Sandwiching Polynomials for Geometric Concepts with Low Intrinsic Dimension

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

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

Arsen Vasilyan ARSENVASILYAN@GMAIL.COM
University of Texas at Austin

Editors: Steve Hanneke and Tor Lattimore

Abstract

Recent work has shown the surprising power of low-degree *sandwiching* polynomial approximators in the context of challenging learning settings such as learning with distribution shift, testable learning, and learning with contamination. A pair of sandwiching polynomials approximate a target function in expectation while also providing *pointwise* upper and lower bounds on the function's values. In this paper, we give a new method for constructing low-degree sandwiching polynomials that yield greatly improved degree bounds for several fundamental function classes and marginal distributions. In particular, we obtain degree $\mathrm{poly}(k)$ sandwiching polynomials for functions of k halfspaces under the Gaussian distribution, improving exponentially over the prior $2^{O(k)}$ bound. More broadly, our approach applies to function classes that are low-dimensional and have smooth boundary.

In contrast to prior work, our proof is relatively simple and directly uses the smoothness of the target function's boundary to construct sandwiching Lipschitz functions, which are amenable to results from high-dimensional approximation theory. For low-dimensional polynomial threshold functions (PTFs) with respect to Gaussians, we obtain doubly exponential improvements without applying the FT-mollification method of Kane used in the best previous result.

Keywords: Sandwiching Polynomials, PAC Learning, Testable Learning, Distribution Shift, Contamination

1. Introduction

Polynomial approximation has played a central role in computational learning theory for over three decades (Linial et al., 1993; Kalai et al., 2008; Klivans et al., 2008; Chandrasekaran et al., 2024a; Pinto Jr et al., 2025; Koehler and Wu, 2025). For example, the work of Kalai et al. (2008) showed that if function class $\mathcal{C}$ admits low-degree polynomial approximators with respect to the input marginal $\mathcal{D}$, then $\mathcal{C}$ can be efficiently learned in the agnostic model of learning. Moreover, subsequent results provide evidence that polynomial approximation is essentially *necessary* for efficient agnostic learning with near-optimal error guarantees (Dachman-Soled et al., 2014; Diakonikolas et al., 2021). In this sense, understanding whether a function class admits a good low-degree polynomial approximator is the key step in developing a provably efficient agnostic learner.

Sandwiching polynomials impose a stronger, more structured form of approximation. Fix a distribution $\mathcal{D}$. A *sandwiching pair* consists of two polynomials p_{down} and p_{up} such that (i) they approximate f on average over D, and (ii) they *pointwise bound* the function for every input $\mathbf{x}$, namely

$$p_{\mathrm{down}}(\mathbf{x}) \le f(\mathbf{x}) \le p_{\mathrm{up}}(\mathbf{x}) \qquad \text{for all } \mathbf{x}.$$

Thus, sandwiching polynomials are a special type of approximating polynomials: beyond small average error, the polynomial approximators are required to never "cross" the target function. Recent work has shown that the existence of low-degree sandwiching polynomials is quite powerful and yields efficient algorithms for several notoriously challenging learning tasks including testable learning, learning under distribution shift, and learning with contamination (Rubinfeld and Vasilyan (2023); Gollakota et al. (2023); Goel et al. (2024); Chandrasekaran et al. (2024b); Klivans et al. (2024a, 2025a)).

Despite their growing importance, the sandwiching degree of many natural function classes is still poorly understood. For example, for the class of functions of k halfspaces with respect to the Gaussian distribution, the best previously known degree bound (prior to this work) was $2^{O(k)}$ (Gopalan et al., 2010; Gollakota et al., 2023).

We give a general methodology for constructing sandwiching polynomials, leading to substantially improved degree bounds in several settings. As a concrete example, our approach yields degree-$\widetilde{O}(k^5)$ sandwiching polynomials for the class of functions of k halfspaces under the Gaussian distribution, improving over the previous $2^{O(k)}$ bounds of Gopalan et al. (2010); Gollakota et al. (2023). These prior bounds are based on two different approaches: Gopalan et al. (2010) use a polynomial composition method to handle functions of halfspaces based on structured polynomial approximators for single halfspaces, while Gollakota et al. (2023) use the method of distances combined with the duality between sandwiching approximation and fooling via moment matching. Both approaches yield exponentially worse degree bounds than ours. We also obtain a *doubly* exponential degree improvement for low-dimensional polynomial threshold functions (PTFs).

More generally, our approach applies to any function class that satisfies two broad conditions: (i) *low-dimensionality*, meaning each function depends only on a projection onto a low-dimensional subspace (e.g., the span of k halfspace normals has dimension at most k), and (ii) a σ-*smooth boundary*, meaning that any ρ-neighborhood of the decision boundary has probability mass at most $\sigma\rho$. Thus, we identify a clean structural principle for low-degree sandwiching, rather than relying on class-specific constructions.

In addition, our approach is not limited to Gaussian marginals: we handle a wide range of distributions far beyond the Gaussian case, namely arbitrary strictly subexponential distributions. See Table 1 for a summary of our sandwiching-degree bounds and the prior state of the art. Our new degree bounds for the above function classes lead to state-of-the-art running times for corresponding algorithms in testable learning, learning with distribution shift, and learning with heavy contamination.

1.1. Our Results

Our main result concerns the sandwiching degree of concepts with low intrinsic dimension and smooth boundary with respect to a strictly subexponential distribution. The sandwiching degree of a concept f relative to a $\mathcal{D}$ is formally defined as follows.

Definition 1 (Sandwiching Degree) *For a function $f : \mathbb{R}^d \to \{\pm 1\}$, a distribution $\mathcal{D}$ over $\mathbb{R}^d$, $\epsilon \in (0,1)$ and $s \geq 1$, we say that the (ϵ, s)-sandwiching degree of f with respect to $\mathcal{D}$ is ℓ if there are polynomials $p_{\mathrm{up}}, p_{\mathrm{down}}$ of degree at most ℓ such that the following conditions hold:*

1. $p_{\mathrm{down}}(\mathbf{x}) \leq f(\mathbf{x}) \leq p_{\mathrm{up}}(\mathbf{x})$ *for all* $\mathbf{x} \in \mathbb{R}^d$.
2. $\|p_{\mathrm{up}} - p_{\mathrm{down}}\|_{\mathcal{D},s} := (\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|p_{\mathrm{up}}(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x})|^s])^{1/s} \leq \epsilon$.

The sandwiching degree of a concept class $\mathcal{F}$ is defined as the supremum of the corresponding sandwiching degrees of its elements.

Theorem 2 (Main Theorem) *The (ϵ, s)-sandwiching degree of concepts with intrinsic dimesion k and σ-smooth boundary with respect to a γ-strictly subexponential distribution $\mathcal{D}$ is:*

$$\ell(\epsilon, s) \;\leq\; \widetilde{O}\left(\frac{\sigma k^{3/2} s}{(\epsilon/2)^{s+1}}\right)^{1+1/\gamma}.$$

Concept Class	This Work	Prior Work	References
Intersections of k Halfspaces	$\widetilde{O}(k^3)$	$O(k^6)$	Gopalan et al. (2010)
Functions of k Halfspaces	$\widetilde{O}(k^5)$	$\exp O(k)$	Gopalan et al. (2010) Klivans and Meka (2013) Gollakota et al. (2023)
Convex Sets in k dimensions	$\widetilde{O}(k^5)$	None*	De et al. (2023)
Degree-q PTFs in k dimensions	$\widetilde{O}(q^6 k^5)$	$\exp\exp O(q)$ or worse†	Kane (2011) Slot et al. (2024)
Functions of t concepts each with intrinsic dimension k and σ-smooth boundary‡	$\widetilde{O}(\sigma^2 t^5 k^3)$	None	–

* The work of De et al. (2023) together with results by Gopalan et al. (2010) imply the existence of upper sandwiching polynomials of degree $\exp \widetilde{O}(k)$ for k-dimensional convex sets that are truncated into a ball of radius $\mathrm{poly}(k)$.

† For PTFs, the dependence on the degree q is not made explicit in prior work (Slot et al., 2024) using FT-mollification due to Kane (2011), but is at least doubly exponential. Note that their bounds are independent of the intrinsic dimension k.

‡ These concepts may depend on different k-dimensional subspaces. Boundary smoothness is w.r.t. the Gaussian.

Table 1: Comparison between upper bounds from this work and the best known bounds in previous work on the $(\epsilon = 0.1, s = O(1))$-sandwiching degree of various geometric concept classes with respect to the standard Gaussian distribution.

In Table 1, we summarize the implications of our main result on the sandwiching degree of several geometric concept classes under the Gaussian distribution (see also Appendix C). Each of the entries in the table corresponds to the state-of-the-art results before and after our work for each of the problems outlined in Appendix D. These results are obtained by combining our main sandwiching degree bound with bounds for the boundary smoothness parameter of the corresponding classes (see Appendix C).

Our result for functions of k halfspaces also works beyond the Gaussian distribution, as long as the distribution $\mathcal{D}$ is strictly subexponential and anticoncentrated in every direction (see Theorem 23).

1.2. Our Techniques

Sandwiching Geometric Concepts by Lipschitz Functions Our first step (Lemma 9) establishes the existence of two Lipschitz functions $f_{\text{up}}, f_{\text{down}} : \mathbb{R}^d \to [-1, 1]$ satisfying

$$f_{\text{down}}(\mathbf{x}) \le f(\mathbf{x}) \le f_{\text{up}}(\mathbf{x}), \tag{1}$$

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[f_{\text{up}}(\mathbf{x}) - f_{\text{down}}(\mathbf{x})] \le \epsilon, \tag{2}$$

where f is a geometric concept and $\mathcal{D}$ is a strictly subexponential distribution.

We define one-sided relaxations $f^{+\rho}$ and $f^{-\rho}$, where $f^{+\rho}$ (resp. $f^{-\rho}$) equals 1 on points within distance ρ of the interior (resp. exterior) of f. We then construct f_{up} as a $(1/\rho)$-Lipschitz interpolation between f and $f^{+\rho}$; such an interpolation exists since their decision boundaries are separated by distance ρ. The function f_{down} is constructed analogously using $f^{-\rho}$. By construction, this immediately yields the pointwise sandwiching property.

To establish the approximation guarantee, we use the fact that if f has σ-smooth boundary, then the ρ-neighborhood of its decision boundary has measure at most $\sigma\rho$ under $\mathcal{D}$. This bounds the expected gap between $f^{+\rho}$ and $f^{-\rho}$, and hence between f_{up} and f_{down}, with ρ chosen appropriately.

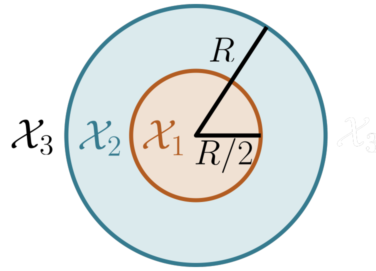

Figure 1: Our upper sandwiching polynomial for $f_{\rm up}$ is of the form $p_{\rm up} = p_1(\mathbf{x})+p_2(\mathbf{x})+\epsilon$. In region $\mathcal{X}_1$, p_1 is a pointwise ϵ-approximator for $f_{\rm up}$ and $0 \le p_2(\mathbf{x}) \le \epsilon$, so $f_{\rm up}(\mathbf{x}) \le p_{\rm up}(\mathbf{x}) \le f_{\rm up}(\mathbf{x}) + 3\epsilon$. In region $\mathcal{X}_2$, p_1 is still an pointwise ϵ-approximator for $f_{\rm up}$ and $p_2(\mathbf{x}) > 0$, so $f_{\rm up}(\mathbf{x}) \le p_1(\mathbf{x})+\epsilon \le p_{\rm up}(\mathbf{x})$. In region $\mathcal{X}_3$ we have $p_2(\mathbf{x}) \ge 1 + |p_1(\mathbf{x})| \ge f_{\rm up}(\mathbf{x}) + |p_1(\mathbf{x})|$, so $p_{\rm up}(\mathbf{x}) \ge f_{\rm up}(\mathbf{x})$.

Sandwiching Polynomials for Lipschitz Functions Next, we construct polynomial sandwiching approximations for $f_{\rm up}$ and $f_{\rm down}$. We focus on $f_{\rm up}$. By multivariate Jackson's theorem (Newman and Shapiro, 1964), there exists a polynomial p_1 that uniformly approximates $f_{\rm up}$ within a ball of radius R, leveraging the Lipschitzness of $f_{\rm up}$. A result of Ben-David et al. (2018) further controls the growth of p_1 outside this region, which is essential for bounding expectations under strictly subexponential distributions $\mathcal{D}$.

While these approximation tools are also used in Chandrasekaran et al. (2025) for learning Lipschitz neural networks under distribution shift, their approach does not yield sandwiching polynomials. To obtain an explicit upper polynomial, we construct a polynomial p_2 that is at most ϵ within a ball of radius $R/2$ and dominates p_1 outside the ball of radius R (see Figure 1). Setting $p_{\rm up} = p_1 + p_2 + \epsilon$ then yields a valid upper sandwiching polynomial. A symmetric construction gives the lower polynomial.

Comparison to Gopalan et al. (2010) The work of Gopalan et al. (2010) constructs sandwiching polynomials for arbitrary functions of k halfspaces

$$f(\mathbf{x}) = G(\mathrm{sign}(\mathbf{w}_1 \cdot \mathbf{x} - \tau_1), \mathrm{sign}(\mathbf{w}_2 \cdot \mathbf{x} - \tau_2), \ldots, \mathrm{sign}(\mathbf{w}_k \cdot \mathbf{x} - \tau_k))$$

under the Gaussian distribution. Their approach begins with the one-dimensional sandwiching polynomials of Diakonikolas et al. (2010a), which approximate the sign function pointwise within a bounded region around the origin, except near the discontinuity, and always lie above or below the sign function.

