Testing Distributional Assumptions of Learning Algorithms

Ronitt Rubinfeld[*]
ronitt@csail.mit.edu
MIT
Cambridge, Massachusetts, USA

Arsen Vasilyan[†]
vasilyan@mit.edu
MIT
Cambridge, Massachusetts, USA

ABSTRACT

There are many important *high dimensional* function classes that have fast agnostic learning algorithms when strong assumptions on the distribution of examples can be made, such as Gaussianity or uniformity over the domain. But how can one be sufficiently confident that the data indeed satisfies the distributional assumption, so that one can trust in the output quality of the agnostic learning algorithm? We propose a model by which to systematically study the design of *tester-learner pairs* $(\mathcal{A}, \mathcal{T})$, such that if the distribution on examples in the data passes the tester $\mathcal{T}$ then one can *safely trust* the output of the agnostic learner $\mathcal{A}$ on the data.

To demonstrate the power of the model, we apply it to the classical problem of agnostically learning halfspaces under the standard Gaussian distribution and present a tester-learner pair with a combined run-time of $n^{\tilde{O}(1/\epsilon^4)}$. This qualitatively matches that of the best known ordinary agnostic learning algorithms for this task. In contrast, finite sample Gaussian distribution testers do not exist for the L_1 and EMD distance measures. Previously it was known that half-spaces are well-approximated with low-degree polynomials relative to the Gaussian distribution. A key step in our analysis is showing that this is the case even relative to distributions whose low-degree moments approximately match those of a Gaussian.

We also go beyond spherically-symmetric distributions, and give a tester-learner pair for halfspaces under the uniform distribution on $\{0,1\}^n$ with combined run-time of $n^{\tilde{O}(1/\epsilon^4)}$. This is achieved using polynomial approximation theory and critical index machinery of [Diakonikolas, Gopalan, Jaiswal, Servedio, and Viola 2009].

Can one design agnostic learning algorithms under distributional assumptions and count on future technical work to produce, as a matter of course, tester-learner pairs with similar run-time? Our answer is a resounding no, as we show there exist some well-studied settings for which $2^{\tilde{O}(\sqrt{n})}$ run-time agnostic learning algorithms are available, yet the combined run-times of tester-learner pairs must be as high as $2^{\Omega(n)}$. On that account, the design of tester-learner pairs is a research direction in its own right independent of standard agnostic learning. To be specific, our lower bounds apply to the problems of agnostically learning convex sets under the Gaussian distribution and for monotone Boolean functions under the uniform distribution over $\{0,1\}^n$.

CCS CONCEPTS

• **Theory of computation → Machine learning theory**; **Streaming, sublinear and near linear time algorithms**.

KEYWORDS

agnostic learning, distribution testing, learning theory

ACM Reference Format:
Ronitt Rubinfeld and Arsen Vasilyan. 2023. Testing Distributional Assumptions of Learning Algorithms. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC '23), June 20–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3564246.3585117

1 INTRODUCTION

1.1 Motivation

Suppose one wants to learn from i.i.d. example-label pairs, but some unknown fraction of labels are corrupted by an adversary. The well-studied field of *agnostic learning* seeks to develop learning algorithms that are robust to such corruptions[1]. Agnostic learning can be notoriously harder than standard learning (see for example [19, 29, 39, 40]). Nevertheless, there are many important *high dimensional* function classes that do have fast agnostic learning algorithms, including halfspaces, convex sets and monotone Boolean functions. However, these learning algorithms make strong assumptions about the underlying distribution on examples, such as Gaussianity or uniformity over $\{0,1\}^n$.

Thus, to be confident in such a learning algorithm one needs to be confident in the distributional assumptions. In some cases, users can attain confidence in their distributional assumptions by creating their own set of examples which conform to the distribution, and querying labels for these examples. Yet, this approach requires query access, which is often unavailable. Is there a way to ascertain that the examples are indeed coming from a distribution for which the learning algorithm will give a robust answer?

We propose to systematically study the design of *tester-learner pairs* $(\mathcal{A}, \mathcal{T})$, such that *tester* $\mathcal{T}$ *tests the distributional assumptions of agnostic learner* $\mathcal{A}$. In other words, the tester-learner pair is to be designed such that if the distribution on examples in the data pass the tester, then one can *safely use the learner on the data.* By considering the most basic requirements that such a pair ought to satisfy, we propose a new model that makes the following end-to-end requirements on a tester-learner pair $(\mathcal{A}, \mathcal{T})$:

- **Composability:** For any example-label distribution, it should be unlikely that simultaneously (i) the tester $\mathcal{T}$ accepts but

[*] Ronitt Rubinfeld was supported in part by NSF awards CCF-2006664, DMS-2022448 and Fintech@CSAIL.

[†] Arsen Vasilyan was supported in part by NSF awards CCF-1565235, CCF-1955217, DMS-2022448, Big George Fellowship and Fintech@CSAIL.

STOC '23, June 20–23, 2023, Orlando, FL, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9913-5/23/06.
https://doi.org/10.1145/3564246.3585117

[1] See [11] for more on how exactly agnostic learning algorithms yield algorithms that are resilient to adversarial noise in labels.

  (ii) the learner $\mathcal{A}$ outputs something not satisfying the agnostic learning guarantee.
- **Completeness:** If the distribution on examples conforms to the distributional assumption, tester $\mathcal{T}$ will likely accept.
- The performance of the tester-learner pair is judged by the combined run-time of $\mathcal{A}$ and $\mathcal{T}$.

See Section 2.2 for the fully formal definition and see Subsection 1.3 for more comments.

We emphasize that assumptions on the distribution of examples are in fact made in a very large number of works on agnostic learning [2]. Here is an incomplete list of such papers that only scratches the surface: [4, 9, 10, 15–17, 23, 25, 30–32, 38, 41, 43, 44, 49, 52, 57]. Hence, we think it is important to understand to what extent these distributional assumptions can be tested.

Perhaps surprisingly, in spite of how natural this definition is, nothing was previously known on how well it can be achieved for various well-studied problems. The gamut of open possibilities included the most optimistic one: that for all these problems one can test the assumption with very small overhead relative to the existing agnostic learning algorithms. It also included the most pessimistic one: that for all these problems one can test the assumption only at a very steep additional cost in terms of run-time. We note that such steep additional cost would indeed be payed if one were to use existing identity testers of n-dimensional distributions, as these testers have run-times of $2^{\Omega(n)}$ (see below for more information on this).

We commence the charting of the landscape of these possibilities. We find that neither of these extreme possibilities holds in general. On one hand, we find that for some natural problems the most optimistic possibility does materialize and there is a tester-learner pair whose run-time is of the same order as that of the best known agnostic learning algorithm. Specifically, for agnostically learning the class of half-spaces with respect to standard[3] Gaussian distribution, we design a tester-learner pair $(\mathcal{A}, \mathcal{T})$ with combined run-time of $n^{\tilde{O}(1/\epsilon^4)}$. This run-time qualitatively matches the run-time of $n^{\tilde{O}(1/\epsilon^2)}$ [20, 43] achieved by the best algorithm[4] and the statistical query lower bound of $n^{\Omega(1/\epsilon^2)}$ by [24, 26, 34]. We also go beyond spherically-symmetric distributions, and give a tester-learner pair for halfspaces under the uniform distribution on $\{0,1\}^n$ with combined run-time of $n^{\tilde{O}(1/\epsilon^4)}$. For this setting, please see the full version of this work for the precise statement of the theorem and the proof. Here also, the run-time qualitatively matches the run-time of $n^{\tilde{O}(1/\epsilon^2)}$ [20, 43] achieved by the best algorithm. Additionally, we remark that positive results in our framework extend to function classes beyond halfspaces and, as a proof of concept, we give a simple tester-learner pair for agnostically learning decision lists[5] under uniform distribution on $\{0,1\}^n$ (see the full version of this work). Also see [40] for some intractability results on distribution-free learning of decision lists.

On the other hand, for some other natural problems, we show that the most pessimistic scenario holds and the additional requirement of testing the distributional assumption comes at a steep price in terms of run-time. Specifically:

- A well-known algorithm of [49] agnostically learns convex sets under the Gaussian distribution with a run-time of $n^{\tilde{O}(\sqrt{n}/\epsilon^4)}$. We show that if a tester $\mathcal{T}$ tests the distributional assumption of this algorithm, then $\mathcal{T}$ has run-time of $2^{\Omega(n)}$. More generally, *any* tester-learner pair for this task requires $2^{\Omega(n)}$ run-time combined.
- A well-known algorithm of [12, 43] agnostically learns monotone Boolean functions under uniform distribution over $\{0,1\}^n$ with a run-time of $2^{\tilde{O}\left(\frac{\sqrt{n}}{\epsilon^2}\right)}$. We show that if a tester $\mathcal{T}$ tests the distributional assumption of this algorithm, then $\mathcal{T}$ has run-time of $2^{\Omega(n)}$. Again, *any* tester-learner pair for this task requires $2^{\Omega(n)}$ run-time combined.

We emphasize that these lower bounds exhibit natural problems where there is a dramatic gap between standard agnostic learning run-time and the run-time of the best tester-learner pair. Therefore, there is provably no general method that allows one to automatically convert standard agnostic learning algorithms into tester-learner pairs with low run-time overhead. Please see the full version of this work for the precise statements of these intractability results and the proofs.

Additionally, lower bounds for tester-learner pairs can imply lower bounds for standard agnostic learning: Specifically, our lower bounds imply that agnostic learning of monotone functions under distributions $\frac{1}{2^{n^{0.99}}}$-close[6] to $n^{0.99}$-wise independent distributions requires $2^{\Omega(n)}$ run-time. The reason is that by [2, 3, 53] one can test $n^{0.99}$-wise independence up to error $\frac{1}{2^{n^{0.99}}}$ in time $2^{\tilde{O}(n^{0.99})}$, and therefore the existence of such an algorithm would contradict our general lower bound for tester-learner pairs. As there are $2^{\tilde{O}(\sqrt{n}/\epsilon^2)}$ time learners for monotone functions over the uniform distribution [12, 43], this lower bound highlights the sensitivity of agnostic learners to the assumption on the input distribution.

Distribution Testing Perspective. Existing work on identity testing of n-dimensional distributions has focused on testing with respect to very strict distance measures (i.e. TV distance, earth-mover distance, etc.). On one hand this yields strong general-purpose guarantees on distributions accepted by the tester – it is hard to think of a situation where closeness in TV distance is unsatisfactory. On the

[2] The reason for this ubiquity of distributional assumptions in high-dimensional agnostic learning is that with no assumption at all on the distribution the task of agnostic learning is usually intractable. For example (i) The task of learning indicators of convex sets over $\mathbb{R}^n$ cannot be achieved with finite number of samples if nothing is assumed about the distribution. If the distribution is assumed to be Gaussian, this task can be achieved with run-time of $n^{\tilde{O}(\sqrt{n}/\epsilon^4)}$ [49]. (ii) If one is unwilling to make any distributional assumption, no agnostic learning algorithm for halfspaces with run-time of $2^{o(n)}$ is known despite decades of research (also see [19, 29, 39] for some known hardness results). However, as we mentioned if the examples are distributed according to the standard Gaussian, a dramatically faster run-time of $n^{\tilde{O}(1/\epsilon^2)}$ is achievable [20, 43].

[3] Note that the case of Gaussian distribution with arbitrary known mean and covariance reduces to the case of standard Gaussian via a change of coordinates.

[4] However, note that the work of [18] shows how to obtain an even faster run-time of poly $\left(n, \frac{1}{\epsilon}\right)$ if one is willing to settle for a weaker guarantee than the standard agnostic learning guarantee. Specifically, for any absolute constant μ, [18] gives a predictor, such that, if the best halfspace has error opt, the predictor of [18] will have error of at most $(1+\mu)\text{opt} + \epsilon$ (note that standard agnostic learning requires an error bound of $\text{opt} + \epsilon$). In this work we only consider standard agnostic learning.

[5] For this example, a *decision list* is a special case of a decision tree corresponding to a path. More formally, for some ordering of the variables $x_{\pi(1)}, \ldots, x_{\pi(n)}$, values $v_1, \ldots, v_n$ and bits $b_1, \ldots, b_n$, a decision list does the following: For $i = 1$ to n, if $x_{\pi(i)} = b_{\pi(i)}$ output $v_{\pi(i)}$, else continue. A more general definition is given in [55].

[6] In total variation distance.

other hand, in n dimensions this leads to run-times of $2^{\Omega(n)}$. As a concrete example, distinguishing the uniform distribution over $\{0,1\}^n$ from a distribution that is ϵ-far from it in total variation distance requires a run-time of $\Theta\left(\frac{1}{\epsilon^2}2^{n/2}\right)$ (see text [14]).

Yet, run-times of $2^{\Omega(n)}$ can be prohibitive. Indeed, as we explained above, the theory of n-dimensional agnostic learning aims at developing algorithms with run-times of $2^{o(n)}$ or even $n^{O_\epsilon(1)}$. If one were to combine these algorithms with a $2^{\Omega(n)}$-run-time distribution tester, the total run-time would rise precipitously.

