Smoothed Analysis for Learning Concepts with Low Intrinsic Dimension

Gautam Chandrasekaran GAUTAMC@CS.UTEXAS.EDU
University of Texas at Austin

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

Vasilis Kontonis VASILIS@CS.UTEXAS.EDU
University of Texas at Austin

Raghu Meka RAGHUM@CS.UCLA.EDU
University of California, Los Angeles

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

Editors: Shipra Agrawal and Aaron Roth

Abstract

In the well-studied agnostic model of learning, the goal of a learner– given examples from an arbitrary joint distribution on $\mathbb{R}^d \times \{\pm 1\}$– is to output a hypothesis that is competitive (to within ϵ) of the best fitting concept from some class. In order to escape strong hardness results for learning even simple concept classes in this model, we introduce a smoothed analysis framework where we require a learner to compete only with the best classifier that is robust to small random Gaussian perturbation.

This subtle change allows us to give a wide array of learning results for any concept that (1) depends on a low-dimensional subspace (aka multi-index model) and (2) has a bounded Gaussian surface area. This class includes functions of halfspaces and (low-dimensional) convex sets, cases that are only known to be learnable in non-smoothed settings with respect to highly structured distributions such as Gaussians.

Perhaps surprisingly, our analysis also yields new results for traditional non-smoothed frameworks such as learning with margin. In particular, we obtain the first algorithm for agnostically learning intersections of k-halfspaces in time $k^{\mathrm{poly}(\frac{\log k}{\epsilon\gamma})}$ where γ is the margin parameter. Before our work, the best-known runtime was exponential in k (Arriaga and Vempala, 1999a).

Keywords: PAC Learning; Agnostic Learning; Margin; Halfspace; Geometric Concepts; Gaussian Surface Area

1. Introduction

In the (agnostic) PAC learning model Valiant (1984a,b); Haussler (1992); Kearns et al. (1994), a learner is given access to random labeled examples and has to compute a classifier that performs approximately as well as the best classifier in a target concept class. More precisely, for an instance distribution D over $\mathbb{R}^d \times \{\pm 1\}$ and a concept class $\mathcal{F}$, the optimal error is defined as $\mathrm{opt} = \inf_{f \in \mathcal{F}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]$. Without assumptions about the feature distribution and/or the label generating process, learning is known to be computationally hard Kharitonov (1993); Guruswami and Raghavendra (2006); Dachman-Soled et al. (2008); Khot and Saket (2008); Feldman et al. (2009); Klivans and Sherstov (2009); Diakonikolas et al. (2011); Feldman et al. (2011);

Daniely and Vardi (2021). In particular, even learning halfspaces (linear classifiers) is intractable without assumptions Kalai et al. (2005); Guruswami and Raghavendra (2006); Feldman (2006); Daniely (2016).

In order to bypass these hardness results, a body of research has focused on beyond worst case learning. The most common approaches are: (1) making distributional assumptions about the underlying feature distribution, e.g., that it is Gaussian or uniform on the hypercube Linial et al. (1993); Long (2003); Kalai et al. (2008); Klivans et al. (2008); Gopalan et al. (2008); Diakonikolas et al. (2021); Kalai et al. (2009), or (2) assuming that the labels are not generated adversarially Awasthi et al. (2015, 2016, 2017); Diakonikolas et al. (2019a, 2020); Chen et al. (2020); Zhang et al. (2020); Diakonikolas et al. (2022).

Our Smoothed Learning Model In this work, we depart from those paradigms, and instead of explicitly imposing structure on the feature or the label distributions we relax the notion of optimality. Inspired by the seminal works Spielman and Teng (2004); Spielman (2005) on the smoothed-complexity of algorithms, we require the learner to compete against the minimum possible error over classifiers that have been translated by a small Gaussian perturbation. Formally, we have the following definition:

Definition 1 (Smoothed Agnostic Learning) *Fix $\epsilon, \sigma > 0$ and $\delta \in (0,1)$. Let $\mathcal{F}$ be a class of Boolean concepts and let $\mathbb{D}$ be a class of distributions over $\mathbb{R}^d$. Let D be a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ such that its $\mathbf{x}$-marginal $D_{\mathbf{x}} \in \mathbb{D}$. We say that the algorithm $\mathcal{A}$ learns $\mathcal{F}$ in the σ-smoothed setting if, after receiving i.i.d. samples from D, $\mathcal{A}$ outputs a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ such that, with probability at least $1-\delta$, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \mathrm{opt}_\sigma + \epsilon$, where*

$$\mathrm{opt}_\sigma = \inf_{f \in \mathcal{F}} \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}} \left[\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D} [f(\mathbf{x} + \sigma \mathbf{z}) \neq y] \right]. \tag{1}$$

We observe that by taking $\sigma = 0$ in Definition 1 we recover the standard definition of agnostic learning. On the other extreme as $\sigma \to \infty$, every concept is evaluated on a random input unrelated to the label y and the error essentially does not depend on the concept f. The smoothed agnostic learning of Definition 1 is therefore an interpolation between the case where the instance distribution D and the optimal classifier can be arbitrarily coupled (which corresponds to agnostic learning and $\sigma = 0$) and completely decoupled (when $\sigma = \infty$). This decoupling allows us to avoid worst-case concepts that can encode complexity-theoretic primitives.

Learning Concepts with Low Intrinsic Dimension We focus on the general class of concepts with low intrinsic dimension, i.e., that implicitly depend on few relevant directions (these are also known as linear or subspace juntas Vempala and Xiao (2011); De et al. (2019, 2021)). More precisely, a concept f is of low intrinsic dimension if there exists an — *unknown to the learner* — subspace V of dimension at most k such that f only depends on the projection of $\mathbf{x}$ onto V, i.e., $f(\mathbf{x}) = f(\mathrm{proj}_V \mathbf{x})$ for all $\mathbf{x}$. We will also use the term "low-dimensional" for such concepts. Perhaps *the* most well-studied low-dimensional concept class is that of halfspaces or linear threshold functions Rosenblatt (1962); Minsky and Papert (1988), where $k = 1$. Another popular low-dimensional class that has been extensively studied is intersections of k halfspaces Blum and Kannan (1993); Arriaga and Vempala (1999a); Klivans and Servedio (2004); Klivans et al. (2008); Vempala (2010). More broadly, in Definition 2 we define a general class of low dimensional concepts with "well-behaved" decision boundary that includes the previous mentioned classes (and

more) as special cases. Essentially all efficient algorithms in prior work for learning such concepts (in fact even for learning halfspaces) rely on strong assumptions, such as Gaussianity (Kalai et al., 2008; Klivans et al., 2008). We investigate whether it is possible to design efficient learning algorithms in the smoothed setting of Definition 2 for natural concept classes while weakening the distributional assumptions that have been used so far in the literature:

Can we relax the strong distributional assumptions (such as Gaussianity) required by previous works and still obtain comparable efficient algorithms in the smoothed setting?

We answer the above question positively and show that efficient smoothed learning is possible assuming only that the feature distribution is concentrated (e.g., bounded or sub-gaussian). In particular, our results in the smoothed setting establish learnability under discrete distributions that are commonly used in hardness constructions in the standard agnostic setting (see, e.g., Daniely and Vardi (2021)). At the same time, we show that our smoothed learning model improves and generalizes prior models such as learning with margin. In fact, for standard non-smoothed settings such as learning intersections of k-halfspaces with margin, we are able to obtain significant improvements over the prior works as corollaries of our smoothed learning results.

1.1. Our Results

In this section we present our main contributions and discuss the connections of the smoothed learning model of Definition 1 with other models.

Measure of Complexity: Gaussian Surface Area As mentioned above, we require that the concept class is low-dimensional, i.e., that it depends on few relevant directions. Moreover, we assume that it has bounded Gaussian Surface Area (GSA). The GSA of a boolean function f, denoted from now on as $\Gamma(f)$, is defined to be the surface area of its decision boundary weighted by the Gaussian density, see Definition 19 for a formal definition. In the context of learning theory, GSA was first used in Klivans et al. (2008) where it was shown that concepts with bounded GSA admit efficient learning algorithms under Gaussian marginals. Since then, GSA has played a significant role as a complexity measure in learning theory and related fields; see, e.g., Kane (2011); Neeman (2014); Kontonis et al. (2019); De et al. (2021).

Definition 2 (Low-Dimensional, Bounded Surface Area Concepts) *For $k \in \mathbb{N}$ and $\Gamma > 0$, a concept $f : \mathbb{R}^d \mapsto \{\pm 1\}$ belongs in the class $\mathcal{F}(k, \Gamma)$ if:*

1. *There exists a subspace U of dimension at most k such that $f(\mathbf{x}) = f(\mathrm{proj}_U(\mathbf{x}))$.*
2. *The Gaussian Surface Area of f, $\Gamma(f)$ is at most Γ.*
3. *For every $\mathbf{t} \in \mathbb{R}^d$ and $r > 0$, the function $f(r\mathbf{x} + \mathbf{t}) \in \mathcal{F}(k, \Gamma)$.*

Remark 3 *(1) While we are using GSA as a complexity measure, we stress that we do* **not** *assume that the $\mathbf{x}$-marginal distribution is Gaussian. (2) The invariance under scaling and translation (the third property of Definition 2) is a mild technical assumption that is satisfied by all classes that we have discussed so far (halfspaces and functions of halfspaces, ptfs, etc.), see also Lemma 20.*

We note that halfspaces belong in $\mathcal{F}(1, O(1))$, intersections of k halfspaces in $\mathcal{F}(k, O(\sqrt{\log k}))$, and k-dimensional polynomial threshold functions of degree ℓ in $\mathcal{F}(k, O(\ell))$. Moreover, Definition 2 also contains non-parametric classes: for example, $\mathcal{F}(k, O(k^{1/4}))$ includes all convex bodies in k dimensions, see Lemma 20. We remark that low-dimensional functions similar to those in Definition 2 are also referred to (usually when the functions are real-valued) as Multi-index Models (MiMs) — a common modeling assumption to avoid the curse of dimensionality in statistics Friedman et al. (1981); Huber (1985); Li (1991); Hall and Li (1993); Xia et al. (2002); Xia (2008).

1.1.1. MAIN RESULTS: SMOOTHED AGNOSTIC LEARNING UNDER CONCENTRATION

We show that we can efficiently learn assuming only concentration properties for the $\mathbf{x}$-marginal. More precisely, we assume that the distribution has sub-gaussian tails, i.e., for every unit direction $\mathbf{v}$ it holds $\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[|\mathbf{v}\cdot\mathbf{x}| \geq t] \leq \exp(-\Omega(t^2))$.

Theorem 4 (Sub-Gaussian – Informal, see also Theorem 17) *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with sub-gaussian $\mathbf{x}$-marginal. There exists an algorithm that learns the class $\mathcal{F}(k, \Gamma)$ in the σ-smoothed setting with $N = d^{\mathrm{poly}(\frac{k\Gamma}{\sigma\epsilon})} \log(\frac{1}{\delta})$ samples and $\mathrm{poly}(d, N)$ runtime.*

We remark that our result works even under weaker tail assumptions: in particular it suffices that the tails are strictly sub-exponential, see Definition 15 and Theorem 17.

We observe that the runtime of Theorem 4 for learning a single halfspace (where $k = 1$) in the smoothed setting qualitatively matches the best known runtime for agnostic learning under Gaussian marginals. For concepts with bounded Gaussian surface area, in Klivans et al. (2008), under the assumption that the $\mathbf{x}$-marginal is Gaussian, an algorithm with $d^{\mathrm{poly}(\Gamma/\epsilon)}$ runtime is given. When the intrinsic dimension $k = O(1)$, our results in the smoothed setting achieve the same runtime and only require sub-gaussian tails. By a simple reduction to learning parities on the hypercube, see Theorem 68, we obtain a Statistical Query (SQ) lower bound of $d^{\Omega(\min(k,\Gamma))}$ for learning over sub-gaussian marginals, showing that in some cases the exponential dependency on the surface area or the intrinsic dimension to learn $\mathcal{F}(k, \Gamma)$ is unavoidable.

Our second result shows that we can significantly improve the runtime when the marginals are bounded. Bounded marginals is a common assumption especially since it is often used together with geometric margin assumptions. At a high-level, in our smoothed learning setting having bounded $\|\mathbf{x}\|_2$ means that the ratio $\|\mathbf{x}\|_2/\sigma$ is more well behaved in the sense that the adversary, who picks $\mathbf{x}$, cannot overpower the smoothing noise σ (see Definition 1). Observe that if the adversary is allowed to select $\mathbf{x}$ with arbitrarily large norm, the effect of Gaussian noise in Definition 1 is negligible and we return to the standard agnostic setting.

Theorem 5 (Bounded – Informal, see also Theorem 18) *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with $\mathbf{x}$-marginal bounded in the unit ball. There exists an algorithm that learns the class $\mathcal{F}(k, \Gamma)$ in the σ-smoothed setting with $N = k^{\mathrm{poly}(\frac{\Gamma}{\epsilon\sigma})} \log(\frac{1}{\delta})$ samples and $\mathrm{poly}(d, N)$ runtime.*

Using our theorem and bounds on the Gaussian surface area we readily obtain corollaries for specific classes. For example, we efficiently learn intersections of k-halfspaces with $k^{\mathrm{poly}(\log k/(\sigma\epsilon))}$ samples and arbitrary k-dimensional convex bodies with $k^{\mathrm{poly}(k/(\sigma\epsilon))}$ samples.

1.1.2. APPLICATIONS

In this section we present several applications of our general smoothed learning results in standard agnostic learning settings that have been considered in the literature. In many cases we obtain significant improvements over the best-known results.

Agnostic Learning with Margin Our smoothed learning model is related to margin-based learning (originally defined in Ben-David and Simon (2000)) because, at a high-level, it incentivizes the adversary not to place points very close to the decision boundary to create non-trivial instances. In (agnostic) learning of a class $\mathcal{C}$ with γ-margin the feature distribution is typically assumed to be bounded and the goal is to compute a classifier with error

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \underbrace{\inf_{f\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}\left[\sup_{\|\mathbf{u}\|_2\leq\gamma} \mathbb{1}\{f(\mathbf{x}+\mathbf{u}) \neq y\}\right]}_{\text{margin-opt}_\gamma} + \epsilon\,. \tag{2}$$

We show that for any concept class with intrinsic dimension k, for $\sigma = \Omega(\gamma/\sqrt{k\log(1/\epsilon)})$, it holds $\mathrm{opt}_\sigma \leq \text{margin-opt}_\gamma + \epsilon$. Therefore, any learning algorithm for the smoothed learning setting can be directly used to learn in the γ-margin setting. For the special case of intersections of k-halfspaces we show that the gap between margin-opt$_\gamma$ and opt_σ is ϵ by choosing $\sigma = \Omega(\gamma/\sqrt{\log k \log(1/\epsilon)})$. Using this fact and Theorem 5 we obtain the following corollary.

Corollary 6 (Intersections of k-halfspaces with γ-margin) *Let D be a distribution on $\mathbb{R}^d\times\{\pm1\}$ whose $\mathbf{x}$-marginal is bounded in the unit ball and let $\mathcal{C}$ be the class of intersections of k-halfspaces. There exists an algorithm that draws $N = k^{\mathrm{poly}(\log k/\gamma\epsilon)}\log(\frac{1}{\delta})$ samples, runs in $\mathrm{poly}(d, N)$ time and computes a hypothesis h such that, with probability at least $1-\delta$, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \text{margin-opt}_\gamma + \epsilon$.*

We remark that, prior to our work, the best known runtime for learning intersections of k-halfspaces with γ-margin in the agnostic setting from Arriaga and Vempala (1999b) was exponential in the number of halfspaces that is $k^{\mathrm{poly}(\frac{k}{\gamma\epsilon})}$. Quasi-polynomial results similar to that of Corollary 6 were only known in the noiseless setting Klivans and Servedio (2004). Beyond intersections of halfspaces with γ-margin, we obtain new results for other classes such as polynomial threshold functions and general convex sets, see Section B.2 for more details.

Agnostic Learning under Smoothed Distributions We conclude with some applications of our framework to the (different) scenario where the *marginal distribution itself is smoothed*. For example, in Kane et al. (2013) sub-Gaussian marginals are smoothed by additive Gaussian noise; i.e., for some sub-Gaussian distribution D a sample from the smoothed distribution D_τ is generated as $\mathbf{x}+\tau\mathbf{z}$ for $\mathbf{x}\sim D$ and $\mathbf{z}\sim\mathcal{N}$. We remind the reader that our smoothed learning model of Definition 1 does not try to make the $\mathbf{x}$-marginal more benign by a Gaussian convolution as is done in smoothed distribution learning settings Kalai and Teng (2008); Kalai et al. (2009); Kane et al. (2013). In our model, the learner observes i.i.d. examples from the original marginal $D_\mathbf{x}$ and not from the convolution $D_\mathbf{x}+\sigma\mathcal{N}$. Perhaps surprisingly, we show that Theorem 4 can be used to significantly improve the results of Kane et al. (2013) and other results for learning with smoothed marginals:

Corollary 7 (Informal, see also Theorem 36) *Let D_τ be a smoothed sub-Gaussian distribution. There exists an algorithm that agnostically learns the class $\mathcal{F}(k,\Gamma)$ with $N = d^{\mathrm{poly}(\frac{k\Gamma}{\tau\epsilon})}\log(\frac{1}{\delta})$ samples and $\mathrm{poly}(d,N)$ runtime.*

We remark that Corollary 7 (i) generalizes the results of Kane et al. (2013) to any class of k-dimensional concepts with bounded surface area and (ii) yields an exponential improvement over Kane et al. (2013) where the runtime is doubly exponenential in k, i.e., $d^{\log\log(k/(\tau/\epsilon))^{\widetilde{O}(k)}\mathrm{poly}(1/(\tau\epsilon))}$.

Agnostic Learning under Anti-concentration Finally, another important direction considered in the literature is making structural assumptions such as anti-concentration over the feature distribution. In particular, in Gollakota et al. (2023) apart from sub-gaussian tails the distribution is assumed to satisfy anti-concentration over slabs, i.e., for any unit vector $\mathbf{v}$ and interval I it holds that $\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[\mathbf{v}\cdot\mathbf{x}\in I]\leq O(|I|)$, where $|I|$ is the length of the interval. In Gollakota et al. (2023) an algorithm for learning any function of a *constant number* of halfspaces is given with runtime $d^{\mathrm{poly}(1/\epsilon)}$. Using Theorem 4 we are able to obtain efficient algorithms for agnostic learning under concentration and anti-concentration for functions of any number of halfspaces.

Corollary 8 (Informal, see also Theorem 33) *Let D be a distribution on $\mathbb{R}^d\times\{\pm1\}$ whose $\mathbf{x}$-marginal is sub-Gaussian and anti-concentrated. There exists an algorithm that agnostically learns arbitrary functions of k halfspaces with $N = d^{\mathrm{poly}(\frac{k}{\epsilon})}\log(\frac{1}{\delta})$ samples and $\mathrm{poly}(d,N)$ runtime.*

1.2. Technical Overview

Our main plan is to use low-degree polynomials that can be efficiently optimized via L_1-regression, similar to the works of Kalai et al. (2005); Klivans et al. (2008). In general, in the agnostic setting, one has to construct a polynomial $p(\mathbf{x})$ that achieves almost optimal L_1 error with the label y. To do this, we have to prove that for every concept f in the class, there exists a low-degree polynomial p such that $\mathbf{E}_{\mathbf{x}\sim D_{\mathbf{x}}}[|p(\mathbf{x})-f(\mathbf{x})|]\leq\epsilon$.

In the distribution-specific setting, i.e., when $\mathbf{x}$ comes from the Gaussian or the uniform distribution on the hypercube, it is known that such a polynomial of degree $\mathrm{poly}(\Gamma/\epsilon)$ exists Klivans et al. (2008). However, without assumptions on D, low-degree polynomial approximations of f do not exist even when the f is a simple concept such as a linear threshold function.

Polynomial Approximation in the Low-Dimensional Space Our high-level plan is to treat the smoothed learning setting as a non-worst-case approximation setting and show that given some f, with high probability over the smoothing $\mathbf{z}$, the translated concept $\mathbf{x}\mapsto f(\mathbf{x}+\sigma\mathbf{z})$ will have a low-degree polynomial approximation. For simplicity, in this sketch, we will assume that $\sigma=1$. The general case can be found in the full proof; see Section 3.1 and also Remark 10. We will construct a family of polynomials $p_{\mathbf{z}}(\mathbf{x})$ such that their expected L_1 error over the smoothing $\mathbf{z}$ is small:

$$\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\Big[\mathop{\mathbf{E}}_{\mathbf{x}\sim D_{\mathbf{x}}}[|p_{\mathbf{z}}(\mathbf{x})-f(\mathbf{x}+\mathbf{z})|]\Big]\leq\epsilon\,.$$

We observe that since every $f(\mathbf{x})$ depends only on a k-dimensional space U, the projection of the input $\mathbf{x}$ down to U is just a linear transformation that does not affect the degree of polynomial approximation. Therefore, from now on, we may assume $\mathbf{x}$ lies in the k-dimensional space U and construct our polynomial approximation there.

Duality Between Input and Smoothing Parameter Our first step is to think of the smoothing random variable as the actual input to the function and treat $\mathbf{x}$ as a fixed parameter. Therefore, as a function of $\mathbf{z}$, we now have to approximate the translated function $f_{\mathbf{x}}(\mathbf{z}) = f(\mathbf{x}+\mathbf{z})$. Even though $\mathbf{z}$ is not available to the learner, when we think of $f_{\mathbf{x}}(\mathbf{z})$ as a function of the Gaussian noise random variable, we can utilize strong approximation results known under the Gaussian. In particular, we can replace the boolean function $f_{\mathbf{x}}(\mathbf{z})$ by its smooth approximation given by the Ornstein-Uhlenbeck operator defined as $T_\rho f_{\mathbf{x}}(\mathbf{z}) = \mathbf{E}_{\mathbf{s}\sim\mathcal{N}}[f_{\mathbf{x}}(\sqrt{1-\rho^2}\cdot\mathbf{z}+\rho\,\mathbf{s})]$.

Using the fact that the concept class of Definition 2 is closed under translation, we have that, since $\Gamma(f)\leq\Gamma$, the GSA of the translated concept $f_{\mathbf{x}}(\mathbf{z})$ as a function of $\mathbf{z}$ is also at most Γ. Using this fact and a result from Ledoux and Pisier (see Lemma 12) that bounds the L_1 approximation error of the Ornstein-Uhlenbeck noise operator, we obtain that with $\rho=\mathrm{poly}(\epsilon/\Gamma)$ it holds that

$$\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}[|T_\rho f_{\mathbf{x}}(\mathbf{z})-f_{\mathbf{x}}(\mathbf{z})|]\leq\epsilon\,.$$

So far, we replaced $f_{\mathbf{x}}$ with $T_\rho f_{\mathbf{x}}$, but have we made progress? We observe that $T_\rho f_{\mathbf{x}}(\mathbf{z}) = \mathbf{E}_{\mathbf{s}\sim\mathcal{N}}[f(\mathbf{x}+\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})]$. The variable $\mathbf{x}$, which is supposed to be input of the polynomial, is still in the function f. Without distributional assumptions on $D_{\mathbf{x}}$ the degree to approximate f can be arbitrarily large.

From Approximating $f(\cdot)$ to Approximating Density Ratios To avoid approximating the concept f, we observe that we can express the Ornstein-Uhlenbeck operator as follows:

$$T_\rho f_{\mathbf{x}}(\mathbf{z}) = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}(\mathbf{x}/\rho,\mathbf{I})}\left[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})\right] = \mathop{\mathbf{E}}_{\mathbf{s}\sim Q}\left[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})\cdot\frac{\mathcal{N}(\mathbf{s};\mathbf{x}/\rho,\mathbf{I})}{Q(\mathbf{s})}\right],$$

where $Q(\mathbf{s})$ is a distribution that we carefully design. We have managed to decouple the variable $\mathbf{x}$ from the function $f(\cdot)$, and now the task is to create a polynomial approximation of the density ratio $\frac{\mathcal{N}(\mathbf{s};\mathbf{x}/\rho,\mathbf{I})}{Q(\mathbf{s})}$, which — at the very least — is a continuous function of $\mathbf{x}$. For this to be possible, we need that the ratio of densities has a bounded L_1 norm with respect to $\mathbf{x}\sim D_{\mathbf{x}}$. When $\mathbf{x}$ is bounded, we can simply select Q to be the standard Gaussian; see Proposition 9. For sub-Gaussian (or strictly sub-exponential) marginals, we select a distribution Q with heavier (exponential) tails than $D_{\mathbf{x}}$. For this overview, we focus on the case of bounded marginals and refer to Section 3.2 for the more general result.

