Testable Learning with Distribution Shift

Adam R. Klivans* UT Austin
Konstantinos Stavropoulos† UT Austin
Arsen Vasilyan‡ MIT

May 22, 2024

Abstract

We revisit the fundamental problem of learning with distribution shift, in which a learner is given labeled samples from training distribution $\mathcal{D}$, *unlabeled* samples from test distribution $\mathcal{D}'$ and is asked to output a classifier with low test error. The standard approach in this setting is to bound the loss of a classifier in terms of some notion of distance between $\mathcal{D}$ and $\mathcal{D}'$. These distances, however, seem difficult to compute and do not lead to efficient algorithms.

We depart from this paradigm and define a new model called *testable learning with distribution shift*, where we *can* obtain provably efficient algorithms for certifying the performance of a classifier on a test distribution. In this model, a learner outputs a classifier with low test error whenever samples from $\mathcal{D}$ and $\mathcal{D}'$ pass an associated test; moreover, the test must accept (with high probability) if the marginal of $\mathcal{D}$ equals the marginal of $\mathcal{D}'$. We give several positive results for learning well-studied concept classes such as halfspaces, intersections of halfspaces, and decision trees when the marginal of $\mathcal{D}$ is Gaussian or uniform on $\{\pm 1\}^d$. Prior to our work, no efficient algorithms for these basic cases were known without strong assumptions on $\mathcal{D}'$.

For halfspaces in the realizable case (where there exists a halfspace consistent with both $\mathcal{D}$ and $\mathcal{D}'$), we combine a moment-matching approach with ideas from active learning to simulate an efficient oracle for estimating disagreement regions. To extend to the non-realizable setting, we apply recent work from testable (agnostic) learning. More generally, we prove that any function class with low-degree $\mathcal{L}_2$-sandwiching polynomial approximators can be learned in our model. Since we require $\mathcal{L}_2$-sandwiching (instead of the usual $\mathcal{L}_1$ loss), we cannot directly appeal to convex duality and instead apply constructions from the pseudorandomness literature to obtain the required approximators. We also provide lower bounds to show that the guarantees we obtain on the performance of our output hypotheses are best possible up to constant factors, as well as a separation showing that realizable learning in our model is incomparable to (ordinary) agnostic learning.

*klivans@cs.utexas.edu. Supported by NSF award AF-1909204 and the NSF AI Institute for Foundations of Machine Learning (IFML).

†kstavrop@cs.utexas.edu. Supported by NSF award AF-1909204, the NSF AI Institute for Foundations of Machine Learning (IFML) and by scholarships from Bodossaki Foundation and Leventis Foundation.

‡vasilyan@mit.edu. 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.

1 Introduction

Mitigating distribution shift remains one of the major challenges of machine learning. Training distributions can deviate significantly from test distributions, and pre-trained models are commonly deployed without a precise understanding of these differences. In such cases, a model may have poor performance with potentially dangerous consequences. For example, several recent studies in the AI/healthcare community highlight the lack of generalization among many AI models trained to detect disease (e.g., skin cancer or pneumonia), often due to distribution shift. As such, developing best practices for using these models in a clinical setting remains a vexing and difficult problem [ZBL+18, WOD+21, TCK+22].

The computational landscape of traditional supervised learning— where training sets and tests are drawn from the same distribution— is by now well understood. There is a rich literature of efficient algorithms and computational hardness results for broad sets of concept classes and distributions. In contrast, little is known in terms of efficient algorithms for classification in the context of distribution shift or domain adaptation. The most common approach is to prove a generalization bound in terms of some notion of distance between $\mathcal{D}$ and $\mathcal{D}'$ [BDBCP06, BDBC+10, MMR09]. These distances, however, involve an enumeration of all functions in the underlying concept class and seem difficult to compute. Other recent work requires oracles for empirical risk minimization [GKKM20, KK21] or the existence of distribution-free reliable learners, which are believed to require superpolynomial time for even simple concept classes (e.g., reliably learning conjunctions is known to be harder than PAC learning DNF formulas) [KK21, Section 4.2].

In this work we define a new model called *testable learning with distribution shift* (TDS learning) and show that this model does admit efficient algorithms for several well-studied concept classes and distributions. Inspired by recent work in testable learning [RV23, GKK23, GKSV23a, GKSV23b], we allow a learner to reject unless $\mathcal{D}$ and $\mathcal{D}'$ pass an efficiently computable test. Whenever the test accepts, the learner outputs a classifier that is assured to have low error with respect to $\mathcal{D}'$. Further, we require that the test accept with high probability whenever the marginal of $\mathcal{D}$ equals the marginal of $\mathcal{D}'$. This approach allows us to take no assumptions on $\mathcal{D}'$ whatsoever and still provide meaningful guarantees.

It is easy to see that TDS learning generalizes the traditional PAC model of learning, and, moreover, TDS learning seems considerably more challenging. For example, even an algorithm to amplify the success probability of a TDS learner is nontrivial, since we do not get to see labeled examples from $\mathcal{D}'$ (we show how to do this in Appendix C). It is also tempting to apply property testing algorithms in this setting to "detect" when $\mathcal{D}$ is "close" to $\mathcal{D}'$, but even for simple cases, distribution testing requires an exponential (in the dimension) number of samples (see e.g. [Can22]). While testable learning and TDS learning both encounter similar issues, they are fundamentally distinct models. Specifically, the realizable setting, where there exists a classifier with zero train and test loss, is a trivial case in testable learning. We further discuss separations among these models in Section 1.3.

1.1 Our Results

Here we formally define TDS learning and summarize our main results. For readability, we have placed some notation and basic definitions in Section 3.

Learning Setup. Let $\mathcal{C}$ be a function class over $\mathbb{R}^d$ and $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. Suppose $\mathcal{A}$ is given as input a set S_{train} consisting of i.i.d. examples from $\mathcal{D}$ labelled by some $f \in \mathcal{C}$, together with a set of i.i.d. unlabelled examples X_{test} from some distribution $\mathcal{D}_{\mathcal{X}}^{\text{test}}$ over $\mathbb{R}^d$. The algorithm $\mathcal{A}$ is allowed to either output REJECT or (ACCEPT, $\widehat{f}$) for some concept $\widehat{f}$. The algorithm $\mathcal{A}$ is a **TDS-learning algorithm** for $\mathcal{C}$ under distribution $\mathcal{D}$ if it satisfies the following two properties:

1. **Soundness.** With probability $1 - \delta$, if the algorithm $\mathcal{A}$ outputs (ACCEPT, $\widehat{f}$), then hypothesis $\widehat{f}$ satisfies $\mathbb{P}_{\mathbf{x} \in \mathcal{D}_{\mathcal{X}}^{\text{test}}}[f(\mathbf{x}) \neq \widehat{f}(\mathbf{x})] \leq \epsilon$.

2. **Completeness.** If $\mathcal{D}_{\mathcal{X}}^{\text{test}} = \mathcal{D}$, then with probability $1 - \delta$, the algorithm $\mathcal{A}$ outputs (ACCEPT, $\widehat{f}$).

TDS Learning: the Agnostic Setting. Sometimes the training data or the testing data cannot be captured perfectly by any function in the function class $\mathcal{C}$ and, instead, follow labeled distributions $\mathcal{D}_{\mathcal{XY}}^{\text{train}}, \mathcal{D}_{\mathcal{XY}}^{\text{test}}$, where the marginal of $\mathcal{D}_{\mathcal{XY}}^{\text{train}}$ is $\mathcal{D}_{\mathcal{X}}^{\text{train}} = \mathcal{D}$ and $\mathcal{D}_{\mathcal{XY}}^{\text{train}}, \mathcal{D}_{\mathcal{XY}}^{\text{test}}$ are otherwise arbitrary. We extend our setup to apply in this setting as well. To this end, a key quantity is the **smallest sum** of expected training error and expected test error among all functions in the concept class $\mathcal{C}$, i.e. $\lambda = \min_{f \in \mathcal{C}} \operatorname{err}(f; \mathcal{D}_{\mathcal{XY}}^{\text{train}}) + \operatorname{err}(f; \mathcal{D}_{\mathcal{XY}}^{\text{test}})$, where $\operatorname{err}(f; \mathcal{D}_{\mathcal{XY}}) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathcal{XY}}}[y \neq f(\mathbf{x})]$. We denote this quantity as λ, and note that it is standard in the domain adaptation literature (see, e.g., [BDBCP06, BCK+07, BDBC+10, DLLP10]).

With this definition at hand, we modify the soundness condition to require that with probability $1 - \delta$, if the algorithm $\mathcal{A}$ outputs (ACCEPT, $\widehat{f}$), then hypothesis $\widehat{f}$ satisfies $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathcal{XY}}^{\text{test}}}[y \neq \widehat{f}(\mathbf{x})] \leq O(\lambda) + \epsilon$. In Theorem 2.13, we show that a dependence of $\Omega(\lambda)$ is unavoidable.

Proposition 1.1 (Informal)**.** *No TDS learning algorithm can have an error guarantee better than* $\Omega(\lambda) + \epsilon$.

Results. We show that TDS learning can be achieved efficiently for a number of natural high-dimension function classes. These include halfspaces, decision trees, intersections of halfspaces and low-depth formulas. See Table 1 for the full list.

	Function class	Training Distribution	TDS Setting	Run-time
1	Homogeneous halfspaces	Isotropic Log-Concave	Agnostic	$\operatorname{poly}(d/\epsilon)$ (Theorem 2.1)
2	General halfspaces	Standard Gaussian	Realizable	$d^{O(\log 1/\epsilon)}$ (Theorem 2.7)
3	General halfspaces	Standard Gaussian Uniform on $\{\pm 1\}^d$	Agnostic	$d^{\tilde{O}(1/\epsilon^2)}$ (Corollary 6.9)
4	Intersection of ℓ halfspaces	Standard Gaussian Uniform on $\{\pm 1\}^d$	Agnostic	$d^{\tilde{O}(\ell^6/\epsilon^2)}$ (Corollary 6.9)
5	Decision trees of size s	Uniform on $\{\pm 1\}^d$	Agnostic	$d^{O(\log(s/\epsilon))}$ (Corollary 6.6)
6	Formulas of size s, depth ℓ	Uniform on $\{\pm 1\}^d$	Agnostic	$d^{\sqrt{s}\cdot O(\log(s/\epsilon))^{\frac{5\ell}{2}}}$ (Corollary 6.7)

Table 1: Our TDS learning results for various function classes. Since agnostic TDS learning is more general than realizable TDS learning, algorithms for the agnostic setting also apply to the realizable setting.

Given the abundance of positive results, it is natural to ask whether TDS learning can always be achieved efficiently for any function class $\mathcal{F}$ that can be efficiently PAC-learned under a distribution $\mathcal{D}$. We answer this question in the negative by proving separations between TDS learning and PAC learning. Our separations hold for the natural and well-studied function classes of monotone functions over $\{\pm 1\}^d$ and convex sets over $\mathbb{R}^d$ (under uniform distribution on $\{\pm 1\}^d$ and the standard Gaussian distribution respectively). Even though for these function classes there are well-known PAC-learning algorithms [BT96, KOS08] that run in time $2^{\tilde{O}(\sqrt{d}\operatorname{poly}(1/\epsilon))}$, we show that any TDS-learning algorithm for these function classes needs to run in time $2^{\Omega(d)}$.

1.2 Techniques

Here we summarize the technical ideas that we use to develop the TDS learning algorithms in Table 1.

Moment Matching/Sandwiching Polynomials. We present a general approach for obtaining TDS learning algorithms for a wide variety of function classes via a **moment matching** approach. In brief, the algorithm for this approach is as follows:

- Estimate all the degree-k moments of $\mathcal{D}_{\mathcal{X}}^{\text{test}}$ up to a high accuracy. REJECT if some of the moments are not close to the corresponding moments of $\mathcal{D}$.
- Otherwise, fit the best degree-k polynomial p on the training data, and output (ACCEPT, $\text{sign}(p)$).

This algorithm above runs in time $d^{O(k)}$, and we show that this algorithm is a valid TDS-learning algoithm for the wide class of functions whose **$\mathcal{L}_2$-sandwiching degree** is bounded by k, which we define as follows: For an approximation parameter ϵ, the $\mathcal{L}_2$-sandwiching degree of a function f is the smallest degree for a pair of polynomials p_{down} and p_{up} satisfying: i) $p_{\text{down}}(\mathbf{x}) \leq f(\mathbf{x}) \leq p_{\text{up}}(\mathbf{x})$ for all $\mathbf{x}$ in the learning domain and ii) $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(p_{\text{up}}(\mathbf{x}) - p_{\text{down}}(\mathbf{x}))^2] \leq \epsilon$.

The related notion of $\mathcal{L}_1$-sandwiching was recently used to obtain several results in testable learning [GKK23]. These results, however, do not seem to apply to TDS learning. Instead, we prove a "transfer lemma" showing that we can relate the test error under $\mathcal{D}_{\mathcal{X}}^{\text{test}}$ of a polynomial to its training error under $\mathcal{D}$ by leveraging the simple fact that the squared loss between two polynomials is itself a polynomial. As such, low-degree moment matching between the training and test marginals ensures that the squared loss between any pair of low-degree polynomials is approximately preserved (Lemma 2.8). Absolute loss cannot be computed by a low-degree polynomial, ruling out this type of transfer lemma based on $\mathcal{L}_1$-sandwiching.

Even though we need the more stringent property of small $\mathcal{L}_2$-sandwiching degree, we show that constructions from works in the pseudorandomness literature that explicitly construct $\mathcal{L}_1$-sandwiching polynomials (e.g., [DGJ+10] and [GOWZ10]) can be extended to bound the $\mathcal{L}_2$-sandwiching degree. This allows us to obtain efficient TDS learning algorithms for the classes of intersections of halfspaces, decision trees and small-depth formulas (see lines 3-6 in Table 1). We also note that this technique yields TDS learning algorithms not only in the realizable setting, but also in the agnostic setting.

Beyond Moment Matching. It is a natural question whether it is possible to beat the moment-matching approach. We answer this question in the affirmative by showing that for the class of halfspaces this is indeed possible. It is a standard fact that one needs polynomials of degree $\tilde{\Omega}(1/\epsilon^2)$ to ϵ-approximate halfspaces up to $\mathcal{L}_1$ error better than ϵ under the standard Gaussian distribution. Therefore the moment-matching approach requires a run-time of at least $d^{\tilde{\Omega}(1/\epsilon^2)}$ to TDS learn halfspaces under the standard Gaussian. We overcome this obstacle and give a TDS learning algorithm for halfspaces that runs in time $d^{O(\log(\frac{1}{\epsilon}))}$ (Line 2, Table 1).

One ingredient we use to design our algorithm is what we call TDS learning via the **disagreement region** method. Suppose we are able to recover the parameters of a halfspace f^* up to some accuracy β. Then, for some points $\mathbf{x}$ in $\mathbb{R}^d$ we will know $f^*(\mathbf{x})$ with certainty, but for some others we will not. We say that the latter points form the disagreement region, and it gets smaller as β decreases. The idea is to (i) use the training data to recover the parameters of halfspace f^* up to such high accuracy β that the probability that a Gaussian sample falls into the disagreement region is very small (ii) make sure that the recovered halfspace $\widehat{f}$ generalizes on the testing dataset by checking that only a small fraction of the testing dataset falls into the uncertainty region. We note that this notion of disagreement region is also widely used in active learning (see discussion in Section 2.2.1).

Although the disagreement region method gives an efficient algorithm for homogeneous (i.e. origin-centered) halfspaces (Proposition 5.1), it fails for general halfspaces. Indeed, in Section 2.2.2 we show that

for general halfspaces under the standard Gaussian distribution the disagreement region method requires $2^{\Omega(d)}$ samples. We design a $d^{O(\log(1/\epsilon))}$-time TDS learning algorithm for general halfspaces under the Gaussian distribution by **combining** the moment matching approach with the disagreement region approach:

- Suppose the halfspace f^* is not too biased, i.e. among $d^{O(\log(1/\epsilon))}$ training samples we see labels with values of both $+1$ and -1. We show that the parameters of such a halfspace can be recovered up to a very high accuracy using only $d^{O(\log(1/\epsilon))}$ additional training samples. This allows us to leverage the disagreement region method to achieve TDS learning.
- Otherwise, the halfspace f^* is highly biased and it almost always takes the same label L on a Gaussian input. For such halfspaces there is no hope to recover their parameters with $d^{O(\log(1/\epsilon)}$ samples. Yet, we show that using the moment-matching approach with degree parameter k of only $O(\log(1/\epsilon))$ allows us to certify that even under the test distribution $\mathcal{D}^{\text{test}}_{\mathcal{X}}$ the halfspace f^* will be biased and very likely to take the label L. Therefore, a predictor $\widehat{f}$ that assigns the label L to all points in $\mathbb{R}^d$ will generalize.

Techniques from Testable Learning. Additionally, in the setting of **agnostic** TDS learning we give an algorithm for the class of homogeneous (i.e. origin-centered) halfspaces under any isotropic log-concave distribution (see line 1 in Table 1). We achieve this using techniques from testable learning [GKK23, GKSV23a]. The first phase of our TDS learning algorithm uses an approximate agnostic learning algorithm for halfspaces [ABL17, DKTZ20] in order to obtain a vector $\widehat{\mathbf{v}}$, such that the homogeneous halfspace defined by $\widehat{\mathbf{v}}$ has error $O(\lambda) + \epsilon$ in the training dataset. Since the training distribution $\mathcal{D}$ is isotropic and log-concave, this means that the angle between $\widehat{\mathbf{v}}$ and the vector $\mathbf{v}$, defining the halfspace with optimal combined error on the training and testing datasets, is also at most $O(\lambda) + \epsilon$. Finally, we apply one of the core procedures from [GKK23, GKSV23a] in order to ensure that every halfspace defined by a vector $\mathbf{v}'$ that forms an angle of at most $O(\lambda) + \epsilon$ with $\widehat{\mathbf{v}}$ agrees on at least $1 - O(\lambda) - \epsilon$ fraction of the testing dataset with the halfspace defined by the vector $\widehat{\mathbf{v}}$. This allows us to certify that the halfspace defined by the vector $\widehat{\mathbf{v}}$ will indeed generalize to the testing distribution. Note that we can use tools from testable learning to remove the assumption on the training marginal; the algorithm would instead run a test that accepts when both $\mathcal{D}^{\text{train}}_{\mathcal{X}}$ and $\mathcal{D}^{\text{test}}_{\mathcal{X}}$ equal the target $\mathcal{D}$ without any assumptions on $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ and $\mathcal{D}^{\text{test}}_{\mathcal{XY}}$ (see also Remark 2.4). For clarity of exposition, we postpone formal statements composing the two models to future work.

1.3 Related Work

Domain Adaptation. The field of *domain adaptation* has received significant attention over the past two decades (see [BDBCP06, BCK+07, MMR09, BDBC+10, DLLP10, RMH+20] and references therein). Similar to our learning setting, domain adaptation considers scenarios where the learner has access to labeled training and unlabeled test examples and is asked to output a hypothesis with low test error without, however, being allowed to reject. [BDBCP06, BCK+07, MMR09] bound the test error of an empirical risk minimizer of training data by a sum of the parameter λ and some notion of distance between the training and test marginals (discrepancy or d_A distances) which is statistically efficient to compute using unlabeled test and training examples. This implies a statistically efficient TDS learning algorithm with error $2\lambda + \epsilon$ (Appendix A). All known algorithms for computing discrepancy distance or d_A distance, however, require exponential time even for basic classes such as halfspaces and decision trees. By allowing the learning algorithm to reject, we design computationally efficient TDS learning algorithms with error $O(\lambda) + \epsilon$ without explicitly computing the discrepancy distance.

PQ Learning. Among the learning models that capture settings with distribution shift, PQ learning (see [GKKM20] and [KK21]) is most relevant to TDS learning. In PQ learning, the learner has access to labeled

training data and unlabeled test data and must output a classifier h and a set $\mathbf{X}$. The classifier needs to minimize the following two criteria simultaneously: (1) the test error of the hypothesis h on test data points that fall into the region $\mathbf{X}$ (in other words, $\mathbf{X}$ is the region where one is confident in the predictions of the hypothesis h for test data) and (2) the probability that a training example falls into $\mathbf{X}$. [GKKM20] show that any concept class that can be agnostically learned in the distribution-free setting can be PQ learned. [KK21] improve this reduction by showing that PQ learning is equivalent to distribution-free reliable agnostic learning (see [KKM12]). The complexity of reliable learning is known to be "in between" agnostic learning and PAC learning. In particular, reliably learning conjunctions implies PAC learning DNF formulas. In Appendix B, we show that PQ learning actually implies TDS learning.

Testable Learning. Although conceptually our definition of TDS learning is inspired by the recent line of work in testable learning [RV23, GKK23, GKSV23a, GKSV23b], the two frameworks address very different issues. Testable learning does not address distribution shift, as it assumes that the training and testing distributions are the same distribution $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$. What the framework of testable learning does (indirectly) test is whether $\mathcal{D}^{\text{train}}_{\mathcal{X}}$ satisfies a certain assumption (e.g. Gaussianity) in order to make sure the learning algorithm gives a hypothesis $\widehat{f}$ that satisfies the agnostic learning guarantee.

As noted in [RV23], in the realizable setting one can trivially satisfy the definition of testable learning by drawing a fresh set of samples and using them to validate the hypothesis $\widehat{f}$. Due to this, existing work on testable learning [RV23, GKK23, GKSV23a, GKSV23b] focuses on the agnostic setting, where such validation procedure cannot be applied (see [RV23] for further detail). In contrast to this, even in the realizable setting, no such validation procedure exists for TDS learning, as indicated by our separations between PAC learning and TDS learning for monotone functions and convex sets (see Section 1.1). In fact, for monotone functions and convex sets, realizable TDS learning is harder than agnostic learning as well.

Furthermore, there are cases where realizable TDS learning is easier than agnostic learning (and, therefore, easier than testable agnostic learning). Here are two examples:

1. Due to statistical query lower bounds and cryptographic hardness results [GGK20, DKZ20, DKPZ21, DKR23], the run-time required to agnostically learn a halfspace under the standard Gaussian distribution is believed to be $d^{\Omega(1/\epsilon^2)}$. In contrast to this, in this work we show that realizable TDS learning of halfspaces with respect to the Gaussian distribution can be achieved using only $d^{O(\log 1/\epsilon))}$ run-time.

2. The agnostic learning of parity functions, even under the uniform distribution on $\{\pm 1\}^d$, is believed to require $2^{\Omega(\frac{d}{\text{poly}\log d})}$ time. In strong contrast with this, the class of parity functions can be TDS-learned in the realizable setting using only $\text{poly}(d/\epsilon)$ time under any distribution over $\{\pm 1\}^d$. This follows from the PQ-learning algorithm of [KK21], together with the connection between PQ learning and TDS learning (Appendix B).

Overall, we conclude that realizable TDS learning is incomparable to regular agnostic learning. In particular, there are examples where realizable TDS learning is easier than testable agnostic learning. Moreover, realizable TDS learning is harder than PAC learning, where distributional assumptions can be verified through validation.

Acknowledgements.

We thank Aravind Gollakota for insightful discussions during early stages of this project and for feedback on a draft of this manuscript. We also thank Varun Kanade for helpful discussions regarding PQ-learning.

2 Technical Overview

2.1 TDS Learning of Homogeneous Halfspaces

We provide an efficient TDS learner for the class of homogeneous halfspaces over $\mathbb{R}^d$ with respect to any given isotropic log-concave distribution that achieves error $O(\lambda)+\epsilon$, by applying results from prior work in the literature of testable learning (see [GKSV23a, GKSV23b]) and agnostic learning (see [Dan15, ABL17, DKTZ20]). We provide the following theorem and a proof sketch. The full proof can be found in Section 4.

Theorem 2.1 (Agnostic TDS learning of Halfspaces)**.** *Let $\mathcal{C}$ be the class of origin-centered halfspaces over $\mathbb{R}^d$ and $C>0$ a sufficiently large universal constant. Let $\mathcal{A},\mathcal{T}$ be as defined in Propositions 2.2 and 4.2. Let $m_{\mathcal{A}}$ be the sample complexity of $\mathcal{A}(\epsilon/C,\delta/4)$ and $m_{\mathcal{T}}=\frac{Cd^4}{\epsilon^2\delta}$. Then, there is an algorithm (Algorithm 1) that, given inputs $S_{\mathrm{train}}, X_{\mathrm{test}}$ of sizes $|S_{\mathrm{train}}|\geq m_{\mathcal{A}}$ and $|X_{\mathrm{test}}|\geq m_{\mathcal{T}}$ is a TDS learning algorithm for $\mathcal{C}$ w.r.t. any isotropic log-concave distribution $\mathcal{D}$ with error $O(\lambda)+\epsilon$ and run-time $\mathrm{poly}(d,\frac{1}{\epsilon})\log(\frac{1}{\delta})$, where is the accuracy parameter and δ is the failure probability.*

Leveraging training data. We first use an efficient agnostic learner on training data to recover a halfspace $\widehat{f}:\mathbf{x}\to\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x})$ with low training error. For example, we may use a (polynomial time) algorithm by [DKTZ20] (Proposition 4.2) that outputs $\widehat{f}$ with $\mathrm{err}(\widehat{f};\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}})\leq O(\eta)+\epsilon$ whenever the training marginal is isotropic log-concave (η is the optimal training error). There are other similar results in the literature of agnostic learning (e.g., see [ABL17]), but we use [DKTZ20] as it is more convenient for our setting.

Approximate parameter recovery. Let $\mathbf{v}^*$ be the parameter vector corresponding to the halfspace f^* that minimizes the common train and test error, i.e., $\mathrm{err}(f^*;\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}})+\mathrm{err}(f^*;\mathcal{D}^{\mathrm{test}}_{\mathcal{XY}})=\lambda$. Then, we have $\mathbb{P}_{\mathcal{D}^{\mathrm{train}}_{\mathcal{X}}}[\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x})\neq\mathrm{sign}(\mathbf{v}^*\cdot\mathbf{x})]\leq\mathrm{err}(\widehat{f};\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}})+\mathrm{err}(f^*;\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}})\leq O(\eta)+\epsilon+\lambda=O(\lambda)+\epsilon$. Since $\mathcal{D}^{\mathrm{train}}_{\mathcal{X}}=\mathcal{D}$ is isotropic log-concave, it is known that the disagreement over $\mathcal{D}^{\mathrm{train}}_{\mathcal{X}}$ between two halfspaces is proportional to the angular distance between their parameters, i.e., $\measuredangle(\widehat{\mathbf{v}},\mathbf{v}^*)=O(\mathbb{P}_{\mathcal{D}^{\mathrm{train}}_{\mathcal{X}}}[\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x})\neq\mathrm{sign}(\mathbf{v}^*\cdot\mathbf{x})])$, which we have bounded by $O(\lambda+\epsilon)$.

Testing phase. We have shown that $\widehat{\mathbf{v}}$ is geometrically close to $\mathbf{v}^*$, which achieves test error at most λ, by definition. It remains to certify that the test marginal behaves like an isotropic log-concave distribution with respect to $\widehat{\mathbf{v}}$, i.e., for a large enough set of i.i.d. examples X_{test} from $\mathcal{D}^{\mathrm{test}}_{\mathcal{X}}$ and for any $\mathbf{v}'\in\mathbb{S}^{d-1}$ we have that $\frac{1}{|X_{\mathrm{test}}|}\sum_{\mathbf{x}\in X_{\mathrm{test}}}\mathbb{1}\{\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x})\neq\mathrm{sign}(\mathbf{v}'\cdot\mathbf{x})\}:=\mathbb{P}_{X_{\mathrm{test}}}[\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x})\neq\mathrm{sign}(\mathbf{v}'\cdot\mathbf{x})]=O(\measuredangle(\widehat{\mathbf{v}},\mathbf{v}'))$, because then we will be able to bound the empirical test error of $\widehat{f}$ by $\lambda+O(\measuredangle(\widehat{\mathbf{v}},\mathbf{v}^*))$, which is $O(\lambda+\epsilon)$. The result then would follow by standard VC dimension arguments.

It turns out that recent work by [GKSV23b] on testable learning has provided an efficient tester that achieves exactly what we need. Note that the proof of the following proposition (Lemma 3.1 in [GKSV23b]) is nontrivial, requiring estimation of low-order moments and careful conditioning. We can apply this to our setting, because it only requires access to the marginal distribution.

Proposition 2.2 (Testably Bounding Halfspace Disagreement, Lemma 3.1 in [GKSV23b])**.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$, $\mathbf{v}_1\in\mathbb{S}^{d-1}$, $\theta\in(0,\pi/4]$, $\delta\in(0,1)$ and $C>0$ a sufficiently large universal constant. Then, there is an algorithm $\mathcal{T}(\theta,\delta)$ that, upon drawing at least $\frac{Cd^4}{\theta^2\delta}$ examples X from $\mathcal{D}$ and in time $\mathrm{poly}(d,\frac{1}{\theta},\frac{1}{\delta})$ either accepts or rejects and satisfies the following.*

(a) If $\mathcal{T}$ accepts, then for any $\mathbf{v}_2\in\mathbb{R}^d$ with $\measuredangle(\mathbf{v}_1,\mathbf{v}_2)\leq\theta$, it holds

$$\mathbb{P}_{\mathbf{x}\sim X}[\mathrm{sign}(\mathbf{v}_1\cdot\mathbf{x})\neq\mathrm{sign}(\mathbf{v}_2\cdot\mathbf{x})]\leq C\measuredangle(\mathbf{v}_1,\mathbf{v}_2)$$

(*b*) *If $\mathcal{D}$ is isotropic log-concave, then $\mathcal{T}$ accepts with probability at least $1-\delta$.*

Remark 2.3. We note that, in fact, the original version of Proposition 2.2 in [GKSV23b] does not require the target marginal to be known, but works universally for any isotropic log-concave distribution (as well as distributions with heavier tails). This implies that the completeness criterion that Algorithm 1 satisfies is actually much stronger: for an appropriate choice of the absolute constant C, Algorithm 1 can be made to accept whenever $\mathcal{D}_{\mathcal{X}}^{\mathrm{test}}$ is isotropic log-concave (and not necessarily equal to the training marginal).

Remark 2.4. Moreover, we point out that we can apply results from [GKSV23b] and substitute algorithm $\mathcal{A}$ with a universal tester-learner for halfspaces. This enables us to remove the assumption that $\mathcal{D}_{\mathcal{X}}^{\mathrm{train}}$ is some fixed isotropic log-concave distribution, and the final algorithm would accept with high probability whenever $\mathcal{D}_{\mathcal{X}}^{\mathrm{train}}$ is isotropic strongly log-concave *and* $\mathcal{D}_{\mathcal{X}}^{\mathrm{test}}$ is isotropic log-concave. In that sense, TDS learning composes well with (universally) testable learning. For sake of presentation, however, we leave formal compositional arguments to future work.

2.2 TDS Learners for General Halfspaces

2.2.1 Warm-Up: Disagreement-Based TDS Learning

We provide a general TDS learner for the realizable setting, based on the notion of disagreement regions from active learning. Not only is this approach interesting in and of itself, but it will also be useful in Section 2.2.2 where we present our main result for TDS learning of general halfspaces in the realizable setting. The main idea is to testably bound the probability that a test example falls in some region $\mathbf{D}$, whose mass with respect to the target distribution becomes smaller as the number of training samples increases and, also, the output of the training algorithm achieves low error on any distribution that assigns small mass to $\mathbf{D}$. It turns out that the quantity $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\in\mathbf{D}]$, where $\mathcal{D}$ is some given distribution over a space $\mathcal{X}\subseteq\mathbb{R}^d$, is a well-studied notion in the literature of active learning (see [CAL94, Han09, BBL06, Han11, Han14, BHV10, Han07] and references therein). We now provide a formal definition for the disagreement region.

Definition 2.5 (Disagreement Region)**.** Let $\mathcal{X}\subseteq\mathbb{R}^d$, $\mathcal{D}$ a distribution over $\mathcal{X}$ and $\mathcal{C}$ a concept class of functions that map $\mathcal{X}$ to $\{\pm 1\}$. For $\epsilon>0$ and $f\in\mathcal{C}$, we define the ϵ-disagreement region of f under $\mathcal{D}$, $\mathbf{D}_{\epsilon}(f;\mathcal{D})$ as the subset of $\mathcal{X}$ such that if $\mathbf{x}\in\mathbf{D}_{\epsilon}(f;\mathcal{D})$, then there are $f_1,f_2\in\mathcal{C}$ with $\mathrm{err}(f_1,f;\mathcal{D})\leq\epsilon$, and $\mathrm{err}(f_2,f;\mathcal{D})\leq\epsilon$ and $f_1(\mathbf{x})\neq f_2(\mathbf{x})$.