From the distribution testing perspective, this work studies *application-targeted testers* that, in favor of much faster run-time, forgo the general-purpose guarantees provided by these strict distance measures. The application domain which this work considers is the testing of distributional assumptions made by agnostic learning algorithms. Here, the application-targeted testers are developed with a view towards special-purpose guarantees sufficient to ensure that the learning algorithms are still robust. For some problems in this domain – this work shows – the use of general-purpose testers can indeed be circumvented, with a dramatic gain in run-time.

In general, surprisingly little is known about such application-targeted testers and we hope more application-targeted distribution testers can be developed for other domains.

Brief Comparison with Distribution-Free Agnostic Learning. Recall that an agnostic learning algorithm is *distribution-free* if it succeeds regardless of the distribution on examples. Designing such algorithms has proven to be intractable for many function classes (see for example [19, 29, 39, 40]). This intractability has prompted the study of agnostic learning algorithms under distributional assumptions.

The model we introduce in this work is intermediate between distribution-free agnostic learning and agnostic learning under a distributional assumption. While the learning algorithm is not required to satisfy the agnostic learning guarantee under every single distribution on example, the testing algorithm needs to alerts us whenever the learning algorithm does fail to satisfy this guarantee.

Incidentally, when using a tester-learner pair, whenever the testing algorithm rejects, the user can choose to then run a slow distribution-free agnostic learning algorithm. Overall, this strategy yields a learning algorithm that always satisfies the agnostic guarantee, and additionally runs fast whenever the distributional assumption does hold, thereby adapting to the distribution on examples.

Recent Followup Work [37]. In an exciting new development we were contacted regarding a follow up work [37] that builds on an earlier version of this paper. [37] develops novel techniques for the design and analysis of tester-learner pairs that leverage connections with the notion of fooling a function class from the field of pseudorandomness. This allows [37] to

- Give tester-learner pairs for more general function classes, such as intersections of halfspaces.
- Handle more general classes of distributional assumptions, such as strictly subexponential distributions in $\mathbb{R}^n$ and uniform over $\{0,1\}^n$.
- Present a new connection between the notion of tester-learner pairs and Rademacher complexity.
- Improve on our run-time for tester-learner pairs for halfspaces under the Gaussian distribution on $\mathbb{R}^n$. Specifically, they give a bound of $n^{\tilde{O}(1/\epsilon^2)}$ which improves upon our bound of $n^{\tilde{O}(1/\epsilon^4)}$. Their tighter bound also matches the known statistical query lower bounds [24, 26, 34].

We would like to note that tester-learner pairs for halfspaces under the uniform distribution on $\{0,1\}^n$ is concurrent work with [37] (they give a faster run-time of $n^{\tilde{O}(1/\epsilon^2)}$ for this problem and also give more general results as explained above). The earlier version of our work (which they build upon) already contained the other results presented in our current version, i.e. (i) the definition of tester-learner pairs (ii) the tester learner pair for half-spaces under the Gaussian distribution with run-time $n^{\tilde{O}(1/\epsilon^4)}$ (Theorem 5) (iii) the intractability results for tester-learner pairs in for the class of convex sets under the Gaussian distribution in $\mathbb{R}^n$ and the class of monotone functions under the uniform distribution in $\{\pm 1\}^n$.

1.2 Our Techniques

1.2.1 Tester-Learner Pair for Agnostically Learning Halfspaces under Gaussian Distribution. We first give an overview of our tester-learner pair $(\mathcal{A},\mathcal{T})$ with combined run-time of $n^{\tilde{O}(1/\epsilon^4)}$ for the class of half-spaces with respect to standard Gaussian distribution. We also discuss the techniques we use to analyze it. See Sections 3, 5 and 6 for complete details.

A natural first approach would be to try to take advantage of the literature on testing and learning distributions. However, almost all results we are aware of on testing and learning high-dimensional distributions (without assuming the distribution already belongs to some highly restricted family as in [13]) require a number of samples that is exponentially large in the dimension. It follows from well-known techniques that Gaussianity over an infinite domain cannot be tested with respect to total variation distance in finite samples. Potentially, one could obtain a tester-learner pair for Gaussianity with respect to the earth-mover distance via the tester[7] of [5], yielding a tester of run-time $2^{\tilde{O}(n)}$. However one can see that, in earth-mover distance, no significantly better (i.e. $2^{o(n)}$) bound can be obtained[8]. Such enormous run-times far exceed the run-times that can be achieved for agnostically learning halfspaces.

Previously it was known that half-spaces are well-approximated with low-degree polynomials relative to the Gaussian distribution. A key step in our analysis is showing that this is the case even relative to distributions whose low-degree moments approximately match those of a Gaussian. One of our ideas is to start with a proof of the exact Gaussian case and modify it so it only relies on low-degree properties of the distribution. We are aware of three distinct proofs of this exact Gaussian case in the literature:

(1) The method of [43] that uses specific facts about Hermite polynomials.
(2) The noise sensitivity method of [49]. This method also uses Hermite polynomials to argue that functions that tend to

[7] This tester requires that the distribution is confined to a box $[-B,B]^n$, but this by itself is not a devastating problem, since most of probability mass of a Gaussian is confined to such a box.

[8] Even when truncating the distribution to a box around the origin.

be stable to perturbations of their input tend to be well-approximated by low-degree polynomials.

(3) The method of [20] that, in order to approximate a halfspace $\mathrm{sign}(\boldsymbol{v}\cdot\boldsymbol{x}+\theta)$, constructs a polynomial $P(\boldsymbol{v}\cdot\boldsymbol{x})$ that approximates this halfspace tightly for values of $|\boldsymbol{v}\cdot\boldsymbol{x}|$ that are not too large. It is then argued that large values of $|\boldsymbol{v}\cdot\boldsymbol{x}|$ do not contribute much to the total L_1 error of the polynomial because its contribution is weighted by a rapidly decaying Gaussian weight.

As Hermite polynomials are the unique family of polynomials orthogonal under the Gaussian distribution, the proof strategies of [43] and [49] seem highly specialized to the distribution being exactly Gaussian. Because of this, a method similar to the one of [20] is the one serving as our starting point.

This method needs to be modified in a thoroughgoing way in order to rely merely on the low-degree moments of the distribution being close to those of Gaussian. For instance, a very easy-to-show property of the n-dimensional standard Gaussian distribution is its anti-concentration when projected on any direction. This property becomes much less obvious once one is only promised that low-degree moments of the distribution are close to those of Gaussian, which is something we do show. We note that this step of our proof is similar in spirit to the work of [45] that introduces a notion of low-degree certified anti-concentration and shows it for various distributions. Our proofs use extensively tools from polynomial approximation theory.

Given these ideas, our tester-learner pair does the following. The tester estimates the low-degree moments of the distribution and compares them to the corresponding moments of the standard Gaussian. It follows then that halfspaces are well-approximated by low-degree polynomials with respect to this distribution. The learning algorithm takes advantage of this by performing low-degree polynomial L_1 regression similar to the one used in [43].

A technical complication, which we deal with, is that both our tester and learner work with a truncated version of the distribution. In other words, they discard the examples whose coordinates are too large. This guarantees to us that we can actually produce estimates for the moments of the truncated distribution (if distribution is not truncated, moments could even be infinite).

Note that our arguments use strongly the fact that we are working with halfspaces and not with some arbitrary function class that is well-approximated by low-degree polynomials under the Gaussian distribution. This is due to how we use the concentration and anti-concentration properties of the distribution. In a certain sense this is necessary, as shown by our intractability results for indicators of convex sets. Even though these functions are also well-approximated by low-degree polynomials [49], for them a similar method based on estimating low-degree moments will provably not succeed. This underscores that designing tester-learner pairs can be subtle and does not generally follow by mere extension of already existing analyses of agnostic learning algorithms.

1.2.2 Tester-Learner Pair for Agnostically Learning Halfspaces under Uniform Distribution on $\{\pm 1\}^n$.

We now discuss the techniques used to give our tester-learner pair for halfspaces under the uniform distribution on $\{\pm 1\}^n$. As we mentioned, the run-time we show here is $n^{\tilde{O}(1/\epsilon^4)}$ and this is concurrent work with [37], who use other techniques. See the full version of this work for complete details.

Our tester tests $\mathrm{poly}(1/\epsilon)$-wise independence of the input distribution with respect to the TV distance using [2, 3, 53]. The learning algorithm uses the low-degree polynomial L_1 regression of [43]. To show that these two algorithms indeed form a valid tester-learner pair we show that every halfspace is well-approximated by a low-degree polynomial relative to any $\mathrm{poly}(1/\epsilon)$-wise independent distribution.

Suppose for a halfspace $\mathrm{sign}(\boldsymbol{v}\cdot\boldsymbol{x}+\theta)$ it is the case that the norm of the vector $\boldsymbol{v}$ is well-distributed among all the coordinates. Then, by Berry-Esseen theorem, for $\boldsymbol{x}$ that is uniform over $\{\pm 1\}^n$ the inner product $\boldsymbol{v}\cdot\boldsymbol{x}$ is distributed similarly to a Gaussian. Roughly, we use this to argue that if $\boldsymbol{x}$ is merely $\mathrm{poly}(1/\epsilon)$-wise independent then $\boldsymbol{v}\cdot\boldsymbol{x}$ has low-degree moments close to those of a Gaussian. This allows us to use methods similar to the ones we use to give tester-learner pairs for halfspaces under the standard Gaussian distribution.

Finally, we handle halfspaces $\mathrm{sign}(\boldsymbol{v}\cdot\boldsymbol{x}+\theta)$ for whom the norm of the vector $\boldsymbol{v}$ is not well-spread across all the coordinates. We use the *critical index* machinery of [20] to handle such halfspaces.

1.2.3 Intractability Results.

Finally, we discuss the techniques used to show that $2^{\Omega(n)}$ samples are required by (i) any tester-learner pair for learning indicator functions of convex sets under the standard Gaussian on $\mathbb{R}^n$ (ii) any tester-learner pair for learning monotone functions under the uniform distribution on $\{0,1\}^n$. See See the full version of this work for complete details.

From technical standpoint, we find these lower bounds surprising: The mentioned standard agnostic learning algorithms in these settings rely on low-degree polynomial regression. This suggests that testing low-degree moments of the distribution (as we did for halfspaces) ought to lead to the development of a fast tester-learner pair. Yet, the lower bounds show that this can not be done.

We now roughly explain how we prove these lower bounds. Let us focus on the lower bound for tester-learner pairs for convex sets under standard Gaussian distribution (the lower bound for monotone functions is similar). Take samples $z_1,\cdots,z_M$ from the standard Gaussian, and let D be the uniform distribution on $\{z_1,\cdots,z_M\}$. The first idea is to show that the tester will have a hard time distinguishing D from the standard Gaussian if it uses much fewer than M samples[9]. The second idea is to show that (very likely over the choice of $z_1,\cdots,z_M$) one can obtain, by excluding only a small fraction of elements from $\{z_1,\cdots,z_M\}$, a subset Q of them such that no point in Q is in the convex hull of the other points in Q. Once we have such a set, we essentially[10] define our hard-to-learn convex set to be the convex hull of a random subset of Q, and this convex set will not contain any other elements of Q because no member of Q is in the convex hull of the rest. In this way, unless a learner has seen a large fraction of the elements in Q already, it has no way of predicting whether a previously unseen element in Q belongs to the random convex set. We note that our

[9] Out actual argument also takes into account that the tester sees labels and not only examples.

[10] This is an oversimplification, as one still needs to figure out what to do with elements outside Q. We show that, for all these elements, we can either include them into or exclude them from the convex set in such a way as to reveal no information about which of the points in Q were included in the convex set.

argument is somewhat similar to well-known arguments proving impossibility of approximation of the volume of a convex set via a deterministic algorithm [6, 28].

1.3 Comments on the Model

1.3.1 What about Cross-Validation? In case of realizable learning (i.e. you are promised there is no noise) a common approach to verifying success is via checking prediction error rate on fresh data and making sure it is not too high. Does this idea allow one to construct a tester $\mathcal{T}$ for the distributional assumption of some agnostic learner $\mathcal{A}$? Such tester would (i) run $\mathcal{A}$ to obtain a predictor $\hat{f}$ (ii) test the success rate of $\hat{f}$ on fresh example-label pairs (iii) accept or reject based on the success rate.

As was mentioned in the discussion of our intractability results, there cannot be a general low-overhead method of transforming standard agnostic learning algorithms into tester-learner pairs, because of our intractability results. Therefore, in particular, there cannot be such a method based on cross-validation.

Intuitively, the reason is the following. Suppose you run the learning algorithm, setting the closeness parameter ϵ to 0.01, then check the success of the predictor on fresh data and find that the generalization error is close to 0.25. This could potentially be consistent with the two following situations: (1) there is a function in the concept class with close to zero generalization error, but the learning algorithm gave a poor predictor due to a violation of the distributional assumption (2) the distributional assumption holds, but every function in the concept class has generalization error of at least 0.24. The *composability* criterion tells you that in case (1) you should reject, but the *completeness* criterion tells you that in case (2) you should accept. Overall, there is no way to tell from generalization error alone which of the two situations you are in, so there is no way to know if you should accept or reject.

