ for all but $\frac{1}{5}$ of the probability mass.

The pdf $p : \mathbb{R}^q \to \mathbb{R}$ of a q-dimensional random vector from the distribution $\mathcal{N}_{q \times q}(0, \Sigma)$ is

$$p(x) = (2\pi)^{-\frac{q}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}x^T \Sigma^{-1} x}.$$

Therefore, the ratio function $r : \mathbb{R}^q \to \mathbb{R}$ between the pdfs of $\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{w}$ and of $\mathbf{y}_{\mathbf{X}}$ is

$$r(x) = \det(\tfrac{1}{n}\mathbf{X}^*\mathbf{X})^{-\frac{1}{2}} e^{\frac{1}{2}x^T((\frac{1}{n}\mathbf{X}^*\mathbf{X})^{-1} - I)x}.$$

Note that

$$x^T((\tfrac{1}{n}\mathbf{X}^*\mathbf{X})^{-1} - I)x \le \|(\tfrac{1}{n}\mathbf{X}^*\mathbf{X})^{-1} - I\|\|x\|_2^2 = \|\tfrac{1}{n}\mathbf{X}^*\mathbf{X} - I\|\|x\|_2^2,$$

so by Lemma F.3 with probability at least $1 - 2e^{-t^2/2}$ we have

$$r(x) \le e^{\frac{q}{2}\left(\frac{(\sqrt{q}+t)^2}{n} + 2\frac{\sqrt{q}+t}{\sqrt{n}}\right) + 3\frac{\sqrt{q}+t}{\sqrt{n}}}\|x\|_2^2.$$

By a union bound, for $U \sim \mathcal{N}(0, I_{n\times n})^u$, $u \in \mathbb{N}$ with $u \ge q$, the above inequality for $r(x)$ is true for all subsets of U of size q, with probability at least $1 - u^q 2e^{-t^2/2}$. Fix $q = n^{\frac{1}{3}}/(50(\ln(u))^{\frac{1}{3}})$ and $t = 2\sqrt{q\ln(u)}$. Then $u^q 2e^{-t^2/2} \le 2u^{-q}$, which is $< 1/4$ for any sufficiently large n. When $\|x\|_2^2 \le 3q$ then for large n, $r(x) \le e^{74/625} < \frac{6}{5}$. To complete the proof, it suffices to show that when $x \sim \mathcal{N}(0, I_{q\times q})$, the probability that $\|x\|_2^2 > 3q$ is at most $\frac{1}{5}2^{-q}$. The random variable $\|x\|_2^2$ has a χ^2 distribution with q degrees of freedom and expected value $\mathbb{E}\|x\|_2^2 = \sum_{i=1}^q \mathbb{E}x_i^2 = q$. Standard concentration bounds for χ^2 variables imply that

$$\Pr_{x \sim \mathcal{N}(0, I_{q\times q})}[\|x\|_2^2 > 3q] \le e^{-\frac{4}{3}q} < \tfrac{1}{5}2^{-q},$$

as we wanted to show. Thus, Lemma B.1 implies $\mathrm{err}^*(\mathrm{DT}_q, \mathrm{Fair}(\pi, \pi', U)) > 1/4$ holds whenever this $r(x)$ inequality is satisfied for all subsets of U of size q; we have shown this happens with probabiliity greater than $3/4$, so we must have $d_{active} \ge q$. $\square$

If we are only interested in bounding d_{coarse}, the proof can be somewhat simplified. Specifically, taking $\delta = n^{-K}$ in Lemma F.2 implies that with probability at least $1 - n^{-K}$,

$$\|\mathbb{P}_{D|X} - \mathbb{P}_{D'|X}\| \le \|\mathbb{P}_{E|X} - \mathbb{P}_{E'|X}\| \le 1/4,$$

which suffices for the claim that $d_{coarse} = \Omega(K)$, where $K = \Omega(\sqrt{n/K\log(n)})$: in particular, $d_{coarse} = \Omega((n/\log(n))^{1/3})$.