In the literature of active learning, the quantity of interest is called the disagreement coefficient and is defined for a concept class $\mathcal{C}$ and a distribution $\mathcal{D}$ as follows (see, e.g., [Han14]).

$$\theta(\epsilon)=\sup_{f\in\mathcal{C}}\sup_{\epsilon'>\epsilon}\frac{\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\in\mathbf{D}_{\epsilon'}(f;\mathcal{D})]}{\epsilon'} \tag{2.1}$$

In particular, for active learning, is is crucial that $\theta(\epsilon)$ is asymptotically bounded by a slowly increasing function of $1/\epsilon$ (e.g., $O(\log(1/\epsilon))$), since bounds on the disagreement coefficient directly provide rates on the label complexity of disagreement-based active learning, up to logarithmic factors [Han11]. In our setting, meaningful results are obtained even when $\theta(\epsilon)=O(1/\epsilon^{1-c})$ for any constant $c\in(0,1)$. Moreover, we also focus on the dependence of the disagreement coefficient on other relevant parameters, like the dimension d. To emphasize this, in what follows, we will use the notation $\theta(\epsilon,d)$ to refer to the disagreement coefficient. We obtain the following result, which implies, for example, a polynomial improvement in the sample complexity bound of realizable TDS learning of homogeneous halfspaces w.r.t. the Gaussian compared to the TDS learner we proposed in Theorem 2.1 for the agnostic setting (see also Section 5.1).

Theorem 2.6 (Disagreement-Based TDS learning)**.** *Let $\mathcal{C}$ be the class of concepts that map $\mathcal{X}\subseteq\mathbb{R}^d$ to $\{\pm1\}$ with VC dimension $\mathrm{VC}(\mathcal{C})$, let $\mathcal{D}$ a distribution over $\mathcal{X}$ and $C>0$ a sufficiently large universal*

constant. Suppose that we have access to an ERM oracle for PAC learning $\mathcal{C}$ under $\mathcal{D}$ and membership access to $\mathbf{D}_{\epsilon'}(f;\mathcal{D})$ for any given $f \in \mathcal{C}$ and $\epsilon' > 0$. Then, there is an algorithm (Algorithm 3) that given inputs of sizes $|S_{\mathrm{train}}| \geq C\frac{\mathrm{VC}(\mathcal{C})}{\epsilon'}\log(\frac{1}{\epsilon'\delta})$ and $|X_{\mathrm{test}}| \geq C(\frac{\mathrm{VC}(\mathcal{C})}{\epsilon}+\frac{1}{\epsilon^2})\log(\frac{1}{\epsilon\delta})$ is a TDS learning algorithm for $\mathcal{C}$ w.r.t. $\mathcal{D}$ that calls the ϵ'-ERM oracle once and the ϵ'-membership oracle $|S_{\mathrm{train}}|$ times, where ϵ is the accuracy parameter, δ is the failure probability and ϵ' such that $\epsilon' \cdot \theta(\epsilon', d) \leq \epsilon/2$.

2.2.2 Beyond Disagreement: TDS Learners for General Halfspaces

We give a TDS-learning algorithm for the class of halfspaces under the standard Gaussian distribution. The algorithm runs in quasi-polynomial time in all relevant parameters and, contrary to the case of homogeneous halfspaces, works in a setting where efficient parameter recovery is not possible. This happens because when a general halfspace has arbitrarily large bias, it is possible, for example, that all of the training examples have the same label.

In particular, applying a pure disagreement-based TDS learning framework (Theorem 2.6) in the case of *general halfspaces* can only give exponential-time algorithms for this problem. To illustrate this, imagine that the ground truth is a general halfspace with bias $\tau = \sqrt{d}$ but unknown direction $\mathbf{v} \in \mathbb{S}^{d-1}$. Then, any general halfspace $\mathbf{x} \mapsto \mathrm{sign}(\mathbf{v}' \cdot \mathbf{x} - \tau)$ with the same bias is $\exp(-\Omega(d))$-close to the ground truth with respect to the Gaussian distribution, due to standard Gaussian concentration, i.e., $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\mathrm{sign}(\mathbf{v}\cdot\mathbf{x} - \tau) \neq \mathrm{sign}(\mathbf{v}'\cdot\mathbf{x}-\tau)] \leq \mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}-\tau) \neq \mathrm{sign}(-\mathbf{v}\cdot\mathbf{x}-\tau)]$, which is upper bounded by $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[|\mathbf{v}\cdot\mathbf{x}| > \sqrt{d}] \leq 2\exp(-d/2)$. Let $\epsilon' = 2\exp(-d/2)$. Suppose that ERM returns a halfspace $\widehat{f}$ that is ϵ'-close to the ground truth but has bias τ. Any $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\|_2 \geq \sqrt{d}$, falls within the disagreement region $\mathbf{D}_{\epsilon'}(\widehat{f};\mathcal{N}(0,I_d))$ and therefore $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\mathbf{x} \in \mathbf{D}_{\epsilon'}(\widehat{f};\mathcal{N}(0,I_d))]$ is constant. This implies that running the ERM oracle on training data even up to exponentially small accuracy $\epsilon' = \exp(-\Omega(d))$ does not meet the requirement of Theorem 2.6 (see also [EYW12]) that the disagreement coefficient is bounded as $\epsilon' \cdot \theta(\epsilon', d) \leq \epsilon/2$.

In order to overcome this obstacle, we perform a case analysis that depends on the bias of the unknown halfspace. If the bias is bounded, then we may use a disagreement-based approach, since we can approximately recover the true parameters of the unknown halfspace using training data and it suffices to verify that the test distribution does not amplify the error between any pair of halfspaces close to the obtained approximations of the true parameters. Now, consider the case when the bias is large. We may assume without loss of generality the constant hypothesis $+1$ has low training error (since the ground truth has large bias and the marginal is Gaussian). If we can certify that the test marginal is sufficiently concentrated in every direction, then this hypothesis must also have small test error. To certify concentration for the test distribution's marginals, we use a moment-matching approach. Checking the moment matching condition only up to degree $O(\log(\epsilon))$ turns out to be sufficient to certify the type of concentration we need. We thus obtain a quasi-polynomial TDS learning algorithm for general halfspaces with respect to the Gaussian distribution. Since the probability of success can be amplified through repetition (see Proposition C.1), we provide a result with constant failure probability. For the full proof, see Section 5.2.

Theorem 2.7 (TDS learning of General Halfspaces)**.** *Let $\mathcal{C}$ be the class of general halfspaces over $\mathbb{R}^d$ and $C > 0$ a sufficiently large universal constant. Then, there is an algorithm (Algorithm 4) that, given inputs of size $|S_{\mathrm{train}}| = |X_{\mathrm{test}}| = Cd^{C\log 1/\epsilon}$ is a TDS learning algorithm for $\mathcal{C}$ w.r.t. $\mathcal{N}(0, I_d)$ with run-time $d^{O(\log 1/\epsilon)}$, where ϵ is the accuracy parameter, and the failure probability δ is at most* 0.01.

Compared to Theorem 2.6, our approach here incurs an increase in the amount of test samples required (from $\mathrm{poly}(d, 1/\epsilon)$ to $d^{O(\log(1/\epsilon))}$, used for moment matching) but significantly decreases the amount of training samples required (from $\exp(\Omega(d))$ to $d^{O(\log(1/\epsilon))}$).

2.3 TDS Learning through Moment Matching

In the previous section, we provided a TDS learner for general halfspaces in the realizable setting that requires ideas beyond parameter recovery and testably bounding the probability of falling in the disagreement region. Crucially, Theorem 2.7 uses a moment-matching approach in the case when the bias of the unknown halfspaces is large. As is explained in this section, we show that the moment-matching approach can actually provide a generic result which demonstrates that $\mathcal{L}_2$-sandwiching (see Definition 3.1) implies TDS learning, even in the non-realizable setting. We also instantiate our framework to several important concept classes (halfspace intersections, decision trees and Boolean formulas) with respect to the Gaussian and uniform distributions, by applying constructions from pseudorandomness literature to bound the $\mathcal{L}_2$-sandwiching degree of each of these classes and acquire entries 3-6 in Table 1.

We provide a general theorem, which demonstrates that $\mathcal{L}_2$-sandwiching implies TDS learning under some additional natural assumptions about the target marginal distribution, which are satisfied by the standard Gaussian distribution over $\mathbb{R}^d$ and the uniform distribution on $\{\pm 1\}^d$. While it is known that $\mathcal{L}_1$-sandwiching implies testable learning (see [GKK23]), we require the stronger notion of $\mathcal{L}_2$-sandwiching. In particular, while $\mathcal{L}_1$-sandwiching would (testably) imply the existence of low degree polynomials with low test error, we do not get to see labeled examples from $\mathcal{D}^{\text{test}}_{\mathcal{XY}}$. Moreover, we cannot a priori assume that the output of the training algorithm is a sandwiching polynomial, even if we know one exists.

In our analysis, we crucially use the fact that the square of the difference between two polynomials is itself a polynomial whose coefficients and degree are bounded by the degree and coefficient bounds of the original polynomials. Crucially, this enables us to use the following transfer lemma which relates the squared distance between polynomials under the test distribution to their squared distance under the training distribution. In what follows, we use the notation $\mathbf{x}^\alpha = \prod_{i\in[d]} \mathbf{x}_i^{\alpha_i}$, where $\alpha \in \mathbb{N}^d$.

Lemma 2.8 (Informal, Transfer Lemma for Square Loss, see Lemma 6.2)**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \subseteq \mathbb{R}^d$ and X_{test} a (multi)set of points in $\mathbb{R}^d$. If $\mathbb{E}_{\mathbf{x}\sim X_{\text{test}}}[\mathbf{x}^\alpha] \approx \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}^\alpha]$ for all $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \le 2k$, then for any degree k polynomials p_1, p_2 with bounded coefficients, it holds*

$$\frac{1}{|X_{\text{test}}|}\sum_{\mathbf{x}\in X_{\text{test}}}(p_1(\mathbf{x})-p_2(\mathbf{x}))^2 \approx \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[(p_1(\mathbf{x})-p_2(\mathbf{x}))^2]$$

Moreover, we use the fact that, due to the $\mathcal{L}_2$-sandwiching assumption, we can bound quantities of the form $\mathbb{E}[(p(\mathbf{x})-f(\mathbf{x}))^2]$ for $f \in \mathcal{C}$ from above by $O(\mathbb{E}[(p(\mathbf{x})-p_{\text{down}}(\mathbf{x}))^2] + \mathbb{E}[(p_{\text{down}}(\mathbf{x})-p_{\text{up}}(\mathbf{x}))^2])$, irrespective of the distribution that the expectations are taken over. Over the training distribution, the quantity $\mathbb{E}_{\mathcal{D}}[(p_{\text{down}}(\mathbf{x})-p_{\text{up}}(\mathbf{x}))^2]$ is small via the definition of $\mathcal{L}_2$-sandwiching degree, and the quantity $\mathbb{E}_{\mathcal{D}}[(p(\mathbf{x})-f(\mathbf{x}))^2]$ because p is obtained from $\mathcal{L}_2$ polynomial regression. If $p, p_{\text{down}}, p_{\text{up}}$ are all low degree and the dataset X_{test} matches low-degree moments with $\mathcal{D}$, then we may apply Lemma 2.8 to bound $\frac{1}{|X_{\text{test}}|}\sum_{\mathbf{x}\in X_{\text{test}}}[(p(\mathbf{x})-f(\mathbf{x}))^2]$. Once it is shown that p fits f well on the testing dataset X_{test}, standard generalization bounds allows us to conclude that it will also predict f well on the testing distribution. Therefore, by running polynomial regression on training data to obtain p and testing whether the empirical test moments match the moments of the training distribution, we acquire the following result, whose proof can be found in Section 6.

Theorem 2.9 ($\mathcal{L}_2$-sandwiching implies TDS Learning)**.** *Let $\mathcal{D}$ be a distribution over a set $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ be a concept class. Let $\epsilon, \delta \in (0,1)$, $\epsilon' = \epsilon/100$ $\delta' = \delta/2$ and assume that the following are true.*

(i) ($\mathcal{L}_2$-Sandwiching) The ϵ'-approximate $\mathcal{L}_2$ sandwiching degree of $\mathcal{C}$ under $\mathcal{D}$ is at most k with coefficient bound B.

(*ii*) *(Moment Concentration) If* $X \sim \mathcal{D}^{\otimes m}$ *and* $m \geq m_{\mathrm{conc}}$ *then, with probability at least* $1 - \delta'$*, we have that for any* $\alpha \in \mathbb{N}^d$ *with* $\|\alpha\|_1 \leq k$ *it holds* $|\mathbb{E}_{\mathcal{D}}[\mathbf{x}^\alpha] - \frac{1}{|X|}\sum_{\mathbf{x}\in X}\mathbf{x}^\alpha| \leq \frac{\epsilon'}{B^2 d^{4k}}$.

(*iii*) *(Generalization) If* $S \sim \mathcal{D}_{\mathcal{XY}}^{\otimes m}$ *where* $\mathcal{D}_{\mathcal{XY}}$ *is any distribution over* $\mathcal{X} \times \{\pm 1\}$ *such that* $\mathcal{D}_{\mathcal{X}} = \mathcal{D}$ *and* $m \geq m_{\mathrm{gen}}$ *then, with probability at least* $1 - \delta'$ *we have that for any degree-*k *polynomial* p *with coefficient bound* B *it holds* $|\mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y - p(\mathbf{x}))^2] - \frac{1}{|S|}\sum_{(\mathbf{x},y)\in S}[(y - p(\mathbf{x}))^2]| \leq \epsilon'$.

Then, there is an algorithm (Algorithm 5) that, upon receiving $m_{\mathrm{train}} \geq m_{\mathrm{gen}}$ *labelled samples* S_{train} *from the training distribution and* $m_{\mathrm{test}} \geq C \cdot \frac{d^k + \log(1/\delta)}{\epsilon^2} + m_{\mathrm{conc}}$ *unlabelled samples* X_{test} *from the test distribution (where* $C > 0$ *is a sufficiently large universal constant), runs in time* $\mathrm{poly}(|S_{\mathrm{train}}|, |X_{\mathrm{test}}|, d^k)$ *and TDS learns* $\mathcal{C}$ *with respect to* $\mathcal{D}$ *up to error* $32\lambda + \epsilon$ *and with failure probability* δ.

2.4 Lower Bounds

We provide three lower bounds for TDS learning. The first one shows that TDS learning the class of monotone functions over $\{\pm 1\}^d$ with respect to the uniform distribution requires an exponential number of examples from either the training or the test distribution, which implies a separation with regular agnostic learning. The second lower bound shows that TDS learning the class of indicators of convex sets also requires an exponential in the dimension number of samples. The third lower bound demonstrates that a linear dependence on the error term λ (as defined in Equation (3.1)) is necessary for TDS learning in the non-realizable setting.

2.4.1 Lower Bound for Monotone Functions and Convex Sets in Realizable Setting

Recent work on testable learning (which is a generalization of the classical agnostic learning framework, see [RV23, GKK23]) has demonstrated that the class of monotone functions over $\{\pm 1\}^d$ cannot be testably learned with respect to the uniform distribution unless the learner draws at least $2^{\Omega(d)}$ training samples. Since the class of monotone functions can be agnostically learned in time $2^{\tilde{O}(\sqrt{d})}$ with respect to the uniform distribution over the hypercube $\{\pm 1\}^d$, this implies that testable (agnostic) learning is strictly harder than regular agnostic learning. We show that the lower bound of $2^{\Omega(d)}$ extends to the problem of TDS learning monotone functions even in the realizable setting. Recall that we have shown that we can TDS learn halfspaces with respect to the standard Gaussian distribution in the realizable setting in time $d^{O(\log(1/\epsilon))}$ (Theorem 2.7) but it is known that, for agnostic learning, any SQ algrorithm for the problem requires time $d^{\Omega(1/\epsilon^2)}$ (see [GGK20, DKZ20, DKPZ21]). Therefore, we conclude that realizable TDS learning and agnostic learning are incomparable. We now provide our lower bound. For the proof, see Section 7.

Theorem 2.10 (Hardness of TDS Learning Monotone Functions)**.** *Let the accuracy parameter* ϵ *be at most* 0.1 *and the success probability parameter* δ *also be at most* 0.1*. Then, in the realizable setting, any TDS learning algorithm for the class of monotone functions over* $\{\pm 1\}^d$ *with accuracy parameter requires either* $2^{0.04d}$ *training samples or* $2^{0.04d}$ *testing samples for all sufficiently large values of* d*.*

We now provide a lower bound for convex sets (see also Section 7). Since the class of indicators of convex sets can be agnostically learned in time $2^{\tilde{O}(\sqrt{d})}$ with respect to the Standard Gaussian on $\mathbb{R}^d$, the following theorem implies yet another separation between agnostic learning and realizable TDS learning in the distribution specific setting under the Gaussian distribution for a well-studied concept class.

Theorem 2.11 (Hardness of TDS Learning Convex Sets)**.** *Let the accuracy parameter* ϵ *be at most* 0.1 *and the success probability parameter* δ *also be at most* 0.1*. Then, in the realizable setting, any TDS learning algorithm for the class of indicators of convex sets under the standard Gaussian distribution on* $\mathbb{R}^d$ *requires either* $2^{0.04d}$ *training samples or* $2^{0.04d}$ *testing samples for all sufficiently large values of* d*.*

Remark 2.12. In Proposition B.3 of the Appendix, we show that TDS learning is not harder than PQ learning (which is a related learning primitive, see [GKKM20, KK21]). [KK21] show that the class of parities over $\{\pm 1\}^d$ can be efficiently PQ learned, which provides another example where TDS learning is easier than agnostic learning.

2.4.2 Lower Bound for the Error Guarantee in the Agnostic Setting

We now focus on the agnostic setting and provide an information theoretic lower bound on the error upon acceptance. Our lower bound is simple and demonstrates that a linear dependence on the error factor λ (see Equation (3.1)) is unavoidable for TDS learning.

Theorem 2.13 (Lower Bound for the Error in the Agnostic Setting)**.** *Let $\mathcal{X}$ be any domain, $\mathcal{D}$ a distribution over $\mathcal{X}$ and $\mathcal{C}$ a class of concepts that map $\mathcal{X}$ to $\{\pm 1\}$ that is closed under complement, i.e., if $f \in \mathcal{C}$ then $-f \in \mathcal{C}$. Then, for any $\eta \in (0, 1/2)$, any $\epsilon \in (0, \eta/2)$ and $\delta \in (0, 1/3)$, no TDS learning algorithm for $\mathcal{C}$ w.r.t. $\mathcal{D}$ with finite sample complexity and failure probability δ, can have an error guarantee better than $\lambda(1-2\eta) + \epsilon = \Omega(\lambda) + \epsilon$.*

Proof. Let $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ denote the training distribution and $\mathcal{D}^{\text{test}}_{\mathcal{XY}}$ the test distribution, which are both over $\mathcal{X} \times \{\pm 1\}$. Suppose that for $\eta \in (0, 1/2)$ and $\epsilon \in (0, \eta/2)$ there exists an algorithm $\mathcal{A}$, that, upon acceptance and with probability at least $1 - \delta$, outputs $\widehat{f} \in \mathcal{C}$ with $\text{err}(\widehat{f}; \mathcal{D}^{\text{test}}_{\mathcal{XY}}) \leq \lambda(1-2\eta) + \epsilon$ ($\lambda = \lambda(\mathcal{C}; \mathcal{D}^{\text{train}}_{\mathcal{XY}}, \mathcal{D}^{\text{test}}_{\mathcal{XY}})$, see Equation (3.1)). Let $C > 0$ be a sufficiently large universal constant.

We consider the following algorithm $\mathcal{T}$. Algorithm $\mathcal{T}$ uses an oracle to $\mathcal{A}$ and accepts or rejects according to the following criteria.

- If $\mathcal{A}$ rejects, then $\mathcal{T}$ rejects.
- If $\mathcal{A}$ accepts and outputs $\widehat{f} \in \mathcal{C}$, then $\mathcal{T}$ draws $\frac{C}{\eta^2}\log(1/\delta)$ examples $S_{\mathcal{T}}$ from $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ and rejects if $\mathbb{P}_{(\mathbf{x},y)\in S_{\mathcal{T}}}[\widehat{f}(\mathbf{x}) \neq y] > 3\eta/4$. Otherwise, $\mathcal{T}$ accepts.

Fix some $f \in \mathcal{C}$ and let $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ be the distribution over $\mathcal{X} \times \{\pm 1\}$ whose marginal on $\mathcal{X}$ is $\mathcal{D}$ and the labels are generated as $y(\mathbf{x}) = f(\mathbf{x})$. Consider the following two cases about $\mathcal{D}^{\text{test}}_{\mathcal{XY}}$.

Case 1. First, suppose that $\mathcal{D}^{\text{test}}_{\mathcal{XY}}$ has $\mathcal{D}$ as marginal on $\mathcal{X}$ and $y(\mathbf{x}) = f(\mathbf{x})$. Then, $\mathcal{A}$ accepts with probability at least $1-\delta$, due to completeness. We have $\lambda = 0$ (attained by f) and, hence, upon acceptance, $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[\widehat{f}(\mathbf{x}) \neq y] = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{test}}_{\mathcal{XY}}}[\widehat{f}(\mathbf{x}) \neq y] \leq \epsilon \leq \eta/2$ with probability at least $1-\delta$. By a Hoeffding bound, we then have that $\mathcal{T}$ must accept with probability at least $1-\delta$. Overall, $\mathcal{T}$ accepts with probability at least $1 - 3\delta > 1/2$.

Case 2. Second, suppose that $\mathcal{D}^{\text{test}}_{\mathcal{XY}}$ has $\mathcal{D}$ as marginal on $\mathcal{X}$ and $y(\mathbf{x}) = -f(\mathbf{x})$. Then, we have that $\lambda = 1$ (because for any point $\mathbf{x} \in \mathcal{X}$, any classifier will either classify $\mathbf{x}$ incorrectly under $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ or under $\mathcal{D}^{\text{test}}_{\mathcal{XY}}$). By assumption, we have $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{test}}_{\mathcal{XY}}}[\widehat{f}(\mathbf{x}) \neq y] \leq \lambda(1-2\eta) + \epsilon \leq 1 - 2\eta + \epsilon$ with probability at least $1 - 2\delta$ (by completeness and soundness). Since the test labels are the negation of the train labels, we have $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{test}}_{\mathcal{XY}}}[\widehat{f}(\mathbf{x}) \neq y] = 1 - \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[\widehat{f}(\mathbf{x}) \neq y]$, and $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[\widehat{f}(\mathbf{x}) \neq y] \geq 2\eta - \epsilon \geq \eta$ (since $\epsilon \leq \eta/2$). By a Hoeffding bound, $\mathcal{T}$ will reject with probability at least $1 - 3\delta > 1/2$.

We have reached a contradiction, because in both cases, the input of $\mathcal{T}$ does not depend on the test labels, and everything else remains the same in both cases. Therefore, $\mathcal{T}$ should have the same behavior in both cases and we conclude that the algorithm $\mathcal{A}$ cannot exist as defined. □

Remark 2.14. While the above lower bound demonstrates that the error of a TDS learning algorithm can be necessarily high in certain settings, we emphasize that the construction corresponds to a contrived case where the training distribution does not provide enough information about the test distribution and, therefore, any meaningful notion of learning should be hopeless (see also [BDU12]).

3 Notation and Basic Definitions

We let $\mathcal{X} \subseteq \mathbb{R}^d$ and, in particular, $\mathcal{X}$ will either be the d-dimensional hypercube $\{\pm 1\}^d$ or the d-dimensional Euclidean space $\mathbb{R}^d$. For a distribution $\mathcal{D}$ over $\mathcal{X}$, 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). We let $\mathbb{R}_+ = (0, \infty)$.

For a function $p : \mathcal{X} \to \mathbb{R}$ and $r \in \mathbb{N}$, we define the $\mathcal{L}_r$ norm of p under $\mathcal{D}$ as $\|p\|_{\mathcal{L}_r(\mathcal{D})} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[p(\mathbf{x})^r]^{\frac{1}{r}}$. For $\mathbf{x} \in \mathcal{X}$ where $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d)$ and for $\alpha \in \mathbb{N}^d$, we denote with $\mathbf{x}^\alpha$ the product $\prod_{i\in[d]} \mathbf{x}_i^{\alpha_i}$, $\mathrm{M}_\alpha = \mathbb{E}[\mathbf{x}^\alpha]$ and $\|\alpha\|_1 = \sum_{i\in[d]} \alpha_i$. For a polynomial p over $\mathbb{R}^d$ and $\alpha \in \mathbb{N}^d$, we denote with p_α the coefficient of p corresponding to $\mathbf{x}^\alpha$, i.e., we have $p(\mathbf{x}) = \sum_{\alpha\in\mathbb{N}^d} p_\alpha \mathbf{x}^\alpha$. If p is a polynomial over $\{\pm 1\}^d$, then we can always express it in a unique multilinear form, so we will only use coefficients p_α with $\alpha \in \{0, 1\}^d$, i.e., $p(\mathbf{x}) = \sum_{\alpha\in\{0,1\}^d} p_\alpha \mathbf{x}^\alpha$. We define the degree of p and denote $\deg(p)$ the maximum degree of a monomial whose coefficient in p is non-zero, i.e., $\deg(p) = \max\{\|\alpha\|_1 : p_\alpha \neq 0\}$.

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 $\|\mathbf{v}_1\|_2\|\mathbf{v}_2\|_2 \cos(\theta) = \mathbf{v}_1 \cdot \mathbf{v}_2$. For $\mathbf{v} \in \mathbb{R}^d, \tau \in \mathbb{R}$, we call a function of the form $\mathbf{x} \mapsto \mathrm{sign}(\mathbf{v} \cdot \mathbf{x})$ an origin-centered (or homogeneous) halfspace and a function of the form $\mathbf{x} \mapsto \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} - \tau)$ a general halfspace over $\mathbb{R}^d$.

$\mathcal{L}_2$-sandwiching degree. We now define the notion of $\mathcal{L}_2$-sandwiching polynomials for a function f with respect to a distribution $\mathcal{D}$, i.e., a pair of polynomials such that one of them is pointwise above f, the other one is pointwise below f and the $\mathcal{L}_2$ distance between the two polynomials with respect to $\mathcal{D}$ is small. While the notion of L_1 sandwiching polynomials is standard in the literature of pseudorandomness (see, e.g., [Baz09]) and has applications to testable learning (see [GKSV23a]), in order to obtain our main results, we make use of the stronger notion of $\mathcal{L}_2$-sandwiching polynomials which we define below.

Definition 3.1 ($\mathcal{L}_2$-sandwiching polynomials)**.** Consider a product set $\mathcal{X}$ and a distribution $\mathcal{D}$ over $\mathcal{X}$. For $\epsilon > 0$ and $f : \mathcal{X} \to \{\pm 1\}$, we say that the polynomials $p_{\mathrm{up}}, p_{\mathrm{down}} : \mathcal{X} \to \mathbb{R}$ are ϵ-approximate $\mathcal{L}_2$-sandwiching polynomials for f under $\mathcal{D}$ if the following are true.

1. $p_{\mathrm{down}}(\mathbf{x}) \leq f(\mathbf{x}) \leq p_{\mathrm{up}}(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{X}$.
2. $\|p_{\mathrm{up}} - p_{\mathrm{down}}\|^2_{\mathcal{L}_2(\mathcal{D})} \leq \epsilon$

Moreover, for $\epsilon > 0$, a concept class $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ and $k, B > 0$, we say that the ϵ-approximate $\mathcal{L}_2$-sandwiching degree of $\mathcal{C}$ under $\mathcal{D}$ is at most k and with coefficient bound B if for any $f \in \mathcal{C}$ there are ϵ-approximate $\mathcal{L}_2$-sandwiching polynomials $p_{\mathrm{up}}, p_{\mathrm{down}}$ for f such that $\deg(p_{\mathrm{up}}), \deg(p_{\mathrm{down}}) \leq k$ and each of the coefficients of $p_{\mathrm{up}}, p_{\mathrm{down}}$ are absolutely bounded by B.

Learning Setup. Consider $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}, \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}$ to be distributions over $\mathcal{X} \times \{\pm 1\}$ and let $\mathcal{D}^{\mathrm{train}}_{\mathcal{X}}, \mathcal{D}^{\mathrm{test}}_{\mathcal{X}}$ be the corresponding marginal distributions on $\mathcal{X} \subseteq \mathbb{R}^d$. Our tester-learners receive labelled examples from $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$ and unlabelled examples from $\mathcal{D}^{\mathrm{test}}_{\mathcal{X}}$ and their goal is to produce a hypothesis with low error on $\mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}$ or

potentially reject but only if distribution shift is detected. Given a hypothesis class $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$, $h_1, h_2 : \mathcal{X} \to \{\pm 1\}$ and distributions $\mathcal{D}_{\mathcal{XY}}, \mathcal{D}^{\text{train}}_{\mathcal{XY}}, \mathcal{D}^{\text{test}}_{\mathcal{XY}}$ over $\mathcal{X} \times \{\pm 1\}$, we define $\text{err}(h_1; \mathcal{D}_{\mathcal{XY}}) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathcal{XY}}}[y \neq h_1(\mathbf{x})]$ and $\text{err}(h_1, h_2; \mathcal{D}_{\mathcal{X}}) = \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$ as well as the following quantity, which is standard in the domain adaptation literature (see, e.g., [BDBCP06, BCK+07, BDBC+10, DLLP10]).

$$\lambda(\mathcal{C}; \mathcal{D}^{\text{train}}_{\mathcal{XY}}, \mathcal{D}^{\text{test}}_{\mathcal{XY}}) := \min_{f\in\mathcal{C}}\{\text{err}(f; \mathcal{D}^{\text{train}}_{\mathcal{XY}}) + \text{err}(f; \mathcal{D}^{\text{test}}_{\mathcal{XY}})\}, \text{ attained by } f^* \in \mathcal{C} \tag{3.1}$$

Observe that parameter λ becomes small whenever the training and test errors can be simultaneously minimized by a common classifier in $\mathcal{C}$. Clearly, if there is no relationship between the training and test distributions, then using data from the training distribution does not reveal any information about the test distribution and, therefore, learning is hopeless (see also Theorem 2.13). We will assume (as is common in the domain adaptation literature) that the parameter λ is a valid choice for quantifying the relationship between the training and test distributions, in the sense that considering λ to be small is not unrealistic. In particular, we will partly focus on the following setting where λ is zero. To distinguish between the two settings, we say that we are in the **agnostic setting** when $\lambda \geq 0$ (arbitrary) and in the **realizable setting** when $\lambda = 0$. When $\lambda = 0$, there exists a classifier in $\mathcal{C}$ that achieves both zero training loss and test loss and we therefore refer to this setting as realizable. Another (slightly more specific) way to view the realizable setting is by considering the labelled distribution $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ (resp. $\mathcal{D}^{\text{test}}_{\mathcal{XY}}$) formed as follows: for some $f^* \in \mathcal{C}$, draw an example $\mathbf{x}$ from $\mathcal{D}^{\text{train}}_{\mathcal{X}}$ (resp. $\mathcal{D}^{\text{test}}_{\mathcal{X}}$) and form the pair $(\mathbf{x}, y) \sim \mathcal{D}^{\text{train}}_{\mathcal{XY}}$ (resp. $(\mathbf{x}, y) \sim \mathcal{D}^{\text{test}}_{\mathcal{XY}}$) by setting $y = f^*(\mathbf{x})$. We now provide a formal definition of our learning model.

