Proceedings of Machine Learning Research vol 196:1–35, 2024 37th Annual Conference on Learning Theory

Learning Intersections of Halfspaces with Distribution Shift: Improved Algorithms and SQ Lower Bounds

Adam Klivans KLIVANS@UTEXAS.EDU
University of Texas at Austin

Konstantinos Stavropoulos KSTAVROP@UTEXAS.EDU
University of Texas at Austin

Arsen Vasilyan VASILYAN@MIT.EDU
MIT

Editors: Shipra Agrawal and Aaron Roth

Abstract

Recent work of Klivans, Stavropoulos, and Vasilyan initiated the study of *testable learning with distribution shift* (TDS learning), where a learner is given labeled samples from training distribution $\mathcal{D}$, unlabeled samples from test distribution $\mathcal{D}'$, and the goal is to output a classifier with low error on $\mathcal{D}'$ whenever the training samples pass a corresponding test. Their model deviates from all prior work in that no assumptions are made on $\mathcal{D}'$. Instead, the test must accept (with high probability) when the marginals of the training and test distributions are equal.

Here we focus on the fundamental case of intersections of halfspaces with respect to Gaussian training distributions and prove a variety of new upper bounds including a $2^{(k/\epsilon)^{O(1)}}\mathsf{poly}(d)$-time algorithm for TDS learning intersections of k homogeneous halfspaces to accuracy ϵ (prior work achieved $d^{(k/\epsilon)^{O(1)}}$). We work under the mild assumption that the Gaussian training distribution contains at least an ϵ fraction of both positive and negative examples (ϵ-balanced). We also prove the first set of SQ lower-bounds for any TDS learning problem and show (1) the ϵ-balanced assumption is necessary for $\mathsf{poly}(d, 1/\epsilon)$-time TDS learning for a single halfspace and (2) a $d^{\tilde{\Omega}(\log 1/\epsilon)}$ lower bound for the intersection of two general halfspaces, even with the ϵ-balanced assumption.

Our techniques significantly expand the toolkit for TDS learning. We use dimension reduction and coverings to give efficient algorithms for computing a *localized* version of discrepancy distance, a key metric from the domain adaptation literature.

Keywords: testable learning, intersections of halfspaces, PAC learning, distribution shift, domain adaptation

1. Introduction

Distribution shift continues to be a major barrier for deploying AI models, especially in the health and bioscience domains. By far the most common approach to modeling distribution shift (or domain adaptation) is to bound the performance of a classifier in terms of some notion of distance between the training and test distributions (Ben-David et al., 2006; Mansour et al., 2009). These distances, however, are computationally intractable to estimate, as they are defined in terms of an enumeration over all classifiers from some class. As such, learners constrained to run in polynomial-time obtain no guarantees on the performance of a classifier (without making strong assumptions on the test distribution).

A recent work of Klivans, Stavropoulos, and Vasilyan (Klivans et al., 2023) departs from this paradigm and defines a model of *testable learning with distribution shift* (TDS learning). In this model, a learner first runs a test on labeled samples drawn from training distribution $\mathcal{D}$ and *unlabeled* samples drawn from test distribution $\mathcal{D}'$. No assumptions are made on $\mathcal{D}'$. If the test accepts, the learner outputs a classifier that is guaranteed to have low error with respect to $\mathcal{D}'$. Further, the test must accept (with high probability) whenever the marginal of $\mathcal{D}$ equals the marginal of $\mathcal{D}'$. It is clear that this model generalizes the traditional PAC model of learning (where $\mathcal{D}$ always equals $\mathcal{D}'$), and, as described in Klivans et al. (2023), obtaining efficient algorithms seems considerably more challenging. Giving positive results for TDS learning with running times that match known results in the traditional PAC model is therefore a best-case scenario.

1.1. Our Results

Here we focus on the intensely studied problem of learning intersections of halfspaces (or halfspace intersections) with respect to Gaussian distributions, where large gaps exist between the best known algorithms for TDS learning versus ordinary PAC learning. Our main contribution is a set of new positive results all of which greatly improve on prior work in TDS learning and in some cases match the best known bounds for PAC learning (see Tables 1 and 2 for precise statements of bounds). Our algorithm assumes that the training distribution contains at least an ϵ fraction of both positive and negative examples (ϵ-balanced), which turns out to be necessary, as we describe below.

Indeed, we provide the first set of SQ lower bounds for *any* problem in TDS learning (that was not already known in the traditional PAC model of learning). We show that no polynomial-time SQ algorithm can TDS learn a single halfspace unless the training distribution is ϵ-balanced. Further, we prove that no polynomial-time SQ algorithm can TDS learn the intersection of two general halfspaces, even if we assume the training distribution is ϵ-balanced. Taken together, these results considerably narrow the gap between efficient TDS learnability and PAC learnability for halfspace-based learning.

1.2. Techniques

TDS Learning through Covering the Solution Space. Our upper bounds are based on the idea of constructing a set of candidate output hypotheses that has three properties: (1) it has small size, (2) it contains one hypothesis with low test error and (3) all of the hypotheses in the set have low training error. Once such a cover is constructed, a small set of unlabeled data from the test distribution is sufficient to ensure that all of the members of the cover have low training error. This is possible by estimating the discrepancy distance between the test marginal and the Gaussian, but only with respect to the members of the cover, i.e., estimating the maximum probability of disagreement between pairs of elements of the cover under the test marginal. Since the cover is small (by (1)), this can be done efficiently and since all of the hypotheses have low training error (by (3)), the test should accept in the absence of distribution shift. If the test accepts, then all of the members have low disagreement with one hypothesis with low test error (by (2)) and they, hence, have low test error as well. The learner may then output any member of the cover.

Constructing Covers for Halfspace Intersections. Our method for covering the solution space for TDS learning halfspace intersections is based on two main ingredients. The first ingredient is access to an algorithm that uses training data and retrieves a low-dimensional subspace that is

	Type of Intersection	Run-time	Test Set Size	Reference
1	Homogeneous	$\mathrm{poly}(d)2^{\mathrm{poly}(\frac{k}{\epsilon})}$	$\mathrm{poly}(dk/\epsilon)$	Corollary 3
2	Homogeneous	$(\frac{dk}{\epsilon})^{O(k)} + d(\frac{k}{\epsilon})^{O(k^2)}$	$\mathrm{poly}(dk/\epsilon)$	Corollary 3
3	General	$d^{\mathrm{poly}(k/\epsilon)}$	$d^{\mathrm{poly}(k/\epsilon)}$	Klivans et al. (2023)
4	General	$d^3 2^{\mathrm{poly}(k/\epsilon)} +$ $d^{O(\log(\frac{k}{\epsilon}))}(\frac{k}{\epsilon})^{O(k^2)}$	$d^{O(\log(\frac{k}{\epsilon}))}$	Corollary 6
5	Homogeneous Non-Degenerate	$\mathrm{poly}(d)(\frac{k}{\epsilon})^{O(k^2)}$	$\mathrm{poly}(dk/\epsilon)$	Corollary 30

Table 1: Upper Bounds for TDS Learning ϵ-Balanced Intersections of k Halfspaces under $\mathcal{N}_d$. All bounds here improve on the best previous bound in row three. For *noise-free PAC learning* intersections of k halfspaces can be learned in time $(dk/\epsilon)^{O(k)}$ (Vempala, 2010b) and is the best known bound for small k. We nearly match this bound in row two above and provide an incomparable result in row four. In row five, we improve on all of these bounds under a non-degeneracy assumption on the intersection of halfspaces; see the Related Work section for a discussion.

	Halfspace Type	Assumption on Intersection	SQ Complexity
1	Homogeneous	Arbitrary	$\mathrm{poly}(d/\epsilon)$, for $k=1$
2	Homogeneous	Arbitrary	$d^{\omega_\epsilon(1)}$, for $k \geq 2$
3	Homogeneous	ϵ-Balanced	$\mathrm{poly}(d/\epsilon)$, for $k = \Theta(1)$
4	General	Arbitrary	$d^{\tilde{\Theta}(\log(1/\epsilon))}$, for $k=1$
5	General	ϵ-Balanced & $\Theta(1)$-non-degenerate	$d^{\tilde{\Theta}(\log(\frac{1}{\epsilon}))}$, for $k \geq 2$, $k = \Theta(1)$

Table 2: Statistical Query complexity (upper and lower) bounds for TDS Learning k-Halfspace Intersections under $\mathcal{N}_d$. No prior SQ lower bounds for any TDS learning problem were known. For the balance assumption, see Definition 17. For the non-degeneracy assumption, see Definition 26. Row 1 and the upper bound of row 4 are from Klivans et al. (2023). All other results are from this work: Theorem 12 (row 2), Corollary 30 (row 3), Theorem 8 (row 4), Theorem 15 (row 5, lower bound), Corollary 32 (row 5, upper bound). The lower bounds of rows 4, 5 hold for $d = O(\epsilon^{-1/4})$.

guaranteed to approximately contain (in terms of angular distance) each of the normal vectors that define the ground truth intersection. See the Related Work section for a more detailed discussion on subspace recovery algorithms. The second ingredient is a local halfspace disagreement tester, namely, a tester that takes as input a vector (and unlabelled test data) and certifies that all of the vectors that are geometrically close to the input define halfspaces with low disagreement to the one defined by the input under the test distribution. Such testers have been proposed in the literature of testable learning Gollakota et al. (2023a,b) and TDS learning Klivans et al. (2023), but, we provide an additional one for the case of general halfspaces. Equipped with both of these ingredients, we use a Euclidean cover for the sphere in the low-dimensional subspace retrieved and run the disagreement tester on each vector in the cover. We form a cover of the solution space with the desired properties by forming all possible intersections of halfspaces with normals in the Euclidean cover and keeping only those with low training error.

For general halfspaces, we also use an additional moment-matching tester which ensures that halfspaces with very high bias can be safely omitted from the output hypothesis, because the test distribution is certified to be sufficiently concentrated in every direction. This is important, because the training data does not reveal enough information for such halfspaces and, hence, it is not guaranteed that their normals will be approximately contained in the retrieved subspace.

SQ Lower Bounds for TDS Learning from Lower Bounds for NGCA. We prove our statistical query (SQ) lower bounds by reducing appropriate distribution testing problems to TDS learning. The distribution testing problems we consider fall in the category of Non-Gaussian Component Analysis (NGCA) where a distinguisher has access to an unknown distribution and is asked to distringuish whether the distribution is Gaussian or it is Gaussian in all but one hidden direction where the marginal satisfies certain problem-specific conditions. Diakonikolas et al. (2023a) provide SQ lower bounds for various instantiations of the problem.

We show that a TDS learner for general halfspaces can distinguish the Gaussian from any distribution that has some non-negligible mass far from the origin along some hidden direction. We then construct a distribution that is Gaussian in all but one direction along which the marginal (1) exactly matches moments with the standard Gaussian up to some degree and (2) assigns non-negligible mass far from the origin. To show approximate moment matching, we use a mass transportation argument and for exact moment matching, we use an argument based on the theory of Linear Programming from Diakonikolas et al. (2023b). Under these conditions, a generic tool from Diakonikolas et al. (2023a) implies an SQ lower bound for the distinguishing problem we constructed and hence an SQ lower bound for TDS learning. A similar construction gives a lower bound for intersections of two general halfspaces. For intersections of two homogeneous halfspaces, we reduce the problem of anti-concentration detection (whose SQ lower bound is given in Diakonikolas et al. (2023a)) to the corresponding TDS learning problem.

1.3. Related Work

Intersections of Halfspaces Learning intersections of halfspaces continues to be an important benchmark for algorithm design in learning theory with a long history of prior work (Long and Warmuth, 1994; Blum and Kannan, 1997; Klivans et al., 2004; Klivans and Sherstov, 2009; Klivans et al., 2009, 2008; Vempala, 2010b,a; Gopalan et al., 2012; Kane et al., 2013; Diakonikolas et al., 2018). Finding a fully polynomial-time algorithm for learning the intersection of k halfspaces in

d dimensions to accuracy ϵ remains a notorious open problem, even in the case of noise-free PAC learning with respect to Gaussian marginals.

The most relevant works here are Vempala (2010b) and Vempala (2010a) which both attempt to recover the subspace spanned by the k normals of the relevant halfspaces. This type of subspace recovery is a crucial ingredient for our work here, as we describe in the Techniques subsection above. In Vempala (2010b), an algorithm with running time and sample complexity $(dk/\epsilon)^{O(k)}$ is given for noise-free PAC learning with respect to log-concave marginals. In a follow-up work Vempala (2010a) claims an improved bound of $(k/\epsilon)^{O(k)}\mathsf{poly}(d)$. Unfortunately, this proof has a gap. In Appendix C.1 we provide a complete proof of a weaker result using the approach of Vempala (2010a), namely we obtain a $2^{O(k^2/\epsilon^2)}\mathsf{poly}(\mathsf{d},\mathsf{k})$ time algorithm for intersections of homogeneous halfspaces. If we take a non-degeneracy assumption on the ground truth intersection (see Appendix C.2), we prove that the gap can be fixed and we recover the $(k/\epsilon)^{O(k)}\mathsf{poly}(d)$ bound.

For large values of k, the best known bound of $d^{\tilde{O}(\log k/\epsilon^2)}$ for PAC or agnostic learning is due to Klivans et al. (2008), obtained using the Gaussian surface area/Hermite analysis approach. For TDS learning, Klivans et al. (2023) gave an algorithm with running time $d^{\tilde{O}(k^6/\epsilon^2)}$ that is improper and outputs a polynomial threshold function as the final hypothesis. In addition to improving their bounds on run-time (as described in Table 1), the algorithm we present here is proper: our learner gives an intersection of k halfspaces as its output hypothesis.

Distribution Shift/Domain Adaptation The field of domain adaptation considers problems very similar to the model introduced here. A learner is presented with labeled training samples, unlabeled test samples, and is required to output a classifier with low test error. The learner in traditional domain adaptation, however, is not allowed to reject. The area is too broad for us to survey here, and we refer the reader to Redko et al. (2020) and references therein. We highlight the works of Ben-David et al. (2006) and Mansour et al. (2009), which provide sample complexity upper bounds for domain adaptation in terms of *discrepancy distance*. It is proved in Klivans et al. (2023) that the notion of discrepancy distance also provides sample complexity guarantees for TDS learning. The first set of efficient algorithms for domain adaptation without taking strong assumptions on the test distribution were given by Klivans et al. (2023). We also note related work due to Goldwasser et al. (2020); Kalai and Kanade (2021); Goel et al. (2023) on PQ learning, a model formally shown to be harder than TDS learning in Klivans et al. (2023).

Testable Learning Although both the Testable Learning framework due to Rubinfeld and Vasilyan (2023) and TDS learning allow a learner to reject unless a training set passes a test, the models address very different issues and are formally incomparable. In testable learning, the goal is to certify that an *agnostic* learner has succeeded (or reject). In particular, (1) testable learning is trivial in the realizable (noise-free) framework (recall in this paper we work exclusively in a noise-free setting) and (2) testable learning does not allow for distribution shift. For a further comparison of the models see Klivans et al. (2023). We do make use of some general techniques from testable learning, as we describe in the Techniques section.

1.4. Preliminaries

For $\mathbf{v} \in \mathbb{R}^d, \tau \in \mathbb{R}$, we call a function of the form $\mathbf{x} \mapsto \mathrm{sign}(\mathbf{v} \cdot \mathbf{x})$ a homogeneous halfspace and a function of the form $\mathbf{x} \mapsto \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + \tau)$ a general halfspace over $\mathbb{R}^d$. An intersection of halfspaces is a function from $\mathbb{R}^d$ to $\{\pm 1\}$ of the form $\mathbf{x} \mapsto 2 \wedge_{i \in [k]} \mathbb{1}\{\mathbf{w}^i \cdot \mathbf{x} + \tau^i \geq 0\} - 1$, where $\mathbf{w}^i$ are called the normals of the intersection and τ^i the corresponding thresholds. Let $\mathcal{N}_d$ be the standard Gaussian in d dimensions. For a subspace $\mathcal{U}$, let $\mathrm{proj}_{\mathcal{U}}(\mathbf{w})$ be the orthogonal projection of a vector $\mathbf{w}$ on the subspace $\mathcal{U}$.

Learning Setup. We focus on the framework of **testable learning with distribution shift (TDS learning)** defined by Klivans et al. (2023). In particular, for a concept class $\mathcal{C} \subseteq \{\mathbb{R}^d \to \{\pm 1\}\}$, the learner $\mathcal{A}$ is given $\epsilon, \delta \in (0,1)$, a set S_{train} of labelled examples of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{D} = \mathcal{N}_d$ and $f^* \in \mathcal{C}$, as well a set X_{test} of unlabelled examples from an arbitrary test distribution $\mathcal{D}'$ and is asked to output a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ with the following guarantees.

(a) (Soundness.) With probability at least $1 - \delta$ over the samples $S_{\text{train}}, X_{\text{test}}$ we have:

If $\mathcal{A}$ accepts, then the output h satisfies $\mathbb{P}_{\mathbf{x} \sim \mathcal{D}'}[f^*(\mathbf{x}) \neq h(\mathbf{x})] \leq \epsilon$.

(b) (Completeness.) Whenever $\mathcal{D}' = \mathcal{N}_d$, $\mathcal{A}$ accepts w.p. at least $1 - \delta$ over $S_{\text{train}}, X_{\text{test}}$.

If the learner $\mathcal{A}$ enjoys the above guarantees, then $\mathcal{A}$ is called an (ϵ, δ)-TDS learner for $\mathcal{C}$ w.r.t. $\mathcal{N}_d$. Since the probability of success can be amplified through repetition (see (Klivans et al., 2023, Proposition C.1)), in what follows, we will provide algorithms with constant failure probability.

2. Proper TDS learners for Halfspace Intersections

2.1. Warm-up: Intersections of Homogeneous Halfspaces

Our first main result concerns the problem of TDS learning intersections of homogeneous halfspaces with respect to the Gaussian distribution. For a single homogeneous halfspace Klivans et al. (2023) showed that there is a fully polynomial-time TDS learner under Gaussian marginals. The learner crucially relied on the approximate recovery of the normal vector corresponding to the ground truth halfspace in terms of angular distance using training data. After obtaining a vector that is geometrically close to the ground truth, the learner used unlabelled test data to certify that any halfspace near the recovered one (and, hence, also the ground truth) does not significantly disagree with the recovered halfspace on the test distribution. Such a certificate can be obtained through appropriate localized testers that rely on low-degree moment estimation (introduced in the testable learning literature, see Gollakota et al. (2023a,b)).

We significantly generalize this approach beyond the case of a single halfspace and obtain improved TDS learners for intersections of any number of homogeneous halfspaces (as well as general halfspaces in Section 2.2). Our approach is once more to recover some information about the ground truth that can be measured in geometric terms. In particular, the appropriate notion of geometric recovery for the case of halfspace intersections is approximate subspace retrieval, namely, recovering a subspace that approximately contains all of the normals to the ground truth intersection, as defined below.