We observe that the approximating function has to be polynomial in $\mathbf{x}$ but can be an arbitrary function of $\mathbf{z}$ and $\mathbf{s}$. Therefore, we select a weighted combination of polynomials (that is still a polynomial in $\mathbf{x}$ but not a polynomial in $\mathbf{z}$):

$$p_{\mathbf{z}}(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{s}\sim Q}[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})\;q(\mathbf{x},\mathbf{s})]\,.$$

To bound the L_1 distance of $\mathrm{T}_\rho f_{\mathbf{x}}(\mathbf{z})$ and $p_{\mathbf{z}}(\mathbf{x})$, since f is boolean and, in particular, bounded, it suffices to show that the polynomial $q(\mathbf{x},\mathbf{s})$ approximates the ratio of normals $\mathcal{N}(\mathbf{s};\mathbf{x}/\rho,\mathbf{I})/\mathcal{N}(\mathbf{s})$. We construct an explicit polynomial approximation of this ratio using the Taylor expansion of the exponential function and show that a degree roughly $\mathrm{poly}(\log(1/\epsilon)/\rho)$ suffices; see Lemma 14. By our choice of ρ, we conclude that the degree of the family of polynomials $p_{\mathbf{z}}$ that we construct is at most $\mathrm{poly}(\Gamma/\epsilon)$.

Dimension Reduction and Polynomial Regression Having constructed polynomial approximations with high probability over the smoothing random variable $\mathbf{z}$, we can use the standard L_1 polynomial regression algorithm; see Kalai et al. (2008); Klivans et al. (2008). For the case of bounded marginals, we show that we can also perform a dimension-reduction preprocessing step by a random projection. Even though the class of concepts of Definition 2 is non-parametric, we show that under bounded GSA, it is possible to reduce the dimension to $\mathrm{poly}(k\Gamma/\epsilon)$; see Section 3.3.

2. Preliminaries and Notation

Notation We use small boldface characters for vectors and capital bold characters for matrices. We use $[d]$ to denote the set $\{1, 2, \ldots, d\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, $\mathbf{x}_i$ denotes the i-th coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^d \mathbf{x}_i^2}$ the ℓ_2 norm of $\mathbf{x}$. We use $\mathbf{x} \cdot \mathbf{y} := \sum_{i=1}^n \mathbf{x}_i \mathbf{y}_i$ as the inner product between them. We use $\mathbb{1}\{E\}$ to denote the indicator function of some event E. We use $\mathbf{E}_{\mathbf{x}\sim D}[f(\mathbf{x})]$ for the expectation of $f(\mathbf{x})$ according to the distribution D and $\mathbf{Pr}_D[E]$ for the probability of event E under D. For simplicity, we may omit the distribution when it is clear from the context. For $\mu \in \mathbb{R}^d, \mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, we denote by $\mathcal{N}(\mu, \mathbf{\Sigma})$ the d-dimensional Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$. We simply use $\mathcal{N}$ for the standard normal distribution. In cases where the dimension is not clear from the context we shall use $\mathcal{N}_k$ to denote the standard normal on k-dimensions. For $(\mathbf{x}, y)$ distributed according to D, we denote $D_{\mathbf{x}}$ to be the marginal distribution of $\mathbf{x}$.

3. Smoothed Agnostic Learning under Concentration

In this section, we present our algorithms for smoothed learning under bounded and (strictly) sub-exponential marginals. The polynomial approximation results are in Section 3.1 for bounded marginals and in Section 3.2 for strictly sub-exponential marginals. In Section 3.3 we present our algorithmic results and the dimension-reduction process for learning under bounded-marginals.

3.1. Polynomial Approximation: Bounded Marginals

In this section we present and prove our main polynomial approximation result for bounded marginals showing that, in expectation over the noise variable $\mathbf{z}$, there exists some polynomial $p_{\mathbf{z}}(\mathbf{x})$ that approximates the translated concept function $f(\mathbf{x}+\mathbf{z})$. The proof of Proposition 9 is split into two steps. Similar to our discussion in Section 1.2, we first fix $\mathbf{x}$ and try to approximate $f_{\mathbf{x}}(\mathbf{z})$. The first step is to replace f by its smoothed version $T_\rho f_{\mathbf{x}}$ (see Definition 11) and show that it is close to $f_{\mathbf{x}}$. The second step, see Lemma 13, is to construct a polynomial approximation of $T_\rho f_{\mathbf{x}}$ (similar to the way we constructed polynomial approximations to the Hermite coefficients of $f_{\mathbf{x}}$ in Section 1.2).

Proposition 9 (Polynomial Approximation of Random Translations) *Fix $\epsilon > 0$ and sufficiently large universal constant $C > 0$. Let D be a distribution on $\mathbb{R}^d$ such that all points $\mathbf{x}$ in the support of D have $\|\mathbf{x}\|_2 \le R$. Let $f \in \mathcal{F}(k, \Gamma)$. There exists a family of polynomials $p_{\mathbf{z}}$ parameterized by $\mathbf{z}$ of degree at most $C(\Gamma/\epsilon)^4 R^2 \log(1/\epsilon)$ such that $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}} \mathbf{E}_{\mathbf{x}\sim D}\left[|p_{\mathbf{z}}(\mathbf{x}) - f(\mathbf{x}+\mathbf{z})|\right]$ is at most ϵ, and every coefficient of $p_{\mathbf{z}}$ is bounded by $d^{C\left((\Gamma/\epsilon)^4 R^2 \log(1/\epsilon)\right)^2}$.*

Remark 10 *We remark that in Proposition 9 we have assumed that $\sigma = 1$ to simplify notation. Using the fact that the surface area bound of the concepts of Definition 2 is invariant under translation*

and positive scaling, we can apply Proposition 9 with $R' = R/\sigma$ for the function $\mathbf{x} \mapsto f(\sigma(\frac{\mathbf{x}}{\sigma} + \mathbf{z}))$ and obtain a polynomial of degree $\widetilde{O}((\Gamma/\epsilon)^4(R/\sigma)^2)$. See also Theorem 43.

Proof We use the following Gaussian noise operator to transform $f(\cdot)$ into a smooth function that is easier to approximate.

Definition 11 (Ornstein-Uhlenbeck Noise Operator) *Let $k \in \mathbb{N}$ and $\rho \in [0,1]$. We define the Ornstein-Uhlenbeck operator $T_\rho : \{\mathbb{R}^d \to \mathbb{R}\} \to \{\mathbb{R}^d \to \mathbb{R}\}$ that maps $f : \mathbb{R}^d \to \mathbb{R}$ to the function $T_\rho f : \mathbb{R}^d \to \mathbb{R}$ with $T_\rho f(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[f(\sqrt{1-\rho^2}\cdot\mathbf{x} + \rho\cdot\mathbf{z})]$.*

We will use the following result showing that under the assumption that some function g has bounded GSA, the Ornstein-Uhlenbeck operator $T_\rho g$ yields a good approximation to g in L_1.

Lemma 12 (Pisier (1986); Ledoux (1994)) *Let $\rho \in [0,1]$ and consider a function $f : \mathbb{R}^d \to \{\pm 1\}$. It holds $\mathbf{E}_{\mathbf{z}}\left[|T_\rho f(\mathbf{z}) - f(\mathbf{z})|\right] \le 2\sqrt{\pi\rho}\cdot\Gamma(f)$.*

Let $f_{\mathbf{x}}$ be the translated function defined as $f_{\mathbf{x}}(\mathbf{z}) = f(\mathbf{x}+\mathbf{z})$. From Lemma 12, we have $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[|T_\rho f_{\mathbf{x}}(\mathbf{z}) - f(\mathbf{z}+\mathbf{x})|\right] \le 2\sqrt{\pi\rho}\cdot\Gamma$.

Choosing $\rho = O(\epsilon^2/\Gamma^2)$ makes this error at most $\epsilon/2$. We now approximate $T_\rho f_{\mathbf{x}}$ using a polynomial. To do this we prove the following result. We provide a proof sketch here, and refer to the Supplementary Material for the details and the formal statement, see Lemma 37.

Lemma 13 (Approximating the Ornstein-Uhlenbeck Smoothed Concept $T_\rho f_{\mathbf{x}}(\cdot)$) *Let D be a distribution on $\mathbb{R}^d$ with every point $\mathbf{x}$ in the support of D having $\|\mathbf{x}\|_2$ at most R. Let $f : \mathbb{R}^d \to \{\pm 1\}$ and $f_{\mathbf{x}} : \mathbb{R}^d \to \mathbb{R}$ be defined as $f_{\mathbf{x}}(\mathbf{z}) = f(\mathbf{x}+\mathbf{z})$. Then, for any $\epsilon > 0$, there exist polynomials $p_{\mathbf{z}}$ parameterized by $\mathbf{z}$ for degree at most $O((R/\rho)^2\log(1/\epsilon))$, such that $\mathbf{E}_{\mathbf{x}\sim D}\,\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[|p_{\mathbf{z}}(\mathbf{x}) - T_\rho f_{\mathbf{x}}(\mathbf{z})|\right] \le \epsilon$.*

Before we prove Lemma 13 we use it to conclude the proof of Proposition 9. From Lemma 13, we get a polynomial $p_{\mathbf{z}}$ of degree $C(\Gamma/\epsilon)^4R^2\log(1/\epsilon)$ such that $\mathbf{E}_{\mathbf{x}\sim D}\,\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[|T_\rho f_{\mathbf{x}}(\mathbf{z}) - p_{\mathbf{z}}(\mathbf{x})|\right] \le \epsilon/2$ where C is a large universal constant. The coefficients of $p_{\mathbf{z}}$ are bounded by $d^{C\left((\Gamma/\epsilon)^4R^2\log(1/\epsilon)\right)^2}$. By a triangle inequality, we get $\mathbf{E}_{\mathbf{x}\sim D}\,\mathbf{E}_{\mathbf{z}\sim\mathcal{N}_k}\left[|p_{\mathbf{z}}(\mathbf{x}) - f(\mathbf{z}+\mathbf{x})|\right] \le \epsilon$.

Sketch of the Proof of Lemma 13 We observe that $T_\rho f_{\mathbf{x}}(\mathbf{z}) = \mathbf{E}_{\mathbf{s}\sim\mathcal{N}}[f(\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{s})]$ has the variable $\mathbf{x}$ inside f. Recall that our goal is to construct a polynomial in $\mathbf{x}$ and, since we have no control over f (which can possibly be very hard to approximate pointwise with a polynomial), we decouple f and $\mathbf{x}$ in the expression of $T_\rho f_{\mathbf{x}}$ by writing the function as an expectation over a Gaussian centered at $\mathbf{x}/\rho$.

$$T_\rho f_{\mathbf{x}}(\mathbf{z}) = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\left[f(\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{s})\right] = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}(\mathbf{x}/\rho,\mathbf{I})}\left[f(\sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{s})\right],$$

Next, we can recenter the expectation around zero and express the Ornstein-Uhlenbeck operator as follows:

$$T_\rho f_{\mathbf{x}}(\mathbf{z}) = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\left[f(\sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{s})\cdot\frac{\mathcal{N}(\mathbf{s};\mathbf{x}/\rho,\mathbf{I})}{N(\mathbf{s};\mathbf{0},\mathbf{I})}\right] = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\left[f(\sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{s})\cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2\rho^2} + (\mathbf{x}/\rho)\cdot\mathbf{s}}\right].$$

To construct our polynomial, we now approximate $e^{-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+(\mathbf{x}/\rho)\cdot\mathbf{s}}$ using the 1-dimensional Taylor expansion of the exponential function $q(\mathbf{x},\mathbf{s}) = q_m\big(-\frac{\|\mathbf{x}\|_2^2}{2\rho^2} + (\mathbf{x}/\rho)\cdot\mathbf{s}\big)$ where $q_m(t) = 1 + \sum_{i=1}^{m-1}\frac{t^i}{i!}$ is the degree $m-1$ Taylor approximation of e^x. Thus, our final polynomial $p_{\mathbf{z}}(\mathbf{x})$ is

$$p_{\mathbf{z}}(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\big[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})\cdot q(\mathbf{x},\mathbf{s})\big]\,. \tag{3}$$

Let $\Delta(\mathbf{x})$ be defined as the error term $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[|p_{\mathbf{z}}(\mathbf{x}) - T_\rho(f_{\mathbf{x}}(\mathbf{z}))|]$. We have that

$$\begin{aligned}\Delta(\mathbf{x}) &= \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\Big[\mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\big[|f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})|\cdot\big|q(\mathbf{x},\mathbf{s}) - e^{-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+(\mathbf{x}/\rho)\cdot\mathbf{s}}\big|\big]\Big]\\ &\le \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\Big[\big|q(\mathbf{x},\mathbf{s}) - e^{-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+(\mathbf{x}/\rho)\cdot\mathbf{s}}\big|\Big]\,,\end{aligned} \tag{4}$$

where for the inequality we used the fact that $|f(\mathbf{x})| = 1$ for all $\mathbf{x}$. We now observe that when $\mathbf{s}\sim\mathcal{N}$ the random variable $-\|\mathbf{x}\|_2^2/(2\rho^2) + (\mathbf{x}/\rho)\cdot\mathbf{s}$ is distributed as $\mathcal{N}(-\alpha^2/2,\alpha^2)$, where $\alpha = -\|\mathbf{x}\|_2^2/\rho^2$. Therefore, we have reduced the original polynomial approximation problem to showing that the Taylor expansion of the exponential function converges fast in L_1 to e^x with respect to $\mathcal{N}(-\alpha^2/2,\alpha^2)$. The proof of the following lemma is technical and can be found in the Supplementary Materical (see Lemma 37). Here we give a heuristic argument.

Lemma 14 (Approximation of e^x with respect to $\mathcal{N}(-\alpha^2/2,\alpha^2)$) *Fix $\alpha > 0$ and sufficiently large universal constant $C > 0$. Let p be the polynomial $p(x) = \sum_{i=0}^{m-1}\frac{x^i}{i!}$ with $m = C\alpha^2\log(1/\epsilon)$. We have that $\mathbf{E}_{x\sim\mathcal{N}(-\alpha^2/2,\alpha^2)}[|e^x - p(x)|] \le \epsilon$.*

Proof [Sketch] We first observe that since the Gaussian has mean $-\alpha^2/2$ and variance α^2 using the strong concentration of the Gaussian (whose tail decays faster than the exponential growth of e^x and its Taylor expansion, see Lemma 37 for more details) we may assume that we only have to approximate the exponential function in the interval $[-\alpha^2/2 - O(\alpha\sqrt{\log(1/\epsilon)}), -\alpha^2/2 + O(\alpha\sqrt{\log(1/\epsilon)})]$. By Taylor's theorem we have that for any interval $[a,b]$ it holds that $|p(x)-e^x| \le e^b\max(|a|,|b|)^m/m!$. Therefore, we have that by picking degree $m = O(\alpha^2\log(1/\epsilon))$ we can make the error of the Taylor expansion at most ϵ. ■

Using Lemma 14 with $\alpha = \|\mathbf{x}/\rho\|_2$ we obtain that with degree $O((R/\rho)^2\log(1/\epsilon))$ the L_1 error of the polynomial $q(\mathbf{x},\mathbf{s})$ in Equation (4) is at most ϵ. To bound the coefficients of the polynomial $p_{\mathbf{z}}(\mathbf{x})$ we use the fact that $f(\mathbf{x})$ is boolean (and therefore bounded) and the fact that the input of the Taylor expansion in $q(\mathbf{x},\mathbf{s})$ is bounded. For the full proof, see the Supplementary Material. ■

3.2. Polynomial Approximation: Strictly Sub-Exponential Marginals

In this section we prove our polynomial approximation for the more general class of Strictly Sub-Exponential distributions, defined as follows.

Definition 15 (Strictly Sub-exponential Distributions) *A distribution D on $\mathbb{R}^d$ is (α,λ)-strictly sub-exponential for $\alpha,\lambda > 0$ if for all $\|\mathbf{v}\|_2 = 1$, $\mathbf{Pr}_{\mathbf{x}\sim D}[|\mathbf{x}\cdot\mathbf{v}| > t] \le 2\cdot e^{-(t/\lambda)^{1+\alpha}}$.*

Our main goal in this section is to prove the following polynomial approximation result which is a generalization of Proposition 9. We refer to Lemma 50 in the appendix for the formal statement.

Proposition 16 (Polynomial Approximation: Strictly Sub-Exponential Marginals) *Let C be a large universal constant. Let D be a distribution on $\mathbb{R}^k$ that is (α,λ)-strictly subexponential. Let $f : \mathbb{R}^k \mapsto \{\pm 1\}$ be a boolean function in $\mathcal{F}(k,\Gamma)$. Then there exist polynomials $p_{\mathbf{z}}$ of degree at most $\left(C\lambda k\Gamma^2 \log(1/\epsilon)/\epsilon^2\right)^{64(1+1/\alpha)^3}$, parameterized by $\mathbf{z}$ whose (expected) L_1 error is $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}_k}\mathbf{E}_{\mathbf{u}\sim D}\left[|p_{\mathbf{z}}(\mathbf{u}) - f(\mathbf{z}+\mathbf{u})|\right] \leq \epsilon$.*

The main proof idea is similar to that of Proposition 9. However, there are significantly more technical hurdles in constructing the approximating polynomial for this case and we will only highlight some of the main differences and refer to the Supplementary Material for the full proof. Similar to the proof of Proposition 9, by using the result of Ledoux and Pisier Lemma 12 we obtain that it suffices to approximate the function $T_\rho f_{\mathbf{x}}(\mathbf{z})$ with some polynomial $p_{\mathbf{z}}(\mathbf{x})$. Since f is low-dimensional (see Definition 2) we can write $f(\mathbf{x}) = f(\mathbf{U}^T\mathbf{U}\mathbf{x})$ for some $k \times d$ projection matrix $\mathbf{U}$. Since the polynomial regression algorithm is able to learn this linear transformation, from now on we assume that f is an explicit k dimensional function $f(\mathbf{u}) : \mathbb{R}^k \mapsto \{\pm 1\}$. We will show that there exists a polynomial of degree at most $(C\lambda k\log(1/\epsilon)/\rho)^{64(1+1/\alpha)^3}$ that approximates $T_\rho f_{\mathbf{u}}(\mathbf{z})$. Similar to the proof of Proposition 9, the first step is to re-write the expression of $T_\rho f_{\mathbf{u}}(\mathbf{z})$ so that $\mathbf{u}$ does not appear inside the target function f. We observe that for any distribution Q we have

$$\begin{aligned} T_\rho f_{\mathbf{u}}(\mathbf{z}) &= \mathop{\mathbf{E}}_{\mathbf{s}\sim Q}\left[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})\cdot\frac{\mathcal{N}(\mathbf{s};\mathbf{u}/\rho,I)}{Q(\mathbf{s})}\right] \\ &= e^{-\frac{\|\mathbf{u}\|_2^2}{2\rho^2}}\mathop{\mathbf{E}}_{\mathbf{s}\sim Q}\left[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})\cdot e^{-\frac{\|\mathbf{s}\|_2^2}{2}-\log Q(\mathbf{s})}e^{(\mathbf{u}/\rho)\cdot\mathbf{s}}\right]. \end{aligned}$$

We observe that we can no longer take Q to be a Gaussian (like we did in Proposition 9) because when $\mathbf{u}$ has weaker tails than the normal density the $\mathbf{E}_{\mathbf{u}\sim D}[(\frac{\mathcal{N}(\mathbf{s};\mathbf{u}/\rho,I)}{Q(\mathbf{s})})^2] = +\infty$. To avoid this we take Q to be the distribution on $\mathbb{R}^k$ with probability distribution function $Q(\mathbf{s}) \propto e^{-\|\mathbf{s}\|_1}$ which has exponential tails. We show, see Lemma 53 in Supplementary Material, that $\mathbf{E}_{\mathbf{x}\sim Q}[(\frac{\mathcal{N}(\mathbf{x};\mathbf{u},\mathbf{I})}{Q(\mathbf{x})})^2] \leq C^k e^{C\|\mathbf{u}\|_1}$. Beyond working with the exponential reweighting function, another technical complication is that we now have to carefully create a polynomial approximation over a strictly sub-exponential distribution for the function $e^{-\|\mathbf{s}\|_2^2}$, see Lemma 52 in Supplementary Material. To do this we use a tighter polynomial approximation using Chebyshev polynomials.

3.3. Efficient Algorithms for Learning under Concentration

Given the polynomial approximation construction of the previous sections one can directly run L_1 polynomial regression to minimize $\mathbf{E}_{(\mathbf{x},y)\sim D}[|p(\mathbf{x})-y|]$ similar to Kalai et al. (2008). We now state our main theorem for strictly sub-exponential distributions.

Theorem 17 *Let $k \in \mathbb{Z}_+$ and $\epsilon,\delta,\sigma \in (0,1)$. Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that the marginal distribution is (α,λ)-strictly subexponential. There exists an algorithm that draws $N = d^{\mathrm{poly}((k\lambda\Gamma/(\sigma\epsilon))^{(1+1/\alpha)^3})}$ samples, runs in time $\mathrm{poly}(d,N)$, and computes a hypothesis $h(\mathbf{x})$ such that, with probability at least $1-\delta$, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq h(\mathbf{x})] \leq \mathrm{opt}_\sigma + \epsilon$.*

In the case of bounded marginals, we can significantly reduce the runtime of the algorithm by performing a dimension reduction via a random Gaussian projection similar to the works of Arriaga and Vempala (1999a) and Klivans and Servedio (2004). We show that when the $\mathbf{x}$-marginal of

the distribution is bounded then we can perform a random projection to reduce dimension down to $\mathrm{poly}(k\Gamma/\epsilon)$ for the class of concepts of Definition 2. Assuming that $f \in \mathcal{F}(k,\Gamma)$ we have that there exists a $k \times d$ matrix $\mathbf{U}$ such that $f(\mathbf{x}) = f(\mathbf{U}^T\mathbf{U}\mathbf{x})$. Let $\mathbf{A}$ be the random projection matrix. It suffices to show that concept $f(\mathbf{A}\mathbf{x})$ is close in L_1 to the original concept $f(\mathbf{x})$. We once again use the fact that we can exchange the order of expectation so that we are able to use the properties of the random Gaussian smoothing. We show, see Lemma 46 in the Supplementary Material, that for every $f \in \mathcal{F}(k,\Gamma)$ it holds that $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\big[|f(\mathbf{u}+\mathbf{z}) - f(\mathbf{v}+\mathbf{z})|\big] \leq O(\Gamma\cdot\|\mathbf{u}-\mathbf{v}\|_2)$. Therefore, we obtain that a random projection down to $\mathrm{poly}(k\Gamma/\epsilon)$ dimensions will imply that $\mathbf{E}_{\mathbf{x}\sim D_\mathbf{x}}\,\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[|f(\mathbf{A}\mathbf{x}+\mathbf{z}) - f(\mathbf{x}+\mathbf{z})|] \leq \epsilon$. By performing polynomial regression in the low-dimensional space we obtain the following improved runtime for bounded $\mathbf{x}$-marginals.

Theorem 18 *Let $k \in \mathbb{Z}_+$ and $\epsilon,\delta,\sigma \in (0,1)$. Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is bounded in the unit ball. There is an algorithm that draws $N = k^{\tilde{O}\left((\Gamma/\epsilon)^4(1/\sigma^2)\right)}\log(\frac{1}{\delta})$ samples, runs in time $\mathrm{poly}(d,N)$, and computes a hypothesis $h(\mathbf{x})$ such that, with probability at least $1-\delta$, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq h(\mathbf{x})] \leq \mathrm{opt}_\sigma + \epsilon$.*

4. Applications and Connections with Other Models

In this section, we show connections between our model of smoothed learning and three important models that have been previously studied: (1) learning with margin, (2) learning under smoothed distributions and (3) learning with concentration and anti-concentration. We briefly discuss (1) and (2) and defer (3) and other details to the Supplementary Material, see Section B.