They lift these univariate polynomials to $\mathbb{R}^d$ by composing them with the linear forms $\mathbf{w}_i \cdot \mathbf{x} - \tau_i$, and then combine the resulting approximations via addition and multiplication according to the structure of G. The resulting sandwiching degree is $\mathrm{poly}(s)$, where s is the size of G when viewed as a decision tree. Since s can be exponential in the number k of inputs to G, this yields an $\exp(O(k))$ bound on the sandwiching degree.

In contrast, our approach is inherently high-dimensional and relies on tools from multivariate polynomial approximation theory, rather than composing one-dimensional sandwiching constructions. This allows us to obtain an exponential improvement in the sandwiching degree over Gopalan et al. (2010) for functions of halfspaces.

1.3. Related Work

Pseudorandomness Sandwiching polynomials play a central role in pseudorandomness (Hatami and Hoza, 2023). The goal is to derandomize randomized algorithms by replacing a high-entropy *base* distribution $\mathcal{D}$ (e.g., the uniform distribution on $\{\pm 1\}^n$) with a *pseudorandom* distribution $\mathcal{D}'$ that is generated from a short random seed, i.e., $\mathcal{D}'$ is the output distribution of a map $G : \{0,1\}^r \to \Omega$ for small r. We say that $\mathcal{D}'$ *fools* a class $\mathcal{F}$ of test functions if it preserves their expectations, namely

$$\big|\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}'}[f(\mathbf{x})]\big| \le \epsilon \quad \text{for all } f \in \mathcal{F}.$$

The *sandwiching degree* ℓ of $\mathcal{F}$ governs the seed length of a natural family of PRGs that fool $\mathcal{F}$ by ensuring that $\mathcal{D}'$ matches the moments of $\mathcal{D}$ up to degree ℓ (see Theorem 51, due to Bazzi (2009); Gollakota et al. (2023)).

Most prior work bounds sandwiching degree primarily for $\mathcal{D}$ equal to the uniform distribution on the Boolean cube (Bazzi, 2009; Razborov, 2009; Gopalan et al., 2010; Diakonikolas et al., 2010a,b; Braverman, 2011; Tal, 2017), though there are also results for continuous distributions such as the Gaussian (Kane, 2011; Gopalan et al., 2010). For each class in Table 1, our bounds yield improved moment-matching PRGs (see Section D.2). While stronger pseudorandomness techniques can achieve even shorter seeds (Meka and Zuckerman, 2010; O'Donnell et al., 2022), they typically do not yield explicit sandwiching-degree guarantees and often define generators via conditions that appear difficult to verify. In contrast, moment matching is efficiently checkable, and sandwiching-degree bounds are exactly what our learning-theoretic applications require.

Learning via Polynomial Approximation Low-degree polynomial approximation has been a core tool in computational learning theory for over three decades (Linial et al., 1993; Kalai et al., 2008; Klivans et al., 2008; Chandrasekaran et al., 2024a; Pinto Jr et al., 2025; Koehler and Wu, 2025). A classic result of Linial et al. (1993) shows that $\mathcal{L}_2$ approximation—the existence of a low-degree polynomial p with small squared error $\mathbb{E}[(p(\mathbf{x}) - f(\mathbf{x}))^2]$ under the input marginal—yields efficient learning. Subsequent work Klivans et al. (2008) showed that $\mathcal{L}_1$ approximation is sufficient even for agnostic learning. Moreover, later results provide evidence that $\mathcal{L}_1$ approximation is essentially *necessary* if one wants efficient learning with near-optimal error (Dachman-Soled et al., 2014; Diakonikolas et al., 2021).

A notable line of work ties polynomial approximation under the Gaussian distribution to geometric measure theory. In particular, Klivans et al. (2008) related Gaussian *surface area* bounds for a concept class to the existence of low-degree approximating polynomials, and hence to efficient agnostic learning under the Gaussian. Their argument combines approximation properties of the Ornstein–Uhlenbeck semigroup (Pisier, 1986; Ledoux, 1994) with surface-area bounds for several geometric classes (Ball, 1993; Nazarov, 2003). Surface area can be viewed as a limiting "boundary smoothness" parameter: it measures the first-order rate at which a random point falls within distance ρ of the boundary as $\rho \to 0$.

We can view our work as giving an analogue of the "learning via Gaussian surface area" paradigm but for *reliable* learning primitives, where the relevant notion is not merely approximation in expectation but the stronger requirement of *sandwiching*. Quantitatively, our bounds incur an explicit polynomial dependence on the intrinsic dimension parameter k of the concept class. This dependence arises from our main approximation tool (Newman and Shapiro, 1964). For example, for intersections of k halfspaces, Klivans et al. (2008) obtains only logarithmic dependence on k, whereas our bounds scale as $\mathrm{poly}(k)$, but in exchange we obtain sandwiching guarantees (and, more generally, $\mathcal{L}_s$-sandwiching for arbitrary $s \ge 1$) rather than approximation in expectation.

Learning via Sandwiching Bounds on the sandwiching degree yield efficient algorithms for a range of recently defined learning primitives that impose strong reliability requirements (Goldwasser et al., 2020; Rubinfeld and Vasilyan, 2023; Gollakota et al., 2023; Klivans et al., 2024a; Goel et al., 2024; Chandrasekaran et al., 2024b; Klivans et al., 2025a). For many of these tasks, $\mathcal{L}_1$-sandwiching is sufficient. In contrast, *current* algorithmic frameworks for PQ learning rely essentially on $\mathcal{L}_2$-sandwiching, and $\mathcal{L}_2$ bounds are also analytically convenient and may be useful beyond PQ learning. This is where the flexibility of our results matters: we obtain sandwiching guarantees in $\mathcal{L}_s$ for *arbitrary* orders $s \ge 1$.

The first appearance of sandwiching polynomial approximation in the context of efficient learning algorithms is due to Klivans and Meka (2013). Their approach combines the duality between sandwiching and fooling via moment matching (Bazzi, 2009) with the method of distances (Klebanov and Rachev, 1996; Zolotarev, 1984; Rachev et al., 2013). Because their guarantee holds for general log-concave marginals, it does not require strict subexponential tails; however, the resulting degree bound is doubly exponential in k. By contrast, our bounds are $\mathrm{poly}(k)$.

More recently, Gollakota et al. (2023) built on Klivans and Meka (2013) in the context of testable learning, obtaining sandwiching polynomials of degree $\exp(O(k))$ with bounded coefficients for functions of k halfspaces under any strictly subexponential and anticoncentrated distribution. Our bounds improve exponentially over Gollakota et al. (2023).

Learning via Boundary Smoothness The notion of *boundary smoothness* was introduced by Chandrasekaran et al. (2024b) in the context of TDS learning. Using this condition, they gave a *realizable* TDS learner for convex sets of intrinsic dimension k under Gaussian training marginals, with running time $\mathrm{poly}(d) \cdot 2^{\mathrm{poly}(k/\epsilon)}$. Their guarantee, however, assumes that both the training and test labels are generated by the *same* target convex set (i.e., realizability), and it additionally imposes a non-degeneracy condition requiring the positive region to have non-negligible Gaussian mass.

In contrast, our results yield a TDS learner running in time $d^{\mathrm{poly}(k/\epsilon)}$ that achieves *near-optimal error guarantees* in the fully agnostic setting (see Definition 39), which is the state of the art for this level of reliability.

At a high level, Chandrasekaran et al. (2024b) combines a dimension-reduction step from classical learning theory (Vempala, 2010) with a distribution-shift tester: after recovering a k-dimensional subspace, the tester forms a grid in this subspace and verifies that the test distribution assigns approximately the same mass to each grid cell as the Gaussian training marginal.

Testable Learning under Relaxed Error Guarantees As discussed above, the best known upper bounds on the computational complexity of testable agnostic learning (Definition 34) and TDS learning (Definition 40) are obtained via sandwiching approximation. There are, however, faster algorithms for both frameworks under *relaxed error guarantees* (Gollakota et al., 2024b; Diakonikolas et al., 2023; Gollakota et al., 2024a; Diakonikolas et al., 2024; Klivans et al., 2024b; Chandrasekaran et al., 2024b).

For testable learning, these improved runtimes are currently known only for halfspaces. In the TDS setting, some results extend beyond halfspaces, but either incur substantially weaker error guarantees (for example, Chandrasekaran et al. (2024b) obtain guarantees for intersections of k halfspaces with error that degrades exponentially in k), or require realizability assumptions, namely that the labels for both the training and test distributions are generated by the same function in the target concept class.

2. Preliminaries

For a function $g : \mathbb{R}^d \to \mathbb{R}$ and a distribution $\mathcal{D}$ over $\mathbb{R}^d$, we define: $\|g\|_{\mathcal{D},s} := (\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|g(\mathbf{x})|^s])^{\frac{1}{s}}$. For a set $K \subseteq \mathbb{R}^d$ and $\mathbf{x} \in \mathbb{R}^d$, we define $\mathsf{dist}(\mathbf{x}, K) = \inf_{\mathbf{x}'\in K} \|\mathbf{x}-\mathbf{x}'\|_2$. We denote with $[t]_a^b = \max\{a, \min\{t, b\}\}$ the (a, b)-clipping function. A *concept* is a function $f : \mathbb{R}^d \to \{\pm 1\}$, while a *concept class* is a set of concepts $\mathcal{F} \subseteq \{\mathbb{R}^d \to \{\pm 1\}\}$. We denote with $\mathcal{N}(0, \mathbf{I}_{d\times d})$ or simply $\mathcal{N}_d$ the standard Gaussian distribution in d dimensions. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote with x_i the i-th coordinate of $\mathbf{x}$. A polynomial p over $\mathbb{R}^d$ of degree ℓ is a function of the form $p(\mathbf{x}) = \sum_{\mathcal{I}\in\mathbb{N}^d} c_p(\mathcal{I})\, \mathbf{x}^{\mathcal{I}}$ where $x^{\mathcal{I}} = \prod_{i\in[d]} x_i^{\mathcal{I}_i}$ and $c_p(\mathcal{I}) = 0$ for any $\mathcal{I}$ with $\|\mathcal{I}\|_1 > \ell$. We denote with $\|p\|_{\mathsf{coef}}$ the quantity $\sum_{\mathcal{I}\in\mathbb{N}^d} |c_p(\mathcal{I})|$, i.e., the sum of the absolute values of the coefficients of p.

We will use dilation and erosion operations on concepts; these correspond, respectively, to taking the Minkowski sum and Minkowski subtraction of the positive regions with a Euclidean ball of fixed radius as defined below.

Definition 3 (Dilation and Erosion) *Let $f : \mathbb{R}^d \to \{\pm 1\}$. For $\rho \geq 0$, we define the ρ-dilation $f^{+\rho}$, as well as the ρ-erosion $f^{-\rho}$ of f as follows.*

$$f^{+\rho}(\mathbf{x}) = \sup_{\mathbf{z}:\|\mathbf{z}\|_2 \leq \rho} f(\mathbf{x}+\mathbf{z}), \quad f^{-\rho}(\mathbf{x}) = \inf_{\mathbf{z}:\|\mathbf{z}\|_2 \leq \rho} f(\mathbf{x}+\mathbf{z})$$

The boundary smoothness parameter relative to a distribution $\mathcal{D}$ is formally defined as follows using the dilation and erosion operations.

Definition 4 (Smooth Boundary) *Let $f : \mathbb{R}^d \to \{\pm 1\}$ and $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. For $\sigma \geq 1$, we say that f has σ-smooth boundary with respect to $\mathcal{D}$ if for any $\rho \geq 0$*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\frac{f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})}{2}\right] \leq \sigma\rho$$

Remark 5 *Note that the quantity $\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})]$ equals the probability that a sample $\mathbf{x}$ from $\mathcal{D}$ is ρ-close to the boundary of f, i.e., there exists $\mathbf{z} : \|\mathbf{z}\|_2 \leq \rho$ such that $f(\mathbf{x}+\mathbf{z}) \neq f(\mathbf{x})$.*

Our results require the following concentration assumption on the distribution $\mathcal{D}$.

Definition 6 (Strictly Subexponential Distributions) *For $\alpha, \beta, \gamma > 0$, we say that a distribution $\mathcal{D}$ over $\mathbb{R}^d$ is γ-strictly subexponential with parameters α, β if for any $\mathbf{w} \in \mathbb{S}^{d-1}$ and any $r \geq 0$,*

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[|\mathbf{w}\cdot\mathbf{x}| \geq r] \leq \alpha\,\exp(-\beta\, r^{1+\gamma}).$$

In what follows, we may omit the parameters α, β and absorb them under big-O notation.

3. Main Result

Our main result is the existence of low-degree sandwiching polynomials for a wide range of pairs of concept classes and distributions captured by the following assumption.

Assumption 7 (Valid Instances) *We will consider concept classes $\mathcal{F}$ and target distribution $\mathcal{D}$ over $\mathbb{R}^d$ that satisfy the following properties with parameters $k, d \in \mathbb{N}, \sigma \geq 1, \gamma > 0$, where $k \leq d$:*

1. *(Low Intrinsic Dimension) For any $f \in \mathcal{F}$, we have $f : \mathbb{R}^d \to \{\pm 1\}$ and there are $F : \mathbb{R}^k \to \{\pm 1\}$ and $\mathbf{W} \in \mathbb{R}^{k\times d}$ with $\mathbf{W}\mathbf{W}^\top = \mathbf{I}_{k\times k}$ such that $f(\mathbf{x}) = F(\mathbf{W}\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^d$.*

2. *(Smooth Boundary) Any $f \in \mathcal{F}$ has σ-smooth boundary with respect to $\mathcal{D}$.*

3. *(Strictly Subexponential Tails) The distribution $\mathcal{D}$ is γ-strictly subexponential.*

The first condition states that every function in the class depends on a potentially different low-dimensional subspace $\mathcal{W}$, meaning that its values remain unchanged when the input $\mathbf{x}$ moves in some direction that is perpendicular to the subspace $\mathcal{W}$. The second condition ensures that the probability of a sample from $\mathcal{D}$ lying ρ-near the decision boundary of f scales linearly with the distance ρ, with rate σ. Finally, the third condition ensures that the distribution $\mathcal{D}$ is sufficiently concentrated in every direction.