Definition 1 (Approximate Subspace Retrieval for Homogeneous Halfspaces) *We say that algorithm $\mathcal{A}$ (ϵ, δ)-retrieves the relevant subspace for $\mathcal{C}$ (whose elements are homogeneous halfspace intersections) under $\mathcal{N}_d$ if $\mathcal{A}$, upon receiving at least $m_{\mathcal{A}}$ examples of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}_d$ and $f^* \in \mathcal{C}$, outputs, w.p. at least $1 - \delta$ a subspace $\mathcal{U}$ such that for any normal $\mathbf{w}$ of f^* we have $\| \mathrm{proj}_{\mathcal{U}} \mathbf{w} \|_2 \geq 1 - \epsilon$.*

It turns out that the idea of approximate subspace retrieval has been explored in the literature of standard PAC learning, as it can be used to provide strong PAC learning guarantees and proper algorithms. We may, therefore, use existing results on approximate subspace retrieval (see Appendix C) as a first step of our TDS learning algorithm. Once we have obtained a low-dimensional subspace that approximately contains all the normals, we (1) generate a small cover of the candidate solution space, (2) acquire (using unlabeled test examples) a certificate that the cover contains a hypothesis with low test error and (3) bound the discrepancy distance (notion from domain adaptation) of the test marginal with the Gaussian, but only with respect to the candidate solution space. We obtain the following result, whose full proof can be found in Section D.1.

Theorem 2 (TDS Learning Intersections of Homogeneous Halfspaces) *Let $\mathcal{C}$ be a class whose elements are intersections of k homogeneous halfspaces on $\mathbb{R}^d$, $\epsilon \in (0,1)$ and $C \geq 1$ a sufficiently large constant. Assume that $\mathcal{A}$ $(\frac{\epsilon^3}{Ck^3}, 0.01)$-retrieves the relevant subspace for $\mathcal{C}$ under $\mathcal{N}_d$ with sample complexity $m_{\mathcal{A}}$. Then, there is an algorithm (Algorithm 3) that $(\epsilon, \delta = 0.02)$-TDS learns the class $\mathcal{C}$, using $m_{\mathcal{A}} + \tilde{O}(\frac{dk^2}{\epsilon^2})$ labeled training examples and $\tilde{O}(\frac{dk^2}{\epsilon^2})$ unlabelled test examples, calls $\mathcal{A}$ once, and uses additional time $\tilde{O}(\frac{d^3k^2}{\epsilon^2}) + d(k/\epsilon)^{O(k^2)}$.*

Algorithm 1: Proper TDS Learner for Homogeneous Halfspace Intersections

---

Input: Labelled set S_{train}, unlabelled set X_{test}, parameter ϵ

Set $\epsilon' = \frac{\epsilon^{3/2}}{Ck^{3/2}}$ and $\epsilon'' = \frac{\epsilon^6}{Ck^7}$ for some sufficiently large universal constant $C \geq 1$.

Run algorithm $\mathcal{A}$ on the set S_{train} and let $(\mathbf{v}^1, \dots, \mathbf{v}^k)$ be its output.

Let $\mathcal{U}$ be the subspace spanned by $(\mathbf{v}^1, \dots, \mathbf{v}^k)$ and consider the following sparse cover of $\mathcal{U}$:

$\mathcal{U}_{\epsilon''} = \{\frac{\mathbf{u}}{\|\mathbf{u}\|_2} : \mathbf{u} = \epsilon'' \sum_{i=1}^k j_i \mathbf{v}^i, j_i \in \mathbb{Z} \cap [-\frac{1}{\epsilon''}, \frac{1}{\epsilon''}], \|\mathbf{u}\|_2 \neq 0\}$

Reject and terminate if $\| \operatorname{Var}_{\mathbf{x}\sim X}(\mathbf{x})\|_2 \geq 2$.

for $\mathbf{u} \in \mathcal{U}_{\epsilon''}$ **do**

  Reject and terminate if $\mathbb{P}_{\mathbf{x}\sim X}[|\mathbf{u}\cdot\mathbf{x}| \leq 2\epsilon'^{2/3}] > 5\epsilon'^{2/3}$.

end

Let $\mathcal{F}$ contain the concepts $f : \mathbb{R}^d \to \{\pm 1\}$ of the form $f(\mathbf{x}) = 2\bigwedge_{i=1}^k \mathbb{1}\{\mathbf{u}^i \cdot \mathbf{x} \geq 0\} - 1$, where $\mathbf{u}^1, \dots, \mathbf{u}^k \in \mathcal{U}_{\epsilon''}$ and $\mathbb{P}_{(\mathbf{x},y)\sim S_{\text{train}}}[y \neq f(\mathbf{x})] \leq \epsilon/5$.

Reject and terminate if $\max_{f_1,f_2\in\mathcal{F}} \mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f_1(\mathbf{x}) \neq f_2(\mathbf{x})] > \epsilon/2$.

Otherwise, output $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$ for some $\widehat{f} \in \mathcal{F}$.

Before proving Theorem 2, we first describe how we can obtain the above algorithm $\mathcal{A}$.

Approximate Subspace Retrieval. To approximately recover the relevant subspace, we apply results from PAC learning (see Vempala (2010a,b)), which we state in Appendix C. For example, Vempala (2010a) uses a Gaussian variance reduction lemma (see Lemma 18) which states that if we truncate the Gaussian distribution on the positive region of some intersection of homogeneous halfspaces, then the variance of the resulting distribution along the directions that define the normals of the intersection is bounded away below 1 (for directions orthogonal to the span of the normals, the variance is 1). Unfortunately, in the original proof of Vempala (2010a), a (crucial) approximate version of the variance reduction lemma (similar to the last part of Lemma 18) is missing and hence it is not clear whether the claimed approximate subspace retrieval result is true. We provide in Sections C.1 and C.2 a full proof of the subspace retrieval lemma, but with the following caveat: we either (1) incur complexity that is exponential in $\mathrm{poly}(k/\epsilon)$ (see Section C.1) or (2) require some non-degeneracy assumption (see Section C.2).

We now give an overview of the proof of Theorem 2.

Stage I: Acquiring a Good Cover. A *good cover* is a list $\mathcal{F}$ of candidate hypotheses (i.e., halfspace intersections) that is guaranteed to contain some intersection with low test error *and* only contains intersections with low training error. We construct such a cover as follows.

1. Once we have obtained a(n orthonormal basis for a) subspace $\mathcal{U}$ such that every normal to the ground truth intersection is geometrically close to some vector in $\mathcal{U}$, we exhaustively cover the unit sphere in $\mathcal{U}$ (see Lemma 20) to obtain a list $\mathcal{U}'$ of $((\frac{k}{\epsilon})^{O(k)})$ candidate unit vectors that is guaranteed to contain, for each normal $\mathbf{w}$ of the ground truth intersection, some element $\mathbf{u}$, such that the angle between $\mathbf{w}$ and $\mathbf{u}$ is small.

2. We then certify that for each element $\mathbf{u}$ of $\mathcal{U}'$, all of the halfspaces whose normals are geometrically close to $\mathbf{u}$ have low disagreement with the halfspace defined by $\mathbf{u}$ on the *test distribution*. Such a certificate can be obtained by using tools (Lemma 21) from the literature of testable learning (see Gollakota et al. (2023a,b)); in fact we may use, here, the same tools that Klivans et al. (2023) utilized to obtain TDS learners for single homogeneous halfspaces.

3. We construct $\mathcal{F}$ by including all possible intersections, of at most k elements from $\mathcal{U}'$, that have low training error. Note that there is one element f in $\mathcal{F}$ such that its normals are (one-by-one) geometrically close to the normals of the ground truth. The previous test has ensured that f has low test error, since the probability that any halfspace in f disagrees with the corresponding true one is small.

Stage II: Estimating Discrepancy Distance. It remains to pick an element from $\mathcal{F}$ with low test error. However, we have only shown that there is one (unknown) element f in $\mathcal{F}$ with low test error. Note that since all of the elements of $\mathcal{F}$ have low training error, then the disagreement between each pair of elements in $\mathcal{F}$ should be small under the training marginal (and the test marginal as well if there was no distribution shift). Therefore, as a last step, we test that the disagreement between any pair of hypotheses in $\mathcal{F}$ is small under test data; otherwise, it is safe to reject. If the test accepts, all of the elements in $\mathcal{F}$ should also have low test error (since they mostly agree with f under test data). We stress that this last test corresponds to estimating the discrepancy distance between the test marginal $\mathcal{D}'$ and the Gaussian with respect to $\mathcal{F}$, i.e., the quantity

$$\mathrm{d}_{\mathrm{disc}}(\mathcal{D}', \mathcal{N}; \mathcal{F}) = \sup_{f_1, f_2 \in \mathcal{F}} \left| \mathbb{P}_{\mathbf{x} \sim \mathcal{D}'}[f_1(\mathbf{x}) \neq f_2(\mathbf{x})] - \mathbb{P}_{\mathbf{x} \sim \mathcal{N}_d}[f_1(\mathbf{x}) \neq f_2(\mathbf{x})] \right|$$

The discrepancy distance is a standard notion in domain adaptation (see, e.g., Mansour et al. (2009)), but involves an enumeration and it can be hard to compute. Since we only compute it with respect to a small set of candidate hypotheses, we can afford to brute force search over all pairs of functions. Combining our Theorem 2 with tools for approximate subspace retrieval (see Appendix C), we obtain the following upper bounds. For a more detailed version of the bounds, see Corollary 30.

Corollary 3 *The class of ϵ-balanced intersections of k homogeneous halfspaces on $\mathbb{R}^d$ can be ϵ-TDS learned in time $\mathrm{poly}(d)2^{\mathrm{poly}(k/\epsilon)}$ using $\mathrm{poly}(d)2^{\mathrm{poly}(k/\epsilon)}$ training examples and $\mathrm{poly}(dk/\epsilon)$ test examples. Moreover, it can be ϵ-TDS learned in time $(\frac{dk}{\epsilon})^{O(k)} + d(\frac{k}{\epsilon})^{O(k^2)}$ using $\tilde{O}(d)(\frac{k}{\epsilon})^{O(k)}$ training examples and $\mathrm{poly}(dk/\epsilon)$ test examples.*

2.2. Intersections of General Halfspaces

In the case of intersections of general halfspaces, we use a similar approach. However, the notion of approximate subspace retrieval of Definition 1 is too strong in this case, as there might be halfspaces that have very high bias and, therefore, it is not possible to obtain enough information about them unless we use a vast amount of training data. We, therefore, define the following relaxed version of approximate subspace retrieval, also used for PAC learning (see Vempala (2010a)).

Definition 4 (Approximate Subspace Retrieval for General Halfspaces) *We say that the algorithm $\mathcal{A}$ (ϵ, δ, T)-retrieves the relevant subspace for $\mathcal{C}$ (whose elements are halfspace intersections) under $\mathcal{N}_d$ if $\mathcal{A}$, upon receiving at least $m_{\mathcal{A}}$ examples of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}_d$ and $f^* \in \mathcal{C}$, outputs, w.p. at least $1-\delta$ a subspace $\mathcal{U}$ such that for any normal $\mathbf{w}$ corresponding to a halfspace $\{\mathbf{x} : \mathbf{w}\cdot\mathbf{x} + \tau \geq 0\}$ of f^* such that $\tau \leq T$, we have $\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}\|_2 \geq 1-\epsilon$.*

The notion of approximate subspace retrieval of Definition 4 is sufficient to design efficient PAC learners, since the halfspaces with large thresholds can be omitted without incurring a significant increase on the error under the training distribution (which, for PAC learning, is the same as the test distribution). In TDS learning, however, the test marginal is allowed to assign non-negligible mass to the unseen region of a hidden halfspace. In fact, this is a source of lower bounds for TDS learning as we show in Theorems 8 and 15.

Prior work on TDS learning (Klivans et al., 2023) focusing on the case of a single general halfspace, used a moment matching tester to ensure that the test marginal does not assign considerable mass to the unseen region of significantly biased halfspaces (as is the case under the Gaussian). Such tests incur a complexity of $d^{\Theta(\log(\frac{1}{\epsilon}))}$, which is essentially unavoidable (see Theorem 8). Note that by assuming that the ground truth is balanced (Definition 17), one can bypass the lower bound of Theorem 8 for TDS learning a single general halfspace. This is not the case, however, for intersections of even 2 general halfspaces (see Theorem 15), where the lower bound of $d^{\tilde{\Omega}(\log(1/\epsilon))}$ persists even under the balanced concepts assumption.

For TDS learning general halfspaces, we adopt a similar moment matching approach as the one used for a single general halfspace (see Klivans et al. (2023)) to ensure that the normals of the ground truth that are not represented by any element of the retrieved subspace (due to high bias) are not important even under the test distribution. Moreover, in order to acquire a certificate that we have a good cover (as per the previous section), we design a local halfspace disagreement tester that works even for general halfspaces (see Lemma 22). We obtain the following result (see Section D.2).

Theorem 5 (TDS Learning Intersections of General Halfspaces) *Let $\mathcal{C}$ be a class whose elements are intersections of k general halfspaces on $\mathbb{R}^d$, $\epsilon \in (0,1)$ and $C \geq 1$ a sufficiently large constant. Assume that $\mathcal{A}$ $(\frac{\epsilon^3}{Ck^3}, 0.01, 3\log^{1/2}(\frac{10k}{\epsilon}))$-retrieves the relevant subspace for $\mathcal{C}$ under $\mathcal{N}_d$ with sample complexity $m_{\mathcal{A}}$. Then, there is an algorithm (Algorithm 4) that $(\epsilon, \delta = 0.02)$-TDS learns the class $\mathcal{C}$, using $m_{\mathcal{A}} + \tilde{O}(\frac{dk^2}{\epsilon^2})$ labelled training examples and $d^{O(\log(k/\epsilon))}$ unlabelled test examples, calls $\mathcal{A}$ once and uses additional time $d^{O(\log(k/\epsilon))}(k/\epsilon)^{O(k^2)}$.*

We once more combine our Theorem 5 with results on approximate subspace retrieval (see Appendix C), to obtain the following upper bounds (see also Corollary 32).

Corollary 6 *The class of ϵ-balanced intersections of k general halfspaces on $\mathbb{R}^d$ can be ϵ-TDS learned in time $d^3 2^{\text{poly}(k/\epsilon)} + d^{O(\log(k/\epsilon))}(k/\epsilon)^{O(k^2)}$ using $\tilde{O}(d)2^{\text{poly}(k/\epsilon)}$ training examples and $d^{O(\log(k/\epsilon))}$ test examples.*

3. Statistical Query Lower Bounds

We will now provide a number of lower bounds for TDS learning in the statistical query model originally defined by Kearns (1998), which has been a standard framework for proving computational lower bounds in machine learning, and is known to capture most commonly used algorithmic techniques like gradient descent, moment methods, etc. (see, for example, Feldman et al. (2017a,b).

Definition 7 (Statistical Query Model) *Let $\varphi > 0$ and $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. We say that an algorithm $\mathcal{A}$ is a statistical query algorithm (SQ algorithm) with tolerance φ if $\mathcal{A}$ only has access to $\mathcal{D}$ through making a number of (adaptive) bounded queries of the form $q : \mathbb{R}^d \to [-1, 1]$, for each of which it receives a value $v \in \mathbb{R}$ with $|v - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[q(\mathbf{x})]| \leq \varphi$.*

Our approach is to reduce appropriate distribution testing problems to TDS learning and then show that these problems cannot be efficiently solved in the SQ framework, by applying recent results from Diakonikolas et al. (2023a) on Non-Gaussian Component Analysis.

3.1. General Halfspaces: A Tight Lower Bound

We prove the following theorem which gives a tight lower bound for TDS learning general halfspaces with respect to the Gaussian distribution in the SQ framework, since the lower bound matches the recent corresponding upper bound of Klivans et al. (2023).

Theorem 8 (SQ Lower Bound for TDS Learning a Single Halfspace) *For $\epsilon > 0$, set $d = \epsilon^{-1/4}$. Then, for all sufficiently small ϵ, the following is true. Let $\mathcal{A}$ be a TDS learning algorithm for general halfspaces over $\mathbb{R}^d$ w.r.t. $\mathcal{N}_d$, with accuracy parameter ϵ and success probability at least 0.95. Further, suppose that $\mathcal{A}$ obtains at most $d^{\frac{\log 1/\epsilon}{\log\log 1/\epsilon}}$ samples from the training distribution and accesses the testing distribution via $2^{d^{o(1)}}$ SQ queries of precision $\varphi > 0$ (the SQ queries are allowed to depend on the training samples). Then, the tolerance φ has to be at most $d^{-\Omega(\frac{\log 1/\epsilon}{\log\log 1/\epsilon})}$.*

We first define an appropriate distribution testing problem which can be reduced to TDS learning general halfspaces. In particular, the distribution testing problem we define amounts to testing whether a distribution to which we have sample access assigns too much mass to some halfspace compare to the mass assigned by the Gaussian.

Definition 9 (Biased Halfspace Detection Problem) *Let $0 \leq \alpha \leq \beta \leq 1$ The (α, β)-biased halfspace detection problem is the task of distinguishing the d-dimensional standard Gaussian distribution from any distribution $\mathcal{D}$ over $\mathbb{R}^d$ for which there exist $\mathbf{v}$ in $\mathbb{R}^d$ and τ in $\mathbb{R}$ satisfying*

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\cdot\mathbf{v} \geq \tau] \geq \beta \quad \textit{and} \quad \mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\mathbf{x}\cdot\mathbf{v} \geq \tau] \leq \alpha$$

The idea is that if one has a TDS learner for general halfspaces, then the TDS learner must also work when the training examples are drawn from a Gaussian and labelled by the constant hypothesis

-1. In this case, the learner cannot extract any information about the training data, except from the fact that they correspond to a halfspace with very high bias (but the direction remains completely unspecified). If the test distribution assigns a lot of mass on the positive region of the halfspace, then the error would be large and the TDS learner will reject. On the other hand, if the test distribution is the Gaussian, the TDS learner will accept. Hence, the TDS learner would solve the biased halfspace detection problem. We obtain the following quantitative result, whose formal proof can be found in Section E.1.

Proposition 10 (Biased Halfspace Detection via TDS Learning) *Let $\mathcal{A}$ be a TDS learning algorithm for general halfspaces over $\mathbb{R}^d$ w.r.t. $\mathcal{N}_d$ with accuracy parameter ϵ and success probability at least 0.95. Suppose $\mathcal{A}$ obtains at most m samples from the training distribution and accesses the test distribution via N SQ queries of tolerance φ (the SQ queries are allowed to depend on the training samples). Then, there exists an algorithm $(\frac{1}{100m}, 10\epsilon)$-biased halfspace detection that uses $N+1$ SQ queries of tolerance $\min(\varphi, \epsilon)$ and has success probability at least 0.8.*

In order to complete the proof of Theorem 8, it remains to show that the biased halfspace detection problem is hard in the SQ framework. To this end, we use a powerful tool from recent work on Non-Gaussian Component Analysis by Diakonikolas et al. (2023a), which states that distinguishing the Gaussian from a distribution which is Gaussian in all but one hidden direction is hard for SQ algorithms, whenever the marginal in this direction is guaranteed to match the low degree moments of the Gaussian (see Theorem 37). For our purposes, it is sufficient to construct a one-dimensional distribution that matches low degree moments with the standard Gaussian, but assigns non negligible mass far from the origin. We obtain the following result whose proof can be found in Section E.1.