1.3.2 Label-Aware vs Label-Oblivious Testers. We say the tester $\mathcal{T}$ is *label-aware* if it makes use of the labels given to it (and not only the examples). Otherwise, we call it *label-oblivious*. We feel that label-obliviousness makes a testing algorithm fit better with the existing literature on testing properties of distributions, because algorithms in this line of work decide to accept or reject a distribution based only on samples from it (and no side information such as labels). However, this condition is not strictly necessary for verifying success. Due to these considerations, our impossibility results are against more general label-aware testers, while the tester given in this paper is label-oblivious.

1.4 Related Work

1.4.1 Agnostic Learning under Distributional Assumptions Using Low-Degree Polynomial Regression. Since the introduction of the agnostic learning model [42, 46] there has been an explosion of work in agnostic learning. Making assumptions on the distribution on examples has been ubiquitous in this line of work. So has been the use of low-degree polynomial regression as one of the main tools. Previous to the work of [43], there existed an extensive body of work on using low-degree polynomial regression for learning under distributional assumptions, including [1, 12, 33, 48, 50, 51]. The work of [43] building on [46] proposed to use low-degree polynomial L^1 regression to obtain *agnostic* learning algorithms for halfspaces under distribution assumptions, as well as extended these previously studied low-degree regression algorithms into the agnostic setting. Further work used low degree polynomial L^1 regression to obtain agnostic learning algorithms for many more problems, again under various distributional assumptions [4, 9, 10, 15–17, 23, 25, 30–32, 38, 41, 44, 49, 52, 57].

1.4.2 Learning Halfspaces. See the work of [25] and references therein, for a historical discussion about the problem of learning halfspaces, as well as some up-to-date references regarding some problems connected to the one studied here.

1.4.3 Polynomial Approximation Theory. Polynomial approximation theory has been used extensively as a tool for studying halfspaces. Among other work, see [18, 20, 25, 27, 43, 47].

1.4.4 Other Works in Testing Distributions. There is a large body of literature on finite sample guarantees for property testing of distributions. Algorithms developed within this framework are given samples of an input distribution and aim to distinguish the case in which the distribution has a specified property, from the case in which the distribution is far (in a reasonable distance metric) from any distribution with that property. Properties of interest include whether the distribution is uniform, independent, monotone, has high entropy or is supported by a large number of distinct elements. We mention a few specific results that are closest to the results in this work: Let p be a distribution on a discrete domain of size M. For a “known" distribution q (where the algorithm knows the value of q on every element of the domain, and does not need samples from it – e.g., when q is the uniform distribution), distinguishing whether p is the same as q from the case where p is ϵ-far (in L_1 norm) from q requires $\Theta(\sqrt{M}/\epsilon^2)$ samples [7, 8, 21, 22, 35, 54]. For a more in depth discussion of the history and results in this area, see the monograph by Canonne [14].

1.4.5 Other Models of Trusting Agnostic Learners. The work of Goldwasser, Rothblum, Shafer and Yehudayoff considers the question of how an untrusted prover can convince a learner that a hypothesis is approximately correct, and show that significantly less data is needed than that required for agnostic learning [36].

2 PRELIMINARIES

2.1 Standard Definitions

The definition of agnostic learning is as follows:

Definition 1. An algorithm $\mathcal{A}$ is an *agnostic* (ϵ, δ)*-learning algorithm* for function class $\mathcal{F}$ relative to the distribution D, if given access to i.i.d. example-label pairs (x, y) distributed according to D_{pairs}, with the marginal distribution on the examples equal to D, the algorithm $\mathcal{A}$ with probability at least $1 - \delta$ outputs a circuit computing a function $\hat{f}$, such that

$$\Pr_{(x,y)\in_R D_{\text{pairs}}}[y \neq \hat{f}(x)] \leq \min_{f\in\mathcal{F}}\left(\Pr_{(x,y)\in_R D_{\text{pairs}}}[f(x) \neq y]\right) + \epsilon.$$

The quantity $\Pr_{(x,y)\in_R D_{\text{pairs}}}[f(x) \neq y]$ is often called the *generalization error* of $\hat{f}$ (a.k.a. *out-of-sample error* or *risk*).

The following is standard theorem about agnostic learning from ℓ_1-approximation. The proof is implicit in [43] and this theorem has been implicitly used in much subsequent work (see Subsection 1.4 for references). Let U be some domain we are working over.

Theorem 2. *Let $\{g_1, \cdots g_N\}$ be a collection of real-valued functions over U that can be evaluated in time T. Then, for every $\epsilon > 0$, there is a learning algorithm $\mathcal{A}$ for which the following is true. Let D be any distribution over U and let $\mathcal{F}$ be any class of Boolean functions over U, such that every element of $\mathcal{F}$ is ϵ-approximated in L^1 norm relative to the distribution D by some element of span $(g_1, \cdots, g_N)$. Then, $\mathcal{A}$ agnostically (ϵ, δ)-learns $\mathcal{F}$ relative to D. The algorithm $\mathcal{A}$ uses $\tilde{O}\left(\frac{N}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ samples and uses run-time polynomial in this number of samples and T.*

We will also need the definition of k-wise independent distributions:

Definition 3. A distribution of a random variable $\boldsymbol{x}$ over $\{\pm 1\}^n$ is called k-wise independent (a.k.a. k-wise uniform) if for any size-k subset S of $\{1, \cdots, n\}$ the distribution of $\{x_i : i \in S\}$ is uniform over $\{\pm 1\}^k$.

2.2 New Definition: Testing Distributional Assumptions of a Learning Algorithm

Definition 4. Let $\mathcal{A}$ be an agnostic (ϵ, δ_1)-learning algorithm for function class $\mathcal{F}$ relative to the distribution D. We say that an algorithm $\mathcal{T}$ is a *tester for the distributional assumption* of $\mathcal{A}$ if

(1) *(Composability)* Suppose a distribution D_{pairs} on example-label pairs is such that, given access to i.i.d. labeled examples from it, the algorithm $\mathcal{T}$ outputs "Yes" with probability at least 1/4. Then $\mathcal{A}$, given access to i.i.d. labeled examples from the same distribution D_{pairs}, will with probability at least $1 - \delta_1$ output a circuit computing a function $\hat{f}$, such that

$$\Pr_{(x,y)\in_R D_{\text{pairs}}}[y \neq \hat{f}(x)] \leq \min_{f\in\mathcal{F}}\left(\Pr_{(x,y)\in_R D_{\text{pairs}}}[f(x) \neq y]\right) + \epsilon.$$

(2) *(Completeness)* Suppose D_{pairs} is such that the marginal distribution on examples equals to D. Then, given i.i.d. example-label pairs from D_{pairs}, tester $\mathcal{T}$ outputs "Yes" with probability at least 3/4.

If this definition is satisfied, then we say that ($\mathcal{A}$,$\mathcal{T}$) form a tester-learner pair.

Constants 1/4 and 3/4 in the definition above can without loss of generality be replaced with any other pair of constants $1 - \delta_2$ and $1 - \delta_3$ with $\delta_2 \in (0, 1)$ and $\delta_3 \in (\delta_2, 1)$. See the full version of this work for the proof via a standard repetition argument.

3 AN EFFICIENT TESTER-LEARNER PAIR FOR LEARNING HALFSPACES

We now describe our tester-learner pair for learning halfspaces under the Gaussian distribution. Roughly, the testing algorithm checks that the low-degree moments of the distribution on examples are close enough to those of the standard Gaussian distribution. The learning algorithm uses a low-degree polynomial regression (similarly to [43]).

As explained earlier, both of the algorithms ignore examples whose absolute value is too high, which allows them to obtain accurate estimates of distribution moments.

Tester-learner pair for learning halfspaces:

- Let $C_1, \cdots, C_4$ be a collection of constants to be tuned appropriately. Define $d := 2\left\lfloor \frac{1}{2\epsilon^4}\ln^3\left(\frac{1}{\epsilon}\right)\right\rfloor$, $\Delta := \left\lfloor \frac{1}{\epsilon^4}\ln^4\left(\frac{1}{\epsilon}\right)\right\rfloor$, $t := C_1\Delta\ln\Delta\sqrt{\log n} + \sqrt{2\ln\left(\frac{C_2 n}{\epsilon}\right)}$, $N_1 := \left\lceil n^{C_3 d}\right\rceil$ and $N_2 := \left\lceil t^{2\Delta} n^{C_4\Delta}\right\rceil$.
- **Learning algorithm $\mathcal{A}$.** Given access to i.i.d. labeled samples $(\boldsymbol{x}, y) \in \mathbb{R}^n \times \{\pm 1\}$ from an unknown distribution:
  (1) Obtain N_1 many labeled samples $(\boldsymbol{x}_i, y_i)$.
  (2) Discard all the samples $(\boldsymbol{x}_i, y_i)$ for which the absolute value of some coordinate $\left|(\boldsymbol{x}_i)_j\right|$ is greater than t.
  (3) Run the algorithm of Theorem 2 on the remaining samples, with accuracy parameter $\frac{\epsilon}{10}$, allowed failure probability $\frac{1}{20}$, and taking the set of $\{g_i\}$ to be the set of monomials of degree at most d, i.e. the set $\left\{\prod_{j=1}^n x_j^{\alpha_j} : \sum_j \alpha_j \leq d\right\}$. This gives us a circuit computing predictor $\hat{f}$. Form a new predictor $\hat{f}'$ that given $\boldsymbol{x}$ outputs (i) $\hat{f}(\boldsymbol{x})$ if for all $j \in [n]$, the value of $\left|(\boldsymbol{x}_i)_j\right|$ is at most t. (ii) 1 if[11] for some $j \in [n]$, the value of $\left|(\boldsymbol{x}_i)_j\right|$ exceeds t.
- **Testing algorithm $\mathcal{T}$.** Given access to i.i.d. labeled samples $\boldsymbol{x} \in \mathbb{R}^n$ from an unknown distribution:
  (1) For each $j \in [n]$:
    (a) Estimate $\Pr\left[\left|x_j\right| > t\right]$ up to additive $\frac{\epsilon}{30n}$ with error probability $\frac{1}{100n}$.
    (b) If the estimate is at least $\frac{\epsilon}{10n}$, output **No** and terminate.
  (2) Draw N_2 fresh samples $\{\boldsymbol{x}_i\}$, and discard the ones for which the absolute value of some coordinate $\left|(\boldsymbol{x}_i)_j\right|$ is greater than t.
  (3) For every monomial $\prod_{j=1}^n x_j^{\alpha_j}$ of degree at most Δ, compute its empirical expectation w.r.t. the samples $\{\boldsymbol{x}_i\}$. If for any of them resulting value is not within $\frac{1}{2n^\Delta}$ of

$$\mathbb{E}_{z\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{j=1}^n x_j^{\alpha_j}\right] = \prod_{j=1}^n\left((\alpha_j - 1)!! \cdot \mathbb{1}_{\alpha_j \text{ is even}}\right),$$

  output **No** and terminate.
  (4) Output **Yes.**

The following theorem shows that the above algorithms indeed satisfy the criteria for a tester-learner pair for learning halfspaces under the Gaussian distribution:

Theorem 5 (Tester-learner pair for learning halfspaces under Gaussian distribution). *Suppose the values $C_1, \cdots, C_4$ present in algorithms $\mathcal{A}$ and $\mathcal{T}$ are chosen to be sufficiently large absolute constants, also assume n and $\frac{1}{\epsilon}$ are larger than some sufficiently large absolute constant. Then, the algorithm $\mathcal{A}$ is an agnostic $(O(\epsilon), 0.1)$-learner for the function class of linear threshold functions over $\mathbb{R}^n$*

[11] This one's arbitrary. Can also output 0 in this case.

under distribution $\mathcal{N}(0, I_{n\times n})$ and the algorithm $\mathcal{T}$ is an assumption tester for $\mathcal{A}$. The algorithms $\mathcal{A}$ and $\mathcal{T}$ both require only $n^{\tilde{O}\left(\frac{1}{\epsilon^4}\right)}$ samples and run-time. Additionally, The tester $\mathcal{T}$ is label-oblivious.

Note that an $(O(\epsilon), 0.1)$-learner can be made an agnostic (ϵ, δ_1)-learner for any fixed constant δ_1 and still require only $n^{\tilde{O}\left(\frac{1}{\epsilon^4}\right)}$ samples and run-time via a standard repeat-and-check argument. The tester $\mathcal{T}$ for the original learner will remain an assumption tester for the new learner.

The proof of correctness of the above tester-learner pair for halfspaces makes use of the following lemmas, which will be proved in Section 5. Lemma 6 states that as long as the low-degree moments of a distribution are similar to the corresponding moments of the Gaussian distribution, then the distribution is concentrated and anti-concentrated when projected onto any direction. Lemma 7 states that as long as distribution D satisfies the "nice" properties of concentration and anti-concentration, then any halfspace can be approximated by a low-degree polynomial with respect to distribution D. Taken together, these lemmas will be used to show that for any distribution D, if the moments of D look similar to moments of the Gaussian distribution, then halfspaces are well-approximated by low degree polynomials under D.