Definition 3.2 (Testable Learning with Distribution Shift (TDS Learning))**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ and consider a distribution $\mathcal{D}$ over $\mathcal{X}$ and a concept class $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$. For some $\psi : [0, 1] \to [0, 1]$ and $\epsilon, \delta \in (0, 1)$, we say that an algorithm $\mathcal{A}$ testably learns $\mathcal{C}$ with distribution shift w.r.t. $\mathcal{D}$ up to error $\psi(\lambda) + \epsilon$ and probability of failure δ if the following is true. For any distributions $\mathcal{D}^{\text{train}}_{\mathcal{XY}}, \mathcal{D}^{\text{test}}_{\mathcal{XY}}$ over $\mathcal{X} \times \{\pm 1\}$ such that $\mathcal{D}^{\text{train}}_{\mathcal{X}} = \mathcal{D}$, algorithm $\mathcal{A}$, upon receiving a large enough set of labelled samples S_{train} from the training distribution $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ and a large enough set of unlabelled samples X_{test} from the test distribution $\mathcal{D}^{\text{test}}_{\mathcal{X}}$, either rejects $(S_{\text{train}}, X_{\text{test}})$ or accepts and outputs a hypothesis $h : \mathcal{X} \to \{\pm 1\}$ with the following guarantees.

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

If A accepts, then the output h satisfies $\text{err}(h; \mathcal{D}^{\text{test}}_{\mathcal{XY}}) \leq \psi(\lambda) + \epsilon$.

(b) (Completeness.) Whenever $\mathcal{D}^{\text{test}}_{\mathcal{X}} = \mathcal{D}^{\text{train}}_{\mathcal{X}}$, A accepts with probability at least $1 - \delta$ over the samples $S_{\text{train}}, X_{\text{test}}$.

In particular, we say that $\mathcal{A}$ testably learns $\mathcal{C}$ with distribution shift w.r.t. $\mathcal{D}$ in the realizable setting, if $\mathcal{A}$ is required to satisfy the above guarantees only when $\mathcal{D}^{\text{train}}_{\mathcal{XY}}, \mathcal{D}^{\text{test}}_{\mathcal{XY}}$ and $\mathcal{C}$ are realizable (where $\lambda = 0 = \psi(\lambda)$).

4 TDS Learning of Homogeneous Halfspaces

We now provide a proof of Theorem 2.1, which we restate here for convenience.

Theorem 4.1 (Agnostic TDS learning of Halfspaces)**.** *Let $\mathcal{C}$ be the class of origin-centered halfspaces over $\mathbb{R}^d$ and $C > 0$ a sufficiently large universal constant. Let $\mathcal{A}, \mathcal{T}$ be as defined in Propositions 2.2 and 4.2. Let $m_{\mathcal{A}}$ be the sample complexity of $\mathcal{A}(\epsilon/C, \delta/4)$ and $m_{\mathcal{T}} = \frac{Cd^4}{\epsilon^2\delta}$. Then, Algorithm 1, given inputs $S_{\text{train}}, X_{\text{test}}$ of sizes $|S_{\text{train}}| \geq m_{\mathcal{A}}$ and $|X_{\text{test}}| \geq m_{\mathcal{T}}$ is a TDS learning algorithm for $\mathcal{C}$ w.r.t. any isotropic log-concave distribution $\mathcal{D}$ with error $O(\lambda) + \epsilon$ and run-time $\text{poly}(d, \frac{1}{\epsilon}) \log(\frac{1}{\delta})$, where ϵ is the accuracy parameter and δ is the failure probability.*

Algorithm 1: Agnostic TDS Learning of Halfspaces

Input: Sets S_{train} from $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$, X_{test} from $\mathcal{D}^{\mathrm{test}}_{\mathcal{X}}$, parameters $\epsilon > 0$, $\delta \in (0,1)$
Set $\epsilon' = \epsilon/C$ where C is some sufficiently large universal constant.
Let $m_{\mathcal{A}}$ be the sample complexity of $\mathcal{A}(\epsilon', \delta/4)$.
Split S_{train} to S_1, S_2 with sizes $m_{\mathcal{A}}, \frac{C}{\epsilon^2}\log(1/\delta)$
Run $\mathcal{A}(\epsilon', \delta/4)$ on S_1 and obtain $\widehat{\mathbf{v}} \in \mathbb{S}^{d-1}$
Let $\widehat{\epsilon} = \mathbb{P}_{(\mathbf{x},y)\sim S_2}[\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}) \neq y]$.
Run $\mathcal{T}(\widehat{\epsilon}, \delta/2)$ on X_{test}.
Reject and terminate if $\mathcal{T}$ rejects.
Otherwise, output $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$ with $\widehat{f} : \mathbf{x} \to \mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x})$.

In order to prove the above theorem, we make use of the following result from [DKTZ20].

Proposition 4.2 (Agnostic Learning of Homogeneous Halfspaces, Theorem 3.1 in [DKTZ20])**.** *Let $\mathcal{D}_{\mathcal{XY}}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ such that its marginal on $\mathbb{R}^d$ is isotropic log-concave. Then there is an algorithm $\mathcal{A}$ such that for any $\epsilon > 0$ and $\delta \in (0,1)$, $\mathcal{A}(\epsilon, \delta)$, upon drawing $m = \tilde{O}(\frac{d}{\epsilon^4}\log(1/\delta))$ independent examples from $\mathcal{D}_{\mathcal{XY}}$ and in time $\mathrm{poly}(d, 1/\epsilon)\cdot\log(1/\delta)$, outputs $\widehat{\mathbf{v}} \in \mathbb{S}^{d-1}$ such that, with probability at least $1-\delta$, the corresponding halfspace has error at most $O(\eta)+\epsilon$, where η is the error of the optimal halfspace on $\mathcal{D}_{\mathcal{XY}}$.*

We also use the following fact about isotropic log-concave distributions.

Fact 4.3. $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{v}^*\cdot\mathbf{x})] = \Theta(\measuredangle(\widehat{\mathbf{v}}, \mathbf{v}^*))$*, when $\mathcal{D}$ is isotropic log-concave.*

Proof. Suppose that S_{train} is a set of m_{train} independent samples from $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$, where the marginal of $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$ on $\mathbb{R}^d$ is the standard Gaussian distribution. Let also X_{test} be a set of m_{test} independent unlabelled samples from $\mathcal{D}^{\mathrm{test}}_{\mathcal{X}}$. In what follows, let $\epsilon' = \epsilon/C$ and let $C > 0$ be a sufficiently large universal constant. Let also $m_{\mathcal{A}}$ be the sample complexity of $\mathcal{A}(\epsilon', \delta/4)$ and $m_{\mathcal{T}} = \frac{Cd^4}{\epsilon^2\delta}$.

Soundness. Suppose that the algorithm accepts. Let $\mathbf{v}^* \in \mathbb{S}^{d-1}$ define the halfspace f^* that achieves $\mathrm{err}(f^*; \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) + \mathrm{err}(f^*; \mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}) = \lambda$. Note that since $|S_2| \geq \frac{C}{\epsilon^2}\log(1/\delta)$, we have that $\widehat{\epsilon} \leq \mathrm{err}(\widehat{f}; \mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}) + \epsilon'$. By Proposition 4.2, since $|S_1| \geq m_{\mathcal{A}}$ we have $\mathrm{err}(\widehat{f}; \mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}) \leq \eta + \epsilon'$, where $\eta \in (0,1)$ is the error of the optimum halfspace, say $f : \mathbf{x} \mapsto \mathrm{sign}(\mathbf{v}\cdot\mathbf{x})$ on $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$. Note that $\eta \leq \lambda$. We have that $\mathrm{err}(\widehat{f}, f; \mathcal{D}^{\mathrm{train}}_{\mathcal{X}}) \leq \mathrm{err}(\widehat{f}; \mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}) + \mathrm{err}(f; \mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}) \leq 2\eta + \epsilon'$. Therefore, due to Fact 4.3, and since $\mathcal{D}^{\mathrm{train}}_{\mathcal{X}} = \mathcal{D}$, we obtain $\measuredangle(\widehat{\mathbf{v}}, \mathbf{v}) \leq 2C'\eta + C'\epsilon'$ for some sufficiently large $C' > 0$ (with $C \gg C'$).

Moreover, we have that $\mathrm{err}(f^*; \mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}) \leq \lambda$ and, hence $\mathrm{err}(f^*, f; \mathcal{D}^{\mathrm{train}}_{\mathcal{X}}) \leq \lambda + \eta$. We now apply Proposition 2.2, to obtain $\mathrm{err}(\widehat{f}, f^*; X_{\mathrm{test}}) \leq \sqrt{C}\measuredangle(\widehat{\mathbf{v}}, \mathbf{v}^*)$. Since $|X_{\mathrm{test}}| \geq \frac{\sqrt{C}}{\epsilon^2}\log(1/\delta)$, due to standard VC dimension arguments, we have $\mathrm{err}(\widehat{f}, f^*; \mathcal{D}^{\mathrm{test}}_{\mathcal{X}}) \leq \sqrt{C}\measuredangle(\widehat{\mathbf{v}}, \mathbf{v}^*) + \epsilon'$. By Fact 4.3, $\measuredangle(\widehat{\mathbf{v}}, \mathbf{v}^*) \leq C'\mathrm{err}(\widehat{f}, f^*; \mathcal{D}^{\mathrm{train}}_{\mathcal{X}})$. Therefore, with probability at least $1-\delta$, we have

$$\begin{aligned}\mathrm{err}(\widehat{f}; \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) &\leq \mathrm{err}(\widehat{f}, f^*; \mathcal{D}^{\mathrm{test}}_{\mathcal{X}}) + \mathrm{err}(f^*; \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) \leq \sqrt{C}\mathrm{err}(\widehat{f}, f^*; \mathcal{D}^{\mathrm{train}}_{\mathcal{X}}) + \epsilon' + \lambda \\ &\leq \sqrt{C}\mathrm{err}(\widehat{f}, f; \mathcal{D}^{\mathrm{train}}_{\mathcal{X}}) + \sqrt{C}\mathrm{err}(f, f^*; \mathcal{D}^{\mathrm{train}}_{\mathcal{X}}) + \epsilon' + \lambda \\ &\leq C\lambda + C\epsilon' \leq \epsilon\end{aligned}$$

Completeness. Readily follows from Proposition 2.2 and $|X_{\mathrm{test}}| \geq m_{\mathcal{T}}$. □

5 Realizable TDS Learning

5.1 Disagreement-Based TDS Learners

In this section, we prove Theorem 2.6. First, we prove the following a special version regarding realizable TDS learning of homogeneous halfspaces with respect to the Gaussian distribution.

Proposition 5.1 (TDS learning of Homogeneous Halfspaces)**.** *Let $\mathcal{C}$ be the class of origin-centered halfspaces over $\mathbb{R}^d$ and $C > 0$ a sufficiently large universal constant. Then, Algorithm 2, given inputs $S_{\mathrm{train}}, X_{\mathrm{test}}$ of sizes $|S_{\mathrm{train}}| \geq C(\frac{d}{\epsilon})^{3/2}\log(\frac{1}{\epsilon\delta})$ and $|X_{\mathrm{test}}| \geq C(\frac{d}{\epsilon}+\frac{1}{\epsilon^2})\log(\frac{1}{\epsilon\delta})$ is a TDS learning algorithm for $\mathcal{C}$ w.r.t. the standard Gaussian distribution $\mathcal{N}(0, I_d)$ with run-time $\mathrm{poly}(d, 1/\epsilon)\log(\frac{1}{\delta})$, where ϵ is the accuracy parameter and δ is the failure probability.*

Algorithm 2: TDS Learning of Homogeneous Halfspaces

Input: Sets S_{train} from $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$, X_{test} from $\mathcal{D}^{\mathrm{test}}_{\mathcal{X}}$, parameter $\epsilon > 0$

Set $\epsilon' = \epsilon^{3/2}/(10d^{1/2})$.

Run the Empirical Risk Minimization algorithm on S_{train} up to error ϵ', i.e., compute a vector $\widehat{\mathbf{v}} \in \mathbb{S}^{d-1}$ with $\widehat{\mathbf{v}} = \arg\min_{\mathbf{v}'\in\mathbb{S}^{d-1}} \mathbb{P}_{(\mathbf{x},y)\in S_{\mathrm{train}}}[y \neq \mathrm{sign}(\mathbf{v}'\cdot\mathbf{x})]$

Let $\mathcal{V} = \{\mathbf{v}' \in \mathbb{S}^{d-1} : \|\mathbf{v}' - \widehat{\mathbf{v}}\|_2 \leq \epsilon'\}$.

For each $\mathbf{x} \in X_{\mathrm{test}}$, compute the following quantities.

$$\mathbf{v}^+_{\mathbf{x}} = \arg\max_{\mathbf{v}'\in\mathcal{V}} \mathbf{v}'\cdot\mathbf{x} \text{ and } \mathbf{v}^-_{\mathbf{x}} = \arg\min_{\mathbf{v}'\in\mathcal{V}} \mathbf{v}'\cdot\mathbf{x}$$

Reject and terminate if $\mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathrm{sign}(\mathbf{v}^+_{\mathbf{x}}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{v}^-_{\mathbf{x}}\cdot\mathbf{x})] > 3\epsilon/4$.

Otherwise, output $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$ with $\widehat{f} : \mathbf{x} \mapsto \mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x})$.

We will use the following fact about the Gaussian distribution.

Fact 5.2. *For any $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{S}^{d-1}$ we have $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\mathrm{sign}(\mathbf{v}_1\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{v}_2\cdot\mathbf{x})] = \measuredangle(\mathbf{v}_1, \mathbf{v}_2)/\pi$.*

Proof of Proposition 5.1. Suppose that S_{train} is a set of m_{train} independent samples from $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$, where the marginal of $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$ on $\mathbb{R}^d$ is the standard Gaussian distribution. Let also X_{test} be a set of m_{test} independent unlabelled samples from $\mathcal{D}^{\mathrm{test}}_{\mathcal{X}}$. In what follows, let $\epsilon' = \epsilon^{3/2}/(8d^{1/2})$.

Soundness. When the algorithm accepts, we have that $\mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathrm{sign}(\mathbf{v}^+_{\mathbf{x}}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{v}^-_{\mathbf{x}}\cdot\mathbf{x})] \leq \frac{3\epsilon}{2}$. By standard VC dimension arguments and Fact 5.2, after running the Empirical Risk Minimization algorithm on training data, as long as $m_{\mathrm{train}} \geq C\frac{d}{\epsilon'}\log(1/(\delta\epsilon'))$, we have $\|\widehat{\mathbf{v}} - \mathbf{v}\|_2 \leq \epsilon'$. Therefore, both $\mathbf{v}$ and $\widehat{\mathbf{v}}$ are within $\mathcal{V} = \{\mathbf{v}' \in \mathbb{S}^{d-1} : \|\mathbf{v}' - \widehat{\mathbf{v}}\|_2 \leq \epsilon'\}$. By the definition of $\mathbf{v}^+_{\mathbf{x}}$ and $\mathbf{v}^-_{\mathbf{x}}$, we have the following.

$$\mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{v}\cdot\mathbf{x})] \leq \mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathrm{sign}(\mathbf{v}^+_{\mathbf{x}}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{v}^-_{\mathbf{x}}\cdot\mathbf{x})] \leq 3\epsilon/4 \tag{5.1}$$

Moreover, we have $\mathrm{err}(\widehat{f}; \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) = \mathbb{E}[\mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{v}\cdot\mathbf{x})]]$, where the expectation is over $X_{\mathrm{test}} \sim (\mathcal{D}^{\mathrm{test}}_{\mathcal{X}})^{\otimes m_{\mathrm{test}}}$. By standard VC dimension arguments, we have that, with probability at least $1-\delta/2$, $\mathrm{err}(\widehat{f}; \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) = \mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{v}\cdot\mathbf{x})] + \epsilon/4$ whenever $m_{\mathrm{test}} \geq C\frac{d}{\epsilon}\log(1/(\delta\epsilon))$. Therefore, with probability at least $1-\delta$ (union bound over two bad events), upon acceptance, we have $\mathrm{err}(\widehat{f}; \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) \leq \epsilon$.

Completeness. For completeness, we assume that X_{test} is drawn from $\mathcal{N}(0, I_d)$. Observe that $\mathcal{V}$ does not depend on X_{test} (since it is formed only using training data). Therefore, we may apply a standard Hoeffding bound to ensure that with probability at least $1-\delta$, whenever $m_{\text{test}} \geq C\frac{1}{\epsilon^2}\log(1/(\delta))$, we have

$$\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[\text{sign}(\mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x}) \neq \text{sign}(\mathbf{v}_{\mathbf{x}}^- \cdot \mathbf{x})] \leq \mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\text{sign}(\mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x}) \neq \text{sign}(\mathbf{v}_{\mathbf{x}}^- \cdot \mathbf{x})] + \epsilon/4$$

It remains to bound $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\text{sign}(\mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x}) \neq \text{sign}(\mathbf{v}_{\mathbf{x}}^- \cdot \mathbf{x})]$ by $\epsilon/2$. We observe that, since $\mathbf{v}^+, \mathbf{v}^- \in \mathcal{V}$, we have $\mathbf{v}_{\mathbf{x}}^- \cdot \mathbf{x} \geq \mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x} - \|\mathbf{v}_{\mathbf{x}}^+ - \mathbf{v}_{\mathbf{x}}^-\|_2\|\mathbf{x}\|_2 \geq \mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x} - \epsilon'\|\mathbf{x}\|_2 \geq \widehat{\mathbf{v}} \cdot \mathbf{x} - \epsilon'\|\mathbf{x}\|_2$ by the definition of $\mathbf{v}_{\mathbf{x}}^+$ and $\mathbf{v}_{\mathbf{x}}^-$. We similarly have $\mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x} \leq \widehat{\mathbf{v}} \cdot \mathbf{x} + \epsilon'\|\mathbf{x}\|_2$.

Therefore the probability that $\text{sign}(\mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x}) \neq \text{sign}(\mathbf{v}_{\mathbf{x}}^- \cdot \mathbf{x})$ is upper bounded by the probability that $|\widehat{\mathbf{v}} \cdot \mathbf{x}| \leq \epsilon'\|\mathbf{x}\|_2$ (since, otherwise, both $\mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x}$ and $\mathbf{v}_{\mathbf{x}}^- \cdot \mathbf{x}$ have the same sign). In particular

$$\begin{aligned}
\mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\text{sign}(\mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x}) \neq \text{sign}(\mathbf{v}_{\mathbf{x}}^- \cdot \mathbf{x})] &\leq \mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[|\widehat{\mathbf{v}} \cdot \mathbf{x}| \leq \epsilon'\|\mathbf{x}\|_2] \\
&\leq \mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\|\mathbf{x}\|_2 > \sqrt{4d/\epsilon}] + \mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[|\widehat{\mathbf{v}} \cdot \mathbf{x}| \leq \epsilon'\sqrt{4d/\epsilon}] \\
&\leq \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\|\mathbf{x}\|_2^2]\epsilon}{4d} + \mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[|\widehat{\mathbf{v}} \cdot \mathbf{x}| \leq \epsilon'\sqrt{4d/\epsilon}]
\end{aligned}$$

We obtain the final inequality by applying Markov's inequality. Since $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\|\mathbf{x}\|_2^2] = d$ and the one-dimensional Gaussian density is upper bounded by $(2\pi)^{-1}$, we have the following bound.

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\text{sign}(\mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x}) \neq \text{sign}(\mathbf{v}_{\mathbf{x}}^- \cdot \mathbf{x})] \leq \frac{\epsilon}{4} + \frac{2}{\sqrt{2\pi}}\epsilon'\sqrt{4d/\epsilon} \leq \epsilon/2\,,$$

since $\epsilon' \leq \epsilon^{3/2}/(8d^{1/2})$. This completes the proof. □

We now prove Theorem 2.6, which we restate here for convenience.

Theorem 5.3 (Disagreement-Based TDS learning)**.** *Let $\mathcal{C}$ be the class of concepts that map $\mathcal{X} \subseteq \mathbb{R}^d$ to $\{\pm 1\}$ with VC dimension $\text{VC}(\mathcal{C})$, let $\mathcal{D}$ a distribution over $\mathcal{X}$ and $C > 0$ a sufficiently large universal constant. Suppose that we have access to an ERM oracle for PAC learning $\mathcal{C}$ under $\mathcal{D}$ and membership access to $\mathbf{D}_{\epsilon'}(f;\mathcal{D})$ for any given $f \in \mathcal{C}$ and $\epsilon' > 0$. Then, Algorithm 3, given inputs of sizes $|S_{\text{train}}| \geq C\frac{\text{VC}(\mathcal{C})}{\epsilon'}\log(\frac{1}{\epsilon'\delta})$ and $|X_{\text{test}}| \geq C(\frac{\text{VC}(\mathcal{C})}{\epsilon} + \frac{1}{\epsilon^2})\log(\frac{1}{\epsilon\delta})$ is a TDS learning algorithm for $\mathcal{C}$ w.r.t. $\mathcal{D}$ that calls the ϵ'-ERM oracle once and the ϵ'-membership oracle $|S_{\text{train}}|$ times, where ϵ is the accuracy parameter, δ is the failure probability and ϵ' such that $\epsilon' \cdot \theta(\epsilon', d) \leq \epsilon/2$.*

Algorithm 3: Disagreement-Based TDS Learning

Input: Sets S_{train} from $\mathcal{D}_{\mathcal{XY}}^{\text{train}}$, X_{test} from $\mathcal{D}_{\mathcal{X}}^{\text{test}}$, parameter $\epsilon > 0$
Set $\epsilon' > 0$ such that $\epsilon' \cdot \theta(\epsilon', d) \leq \epsilon/2$.
Run the Empirical Risk Minimization algorithm on S_{train} up to error ϵ', i.e., compute $\widehat{f} \in \mathcal{C}$ with $\widehat{f} = \arg\min_{f'\in\mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\in S_{\text{train}}}[y \neq f'(\mathbf{x})]$
Let $\mathbf{D}_{\epsilon'}(\widehat{f};\mathcal{D})$ be as in Definition 2.5.
Reject and terminate if $\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}[\mathbf{x} \in \mathbf{D}_{\epsilon'}(\widehat{f};\mathcal{D})] > \epsilon/2$.
Otherwise, output $\widehat{f}$.

Proof of Theorem 2.6. Suppose that S_{train} is a set of m_{train} independent samples from $\mathcal{D}_{\mathcal{XY}}^{\text{train}}$, where the marginal of $\mathcal{D}_{\mathcal{XY}}^{\text{train}}$ on $\mathcal{X}$ is the distribution $\mathcal{D}$. Let also X_{test} be a set of m_{test} independent unlabelled samples from $\mathcal{D}_{\mathcal{X}}^{\text{test}}$. In what follows, let $\epsilon' > 0$ such that $\epsilon'\theta(\epsilon', d) \leq \epsilon/2$. The proof follows a similar recipe as the one of Proposition 5.1. For the following, let $f^* \in \mathcal{C}$ be the label generating function.

Soundness. Suppose that the algorithm accepts. Then, $\mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathbf{x} \in \mathbf{D}_{\epsilon'}(\widehat{f}; \mathcal{D})] \leq \epsilon/2$. Since $\widehat{f}$ is an minimizes the empirical error on training data, by standard VC arguments, we have that $\mathrm{err}(\widehat{f}, f^*; \mathcal{D}) \leq \epsilon/2$, whenever $m_{\mathrm{train}} \geq C\frac{\mathrm{VC}(\mathcal{C})}{\epsilon'}\log(\frac{1}{\epsilon'\delta})$, since $\mathcal{D}^{\mathrm{train}}_{\mathcal{X}} = \mathcal{D}$ by assumption. Therefore, by the definition of $\mathbf{D}_{\epsilon'}(\widehat{f}; \mathcal{D})$, for any $\mathbf{x} \notin \mathbf{D}_{\epsilon'}(\widehat{f}; \mathcal{D})$, we have $\widehat{f}(\mathbf{x}) = f^*(\mathbf{x})$. Therefore, we have

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim X_{\mathrm{test}}}[\widehat{f}(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \mathop{\mathbb{P}}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathbf{x} \in \mathbf{D}_{\epsilon'}(\widehat{f}; \mathcal{D})] \leq \epsilon/2$$

Whenever $m_{\mathrm{test}} \geq C\frac{\mathrm{VC}(\mathcal{C})}{\epsilon}\log(\frac{1}{\epsilon\delta})$, we have $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}^{\mathrm{test}}_{\mathcal{X}\mathcal{Y}}}[y \neq f^*(\mathbf{x})] \leq \mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\widehat{f}(\mathbf{x}) \neq f^*(\mathbf{x})] + \epsilon/2 \leq \epsilon$.

Completeness. Suppose that $\mathcal{D}^{\mathrm{test}}_{\mathcal{X}} = \mathcal{D}$. Then, by a standard Hoeffding bound, we have that whenever $m_{\mathrm{test}} \geq C\frac{1}{\epsilon}\log(1/\delta)$, we have $\mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathbf{x} \in \mathbf{D}_{\epsilon'}(\widehat{f}; \mathcal{D})] \leq \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{D}_{\epsilon'}(\widehat{f}; \mathcal{D})] + \epsilon/2$ with probability at least $1-\delta$ and $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{D}_{\epsilon'}(\widehat{f}; \mathcal{D})] \leq \epsilon'\theta(\epsilon', d) \leq \epsilon/2$, by the choice of ϵ'. □

5.2 TDS Learner for General Halfspaces

We now prove Theorem 2.7 which we restate here for convenience.

Theorem 5.4 (TDS learning of General Halfspaces)**.** *Let $\mathcal{C}$ be the class of general halfspaces over $\mathbb{R}^d$ and $C > 0$ a sufficiently large universal constant. Then, Algorithm 4, given inputs of size $|S_{\mathrm{train}}| = |X_{\mathrm{test}}| = Cd^{C\log 1/\epsilon}$ is a TDS learning algorithm for $\mathcal{C}$ w.r.t. the standard Gaussian distribution $\mathcal{N}(0, I_d)$ with runtime $d^{O(\log 1/\epsilon)}$, where ϵ is the accuracy parameter, and the failure probability δ is at most* 0.01.

Algorithm 4: TDS Learning of General Halfspaces

Input: Sets S_{train} from $\mathcal{D}^{\mathrm{train}}_{\mathcal{X}\mathcal{Y}}$, X_{test} from $\mathcal{D}^{\mathrm{test}}_{\mathcal{X}}$, parameter $\epsilon > 0$

Set $T = 2^{C_1^2\log\frac{1}{\epsilon}+1}$, $k = C_1\log\frac{1}{\epsilon}$, $\Delta = \frac{\epsilon}{d^{C_2k}}$ and $\beta = \frac{\epsilon^2}{C_3d^{C_3}}$.

if $\mathbb{P}_{(\mathbf{x},y)\sim S_{\mathrm{train}}}[y \neq b] \leq \frac{1}{T}$ for some $b \in \{\pm 1\}$ (large bias case) **then**

  For each $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \leq k$, compute the quantity $\widehat{\mathrm{M}}_\alpha = \mathbb{E}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathbf{x}^\alpha]$.

  Reject and terminate if $|\widehat{\mathrm{M}}_\alpha - \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[\mathbf{x}^\alpha]| > \Delta$ for some α with $\|\alpha\|_1 \leq k$.

  Otherwise, output $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$ and terminate, where $\widehat{f} : \mathbf{x} \mapsto b$ ($\widehat{f}$ constant).

else

  Set $\widehat{\mathbf{v}} = \frac{\mathbb{E}_{(\mathbf{x},y)\sim S_{\mathrm{train}}}[y\mathbf{x}]}{\|\mathbb{E}_{(\mathbf{x},y)\sim S_{\mathrm{train}}}[y\mathbf{x}]\|_2}$.

  Let $\mathcal{T} = \{\widehat{\mathbf{v}}\cdot\mathbf{x} : (\mathbf{x}, y) \in S_{\mathrm{train}}\}$.

  Set $\widehat{\tau} = \arg\min_{\tau\in\mathcal{T}} \mathbb{P}_{(\mathbf{x},y)\in S_{\mathrm{train}}}[f^*(\mathbf{x}) \neq \mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x} - \tau')]$,

  Let $\mathcal{V} = \{(\mathbf{v}', \tau') : \|\mathbf{v}' - \widehat{\mathbf{v}}\|_2 \leq \beta, |\tau' - \widehat{\tau}| \leq \beta\}$.

  For each $\mathbf{x} \in X_{\mathrm{test}}$, compute the following quantities.

$$(\mathbf{v}^+_{\mathbf{x}}, \tau^+_{\mathbf{x}}) = \mathop{\arg\max}_{(\mathbf{v}',\tau')\in\mathcal{V}} \mathbf{v}'\cdot\mathbf{x} - \tau' \text{ and } (\mathbf{v}^-_{\mathbf{x}}, \tau^-_{\mathbf{x}}) = \mathop{\arg\min}_{(\mathbf{v}',\tau')\in\mathcal{V}} \mathbf{v}'\cdot\mathbf{x} - \tau'$$

  Reject and terminate if $\mathbb{P}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathrm{sign}(\mathbf{v}^+_{\mathbf{x}}\cdot\mathbf{x} - \tau^+_{\mathbf{x}}) \neq \mathrm{sign}(\mathbf{v}^-_{\mathbf{x}}\cdot\mathbf{x} - \tau^-_{\mathbf{x}})] > 10\epsilon$.

  Otherwise, output $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$ with $\widehat{f} : \mathbf{x} \mapsto \mathrm{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x} - \widehat{\tau})$.

end if

Suppose the ground-truth halfspace $f^*(\mathbf{x}) = \mathrm{sign}(\mathbf{x}\cdot\mathbf{v} - \tau)$ is determined by a unit vector $\mathbf{v} \in \mathbb{R}^d$ and a value $\tau \in \mathbb{R}$. We will need the following showing that if a halfspace not too biased under the standard Gaussian distribution, then it is possible to recover the parameters of the halfspace up to a very high accuracy. See Subsection 5.2.3 for the proof.

Proposition 5.5 (Parameter recovery for halfspaces)**.** *For a sufficiently large absolute constant $C > 0$, the following is true. For every $\beta, \gamma \in (0,1)$ and integer d, let S_{train} be a set of $C(\frac{d}{\beta\gamma})^C$ i.i.d samples from a distribution $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ such that $\mathcal{D}^{\text{train}}_{\mathcal{X}} = \mathcal{N}(0, I_d)$ and the labels are given by an unknown halfspace $f : \mathbf{x} \mapsto \text{sign}(\mathbf{v} \cdot \mathbf{x} - \tau)$. Additionally, assume that the halfspace f satisfies $\mathbb{P}_{\mathbf{x} \in \mathcal{N}(0,I_d)}[f^*(\mathbf{x}) = -1] \geq \gamma$ and $\mathbb{P}_{\mathbf{x} \in \mathcal{N}(0,I_d)}[f^*(\mathbf{x}) = 1] \geq \gamma$. Let $\mathcal{T} = \{\widehat{\mathbf{v}} \cdot \mathbf{x} : (\mathbf{x}, y) \in S_{\text{train}}\}$ and set*

$$\widehat{\mathbf{v}} = \frac{\sum_{(\mathbf{x},y) \in S_{\text{train}}} \mathbf{x}y}{\|\sum_{(\mathbf{x},y) \in S_{\text{train}}} \mathbf{x}y\|_2} \text{ and } \widehat{\tau} = \arg\min_{\tau' \in \mathcal{T}} \mathbb{P}_{(\mathbf{x},y) \in S_{\text{train}}}[f^*(\mathbf{x}) \neq \text{sign}(\widehat{\mathbf{v}} \cdot \mathbf{x} - \tau')].$$

Then, with probability at least $1 - 1/1000$ we have $\|\mathbf{v} - \widehat{\mathbf{v}}\|_2 \leq \beta$ and $|\tau - \widehat{\tau}| \leq \beta$.

We also highlight two technical lemmas that we use for the analysis of Algorithm 4. Our first technical lemma insures that if f is a halfspace that very likely assigns the same label to samples from the Gaussian distribution, then f also very likely assigns the same label to samples form a distribution whose low-degree moments match those of a Gaussian. This lemma will be useful for proving the soundness of Algorithm 4, and is proven in Section 5.2.2. (Recall that for $\mathbf{x} \in \mathbb{R}^d$ we denote $\prod_{i=1}^n x_i^{\alpha_i}$ as $\mathbf{x}^\alpha$.)