Proposition 11 (SQ Lower Bound for Biased Halfspace Detection) *For $\epsilon > 0$, set $d = \frac{1}{\epsilon^{1/4}}$. Then, for all sufficiently small ϵ, the following is true. Suppose that $\mathcal{A}$ is an SQ algorithm for $(d^{-\ln(1/\epsilon)}, 10\epsilon)$-biased halfspace detection problem over $\mathbb{R}^d$, and $\mathcal{A}$ has a success probability of at least $2/3$. Then, $\mathcal{A}$ either has to use SQ tolerance of $d^{-\Omega(\frac{\log 1/\epsilon}{\log\log 1/\epsilon})}$, or make $2^{d^{\Omega(1)}}$ SQ queries.*

3.2. Intersections of Two Homogeneous Halfspaces

The following theorem demonstrates that, although TDS learning a single homogeneous halfspace with respect to the Gaussian distribution admits fully polynomial time algorithms (see Klivans et al. (2023)), for intersections of two homogeneous halfspaces, there is no polynomial-time SQ algorithm. Notably, the construction corresponds to a highly unbalanced intersection, so the lower bound does not hold for the problem of TDS learning balanced intersections.

Theorem 12 (SQ Lower Bound for TDS Learning Two Homogeneous Halfspaces) *Let $\epsilon > 0$ with $\epsilon \in (0, 1/10)$ and let $\mathcal{A}$ be a TDS learning algorithm for learning intersections of 2 homogeneous halfspaces over $\mathbb{R}^d$ w.r.t. $\mathcal{N}_d$ with accuracy ϵ and success probability at least 0.95. Then $\mathcal{A}$ either makes some query of tolerance $\varphi = d^{-\omega_\epsilon(1)}$ to the test distribution or runs in time $d^{\omega_\epsilon(1)}$.*

To prove our result, we use an SQ lower bound for detecting anti-concentration (AC) from Diakonikolas et al. (2023a).

Theorem 13 (SQ Lower Bound for Detecting AC, Theorem 1.10 in Diakonikolas et al. (2023a)) *Let $\epsilon \in (0, 1/2)$. Any SQ algorithm with SQ access to either (1) $\mathcal{N}_d$ or (2) some distribution $\mathcal{D}'$ that assigns mass at least ϵ on some subspace of dimension $d-1$ and distinguishes the two cases w.p. at least $2/3$, either uses $2^{d^{\Omega(1)}}$ queries, or uses a query with tolerance at most $d^{-\omega_\epsilon(1)}$.*

It remains to reduce the AC detection problem to the problem of TDS learning intersections of two homogeneous halfspaces. The idea is to use an intersection of two almost opposite halfspaces, whose positive region effectively coincides with half of the subspace where $\mathcal{D}'$ has non negligible mass. Therefore, upon acceptance, the output function should take the value 1 with non-negligible probability only if the unknown distribution is $\mathcal{D}'$, which implies that we have solved the distinguishing problem. See Section E.2 for a proof.

Remark 14 *Under the balance assumption, our algorithms achieve polynomial-time performance for learning intersections of $k = O(1)$ homogeneous halfspaces (see Corollary 30). This demonstrates the importance of the balance condition on the training data.*

3.3. Balanced Intersections of Two General Halfspaces

We now provide an SQ lower bound for TDS learning balanced (see Definition 17) intersections of two general halfspaces. The lower bound demonstrates that the balance condition cannot always mitigate the obstacles of TDS learning due to hard examples that are trivial for PAC learning. In particular, the hard example here is an intersection of two halfspaces, where one of them is known and the other one is orthogonal to the first and is effectively irrelevant for the intersection under the Gaussian measure. For PAC learning, this implies that the second halfspace can be safely ignored, but for TDS learning, the hidden halfspace is a source of SQ lower bounds as demonstrated below.

Theorem 15 (SQ Lower bound for TDS Learning Halfspace Intersections) *For $\epsilon > 0$, set $d = \epsilon^{-1/4}$. Then, for all sufficiently small ϵ, the following is true. Let $\mathcal{A}$ be a TDS learning algorithm for $\frac{1}{3}$-balanced intersections of 2 general halfspaces over $\mathbb{R}^d$ w.r.t. $\mathcal{N}_d$, with accuracy parameter ϵ and success probability at least 0.95. Further, suppose that $\mathcal{A}$ obtains at most $d^{\frac{\log 1/\epsilon}{\log\log 1/\epsilon}}$ samples from the training distribution and accesses the testing distribution via $2^{d^{o(1)}}$ SQ queries of precision $\varphi > 0$ (the SQ queries are allowed to depend on the training samples). Then, the tolerance φ has to be at most $d^{-\Omega(\frac{\log 1/\epsilon}{\log\log 1/\epsilon})}$.*

The idea is similar to the one used for the proof of Theorem 8. We once more prove a general reduction of the biased halfspace detection problem to TDS learning. The hard instance corresponds once more (as for the proof of Theorem 8) to the detection problem where the unknown distribution is either (1) the standard Gaussian or (2) some distribution $\mathcal{D}'$ that assigns non-trivial mass in the negative region of a halfspace $H_1 = \{\mathbf{x} : \mathbf{v} \cdot \mathbf{x} + \tau \geq 0\}$ for some appropriately large τ.

The reduction of the hard instance to TDS learning follows closely the proof of Proposition 10 (see Appendix E.1.1), but we run the TDS algorithm twice, once using training data of the form $(\mathbf{x}, \mathrm{sign}(\mathbf{u} \cdot \mathbf{x}))$ with $\mathbf{x} \sim \mathcal{N}_d$ and $\mathbf{u}$ some random vector in $\mathbb{S}^{d-1}$ and another one with training data of the form $(\mathbf{x}, \mathrm{sign}(-\mathbf{u} \cdot \mathbf{x}))$, $\mathbf{x} \sim \mathcal{N}_d$.

For each of the executions of the TDS algorithm, the training data are consistent (w.h.p.) with the unknown intersection defined by the halfspaces $H_1 = \{\mathbf{x} : \mathbf{v} \cdot \mathbf{x} + \tau \geq 0\}$ and $H_2 = \{\mathbf{x} : \mathbf{u} \cdot \mathbf{x} \geq 0\}$ (or $\bar{H}_2 = \{\mathbf{x} : -\mathbf{u} \cdot \mathbf{x} \geq 0\}$). If the TDS algorithm rejects, then we have a certificate that the marginal was not the Gaussian. If the TDS algorithm accepts, then we may use one SQ query for the probability that the output function is positive. If $\mathcal{D}'$ was the Gaussian, then this probability should be very close to $1/2$. Otherwise, it should be bounded away from $1/2$ for at least one of the executions ($\mathcal{D}'$ assigns non-trivial mass in the negative region of H_1, so it must assign non-trivial mass to either $H_2 \setminus H_1$ or $\bar{H}_2 \setminus H_1$). Hence, the pair of our SQ queries (one for each execution) will indicate the answer to the biased halfspace detection problem.

Remark 16 *Note that the lower bound of Theorem 15 holds even for the problem of TDS learning* 2*-non-degenerate intersections of two halfspaces (according to Definition 26). Under the non-degeneracy assumption, our algorithms achieve improved performance (see Corollary 32) and, in particular, the lower bound of Theorem 15 is essentially tight (*$d^{\tilde{\Theta}(\log(1/\epsilon))}$*) for TDS learning* $\Theta(1)$*-non-degenerate,* $\mathrm{poly}(\epsilon)$*-balanced intersections of* $k = O(1)$ *halfspaces.*

Acknowledgments

We thank the anonymous reviewers of COLT 2024 for their constructive feedback. Adam Klivans was supported by NSF award AF-1909204 and the NSF AI Institute for Foundations of Machine Learning (IFML). Konstantinos Stavropoulos was supported by the NSF AI Institute for Foundations of Machine Learning (IFML) and by scholarships from Bodossaki Foundation and Leventis Foundation. Arsen Vasilyan was supported in part by NSF awards CCF-2006664, DMS-2022448, CCF-1565235, CCF-1955217, CCF-2310818, Big George Fellowship and Fintech@CSAIL. Work done in part while visiting UT Austin.

References

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Avrim Blum and Ravindran Kannan. Learning an intersection of a constant number of halfspaces over a uniform distribution. *J. Comput. Syst. Sci.*, 54(2):371–380, 1997. doi: 10.1006/JCSS.1997.1475. URL https://doi.org/10.1006/jcss.1997.1475.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In Ilias Diakonikolas, David Kempe, and Monika Henzinger, editors, *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1061–1073. ACM, 2018. doi: 10.1145/3188745.3188754. URL https://doi.org/10.1145/3188745.3188754.

Ilias Diakonikolas, Daniel Kane, Lisheng Ren, and Yuxin Sun. Sq lower bounds for non-gaussian component analysis with weaker assumptions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. Sq lower bounds for learning mixtures of separated and bounded covariance gaussians. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2319–2349. PMLR, 12–15 Jul 2023b. URL https://proceedings.mlr.press/v195/diakonikolas23b.html.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2): 1–37, 2017a.

Vitaly Feldman, Cristobal Guzman, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1265–1277. SIAM, 2017b.

Tadahisa Funaki and Kou Toukairin. Dynamic approach to a stochastic domination: the fkg and brascamp-lieb inequalities. *Proceedings of the American Mathematical Society*, 135(6):1915–1922, 2007.

Surbhi Goel, Steve Hanneke, Shay Moran, and Abhishek Shetty. Adversarial resilience in sequential prediction via abstention. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. *Advances in Neural Information Processing Systems*, 33:15859–15870, 2020.

Aravind Gollakota, Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. An efficient tester-learner for halfspaces. *arXiv preprint arXiv:2302.14853*, 2023a.

Aravind Gollakota, Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Tester-learners for halfspaces: Universal algorithms. *37th Conference on Neural Information Processing Systems (NeurIPS 2023, to appear).*, 2023b.

Parikshit Gopalan, Adam R. Klivans, and Raghu Meka. Learning functions of halfspaces using prefix covers. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, volume 23 of *JMLR Proceedings*, pages 15.1–15.10. JMLR.org, 2012. URL http://proceedings.mlr.press/v23/gopalan12/gopalan12.pdf.

Adam Tauman Kalai and Varun Kanade. Efficient learning with arbitrary covariate shift. In *Algorithmic Learning Theory*, pages 850–864. PMLR, 2021.

Daniel M. Kane, Adam R. Klivans, and Raghu Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 522–545. JMLR.org, 2013. URL http://proceedings.mlr.press/v30/Kane13.html.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Adam Klivans, Philip Long, and Alex Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. pages 588–600, 01 2009. ISBN 978-3-642-03684-2. doi: 10.1007/978-3-642-03685-9_44.

Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.*, 68(4):808–840, 2004. doi: 10.1016/J.JCSS.2003.11.002. URL https://doi.org/10.1016/j.jcss.2003.11.002.

Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. 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*, pages 541–550. IEEE Computer Society, 2008. doi: 10.1109/FOCS.2008.64. URL https://doi.org/10.1109/FOCS.2008.64.

Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. *arXiv preprint arXiv:2311.15142*, 2023.

Philip M. Long and Manfred K. Warmuth. Composite geometric concepts and polynomial predictability. *Inf. Comput.*, 113(2):230–252, 1994. doi: 10.1006/INCO.1994.1071. URL https://doi.org/10.1006/inco.1994.1071.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*, Montréal, Canada, 2009. URL http://www.cs.nyu.edu/~mohri/postscript/nadap.pdf.

Raghu Meka. personal communication, 2010.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

Ronitt Rubinfeld and Arsen Vasilyan. Testing distributional assumptions of learning algorithms. *Proceedings of the fifty-fifth annual ACM Symposium on Theory of Computing*, 2023.

Santosh S Vempala. Learning convex concepts from gaussian distributions with pca. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 124–130. IEEE, 2010a.

Santosh S Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *Journal of the ACM (JACM)*, 57(6):1–14, 2010b.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Appendix A. Notation and Basic Definitions

We let $\mathbb{R}^d$ be the d-dimensional Euclidean space. For a distribution $\mathcal{D}$ over $\mathbb{R}^d$, we use $\mathbb{E}_{\mathcal{D}}$ (or $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}$) to refer to the expectation over distribution $\mathcal{D}$ and for a given (multi)set X, we use $\mathbb{E}_X$ (or $\mathbb{E}_{\mathbf{x}\sim X}$) to refer to the expectation over the uniform distribution on X (i.e., $\mathbb{E}_{\mathbf{x}\sim X}[g(\mathbf{x})] = \frac{1}{|X|}\sum_{\mathbf{x}\in X} g(\mathbf{x})$, counting possible duplicates separately). For $\mathbf{x} \in \mathbb{R}^d$ where $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d)$ and for $\alpha \in \mathbb{N}^d$, we denote with $\mathbf{x}^\alpha$ the product $\prod_{i\in[d]} \mathbf{x}_i^{\alpha_i}$. We denote with $\mathbb{S}^{d-1}$ the $d-1$ dimensional sphere on $\mathbb{R}^d$. For any $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$, we denote with $\mathbf{v}_1 \cdot \mathbf{v}_2$ the inner product between $\mathbf{v}_1$ and $\mathbf{v}_2$ and we let $\measuredangle(\mathbf{v}_1, \mathbf{v}_2)$ be the angle between the two vectors, i.e., the quantity $\theta \in [0, \pi]$ such that $\|\mathbf{v}_1\|_2\|\mathbf{v}_2\|_2 \cos(\theta) = \mathbf{v}_1 \cdot \mathbf{v}_2$. Let $\mathrm{Var}_{\mathbf{x}}(\mathbf{v} \cdot \mathbf{x})$ denotes the variance of random variable $\mathbf{v} \cdot \mathbf{x}$, for some vector $\mathbf{v} \in \mathbb{R}^d$. For $\mathbf{v} \in \mathbb{R}^d, \tau \in \mathbb{R}$, we call a function of the form $\mathbf{x} \mapsto \mathrm{sign}(\mathbf{v} \cdot \mathbf{x})$ a homogeneous halfspace and a function of the form $\mathbf{x} \mapsto \mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+\tau)$ a general halfspace over $\mathbb{R}^d$. An intersection of halfspaces is a function from $\mathbb{R}^d$ to $\{\pm 1\}$ of the form $\mathbf{x} \mapsto 2\wedge_{i\in[k]} \mathbb{1}\{\mathbf{w}^i \cdot \mathbf{x} + \tau^i \geq 0\} - 1$, where $\mathbf{w}^i$ are called the normals of the intersection and τ^i the corresponding thresholds.

Learning Setup. We focus on the framework of **testable learning with distribution shift (TDS learning)** defined by Klivans et al. (2023). In particular, for a concept class $\mathcal{C} \subseteq \{\mathbb{R}^d \to \{\pm 1\}\}$, the learner $\mathcal{A}$ is given $\epsilon, \delta \in (0, 1)$, a set S_{train} of labelled examples of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{D} = \mathcal{N}_d$ and $f^* \in \mathcal{C}$, as well a set X_{test} of unlabelled examples from an arbitrary test distribution $\mathcal{D}'$ and is asked to output a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ with the following guarantees.

(a) (Soundness.) With probability at least $1 - \delta$ over the samples $S_{\text{train}}, X_{\text{test}}$ we have:

If $\mathcal{A}$ accepts, then the output h satisfies $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}'}[f^*(\mathbf{x}) \neq h(\mathbf{x})] \leq \epsilon$.

(b) (Completeness.) Whenever $\mathcal{D}' = \mathcal{N}_d$, $\mathcal{A}$ accepts w.p. at least $1 - \delta$ over $S_{\text{train}}, X_{\text{test}}$.

If the learner $\mathcal{A}$ enjoys the above guarantees, then $\mathcal{A}$ is called an (ϵ, δ)-TDS learner for $\mathcal{C}$ w.r.t. $\mathcal{N}_d$. Since the probability of success can be amplified through repetition (see (Klivans et al., 2023, Proposition C.1)), in what follows, we will provide algorithms with constant failure probability.

For our upper bounds, we will make use of a balanced concepts condition, whose importance we justify through appropriate lower bounds (see Sections 3.1 and 3.2). In particular, we will assume that the ground truth $(\mathcal{D}, f^*)$ is sufficiently balanced, meaning that positive and negative examples from the training data both have sufficiently large frequency.

Definition 17 (Balance Condition) *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \{\pm 1\}$. For $\eta \in (0, 1/2]$, we say that f is η-balanced with respect to $\mathcal{D}$ if*

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x}) = 1] \in [\eta, 1 - \eta]$$

For a concept class $\mathcal{C} \subseteq \{\mathbb{R}^d \to \{\pm 1\}\}$, we denote with $\mathcal{C}_\eta$ the η-balanced version of $\mathcal{C}$, i.e., the subset of $\mathcal{C}$ that contains the elements that are η-balanced.

Note that the algorithm can check whether the ground truth is balanced using training data and, therefore, detect possible failure due to imbalance (i.e., the condition is testable).

Appendix B. Additional Tools

Our positive results build on the dimension reduction technique of Vempala (2010a) for PAC learning intersections of halfspaces and low-dimensional convex sets through principal component analysis (PCA), which is based on the following Gaussian variance reduction lemma. Note that although the first two parts of the lemma were known (see e.g., Vempala (2010a)), the last part (which gives variance reduction for any vector that has some correlation with a normal) is proven here. In fact, this more general form of the lemma is important even for the results in Vempala (2010a) (although it is missing from the original paper).

Lemma 18 (Variance Reduction, variant of Lemma 4.7 in Vempala (2010a)) *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be an intersection of halfspaces and let $\mathcal{N}_d|_{\mathcal{K}}$ be the truncation of the standard Gaussian distribution in d dimensions $\mathcal{N}_d$ to $\mathcal{K}$. For any $\mathbf{u} \in \mathbb{S}^{d-1}$, we have $\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}(\mathbf{u}\cdot\mathbf{x}) \leq 1$. Moreover, if for some $T \in \mathbb{R}$ the halfspace $\{\mathbf{x} : \mathbf{u}\cdot\mathbf{x} + T \geq 0\}$ is one of the defining halfspaces of the intersection then, we have variance reduction along $\mathbf{u}$, i.e., $\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}(\mathbf{u}\cdot\mathbf{x}) \leq 1 - \frac{1}{C}e^{-\frac{1}{2}(\max\{0,T\})^2}$ for a sufficiently large universal constant $C > 0$. Furthermore, for any $\epsilon \in (0, \frac{1}{4})$ and any $\mathbf{u}' \in \mathbb{S}^{d-1}$ with $\mathbf{u}\cdot\mathbf{u}' \geq \epsilon$, for a sufficiently large constant $C' > 0$, if $\eta = \mathbb{P}_{\mathcal{N}_d}[\mathbf{x} \in \mathcal{K}]$ we have*

$$\mathop{\mathrm{Var}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}(\mathbf{u}'\cdot\mathbf{x}) \leq 1 - \big(\eta e^{-T^2}/2\big)^{\frac{C'}{\epsilon^2}}$$

Proof The first two parts follow from Cafarelli's theorem, see e.g. Theorem 3.1 in Funaki and Toukairin (2007) where one may set the function ψ to be a quadratic function within the interval $(-T, \infty)$ and either 0 outside it when $T < 0$ or a linear function tangent to the graph of $y = x^2$ at the point $x = T$ if $T \geq 0$[1].