Lemma 6 (*Low degree moment lemma for distributions.*)**.** *Suppose D is a distribution over $\mathbb{R}^n$ and Δ is an even positive integer, such that for every monomial $\prod_{i=1}^n x_i^{\alpha_i}$ of degree at most Δ we have*

$$\left|\mathbb{E}_{\boldsymbol{x}\sim D}\left[\prod_{i=1}^n x_i^{\alpha_i}\right] - \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{i=1}^n x_i^{\alpha_i}\right]\right| \leq \frac{1}{n^{\Delta}}.$$

Further, assume that $\Delta \geq \frac{1}{\epsilon^4}\ln^4\left(\frac{1}{\epsilon}\right)$. Then, for every unit vector $\boldsymbol{v}$, the random variable $\boldsymbol{v}\cdot\boldsymbol{x}$ (with $\boldsymbol{x} \in_R D$) has the following properties

- ***Concentration:*** *For any even positive integer $d \leq \Delta$, we have $\left(\mathbb{E}_{\boldsymbol{x}\in_R D}\left[|\boldsymbol{v}\cdot\boldsymbol{x}|^d\right]\right)^{1/d} \leq 2\sqrt{d}$.*
- ***Anti-concentration:*** *for any real y, we have*
$$\Pr_{\boldsymbol{x}\in_R D}\left[\boldsymbol{v}\cdot\boldsymbol{x} \in [y, y+\epsilon]\right] \leq O(\epsilon).$$

Lemma 7 (*Low degree approximation lemma for halfspaces.*)**.** *Suppose D is a distribution on $\mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^n$ is a unit vector, such that for some positive real parameters α, γ, ϵ and a positive integer parameter d_0 we have*

- ***Anti-concentration:*** *for any real y, we have*
$$\Pr_{\boldsymbol{x}\in_R D}\left[\boldsymbol{v}\cdot\boldsymbol{x} \in [y, y+\epsilon]\right] \leq \alpha,$$
- ***Concentration:***
$$\left(\mathbb{E}_{\boldsymbol{x}\in_R D}\left[|\boldsymbol{v}\cdot\boldsymbol{x}|^{d_0}\right]\right)^{1/d_0} \leq \beta,$$
for some $\beta \geq 1$.

Also assume $d_0 > \frac{5\beta}{\epsilon^2}$ and that ϵ is smaller than some sufficiently small absolute constant. Then, for every $\theta \in \mathbb{R}$ and there is a polynomial $P(x)$ of degree at most $\frac{2\beta}{\epsilon^2} + 1$ such that

$$E_{\boldsymbol{x}\in_R D}\left[|P(\boldsymbol{v}\cdot\boldsymbol{x}) - sign(\boldsymbol{v}\cdot\boldsymbol{x} - \theta)|\right] = O\left(\alpha + \epsilon + \frac{(8\beta)^{\frac{2\beta}{\epsilon^2}+1}}{2^{d_0}}\right).$$

Each coefficient of the polynomial P has magnitude of at most $O\left(2^{\frac{4\beta}{\epsilon^2}}\right)$.

4 TECHNICAL PRELIMINARIES

4.1 Polynomial Approximation Theory

We will need some standard facts about Chebychev polynomials and approximation of functions using them. See, for example, the text [56] for comprehensive treatment of this topic. First, we define Chebychev polynomials and present relevant facts about them. On the interval $[-1, 1]$ the k-th Chebychev polynomial can be defined as[12] $T_k(x) := \cos(k \arccos(x))$.

For any $k \geq 0$, the polynomial $T_k(x)$ maps $[-1, 1]$ to $[-1, 1]$ (this follows immediately from the definition). Also, it is known that the Chebyshev polynomials satisfy a recurrence relation

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x),$$

with the first two polynomials being $T_0(x) = 1$ and $T_1(x) = x$.

To present a standard theorem from text [56] about approximating functions with Chebyshev polynomials, we will need the standard notions of Lipschitz continuity and of bounded variation functions. A function f is said to be Lipschitz continuous on $[-1, 1]$ if there is some C so for any $x, y \in [-1, 1]$ we have that $|f(x) - f(y)| \leq C|x - y|$. For a differentiable function $f : [-w, w] \to \mathbb{R}$, the *total variation of* f is the L_1 norm of it's derivative, i.e.

$$\int_{-w}^{w}\left|\frac{df(x)}{dx}\right| dx.$$

If f has a single discontinuity at some point a and is differentiable everywhere else, then the total variation of f is defined as the sum of the following three terms (i) $\int_{-w}^{a}\left|\frac{df(x)}{dx}\right| dx$, (ii) the magnitude of the discontinuity at a and (iii) $\int_{a}^{w}\left|\frac{df(x)}{dx}\right| dx$. Analogously, the definition extends to functions that are differentiable outside of finitely many discontinuities[13]. We say "f is of bounded variation V" if the total variation of f is at most V.

We are now ready to state the following theorem about approximating functions using Chebyshev polynomials:

Theorem 8 (Consequence of Theorem 7.2 in the text [56] (see also Theorem 3.1 on page 19 in the text [56])). *Let f be Lipschitz continuous on $[-1, 1]$ and suppose the derivative f' is of bounded variation V. Define for $k \geq 0$*

$$a_k := \frac{1 + \mathbb{1}_{k>0}}{\pi}\int_{-1}^{1}\frac{f(x)T_k(x)}{\sqrt{1-x^2}}\,dx.$$

Then, for any $d \geq 0$ we have

$$\max_{x\in[-1,1]}\left|f(x) - \sum_{k=0}^{d} a_k T_k(x)\right| = O\left(\frac{V}{d}\right).$$

The partial sums $\sum_{k=0}^{d} a_k T_k$ are called Chebyshev projections.

[12] One needs to check that $\cos(k\alpha)$ is indeed a polynomial in $\cos\alpha$, which follows by writing $\cos(k\alpha) = \frac{e^{ik\alpha}+e^{-ik\alpha}}{2} = \frac{1}{2}\left((\cos\alpha + i\sin\alpha)^k + (\cos\alpha - i\sin\alpha)^k\right)$, expanding, observing that terms involving odd powers of $\sin\alpha$ cancel out, and using the identity $\sin^2\alpha = 1 - \cos^2\alpha$.

[13] It is also standard to consider more general functions, but we will not need that.

5 PROVING THE TWO MAIN LEMMAS (6,7) VIA POLYNOMIAL APPROXIMATION THEORY

5.1 Propositions Useful for Proving Both Main Lemmas

Here we will present proposition that will be useful for proving both Lemma 6 and 7. We start with an observation that bounds the magnitude of the coefficients of Chebyshev polynomials.

Observation 9. *Let $f : \mathbb{R} \to [-1,1]$ be a Lipschitz continuous function. Let $d \geq 1$ be an integer, let $w \geq 1$ be a real number, and let $f_d(x) := \sum_{k=0}^{d} a_k T_k(\frac{x}{w})$, where $a_k := \frac{1+\mathbb{1}_{k>0}}{\pi}\int_{-1}^{1}\frac{f(wy)T_k(y)}{\sqrt{1-y^2}}\,dy$. Then, the largest coefficient from among all the monomials of $f_d(x)$ has value of at most $O\left(d3^d\right)$.*

Proof. See the full version of this work for the proof. □

Proving both lemmas, we will be approximating certain functions using Chebyshev polynomials re-scaled to the window $[-w,w]$. The following proposition lets us bound the error between function f and its low-degree polynomial approximation, contributed by the region $(-\infty,w)\cup(w,+\infty)$.

Proposition 10. *Let f be a Lipschitz continuous function $\mathbb{R} \to [-1,1]$. Let $d \geq 1$ be an integer and $w \geq 1$ be real-valued, and let $f_d(x) := \sum_{k=0}^{d} a_k T_k(\frac{x}{w})$, where $a_k := \frac{1+\mathbb{1}_{k>0}}{\pi}\int_{-1}^{1}\frac{f(wy)T_k(y)}{\sqrt{1-y^2}}\,dy$. Then, for any distribution D, it is the case that*

$$\mathbb{E}_{x\in_R D}\left[|f(x)-f_d(x)|\,\mathbb{1}_{|x|>w}\right] \leq O\left(4^d\,\mathbb{E}_{x\in_R D}\left[|x|^d\,\mathbb{1}_{|x|>w}\right]\right).$$

Proof. See the full version of this work for complete details. □

The following proposition, in turn, allows us to bound the expression we encounter in Proposition 10 in terms of a bound on the moments of distribution D.

Proposition 11. *Let D be a distribution on $\mathbb{R}$ and $d_0 \in \mathbb{Z}^{>0}$ such that*

$$\left(\mathbb{E}_{x\in_R D}\left[|x|^{d_0}\right]\right)^{1/d_0} \leq \beta.$$

Then, for any $k \in \mathbb{Z}\cap[0,d_0/2]$ and $w \in \mathbb{R}^+$ we have

$$\mathbb{E}_{x\in_R D}\left[|x|^k\,\mathbb{1}_{|x|>w}\right] \leq 2w^k\left(\frac{\beta}{w}\right)^{d_0}$$

Proof. See the full version of this work for the proof. □

5.2 Proof of Low Degree Moment Lemma for Distributions(Lemma 6)

Let us recall the setting of Lemma 6. D is a distribution over $\mathbb{R}^n$ and Δ is an even positive integer, such that for every monomial $\prod_{i=1}^{n} x_i^{\alpha_i}$ of degree at most Δ we have

$$\left|\mathbb{E}_{\boldsymbol{x}\sim D}\left[\prod_{i=1}^{n} x_i^{\alpha_i}\right]-\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{i=1}^{n} x_i^{\alpha_i}\right]\right| \leq \frac{1}{n^{\Delta}}.$$

Further, we have that $\Delta \geq \frac{1}{\epsilon^4}\ln^4\left(\frac{1}{\epsilon}\right)$. Then, we would like to show that for every unit vector $\boldsymbol{v}$, the random variable $\boldsymbol{v}\cdot\boldsymbol{x}$ (with $\boldsymbol{x}\in_R D$) has the following properties

- **Concentration:** For any even integer $d \leq \Delta$, we have $\left(\mathbb{E}_{\boldsymbol{x}\in_R D}\left[|\boldsymbol{v}\cdot\boldsymbol{x}|^d\right]\right)^{1/d} \leq 2\sqrt{d}$.
- **Anti-concentration:** for any real-valued parameter $w \geq 1$, for any real y, we have

$$\Pr_{\boldsymbol{x}\in_R D}\left[\boldsymbol{v}\cdot\boldsymbol{x}\in[y,y+\epsilon]\right] \leq O(\epsilon).$$

We start with the following observation saying that if moments of a distribution D are similar to standard Gaussian, then the expectation of a polynomial of a form $(\boldsymbol{v}\cdot\boldsymbol{x})^d$ for D is similar to the same expectation under standard Gaussian.

Observation 12. *Suppose D is a distribution over $\mathbb{R}^n$ and Δ is a positive integer, such that for every monomial $\prod_{i=1}^{n} x_i^{\alpha_i}$ of degree at most Δ we have $\left|\mathbb{E}_{\boldsymbol{x}\sim D}\left[\prod_{i=1}^{n} x_i^{\alpha_i}\right]-\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,1)}\left[\prod_{i=1}^{n} x_i^{\alpha_i}\right]\right| \leq \frac{1}{n^{\Delta}}$. Then, for any unit vector $\boldsymbol{v}$ and integer $d \leq \Delta$ we have*

$$\left|\mathbb{E}_{\boldsymbol{x}\in_R D}\left[(\boldsymbol{v}\cdot\boldsymbol{x})^d\right]-\mathbb{E}_{\boldsymbol{x}\in_R\mathcal{N}(0,I_{n\times n})}\left[(\boldsymbol{v}\cdot\boldsymbol{x})^d\right]\right| \leq \frac{n^d}{n^{\Delta}}.$$

Proof. See the full version of this work for the proof. □

Let us now show the concentration property. Let d be even. Recall that for even d we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[(\boldsymbol{v}\cdot\boldsymbol{x})^d\right] = \mathbb{E}_{x'\sim\mathcal{N}(0,1)}\left[(x')^d\right] = (d-1)!! \leq d^{d/2}.$$

This, together with Observation 12 implies

$$\left(\mathbb{E}_{\boldsymbol{x}\sim D}\left[(\boldsymbol{v}\cdot\boldsymbol{x})^d\right]\right)^{1/d} \leq \left(d^{d/2}+\frac{n^d}{n^{\Delta}}\right)^{1/d} = \sqrt{d}\left(1+\frac{n^{d-\Delta}}{d^{d/2}}\right)^{1/d} \leq 2\sqrt{d},$$

which is the *concentration* property we wanted to show.