G Testing Semi-Supervised Learning Assumptions

We now consider testing of common assumptions made in semi-supervised learning [15], where unlabeled data, together with assumptions about how the target function and data distribution relate, are used to constrain the search space. As mentioned in Section 5, one such assumption we can test using our generic disjoint-unions tester is the cluster assumption, that if data lies in N identifiable clusters, then points in the same cluster should have the same label. We can in fact achieve the following tighter bounds:

Theorem G.1. *We can test the cluster assumption with active testing using $O(N/\epsilon)$ unlabeled examples and $O(1/\epsilon)$ queries.*

Proof. Let p_{i1} and p_{i0} denote the probability mass on positive examples and negative examples respectively in cluster i, so $p_{i1} + p_{i0}$ is the total probabilty mass of cluster i. Then $dist(f, \mathcal{P}) = \sum_i \min(p_{i1}, p_{i0})$. Thus, a simple tester is to draw a random example x, draw a random example y from x's cluster, and check if $f(x) = f(y)$. Notice that with probability *exactly* $dist(f, \mathcal{P})$, point x is in the minority class of its own cluster, and conditioned on this event, with probability at least $1/2$, point y will have a different label. It thus suffices to repeat this process $O(1/\epsilon)$ times. One complication is that as stated, this process might require a large *unlabeled* sample, especially if x belongs to a cluster i such that $p_{i0} + p_{i1}$ is small, so that many draws are needed to find a point y in x's cluster. To achieve the given *unlabeled* sample bound, we initially draw an unlabeled sample of size $O(N/\epsilon)$ and simply perform the above test on the uniform distribution U over that sample, with distance parameter $\epsilon/2$. Standard sample complexity bounds [62] imply that $O(N/\epsilon)$ unlabeled points are sufficient so that if $dist_D(f, \mathcal{P}) \ge \epsilon$ then with high probability, $dist_U(f, \mathcal{P}) \ge \epsilon/2$. $\square$

We now consider the property of a function having a large margin with respect to the underlying distribution: that is, the distribution D and target f are such that any point in the support of $D|_{f=1}$ is at distance γ or more from any point in the support of $D|_{f=0}$. This is a common property assumed in graph-based and nearest-neighbor-style semi-supervised learning algorithms [15]. Note that we are not additionally requiring the target to be a linear separator or have any special functional form. For scaling, we assume that points lie in the unit ball in R^d, where we view d as constant and $1/\gamma$ as our asymptotic parameter. Since we are not assuming any specific functional form for the target, the number of labeled examples needed for *learning* could be as large as $\Omega(1/\gamma^d)$ by having a distribution with support over $\Omega(1/\gamma^d)$ points that are all at distance γ from each other (and therefore can be labeled arbitrarily). Furthermore, passive testing would require $\Omega(1/\gamma^{d/2})$ samples as this specific case encodes the cluster-assumption setting with $N = \Omega(1/\gamma^d)$ clusters. We will be able to perform active testing using only $O(1/\epsilon)$ label requests.

First, one distinction between this and other properties we have been discussing is that it is a property of the *relation* between the target function f and the distribution D; i.e., of the combined distribution $D_f = (D, f)$ over labeled examples. As a result, the natural notion of *distance* to this property is in terms of the variation distance of D_f to the closest D_* satisfying the property. As a simple example illustrating the issue, consider $X = [0, 1]$, a target f that is negative on $[0, 1/2)$ and positive on $[1/2, 1]$, and a distribution D that is uniform but where the region $[1/2, 1/2 + \gamma]$ is downweighted to have total probability mass only $1/2^n$. Such a D_f is $1/2^n$-close to the property under variation distance, but would be nearly $1/2$-far from the property if the only operation allowed were to change the function f. A second issue is that we will have to also allow some amount of slack on the γ parameter as well. Specifically, our tester will distinguish the case that D_f indeed has margin γ from the case that the D_f is ϵ-far from having margin γ' where $\gamma' = \gamma(1 - 1/c)$ for some constant $c > 1$; e.g., think of $\gamma' = \gamma/2$. This slack can also be seen to be necessary (see discussion following the proof of Theorem 5.2). In particular, we have the following.