For the last part, we will introduce an artificial halfspace in the direction of $\mathbf{u}'$ and we will link the variance in the direction of $\mathbf{u}'$ under the truncation of the Gaussian on the initial intersection to the variance under the new (artificial) truncation. In particular, let $\mathcal{K}'$ be the set $\mathcal{K} \cap \{\mathbf{u}'\cdot\mathbf{x} + \theta \geq 0\}$, where $\theta > 0$ is a parameter of our choice. We then have $\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}'}}(\mathbf{x}) \leq 1 - \frac{1}{C}\exp(-\theta^2/2)$, by the previous part of the lemma. However, we are interested in the quantity $\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}(\mathbf{x})$. We have the following

$$\begin{aligned}
\mathop{\mathrm{Var}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}(\mathbf{x}) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}[(\mathbf{u}'\cdot\mathbf{x})^2] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}[\mathbf{u}'\cdot\mathbf{x}]^2 \\
&= \underbrace{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}[(\mathbf{u}'\cdot\mathbf{x})^2\,\mathbb{1}\{\mathbf{x}\in\mathcal{K}'\}]}_{s_1} + \underbrace{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}[(\mathbf{u}'\cdot\mathbf{x})^2\,\mathbb{1}\{\mathbf{x}\notin\mathcal{K}'\}]}_{s_2} \\
&\quad - (\underbrace{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}[(\mathbf{u}'\cdot\mathbf{x})\,\mathbb{1}\{\mathbf{x}\in\mathcal{K}'\}]}_{\mu_1} + \underbrace{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}[(\mathbf{u}'\cdot\mathbf{x})\,\mathbb{1}\{\mathbf{x}\notin\mathcal{K}'\}]}_{\mu_2})^2
\end{aligned}$$

1. This choice of ψ is due to Raghu Meka (Meka, 2010).

For the first term s_1, we have $s_1 \le \mathbb{E}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}'}}[(\mathbf{u}'\cdot\mathbf{x})^2]$. For the second term s_2, we have

$$
\begin{aligned}
s_2 &= \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{N}_d}[(\mathbf{u}'\cdot\mathbf{x})^2\,\mathbb{1}\{\mathbf{x}\in\mathcal{K}\setminus\mathcal{K}'\}]}{\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\mathbf{x}\sim\mathcal{K}]} \\
&\le \frac{1}{\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}]}\cdot \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_d}\Big[(\mathbf{u}'\cdot\mathbf{x})^2\,\mathbb{1}\Big\{\mathbf{u}'\cdot\mathbf{x}+\theta<0, \mathbf{v}\cdot\mathbf{x} > \underbrace{\frac{\theta}{\tan\cos^{-1}\epsilon}-\frac{T}{\sin\cos^{-1}\epsilon}}_{\gamma}\Big\}\Big],
\end{aligned}
$$

where the inequality follows from the fact that for any $\mathbf{x}\in\mathcal{K}$ we have $\mathbf{u}\cdot\mathbf{x}+T\ge 0$ and for any $\mathbf{x}\notin\mathcal{K}'$ we have $\mathbf{u}'\cdot\mathbf{x}+\theta<0$, where $\mathbf{v}=\frac{\mathbf{u}-(\mathbf{u}\cdot\mathbf{u}')\mathbf{u}'}{\|\mathbf{u}-(\mathbf{u}\cdot\mathbf{u}')\mathbf{u}'\|_2}$. Hence, by bounding the Gaussian integral of the above inequality (note that $\mathbf{u}'\perp\mathbf{v}$), we obtain that for some sufficiently large constant $C'>0$ we have $s_2\le\frac{1}{\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}]}C'\theta^2 e^{-\frac{1}{2}\theta^2-\frac{1}{2}\gamma^2}$. For the term μ_1 we have

$$
\begin{aligned}
\mu_1 &= \mathop{\mathbb{E}}_{\mathcal{N}_d|_{\mathcal{K}'}}[\mathbf{u}'\cdot\mathbf{x}]\cdot\big(1-\mathop{\mathbb{P}}_{\mathcal{N}_d|_{\mathcal{K}}}[\mathbf{x}\notin\mathcal{K}']\big) \\
&= \mathop{\mathbb{E}}_{\mathcal{N}_d|_{\mathcal{K}'}}[\mathbf{u}'\cdot\mathbf{x}]\cdot\Big(1-\underbrace{\frac{\mathbb{P}_{\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}\setminus\mathcal{K}']}{\mathbb{P}_{\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}]}}_{\xi}\Big)
\end{aligned}
$$

Therefore, we have that $\mu_1^2\ge\mathbb{E}_{\mathcal{N}_d|_{\mathcal{K}'}}[\mathbf{u}'\cdot\mathbf{x}]^2-2\xi\,\mathbb{E}_{\mathcal{N}_d|_{\mathcal{K}'}}[\mathbf{u}'\cdot\mathbf{x}]$. Additionally, we have that $\mathbb{E}_{\mathcal{N}_d|_{\mathcal{K}'}}[\mathbf{u}'\cdot\mathbf{x}]=\frac{1}{\mathbb{P}_{\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}']}\cdot\mathbb{E}_{\mathcal{N}_d}[(\mathbf{u}'\cdot\mathbf{x})\,\mathbb{1}\{\mathbf{x}\in\mathcal{K}'\}]\le\frac{1}{(1-\xi)\,\mathbb{P}_{\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}]}(\mathbb{E}_{\mathcal{N}_d}[(\mathbf{u}'\cdot\mathbf{x})^2\,\mathbb{1}\{\mathbf{x}\in\mathcal{K}'\}])^{1/2}$ which implies that $\mu_1^2\ge\mathbb{E}_{\mathcal{N}_d|_{\mathcal{K}'}}[\mathbf{u}'\cdot\mathbf{x}]^2-\frac{2\xi}{(1-\xi)\,\mathbb{P}_{\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}]}$. Note that the quantity $\mathbb{P}_{\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}\setminus\mathcal{K}']$ is bounded by $\mathbb{P}_{\mathcal{N}_d}[\mathbf{u}'\cdot\mathbf{x}+\theta<0,\mathbf{v}\cdot\mathbf{x}>\gamma]\le e^{-\frac{1}{2}\theta^2-\frac{1}{2}\gamma^2}$.

The term $2\mu_1\mu_2$ can be bounded similarly (observe that $\mu_2\le s_2^{1/2}$). Hence, overall, we have

$$
\mathop{\mathrm{Var}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}(\mathbf{u}'\cdot\mathbf{x})\le\mathop{\mathrm{Var}}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}'}}(\mathbf{u}'\cdot\mathbf{x})+\Big(\frac{C'\theta^2}{\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}]}+\frac{C'}{\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}]^2}\Big)\cdot e^{-\frac{1}{2}\theta^2-\frac{1}{2}\gamma^2}
$$

Recall that $\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}'}}(\mathbf{u}'\cdot\mathbf{x})\le 1-\frac{1}{C}e^{-\frac{1}{2}\theta^2}$ and hence by picking $\theta=C''\frac{T+\log^{1/2}(1/\eta)}{\epsilon}$, where $\eta=\mathbb{P}_{\mathcal{N}_d}[\mathbf{x}\in\mathcal{K}]$ and $C''\ge 1$ some sufficiently large constant, we have $\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}(\mathbf{u}'\cdot\mathbf{x})\le 1-\frac{1}{2C}e^{-\frac{1}{2}\theta^2}$. This concludes the proof of Lemma 18. ■

We will also make use of the following lemma regarding the sample complexity of estimating the expectation and covariance matrix of a log-concave distribution. Note that the truncation of the standard Gaussian on any convex set is log-concave and has variance at most 1 in every direction.

Lemma 19 (Mean and Covariance Estimation, see Lemma 4.2 in Vempala (2010a)) *Let $C>0$ be a sufficiently large universal constant, let $\gamma>0,\delta\in(0,1)$, let $\mathcal{D}$ be some log-concave distribution over $\mathbb{R}^d$ such that the variance in every direction is bounded by 1 and let X be a set of i.i.d. samples from $\mathcal{D}$ of size $|X|\ge C\cdot\frac{d}{\gamma^2}\log^2(d/\delta)$. Then, with probability at least $1-\delta$, we have*

$$
\big\|\mathop{\mathbb{E}}_{\mathbf{x}\sim X}[\mathbf{x}]-\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}]\big\|_2\le\gamma \text{ and } \big\|\mathop{\mathrm{Var}}_{\mathbf{x}\sim X}(\mathbf{x})-\mathop{\mathrm{Var}}_{\mathbf{x}\sim\mathcal{D}}(\mathbf{x})\big\|_2\le\gamma
$$

The following lemma is a standard argument that provides a sparse cover of the k-dimensional sphere and will be useful in order to exhaustively search in the low-dimensional subspace.

Lemma 20 (Sparse Cover w.r.t. Angular Distance) *Let $\mathcal{U}$ be a linear subspace spanned by the vectors $(\mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^k)$. For $\epsilon \in (0, \frac{1}{k})$, let $\mathcal{U}_\epsilon = \{\frac{\mathbf{u}}{\|\mathbf{u}\|_2} : \mathbf{u} = \epsilon \sum_{i=1}^k j_i \mathbf{v}^i, j_i \in \mathbb{Z} \cap [-\frac{1}{\epsilon}, \frac{1}{\epsilon}]\}$. Then, for any $\mathbf{v} \in \mathcal{U}$, there is $\mathbf{u} \in \mathcal{U}_\epsilon$ such that $\measuredangle(\mathbf{v}, \mathbf{u}) \le 6(k\epsilon)^{1/4}$ and $|\mathcal{U}_\epsilon| \le (\frac{2}{\epsilon})^k$.*

Proof of Lemma 20, see Vempala (2010b). Let $\mathbf{v} \in \mathcal{U}$, which we assume w.l.o.g. to have unit norm (since we only focus on angular distance). We have $\mathbf{v} = \sum_{i\in[k]} \lambda_i \mathbf{v}^i$ with $\sum_{i\in[k]} \lambda_i^2 = 1$ and $\lambda_i \in [-1, 1]$. For each i, there exists $j_i \in \mathbb{Z} \cap [-\frac{1}{\epsilon}, \frac{1}{\epsilon}]$ such that $|\lambda_i - \epsilon j_i| \le \epsilon$. Therefore, if $\mathbf{u} = \sum_{i\in[k]} \epsilon j_i \mathbf{v}^i$, then we have $\mathbf{v} \cdot \mathbf{u} \ge 1 - k\epsilon$ and $\|\mathbf{u}\|_2 \le 1 + 3\sqrt{k\epsilon}$, which implies that $\cos(\mathbf{u}, \mathbf{v}) \ge \frac{1-k\epsilon}{1+3\sqrt{k\epsilon}} \ge 1 - 4\sqrt{k\epsilon}$ and therefore $\measuredangle(\mathbf{u}, \mathbf{v}) \le 6(k\epsilon)^{1/4}$. ■

We will need the following result from Gollakota et al. (2023a) which provides a tester which ensures that any homogeneous halfspace with normal that is geometrically close to some given vector $\widehat{\mathbf{w}}$ has low disagreement with the halfspace corresponding to $\widehat{\mathbf{w}}$ under the tested marginal.

Lemma 21 (Tester for Local Halfspace Disagreement, see Gollakota et al. (2023a)) *Let $C > 0$ be a sufficiently large universal constant. There is a tester that for any $\epsilon, \delta \in (0, \frac{1}{2})$, any $\widehat{\mathbf{w}} \in \mathbb{S}^{d-1}$ and any (multi)set X of points in $\mathbb{R}^d$, runs in time $O(d^3 + d^2|X|)$ and satisfies the following.*

(a) *(Soundness.) If the tester accepts, then for any $\mathbf{w} \in \mathbb{S}^{d-1}$, with $\measuredangle(\mathbf{w}, \widehat{\mathbf{w}}) \le \epsilon$ we have*

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim X}[\text{sign}(\mathbf{w}\cdot\mathbf{x}) \ne \text{sign}(\widehat{\mathbf{w}}\cdot\mathbf{x})] \le C \cdot \epsilon^{\frac{2}{3}}$$

(b) *(Completeness.) Whenever X consists of $m \ge C(\frac{1}{\epsilon^{4/3}}\log(1/\delta) + d\log^2(d/\delta))$ independent samples from $\mathcal{N}_d$, the tester accepts w.p. at least $1 - \delta$.*

Proof of Lemma 21, combination of Propositions 3.2, 3.3 and 4.5 in Gollakota et al. (2023a). The tester does the following.

1. Compute $\mathbb{P}_{\mathbf{x}\sim X}[|\widehat{\mathbf{w}}\cdot\mathbf{x}| \le 2\epsilon^{2/3}]$ and **reject** if its value is greater than $5\epsilon^{2/3}$.
2. Compute the largest eigenvalue of the covariance matrix $\text{Var}_{\mathbf{x}\sim X}(\mathbf{x})$ and **reject** if its value is greater than 2.
3. Otherwise, **accept**.

Soundness. If the tester accepts, then we have the following. Suppose that $\mathbf{w} \ne \widehat{\mathbf{w}}$ (otherwise, the proof is trivial). Let $\mathbf{v} = \frac{\mathbf{w} - (\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}}{\|\mathbf{w} - (\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}\|_2}$ (so $\mathbf{v}$ orthogonal to $\widehat{\mathbf{w}}$). Observe that for any $\mathbf{x}$ with $\text{sign}(\mathbf{w}\cdot\mathbf{x}) \ne \text{sign}(\widehat{\mathbf{w}}\cdot\mathbf{x})$ and $|\widehat{\mathbf{w}}\cdot\mathbf{x}| > 2\epsilon^{2/3}$, it holds that $|\mathbf{v}\cdot\mathbf{x}| \ge \frac{2\epsilon^{2/3}}{\tan\epsilon}$, since we have $|\mathbf{v}\cdot\mathbf{x}| = \frac{|\mathbf{w}\cdot\mathbf{x}| + |\mathbf{w}\cdot\widehat{\mathbf{w}}|\cdot|\widehat{\mathbf{w}}\cdot\mathbf{x}|}{\|\mathbf{w} - (\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}\|_2}$, where $\mathbf{w}\cdot\mathbf{x} \ge 0$, $\mathbf{w}\cdot\widehat{\mathbf{w}} \ge \cos\epsilon$ and $\|\mathbf{w} - (\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}\|_2 \le \sin\epsilon$. Therefore, we obtain the following by additionally using Chebyshev's inequality.

$$\begin{aligned}
\mathop{\mathbb{P}}_{\mathbf{x}\sim X}[\text{sign}(\mathbf{w}\cdot\mathbf{x}) \ne \text{sign}(\widehat{\mathbf{w}}\cdot\mathbf{x})] &\le \mathop{\mathbb{P}}_{\mathbf{x}\sim X}[|\widehat{\mathbf{w}}\cdot\mathbf{x}| \le 2\epsilon^{2/3}] + \mathop{\mathbb{P}}_{\mathbf{x}\sim X}[|\mathbf{v}\cdot\mathbf{x}| \ge 2\epsilon^{2/3}/\tan\epsilon] \\
&\le 5\epsilon^{2/3} + \frac{(\tan\epsilon)^2\,\mathbb{E}_{\mathbf{x}\sim X}[(\mathbf{v}\cdot\mathbf{x})^2]}{4\epsilon^{4/3}} \\
&\le 5\epsilon^{2/3} + 2\epsilon^{2-\frac{4}{3}} = 7\epsilon^{2/3}
\end{aligned}$$

Completeness. For completeness, assume that X consists of m i.i.d. Gaussian examples. We have that $\mathbb{E}_X[\mathbb{P}_{\mathbf{x}\sim X}[|\widehat{\mathbf{w}}\cdot\mathbf{x}| \leq 2\epsilon^{2/3}]] = \mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[|\widehat{\mathbf{w}}\cdot\mathbf{x}| \leq 2\epsilon^{2/3}] \leq 4\epsilon^{2/3}$. By using a standard Hoeffding bound, we have that the first test will accept with probability at least $1-2\delta$ as long as $m \geq \frac{C}{\epsilon^{4/3}}\log(1/\delta)$ and C is sufficiently large. Moreover, by Lemma 19, as long as $m \geq C\cdot d\cdot\log^2(d/\delta)$, we have that the largest eigenvalue of $\mathrm{Var}_{\mathbf{x}\sim X}(\mathbf{x})$ is at most 2 (since $\|\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d}(\mathbf{x})\|_2 = 1$). ■

We also prove the following generalization of Lemma 21 for general halfspaces.

Lemma 22 (Tester for Local Halfspace Disagreement: General Halfspaces) *Let $C > 0$ be a sufficiently large universal constant. There is a tester that for any $\epsilon, \delta \in (0, \frac{1}{2})$ and $T > 0$, any $\widehat{\mathbf{w}} \in \mathbb{S}^{d-1}, \widehat{\tau} \in [-T, T]$ and any (multi)set X of points in $\mathbb{R}^d$, runs in time $O(d^3 + d^2|X|)$ and*

(a) *(Soundness.) If the tester accepts, then for any $\mathbf{w} \in \mathbb{S}^{d-1}$, $\tau \in \mathbb{R}$, with $\measuredangle(\mathbf{w}, \widehat{\mathbf{w}}) \leq \epsilon$ and $|\tau - \widehat{\tau}| \leq \epsilon$ we have*

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim X}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+\tau) \neq \mathrm{sign}(\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau})] \leq C\epsilon T + C\epsilon^{\frac{2}{3}}$$

(b) *(Completeness.) Whenever X consists of $m \geq C((\frac{1}{T^2\epsilon^2} + \frac{1}{\epsilon^{4/3}})\log(1/\delta) + d\log^2(d/\delta))$ independent samples from $\mathcal{N}_d$, the tester accepts w.p. at least $1-\delta$.*

Proof of Lemma 22. The tester does the following for $\gamma = 10(\epsilon T + \epsilon^{2/3})$.

1. Compute $\mathbb{P}_{\mathbf{x}\sim X}[|\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau}| \leq \gamma]$ and **reject** if its value is greater than 5γ.
2. Compute the largest eigenvalue of the covariance matrix $\mathrm{Var}_{\mathbf{x}\sim X}(\mathbf{x})$ and **reject** if its value is greater than 2.
3. Otherwise, **accept**.

Soundness. If the tester accepts, then we have the following. Suppose that $\mathbf{w} \neq \widehat{\mathbf{w}}$ (otherwise, the proof is trivial). Let $\mathbf{v} = \frac{\mathbf{w}-(\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}}{\|\mathbf{w}-(\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}\|_2}$ (so $\mathbf{v}$ orthogonal to $\widehat{\mathbf{w}}$). Observe that for any $\mathbf{x}$ with $\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+\tau) \neq \mathrm{sign}(\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau})$ and $|\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau}| > \gamma$, we have the following.

$$\begin{aligned}
|\mathbf{v}\cdot\mathbf{x}| &= \frac{|\mathbf{w}\cdot\mathbf{x} - (\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}\cdot\mathbf{x}|}{\|\mathbf{w}-(\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}\|_2} \\
&= \frac{|\mathbf{w}\cdot\mathbf{x}+\tau-\tau+\widehat{\tau}(\mathbf{w}\cdot\widehat{\mathbf{w}}) - (\mathbf{w}\cdot\widehat{\mathbf{w}})(\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau})|}{\|\mathbf{w}-(\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}\|_2} \\
&\geq \frac{|\mathbf{w}\cdot\mathbf{x}+\tau| + |(\mathbf{w}\cdot\widehat{\mathbf{w}})(\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau})| - |\tau-\widehat{\tau}(\mathbf{w}\cdot\widehat{\mathbf{w}})|}{\|\mathbf{w}-(\mathbf{w}\cdot\widehat{\mathbf{w}})\widehat{\mathbf{w}}\|_2},
\end{aligned}$$

where for the first equality we add and subtract the terms τ and $\widehat{\tau}(\mathbf{w}\cdot\widehat{\mathbf{w}})$ and for the inequality we use the fact that the signs of the halfspaces are opposite. Moreover, since we have $|\mathbf{w}\cdot\mathbf{x}+\tau| \geq 0$, $|\mathbf{w}\cdot\widehat{\mathbf{w}}| \geq \cos\epsilon$, $|\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau}| > \gamma$ and $|\widehat{\tau}-\tau| \leq \epsilon$, $|\widehat{\tau}| \leq T$, we obtain the following.

$$|\mathbf{v}\cdot\mathbf{x}| \geq \frac{\gamma\cos\epsilon - T|1-\cos\epsilon| - \epsilon}{\sin\epsilon} \geq \frac{\gamma\cos\epsilon - \epsilon(T+1)}{\sin\epsilon} \geq \frac{\gamma}{\tan\epsilon} - (T+1) =: \beta$$

Therefore, we obtain the following by additionally using Chebyshev's inequality.