Now, we proceed to the *anti-concentration* property. Recall that for this property we need to bound $\Pr_{\boldsymbol{x}\in_R D}\left[\boldsymbol{v}\cdot\boldsymbol{x}\in[y,y+\epsilon]\right]$. To this end, we first approximate $\mathbb{1}_{z\in[y,y+\epsilon]}$ using the following function

$$g(z) := \begin{cases} 0 & \text{if } z \leq y-\epsilon, \\ \frac{z-(y-\epsilon)}{\epsilon} & \text{if } z\in[y-\epsilon,y], \\ 1 & \text{if } z\in[y,y+\epsilon], \\ \frac{(y+2\epsilon)-z}{\epsilon} & \text{if } z\in[y+\epsilon,y+2\epsilon], \\ 0 & \text{if } z \geq y+2\epsilon. \end{cases} \tag{1}$$

The key properties of g are (i) $g(z) \geq \mathbb{1}_{z\in[y,y+\epsilon]}$ (ii) $g(z)\in[0,1]$ (ii) $g(z)$ is Lipschitz continuous (iii) the derivative $g'(z)$ is of bounded variation of $\frac{4}{\epsilon}$ (because the function has four discontinuities, each of magnitude $1/\epsilon$ and it stays constant in-between the discontinuities).

Let $w \geq 1$ be real-valued and d be an integer in $[1,\Delta/2]$, to be chosen later and let $g_d(x) := \sum_{k=0}^{d} a_k T_k(\frac{x}{w})$, where $a_k := \frac{1+\mathbb{1}_{k>0}}{\pi}\int_{-1}^{1}\frac{g(wy)T_k(y)}{\sqrt{1-y^2}}\,dy$. Observation 13 and propositions 14 and

15 are stated and proven below, and we use them no to get the following bound:

$$\Pr_{\boldsymbol{x}\in_R D}[\boldsymbol{v}\cdot\boldsymbol{x}\in[y,y+\epsilon]]\le\mathbb{E}_{\boldsymbol{x}\in_R D}[g(\boldsymbol{v}\cdot\boldsymbol{x})]\le$$

$$\overbrace{\mathop{\mathbb{E}}_{\boldsymbol{x}\in_R\mathcal{N}(0,I_{n\times n})}[g(\boldsymbol{v}\cdot\boldsymbol{x})]}^{O(\epsilon)\text{ by Observation 13}}+\overbrace{\mathop{\mathbb{E}}_{\boldsymbol{x}\in_R\mathcal{N}(0,I_{n\times n})}[|g_d(\boldsymbol{v}\cdot\boldsymbol{x})-g(\boldsymbol{v}\cdot\boldsymbol{x})|]}^{\mathrm{O}\left(4^d w^d\left(\frac{2\sqrt{\Delta}}{w}\right)^{\Delta}+\frac{w}{\epsilon d}\right)\text{ by Proposition 14}}+$$

$$+\overbrace{\left|\mathbb{E}_{\boldsymbol{x}\in_R D}[g_d(\boldsymbol{v}\cdot\boldsymbol{x})]-\mathbb{E}_{\boldsymbol{x}\in_R\mathcal{N}(0,I_{n\times n})}[g_d(\boldsymbol{v}\cdot\boldsymbol{x})]\right|}^{\mathrm{O}\left(4^d\frac{n^d}{n^{\Delta}}\right)\text{ by Proposition 15}}+$$

$$+\quad\underbrace{\mathbb{E}_{\boldsymbol{x}\in_R D}[|g(\boldsymbol{v}\cdot\boldsymbol{x})-g_d(\boldsymbol{v}\cdot\boldsymbol{x})|]}_{\mathrm{O}\left(4^d w^d\left(\frac{2\sqrt{\Delta}}{w}\right)^{\Delta}+\frac{w}{\epsilon d}\right)\text{ by Proposition 14}}\quad=$$

$$=O\left(\epsilon+4^d w^d\left(\frac{2\sqrt{\Delta}}{w}\right)^{\Delta}+\frac{w}{\epsilon d}+4^d\frac{n^d}{n^{\Delta}}\right).$$

Now, recall we assumed without loss of generality that Δ equals to $\frac{1}{\epsilon^4}\ln^4\left(\frac{1}{\epsilon}\right)$, so taking[14] $d=\frac{1}{10\epsilon^4}\ln^2\left(\frac{1}{\epsilon}\right)$ and $w=\frac{10}{\epsilon^2}\ln^2\left(\frac{1}{\epsilon}\right)$ we get

$$\Pr_{\boldsymbol{x}\in_R D}[\boldsymbol{v}\cdot\boldsymbol{x}\in[y,y+\epsilon]]\le$$

$$O\left(\epsilon+4^d w^d\left(\frac{2\sqrt{\Delta}}{w}\right)^{\Delta}+\frac{w}{\epsilon d}+4^d\frac{n^d}{n^{\Delta}}\right)=$$

$$O\left(\epsilon+\left(\frac{40}{\epsilon^2}\ln^2\left(\frac{1}{\epsilon}\right)\right)^{\frac{1}{10\epsilon^4}\ln^2\left(\frac{1}{\epsilon}\right)}\left(\frac{1}{5}\right)^{\frac{1}{\epsilon^4}\ln^4\left(\frac{1}{\epsilon}\right)}+\right.$$

$$\left.+4^{\frac{1}{10\epsilon^4}\ln^2\left(\frac{1}{\epsilon}\right)}\frac{1}{n^{\frac{1}{\epsilon^4}\ln^4\left(\frac{1}{\epsilon}\right)-\frac{1}{10\epsilon^4}\ln^2\left(\frac{1}{\epsilon}\right)}}\right)=O(\epsilon).$$

The only thing left to do is to prove the observations referenced above.

Observation 13. *For the function g as defined in Equation 1, we have*

$$\mathbb{E}_{\boldsymbol{x}\in_R\mathcal{N}(0,I_{n\times n})}[g(\boldsymbol{v}\cdot\boldsymbol{x})]=O(\epsilon)$$

Proof. The function g has a range of $[0,1]$ and is supported on $[y-\epsilon,y+3\epsilon]$. Also, $\boldsymbol{v}\cdot\boldsymbol{x}$ is distributed as a standard one-dimensional Gaussian. Therefore, the probability that $\boldsymbol{v}\cdot\boldsymbol{x}$ lands in $[y-\epsilon,y+3\epsilon]$, is at most $O(\epsilon)$, which finishes the proof. □

Proposition 14. *Suppose D is a distribution over $\mathbb{R}^n$ and Δ is a positive integer, such that for every monomial $\prod_{i=1}^n x_i^{\alpha_i}$ of degree at most Δ we have $\left|\mathbb{E}_{\boldsymbol{x}\sim D}\left[\prod_{i=1}^n x_i^{\alpha_i}\right]-\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{i=1}^n x_i^{\alpha_i}\right]\right|\le\frac{1}{n^{\Delta}}$. Let d be an integer in $[1,\Delta/2]$, let $w\ge 1$ be a real-valued parameter, and suppose $g:[-w,w]\longrightarrow[-1,1]$ is a Lipschitz function whose derivative g' is of Bounded variation V, and let $g_d(x):=\sum_{k=0}^d a_k T_k(\frac{x}{w})$, where $a_k:=\frac{1+\mathbb{1}_{k>0}}{\pi}\int_{-1}^1\frac{g(wy)T_k(y)}{\sqrt{1-y^2}}dy$. Then, it is the case that*

$$\mathbb{E}_{\boldsymbol{x}\in_R D}[|g(\boldsymbol{v}\cdot\boldsymbol{x})-g_d(\boldsymbol{v}\cdot\boldsymbol{x})|]\le O\left(4^d w^d\left(\frac{2\sqrt{\Delta}}{w}\right)^{\Delta}+\frac{Vw}{d}\right).$$

Proof. Proposition 10 and Proposition 11 imply

$$\mathbb{E}_{\boldsymbol{x}\sim D}\left[|g(\boldsymbol{v}\cdot\boldsymbol{x})-g_d(\boldsymbol{v}\cdot\boldsymbol{x})|\,\mathbb{1}_{|\boldsymbol{v}\cdot\boldsymbol{x}|>w}\right]\le$$

$$O\left(4^d\mathbb{E}_{\boldsymbol{x}\in_R D}\left[|\boldsymbol{v}\cdot\boldsymbol{x}|^d\,\mathbb{1}_{|\boldsymbol{v}\cdot\boldsymbol{x}|>w}\right]\right)\le 4^d w^d\left(\frac{2\sqrt{\Delta}}{w}\right)^{\Delta}\frac{\Delta}{\Delta-d}.$$

To use Theorem 8, we need to bound the total variation of the function $\frac{\mathrm{d}g(wz)}{\mathrm{d}z}=wg'(wz)$. Inspecting the definition of total variation, we see that $g'(wz)$ has the same total variation as $g'(z)$, which is at most V. Therefore, the total variation of $\frac{\mathrm{d}g(wz)}{\mathrm{d}z}$ is at most Vw. Thus, we have by Theorem 8 that

$$\mathbb{E}_{\boldsymbol{x}\sim D}\left[|g(\boldsymbol{v}\cdot\boldsymbol{x})-g_d(\boldsymbol{v}\cdot\boldsymbol{x})|\,\mathbb{1}_{|\boldsymbol{v}\cdot\boldsymbol{x}|\le w}\right]\le$$

$$\max_{z\in[-w,w]}|g(z)-g_d(z)|\le O\left(\frac{Vw}{d}\right).$$

Summing the two equations above and recalling that $d\le\Delta/2$, our proposition follows. □

Proposition 15. *Suppose D is a distribution over $\mathbb{R}^n$ and Δ is a positive integer, such that for every monomial $\prod_{i=1}^n x_i^{\alpha_i}$ of degree at most Δ we have $\left|\mathbb{E}_{\boldsymbol{x}\sim D}\left[\prod_{i=1}^n x_i^{\alpha_i}\right]-\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,1)}\left[\prod_{i=1}^n x_i^{\alpha_i}\right]\right|\le\frac{1}{n^{\Delta}}$. Let $g:\mathbb{R}\longrightarrow[-1,1]$ be a Lipschitz continuous function, and $g_d(x):=\sum_{k=0}^d a_k T_k(\frac{x}{w})$,where $a_k:=\frac{1+\mathbb{1}_{k>0}}{\pi}\int_{-1}^1\frac{f(wy)T_k(y)}{\sqrt{1-y^2}}dy$. Then*

$$\left|\mathbb{E}_{\boldsymbol{x}\in_R D}[g_d(\boldsymbol{v}\cdot\boldsymbol{x})]-\mathbb{E}_{\boldsymbol{x}\in_R\mathcal{N}(0,I_{n\times n})}[g_d(\boldsymbol{v}\cdot\boldsymbol{x})]\right|=O\left(4^d\frac{n^d}{n^{\Delta}}\right).$$

Proof. Observation 9 implies that $g_d(z)$ is a degree d polynomial, whose largest coefficient is at most $d3^d$. Using Observation 12 for each of these monomials, we get

$$\left|\mathbb{E}_{\boldsymbol{x}\in_R D}[g_d(\boldsymbol{v}\cdot\boldsymbol{x})]-\mathbb{E}_{\boldsymbol{x}\in_R\mathcal{N}(0,I_{n\times n})}[g_d(\boldsymbol{v}\cdot\boldsymbol{x})]\right|\le$$

$$O\left(d^2 3^d\right)\frac{n^d}{n^{\Delta}}=O\left(4^d\frac{n^d}{n^{\Delta}}\right).$$

□

5.3 Proof of Low Degree Approximation Lemma for Halfspaces (Lemma 7)

Let us recall what we need to show to prove Lemma 7. Without loss of generality, we assume we are in one dimension. D is a distribution on $\mathbb{R}$, such that for some positive real parameters α,γ,ϵ and a positive integer parameter d_0 we have

- **Anti-concentration:** for any real y, we have
$$\Pr_{x\in_R D}[x\in[y,y+\epsilon]]\le\alpha$$
,
- **Concentration:** $\left(\mathbb{E}_{x\in_R D}\left[|x|^{d_0}\right]\right)^{1/d_0}\le\beta$, for some $\beta\ge 1$.

Also we have $d_0>\frac{5\beta}{\epsilon^2}$ and that ϵ is smaller than some sufficiently small absolute constant. Then, for every $\theta\in\mathbb{R}$ we would like to show there is a polynomial $P(x)$ of degree at most $\frac{2\beta}{\epsilon^2}+1$ such that

$$E_{x\in_R D}[|P(x)-\mathrm{sign}(x-\theta)|]=O\left(\alpha+\epsilon+\frac{(8\beta)^{\frac{2\beta}{\epsilon^2}+1}}{2^{d_0}}\right).$$

[14] We also check that (taking ϵ small enough) d is indeed in $[1,\Delta/2]$, as was required earlier.

Let $w > 1$ and $d \in Z^+$ be parameters, values of which will be set later. We will approximate the sign function with a polynomial in the following two steps:

- Approximate $\text{sign}(x-\theta)$ by a continuous function
$$f(x) := \begin{cases} 1 & \text{if } \frac{x-\theta}{\epsilon} > 1, \\ -1 & \text{if } \frac{x-\theta}{\epsilon} < -1, \\ \frac{x-\theta}{\epsilon} & \text{otherwise.} \end{cases}$$
- For a parameter d, approximate $f(x)$ by
$$f_d(x) := \sum_{k=0}^{d} a_k T_k(\frac{x}{w}),$$
where
$$a_k := \frac{1 + \mathbb{1}_{k>0}}{\pi} \int_{-1}^{1} \frac{f(wy)T_k(y)}{\sqrt{1-y^2}}\, dy.$$

First, we observe that f is a good approximator for $\text{sign}(x-\theta)$ with respect to D.