Lemma 5.6. *When C_1 and C_2 both exceed some specific absolute constant, the following holds. Let k and T be defined as in Algorithm 4. Suppose, the set X_{test} is such that for every collection of non-negative integers $(\alpha_1, \cdots, \alpha_d)$ satisfying $\sum_i \alpha_i \leq k$ we have*

$$\left| \mathop{\mathbb{E}}_{\mathbf{x} \sim X_{\text{test}}} [\mathbf{x}^\alpha] - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{N}(0,I_d)} [\mathbf{x}^\alpha] \right| \leq \frac{\epsilon}{d^{C_2 k}}. \tag{5.2}$$

Also, suppose the function $f^*(\mathbf{x}) = \text{sign}(\mathbf{x} \cdot \mathbf{v} - \tau)$ and the value $L \in \{\pm 1\}$ are such that

$$\mathop{\mathbb{P}}_{\mathbf{x} \sim \mathcal{N}(0,1)}[f^*(\mathbf{x}) \neq L] \leq \frac{2}{T}. \tag{5.3}$$

Then, it is the case that

$$\mathop{\mathbb{P}}_{\mathbf{x} \sim X_{\text{test}}}[f^*(\mathbf{x}) \neq L] \leq O(\epsilon). \tag{5.4}$$

Our second technical lemma bounds, for $\mathbf{x}$ chosen from the standard Gaussian, the probability that one is unsure about $f^*(\mathbf{x}) = \text{sign}(\mathbf{v} \cdot \mathbf{x} - \tau)$ when one only has approximate estimates for $\widehat{\mathbf{v}}$ and $\widehat{\tau}$ for $\mathbf{v}$ and τ respectively. This lemma will be useful for proving the completeness of Algorithm 4, and is proven in Section 5.2.1.

Lemma 5.7. *There is some absolute constant K_1, such that for every positive integer d and $\beta \in (0,1)$, the following holds. Let $\widehat{\mathbf{v}}$ be any unit vector in $\mathbb{R}^d$ and $\widehat{\tau}$ be in $\mathbb{R}$. Then, we have for $\mathcal{V} = \{(\mathbf{v}', \tau') : \|\mathbf{v}' - \widehat{\mathbf{v}}\|_2 \leq \beta, |\tau' - \widehat{\tau}| \leq \beta\}$*

$$\mathop{\mathbb{P}}_{\mathbf{x} \sim \mathcal{N}(0,I_d)} \left[\text{sign}\left(\max_{(\mathbf{v}',\tau') \in \mathcal{V}} \mathbf{v}' \cdot \mathbf{x} - \tau' \right) \neq \text{sign}\left(\min_{(\mathbf{v}',\tau') \in \mathcal{V}} \mathbf{v}' \cdot \mathbf{x} - \tau' \right) \right] \leq K_1 d^{K_1} \sqrt{\beta} \tag{5.5}$$

5.2.1 Proof of Soundness.

In this subsection we show that if Algorithm 4 accepts then the output $\widehat{f}$ of our algorithm will generalize on the distribution $\mathcal{D}^{\text{test}}_{\mathcal{X}}$.

Proposition 5.8 (Soundness)**.** *For any sufficiently large absolute constant C, the following is true. For any distribution $\mathcal{D}^{\text{test}}_{\mathcal{X}}$ and any halfspace $f = \text{sign}(\widehat{\mathbf{v}} \cdot \mathbf{x} - \widehat{\tau})$, the following is true. It can happen with probability only at most $\frac{1}{100}$ that Algorithm 4 gives an output (ACCEPT, $\widehat{f}$) for some predictor $\widehat{f}$, but it is not the case that*

$$\mathop{\mathbb{P}}_{\mathbf{x} \sim \mathcal{D}^{\text{test}}_{\mathcal{X}}}[f^*(\mathbf{x}) \neq \widehat{f}(\mathbf{x})] \leq O(\epsilon).$$

To prove this proposition, we first need to prove Lemma 5.6.

Proof of Lemma 5.6. First of all, we claim that Equation 5.3 implies that

$$|\tau| \geq \sqrt{\frac{1}{2}\log\frac{T}{2}} \tag{5.6}$$

Indeed, we have

$$\frac{2}{T} \geq \frac{1}{\sqrt{2\pi}}\int_{|\tau|}^{\infty} e^{-z^2/2}dz \geq |\tau|e^{-2|\tau|^2} \geq e^{-2|\tau|^2},$$

where the last inequality holds because for sufficiently large C_1 the value of T and therefore $|\tau|$ is sufficiently large and exceeds 1.

Recall that $\mathbf{v}$ is assumed to be a unit vector in $\mathbb{R}^d$. Assume, without loss of generality, that $L = -1$, and therefore $\tau > 0$. We have

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim X_{\text{test}}}[\text{sign}(\mathbf{x}\cdot\mathbf{v}-\tau) \neq -1] = \mathop{\mathbb{P}}_{\mathbf{x}\sim X_{\text{test}}}[\mathbf{x}\cdot\mathbf{v} \geq \tau] \leq \frac{\mathbb{E}_{\mathbf{x}\sim X_{\text{test}}}[(\mathbf{x}\cdot\mathbf{v})^k]}{\tau^k}. \tag{5.7}$$

To use this inequality, we need to upper-bound $\mathbb{E}_{\mathbf{x}\sim X_{\text{test}}}[(\mathbf{x}\cdot\mathbf{v})^k]$. Since $\mathbf{v}$ is a unit vector, every (of at most d^k) terms of the polynomial mapping $\mathbf{x} \in \mathbb{R}^d$ to $(\mathbf{x}\cdot\mathbf{v})^k$ has coefficient at most 1. This, together with Equation 5.2 and the triangle inequality, allows us to conclude that

$$\left|\mathop{\mathbb{E}}_{\mathbf{x}\sim X_{\text{test}}}[(\mathbf{x}\cdot\mathbf{v})^k] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[(\mathbf{x}\cdot\mathbf{v})^k]\right| \leq d^k\frac{\epsilon}{d^{C_2 k}}.$$

Now, since $\mathbf{v}$ is a unit vector, we have $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}[(\mathbf{x}\cdot\mathbf{v})^k] = k!! \leq k^k$. Combining this with the equation above, and Equation 5.7 and then substituting Equation 5.6 and the values of k and T we get:

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim X_{\text{test}}}[\text{sign}(\mathbf{x}\cdot\mathbf{v}-\tau) \neq -1] \leq \frac{1}{|\tau|^k}\left(k^{k/2} + d^k\frac{\epsilon}{d^{C_2 k}}\right) \leq \frac{1}{\left(\frac{1}{2}C_1^2\log\frac{1}{\epsilon}\right)^{C_1\log\frac{1}{\epsilon}}}\left(\left(C_1\log\frac{1}{\epsilon}\right)^{C_1\log\frac{1}{\epsilon}} + d^k\frac{\epsilon}{d^{C_2 k}}\right)$$

We see that when C_1 and C_2 both exceed some absolute constant, the above expression is at most ϵ, which completes the proof. □

Having proven Lemma 5.6, we are now ready to prove Proposition 5.8.

Proof of Proposition 5.8. First, suppose the algorithm outputs (ACCEPT, L) for some $L \in \{\pm 1\}$ via Step 5. For the algorithm to reach this step, it has to be that

$$\mathop{\mathbb{P}}_{\mathbf{x}\in S}[f^*(\mathbf{x}) \neq L] \leq \frac{1}{T},$$

Via Hoeffding's inequality, if C is sufficiently large then with probability at least $1 - \frac{1}{1000}$ it holds that

$$|\mathop{\mathbb{P}}_{\mathbf{x}\in S}[f^*(\mathbf{x}) \neq L] - \mathop{\mathbb{P}}_{\mathbf{x}\in S}[f^*(\mathbf{x}) \neq L]| \leq \frac{1}{2T}, \tag{5.8}$$

and combining the two equations above

$$\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[f^*(\mathbf{x}) \neq L] \leq \frac{2}{T}.$$

Furthermore, for the algorithm not output REJECT in Step 4, it has to be the case that for every collection of non-negative integers $(\alpha_1, \cdots, \alpha_d)$ satisfying $\sum_i \alpha_i \leq k$ we have

$$\left| \mathop{\mathbb{E}}_{\mathbf{x} \sim X_{\text{test}}} [\mathbf{x}^\alpha] - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{N}(0, I_d)} [\mathbf{x}^\alpha] \right| > \frac{\epsilon}{d^{C_2 k}}.$$

Overall, this allows us to apply Lemma 5.6 to conclude that

$$\mathop{\mathbb{P}}_{\mathbf{x} \sim X_{\text{test}}} [f^*(\mathbf{x}) \neq L] \leq O(\epsilon),$$

and, for a sufficiently large absolute constant C, with probability at least $1 - \frac{1}{1000}$, this is only possible if

$$\mathop{\mathbb{P}}_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}^{\text{test}}} [f^*(\mathbf{x}) \neq L] \leq O(\epsilon),$$

which finishes the proof for the case when the algorithm accepts in Step 5.

Now, suppose the algorithm accepts in Step 13. For the algorithm to reach this step, it has to be that

$$\mathop{\mathbb{P}}_{\mathbf{x} \in S} [f^*(\mathbf{x}) \neq L] > \frac{1}{T},$$

And together with Equation 5.8, this implies that

$$\mathop{\mathbb{P}}_{\mathbf{x} \in \mathcal{N}(0, I_d)} [f^*(\mathbf{x}) \neq L] > \frac{1}{2T}.$$

For such f^* we can apply Proposition 5.5 and conclude that with probability at least $1 - 1/1000$ the values of $\widehat{\mathbf{v}}$ and τ obtained in Algorithm 4 satisfy

$$\|\mathbf{v} - \widehat{\mathbf{v}}\|_2 \leq \left(\frac{\epsilon}{C_3 d^{C_3}} \right)^2 = \beta, \tag{5.9}$$

$$|\tau - \widehat{\tau}| \leq \left(\frac{\epsilon}{C_3 d^{C_3}} \right)^2 = \beta, \tag{5.10}$$

where the last equality is by the definition of β. Now, since the algorithm did not reject in Step 12, it must be the case that the fraction of elements in X_{test} that satisfy $\text{sign}(\mathbf{v}_{\mathbf{x}}^+ \cdot \mathbf{x} - \tau_{\mathbf{x}}^+) \neq \text{sign}(\mathbf{v}_{\mathbf{x}}^- \cdot \mathbf{x} - \tau_{\mathbf{x}}^-)$ is at most 10ϵ. If C is a sufficiently large absolute constant, the standard Hoeffding inequality tells us that for this to happen with probability larger than $1/1000$ it has to be the case that

$$\mathop{\mathbb{P}}_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}^{\text{test}}} \left[\text{sign} \left(\max_{(\mathbf{v}', \tau') \in \mathcal{V}} \mathbf{v}' \cdot \mathbf{x} - \tau' \right) \neq \text{sign} \left(\min_{(\mathbf{v}', \tau') \in \mathcal{V}} \mathbf{v}' \cdot \mathbf{x} - \tau' \right) \right] \leq 11\epsilon.$$

Whenever the event above occurs, since $\mathcal{V} = \{(\mathbf{v}', \tau') : \|\mathbf{v}' - \widehat{\mathbf{v}}\|_2 \leq \beta, |\tau' - \widehat{\tau}| \leq \beta\}$ we can use Equations 5.9 and 5.10 to conclude $\text{sign}(\mathbf{v} \cdot \mathbf{x} - \tau) = \text{sign}(\widehat{\mathbf{v}} \cdot \mathbf{x} - \widehat{\tau})$. Therefore,

$$\mathop{\mathbb{P}}_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}^{\text{test}}} [\text{sign}(\mathbf{v} \cdot \mathbf{x} - \tau) \neq \text{sign}(\widehat{\mathbf{v}} \cdot \mathbf{x} - \widehat{\tau})] \leq 11\epsilon$$

This completes the proof of soundness of Algorithm 4. □

5.2.2 Proof of Completeness.

The second proposition shows that if the testing distribution is the standard Gaussian, then the algorithm will likely accept. Together, propositions 5.8 and 5.9 yield Theorem 2.7.

Proposition 5.9 (Completeness)**.** *For sufficiently large value of the absolute constants C and C_3 and for any halfspace $f = \mathrm{sign}(\widehat{\mathbf{v}} \cdot \mathbf{x} - \widehat{\tau})$, suppose the testing distribution $\mathcal{D}_{\mathcal{X}}^{\mathrm{test}}$ is the standard Gaussian distribution. Then, with probability at least $1 - \frac{1}{100}$ Algorithm 4 will accept, i.e. output (ACCEPT, $\widehat{f}$) for some $\widehat{f}$.*

To prove this proposition, we first need to prove Lemma 5.7.

Proof of Lemma 5.7. We have $\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}[\|\mathbf{x}\|_2^2] = d$. Therefore, by Markov's inequality, we have

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}\left[\|\mathbf{x}\|_2 > \frac{\sqrt{d}}{\sqrt{\beta}}\right] = \mathbb{P}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}\left[\|\mathbf{x}\|_2^2 > \frac{d}{\beta}\right] \leq \beta \tag{5.11}$$

Additionally, from the bound of $\frac{1}{\sqrt{2\pi}}$ on the density of standard Gaussian in one dimension, we get:

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}\left[|\widehat{\mathbf{v}} \cdot \mathbf{x} - \widehat{\tau}| \leq 100\sqrt{\beta d} + \beta\right] \leq \frac{200\sqrt{\beta d} + 2\beta}{\sqrt{2\pi}} \tag{5.12}$$

If it holds that $\|\mathbf{x}\|_2 \leq \frac{\sqrt{d}}{\sqrt{\beta}}$, we have for every $\mathbf{v}'$ satisfying $\|\mathbf{v}' - \widehat{\mathbf{v}}\|_2 \leq \beta$ and any τ' satisfying $|\tau' - \widehat{\tau}| \leq \beta$ that

$$|\mathbf{v}' \cdot \mathbf{x} - \tau' - (\widehat{\mathbf{v}} \cdot \mathbf{x} - \widehat{\tau})| \leq \sqrt{d\beta} + \beta$$

Therefore, if it is also the case that $|\widehat{\mathbf{v}} \cdot \mathbf{x} - \widehat{\tau}| > 100\sqrt{\beta d} + \beta$, then we have

$$\mathrm{sign}\left(\mathbf{v}' \cdot \mathbf{x} - \tau'\right) = \mathrm{sign}\left(\widehat{\mathbf{v}} \cdot \mathbf{x} - \widehat{\tau}\right)$$

This allows us to conclude that

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}\left[\mathrm{sign}\left(\max_{(\mathbf{v}', \tau') \in \mathcal{V}} \mathbf{v}' \cdot \mathbf{x} - \tau'\right) \neq \mathrm{sign}\left(\min_{(\mathbf{v}', \tau') \in \mathcal{V}} \mathbf{v}' \cdot \mathbf{x} - \tau'\right)\right] \leq$$
$$\mathbb{P}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}\left[\|\mathbf{x}\|_2 > \frac{\sqrt{d}}{\sqrt{\beta}}\right] + \mathbb{P}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}\left[|\widehat{\mathbf{v}} \cdot \mathbf{x} - \widehat{\tau}| \leq 100\sqrt{\beta d} + \beta\right] \leq \beta + \frac{200\sqrt{\beta d} + 2\beta}{\sqrt{2\pi}},$$

where in the end we substituted Equation 5.11 and Equation 5.12. Recalling that for $\beta \in (0, 1)$ we have $\beta < \sqrt{\beta}$ and picking K_1 to be a sufficiently large absolute constant, our proposition follows from the inequality above. □

Having proven Lemma 5.7, we are now ready to prove Proposition 5.9.

Proof of Proposition 5.9. There are two ways for the algorithm to output REJECT: through Step 4 and through Step 12. We will argue neither takes place. From standard Gaussian concentration, if C is a sufficiently large absolute const ant, with probability at least $1 - \frac{1}{1000}$ the algorithm will not output REJECT in Step 4.

We now proceed to ruling out the possibility that the algorithm outputs REJECT in Step 12. For the algorithm to reach step Step 12, it is necessary that

$$\mathbb{P}_{\mathbf{x} \in S}[f^*(\mathbf{x}) \neq L] > \frac{1}{T},$$

Via Hoeffding's inequality, if C is sufficiently large then with probability at least $1 - \frac{1}{1000}$ it holds that $|\mathbb{P}_{\mathbf{x}\in S}[f^*(\mathbf{x}) \neq L] - \mathbb{P}_{\mathbf{x}\in S}[f^*(\mathbf{x}) \neq L]| \leq \frac{1}{2T}$, which together with the equation above implies that

$$\mathbb{P}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[f^*(\mathbf{x}) \neq L] > \frac{1}{2T}.$$

For such f^* we can apply Proposition 5.5 and conclude that with probability at least $1 - 1/1000$ the values of $\widehat{\mathbf{v}}$ and τ obtained in Algorithm 4 satisfy

$$\|\mathbf{v} - \widehat{\mathbf{v}}\|_2 \leq \left(\frac{\epsilon}{C_3 d^{C_3}}\right)^2 = \beta, \tag{5.13}$$

$$|\tau - \widehat{\tau}| \leq \left(\frac{\epsilon}{C_3 d^{C_3}}\right)^2 = \beta, \tag{5.14}$$

Recall that $\mathcal{V} = \{(\mathbf{v}', \tau') : \|\mathbf{v}' - \widehat{\mathbf{v}}\|_2 \leq \beta, |\tau' - \widehat{\tau}| \leq \beta\}$. The equation above together with Lemma 5.7 implies that

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,I_d)}\left[\operatorname{sign}\left(\max_{(\mathbf{v}',\tau')\in\mathcal{V}} \mathbf{v}'\cdot\mathbf{x} - \tau'\right) \neq \operatorname{sign}\left(\min_{(\mathbf{v}',\tau')\in\mathcal{V}} \mathbf{v}'\cdot\mathbf{x} - \tau'\right)\right] \leq K_1 d^{K_1}\frac{\epsilon}{C_3 d^{C_3}} \leq \epsilon,$$

where the last inequality holds for sufficiently large value of C_3. Combining the inequality above with the standard Hoeffding bound and recalling that $\mathcal{D}_{\mathcal{X}}^{\text{test}} = \mathcal{N}(0, I_d)$, we see that with probability at least $1 - \frac{1}{1000}$,

$$\mathbb{P}_{\mathbf{x}\sim X_{\text{test}}}\left[\operatorname{sign}\left(\max_{(\mathbf{v}',\tau')\in\mathcal{V}} \mathbf{v}'\cdot\mathbf{x} - \tau'\right) \neq \operatorname{sign}\left(\min_{(\mathbf{v}',\tau')\in\mathcal{V}} \mathbf{v}'\cdot\mathbf{x} - \tau'\right)\right] \leq 2\epsilon,$$

In conclusion, we see that the inequality above implies that the algorithm does not output REJECT in Step 12. This completes our proof. □

5.2.3 Parameter recovery.

Here we prove Proposition 5.5, which was used in the proofs of Proposition 5.8 and Proposition 5.9, thereby finishing the proof of Theorem 2.7. Let us first recall the setting of Proposition 5.5. For a unit vector $\mathbf{v}$ in $\mathbb{R}^d$ and $\tau \in \mathbb{R}$ satisfying

$$\min\left(\mathbb{P}_{x\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x} - \tau > 0], \mathbb{P}_{x\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x} - \tau < 0]\right) \geq \eta,$$

S_{train} is a set of $C\left(\frac{d}{\eta\beta}\right)^C$ i.i.d samples from a distribution $\mathcal{D}_{\mathcal{X}\mathcal{Y}}^{\text{train}}$ with $\mathcal{X}$-marginal distributed as standard Gaussian and $\mathcal{Y}$-marginal given by the halfspace $f = \operatorname{sign}(\mathbf{v}\cdot\mathbf{x} - \tau)$. The absolute constant C is assumed to be sufficiently large. We let $\mathcal{T} = \{\widehat{\mathbf{v}}\cdot\mathbf{x} : (\mathbf{x}, y) \in S_{\text{train}}\}$ and set

$$\widehat{\mathbf{v}} = \frac{\sum_{(\mathbf{x},y)\in S_{\text{train}}} \mathbf{x}y}{\|\sum_{(\mathbf{x},y)\in S_{\text{train}}} \mathbf{x}y\|_2}$$

$$\widehat{\tau} = \operatorname*{arg\,min}_{\tau'\in\mathcal{T}} \mathbb{P}_{(\mathbf{x},y)\in S_{\text{train}}}[f^*(\mathbf{x}) \neq \operatorname{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x} - \tau')].$$

We would like to prove that with probability at least $29/30$ we have

$$\|\mathbf{v} - \widehat{\mathbf{v}}\|_2 \leq \beta,$$

$$|\tau - \widehat{\tau}| \leq \beta.$$

The following proposition tells us that the first inequality above is likely to hold:

Proposition 5.10 (Recovery of normal vector for halfspaces)**.** *For a sufficiently large absolute constant C, and every $\eta, \beta \in (0,1)$ and integer d, the following holds. Let S_{train} is a set of at least $C\left(\frac{d}{\eta\beta}\right)^C$ i.i.d samples from a distribution $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ with $\mathcal{X}$-marginal distributed as standard Gaussian and $\mathcal{Y}$-marginal given by the halfspace $f = \text{sign}(\mathbf{v}\cdot\mathbf{x} - \tau)$. For every unit vector $\mathbf{v}$ in $\mathbb{R}^d$ and $\tau \in \mathbb{R}$ satisfying*

$$\min\left(\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x} - \tau > 0], \mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x} - \tau < 0]\right) \geq \eta,$$

The vector $\widehat{\mathbf{v}} = \frac{\sum_{(\mathbf{x},y)\in S_{\text{train}}} \mathbf{x}y}{\|\sum_{(\mathbf{x},y)\in S_{\text{train}}} \mathbf{x}y\|_2}$ with probability at least $1 - \frac{1}{2000}$ satisfies:

$$\|\mathbf{v} - \widehat{\mathbf{v}}\|_2 \leq \beta,$$

Once this stage is accomplished, the next proposition tells us that we can recover the offset τ.

Proposition 5.11 (Offset recovery for halfspaces)**.** *For a sufficiently large absolute constant C, and every $\eta, \gamma \in (0,1)$ and integer d, the following holds. Let S_{train} is a set of at least $C\left(\frac{d}{\eta\gamma}\right)^C$ i.i.d samples from a distribution $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ with $\mathcal{X}$-marginal distributed as standard Gaussian and $\mathcal{Y}$-marginal given by the halfspace $f = \text{sign}(v\cdot\mathbf{x} - \tau)$. For every unit vector $\mathbf{v}$ in $\mathbb{R}^d$ and $\tau \in \mathbb{R}$ satisfying*

$$\min\left(\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x} - \tau > 0], \mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x} - \tau < 0]\right) \geq \eta,$$

Then, with probability at least $1 - \frac{1}{2000}$, for every unit vector $\widehat{\mathbf{v}}$ that forms an angle of at most γ with $\mathbf{v}$ the value

$$\widehat{\tau} = \mathop{\arg\min}_{\tau'\in\mathbb{R}} \mathop{\mathbb{P}}_{(\mathbf{x},y)\in S_{\text{train}}}[f^*(\mathbf{x}) \neq \text{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x} - \tau')].$$

satisfies

$$|\tau - \widehat{\tau}| \leq O\left(\frac{1}{\eta^{50}}\sqrt{\gamma}\right).$$

Formally, Proposition 5.5 follows from the two propositions above as follows. One first uses Proposition 5.10 to conclude that, for any absolute constant C_5, there is a value of the absolute constant C for which with probability $1 - \frac{1}{2000}$ a vector $\widehat{\mathbf{v}}$ that satisfies $\|\mathbf{v} - \widehat{\mathbf{v}}\| \leq \frac{1}{C_5}\beta^2\eta^{100}$. This implies that the angle between $\mathbf{v}$ and $\widehat{\mathbf{v}}$ is upper-bounded by $\frac{10}{C_5}\beta^2\eta^{100}$. Then, if the absolute constant C_5 is large enough, if we use Proposition 5.11, then with probability $1 - \frac{1}{2000}$ the value $\widehat{\tau}$ satisfies $|\tau - \widehat{\tau}| \leq \beta$, finishing the proof of Proposition 5.5.

Now, proceed to prove the two propositions above. We start with Proposition 5.10.

Proof of Proposition 5.10. Let $\{\mathbf{e}_1, \cdots \mathbf{e}_{d-1}\}$ form an orthonormal basis for the subspace orthogonal to $\mathbf{v}$. Since all the projections $\{\mathbf{v}\cdot\mathbf{x}, \mathbf{e}_1\cdot\mathbf{x}, \cdots, \mathbf{e}_{d-1}\cdot\mathbf{x}\}$ are independent standard Gaussians and $f^*(\mathbf{x}) = \text{sign}(\mathbf{v}\cdot\mathbf{x} - \tau)$ we have for all i

$$\mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[\mathbf{e}_i\cdot\mathbf{x}f^*(\mathbf{x})] = 0.$$

At the same time

$$\mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x}f(\mathbf{x})] = \int_{t=-\infty}^{+\infty} t\,\text{sign}(t-\tau)\frac{1}{\sqrt{2\pi}}\,dt =$$
$$\int_{t\in[-|\tau|,|\tau|]} t\,\text{sign}(t-\tau)\frac{1}{\sqrt{2\pi}}\,dt + \int_{t\in[-\infty,-|\tau|]\cup[|\tau|,+\infty]} t\,\text{sign}(t-\tau)\frac{1}{\sqrt{2\pi}}\,dt = \frac{2}{\sqrt{2\pi}}\int_{t=|\tau|}^{\infty} t\,dt$$

For some positive absolute constant K_2, the final expression above is at least $K_2 \mathbb{P}_{t\sim N(0,1)}[t > \tau]$, because if $|\tau| > 1$, then one can lower-bound the expression above by $\frac{2}{\sqrt{2\pi}} \int_{t=|\tau|}^{\infty} dt$. On the other hand, if $|\tau| \in [0, 1]$, then the expression on the right side is at least $\frac{2}{\sqrt{2\pi}} \int_{t=1}^{\infty} dt$ which is a positive absolute constant, while $\mathbb{P}_{t\sim N(0,1)}[t > \tau]$ is always upper-bounded by 1. Overall, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x}f^*(\mathbf{x})] &\geq K_2 \mathop{\mathbb{P}}_{t\sim N(0,1)}[t > \tau] \\
&= K_2 \min\left(\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x} - \tau > 0], \mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}[\mathbf{v}\cdot\mathbf{x} - \tau < 0]\right) \\
&\geq K_2\eta.
\end{aligned}$$

Now, we bound the variance of $\mathbf{x}f^*(\mathbf{x})$. Since $f^*(\mathbf{x}) \in \{\pm 1\}$, we have

$$\mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}\left[(\mathbf{e}_i\cdot\mathbf{x}f^*(\mathbf{x}))^2\right] = \mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}\left[(\mathbf{e}_i\cdot\mathbf{x})^2\right] = 1,$$

$$\mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}\left[(\mathbf{v}\cdot\mathbf{x}f^*(\mathbf{x}))^2\right] = \mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}\left[(\mathbf{v}\cdot\mathbf{x})^2\right] = 1.$$

This allows us to use the Chebychev's inequality together with the union bound to conclude that with probability at least $1 - \frac{1}{2000}$ we have for all i

$$|E_{\mathbf{x}\in S}[\mathbf{e}_i\cdot\mathbf{x}f^*(\mathbf{x})]| \leq \sqrt{\frac{60d}{N}},$$

and also

$$E_{\mathbf{x}\in S}[\mathbf{v}\cdot\mathbf{x}f^*(\mathbf{x})] \geq K_2\eta - \sqrt{\frac{60d}{N}},$$

Recalling that $\widehat{\mathbf{v}} = \frac{\sum_{\mathbf{x}\in S_1}\mathbf{x}f^*(\mathbf{x})}{\|\sum_{\mathbf{x}\in S_1}\mathbf{x}f^*(\mathbf{x})\|_2} = \frac{\mathbb{E}_{\mathbf{x}\in S_1}\mathbf{x}f^*(\mathbf{x})}{\|\mathbb{E}_{\mathbf{x}\in S_1}\mathbf{x}f^*(\mathbf{x})\|_2}$, we see that

$$|\widehat{\mathbf{v}}\cdot\mathbf{e}_i| \leq \frac{\sqrt{\frac{60d}{N}}}{K_2\eta - \sqrt{\frac{60d}{N}}}$$

Substituting $N = C(\frac{d}{\eta\beta})^C$, and letting C be a sufficiently large absolute constant, we obtain from above implies that $|\widehat{\mathbf{v}}\cdot\mathbf{e}_i| \leq \frac{\beta}{10\sqrt{d}}$. Since $\|\widehat{\mathbf{v}}\| = 1$ we have

$$1 \geq |\widehat{\mathbf{v}}\cdot\mathbf{v}| \geq \sqrt{1 - \frac{\beta}{10}} \geq 1 - \frac{\beta}{10},$$

we also see that taking C to be a sufficiently large absolute constant also ensures that $\widehat{\mathbf{v}}\cdot\mathbf{v} > 0$, so overall we get

$$\|\widehat{\mathbf{v}} - \mathbf{v}\| \leq \beta,$$

which finishes the proof. □

In order to prove Proposition 5.11, we will need a proposition that relates the following two quantities: (1) the difference in offsets τ_1 and τ_2 of two halfspaces (2) The probability that these two hafspaces disagree on a point drawn from the standard Gaussian.

Proposition 5.12. *There is some absolute constant K_1 such that for any pair of unit vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$ and a pair of real numbers τ_1, τ_2, letting γ denote the angle between $\mathbf{v}_1$ and $\mathbf{v}_2$, the following holds. Suppose $\gamma < \pi/4$, then*

$$\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}\left[\operatorname{sign}\left(\mathbf{v}_1\cdot\mathbf{x}-\tau_1\right)\neq\operatorname{sign}\left(\mathbf{v}_2\cdot\mathbf{x}-\tau_2\right)\right]\geq\frac{1}{K_1}e^{-\tau_1^2/2}\min\left(\left|\tau_1-\frac{\tau_2}{\cos\gamma}\right|,\frac{1}{|\tau_1|+1}\right) \tag{5.15}$$

It is also the case that

$$\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}\left[\operatorname{sign}\left(\mathbf{v}_1\cdot\mathbf{x}-\tau_1\right)\neq\operatorname{sign}\left(\mathbf{v}_2\cdot\mathbf{x}-\tau_1\cos\gamma\right)\right]\leq K_1\sqrt{\gamma} \tag{5.16}$$

Proof. To prove this, we first show that for any $z \in \mathbb{R}$, conditioned on $\mathbf{v}_1 \cdot \mathbf{x} = z_1$ the distribution of $\mathbf{v}_2 \cdot \mathbf{x}$ is $\mathcal{N}(z_1 \cos\gamma, \sin\gamma)$. Indeed, let $\mathbf{v}_3$ be the unit vector that one obtains by first projecting $\mathbf{v}_2$ into the subspace perpendicular to $\mathbf{v}_1$, and then normalizing the resulting vector to have unit norm. This means $\mathbf{v}_3$ is orthogonal to $\mathbf{v}_1$ and we have

$$\mathbf{v}_2 = \mathbf{v}_1\cos\gamma + \mathbf{v}_3\sin\gamma.$$

Therefore

$$\mathbf{x}\cdot\mathbf{v}_2 = \mathbf{x}\cdot\mathbf{v}_1\cos\gamma + \mathbf{x}\cdot\mathbf{v}_3\sin\gamma$$

Now, since $\mathbf{x}\cdot\mathbf{v}_1$ and $\mathbf{x}\cdot\mathbf{v}_3$ are distributed as i.i.d. one-dimensional standard Gaussians. Thus, conditioning on $\mathbf{x}\cdot\mathbf{v}_1 = z_1$ we get that $\mathbf{x}\cdot\mathbf{v}_2$ is distributed as $\mathcal{N}(z\cos\gamma, \sin\gamma)$.