Learning with Margin We show that any algorithm for smoothed agnostic learning can be directly used to learn in the agnostic setting with margin. For the formal definition of agnostic learning with γ-margin we refer to Equation (2) and Definition 22. We denote by $\partial_\gamma f$ all points that are in distance at most γ from the decision boundary. We observe (see Lemma 25) that opt_σ is not much larger than margin-opt$_\gamma$, as long we have that for any $\mathbf{x} \notin \partial_\gamma$ it holds that the value of f is unlikely to change by the random perturbation:

$$\mathrm{opt}_\sigma \leq \text{margin-opt}_\gamma + \sup_{\mathbf{x}\notin\partial_\gamma f} \mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x}+\sigma\mathbf{z}) \neq f(\mathbf{x})]\,.$$

For any boolean concept f, we show (see Lemma 26) that as long as σ is smaller than $\frac{\gamma}{\sqrt{k\log(1/\epsilon)}}$ it holds that $\sup_{\mathbf{x}\notin\partial_\gamma f}\mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x}+\sigma\mathbf{z}) \neq f(\mathbf{x})] \leq \epsilon$. While this holds in full generality, for specific concept classes we are able to provide better bounds. In particular, for intersections of k halfspaces we show, see Lemma 27, that picking $\sigma = \gamma/\sqrt{\log k/\epsilon}$ suffices. Therefore, using Theorem 43 we readily obtain the agnostic learning result for intersections of k-halfspaces of Corollary 6.

Agnostic Learning with Distributional Assumptions As mentioned, our smoothed agnostic model generalizes agnostic learning with distributional assumptions. We denote by opt the standard optimal agnostic error under a distribution D. We see (see Lemma 30 in the Supplementary Material) that $\mathrm{opt}_\sigma \leq \mathrm{opt} + \mathbf{Pr}_{\mathbf{x}\sim D_\mathbf{x},\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x}+\sigma\mathbf{z}) \neq f(\mathbf{x})]$. For the case of distribution smoothing we have that the smoothed distribution D_τ is the convolution of $D_\mathbf{x} + \mathcal{N}(\mathbf{0},\tau^2\mathbf{I})$. In that case we have that $\mathbf{E}_{\mathbf{x}\sim D_\mathbf{x},\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x}+\tau\mathbf{z}_1+\sigma\mathbf{z}_2) \neq f(\mathbf{x}+\tau\mathbf{z}_1)] \leq O(\frac{\sigma\Gamma\sqrt{k}}{\tau})$. Therefore, by choosing $\sigma = O(\epsilon\tau/(\Gamma\sqrt{k}))$, we obtain that the gap between opt_σ and opt is at most ϵ. For this value of σ,

we are able to recover the strong results of Corollary 7 which yields an exponential improvement over the prior work Kane et al. (2013).

5. Conclusion and Open Problems

In this work we introduce a new beyond worst-case model for agnostic learning and show that it is possible to obtain efficient algorithms with runtime that were previously known only under very strong distributional assumptions, e.g., Gaussianity. Moreover, we show that our framework and results generalize over several settings considered in the literature — often improving the best known results significantly (e.g., for the fundamental problem of learning intersections of k halfspaces with margin). There are many interesting open questions in smoothed agnostic learning: Can we improve the runtime of Theorem 4 and remove or make milder the exponential dependency on the intrinsic dimension k? Is it possible to generalize the result beyond (strictly) sub-exponential tails? It seems that when the adversary is left completely unrestricted to pick instances with arbitrarily large norm $\|\mathbf{x}\|$, the effect of Gaussian smoothing of Definition 1 is negligible. What are the weakest assumptions on the $\mathbf{x}$-marginal that enable learnability?

Acknowledgments

We thank the anonymous reviewers of COLT 2024 for their valuable feedback. Gautam Chandrasekaran was supported by the NSF AI Institute for Foundations of Machine Learning (IFML). Adam Klivans was supported by NSF award AF-1909204 and the NSF AI Institute for Foundations of Machine Learning (IFML). Vasilis Kontonis was supported by the NSF AI Institute for Foundations of Machine Learning (IFML). Raghu Meka was supported by NSF Collaborative Research: Award 2217033. Konstantinos Stavropoulos was supported by the NSF AI Institute for Foundations of Machine Learning (IFML) and by scholarships from Bodossaki Foundation and Leventis Foundation.

References

Amol Aggarwal and Josh Alman. Optimal-degree polynomial approximations for exponentials and gaussian kernel density estimation. In *Proceedings of the 37th Computational Complexity Conference*, CCC '22, Dagstuhl, DEU, 2022. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 9783959772419. doi: 10.4230/LIPIcs.CCC.2022.22. URL https://doi.org/10.4230/LIPIcs.CCC.2022.22.

R. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 616–623, New York, NY, 1999a.

R.I. Arriaga and S. Vempala. An algorithmic theory of learning: robust concepts and random projection. In *40th Annual Symposium on Foundations of Computer Science (Cat. No.99CB37039)*, pages 616–623, 1999b. doi: 10.1109/SFFCS.1999.814637.

P. Awasthi, M. F. Balcan, N. Haghtalab, and R. Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 167–190, 2015.

P. Awasthi, M. F. Balcan, N. Haghtalab, and H. Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pages 152–192, 2016.

P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6):50:1–50:27, 2017.

K. Ball. The Reverse Isoperimetric Problem for Gaussian Measure. *Discrete and Computational Geometry*, 10:411–420, 1993.

S. Ben-David and H. U. Simon. Efficient learning of linear perceptrons. *Advances in Neural Information Processing Systems 14*, 2000.

Avrim Blum and Ravi Kannan. Learning an intersection of k halfspaces over a uniform distribution. In *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*, pages 312–320. IEEE, 1993.

M. Brennan, G. Bresler, S. B. Hopkins, J. Li, and T. Schramm. Statistical query algorithms and low-degree tests are almost equivalent. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021.

S. Chen, F. Koehler, A. Moitra, and M. Yau. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

D. Dachman-Soled, H. Lee, T. Malkin, R. Servedio, A. Wan, and H. Wee. Optimal cryptographic hardness of learning monotone functions. In *Proc. 35th International Colloquium on Algorithms, Languages and Programming (ICALP)*, pages 36–47, 2008.

A. Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual Symposium on Theory of Computing, STOC 2016*, pages 105–117, 2016.

A. Daniely and G. Vardi. From local pseudorandom generators to hardness of learning. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1358–1394. PMLR, 2021. URL http://proceedings.mlr.press/v134/daniely21a.html.

A. De, E. Mossel, and J. Neeman. Robust testing of low dimensional functions. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 584–597. ACM, 2021.

Anindya De, Elchanan Mossel, and Joe Neeman. Is your function low dimensional? In *Conference on Learning Theory*, pages 979–993. PMLR, 2019.

I. Diakonikolas, R. O'Donnell, R. Servedio, and Y. Wu. Hardness results for agnostically learning low-degree polynomial threshold functions. In *SODA*, pages 1590–1606, 2011.

I. Diakonikolas, T. Gouleakis, and C. Tzamos. Distribution-independent pac learning of halfspaces with massart noise. In *Advances in Neural Information Processing Systems 32, NeurIPS*. 2019a.

I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory, COLT*, 2020.

I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021.

I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Learning general halfspaces with general massart noise under the gaussian distribution. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 874–885. ACM, 2022.

Ilias Diakonikolas, Daniel Kane, and Pasin Manurangsi. Nearly tight bounds for robust proper learning of halfspaces with a margin. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

V. Feldman. Optimal hardness results for maximizing agreements with monomials. In *IEEE Conference on Computational Complexity*, pages 226–236. IEEE Computer Society, 2006. ISBN 0-7695-2596-2. doi: 10.1109/CCC.2006.31.

V. Feldman. Statistical query learning. In *Encyclopedia of Algorithms*, pages 2090–2095. 2016.

V. Feldman, V. Guruswami, P. Raghavendra, and Y. Wu. Agnostic learning of monomials by halfspaces is hard. In *FOCS*, pages 385–394, 2009.

V. Feldman, H. Lee, and R. Servedio. Lower bounds and hardness amplification for learning shallow monotone formulas. In *COLT*, pages 273–292, 2011.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM*, 64(2):8:1–8:37, 2017a.

V. Feldman, C. Guzman, and S. S. Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In Philip N. Klein, editor, *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017*, pages 1265–1277. SIAM, 2017b.

J. H. Friedman, M. Jacobson, and W. Stuetzle. Projection Pursuit Regression. *J. Am. Statist. Assoc.*, 76:817, 1981. doi: 10.2307/2287576.

Aravind Gollakota, Adam R. Klivans, and Pravesh K. Kothari. A moment-matching approach to testable learning and a new characterization of rademacher complexity. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 1657–1670, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399135. doi: 10.1145/3564246.3585206. URL https://doi.org/10.1145/3564246.3585206.

P. Gopalan, A. Kalai, and A. Klivans. Agnostically learning decision trees. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 527–536, 2008.

V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. In *Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 543–552. IEEE Computer Society, 2006.

P. Hall and K.-C. Li. On almost Linearity of Low Dimensional Projections from High Dimensional Data. *The Annals of Statistics*, 21(2):867 – 889, 1993.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

P. J. Huber. Projection Pursuit. *The Annals of Statistics*, 13(2):435 – 475, 1985.

A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2005.

A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008. Special issue for FOCS 2005.

Adam Tauman Kalai and Shang-Hua Teng. Decision trees are pac-learnable from most product distributions: a smoothed analysis. *arXiv preprint arXiv:0812.0933*, 2008.

Adam Tauman Kalai, Alex Samorodnitsky, and Shang-Hua Teng. Learning and smoothed analysis. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 395–404. IEEE, 2009.

D. M. Kane. The gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. *Computational Complexity*, 20(2):389–412, 2011.

Daniel Kane. The gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. volume 20, pages 205–210, 06 2010. doi: 10.1109/CCC.2010.27.

Daniel Kane, Adam Klivans, and Raghu Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 522–545, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL https://proceedings.mlr.press/v30/Kane13.html.

M. Kearns, R. Schapire, and L. Sellie. Toward Efficient Agnostic Learning. *Machine Learning*, 17 (2/3):115–141, 1994.

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

M. Kharitonov. Cryptographic hardness of distribution-specific learning. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of Computing*, STOC '93. Association for Computing Machinery, 1993. ISBN 0897915917. doi: 10.1145/167088.167197.

S. Khot and R. Saket. On hardness of learning intersection of two halfspaces. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 345–354, 2008.

A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via Gaussian surface area. In *Proc. 49th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 541–550, Philadelphia, Pennsylvania, 2008.

Adam R. Klivans and Rocco A. Servedio. Learning intersections of halfspaces with a margin. In John Shawe-Taylor and Yoram Singer, editors, *Learning Theory*, pages 348–362, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

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

V. Kontonis, C. Tzamos, and M. Zampetakis. Efficient truncated statistics with unknown truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595. IEEE, 2019.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.

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

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform and learnability. *Journal of the ACM*, 40(3):607–620, 1993.

P. Long. An upper bound on the sample complexity of PAC learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–234, 2003.

M. Minsky and S. Papert. *Perceptrons: an introduction to computational geometry (expanded edition)*. MIT Press, Cambridge, MA, 1988.

J. Neeman. Testing surface area with arbitrary accuracy. In *Symposium on Theory of Computing, STOC 2014, 2014*, pages 393–397. ACM, 2014.

G. Pisier. Probabilistic methods in the geometry of Banach spaces. In *Lecture notes in Math.*, pages 167–241. Springer, 1986.

F. Rosenblatt. *Principles of neurodynamics*. Springer-Verlag, New York, 1962.

Daniel A Spielman. The smoothed analysis of algorithms. In *Fundamentals of Computation Theory: 15th International Symposium, FCT 2005, Lübeck, Germany, August 17-20, 2005. Proceedings 15*, pages 17–18. Springer, 2005.

Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984a.

L. G. Valiant. A theory of the learnable. In *Proc. 16th Annual ACM Symposium on Theory of Computing (STOC)*, pages 436–445. ACM Press, 1984b.

S. Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *J. ACM*, 57(6):32:1–32:14, 2010.

Santosh S Vempala and Ying Xiao. Structure from local optima: Learning subspace juntas via higher order pca. *arXiv preprint arXiv:1108.3329*, 2011.

Y. Xia. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484):1631–1640, 2008.

Y. Xia, H. Tong, W. K. Li, and L. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):363–410, 2002.

C. Zhang, J. Shen, and P. Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

Appendix A. Gaussian Surface Area

Here we give the formal definition of Gaussian Surface Area of a concept and present some known bounds for the concept classes that we consider in this work.

Definition 19 (Gaussian Surface Area) *Let f be a boolean function, $\Gamma(f)$ is the Gaussian surface area of the set $A_f = \{\mathbf{u} \in \mathbb{R}^k : f(\mathbf{u}) = 1\}$, i.e., $\Gamma(f)$ is the following quantity.*

$$\Gamma(f) = \liminf_{\delta \to 0} \frac{1}{\delta} \Pr_{\mathbf{z} \sim \mathcal{N}(0, I_k)} \left[\mathbf{z} \in A_f^\delta \setminus A_f\right], \text{ where } A_f^\delta = \{\mathbf{u} : \min_{\mathbf{v} \in A_f} \|\mathbf{u} - \mathbf{v}\|_2 \leq \delta\}$$

We use the following bounds on Gaussian surface area in our results.

Lemma 20 (Bounds on Gaussian surface area of various functions) *The following are bounds on the Gaussian surface area of some common classes of functions:*

1. *If f is a halfspace, then $\Gamma(f) \leq O(1)$Klivans et al. (2008),*
2. *If f is an intersection of k halfspaces, then $\Gamma(f) \leq O(\sqrt{\log k})$(Klivans et al. (2008), due to Nazarov),*
3. *If f is an arbitrary boolean function of k halfspaces, then $\Gamma(f) \leq O(k)$ (folklore, see also Lemma 21),*
4. *If f is the degree ℓ polynomial threshold function(PTF), then $\Gamma(f) \leq O(\ell)$Kane (2010),*
5. *if f is an arbitrary convex set on k variables, then $\Gamma(f) \leq O(k^{1/4})$Ball (1993).*

Lemma 21 (Gaussian Surface Area of functions of k halfspaces) *Let f be a boolean function on k halfspaces. Then, we have that $\Gamma(f) \leq O(k)$.*

Proof Since f is a boolean function on k halfspaces, we have that for any input $\mathbf{x}$, $f(\mathbf{x}) = g\big(h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_k(\mathbf{x})\big)$ where $h_1, \ldots, h_k$ are halfspaces on $\mathbb{R}^d$ and g is an arbitrary boolean function. For any function h, let $A_h = \{\mathbf{x} \in \mathbb{R}^d : h(\mathbf{x}) = 1\}$. For any set S, let S^δ denote the set of points at distance at most δ from S. For a boolean function h, let h^c denote it's complement.

Observe that $A_f^\delta \setminus A_f \subseteq \big(\bigcup_{i=1}^k (A_{h_i}^\delta \setminus A_{h_i})\big) \cup \big(\bigcup_{i=1}^k (A_{h_i^c}^\delta \setminus A_{h_i^c})\big)$. Then, we have that

$$\Gamma(f) = \liminf_{\delta \to 0} \frac{1}{\delta} \Pr_{\mathbf{z} \sim \mathcal{N}(0, I_k)} \left[\mathbf{z} \in A_f^\delta \setminus A_f\right] \leq \sum_{i=1}^k \big(\Gamma(h_i) + \Gamma(h_i^c)\big) \leq O(k)$$

where the first inequality comes from a union bound and the second comes from the bound on surface area of a single halfspace(Lemma 20). ■

Appendix B. Applications and Connections with Other Models

In this section, we show connections between our model of smoothed learning and three important models that have been previously studied: (1) learning with margin, (2) learning under smoothed distributions and (3) learning with concentration and anticoncentration. We prove that our results straightforwardly imply improved results in each of these models. For example, we give the first quasi-polynomial algorithm for agnostically learning intersections of halfspaces with margin.

Concept Class	Bounded	Sub-Gaussian
Intersections of k halfspaces	$\mathrm{poly}(d) \cdot k^{\mathrm{poly}(\frac{\log k}{\epsilon\sigma})}$	$d^{\mathrm{poly}(\frac{k}{\sigma\epsilon})}$
Arbitrary functions of k halfspaces	$\mathrm{poly}(d) \cdot k^{\mathrm{poly}(\frac{k}{\gamma\epsilon})}$	$d^{\mathrm{poly}(\frac{k}{\sigma\epsilon})}$
k-dimensional, ℓ-degree PTFs	$\mathrm{poly}(d) \cdot k^{\mathrm{poly}(\frac{k\ell}{\gamma\epsilon})}$	$d^{\mathrm{poly}(\frac{k\ell}{\sigma\epsilon})}$
k-dimensional convex sets	$\mathrm{poly}(d) \cdot k^{\mathrm{poly}(\frac{k}{\gamma\epsilon})}$	$d^{\mathrm{poly}(\frac{k}{\sigma\epsilon})}$

Table 1: Our results on smoothed agnostic learning.

B.1. Notation

We introduce some notation for that we use in this section. For distribution D on $\mathbb{R}^d \times \{\pm 1\}$ and function $f : \mathbb{R}^d \to \{\pm 1\}$, let $\mathrm{err}(f, D) = \mathbf{E}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]$ and let $\mathrm{err}_\sigma(f, D) = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}} E_{(\mathbf{x},y)\sim D}[f(\mathbf{x} + \sigma\mathbf{z}) \neq y]$.

B.2. Learning with Margin

In this section we investigate the connection of our smoothed learning model with (agnostic) learning with margin. We first define the model. As discussed in the introduction, this model disincentivizes the adversary from placing points close to the boundary of the function in a bid to create worst case instances.

Definition 22 (Agnostic Learning with Margin) *Fix $\epsilon, \gamma > 0$ and $\delta \in (0, 1)$. Let $\mathcal{F} \subseteq \{\mathbb{R}^d \to \pm 1\}$ be a class of Boolean concepts and let $\mathbb{D}$ be a class of distributions over $\mathbb{R}^d$. Consider D to be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ such that $D_{\mathbf{x}} \in \mathbb{D}$. We say that the algorithm $\mathcal{A}$ learns $\mathcal{F}$ in the γ-margin setting if, after receiving i.i.d. samples from D, $\mathcal{A}$ outputs a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ such that, with probability at least $1 - \delta$, over the samples it holds*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y \neq h(\mathbf{x})] \leq \inf_{f\in\mathcal{F}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left[\sup_{\|\mathbf{u}\|_2\leq\gamma} \mathbb{1}\{f(\mathbf{x}+\mathbf{u}) \neq y\}\right] + \epsilon\,. \tag{5}$$

Remark 23 (Other definitions of agnostic learning with margin) *We now highlight connections to other previously studied notions of margin. An equivalent model to Definition 22 is to define the γ-margin optimal error $\inf_{f\mathcal{F}} \mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbb{1}\{f(\mathbf{x}) \neq y \text{ or } \mathbf{x} \in \partial_\gamma f\}]$, see, e.g., Diakonikolas et al. (2019b). Another related model is defined in Klivans and Servedio (2004) where they define the set $\mathcal{F}_\gamma$ containing all functions $f \in \mathcal{F}$ that have γ-margin with respect to $D_{\mathbf{x}}$, i.e., $f \in \mathcal{F}_\gamma$ if $f \in \mathcal{F}$ and $\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[x \in \partial_\gamma f] = 0$ and define the (margin) optimal error as $\inf_{f\in\mathcal{F}_\gamma} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]$. We remark that our result readily applies to this variant as well.*

A key tool that we use in our reductions is the notion of *Gaussian sensitivity* of a function at a point, which we now define.

Definition 24 (Gaussian Sensitivity at $\mathbf{x}$) *Let $f : \mathbb{R}^d \mapsto \{\pm 1\}^d$ be a boolean function. We define the Gaussian σ-sensitivity of f at $\mathbf{x}$ as $\mathbb{S}(\mathbf{x}; \sigma, f) \coloneqq \mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x} + \sigma\mathbf{z}) \neq f(\mathbf{x})]$.*

We now prove our reduction. We show that a learner for the smooth agnostic model of Definition 1 can readily give an algorithm for agnostic learning with margin if we have bounds on the Gaussian sensitivity at all points not in a γ neighbourhood of the function's surface.

Lemma 25 (From Margin to Smooth Agnostic) *Fix some boolean function $f : \mathbb{R}^d \mapsto \{\pm 1\}^d$ and some distribution D over labeled examples on $\mathbb{R}^d \times \{\pm 1\}$. We say that a point $\mathbf{x}$ lies in the γ-boundary of f, $\mathbf{x} \in \partial_\gamma f$, if there exists $\mathbf{u}$ with $\|\mathbf{u}\|_2 \leq \gamma$, such that $f(\mathbf{x}+\mathbf{u}) \neq f(\mathbf{x})$. It holds*

$$\mathrm{err}_\sigma(f, D) \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left[\sup_{\|\mathbf{u}\|_2 \leq \gamma} \mathbb{1}\{f(\mathbf{x}+\mathbf{u}) \neq y\}\right] + \sup_{\mathbf{x}\notin\partial_\gamma f} \mathbb{S}(\mathbf{x}; \sigma, f)\,.$$

Proof We first observe that we change the order of expectations in the smoothed error of f and consider $\mathbf{E}_{(\mathbf{x},y)\sim D}\,\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[\mathbb{1}\{f(\mathbf{x}+\sigma\mathbf{z}) \neq y\}\right]$. Now it suffices to show that for every $(\mathbf{x}, y)$ it holds

$$\mathop{\mathbf{Pr}}_{\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x}+\sigma\mathbf{z}) \neq y] \leq \sup_{\|\mathbf{u}\|_2 \leq \gamma} \mathbb{1}\{f(\mathbf{x}+\mathbf{u}) \neq y\} + \sup_{\mathbf{x}\notin\partial_\gamma f} \mathbb{S}(\mathbf{x}; \sigma, f)\,.$$

For any labeled example $(\mathbf{x}, y)$ where $\mathbf{x} \in \partial_\gamma f$ we have that for any $y \in \{\pm 1\}$ there exists a $\mathbf{u}$ with $\|\mathbf{u}\|_2 \leq \gamma$ such that $f(\mathbf{x}+\mathbf{u}) \neq y$: if $y \neq f(\mathbf{x})$ then pick $\mathbf{u} = \mathbf{0}$ and if $y = f(\mathbf{x})$ there exists $\mathbf{u}$ with $\|\mathbf{u}\|_2 \leq \gamma$ such that $f(\mathbf{x}+\mathbf{u}) \neq f(\mathbf{x})$ and therefore $f(\mathbf{x}+\mathbf{u}) \neq y$. When $\mathbf{x} \notin \partial_\gamma f$ we have that $\sup_{\|\mathbf{u}\|_2 \leq \gamma} \mathbb{1}\{f(\mathbf{x}+\mathbf{u}) \neq y\} = \mathbb{1}\{f(\mathbf{x}) \neq y\}$. By the triangle inequality we have $\mathbb{1}\{f(\mathbf{x}+\sigma\mathbf{z}) \neq y\} - \mathbb{1}\{f(\mathbf{x}) \neq y\} \leq \mathbb{1}\{f(\mathbf{x}+\sigma\mathbf{z}) \neq f(\mathbf{x})\}$. Therefore, by taking the expectation with respect to $\mathbf{z} \sim \mathcal{N}$ we obtain the result. ■

Before, we can use the above lemma to get results in the margin setting, we need to bound the Gaussian sensitivity(at points γ distance away from the surface) of various concepts. First, we bound this quantity for an arbitrary function. We then obtain stronger bounds for the class of intersections of halfspaces.

Lemma 26 (Gaussian Sensitivity of arbitrary functions under γ-margin) *Let $f : \mathbb{R}^k \mapsto \{\pm 1\}$ be a Boolean function. It holds $\sup_{\mathbf{x}\notin\partial_\gamma f} \mathbb{S}(\mathbf{x}; \sigma, f) \leq e^{-(\gamma/\sigma)^2/5+k}$. Equivalently for $\sigma = \gamma/\sqrt{5k} + 1/\sqrt{\log(1/\epsilon)}$, it holds $\sup_{\mathbf{x}\notin\partial_\gamma f} \mathbb{S}(\mathbf{x}; \sigma, f) \leq \epsilon$.*

Proof We first observe that, since $\mathbf{x} \notin \partial_\gamma f$, the sign of f will not change as long as the perturbation $\sigma\mathbf{z}$ has norm smaller than γ. Thus, we can bound $\mathbb{S}(\mathbf{x}; \sigma, f)$ above by $\mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[\sigma\|\mathbf{z}\|_2 \geq \gamma]$. By using a standard tail bound for the χ^2 distribution, see, e.g., Lemma 1 Laurent and Massart (2000), we obtain that $\mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[\sigma\|\mathbf{z}\|_2 \geq \gamma] \leq \exp(-(\gamma/\sigma)^2/5 + k)$. ■

For the sensitivity in the above bound to be less than ϵ, we need σ to be less than $\gamma/\sqrt{k} + \gamma/\sqrt{\log(1/\epsilon)}$. Recall that our smoothed learner's runtime scales exponentially in $1/\sigma$. Thus, we pay $\mathrm{poly}(k)$ in the exponent for arbitrary functions. We now prove the improved bound for intersections of halfspaces. Note for any $\epsilon > 0$, we have that for all $\sigma < \gamma/\sqrt{2\log(k/\epsilon)}$, the sensitivity at points at least γ distance from the surface is bounded by ϵ. This is a major improvement over the case of general functions and this in turn implies our quasi-polynomial time algorithm for agnostically learning intersections of halfspaces with margin(as $1/\sigma$ is now only $\mathrm{poly}(\log k)$).