Theorem 5.2 (Restated). *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.*

Proof. First, partition the input space X (the unit ball in R^d) into regions $R_1, R_2, \ldots, R_N$ of diameter at most $\gamma/(2c)$. By a standard volume argument, this can be done using $N = O(1/\gamma^d)$ regions (absorbing "c" into the $O()$). Next, we run the cluster-property tester on these N regions, with distance parameter $\epsilon/4$. Clearly, if the cluster-tester rejects, then we can reject as well. Thus, we may assume below that the total impurity within individual regions is at most $\epsilon/4$.

Now, consider the following weighted graph G_γ. We have N vertices, one for each of the N regions. We have an edge (i, j) between regions R_i and R_j if $\text{diam}(R_i \cup R_j) < \gamma$. We define the *weight* $w(i, j)$ of this edge to be $\min(D[R_i], D[R_j])$ where $D[R]$ is the probability mass in R under distribution D. Notice that if there is no edge between region R_i and R_j, then by the triangle inequality every point in R_i must be at distance at least γ' from every point in R_j. Also, note that each vertex has degree $O(c^d) = O(1)$, so the total weight over all edges is $O(1)$. Finally, note that while algorithmically we do not know the edge weights precisely, we can estimate all edge weights to $\pm\epsilon/(4M)$, where $M = O(N)$ is the total number of edges, using the unlabeled sample size bounds given in the Theorem statement. Let $\tilde{w}(i, j)$ denote the estimated weight of edge (i, j).

Let $E_{witness}$ be the set of edges (i, j) such that one endpoint is majority positive and one is majority negative. Note that if D_f satisfies the γ-margin property, then every edge in $E_{witness}$ has weight 0. On the other hand, if D_f is ϵ-far from the γ'-margin property, then the total weight of edges in $E_{witness}$ is at least $3\epsilon/4$. The reason is that otherwise one could convert D_f to D'_f satisfying the margin condition by zeroing out the probability mass in the lightest endpoint of every edge $(i, j) \in E_{witness}$, and then for each vertex, zeroing out the probability mass of points in the minority label of that vertex. (Then, renormalize to have total probability 1.) The first step moves distance at most $3\epsilon/4$ and the second step moves distance at most $\epsilon/4$ by our assumption of success of the cluster-tester. Finally, if the true total weight of edges in $E_{witness}$ is at least $3\epsilon/4$ then the sum of their estimated weights $\tilde{w}(i, j)$ is at least $\epsilon/2$. This implies we can perform our test as follows. For $O(1/\epsilon)$ steps, do:

1. Choose an edge (i, j) with probability proportional to $\tilde{w}(i, j)$.

2. Request the label for a random $x \in R_i$ and $y \in R_j$. If the two labels disagree, then reject.

If D_f is ϵ-far from the γ'-margin property, then each step has probability $\tilde{w}(E_{witness})/\tilde{w}(E) = O(\epsilon)$ of choosing a witness edge, and conditioned on choosing a witness edge has probability at least $1/2$ of detecting a violation. Thus, overall, we can test using $O(1/\epsilon)$ labeled examples and $O(1/(\gamma^{2d}\epsilon^2))$ unlabeled examples. $\qquad\square$

On the necessity of slack in testing the margin assumption: Consider an instance space $\mathcal{X} = [0,1]^2$ and two distributions over labeled examples D_1 and D_2. Distribution D_1 has probability mass $1/2^{n+1}$ on positive examples at location $(0, i/2^n)$ and negative examples at $(\gamma', i/2^n)$ for each $i = 1, 2, \ldots, 2^n$, for $\gamma' = \gamma(1 - 1/2^{2n})$. Notice that D_1 is $1/2$-far from the γ-margin property because there is a matching between points in the support of $D_1|_{f=1}$ and points in the support of $D_1|_{f=0}$ where the matched points have distance less than γ. On the other hand, for each $i = 1, 2, \ldots, 2^n$, distribution D_2 has probability mass $1/2^n$ at either a positive point $(0, i/2^n)$ *or* a negative point $(\gamma', i/2^n)$, chosen at random, but zero probability mass at the other location. Distribution D_2 satisfies the γ-margin property, and yet D_1 and D_2 cannot be distinguished using a polynomial number of unlabeled examples.