Our observation allows us to write:

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{N}(0,I_d)}\left[\operatorname{sign}\left(\mathbf{v}_1\cdot\mathbf{x}-\tau_1\right)\neq\operatorname{sign}\left(\mathbf{v}_2\cdot\mathbf{x}-\tau_2\right)\right] &= \\ \mathop{\mathbb{P}}_{z_1,z_2\in\mathcal{N}(0,1)}\left[\operatorname{sign}\left(z_1-\tau_1\right)\neq\operatorname{sign}\left(z_1\cos\gamma+z_2\sin\gamma-\tau_2\right)\right] &= \\ \mathop{\mathbb{P}}_{z_1,z_2\in\mathcal{N}(0,1)}\left[\operatorname{sign}\left(z_1-\tau_1\right)\neq\operatorname{sign}\left(z_1+z_2\tan\gamma-\tau_2/\cos\gamma\right)\right] &\end{aligned} \tag{5.17}$$

Let us first focus on the case when $\gamma \in [0, \pi/2)$. We see that

$$\begin{aligned}\mathop{\mathbb{P}}_{z_1,z_2\in\mathcal{N}(0,1)}\left[\operatorname{sign}\left(z_1-\tau_1\right)\neq\operatorname{sign}\left(z_1+z_2\tan\gamma-\tau_2/\cos\gamma\right)\right]\geq & \\ \frac{1}{2}\mathop{\mathbb{P}}_{z_1\in\mathcal{N}(0,1)}\left[\operatorname{sign}\left(z_1-\tau_1\right)\neq\operatorname{sign}\left(z_1-\tau_2/\cos\gamma\right)\right] &\end{aligned} \tag{5.18}$$

The reason that inequality above is true is that, conditioned on a specific value of z_1, if $z_1 > \tau_2/\cos\gamma$, then $z_1 + z_2\tan\gamma - \tau_2$ is more likely to be positive than negative. At the same time, if $z_1 < \tau_2/\cos\gamma$, then $z_1 + z_2\tan\gamma - \tau_2$ is more likely to be negative than positive.

We lower-bound the probability above as follows. Let A be the interval of $\mathbb{R}$ defined as follows:

$$A := \left\{z\in\mathbb{R}:\quad \operatorname{sign}(z-\tau_1)\neq\operatorname{sign}(z-\tau_2/\cos\gamma)\ \ \&\ \ |z-\tau_1|\leq\frac{1}{|\tau_1|+1}\right\}$$

We have

$$\begin{aligned}\mathop{\mathbb{P}}_{z_1\in\mathcal{N}(0,1)}\left[\operatorname{sign}\left(z_1-\tau_1\right)\neq\operatorname{sign}\left(z_1-\tau_2/\cos\gamma\right)\right]\geq\mathop{\mathbb{P}}_{z_1\in\mathcal{N}(0,1)}\left[z_1\in A\right]\geq & \\ \geq\min\left(\left|\tau_1-\frac{\tau_2}{\cos\gamma}\right|,\frac{1}{|\tau_1|+1}\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(|\tau_1|-\frac{1}{|\tau_1|+1}\right)^2} & \\ \geq\Omega(1)\cdot\min\left(\left|\tau_1-\frac{\tau_2}{\cos\gamma}\right|,\frac{1}{|\tau_1|+1}\right)e^{-\tau_1^2/2}, &\end{aligned} \tag{5.19}$$

which, combined with Equations 5.17 and 5.18, finishes the proof of Equation 5.15.

Now, we proceed to proving Equation 5.16. We proceed as follows:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{z_1,z_2\in\mathcal{N}(0,1)}\left[\operatorname{sign}\left(z_1-\tau_1\right)=\operatorname{sign}\left(z_1+z_2\tan\gamma-\tau_1\right)\right]\\
&\qquad\geq \mathop{\mathbb{P}}_{z_1,z_2\in\mathcal{N}(0,1)}\left[|z_1-\tau_1|>\sqrt{\tan\gamma}\ \ \&\ \ |z_2|<\frac{1}{\sqrt{\tan\gamma}}\right]\\
&\qquad\geq 1-O(1)\cdot\sqrt{\tan\gamma}-O(1)\int_{\frac{1}{\sqrt{\tan\gamma}}}^{\infty}e^{-z^2/2}\,dz\\
&\qquad= 1-O(\sqrt{\tan\gamma})=1-O(\sqrt{\gamma}),
\end{aligned}$$

which, when combining with with Equation 5.17 and substituting $\tau_2=\tau_1\cos\gamma$, proves Equation 5.16. □

Having proven Proposition 5.12, we are now ready to prove Proposition 5.11.

Proof of Proposition 5.11. Recall that $\mathcal{T}=\{\widehat{\mathbf{v}}\cdot\mathbf{x}:(\mathbf{x},y)\in S_{\text{train}}\}$. We see for τ' between two neighboring elements of $\mathcal{T}$ the value of $\mathbb{P}_{\mathbf{x}\in\mathbb{N}(0,I)}[f^*(\mathbf{x})\neq\operatorname{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}-\tau')]$ stays the same. Therefore

$$\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{T}}[f^*(\mathbf{x})\neq\operatorname{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}-\widehat{\tau})]=\min_{\tau'\in\mathcal{T}}\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{T}}[f^*(\mathbf{x})\neq\operatorname{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}-\tau')]=\min_{\tau'\in\mathbb{R}}\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{T}}[f^*(\mathbf{x})\neq\operatorname{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}-\tau')]. \tag{5.20}$$

Since the function class $\{\operatorname{sign}(\mathbf{v}'\cdot\mathbf{x}-\tau':\ \mathbf{v}'\in\mathbb{R}^d,\ \tau'\in\mathbb{R}\}$ has a VC dimension of $d+1$, the standard VC bound tells us that for sufficiently large absolute constant C with probability at least $1-\frac{1}{2000}$ we have for every $\tau'\in\mathbb{R}$ and unit vector $\widehat{\mathbf{v}}$ that

$$\left|\mathop{\mathbb{P}}_{\mathbf{x}\in\mathbb{N}(0,I)}[f^*(\mathbf{x})\neq\operatorname{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}-\tau']-\mathop{\mathbb{P}}_{\mathbf{x}\in\mathcal{T}}[f^*(\mathbf{x})\neq\operatorname{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}-\tau')]\right|\leq\sqrt{\gamma} \tag{5.21}$$

From Equation 5.16 in Proposition 5.12 we have that

$$\min_{\tau'\in\mathbb{R}}\mathop{\mathbb{P}}_{\mathbf{x}\in\mathbb{N}(0,I)}[f^*(\mathbf{x})\neq\operatorname{sign}(\widehat{\mathbf{v}}\cdot\mathbf{x}-\tau')]\leq K_1\sqrt{\gamma}\leq O(\sqrt{\gamma}) \tag{5.22}$$

We now upper-bound $|\tau|$ in terms η as follows:

$$|\tau|\leq 10\sqrt{\log\frac{1}{\eta}}, \tag{5.23}$$

For $|\tau|<1$, this is immediate, because the probability that the Gaussian exceeds one standard deviation in a given direction is at least $1/10$. For $|\tau|\geq 1$, we write

$$\eta\geq\int_{|\tau|}^{\infty}e^{-t^2/2}\,dt\geq\frac{1}{|\tau|}e^{-(|\tau|+1/|\tau|)^2/2}\geq\frac{1}{e^2}\cdot\frac{1}{|\tau|}e^{-|\tau|^2/2},$$

which proves Equation 5.23.

Taking Equation 5.15 in Proposition 5.12 and substituting Equation 5.23 we get

$$\mathop{\mathbb{P}}_{x\in\mathcal{N}(0,I_d)}\left[f(x)\neq\operatorname{sign}\left(\widehat{\mathbf{v}}\cdot x-\widehat{\tau}\right)\right]\geq\frac{1}{K_1}e^{-\tau_1^2/2}\min\left(\left|\tau-\frac{\widehat{\tau}}{\cos\gamma}\right|,\frac{1}{|\tau|+1}\right)\geq\frac{\eta^{50}}{K_1}\min\left(\left|\tau-\frac{\widehat{\tau}}{\cos\gamma}\right|,1\right)$$

Combining the above with Equation 5.20, Equation 5.21 and Equation 5.22 we get

$$\left|\tau - \frac{\widehat{\tau}}{\cos\gamma}\right| \le \frac{K_1}{\eta^{50}}(O(\sqrt{\gamma}) + \sqrt{\gamma}) \le O(\sqrt{\gamma}/\eta^{50}).$$

Finally, we see that

$$|\tau - \widehat{\tau}| \le \left|\tau - \frac{\widehat{\tau}}{\cos\gamma}\right| + \left|\widehat{\tau} - \frac{\widehat{\tau}}{\cos\gamma}\right| \le O(\sqrt{\gamma}/\eta^{50}) + O(\sqrt{\log(1/\eta)}\gamma^2) = O(\sqrt{\gamma}/\eta^{50}).$$

This completes the proof of Proposition 5.11. □

6 TDS Learning Through Moment Matching

6.1 $\mathcal{L}_2$-Sandwiching Implies TDS Learning

We now prove Theorem 2.9 which we restate here for convenience.

Theorem 6.1 ($\mathcal{L}_2$-sandwiching implies TDS Learning)**.** *Let $\mathcal{D}$ be a distribution over a set $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ be a concept class. Let $\epsilon, \delta \in (0,1)$, $\epsilon' = \epsilon/100$ $\delta' = \delta/2$ and assume that the following are true.*

(i) ($\mathcal{L}_2$-Sandwiching) The ϵ'-approximate $\mathcal{L}_2$-sandwiching degree of $\mathcal{C}$ under $\mathcal{D}$ is at most k with coefficient bound B.

(ii) (Moment Concentration) If $X \sim \mathcal{D}^{\otimes m}$ and $m \ge m_{\mathrm{conc}}$ then, with probability at least $1-\delta'$, we have that for any $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \le k$ it holds $|\mathbb{E}_X[\mathbf{x}^\alpha] - \mathbb{E}_{\mathcal{D}}[\mathbf{x}^\alpha]| \le \frac{\epsilon'}{B^2 d^{4k}}$.

(iii) (Generalization) If $S \sim \mathcal{D}_{\mathcal{XY}}^{\otimes m}$ where $\mathcal{D}_{\mathcal{XY}}$ is any distribution over $\mathcal{X} \times \{\pm 1\}$ such that $\mathcal{D}_{\mathcal{X}} = \mathcal{D}$ and $m \ge m_{\mathrm{gen}}$ then, with probability at least $1-\delta'$ we have that for any degree-k polynomial p with coefficient bound B it holds $|\mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2] - \mathbb{E}_S[(y-p(\mathbf{x}))^2]| \le \epsilon'$.

Then, Algorithm 5, upon receiving $m_{\mathrm{train}} \ge m_{\mathrm{gen}}$ labelled samples S_{train} from the training distribution and $m_{\mathrm{test}} \ge C \cdot \frac{d^k + \log(1/\delta)}{\epsilon^2} + m_{\mathrm{conc}}$ unlabelled samples X_{test} from the test distribution (where $C > 0$ is a sufficiently large universal constant), runs in time $\mathrm{poly}(|S_{\mathrm{train}}|, |X_{\mathrm{test}}|, d^k)$ and TDS learns $\mathcal{C}$ with respect to $\mathcal{D}$ up to error $32\lambda + \epsilon$ and with failure probability δ.

One key ingredient of the proof of Theorem 2.9 is the following transfer lemma which states that moment matching implies that the empirical squared loss between two polynomials on the test distribution is close to their expected squared loss under the target distribution.

Lemma 6.2 (Transfer Lemma for Square Loss)**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \subseteq \mathbb{R}^d$ and X_{test} a (multi)set of points in $\mathbb{R}^d$. If $|\mathbb{E}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathbf{x}^\alpha] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}^\alpha]| \le \Delta$ for all $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \le 2k$, then for any degree k polynomials p_1, p_2 with coefficients that are absolutely bounded by B, it holds*

$$\left|\mathop{\mathbb{E}}_{\mathbf{x}\sim X_{\mathrm{test}}}[(p_1(\mathbf{x}) - p_2(\mathbf{x}))^2] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[(p_1(\mathbf{x}) - p_2(\mathbf{x}))^2]\right| \le B^2 \cdot d^{4k} \cdot \Delta$$

Proof. The polynomials p_1, p_2 all have degree at most k and coefficients that are absolutely bounded by B. Therefore, the polynomial $(p_1 - p_2)^2$ has degree at most $2k$ and coefficients that are absolutely bounded by $B^2 d^{2k}$. Let $p' = (p_1 - p_2)^2 = \sum_{\alpha:\|\alpha\|_1 \le 2k} p'_\alpha \mathbf{x}^\alpha$ (with $|p'_\alpha| \le B^2 d^{2k}$ as argued above) which gives the following.

$$\|p_1 - p_2\|^2_{\mathcal{L}_2(X_{\mathrm{test}})} = \mathop{\mathbb{E}}_{\mathbf{x}\sim X_{\mathrm{test}}}\left[(p_1(\mathbf{x}) - p_2(\mathbf{x}))^2\right] = \mathop{\mathbb{E}}_{\mathbf{x}\sim X_{\mathrm{test}}}\left[p'(\mathbf{x})\right]$$

Algorithm 5: TDS Learning through Moment Matching

Input: Sets S_{train} from $\mathcal{D}_{\mathcal{XY}}^{\mathrm{train}}$, X_{test} from $\mathcal{D}_{\mathcal{X}}^{\mathrm{test}}$, parameters $\epsilon > 0, \delta \in (0,1), k \in \mathbb{N}, B > 0$

Set $\epsilon' = \epsilon/100$, $\delta' = \delta/2$ and $\Delta = \frac{\epsilon'}{B^2 d^{4k}}$

For each $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \leq 2k$, compute the quantity

$\widehat{\mathrm{M}}_\alpha = \mathbb{E}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathbf{x}^\alpha] = \mathbb{E}_{\mathbf{x}\sim X_{\mathrm{test}}}\left[\prod_{i\in[d]} \mathbf{x}_i^{\alpha_i}\right]$

Reject and terminate if $|\widehat{\mathrm{M}}_\alpha - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}^\alpha]| > \Delta$ for some α with $\|\alpha\|_1 \leq 2k$.

Otherwise, solve the following least squares problem on S_{train} up to error ϵ'

$$\min_{p} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim S_{\mathrm{train}}} \left[(y - p(\mathbf{x}))^2\right]$$
$$\text{s.t. } p \text{ is a polynomial with degree at most } k$$
$$\text{each coefficient of } p \text{ is absolutely bounded by } B$$

Let $\widehat{p}$ be an ϵ'-approximate solution to the above optimization problem.

Accept and output $h : \mathcal{X} \to \{\pm 1\}$ where $h : \mathbf{x} \mapsto \mathrm{sign}(\widehat{p}(\mathbf{x}))$.

It remains to relate $\mathbb{E}_{\mathbf{x}\sim X_{\mathrm{test}}}[p'(\mathbf{x})]$ to $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[p'(\mathbf{x})]$, which follows by the moment-matching assumption.

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{\mathbf{x}\sim X_{\mathrm{test}}}\left[p'(\mathbf{x})\right] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[p'(\mathbf{x})\right]\right| &= \left|\sum_{\alpha:\|\alpha\|_1\leq 2k} p'_\alpha \left(\mathop{\mathbb{E}}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathbf{x}^\alpha] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}^\alpha]\right)\right| \\
&\leq \sum_{\alpha:\|\alpha\|_1\leq 2k} |p'_\alpha| \cdot \left|\mathop{\mathbb{E}}_{\mathbf{x}\sim X_{\mathrm{test}}}[\mathbf{x}^\alpha] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}^\alpha]\right| \\
&= \sum_{\alpha:\|\alpha\|_1\leq 2k} |p'_\alpha| \cdot \left|\widehat{\mathrm{M}}_\alpha - \mathrm{M}_\alpha\right| \\
&\leq d^{2k} \cdot B^2 \cdot d^{2k} \cdot \Delta\,,
\end{aligned}$$

which concludes the proof of the lemma. □

We are now ready to prove Theorem 2.9.

Proof of Theorem 2.9. For the following, let $\mathcal{D}_{\mathcal{XY}}^{\mathrm{train}}$ be the training distribution, $\mathcal{D}_{\mathcal{XY}}^{\mathrm{test}}$ the test distribution (both over $\mathcal{X} \times \{\pm 1\}$) and $\mathcal{D}_{\mathcal{X}}^{\mathrm{train}}, \mathcal{D}_{\mathcal{X}}^{\mathrm{test}}$ the corresponding marginal distributions over $\mathcal{X}$. We assume that $\mathcal{D}_{\mathcal{X}}^{\mathrm{train}} = \mathcal{D}$. Let $m_{\mathrm{train}} = |S_{\mathrm{train}}|$ and $m_{\mathrm{test}} = |X_{\mathrm{test}}|$, $\epsilon' = \epsilon/100$, $\delta' = \delta/2$, k, B as defined in condition (i). We also set $\Delta = \frac{\epsilon'}{B^2 d^{4k}}$ and m_{conc} as defined in condition (ii), as well as m_{gen} as defined in (iii).

Soundness. Suppose that Algorithm 5 accepts and outputs $h = \mathrm{sign}(\widehat{p})$. For the following, let $\lambda_{\mathrm{train}} = \mathrm{err}(f^*; \mathcal{D}_{\mathcal{XY}}^{\mathrm{train}})$ and $\lambda_{\mathrm{test}} = \mathrm{err}(f^*; \mathcal{D}_{\mathcal{XY}}^{\mathrm{test}})$ (where we have $\lambda = \lambda_{\mathrm{train}} + \lambda_{\mathrm{test}}$). We can bound the error of the hypothesis h on $\mathcal{D}_{\mathcal{XY}}^{\mathrm{test}}$ as follows

$$\begin{aligned}
\mathrm{err}(h; \mathcal{D}_{\mathcal{XY}}^{\mathrm{test}}) &\leq \mathrm{err}(f^*; \mathcal{D}_{\mathcal{XY}}^{\mathrm{test}}) + \mathrm{err}(f^*, h; \mathcal{D}_{\mathcal{X}}^{\mathrm{test}}) \\
&= \lambda_{\mathrm{test}} + \mathbb{E}[\mathrm{err}(f^*, h; X_{\mathrm{test}})]\,,
\end{aligned}$$

where the expectation above is over $X_{\mathrm{test}} \sim (\mathcal{D}_{\mathcal{X}}^{\mathrm{test}})^{\otimes m_{\mathrm{test}}}$. Denote $\mathrm{err}(h; \mathcal{D}_{\mathcal{XY}}^{\mathrm{test}}) = \mathbb{P}_{\mathcal{D}_{\mathcal{XY}}^{\mathrm{test}}}[y \neq h(\mathbf{x})]$ and $\mathrm{err}(h_1, h_2; \mathcal{D}_{\mathcal{X}}^{\mathrm{test}}) = \mathbb{P}_{\mathcal{D}_{\mathcal{X}}^{\mathrm{test}}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$ and use the fact that for random variables $y_1, y_2, y_3 \in \{\pm 1\}$, it holds $\mathbb{P}[y_1 \neq y_2] \leq \mathbb{P}[y_1 \neq y_3] + \mathbb{P}[y_2 \neq y_3]$. Since h is the sign of a polynomial with degree at most

$k = k(\epsilon')$ (see Algorithm 5) and the class of functions of this form has VC dimension at most d^k (e.g., by viewing it as the class of halfspaces in d^k dimensions) we have that whenever $m_{\text{test}} \geq C \cdot \frac{d^k + \log(1/\delta')}{\epsilon'^2}$ for some sufficiently large universal constant $C > 0$ the following is true with probability at least $1 - \delta'$ over the distribution of X_{test}.

$$\mathbb{E}[\text{err}(f^*, h; X_{\text{test}})] \leq \text{err}(f^*, h; X_{\text{test}}) + \epsilon'$$

Therefore, it is sufficient to bound the quantity $\text{err}(f^*, h; X_{\text{test}})$. We now observe the following simple fact.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x} \sim X_{\text{test}}}[(f^*(\mathbf{x}) - \widehat{p}(\mathbf{x}))^2] &\geq \mathop{\mathbb{P}}_{X_{\text{test}}}[f^*(\mathbf{x}) = 1, \widehat{p}(\mathbf{x}) < 0] + \mathop{\mathbb{P}}_{X_{\text{test}}}[f^*(\mathbf{x}) = -1, \widehat{p}(\mathbf{x}) \geq 0] \\
&= \mathop{\mathbb{P}}_{X_{\text{test}}}[f^*(\mathbf{x}) \neq \operatorname{sign} \widehat{p}(\mathbf{x})] \\
&= \text{err}(f^*, h; X_{\text{test}})
\end{aligned}$$

Therefore, we have $\text{err}(f^*, h; X_{\text{test}}) \leq \|f^* - \widehat{p}\|^2_{\mathcal{L}_2(X_{\text{test}})}$. Let $p_{\text{up}}, p_{\text{down}}$ be ϵ'-approximate $\mathcal{L}_2$ sandwiching polynomials for f^* of degree at most $k = k(\epsilon')$ and with coefficient bound $B = B(\epsilon')$. The right hand side can be bounded as follows.

$$\begin{aligned}
\|f^* - \widehat{p}\|_{\mathcal{L}_2(X_{\text{test}})} &\leq \|f^* - p_{\text{down}}\|_{\mathcal{L}_2(X_{\text{test}})} + \|p_{\text{down}} - \widehat{p}\|_{\mathcal{L}_2(X_{\text{test}})} \\
&\leq \|p_{\text{up}} - p_{\text{down}}\|_{\mathcal{L}_2(X_{\text{test}})} + \|p_{\text{down}} - \widehat{p}\|_{\mathcal{L}_2(X_{\text{test}})}
\end{aligned}$$

In the last inequality, we used the fact that $p_{\text{down}}(\mathbf{x}) \leq f^*(\mathbf{x}) \leq p_{\text{up}}(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$. We will now compare $\|p_{\text{up}} - p_{\text{down}}\|_{\mathcal{L}_2(X_{\text{test}})}$ to $\|p_{\text{up}} - p_{\text{down}}\|_{\mathcal{L}_2(\mathcal{D})}$ (and, similarly, $\|p_{\text{down}} - \widehat{p}\|_{\mathcal{L}_2(X_{\text{test}})}$ to $\|p_{\text{down}} - \widehat{p}\|_{\mathcal{L}_2(\mathcal{D})}$) using the transfer lemma (Lemma 6.2). The polynomials $p_{\text{up}}, p_{\text{down}}, \widehat{p}$ all have degree at most k and coefficients that are absolutely bounded by B. Moreover, since Algorithm 5 has accepted, we have that for any $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \leq 2k$, the following is true

$$\left|\widehat{\mathrm{M}}_\alpha - \mathrm{M}_\alpha\right| \leq \Delta\,, \tag{6.1}$$

where $\widehat{\mathrm{M}} = \mathbb{E}_{\mathbf{x} \sim X_{\text{test}}}[\mathbf{x}^\alpha]$ (recall that $\mathbf{x}^\alpha = \prod_{i \in [d]} \mathbf{x}_i^{\alpha_i}$), $\mathrm{M} = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\mathbf{x}^\alpha]$ and $\Delta = \frac{\epsilon'}{B^2 d^{4k}}$. Therefore, by applying Lemma 6.2, we obtain that $\|p_{\text{up}} - p_{\text{down}}\|_{\mathcal{L}_2(X_{\text{test}})} \leq \|p_{\text{up}} - p_{\text{down}}\|_{\mathcal{L}_2(\mathcal{D})} + \sqrt{\epsilon'}$ and, similarly, $\|p_{\text{down}} - \widehat{p}\|_{\mathcal{L}_2(X_{\text{test}})} \leq \|p_{\text{down}} - \widehat{p}\|_{\mathcal{L}_2(\mathcal{D})} + \sqrt{\epsilon'}$.

We have assumed that $p_{\text{up}}, p_{\text{down}}$ are ϵ'-approximate $\mathcal{L}_2$ sandwiching polynomials for f^* and, therefore $\|p_{\text{up}} - p_{\text{down}}\|_{\mathcal{L}_2(\mathcal{D})} = \sqrt{\|p_{\text{up}} - p_{\text{down}}\|^2_{\mathcal{L}_2(\mathcal{D})}} \leq \sqrt{\epsilon'}$ (see Definition 3.1). We bound the quantity $\|p_{\text{down}} - \widehat{p}\|_{\mathcal{L}_2(\mathcal{D})}$ as follows.

$$\begin{aligned}
\|p_{\text{down}} - \widehat{p}\|_{\mathcal{L}_2(\mathcal{D})} &\leq \|p_{\text{down}} - f^*\|_{\mathcal{L}_2(\mathcal{D})} + \|f^* - \widehat{p}\|_{\mathcal{L}_2(\mathcal{D})} \\
&\leq \|p_{\text{up}} - p_{\text{down}}\|_{\mathcal{L}_2(\mathcal{D})} + \|f^* - \widehat{p}\|_{\mathcal{L}_2(\mathcal{D})} && \text{(since } p_{\text{down}} \leq f^* \leq p_{\text{up}}\text{)} \\
&\leq \sqrt{\epsilon'} + \|f^* - \widehat{p}\|_{\mathcal{L}_2(\mathcal{D})} && (6.2)
\end{aligned}$$

Recall that $\|f^* - \widehat{p}\|^2_{\mathcal{L}_2(\mathcal{D})} = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[(\widehat{p}(\mathbf{x}) - f^*(\mathbf{x}))^2]$. By assumption, $\mathcal{D}^{\text{train}}_{\mathcal{X}} = \mathcal{D}$ and therefore $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[(\widehat{p}(\mathbf{x}) - f^*(\mathbf{x}))^2] = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}^{\text{train}}_{\mathcal{X}}}[(\widehat{p}(\mathbf{x}) - f^*(\mathbf{x}))^2]$. Moreover, we can view the expectation to be over the joint distribution $(\mathbf{x}, y) \sim \mathcal{D}^{\text{train}}_{\mathcal{X}\mathcal{Y}}$ (coupling of $\mathbf{x}$ and y), but the variable y is ignored, i.e., $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}^{\text{train}}_{\mathcal{X}}}[(\widehat{p}(\mathbf{x}) - f^*(\mathbf{x}))^2] = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}^{\text{train}}_{\mathcal{X}\mathcal{Y}}}[(\widehat{p}(\mathbf{x}) - f^*(\mathbf{x}))^2]$. We can bound the latter term as follows.

$$\begin{aligned}
\mathop{\mathbb{E}}_{(\mathbf{x}, y) \sim \mathcal{D}^{\text{train}}_{\mathcal{X}\mathcal{Y}}}[(\widehat{p}(\mathbf{x}) - f^*(\mathbf{x}))^2]^{1/2} &= \mathop{\mathbb{E}}_{(\mathbf{x}, y) \sim \mathcal{D}^{\text{train}}_{\mathcal{X}\mathcal{Y}}}[(\widehat{p}(\mathbf{x}) - y + y - f^*(\mathbf{x}))^2]^{1/2} \\
&\leq \mathop{\mathbb{E}}_{\mathcal{D}^{\text{train}}_{\mathcal{X}\mathcal{Y}}}[(\widehat{p}(\mathbf{x}) - y)^2]^{1/2} + \mathop{\mathbb{E}}_{\mathcal{D}^{\text{train}}_{\mathcal{X}\mathcal{Y}}}[(y - f^*(\mathbf{x}))^2]^{1/2}
\end{aligned}$$

For the term $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[(\widehat{p}(\mathbf{x})-y)^2]$, we use condition (iii) to have with probability at least $1-\delta'$, $|\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[(\widehat{p}(\mathbf{x})-y)^2]-\mathbb{E}_{(\mathbf{x},y)\sim S_{\text{train}}}[(\widehat{p}(\mathbf{x})-y)^2]|\le\epsilon'$ whenever $m_{\text{train}}\ge m_{\text{gen}}$. We now use the fact that $\widehat{p}$ is an ϵ'-approximate solution to the least squares problem defined in Algorithm 5 and have the following bound

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim S_{\text{train}}}[(\widehat{p}(\mathbf{x})-y)^2]^{1/2}\le\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim S_{\text{train}}}[(p_{\text{down}}(\mathbf{x})-y)^2]^{1/2}+\sqrt{\epsilon'}$$

Therefore, due to the generalization condition we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[(\widehat{p}(\mathbf{x})-y)^2]^{1/2} &\le \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[(p_{\text{down}}(\mathbf{x})-y)^2]^{1/2}+3\sqrt{\epsilon'}\\
&\le \|p_{\text{down}}-f^*\|_{\mathcal{L}_2(\mathcal{D}^{\text{train}}_{\mathcal{X}})}+\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[(y-f^*(\mathbf{x}))^2]^{1/2}+3\sqrt{\epsilon'}\\
&\le \|p_{\text{down}}-p_{\text{up}}\|_{\mathcal{L}_2(\mathcal{D})}+\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[(y-f^*(\mathbf{x}))^2]^{1/2}+3\sqrt{\epsilon'}\\
&\le \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[(y-f^*(\mathbf{x}))^2]^{1/2}+4\sqrt{\epsilon'}
\end{aligned}$$

Therefore, we have shown that $\|f^*-\widehat{p}\|_{\mathcal{L}_2(\mathcal{D})}\le 4\,\mathbb{E}_{\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[(y-f^*(\mathbf{x}))^2]^{1/2}+2\sqrt{\epsilon'}$. Note that $\mathbb{E}_{\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[(y-f^*(\mathbf{x}))^2]=4\,\mathbb{P}_{\mathcal{D}^{\text{train}}_{\mathcal{XY}}}[y\ne f^*(\mathbf{x})]=4\lambda_{\text{train}}$. Therefore, $\|f^*-\widehat{p}\|_{\mathcal{L}_2(\mathcal{D})}\le 4\sqrt{\lambda_{\text{train}}}+4\sqrt{\epsilon'}$. By Equation (6.2), this implies $\|p_{\text{down}}-\widehat{p}\|_{\mathcal{L}_2(\mathcal{D})}\le 4\sqrt{\lambda_{\text{train}}}+5\sqrt{\epsilon'}$, which in turn implies $\|p_{\text{down}}-\widehat{p}\|_{\mathcal{L}_2(X_{\text{test}})}\le 4\sqrt{\lambda_{\text{train}}}+7\sqrt{\epsilon'}$. We overall obtain the following bound.

$$\begin{aligned}
\text{err}(h;\mathcal{D}^{\text{test}}_{\mathcal{XY}}) &\le \lambda_{\text{test}}+(4\lambda^{1/2}_{\text{train}}+7\sqrt{\epsilon'})^2\\
&\le \lambda_{\text{test}}+32\lambda_{\text{train}}+100\epsilon'\\
&\le 32\lambda+\epsilon && \text{(since } \epsilon'=\epsilon/100 \text{ and } \lambda_{\text{test}}\ge 0)
\end{aligned}$$

Note that, in fact, we have also demonstrated that upon acceptance, the following is true.

$$\text{err}(f^*,h;\mathcal{D}^{\text{test}}_{\mathcal{X}})\le 32\lambda_{\text{train}}+\epsilon$$

The results above holds with probability at least $1-3\delta'=1-\delta$ (union bound over two bad events).

Completeness. For completeness, it is sufficient to ensure that $m_{\text{test}}\ge m_{\text{conc}}$, because then, the probability of acceptance is at least $1-\delta$, due to condition (ii), as required. □

6.2 Applications

In this section, we apply our main result in Theorem 2.9 to obtain a number of TDS learners for important concept classes with respect to Gaussian and Uniform target marginals. In particular, we will use the following corollary, which follows by Theorem 2.9 and some simple properties of the Gaussian and Uniform distributions (see Lemmas D.1 and D.2).

Corollary 6.3. *Let $\mathcal{D}$ be either the standard Gaussian in d dimensions or the uniform distribution over the d-dimensional hypercube. Let $\mathcal{C}$ be a concept class whose ϵ-approximate sandwiching degree with respect to $\mathcal{D}$ is k. Then, there is an algorithm that runs in time $d^{O(k)}$ and TDS learns $\mathcal{C}$ up to error $32\lambda+O(\epsilon)$ and failure probability at most 0.1.*

Boolean Classes. We now bound the $\mathcal{L}_2$ sandwiching degree of bounded size Decision trees and bounded size and depth Boolean Formulas.