Lemma 27 (Gaussian Sensitivity of Intersections of k-halfspaces with γ-margin) *Let $f : \mathbb{R}^d \mapsto \{\pm 1\}$ be an intersection of k-halfspaces and D a distribution over $\mathbb{R}^d$. It holds that*

$$\sup_{\mathbf{x}\notin\partial_\gamma f} \mathbb{S}(\mathbf{x}; \sigma, f) \leq k \exp(-\gamma^2/(2\sigma^2))\,.$$

Proof We split the proof in two cases. We first assume that the point $\mathbf{x}$ lies inside the intersection of halfspaces, i.e., $f(\mathbf{x}) = +1$. Denote by $l_i(\mathbf{x}) = \mathbf{w}_i \cdot \mathbf{x} + t_i$ the linear functions defining the intersection of halfspaces, i.e., $f(\mathbf{x}) = 1$ if and only if $l_i(\mathbf{x}) \geq 0$ for all $i = 1, \ldots, k$. The probability that $\mathbf{x} + \sigma\mathbf{z}$ is classified differently (i.e., $f(\mathbf{x} + \sigma\mathbf{z}) = -1$) is then

$$\mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x}+\sigma\mathbf{z}) \neq f(\mathbf{x})] = \mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}\left[\bigcup_{i=1}^{k}\{l_i(\mathbf{x}+\sigma\mathbf{z}) < 0\}\right] \leq \sum_{i=1}^{k} \mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}\left[l_i(\mathbf{x}) + \sigma\mathbf{z}\cdot\mathbf{w}_i < 0\right]. \quad (6)$$

For an $\mathbf{x}$ inside the intersection, it must be that its distance from all faces of the polytope is at least γ as otherwise the γ-margin assumption would be violated. Therefore, we have that for all $i = 1, \ldots, k$ it holds that $l_i(\mathbf{x}) \geq \gamma\|\mathbf{w}_i\|_2$. Therefore, we have that $\mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[l_i(\mathbf{x}) + \sigma\mathbf{z}\cdot\mathbf{w}_i < 0] \leq \mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[\|\mathbf{w}_i\|_2\gamma + \sigma\mathbf{z}\cdot\mathbf{w}_i < 0] = \mathbf{Pr}_{t\sim\mathcal{N}}[t < -\gamma/\sigma] \leq e^{-\gamma^2/(2\sigma^2)}$, where we used the tail bound for the 1-dimension normal density. Therefore, when $f(\mathbf{x}) = +1$ using Equation (6) we conclude that

$$\mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x}+\sigma\mathbf{z}) \neq f(\mathbf{x})] \leq k\exp(-\gamma^2/(2\sigma^2)).$$

We now move on to the case where $f(\mathbf{x}) = -1$. Let $C = \{\mathbf{x} : f(\mathbf{x}) = +1\}$ be the convex set corresponding to the intersection, in this case we have

$$\mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x}+\sigma\mathbf{z}) \neq f(\mathbf{x})] = \mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[\mathbf{x}+\sigma\mathbf{z} \in C] = \mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}(\mathbf{x},\sigma^2 I)}[\mathbf{z} \in C] = \int \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|\mathbf{x}-\mathbf{z}\|_2^2}{2\sigma^2}} \mathbb{1}\{\mathbf{z} \in C\} \mathrm{d}\mathbf{z}.$$

Let $\pi_C(\mathbf{x})$ be the projection of $\mathbf{x}$ (that lies outside of C) onto C. Since C is convex, for any $\mathbf{z} \in C$, we have that $\|\mathbf{x} - \pi_C(\mathbf{x})\|_2^2 + \|\pi_C(\mathbf{x}) - \mathbf{z}\|_2^2 \leq \|\mathbf{x} - \mathbf{z}\|_2^2$. Therefore, it holds

$$\begin{aligned}\int \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|\mathbf{x}-\mathbf{z}\|_2^2}{2\sigma^2}} \mathrm{d}\mathbf{z} &\leq e^{-\frac{\|\mathbf{x}-\pi_C(\mathbf{x})\|_2^2}{2\sigma^2}} \int \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|\pi_C(\mathbf{x})-\mathbf{z}\|_2^2}{2\sigma^2}} \mathrm{d}\mathbf{z} \\ &\leq e^{-\frac{\gamma^2}{2\sigma^2}} \mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}(\pi_C(\mathbf{x}),\sigma^2 I)}[\mathbf{z} \in C] \leq e^{-\frac{\gamma^2}{2\sigma^2}},\end{aligned}$$

where for the penultimate inequality we used the margin assumption which implies that $\mathbf{x}$ must lie γ-far from the boundary of C. ■

We are now ready to state and prove our main theorem about agnostic learning with geometric margin γ.

Theorem 28 (Learning intersections of Halfspaces with γ-margin) *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ where the $\mathbf{x}$-marginal is bounded in the unit ball. Let $\mathcal{F}$ be the class of intersections of k halfspaces. Then, there exists an algorithm that learns $\mathcal{F}$ in the γ-margin setting that takes $N = k^{\mathrm{poly}(\log k/\epsilon\gamma)}\log(1/\delta)$ samples, runs in time $\mathrm{poly}(d, N)$ and with probability at least $1 - \delta$, over the samples it holds that*

$$\Pr_{(\mathbf{x},y)\sim D} \leq \inf_{f\in\mathcal{F}} \mathbf{E}_{(\mathbf{x},y)\sim D)}[\sup_{\|\mathbf{u}\|_2\leq\gamma} \mathbb{1}\{f(\mathbf{x}+\mathbf{u}) \neq y\}] + \epsilon.$$

Proof Let f^* be the optimal hypothesis that minimizes the γ-margin error. From Lemma 27, we have that $\sup_{\mathbf{x}\notin\partial_\gamma f^*} \mathbb{S}(\mathbf{x}; \sigma, f) \leq \epsilon/2$ when $\sigma = \gamma/\sqrt{2\log(2k/\epsilon)}$. The result is then implied by Theorem 43 and Lemma 25. ■

Theorem 29 (Learning $\mathcal{F}(k,\Gamma)$ with γ-margin) *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ where the $\mathbf{x}$-marginal is bounded in the unit ball. Then, there exists an algorithm that learns $\mathcal{F}$ in the γ-margin setting that takes $N = k^{\mathrm{poly}(k/\epsilon\gamma)} \log(1/\delta)$ samples, runs in time $\mathrm{poly}(d, N)$ and with probability at least $1-\delta$, over the samples it holds that*

$$\Pr_{(\mathbf{x},y)\sim D} \leq \inf_{f\in\mathcal{F}(k,\Gamma)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D)} [\sup_{\|\mathbf{u}\|_2\leq\gamma} \mathbb{1}\{f(\mathbf{x}+\mathbf{u}) \neq y\}] + \epsilon\,.$$

Proof Let f^* be the function that achieves optimal γ-margin error. From Lemma 26, we have that $\sup_{\mathbf{x}\notin\partial_\gamma f^*} \mathbb{S}(\mathbf{x};\sigma,f) \leq \epsilon/2$ when $\sigma = \gamma/\sqrt{k\log(2/\epsilon}$. Our result is then implied by Lemma 25 and Theorem 43. ■

Concept Class	Runtime	Model	Source
Intersections of k halfspaces	$d \cdot k^{\tilde{O}(k/\epsilon\gamma^2)}$	Agnostic	Arriaga and Vempala (1999b)
Intersections of k halfspaces	$\mathrm{poly}(d) \cdot k^{\mathrm{poly}(\log k/(\gamma\epsilon))}$	Agnostic	This work
Arbitrary functions of k halfspaces	$\mathrm{poly}(d) \cdot k^{\mathrm{poly}(k/(\gamma\epsilon))}$	Agnostic	This work
k dimensional Convex sets	$\mathrm{poly}(d) \cdot k^{\mathrm{poly}(k/(\gamma\epsilon))}$	Agnostic	This work

Table 2: Our results on distributions with geometric margin γ

B.3. Distribution Specific Learning

In this section, we study the classic setting of agnostic learning with respect to specific distributions. Perhaps surprisingly, our smoothed learning model implies various new results in this standard model.

First, we prove the following lemma that connects the smoothed error to the true error of a function.

Lemma 30 (From Distribution Specific Agnostic to Smooth Agnostic) *Fix some boolean function $f : \mathbb{R}^d \mapsto \{\pm 1\}^d$ and some distribution D over labeled examples on $\mathbb{R}^d \times \{\pm 1\}$. It holds*

$$\mathrm{err}_\sigma(f, D) \leq \mathrm{err}(f, D) + \mathop{\mathbf{E}}_{\mathbf{x}\sim D_{\mathbf{x}}} [\mathbb{S}(\mathbf{x};\sigma,f)]\,.$$

Proof A triangle inequality implies that for any vectors $\mathbf{x}, \mathbf{z}$ and label y, we have that $\mathbb{1}\{y \neq f(\mathbf{x}+\sigma\mathbf{z})\} \leq \mathbb{1}\{f(\mathbf{x}) \neq y\} + \mathbb{1}\{f(\mathbf{x}) \neq f(\mathbf{x}+\sigma\mathbf{z})\}$. Taking expectation of $(\mathbf{x}, y)$ over D and expectation of $\mathbf{z}$ over $\mathcal{N}(0, I_d)$ implies the result. ■

Henceforth, we refer to the quantity $\mathbf{E}_{\mathbf{x}\sim D}[\mathbb{S}(\mathbf{x};\sigma,f)]$ as the *expected sensitivity* of the function f with respect to distribution D(or *expected sensitivity* of f for brevity).

B.3.1. LEARNING WITH ANTI-CONCENTRATION AND CONCENTRATION

In this section, we consider the problem of agnostic learning when the marginal has both concentration and anti concentration. In particular, we define the following notion of anti concentration.

Definition 31 (M-anti-concentrated distributions) *We say that a distribution D on $\mathbb{R}^d$ is M-anti-concentrated if for all $\|\mathbf{v}\|_2 = 1$ and continuous intervals $T \subseteq \mathbb{R}$, we have $\mathbf{Pr}[\mathbf{x} \cdot \mathbf{v} \in T] \leq \tau|T|$*

We consider M-anti-concentrated (α, λ)-strictly subexponential distributions in this section. Let $\mathcal{F}$ be the class of arbitrary functions of k halfspaces. We prove that we can bound the expected sensitivity of functions in $\mathcal{F}$ with respect to anti concentrated distributions.

Lemma 32 (Expected Gaussian Sensitivity of functions of halspaces with anticoncentration) *Let D be an M-anti concentrated distribution on $\mathbb{R}^d$ and let f be a boolean function on k halfspaces. Then, for any $\epsilon > 0$, we have that $\mathbf{E}_{\mathbf{x}\sim D}[\mathbb{S}(\mathbf{x}; \sigma, f)] \leq k(2M\epsilon + e^{-\epsilon^2/(2\sigma^2)})$*

Proof Let f be the function $f(\mathbf{x}) = g\big(\mathrm{sign}(\mathbf{w}_1 \cdot \mathbf{x} - b_1), \ldots, \mathrm{sign}(\mathbf{w}_k \cdot \mathbf{x} - b_k)\big)$ with $\|\mathbf{w}_i\|_2 = 1$ for all $i \in [k]$ and g being an arbitrary boolean function on k variables. We can assume the bound on the norms of the weights without loss of generality by renormalizing. Let S be the set of all vectors $\mathbf{y} \in \mathbb{R}^d$ such that there exists an $i \in [k]$ such that $\mathbf{w}_i \cdot \mathbf{y} \in [b_i - \epsilon, b_i + \epsilon]$. From M-anticoncentration of D and a union bound, we have that $\mathbf{Pr}_{\mathbf{x}\sim D}[\mathbf{x} \in S] \leq k(2M\epsilon)$.

We now bound $\mathbb{S}(\mathbf{x}; \sigma, f)$ for $\mathbf{x} \notin S$. By definition of S, we have that $|\mathbf{w}_i \cdot x - b_i| \geq \epsilon$. Thus, using a union bound and the tail bound for the 1-dimensional normal density, we have that

$$\mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}[f(\mathbf{x} + \sigma\mathbf{z}) \neq f(\mathbf{x})] \leq ke^{-\epsilon^2/(2\sigma^2)}$$

Thus, we have $\mathbf{E}_{\mathbf{x}\sim D}[\mathbb{S}(\mathbf{x}; \sigma, f)] \leq \mathbf{Pr}_{\mathbf{x}\sim D}[\mathbf{x} \in S] + ke^{-\epsilon^2/(2\sigma^2)}$ which completes the proof since the first term is bounded by $2kM\epsilon$. ■

We are now ready to give an algorithm to learn arbitrary boolean functions of k halfspaces under strictly subexponential distributions with anti-concentration. We use our previous results on smooth learning with respect to strictly subexponential distributions and the bound we proved above on the expected sensitivity of these functions under anti concentrated distributions.

Theorem 33 (Agnostic Learning of functions of k halfspaces under anti-concentration) *Let $k \in \mathbb{Z}_+$, $\epsilon, \delta \in (0, 1)$ and $M \in \mathbb{R}$. Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that the marginal distribution is M-anticoncentrated (α, λ)-strictly subexponential. Let $\mathcal{F}$ be the class of all functions on k halfspaces. There exists an algorithm that draws $N = d^{\mathrm{poly}((\lambda Mk/\epsilon)^{(1+1/\alpha)^3})} \log(1/\delta)$ samples, runs in time $\mathrm{poly}(d, N)$ and computes a hypothesis $h(\mathbf{x})$ such that, with probability at least $1 - \delta$, it holds*

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

Proof Let $\sigma = (\epsilon/(8\sqrt{2}Mk))\sqrt{\log(8/\epsilon)}$. The algorithm is simple. Run the algorithm from Theorem 56 to get a hypothesis h with error at most $\mathrm{opt}_\sigma + \epsilon/2$. Since $\Gamma(\mathcal{F}) \leq O(k)$, the algorithm uses $N = d^{\mathrm{poly}((\lambda Mk/\epsilon)^{(1+1/\alpha)^3})} \log(1/\delta)$ samples and runs in time $\mathrm{poly}(d, N)$. Using Lemma 32 with parameters σ and error $\epsilon/8Mk$ implies that $\mathbf{E}_{\mathbf{x}\sim D_\mathbf{x}}[\mathbb{S}(\mathbf{x}; \sigma, f)] \leq \epsilon/2$. From Lemma 30, we get that the error of the classifier is at most $\min_{f\in\mathcal{F}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq f(\mathbf{x})] + \epsilon$. ■

Concept Class	Runtime	Model	Source
Arbitrary functions on k halfspaces	$\mathrm{poly}(d) \cdot d^{1/\epsilon^2}, k = O(1)$	Agnostic	Gollakota et al. (2023)
Arbitrary functions on k halfspaces	$\mathrm{poly}(d) \cdot d^{\mathrm{poly}(k/\epsilon)}$	Agnostic	This work

Table 3: Our results on strictly sub-exponential distributions with anticoncentration

B.3.2. LEARNING UNDER SMOOTHED DISTRIBUTIONS

Finally, we consider learning distribution that have been convolved with a Gaussian. This model was studied in Kane et al. (2013). This is a natural beyond the worst-case model where the input distribution is smoothed to make learnability easier. The problem of agnostically learning functions of halfspaces under smoothed distributions was studied in Kane et al. (2013) where they obtained a runtime that was double-exponential in the number of halfspaces. We significantly improve this by reducing the runtime to only depend exponentially on the number of halfspaces.

For any distribution D on $\mathbb{R}^d$, we denote the convolution $D * \mathcal{N}(0, \tau^2 I_d)$ as D_τ. We call this a τ-smoothed distribution. We argue that the expected sensitivity of functions in $\mathcal{F}(k, \Gamma)$ with respect to τ-smoothed distributions strictly subexponential distributions is small .

Lemma 34 (Bounding Expected Sensitivity of Smoothed Distributions) *Let $f : \mathbb{R}^d \mapsto \{\pm 1\}$ be a function in $\mathcal{F}(k, \Gamma)$ and D be an (α, λ)-strictly subexponential distribution on $\mathbb{R}^d$. Then, we have that $\mathbf{E}_{\mathbf{x} \sim D_\tau}[[\mathbb{S}(\mathbf{x}; \sigma, f)] \leq O(\sigma \Gamma \sqrt{k}/\tau)$.*

Proof Let f_τ be the function defined as $f_\tau(\mathbf{u}) = f(\tau \mathbf{u})$ The quantity we want to bound is equal to $\mathbf{E}_{\mathbf{x} \sim D} \mathbf{E}_{\mathbf{z}_1, \mathbf{z}_2 \sim \mathcal{N}} [|f_\tau((1/\tau)\mathbf{x} + \mathbf{z}_1 + (\sigma/\tau)\mathbf{z}_2) - f_\tau((1/\tau)\mathbf{x} + \mathbf{z}_1)|]$. We bound the inner expectation pointwise for any $\mathbf{x}$. Since $\mathcal{F}(k, \Gamma)$ is closed under shifts and the following argument only relies on the bound on surface area, we assume that $\mathbf{x} = 0$ without loss of generality.

Since f_τ is k dimensional, we know that there exists an orthonormal matrix $\mathbf{P}$ such that $f_\tau(\mathbf{x}) = f_\tau(\mathbf{P}\mathbf{P}^T\mathbf{x})$. Let g be the function on $\mathbb{R}^k$ such that $g(\mathbf{u}) = f_\tau(\mathbf{P}\mathbf{u})$ for some orthonormal matrix $\mathbf{P}$. Clearly, $f_\tau(\mathbf{x}) = g(\mathbf{P}^T\mathbf{x})$. Using Lemma 45, we get that $\Gamma(g) \leq \Gamma$. Thus, the quantity we want to bound is $\mathbf{E}_{\mathbf{z}_1, \mathbf{z}_2 \sim \mathcal{N}}[|g(\mathbf{P}^T\mathbf{z}_1 + (\sigma/\tau)\mathbf{P}^T\mathbf{z}_2) - g(\mathbf{P}^T\mathbf{z}_1)|]$. Using Lemma 46, we can bound this term by $O\big(\Gamma \, \mathbf{E}_{\mathbf{z} \sim \mathcal{N}(0, I_k)}[\|(\sigma/\tau)\mathbf{z}\|_2]]\big) \leq O(\sigma \Gamma \sqrt{k}/\tau)$. This completes the proof. ■

We also need the following lemma that proves that a τ-smoothed strictly subexponential distribution is also strictly subexponential.

Lemma 35 *Let D be an (α, λ)-strictly subexponential distribution on $\mathbb{R}^d$ and $C > 0$ be a sufficiently large universal constant. Then, for any $\tau > 0$, we have that D_τ is $(\min(\alpha, 1), C \cdot \max(\lambda, \tau))$-strictly subexponential.*

Proof Let $s = \min(\alpha, 1)$ and let $r = \max(\lambda, \tau)$. For any unit norm vector $\mathbf{v}$ and $t > 0$, we have that $\mathbf{Pr}_{\mathbf{y} \sim D_\tau}[|\mathbf{v} \cdot \mathbf{y}| \geq t]$ is upper bounded by the maximum of the sum of $\mathbf{Pr}_{\mathbf{z} \sim \mathcal{N}}[|\mathbf{v} \cdot \tau \mathbf{z}| \geq t]$ and $\mathbf{Pr}_{\mathbf{x} \sim D}[|\mathbf{v} \cdot \mathbf{x}| \geq t]$ and 1. Let $f(t) = 2 \cdot e^{-t^2/2\tau^2}$ and let $g(t) = 2 \cdot e^{-(t/\lambda)^{1+\alpha}}$. Let $q = \max(1, 4r)$. We have that $f(t), g(t) \leq 1/4$ when $t \geq q$. Thus, we have that $\mathbf{Pr}_{\mathbf{y} \sim D_\tau}[|\mathbf{v} \cdot \mathbf{y} \geq q|] \leq 2(f(t) + g(t)) \leq 1$. Let $h(t) = 2 \cdot e^{-(t \ln 2/q)^{1+s}} \geq 2 \cdot e^{-\ln 2 (t/q)^{1+s}}$. Clearly, we have that

$h(q) \geq 1/2 \geq 2 \cdot \max(f(q), g(q))$. It is straightforward to see that h decreases slower than $2f$ and $2g$. Thus, we have that $2h(t) \geq \max(2 \cdot (f(t) + g(t)), 1)$ for all $t \geq 0$. This implies that $\mathbf{Pr}_{\mathbf{y} \sim D_\tau} \leq 2 \cdot e^{-(t \ln 2/q)^{1+s}}$ for all $t >\geq 0$. This implies that D_τ is $(\min(\alpha, 1), (4/\ln 2)\max(\lambda, \tau))$-strictly subexponential.

■

Now, we are ready to state and prove the main result of this section regarding agnostic learning of functions with bounded surface area under τ-smoothed strictly subexponential distributions.