We are now ready to state our main theorem.

Theorem 8 (Main Theorem, restated) *Let $\mathcal{F}$ be some concept class that satisfies Assumption 7 with parameters k, d, σ with respect to a γ-strictly subexponential distribution $\mathcal{D}$ over $\mathbb{R}^d$. Then, for any $\epsilon \in (0,1)$ and $s \geq 1$, the (ϵ, s)-sandwiching degree of $\mathcal{F}$ satisfies*

$$\ell(\epsilon, s) \;\leq\; \widetilde{O}\left(\frac{\sigma k^{3/2} s}{(\epsilon/2)^{s+1}}\right)^{1+1/\gamma}.$$

Moreover, the sum of the absolute values of the coefficients of the corresponding sandwiching polynomials is upper bounded by B where $\log B = \widetilde{O}(\ell(\epsilon, s)) \cdot \mathrm{poly}(\log d)$.

Note that the above theorem works for any choice of $s \geq 1$, meaning that it establishes the existence of low-degree sandwiching polynomials $p_{\text{up}}, p_{\text{down}}$ with the property $\|p_{\text{up}} - p_{\text{down}}\|_{\mathcal{D},s} \leq \epsilon$, for any choice of s. Choosing larger values for s gives stronger notions of approximation, which is sometimes useful for downstream applications.

For example, the first positive results for TDS learning (Definition 40, see Klivans et al. (2024a)) were based on $\mathcal{L}_2$-sandwiching, although subsequent work showed that $\mathcal{L}_1$-sandwiching suffices Chandrasekaran et al. (2024b). For PQ learning (Definition 42), Goel et al. (2024) showed that $\mathcal{L}_2$-sandwiching suffices for efficiency, but it remains an open question whether a similar result can be obtained via $\mathcal{L}_1$-sandwiching. Our work demonstrates that, for a wide range of instances, the gap between the sandwiching degree of different orders is much narrower than previously believed.

Concretely, we obtain the first $\mathcal{L}_2$-sandwiching degree bound for polynomial threshold functions. Hence, we provide the first non-trivial result for PQ learning of the class of PTFs with intrinsic dimension k and degree q such that $q^6 k^3 = O(d^{1-c})$ for any $c > 0$ with respect to the Gaussian distribution (see Theorems 32 and 43).[1] Note that bounds on the $\mathcal{L}_1$-sandwiching of PTFs with respect to both the Gaussian distribution, as well as the uniform distribution on the boolean hypercube have been provided in prior work Diakonikolas et al. (2010b); Kane (2011); Slot et al. (2024).

In the rest of this section, we provide the proof of Theorem 8.

3.1. Full-Dimensional Case

We will first prove Theorem 8 in the special case where $k = d$. Throughout this section, we consider $\mathcal{F}, \mathcal{D}$ to be as specified in the premise of Theorem 8, with $k = d$. The main idea for the proof is to sandwich an arbitrary element f of $\mathcal{F}$ by functions $f_{\text{up}}, f_{\text{down}}$ that are Lipschitz and are close to f in expectation. It will then be sufficient to provide sandwiching polynomials for $f_{\text{up}}, f_{\text{down}}$. The existence of such functions $f_{\text{up}}, f_{\text{down}}$ is ensured by the following lemma.

Lemma 9 *Let f be a function with σ-smooth boundary w.r.t. $\mathcal{D}$, $\epsilon \in (0,1)$ and $s \geq 1$. Then, there exist functions $f_{\text{up}}, f_{\text{down}} : \mathbb{R}^d \to \mathbb{R}$ that are L-Lipschitz for $L = 2\,\sigma\,(\frac{2}{\epsilon})^s$ such that for any $\mathbf{x} \in \mathbb{R}^d$ and any $s \geq 1$, we have:*

$$f_{\text{down}}(\mathbf{x}) \leq f(\mathbf{x}) \leq f_{\text{up}}(\mathbf{x}) \tag{3}$$

$$\|f_{\text{up}} - f_{\text{down}}\|_{\mathcal{D},s} \leq \epsilon \tag{4}$$

Proof We let $\rho = \frac{1}{\sigma} \cdot (\frac{\epsilon}{2})^s$ and define the following auxiliary sets.

$$S_{\text{in}} := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = 1\}, \quad S_{\text{far-in}} := \{\mathbf{x} \in \mathbb{R}^d : f^{-\rho}(\mathbf{x}) = 1\}$$
$$S_{\text{out}} := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = -1\}, \quad S_{\text{far-out}} := \{\mathbf{x} \in \mathbb{R}^d : f^{+\rho}(\mathbf{x}) = -1\}$$

1. For $q^6 k^3 = \Omega(d)$, the sandwiching degree becomes $\widetilde{\Omega}(d)$ (see Theorem 32), and a covering-based algorithm improves upon the result provided by Theorem 43.

We also define the functions $g_{\text{up}}, g_{\text{down}} : \mathbb{R}^d \to \mathbb{R}$ as follows.

$$g_{\text{up}}(\mathbf{x}) = \mathsf{dist}(\mathbf{x}, S_{\text{far-out}}) - \mathsf{dist}(\mathbf{x}, S_{\text{in}})$$
$$g_{\text{down}}(\mathbf{x}) = \mathsf{dist}(\mathbf{x}, S_{\text{out}}) - \mathsf{dist}(\mathbf{x}, S_{\text{far-in}})$$

Observe that the functions $g_{\text{up}}, g_{\text{down}}$ are both 2-Lipschitz, regardless of the sets S_{in}, S_{out}, $S_{\text{far-in}}$, $S_{\text{far-out}}$. Therefore, the following choices for $f_{\text{up}}, f_{\text{down}}$ are L-Lipschitz for $L = 2/\rho = 2\sigma(2/\epsilon)^s$.

$$f_{\text{up}}(\mathbf{x}) = \left[\frac{g_{\text{up}}(\mathbf{x})}{\rho}\right]_{-1}^{+1} \qquad f_{\text{down}}(\mathbf{x}) = \left[\frac{g_{\text{down}}(\mathbf{x})}{\rho}\right]_{-1}^{+1}$$

We will show that the following property is true.

$$f^{-\rho}(\mathbf{x}) \le f_{\text{down}}(\mathbf{x}) \le f(\mathbf{x}) \le f_{\text{up}}(\mathbf{x}) \le f^{+\rho}(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathbb{R}^d \tag{5}$$

We will focus on f_{up} since the argument for f_{down} is analogous.

- ($f(\mathbf{x}) \le f_{\text{up}}(\mathbf{x})$). If $f(\mathbf{x}) = 1$, then $\mathbf{x} \in S_{\text{in}}$ and $\mathsf{dist}(\mathbf{x}, S_{\text{far-out}}) \ge \rho$. Therefore, $g_{\text{up}}(\mathbf{x}) \ge \rho$ and $f_{\text{up}}(\mathbf{x}) = f(\mathbf{x}) = 1$ for any $\mathbf{x} \in S_{\text{in}}$. For $\mathbf{x} \notin S_{\text{in}}$, we have $f(\mathbf{x}) = -1 \le f_{\text{up}}(\mathbf{x})$.
- ($f_{\text{up}}(\mathbf{x}) \le f^{+\rho}(\mathbf{x})$). If $f^{+\rho}(\mathbf{x}) = -1$, then $\mathbf{x} \in S_{\text{far-out}}$ and $\mathsf{dist}(\mathbf{x}, S_{\text{in}}) \ge \rho$. Therefore, $g_{\text{up}}(\mathbf{x}) \le -\rho$ and $f_{\text{up}}(\mathbf{x}) = -1 = f^{+\rho}(\mathbf{x})$. If $f^{+\rho}(\mathbf{x}) = 1$, then $f_{\text{up}}(\mathbf{x}) \le 1 \le f^{+\rho}(\mathbf{x})$.

It remains to bound the quantity $\|f_{\text{up}} - f_{\text{down}}\|_{\mathcal{D},s}$. We have the following:

$$\|f_{\text{up}} - f_{\text{down}}\|_{\mathcal{D},s} \le \|f^{+\rho} - f^{-\rho}\|_{\mathcal{D},s} = \Big(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[\Big|f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})\Big|^s\Big]\Big)^{1/s}$$

Observe that for any $\mathbf{x}$ we have $|f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})|^s \in \{0, 2^s\}$, and therefore we may use the following simplification:

$$|f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})|^s = 2^s \cdot \frac{f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})}{2}$$

Therefore, we overall obtain:

$$\begin{aligned}\|f_{\text{up}} - f_{\text{down}}\|_{\mathcal{D},s} &\le 2\cdot\Big(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[\frac{f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})}{2}\Big]\Big)^{1/s} \\ &\le 2\cdot(\sigma\rho)^{1/s} = \epsilon\,,\end{aligned}$$

where the second inequality follows from Assumption 7. ■

So far, we have obtained a pair of functions that sandwich f and are Lipschitz continuous, but are not necessarily polynomials. To obtain low-degree sandwiching polynomials, we apply the following lemma, first observed in Chandrasekaran et al. (2025) in the context of learning Lipschitz neural networks under distribution shift. Its proof combines a classical approximation-theoretic result—multivariate Jackson's theorem Newman and Shapiro (1964)—with a bound from Ben-David et al. (2018). In particular, Jackson's theorem guarantees the existence of low-degree $\mathcal{L}_\infty$-approximators for Lipschitz functions over bounded domains, while Ben-David et al. (2018) controls the coefficients of such polynomials, ensuring bounded behavior outside the approximation domain.

Lemma 10 (Chandrasekaran et al. (2025); Newman and Shapiro (1964); Ben-David et al. (2018)) *Let $g : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz and let $\epsilon \in (0,1)$, $R \geq 1$. Then, there exists polynomial p of degree $\ell = O(LRd/\epsilon)$ with coefficients bounded by $(d\ell)^{O(\ell)}$ in absolute value and with the following properties:*

$$|g(\mathbf{x}) - p(\mathbf{x})| \leq \epsilon, \textit{ for all } \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq R \tag{6}$$

$$|p(\mathbf{x})| \leq (d\ell)^{O(\ell)} \left(\frac{\|\mathbf{x}\|_2}{R}\right)^{\ell}, \textit{ for all } \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 > R \tag{7}$$

The polynomials from the above lemma are used in Chandrasekaran et al. (2025) to obtain a TDS learning algorithm for Lipschitz neural networks under strictly subexponential distributions. To achieve this, they crucially use the fact that the tails of the target distribution are strictly subexponential, to ensure that the expectation of the polynomial approximators outside the approximation domain vanishes for some choice of the parameters ℓ, R, while still satisfying the condition $\ell = O(LRd/\epsilon)$.

In particular, observe that for γ-strictly subexponential distributions and p as in Lemma 10, we have $\mathbb{E}[(p(\mathbf{x}))^2] \leq (d\ell)^{O(\ell)}$ and $\mathbb{P}[\|\mathbf{x}\|_2 > R] \leq \exp(-\Omega(R/\sqrt{d})^{1+\gamma})$. Therefore:

$$\mathbb{E}[|p(\mathbf{x})| \, \mathbb{1}\{\|\mathbf{x}\|_2 > R\}] \overset{\ell\to\infty}{\to} 0$$

We build on the approach of Chandrasekaran et al. (2025) by taking the sum $p_1 + p_2 + \epsilon$ of the pointwise approximator p_1 provided by Lemma 10 and a closed-form polynomial p_2 that vanishes near the origin, and dominates p_1 outside the domain of approximation. This ensures that the polynomial $p_1 + p_2 + \epsilon$ is an upper sandwiching polynomial for our Lipschitz function f_{up} (see Figure 1). A symmetric argument provides a lower sandwiching polynomial.

Proof of Theorem 8 for $k = d$. Let $f \in \mathcal{F}$. We will use the functions $f_{\text{up}}, f_{\text{down}}$ from Lemma 9. In particular, it suffices to show the existence of an upper sandwiching polynomial for f_{up} and a lower sandwiching polynomial for f_{down}. We focus on f_{up}, since the proof for f_{down} will follow from a symmetric argument, i.e., by applying the argument below to $-f$.

Let R be some large enough parameter which will be chosen later. Let $L = 2\sigma(2/\epsilon)^s$, and let p_1 be a polynomial of degree $\ell_1 = C_1 LRd/\epsilon$ with the properties specified in Lemma 10. We set $\ell_2 = C_2(\ell_1 \log(d\ell_1) + \log(1/\epsilon))$ and choose the polynomial p_{up} as follows:

$$p_{\text{up}}(\mathbf{x}) = p_1(\mathbf{x}) + p_2(\mathbf{x}) + \epsilon, \quad \text{where } p_2(\mathbf{x}) = \epsilon \left(\frac{2\|\mathbf{x}\|_2}{R}\right)^{2\ell_2} \tag{8}$$

For appropriate choices for the universal constants $C_1, C_2 \geq 1$ we have the following:

- For any $\mathbf{x}$ with $\|\mathbf{x}\|_2 \leq R/2$, we have $p_{\text{up}}(\mathbf{x}) \leq p_1(\mathbf{x}) + 2\epsilon \leq f_{\text{up}}(\mathbf{x}) + 3\epsilon$ and $p_{\text{up}}(\mathbf{x}) \geq f_{\text{up}}(\mathbf{x})$, since $|p_1(\mathbf{x}) - f_{\text{up}}(\mathbf{x})| \leq \epsilon$ and $p_2(\mathbf{x}) \in [0, \epsilon]$.
- For any $\mathbf{x}$ with $\|\mathbf{x}\|_2 \in (R/2, R]$ we have $p_{\text{up}}(\mathbf{x}) \geq f_{\text{up}}(\mathbf{x})$, since $|p_1(\mathbf{x}) - f_{\text{up}}(\mathbf{x})| \leq \epsilon$.
- For any $\mathbf{x}$ with $\|\mathbf{x}\|_2 > R$ we have $p_{\text{up}}(\mathbf{x}) \geq f_{\text{up}}(\mathbf{x})$, since $p_2(\mathbf{x}) \geq 1 + |p_1(\mathbf{x})|$.