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathbf{x}\sim X}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+\tau)\neq\mathrm{sign}(\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau})] &\leq \mathop{\mathbb{P}}_{\mathbf{x}\sim X}[|\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau}|\leq\gamma]+\mathop{\mathbb{P}}_{\mathbf{x}\sim X}[|\mathbf{v}\cdot\mathbf{x}|\geq\beta] \\ &\leq 3\gamma+\frac{\mathbb{E}_{\mathbf{x}\sim X}[(\mathbf{v}\cdot\mathbf{x})^2]}{\beta^2} \\ &\leq 3\gamma+\frac{2}{\beta^2}\leq C'\gamma\,,\end{aligned}$$

for a sufficiently large constant $C'>0$, due to the choice of γ.

Completeness. For completeness, assume that X consists of m i.i.d. Gaussian examples. We have that $\mathbb{E}_X[\mathbb{P}_{\mathbf{x}\sim X}[|\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau}|\leq\gamma]]=\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[|\widehat{\mathbf{w}}\cdot\mathbf{x}+\widehat{\tau}|\leq\gamma]\leq 2\gamma$. By using a standard Hoeffding bound, we have that the first test will accept with probability at least $1-2\delta$ as long as $m\geq\frac{C}{\gamma^2}\log(1/\delta)$ and C is sufficiently large. Moreover, by Lemma 19, as long as $m\geq C\cdot d\cdot\log^2(d/\delta)$, we have that the largest eigenvalue of $\mathrm{Var}_{\mathbf{x}\sim X}(\mathbf{x})$ is at most 2 (since $\|\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d}(\mathbf{x})\|_2=1$). ■

Finally, we state the following result from Klivans et al. (2023), which demonstrates that any high bias halfspace behaves as a constant function with respect to any distribution that matches sufficiently many moments up to sufficiently small accuracy with the Gaussian distribution.

Lemma 23 (Concentration via Moment Matching, see Lemma 5.6 in Klivans et al. (2023)) *Let $\epsilon>0$. Suppose that X is a set of points in $\mathbb{R}^d$ such that the empirical moments of bounded degree the uniform distribution over X approximately match the corresponding moments of the standard Gaussian, i.e., $|\mathbb{E}_{\mathbf{x}\sim X}[\mathbf{x}^\alpha]-\mathbb{E}_{\mathbf{x}\sim\mathcal{N}_d}[\mathbf{x}^\alpha]|\leq d^{-\log(1/\epsilon)}$ for any $\alpha\in\mathbb{N}^d$ s.t. $\|\alpha\|_1\leq\log(1/\epsilon)$. Then, for any $\mathbf{w}\in\mathbb{S}^{d-1}$ and $\tau\in\mathbb{R}$, with $|\tau|\geq 3\sqrt{\log(1/\epsilon)}$ we have that*

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+\tau)\neq\mathrm{sign}(\tau)]\leq\epsilon$$

Appendix C. Approximate Subspace Retrieval

In this section we provide a number of subspace retrieval lemmas, originally from Vempala (2010a) (see Sections C.1 and C.2) and Vempala (2010b) (see Section C.3). For the subspace retrieval lemma from Vempala (2010a), we provide a detailed proof here, but we incur an exponential dependence on $1/\epsilon^2$. In fact, it is not clear whether our analysis can be improved, since the original proof by Vempala (2010a) has a gap and, unless a stronger version of Lemma 18 is proven, the complexity of the algorithm in Vempala (2010a) should involve a term of $2^{\mathrm{poly}(k/\epsilon)}$ as well. To circumvent this obstacle, we also provide a fully polynomial upper bound, under some non-degeneracy assumption (see Section C.2).

C.1. Subspace Retrieval through PCA for Balanced Intersections

In this section, we will present a proof of Lemma 24, which was originally proven by Vempala (2010a). The idea of the proof is not novel, but we provide a detailed and complete version of it for concreteness. We restate the lemma here for convenience.

Lemma 24 (Subspace Retrieval, modification from Vempala (2010a)) *Let $C \geq 1$ be a sufficiently large universal constant. Let $\mathcal{C}$ be the class of intersections of k general halfspaces on $\mathbb{R}^d$, $\epsilon \in (0,1)$, $T > 0$ and $\eta \in (0, 1/2]$. Let S be a set of at least $dk^4(1/\eta)^{C/\epsilon^2}2^{CT^2/\epsilon^2}\log^2(d/\delta)$ labelled examples of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}_d$ and $f^* \in \mathcal{C}_\eta$ is an η-unbiased intersection which is defined by the normal vectors $(\mathbf{w}^1, \ldots, \mathbf{w}^k)$ and the corresponding thresholds $(\tau^1, \ldots, \tau^k)$. Then, with probability at least $1-\delta$, the subspace $\mathcal{U}$ spanned by the k-smallest variance orthogonal components of the positive examples $S^+ = \{\mathbf{x} : (\mathbf{x}, 1) \in S\}$ approximately includes all of the normal vectors corresponding to bounded thresholds, i.e., for any $i \in [k]$ if $\tau^i \leq T$, then $\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}^i\|_2 \geq 1 - \epsilon$.*

Algorithm 2: Subspace Retrieval through PCA

Input: Labelled set S_{train}, parameter k
Output: Orthonormal basis $(\mathbf{v}^1, \ldots, \mathbf{v}^k)$
Let S^+_{train} be the subset of S_{train} corresponding to positive examples.
Run Principal Component Analysis on $S^+_{\text{train}} = \{\mathbf{x} : (\mathbf{x}, 1) \in S_{\text{train}}\}$ and let $\mathbf{v}^1, \ldots, \mathbf{v}^k$ be the k smallest-variance orthogonal components (i.e., the right singular vectors corresponding to the k smallest singular values of the $(|S^+_{\text{train}}| \times d)$-dimensional sample matrix).
Output $(\mathbf{v}^1, \ldots, \mathbf{v}^k)$ and terminate.

For the proof, we will use the following strong theorem which ensures that the subspace retrieved by PCA on the empirical distribution will be geometrically close to the true corresponding subspace, as long as there is a spectral gap in the covariance matrix of the true distribution.

Proposition 25 (Davis-Kahan, modification of Theorem 2 in Yu et al. (2015)) *Let $M \in \mathbb{R}^{d\times d}$ and $\widehat{M} \in \mathbb{R}^{d\times d}$ be symmetric matrices such that for some $k \in [d]$, the gap between the k-th smallest eigenvalue of M and the $(k+1)$-th smallest eigenvalue of M is positive, i.e., $\lambda_{k+1} - \lambda_k > 0$. Let $\mathbf{v}^1, \ldots, \mathbf{v}^k$ be the eigenvectors of M corresponding to the k smallest eigenvalues and, similarly, $\mathbf{u}^1, \ldots, \mathbf{u}^k$ the k smallest eigenvectors of $\widehat{M}$. Then we have that*

$$\sum_{i\in[k]} \sin^2(\measuredangle(\mathbf{v}^i, \mathbf{u}^i)) \leq \frac{4k\|M - \widehat{M}\|_2^2}{(\lambda_{k+1} - \lambda_k)^2}$$

Let $\mathcal{W}$ be the span of $(\mathbf{w}^1, \ldots, \mathbf{w}^k)$ and note that every direction orthogonal to $\mathcal{W}$ has variance 1 under $\mathcal{N}_d|_\mathcal{K}$. Let $\gamma = (1/\eta)^{C/\epsilon^2}2^{CT^2/\epsilon^2}$ and let $\mathcal{W}_\gamma$ be the subspace of $\mathcal{W}$ such that for every direction $\mathbf{u}$ orthogonal to $\mathcal{W}_\gamma$, we have $\operatorname{Var}_{\mathbf{x}\sim\mathcal{N}_d|_\mathcal{K}}(\mathbf{u}\cdot\mathbf{x}) > 1-\gamma$ and $\mathcal{W}_\gamma$ is spanned by an orthonormal basis $(\mathbf{z}^1, \ldots, \mathbf{z}^\ell)$ with $\operatorname{Var}_{\mathbf{x}\sim\mathcal{N}_d|_\mathcal{K}}(\mathbf{z}^i\cdot\mathbf{x}) \leq 1-\gamma$. In other words, $\mathcal{W}_\gamma$ is the span of the eigenvectors of the covariance matrix M of $\mathcal{N}_d|_\mathcal{K}$ whose corresponding eigenvalues are at most $1-\gamma$. Note that since $\dim(\mathcal{W}) \leq k$ and $\mathcal{W}_\gamma \subseteq \mathcal{W}$, we have $\ell \leq k$. Let $0 \leq \lambda_1 \leq \cdots \leq \lambda_\ell \leq 1-\gamma < \lambda_{\ell+1} \leq \cdots \leq \lambda_k \leq 1 = \lambda_{k+1}$ be the $k+1$ smallest eigenvalues of the covariance matrix of $\mathcal{N}_d|_\mathcal{K}$. Since there is a γ gap between λ_ℓ and λ_{k+1}, there is some $j \in [\ell, k]$ such that $\lambda_{j+1} - \lambda_j > \frac{\gamma}{k}$.

Let $\mathcal{U}$ be the subspace corresponding to the k smallest eigenvectors of the empirical covariance matrix $\widehat{M}$ of the set of positive examples S^+. Since $|S| \geq \frac{1}{\eta^2}\log(1/\delta)$, due to a Hoeffding bound, we have that with probability at least $1-\delta/10$, $|S^+| \geq \frac{\eta}{2}|S| \geq dk^4(1/\eta)^{C/\epsilon^2}2^{CT^2/\epsilon^2}\log^2(d/\delta)$. We can therefore apply Lemma 19 to $\mathcal{N}_d|_\mathcal{K}$ (which is log-concave) to obtain that $\|M - \widehat{M}\|_2 \leq \frac{\gamma\epsilon}{2C'k^2}$. Let $\mathcal{U}_\ell$ be the subspace of $\mathcal{U}$ corresponding to the ℓ smallest eigenvalues of $\widehat{M}$, and let $(\mathbf{v}^1, \dots, \mathbf{v}^\ell)$ be the corresponding eigenvectors. By Proposition 25, we have that

$$\sum_{i\in[\ell]} \sin^2(\measuredangle(\mathbf{v}^i, \mathbf{z}^i)) \leq \epsilon/(C'\sqrt{k}) \tag{1}$$

Let $i \in [k]$ such that $\tau^i \leq T$. We analyze $\mathbf{w}^i$ in two orthogonal components, $\mathbf{w}$ and $\mathbf{w}'$, where $\mathbf{w}$ is the normalized projection of $\mathbf{w}^i$ on $\mathcal{W}_\gamma$ and $\mathbf{w}'$ is therefore orthogonal to $\mathcal{W}_\gamma$. Since $\mathbf{w}'$ is orthogonal to $\mathcal{W}_\gamma$, by the definition of $\mathcal{W}_\gamma$, we have $\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d}(\mathbf{w}' \cdot \mathbf{x}) > 1 - \gamma$. By Lemma 24, this implies that $\mathbf{w}^i \cdot \mathbf{w}' < C'' \frac{1+T+\log^{1/2}(1/\eta)}{\log^{1/2}(1/\gamma)}$. Therefore, $\measuredangle(\mathbf{w}^i, \mathbf{w}) \leq 2C'' \frac{1+T+\log^{1/2}(1/\eta)}{\log^{1/2}(1/\gamma)}$. Moreover, by Equation (1), we have that $\measuredangle(\mathbf{w}, \mathrm{proj}_{\mathcal{U}_\ell} \mathbf{w}) \leq \epsilon/10$. Since $2C'' \frac{1+T+\log^{1/2}(1/\eta)}{\log^{1/2}(1/\gamma)} \leq \epsilon/10$ by the choice of γ, we obtain the desired result.

C.2. Subspace Retrieval through PCA under a Non-Degeneracy Assumption

In the previous subsection we provided a detailed proof of the subspace retrieval lemma which was originally proven in Vempala (2010a), incurring, however, an exponential dependence on $1/\epsilon^2$. Here, we define a technical assumption on the concept class considered which is sufficient to provide a fully polynomial result for subspace retrieval. Despite its technicality, the non-degeneracy condition is satisfied by the constructions we use for our lower bounds, which implies that under the non-degeneracy condition, our upper and lower bounds are directly comparable (and tight in some regimes).

Definition 26 (Non-Degeneracy Condition) *Let $\mathcal{K}$ be an intersection of halfspaces in $\mathbb{R}^d$ and $\mathcal{N}_d|_\mathcal{K}$ be the truncation of the standard Gaussian to $\mathcal{K}$. For $\beta \geq 1$, we say that $\mathcal{K}$ is β-non-degenerate if the following is true. For every subspace $\mathcal{W}$ spanned by some of the normals of $\mathcal{K}$ and for every vector $\mathbf{w} \in \mathbb{S}^{d-1}$ that is a normal to $\mathcal{K}$ with non-zero projection $\mathbf{w}' \in \mathbb{R}^d \setminus \{0\}$ onto the subspace orthogonal to $\mathcal{W}$ we have*

$$\underset{\mathbf{x}\sim\mathcal{N}_d}{\mathrm{Var}}(\widehat{\mathbf{w}}' \cdot \mathbf{x}) - \underset{\mathbf{x}\sim\mathcal{N}_d|_\mathcal{K}}{\mathrm{Var}}(\widehat{\mathbf{w}}' \cdot \mathbf{x}) \geq \Big(\underset{\mathbf{x}\sim\mathcal{N}_d}{\mathrm{Var}}(\mathbf{w} \cdot \mathbf{x}) - \underset{\mathbf{x}\sim\mathcal{N}_d|_\mathcal{K}}{\mathrm{Var}}(\mathbf{w} \cdot \mathbf{x})\Big)^\beta, \text{ where } \widehat{\mathbf{w}}' = \mathbf{w}'/\|\mathbf{w}'\|_2$$

For any class $\mathcal{C}$ of halfspace intersections on $\mathbb{R}^d$, we denote with $\mathcal{C}^\beta$ the β-non-degenerate version of $\mathcal{C}$, i.e., the subset of $\mathcal{C}$ that contains the elements that are β-non-degenerate.

The condition defined above states that each normal $\mathbf{w}$ of the intersection has either zero or non-trivial relative influence on subspaces orthogonal to the span $\mathcal{W}'$ of any subset of the normals. The influence is measured in terms of the variance reduction along the residual direction $\mathbf{w} - \mathrm{proj}_{\mathcal{W}'}(\mathbf{w})$. In particular, in light of the third part of Lemma 18, for intersections of two halfspaces, the non-degeneracy condition is satisfied whenever the two halfspaces of the intersection have normals either pointing to the exact same direction or have sufficiently large angular distance (but nothing in between). This enables one to circumvent the need for a strong quantitative statement relating (1) the angle between some vector $\mathbf{u}$ and a normal with (2) the variance reduction along $\mathbf{u}$, which is the source of the exponential dependence of $2^{1/\epsilon^2}$. With an analysis similar to the one of Section C.1, we obtain the following subspace retrieval result.

Lemma 27 (Subspace Retrieval under Non-Degeneracy, see Vempala (2010a)) *Let $C \geq 1$ be a sufficiently large universal constant. Let $\mathcal{C}$ be the class of intersections of k general halfspaces on $\mathbb{R}^d$, $\epsilon \in (0,1)$, $T \geq 0$ and $\beta \geq 1, \eta \in (0, 1/2]$. Let S be a set of at least $\frac{Cdk^4}{\epsilon^2\eta^2} e^{\beta T^2} \log^2(d/\delta)$ labelled examples of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}_d$ and $f^* \in \mathcal{C}_\eta^\beta$ is an η-unbiased and β-non-degenerate intersection which is defined by the normal vectors $(\mathbf{w}^1, \dots, \mathbf{w}^k)$ and the corresponding thresholds $(\tau^1, \dots, \tau^k)$. Then, with probability at least $1-\delta$, the subspace $\mathcal{U}$ spanned by the k-smallest variance orthogonal components of the positive examples $S^+ = \{\mathbf{x} : (\mathbf{x}, 1) \in S\}$ approximately includes all of the normal vectors corresponding to bounded thresholds, i.e., for any $i \in [k]$ if $\tau^i \leq T$, then $\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}^i\|_2 \geq 1 - \epsilon$.*

C.3. Subspace Retrieval through Polar Planes algorithm

We now present the following lemma from Vempala (2010b) which provides another algorithm for approximately retrieving the relevant subspace for homogeneous intersections whose runtime is not exponential in $1/\epsilon$, even without making a non-degeneracy assumption. The lemma follows from combining Theorem 4 and Lemma 3 from Vempala (2010b).

Lemma 28 (Subspace Retrieval through Polar Planes, from Vempala (2010b)) *Consider $\mathcal{C}$ to be the class of intersections of k homogeneous halfspaces on $\mathbb{R}^d$, $\epsilon \in (0,1)$ and $\eta \in (0, 1/2]$. Let S be a set of at least $m = d(\frac{k}{\epsilon\eta})^{O(k)} \log(1/\delta)$ labelled examples of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}_d$ and $f^* \in \mathcal{C}_\eta$ is an η-balanced intersection which is defined by the normal vectors $(\mathbf{w}^1, \dots, \mathbf{w}^k)$. There is an algorithm (Polar Planes from Vempala (2010b)) that on input S, returns, w.p. at least $1-\delta$, an orthonormal basis for a subspace $\mathcal{U}$ of dimension k that approximately includes all of the normal vectors, i.e., for any $i \in [k]$, we have $\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}^i\|_2 \geq 1 - \epsilon$, in time $(\frac{dk}{\epsilon\eta})^{O(k)}$.*

Appendix D. TDS Learning Intersections of Halfspaces

We now provide full proofs for all of our upper bounds, assuming the balanced concepts condition (Definition 17), both with and without assuming the non-degeneracy condition (Definition 26).

D.1. Homogeneous Halfspace Intersections

We prove our result on learning intersections of homogeneous halfspaces, which we restate here for convenience.

Theorem 29 (TDS Learning Intersections of Homogeneous Halfspaces) *Let $\mathcal{C}$ be a class whose elements are intersections of k homogeneous halfspaces on $\mathbb{R}^d$, $\epsilon \in (0, 1)$ and $C \geq 1$ a sufficiently large constant.*

- *Assume that there is an algorithm $\mathcal{A}$ that upon receiving at least $m_{\mathcal{A}}$ examples from a training distribution of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}_d$ and $f^* \in \mathcal{C}$, outputs, with probability at least 0.99 an orthonormal basis for a subspace $\mathcal{U}$ such that for any normal $\mathbf{w}$ of f^* we have $\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}\|_2 \geq 1 - (\frac{k}{C\epsilon})^3$.*

Then, there is an algorithm (Algorithm 3) that $(\epsilon, \delta = 0.02)$-TDS learns the class $\mathcal{C}$, using $m_{\mathcal{A}} + \tilde{O}(\frac{dk^2}{\epsilon^2})$ labelled training examples and $\tilde{O}(\frac{dk^2}{\epsilon^2})$ unlabelled test examples, calls $\mathcal{A}$ once and uses additional time $\tilde{O}(\frac{d^3k^2}{\epsilon^2}) + d(k/\epsilon)^{O(k^2)}$.