Proposition 16. *If D is a distribution over $\mathbb{R}$ such that for every $x_0 \in \mathbb{R}$ we have $\Pr_{x \in_R D}[x \in [x_0, x_0+\epsilon]] \leq \alpha$, then (with $f(x)$ defined as above) we have*
$$\mathbb{E}_{x \in_R D}[|f(x) - sign(x-\theta)|] \leq 2\alpha.$$

Proof. The two functions differ only on $[\theta-\epsilon, \theta+\epsilon]$, with the absolute value of difference being at most 1. Since the distribution D cannot have probability mass more than 2α in this interval, the proposition follows. □

Secondly, we show that f_d is a good approximator to f with respect to D, within the region $[-w, w]$.

Proposition 17. *For any distribution D, we have*
$$\mathbb{E}_{x \in_R D}\left[|f(x) - f_d(x)|\, \mathbb{1}_{|x| \leq w}\right] \leq O\left(\frac{w}{\epsilon d}\right)$$

Proof. Using Theorem 8 we have
$$\mathbb{E}_{x \in_R D}\left[|f(x) - f_d(x)|\, \mathbb{1}_{|x| \leq w}\right] \leq$$
$$\max_{x \in [-w,w]} |f(x) - f_d(x)| = \max_{y \in [-1,1]} |f(wy) - f_d(wy)| = O\left(\frac{w}{\epsilon d}\right).$$
□

Now, we put all the relevant propositions together to show the lemma. Using Propositions 10 and 11, we see that if we have $d \in \mathbb{Z} \cap [1, d_0/2]$ then
$$\mathbb{E}_{x \in_R D}\left[|f(x) - f_d(x)|\, \mathbb{1}_{|x| > w}\right] \leq$$
$$O\left(4^d \mathbb{E}_{x \in_R D}\left[|x|^d\, \mathbb{1}_{|x|>w}\right]\right) \leq O\left(4^d 2 w^d \left(\frac{\beta}{w}\right)^{d_0}\right)$$

Together with Proposition 17, this implies that
$$\mathbb{E}_{x \in_R D}[|f(x) - f_d(x)|] \leq O\left(4^d 2 w^d \left(\frac{\beta}{w}\right)^{d_0}\right) + O\left(\frac{w}{\epsilon d}\right)$$

This, in turn, together with Proposition 16 implies that
$$E_{x \in_R D}[|f_d(x) - \text{sign}(x-\theta)|] = O\left(\alpha + \frac{w}{\epsilon d} + 4^d w^d \left(\frac{\beta}{w}\right)^{d_0}\right).$$

Taking[15] $w = 2\beta$ and $d = \left\lceil \frac{2\beta}{\epsilon^2} \right\rceil$ we get
$$E_{x \in_R D}[|f_d(x) - \text{sign}(x-\theta)|] =$$
$$O\left(\alpha + \epsilon + \frac{(8\beta)^{\left\lceil \frac{2\beta}{\epsilon^2} \right\rceil}}{2^{d_0}}\right) = O\left(\alpha + \epsilon + \frac{(8\beta)^{\frac{2\beta}{\epsilon^2}+1}}{2^{d_0}}\right).$$

Finally, we note that by Observation 9 we have that each coefficient of the polynomial f_d has a magnitude of at most $O(d3^d) = O\left(4^{\frac{2\beta}{\epsilon^2}}\right)$. This completes the proof of the low degree approximation lemma for halfspaces (Lemma 7).

6 PROOF OF MAIN THEOREM VIA TWO MAIN LEMMAS

6.1 Truncated Gaussian has Moments Similar to Gaussian

Recall that our tester truncates the samples and checks that low-degree moments are close to the corresponding moments of a Gaussian. If the distribution is indeed Gaussian, the following proposition shows that this truncation step does not distort the moments too much.

Proposition 18. *Let $\prod_{i=1}^n x_i^{\alpha_i}$ be a monomial of degree at most Δ and t a real number in $\left[2\sqrt{\Delta}+1, +\infty\right)$. Then we have*
$$\left|\mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(0, I_{n \times n})}\left[\prod_{i=1}^{n} x_i^{\alpha_i} \middle| \forall i : |x_i| \leq t\right] - \right.$$
$$\left. - \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(0, I_{n \times n})}\left[\prod_{i=1}^{n} x_i^{\alpha_i}\right]\right| \leq O\left(2^\Delta \Delta^{\frac{\Delta+2}{2}} t^\Delta e^{-\frac{t^2}{2}}\right).$$

Proof. For the proof, we refer the reader to the full version of this work. □

6.2 Finishing the Proof of Theorem 5

In this subsection we finish the proof of Theorem 5, using the low degree moment lemma for distributions (Lemma 6) and the low degree approximation lemma for halfspaces (Lemma 7). The main thing left to do is to address issues relating to truncation of samples in the learning and testing algorithms.

We now restate the theorem. The values $C_1, \cdots, C_4$ present in algorithms $\mathcal{A}$ and $\mathcal{T}$ (in the beginning of Section 3) are chosen to be sufficiently large absolute constants, and also n and $\frac{1}{\epsilon}$ are larger than some sufficiently large absolute constant. Then, we need to show that the algorithm $\mathcal{A}$ is an agnostic $(O(\epsilon), 0.1)$-learner for the function class of linear threshold functions over $\mathbb{R}^n$ under distribution $\mathcal{N}(0, I_{n \times n})$ and the algorithm $\mathcal{T}$ is an assumption tester for $\mathcal{A}$. We also need to show that $\mathcal{A}$ and $\mathcal{T}$ require only $n^{\tilde{O}\left(\frac{1}{\epsilon^4}\right)}$ samples and run-time.

Bounds on the run-time and sample complexity of our algorithms follow directly from our choice of parameters.

[15] Recall that to do all this we needed that d is in $[1, d_0/2]$. Recall that by an assumption of the lemma we are proving we have $d_0 > \frac{5\beta}{\epsilon^2}$ and $\beta \geq 1$. Therefore, for ϵ smaller than some sufficiently small absolute constant we indeed have $\left\lceil \frac{2\beta}{\epsilon^2} \right\rceil \in [1, d_0/2]$.

- The learner $\mathcal{A}$ draws $N_1 := n^{\tilde{O}\left(\frac{1}{\epsilon^4}\right)}$ samples, then performs a computation running in time polynomial in (i) N_1 (ii) the number of monomials $\prod_{j=1}^{n} x_j^{\alpha_j}$ of degree at most d, which is $O\left(n^d\right)$ (this includes the run-time consumed by the algorithm of Theorem 2). Overall, the learner $\mathcal{A}$ uses $n^{\tilde{O}\left(\frac{1}{\epsilon^4}\right)}$ samples and run-time.
- The tester $\mathcal{T}$ first performs estimations of values $\Pr\left[\left|x_j\right| > t\right]$ up to additive $\frac{\epsilon}{30n}$ with error probability $\frac{1}{100n}$, which in total require $poly\left(\frac{n}{\epsilon}\right)$ samples and run-time. Then, the tester $\mathcal{T}$ obtains $N_2 := \lceil t^{\Delta} n^{C_4 \Delta} \rceil$ samples (where $\Delta := \left\lfloor \frac{1}{\epsilon^4} \ln^4\left(\frac{1}{\epsilon}\right) \right\rfloor$ and $t := C_1 \Delta \ln \Delta \sqrt{\log n} + \sqrt{2 \ln\left(\frac{C_2 n}{\epsilon}\right)}$) and performs a polynomial time computation with them. We see that $t = O\left(poly\left(\frac{1}{\epsilon}, n\right)\right)$ and therefore $N_2 = n^{\tilde{O}\left(\frac{1}{\epsilon^4}\right)}$. Finally, the tester $\mathcal{T}$ runs a computation running in time polynomial in (i) N_2 and (ii) the number of monomials $\prod_{j=1}^{n} x_j^{\alpha_j}$ of degree at most Δ, which is $O\left(n^{\Delta}\right)$. Overall, we get that the run-time and sample complexity of $\mathcal{T}$ is $n^{\tilde{O}\left(\frac{1}{\epsilon^4}\right)}$.

Proposition 19. *The following proposition uses the low degree approximation lemma for halfspaces (Lemma 7) to argue that, under certain regularity conditions on the distribution D, the learning algorithm satisfies the agnostic learning guarantee. Suppose the $C_1, \cdots, C_4$ are chosen to be sufficiently large absolute constants, n and $\frac{1}{\epsilon}$ are larger than some sufficiently large absolute constant. Suppose D is a distribution over $\mathbb{R}^n$ such that it the following properties hold*

- ***Good tail:*** *We have* $\Pr_{\boldsymbol{x} \in_R D}\left[\exists i \in [n] : |x_i| > t\right] \le \frac{\epsilon}{5}$.
- **Concentration along any direction for truncated distribution:** For any unit vector $\boldsymbol{v}$ we have
$$\left(\mathbb{E}_{\boldsymbol{x} \in_R D}\left[\left|\boldsymbol{v} \cdot \boldsymbol{x}\right|^d \middle| \forall i \in [n] : |x_i| \le t\right]\right)^{1/d} \le 2\sqrt{d}.$$
- **A*nti-concentration along* any direction for truncated distribution*:*** For any unit vector $\boldsymbol{v}$ and for any real y, we have
$$\Pr_{\boldsymbol{x} \in_R D}\left[\boldsymbol{v} \cdot \boldsymbol{x} \in [y, y+\epsilon] \middle| \forall i \in [n] : |x_i| \le t\right] \le O(\epsilon).$$

Then, the algorithm $\mathcal{A}$ is an agnostic $(O(\epsilon), 0.1)$-learner for the function class of linear threshold functions over $\mathbb{R}^n$ under distribution D with failure probability at most $\frac{1}{20}$.

Proof. Let $D_{\text{truncated}}$ be the distribution of $\boldsymbol{x}$ drawn from D conditioned on $|x_i| \le t$ for all i. We see that the premises of this proposition imply that the distribution $D_{\text{truncated}}$ satisfies the premises of the low degree approximation lemma for halfspaces(Lemma 7) with parameters $d_0 = d$, $\alpha = O(\epsilon)$ and $\beta = 2\sqrt{d}$. Taking ϵ smaller than some absolute constant ensures that the condition $d > \frac{5\beta}{\epsilon^2} = \frac{10\sqrt{d}}{\epsilon^2}$ is also satisfied.

The low degree approximation lemma for halfspaces(Lemma 7) then allows us to conclude that for every $\theta \in \mathbb{R}$ and for any $w \ge 1$ there is a polynomial $P(x)$ of degree at most d such that
$$E_{\boldsymbol{x} \in_R D_{\text{truncated}}}\left[\left|\text{sign}(\boldsymbol{v} \cdot \boldsymbol{x} - \theta) - P(\boldsymbol{v} \cdot \boldsymbol{x})\right|\right] = O\left(\epsilon + \frac{\left(16\sqrt{d}\right)^{\frac{4\sqrt{d}}{\epsilon^2}+1}}{2^d}\right)$$

Recalling that $d := 2\left\lfloor \frac{1}{2\epsilon^4} \ln^3\left(\frac{1}{\epsilon}\right) \right\rfloor$ so we get that
$$E_{\boldsymbol{x} \in_R D_{\text{truncated}}}\left[\left|\text{sign}(\boldsymbol{v} \cdot \boldsymbol{x} - \theta) - P(\boldsymbol{v} \cdot \boldsymbol{x})\right|\right] = O\left(\epsilon + \frac{\left(O\left(\frac{1}{\epsilon^2} \ln^{1.5}\left(\frac{1}{\epsilon}\right)\right)\right)^{O\left(\frac{1}{\epsilon^4} \ln^{1.5}\left(\frac{1}{\epsilon}\right)\right)}}{2^{\Omega\left(\frac{1}{\epsilon^4} \ln^3\left(\frac{1}{\epsilon}\right)\right)}}\right).$$

For ϵ smaller than some sufficiently small absolute constant, the above is $O(\epsilon)$.

