Smoothed Analysis for Learning Concepts with Low Intrinsic Dimension

author names withheld

Editor: Under Review for COLT 2024

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 Gaussians (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 margins 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 learn efficiently 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 $\mathbb{R}^d \times \{\pm 1\}$ 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 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 $\mathbb{R}^d \times \{\pm 1\}$ 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 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:

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

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 $\mathbf{z}$):

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

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 $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[|T_\rho f_{\mathbf{x}}(\mathbf{z}) - f(\mathbf{z}+\mathbf{x})|\right] \leq 2\sqrt{\pi\rho}\cdot\Gamma$.

Choosing $\rho = O(\epsilon^2/\Gamma^2)$ makes this error at most $\epsilon/2$. We now approximate $T_\rho f_{\mathbf{x}}$ 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 $\mathbf{E}_{\mathbf{x}\sim D}\,\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[|p_{\mathbf{z}}(\mathbf{x}) - T_\rho f_{\mathbf{x}}(\mathbf{z})|\right] \leq \epsilon$.*

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 $\mathbf{E}_{\mathbf{x}\sim D}\,\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}\left[|T_\rho f_{\mathbf{x}}(\mathbf{z}) - p_{\mathbf{z}}(\mathbf{x})|\right] \leq \epsilon/2$ 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 $\mathbf{E}_{\mathbf{x}\sim D}\,\mathbf{E}_{\mathbf{z}\sim\mathcal{N}_k}\left[|p_{\mathbf{z}}(\mathbf{x}) - f(\mathbf{z}+\mathbf{x})|\right] \leq \epsilon$.

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 $q(\mathbf{x},\mathbf{s}) = q_m\left(-\frac{\|\mathbf{x}\|_2^2}{2\rho^2}+(\mathbf{x}/\rho)\cdot\mathbf{s}\right)$ 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

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

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

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

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 $\mathbf{E}_{x\sim\mathcal{N}(-\alpha^2/2,\alpha^2)}[|e^x - p(x)|] \leq \epsilon$.*

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 $|p(x) - e^x| \leq e^b \max(|a|, |b|)^m/m!$. 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 $\|\mathbf{v}\|_2 = 1$, $\mathbf{Pr}_{\mathbf{x}\sim D}[|\mathbf{x} \cdot \mathbf{v}| > t] \leq 2 \cdot e^{-(t/\lambda)^{1+\alpha}}$.*

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

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

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?

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.