Algorithm 3: Proper TDS Learner for Homogeneous Halfspace Intersections

Input: Labelled set S_{train}, unlabelled set X_{test}, parameter ϵ
Set $\epsilon' = \frac{\epsilon^{3/2}}{Ck^{3/2}}$ and $\epsilon'' = \frac{\epsilon^6}{Ck^7}$ for some sufficiently large universal constant $C \geq 1$.
Run algorithm $\mathcal{A}$ on the set S_{train} and let $(\mathbf{v}^1, \dots, \mathbf{v}^k)$ be its output.
Let $\mathcal{U}$ be the subspace spanned by $(\mathbf{v}^1, \dots, \mathbf{v}^k)$ and consider the following sparse cover of $\mathcal{U}$:
$\mathcal{U}_{\epsilon''} = \{\frac{\mathbf{u}}{\|\mathbf{u}\|_2} : \mathbf{u} = \epsilon'' \sum_{i=1}^k j_i \mathbf{v}^i, j_i \in \mathbb{Z} \cap [-\frac{1}{\epsilon''}, \frac{1}{\epsilon''}], \|\mathbf{u}\|_2 \neq 0\}$
Reject and terminate if $\| \operatorname{Var}_{\mathbf{x}\sim X}(\mathbf{x})\|_2 \geq 2$.
for $\mathbf{u} \in \mathcal{U}_{\epsilon''}$ **do**
  Reject and terminate if $\mathbb{P}_{\mathbf{x}\sim X}[|\mathbf{u}\cdot\mathbf{x}| \leq 2\epsilon'^{2/3}] > 5\epsilon'^{2/3}$.
end
Let $\mathcal{F}$ contain the concepts $f : \mathbb{R}^d \to \{\pm 1\}$ of the form $f(\mathbf{x}) = 2\bigwedge_{i=1}^k \mathbb{1}\{\mathbf{u}^i \cdot \mathbf{x} \geq 0\} - 1$,
where $\mathbf{u}^1, \dots, \mathbf{u}^k \in \mathcal{U}_{\epsilon''}$ and $\mathbb{P}_{(\mathbf{x},y)\sim S_{\text{train}}}[y \neq f(\mathbf{x})] \leq \epsilon/5$.
Reject and terminate if $\max_{f_1,f_2\in\mathcal{F}} \mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f_1(\mathbf{x}) \neq f_2(\mathbf{x})] > \epsilon/2$.
Otherwise, output $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$ for some $\widehat{f} \in \mathcal{F}$.

Proof of Theorem 2. Let S_{train} be a set of m_{train} samples from the training distribution, i.e., of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{D} = \mathcal{N}_d$ and let X_{test} be a set of m_{test} samples from the test distribution $\mathcal{D}'$. Let $C > 0$ be a sufficiently large universal constant. Let $f^* : \mathbb{R}^d \to \{\pm 1\}$ denote the ground truth, i.e., the intersection of k homogeneous halfspaces

$$f^*(\mathbf{x}) = 2\wedge_{i\in[k]} \mathbb{1}\{\mathbf{w}^i \cdot \mathbf{x} \geq 0\} - 1\,,\text{ for some } \mathbf{w}^1, \dots, \mathbf{w}^k \in \mathbb{S}^{d-1}$$

In the following, we will say that an event holds with high probability if it holds with probability sufficiently close to 1 so that union bounding over all the bad events gives a probability of failure of at most 0.01. This is possible by choosing C to be a sufficiently large constant.

Soundness. To prove soundness, suppose that the tests have accepted. We first use the approach of Vempala (2010a) to show that using training data, we can retrieve a subspace that is geometrically close to the normal subspace of the ground truth. Let C', C'' be sufficiently large universal constants.

In particular, the guarantee for algorithm $\mathcal{A}$ implies that the retrieved subspace $\mathcal{U}$ has the property that for any $i \in [k]$ we have $\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}^i\|_2 \geq 1 - (\frac{\epsilon}{C'k})^3$ with high probability, as long as $m_{\text{train}} \geq m_{\mathcal{A}}$. Let $\mathbf{w}^i_{\mathcal{U}} = \frac{\operatorname{proj}_{\mathcal{U}} \mathbf{w}^i}{\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}^i\|_2}$. Then, we have $\measuredangle(\mathbf{w}^i, \mathbf{w}^i_{\mathcal{U}}) \leq \frac{4\epsilon^{3/2}}{C'k^{3/2}}$. Due to Lemma 20, there is a vector $\mathbf{u}^i \in \mathcal{U}_{\epsilon''}$ with $\measuredangle(\mathbf{u}^i, \mathbf{w}^i_{\mathcal{U}}) \leq \frac{\epsilon^{3/2}}{C'k^{3/2}}$, whenever $\epsilon'' \leq \frac{\epsilon^6}{6^4 C'k^7}$, in which case, $|\mathcal{U}_{\epsilon''}| \leq (\frac{2\cdot 6^4 C'k^7}{\epsilon^6})^k$. Therefore, for any $i \in [k]$ we have some vector $\mathbf{u}^i$ in the cover $\mathcal{U}_{\epsilon''}$ that is close to the normal $\mathbf{w}^i$, i.e., $\measuredangle(\mathbf{u}^i, \mathbf{w}^i) \leq (\frac{5\epsilon}{C'k})^{3/2}$.

Consider now the hypothesis $f(\mathbf{x}) = 2\wedge_{i\in[k]} \mathbb{1}\{\mathbf{u}^i\cdot\mathbf{x} \geq 0\} - 1$. If suffices to show that f belongs in the set $\mathcal{F}$ of candidate concepts and that f has small test error $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \epsilon/4$, because then for any other candidate concept $f' \in \mathcal{F}$, we know that it disagrees with f only on a small fraction of test points and, hence, we will have $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f'(\mathbf{x}) \neq f^*(\mathbf{x})] \leq 3\epsilon/4$. By standard VC dimension arguments, this would imply that, whenever $m_{\text{test}} \geq C\frac{dk\log k}{\epsilon^2}$, with high probability, the test error of any element of $\mathcal{F}$ satisfies $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}'}[f'(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \epsilon$.

We appeal to the tester for local halfspace disagreement of Lemma 21 in order to demonstrate that $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \epsilon/4$. In particular, we have that

$$\begin{aligned}\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f(\mathbf{x}) \neq f^*(\mathbf{x})] &\leq k \mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[\text{sign}(\mathbf{u}^i \cdot \mathbf{x}) \neq \text{sign}(\mathbf{w}^i \cdot \mathbf{x})] \\ &\leq C'' k(\measuredangle(\mathbf{u}^i, \mathbf{w}^i))^{2/3} \leq \epsilon/4\end{aligned}$$

Finally, we show that the hypothesis f lies within $\mathcal{F}$. In particular, $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x}) \neq f^*(\mathbf{x})] \leq k\,\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\text{sign}(\mathbf{u}^i \cdot \mathbf{x}) \neq \text{sign}(\mathbf{w}^i \cdot \mathbf{x})] = O(k\measuredangle(\mathbf{u}^i, \mathbf{w}^i))$, which is bounded by $\epsilon/10$ by choosing the constant C' appropriately. By standard VC dimension arguments, we therefore have that $\mathbb{P}_{\mathbf{x}\sim S_{\text{train}}}[f(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \epsilon/5$ as long as $m_{\text{train}} \geq \frac{Cdk\log k}{\epsilon^2}$.

Completeness. To prove completeness, suppose that $\mathcal{D}' = \mathcal{N}_d$. Since $\mathcal{U}_{\epsilon''}, \mathcal{F}$ do not depend on X_{test}, we can use Hoeffding bounds to bound the probability of rejection, as well as union bounds over $\mathcal{F} \times \mathcal{F}$ accordingly. In particular, the tester of Lemma 21 will accept with high probability as long as $m_{\text{test}} \geq C\frac{1}{\epsilon'^{4/3}} + Cd\log^2 d = O(\frac{k^2}{\epsilon^2} + d\log^2 d)$ and the tester of the disagreement probabilities of pairs in $\mathcal{F}$ will accept (due to standard Hoeffding and union bounds) with high probability whenever $m_{\text{test}} \geq C\frac{1}{\epsilon^2}\log|\mathcal{F}| = O(\frac{k^2}{\epsilon^2}\log(\frac{k}{\epsilon}))$ (since $|\mathcal{F}| = (k/\epsilon)^{O(k^2)}$ as we need to choose k normals from $\mathcal{U}_{\epsilon''}$). ■

By combining Theorem 29 with Lemmas 24, 27 and 28 we obtain the following bounds for TDS learning homogeneous halfspace intersections.

Corollary 30 (TDS Learning Bounds for Homogeneous Halfspace Intersections) *Let $\eta \in (0, \frac{1}{2})$, $\epsilon > 0$, $\beta \geq 1$ and let $\mathcal{C}$ be the class of intersections of k homogeneous halfspaces on $\mathbb{R}^d$.*

(a) There is an $(\epsilon, \delta = 0.02)$-TDS learner for the class $\mathcal{C}_\eta$ of η-balanced intersections that uses $\tilde{O}(d)(\frac{k}{\epsilon\eta})^{O(\frac{k^6}{\epsilon^6})}$ labelled training examples, $\tilde{O}(\frac{dk^2}{\epsilon^2})$ unlabelled test examples and runs in time $\tilde{O}(d^3)(\frac{k}{\epsilon\eta})^{O(\frac{k^6}{\epsilon^6})}$.

(b) There is an $(\epsilon, \delta = 0.02)$-TDS learner for the class $\mathcal{C}_\eta^\beta$ of η-balanced and β-non-degenerate intersections that uses $\tilde{O}(d) \cdot \frac{1}{\eta^2} \cdot (\frac{k}{\epsilon})^{O(\beta)}$ labelled training examples, $\tilde{O}(\frac{dk^2}{\epsilon^2})$ unlabelled test examples and runs in time $\tilde{O}(d^3) \cdot \frac{1}{\eta^2} \cdot (\frac{k}{\epsilon})^{O(\beta)} + d(k/\epsilon)^{O(k^2)}$.

(c) There is an $(\epsilon, \delta = 0.02)$-TDS learner for the class $\mathcal{C}_\eta$ of η-balanced intersections that uses $\tilde{O}(d)(\frac{k}{\epsilon\eta})^{O(k)}$ labelled training examples, $\tilde{O}(\frac{dk^2}{\epsilon^2})$ unlabelled test examples and runs in time $(\frac{dk}{\epsilon\eta})^{O(k)} + d(k/\epsilon)^{O(k^2)}$.

D.2. General Halfspace Intersections

We now prove our positive results on learning intersections of general halfspaces.

Theorem 31 (TDS Learning Intersections of General Halfspaces) *Let $\mathcal{C}$ be a class whose elements are intersections of k general halfspaces on $\mathbb{R}^d$, $\epsilon, T \in (0, 1)$ and $C \geq 1$ a sufficiently large constant.*

- *Assume that there is an algorithm $\mathcal{A}$ that upon receiving at least $m_{\mathcal{A}}$ examples of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}_d$ and $f^* \in \mathcal{C}$, outputs, with probability at least 0.99 an orthonormal basis for a subspace $\mathcal{U}$ such that for any normal $\mathbf{w} \in \mathbb{S}^{d-1}$ that corresponds to some halfspace $\{\mathbf{x} : \mathbf{w} \cdot \mathbf{x} + \tau \geq 0\}$ of f^* with threshold $\tau \leq T$ we have $\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}\|_2 \geq 1 - (\frac{k}{C\epsilon})^3$.*

Then, there is an algorithm (Algorithm 4) that $(\epsilon, \delta = 0.02)$-TDS learns the class $\mathcal{C}$, using $m_{\mathcal{A}} + \tilde{O}(\frac{dk^2}{\epsilon^2})$ labelled training examples and $d^{O(\log(k/\epsilon))}$ unlabelled test examples, calls $\mathcal{A}$ once and uses additional time $d^{O(\log(k/\epsilon))}(k/\epsilon)^{O(k^2)}$.

Algorithm 4: Proper TDS Learner for General Halfspace Intersections

---

Input: Labelled set S_{train}, unlabelled set X_{test}, parameter ϵ

Set $T = 3\log^{1/2}(\frac{10k}{\epsilon})$, $r \geq \log(10k/\epsilon)$, $\Delta = d^{-r}$, $\epsilon' = \frac{\epsilon^{3/2}}{Ck^{3/2}}$ and $\epsilon'' = \frac{\epsilon^6}{Ck^{3/2}}$, where $C \geq 1$ is a sufficiently large constant.

Reject and terminate if for some $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \leq r$ it holds $|\mathbb{E}_{\mathbf{x}\sim X_{\text{test}}}[\mathbf{x}^\alpha] - \mathbb{E}_{\mathbf{x}\sim\mathcal{N}}[\mathbf{x}^\alpha]| > \Delta$

Run algorithm $\mathcal{A}$ on set S_{train} and let $(\mathbf{v}^1, \dots, \mathbf{v}^k)$ be its output.

Let $\mathcal{U}$ be the subspace spanned by $(\mathbf{v}^1, \dots, \mathbf{v}^k)$ and consider the following sparse cover of $\mathcal{U}$: $\mathcal{U}_{\epsilon''} = \{\frac{\mathbf{u}}{\|\mathbf{u}\|_2} : \mathbf{u} = \epsilon'' \sum_{i=1}^k j_i \mathbf{v}^i, j_i \in \mathbb{Z} \cap [-\frac{1}{\epsilon''}, \frac{1}{\epsilon''}], \|\mathbf{u}\|_2 \neq 0\}$

Let $\mathcal{T}_{\epsilon'} = \{j\epsilon' : j \in \mathbb{Z} \cap [-\frac{T}{\epsilon'}, \frac{T}{\epsilon'}]\}$ be a cover of the candidate halfspace biases.

Reject and terminate if $\|\operatorname{Var}_{\mathbf{x}\sim X}(\mathbf{x})\|_2 \geq 2$.

for $(\mathbf{u}, \theta) \in \mathcal{U}_{\epsilon''} \times \mathcal{T}_{\epsilon'}$ **do**

  Reject and terminate if $\mathbb{P}_{\mathbf{x}\sim X}[|\mathbf{u} \cdot \mathbf{x} + \theta| \leq 2\epsilon'^{2/3}] > 5\epsilon'^{2/3}$.

end

Let $\mathcal{F}$ contain the concepts $f : \mathbb{R}^d \to \{\pm 1\}$ of the form $f(\mathbf{x}) = 2\bigwedge_{i=1}^k \mathbb{1}\{\mathbf{u}^i \cdot \mathbf{x} + \theta^i \geq 0\} - 1$, where $(\mathbf{u}^1, \theta^1), \dots, (\mathbf{u}^k, \theta^k) \in \mathcal{U}_{\epsilon''} \times \mathcal{T}_{\epsilon'}$ and $\mathbb{P}_{(\mathbf{x},y)\sim S_{\text{train}}}[y \neq f(\mathbf{x})] \leq \epsilon/5$.

Reject and terminate if $\max_{f_1, f_2 \in \mathcal{F}} \mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f_1(\mathbf{x}) \neq f_2(\mathbf{x})] > \epsilon/2$.

Otherwise, output $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$ for some $\widehat{f} \in \mathcal{F}$.

Proof of Theorem 5. The proof is similar to the one of Theorem 2, but since the intersections are general, there are some additional complications. Let once more S_{train} be a set of m_{train} samples from the training distribution, i.e., of the form $(\mathbf{x}, f^*(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{D} = \mathcal{N}_d$ and let X_{test} be a set of m_{test} samples from the test distribution $\mathcal{D}'$. Let $C > 0$ be a sufficiently large universal constant. Let $f^* : \mathbb{R}^d \to \{\pm 1\}$ denote the ground truth, i.e., the intersection of k halfspaces

$$f^*(\mathbf{x}) = 2 \wedge_{i\in[k]} \mathbb{1}\{\mathbf{w}^i \cdot \mathbf{x} + \tau^i \geq 0\} - 1\,,\ \text{for } \mathbf{w}^1, \dots, \mathbf{w}^k \in \mathbb{S}^{d-1} \text{ and } \tau^1, \dots, \tau^k \in \mathbb{R}$$

In the following, we will say that an event holds with high probability if it holds with probability sufficiently close to 1 so that union bounding over all the bad events gives a probability of failure of at most 0.01. This is possible by choosing C to be a sufficiently large constant.

Soundness. Suppose that the tests have accepted. We will once more use the subspace retrieval lemma from Vempala (2010a), but this time we will use a version (Lemma 24) that works for arbitrary halfspace intersections. We pick $T = 3\sqrt{\log(10k/\epsilon)}$, $r \geq \log(10k/\epsilon)$ and $C', C'' > 0$ sufficiently large universal constants.

Due to Lemma 24, the retrieved subspace $\mathcal{U}$ has the property that, with high probability, for any $i \in [k]$ with $\tau^i \leq T$ we have $\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}^i\|_2 \geq 1 - (\frac{\epsilon}{C'k})^3$, as long as $m_{\text{train}} \geq m_{\mathcal{A}}$. Consider once more $\mathbf{w}^i_{\mathcal{U}} = \frac{\operatorname{proj}_{\mathcal{U}} \mathbf{w}^i}{\|\operatorname{proj}_{\mathcal{U}} \mathbf{w}^i\|_2}$. We have $\measuredangle(\mathbf{w}^i, \mathbf{w}^i_{\mathcal{U}}) \leq \frac{4\epsilon^{3/2}}{C'k^{3/2}}$ and for some $\mathbf{u}^i \in \mathcal{U}_{\epsilon''}$, we have $\measuredangle(\mathbf{u}^i, \mathbf{w}^i_{\mathcal{U}}) \leq \frac{\epsilon^{3/2}}{C'k^{3/2}}$, whenever $\epsilon'' \leq \frac{\epsilon^6}{6^4 C'k^7}$ (which implies $|\mathcal{U}_{\epsilon''}| \leq (\frac{2\cdot 6^4 C'k^7}{\epsilon^6})^k$). Therefore, for any $i \in [k]$ that corresponds to a halfspace with bounded bias $\tau^i \leq T$, we have $\measuredangle(\mathbf{u}^i, \mathbf{w}^i) \leq (\frac{5\epsilon}{C'k})^{3/2}$. Moreover, for any such i, there is some $\theta^i \in \mathcal{T}_{\epsilon'}$ that is either close to the i-th threshold ($|\theta^i - \tau^i| \leq \epsilon'$) or they are both large enough ($\tau^i \leq -T$ and $\theta^i = -T$). Assume without loss of generality that $\{i \in [k] : \tau^i \leq T\} = [\ell]$ for some $\ell \leq k$.

Consider now the hypothesis $f(\mathbf{x}) = 2 \wedge_{i\in[\ell]} \mathbb{1}\{\mathbf{u}^i \cdot \mathbf{x} + \theta^i \geq 0\} - 1$. Once more, it suffices to show that f belongs in the set $\mathcal{F}$ of candidate concepts and that f has small test error $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \epsilon/4$, because then for any other candidate concept $f' \in \mathcal{F}$, we know that it disagrees with f only on a small fraction of test points and, hence, we will have $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f'(\mathbf{x}) \neq f^*(\mathbf{x})] \leq 3\epsilon/4$. By standard VC dimension arguments, this would imply that, whenever $m_{\text{test}} \geq C\frac{dk\log k}{\epsilon^2}$, with high probability, the test error of any element of $\mathcal{F}$ satisfies $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}'}[f'(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \epsilon$.