Theorem 36 (Agnostic learning under smoothed distributions) *Let $k \in \mathbb{Z}_+$, $\epsilon, \delta \in (0, 1)$ and $\tau, \Gamma \in \mathbb{R}_+$. Let D' be an (α, λ)-strictly subexponential distribution on $\mathbb{R}^d$. Let D be distribution on $R^d \times \{\pm 1\}$ with marginal D'_τ. Then, there exists an algorithm that draws $N = d^{\mathrm{poly}\left((\lambda k \Gamma/(\tau\epsilon))^{(1+1/\alpha)^3}\right)} \log(1/\delta)$ samples, runs in time $\mathrm{poly}(d, N)$ and computes a hypothesis $h(\mathbf{x})$ such that, with probability at least $1 - \delta$, it holds*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y \neq h(\mathbf{x})] \leq \min_{f \in \mathcal{F}(k,\Gamma)} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y \neq f(\mathbf{x})] + \epsilon$$

Proof From Lemma 35, we know that D'_τ is $(\min(\alpha, 1), C\max(\lambda, \tau))$-strictly subexponential for large universal constant C. For all $f \in \mathcal{F}(k, \Gamma)$, Lemma 34 implies that $\mathbf{E}_{\mathbf{x} \sim D'_\tau}[\mathbb{S}(\mathbf{x}; \sigma, f) \leq \epsilon/2$ when $\sigma = \tau/2\Gamma\sqrt{k}$. Now, we use Lemma 30 and Theorem 56 to obtain that there exists an algorithm that draws $N = d^{\mathrm{poly}\left((k\Gamma/\tau\epsilon)^{(1+1/\alpha)^3}\right)}$, runs in time $\mathrm{poly}(d, N)$ and outputs a hypothesis h that has error at most $\min_{f \in \mathcal{F}(k,\Gamma)} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq f(\mathbf{x})] + \epsilon$. ■

Appendix C. Details of Section 3

C.1. Polynomial approximation

Lemma 37 (Approximating the Ornstein-Uhlenbeck Smoothed Concept $T_\rho f_{\mathbf{x}}(\cdot)$) *Let $C > 0$ be some large universal constant. Let D be a distribution on $\mathbb{R}^d$ with every point $\mathbf{x}$ in the support of D having $\|\mathbf{x}\|_2$ at most R. Let $f : \mathbb{R}^d \to \{\pm 1\}$ and $f_{\mathbf{x}} : \mathbb{R}^d \to \mathbb{R}$ be defined as $f_{\mathbf{x}}(\mathbf{z}) = f(\mathbf{x} + \mathbf{z})$. Then, for any $\epsilon > 0$, there exists a polynomial $p_{\mathbf{z}}$ parameterized by $\mathbf{z}$ such that:*

1. *It holds that $\mathbf{E}_{\mathbf{x}\sim D} \mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[\left|p_{\mathbf{z}}(\mathbf{x}) - T_\rho f_{\mathbf{x}}(\mathbf{z})\right|\right] \leq \epsilon$,*
2. *The degree of $p_{\mathbf{z}}$ is at most $C(R/\rho)^2 \log(1/\epsilon)$, and every coefficient of $p_{\mathbf{z}}$ is bounded in absolute value by $d^{C\left((R/\rho)^2 \log(1/\epsilon)\right)^2}$.*

Proof We observe that $T_\rho f_{\mathbf{x}}(\mathbf{z}) = \mathbf{E}_{\mathbf{s}\sim\mathcal{N}}[f(\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{s})]$ has the variable $\mathbf{x}$ inside f. Recall that our goal is to construct a polynomial in $\mathbf{x}$ and, since we have no control over f (which can possibly be very hard to approximate pointwise with a polynomial), we decouple f and $\mathbf{x}$ in the expression of $T_\rho f_{\mathbf{x}}$ by writing the function as an expectation over a Gaussian centered at $\mathbf{x}/\rho$.

$$T_\rho f_{\mathbf{x}}(\mathbf{z}) = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\left[f(\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{s})\right] = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}(\mathbf{x}/\rho,\mathbf{I})}\left[f(\sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{s})\right],$$

Next, we can recenter the expectation around zero and express the Ornstein-Uhlenbeck operator as follows:

$$T_\rho f_{\mathbf{x}}(\mathbf{z}) = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\left[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})\cdot\frac{\mathcal{N}(\mathbf{s};\mathbf{x}/\rho,\mathbf{I})}{N(\mathbf{s};\mathbf{0},\mathbf{I})}\right] = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\left[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})\cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+(\mathbf{x}/\rho)\cdot\mathbf{s}}\right].$$

To construct our polynomial, we now approximate $e^{-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+(\mathbf{x}/\rho)\cdot\mathbf{s}}$ using the 1-dimensional Taylor expansion of the exponential function $q(\mathbf{x},\mathbf{s}) = q_m\big(-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+(\mathbf{x}/\rho)\cdot\mathbf{s}\big)$ where $q_m(t) = 1+\sum_{i=1}^{m-1}\frac{t^i}{i!}$ is the degree $m-1$ Taylor approximation of e^x. Thus, our final polynomial $p_{\mathbf{z}}(\mathbf{x})$ is

$$p_{\mathbf{z}}(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\big[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})\cdot q(\mathbf{x},\mathbf{s})\big].$$

Let $\Delta(\mathbf{x})$ be defined as the error term $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[|p_{\mathbf{z}}(\mathbf{x}) - T_\rho(f_{\mathbf{x}}(\mathbf{z}))|]$. We have that $\Delta(\mathbf{x})$ is equal to

$$\begin{aligned}\Delta(\mathbf{x}) &= \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\Big[\mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\big[|f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{s})|\cdot\big|q(\mathbf{x},\mathbf{s}) - e^{-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+(\mathbf{x}/\rho)\cdot\mathbf{s}}\big|\big]\Big]\\ &\le \mathop{\mathbf{E}}_{\mathbf{s}\sim\mathcal{N}}\Big[\big|q(\mathbf{x},\mathbf{s}) - e^{-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+(\mathbf{x}/\rho)\cdot\mathbf{s}}\big|\Big],\end{aligned}$$

where for the inequality we used the fact that $|f(\mathbf{x})| = 1$ for all $\mathbf{x}$. We now observe that when $\mathbf{s}\sim\mathcal{N}$ the random variable $-\|\mathbf{x}\|_2^2/(2\rho^2)+(\mathbf{x}/\rho)\cdot\mathbf{s}$ is distributed as $\mathcal{N}(-\alpha^2/2,\alpha^2)$, where $\alpha = -\|\mathbf{x}\|_2^2/\rho^2$. Therefore, we have reduced the original polynomial approximation problem to showing that the Taylor expansion of the exponential function converges fast in L_1 to e^x with respect to $\mathcal{N}(-\alpha^2/2,\alpha^2)$.

Lemma 38 (Approximation of e^x with respect to $\mathcal{N}(-a^2/2,a^2)$) *Fix $a>0$ and sufficiently large universal constant $C>0$. Let p be the polynomial $p(x) = \sum_{i=0}^{m-1}\frac{x^i}{i!}$ with $m = Ca^2\log(1/\epsilon)$. We have that $\mathbf{E}_{x\sim\mathcal{N}(-a^2/2,a^2)}[|e^x - p(x)|]\le\epsilon$.*

Proof Let $\Delta = \mathbf{E}_{x\sim D}[|e^x - p(x)|] = \mathbf{E}_{t\sim\mathcal{N}(0,1)}[|e^{-a^2/2+at} - p(-a^2/2+at)|]$. We have that Δ can be bounded as the sum of the following two terms:

$$\Delta_1 = \mathop{\mathbf{E}}_{t\sim\mathcal{N}_1}\Big[2e^{|-\frac{a^2}{2}+at|}\cdot\mathbb{1}\{|t|>T\}\Big] \text{ and } \Delta_2 = \mathop{\mathbf{E}}_{t\sim\mathcal{N}_1}\left[e^{|-\frac{a^2}{2}+at|}\cdot\frac{\big|-\frac{a^2}{2}+at\big|^m}{(m!)}\cdot\mathbb{1}\{|t|\le T\}\right].$$

The second term's bound comes from the fact that $|p_e(x) - e^x|\le\frac{e^{|x|}}{m!}\cdot|x|^m$. The first term's bound follows from the fact that $|p_e(x) - e^x|\le 2e^{|x|}$ which is true because of the following fact whose proof can be found in Section C.1.

Lemma 39 *For any $m\in\mathbb{N}$, let $p_m:\mathbb{R}\to\mathbb{R}$ be defined as the degree m Taylor expansion of e^x. That is, $p_m(x) = 1+\sum_{i=1}^m\frac{x^i}{i!}$. Then we have $|p_m(x)|\le e^{|x|}$ for all $x\in\mathbb{R}$.*

We first bound Δ_1. We have that

$$\begin{aligned}\Delta_1 &\le 2\sqrt{\mathop{\mathbf{E}}_{t\sim\mathcal{N}_1}[e^{|-a^2+2at|}]\Pr_{t\sim\mathcal{N}_1}[|t|>T]} \le 4\sqrt{\mathop{\mathbf{E}}_{t\sim\mathcal{N}_1}[e^{-a^2+2at}]\Pr_{t\sim\mathcal{N}_1}[|t|>T]}\\ &\le 4\sqrt{e^{Ca^2-T^2/2}}\le\epsilon/2,\end{aligned}$$

when $T = C'a\log(1/\epsilon)$ for large constant C'. The first inequality follows by applying Cauchy-Schwartz. The second follows from the symmetry of the Gaussian distibution. We get the third inequality by using the following fact (see Section C.1 for the proof) and bounds on the tail of the Gaussian distribution.

Lemma 40 *There exists a large enough universal constant C such that for every $u \in \mathbb{R}$ it holds that* $\mathbf{E}_{x\sim\mathcal{N}_1}\left[\left(\frac{\mathcal{N}(x;u,1)}{\mathcal{N}(x;0,1)}\right)^2\right] \leq e^{Cu^2}$.

We now bound Δ_2. We have that $\left|-\frac{a^2}{2}+at\right|$ is atmost $\frac{a^2}{2}+aT$ when $|t|\leq T$. Thus, we obtain that

$$\Delta_2 \leq e^{\frac{a^2}{2}+aT}\frac{\left(\frac{a^2}{2}+aT\right)^m}{(m!)} \leq e^{Ca^2}\left(\frac{Ca^2\log(1/\epsilon)}{m}\right)^m \leq \epsilon/2\,,$$

when $m = C''a^2\log(1/\epsilon)$ for large constant C''. Thus, the final error of our polynomial is at most ϵ. ∎

We set $a = \|\mathbf{x}/\rho\|_2$ in the above claim. Thus, the degree m of our taylor expansion is bounded by $O((R/\rho)^2\log(1/\epsilon))$. Thus, we obtain that the final error of our polynomial is

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim D}\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_k}\left[|(T_\rho f_{\mathbf{x}}(\mathbf{z}) - p_{\mathbf{z}}(\mathbf{x}))|\right] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[\Delta_1(\mathbf{x})\right] \leq \epsilon.$$

The degree of our polynomial $p_{\mathbf{z}}(\mathbf{u})$ is $O((R/\rho)^2\log(1/\epsilon))$. We now bound the coefficients of $p_{\mathbf{z}}(\mathbf{x})$. Since f is bounded by 1, it suffices to bound the coefficients of $\mathbf{x}$ in the polynomial $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\left[q(\mathbf{x},\mathbf{s})\right]$ Since $q(\mathbf{x},\mathbf{s}) = p_e\left(-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+\langle(\mathbf{x}/\rho),\mathbf{s}\rangle\right)$, it is a composition of a univariate polynomial p_e and a multivariate polynomial $r(\mathbf{x},\mathbf{s}) = -\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+\langle(\mathbf{x}/\rho),\mathbf{s}\rangle$. We use the following fact about the coefficients of the composition of polynomials. The proof can be found in Section C.1.

Lemma 41 *Let $p_1(x)$ be a polynomial on $\mathbb{R}$ having degree ℓ_1 and coefficients bounded by t_1. Let $p_2(\mathbf{x})$ be a polynomial on $\mathbb{R}^d$ of degree ℓ_2 with coefficients bounded by t_2. The polynomial $q(\mathbf{x}) = p_1(p_2(\mathbf{x}))$ has coefficients bounded by $t_1t_2^{\ell_1}\cdot(Cd)^{2\ell_1\ell_2}$ for large constant C.*

Thus for a fixed $\mathbf{x}$, we have obtain that the coefficients of $q(\mathbf{u},\mathbf{x})$ are bounded by $\|\mathbf{x}\|_\infty^m\cdot(C_1k)^m$ for large enough constant C_1. Thus, the coefficients of $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\left[q(\mathbf{u},\mathbf{x})\right]$ are bounded by $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\left[\|\mathbf{x}\|_\infty^m\cdot(C_1k)^m\right]$ which is bounded above by $(C_2d)^{m^2}$ for large enough C_2. This follows from the fact that $\|\mathbf{x}\|_\infty^m \leq \sum_{i=1}^k|\mathbf{x}_i|^m$ and the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\left[|\mathbf{x}_i|^m\right]$ is atmost $(C_3m)^m$ for some constant C_3. Thus, we finally conclude that the coefficients of $p_{\mathbf{z}}(\mathbf{u})$ are bounded above by $d^{C\left((R/\rho)^2\log(1/\epsilon)\right)^2}$ for some large constant C. ∎

Proof [Proof of Lemma 39] For any $x\in\mathbb{R}$, $|p_m(x)| \leq p_m(|x|)$. We prove the claim by induction on m. Clearly, $p_0(x) = 1 \leq e^{|x|}$. Now, by induction, we have $|\frac{\mathrm{d}}{\mathrm{d}x}p_m(x)| = |p_{m-1}(x)| \leq e^{|x|} \leq \frac{\mathrm{d}}{\mathrm{d}x}e^x$ where the last inequality holds for all $x > 0$. Also, we have $p_m(0) = e^0$. Thus, we get that $|p_m(x)| \leq e^{|x|}$. ∎

Proof [Proof of Lemma 40] The proof is below is a obtained by completing the squares.

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}_1}\left[\left(\frac{\mathcal{N}(x;u,1)}{\mathcal{N}(x;0,1)}\right)^2\right] = \int_{z\in\mathbb{R}} e^{-\frac{1}{2}\left(2u^2+z^2-4uz\right)}dz$$
$$= \int_{z\in\mathbb{R}} e^{u^2}\cdot e^{-\frac{1}{2}(2u-z)^2}dz \le e^{Cu^2}\,.$$

■

Lemma 42 *Let $p(\mathbf{u})$ be a polynomial on $\mathbb{R}^k$ having degree ℓ and coefficients bounded by t. Then $q(\mathbf{u}) = (p(\mathbf{u}))^m$ has coefficients bounded by $t^m \cdot (Ck)^{\ell m}$ for large constant C.*

Proof The number of monomials in a polynomial of degree ℓ on $\mathbb{R}^k$ is at most $\sum_{i=0}^{\ell}\left(\!\binom{k}{i}\!\right) = \binom{k+\ell}{k} \le (e\cdot(k+1))^{\ell}$. Expanding $q(\mathbf{x}) = (p(\mathbf{x}))^m$ as a sum, we get $(Ck)^{\ell m}$ product terms for large constant C. Each of the terms in this sum have coeffients bounded by t^m. Since every monomial is formed as a sum of a subset of these terms, we get that the coefficients of each monomial in q is bounded by $t^m \cdot (Ck)^{\ell m}$ ■

Proof [Proof of Lemma 41] Let $p_1(x) = \sum_{i=0}^{\ell_1} c_i x^i$. Thus, $q(\mathbf{x}) = \sum_{i=0}^{\ell_1} c_i \left(p_2(\mathbf{x})\right)^i$. For any monomial of q, the contribution to the coefficient from each term in the previous sum is atmost $t_1 \cdot t_2^{\ell_1} \cdot (Ck)^{\ell_1\ell_2}$ for large constant C from Lemma 42. Thus, the coefficients are bounded by $\ell_1 t_1 \cdot t_2^{\ell_1} \cdot (Ck)^{\ell_1\ell_2} \le t_1 t_2^{\ell_1} \cdot (Ck)^{2\ell_1\ell_2}$ ■

C.2. Random Projection and Polynomial Regression

We are now ready to implement the second and third steps in our plan. The algorithmic idea is simple: reduce dimension by applying a random projection and then run polynomial regression in this low dimension space.

Algorithm 1 Agnostic Learner for Smooth Boolean Concepts with Random Projection

Input: Labeled Dataset $S = \{(\mathbf{x}_i, y_i)\}_{i\in[N]}$, degree ℓ, subspace dimension m
Output: Hypothesis h

Sample Random Matrix $\mathbf{R} \in \mathbb{R}^{m\times d}$ with each entry sampled from $\frac{\mathcal{N}(0,1)}{\sqrt{m}}$
Find polynomial p of degree at most ℓ such that p minimizes $\frac{1}{N}\sum_{i=1}^{N}|p(\mathbf{R}\mathbf{x}_i) - y_i|$
Choose $t \in [-1,1]$ such that $\sum_{i=1}^{N} \mathbb{1}[\mathrm{sign}\big(p(\mathbf{R}\mathbf{x}_i) - t\big) \neq y_i]$ is minimized.
Output hypothesis h, such that $h(\mathbf{x}) = \mathrm{sign}\big(p(\mathbf{R}\mathbf{x}) - t\big)$

The above algorithm is run with degree and subspace dimension chosen according to the error we target. We boost the success probability using a standard technique of repeating the algorithm multiple times and choosing the hypothesis that performs best on an independently chosen validation set. Our choice of the subspace dimension m depends on the intrinsic dimension of our function f. The degree ℓ we choose depends on the degree bound we obtain from Proposition 9.

For projection matrix $\mathbf{P} \in \mathbb{R}^{d\times k}$, let $f(\mathbf{x}) = f(\mathbf{P}\mathbf{P}^T\mathbf{x}) = g(\mathbf{P}^T\mathbf{x})$ for all $\mathbf{x}$. We argue that with good probability over the random projection matrix $\mathbf{R}$, we have that $\|\mathbf{P}^T\mathbf{x} - (\mathbf{P}^T\mathbf{R}^T\mathbf{R}\mathbf{x})\|_2$ is bounded(Lemma 44) for all sample points $(\mathbf{x}, y)$ in our input data set. Using this and the bounds on surface area, we argue that $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}} |f(\mathbf{x}+\sigma\mathbf{z}) - f(\mathbf{R}^T\mathbf{R}\mathbf{x}+\sigma\mathbf{z})|$ is bounded for all $\mathbf{x}$ in the data set. Thus, we can treat $\mathbf{R}\mathbf{x}$ as our inputs and the work in the m-dimensional space. From here, the proof is straightforward and uses ideas similar to Kalai et al. (2005). We now describe in full detail the proof of our main theorem.

Theorem 43 *Let $k \in \mathbb{Z}_+$ and $\epsilon, \delta, \sigma \in (0,1)$. Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is bounded in the unit ball. There exists an algorithm that draws $N = k^{\tilde{O}\left((\Gamma/\epsilon)^4(1/\sigma^2)\right)} \log(1/\delta)$ samples, runs in time $\mathrm{poly}(d, N)$, and computes a hypothesis $h(\mathbf{x})$ such that, with probability at least $1-\delta$, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq h(\mathbf{x})] \leq \mathrm{opt}_\sigma + \epsilon$.*

Proof The algorithm is as follows: run Algorithm 1 $O\big(\log(1/\delta)/\epsilon\big)$ times with the same random matrix $\mathbf{R}$ and fresh samples each time with parameters N, ℓ, k which will be chosen later. Output the hypothesis that has the lowest error on a validation set of size $O\big(\log(1/\delta)/\epsilon^2\big)$. Clearly the run time is dominated by the time required for polynomial regression which is at most $\mathrm{poly}(d, N)$.

We now argue the correctness of the algorithm. Let $f^* \in \mathcal{F}(k, \Gamma)$ be the optimal function that achieves opt_σ. There exists some orthonormal matrix $\mathbf{P} \in \mathbb{R}^{d\times k}$ such that $f^*(\mathbf{u}) = f^*(\mathbf{P}\mathbf{P}^T\mathbf{u})$ for all $\mathbf{u} \in \mathbb{R}^d$. We say that matrix $\mathbf{R}$ is (α)-*good* for a set S if $\|\mathbf{P}^T\mathbf{x} - \mathbf{P}^T\mathbf{R}^T\mathbf{R}\mathbf{x}\|_2 \leq \alpha$ for all $\mathbf{x} \in S$. We crucially use the following claim which we prove in the end of this section.

Lemma 44 *Let $S \subseteq \mathbb{R}^d$ with $\|\mathbf{x}\|_2 \leq B$ for all $\mathbf{x} \in S$ and $\mathbf{W} \in \mathbb{R}^{k\times d}$ with $\|\mathbf{W}\|_2 \leq \lambda$. Sample a $m\times d$ random matrix $\mathbf{R}$ with every entry sampled from $\frac{\mathcal{N}(0,1)}{\sqrt{m}}$ where $m = O\big((B\lambda)^2 k \log(|S|/\delta)/\epsilon^2\big)$. Then, with probability $1-\delta$, we have that $\forall \mathbf{x} \in S$, $\|\mathbf{W}\mathbf{x} - (\mathbf{W}\mathbf{R}^T\mathbf{R}\mathbf{x})\|_2 \leq \epsilon$*

We choose the dimension m in such a way that Lemma 44 implies with probability at least $1 - (\delta\epsilon/16)$ that the random gaussian matrix $\mathbf{R}$ is $(\epsilon/(32\Gamma))$-*good* for a given set S of size N. Having chosen m in this way, it is easy to see that with probability at least $1-(\delta/2)$, the random matrix $\mathbf{R}$ is $(\epsilon/(32\Gamma))$-*good* for at least $1-(\epsilon/8)$ mass of the datasets S drawn from $D^{\otimes N}$. Henceforth, we assume that $\mathbf{R}$ satisfies the above property. From here, our analysis is similar to the proof of Theorem 5 of Kalai et al. (2005) accompanied with applications of Lemma 44 and Proposition 9.

Consider the sample dataset $S = \{(\mathbf{x}_i, y_i)\}_{i\in[N]}$ of size N in a single run of Algorithm 1. Let p_S be the polynomial chosen by the algorithm and let h_S be the corresponding hypothesis that the algorithm outputs. From the proof of Theorem 5 of Kalai et al. (2005), we have that $\frac{1}{N}\sum_{i=1}^N \mathbb{1}[h_S(\mathbf{x}_i) \neq y_i] \leq \min\big(\frac{1}{2N}\sum_{i=1}^N |p_S(\mathbf{R}\mathbf{x}_i) - y_i|, 1\big)$. We now bound $\mathbf{E}_{S\sim D^{\otimes N}}\left[\mathbb{1}[h_S(\mathbf{x}_i) \neq y_i]\right]$ by bounding the expectation of the right hand side. We have that

$$\min\left(\frac{1}{2N}\sum_{i=1}^N |p_S(\mathbf{R}\mathbf{x}_i) - y_i|, 1\right) \leq \min\left(\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\left[\frac{1}{2N}\sum_{i=1}^N |p_{\mathbf{z}}(\mathbf{R}\mathbf{x}_i) - y_i|\right], 1\right)$$
$$\leq \underbrace{\min\left(\mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\frac{1}{2N}\sum_{i=1}^N |p_{\mathbf{z}}(\mathbf{R}\mathbf{x}_i) - f^*(\mathbf{x}_i+\sigma\mathbf{z})|\right], 1\right)}_{\Delta_1(S)} + \underbrace{\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\left[\frac{1}{2N}\sum_{i=1}^N |f^*(\mathbf{x}_i+\sigma\mathbf{z}) - y_i|\right]}_{\Delta_2(S)}$$

The first inequality follows from the fact that p_S is the minimizer of the error and thus beats any polynomial $p_{\mathbf{z}}$ which we choose later. The second is a triangle inequality. We bound the two terms separately. For any function f, let f_σ be the function defined as $f_\sigma(\mathbf{x}) = f(\sigma\mathbf{x})$. Let g^* be the function on $\mathbb{R}^k$ defined as $g^*(\mathbf{u}) = f^*(\mathbf{P}\mathbf{u})$. Recall that $f^*(\mathbf{x}) = f^*(\mathbf{P}\mathbf{P}^T\mathbf{x}) = g^*(\mathbf{P}^T\mathbf{x})$. We use the following claim to bound the surface area of g^* by Γ. The proof is available in the end of this section.

Lemma 45 *Let $f : \mathbb{R}^d \to \{\pm 1\}$ be a function such that $f(\mathbf{x}) = f(\mathbf{P}\mathbf{P}^T\mathbf{x})$ for some orthonormal matrix $\mathbf{P} \in \mathbb{R}^{d\times k}$. Define $g : \mathbb{R}^k \to \{\pm 1\}$ to be the function $g(\mathbf{y}) = f(\mathbf{P}\mathbf{x})$. Then, we have that $\Gamma(g) = \Gamma(f)$.*

First consider $\Delta_1(S)$. We define the terms

$$\Delta_{11}(S) = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\frac{1}{2N}\sum_{i=1}^{N}|p_{\mathbf{z}}(\mathbf{R}\mathbf{x}_i) - f^*(\mathbf{R}^T\mathbf{R}\mathbf{x}_i + \sigma\mathbf{z})|\right]$$

and

$$\Delta_{12}(S) = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\frac{1}{2N}\sum_{i=1}^{N}|f^*(\mathbf{R}^T\mathbf{R}\mathbf{x}_i + \sigma\mathbf{z}) - f^*(\mathbf{x}_i + \sigma\mathbf{z})|\right]$$

to help us bound $\Delta_1(S)$.

Observe that $\Delta_{11}(S) = \mathbf{E}_{z\sim\mathcal{N}}\left[\frac{1}{2N}\sum_{i=1}^{N}|p_{\mathbf{z}}(\mathbf{R}\mathbf{x}_i) - g^*(\mathbf{P}^T\mathbf{R}^T\mathbf{R}\mathbf{x}_i + \sigma\mathbf{P}^T\mathbf{z})|\right]$. Conditioning on the event that $\mathbf{R}$ is $(\epsilon/(32\Gamma))$-*good* for S, we have that $\|\mathbf{P}^T\mathbf{R}^T\mathbf{R}\mathbf{x}_i\|_2 \le \|\mathbf{P}^T\mathbf{x}_i\|_2 + \epsilon$ for all $i \in [N]$. Recall that $g^*(\mathbf{y} + \sigma\mathbf{z}) = g^*_\sigma(\mathbf{y}/\sigma + \mathbf{z})$. Using the fact that $\mathcal{F}(k,\Gamma)$ is closed under scaling, we have that $\Gamma(g^*_\sigma) \le \Gamma$. Thus, using Proposition 9, we obtain a polynomial $q_{\mathbf{z}}$ of degree $\ell = O\big((\Gamma/\epsilon)^4(1/\sigma^2)\log(1/\epsilon)\big)$ such that $\mathbf{E}_{\mathbf{x}\sim D_S}\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\big[|g^*_\sigma\big((\mathbf{P}^T\mathbf{R}^T\mathbf{R}/\sigma)\mathbf{x} + \mathbf{P}^T\mathbf{z}\big) - q_{\mathbf{z}}(\mathbf{P}^T\mathbf{R}^T\mathbf{R}\mathbf{x}/\sigma)|\big] \le \epsilon/4$ where D_S is the uniform distribution over the samples S. Let $p_{\mathbf{z}}$ be the polynomial $p_{\mathbf{z}}(\mathbf{u}) = q_{\mathbf{z}}(\mathbf{P}^T\mathbf{R}^T\mathbf{u}/\sigma)$. Clearly, $p_{\mathbf{z}}(\mathbf{R}\mathbf{x}) = q_{\mathbf{z}}(\mathbf{P}^T\mathbf{R}^T\mathbf{R}\mathbf{x}/\sigma)$. Thus, we obtain that $\Delta_{11}(S) \le \epsilon/8$.