So far, we have established that $f_{\text{up}}(\mathbf{x}) \leq p_{\text{up}}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$. We will now show that the quantity $\|p_{\text{up}} - f_{\text{up}}\|_{\mathcal{D},s}$ is bounded by $O(\epsilon)$ for some appropriate choice of $R = \widetilde{O}((Ls/\epsilon)^{1/\gamma} d^{0.5+1.5/\gamma})$.[2]

2. If we wish to achieve error ϵ instead of $O(\epsilon)$, we may just substitute $\epsilon' = \epsilon/C$ for some universal constant $C \geq 1$.

Due to the properties of $p_{\rm up}$ in the region $\|\mathbf{x}\|_2 \le R/2$, it suffices to account for the region $\|\mathbf{x}\|_2 > R/2$. In particular, we have:

$$\begin{aligned}\|p_{\rm up} - f_{\rm up}\|_{\mathcal{D},s} &\le 3\epsilon + \Big(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[(p_{\rm up}(\mathbf{x}) - f_{\rm up}(\mathbf{x}))^s\, \mathbb{1}\{\|\mathbf{x}\|_2 > R/2\}\Big]\Big)^{1/s} \\ &\le 3\epsilon + \Big(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[(p_{\rm up}(\mathbf{x}) - f_{\rm up}(\mathbf{x}))^{2s}\Big]\Big)^{\frac{1}{2s}}\Big(\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}\Big[\|\mathbf{x}\|_2 > R/2\Big]\Big)^{\frac{1}{2s}} \\ &= 3\epsilon + \|p_{\rm up} - f_{\rm up}\|_{\mathcal{D},2s}\cdot\Big(\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}\Big[\|\mathbf{x}\|_2 > R/2\Big]\Big)^{\frac{1}{2s}} \end{aligned} \tag{9}$$

where the second inequality follows from the Cauchy-Schwarz inequality.

Since the distribution $\mathcal{D}$ is γ-strictly subexponential, we have the following for some constants $\alpha, \beta > 0$:

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}\Big[\|\mathbf{x}\|_2 > R/2\Big] &\le \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}\Big[\|\mathbf{x}\|_\infty > R/(2\sqrt{d})\Big] \\ &\le d\cdot \sup_{\mathbf{w}\in\mathbb{S}^{d-1}} \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}\Big[|\mathbf{w}\cdot\mathbf{x}| > R/(2\sqrt{d})\Big] \\ &\le \alpha\, d\cdot\exp\Big(-\beta\Big(\frac{R}{2\sqrt{d}}\Big)^{1+\gamma}\Big)\end{aligned} \tag{10}$$

By combining the above tail bound with the bounds on $|p_1(\mathbf{x})|$ and $|p_2(\mathbf{x})|$ (see Lemma 10 and (8)), we also obtain the following bound:

$$\|p_{\rm up} - f_{\rm up}\|_{\mathcal{D},2s} \le (d\ell_2)^{O(d\ell_2)} = \exp\Big(\widetilde{O}\Big(\frac{LRd}{\epsilon}\Big)\Big) \tag{11}$$

Therefore we may combine (9), (10), and (11) to obtain that there is some appropriate choice for R with $R = \widetilde{O}((Ls/\epsilon)^{1/\gamma} d^{0.5+1.5/\gamma})$ such that $\|p_{\rm up} - f_{\rm up}\|_{\mathcal{D},s} \le O(\epsilon)$. The overall bound on the degree of $p_{\rm up}$ is:

$$\deg(p_{\rm up}) = 2\ell_2 = \widetilde{O}\left(\frac{\sigma d^{3/2} s}{(\epsilon/2)^{s+1}}\right)^{1+1/\gamma},$$

as desired. ■

3.2. Concepts with Low Intrinsic Dimension

To complete the proof of Theorem 8, it remains to account for the case where $k \le d$. To this end, we show that for a function $f(\mathbf{x}) = F(\mathbf{W}\mathbf{x})$ where $\mathbf{W}\mathbf{W}^\top = \mathbf{I}_{k\times k}$, the boundary smoothness parameter of f is equal to the boundary smoothness parameter associated with F. Our result then follows from the observation that for any polynomial p, the function $p(\mathbf{W}\mathbf{x})$ is a polynomial of the same degree.

Proposition 11 *Let $f \in \mathcal{F}$, where $\mathcal{F}$ satisfies Assumption 7 with parameters k, d, σ with respect to a distribution $\mathcal{D}$ over $\mathbb{R}^d$. Let $F : \mathbb{R}^k \to \{\pm 1\}$ be such that $f(\mathbf{x}) = F(\mathbf{W}\mathbf{x})$ where $\mathbf{W} \in \mathbb{R}^{k\times d}$ and $\mathbf{W}\mathbf{W}^\top = \mathbf{I}_{k\times k}$. Then F has σ-smooth boundary with respect to the distribution of $\mathbf{W}\mathbf{x}$ when $\mathbf{x}\sim\mathcal{D}$.*

Proof It suffices to show that $f^{\pm\rho}(\mathbf{x}) = F^{\pm\rho}(\mathbf{W}\mathbf{x})$ for all $\mathbf{x}\in\mathbb{R}^d$. We have the following:

$$\begin{aligned} f^{+\rho}(\mathbf{x}) &= \sup_{\mathbf{z}\in\mathbb{R}^d:\|\mathbf{z}\|_2\le\rho} f(\mathbf{x}+\mathbf{z}) \\ &= \sup_{\mathbf{z}\in\mathbb{R}^d:\|\mathbf{z}\|_2\le\rho} F(\mathbf{W}\mathbf{x}+\mathbf{W}\mathbf{z}) \\ &= \sup_{\mathbf{t}\in\mathbb{R}^k:\|\mathbf{t}\|_2\le\rho} F(\mathbf{W}\mathbf{x}+\mathbf{t}) = F^{+\rho}(\mathbf{W}\mathbf{x}),\end{aligned}$$

where we have used the fact that the image of the unit ball in d dimensions under $\mathbf{W}$ is the unit ball in k dimensions. Similarly, we can show that $f^{-\rho}(\mathbf{x}) = F^{-\rho}(\mathbf{W}\mathbf{x})$. ■

Proof of Theorem 8. In Section 3.1 we proved Theorem 8 in the special case where $k = d$. Let $f \in \mathcal{F}$ with $f(\mathbf{x}) = F(\mathbf{W}\mathbf{x})$. By the $k = d$ case, there exist (ϵ, s)-sandwiching polynomials $P_{\text{up}}, P_{\text{down}}$ for F with respect to the distribution of $\mathbf{W}\mathbf{x}$ when $\mathbf{x} \sim \mathcal{D}$, with the desired degree bound ℓ.

Define $p_{\text{up}}(\mathbf{x}) = P_{\text{up}}(\mathbf{W}\mathbf{x})$ and $p_{\text{down}}(\mathbf{x}) = P_{\text{down}}(\mathbf{W}\mathbf{x})$. Then $p_{\text{up}}, p_{\text{down}}$ are (ϵ, s)-sandwiching polynomials for f with respect to $\mathcal{D}$. Moreover, we have $\|p_{\text{up}}\|_{\mathsf{coef}}, \|p_{\text{down}}\|_{\mathsf{coef}} \leq (d\ell)^{O(\ell)}$, because the coefficients of the linear polynomial $p_{\text{lin}}(\mathbf{x}) = \mathbf{W}\mathbf{x}$ satisfy $\|p_{\text{lin}}\|_{\mathsf{coef}} \leq \text{poly}(d)$, and we also have $\|p_{\text{up}}\|_{\mathsf{coef}} \leq \text{poly}(\|P_{\text{up}}\|_{\mathsf{coef}}, (\|p_{\text{lin}}\|_{\mathsf{coef}})^{O(\deg(P_{\text{up}}))})$, where $\|P_{\text{up}}\|_{\mathsf{coef}} \leq (k\ell)^{O(\ell)}$. ■

Acknowledgments

The authors were supported in part by the NSF AI Institute for Foundations of Machine Learning (IFML). Konstantinos Stavropoulos was also supported in part by the 2025 Apple Scholars in AI/ML PhD fellowship.

References

Keith Ball. The reverse isoperimetric problem for gaussian measure. *Discrete Comput. Geom.*, 10(4): 411–420, dec 1993. ISSN 0179-5376. doi: 10.1007/BF02573986. URL https://doi.org/10.1007/BF02573986.

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

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

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.

Shalev Ben-David, Adam Bouland, Ankit Garg, and Robin Kothari. Classical lower bounds from quantum upper bounds. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 339–349. IEEE, 2018.

Sergey G Bobkov. Some generalizations of prokhorov's results on khinchin-type inequalities for polynomials. *Theory of Probability & Its Applications*, 45(4):644–647, 2001.

Mark Braverman. Poly-logarithmic independence fools bounded-depth boolean circuits. *Communications of the ACM*, 54(4):108–115, 2011.

Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. Pac learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002. ISSN 0304-3975. doi: https://doi.org/10.1016/S0304-3975(01)00403-0. URL https://www.sciencedirect.com/science/article/pii/S0304397501004030. Algorithmic Learning Theory.

Gautam Chandrasekaran, Adam Klivans, Vasilis Kontonis, Raghu Meka, and Konstantinos Stavropoulos. Smoothed analysis for learning concepts with low intrinsic dimension. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 876–922. PMLR, 2024a.

Gautam Chandrasekaran, Adam Klivans, Vasilis Kontonis, Konstantinos Stavropoulos, and Arsen Vasilyan. Efficient discrepancy testing for learning with distribution shift. *Advances in Neural Information Processing Systems*, 37:137263–137308, 2024b.

Gautam Chandrasekaran, Adam Klivans, Lin Lin Lee, and Konstantinos Stavropoulos. Learning neural networks with distribution shift: Efficiently certifiable guarantees. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ed7zI29lRF.

Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 498–511. SIAM, 2014.

Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 105–117, 2016.

Anindya De, Shivam Nadimpalli, and Rocco A Servedio. Gaussian approximation of convex sets by intersections of halfspaces. *arXiv preprint arXiv:2311.08575*, 2023.

Ilias Diakonikolas and Daniel Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. In *Conference on Learning Theory*, pages 4258–4282. PMLR, 2022.

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

Ilias Diakonikolas, Daniel M Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 11–20. IEEE, 2010b.

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.

Ilias Diakonikolas, Daniel Kane, Pasin Manurangsi, and Lisheng Ren. Cryptographic hardness of learning halfspaces with massart noise. In *Advances in Neural Information Processing Systems*, 2022.

Ilias Diakonikolas, Daniel Kane, Vasilis Kontonis, Sihan Liu, and Nikos Zarifis. Efficient testable learning of halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 36, 2023.

Ilias Diakonikolas, Daniel Kane, Sihan Liu, and Nikos Zarifis. Testable learning of general halfspaces with adversarial label noise. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1308–1335. PMLR, 2024.

Herbert Federer. *Geometric measure theory*. Springer, 1969.

Dietmar Ferger. Optimal constants in the marcinkiewicz–zygmund inequalities. *Statistics & Probability Letters*, 84:96–101, 2014. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2013.09.029. URL https://www.sciencedirect.com/science/article/pii/S0167715213003271.

Surbhi Goel, Abhishek Shetty, Konstantinos Stavropoulos, and Arsen Vasilyan. Tolerant algorithms for learning with arbitrary covariate shift. *arXiv preprint arXiv:2406.02742*, 2024.

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

Aravind Gollakota, Adam R Klivans, 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.

Aravind Gollakota, Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Tester-learners for halfspaces: Universal algorithms. *Advances in Neural Information Processing Systems*, 36, 2024a.

Aravind Gollakota, Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. An efficient tester-learner for halfspaces. *The Twelfth International Conference on Learning Representations*, 2024b.

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.

Pooya Hatami and William Hoza. Theory of unconditional pseudorandom generators. In *Electron. Colloquium Comput. Complex., TR23-019*, volume 1, page 1, 2023.

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

Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

Daniel M. Kane. k-independent gaussians fool polynomial threshold functions, 2011.

Daniel M Kane. The average sensitivity of an intersection of half spaces. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 437–440, 2014.

Michael J. Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM J. Comput.*, 22(4): 807–837, 1993. doi: 10.1137/0222052. URL https://doi.org/10.1137/0222052.

LB Klebanov and ST Rachev. Proximity of probability measures with common marginals in a finite number of directions. *Lecture Notes-Monograph Series*, pages 162–174, 1996.

Adam Klivans and Raghu Meka. Moment-matching polynomials. *arXiv preprint arXiv:1301.0820*, 2013.

Adam Klivans, Konstantinos Stavropoulos, Kevin Tian, and Arsen Vasilyan. The power of iterative filtering for supervised learning with (heavy) contamination. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=E7knuYAvpt.

Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Learning constant-depth circuits in malicious noise models. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3253–3263. PMLR, 30 Jun–04 Jul 2025b. URL https://proceedings.mlr.press/v291/klivans25a.html.

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.

Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. *The Thirty Seventh Annual Conference on Learning Theory*, 2024a.

Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Learning intersections of halfspaces with distribution shift: Improved algorithms and sq lower bounds. *The Thirty Seventh Annual Conference on Learning Theory*, 2024b.

Frederic Koehler and Beining Wu. Constructive approximation under carleman's condition, with applications to smoothed analysis. *arXiv preprint arXiv:2512.04371*, 2025.