Lemma 6.4 ($\mathcal{L}_2$ sandwiching degree of Decision Trees)**.** *Let $\mathcal{D}$ be the uniform distribution over the hypercube $\mathcal{X} = \{\pm 1\}^d$. For $s \in \mathbb{N}$, let $\mathcal{C}$ be the class of Decision Trees of size s. Then, for any $\epsilon > 0$ the $\mathcal{L}_2$ sandwiching degree of $\mathcal{C}$ is at most $k = O(\log(s/\epsilon))$.*

Proof. Let $f \in \mathcal{C}$ be a decision tree of size s. Consider the polynomials $p_{\mathrm{up}}, p_{\mathrm{down}}$ over $\{\pm 1\}^d$ which correspond to the following truncated decision trees. For p_{up}, we truncate f at depth k and substitute the internal nodes at depth k with leaf nodes labelled 1. Then, p_{up} corresponds to a sum of polynomials of degree at most k, each corresponding to a root-to-leaf path in the truncated decision tree. Clearly, $p_{\mathrm{up}} \geq f$ and p_{up} has degree k. We have that $\mathbb{E}_{\mathcal{D}}[(p_{\mathrm{up}}(\mathbf{x}) - f(\mathbf{x}))^2]$ is upper bounded by a constant multiple of the probability that p_{up} takes the value 1, while $f(\mathbf{x})$ takes the value -1, since p_{up} is itself a Boolean-valued function (it is a decision tree). The probability that this happens is at most $s \cdot 2^{-k} = O(\epsilon)$ for $k = O(\log(s/\epsilon))$. We obtain p_{down} by a symmetric argument. □

For the following lemma, we make use of an upper bound for the pointwise distance between a Boolean formula and the best approximating low-degree polynomial from [OS03] (which readily implies the existence of low-degree $\mathcal{L}_2$ sandwiching polynomials).

Lemma 6.5 ($\mathcal{L}_2$ sandwiching degree of Boolean Formulas, modification of Theorem 6 in [OS03])**.** *Let $\mathcal{D}$ be the uniform distribution over the hypercube $\mathcal{X} = \{\pm 1\}^d$. For $s, \ell \in \mathbb{N}$, let $\mathcal{C}$ be the class of Boolean formulas of size at most s and depth at most ℓ. Then, for any $\epsilon > 0$ the $\mathcal{L}_2$ sandwiching degree of $\mathcal{C}$ is at most $k = (C \log(s/\epsilon))^{5\ell/2}\sqrt{s}$, for some sufficiently large universal constant $C > 0$.*

Proof. Let $f \in \mathcal{C}$ be an formula of size s and depth ℓ. We first construct a polynomial p that satisfies $|p(\mathbf{x}) - f(\mathbf{x})| \leq \sqrt{\epsilon}/2$ for any $\mathbf{x} \in \{\pm 1\}^d$. This corresponds to a slight modification of the proof of Theorem 6 in [OS03], where the basis of the inductive construction of p (see Lemma 10 in [OS03]) is an $O(\sqrt{\epsilon}/s^3)$ bound (instead of the original $1/s^3$ bound) for the (trivial) approximation of a single variable $\mathbf{x}_i$ by itself. The degree of p is indeed upper bounded by $(C \log(s/\epsilon))^{5\ell/2}\sqrt{s}$ and we may obtain $p_{\mathrm{up}}, p_{\mathrm{down}}$ by setting $p_{\mathrm{up}}(\mathbf{x}) = p(\mathbf{x}) + \sqrt{\epsilon}/2$ and $p_{\mathrm{down}} = p(\mathbf{x}) - \sqrt{\epsilon}/2$. Clearly, $p_{\mathrm{down}}(\mathbf{x}) \leq f(\mathbf{x}) \leq p_{\mathrm{up}}(\mathbf{x})$ and $|p_{\mathrm{up}}(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x})| = \sqrt{\epsilon}$ for all $\mathbf{x} \in \{\pm 1\}^d$. Therefore $\|p_{\mathrm{up}} - p_{\mathrm{down}}\|^2_{\mathcal{L}_2(\mathcal{D})} \leq \epsilon$. □

We obtain the following results for agnostic TDS learning of boolean concept classes.

Corollary 6.6 (TDS Learner for Decision Trees)**.** *Let $\mathcal{D}$ be the uniform distribution over the hypercube in d dimensions. Then, there is an algorithm that runs in time $d^{O(\log(s/\epsilon))}$ and TDS learns Decision Trees of size s with respect to $\mathrm{Unif}(\{\pm 1\}^d)$ up to error $32\lambda + O(\epsilon)$.*

Corollary 6.7 (TDS Learner for Boolean Formulas)**.** *Let $\mathcal{D}$ be the uniform distribution over the hypercube in d dimensions and $C > 0$ some sufficiently large universal constant. Then, there is an algorithm that runs in time $d^{\sqrt{s}(C \log(s/\epsilon))^{5\ell/2}}$ and TDS learns Boolean formulas of size at most s and depth at most ℓ with respect to $\mathrm{Unif}(\{\pm 1\}^d)$ up to error $32\lambda + O(\epsilon)$.*

Intersections and Decision Trees of Halfspaces. We now provide an upper bound for the $\mathcal{L}_2$-sandwiching degree of Decision Trees of halfspaces, which does not merely follow from a bound on the $\mathcal{L}_\infty$ approximate degree and, in particular, holds under both the Gaussian distribution and the Uniform over the hypercube. The following lemma is based on a powerful result from pseudorandomness literature (Theorem 10.4 from [GOWZ10]) which was originally used to provide a bound for the $\mathcal{L}_1$-sandwiching degree of decision trees of halfspaces, but, as we show, also provides a bound on the $\mathcal{L}_2$-sandwiching degree with careful manipulation.

Lemma 6.8 ($\mathcal{L}_2$-sandwiching degree of Intersections and Decision Trees of Halfspaces)**.** *Let $\mathcal{D}$ be either the uniform distribution over the hypercube $\mathcal{X} = \{\pm 1\}^d$ or the multivariate Gaussian distribution $\mathcal{N}(0, I_d)$ over $\mathcal{X} = \mathbb{R}^d$. For $\ell \in \mathbb{N}$, let also $\mathcal{C}$ be the class of concepts that can be expressed as an intersection of ℓ halfspaces on $\mathcal{X}$. Then, for any $\epsilon > 0$ the $\mathcal{L}_2$ sandwiching degree of $\mathcal{C}$ is at most $k = \widetilde{O}(\frac{\ell^6}{\epsilon^2})$. For Decision Trees of halfspaces of size s and depth ℓ, the bound is $k = \widetilde{O}(\frac{s^2\ell^6}{\epsilon^2})$.*

The above result implies the following corollary.

Corollary 6.9 (TDS Learner for Intersections and Decision Trees of Halfspaces)**.** *Let $\mathcal{D}$ be either the standard Gaussian in $\mathbb{R}^d$ or the uniform distribution over the hypercube in d dimensions. Then, there is an algorithm that runs in time $d^{\widetilde{O}(\ell^6/\epsilon^2)}$ and TDS learns intersections of ℓ halfspaces with respect to $\mathcal{D}$ up to error $32\lambda + O(\epsilon)$. For Decision Trees of halfspaces with size s and depth ℓ the bound is $d^{\widetilde{O}(s^2\ell^6/\epsilon^2)}$.*

In order to apply the structural result we need from [GOWZ10], we first provide a formal definition for the notion of hypercontractivity.

Definition 6.10 (Hypercontractivity)**.** Let $\mathcal{D}_1$ be a distribution over $\mathbb{R}$ and let $T \in \mathbb{N}, T > 2, \eta \in (0, 1)$. We say that $\mathcal{D}_1$ is $(T, 2, \eta)$-hypercontractive if $\mathbb{E}[x^T] < \infty$ and for any $a \in \mathbb{R}$ we have

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_1}[(a + \eta x)^T]^{1/T} \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_1}[(a + \eta x)^2]^{1/2}$$

The following result can be used to show Lemma 6.8.

Proposition 6.11 (Modification of Theorem 10.4 from [GOWZ10])**.** *Let $r \in \mathbb{N}$, $\sigma \in (0, 1)$, $T \in \mathbb{N}$, $\eta > 0$ and $t > 4$ be parameters and consider $\mathcal{D}$ to be a product distribution over $\mathcal{X} \subseteq \mathbb{R}^d$ such that each of its independent coordinates is $(4, 2, \eta)$-hypercontractive, and $(T, 2, 4/t)$-hypercontractive. Suppose that $T \geq Cr\log(rt)$ for some sufficiently large universal constant $C > 0$ and T is even. Then, for any function of the form $h : \mathcal{X} \to \mathbb{R}$, $h(\mathbf{x}) = \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \geq \tau\}$, where $\mathbf{w} \in \mathbb{R}^d$ and $\tau \in \mathbb{R}$, there is a polynomial $p : \mathcal{X} \to \mathbb{R}$ such that the following are true.*

(i) The degree of p is at most $k = \mathrm{poly}(\log t, \frac{1}{\eta}) \cdot \frac{1}{\sigma} + O(\frac{T}{r})$.

(ii) For any $\mathbf{x} \in \mathcal{X}$ we have $p(\mathbf{x}) \geq h(\mathbf{x})$.

(iii) The expected distance between p and h is bounded by $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[p(\mathbf{x}) - h(\mathbf{x})] \leq O(\sigma^{\frac{1}{2}} + \frac{rt\log(rt)}{T})$.

(iv) The values of p are upper bounded with high probability, i.e., $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[p(\mathbf{x}) > 1 + \frac{1}{r^2}] \leq 2^{-T/r}$.

(v) The $L_{2r}(\mathcal{D})$ norm of p is bounded by $\|p\|_{L_{2r}(\mathcal{D})} \leq 1 + \frac{2}{r^2}$.

Proof of Lemma 6.8. Let $f \in \mathcal{C}$ be an intersection of ℓ halfspaces over $\mathcal{X}$, i.e., f can be written in the following form

$$f(\mathbf{x}) = 2\prod_{j=1}^{\ell} h_j(\mathbf{x}) - 1, \text{ where } h_j(\mathbf{x}) = \mathbb{1}\{\mathbf{w}_j \cdot \mathbf{x} + \tau_j\} \text{ for some } \mathbf{w}_j \in \mathbb{R}^d, \tau_j \in \mathbb{R}$$

Note that if f is a Decision Tree of halfspaces of size s and depth ℓ, then f can be written as a sum of at most s intersections of ℓ halfspaces and it suffices to use accuracy parameter ϵ/s for each intersection.

Back to the case where f is an intersection of ℓ halfspaces, we will apply Proposition 6.11 in a way similar to the proof of Lemma 10.1 in [GOWZ10]. However, our goal here is to show that Proposition 6.11 implies the existence of $\mathcal{L}_2$ (rather than $\mathcal{L}_1$) sandwiching polynomials for f. We use the following standard fact about the Gaussian and Uniform distributions.

Claim (Hypercontractivity of Gaussian and Uniform marginals, see e.g. [KS88, Wol07, GOWZ10]). *If $\mathcal{D}$ is either the standard Gaussian $\mathcal{N}(0, I_d)$ over $\mathbb{R}^d$ or the uniform distribution over the hypercube $\{\pm 1\}^d$, then, for some universal constant $C > 0$, each of the coordinates of $\mathcal{D}$ is $(\lceil Ct^2 \rceil, 2, \frac{4}{t})$-hypercontractive for any $t > 0$ and, in particular, each one is also $(4, 2, \frac{1}{\sqrt{3}})$-hypercontractive.*

We may apply Proposition 6.11 for each h_j with parameters $r = 2\ell, \sigma = \frac{\epsilon^2}{C\ell^4}, t = C\frac{\ell^3}{\epsilon}\log(\ell/\epsilon)$, $\eta = 1/\sqrt{3}$ and $T = Ct^2$, for some sufficiently large universal constant C to obtain a polynomial p_j of degree $k = \widetilde{O}(\frac{\ell^5}{\epsilon^2})$ such that the following are true.

$$p_j(\mathbf{x}) \geq h_j(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{X} \tag{6.3}$$

$$\epsilon_1 := \mathop{\mathbb{E}}_{\mathcal{D}}[p_j(\mathbf{x}) - h_j(\mathbf{x})] = O\Big(\frac{\epsilon}{\ell^2}\Big) \tag{6.4}$$

$$\epsilon_2 := \mathop{\mathbb{P}}_{\mathcal{D}}\Big[p_j(\mathbf{x}) > 1 + \frac{1}{4\ell^2}\Big] \leq 2^{-\Omega(\frac{\ell^5}{\epsilon^2}\log^2(\ell/\epsilon))} \tag{6.5}$$

$$\|p_j\|_{L_{4m}(\mathcal{D})} \leq 1 + \frac{1}{2\ell^2} \tag{6.6}$$

We will now construct a polynomial p_{up} of degree $\widetilde{O}(\frac{\ell^6}{\epsilon^2})$ such that $p_{\mathrm{up}}(\mathbf{x}) \geq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ and also $\mathbb{E}_{\mathcal{D}}[(p_{\mathrm{up}}(\mathbf{x}) - f(\mathbf{x}))^2] \leq \epsilon/4$. This implies the existence of a corresponding polynomial p_{down} with $p_{\mathrm{down}}(\mathbf{x}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ and $\mathbb{E}_{\mathcal{D}}[(p_{\mathrm{up}}(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x}))^2] \leq \epsilon$ via a symmetric argument. Our proof consists of a hybrid argument similar to the one used in the proof of Lemma 10.1 in [GOWZ10], modified to provide a bound for the $\mathcal{L}_2$ error of approximation.

We pick $p_{\mathrm{up}} = 2p - 1$, where $p = \prod_{j=1}^{\ell} p_j$. Let $p^{(0)} = \prod_{j=1}^{\ell} h_j$, $p^{(i)} = (\prod_{j=1}^{i} p_j)(\prod_{j=i+1}^{\ell} h_j)$ and $p^{(\ell)} = p$. We then have the following.

$$\begin{aligned}
\|p - h\|_{\mathcal{L}_2(\mathcal{D})} = \|p^{(\ell)} - p^{(0)}\|_{\mathcal{L}_2(\mathcal{D})} &\leq \sum_{i=1}^{\ell} \|p^{(i)} - p^{(i-1)}\|_{\mathcal{L}_2(\mathcal{D})} \\
&= \sum_{i=1}^{\ell} \Big\| \Big(\prod_{j=1}^{i-1} p_j\Big)\Big(\prod_{j=i+1}^{\ell} h_j\Big)(p_i - h_i)\Big\|_{\mathcal{L}_2(\mathcal{D})} \\
&\leq \sum_{i=1}^{\ell} \Big\| \Big(\prod_{j\neq i} p_j\Big)(p_i - h_i)\Big\|_{\mathcal{L}_2(\mathcal{D})} \qquad \text{(by property (6.3))}
\end{aligned}$$

For any fixed $i \in [\ell]$ we have

$$\begin{aligned}
\Big\| \Big(\prod_{j\neq i} p_j\Big)(p_i - h_i)\Big\|^2_{\mathcal{L}_2(\mathcal{D})} &= \mathop{\mathbb{E}}_{\mathcal{D}}\Big[\Big(\prod_{j\neq i} p_j^2(\mathbf{x})\Big)(p_i(\mathbf{x}) - h_i(\mathbf{x}))^2\Big] \\
&\leq \mathop{\mathbb{E}}_{\mathcal{D}}\Big[\Big(\prod_{j\neq i} p_j^2(\mathbf{x})\Big)(p_i(\mathbf{x}) - h_i(\mathbf{x}))p_i(\mathbf{x})\Big] \qquad \text{(since } h_i \geq 0 \text{ and } p_i \geq h_i\text{)}
\end{aligned}$$

In order to bound the quantity $\mathbb{E}_{\mathcal{D}}[(\prod_{j\neq i} p_j^2)(p_i - h_i)p_i]$, we split the expectation according to the event $\mathcal{E}$ that $(\prod_{j\neq i} p_j)\sqrt{p_i} < 2$. In particular, we have that $\mathbb{E}_{\mathcal{D}}[(\prod_{j\neq i} p_j^2)(p_i - h_i)p_i\, \mathbb{1}\{\mathcal{E}\}]$ is at most $4\epsilon_1$ by property (6.4) and $\mathbb{E}_{\mathcal{D}}[(\prod_{j\neq i} p_j^2)(p_i - h_i)p_i\, \mathbb{1}\{\neg\mathcal{E}\}]$ is bounded as follows.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{D}}\Big[\Big(\prod_{j\neq i} p_j^2(\mathbf{x})\Big)(p_i(\mathbf{x}) - h_i(\mathbf{x}))p_i(\mathbf{x})\, \mathbb{1}\Big\{\Big(\prod_{j\neq i} p_j(\mathbf{x})\Big)\sqrt{p_i(\mathbf{x})} \geq 2\Big\}\Big] \leq \\
&\qquad \leq \mathop{\mathbb{E}}_{\mathcal{D}}\Big[\Big(\prod_{j\in[\ell]} p_j^2(\mathbf{x})\Big)\, \mathbb{1}\Big\{\Big(\prod_{j\neq i} p_j(\mathbf{x})\Big)\sqrt{p_i(\mathbf{x})} \geq 2\Big\}\Big] \qquad \text{(by property (6.3))}
\end{aligned}$$

We now observe that whenever $(\prod_{j\neq i} p_j(\mathbf{x}))\sqrt{p_i(\mathbf{x})} \geq 2$, there must exist some index j' such that $p_{j'}(\mathbf{x}) > 1 + \frac{1}{4\ell^2}$ and, therefore, we can further bound the above quantity by the following one.

$$\mathop{\mathbb{E}}_{\mathcal{D}}\Big[\sum_{j'=1}^{\ell} \mathbb{1}\Big\{p_{j'}(\mathbf{x}) > 1 + \frac{1}{4\ell^2}\Big\}\Big(\prod_{j\in[\ell]} p_j^2(\mathbf{x})\Big)\Big] = \sum_{j'=1}^{\ell} \mathop{\mathbb{E}}_{\mathcal{D}}\Big[\mathbb{1}\Big\{p_{j'}(\mathbf{x}) > 1 + \frac{1}{4\ell^2}\Big\}\Big(\prod_{j\in[\ell]} p_j^2(\mathbf{x})\Big)\Big]$$

In the above expression we used linearity of expectation. We now apply Hölder's inequality and obtain the bound $\sum_{j'=1}^{\ell}(\mathbb{P}_{\mathcal{D}}[p_{j'}(\mathbf{x}) > 1 + \frac{1}{4\ell^2}])^{\frac{1}{2}} \prod_{j=1}^{\ell}\big(\mathbb{E}_{\mathcal{D}}[p_j^{4\ell}(\mathbf{x})]\big)^{\frac{1}{2\ell}}$. Due to properties (6.5) and (6.6), we finally have the bound $\ell\sqrt{\epsilon_2}\cdot\prod_{j=1}^{\ell}\|p_j\|_{L_{4\ell}}^2 \leq \ell\sqrt{\epsilon_2}(1+\frac{1}{2\ell^2})^{2\ell} \leq 3\ell\sqrt{\epsilon_2}$. Therefore, in total, we have $\|p-h\|_{L_2(\mathcal{D})}^2 \leq 4\ell^2\epsilon_1 + 3\ell^3\epsilon_2 \leq \epsilon$, which implies that $\|p_{\mathrm{up}} - f\|_{\mathcal{L}_2(\mathcal{D})} \leq \epsilon$ and $p_{\mathrm{up}} \geq f$. □

7 Lower Bounds

7.1 Lower Bound for Realizable TDS Learning of Monotone Functions

We now prove Theorem 2.10, which we restate here for convenience.

Theorem 7.1 (Hardness of TDS Learning Monotone Functions)**.** *Let the accuracy parameter ϵ be at most 0.1 and the success probability parameter δ also be at most 0.1. Then, in the realizable setting, any TDS learning algorithm for the class of monotone functions over $\{\pm1\}^d$ with accuracy parameter ϵ and success probability at least $1-\delta$ requires either $2^{0.04d}$ training samples or $2^{0.04d}$ testing samples for all sufficiently large values of d.*

We will need the following standard fact, see for example [RV23] for a proof:

Fact 7.2. *For any distribution D over any domain, let multisets T_1 and T_2 be sampled as follows:*

1. *Set T_1 is N i.i.d. samples from D.*
2. *First, multiset S is formed by taking M i.i.d. samples from D. Then, multiset T_2 is formed by taking N i.i.d. uniform elements from S.*

Then, the statistical distance between the distributions of T_1 and T_2 is at most $\frac{N^2}{M}$.

Now, we prove Theorem 2.10.

Proof of Theorem 2.10. We fix $\delta \leq 0.1$ and also fix $\epsilon \leq 0.1$. Let $\mathcal{A}$ be an algorithm that takes $N \leq 2^{0.04d}$ testing samples and $N \leq 2^{0.04d}$ training samples, and either outputs REJECT, or (ACCEPT, $\widehat{f}$) for a function $\widehat{f}:\{\pm1\}^d \to \{\pm1\}$. We argue that for, a sufficiently large d, the algorithm $\mathcal{A}$ will fail to be a TDS-learning algorithm for monotone functions over $\{\pm1\}^d$.

Let f be some function mapping $\{\pm1\}^d \to \{\pm1\}$ and let a multiset S consist of elements in $\{\pm1\}^d$. We define $\mathcal{T}(f,S)$ to be a random variable supported on $\{\text{Yes}, \text{No}\}$ determined as follows (informally, if $\mathcal{A}$ is a TDS-learner for monotone functions, then $\mathcal{T}(f,S)$ will allow us to distinguish a uniform distribution over S from the uniform distribution over $\{\pm1\}^d$):

1. Let $S_{\text{train}} \subset \{\pm1\}^d \times \{\pm1\}$ consist of N pairs $(\mathbf{x}, f(\mathbf{x}))$, where $\mathbf{x}$ are drawn i.i.d. uniformly from $\{\pm1\}^d$.
2. Let X_{test} consist of N i.i.d. uniform samples from set S.
3. The algorithm $\mathcal{A}$ is run on $(S_{\text{train}}, X_{\text{test}})$.

4. If $\mathcal{A}$ outputs REJECT, then output $\mathcal{T}(f, S)$ =No.

5. If $\mathcal{A}$ outputs (ACCEPT, $\widehat{f}$), then

   (a) Obtain a new set X_2 of 10000 i.i.d. uniform samples from S.

   (b) If, on the majority of points $\mathbf{x}$ in X_2, we have $\widehat{f}(\mathbf{x}) = 1$, then output No.

   (c) Otherwise, output Yes.

For a multiset S consisting of elements in $\{\pm 1\}^d$, let f_S be the monotone function defined as follows:

$$f_S(\mathbf{x}) := \begin{cases} +1 & \text{if there exists } \mathbf{z} \in S : \ \mathbf{x} \succeq \mathbf{z}, \\ -1 & \text{otherwise}. \end{cases}$$

First, we observe that if $\mathcal{A}$ is indeed a (ϵ, δ)-TDS learning algorithm for monotone functions over $\{\pm 1\}^d$, then:

- $\mathcal{T}(-1, \{\pm 1\}^d)$=Yes with probability at least $\frac{2}{3}$ (from here on, by -1 we mean the function that maps every element in $\{\pm 1\}^d$ into -1). This is true because, by the definition of a TDS learner, since S_{train} comes from the uniform distribution over $\{\pm 1\}^d$, with probability at least $1 - 2\delta = 0.8$ the algorithm $\mathcal{A}$ will output (ACCEPT, $\widehat{f}$) for some $\widehat{f}$ satisfying $\mathbb{P}_{\mathbf{x} \sim \{\pm 1\}^d}[\widehat{f}(x) \neq -1] \leq \epsilon = 0.1$. Then, via a standard Hoeffding bound, with probability at least 0.9 on the majority of elements $\mathbf{x}$ in X_2 we have $\widehat{f}(\mathbf{x}) = -1$ and then $\mathcal{T}(-1, \{\pm 1\}^d)$=Yes.

- For any multiset S with elements in $\{\pm 1\}^d$, we have $\mathcal{T}(f_S, S)$ =No with probability at least $\frac{2}{3}$. Indeed, from the definition of a TDS learning algorithm, we see that, with probability at least $1 - \delta = 0.9$, the algorithm $\mathcal{A}$ will either output

  – Output reject, in which case $\mathcal{T}(f_S, S)$ =No.

  – Output (ACCEPT, $\widehat{f}$) with $\mathbb{P}_{\mathbf{x} \sim S}[\widehat{f}(\mathbf{x}) \neq f_S(\mathbf{x})] \leq \epsilon = 0.1$. But we know that f_S takes values $+1$ on all elements in S. Therefore, $\mathbb{P}_{\mathbf{x} \sim S}[\widehat{f}(\mathbf{x}) \neq f_S(\mathbf{x})] \leq 0.1$. Then, via a standard Hoeffding bound, with probability at least 0.9 on the majority of elements $\mathbf{x}$ in X_2 we have $\widehat{f}(\mathbf{x}) = +1$ and then $\mathcal{T}(f_S, S)$=No.

In particular, if S is obtained by picking $M = 2^{0.1d}$ i.i.d. elements from $\{\pm 1\}^d$, we have

$$\left| \mathop{\mathbb{P}}_{\substack{S \sim \text{Unif}(\{\pm 1\}^d)^{\otimes M} \\ \text{Randomness of } \mathcal{T}}} [\mathcal{T}(f_S, S) = \text{Yes}] - \mathop{\mathbb{P}}_{\text{Randomness of } \mathcal{T}} [\mathcal{T}(-1, \{\pm 1\}^d) = \text{Yes}] \right| > \frac{1}{3}. \tag{7.1}$$

The rest of the proof argues, via a hybrid argument, that this is impossible. To be specific, we claim that for sufficiently large d the following two inequalities must hold

$$\left| \mathop{\mathbb{P}}_{\substack{S \sim \text{Unif}(\{\pm 1\}^d)^{\otimes M} \\ \text{Randomness of } \mathcal{T}}} [\mathcal{T}(-1, S) = \text{Yes}] - \mathop{\mathbb{P}}_{\text{Randomness of } \mathcal{T}} [\mathcal{T}(-1, \{\pm 1\}^d) = \text{Yes}] \right| \leq \frac{N^2}{M}. \tag{7.2}$$

$$\left| \mathop{\mathbb{P}}_{\substack{S \sim \text{Unif}(\{\pm 1\}^d)^{\otimes M} \\ \text{Randomness of } \mathcal{T}}} [\mathcal{T}(f_S, S) = \text{Yes}] - \mathop{\mathbb{P}}_{\substack{S \sim \text{Unif}(\{\pm 1\}^d)^{\otimes M} \\ \text{Randomness of } \mathcal{T}}} [\mathcal{T}(-1, S) = \text{Yes}] \right| \leq 2 \left(\frac{3}{4} \right)^d MN. \tag{7.3}$$

We observe that Equation 7.2 follows immediately from Fact 7.2, because if Equation 7.2 didn't hold, then we would be able to achieve advantage greater than $\frac{M}{N^2}$ when distinguishing N i.i.d. uniform samples from $\{\pm 1\}^d$ from N i.i.d. uniform examples from S.

Now we prove Equation 7.3. Let $S_{\text{train}}^{\mathcal{T}(f_S,S)}$ denote the collection of pairs $\{(\mathbf{x}, f_S(\mathbf{x}))\}$ sampled in Step 1 of $\mathcal{T}(f_S, S)$. Analogously, let $S_{\text{train}}^{\mathcal{T}(-1,S)}$ denote the collection of pairs $(\mathbf{x}, -1)$ in set used in procedure $\mathcal{T}(-1, S)$. In either case, the elements in $S_{\text{train}}^{\mathcal{T}(f_S,S)}$ and $S_{\text{train}}^{\mathcal{T}(-1,S)}$ are i.i.d. uniformly random elements in $\{\pm 1\}^d$. Let $E^{\mathcal{T}(-1,S)}$ be the event, over the choice of S and the choice of $S_{\text{train}}^{\mathcal{T}(-1,S)}$, that for every $(\mathbf{x}, -1) \in S_{\text{train}}^{\mathcal{T}(-1,S)}$ there is no $\mathbf{z}$ in S satisfying $\mathbf{x} \succeq \mathbf{z}$. Analogously, let $E^{\mathcal{T}(f_S,S)}$ be the event, over the choice of S and the choice of $S_{\text{train}}^{\mathcal{T}(f_S,S)}$, that for every $(\mathbf{x}, f_S(\mathbf{x})) \in S_{\text{train}}^{\mathcal{T}(f_S,S)}$ there is no $\mathbf{z}$ in S satisfying $\mathbf{x} \succeq \mathbf{z}$. We observe that

$$\mathop{\mathbb{P}}_{\substack{S\sim\text{Unif}(\{\pm1\}^d)^{\otimes M}\\ \text{Randomness of } \mathcal{T}}}\left[\mathcal{T}(f_S, S) = \text{Yes} \middle| E^{\mathcal{T}(f_S,S)}\right] = \mathop{\mathbb{P}}_{\substack{S\sim\text{Unif}(\{\pm1\}^d)^{\otimes M}\\ \text{Randomness of } \mathcal{T}}}\left[\mathcal{T}(-1, S) = \text{Yes} \middle| E^{\mathcal{T}(-1,S)}\right] \tag{7.4}$$

which is true because, subject to $E^{\mathcal{T}(f_S,S)}$ or $E^{\mathcal{T}(-1,S)}$, the function f_S takes values of -1 on every element $\mathbf{x}$ in $S_{\text{train}}^{\mathcal{T}(f_S,S)}$ and $S_{\text{train}}^{\mathcal{T}(-1,S)}$ respectively. We also see that the random variables $(S, S_{\text{train}}^{\mathcal{T}(f_S,S)})$ and $(S, S_{\text{train}}^{\mathcal{T}(-1,S)})$ are identically distributed (conditioned on $E^{\mathcal{T}(f_S,S)}$ and $E^{\mathcal{T}(-1,S)}$ respectively). We also observe that

$$\mathop{\mathbb{P}}_{\substack{S\sim\text{Unif}(\{\pm1\}^d)^{\otimes M}\\ \text{Randomness of } \mathcal{T}}}\left[E^{\mathcal{T}(f_S,S)}\right] = \mathop{\mathbb{P}}_{\substack{S\sim\text{Unif}(\{\pm1\}^d)^{\otimes M}\\ \text{Randomness of } \mathcal{T}}}\left[E^{\mathcal{T}(-1,S)}\right] \leq \left(\frac{3}{4}\right)^d MN, \tag{7.5}$$

where the equality of the two probabilities follows immediately by definition, and the upper bound of $\left(\frac{3}{4}\right)^d MN$ is true for the following reason. Let $\mathbf{z}$ and $\mathbf{x}$ be a pair of i.i.d. uniformly random elements in $\{\pm 1\}^d$, then $\mathbb{P}[\mathbf{x} \succeq \mathbf{z}] = \left(\frac{3}{4}\right)^d$ as each bit of $\mathbf{x}$ and $\mathbf{z}$ are independent and for each of the bits we have $\mathbf{x} \geq \mathbf{z}$ with probability exactly $3/4$. Now, taking a union bound over every $(\mathbf{x}, -1) \in S_{\text{train}}^{\mathcal{T}(-1,S)}$ and $\mathbf{z} \in S$, we obtain the bound in Equation 7.5.

Overall, combining Equation 7.2 with Equation 7.3 and substituting $N \leq 2^{0.04d}$ and $M = 2^{0.1d}$ we get

$$\left|\mathop{\mathbb{P}}_{\substack{S\sim\text{Unif}(\{\pm1\}^d)^{\otimes M}\\ \text{Randomness of } \mathcal{T}}}[\mathcal{T}(f_S, S) = \text{Yes}] - \mathop{\mathbb{P}}_{\text{Randomness of } \mathcal{T}}[\mathcal{T}(-1, \{\pm 1\}^d) = \text{Yes}]\right| \leq \frac{N^2}{M} + 2\left(\frac{3}{4}\right)^d MN = 2^{-\Omega(d)},$$

which is in contradiction with Equation 7.1 for a sufficiently large value of d. This proves that $\mathcal{A}$ is not a (ϵ, δ)-TDS learning algorithm for monotone functions. □

7.2 Lower Bound for Realizable TDS Learning of Convex Sets

We now prove Theorem 2.11 which we restate here for convenience.