As a first step, we will show that the ground truth is close to the intersection corresponding to the bounded bias halfspaces with respect to both the training and the test examples, i.e., that for $\tilde{f}^*(\mathbf{x}) = 2 \wedge_{i\in[\ell]} \mathbb{1}\{\mathbf{w}^i \cdot \mathbf{x} + \tau^i \geq 0\} - 1$ we have $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f^*(\mathbf{x}) \neq \tilde{f}^*(\mathbf{x})] \leq \epsilon/8$ and $\mathbb{P}_{\mathbf{x}\sim S_{\text{train}}}[f^*(\mathbf{x}) \neq \tilde{f}^*(\mathbf{x})] \leq \epsilon/10$. This is important, because we can then relate f, f^* through $\tilde{f}^*$. Since the moment-matching test has accepted, by Lemma 23, as long as $r \geq \log(10k/\epsilon)$ and $T \geq 3\sqrt{\log(10k/\epsilon)}$, for any $i > \ell$, we have that $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[\operatorname{sign}(\mathbf{w}^i \cdot \mathbf{x} + \tau^i) \neq 1] \leq \frac{\epsilon}{10k}$. Therefore, $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f^*(\mathbf{x}) \neq \tilde{f}^*(\mathbf{x})] \leq \sum_{i>\ell} \mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[\operatorname{sign}(\mathbf{w}^i \cdot \mathbf{x} + \tau^i) \neq 1] \leq \epsilon/8$, due to a union bound (and the fact that the only possibility that f^* and $\tilde{f}^*$ differ is if some of the omitted halfspaces in $\tilde{f}^*$ becomes negative). Similarly, for S_{train}, the claim follows with high probability by a standard Hoeffding bound (f^* and $\tilde{f}^*$ do not depend on S_{train}), as long as $|S_{\text{train}}| \geq C\frac{k^2}{\epsilon^2}$.

We will now bound the quantity $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f(\mathbf{x}) \neq \tilde{f}^*(\mathbf{x})]$ by $\epsilon/8$. Observe that in the case that $|\tau^i| \geq T$, then, by Lemma 23 (as argued above), the corresponding halfspace is constant with probability at least $1 - \epsilon/(10k)$ and the same is true for $\theta^i = T$. Therefore, we may safely omit these terms from f and $\tilde{f}^*$ by only incurring an error of at most $\epsilon/10$. For the remaining terms, we appeal to the tester for local (general) halfspace disagreement of Lemma 22 in order to show that $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[f(\mathbf{x}) \neq \tilde{f}^*(\mathbf{x})] \leq \epsilon/8$. In particular, we have that

$$\begin{aligned}
\mathop{\mathbb{P}}_{\mathbf{x}\sim X_{\text{test}}}[f(\mathbf{x}) \neq \tilde{f}^*(\mathbf{x})] &\leq k \mathop{\mathbb{P}}_{\mathbf{x}\sim X_{\text{test}}}[\operatorname{sign}(\mathbf{u}^i \cdot \mathbf{x} + \theta^i) \neq \operatorname{sign}(\mathbf{w}^i \cdot \mathbf{x} + \tau^i)] \\
&\leq C''k(\measuredangle(\mathbf{u}^i, \mathbf{w}^i))^{2/3} + C''k(\measuredangle(\mathbf{u}^i, \mathbf{w}^i))\log^{1/2}(1/\epsilon) \\
&\leq \epsilon/8
\end{aligned}$$

Finally, we show that the hypothesis f lies within $\mathcal{F}$. In particular, $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x}) \neq \tilde{f}^*(\mathbf{x})] \leq k\,\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\operatorname{sign}(\mathbf{u}^i \cdot \mathbf{x} + \theta^i) \neq \operatorname{sign}(\mathbf{w}^i \cdot \mathbf{x} + \tau^i)] = O(kT\measuredangle(\mathbf{u}^i, \mathbf{w}^i))$, which is bounded by $\epsilon/20$ by choosing the constant C' appropriately. By standard VC dimension arguments, we therefore have that $\mathbb{P}_{\mathbf{x}\sim S_{\text{train}}}[f(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \epsilon/5$ as long as $m_{\text{train}} \geq \frac{Cdk\log k}{\epsilon^2}$.

Completeness. To prove completeness, suppose that $\mathcal{D}' = \mathcal{N}_d$. Since $\mathcal{U}_{\epsilon''}, \mathcal{F}$ do not depend on X_{test}, we can use Hoeffding bounds to bound the probability of rejection, as well as union bounds over $\mathcal{F} \times \mathcal{F}$ accordingly. In particular, the tester of Lemma 22 will accept with high probability as long as $m_{\text{test}} \geq C\frac{1}{\epsilon'^{4/3}} + Cd\log^2 d = O(\frac{k^2}{\epsilon^2} + d\log^2 d)$ and the tester of the disagreement probabilities of pairs in $\mathcal{F}$ will accept (due to standard Hoeffding and union bounds) with high probability whenever $m_{\text{test}} \geq C\frac{1}{\epsilon^2}\log|\mathcal{F}| = O(\frac{k^2}{\epsilon^2}\log(\frac{k}{\epsilon}))$ (since $|\mathcal{F}| = (k/\epsilon)^{O(k^2)}$ as we need to choose k normals from $\mathcal{U}_{\epsilon''}$ and k elements from $\mathcal{T}_{\epsilon'}$). For the moment matching tester, we require that $m_{\text{test}} \geq Cd^{4\log(k/\epsilon)}$, since the tester would then have to accept with high probability (see also Lemma D.1 in Klivans et al. (2023)). ■

By combining Theorem 31 with Lemmas 24, 27 and 28 we obtain the following bounds for TDS learning general halfspace intersections.

Corollary 32 (TDS Learning Bounds for General Halfspace Intersections) *Let $\eta \in (0, \frac{1}{2})$, $\epsilon > 0$, $\beta \geq 1$ and let $\mathcal{C}$ be the class of intersections of k general halfspaces on $\mathbb{R}^d$.*

(*a*) *There is an $(\epsilon, \delta = 0.02)$-TDS learner for the class $\mathcal{C}_\eta$ of η-balanced intersections that uses $\tilde{O}(d)(\frac{k}{\epsilon\eta})^{O(\frac{k^6}{\epsilon^6})}$ labelled training examples, $d^{O(\log(k/\epsilon))}$ unlabelled test examples and runs in time $\tilde{O}(d^3)(\frac{k}{\epsilon\eta})^{O(\frac{k^6}{\epsilon^6})} + d^{O(\log(k/\epsilon))}(k/\epsilon)^{O(k^2)}$.*

(*b*) *There is an $(\epsilon, \delta = 0.02)$-TDS learner for the class $\mathcal{C}_\eta^\beta$ of η-balanced and β-non-degenerate intersections that uses $\tilde{O}(d) \cdot \frac{1}{\eta^2} \cdot (\frac{k}{\epsilon})^{O(\beta)}$ labelled training examples, $d^{O(\log(k/\epsilon))}$ unlabelled test examples and runs in time $\tilde{O}(d^3) \cdot \frac{1}{\eta^2} \cdot (\frac{k}{\epsilon})^{O(\beta)} + d^{O(\log(1/\epsilon))}(k/\epsilon)^{O(k^2)}$.*

Appendix E. SQ Lower Bounds for TDS Learning

E.1. SQ Lower Bounds for TDS Learning General Halfspaces

In this section, we provide the proof of the SQ lower bound for TDS learning general halfspaces. Recall that the proof consists of two main steps. First, we reduce the problem of biased halfspace detection of Definition 9 to TDS learning halfspaces and then we show that the bias halfspace detection problem is hard in the SQ framework.

E.1.1. DETECTING BIASED HALFSPACES THROUGH TDS LEARNING

For the first ingredient we use the following proposition which we restate here for convenience.

Proposition 33 (Biased Halfspace Detection via TDS Learning) *Let $\mathcal{A}$ be a TDS learning algorithm for general halfspaces over $\mathbb{R}^d$ w.r.t. $\mathcal{N}_d$ with accuracy parameter ϵ and success probability at least 0.95. Suppose $\mathcal{A}$ obtains at most m samples from the training distribution and accesses the test distribution via N SQ queries of tolerance φ (the SQ queries are allowed to depend on the training samples). Then, there exists an algorithm $(\frac{1}{100m}, 10\epsilon)$-biased halfspace detection that uses $N + 1$ SQ queries of tolerance $\min(\varphi, \epsilon)$ and has success probability at least 0.8.*

Proof Without loss of generality, suppose that the algorithm $\mathcal{A}$ uses exactly m samples from the training distribution. We use the following algorithm that uses the TDS learning algorithm $\mathcal{A}$.

- **Given:** Statistical query access to distribution $\mathcal{D}$ over $\mathbb{R}^d$ with tolerance $\min(\varphi, \epsilon)$.
- **Output:** "Accept" or "Reject".

1. Generate $S_{\text{train}} \subset \mathbb{R}^d \times \{\pm 1\}$, of pairs ($\mathbf{x}^i$,-1), where each $\mathbf{x}^i$ is sampled from $\mathcal{N}_d$.
2. Run the TDS learning algorithm $\mathcal{A}$ on the training set S_{train}. Every time $\mathcal{A}$ makes an SQ query to the test distribution, make the same SQ query to $\mathcal{D}$, and return $\mathcal{A}$ the result.
3. If $\mathcal{A}$ returns "Reject", then our algorithm also returns "Reject" and terminates.
4. Otherwise, $\mathcal{A}$ outputs "Accept" and a classifier $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$.
5. Using an SQ query, let $\widehat{\lambda}$ be an estimate up to additive error $\min(\varphi, \epsilon)$ of $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}\left[\widehat{f}(\mathbf{x}) = 1\right]$.
6. If $\widehat{\lambda} > 4\epsilon$, then output "Reject" and terminate.
7. Otherwise, output "Accept" and terminate.

First, we argue that if $\mathcal{D}$ is $\mathcal{N}_d$, then the algorithm above will output "Accept" with probability at least 0.8. For arbitrarily chosen unit vector $\mathbf{w}$, as a parameter τ grows to infinity, the statistical distance between $S_{\text{train}} = \left\{(\mathbf{x}^i, -1)\right\}$ and the set $S'_{\text{train}} = \left\{(\mathbf{x}^i, \text{sign}\left(\mathbf{w}\cdot\mathbf{x}^i - \tau\right))\right\}$ goes to zero. If $\mathcal{A}$ is given S'_{train} and $\mathcal{D} = \mathcal{N}_d$, then the definition of TDS learning requires $\mathcal{A}$ with probability at least 0.95 to accept and output a hypothesis $\widehat{f}$ satisfying $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}\left[\widehat{f}(\mathbf{x}) \neq \text{sign}\left(\mathbf{w}\cdot\mathbf{x} - \tau\right)\right] \leq \epsilon$. Taking the parameter τ to be sufficiently large, we see that if $\mathcal{A}$ is given $S_{\text{train}} = \left\{(\mathbf{x}^i, -1)\right\}$ and $\mathcal{D} = \mathcal{N}_d$, then with probability at least 0.94 the algorithm $\mathcal{A}$ accepts and outputs a hypothesis $\widehat{f}$ satisfying $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}\left[\widehat{f}(\mathbf{x}) \neq -1\right] \leq 2\epsilon$. Therefore, the estimate $\widehat{\lambda}$ will be at most 3ϵ, and we will thus output "Accept".

Now, suppose $\mathcal{D}$ is such that for some unit vector $\mathbf{v}$ and $\tau \in \mathbb{R}$ we have $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\cdot\mathbf{v} \geq \tau] \geq 10\epsilon$ and $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\mathbf{x}\cdot\mathbf{v} \geq \tau] \leq \frac{1}{100m}$. Besed on the set $S_{\text{train}} = \left\{(\mathbf{x}^i, -1)\right\}$, define the set S''_{train} as $S''_{\text{train}} = \left\{(\mathbf{x}^i, \text{sign}\left(\mathbf{v}\cdot\mathbf{x}^i - \tau\right))\right\}$. If the algorithm $\mathcal{A}$ were given the set S''_{train} instead of S_{train} as the training set, then the definition of TDS learning would require $\mathcal{A}$ with probability at least 0.95 either to output "Reject" or give a hypothesis $\widehat{f}$ satisfying $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}\left[\widehat{f}(\mathbf{x}) \neq \text{sign}\left(\mathbf{v}\cdot\mathbf{x} - \tau\right)\right] \leq \epsilon$. Since $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_d}[\mathbf{x}\cdot\mathbf{v} \geq \tau] \leq \frac{1}{100m}$ and $|S_{\text{train}}| = |S''_{\text{train}}| = m$, we see via a union bound that that the statistical distance between S_{train} and S''_{train} is at most 0.01. Thus, in the algorithm above, the algorithm $\mathcal{A}$ with probability at least 0.94 indeed either outputs "Reject" or gives a hypothesis $\widehat{f}$ satisfying $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}\left[\widehat{f}(\mathbf{x}) \neq \text{sign}\left(\mathbf{v}\cdot\mathbf{x} - \tau\right)\right] \leq \epsilon$. In the former case, our algorithm will also output "Reject". In the latter case we will have $\widehat{\lambda} > 9\epsilon$, since $\mathcal{D}$ is such that $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\cdot\mathbf{v} \geq \tau] \geq 10\epsilon$. Therefore, in this case too our algoirthm outputs "Reject", which completes the proof. ■

E.1.2. LOWER BOUNDS FOR DETECTING BIASED HALFSPACES

We now provide a proof for the second ingredient, namely, that no efficient SQ algorithm can solve the problem of detecting biased halfspaces, i.e., the following proposition (restated here for convenience).

Proposition 34 (SQ Lower Bounds for Biased Halfspace Detection) *For $\epsilon > 0$, set $d = \frac{1}{\epsilon^{1/4}}$. Then, for all sufficiently small ϵ, the following is true. Suppose $\mathcal{A}$ is an SQ algorithm for $(d^{-\ln(1/\epsilon)}, 10\epsilon)$-biased halfspace detection problem over $\mathbb{R}^d$, and $\mathcal{A}$ has a success probability of at least $2/3$. Then, $\mathcal{A}$ either has to use SQ tolerance of $d^{-\Omega(\frac{\log 1/\epsilon}{\log\log 1/\epsilon})}$, or make $2^{d^{\Omega(1)}}$ SQ queries.*

To prove the above claim, we first construct a one-dimensional distribution $\mathcal{D}_1$ that approximately matches the low-degree moments of $\mathcal{N}_d$, while having a lot of probability mass above a certain threshold.

Proposition 35 *For $\epsilon > 0$, let k_0 be defined as $k_0 = \frac{\ln 1/\epsilon}{100 \ln\ln 1/\epsilon}$. If ϵ is sufficiently small, then there exists a distribution $\mathcal{D}_1$ supported on a finite subset of $\mathbb{R}$, satisfying*

$$\left| \mathop{\mathbb{E}}_{x\sim\mathcal{D}_1}\left[x^i\right] - \mathop{\mathbb{E}}_{x\sim\mathcal{N}_1}\left[x^i\right] \right| \le \frac{1}{k_0^{10k_0}},$$

for every $i \in \{0, \cdots, 10k_0\}$ while also satisfying $\mathbb{P}_{x\sim\mathcal{D}_1}[x \ge t] \ge 12\epsilon$, for some t for which $\mathbb{P}_{x\sim\mathcal{N}_1}[x \ge t] \le \epsilon^{\frac{1}{4}\ln 1/\epsilon}$.

Proof We will first construct a distribution $\mathcal{D}'_1$ that satisfies the conditions above, but does not have finite support. Afterwards, we will discretize $\mathcal{D}'_1$.

We take $t := \ln 1/\epsilon$ and observe that

$$\begin{aligned} \mathop{\mathbb{P}}_{x\sim\mathcal{N}_1}[x \ge t] &= \frac{1}{\sqrt{2\pi}} \int_{\ln 1/\epsilon}^{\infty} e^{-x^2/2}\, dx \le \frac{e^{-(\ln 1/\epsilon)^2/2}}{\sqrt{2\pi}} \int_0^{\infty} e^{-x \ln 1/\epsilon}\, dx \\ &= \underbrace{\frac{e^{-(\ln 1/\epsilon)^2/2}}{\sqrt{2\pi}\ln 1/\epsilon} \le \epsilon^{\frac{1}{4}\ln 1/\epsilon}}_{\text{For } \epsilon \text{ sufficiently small.}}. \end{aligned} \tag{2}$$

Let τ be the real number for which $\mathbb{P}_{x\sim\mathcal{N}_1}\left[x \in [0, \tau]\right] = 13\epsilon$. From Equation 2, we see that for all sufficiently small ϵ it is the case that $\tau < \epsilon$. We define $\mathcal{D}'_1$ the following way: to sample $z \sim \mathcal{D}'_1$ (i) sample $x \sim \mathcal{N}_1$ (ii) if $x \in [0, \tau]$, then $z = t$ (iii) otherwise, $z = x$. Since $\mathbb{P}_{x\sim\mathcal{N}_1}\left[x \in [0, \tau]\right] = 13\epsilon$, we see that $\mathbb{P}_{x\sim\mathcal{D}'_1}\left[x \ge t\right] \ge 13\epsilon$. Furthermore, we see that for every $i \in \{0, \cdots, 10k_0\}$

$$\begin{aligned} \left| \mathop{\mathbb{E}}_{x\sim\mathcal{D}'_1}\left[x^i\right] - \mathop{\mathbb{E}}_{x\sim\mathcal{N}_1}\left[x^i\right] \right| \le t^{k_0} \mathop{\mathbb{P}}_{x\sim\mathcal{N}_1}\left[x \in [0, \tau]\right] = 12\epsilon \cdot (\ln 1/\epsilon)^{\frac{\ln 1/\epsilon}{100 \ln\ln 1/\epsilon}} = \\ = 12\epsilon^{0.99} \underbrace{\le \frac{1}{2} \cdot \left(\frac{100 \ln\ln 1/\epsilon}{\ln 1/\epsilon}\right)^{\frac{\ln 1/\epsilon}{10 \ln\ln 1/\epsilon}}}_{\text{For } \epsilon \text{ sufficiently small.}} = \frac{1}{2k_0^{10k_0}}. \end{aligned}$$

Overall, we have so far shown that $\mathbb{P}_{x\sim\mathcal{D}'_1}\left[x \ge t\right] \ge 13\epsilon$ and $\left|\mathbb{E}_{x\sim\mathcal{D}'_1}\left[x^i\right] - \mathbb{E}_{x\sim\mathcal{N}_1}\left[x^i\right]\right| < \frac{1}{2k_0^{10k_0}}$, but $\mathcal{D}'_1$ is not supported on a finite subset of $\mathbb{R}$. We will now construct a finitely-supported distribution $\mathcal{D}_1$ via the probabilistic method. Obtain $\mathcal{D}_1$ as the empirical distribution over K i.i.d. samples from $\mathcal{D}'_1$. Since all moments of $\mathcal{D}'_1$ are bounded, as K grows to infinity, for all $i \in \{0, \cdots, 10k_0\}$ the quantity $\mathbb{E}_{x\sim\mathcal{D}_1}\left[x^i\right]$ converges in probability to $\mathbb{E}_{x\sim\mathcal{D}'_1}\left[x^i\right]$, and the quantity $\mathbb{P}_{x\sim\mathcal{D}_1}\left[x \ge t\right]$ converges in probility to $\mathbb{P}_{x\sim\mathcal{D}'_1}\left[x \ge t\right]$. Thus, for a sufficiently large K, we have $\mathbb{P}_{x\sim\mathcal{D}_1}\left[x \ge t\right] \ge 12\epsilon$ and $\left|\mathbb{E}_{x\sim\mathcal{D}_1}\left[x^i\right] - \mathbb{E}_{x\sim\mathcal{N}_1}\left[x^i\right]\right| < \frac{1}{k_0^{10k_0}}$, with non-zero probability over the choice of $\mathcal{D}_1$, which completes the proof. ■

We now apply the following theorem which is implicit in Diakonikolas et al. (2023b) to obtain a distribution $\mathcal{D}$ over $\mathbb{R}$ that has a lot of probability mass above a certain threshold and whose moments match $\mathcal{N}_1$ exactly.