Michel Ledoux. Semigroup proofs of the isoperimetric inequality in euclidean and gauss space. *Bulletin des sciences mathématiques*, 118(6):485–510, 1994.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM (JACM)*, 40(3):607–620, 1993.

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

Raghu Meka and David Zuckerman. Pseudorandom generators for polynomial threshold functions. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, pages 427–436, 2010.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, second edition, 2018.

Fedor Nazarov. On the maximal perimeter of a convex set in with respect to a gaussian measure. In *Geometric Aspects of Functional Analysis: Israel Seminar 2001-2002*, pages 169–187. Springer, 2003.

D. J. Newman and H. S. Shapiro. *Jackson's Theorem in Higher Dimensions*, pages 208–219. Springer Basel, Basel, 1964.

Ryan O'Donnell, Rocco A Servedio, and Li-Yang Tan. Fooling polytopes. *ACM Journal of the ACM (JACM)*, 69(2):1–37, 2022.

Renato Ferreira Pinto Jr, Cassandra Marcussen, Elchanan Mossel, and Shivam Nadimpalli. Learning and testing convex functions. *arXiv preprint arXiv:2511.11498*, 2025.

Gilles Pisier. Probabilistic methods in the geometry of banach spaces. *Lecture Notes in Mathematics*, pages 167–241, 1986.

Svetlozar T Rachev, Lev B Klebanov, Stoyan V Stoyanov, and Frank Fabozzi. *The methods of distances in the theory of probability and statistics*, volume 10. Springer, 2013.

Alexander Razborov. A simple proof of bazzi's theorem. *ACM Transactions on Computation Theory (TOCT)*, 1(1):1–5, 2009.

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

Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8: 45, 2014.

Lucas Slot, Stefan Tiegel, and Manuel Wiedmer. Testably learning polynomial threshold functions. *Advances in Neural Information Processing Systems*, 37:3781–3831, 2024.

Avishay Tal. Tight bounds on the Fourier spectrum of AC^0. In *32nd Computational Complexity Conference (CCC 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Leslie G. Valiant. Learning disjunction of conjunctions. In Aravind K. Joshi, editor, *Proceedings of the 9th International Joint Conference on Artificial Intelligence. Los Angeles, CA, USA, August 1985*, pages 560–566. Morgan Kaufmann, 1985. URL http://ijcai.org/Proceedings/85-1/Papers/107.pdf.

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

Vladimir Mikhailovich Zolotarev. Probability metrics. *Theory of Probability & Its Applications*, 28(2): 278–302, 1984.

Appendix A. Gaussian Surface Area and Rectifiable Boundaries

In Appendix C, we provide sandwiching degree bounds for several fundamental concept classes. To this end, it is important to obtain sharp bounds on the boundary smoothness parameter. Prior work in geometric measure theory as well as learning theory (see, e.g., Nazarov (2003); Klivans et al. (2008)) has focused on the notion of Gaussian surface area, which is essentially an (one-sided) asymptotic version of the boundary smoothness parameter defined as follows.

Definition 12 (Gaussian Surface Area) *Let $f : \mathbb{R}^d \to \{\pm 1\}$. The Gaussian Surface Area of f is defined as follows:*

$$\mathsf{GSA}(f) := \lim_{\rho \to 0^+} \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{N}_d}[f^{+\rho}(\mathbf{x}) - f(\mathbf{x})]}{2\rho}$$

In particular, the following result by Nazarov gives a sharp bound on the Gaussian surface area of halfspace intersections.

Fact 13 (Nazarov, see Klivans et al. (2008)) *Let $f : \mathbb{R}^d \to \{\pm 1\}$ be an intersection of k halfspaces. Then:*

$$\mathsf{GSA}(f) \le \sqrt{2\ln k} + 2$$

In order to be able to transform bounds on the surface area to bounds on boundary smoothness, we require the following regularity condition from geometric measure theory.

Definition 14 (Rectifiable Set (Federer, 1969)) *Let $K \subseteq \mathbb{R}^d$. We say that K is $(d-1)$-rectifiable if there exists a countable collection of Lipschitz continuous maps $\{G_i : \mathbb{R}^{d-1} \to \mathbb{R}^d\}_i$ such that:*

$$\mathcal{H}^{d-1}\Big(K \setminus \bigcup_{i=1}^{\infty} G_i(\mathbb{R}^{d-1})\Big) = 0\,,$$

where $\mathcal{H}^{d-1}$ is the $(d-1)$-dimensional Hausdorff measure.

The above condition defines rectifiable sets as the sets in $\mathbb{R}^d$ that effectively behave like countable unions of surfaces defined by Lipschitz continuous transformations from $\mathbb{R}^{d-1}$ to $\mathbb{R}^d$.

We will use a result from geometric measure theory that equates the Gaussian surface area of a concept to the Gaussian-weighted surface integral of its decision boundary, whenever the decision boundary is rectifiable. We provide the formal definition of the decision boundary, as well as the aforementioned result below.

Definition 15 (Boundary of a Concept) *Let $f : \mathbb{R}^d \to \{\pm 1\}$. We define the boundary ∂f of f to be the set of points $\mathbf{x}$ where for any $\rho > 0$ there exists $\mathbf{z} \in \mathbb{R}^d$ with $\|\mathbf{z}\|_2 \le \rho$ and $f(\mathbf{x}) \neq f(\mathbf{x}+\mathbf{z})$.*

Fact 16 (See Federer (1969)) *Let $f : \mathbb{R}^d \to \{\pm 1\}$ such that ∂f is $(d-1)$-rectifiable. Then we have:*

$$\mathsf{GSA}(f) = \int_{\partial f} \varphi(\mathbf{x})\, d\mathcal{H}^{d-1}(\mathbf{x})\,,$$

where φ is the standard Gaussian density $\varphi(\mathbf{x}) = (2\pi)^{-d/2}\exp(-\|\mathbf{x}\|_2^2/2)$ and $\mathcal{H}^{d-1}$ is the $(d-1)$-dimensional Hausdorff measure. Moreover, if f indicates a convex set then ∂f is $(d-1)$-rectifiable.

Appendix B. Polynomial Concentration under Subexponential Distributions

Throughout this work, we have focused on distributions that are strictly subexponential. For our applications to learning theory, we use the following fact regarding the uniform convergence for polynomials with bounded coefficients, with respect to subexponential distributions.

Fact 17 *Let $\mathcal{P}(\ell, B)$ be the family of polynomials p of degree at most ℓ and $\|p\|_{\mathsf{coef}} \le B$. Let $\mathcal{D}$ be some distribution over $\mathbb{R}^d$ that is subexponential and let $\bar{\mathcal{D}}$ be any labeled distribution over $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbb{R}^d$-marginal is $\mathcal{D}$. Then, the following holds with probability at least $1-\delta$ over a set $\bar{S}$ of $m \ge (B/\epsilon)^{O(1)}(d \cdot \log 1/\delta)^{\widetilde{O}(\ell)}$ samples drawn independently from $\bar{\mathcal{D}}$:*

$$\left| \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}\Big[(y-p(\mathbf{x}))^2\Big] - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\bar{S}}\Big[(y-p(\mathbf{x}))^2\Big] \right| \le \epsilon\,, \quad \textit{for all } p \in \mathcal{P}$$

The proof of Fact 17 is based on standard thresholding arguments (e.g., see the proof of Lemma 59 in Klivans et al. (2024a)) combined with classical generalization results for kernelized regression (see, e.g., Mohri et al. (2018)), as well as the following result on the concentration of monomials under subexponential distributions.

Fact 18 *Let $\mathcal{D}$ be some distribution over $\mathbb{R}^d$ that is subexponential. Then, the following holds with probability at least $1-\delta$ over a set S of $m \ge (1/\epsilon)^2(\log 1/\delta)^{\widetilde{O}(\ell)}$ independent samples from $\mathcal{D}$:*

$$\left| \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\mathbf{x}^{\mathcal{I}}\Big] - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim S}\Big[\mathbf{x}^{\mathcal{I}}\Big] \right| \le \epsilon\,, \quad \textit{for all } \mathcal{I} \in \mathbb{N}^d : \|\mathcal{I}\|_1 \le \ell$$

The proof of Fact 18 is an application of the following result from probability theory.

Theorem 19 (Marcinkiewicz–Zygmund, see Ferger (2014)) *Let $\mathcal{D}$ be a distribution over the reals, let $q \ge 2$, and let S be a set of m i.i.d. examples from $\mathcal{D}$. We have the following:*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^{\otimes m}}\left[\left|\frac{1}{m}\sum_{x\in S}\Big(x - \mathop{\mathbb{E}}_{x'\sim\mathcal{D}}[x']\Big)\right|^q\right] \le \frac{2\,q^{q/2}}{m^{q/2}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left|x - \mathop{\mathbb{E}}_{x'\sim\mathcal{D}}[x']\right|^q\right]$$

In particular, to obtain Fact 18, we may use Markov's inequality on the random variable $|x_i^\ell - \mathbb{E}[x_i^\ell]|^q$, where $\mathbf{x}$ follows some subexponential distribution, and then apply Theorem 19 on the random variable $x = x_i^\ell$.

Appendix C. Sandwiching Degree Bounds for Fundamental Concept Classes

In this section, we provide a suite of new or improved sandwiching degree bounds for several important concept classes.

One especially appealing property of boundary smoothness is that it behaves additively under composition. In particular, the following proposition shows that composing an arbitrary Boolean function with concepts of smooth boundary preserves boundary smoothness, with a smoothness parameter equal to the sum of the individual parameters.

Proposition 20 *Let $g_1, g_2, \ldots, g_k : \mathbb{R}^d \to \{\pm 1\}$ be such that g_i has σ_i-smooth boundary with respect to a distribution $\mathcal{D}$ over $\mathbb{R}^d$. Then, for any $F : \mathbb{R}^k \to \{\pm 1\}$, the function $f(\mathbf{x}) = F(g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_k(\mathbf{x}))$ has $\sum_{i\in[k]} \sigma_i$-smooth boundary with respect to $\mathcal{D}$.*

Proof Let $\mathbf{x} \in \mathbb{R}^d$ such that $f^{+\rho}(\mathbf{x}) \neq f^{-\rho}(\mathbf{x})$. Then, there must be some $\mathbf{z} \in \mathbb{R}^d$ with $\|\mathbf{z}\|_2 \leq \rho$ and some $i \in [k]$ such that $g_i(\mathbf{x}) \neq g_i(\mathbf{x}+\mathbf{z})$. Moreover, we have $g_i^{+\rho}(\mathbf{x}) - g_i^{-\rho}(\mathbf{x}) \in \{0, 2\}$ for all $\mathbf{x}$. Therefore, we have:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[\frac{f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})}{2}\Big] \leq \sum_{i\in[k]} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[\frac{g_i^{+\rho}(\mathbf{x}) - g_i^{-\rho}(\mathbf{x})}{2}\Big] \leq \sum_{i\in[k]} \sigma_i \,,$$

as desired. ■

As we shall see, several fundamental concept classes—including intersections of halfspaces, low-dimensional polynomial threshold functions, and low-dimensional convex sets—have smooth boundary. Together with Theorem 8, the above proposition yields sandwiching degree bounds not only for each of these classes individually, but also for arbitrary Boolean combinations thereof.

C.1. Functions of Halfspaces

We begin by focusing on the fundamental class of halfspaces. The decision boundary of a halfspace is a hyperplane. Therefore, the probability of falling near the boundary of a halfspace is equal to the probability of falling within a narrow band. As such, the smooth boundary condition for halfspaces is satisfied precisely for distributions that are anticoncentrated in the following sense.

Definition 21 (Anticoncentration) *For $\alpha > 0$, we say that a distribution $\mathcal{D}$ over $\mathbb{R}^d$ is α-anticoncentrated in every direction if for any $\mathbf{w} \in \mathbb{S}^{d-1}$, any $t \in \mathbb{R}$, and any $r \geq 0$,*

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[|\mathbf{w}\cdot\mathbf{x} - t| \leq r] \leq \alpha\, r\,.$$

In what follows, we may omit the parameter α and absorb it into universal constants under big-O notation.

The following proposition is immediate from the definition of anticoncentration.

Proposition 22 *Let $\mathcal{D}$ be some distribution over $\mathbb{R}^d$ that is anticoncentrated in every direction. Then, the class of halfspaces has $O(1)$-smooth boundary with respect to $\mathcal{D}$.*

Proof Let $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} - t)$ for some $\mathbf{w} \in \mathbb{S}^{d-1}$ and $t \in \mathbb{R}$. Then, we have:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[\frac{f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})}{2}\Big] = \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[|\mathbf{w}\cdot\mathbf{x} - t| \leq \rho] \leq O(1)\cdot\rho\,,$$

as desired. ■

General Functions of Halfspaces The following result provides an exponential improvement over the best known bound for the $(\epsilon = 0.1, s = 1)$-sandwiching degree of functions of k halfspaces with respect to any anticoncentrated and strictly subexponential distribution Gopalan et al. (2010); Gollakota et al. (2023).

Theorem 23 *Let $\mathcal{F}$ be the class of arbitrary functions of k halfspaces over $\mathbb{R}^d$ and let $\mathcal{D}$ be some distribution over $\mathbb{R}^d$ that is anticoncentrated in every direction and γ-strictly subexponential. Then, the (ϵ, s)-sandwiching degree of $\mathcal{F}$ with respect to $\mathcal{D}$ is*

$$\ell(\epsilon, s) \leq \widetilde{O}\Big(\frac{k^{5/2}s}{(\epsilon/2)^{s+1}}\Big)^{1+1/\gamma}.$$

The above theorem is an immediate corollary of Theorem 8 and the following lemma, which is a combination of Propositions 20 and 22.