Observe that $\Delta_{12}(S)$ is equal to $\mathbf{E}_{z\sim\mathcal{N}}\left[\frac{1}{2N}\sum_{i=1}^{N}|g^*(\mathbf{P}^T\mathbf{R}^T\mathbf{R}\mathbf{x}_i + \sigma\mathbf{z}) - g^*(\mathbf{P}^T\mathbf{x}_i + \sigma\mathbf{z})|\right]$. We use the following lemma(proof in the end of this section) along with the fact that $\mathbf{R}$ is $(\epsilon/(32\Gamma))$-*good* for S to obtain that $\Delta_{12}(S) \le \epsilon/4$.

Lemma 46 *Let $\mathcal{F}$ be a class of binary functions with sufficiently smooth decision boundaries that is close under arbitrary translations, and whose elements have Gaussian Surface area bounded by Γ. Then, for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, we have $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\big[|f(\mathbf{u}+\mathbf{z}) - f(\mathbf{v}+\mathbf{z})|\big] \le 8\cdot\Gamma\cdot\|\mathbf{u}-\mathbf{v}\|_2$.*

From a triangle, inequality, it follows that $\Delta_1(S) \le \min\big(\Delta_{11}(S) + \Delta_{12}(S), 1\big)$. Thus, we get that $\Delta_1(S) \le (3\epsilon/8)$ when $\mathbf{R}$ is $(\epsilon/(32\Gamma))$-*good* for S and at most 1 otherwise. Also recall that the second event happens with probability at most $\epsilon/8$ over S. Thus, we have that $E_{S\sim D^{\otimes N}}[\Delta_1(S)] \le 3\epsilon/8 + \epsilon/8 \le \epsilon/2$. Thus, we have that

$$\mathop{\mathbf{E}}_{S\sim D^{\otimes N}}\left[\min\left(\frac{1}{2N}\sum_{i=1}^{N}|p_S(\mathbf{R}\mathbf{x}_i) - y_i)|, 1\right)\right]$$
$$\le \mathop{\mathbf{E}}_{S\sim D^{\otimes N}}[\Delta_1(S) + \Delta_2(S)] \le \epsilon/2 + \mathop{\mathbf{E}}_{S\sim D^{\otimes N}}[\Delta_2(S)] \le \mathrm{opt}_\sigma + \epsilon/2$$

The final inequality follows from the definition of opt_σ. Thus, we have that

$$\mathop{\mathbf{E}}_{S\sim D^{\otimes N}}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}[h_S(\mathbf{x}_i)\neq y_i]\right]\leq \mathrm{opt}_\sigma+\epsilon/2.$$

Since our hypothesis h_S is a PTF of degree ℓ on m variables, VC theory tells us that for $N=\mathrm{poly}(m^\ell/\epsilon)$, we have that

$$\mathop{\mathbf{E}}_{S\sim D^{\otimes N}}\left[\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y\neq h_S(\mathbf{x})]\right]\leq \mathrm{opt}_\sigma+3\epsilon/4.$$

By Markov's inequality, we have that with probaility atleast $\epsilon/16$ over samples S, we have that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y\neq h_S(\mathbf{x})]\leq \mathrm{opt}_\sigma+7\epsilon/8$. Let $h_1,h_2,\ldots,h_r$ be the r hypotheses outputted on the $r=O(\log(1/\delta)/\epsilon)$ repetitions of algorithm 2. With probability at least $1-\delta/4$, there exists $i\in[r]$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y\neq h_i(\mathbf{x})]\leq \mathrm{opt}_\sigma+7\epsilon/8$. Thus, using the validation set of size $O(\log(1/\delta)/\epsilon^2)$, with probability at least $1-\delta/4$, we choose a hypothesis h such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y\neq h(\mathbf{x})]\leq \mathrm{opt}_\sigma+\epsilon$. Thus with total error probability of at most δ, our algorithm outputs a hypothesis h such that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y\neq h(\mathbf{x})]\leq \mathrm{opt}_\sigma+\epsilon$$

We now calculate the required parameters. The degree ℓ is $O\big((\Gamma/\epsilon)^4(1/\sigma^2)\log(1/\epsilon)\big)$. The dimension m is $O\big(\frac{k\Gamma^2\log N}{(\epsilon\sigma)^2}\log(k/(\delta\epsilon))\big)$ and the number of samples $N=\mathrm{poly}(m^\ell/\epsilon)$. We get that these conditions are satisfied when $N=\mathrm{poly}\left(\left(k\Gamma^2/(\sigma\epsilon)^2\right)^\ell\right)$ and $m\geq\mathrm{poly}\left(\frac{k\Gamma^2\log(k/(\delta\epsilon))}{(\epsilon\sigma)^2}\right)$. ■

Lemma 47 (Arriaga and Vempala (1999b)) *Let $S\subseteq\mathbb{R}^d$ with $\|\mathbf{x}\|_2\leq 1$ for all $\mathbf{x}\in S$ and $\mathbf{w}\in\mathbb{R}^d$ with $\|\mathbf{w}\|_2\leq 1$. Sample a $m\times d$ random matrix $\mathbf{R}$ with every entry sampled from $\frac{\mathcal{N}(0,1)}{\sqrt{m}}$ where $m=O\big(\log(|S|/\delta)/\epsilon^2\big)$. Then, with probability $1-\delta$, we have that $\forall\mathbf{x}\in S$, $|\langle\mathbf{w},\mathbf{x}\rangle-\langle\mathbf{Rw},\mathbf{Rx}\rangle|\leq\epsilon$*

Proof [Proof of Lemma 44] Using Lemma 47 on set S with $\epsilon'=\epsilon/(\sqrt{k}\lambda B)$ and $\delta'=\delta/k$ we get that with probability at least $1-\delta/k$,

$$|\langle\mathbf{W}_i,\mathbf{x}\rangle-\langle\mathbf{RW}_i,\mathbf{Rx}\rangle|\leq\|\mathbf{W}_i\|_2\|\mathbf{x}\|_2\epsilon'\leq\epsilon/\sqrt{k}$$

for fixed $i\in[k]$. The first inequality follows from the definition of norm and the second inequality follows from the fact that $\|\mathbf{W}_i\|_2\leq\|\mathbf{W}\|_2$ for all i. Now, applying a union bound gives us that $|\langle\mathbf{W}_i,\mathbf{x}\rangle-\langle\mathbf{RW}_i,\mathbf{Rx}\rangle|\leq\epsilon/\sqrt{k}$ for all $i\in[k]$. Thus, with probability at least $1-\delta$, we have

$$\|\mathbf{Wx}-\mathbf{WR}^T\mathbf{Rx}\|_2\leq\sqrt{k}\left\|\mathbf{Wx}-\mathbf{WR}^T\mathbf{Rx}\right\|_\infty\leq\epsilon$$

■

Proof [Proof of Lemma 45] Since the Gaussian density is spherically symmetric, we have that the Gaussian surface area is spherically symmetric under rotations. Thus, we can assume that $\mathbf{P}^T=$

$\begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix}$ where $\mathbf{I}$ is the $k \times k$ identity matrix. We have that $f(\mathbf{x}) = f(\mathbf{x}^k)$ where $\mathbf{x}^k_i = \mathbf{x}_i$ for $i \leq k$ and 0 otherwise. Let $A_f = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = 1\}$. Similarly, define $A_g = \{\mathbf{y} \in \mathbb{R}^k : g(\mathbf{x}) = 1\}$. It is easy to see that $A_f = A_g \times \mathbb{R}^{d-k}$. We are now ready to prove the lemma. For any set S, let S^δ denote the set of points at distance at most δ from S. Then, we have that

$$\Gamma(f) = \liminf_{\delta \to 0} \frac{1}{\delta} \mathop{\mathbf{Pr}}_{\mathbf{z} \sim \mathcal{N}(0, I_d)} \left[\mathbf{z} \in A_g^\delta \times \mathbb{R}^{d-k} \setminus A_g \times \mathbb{R}^{d-k}\right] = \liminf_{\delta \to 0} \frac{1}{\delta} \mathop{\mathbf{Pr}}_{\mathbf{z} \sim \mathcal{N}(0, I_k)} \left[\mathbf{z} \in A_g^\delta \setminus A_g\right] = \Gamma(g)$$

■

Proof [Proof of Lemma 46] Let $g(t) = \mathbf{E}_{\mathbf{z} \sim \mathcal{N}}[|f(\mathbf{u}+\mathbf{z}) - f(\mathbf{u} + t \cdot \frac{\mathbf{v}-\mathbf{u}}{\|\mathbf{v}-\mathbf{u}\|_2} + \mathbf{z})|]$. We first observe that, if g is differentiable, then $\int_{t=0}^{\|\mathbf{u}-\mathbf{v}\|_2} g'(t)\, dt = \mathbf{E}_{\mathbf{z} \sim \mathcal{N}}[|f(\mathbf{u}+\mathbf{z}) - f(\mathbf{v}+\mathbf{z})|]$. Therefore, it suffices to show that g' is differentiable and to bound the quantity $g'(t)$ uniformly over $t \in [0, \|\mathbf{u}-\mathbf{v}\|_2]$ by $O(\Gamma)$.

Let $A_f = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = 1\}$ and $A_f(\mathbf{u})$ such that $\mathbf{x} \in A_f(\mathbf{u})$ iff $\mathbf{x} + \mathbf{u} \in A_f$. Recall that we have $f : \mathbb{R}^d \to \{\pm 1\}$ and therefore we may express $g(t)$ as follows, where $\mathbf{w} = \frac{\mathbf{v}-\mathbf{u}}{\|\mathbf{v}-\mathbf{u}\|_2}$.

$$g(t) = 2 \mathop{\mathbf{Pr}}_{\mathbf{z} \sim \mathcal{N}}[\mathbf{z} \in A_f(\mathbf{u}) \triangle A_f(\mathbf{u} + t \cdot \mathbf{w})]$$

where $\triangle$ is the symmetric difference. Let $B(t) = A_f(\mathbf{u}) \triangle A_f(\mathbf{u} + t \cdot \mathbf{w})$. Then, we have the following

$$\begin{aligned} g'(t) &= \lim_{\delta \to 0} \frac{g(t+\delta) - g(t)}{\delta} \\ &= \lim_{\delta \to 0} \frac{2}{\delta} \cdot \left(\mathop{\mathbf{Pr}}_{\mathbf{z} \sim \mathcal{N}}[\mathbf{z} \in B(t+\delta) \setminus B(t)] - \mathop{\mathbf{Pr}}_{\mathbf{z} \sim \mathcal{N}}[\mathbf{z} \in B(t) \setminus B(t+\delta)] \right). \end{aligned}$$

In order to bound $|g'(t)|$ (which is an upper bound for $g'(t)$), we bound the quantity corresponding to the first term in the limit $\lim_{\delta \to 0} \frac{2}{\delta} \cdot \mathbf{Pr}_{\mathbf{z} \sim \mathcal{N}}[\mathbf{z} \in B(t+\delta) \setminus B(t)]$ (and similarly the one corresponding to the other term). In particular, we have $B(t+\delta) \setminus B(t) \subseteq A_f(\mathbf{u} + t \cdot \mathbf{w}) \triangle A_f(\mathbf{u} + (t+\delta) \cdot \mathbf{w})$, which implies that $|g'(t)| \leq 4 \lim_{\delta \to 0} \frac{1}{\delta} \mathbf{Pr}_{\mathbf{z} \sim \mathcal{N}}[\mathbf{z} \in A_f(\mathbf{u} + t \cdot \mathbf{w}) \triangle A_f(\mathbf{u} + (t+\delta) \cdot \mathbf{w})]$. Denote with A_f^δ the set containing all the points with distance at most δ from the boundary of A_f. Denote with $f_{\mathbf{u}}$ the function with $f_{\mathbf{u}}(\mathbf{x}) = f(\mathbf{u}+\mathbf{x})$. Since $\mathcal{F}$ is closed under translations and only contains functions with sufficiently smooth boundaries (see, e.g., (Kane, 2010, Proposition A.3)), we have that

$$\Gamma \geq \Gamma(f_{\mathbf{u}+t\mathbf{w}}) = \lim_{\delta \to 0} \frac{\mathbf{Pr}_{\mathbf{z} \sim \mathcal{N}}[\mathbf{z} \in A^\delta_{f_{\mathbf{u}+t\mathbf{w}}}]}{2\delta}$$

Observe now that $A_f(\mathbf{u} + t \cdot \mathbf{w}) \triangle A_f(\mathbf{u} + (t+\delta) \cdot \mathbf{w}) \subseteq A^\delta_{f_{\mathbf{u}+t\mathbf{w}}}$, which therefore implies the desired bound. ■

Lemma 48 *Let Q_1 be a univariate distribution such that $E_{x \sim Q_1}[|x|^p] \leq (C \cdot p)^p$ where C is a constant associated with the distribution and $p \leq t$. Then,*

$$\mathop{\mathbf{E}}_{x \sim Q_1}[h_t^2(x+u)] \leq t^{5t} \cdot C^{2t} \max(1, u^{2t})$$

Proof We have that

$$h_t(x) = \sqrt{t!} \sum_{m=0}^{\lfloor \frac{t}{2} \rfloor} \frac{(-1)^m}{m!(t-2m)!} \frac{x^{n-2m}}{2^m}.$$

We have

$$\mathop{\mathbf{E}}_{x \sim Q_1} (x+u)^d = \mathop{\mathbf{E}}_{x \sim Q_1} \left[\sum_{i=0}^{d} \binom{d}{i} x^i u^{d-i} \right] \leq 2^d \max_{i=0}^{n} \mathop{\mathbf{E}}_{\mathbf{x} \sim Q_1} \left[|x|^i u^{d-i} \right] \leq 2^d (Cd)^d \max(1, u^d) .$$

This implies that

$$\begin{aligned}
\mathop{\mathbf{E}}_{x \sim Q_1} [h_t^2(x+u)] &\leq \mathop{\mathbf{E}}_{x \sim Q_1} \left[\left(t^{2t} \cdot \sum_{i=0}^{\lfloor \frac{t}{2} \rfloor} \sum_{j=0}^{\lfloor \frac{t}{2} \rfloor} (x+u)^{2t-2i-2j} \right) \right] \\
&\leq t^{2t} \cdot t^2 2^{2t} (2Ct)^{2t} \max(1, u^{2t}) \\
&\leq t^{5t} \cdot C^{2t} \max(1, u^{2t})
\end{aligned}$$

■

Lemma 49 *Let Q be a multivariate product distribution on $\mathbb{R}^k$ such that for any $i \in [k]$ and any $p \in \mathbb{N}$ we have $\mathbf{E}_{\mathbf{w} \sim Q}[|\mathbf{w}_i|^p] \leq (C \cdot p)^p$ for some constant $C > 0$. Let $S \in \mathbb{N}^k$ be a multi index. Then*

$$\mathop{\mathbf{E}}_{\mathbf{w} \sim Q} [H_S^2(\mathbf{w} - \mathbf{u})] \leq C^{|S| \log |S|} \max\{1, \|\mathbf{u}\|_\infty^{2|S|}\}$$

Proof

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{w} \sim Q} [H_S^2(\mathbf{w} - \mathbf{u})] &= \mathop{\mathbf{E}}_{\mathbf{w} \sim Q} \left[\prod_{i \in [k]} h_{S_i}^2(\mathbf{w}_i - \mathbf{u}_i) \right] \\
&= \prod_{i \in [k]} \mathop{\mathbf{E}}_{\mathbf{w} \sim Q} \left[h_{S_i}^2(\mathbf{w}_i - \mathbf{u}_i) \right] \\
&\leq \prod_{i \in [k]} S_i^{5S_i} C^{2S_i} \max\{1, \mathbf{u}_i^{2S_i}\} && \text{(Lemma 48)} \\
&= \exp\left(\sum_{i \in [k]} 5S_i \log(S_i) \right) C^{2\sum_{i \in [k]} \alpha_i} \prod_{i \in [k]} \max\{1, \mathbf{u}_i^{2S_i}\} \\
&\leq C^{|S| \log |S|} \max\{1, \|\mathbf{u}\|_\infty^{2|S|}\} \\
&\qquad (\textstyle\sum_{i \in [k]} S_i = |S|,\ \mathbf{u}_i \leq \|\mathbf{u}\|_\infty,\ \log(S_i) \leq \log(|S|))
\end{aligned}$$

■

Appendix D. Deferred proofs from Section 4

Lemma 50 *Let D be a distribution on $\mathbb{R}^k$ that is (α, λ)-strictly subexponential. Let $f : \mathbb{R}^k \mapsto \{\pm 1\}$ be a boolean function such that for all $\mathbf{r} \in \mathbb{R}^k$ it holds that the GSA of $f_{\mathbf{r}}$ is at most Γ. Then there exists a polynomial $p_{\mathbf{z}}$ parametrized by $\mathbf{z}$ and large universal constant C such that*

1. *The (expected) L_1 error of $p_{\mathbf{z}}$, $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}_k} \mathbf{E}_{\mathbf{u}\sim D}\left[|p_{\mathbf{z}}(\mathbf{u}) - f(\mathbf{z}+\mathbf{u})|\right]$ is at most ϵ,*
2. *The degree of $p_{\mathbf{z}}$ is at most $\left(C\lambda k \Gamma^2 \log(1/\epsilon)/\epsilon^2\right)^{64(1+1/\alpha)^3}$,*
3. *Every coefficients of $p_{\mathbf{z}}$ is bounded(in absolute value) by $e^{\left(C\lambda k \Gamma^2 \log(1/\epsilon)/\epsilon^2\right)^{70(1+1/\alpha)^3}}$.*

Proof From Lemma 12, we have $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}(0,I)}\left[|T_\rho f_{\mathbf{u}}(\mathbf{z}) - f(\mathbf{z}+\mathbf{u})|\right] \leq 2\sqrt{\pi\rho}\cdot\Gamma$. Choosing $\rho = O(\epsilon^2/\Gamma^2)$ makes this error at most $\epsilon/2$.

We now approximate $T_\rho f_{\mathbf{u}}$ using a polynomial.We know that $\mathbf{E}_{\mathbf{u}\sim D}\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[|T_\rho f_{\mathbf{u}}(\mathbf{z}) - f_{\mathbf{u}}(\mathbf{z})|\right]$ is at most $\sqrt{\mathbf{E}_{\mathbf{u}\sim D}\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[\left(T_\rho f_{\mathbf{u}}(\mathbf{z}) - f_{\mathbf{u}}(\mathbf{z})\right)^2\right]}$. Thus, using Lemma 51, we get a polynomial $p_{\mathbf{z}}$ of degree $\left(C\lambda k \log(1/\epsilon)\Gamma^2/\epsilon^2\right)^{64(1+1/\alpha)^3}$ such that $\mathbf{E}_{\mathbf{u}\sim D}\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[|T_\rho f_{\mathbf{u}}(\mathbf{z}) - p_{\mathbf{z}}(\mathbf{u})|\right] \leq \epsilon/2$ where C is a large universal constant. Also recall that the coefficients of $p_{\mathbf{z}}$ are bounded by $e^{(C\lambda k\log(1/\epsilon)/\rho)^{70(1+1/\alpha)^3}}$. From a triangle inequality, we get $\mathbf{E}_{\mathbf{u}\sim D}\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[|p_{\mathbf{z}}(\mathbf{u}) - f(\mathbf{z}+\mathbf{u})\right] \leq \epsilon$. ■

We now construct the polynomial approximator for $T_\rho f_{\mathbf{u}}$ and bound it's error.

Lemma 51 (Polynomial approximation of $T_\rho f_{\mathbf{u}}$) *Let D be a (α, λ)-strictly subexponential distribution on $\mathbb{R}^k$ and $f : \mathbb{R}^k \to \{\pm 1\}$ be any function. Let $f_{\mathbf{u}}$ be the function defined as $f_{\mathbf{u}}(\mathbf{x}) = f(\mathbf{u}+\mathbf{x})$. Then, there exists a polynomial $p_{\mathbf{z}}(\mathbf{u})$ parametrized by $\mathbf{z}$ and large universal constant C such that*

1. *It holds that $\mathbf{E}_{\mathbf{u}\sim D}\mathbf{E}_{\mathbf{z}\sim\mathcal{N}_k}\left[(p_{\mathbf{z}}(\mathbf{u}) - T_\rho f_{\mathbf{u}}(\mathbf{z}))^2\right] \leq \epsilon$,*
2. *The degree of the polynomial $p_{\mathbf{z}}$ is at most $(C\lambda k \log(1/\epsilon)/\rho)^{64(1+1/\alpha)^3}$,*
3. *Every coefficient of $p_{\mathbf{z}}$ is bounded(in absolute value) by $e^{(C\lambda k \log B \log(1/\epsilon)/\rho)^{70(1+1/\alpha)^3}}$.*

Proof Let Q be the distribution on $\mathbb{R}^k$ with probability distribution function $Q(\mathbf{x}) = \frac{1}{2^k}e^{-\sum|\mathbf{x}_i|}$. We have that

$$\begin{aligned}
T_\rho f_{\mathbf{u}}(\mathbf{z}) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_k}\left[f(\mathbf{u} + \sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{x})\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{u}/\rho, I)}\left[f(\sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{x})\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[f(\sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{x}) \cdot \frac{\mathcal{N}(\mathbf{x}; \mathbf{u}/\rho, I)}{Q(\mathbf{x})}\right] \\
&= e^{-\frac{\|\mathbf{u}\|_2^2}{2\rho^2}} \mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[f(\sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{x}) \cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2} - \log Q(\mathbf{x})} e^{(\mathbf{u}/\rho)\cdot\mathbf{x}}\right]
\end{aligned}$$

where the second equality follows by recentering the distribution of $\mathbf{x}$ and the final equality comes from expanding the ratio of the two probability density functions.

We now define a polynomial $p_{\mathbf{z}}(\mathbf{u})$ approximating $T_\rho f_{\mathbf{u}}(\mathbf{z})$. To do this, we approximate $e^{-\frac{\|\mathbf{u}\|_2^2}{2\rho^2}}$ and $e^{\mathbf{u}\cdot\mathbf{x}}$ using polynomials in $\mathbf{u}$. First, we use a polynomial $p_1(\mathbf{u})$ to approximate $e^{\left(\frac{-\|\mathbf{u}\|_2^2}{2\rho^2}\right)}$. This polynomial is given by the following lemma whose proof is at the end of this section. We choose the parameters later.