Theorem 7.3 (Hardness of TDS Learning Convex Sets)**.** *Let the accuracy parameter ϵ be at most 0.1 and the success probability parameter δ also be at most 0.1. Then, in the realizable setting, any TDS learning algorithm for the class of indicators of convex sets under the standard Gaussian distribution on $\mathbb{R}^d$ requires either $2^{0.04d}$ training samples or $2^{0.04d}$ testing samples for all sufficiently large values of d.*

We will need the following standard facts about Gaussian distributions:

Fact 7.4 (Concentration of Gaussian norm, see e.g. Lemma 8.1 in [BM97])**.** *For any $\eta > 0$ it is the case that*

$$\mathbb{P}_{\mathbf{x}\in\mathcal{N}(0,I_d)}\left[d - 2\sqrt{d\ln\left(\frac{2}{\eta}\right)} \le \|\mathbf{x}\|_2^2 \le d + 2\sqrt{d\ln\left(\frac{2}{\eta}\right)} + 2\ln\left(\frac{2}{\eta}\right)\right] \ge 1-\eta$$

Fact 7.5 (Concentration of Gaussian norm. See e.g. [RV23].)**.** *For any $r > 0$ it is the case that*

$$\mathbb{P}_{\mathbf{x}^1,\mathbf{x}^2\in\mathcal{N}(0,I_d)}\left[\|\mathbf{x}^1 - \mathbf{x}^2\|_2 \le r\right] \le \left(\frac{64r^2}{d}\right)^{d/2}$$

Recall that we use $\mathcal{B}_a$ to denote the origin-centered closed ball in $\mathbb{R}^d$ of radius a. Using $\mathrm{conv}(\cdot)$ to denote the convex hull of a set of points, will state the following geometric observation of [RV23] about convex hulls of a collection of point.

Fact 7.6 ([RV23])**.** *For any $a > 0$, let $\{\mathbf{x}^i\}_{i=1}^M$ be a collection of points in $\mathcal{B}_b \setminus \mathcal{B}_a$. If for every pair of points $(\mathbf{x}^i, \mathbf{x}^j)$ the $\|\mathbf{x}^i - \mathbf{x}^j\|_2$ is greater than $2\sqrt{b^2 - a^2}$, then for every i and j we have*

$$\mathrm{conv}(\mathbf{x}^i, \mathcal{B}_a) \cap \mathrm{conv}(\mathbf{x}^i, \mathcal{B}_a) = \mathcal{B}_a$$

and also

$$\mathrm{conv}(\mathbf{x}^1, \cdots, \mathbf{x}^M, \mathcal{B}_a) = \cup_i \mathrm{conv}(\mathbf{x}^i, \mathcal{B}_a).$$

For the rest of the section we will set

$$a = \sqrt{d - 2\sqrt{d\ln\left(\frac{1}{50}\right)}} \qquad b = \sqrt{d + 2\sqrt{d\ln\left(\frac{1}{50}\right)} + 2\ln\left(\frac{1}{50}\right)}, \tag{7.6}$$

and from Fact 7.4 we see that the norm a standard Gaussian vector in $\mathbb{R}^d$ falls in interval (a, b) with probability at least 0.99.

Now, we are ready to prove Theorem 2.11.

Proof of Theorem 2.10. We fix $\delta \le 0.1$ and also fix $\epsilon \le 0.1$. Let $\mathcal{A}$ be an algorithm that takes $N \le 2^{0.04d}$ testing samples and $N \le 2^{0.04d}$ training samples, and either outputs REJECT, or (ACCEPT, $\widehat{f}$) for a function $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$. We argue that for, a sufficiently large d, the algorithm $\mathcal{A}$ will fail to be a TDS-learning algorithm for convex sets under the Gaussian distribution on $\mathbb{R}^d$.

For a set S we will define g_S as the indicator of the convex set $\mathrm{conv}(S \cap (\mathcal{B}_b \setminus \mathcal{B}_a), \mathcal{B}_a)$. And in this section we denote the uniform distribution over S as $\mathbb{U}_S$.

Let f be some function mapping $\mathbb{R}^d \to \{\pm 1\}$ and let a set D be a distribution over $\mathbb{R}^d$. We define $\mathcal{H}(f, D)$ to be a random variable supported on $\{\text{Yes}, \text{No}\}$ determined as follows (informally, if $\mathcal{A}$ is a TDS-learner for convex sets, then $\mathcal{H}(f, D)$ will allow us to distinguish D from the Gaussian distribution over $\mathbb{R}^d$):

1. Let $S_{\text{train}} \subset \mathbb{R}^d \times \{\pm 1\}$ consist of N pairs $(\mathbf{x}, f(\mathbf{x}))$, where $\mathbf{x}$ are drawn i.i.d. from $\mathcal{N}(0, I_d)$.
2. Let X_{test} consist of N i.i.d. uniform samples from D.
3. The algorithm $\mathcal{A}$ is run on $(S_{\text{train}}, X_{\text{test}})$.
4. If $\mathcal{A}$ outputs REJECT, then output $\mathcal{H}(f, S) =$No.

5. If $\mathcal{A}$ outputs (ACCEPT, $\widehat{f}$), then
   (a) Obtain a new set X_2 of 10000 i.i.d. samples from D.
   (b) If, on the majority of points $\mathbf{x}$ in X_2, we have $\widehat{f}(\mathbf{x}) = -1$, then output No.
   (c) Otherwise, output Yes.

First, we observe that if $\mathcal{A}$ is indeed a (ϵ, δ)-TDS learning algorithm for convex sets over $\mathbb{R}^d$ under $\mathcal{N}(0, I_d)$, then:

- $\mathcal{H}(g_\emptyset, \mathcal{N}(0, I_d))$=Yes with probability at least $\frac{2}{3}$ (from here on, by -1 we mean the function that maps every element in $\{\pm 1\}^d$ into -1). This is true because, by the definition of a TDS learner, since S_{train} comes from the uniform distribution over $\mathcal{N}(0, I_d)$, with probability at least $1 - 2\delta = 0.8$ the algorithm $\mathcal{A}$ will output (ACCEPT, $\widehat{f}$) for some $\widehat{f}$ satisfying $\mathbb{P}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}[\widehat{f}(\mathbf{x}) \neq g_\emptyset(\mathbf{x})] \leq \epsilon = 0.1$. Since a was chosen is such manner that $\Pr_{\mathbf{x} \in \mathcal{N}(0, I_d)}[\mathbf{x} \in \mathcal{B}_a] < 0.01$, and $g_\emptyset$ is the indicator function of $\mathcal{B}_a$, we have $\Pr_{\mathbf{x} \in \mathcal{N}(0, I_d)}[g_\emptyset(\mathbf{x}) \neq -1] < 0.01$. Via a union bound, we see that $\mathbb{P}_{\mathbf{x} \sim \mathcal{N}(0, I_d)}[\widehat{f}(\mathbf{x}) \neq -1] \leq 0.11$. Then, via a standard Hoeffding bound, with probability at least 0.9 on the majority of elements $\mathbf{x}$ in X_2 we have $\widehat{f}(\mathbf{x}) = -1$ and then $\mathcal{H}(g_\emptyset, \mathcal{N}(0, I_d))$=Yes.
- For any set S with elements in $\mathbb{R}^d$, we have $\mathcal{H}(g_S, \mathbb{U}_S)$ =No with probability at least $\frac{2}{3}$. Indeed, from the definition of a TDS learning algorithm, we see that, with probability at least $1 - \delta = 0.9$, the algorithm $\mathcal{A}$ will either
  - Output reject, in which case $\mathcal{H}(g_S, \mathbb{U}_S)$ =No.
  - Output (ACCEPT, $\widehat{f}$) with $\mathbb{P}_{\mathbf{x} \sim \mathbb{U}_S}[\widehat{f}(\mathbf{x}) \neq g_S(\mathbf{x})] \leq \epsilon = 0.1$. But we know that g_S takes values $+1$ on all elements in S. Therefore, $\mathbb{P}_{\mathbf{x} \sim \mathbb{U}_S}[\widehat{f}(\mathbf{x}) \neq f_S(\mathbf{x})] \leq 0.1$. Then, via a standard Hoeffding bound, with probability at least 0.9 on the majority of elements $\mathbf{x}$ in X_2 we have $\widehat{f}(\mathbf{x}) = +1$ and then $\mathcal{H}(g_S, \mathbb{U}_S)$=No.

In particular, if S is obtained by picking $M = 2^{0.1d}$ i.i.d. elements from $\mathcal{N}(0, I_d)$, we have

$$\left| \mathop{\mathbb{P}}_{\substack{S \sim \mathcal{N}(0, I_d)^{\otimes M} \\ \text{Randomness of } \mathcal{H}}} [\mathcal{H}(g_S, \mathbb{U}_S) = \text{Yes}] - \mathop{\mathbb{P}}_{\text{Randomness of } \mathcal{H}} [\mathcal{H}(g_\emptyset, \mathcal{N}(0, I_d)) = \text{Yes}] \right| > \frac{1}{3}. \tag{7.7}$$

The rest of the proof argues, via a hybrid argument, that this is impossible. To be specific, we claim that for sufficiently large d the following two inequalities must hold

$$\left| \mathop{\mathbb{P}}_{\substack{S \sim \mathcal{N}(0, I_d)^{\otimes M} \\ \text{Randomness of } \mathcal{H}}} [\mathcal{H}(g_\emptyset, \mathbb{U}_S) = \text{Yes}] - \mathop{\mathbb{P}}_{\text{Randomness of } \mathcal{H}} [\mathcal{H}(g_\emptyset, \mathcal{N}(0, I_d)) = \text{Yes}] \right| \leq \frac{N^2}{M}. \tag{7.8}$$

$$\begin{aligned} &\left| \mathop{\mathbb{P}}_{\substack{S \sim \mathcal{N}(0, I_d)^{\otimes M} \\ \text{Randomness of } \mathcal{H}}} [\mathcal{H}(g_S, \mathbb{U}_S) = \text{Yes}] - \mathop{\mathbb{P}}_{\substack{S \sim \mathcal{N}(0, I_d)^{\otimes M} \\ \text{Randomness of } \mathcal{H}}} [\mathcal{H}(g_\emptyset, \mathbb{U}_S) = \text{Yes}] \right| \\ &\qquad \leq \left(\frac{64(b^2 - a^2)}{d} \right)^{d/2} (M + N)^2. \end{aligned} \tag{7.9}$$

We observe that Equation 7.8 follows immediately from Fact 7.2, because if Equation 7.8 didn't hold, then we would be able to achieve advantage greater than $\frac{M}{N^2}$ when distinguishing N i.i.d. uniform samples from $\mathcal{N}(0, I_d)$ and N i.i.d. uniform examples from S.

Now we prove Equation 7.9. Let $S^{\mathcal{H}(g_S,\mathbb{U}_S)}_{\text{train}}$ denote the collection of pairs $\{(\mathbf{x}, g_S(\mathbf{x}))\}$ sampled in Step 1 of $\mathcal{H}(g_S, \mathbb{U}_S)$. Analogously, let $S^{\mathcal{H}(g_\emptyset,\mathbb{U}_S)}_{\text{train}}$ denote the collection of pairs $(\mathbf{x}, -1)$ in set used in procedure $\mathcal{H}(g_\emptyset, \mathbb{U}_S)$. In either case, the elements in $S^{\mathcal{H}(g_S,\mathbb{U}_S)}_{\text{train}}$ and $S^{\mathcal{H}(g_\emptyset,\mathbb{U}_S)}_{\text{train}}$ are i.i.d. elements from $\mathcal{N}(0, I_d)$. Let $\mathcal{E}^{\mathcal{H}(g_S,\mathbb{U}_S)}$ be the event, over the choice of S and the choice of $S^{\mathcal{H}(g_S,\mathbb{U}_S)}_{\text{train}}$, that for each pair of points $\mathbf{x}^1$ and $\mathbf{x}^2$ in $S \cup \{\mathbf{x} : (\mathbf{x}, g_S(x)) \in S^{\mathcal{H}(g_S,\mathbb{U}_S)}_{\text{train}}\}$ we have $\|\mathbf{x}^1 - \mathbf{x}^2\|_2 > 2\sqrt{b^2 - a^2}$. Analogously, let $\mathcal{E}^{\mathcal{H}(g_S,\mathbb{U}_S)}$ be the event, over the choice of S and the choice of $S^{\mathcal{H}(g_\emptyset,\mathbb{U}_S)}_{\text{train}}$, that for each pair of points $\mathbf{x}^1$ and $\mathbf{x}^2$ in $S \cup \{\mathbf{x} : (\mathbf{x}, g_\emptyset(x)) \in S^{\mathcal{H}(g_\emptyset,\mathbb{U}_S)}_{\text{train}}\}$ we have $\|\mathbf{x}^1 - \mathbf{x}^2\|_2 > 2\sqrt{b^2 - a^2}$.

We first observe that subject to $\mathcal{E}^{\mathcal{H}(g_\emptyset,\mathbb{U}_S)}$ it is the case that for every $\{(\mathbf{x}, g_S(\mathbf{x}))\}$ in $S^{\mathcal{H}(g_S,\mathbb{U}_S)}_{\text{train}}$ it is the case that $g_S = g_\emptyset(x)$. For $\mathbf{x} \in \mathcal{B}_a \cup (\mathbb{R} \setminus \mathcal{B}_b)$ this is immediate because g_S as the indicator of the convex set $\text{conv}(S \cap (\mathcal{B}_b \setminus \mathcal{B}_a), \mathcal{B}_a)$. It remains to show this only for points $(\mathbf{x}, g_S(\mathbf{x})) \in S^{\mathcal{H}(g_S,\mathbb{U}_S)}_{\text{train}}$ that also satisfy $\mathbf{x} \in \mathcal{B}_b \setminus \mathcal{B}_a$. Since $\mathbf{x}$ is outside $\mathcal{B}_a$, we have $g_\emptyset(\mathbf{x}) = -1$ and therefore we would like to show that $g_S(\mathbf{x})$ also equals to -1. This is true because from Fact 7.6 it is the case that if $\mathcal{E}^{\mathcal{H}(g_\emptyset,\mathbb{U}_S)}$ takes place, then for every such $\mathbf{x}$ we have

$$\text{conv}(\mathbf{x}, \mathcal{B}_a) \cap \text{conv}(S \cap (\mathcal{B}_b \setminus \mathcal{B}_a), \mathcal{B}_a) = \text{conv}(\mathbf{x}, \mathcal{B}_a) \cap \left(\bigcup_{\mathbf{z} \in S \cap (\mathcal{B}_b \setminus \mathcal{B}_a)} \text{conv}(\mathbf{z} \cap (\mathcal{B}_b \setminus \mathcal{B}_a), \mathcal{B}_a) \right) =$$
$$\bigcup_{\mathbf{z} \in S \cap (\mathcal{B}_b \setminus \mathcal{B}_a)} (\text{conv}(\mathbf{x}, \mathcal{B}_a) \cap (\text{conv}(\mathbf{z} \cap (\mathcal{B}_b \setminus \mathcal{B}_a), \mathcal{B}_a))) = \mathcal{B}_a,$$

which in particular implies that $\mathbf{x}$ is not in the convex hull $\text{conv}(S \cap (\mathcal{B}_b \setminus \mathcal{B}_a), \mathcal{B}_a)$ and $g_S(\mathbf{x}) = -1$, concluding the proof of our observation.

We therefore conclude that distributions of $(S, S^{\mathcal{H}(g_S,\mathbb{U}_S)}_{\text{train}})$ and $(S, S^{\mathcal{H}(\mathcal{H}(g_\emptyset,\mathbb{U}_S)}_{\text{train}})$ are identically distributed conditioned on $\mathcal{E}^{\mathcal{H}(g_S,\mathbb{U}_S)}$ and $\mathcal{E}^{\mathcal{H}(g_\emptyset,\mathbb{U}_S)}$ respectively, which implies that

$$\mathop{\mathbb{P}}_{\substack{S \sim \mathcal{N}(0,I_d)^{\otimes M} \\ \text{Randomness of } \mathcal{H}}} \left[\mathcal{H}(g_S, \mathbb{U}_S) = \text{Yes} \middle| \mathcal{E}^{\mathcal{H}(g_S,\mathbb{U}_S)} \right] = \mathop{\mathbb{P}}_{\substack{S \sim \mathcal{N}(0,I_d)^{\otimes M} \\ \text{Randomness of } \mathcal{H}}} \left[\mathcal{H}(g_\emptyset, \mathbb{U}_S) = \text{Yes} \middle| \mathcal{E}^{\mathcal{H}(g_\emptyset,\mathbb{U}_S)} \right], \tag{7.10}$$

We also observe that

$$\mathop{\mathbb{P}}_{\substack{S \sim \mathcal{N}(0,I_d)^{\otimes M} \\ \text{Randomness of } \mathcal{H}}} \left[\mathcal{E}^{\mathcal{H}(g_S,\mathbb{U}_S)} \right] = \mathop{\mathbb{P}}_{\substack{S \sim \mathcal{N}(0,I_d)^{\otimes M} \\ \text{Randomness of } \mathcal{H}}} \left[\mathcal{E}^{\mathcal{H}(g_\emptyset,\mathbb{U}_S)} \right] \leq \left(\frac{64(b^2 - a^2)}{d} \right)^{d/2} (M + N)^2, \tag{7.11}$$

where the equality of the two probabilities follows immediately by definition, and the upper bound of $\left(\frac{64(b^2-a^2)}{d}\right)^{d/2} (M + N)^2$ is true by applying Fact 7.5 to each relevant pair of points. Therefore, we obtain the bound in Equation 7.11.

Overall, combining Equation 7.8 with Equation 7.9 and substituting $N \leq 2^{0.04d}$, $M = 2^{0.1d}$ as well as $a = \sqrt{d - 2\sqrt{d \ln\left(\frac{1}{50}\right)}}$ and $b = \sqrt{d + 2\sqrt{d \ln\left(\frac{1}{50}\right)} + 2\ln\left(\frac{1}{50}\right)}$, we obtain

$$\left| \mathop{\mathbb{P}}_{\substack{S \sim \mathcal{N}(0,I_d)^{\otimes M} \\ \text{Randomness of } \mathcal{H}}} [\mathcal{H}(f_S, S) = \text{Yes}] - \mathop{\mathbb{P}}_{\text{Randomness of } \mathcal{H}} [\mathcal{H}(g_\emptyset, \mathcal{N}(0, I_d)) = \text{Yes}] \right| \leq$$
$$\frac{N^2}{M} + \left(\frac{64(b^2 - a^2)}{d} \right)^{d/2} (M + N)^2 = 2^{-0.02d} + \left(O\left(\frac{1}{\sqrt{d}} \right) \right)^{d/2} = 2^{-\Omega(d)},$$

which is in contradiction with Equation 7.7 for a sufficiently large value of d. This proves that $\mathcal{A}$ is not a (ϵ, δ)-TDS learning algorithm for convex sets. □

References

[ABL17] Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63(6):1–27, 2017. 1.2, 2.1, 2.1

[Baz09] Louay MJ Bazzi. Polylogarithmic independence can fool dnf formulas. *SIAM Journal on Computing*, 38(6):2220–2272, 2009. 3

[BBL06] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72, 2006. 2.2.1

[BCK+07] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007. 1.1, 1.3, 3, A, A.1, A

[BDBC+10] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 1, 1.1, 1.3, 3, A

[BDBCP06] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 1, 1.1, 1.3, 3, A, A

[BDU12] Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, 2012. 2.14

[BHV10] Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Machine learning*, 80:111–139, 2010. 2.2.1

[BM97] Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for lucien le cam*, pages 55–87. Springer, 1997. 7.4

[BT96] Nader H Bshouty and Christino Tamon. On the fourier spectrum of monotone functions. *Journal of the ACM (JACM)*, 43(4):747–770, 1996. 1.1

[CAL94] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15:201–221, 1994. 2.2.1

[Can22] Clément Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends® in Communications and Information Theory*, 19(6):1032–1198, 2022. 1

[Dan15] Amit Daniely. A ptas for agnostically learning halfspaces. In *Conference on Learning Theory*, pages 484–502. PMLR, 2015. 2.1

[DGJ+10] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010. 1.2

[DKPZ21] Ilias Diakonikolas, Daniel M Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the sq model. In *Conference on Learning Theory*, pages 1552–1584. PMLR, 2021. 1, 2.4.1

[DKR23] Ilias Diakonikolas, Daniel Kane, and Lisheng Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. In *International Conference on Machine Learning*, pages 7922–7938. PMLR, 2023. 1

[DKTZ20] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Non-convex sgd learns halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 33:18540–18549, 2020. 1.2, 2.1, 2.1, 4, 4.2

[DKZ20] Ilias Diakonikolas, Daniel Kane, and Nikos Zarifis. Near-optimal sq lower bounds for agnostically learning halfspaces and relus under gaussian marginals. *Advances in Neural Information Processing Systems*, 33:13586–13596, 2020. 1, 2.4.1

[DLLP10] Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010. 1.1, 1.3, 3, A

[EYW12] Ran El-Yaniv and Yair Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2), 2012. 2.2.2

[GGK20] Surbhi Goel, Aravind Gollakota, and Adam Klivans. Statistical-query lower bounds via functional gradients. *Advances in Neural Information Processing Systems*, 33:2147–2158, 2020. 1, 2.4.1

[GKK23] Aravind Gollakota, Adam R Klivans, and Pravesh K Kothari. A moment-matching approach to testable learning and a new characterization of rademacher complexity. *Proceedings of the fifty-fifth annual ACM Symposium on Theory of Computing*, 2023. 1, 1.2, 1.2, 1.3, 2.3, 2.4.1

[GKKM20] 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. 1, 1.3, 2.12, B, B.1

[GKSV23a] Aravind Gollakota, Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. An efficient tester-learner for halfspaces. *arXiv preprint arXiv:2302.14853*, 2023. 1, 1.2, 1.3, 2.1, 3

[GKSV23b] 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).*, 2023. 1, 1.3, 2.1, 2.1, 2.2, 2.3, 2.4

[GOWZ10] Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman. Fooling functions of halfspaces under product distributions. In *2010 IEEE 25th Annual Conference on Computational Complexity*, pages 223–234. IEEE, 2010. 1.2, 6.2, 6.2, 6.11, 6.2, 6.2

[Han07] Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th international conference on Machine learning*, pages 353–360, 2007. 2.2.1

[Han09] Steve Hanneke. *Theoretical foundations of active learning*. Carnegie Mellon University, 2009. 2.2.1

[Han11] Steve Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, pages 333–361, 2011. 2.2.1, 2.2.1

[Han14] Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014. 2.2.1, 2.2.1

[KK21] Adam Tauman Kalai and Varun Kanade. Efficient learning with arbitrary covariate shift. In *Algorithmic Learning Theory*, pages 850–864. PMLR, 2021. 1, 1.3, 2, 2.12, B, B.1, B.4, B

[KKM12] Adam Tauman Kalai, Varun Kanade, and Yishay Mansour. Reliable agnostic learning. *Journal of Computer and System Sciences*, 78(5):1481–1495, 2012. 1.3, B

[KOS08] Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. Learning geometric concepts via gaussian surface area. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 541–550. IEEE, 2008. 1.1

[KS88] Wieslaw Krakowiak and Jerzy Szulga. Hypercontraction principle and random multilinear forms. *Probability Theory and Related Fields*, 77(3):325–342, 1988. 6.2

[MMR09] 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. 1, 1.3, A, A, A.3

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. D.1, D.2

[O'D14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014. D.1, D.2

[OS03] Ryan O'Donnell and Rocco A Servedio. New degree bounds for polynomial threshold functions. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 325–334, 2003. 6.2, 6.5, 6.2

[RMH+20] 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. 1.3

[RV23] Ronitt Rubinfeld and Arsen Vasilyan. Testing distributional assumptions of learning algorithms. *Proceedings of the fifty-fifth annual ACM Symposium on Theory of Computing*, 2023. 1, 1.3, 2.4.1, 7.1, 7.5, 7.2, 7.6

[TCK+22] Niels K Ternov, Anders N Christensen, Peter JT Kampen, Gustav Als, Tine Vestergaard, Lars Konge, Martin Tolsgaard, Lisbet R Hölmich, Pascale Guitera, Annette H Chakera, et al. Generalizability and usefulness of artificial intelligence for skin cancer diagnostics: An algorithm validation study. *JEADV Clinical Practice*, 1(4):344–354, 2022. 1

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. D.1, D.1

[WOD+21] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070, 2021. 1

[Wol07] Paweł Wolff. Hypercontractivity of simple random variables. *Studia Mathematica*, 3(180):219–236, 2007. 6.2

[ZBL+18] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018. 1

A Sample Complexity of TDS Learning

In the previous sections, we explored a number of computational aspects of TDS learning, deriving dimension efficient algorithms for several instantiations of our setting. In this section, we focus on the statistical aspects of TDS learning. There are several prior works in the literature of domain adaptation that study the statistical landscape of the problem of learning under shifting distributions (see, e.g., [BDBCP06, BCK+07, MMR09, BDBC+10, DLLP10]). All of the previous generalization upper bounds on this problem involve some discrepancy term, which quantifies the amount of distribution shift, as well as some additional terms that are typically considered small for reasonable settings. For a concept class $\mathcal{C} : \mathcal{X} \to \{\pm 1\}$, considering that the error term λ (see Eq. (3.1)) is small is a standard assumption in domain adaptation (see, e.g., [BDBCP06, BCK+07]). Furthermore, one standard measure of discrepancy is defined as follows.

Definition A.1 (Discrepancy Distance, [BCK+07])**.** Let $\mathcal{X} \subset \mathbb{R}^d$ and let $\mathcal{C}$ be a concept class mapping $\mathcal{X}$ to $\{\pm 1\}$. For distributions $\mathcal{D}, \mathcal{D}'$ over $\mathcal{X}$, we define the discrepancy distance $\mathrm{disc}_{\mathcal{C}}(\mathcal{D}, \mathcal{D}')$ as follows.

$$\mathrm{disc}_{\mathcal{C}}(\mathcal{D}, \mathcal{D}') = \sup_{f, f' \in \mathcal{C}} \left| \mathbb{P}_{\mathcal{D}}[f(\mathbf{x}) \neq f'(\mathbf{x})] - \mathbb{P}_{\mathcal{D}'}[f(\mathbf{x}) \neq f'(\mathbf{x})] \right|$$

In particular, [BDBCP06, BCK+07] observe that for any $f \in \mathcal{C}$ and distributions $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}, \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}$ over $\mathcal{X} \times \{\pm 1\}$ the following is true.

$$\mathrm{err}(f; \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) \leq \mathrm{err}(f; \mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}) + \mathrm{disc}_{\mathcal{C}}(\mathcal{D}^{\mathrm{train}}_{\mathcal{X}}, \mathcal{D}^{\mathrm{test}}_{\mathcal{X}}) + \lambda(\mathcal{C}; \mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}, \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) \tag{A.1}$$

The bound of Eq. (A.1) can be translated to a generalization bound for domain adaptation, through the use Rademacher complexity, whose definition is provided below.

Definition A.2 (Rademacher Complexity)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$, let $\mathcal{D}$ be a distribution over $\mathcal{X}$ and let $\mathcal{C}$ be a concept class mapping $\mathcal{X}$ to $\{\pm 1\}$. For a set of m samples $X = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)})$ drawn independently from $\mathcal{D}$, we define the empirical Rademacher complexity of $\mathcal{C}$ w.r.t. X as follows

$$\widehat{\mathfrak{R}}_X(\mathcal{C}) = \frac{2}{m} \mathbb{E} \sup_{f \in \mathcal{C}} \sum_{j=1}^{m} \sigma_j f(\mathbf{x}^{(j)}), \text{ where the expectation is over } \sigma \sim \mathrm{Unif}(\{\pm 1\}^d)$$

Moreover, we define the Rademacher complexity of $\mathcal{C}$ at m w.r.t. $\mathcal{D}$ as $\mathfrak{R}_m(\mathcal{C}; \mathcal{D}) = \mathbb{E}[\widehat{\mathfrak{R}}_X(\mathcal{C})]$, where the expectation is over $X \sim \mathcal{D}^{\otimes m}$.

Corollaries 6, 7 in [MMR09], demonstrate that the discrepancy between two distributions is upper bounded as follows.

Proposition A.3 (Bounding the Discrepancy, Corollary 7 in [MMR09])**.** *Consider $\mathcal{X} \subseteq \mathbb{R}^d$, a concept class $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}^d\}$ and distributions $\mathcal{D}, \mathcal{D}'$ over $\mathcal{X}$. Then for any $\delta > 0$, $m, m' \in \mathbb{N}$, if X, X' are independent examples from $\mathcal{D}, \mathcal{D}'$, respectively, of sizes m, m', the following is true.*

$$\mathrm{disc}_{\mathcal{C}}(\mathcal{D}, \mathcal{D}') \leq \mathrm{disc}_{\mathcal{C}}(X, X') + 4\widehat{\mathfrak{R}}_X(\mathcal{C}) + 4\widehat{\mathfrak{R}}_{X'}(\mathcal{C}) + 3\left(\log(4/\delta)\right)^{1/2} \sqrt{\frac{1}{m} + \frac{1}{m'}}$$

Combining inequality (A.1) with Proposition A.3 and standard generalization bounds for classification, yields a data-dependent generalization bound for domain adaptation whose only unknown parameter is λ. In our setting this readily implies the following sample complexity upper bound in terms of the Rademacher complexity of the concept class $\mathcal{C}$.

Corollary A.4 (Sample Complexity upper bound for TDS learning)**.** *Let $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ be a hypothesis class and $\mathcal{D}$ a distribution over $\mathcal{X}$ such that $\mathfrak{R}_m(\mathcal{C}; \mathcal{D}) \leq \epsilon/10$. The algorithm that runs the Empirical Risk Minimizer on training data and accepts only when both the empirical discrepancy distance between the training and test unlabelled examples, i.e. $\mathrm{disc}_{\mathcal{C}}(X_{\mathrm{train}}, X_{\mathrm{test}})$, and the Rademacher complexity with respect to the test examples, i.e. $\widehat{\mathfrak{R}}_{X_{\mathrm{test}}}(\mathcal{C})$, are $O(\epsilon)$, is an (ϵ, δ)-TDS learning algorithm for $\mathcal{C}$ up to error $2\lambda + \epsilon$ with sample complexity $O(m + \frac{1}{\epsilon^2}\log(1/\delta))$. Moreover, if there is a concept in $\mathcal{C}$ with zero training error, the same is true up to error $\lambda + \epsilon$.*

We emphasize that, while Corollary A.4 readily follows from prior results in the literature of domain adaptation, it highlights an important distinction between domain adaptation and TDS learning: A TDS learning algorithm, upon acceptance, achieves error that does not scale with the discrepancy between the training and test marginal distributions, but only a term that depends on the quantity λ, which, as we show in Theorem 2.13, is unavoidable.

B PQ Learning and Distribution-Free TDS Learning

In recent years, there has been a vast amount of work on the problem of learning under shifting distributions. One of the most relevant models to TDS learning is PQ learning (see [GKKM20, KK21]), which was defined by [GKKM20]. In this section, we establish a connection between PQ learning and TDS learning and, in particular, we show that TDS learning can be reduced to PQ learning, thereby inheriting all of the existing results in the latter framework. Unfortunately, to the best of our knowledge, most of the positive results on the PQ learning framework make strong assumptions regarding oracle access to solvers of learning primitives that are typically hard to solve. Nonetheless, PQ learning is an important theoretical framework for learning under arbitrary covariate shifts and it is an interesting open question whether our methods can be extended to provide positive results for the not-easier problem of PQ learning.

In the PQ learning framework, a learner outputs a pair $(h, \mathbf{X})$, where $h : \mathcal{X} \to \{\pm 1\}$ is a classifier and $\mathbf{X} \subseteq \mathcal{X}$ is a subset of the feature space where one can be confident on the predictions of h. In particular, the PQ learning model is defined as follows.

Definition B.1 (PQ Learning, [GKKM20, KK21])**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a set and $\mathcal{C} \subseteq \mathcal{X} \to \{\pm 1\}$ a concept class. For $\epsilon, \delta \in (0, 1)$ we say that algorithm $\mathcal{A}$ PQ learns $\mathcal{C}$ up to error ϵ and probability of failure δ if for any distributions $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}, \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}$ over $\mathcal{X} \times \{\pm 1\}$ such that there is some $f^* \in \mathcal{C}$ so that $y = f^*(\mathbf{x})$ for any $(\mathbf{x}, y)$ drawn from either $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$ or $\mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}$, algorithm $\mathcal{A}$, upon receiving a large enough number of labelled samples from $\mathcal{D}^{\mathrm{train}}_{\mathcal{XY}}$ and a large enough number of unlabelled samples from $\mathcal{D}^{\mathrm{test}}_{\mathcal{X}}$, outputs a pair $(h, \mathbf{X})$ such that $h : \mathcal{X} \to \{\pm 1\}$, $\mathbf{X} \subseteq \mathcal{X}$ and with probability at least $1 - \delta$ the following is true.