Lemma 24 (Combination of Propositions 20 and 22) *Let $\mathcal{F}$ be the class of arbitrary functions of k halfspaces over $\mathbb{R}^d$ and let $\mathcal{D}$ be some distribution over $\mathbb{R}^d$ that is anticoncentrated in every direction. Then, $\mathcal{F}$ has $O(k)$-smooth boundary with respect to $\mathcal{D}$.*

Note that Theorem 23 also provides a polynomial improvement over the previous state-of-the-art upper bound on the sandwiching degree of halfspace intersections with respect to the Gaussian distribution for constant ϵ Gopalan et al. (2010); Klivans et al. (2024a). In particular, the best known bound before this work was $O(k^6)$, and Theorem 23 gives a bound of $\widetilde{O}(k^5)$ for the Gaussian distribution ($\gamma = 1$).

Improvements for Intersections of Halfspaces under the Gaussian In the following theorem, we show that we can actually improve the sandwiching bound of k-halfspace intersections with respect to the Gaussian distribution to $\widetilde{O}(k^3)$.

Theorem 25 *Let $\mathcal{F}$ be the class of intersections of k halfspaces over $\mathbb{R}^d$. Then, the (ϵ, s)-sandwiching degree of $\mathcal{F}$ with respect to the standard Gaussian distribution is*

$$\ell(\epsilon, s) \leq \widetilde{O}\left(\frac{k^3 s}{(\epsilon/2)^{2s+2}}\right).$$

The above theorem is a corollary of Theorem 8 combined with the following result which gives a sharp bound on the boundary smoothness parameter of halfspace intersections.

Lemma 26 (Boundary Smoothness of Halfspace Intersections) *Let $\mathcal{F}$ be the class of intersection of k halfspaces over $\mathbb{R}^d$. Then, $\mathcal{F}$ has $O(\sqrt{\log k})$-smooth boundary with respect to the standard Gaussian distribution.*

At the core of Lemma 26 is the following bound follows from a result by Nazarov on the Gaussian surface area of halfspace intersections (see Nazarov (2003); Klivans et al. (2008)) as well as the fact that halfspace intersections have $(d-1)$-rectifiable boundaries. See Appendix A and Kane (2014); De et al. (2023) for relevant discussions.

Fact 27 (Nazarov (see also Nazarov (2003); Klivans et al. (2008))) *Let $f : \mathbb{R}^d \to \{\pm 1\}$ be an intersection of k halfspaces. Let ∂f be the boundary of f, i.e., the set of points $\mathbf{x}$ where for any $\rho > 0$ there exists $\mathbf{z} \in \mathbb{R}^d$ with $\|\mathbf{z}\|_2 \leq \rho$ and $f(\mathbf{x}) \neq f(\mathbf{x} + \mathbf{z})$. Then, we have the following:*

$$\int_{\partial f} \varphi(\mathbf{x})\, d\mathcal{H}^{d-1}(\mathbf{x}) \leq \sqrt{2 \ln k} + 2\,,$$

where φ is the standard Gaussian density $\varphi(\mathbf{x}) = (2\pi)^{-d/2} \exp(-\|\mathbf{x}\|_2^2/2)$ and $\mathcal{H}^{d-1}$ is the $(d-1)$-dimensional Hausdorff measure.

Nazarov's result gives an asymptotic version of Lemma 26, i.e., it shows that for an intersection of k halfspaces f we have:

$$\lim_{\rho \to 0^+} \mathbb{E}\left[\frac{f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})}{2\rho}\right] \leq O(\sqrt{\log k})$$

To obtain a non-asymptotic version of this result, we use the coarea formula, which is an integration tool from geometric measure theory Federer (1969). To apply the coarea formula properly, we define the following alternative notions of dilation and erosion which are specialized to halfspace intersections. Crucially, these operations generate functions that themselves are intersections of the same number of halfspaces.

Definition 28 (Bias-shift Dilation and Erosion) *Let $f : \mathbb{R}^d \to \{\pm 1\}$ be an intersection of k halfspaces, i.e., $f(\mathbf{x}) = 2 \cdot \prod_{i=1}^k \mathbb{1}\{\mathbf{w}_i \cdot \mathbf{x} \le \tau_i\} - 1$ for some $\mathbf{w}_i \in \mathbb{S}^{d-1}, \tau_i \in \mathbb{R}$. For $\rho \ge 0$, we define the ρ-bias-shift dilation $f^{\boxplus \rho}$, as well as the ρ-bias-shift erosion $f^{\boxminus \rho}$ of f as follows.*

$$f^{\boxplus \rho}(\mathbf{x}) = 2 \cdot \prod_{i=1}^k \mathbb{1}\{\mathbf{w}_i \cdot \mathbf{x} \le \tau_i + \rho\} - 1$$
$$f^{\boxminus \rho}(\mathbf{x}) = 2 \cdot \prod_{i=1}^k \mathbb{1}\{\mathbf{w}_i \cdot \mathbf{x} \le \tau_i - \rho\} - 1$$

Moreover, in the following proposition we show that the expected difference between $f^{+\rho}$ and $f^{-\rho}$ is dominated by the difference $f^{\boxplus \rho} - f^{\boxminus \rho}$. This is important for our analysis, because we will essentially integrate the surface integrals of $f^{\boxplus t}, f^{\boxminus t}$ over $t \in [0, \rho]$, and will use the fact that these functions are halfspace intersections to invoke Fact 27.

Proposition 29 *Let $f : \mathbb{R}^d \to \{\pm 1\}$ be an intersection of k halfspaces. We have:*

$$f^{\boxminus \rho}(\mathbf{x}) = f^{-\rho}(\mathbf{x}) \le f(\mathbf{x}) \le f^{+\rho}(\mathbf{x}) \le f^{\boxplus \rho}(\mathbf{x}) \text{ , for all } \mathbf{x} \in \mathbb{R}^d \text{ and } \rho \ge 0$$

Proof Observe that for all $i \in [k]$ we have that $|\mathbf{w}_i \cdot \mathbf{z}| \le \|\mathbf{z}\|_2$. Therefore, we have the following:

$$\begin{aligned} f^{+\rho}(\mathbf{x}) = \sup_{\mathbf{z}:\|\mathbf{z}\|_2 \le \rho} f(\mathbf{x} + \mathbf{z}) &= \sup_{\mathbf{z}:\|\mathbf{z}\|_2 \le \rho} 2 \cdot \prod_{i=1}^k \mathbb{1}\{\mathbf{w}_i \cdot \mathbf{x} \le \tau_i - \mathbf{w}_i \cdot \mathbf{z}\} - 1 \\ &\le \sup_{\mathbf{z}:\|\mathbf{z}\|_2 \le \rho} 2 \cdot \prod_{i=1}^k \mathbb{1}\{\mathbf{w}_i \cdot \mathbf{x} \le \tau_i + \rho\} - 1 \\ &= f^{\boxplus \rho}(\mathbf{x}) \end{aligned}$$

We have shown that $f^{+\rho} \le f^{\boxplus \rho}$. A symmetric argument gives $f^{-\rho} \ge f^{\boxminus \rho}$. Moreover, in the case of erosion, we obtain $f^{-\rho} = f^{\boxminus \rho}$. To see this, let $\mathbf{z}^* = \mathbf{z}^*(\mathbf{x}) = \rho \, \mathbf{w}_{i^*(\mathbf{x})}$ where $i^*(\mathbf{x}) = \arg\max_{i \in [k]}\{\mathbf{w}_i \cdot \mathbf{x} - \tau_i\}$. Suppose that $f(\mathbf{x}) = 1$, since the other case is trivial. We have:

$$\begin{aligned} f^{-\rho}(\mathbf{x}) &= 2 \cdot \inf_{\mathbf{z}:\|\mathbf{z}\|_2 \le \rho} \mathbb{1}\{\mathbf{w}_{i^*} \cdot (\mathbf{x} + \mathbf{z}) \le \tau_{i^*}\} - 1 \\ &= 2 \cdot \mathbb{1}\{\mathbf{w}_{i^*} \cdot (\mathbf{x} + \mathbf{z}^*) \le \tau_{i^*}\} - 1 \\ &= 2 \cdot \mathbb{1}\{\mathbf{w}_{i^*} \cdot \mathbf{x} \le \tau_{i^*} - \rho\} - 1 \\ &= f^{\boxminus \rho}(\mathbf{x}) \, , \end{aligned}$$

since the coordinate i^* corresponds to the smallest gap between $\mathbf{w}_i \cdot \mathbf{x}$ and τ_i among all $i \in [k]$. ∎

We are now ready to prove Lemma 26.

Proof of Lemma 26. Let f be an intersection of k halfspaces, i.e., $f(\mathbf{x}) = 2 \cdot \prod_{i=1}^k \mathbb{1}\{\mathbf{w}_i \cdot \mathbf{x} \le \tau_i\} - 1$ for some $\mathbf{w}_i \in \mathbb{S}^{d-1}$ and $\tau_i \in \mathbb{R}$. Due to Proposition 29, we have the following for any $\rho \ge 0$:

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{N}_d}\Big[\frac{f^{+\rho}(\mathbf{x}) - f^{-\rho}(\mathbf{x})}{2}\Big] \le \mathbb{E}_{\mathbf{x} \sim \mathcal{N}_d}\Big[\frac{f^{\boxplus \rho}(\mathbf{x}) - f^{\boxminus \rho}(\mathbf{x})}{2}\Big]$$

It is now convenient to define the following slack function:

$$\Psi(\mathbf{x}) = \max_{i \in [k]}\{\mathbf{w}_i \cdot \mathbf{x} - \tau_i\}$$

Using this notation, we have:

$$\underset{\mathbf{x}\sim\mathcal{N}_d}{\mathbb{E}}\Big[\frac{f^{\boxplus\rho}(\mathbf{x})-f^{\boxminus\rho}(\mathbf{x})}{2}\Big]=\underset{\mathbf{x}\sim\mathcal{N}_d}{\mathbb{P}}\Big[\Psi(\mathbf{x})\in[-\rho,\rho]\Big]=\int_{\mathbf{x}:|\Psi(\mathbf{x})|\le\rho}\varphi(\mathbf{x})\,d\mathbf{x}$$

Observe that the function Ψ is 1-Lipschitz as a maximum of affine functions. Hence, apart from a measure-zero set ND_Ψ, we have $\|\nabla\Psi(\mathbf{x})\|_2=1$, because it equals $\|\mathbf{w}_i\|_2$ for an active constraint i. We may apply the coarea formula (Theorem 3.2.22 in Federer (1969)) to obtain:

$$\begin{aligned}\int_{\mathbf{x}:|\Psi(\mathbf{x})|\le\rho}\varphi(\mathbf{x})\,d\mathbf{x}&=\int_{t=-\rho}^{\rho}\int_{\mathbf{x}\in\Psi^{-1}(t)\setminus\mathsf{ND}_\Psi}\frac{\varphi(\mathbf{x})}{\|\nabla\Psi(\mathbf{x})\|_2}\,d\mathcal{H}^{d-1}\,dt\\&=\int_{t=-\rho}^{\rho}\int_{\mathbf{x}\in\Psi^{-1}(t)}\varphi(\mathbf{x})\,d\mathcal{H}^{d-1}\,dt\end{aligned}$$

Observe now that, for $t\in[-\rho,\rho]$ with $r=|t|$, the set $\Psi^{-1}(t)=\{\mathbf{x}:\Psi(\mathbf{x})=t\}$ is either equal to $\partial f^{\boxplus r}$ or $\partial f^{\boxminus r}$ depending on whether $t\ge 0$ or $t<0$ (see Definition 15). The function $f^{\boxplus r}$ is always an intersection of halfspaces. The function $f^{\boxminus r}$ is either an intersection of halfspaces, or $\partial f^{\boxminus r}=\emptyset$. Therefore, by applying Fact 27, we obtain:

$$\int_{\mathbf{x}:|\Psi(\mathbf{x})|\le\rho}\varphi(\mathbf{x})\,d\mathbf{x}=\int_{t=-\rho}^{\rho}(\sqrt{2\ln k}+2)\,dt=O(\sqrt{\log k})\,\rho\,,$$

which concludes the proof of Lemma 26. ■

C.2. Convex Sets and Polynomial Threshold Functions

Low-Dimensional Convex Sets Another important geometric concept class is that of low-dimensional convex sets. While the boundary smoothness of convex sets has been studied in prior work (Nazarov, 2003; Chandrasekaran et al., 2024b), there have been no end-to-end results on the sandwiching degree of convex sets. Previous work (De et al., 2023) showed that convex sets in k dimensions under the Gaussian distribution admit upper sandwiching functions representable as intersections of $2^{O(k)}$ halfspaces. Combined with the sandwiching bounds for intersections of halfspaces from Gopalan et al. (2010), this implies the existence of upper sandwiching polynomials for convex sets, albeit of exponential degree. While this approach could in principle be extended to obtain lower sandwiching polynomials as well, our result achieves an exponential improvement by establishing sandwiching degree $\mathrm{poly}(k)$ for convex sets under the Gaussian distribution.

Theorem 30 *Let $\mathcal{F}$ be the class of convex functions in $\mathbb{R}^d$ with intrinsic dimension k. Then, the (ϵ,s)-sandwiching degree of $\mathcal{F}$ with respect to $\mathcal{N}(0,\mathbf{I}_{d\times d})$ is*

$$\ell(\epsilon,s)\le\widetilde{O}\Big(\frac{k^{5/2}s}{(\epsilon/2)^{s+1}}\Big)^{1+1/\gamma}.$$

The above theorem follows from combining Theorem 8 with the following result from Chandrasekaran et al. (2024b).