Lemma 52 *Let $b \in \mathbb{Z}_+$. Let D be a (α,λ)-strictly subexponential distribution on $\mathbb{R}^k$. Then, there exists a polynomial q of degree $O\left((b^2\lambda k\log(1/\epsilon))^{2+2/\alpha}\right)$ such that*

1. *The approximation error $\mathbf{E}_{\mathbf{x}\sim D}\left[\left(q(\mathbf{x}) - e^{\left(-\|\mathbf{x}\|_2^2\right)}\right)^b\right]$ is upper bounded by ϵ*
2. *Every coefficient of q is bounded(in abolute value) by $k^{O\left((b^2\lambda k\log(1/\epsilon))^{2+2/\alpha}\right)}$.*

Second, to approximate $e^{(\mathbf{u}/\rho)\cdot\mathbf{x}}$, we use the function $p_2(\mathbf{u},\mathbf{x}) = p_e((\mathbf{u}/\rho)\cdot\mathbf{x})\mathbb{1}\{\|\mathbf{x}\|_2 \le T\}$ where $p_e(x) = 1 + \sum_{i=1}^{m-1}\frac{x^i}{i!}$ is the degree $m-1$ Taylor approximation of e^x. We choose m and T later. Thus, our final approximation $T_\rho f_{\mathbf{u}}$ is

$$p_{\mathbf{z}}(\mathbf{u}) = p_1(\mathbf{u})\underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{x})\cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2}-\log Q(\mathbf{x})}p_2(\mathbf{u},\mathbf{x})\right]}_{q(\mathbf{u})}.$$

$p_{\mathbf{z}}(\mathbf{u})$ is a polynomial in $\mathbf{u}$ as both p_1 and p_2 are polynomials in $\mathbf{u}$ when $\mathbf{x}$ is fixed and the dependence on $\mathbf{x}$ gets marginalized out making q a polynomial as well. The indicator variable in the definition of p_2 makes our calculations easier and the analysis cleaner. We now want to bound $E_{\mathbf{u}\sim D}\,\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[(p_{\mathbf{z}}(\mathbf{u}) - T_\rho f_{\mathbf{u}}(\mathbf{z}))^2\right]$. To help us analyse the error, we define the "hybrid" function $\tilde{p}_{\mathbf{z}}(\mathbf{u})$ such that

$$\tilde{p}_{\mathbf{z}}(\mathbf{u}) = e^{-\frac{\|\mathbf{u}\|_2^2}{2\rho^2}}\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{x})\cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2}-\log Q(\mathbf{x})}p_2(\mathbf{u},\mathbf{x})\right].$$

We have that

$$\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_k}\left[(T_\rho f_{\mathbf{u}}(\mathbf{z}) - p_{\mathbf{z}}(\mathbf{u}))^2\right] \le 2\cdot\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\Bigg[\underbrace{\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_k}\left[(T_\rho f_{\mathbf{u}}(\mathbf{z}) - \tilde{p}_{\mathbf{z}}(\mathbf{u}))^2\right]}_{\Delta_1(\mathbf{u})} + \underbrace{\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\left[(\tilde{p}_{\mathbf{z}}(\mathbf{u}) - p_{\mathbf{z}}(\mathbf{u}))^2\right]}_{\Delta_2(\mathbf{u})}\Bigg]$$

from the fact that $(a+b)^2 \le 2(a^2+b^2)$. We now bound $\Delta_1(\mathbf{u})$ and $\Delta_2(\mathbf{u})$ separately. We have that

$$\begin{aligned}\Delta_1(\mathbf{u}) &= \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_k}\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[f^2(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{x})e^{2\left(\frac{-\|\mathbf{u}\|_2^2}{2\rho^2}-\frac{\|\mathbf{x}\|_2^2}{2}-\log Q(\mathbf{x})\right)}\left(p_2(\mathbf{u},\mathbf{x}) - e^{(\mathbf{u}/\rho)\cdot\mathbf{x}}\right)^2\right]\\ &= \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_k}\underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[e^{\left(-\frac{\|\mathbf{u}\|_2^2}{\rho^2}-\|\mathbf{x}\|_2^2-2\log\Phi(\mathbf{x})\right)}\left(p_2(\mathbf{u},\mathbf{x}) - e^{(\mathbf{u}/\rho)\cdot\mathbf{x}}\right)^2\right]}_{\tilde{\Delta}_1(\mathbf{u})}\end{aligned}$$

since $|f(\mathbf{x})| = 1$. Observe that $\tilde{\Delta}_1(\mathbf{u})$ can be bounded as the sum of the following two terms.

$$\Delta_{11}(\mathbf{u}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[e^{\left(\frac{-\|\mathbf{u}\|_2^2}{\rho^2} - \|\mathbf{x}\|_2^2 - 2\log Q(\mathbf{x})\right)} \frac{e^{2|(\mathbf{u}/\rho)\cdot\mathbf{x}|}}{(m!)^2} |(\mathbf{u}/\rho)\cdot\mathbf{x}|^{2m} \mathbb{1}\{\|\mathbf{x}\| \le T\}\right]$$

and

$$\Delta_{12}(\mathbf{u}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[e^{\left(\frac{-\|\mathbf{u}\|_2^2}{\rho^2} - \|\mathbf{x}\|_2^2 - 2\log Q(\mathbf{x})\right)} e^{2(\mathbf{u}/\rho)\cdot\mathbf{x}} \mathbb{1}\{\|\mathbf{x}\| > T\}\right]$$

where we used the fact that $|p_e(x) - e^x| \le \frac{e^{|x|}}{m!}\cdot|x|^m$ and the fact that $p_2(\mathbf{u},\mathbf{x}) = 0$ when $\|\mathbf{x}\|_2 \ge T$.

We first bound Δ_{11}. We have that

$$\begin{aligned}\Delta_{11}(\mathbf{u}) &\le 2\cdot \mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[e^{\left(-\frac{\|\mathbf{u}\|_2^2}{\rho^2} - \|\mathbf{x}\|_2^2 - 2\log Q(\mathbf{x})\right) + 2(\mathbf{u}/\rho)\cdot\mathbf{x}}\right] \frac{(T\|\mathbf{u}/\rho\|_2)^{2m}}{(m!)^2} \\ &\le 2\cdot \frac{(T\|\mathbf{u}/\rho\|_2)^{2m}}{(m!)^2}\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[\left(\frac{\mathcal{N}(\mathbf{x};\mathbf{u}/\rho, I)}{Q(\mathbf{x})}\right)^4\right]} \le C_1^k e^{C_1\|\mathbf{u}/\rho\|_1}\frac{(T\|(\mathbf{u}/\rho\|_2)^{2m}}{(m!)^2}\end{aligned}$$

where C_1 is a large constant. The first inequality follows using the fact that Q is symmetric(we replaced $e^{|(\mathbf{u}/\rho)\cdot\mathbf{x}|}$ by $2e^{(\mathbf{u}/\rho)\cdot\mathbf{x}}$) and the second inequality follows from an application of Cauchy Schwartz. The last inequality follows from following claim whose proof is in the end of this section.

Lemma 53 *Define the distribution Q on $\mathbb{R}^k$ with density function $Q(\mathbf{x}) = (1/2)^k \mathbf{e}^{-\|\mathbf{x}\|_1}$. Then, there exists a large universal constant C such that for every vector $\mathbf{u}$, it holds that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[\left(\frac{\mathcal{N}(\mathbf{x};\mathbf{u},\mathbf{I})}{Q(\mathbf{x})}\right)^4\right] \le C^k e^{C\|\mathbf{u}\|_1}.$$

We now compute $\mathbf{E}_{\mathbf{u}\sim D}[\Delta_{11}(\mathbf{u})]$ as we will need it later.

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{u}\sim D}[\Delta_{11}(\mathbf{u})] &\le C_1^k\sqrt{\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[e^{2C_1\|\mathbf{u}/\rho\|_1}\right]\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[\frac{(T\|\mathbf{u}/\rho\|_2)^{4m}}{(m!)^4}\right]} \\ &\le C_1^k\sqrt{\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[e^{2C_1\|\mathbf{u}/\rho\|_1}\right]\sum_{i=1}^k k^m \mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[\frac{(T|\mathbf{u}_i/\rho|)^{4m}}{(m!)^4}\right]} \\ &\le C^{(C\lambda k/\rho)^{3+3/\alpha}}\left(\frac{CkeT\lambda}{m^{\alpha/1+\alpha}\cdot\rho}\right)^{4m} \le \delta\end{aligned}$$

when $m = (C'T\lambda k/\rho)^{3+3/\alpha}\log(1/\delta)$ where C' is a large enough constant. The first inequality follows from an application of the Cauchy-Schwartz inequality. The second inequality uses the fact that $\|\mathbf{u}\|_2 \le \sqrt{k}\|\mathbf{u}\|_\infty \le \sqrt{k}\sum_{i=1}^k |\mathbf{u}_i|$. The third inequality is obtained using Definition 11 and the following claim.

Lemma 54 (Section D.1) *If D on $\mathbb{R}^k$ is (α,λ)-strictly subexponential, then for any constant $b > 0$, we have*

$$\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[e^{b\|\mathbf{u}\|_1}\right] \le C^{(Cb\lambda k)^{3+3/\alpha}}$$

for large enough constant $C > 0$.

We now bound $\Delta_{12}(\mathbf{u})$.

$$\begin{aligned}\Delta_{12}(\mathbf{u}) &\leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[\left(\frac{\mathcal{N}(\mathbf{x};\mathbf{u}/\rho,I)}{Q(\mathbf{x})}\right)^4\right]\cdot \Pr_{\mathbf{x}\sim Q}[\|\mathbf{x}\|_2 > T]}\\ &\leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[\left(\frac{\mathcal{N}(\mathbf{w};\mathbf{u}/\rho,I)}{Q(\mathbf{x})}\right)^4\right]\cdot k\cdot e^{-T/k}}\\ &\leq C_2^k e^{C_2\|\mathbf{u}/\rho\|_1}e^{-T/k}\end{aligned}$$

where the first inequality is Cauchy-Schwartz. The last inequality follows from Lemma 53. The second inequality follows from the following fact about the exponential tail(proof in the end of this section).

Lemma 55 *Let D be the distribution on $\mathbb{R}^k$ with density function $\Phi(\mathbf{x}) = (1/2)^k e^{-\|\mathbf{x}\|_1}$. We have that*

$$\Pr_{\mathbf{x}\sim D}[\|\mathbf{x}\|_2 > T] \leq 2k\cdot e^{-T/k}$$

Thus, we have that

$$\mathop{\mathbf{E}}_{\mathbf{u}\sim D}[\Delta_{12}(\mathbf{u})] \leq C_2^k\cdot \mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[e^{C_2\|\mathbf{u}/\rho\|_1}\right]e^{(-T/k)} \leq C^k\cdot C^{(C\lambda k/\rho)^{3+3/\alpha}}\cdot e^{(-T/k)} \leq \delta$$

when $T = ((C\lambda k/\rho)^{4+4/\alpha}\log(1/\delta))$.

Plugging this into the bound for m, we get $m \leq (C'\lambda k\log(1/\delta)/\rho)^{15(1+1/\alpha)^2}$ for large constant C'. Thus we now get $\mathbf{E}_{\mathbf{u}\sim D}[\Delta_1(\mathbf{u})] \leq \mathbf{E}_{\mathbf{u}\sim D}\big[\Delta_{11}(\mathbf{u}) + \Delta_{12}(\mathbf{u})\big] \leq \epsilon/4$ when $\delta = \epsilon/8$.

We now bound $\Delta_2(\mathbf{u})$. We have that

$$\begin{aligned}\Delta_2(\mathbf{u}) &= \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_k}\big[(\tilde{p}_{\mathbf{z}}(\mathbf{u}) - p_{\mathbf{z}}(\mathbf{u}))^2\big]\\ &= \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_k}\left[\big(p_1(\mathbf{u}) - e^{-\frac{\|\mathbf{u}\|_2^2}{2\rho^2}}\big)^2\cdot\big(\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\big[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{x})\cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2}-\log Q(\mathbf{x})}\cdot p_2(\mathbf{u},\mathbf{x})\big]\big)^2\right]\\ &\leq \big(p_1(\mathbf{u}) - e^{-\frac{\|\mathbf{u}/\rho\|_2^2}{2}}\big)^2\cdot\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_k}\left[\big(\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\big[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{x})\cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2}-\log Q(\mathbf{x})}\cdot p_2(\mathbf{u},\mathbf{x})\big]\big)^2\right]\end{aligned}$$

where the last inequality follows since $\mathbf{u}$ doesn't depend on $\mathbf{z}$. We bound the last term in the product above as

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_k}\left[\big(\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\big[f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{x})\cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2}-\log Q(\mathbf{x})}\cdot p_2(\mathbf{u},\mathbf{x})\big]\big)^2\right]\\ &\quad\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[e^{-\|\mathbf{x}\|_2^2-2\log Q(\mathbf{x})}\right]\cdot\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[\left(1+\sum_{i=1}^{m-1}\frac{((\mathbf{u}/\rho)\cdot\mathbf{x})^i}{i!}\right)^2\mathbb{1}\{\|\mathbf{x}\|_2 < T\}\right]\\ &\quad\leq 2^{2k}\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[e^{-\|\mathbf{x}\|_2^2+2\|\mathbf{x}\|_1}\right]\cdot\left(\sum_{i=0}^{m-1}(T\|\mathbf{u}/\rho\|_2)^i\right)^2 \leq C^k\cdot\left(\sum_{i=0}^{m-1}(T\|\mathbf{u}/\rho\|_2)^i\right)^2\end{aligned}$$

for large enough constant C. The first inequality follows from an application of Cauchy Schwartz and the fact that f is boolean. The second inequality comes from expanding p_2 and conditioning on

the event that $\|\mathbf{x}\|_2 < T$. The last inequality comes from applying Cauchy Schwartz once more and using Lemma 53 with $\mathbf{u}$ set to zero.

Thus, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{u}\sim D}[\Delta_2(\mathbf{u})] &\le C^k \mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[\left(p_1(\mathbf{u}) - e^{-\frac{\|\mathbf{u}/\rho\|_2^2}{2}}\right)^2 \cdot \left(\sum_{i=0}^{m-1}(T\|\mathbf{u}/\rho\|_2)^i\right)^2\right] \\
&\le C^k \sqrt{\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[\left(p_1(\mathbf{u}) - e^{-\frac{\|\mathbf{u}/\rho\|_2^2}{2}}\right)^4\right] \cdot \mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[\left(\sum_{i=0}^{m-1}(T\|\mathbf{u}/\rho\|_2)^i\right)^4\right]} \\
&\le C^k \cdot \delta \cdot \sqrt{\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[(mT^m)^4 \max_{i=0}^{m-1}\|\mathbf{u}\|_2^{4i}\right]} \\
&\le C^k \cdot \delta \cdot (mT^m)^2 (4mk\lambda)^{2m} \\
&\le \epsilon/4
\end{aligned}$$

when δ is chosen accordingly. The second inequality is obtained by applying Cauchy-Schwartz. $p_1(\mathbf{u})$ is chosen such that it has approximation error δ^2 when using Lemma 60 with exponent 4. The penultimate inequality is obtained by using the fact that $\|\mathbf{u}\|_2 \le \sqrt{k}\|\mathbf{u}\|_\infty$ and then using Definition 15. The degree of $p_1(\mathbf{u})$ required to get this error is $O\left((C'm^2k^2\lambda\log(1/\epsilon)/\rho)^{2+2/\alpha}\right)$ where $C' > 0$ is a large enough universal constant.

Putting everything together, we get that $E_{\mathbf{u}\sim D}E_{\mathbf{z}\sim\mathcal{N}}\left[(T_\rho f_{\mathbf{u}}(\mathbf{z}) - p_{\mathbf{z}}(\mathbf{u}))^2\right] \le \epsilon$. The total degree of $p_{\mathbf{z}}(\mathbf{u})$ is $\deg(p_1) + \deg(p_2)$ which is at most $(C\lambda k\log(1/\epsilon)/\rho)^{64(1+1/\alpha)^3}$ for large enough C.

We now bound the coefficients of $p_{\mathbf{z}}$. To do this we first recall the definition of $p_{\mathbf{z}}(\mathbf{u})$ and observe some properties. Recall that $p_{\mathbf{z}}(\mathbf{u}) = p_1(\mathbf{u}) \cdot q(\mathbf{u})$ where

$$q(\mathbf{u}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[f(\sqrt{1-\rho^2}\mathbf{z} + \rho\mathbf{x}) \cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2} - \log Q(\mathbf{x})} \cdot p_2(\mathbf{u}/\rho, \mathbf{x})\right].$$

Thus, $p_{\mathbf{z}}(\mathbf{u})$ is the product of two polynomials. $p_1(\mathbf{u})$ has bounded coefficients as given by *Lemma* 52. We now bound the coefficients of $q(\mathbf{u})$. Since $p_2(\mathbf{u}, \mathbf{x}) = p_e((\mathbf{u}/\rho)\cdot\mathbf{x})\mathbb{1}\{\|\mathbf{x}\|_2 \le T\}$, the $q(\mathbf{u})$ term only picks up non zero values when $\|\mathbf{x}\|_2 \le T$. Note that for each fixed $\mathbf{w}$, the term inside the expectation is a polynomial in $\mathbf{u}$. Thus, proving an absolute bound on the coefficients when $\|\mathbf{x}\|_2 \le T$ will bound the final coefficients of $q(\mathbf{u})$.

We now bound the coefficients of the polynomial $f(\sqrt{1-\rho^2}\mathbf{z}+\rho\mathbf{x})\cdot e^{-\frac{\|\mathbf{x}\|_2^2}{2} - \log Q(\mathbf{x})}\cdot p_2(\mathbf{u}/\rho, \mathbf{x})$ where $\mathbf{x}$ is a fixed vector of norm atmost T. Since $\|\mathbf{x}\|_2 \le T$, we also have that $|\mathbf{x}_i| \le T$ for all $i \in [k]$. We know that f is bounded by 1 and $e^{\left(-\|\mathbf{x}\|_2^2 - 2\log Q(\mathbf{x})\right)} \le e^{k\|\mathbf{x}\|_1} \le e^{k^2T}$. Now, we bound the coefficients of $p_2(\mathbf{u}, \mathbf{x})$. This is the composition of two polynomials, $p_e(x)$ and $(\mathbf{u}/\rho)\cdot\mathbf{x}$. The degree of p_e is m and the coefficients are at most 1. The degree of $(\mathbf{u}/\rho)\cdot\mathbf{x}$ is 1 and coefficients are bounded by T/ρ. Applying Lemma 41, we get that the coefficients of $p_2(\mathbf{u}, \mathbf{x})$ are bounded by $e^{O(mk\log T/\rho)}$. Putting everything together, we obtain that the coefficients of $q(\mathbf{u})$ are bounded by $e^{O(k^2m^2/\rho)}$ as $T \le m$. Putting together the coefficient bounds for $p_1(\mathbf{u})$ and $q(\mathbf{u})$, we get that the coefficients of $p_{\mathbf{z}}$ are bounded by atmost $e^{(C\lambda k\log B\log(1/\epsilon)/\rho)^{70(1+1/\alpha)^3}}$ for large constant C and for any $\mathbf{z}$. ■

D.1. Proof of Theorem 56

We give our algorithm here for completeness. We run polynomial regression and the hypothesis we output is a PTF with an appropriately chosen bias term.

Algorithm 2 Agnostic Learner for Smooth Boolean Concepts

Input: Labeled Dataset $S = \{(\mathbf{x}_i, y_i)\}_{i\in[N]}$, degree ℓ,
Output: Hypothesis h

Find polynomial P of degree at most ℓ such that P minimizes $\frac{1}{N}\sum_{i=1}^{N} |P(\mathbf{x}_i) - y_i|$
Choose $t \in [-1, 1]$ such that $\sum_{i=1}^{N} \mathbb{1}[\mathrm{sign}\big(P(\mathbf{x}_i) - t\big) \neq y_i]$ is minimized.
Output hypothesis h, such that $h(\mathbf{x}) = \mathrm{sign}\big(P(\mathbf{x}) - t\big)$

We are now ready to state and prove the main theorem of this section.

Theorem 56 *Let $k \in \mathbb{Z}_+$ and $\epsilon, \delta, \sigma \in (0, 1)$. Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that the marginal distribution is (α, λ)-strictly subexponential. There exists an algorithm that draws $N = d^{\mathrm{poly}\left((k\lambda\Gamma/(\sigma\epsilon))^{(1+1/\alpha)^3}\right)}$ samples, runs in time $\mathrm{poly}(d, N)$, and computes a hypothesis $h(\mathbf{x})$ such that, with probability at least $1 - \delta$, it holds*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y \neq h(\mathbf{x})] \leq \min_{f\in\mathcal{F}(k,\Gamma)} \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y \neq f(\mathbf{x} + \sigma\mathbf{z})] + \epsilon\,.$$

Proof

Let $\ell = O\left(\left(\lambda k\Gamma^2 \log(1/\epsilon)/(\sigma\epsilon^2)\right)^{64(1+1/\alpha)^3}\right)$ and $N = \mathrm{poly}(d^\ell/\epsilon)$. We denote the marginal distribution of D by $D_{\mathbf{x}}$. The algorithm for this task is simple, repeat Algorithm 2 $r = O(\log(1/\delta)/\epsilon)$ times with degree ℓ and N fresh samples each time. Output the hypothesis that has the minimum loss on an independent validation set of size $O(\log(1/\delta)/\epsilon^2)$. The time required is the time required for polynomial regression which is $\mathrm{poly}(d, N)$.

We now analyze the correctness. Let $f^* \in \mathcal{F}(k, \Gamma)$ that be the function that achieves opt_σ. Using the fact that the function is low dimensional, there exists an orthonormal matrix $\mathbf{P}$ such that $f^*(\mathbf{x}) = f^*(\mathbf{P}\mathbf{P}^T\mathbf{x})$ for all $\mathbf{x}$. This implies that $f^*(\mathbf{x}) = g^*(\mathbf{P}\mathbf{x})$ where g^* is the function on $\mathbb{R}^k$ defined as $g^*(\mathbf{u}) = f^*(\mathbf{P}\mathbf{u})$. Let f_σ be defined as the function defined as $f_\sigma(\mathbf{u}) = f(\sigma\mathbf{u})$. Lemma 45 and the fact that $\mathcal{F}(k, \Gamma)$ is closed under scaling imply that $\Gamma(g^*_\sigma) \leq \Gamma$.

Lemma 57 *If D on $\mathbb{R}^d$ is (α, λ)-strictly subexponential, then for any $\mathbf{A} \in \mathbb{R}^{k\times d}$, the distribution of $\mathbf{y} = \mathbf{A}\mathbf{x}$ when $\mathbf{x} \sim D$ is $\big(\alpha, \lambda\|\mathbf{A}\|_2\big)$-strictly subexponential.*

From the above claim(proof at the end of the section), observe that for $\mathbf{x}$ sampled from $D_{\mathbf{x}}$, the distribution of $\mathbf{P}^T\mathbf{x}/\sigma$ is $(\alpha, \lambda/\sigma)$-strictly subexponential on $\mathbb{R}^k$. Thus, Lemma 50 implies that there exists a polynomial $p_{\mathbf{z}}$ with degree at most ℓ and coefficients at most $k^{\mathrm{poly}(\ell)}$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim D_{\mathbf{x}}} \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\big[|p_{\mathbf{z}}(\mathbf{P}^T\mathbf{x}) - g^*_\sigma\big(\mathbf{P}^T\mathbf{x}/\sigma + \mathbf{z}\big)|\big] \leq \epsilon/2\,.$$

We first analyze the success probability of one run of the algorithm. Let $S = \{(\mathbf{x}_i, y_i)\}_{i\in[N]}$ be the set of samples and let p_S, h_S be the polynomial and hypothesis output by the algorithm.