Thus, we have that for a linear threshold function $\text{sign}\left(\boldsymbol{v} \cdot \boldsymbol{x} - \theta\right)$ there is a degree d multivariate polynomial Q for which
$$\mathbb{E}_{\boldsymbol{x} \in_R D_{\text{truncated}}}\left[\left|\text{sign}\left(\boldsymbol{v} \cdot \boldsymbol{x} - \theta\right) - Q(\boldsymbol{x})\right|\right] \le O(\epsilon)$$
In other words, under $D_{\text{truncated}}$, any linear threshold function $\text{sign}(\boldsymbol{v} \cdot \boldsymbol{x} - \theta)$ is $O(\epsilon)$-approximated in L^1 by something in the span of set of monomials of degree at most d, i.e. the set
$$\left\{\prod_{j=1}^{n} x_j^{\alpha_j} : \sum_j \alpha_j \le d\right\}.$$
Now, Theorem 2. tells us that with probability at least $1 - \frac{1}{20}$ the predictor $\widehat{f}$ given in step 3 has an error of at most $O(\epsilon)$ more than $\text{sign}(\boldsymbol{v} \cdot \boldsymbol{x} - \theta)$ for samples $\boldsymbol{x} \in_R D_{\text{truncated}}$. Overall, recalling the definition of $D_{\text{truncated}}$ we have
$$\Pr_{\boldsymbol{x}, y \in_R D_{\text{pairs}}}\left[\widehat{f}'(\boldsymbol{x}) \ne y\right] \le \Pr_{\boldsymbol{x} \in_R D}\left[\exists i \in [n] : |x_i| > t\right] + $$
$$+ \Pr_{\boldsymbol{x}, y \in_R D_{\text{pairs}}}\left[\widehat{f}(\boldsymbol{x}) \ne y \middle| \forall i \in [n] : |x_i| \le t\right] \le$$
$$\Pr_{\boldsymbol{x}, y \in_R D_{\text{pairs}}}\left[\text{sign}(\boldsymbol{v} \cdot \boldsymbol{x} - \theta) \ne y \middle| \forall i \in [n] : |x_i| \le t\right] + O(\epsilon),$$
which completes the proof. □

Now, the following proposition, using low degree moment lemma for distributions (Lemma 6), tells us that the tester we use (1) is likely accept if the Gaussian assumption indeed holds (2) is likely to reject if the regularity conditions for Proposition 19 do not hold.

Proposition 20. *Suppose the $C_1, \cdots, C_4$ are chosen to be sufficiently large absolute constants, n and $\frac{1}{\epsilon}$ are larger than some sufficiently large absolute constant. Then, there is some absolute constant B, so the tester $\mathcal{T}$ has the following properties:*

1. *If $\mathcal{T}$ is given samples from $\mathcal{N}(0, I_{n \times n})$, it outputs **Yes** with probability at least* 0.9.
2. *The tester $\mathcal{T}$ rejects with probability greater than* 0.9 *any D for which at least one of the following holds:*
   (a) ***Bad tail:*** *We have* $\Pr_{\boldsymbol{x} \in_R D}\left[\exists i \in [n] : |x_i| > t\right] > \frac{\epsilon}{5}$.

(b) ***Failure of concentration along some direction for truncated distribution:*** *there is a unit vector v such that*

$$\left(\mathbb{E}_{\boldsymbol{x}\in_R D}\left[|v\cdot\boldsymbol{x}|^d\,\middle|\,\forall i\in[n]:\ |x_i|\le t\right]\right)^{1/d} > 2\sqrt{d}.$$

(c) ***Failure of anti-concentration along some direction for truncated distribution:*** *there is a unit vector v and real y, for which*

$$\Pr_{\boldsymbol{x}\in_R D}\left[v\cdot\boldsymbol{x}\in[y,y+\epsilon]\,\middle|\,\forall i\in[n]:\ |x_i|\le t\right] > B\epsilon.$$

Proof. First, assume that $\mathcal{T}$ is getting samples from $\mathcal{N}(0,I_{n\times n})$ and let us prove that $\mathcal{T}$ outputs ***Yes*** *with probability at least* 0.9.

Since $t\ge 1$, by we have[16] $\Pr_{z\in\mathcal{N}(0,1)}\left[|z|>t\right]\le O\left(e^{-\frac{t^2}{2}}\right)$. As $t\ge\sqrt{2\ln\left(\frac{C_2 n}{\epsilon}\right)}$, taking C_2 large enough we get

$$\Pr_{z\in\mathcal{N}(0,1)}\left[|z|>t\right]\le\frac{\epsilon}{30n}.$$

Therefore, $\mathcal{N}(0,I_{n\times n})$ passes step 1 of tester $\mathcal{T}$ with probability at least $1-\frac{1}{100}$.

Also, $\Pr_{z\in\mathcal{N}(0,1)}\left[|z|>t\right]\le\frac{\epsilon}{30n}$ implies that

$$\Pr_{\boldsymbol{x}\in\mathcal{N}(0,I_{n\times n})}\left[\forall i\in[n]:\ |x_i|\le t\right]\ge 1-\frac{\epsilon}{30}.$$

Together with a very loose application of the Hoeffding bound, we see that for sufficiently large C_4 with probability at least $1-\frac{1}{100}$ only at most half of the samples are discarded in the step 2 of $\mathcal{T}$. We henceforth assume this indeed was the case. The remaining samples themselves are i.i.d. and distributed according to $\mathcal{N}(0,I_{n\times n})$ conditioned on all coordinates being in $[-t,t]$.

Since all remaining samples have the size of their coordinates bounded by t, the value of a given monomial $\prod_{j=1}^n x_j^{\alpha_j}$ of degree at most Δ evaluated on any of them is in $\left[-t^\Delta,t^\Delta\right]$. Therefore, the Hoeffding bound implies that for sufficiently large C_4 with probability at least $1-\frac{1}{100n^\Delta}$ the empirical average of $\prod_{j=1}^n x_j^{\alpha_j}$ on the (at least $\frac{N_2}{2}$ many) remaining samples is within $\frac{1}{10n^\Delta}$ of

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{i=1}^n x_i^{\alpha_i}\,\middle|\,\forall i:\ |x_i|\le t\right].$$

For sufficiently large C_1, we verify the premise of Proposition 18 that $t\in\left[2\sqrt{\Delta}+1,+\infty\right)$ and therefore have

$$\left|\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{i=1}^n x_i^{\alpha_i}\,\middle|\,\forall i:\ |x_i|\le t\right]-\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{i=1}^n x_i^{\alpha_i}\right]\right|\le O\left(2^\Delta\Delta^{\frac{\Delta+2}{2}}t^\Delta e^{-\frac{t^2}{2}}\right).$$

Now, we have $\frac{d}{dt}\left(\Delta\log t-\frac{t^2}{2}\right)=\frac{\Delta}{t}-t$ which is negative when $t>\sqrt{\Delta}$. As $t\ge C_1\Delta\left(\ln\Delta\sqrt{\log n}\right)>\sqrt{\Delta}$, we have

$$t^\Delta e^{-\frac{t^2}{2}}\le\left(C_1\Delta\left(\ln\Delta\sqrt{\log n}\right)\right)^\Delta\exp\left(-\frac{\left(C_1\Delta\left(\ln\Delta\sqrt{\log n}\right)\right)^2}{2}\right),$$

which together with the preceding inequality implies

$$\left|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{i=1}^n x_i^{\alpha_i}\,\middle|\,\forall i:\ |x_i|\le t\right]-\right.$$
$$\left.-\,\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{i=1}^n x_i^{\alpha_i}\right]\right|\le$$
$$O\left(2^\Delta\Delta^{\frac{\Delta+2}{2}}\left(C_1\Delta\left(\ln\Delta\sqrt{\log n}\right)\right)^\Delta\exp\left(-\frac{\left(C_1\Delta\left(\ln\Delta\sqrt{\log n}\right)\right)^2}{2}\right)\right)$$

for sufficiently large C_1 the above is less than $\frac{1}{10n^\Delta}$. Therefore, in the whole, we have that the empirical average of $\prod_{j=1}^n x_j^{\alpha_j}$ in step 3 of $\mathcal{T}$ is with probability at least $1-\frac{1}{100n^\Delta}$ within $\frac{1}{10n^\Delta}$ of $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{i=1}^n x_i^{\alpha_i}\right]$. Taking a union bound over all monomials $\prod_{j=1}^n x_j^{\alpha_j}$ of degree at most Δ, we see that the step 3 of the tester $\mathcal{T}$ also passes with probability at least $1-\frac{1}{100}$ when it is run on $\mathcal{N}(0,I_{n\times n})$.

Overall, we conclude that the probability $\mathcal{T}$ outputs **No** when given samples from $\mathcal{N}(0,I_{n\times n})$ is at most $\frac{3}{100}<0.1$ as promised.

Now, we shall show that $\mathcal{T}$ will likely output **No** if any of the conditions given in the proposition hold.

If Condition (a) holds, we have $\Pr_{\boldsymbol{x}\in_R D}\left[\exists i\in[n]:\ |x_i|>t\right]>\frac{\epsilon}{5}$, then there is some coordinate i for which $\Pr_{\boldsymbol{x}\in_R D}\left[|x_i|>t\right]>\frac{\epsilon}{5n}$. This coordinate will lead to $\mathcal{T}$ outputting **No** in step 1 with probability at least $1-\frac{1}{100}$.

Now, suppose condition (a) doesn't hold so we have

$$\Pr_{\boldsymbol{x}\in_R D}\left[\exists i\in[n]:\ |x_i|>t\right]\le\frac{\epsilon}{5}$$

but condition (b) or (c) does hold. We would like to show that $\mathcal{T}$ will still likely output **No.** With a very loose application of the Hoeffding bound, for sufficiently large C_4 with probability at least $1-\frac{1}{100}$ only at most half of the samples are discarded in the step 2 of $\mathcal{T}$, which we also assume henceforth. Using the Hoeffding bound again, we see that for sufficiently large C_4 with probability at least $1-\frac{1}{100}$ the empirical expectation of all monomials $\prod_{j=1}^n x_j^{\alpha_j}$ of degree at most Δ is within $\frac{1}{10n^\Delta}$ of

$$\mathbb{E}_{\boldsymbol{x}\in_R D}\left[\prod_{i=1}^n x_i^{\alpha_i}\,\middle|\,\forall i\in[n]:\ |x_i|\le t\right].$$

In other words, with probability at least $1-\frac{1}{100}$ the tester $\mathcal{T}$ will output **No** in step 3, unless we have for all monomials $\prod_{j=1}^n x_j^{\alpha_j}$

[16] Proof: $\int_t^{+\infty}e^{-\frac{x^2}{2}}\,dx\le e^{-\frac{t^2}{2}}\int_t^{+\infty}e^{-\frac{t(x-t)}{2}}\,dx\le\frac{2e^{-\frac{t^2}{2}}}{t}\le O\left(e^{-\frac{t^2}{2}}\right)$.

that

$$\left|\mathbb{E}_{\boldsymbol{x}\in_R D}\left[\prod_{i=1}^{n} x_i^{\alpha_i}\middle|\forall i\in[n]:\ |x_i|\le t\right]-\right.$$

$$\left.-\ \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,I_{n\times n})}\left[\prod_{j=1}^{n} x_j^{\alpha_j}\right]\right|\le$$

$$\frac{1}{2n^{\Delta}}+\frac{1}{10n^{\Delta}}=\frac{3}{5n^{\Delta}}.$$

So, to finish the proof, it is enough to show that the inequality above cannot hold if Condition (b) or Condition (c) holds. This follows from the low degree moment lemma for distributions(Lemma 6), for a sufficiently large choice of B, thereby finishing the proof[17]. □

Finally, we can use the two propositions above to finish the proof of Theorem 5. Bounds on run-time have been shown earlier, so now we need to show correctness. That requires us to show the following two conditions:

(1) **(Composability)** If, given access to i.i.d. labeled samples (x, y) distributed according to D_{pairs}, the algorithm $\mathcal{T}$ outputs "Yes" with probability at least 1/4, then $\mathcal{A}$ will with probability at least 0.9 output a circuit computing a function $\hat{f}$, such that

$$\Pr_{(x,y)\in_R D_{\text{pairs}}}[y\neq\hat{f}(x)]\le$$

$$\min_{f\in\text{halfspaces}}\left(\Pr_{(x,y)\in_R D_{\text{pairs}}}[f(x)\neq y]\right)+O(\epsilon)\,.$$

(2) **(Completeness)** Given access to i.i.d. labeled samples (x, y) distributed according to D_{pairs}, with x itself distributed as a Gaussian over R^n, tester $\mathcal{T}$ outputs "Yes" with probability at least 3/4.

(3) $\mathcal{A}$ is an agnostic learner for halfspaces over $\mathbb{R}^n$ under the Gaussian distribution.

Note that Condition 3 follows from the first two. The completeness condition (i.e. Condition 2) immediately follows from Proposition 20. The composability condition (i.e. Condition 1) follows from Proposition 20 and Proposition 19 in following way. If $\mathcal{T}$ outputs "No" with probability less than 3/4 then conditions (a), (b) and (c) in Proposition 20 should all be violated. This allows us to use Proposition 19 to conclude that $\mathcal{A}$ is an agnostic $(O(\epsilon), 0.1)$-learner for the function class of linear threshold functions over $\mathbb{R}^n$ under distribution D, where D is the marginal distribution of x when (x, y) distributed according to D_{pairs}. This implies the composability condition (i.e. Condition 1 above) and finishes the proof of Theorem 5.

ACKNOWLEDGEMENTS

We wish to thank Jonathan Kelner and Pravesh Kothari for their useful comments and references. We also thank anonymous referees for their comments.

[17] To be explicit: if condition (a) doesn't hold but condition (b) or (c) does hold via union bound the probability that $\mathcal{T}$ will fail to output **No** is at most $\frac{1}{100}+\frac{1}{100}<0.1$ as required.