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{D}^{\mathrm{train}}_{\mathcal{X}}}[\mathbf{x} \notin \mathbf{X}] \leq \epsilon \text{ and } \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}}[h(\mathbf{x}) \neq y \text{ and } \mathbf{x} \in \mathbf{X}] \leq \epsilon$$

We note that the above definition of PQ learning is distribution-free, i.e., the guarantees hold for any distribution and not with respect to a specific target distribution. In Definition 3.2 for TDS learning, the completeness criterion is stated with respect to a particular target distribution that is the same as the training distribution. However, in order to demonstrate a connection between PQ learning and TDS learning, we now define Distribution-Free TDS learning.

Definition B.2 (Distribution-free TDS Learning)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ and consider a concept class $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$. For $\epsilon, \delta \in (0,1)$, we say that an algorithm $\mathcal{A}$ testably learns $\mathcal{C}$ under distribution shifts up to error ϵ and probability of failure δ if the following is true. For any distributions $\mathcal{D}^{\text{train}}_{\mathcal{XY}}, \mathcal{D}^{\text{test}}_{\mathcal{XY}}$ over $\mathcal{X} \times \{\pm 1\}$ such that there is some $f^* \in \mathcal{C}$ such that $y = f^*(\mathbf{x})$ for any $(\mathbf{x}, y)$ drawn from either $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ or $\mathcal{D}^{\text{test}}_{\mathcal{XY}}$, algorithm $\mathcal{A}$, upon receiving a large enough set of labelled samples S_{train} from the training distribution $\mathcal{D}^{\text{train}}_{\mathcal{XY}}$ and a large enough set of unlabelled samples X_{test} from the test distribution $\mathcal{D}^{\text{test}}_{\mathcal{X}}$, either rejects $(S_{\text{train}}, X_{\text{test}})$ or accepts and outputs a hypothesis $h : \mathcal{X} \to \{\pm 1\}$ with the following guarantees.

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

If A accepts, then the output h satisfies $\text{err}(h; \mathcal{D}^{\text{test}}_{\mathcal{XY}}) \leq \epsilon$.

(b) (Completeness.) Whenever $\mathcal{D}^{\text{test}}_{\mathcal{X}} = \mathcal{D}^{\text{train}}_{\mathcal{X}}$, A accepts with probability at least $1 - \delta$ over the samples $S_{\text{train}}, X_{\text{test}}$.

We are now ready to prove that distribution-free TDS learning reduces to PQ learning.

Proposition B.3 (TDS learning via PQ learning)**.** *Algorithm 6 reduces TDS to PQ learning. In particular, for $\epsilon, \delta \in (0,1)$, PQ learning algorithm $\mathcal{A}$ and a concept class $\mathcal{C}$, Algorithm 6, upon receiving $m_P + \frac{C}{\epsilon^2}\log(1/\delta)$ labelled examples S_{train} from the training distribution and $m_Q + \frac{C}{\epsilon^2}\log(1/\delta)$ unlabelled examples X_{test} from the test distribution where m_P, m_Q are such that $\mathcal{A}$ is an $(\epsilon/4, \delta)$-PQ learning algorithm for $\mathcal{C}$ given m_P training and m_Q test examples, (ϵ, δ)-TDS learns $\mathcal{C}$.*

Proof. Let $C > 0$ be a sufficiently large universal constant. For **soundness**, we observe that upon acceptance, we have $\mathbb{P}_{\mathbf{x} \sim X_2}[\mathbf{x} \notin \mathbf{X}]$ and by a Hoeffding bound, since $m_2 \geq \frac{C}{\epsilon_2}\log(1/\delta)$, we have $\mathbb{P}_{\mathbf{x} \sim \mathcal{D}^{\text{test}}_{\mathcal{X}}}[\mathbf{x} \notin \mathbf{X}] \leq 2\epsilon/3$. By using the fact that $\text{err}(h; \mathcal{D}^{\text{test}}_{\mathcal{XY}}) \leq \mathbb{P}_{\mathbf{x} \sim \mathcal{D}^{\text{test}}_{\mathcal{X}}}[\mathbf{x} \in \mathbf{X}] + \mathbb{P}_{\mathbf{x} \sim \mathcal{D}^{\text{test}}_{\mathcal{X}}}[\mathbf{x} \in \mathbf{X}]$ and the guarantee of the PQ learner we obtain $\text{err}(h; \mathcal{D}^{\text{test}}_{\mathcal{XY}}) \leq \epsilon$, with probability at least $1 - \delta$. For **completeness**, we use the definition of PQ learning and a Hoeffding bound to show that with probability at least $1 - \delta$, Algorithm 6 accepts whenever $\mathcal{D}^{\text{test}}_{\mathcal{X}} = \mathcal{D}^{\text{train}}_{\mathcal{X}}$. □

Algorithm 6: TDS learning through PQ learning

Input: Sets S_{train}, X_{test}, parameters $\epsilon, \delta \in (0,1)$, $(\epsilon' = \frac{\epsilon}{4}, \delta)$-PQ learner $\mathcal{A}$
Set $m_1 = m_Q$, $m_2 = \frac{C}{\epsilon^2}\log(1/\delta)$ and split X_{test} in X_1, X_2 with sizes m_1, m_2.
Run algorithm $\mathcal{A}$ on (S_{train}, X_1) and receive output $(h, \mathbf{X})$.
Reject if $\mathbb{P}_{\mathbf{x} \sim X_2}[\mathbf{x} \notin \mathbf{X}] > \epsilon/3$.
Otherwise, output h and terminate.

The simple reduction we provided in Proposition B.3 implies that all of the positive results on PQ learning transfer to TDS learning. Moreover, note that the reduction does not alter the training and test distributions between the corresponding TDS and PQ algorithms and, therefore, would hold even in the distribution specific setting. This is not true, however, about the following corollary which is based on a reduction from PQ learning to reliable agnostic learning, which does not preserve the marginal distributions.

Corollary B.4 (Combination of Theorem 5 in [KK21] and Proposition B.3)**.** *If a concept class $\mathcal{C}$ is distribution -free reliably learnable, then it is TDS learnable in the distribution-free setting.*

We remark that, in fact, (distribution-free) PQ learning is equivalent to (distribution-free) reliable learning (see Theorems 5, 6 in [KK21]). For a definition of reliable learning we refer the reader to [KKM12]. It is known that reliable learning is no harder than agnostic learning and no easier than PAC learning.

C Amplifying success probability

We will now demonstrate that it is possible to amplify the probability of success of a TDS learner through repetition. Note that this is not immediate for TDS learning as it is, for example, in agnostic learning, where one may repeat an agnostic learning algorithm and choose the hypothesis with the smallest error estimate among the outputs of the independent runs. The main obstacle is that test labels are not available. Nonetheless, we obtain the following theorem regarding amplifying the probability of success.

Proposition C.1 (Amplifying Success Probability)**.** *Let $\mathcal{C}$ be a hypothesis class, $\mathcal{D}$ a distribution and suppose $\mathcal{A}$ is a TDS learner for $\mathcal{C}$ with respect to $\mathcal{D}$ with error guarantee $\psi(\lambda) + \epsilon$ and failure probability at most 0.1. Then, there is a TDS learner $\mathcal{A}'$ for $\mathcal{C}$ with respect to $\mathcal{D}$ with error guarantee $4\psi(\lambda) + 4\epsilon$ and failure probability at most δ. In particular, $\mathcal{A}'$ repeats $\mathcal{A}$ for $T = O(\log(\frac{1}{\epsilon\delta}))$ times and rejects if most of the repetitions reject. If most repetitions accept, $\mathcal{A}'$ outputs the hypothesis $h = \mathrm{maj}(h_1, \ldots, h_{T/2})$ (h outputs the majority vote of h_i), where $h_1, \ldots, h_{T/2}$ are the outputs of the first $T/2$ repetitions of $\mathcal{A}$ that accepted.*

Proof. We split the proof into two parts, one for soundness and one for completeness.

Soundness. For soundness, suppose that $\mathcal{A}'$ accepts. We denote with $\widehat{\mathbb{P}}$ (resp. $\widehat{\mathbb{E}}$) the probabilities (resp. expectations) over the randomness of $h_1, \ldots, h_{T/2}$ (which originates to the randomness of the samples given to $\mathcal{A}$) and with $\mathbb{P}$ (resp. $\mathbb{E}$) the probabilities (resp. expectations) over the randomness of a pair $(\mathbf{x}, y)$ drawn from $\mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}$. In what follows, let $\eta = \psi(\lambda) + \epsilon$. We have that for any $i = 1, 2, \ldots, T/2$, $\widehat{\mathbb{P}}[\mathrm{err}(h_i, \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) \leq \eta] \geq 0.9$, by the guarantees of $\mathcal{A}'$. We will show that $\widehat{\mathbb{P}}[\mathrm{err}(h, \mathcal{D}^{\mathrm{test}}_{\mathcal{XY}}) \leq 4\eta] \geq 1 - \delta$ for a sufficiently large $T = O(\log(\frac{1}{\epsilon\delta}))$.

We define $\mathcal{G}_i$ to be the event (over the randomness of h_i) that h_i is 'good', i.e., that $\mathbb{P}[h_i(\mathbf{x}) \neq y] \leq \eta$. We define $\mathbf{Z}$ to be the 'bad' region of $(\mathbf{x}, y)$, i.e., $\mathbf{Z} = \{(\mathbf{x}, y) \in \mathcal{X} \times \{\pm 1\} : \widehat{\mathbb{P}}[h_1(\mathbf{x}) \neq y | \mathcal{G}_1] > 1/3\}$. Note that $\mathbf{Z}$ would be the same even if we substituted $(h_1, \mathcal{G}_1)$ above with an arbitrary $(h_i, \mathcal{G}_i)$.

First, we observe that $\mathbb{P}[h(\mathbf{x}) \neq y] \leq \mathbb{P}[(\mathbf{x}, y) \in \mathbf{Z}] + \mathbb{P}[h(\mathbf{x}) \neq y | (\mathbf{x}, y) \notin \mathbf{Z}]$.

We now observe that $\mathbb{P}[(\mathbf{x}, y) \in \mathbf{Z}] = \mathbb{P}[\widehat{\mathbb{P}}[h_1 \neq y | \mathcal{G}_1] > 1/3] \leq 3\,\mathbb{E}\,\widehat{\mathbb{P}}[h_1(\mathbf{x}) \neq y | \mathcal{G}_1]$ by Markov's inequality. Now, we may swap the expectations to obtain $\mathbb{P}[(\mathbf{x}, y) \in \mathbf{Z}] \leq 3\widehat{\mathbb{E}}[\mathbb{P}[h_1(\mathbf{x}) \neq y] | \mathcal{G}_1] \leq 3\eta$.

So far, we have shown $\mathbb{P}[h(\mathbf{x}) \neq y] \leq 3\eta + \mathbb{P}[h(\mathbf{x}) \neq y | (\mathbf{x}, y) \notin \mathbf{Z}]$. We will bound the probability over $h_1, \ldots, h_{T/2}$ that $\mathbb{P}[h(\mathbf{x}) \neq y | (\mathbf{x}, y) \notin \mathbf{Z}] > \eta$. In particular, we have the following due to Markov's inequality $\widehat{\mathbb{P}}[\mathbb{P}[h(\mathbf{x}) \neq y | (\mathbf{x}, y) \notin \mathbf{Z}] > \eta] \leq \frac{1}{\eta}\widehat{\mathbb{E}}[\mathbb{P}[h(\mathbf{x}) \neq y | (\mathbf{x}, y) \notin \mathbf{Z}]]$. Once more, we may swap the expectations to obtain $\widehat{\mathbb{P}}[\mathbb{P}[h(\mathbf{x}) \neq y | (\mathbf{x}, y) \notin \mathbf{Z}] > \eta] \leq \frac{1}{\eta}\,\mathbb{E}[\widehat{\mathbb{P}}[h(\mathbf{x}) \neq y] | (\mathbf{x}, y) \notin \mathbf{Z}]$.

Moreover, if we fix $(\mathbf{x}, y) \notin \mathbf{Z}$, then $\widehat{\mathbb{P}}[h_i(\mathbf{x}) = y] \geq \widehat{\mathbb{P}}[h_i(\mathbf{x}) = y \text{ and } \mathcal{G}_i] \geq \frac{2}{3} \cdot \frac{9}{10} \geq 3/5$. Because $\widehat{\mathbb{P}}[\mathcal{G}_i] \geq 0.9$ and $\widehat{\mathbb{P}}[h_i(\mathbf{x}) = y | \mathcal{G}_i] \geq 2/3$ whenever $(\mathbf{x}, y) \notin \mathbf{Z}$, by the definition of $\mathbf{Z}$. Therefore, since $h_1, \ldots, h_{T/2}$ are independent, we have that $\widehat{\mathbb{P}}[h(\mathbf{x}) \neq y] \leq \exp(-T/C)$ for some sufficiently large universal constant $C > 0$, for any $(\mathbf{x}, y) \notin \mathbf{Z}$.

Therefore, in total, $\widehat{\mathbb{P}}[\mathbb{P}[h(\mathbf{x}) \neq y | (\mathbf{x}, y) \notin \mathbf{Z}] > \eta] \leq \frac{1}{\eta}\exp(-T/C)$. We set $T = C\ln(\frac{1}{\epsilon\delta}) \geq C\ln(\frac{1}{\eta\delta})$ to obtain $\widehat{\mathbb{P}}[\mathbb{P}[h(\mathbf{x}) \neq y | (\mathbf{x}, y) \notin \mathbf{Z}] > \eta] \leq \delta$ and, hence, with probability at least $1 - \delta$ over the randomness of h we overall have $\mathbb{P}[h(\mathbf{x}) \neq y] \leq 4\eta$.

Completeness. Completeness follows by a standard Hoeffding bound. □

D Auxiliary Propositions

Let $\mathcal{N}(0, I_d)$ denote the standard multivariate Gaussian distribution over $\mathbb{R}^d$ and $\mathrm{Unif}(\{\pm 1\}^d)$ denote the uniform distribution over the hypercube $\{\pm 1\}^d$. For each of these distributions, we show that the sand-

wiching polynomials of any binary concept have coefficients that are absolutely bounded, that the empirical moments concentrate around the true ones and that the empirical squared error of polynomials with bounded degree and coefficients uniformly converges to the true squared error. These properties are used in order to apply Theorem 2.9 to obtain TDS learning algorithms for a number of classes under the Gaussian and Uniform distributions.

D.1 Properties of Gaussian Distribution

We prove the following fact about the Gaussian distribution.

Lemma D.1 (Properties of the Gaussian Distribution)**.** *Let $\mathcal{D}$ be the standard Gaussian $\mathcal{N}(0, I_d)$ over $\mathbb{R}^d$. Then the following are true.*

(*i*) *(Coefficient Bound) Suppose that for some $\epsilon \in (0, 1]$, $k > 0$ and some concept class $\mathcal{C} \subseteq \mathbb{R}^d \to \{\pm 1\}$, the ϵ-approximate L_2 sandwiching degree of $\mathcal{C}$ w.r.t. $\mathcal{N}(0, I_d)$ is at most k. Then, the coefficients of the sandwiching polynomials for $\mathcal{C}$ are absolutely bounded by $B = O(d)^k$.*

(*ii*) *(Concentration) For any $\delta \in (0, 1), \Delta > 0$ and $k > 0$, if X is a set of independent samples from $\mathcal{D}$ with size at least $m_{\text{conc}} = \frac{O(dk)^k}{\Delta^2 \cdot \delta}$ then, with probability at least $1 - \delta$ over the randomness of X, we have that for any $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \leq k$ it holds $|\mathbb{E}_X[\mathbf{x}^\alpha] - \mathbb{E}_{\mathcal{D}}[\mathbf{x}^\alpha]| \leq \Delta$.*

(*iii*) *(Generalization) For any $\epsilon > 0$, $\delta \in (0, 1), B > 0$, $k > 0$, and any distribution $\mathcal{D}_{\mathcal{XY}}$ over $\mathbb{R}^d \times \{\pm 1\}$ whose marginal on $\mathbb{R}^d$ is $\mathcal{D}$, if S is a set of independent samples from $\mathcal{D}_{\mathcal{XY}}$ with size at least $m_{\text{gen}} = \tilde{O}(\frac{B^8}{\epsilon^4 \delta}) \cdot d^{O(k)}$ then, with probability at least $1 - \delta$ over the randomness of S, we have that for any polynomial p of degree at most k and coefficients that are absolutely bounded by B it holds $|\mathbb{E}_S[(y - p(\mathbf{x}))^2] - \mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y - p(\mathbf{x}))^2]| \leq \epsilon$.*

Proof. We will prove each part of the Lemma separately.
Part (i). Suppose that $p_{\text{up}}, p_{\text{down}}$ are 1-sandwiching polynomials for some concept $f \in \mathcal{C}$ with degree at most k. Then, we have the following.

$$\begin{aligned}\|p_{\text{down}}\|_{L_2(\mathcal{D})} &\leq \|p_{\text{up}} - f\|_{L_2(\mathcal{D})} + \|f\|_{L_2\mathcal{D}} \\ &\leq \|p_{\text{up}} - p_{\text{down}}\|_2 + 1 \leq 2\end{aligned}$$

Since $\mathcal{D}$ is the standard Gaussian distribution, the quantity $\|p_{\text{down}}\|^2_{L_2(\mathcal{D})}$ equals to the sum of the squares of the coefficients of the Hermite expansion of p_{down} (see e.g. [O'D14]). Therefore, each Hermite coefficient of p_{down} is absolutely bounded by 2. Each Hermite polynomial of degree at most k has coefficients that are absolutely bounded by 2^k. Since p_{down} has degree at most k, each coefficient of p_{down} is absolutely bounded by $d^{O(k)}$.
Part (ii). Suppose that $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \leq k$. Then, the worst case regarding moment concentration is $\alpha_1 = k$. For a sample X from $\mathcal{D}$, we apply Chebyshev's inequality on the random variable $z = |\mathbb{E}_X[\mathbf{x}_1^k] - \mathbb{E}_{\mathcal{D}}[\mathbf{x}_1^k]|$ and by bounding $\mathbb{E}[z^2]$ by $\mathbb{E}_{\mathcal{D}}[\mathbf{x}_1^{2k}]$ we have that for any $\Delta > 0$, $z \leq \Delta$ with probability at least $1 - \frac{(Ck)^k}{|X|\Delta^2}$, where the randomness is over the random choice of X and $C > 0$ is a sufficiently large universal constant (for bounds on the Gaussian moments, see, e.g., Proposition 2.5.2 in [Ver18]). Since we need the result to hold for all α simultaneously, the result follows by a union bound.
Part (iii). We define $\mathcal{P}$ to be the class of polynomials over $\mathbb{R}^d$ with degree at most k and coefficients that are absolutely bounded by B. Let $T > 0$ to be disclosed later and $m = |S|$. We will first show that with probability at least $1 - \delta/2$ over the choice of S, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{XY}}}[(y - p(\mathbf{x}))^2] \leq \mathop{\mathbb{E}}_{S}[(y - p(\mathbf{x}))^2] + \epsilon \text{ for all } p \in \mathcal{P}$$

We aim to apply some standard uniform convergence argument, but in order to do so we first need to ensure certain boundedness conditions as follows.

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2] = \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2 \cdot \mathbb{1}\{\forall q \in \mathcal{P} : |q(\mathbf{x})| \le T\}] + \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2 \cdot \mathbb{1}\{\exists q \in \mathcal{P} : |q(\mathbf{x})| > T\}]$$

where we have $\mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2 \cdot \mathbb{1}\{\forall q \in \mathcal{P} : |q(\mathbf{x})| \le T\}] \le \mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2 \mid \forall q \in \mathcal{P} : |q(\mathbf{x})| \le T]$. Let $\mathcal{D}'_{\mathcal{XY}}$ be the distribution that corresponds to $\mathcal{D}_{\mathcal{XY}}$ conditioned on the event $\{\forall q \in \mathcal{P} : |q(\mathbf{x})| \le T\}$ and let $S' = \{(\mathbf{x}, y) \in S : |q(\mathbf{x})| \le T, \forall q \in \mathcal{P}\}$. By standard arguments using Rademacher complexity bounds for bounded losses (see, e.g., Theorems 5.5 and 10.3 in [MRT18]) we have that for some sufficiently large universal constant $C > 0$, with probability at least $1 - \delta/4$, we have for any $p \in \mathcal{P}$

$$\mathop{\mathbb{E}}_{\mathcal{D}'_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2] \le \mathop{\mathbb{E}}_{S'}[(y-p(\mathbf{x}))^2] + T^4 \cdot \frac{B + \sqrt{\log(1/\delta)}}{\sqrt{m/C}} \tag{D.1}$$

We now need to link $\mathbb{E}_{S'}[(y-p(\mathbf{x}))^2]$ to $\mathbb{E}_S[(y-p(\mathbf{x}))^2]$. We have the following.

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}[(y-p(\mathbf{x}))^2] &\ge (1 - \mathop{\mathbb{P}}_{S}[\exists q \in \mathcal{P} : |q(\mathbf{x})| > T]) \mathop{\mathbb{E}}_{S'}[(y-p(\mathbf{x}))^2] \\
&\ge \mathop{\mathbb{E}}_{S'}[(y-p(\mathbf{x}))^2] - \mathop{\mathbb{P}}_{S}[\exists q \in \mathcal{P} : |q(\mathbf{x})| > T] \cdot 2T^2 \qquad \text{(since } y \in \{\pm 1\} \text{ and } p \in \mathcal{P}\text{)}
\end{aligned}$$

We will upper bound the quantity $\mathbb{P}_S[\exists q \in \mathcal{P} : |q(\mathbf{x})| > T]$. We have

$$\begin{aligned}
\mathop{\mathbb{P}}_{S}[\exists q \in \mathcal{P} : |q(\mathbf{x})| > T] &= \mathop{\mathbb{P}}_{S}\Big[\exists (q_\alpha)_{\|\alpha\|_1 \le k} \in [-B, B]^{d^k} : \Big|\sum_\alpha q_\alpha \mathbf{x}^\alpha\Big| > T\Big] \\
&\le \sum_{\alpha : \|\alpha\|_1 \le k} \mathop{\mathbb{P}}_{S}\Big[|\mathbf{x}^\alpha| \ge \frac{T}{Bd^k}\Big] \\
&\le \sum_{\alpha : \|\alpha\|_1 \le k} \mathop{\mathbb{P}}_{\mathcal{D}}\Big[|\mathbf{x}^\alpha| \ge \frac{T}{Bd^k}\Big] + \frac{d^k}{\sqrt{2m}} \log\Big(\frac{8}{\delta}\Big), \text{ w.p. at least } 1 - \delta/4
\end{aligned} \tag{D.2}$$

In the last step, we used a standard Chernoff-Hoeffding bound. We now bound $\sum_{\alpha:\|\alpha\|_1 \le k} \mathbb{P}_{\mathcal{D}}[|\mathbf{x}^\alpha| \ge \frac{T}{Bd^k}]$. Recall that $\mathcal{D} = \mathcal{N}(0, I_d)$ and therefore the worst case for α regarding concentration is the case $\alpha_1 = k$. We therefore obtain the following via Gaussian concentration.

$$\begin{aligned}
\sum_{\alpha : \|\alpha\|_1 \le k} \mathop{\mathbb{P}}_{\mathcal{D}}\Big[|\mathbf{x}^\alpha| \ge \frac{T}{Bd^k}\Big] &\le d^k \mathop{\mathbb{P}}_{\mathcal{D}}\Big[|\mathbf{x}_1^k| \ge \frac{T}{Bd^k}\Big] \\
&\le d^k \exp\Big(-\frac{1}{2} \cdot \frac{T^{1/k}}{B^{1/k} d}\Big)
\end{aligned} \tag{D.3}$$

It remains to bound the term $\mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2 \cdot \mathbb{1}\{\exists q \in \mathcal{P} : |q(\mathbf{x})| > T\}]$. By applying the Cauchy-Schwarz inequality, it is sufficient to bound $\sqrt{\mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^4]} \cdot \sqrt{\mathbb{P}_{\mathcal{D}}[\exists q \in \mathcal{P} : |q(\mathbf{x})| > T]}$. For the second term, we use Equation (D.3). For the first term, we have the following for some sufficiently large constant $C > 0$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{D}}[(y-p(\mathbf{x}))^4] &\le 8 + 8 \mathop{\mathbb{E}}_{\mathcal{D}}[p^4(\mathbf{x})] \\
&\le 8 + B^4 d^{4k} \sum_{\|\alpha\|_1 \le 4k} \prod_{i : \alpha_i > 0} \mathop{\mathbb{E}}_{\mathcal{D}}[\mathbf{x}_i^{\alpha_i}] \qquad \text{(since } \deg(p^4) \le 4k \text{ and } |(p^4)_\alpha| \le B^4 d^{4k}\text{)} \\
&\le B^4 d^{8k} (Ck)^{2k} \qquad \text{(since } \mathcal{D} = \mathcal{N}(0, I_d)\text{, see Proposition 2.5.2 in [Ver18])}
\end{aligned}$$

Using the above inequality along with (D.1), (D.2) and (D.3) we obtain that $\mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2]-\mathbb{E}_S[(y-p(\mathbf{x}))^2]$ is upper bounded by the following quantity for some sufficiently large universal constant $C > 0$

$$T^4 \cdot \frac{B+\sqrt{\log(1/\delta)}}{\sqrt{m/C}} + 2T^2 d^k \exp\left(-\frac{1}{2}\cdot\left(\frac{T}{Bd^k}\right)^{1/k}\right) + \\ + 2T^2 \frac{d^k}{\sqrt{2m}}\log\left(\frac{10}{\delta}\right) + B^2 d^{4k}(Ck)^k d^{k/2}\exp\left(-\frac{1}{4}\cdot\left(\frac{T}{Bd^k}\right)^{1/k}\right),$$

which is at most ϵ when we choose m, T as follows for some universal constant $C > 0$ (possibly larger than the previously defined constants for which we used the same letter) for the choice $T = CB(4d)^k k\log(\frac{Bdk}{\epsilon})$ and $m = \frac{C}{\epsilon^2}(B^2+\log(\frac{1}{\delta}))B^8(4d)^{8k}k^8\log(\frac{Bdk}{\epsilon}) = \tilde{O}(\frac{B}{\epsilon^2})\cdot O(d)^{8k}\cdot\log(1/\delta)$.

In order to bound the symmetric difference, we also need to bound the quantity $\mathbb{E}_S[(y-p(\mathbf{x}))^2] - \mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2]$, which we may do following a similar reasoning, but requiring, at times, bounds on quantities that correspond to empirical expectations (instead of expectations over the population distribution). In particular, we will require a bound on $\mathbb{E}_S[(y-p(\mathbf{x}))^4]$, which can be reduced to bounding $\mathbb{E}_S[p^4(\mathbf{x})]$, for which we may use part (ii), demanding $m \geq d^{O(k)}/\delta$ to obtain

$$\mathop{\mathbb{E}}_{S}[p^4(\mathbf{x})] \leq 2B^4 d^{4k}(Ck)^k$$

Overall, this step will introduce the additional requirement that $m \geq \frac{B^8}{\epsilon^4\delta}d^{16k}(Ck)^{4k}\log^2(\frac{1}{\delta})$. Therefore, overall, for $m \geq m_{\text{gen}} = \tilde{O}(\frac{B^8}{\epsilon^2\delta})\cdot d^{O(k)}\cdot\log^2(\frac{1}{\delta})$, we have the desired result. □

D.2 Properties of Uniform Distribution

We prove the following fact about the uniform distribution.

Lemma D.2 (Properties of the Uniform Distribution)**.** *Let $\mathcal{D}$ be the uniform distribution over the hypercube $\text{Unif}(\{\pm1\}^d)$ and $C > 0$ some sufficiently large constant. Then the following are true.*

(*i*) *(Coefficient Bound) Suppose that for some $\epsilon \in (0,1]$, $k > 0$ and some concept class $\mathcal{C} \subseteq \mathbb{R}^d \to \{\pm1\}$, the ϵ-approximate L_2 sandwiching degree of $\mathcal{C}$ w.r.t. $\mathcal{D}$ is at most k. Then, the coefficients of the sandwiching polynomials for $\mathcal{C}$ are absolutely bounded by $B = 2$.*

(*ii*) *(Concentration) For any $\delta \in (0,1), \Delta > 0$ and $k > 0$, if X is a set of independent samples from $\mathcal{D}$ with size at least $m_{\text{conc}} = \frac{Ck}{\Delta^2}\log(\frac{d}{\delta})$ then, with probability at least $1-\delta$ over the randomness of X, we have that for any $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \leq k$ it holds $|\mathbb{E}_X[\mathbf{x}^\alpha] - \mathbb{E}_{\mathcal{D}}[\mathbf{x}^\alpha]| \leq \Delta$.*

(*iii*) *(Generalization) For any $\epsilon > 0$, $\delta \in (0,1)$, $B > 0$, $k > 0$, and any distribution $\mathcal{D}_{\mathcal{XY}}$ over $\mathbb{R}^d \times \{\pm1\}$ whose marginal on $\mathbb{R}^d$ is $\mathcal{D}$, if S is a set of independent samples from $\mathcal{D}_{\mathcal{XY}}$ with size at least $m_{\text{gen}} = \tilde{O}(\frac{1}{\epsilon^2})\cdot B^{O(1)}\cdot d^{O(k)}\cdot\log(\frac{1}{\delta})$ then, with probability at least $1-\delta$ over the randomness of S, we have that for any polynomial p of degree at most k and coefficients that are absolutely bounded by B it holds $|\mathbb{E}_S[(y-p(\mathbf{x}))^2] - \mathbb{E}_{\mathcal{D}_{\mathcal{XY}}}[(y-p(\mathbf{x}))^2]| \leq \epsilon$.*

Proof. We will prove each part of the Lemma separately.
Part (i). Suppose that $p_{\text{up}}, p_{\text{down}}$ are 1-sandwiching polynomials for some concept $f \in \mathcal{C}$ with degree at most k. Then, we have the following.

$$\begin{aligned}\|p_{\text{down}}\|_{L_2(\mathcal{D})} &\leq \|p_{\text{up}} - f\|_{L_2(\mathcal{D})} + \|f\|_{L_2\mathcal{D}} \\ &\leq \|p_{\text{up}} - p_{\text{down}}\|_2 + 1 \leq 2\end{aligned}$$

Since $\mathcal{D}$ is the uniform distribution, the quantity $\|p_{\text{down}}\|^2_{L_2(\mathcal{D})}$ equals to the sum of the squares of the coefficients of p_{down} (see e.g. [O'D14]). Therefore, each coefficient of p_{down} is absolutely bounded by 2.

Part (ii). Suppose that $\alpha \in \{0,1\}^d$ with $\|\alpha\|_1 \leq k$. For a sample X from $\mathcal{D}$, we apply Hoeffding's inequality on the random variable $z = |\mathbb{E}_X[\mathbf{x}^\alpha] - \mathbb{E}_{\mathcal{D}}[\mathbf{x}^\alpha]|$ and by observing that $\mathbf{x}^\alpha \in \{\pm 1\}$ we have that the probability that $z > \Delta$ is at most $2\exp(-|X|\Delta^2/10)$. We obtain the desired result by a union bound.

Part (iii). We define $\mathcal{P}$ to be the class of polynomials over $\{\pm 1\}^d$ with degree at most k and coefficients that are absolutely bounded by B. Let $T > 0$ to be disclosed later and $m = |S|$. We will show that with probability at least $1 - \delta$ over the choice of S, we have

$$|\mathop{\mathbb{E}}_{\mathcal{D}\mathcal{X}\mathcal{Y}}[(y - p(\mathbf{x}))^2] - \mathop{\mathbb{E}}_{S}[(y - p(\mathbf{x}))^2]| \leq \epsilon \text{ for all } p \in \mathcal{P}$$

We apply some standard uniform convergence argument, by observing that $(y - p(\mathbf{x}))^2 \leq 2 + 2B^2 d^k$. In particular by standard arguments using Rademacher complexity bounds for bounded losses (see, e.g., Theorems 5.5 and 10.3 in [MRT18]) we obtain the desired result. □