From the proof of Theorem 5 of Kalai et al. (2005), we have that $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{h_S(\mathbf{x}_i) \neq y_i\} \leq \frac{1}{2N}\sum_{i=1}^{N}|p_S(\mathbf{x}_i) - y_i|$. We now bound the right hand side. We have that

$$\begin{aligned}\frac{1}{2N}\sum_{i=1}^{N}|p_S(\mathbf{x}_i) - y_i| &\leq \frac{1}{2N}\sum_{i=1}^{N}|p_z(\mathbf{P}^T\mathbf{x}_i) - y_i| \\ &\leq \frac{1}{2N}\sum_{i=1}^{N}|f^*(\mathbf{x}_i + \sigma\mathbf{z}) - y_i| + \frac{1}{2N}\sum_{i=1}^{N}|f^*(\mathbf{x}_i + \sigma\mathbf{z}) - p_z(\mathbf{x}_i)|\end{aligned}$$

where $\mathbf{z}$ is an arbitrary vector from $\mathbb{R}^k$. The first follows from the fact that p minimizes the empirical error among polynomials of degree less than ℓ and the last inequality follows from a triangle inequality. Observe that the expectation(with respect to $\mathbf{z} \sim \mathcal{N}(0, I_d)$ and $(\mathbf{x}, y) \sim D$) of the right hand side of the last equality is $\mathrm{opt}_\sigma + \epsilon/2$. Thus, we have that

$$\mathop{\mathbf{E}}_{S\sim D^{\otimes N}}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{h_S(\mathbf{x}_i) \neq y_i\}\right] \leq \mathrm{opt}_\sigma + \epsilon/2$$

Since our hypothesis h_S is a PTF of degree ℓ on m variables, VC theory tells us that for $N = \mathrm{poly}(d^\ell/\epsilon)$, we have that

$$\mathop{\mathbf{E}}_{S\sim D^{\otimes N}}\left[\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y \neq h_S(\mathbf{x})]\right] \leq \mathrm{opt}_\sigma + 3\epsilon/4$$

By Markov's inequality, we have that with probability at least $\epsilon/16$ over samples S, we have that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq h_S(\mathbf{x})] \leq \mathrm{opt}_\sigma + 7\epsilon/8$. Let $h_1, h_2, \ldots, h_r$ be the r hypotheses outputted on the $r = O(\log(1/\delta)/\epsilon)$ repetitions of algorithm 2. With probability at least $1 - \delta/4$, there exists $i \in [r]$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq h_i(\mathbf{x})] \leq \mathrm{opt}_\sigma + 7\epsilon/8$. Thus, using the validation set of size $O(\log(1/\delta)/\epsilon^2)$, with probability at least $1 - \delta/4$, we choose a hypothesis h such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq h(\mathbf{x})] \leq \mathrm{opt}_\sigma + \epsilon$. Thus with total error probability of at most δ, our algorithm outputs a hypothesis h such that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y \neq h(\mathbf{x})] \leq \mathrm{opt}_\sigma + \epsilon$$

■

Lemma 58 *If D on $\mathbb{R}^k$ is (α, λ)-strictly subexponential, then we have*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim D}\left[\|\mathbf{x}\|_2 > T\right] \leq 2k \cdot e^{\left(-(T/k\lambda)^{(1+\alpha)}\right)}.$$

Proof

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim D}\left[\|\mathbf{x}\|_2 > T\right] \leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim D}\left[\sum_{i=1}^{k}|\mathbf{x}_i| > T\right] \leq \sum_{i=1}^{k}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim D}\left[|\mathbf{x}_i| > T/k\right] \leq 2k \cdot e^{\left(-(T/k\lambda)^{(1+\alpha)}\right)}$$

where the second inequality follows from a union bound and the last follows from Definition 15. ■

Proof [Proof of Lemma 53] The proof below is a straightforward calculation by completing the squares.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim Q}\left[\left(\frac{\mathcal{N}(\mathbf{x};\mathbf{u},\mathbf{I})}{Q(\mathbf{x})}\right)^4\right] \\
&= \frac{2^{3k}}{(2\pi)^{2k}}\int_{\mathbf{z}\in\mathbb{R}^k} e^{-(2\|\mathbf{u}\|_2^2+2\|\mathbf{z}\|_2^2-4\mathbf{u}\cdot\mathbf{z}-3\|\mathbf{z}\|_1)}d\mathbf{z} = \frac{2^{3k}}{(2\pi)^{2k}}\cdot\prod_{i=1}^k\int_{\mathbf{z}_i\in\mathbb{R}} e^{-(2\mathbf{u}_i^2+2\mathbf{z}_i^2-4\mathbf{u}_i\mathbf{z}_i-3|\mathbf{z}_i|)}d\mathbf{z}_i \\
&\le (C')^k\cdot\prod_{i=1}^k\int_{\mathbf{z}_i\in\mathbb{R}} e^{-(2\mathbf{u}_i^2+2\mathbf{z}_i^2-4\mathbf{u}_i\mathbf{z}_i-3\mathbf{z}_i)}d\mathbf{z}_i \le (C')^k e^{-9k/2}\cdot\prod_{i=1}^k\int_{\mathbf{z}_i\in\mathbb{R}} e^{6\mathbf{u}_i}e^{-(2\sqrt{2}\mathbf{z}_i-(\sqrt{2}\mathbf{u}_i-3/\sqrt{2}))^2} \\
&\le C^k e^{C\|\mathbf{u}\|_1}
\end{aligned}$$

where C', C are appropriately chosen constants. ■

Lemma 59 *If D on $\mathbb{R}^k$ is (α,λ)-strictly subexponential, then for any constant $b>0$ and vector $\mathbf{v}$ with $\|\mathbf{v}\|_2=1$, we have*

$$\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[e^{b|\mathbf{v}\cdot\mathbf{u}|}\right] \le C^{(Cb\lambda)^{3+3/\alpha}}$$

for large enough constant $C>0$.

Proof

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[e^{b|\mathbf{v}\cdot\mathbf{u}|}\right] &= 1+\sum_{i=1}^\infty\frac{\mathbf{E}_{\mathbf{u}\sim D}\left[b^i|\mathbf{v}\cdot\mathbf{u}|^i\right]}{i!} \le 1+\sum_{i=1}^\infty\frac{b^i\lambda^i(i)^{i/(1+\alpha)}}{(i/e)^i} \le 1+\sum_{i=1}^\infty\left(\frac{be\lambda}{i^{\alpha/(1+\alpha)}}\right)^i \\
&\le 1+\sum_{i=1}^{(2be\lambda)^{1+1/\alpha}}(be\lambda)^i+\sum_{i=1}^\infty\frac{1}{2^i} \le 1+(be\lambda)^{(2be\lambda)^{1+1/\alpha}+1}+\sum_{i=1}^\infty\frac{1}{2^i} \le C^{(Cb\lambda)^{3+3/\alpha}}.
\end{aligned}$$

for some constant $C>0$. The first equality follows from a Taylor expansion, and the rest are straightforward calculations. ■

Proof [Proof of Lemma 54]

$$\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[e^{b\|\mathbf{u}\|_1}\right] \le \mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[\prod_{i=1}^k e^{b|\mathbf{u}_i|}\right] \le \prod_{i=1}^k\left(\mathop{\mathbf{E}}_{\mathbf{u}\sim D}\left[e^{bk|\mathbf{u}_i|}\right]\right)^{1/k} \le C^{(Cb\lambda k)^{3+3/\alpha}}$$

where the penultimate inequality follows from Hölder and the last inequality follows from Lemma 59. ■

Proof [Proof of Lemma 55]

$$\Pr_{\mathbf{x}\sim D}\left[\|\mathbf{x}\|_2>T\right] \le \Pr_{\mathbf{x}\sim D}\left[\sum_{i=1}^k|\mathbf{x}_i|>T\right] \le \sum_{i=1}^k\Pr_{\mathbf{x}\sim D}\left[|\mathbf{x}_i|>T/k\right] \le 2k\cdot e^{-T/k}$$

where the last inequality follows from the tail of a univariate exponential random variable. ■

Proof [Proof of Lemma 57] Let $\mathbf{v}$ be a vector such that $\|\mathbf{v}\|_2 = 1$. Observe that $|\mathbf{v} \cdot (\mathbf{A}\mathbf{x})| = |(\mathbf{A}^T\mathbf{v}) \cdot \mathbf{x}| \leq \|\mathbf{A}^T\mathbf{v}\|_2 |\mathbf{u} \cdot \mathbf{x}| \leq \|\mathbf{A}\|_2 |\mathbf{u} \cdot \mathbf{x}|$ where $\mathbf{u} = \frac{\mathbf{A}^T\mathbf{v}}{\|\mathbf{A}^T\mathbf{v}\|_2}$. We have that $\mathbf{Pr}_{\mathbf{x}\sim D}[|\mathbf{v} \cdot (\mathbf{A}\mathbf{x})| \geq t] \leq \mathbf{Pr}_{\mathbf{x}\sim D}[|\mathbf{u} \cdot \mathbf{x}| \geq t/\|\mathbf{A}\|_2] \leq 2 \cdot e^{-(t/\lambda\|\mathbf{A}\|_2)^{1+\alpha)}}$. Considering the second condition from Definition 15, we have that

$$\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[|\mathbf{v} \cdot \mathbf{A}\mathbf{x}|^m\right]\right)^{1/m} \leq \|\mathbf{A}\|_2 \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[|\mathbf{u} \cdot \mathbf{x}|^m\right]\right)^{1/m} \leq \|\mathbf{A}\|_2 \lambda m^{1/(1+\alpha)}\,.$$

Finally, we have that $\mathbf{E}_{\mathbf{x}\sim D}\left[e^{(|(\mathbf{A}\mathbf{x})\cdot v|/\lambda\|\mathbf{A}\|_2)^{1+\alpha}}\right] \leq \mathbf{E}_{\mathbf{x}\sim D}\left[e^{(|\mathbf{x}\cdot v|/\lambda)^{1+\alpha}}\right] \leq 2\,.$ ■

We now show that under any (α, λ)-strictly subexponential distribution, we can approximate the function $e^{-\|\mathbf{x}\|_2^2}$ by a polynomial of degree $O\left(b^2\lambda(k\log(1/\epsilon))^{1+1/\alpha}\right)$. For this, we use the following theorem from Aggarwal and Alman (2022).

Lemma 60 *For $T > 0$ and error $\epsilon > 0$, there exists a polynomial p such that*

1. $\sup_{x\in[0,T]} |p(x) - e^{(-x)}| \leq \epsilon$
2. $\deg(p) \leq O(\sqrt{T\log(1/\epsilon)})$, *if* $T = \omega(\log 1/\epsilon)$
3. $p(x) = \sum_{i=0}^{\deg(p)} c_i x^i$ *where* $|c_i| \leq e^{\left(C\left(\sqrt{T\log(1/\epsilon)}\right)\right)}$ *for all* $i \leq \deg(p)$. *Here C is a large enough constant.*

The bound on the coefficients is not explicitly stated in Aggarwal and Alman (2022) and hence we calculate them.

Lemma 61 *Every coefficient of the polynomial p in Lemma 60 is bounded(in absolute value) by $e^{C\sqrt{T\log(1/\epsilon)}}$*

Proof To bound the coefficients, we first recall their polynomial. Their polynomial $p(t)$ of degree $\ell = \sqrt{T\log(1/\epsilon)}$ approximating e^{-t} is

$$p(t) = \sum_{r=0}^{\ell} 2^{r-1} A_{r,T/2} Q_r(2t/T - 1)$$

where $Q_r(t) = 2^{1-r}\sum_{s=0}^{\lfloor r/2\rfloor}\binom{r}{2s}(t^2-1)^s t^{r-2s}$ and $|A_{r,\lambda}| \leq C^{2r}$ for large constant C. The asymptotics of $A_{r,\lambda}$ is explicitly stated in Aggarwal and Alman (2022). We now bound the coefficients of the polynomial $Q_r(t)$. We bound the coefficient of t^j in each term of the summation for arbitrary j. The coefficients of $(t^2-1)^s$ are bounded by C_1^s for large constant C_1 by using Lemma 41. From the expression of $Q_r(t)$, we have that each term in the summation contributes $\binom{r}{2s}C_1^s \leq C_2^r$ for large constant C_2. The previous inequality follows from the fact that $s < r$ and $\binom{n}{k} \leq 2^n$. Thus, summing from 0 to $\lceil r/2\rceil$ bounds the final coefficient of t^j by C_3^r for some constant C_3. Thus, the coefficients of $Q_r(2t/T-1)$ are bounded by C_4^r for constant C_4 by using Lemma 41 again. Now since $A_{r,T/2} \leq C^{2r}$, summing from 0 to l gives us that the coefficients are bounded by $e^{C'\ell}$ for large constant C' ■

We can now prove our result about approximating $e^{-\|\mathbf{x}\|_2^2}$ under strictly subexponential distributions.

Proof [Proof of Lemma 52] Let $p = \sum_{i=0}^{\deg(p)} c_i x^i$ be the polynomial obtained from Lemma 60 with error $\epsilon/2$ and $T = \omega(\log(1/\epsilon))$ to be chosen later. Our final polynomial is $q(\mathbf{x}) = p\left(\|\mathbf{x}\|_2^2\right)$. Clearly, $\deg(q) = 2 \cdot \deg(p) = O(\sqrt{T \log(1/\epsilon)})$. We now bound the error.

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[\left(q(\mathbf{x}) - e^{\left(-\|\mathbf{x}\|_2^2\right)}\right)^b\right] &\leq \epsilon/2 + E_{\mathbf{x}\sim D}\left[\left(q(\mathbf{x}) - e^{\left(-\|\mathbf{x}\|_2^2\right)}\right)^b \mathbb{1}\{\|\mathbf{x}\|_2^2 \geq T\}\right] \\
&\leq \epsilon/2 + \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[\left(q(\mathbf{x}) - e^{\left(-\|\mathbf{x}\|_2^2\right)}\right)^{2b}\right] \cdot \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[\mathbb{1}\{\|\mathbf{x}\|_2^2 \geq T\}\right]} \\
&\leq \epsilon/2 + \sqrt{2k \cdot \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[(|q(\mathbf{x})|+1)^{2b}\right] \cdot e^{\left(-(\sqrt{T}/k\lambda)^{(1+\alpha)}\right)}}\,.
\end{aligned}$$

The first inequality follows from the approximation error of p when $\|\mathbf{x}\|_2^2 \leq T$. We use Lemma 58 for the last inequality.

We now bound $\mathbf{E}_{\mathbf{x}\sim D}\left[(|q(\mathbf{x})|+1)^{2b}\right]$. We have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim D}[(|q(\mathbf{x})|+1)^{2b}] &\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[\left(1 + \sum_{i=0}^{\deg(p)} |c_i| \|\mathbf{x}\|_2^{2i}\right)^{2b}\right] \\
&\leq (\deg(p)+1)^{2b} \max_{i=0}^{\deg(p)} |c_i|^{2b} \max_{i=0}^{\deg(p)} \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[\|\mathbf{x}\|_2^{4bi}\right] \\
&\leq e^{\left(Cb\sqrt{T\log(1/\epsilon)}\right)} \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[\left(\sqrt{k}\,\|\mathbf{x}\|_\infty\right)^{4b\sqrt{T\log(1/\epsilon)}}\right] \\
&\leq \sqrt{k} e^{\left(Cb\sqrt{T\log(1/\epsilon)}\right)} \sum_{i=1}^{k} \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\left[(|\mathbf{x}_i|)^{4b\sqrt{T\log(1/\epsilon)}}\right]
\end{aligned}$$

for large enough constant C.

The first inequality follows from the definition of q and the second follows from linearity of expectation and a straightforward calculation: $(\deg(p)+1)^{2b}$ is the total number of terms in the summation when expanded, $\max_{i=0}^{\deg(p)} |c_i|^{2b}$ is an upper bound on any coefficient in the expansion and $\max_{i=0}^{\deg(p)} \mathbf{E}_{\mathbf{x}\sim D}[\|\mathbf{x}\|_2^{4bi}]$ is an upper bound on the expectation of any term in the expansion. The third inequality follows from the fact that $\|\mathbf{x}\|_2 \leq \sqrt{k}\,\|\mathbf{x}\|_\infty$. The second term in the right hand side of the last inequality above can be bounded by using Definition 15. Putting it all together,we get that $k \cdot \mathbf{E}_{\mathbf{x}\sim D}\left[(|q(x)|+1)^{2b}\right] e^{\left(-(\sqrt{T}/k\lambda)^{(1+\alpha)}\right)}$ is bounded by $e^{\left(C'\left(b^2 \log\lambda \log k \log(T\log(1/\epsilon))\sqrt{T\log(1/\epsilon)} - (T/\sqrt{k\lambda})^{(1+\alpha)/2}\right)\right)}$ where C' is a large enough constant.

Choosing $T = O\left(\left(b^2\lambda k \log(1/\epsilon)\right)^{3+3/\alpha}\right)$ makes the total error less than ϵ. Since T is $\omega(\log 1/\epsilon)$, the degree of the final polynomial is $O(\sqrt{T\log(1/\epsilon)})$ which is $O\left((b^2\lambda k \log(1/\epsilon))^{2+2/\alpha}\right)$.

We now bound the coefficients of q. We have that $q(\mathbf{x}) = p(\|\mathbf{x}\|_2^2)$ is the composition of two polynomials, p and $\|\mathbf{x}\|_2^2$. The degree of p is $O(\sqrt{T\log(1/\epsilon)})$ and coefficients bounded by

$e^{O\left(\sqrt{T\log(1/\epsilon)}\right)}$. The degree of $\|\mathbf{x}\|_2^2$ is 2 and it has coefficients equal to 1. Thus, using Lemma 41 with these polynomials, we get that the coefficients of q are bounded by $k^{O\left((b^2\lambda k\log(1/\epsilon))^{2+2/\alpha}\right)}$. ■

Appendix E. SQ Lower Bound for Smoothed Agnostic Learning

E.1. Background on SQ Lower Bounds

Our lower bound applies for the class of Statistical Query (SQ) algorithms. Statistical Query (SQ) algorithms are a class of algorithms that are allowed to query expectations of bounded functions of the underlying distribution rather than directly access samples. Formally, an SQ algorithm has access to the following oracle.

Definition 62 *Let $\mathcal{D}$ be a distribution on labeled examples supported on $X \times \{-1,1\}$, for some domain X. A statistical query is a function $q : X \times \{-1,1\} \to [-1,1]$. We define* $\mathrm{STAT}(\tau)$ *to be the oracle that given any such query $q(\cdot,\cdot)$ outputs a value v such that $|v - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[q(\mathbf{x},y)]| \le \tau$, where $\tau > 0$ is the tolerance parameter of the query.*

The SQ model was introduced by Kearns Kearns (1998) in the context of supervised learning as a natural restriction of the PAC model Valiant (1984a). Subsequently, the SQ model has been extensively studied in many contexts (see, e.g., Feldman (2016) and references therein). The class of SQ algorithms is rather broad and captures a range of known supervised learning algorithms. More broadly, several known algorithmic techniques in machine learning are known to be implementable using SQs. These include spectral techniques, moment and tensor methods, local search (e.g., Expectation Maximization), and many others (see, e.g., Feldman et al. (2017a,b)). Recent work Brennan et al. (2021) has shown a near-equivalence between the SQ model and low-degree polynomial tests.

Statistical Query Dimension To bound the complexity of SQ learning a concept class $\mathcal{C}$, we use the SQ framework for problems over distributions Feldman et al. (2017a).

Definition 63 (Decision Problem over Distributions) *Let D be a fixed distribution and $\mathfrak{D}$ be a family of distributions. We denote by $\mathcal{B}(\mathfrak{D}, D)$ the decision (or hypothesis testing) problem in which the input distribution D' is promised to satisfy either (a) $D' = D$ or (b) $D' \in \mathfrak{D}$, and the goal is to distinguish between the two cases.*

Definition 64 (Pairwise Correlation) *The pairwise correlation of two distributions with probability density functions $D_1, D_2 : \mathbb{R}^n \to \mathbb{R}_+$ with respect to a distribution with density $D : \mathbb{R}^n \to \mathbb{R}_+$, where the support of D contains the supports of D_1 and D_2, is defined as $\chi_D(D_1, D_2) := \int_{\mathbb{R}^n} D_1(\mathbf{x})D_2(\mathbf{x})/D(\mathbf{x})\,\mathrm{d}\mathbf{x} - 1$.*

Definition 65 *We say that a set of s distributions $\mathfrak{D} = \{D_1, \ldots, D_s\}$ over $\mathbb{R}^n$ is (γ,β)-correlated relative to a distribution D if $|\chi_D(D_i, D_j)| \le \gamma$ for all $i \ne j$, and $|\chi_D(D_i, D_i)| \le \beta$ for all i.*

Definition 66 (Statistical Query Dimension) *For $\beta, \gamma > 0$ and a decision problem $\mathcal{B}(\mathfrak{D}, D)$, where D is a fixed distribution and $\mathfrak{D}$ is a family of distributions, let s be the maximum integer such that there exists a finite set of distributions $\mathfrak{D}_D \subseteq \mathfrak{D}$ such that $\mathfrak{D}_D$ is (γ,β)-correlated relative to D and $|\mathfrak{D}_D| \ge s$. The* Statistical Query dimension *with pairwise correlations (γ,β) of $\mathcal{B}$ is defined to be s, and denoted by $\mathrm{SD}(\mathcal{B},\gamma,\beta)$.*

Lemma 67 (Corollary 3.12 of Feldman et al. (2017a)) *Let $\mathcal{B}(\mathfrak{D}, D)$ be a decision problem, where D is the reference distribution and $\mathfrak{D}$ is a class of distributions. For $\gamma, \beta > 0$, let $s = \mathrm{SD}(\mathcal{B}, \gamma, \beta)$. For any $\gamma' > 0$, any SQ algorithm for $\mathcal{B}$ requires queries of tolerance at most $\sqrt{\gamma + \gamma'}$ or makes at least $s\gamma'/(\beta - \gamma)$ queries.*

E.2. SQ Lower Bound on the Dependence on the Dimension and the Smoothing Parameter

Theorem 68 (SQ-Lower Bound) *Fix $d \in \mathbb{N}$. Let $\sigma \in (0,1)$ and $k \in \mathbb{N}$ such that $\sigma \leq O(1/\sqrt{\log k})$. Any SQ algorithm that learns the class of degree k PTFs in the smoothed agnostic setting (with respect to the uniform distribution on the hypercube) with any accuracy $\epsilon < 1/100$ either requires queries with tolerance at most $d^{-\Omega(k)}$ or makes at least $d^{\Omega(k)}$ queries.*

Proof

Given a subset $S \subseteq \{\pm 1\}^d$, we denote by $\chi_S(\mathbf{x})$ the parity function on the subset S, i.e., $\chi_S(\mathbf{x}) = \prod_{i \in S} \mathbf{x}_i$. We can extend the domain of χ_S to all of $\mathbb{R}^d$ as $\chi_S(\mathbf{x}) = \mathrm{sign}(\prod_{i \in S} \mathbf{x}_i)$ which is a degree $|S|$ Polynomial Threshold Function (PTF) and hence has surface area $C|S|$ for some universal constant $C > 0$. Observe that we have that $\chi_S \in \mathcal{F}(k, Ck)$.

For every subset S of $\{0,1\}^d$ of size m, we define the distribution D_S on $\{\pm 1\}^d \times \{\pm 1\}$ to be the distribution of the pair $(\mathbf{x}, \chi_S(\mathbf{x}))$ where $\mathbf{x} \sim U_d$ is drawn uniformly at random from the Boolean hypercube $U_d = \{\pm\}^d$ (we use U_d to denote both the d-dimensional Boolean hypercube and the uniform distribution over it). Moreover, we define N to be the distribution of $(\mathbf{x}, y)$ where $\mathbf{x}$ is drawn uniformly from the Boolean hypercube and y is ± 1 with probability $1/2$. We let $\mathcal{D}_k$ be the set of all distributions D_S for every subset $S \subseteq \{0,1\}^d$ of size at most k. We will show that given a learner for PTFs in the smoothed agnostic setting(under the uniform distribution on the hypercube), we can solve the decision problem $\mathcal{B}(\mathcal{D}_k, N)$. Since all parities are pairwise orthogonal, it is well known that the set of distributions $\mathcal{D}_k$ is $(0,1)$ correlated. Therefore, by Lemma 67 we obtain that any algorithm that solves the decision problem $\mathcal{B}$ either requires a query of tolerance $d^{-\Omega(k)}$ or makes at least $d^{\Omega(k)}$ queries (since the class $\mathcal{D}_k$ contains $\binom{d}{k}$ distributions).

We now show that using an algorithm $\mathcal{A}$ that learns a hypothesis $h(\cdot)$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \mathbf{E}_{\mathbf{z}\sim\mathcal{N}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[\chi_S(\mathbf{x} + \sigma\mathbf{z}) \neq y] + \epsilon$ for some $\epsilon \leq 1/100$ we can solve the k-parity decision problem $\mathcal{B}(\mathcal{D}_k, N)$ defined above. Let $h(\cdot)$ be the hypothesis returned by $\mathcal{A}$. We can perform a statistical query of tolerance $\tau = 1/10$ to obtain an estimate of the error of $h(\cdot)$, $q = \mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbb{1}\{h(\mathbf{x}) \neq y\}]$. If $q \leq 1/2 - 2\tau$ we declare that D corresponds to a k-parity, otherwise we declare that $D = N$. We observe that if D actually corresponds to a k-parity(with set S), then we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\chi_S(\mathbf{x} + \sigma\mathbf{z}) \neq y] &\leq \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}} \mathop{\mathbf{Pr}}_{\mathbf{x}\sim U_d}[\chi_S(\mathbf{x} + \sigma\mathbf{z}) \neq \chi_S(\mathbf{x})] \\ &\leq \max_{\mathbf{x}\in U_d} \mathop{\mathbf{Pr}}_{\mathbf{z}\sim\mathcal{N}}\left[\bigcup_{i=1}^{k} \left\{\sigma|\mathbf{w}_i \cdot \mathbf{z}| \geq 1/2\right\}\right],\end{aligned}$$

where the final inequality follows by the definition of χ_S:: for any $\mathbf{x} \in U_d$ we have that if $\mathrm{sign}(\mathbf{x}_i + \sigma\mathbf{z}_i) \leq 1/2$ for all i then $\chi_S(\mathbf{x}) = \chi_S(\mathbf{x} + \sigma\mathbf{z})$. Using the tail of the Gaussian density, we have that $\mathbf{Pr}_{\mathbf{z}\sim\mathcal{N}}\left[\bigcup_{i=1}^{k}\left\{\sigma|\mathbf{z}| \geq 1/2\right\}\right] \leq k\exp(-\Omega(1/(\sigma)^2)) \leq k\exp(-\Omega(1/\sigma)^2) \leq 1/10$ when $\sigma \leq O(1\sqrt{\log k})$. Therefore, by using a statistical query of tolerance $1/10$ and the learning algorithm $\mathcal{A}$ we can solve the k-parity decision problem. ■