REFERENCES

[1] William Aiello and Milena Mihail. 1991. Learning the Fourier spectrum of probabilistic lists and trees. In *Proceedings of the second annual ACM-SIAM symposium on Discrete algorithms (SODA '91)*. Society for Industrial and Applied Mathematics, USA, 291–299.

[2] Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. 2007. Testing k-wise and almost k-wise independence. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, David S. Johnson and Uriel Feige (Eds.). ACM, 496–505.

[3] Noga Alon, Oded Goldreich, and Yishay Mansour. 2003. Almost k-wise independence versus k-wise independence. *Inf. Process. Lett.* 88, 3 (2003), 107–110.

[4] Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. 2014. The power of localization for efficiently learning linear separators with noise. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, David B. Shmoys (Ed.). ACM, 449–458. https://doi.org/10.1145/2591796.2591839

[5] Khanh Do Ba, Huy L. Nguyen, Huy N. Nguyen, and Ronitt Rubinfeld. 2011. Sublinear Time Algorithms for Earth Mover's Distance. *Theory Comput. Syst.* 48, 2 (2011), 428–442.

[6] Imre Bárány and Zoltán Füredi. 1986. Computing the Volume Is Difficult. In *Proceedings of the 18th Annual ACM Symposium on Theory of Computing, May 28-30, 1986, Berkeley, California, USA*, Juris Hartmanis (Ed.). ACM, 442–447. https://doi.org/10.1145/12130.12176

[7] Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. 2001. Testing Random Variables for Independence and Identity. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*. IEEE Computer Society, 442–451.

[8] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. 2000. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA*. IEEE Computer Society, 259–269.

[9] Eric Blais, Clément L. Canonne, Igor C. Oliveira, Rocco A. Servedio, and Li-Yang Tan. 2015. Learning Circuits with few Negations. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2015) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 40)*, Naveen Garg, Klaus Jansen, Anup Rao, and José D. P. Rolim (Eds.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 512–527. https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2015.512 ISSN: 1868-8969.

[10] E. Blais, R. O'Donnell, and K. Wimmer. 2008. Polynomial regression under arbitrary product distributions. *Machine Learning* (2008). https://doi.org/10.1007/s10994-010-5179-6

[11] Guy Blanc, Jane Lange, Ali Malik, and Li-Yang Tan. 2022. On the power of adaptivity in statistical adversaries. In *Conference on Learning Theory*. PMLR, 5030–5061.

[12] Nader H. Bshouty and Christino Tamon. 1995. On the Fourier spectrum of monotone functions. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing (STOC '95)*. Association for Computing Machinery, New York, NY, USA, 219–228. https://doi.org/10.1145/225058.225125

[13] T Tony Cai and Zongming Ma. 2013. Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli* 19, 5B (2013), 2359–2388.

[14] Clément L. Canonne. 2022. *Topics and Techniques in Distribution Testing: A Biased but Representative Sample*. https://ccanonne.github.io/files/misc/main-survey-fnt.pdf

[15] Clément L. Canonne, Elena Grigorescu, Siyao Guo, Akash Kumar, and Karl Wimmer. 2017. Testing k-Monotonicity. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA (LIPIcs, Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 29:1–29:21. https://doi.org/10.4230/LIPIcs.ITCS.2017.29

[16] Mahdi Cheraghchi, Adam Klivans, Pravesh Kothari, and Homin K. Lee. 2012. Submodular Functions are Noise Stable. In *Proceedings of the 2012 Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 1586–1592. https://doi.org/10.1137/1.9781611973099.126

[17] Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. 2014. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *Proceedings of the 2015 Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 498–511. https://doi.org/10.1137/1.9781611973730.34

[18] Amit Daniely. 2015. A PTAS for Agnostically Learning Halfspaces. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015 (JMLR Workshop and Conference Proceedings, Vol. 40)*, Peter Grünwald, Elad Hazan, and Satyen Kale (Eds.). JMLR.org, 484–502. http://proceedings.mlr.press/v40/Daniely15.html

[19] Amit Daniely. 2016. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, Daniel Wichs and Yishay Mansour (Eds.). ACM, 105–117. https://doi.org/10.1145/2897518.2897520

[20] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. 2009. Bounded Independence Fools Halfspaces. In *50th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2009, October 25-27, 2009, Atlanta, Georgia, USA*. IEEE Computer Society, 171–180. https://doi.org/10.1109/FOCS.2009.68

[21] Ilias Diakonikolas, Themis Gouleakis, Daniel M. Kane, John Peebles, and Eric Price. 2021. Optimal testing of discrete distributions with high probability. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, Samir Khuller and Virginia Vassilevska Williams (Eds.). ACM, 542–555.

[22] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. 2016. Collision-based Testers are Optimal for Uniformity and Closeness. *Electron. Colloquium Comput. Complex.* (2016), 178.

[23] Ilias Diakonikolas, Prahladh Harsha, Adam Klivans, Raghu Meka, Prasad Raghavendra, Rocco A. Servedio, and Li-Yang Tan. 2010. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *Proceedings of the 42nd ACM symposium on Theory of computing - STOC '10*. ACM Press, Cambridge, Massachusetts, USA, 533. https://doi.org/10.1145/1806689.1806763

[24] Ilias Diakonikolas, Daniel Kane, and Nikos Zarifis. 2020. Near-Optimal SQ Lower Bounds for Agnostically Learning Halfspaces and ReLUs under Gaussian Marginals. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[25] Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. 2021. Agnostic Proper Learning of Halfspaces under Gaussian Marginals. In *Proceedings of Thirty Fourth Conference on Learning Theory*. PMLR, 1522–1551. https://proceedings.mlr.press/v134/diakonikolas21b.html ISSN: 2640-3498.

[26] Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. 2021. The Optimality of Polynomial Regression for Agnostic Learning under Gaussian Marginals in the SQ Model. In *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA (Proceedings of Machine Learning Research, Vol. 134)*, Mikhail Belkin and Samory Kpotufe (Eds.). PMLR, 1552–1584.

[27] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. 2020. Non-Convex SGD Learns Halfspaces with Adversarial Label Noise. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/d785bf9067f8af9e078b93cf26de2b54-Abstract.html

[28] György Elekes. 1986. A Geometric Inequality and the Complexity of Computing Volume. *Discret. Comput. Geom.* 1 (1986), 289–292. https://doi.org/10.1007/BF02187701

[29] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. 2006. New Results for Learning Noisy Parities and Halfspaces. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*. IEEE Computer Society, 563–574. https://doi.org/10.1109/FOCS.2006.51

[30] Vitaly Feldman and Pravesh Kothari. 2015. Agnostic learning of disjunctions on symmetric distributions. *The Journal of Machine Learning Research* 16, 1 (Jan. 2015), 3455–3467.

[31] Vitaly Feldman, Pravesh Kothari, and Jan Vondrák. 2017. Tight Bounds on `$\ell_1$` Approximation and Learning of Self-Bounding Functions. In *International Conference on Algorithmic Learning Theory*. PMLR, 540–559. http://proceedings.mlr.press/v76/feldman17a.html ISSN: 2640-3498.

[32] V. Feldman and J. Vondrák. 2015. Tight Bounds on Low-Degree Spectral Concentration of Submodular and XOS Functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. 923–942. https://doi.org/10.1109/FOCS.2015.61 ISSN: 0272-5428.

[33] Merrick L. Furst, Jeffrey C. Jackson, and Sean W. Smith. 1991. Improved learning of AC^0 functions. In *Proceedings of the fourth annual workshop on Computational learning theory (COLT '91)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 317–325.

[34] Surbhi Goel, Aravind Gollakota, and Adam R. Klivans. 2020. Statistical-Query Lower Bounds via Functional Gradients. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[35] Oded Goldreich and Dana Ron. 2000. On Testing Expansion in Bounded-Degree Graphs. *Electron. Colloquium Comput. Complex.* 20 (2000).

[36] Shafi Goldwasser, Guy N. Rothblum, Jonathan Shafer, and Amir Yehudayoff. 2020. Interactive Proofs for Verifying Machine Learning. *Electron. Colloquium Comput. Complex.* (2020), 58.

[37] Aravind Gollakota, Adam R. Klivans, and Pravesh K. Kothari. 2023. A Moment-Matching Approach to Testable Learning and a New Characterization of Rademacher Complexity. In *Proceedings of the 55th annual ACM symposium on Theory of computing (STOC 2023 (to appear))*. Association for Computing Machinery.

[38] Parikshit Gopalan and Rocco A. Servedio. 2010. Learning and Lower Bounds for AC0 with Threshold Gates. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (Lecture Notes in Computer Science)*, Maria Serna, Ronen Shaltiel, Klaus Jansen, and José Rolim (Eds.). Springer, Berlin, Heidelberg, 588–601. https://doi.org/10.1007/978-3-642-15369-3_44

[39] Venkatesan Guruswami and Prasad Raghavendra. 2006. Hardness of Learning Halfspaces with Noise. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*. IEEE Computer Society, 543–552. https://doi.org/10.1109/FOCS.2006.33

[40] Thomas Hancock, Tao Jiang, Ming Li, and John Tromp. 1996. Lower bounds on learning decision lists and trees. *Information and Computation* 126, 2 (1996), 114–122.

[41] Prahladh Harsha, Adam Klivans, and Raghu Meka. 2010. An invariance principle for polytopes. In *Proceedings of the forty-second ACM symposium on Theory of computing (STOC '10)*. Association for Computing Machinery, New York, NY, USA, 543–552. https://doi.org/10.1145/1806689.1806764

[42] David Haussler. 1992. Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications. *Inf. Comput.* 100, 1 (1992), 78–150. https://doi.org/10.1016/0890-5401(92)90010-D

[43] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. 2008. Agnostically Learning Halfspaces. *SIAM J. Comput.* 37, 6 (2008), 1777–1805.

[44] D. M. Kane. 2010. The Gaussian Surface Area and Noise Sensitivity of Degree-d Polynomial Threshold Functions. In *2010 IEEE 25th Annual Conference on Computational Complexity*. 205–210. https://doi.org/10.1109/CCC.2010.27 ISSN: 1093-0159.

[45] Sushrut Karmalkar, Adam R. Klivans, and Pravesh Kothari. 2019. List-decodable Linear Regression. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. 7423–7432.

[46] Michael J. Kearns, Robert E. Schapire, and Linda Sellie. 1994. Toward Efficient Agnostic Learning. *Mach. Learn.* 17, 2-3 (1994), 115–141. https://doi.org/10.1007/BF00993468

[47] Adam R. Klivans, Philip M. Long, and Rocco A. Servedio. 2009. Learning Halfspaces with Malicious Noise. In *Automata, Languages and Programming, 36th International Colloquium, ICALP 2009, Rhodes, Greece, July 5-12, 2009, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 5555)*, Susanne Albers, Alberto Marchetti-Spaccamela, Yossi Matias, Sotiris E. Nikoletseas, and Wolfgang Thomas (Eds.). Springer, 609–621. https://doi.org/10.1007/978-3-642-02927-1_51

[48] A. R. Klivans, R. O'Donnell, and R. A. Servedio. 2002. Learning intersections and thresholds of halfspaces. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.* 177–186. https://doi.org/10.1109/SFCS.2002.1181894 ISSN: 0272-5428.

[49] Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. 2008. Learning Geometric Concepts via Gaussian Surface Area. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*. IEEE Computer Society, 541–550.

[50] N. Linial, Y. Mansour, and N. Nisan. 1989. Constant depth circuits, Fourier transform, and learnability. IEEE Computer Society, 574–579. https://doi.org/10.1109/SFCS.1989.63537

[51] Yishay Mansour. 1992. An O(n$^{\log\log n}$) learning algorithm for DNF under the uniform distribution. In *Proceedings of the fifth annual workshop on Computational learning theory (COLT '92)*. Association for Computing Machinery, New York, NY, USA, 53–61. https://doi.org/10.1145/130385.130391

[52] R. O'Donnell and R. A. Servedio. 2006. Learning monotone decision trees in polynomial time. In *21st Annual IEEE Conference on Computational Complexity (CCC'06)*. 13 pp.–225. https://doi.org/10.1109/CCC.2006.25 ISSN: 1093-0159.

[53] Ryan O'Donnell and Yu Zhao. 2018. On Closeness to k-Wise Uniformity. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2018, August 20-22, 2018 - Princeton, NJ, USA (LIPIcs, Vol. 116)*, Eric Blais, Klaus Jansen, José D. P. Rolim, and David Steurer (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 54:1–54:19. https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2018.54

[54] Liam Paninski. 2008. A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data. *IEEE Trans. Inf. Theory* 54, 10 (2008), 4750–4755.

[55] Ronald L. Rivest. 1987. Learning Decision Lists. *Mach. Learn.* 2, 3 (1987), 229–246. https://doi.org/10.1007/BF00058680

[56] Lloyd N Trefethen. 2019. *Approximation Theory and Approximation Practice, Extended Edition*. SIAM.

[57] K. Wimmer. 2010. Agnostically Learning under Permutation Invariant Distributions. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. 113–122. https://doi.org/10.1109/FOCS.2010.17 ISSN: 0272-5428.

Received 2022-11-07; accepted 2023-02-06