Lemma 31 (Chandrasekaran et al. (2024b)) *The class of convex functions in $\mathbb{R}^d$ with intrinsic dimension k has σ-smooth boundary for $\sigma=O(k\log k)$.*

Low-Dimensional PTFs The last class we consider is that of polynomial threshold functions. Recent work by Slot et al. (2024) building on the approach of Kane (2011) gave a bound on the sandwiching degree of degree-q PTFs that does not depend on the intrinsic dimension k, but the dependence on q is doubly exponential or worse.[3] In the following bound, we achieve an exponential improvement over prior work, as long as $q = \Omega(\log k)$, or a doubly-exponential improvement if $q = \mathrm{poly}(k)$.

Theorem 32 *Let $\mathcal{F}$ be the class of degree-q PTFs in $\mathbb{R}^d$ with intrinsic dimension k. Then, the (ϵ, s)-sandwiching degree of $\mathcal{F}$ with respect to $\mathcal{N}(0, \mathbf{I}_{d\times d})$ is*

$$\ell(\epsilon, s) \leq \widetilde{O}\left(\frac{q^3 k^{5/2} s}{(\epsilon/2)^{s+1}}\right)^{1+1/\gamma}.$$

Once more, our result follows from Theorem 8 combined with a corresponding bound on the boundary smoothness parameter of PTFs from Chandrasekaran et al. (2024b).

Lemma 33 (Chandrasekaran et al. (2024b)) *The class of degree-q PTFs in $\mathbb{R}^d$ with intrinsic dimension k has σ-smooth boundary for $\sigma = O(q^3 k)$.*

Appendix D. Applications

Sandwiching polynomials are a strong and versatile tool. While standard approximation theory provides low-degree polynomials that approximate functions in expectation, sandwiching polynomials additionally constrain the approximated function to lie within a narrow envelope. Crucially, this guarantee holds pointwise over the entire domain, rather than merely in expectation with respect to a given distribution. As a result, sandwiching polynomials remain representative of the target function even under distributions other than the one defining the approximation-in-expectation guarantee. This property has recently led to several concrete applications of sandwiching in learning theory, including learning under distribution shift (Klivans et al., 2024a; Chandrasekaran et al., 2024b), learning with heavy contamination (Klivans et al., 2025a), and testable learning (Rubinfeld and Vasilyan, 2023; Gollakota et al., 2023).

In what follows, we describe a collection of implications of our theorem for learning theory, as well as for pseudorandomness.

D.1. Learning Theory

A recent line of work in learning theory has focused on frameworks for algorithms that generate certificates of their own performance (Goldwasser et al., 2020; Rubinfeld and Vasilyan, 2023; Klivans et al., 2024a). This line of work emerged in response to the strong assumptions that are ubiquitous in learning theory and are often either unverifiable from samples or computationally intractable to check. Across these frameworks, it has been shown that the existence of low-degree sandwiching polynomials suffices for efficient learnability (Gollakota et al., 2023; Klivans et al., 2024a; Goel et al., 2024; Chandrasekaran et al., 2024b).

Moreover, the techniques developed in this line of work have been used to obtain near-optimal upper bounds (Klivans et al., 2025b,a) on the computational complexity of the classical problem of learning under contamination (Valiant, 1985; Kearns and Li, 1993; Bshouty et al., 2002). In fact, leveraging sandwiching polynomials allows these results to extend even to the setting of heavy contamination (Klivans et al., 2025a).

All of our learning-theoretic applications follow from combining Theorem 8 with existing results showing that sufficiently low sandwiching degree implies efficient learnability under the corresponding learning primitive. These results additionally require uniform convergence for the relevant class of polynomials under the target marginal distribution $\mathcal{D}$. Since the sandwiching polynomials provided by Theorem 8 have bounded

3. The dependence on q is not made explicit in prior work, as q is considered to be a constant.

coefficients and bounded degree, and since we restrict attention to (strictly) subexponential distributions, the required uniform convergence guarantees hold in our setting, as discussed in Appendix B.

In the following, we will use the notation $\bar{\mathcal{D}}$ to denote a labeled distribution over $\mathbb{R}^d \times \{\pm 1\}$, where $\mathcal{D}$ is the marginal of $\bar{\mathcal{D}}$ on $\mathbb{R}^d$. We will also use a similar convention for labeled sets of examples $\bar{S}$ and their unlabeled counterparts S.

Testable Learning We first focus on the problem of testable agnostic learning, in which the learner receives labeled samples from an unknown distribution over $\mathbb{R}^d \times \{\pm 1\}$ without any guaranteed assumptions on the underlying data-generating process. While there is strong evidence that some structural assumptions on the $\mathbb{R}^d$-marginal are necessary for efficient learnability even for simple concept classes (Daniely, 2016; Diakonikolas and Kane, 2022; Diakonikolas et al., 2022), testable learning algorithms avoid these hardness barriers by retaining the ability to abstain when their target distributional assumption is violated. In particular, such algorithms either accept and output a hypothesis whose error is near-optimal with respect to a given concept class, or reject by declaring that the underlying assumption does not hold.

Definition 34 (Testable Learning (Rubinfeld and Vasilyan, 2023)) *An algorithm $\mathcal{A}$ is a tester-learner for concept class $\mathcal{F}$ with respect to some target distribution $\mathcal{D}^*$ over $\mathbb{R}^d$ if on input $(\epsilon, \delta, \bar{S})$, where $\epsilon, \delta \in (0,1)$ and $\bar{S}$ is a set of i.i.d. examples from some arbitrary distribution $\bar{\mathcal{D}}$, the algorithm $\mathcal{A}$ either outputs* Reject *or outputs* (Accept, h)*, where $h : \mathbb{R}^d \to \{\pm 1\}$ such that with probability at least $1-\delta$ over $\bar{S}$, and the randomness of $\mathcal{A}$, the following conditions hold.*

1. *(Soundness) Upon acceptance, the error of h is bounded as follows:*

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq h(\mathbf{x})] \leq \min_{f\in\mathcal{F}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})] + \epsilon$$

2. *(Completeness) If $\mathcal{D} = \mathcal{D}^*$, where $\mathcal{D}$ is the marginal of $\bar{\mathcal{D}}$ on $\mathbb{R}^d$, then $\mathcal{A}$ accepts.*

The work of Gollakota et al. (2023) showed that low sandwiching degree implies efficient algorithms for testable learning. Therefore, Theorem 8 implies the following result.

Theorem 35 (Combination of Gollakota et al. (2023) with Theorem 8) *Let $\mathcal{F}$ be some concept class that satisfies Assumption 7 with parameters k, d, σ with respect to a γ-strictly subexponential distribution $\mathcal{D}^*$ over $\mathbb{R}^d$. There is a tester-learner for $\mathcal{F}$ with respect to $\mathcal{D}^*$ that has runtime and sample complexity $d^{\widetilde{O}(\ell)}O(\log 1/\delta)$ where we have:*

$$\ell \;\leq\; \widetilde{O}\left(\frac{\sigma k^{3/2}}{\epsilon^2}\right)^{1+1/\gamma}.$$

One limitation of Definition 34 for testable agnostic learning is that it permits the learner to reject whenever there is even a small violation of the target assumption $\mathcal{D}^*$. As a result, the corresponding tester-learners may be fragile and reject too often, which can limit their practical usefulness. Fortunately, it is possible to design algorithms that are guaranteed to accept with high probability even when the distributional assumption is violated, provided that the marginal distribution is sufficiently close to the target $\mathcal{D}^*$ in total variation distance. Such algorithms are known as tolerant tester-learners and are defined as follows.

Definition 36 (Tolerant Testable Learning (Goel et al., 2024)) *The definition is the same as for testable learning (Definition 34), with the only difference being that the algorithm receives an additional parameter $\tau \in (0,1)$ as input and the soundness and completeness criteria are modified as follows:*

1. *(Soundness) Upon acceptance, the error of h is bounded as follows:*

$$\underset{(\mathbf{x},y)\sim\bar{\mathcal{D}}}{\mathbb{P}}[y \neq h(\mathbf{x})] \leq \min_{f\in\mathcal{F}} \underset{(\mathbf{x},y)\sim\bar{\mathcal{D}}}{\mathbb{P}}[y \neq f(\mathbf{x})] + \tau + \epsilon$$

2. *(Completeness) If $\mathrm{d_{TV}}(\mathcal{D}, \mathcal{D}^*) \leq \tau$, where $\mathcal{D}$ is the marginal of $\bar{\mathcal{D}}$ on $\mathbb{R}^d$, then $\mathcal{A}$ accepts.*

To obtain upper bounds on the computational complexity of tolerant testable learning, an additional distributional assumption on the target marginal $\mathcal{D}^*$ is required. In particular, Klivans et al. (2025a) assume that $\mathcal{D}^*$ is hypercontractive with respect to arbitrary polynomials of any degree. This property is defined below and is satisfied by a broad class of distributions, including all log-concave distributions (Bobkov, 2001; Saumard and Wellner, 2014), as well as the uniform distribution on the Boolean hypercube.

Definition 37 (Hypercontractivity) *A distribution $\mathcal{D}^*$ over $\mathbb{R}^d$ is polynomially hypercontractive if there is a constant $C \geq 1$ such that for any polynomial p over $\mathbb{R}^d$ and any $q \geq 2$ we have*

1. $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[|p(\mathbf{x})|^q] \leq (Cq)^{\ell q}\big(\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[|p(\mathbf{x})|]\big)^q$, *where ℓ is the degree of p.*
2. *The absolute expectations of degree-1 monomials are finite under $\mathcal{D}^*$.*

We obtain the following result by combining a theorem from Klivans et al. (2025a)—which shows that $\mathcal{L}_1$ sandwiching suffices for tolerant testable learning—with Theorem 8.

Theorem 38 (Combination of Klivans et al. (2025a) with Theorem 8) *Let $\mathcal{F}$ be some concept class that satisfies Assumption 7 with parameters k, d, σ with respect to a γ-strictly subexponential and polynomially hypercontractive distribution $\mathcal{D}^*$ over $\mathbb{R}^d$. For any $\tau \geq 0$, there is a τ-tolerant (ϵ, δ)-tester-learner for $\mathcal{F}$ with respect to $\mathcal{D}^*$ that has runtime and sample complexity $d^{\widetilde{O}(\ell)}O(\log 1/\delta)$ where we have:*

$$\ell \;\leq\; \widetilde{O}\left(\frac{\sigma k^{3/2}}{\epsilon^4}\right)^{1+1/\gamma}.$$

Learning with Distribution Shift Another learning problem in which sandwiching polynomials play a central role is learning under distribution shift. In this setting, the learner has access to labeled samples from a training distribution, as well as unlabeled samples from a potentially different test distribution. The goal is to output a hypothesis with low error on the test distribution. Since labels from the test distribution are not observed, this task is only feasible when the training and test labeling functions are suitably related. The following quantity—originally introduced in the domain adaptation literature (Ben-David et al., 2006; Mansour et al., 2009; Ben-David et al., 2010)—quantifies this relationship in terms of a hypothesis that simultaneously achieves low error on both the training and test distributions.

Definition 39 (Error Benchmark in the Distribution Shift Setting) *Let $\bar{\mathcal{D}}, \bar{\mathcal{D}}'$ be two distributions over $\mathbb{R}^d \times \{\pm 1\}$ and $\mathcal{F}$ some concept class over $\mathbb{R}^d$. We consider the following error benchmark:*

$$\lambda(\bar{\mathcal{D}}, \bar{\mathcal{D}}'; \mathcal{F}) = \min_{f\in\mathcal{F}}\Big(\underset{(\mathbf{x},y)\sim\bar{\mathcal{D}}}{\mathbb{P}}[y \neq f(\mathbf{x})] + \underset{(\mathbf{x},y)\sim\bar{\mathcal{D}}'}{\mathbb{P}}[y \neq f(\mathbf{x})]\Big)$$

Note that, in the absence of test labels, the dependence of the test error of the output hypothesis on λ is unavoidable (see Klivans et al. (2024a)). We focus now on testable learning with distribution shift (TDS learning), where the learner is asked to either provide a hypothesis with near-optimal error on the test distribution, or detect the presence of harmful distribution shift.

Definition 40 (TDS Learning (Klivans et al., 2024a)) *An algorithm $\mathcal{A}$ is a TDS-learner for concept class $\mathcal{F}$ with respect to some target distribution $\mathcal{D}^*$ over $\mathbb{R}^d$ if on input $(\epsilon, \delta, \bar{S}, S')$, where $\epsilon, \delta \in (0,1)$, $\bar{S}$ is a set of labeled i.i.d. examples from some training distribution $\bar{\mathcal{D}}$ where $\mathcal{D} = \mathcal{D}^*$ and S' is a set of unlabeled i.i.d. examples from the marginal $\mathcal{D}'$ of some arbitrary test distribution $\bar{\mathcal{D}}'$, the algorithm $\mathcal{A}$ either outputs* Reject *or outputs* (Accept, h)*, where $h : \mathbb{R}^d \to \{\pm 1\}$ such that with probability at least $1-\delta$ over $\bar{S}, S$, and the randomness of $\mathcal{A}$, the following conditions hold.*

1. *(Soundness) Upon acceptance, the test error of h is bounded as follows:*

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}'}[y \neq h(\mathbf{x})] \leq \lambda(\bar{\mathcal{D}}, \bar{\mathcal{D}}'; \mathcal{F}) + \min_{f\in\mathcal{F}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})] + \epsilon$$

2. *(Completeness) If $\mathcal{D}' = \mathcal{D}^*$, where $\mathcal{D}'$ is the marginal of the test distribution $\bar{\mathcal{D}}'$ on $\mathbb{R}^d$, then $\mathcal{A}$ accepts.*

Once more, we may combine Theorem 8 with a result from Chandrasekaran et al. (2024b) to obtain the following theorem.