Theorem 36 (Implicit in Diakonikolas et al. (2023b)) *Let k be a sufficiently large positive integer and let $\mathcal{D}_0$ be a distribution supported on a finite subset of $\mathbb{R}$, and suppose that for every $i \in \{0, \cdots, 10k\}$ we have*

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_0} \left[x^i\right] - \mathop{\mathbb{E}}_{x \sim \mathcal{N}_1} \left[x^i\right] \right| \leq \frac{1}{k^{10k}}, \tag{3}$$

then there exists a distribution $\mathcal{D}_1$ with the same support as $\mathcal{D}_0$ with $\mathbb{E}_{x \sim \mathcal{D}_0} \left[x^i\right] = \mathbb{E}_{x \sim \mathcal{N}_1} \left[x^i\right]$ for every $i \in \{0, \cdots, k\}$, and also satisfying

$$\mathop{\mathbb{P}}_{x \sim \mathcal{D}_1} [x = x_0] \geq 0.9 \mathop{\mathbb{P}}_{x \sim \mathcal{D}_0} [x = x_0]$$

for every x_0 in the support of $\mathcal{D}_0$.

The proof is equivalent to the proof given by Diakonikolas et al. (2023b), but is provided here with slight modifications for completeness. We will need the following fact.

Fact 1 *Let p be a polynomial over $\mathbb{R}$ of degree at most k, and let $\mathbb{E}_{x \sim \mathcal{N}_1} \left[\left(p(x)\right)^2\right] \leq 1$. Then, each coefficient of p has absolute value of at most 2^{k+1}.*

Proof We will use the Hermite polynomials. Recall that for $i = 0, 1, 2, \cdot$ Hermite polynomials $\{H_i\}$ are the unique collection of polynomials over $\mathbb{R}$ that are orthogonal with respect to Gaussian distribution. In other words $\mathbb{E}_{x \in \mathcal{N}_1}[H_i(x) H_j(x)] = 0$ whenever $i \neq j$. In this work, we normalize the Hermite polynomials to further satisfy $\mathbb{E}_{x \in \mathcal{N}_1}[H_i(x) H_i(x)] = 1$. It is a standard fact from theory of orthogonal polynomials that $H_0(x) = 1$, $H_1(x) = x$ and for $i \geq 2$ Hermite polynomials satisfy the following recursive identity:

$$H_{i+1}(x) \cdot \sqrt{(i+1)!} = x H_i(x) \cdot \sqrt{i!} - i \cdot H_{i-1}(x) \cdot \sqrt{(i-1)!}$$

It follows immediately from the recursion relation that Each coefficient of H_i is bounded by 2^i in absolute value. We expand $P(x)$ as a sum of Hermite polynomials[2]:

$$p(x) = \sum_{i=0}^{k} \alpha_i H_i(x) \tag{4}$$

Due to orthogonality of Hermite polynomials, we have:

$$\sum_{i=0}^{k} \alpha_i^2 = \mathop{\mathbb{E}}_{x \in \mathcal{N}(0,1)} [(p(x))^2] \leq 1$$

In particular, this implies that each coefficient α_i is bounded by 1 in absolute value. Combining this with Equation 4, the fact that each coefficient of H_i is bounded by 2^i in absolute value, we see that each coefficient of p is bounded by $\sum_{i=0}^{k} 2^i < 2^{k+1}$ in absolute value. ■

Proof of Theorem 36, implicit in Diakonikolas et al. (2023b). Provided here for completeness.

2. Note that the expansion below is always possible for a degree k polynomial because polynomials of the form H_i have degree at most k and are linearly independent, because they are orthonormal with respect to the standard Gaussian distribution.

We first restate the setting of the theorem. Let k be a sufficiently large positive integer and let $\mathcal{D}_0$ be a distribution supported on a finite subset of $\mathbb{R}$, and suppose that for every $i \in \{0, \cdots, 10k\}$ we have

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_0} \left[x^i\right] - \mathop{\mathbb{E}}_{x \sim \mathcal{N}_1} \left[x^i\right] \right| \leq \frac{1}{k^{10k}}, \tag{5}$$

then we would like to show that there exists a distribution $\mathcal{D}_1$ with the same support as $\mathcal{D}_0$ satisfying $\mathbb{E}_{x \sim \mathcal{D}_0} \left[x^i\right] = \mathbb{E}_{x \sim \mathcal{N}_1} \left[x^i\right]$ for every $i \in \{0, \cdots, k\}$, and also satisfying

$$\Pr_{x \sim \mathcal{D}_1} [x = x_0] \geq 0.9 \Pr_{x \sim \mathcal{D}_0} [x = x_0]$$

for every x_0 in the support of $\mathcal{D}_0$.

Let N denote the number of elements in the support of $\mathcal{D}_0$ and let $\{x_1, \cdots, x_N\}$ be the elements in the support of $\mathcal{D}_0$. Consider the following linear program:

$$\begin{aligned} \text{Find} \quad & \mu_{x_1}, \cdots \mu_{x_N} \\ s.t. \quad & \mathop{\mathbb{E}}_{x \sim \mathcal{D}_0} [\mu_x p(x)] = \mathop{\mathbb{E}}_{x \sim \mathcal{N}_1} [p(x)] \qquad \text{for every polynomial } p \text{ of degree at most } k \\ & \mu_{x_j} \geq 0.9 \qquad \text{for all } j \in \{1, \cdots N\} \end{aligned}$$

If the linear program above is feasible, then the proposition will be satisfied by a distribution $\mathcal{D}_1$ supported on $x_1, \cdots x_N$ that has probability $\mu_{x_j} \Pr_{x \sim \mathcal{D}_0} [x = x_j]$ on each x_j (note that $\mathcal{D}_1$ is indeed a probability distribution because the equality $\sum_j \mu_{x_j} \Pr_{x \sim \mathcal{D}_0} [x = x_j] = 1$ follows by the constraint in the linear program when p is identically equal to 1).

The linear program above is feasible if and only if its dual linear program is infeasible. The dual linear program is as follows:

$$\begin{aligned} \text{Find} \quad & \text{polynomial } p \text{ of degree at most } k, \\ s.t. \quad & p(x_j) \geq 0 \qquad \text{for all } j \in \{1, \cdots, N\}, \\ & \mathop{\mathbb{E}}_{x \sim \mathcal{N}_1} [p(x)] < 0.9 \mathop{\mathbb{E}}_{x \sim \mathcal{D}_0} [p(x)]. \end{aligned} \tag{6}$$

It is now shown that the above is indeed infeasible if $\mathcal{D}_0$ is such that for every $i \in \{0, \cdots, 10k\}$ we have $\left|\mathbb{E}_{x \sim \mathcal{D}_0} \left[x^i\right] - \mathbb{E}_{x \sim \mathcal{N}_1} \left[x^i\right]\right| \leq \frac{1}{k^{10k}}$. For the sake of contradiction, suppose that the linear program above is feasible and is satisfied by some polynomial p. Without loss of generality, assume that $\mathbb{E}_{x \sim \mathcal{N}_1} \left[(p(x))^2\right] = 1$, because otherwise one could rescale p while still satisfying the dual linear program above. By Fact 1 each coefficient of p has absolute value of at most 2^{k+1}. This implies that each coefficient of p^2 has an absolute value of at most 8^{k+1} and each coefficient of p^4 has an absolute value of at most 32^{k+1}. Combining these coefficient bounds with Equation 5, and applying the triangle inequality, we see that

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_0} [p(x)] - \mathop{\mathbb{E}}_{x \sim \mathcal{N}_1} [p(x)] \right| \leq \frac{(k+1)2^{k+1}}{k^{10k}}, \tag{7}$$

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_0} \left[(p(x))^2\right] - \mathop{\mathbb{E}}_{x \sim \mathcal{N}_1} \left[(p(x))^2\right] \right| \leq \frac{(2k+1)8^{k+1}}{k^{10k}}, \tag{8}$$

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_0} \left[(p(x))^4\right] - \mathop{\mathbb{E}}_{x \sim \mathcal{N}_1} \left[(p(x))^4\right] \right| \leq \frac{(4k+1)32^{k+1}}{k^{10k}}. \tag{9}$$

This allows us to upper-bound $\mathbb{E}_{x\sim\mathcal{D}_0}[|p(x)|]$ as follows, where the first inequality follows by Equation 7, the second by the fact that p satisfies the Linear Program 6 and the equality because p is positive on the support of $\mathcal{D}_0$ due satisfying the Linear Program 6.

$$\frac{(k+1)2^{k+1}}{k^{10k}} \geq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_0}[p(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{N}_1}[p(x)] > 0.1 \mathop{\mathbb{E}}_{x\sim\mathcal{D}_0}[p(x)] = 0.1 \mathop{\mathbb{E}}_{x\sim\mathcal{D}_0}[|p(x)|] \tag{10}$$

However, we can also lower-bound $\mathbb{E}_{x\sim\mathcal{D}_0}[|p(x)|]$ in the following way

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_0}[|p(x)|] \geq \overbrace{\frac{\left(\mathbb{E}_{x\sim\mathcal{D}_0}\left[(p(x))^2\right]\right)^{3/2}}{\mathbb{E}_{x\sim\mathcal{D}_0}\left[(p(x))^4\right]}}^{\text{By generalized Holder inequality.}} \geq \overbrace{\frac{\left(\mathbb{E}_{x\sim\mathcal{N}_1}\left[(p(x))^2\right] - \frac{(2k+1)8^{k+1}}{k^{10k}}\right)^{3/2}}{\mathbb{E}_{x\sim\mathcal{N}_1}\left[(p(x))^4\right] + \frac{(4k+1)32^{k+1}}{k^{10k}}}}^{\text{By Equations 8 and 9.}} \geq$$

$$\geq \frac{\left(1 - \frac{(2k+1)8^{k+1}}{k^{10k}}\right)^{3/2}}{(4k+1)32^{k+1}k!! + \frac{(4k+1)32^{k+1}}{k^{10k}}} \geq \frac{1}{k^k}, \quad \text{for sufficiently large } k. \tag{11}$$

where the prior to last inequality follows form the fact that $\mathbb{E}_{x\sim\mathcal{N}_1}\left[(p(x))^4\right] \leq (4k+1)32^{k+1}k!!$, as each coefficient of p^4 is at most 32^{k+1} in absolute value. Overall, we see that Equations 11 and 10 cannot hold simultaneously for a sufficiently large k , contradiction. ■

In order to conclude the proof of Proposition 34, we a tool from Diakonikolas et al. (2023a).

Theorem 37 (Special case of Diakonikolas et al. (2023a)) *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ such that for every $i \in \{0, \cdots, k\}$ we have $\mathbb{E}_{x\sim\mathcal{D}}\left[x^i\right] = \mathbb{E}_{x\sim\mathcal{N}_1}\left[x^i\right]$. For a unit vector $\mathbf{v}$, let $\mathcal{D}_{\mathbf{v}}$ denote the distribution over $\mathbb{R}^d$ such that for $\mathbf{x} \sim \mathcal{D}_{\mathbf{v}}$ (i) the projection $\mathbf{x}\cdot\mathbf{v}$ is distributed as $\mathcal{D}$ (ii) the projection of $\mathbf{x}$ onto the subspace orthogonal to $\mathbf{v}$ is distributed as $\mathcal{N}_{d-1}$ independently from $\mathbf{x}\cdot\mathbf{v}$. Suppose $\mathcal{A}$ is an SQ algorithm that distinguishes with success probability at least $2/3$ the distribution $\mathcal{N}_d$ from $\mathcal{D}_{\mathbf{v}}$, with $\mathbf{v}$ a uniformly random unit vector. Then, $\mathcal{A}$ either needs to use SQ tolerance of $k^{10k}d^{-0.1k}$ or make $2^{d^{\Omega(1)}}$ SQ queries.*

E.1.3. TDS LEARNING GENERAL HALFSPACES IS HARD FOR SQ ALGORITHMS

Finally, we prove Theorem 8 by combining the reduction of Proposition 10 with the SQ lower bound of Proposition 11 to obtain an SQ lower bound for TDS learning of general halfspaces.

Recall that in the setting of Theorem 8 for $\epsilon > 0$, we let d be chosen as $d = \frac{1}{\epsilon^{1/4}}$. Suppose Theorem 8 is false. Then for a sequence of ϵ approaching 0 there is a TDS learning algorithm $\mathcal{A}$ for general halfspaces over $\mathbb{R}^d$ with accuracy parameter ϵ and success probability at least 0.95. The algorithm $\mathcal{A}$ obtains at most $d^{\frac{\log 1/\epsilon}{\log\log 1/\epsilon}}$ samples from the training distribution and accesses the testing distribution via $2^{d^{o(1)}}$ SQ querries of tolerance at least $d^{-o(\frac{\log 1/\epsilon}{\log\log 1/\epsilon})}$.

Combining this with Proposition 10, we see that for an infinite sequence of values of positive ϵ that approaches zero, there exists an algorithm for $(\frac{1}{100}d^{-\frac{\log 1/\epsilon}{\log\log 1/\epsilon}}, 10\epsilon)$-biased halfspace detection that uses $2^{d^{o(1)}}$ SQ querries of tolerance $\min(d^{-o(\frac{\log 1/\epsilon}{\log\log 1/\epsilon})}, \epsilon) = d^{-o(\frac{\log 1/\epsilon}{\log\log 1/\epsilon})}$ and has success probability at least 0.8. However, for sufficiently small values of ϵ, this directly contradicts Proposition 11. This finishes the proof of Theorem 8.

E.2. SQ Lower Bounds for Intersections of two Homogeneous Halfspaces

In order to prove Theorem 12, it suffices to reduce the anti-concentration detection problem of Theorem 13 to TDS learning of two homogeneous halfspaces.

The reduction follows the template of the proof of Proposition 10. In this case, we construct a distringuisher for the AC detection problem (between the two options (1) $\mathcal{N}_d$ and (2) $\mathcal{D}'$ described in Theorem 13) by providing training examples of the form $(\mathbf{x}, -1)$, $\mathbf{x} \sim \mathcal{N}_d$ to the input of the TDS algorithm and the SQ oracle for the unknown distribution as an oracle to the test marginal.

The training data are with high probability consistent with the intersection of the halfspaces $H_1 = \{\mathbf{x} : (\sqrt{\alpha}\mathbf{u} + \sqrt{1-\alpha}\mathbf{v}) \cdot \mathbf{x} \geq 0\}$ and $H_2 = \{\mathbf{x} : (\sqrt{\alpha}\mathbf{u} - \sqrt{1-\alpha}\mathbf{v}) \cdot \mathbf{x} \geq 0\}$, where $\mathbf{v}, \mathbf{u} \in \mathbb{S}^{d-1}$, $V = \{\mathbf{x} : \mathbf{v} \cdot \mathbf{x} = 0\}$ is the subspace where $\mathcal{D}'$ assigns non-negligible mass, $\mathbf{u} \perp \mathbf{v}$ and $\alpha \in (0, 1/2)$ is arbitrarily small (even exponentially in $d, \frac{1}{\epsilon}$). Assume, also, that the mass, under $\mathcal{D}'$, of $V \cap \{\mathbf{x} : \mathbf{u} \cdot \mathbf{x} \geq 0\}$ is greater than the mass of $V \cap \{\mathbf{x} : \mathbf{u} \cdot \mathbf{x} < 0\}$ (otherwise, note that the training data are also consistent w.h.p. with the intersection of the complement of H_1 with the complement of H_2).

Suppose that the TDS algorithm rejects. Then, we have a certificate that the test data are not Gaussian and therefore we are in the case (2) of the distinguishing problem (w.h.p.). If the TDS algorithm accepts and outputs some hypothesis $\widehat{f}$, then we query $\mathbb{P}[\widehat{f}(\mathbf{x}) = 1]$ to the SQ oracle for the test marginal. If the test marginal was the Gaussian, then the value of the query should be very close to 0 (because, upon acceptance, $\widehat{f}$ achieves low error). If the test marginal was $\mathcal{D}'$, then the value of the query should be bounded away from 0, because $\mathcal{D}'$ assigns non-negligible mass to the positive region of the intersection and $\widehat{f}$ must achieve low error. Hence, the value of the query indicates the answer to the distinguishing problem.

E.3. SQ Lower Bounds under Non-Degeneracy Condition

In Section C.2 we define a non-degeneracy condition (Definition 26) which is sufficient to obtain an exponential improvement for the problem of approximately retrieving the relevant subspace (see Lemma 27). This implies improved performance for our TDS learners for halfspace intersections. Importantly, our SQ lower bounds (Theorems 12 and 15) hold even for under the non-degeneracy condition and this enables us to compare our upper and lower bounds under this condition.

For Theorem 12, the unknown intersection of the hard construction is non-degenerate, because it corresponds to an intersection of two halfspaces with normals $\mathbf{w}_1, \mathbf{w}_2$ such that $\mathbf{w}_1, \mathbf{w}_2$ are pointing almost in opposite directions. This implies that after projecting $\mathbf{w}_2$ on the subspace orthogonal to $\mathbf{w}_1$, we obtain a direction $\mathbf{w}'$ such that the halfspace $\{\mathbf{x} : \mathbf{w}' \cdot \mathbf{x} \geq 0\}$ is consistent with all of the points in the interior of the unknown intersection and therefore, by Lemma 18, there is significant variance reduction in the direction of $\mathbf{w}'$. Overall, the constructed intersection is 2-non-degenerate.

For Theorem 15, the construction corresponds to an intersection of two halfspaces with normals $\mathbf{w}_1, \mathbf{w}_2$ such that $\mathbf{w}_1, \mathbf{w}_2$ are pointing (w.h.p. as d increases) in almost orthogonal directions. In this case, we do not apply Lemma 18 directly, because the statement is not tight when the residual vector $\mathbf{u} = \frac{\mathbf{w}_2 - \mathrm{proj}_{\mathbf{w}_1} \mathbf{w}_2}{\|\mathbf{w}_2 - \mathrm{proj}_{\mathbf{w}_1} \mathbf{w}_2\|_2}$ is very close to $\mathbf{w}_2$. Instead, we refer to the proof of Lemma 18, which implies that, if $\mathbf{u} \cdot \mathbf{w}_2$ is sufficiently close to 1, then we have variance reduction along $\mathbf{u}$ that indeed scales proportionally to the variance reduction along $\mathbf{w}_2$ and, hence, the corresponding intersection is 2-non-degenerate.