Theorem 41 (Combination of Chandrasekaran et al. (2024b) with Theorem 8) *Let $\mathcal{F}$ be some concept class that satisfies Assumption 7 with parameters k, d, σ with respect to a γ-strictly subexponential distribution $\mathcal{D}^*$ over $\mathbb{R}^d$. There is a TDS-learner for $\mathcal{F}$ with respect to $\mathcal{D}^*$ that has runtime and sample complexity $(d \log 1/\delta)^{\widetilde{O}(\ell)}$ where we have:*[4]

$$\ell \;\leq\; \widetilde{O}\left(\frac{\sigma k^{3/2}}{\epsilon^2}\right)^{1+1/\gamma}.$$

In TDS learning, the algorithm accepts or rejects on a population basis. One natural question is whether it is possible for a learner to abstain on a per-point basis, meaning that it produces a hypothesis and a selector with the following guarantees.

Definition 42 (PQ Learning (Goldwasser et al., 2020)) *An algorithm $\mathcal{A}$ is a PQ-learner for concept class $\mathcal{F}$ with respect to some target distribution $\mathcal{D}^*$ over $\mathbb{R}^d$ if on input $(\epsilon, \delta, \bar{S}, S')$, where $\epsilon, \eta, \delta \in (0,1)$, $\bar{S}$ is a set of labeled i.i.d. examples from some training distribution $\bar{\mathcal{D}}$ where $\mathcal{D} = \mathcal{D}^*$ and S' is a set of unlabeled i.i.d. examples from the marginal $\mathcal{D}'$ of some arbitrary test distribution $\bar{\mathcal{D}}'$, the algorithm $\mathcal{A}$ either outputs a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ and a selector $g : \mathbb{R}^d \to \{0,1\}$ such that with probability at least $1-\delta$ over $\bar{S}, S$, and the randomness of $\mathcal{A}$, the following conditions hold.*

1. *(Error Rate) The error of h on the selected part of the test distribution is bounded as follows:*

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}'}[y \neq h(\mathbf{x}) \text{ and } g(\mathbf{x}) = 1] \leq O\left(\frac{\lambda(\bar{\mathcal{D}}, \bar{\mathcal{D}}'; \mathcal{F})}{\eta}\right) + \epsilon$$

2. *(Rejection Rate) The rejection rate of g on the training distribution is bounded as follows:*

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[g(\mathbf{x}) = 0] \leq \eta$$

The definition of PQ learning (Goldwasser et al., 2020) predates that of TDS learning (Klivans et al., 2024a), although PQ learning constitutes a strictly more demanding learning primitive. Indeed, PQ learning implies a tolerant version of TDS learning (Klivans et al., 2024a; Goel et al., 2024). However, the first

4. The factor $\log(1/\delta)^{\widetilde{O}(\ell)}$ is inherited from Fact 17. Repetition-based amplification in Klivans et al. (2024a) incurs a constant error loss factor, so we instead obtain failure probability δ directly via Fact 17.

end-to-end efficient algorithms for PQ learning (Goel et al., 2024) were obtained by building on techniques developed for TDS learning (Klivans et al., 2024a), through the use of sandwiching polynomials. Prior to these results, Goldwasser et al. (2020); Kalai and Kanade (2021) established reductions between PQ learning and more classical learning primitives, such as distribution-free reliable agnostic learning. Such primitives, however, are likely computationally intractable even for very simple concept classes, including conjunctions (see Kalai and Kanade (2021)).

We provide the following positive result for PQ learning, which uses—once more—polynomial hypercontractivity, and the result of Goel et al. (2024). Note that the result of Goel et al. (2024) requires the existance of $\mathcal{L}_2$-sandwiching polynomials, instead of standard $\mathcal{L}_1$-sandwiching. However, our techniques yield results for sandwiching of any order $s \geq 1$.

Theorem 43 (Combination of Goel et al. (2024) with Theorem 8) *Let $\mathcal{F}$ be some concept class that satisfies Assumption 7 with parameters k, d, σ with respect to a γ-strictly subexponential and polynomially hypercontractive distribution $\mathcal{D}^*$ over $\mathbb{R}^d$. There is a PQ-learner for $\mathcal{F}$ with respect to $\mathcal{D}^*$ that has runtime and sample complexity $(d \log 1/\delta)^{\widetilde{O}(\ell)}$ where we have:*

$$\ell \;\leq\; \widetilde{O}\left(\frac{\sigma k^{3/2}}{(\epsilon\eta)^{3/2}}\right)^{1+1/\gamma}.$$

Note that the dependence on $1/\epsilon$ here improves upon that in Theorem 41. This improvement arises because, in the PQ learning setting, we allow the learner to return a constant-factor approximate solution, namely an error of $O(\lambda/\eta)$ (which reduces to $O(\lambda)$ when $\eta \approx 1$), whereas our definition of TDS learning requires a strictly tighter error guarantee. In particular, our analysis yields an upper bound on the test error that is proportional to $\|p_{\mathrm{up}} - p_{\mathrm{down}}\|_{\mathcal{D},2}^2$, rather than $\|p_{\mathrm{up}} - p_{\mathrm{down}}\|_{\mathcal{D},2}$. This quadratic dependence is one of the reasons a constant multiplicative error factor appears, but it allows us to work with $(\sqrt{\epsilon}, 2)$-sandwiching polynomials.

It remains unclear whether tighter error guarantees are achievable for PQ learning. In particular, it is not known whether $\mathcal{L}_1$-sandwiching alone suffices, while $\mathcal{L}_2$ approximation is known to incur a constant-factor loss in agnostic learning (Kalai et al., 2008).

Learning with Heavy Contamination Our final learning-theoretic application is learning with heavy contamination. This primitive was recently defined by Klivans et al. (2025a). In the heavy contamination, the input consists mostly of examples that are adversarial, with a constant fraction of clean points.

Definition 44 (Heavily Contaminated (HC) Datasets Klivans et al. (2025a)) *Let $\bar{\mathcal{D}}$ be some distribution over $\mathbb{R}^d \times \{\pm 1\}$ (we think of $\bar{\mathcal{D}}$ as the clean or uncorrupted distribution). We say that a set of samples $\bar{S}_{\mathrm{inp}}$ is generated by $\bar{\mathcal{D}}$ with Q-heavy contamination if it is generated as follows for some $m \leq M$ with $M/m \leq Q$.*

1. *First, a set $\bar{S}_{\mathrm{cln}}$ of m i.i.d. labeled examples from $\bar{\mathcal{D}}$ is drawn.*
2. *Then, an adversary receives $\bar{S}_{\mathrm{cln}}$ and adds $M - m$ arbitrary labeled examples to form $\bar{S}_{\mathrm{inp}}$.*

The error benchmark in the HC setting is defined as follows, and in particular requires that there is a classifier in the considered concept class that classifies almost all of the input examples correctly.

Definition 45 (Error Benchmark in the Heavy Contamination Setting) *Let $\bar{S}_{\mathrm{inp}}$ be a set of labeled examples in $\mathbb{R}^d \times \{\pm 1\}$ and $\mathcal{F}$ some concept class over $\mathbb{R}^d$. We define the following error benchmark:*

$$\mathsf{opt}_{\mathrm{total}}(\bar{S}_{\mathrm{inp}}; \mathcal{F}) = \min_{f \in \mathcal{F}} \frac{1}{|\bar{S}_{\mathrm{inp}}|} \sum_{(\mathbf{x},y) \in \bar{S}_{\mathrm{inp}}} \mathbb{1}\{y \neq f(\mathbf{x})\}$$

Note that empirical risk minimization over VC classes achieves an error guarantee of $Q \cdot \mathsf{opt}_{\text{total}} + \epsilon$ in the HC setting. The more interesting question is whether there are any efficient algorithms that can achieve the same guarantee, given that the input samples are not structured, apart from a small fraction. As we have said, structure is important for efficient learnability, but the question is whether HC datasets are structured enough. We now provide the definition of HC-learning.

Definition 46 (HC-Learning (Klivans et al., 2025a)) *An algorithm $\mathcal{A}$ is an HC-learner for concept class $\mathcal{F}$ with respect to some target distribution $\mathcal{D}^*$ over $\mathbb{R}^d$ if on input $(\epsilon, \delta, Q, \bar{S}_{\text{inp}})$, where $\epsilon, \delta \in (0,1)$, $Q \geq 1$ and $\bar{S}_{\text{inp}}$ is a Q-heavily contaminated set of labeled examples generated by distribution $\bar{\mathcal{D}}$ whose marginal on $\mathbb{R}^d$ is $\mathcal{D} = \mathcal{D}^*$ (as described in Definition 44), the algorithm $\mathcal{A}$ outputs some hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ such that with probability at least $1 - \delta$ over the clean examples in $\bar{S}_{\text{inp}}$, and the randomness of $\mathcal{A}$:*

$$\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq h(\mathbf{x})] \leq Q \cdot \mathsf{opt}_{\text{total}}(\bar{S}_{\text{inp}}; \mathcal{F}) + \epsilon$$

Klivans et al. (2025a) showed that the existence of low-degree $\mathcal{L}_1$-sandwiching polynomials suffices for efficient HC-learning when the clean distribution is hypercontractive. We obtain the following theorem by combining their result with Theorem 8.

Theorem 47 (Combination of Klivans et al. (2025a) with Theorem 8) *Let $\mathcal{F}$ be some concept class that satisfies Assumption 7 with parameters k, d, σ with respect to a γ-strictly subexponential and polynomially hypercontractive distribution $\mathcal{D}^*$ over $\mathbb{R}^d$. There is an HC-learner for $\mathcal{F}$ with respect to $\mathcal{D}^*$ that has runtime and sample complexity $(d \log 1/\delta)^{\widetilde{O}(\ell)}$ where we have:*

$$\ell \;\leq\; \widetilde{O}\left(\frac{\sigma k^{3/2}}{\epsilon^2}\right)^{1+1/\gamma}.$$

D.2. Pseudorandomness

Our work also has implications for pseudorandomness, where the goal is to replace a distribution such as the Gaussian with another distribution that can be generated using a small number r of random bits while preserving certain test statistics. This is central to derandomization, since when r is small, one can efficiently enumerate over all 2^r possible seeds and thereby simulate randomized algorithms deterministically with only a modest overhead (see Hatami and Hoza (2023) and references therein).

A central paradigm in pseudorandomness is to fool a class of test functions using distributions with limited randomness, such as bounded independence or, more generally, moment matching. While more sophisticated techniques usually lead to pseudorandom generators with shorter seed lengths, moment matching leads to simple, explicit pseudorandom generators and has been thoroughly studied in the relevant literature (Bazzi, 2009; Razborov, 2009; Gopalan et al., 2010; Diakonikolas et al., 2010a; Braverman, 2011; Kane, 2011; Diakonikolas et al., 2010b; Tal, 2017; Hatami and Hoza, 2023).

Definition 48 (Moment Matching Distributions) *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. For $\ell \in \mathbb{N}, \Delta > 0$, we define $\mathsf{MM}_{\mathcal{D}}(\ell, \Delta)$ to be the set of distributions $\mathcal{D}'$ such that for any $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_0 \leq \ell$ we have:*

$$\left| \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}^\alpha] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}'}[\mathbf{x}^\alpha] \right| \leq \Delta$$

Having the definition of moment matching at hand, we are ready to give a definition for fooling via moment matching.

Definition 49 (Fooling via Moment Matching) *Let $\mathcal{F}$ be a concept class over $\mathbb{R}^d$ and $\mathcal{D}$ a distribution over $\mathbb{R}^d$. We say that (ℓ, Δ)-moment matching ϵ-fools $\mathcal{F}$ with respect to $\mathcal{D}$ if the following is true for any distribution $\mathcal{D}' \in \mathsf{MM}_{\mathcal{D}}(\ell, \Delta)$:*

$$\left| \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}'}[f(\mathbf{x})] \right| \le \epsilon\,, \quad \textit{for all } f \in \mathcal{F}$$

Our result is the following fooling result for geometric concepts with low instrinsic dimension.

Theorem 50 (Fooling Geometric Concepts with Low Intrinsic Dimension) *Let $\mathcal{F}$ be some concept class that satisfies Assumption 7 with parameters k, d, σ with respect to a γ-strictly subexponential distribution $\mathcal{D}$ over $\mathbb{R}^d$. Then, for any $\epsilon \in (0,1)$ and $\Delta \ge 0$, we have that (ℓ, Δ)-moment matching ϵ'-fools $\mathcal{F}$ with respect to $\mathcal{D}$, where $\epsilon' = \epsilon + \Delta B$ and*

$$\ell \;\le\; \widetilde{O}\left(\frac{\sigma k^{3/2}}{\epsilon^2}\right)^{1+1/\gamma}, \quad B \le \exp\left(\widetilde{O}\left(\frac{\sigma k^{3/2}}{\epsilon^2}\right)^{1+1/\gamma}\right)$$

The result of the above theorem is a corollary of Theorem 8, combined with the following theorem which was first proven by Bazzi (2009) for the case $\Delta = 0$, and then for $\Delta > 0$ by Gollakota et al. (2023). The result states that the sandwiching degree characterizes the degree of fooling via moment matching, and the proof is based on linear programming duality.

Theorem 51 (Bazzi (2009); Gollakota et al. (2023)) *The following hold for any $\epsilon, \ell, \Delta, B > 0$ and $f : \mathbb{R}^d \to \{\pm 1\}$:*

- *If there exist $(\epsilon, 1)$-sandwiching polynomials for f with respect to $\mathcal{D}$ whose degree is at most ℓ and and the absolute values of their coefficients sum to B, then (ℓ, Δ)-moment matching ϵ'-fools f with respect to $\mathcal{D}$, where $\epsilon' = \epsilon + \Delta B$.*
- *Conversely, if (ℓ, Δ)-moment matching ϵ-fools f with respect to $\mathcal{D}$, then there are $(2\epsilon, 1)$-sandwiching polynomials for f with respect to $\mathcal{D}$ whose degree is at most ℓ and the sum of the absolute values of their coefficients is at most $B' = 2\epsilon/\Delta$.[5]*

5. In the case $\Delta = 0$, there is no bound for the coefficients of the sandwiching polynomials in general.