Agnostic Proper Learning of Halfspaces under Gaussian Marginals

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin Madison
Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego
Vasilis Kontonis KONTONIS@WISC.EDU
University of Wisconsin Madison
Christos Tzamos TZAMOS@WISC.EDU
University of Wisconsin Madison
Nikos Zarifis ZARIFIS@WISC.EDU
University of Wisconsin Madison

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study the problem of agnostically learning halfspaces under the Gaussian distribution. Our main result is the *first proper* learning algorithm for this problem whose sample complexity and computational complexity qualitatively match those of the best known improper agnostic learner. Building on this result, we also obtain the first proper polynomial-time approximation scheme (PTAS) for agnostically learning homogeneous halfspaces. Our techniques naturally extend to agnostically learning linear models with respect to other non-linear activations, yielding in particular the first proper agnostic algorithm for ReLU regression.

Keywords: Agnostic Learning, Halfspaces, Proper Learning

1. Introduction

1.1. Background and Motivation

Halfspaces, or Linear Threshold Functions (LTFs), are Boolean functions $f : \mathbb{R}^d \to \{\pm 1\}$ of the form $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - t)$, for some $\mathbf{w} \in \mathbb{R}^d$ (known as the weight vector) and $t \in \mathbb{R}$ (known as the threshold). The function $\mathrm{sign} : \mathbb{R} \to \{\pm 1\}$ is defined as $\mathrm{sign}(u) = 1$ for $u \geq 0$ and $\mathrm{sign}(u) = -1$ otherwise. Halfspaces have arguably been *the* most extensively studied concept class in machine learning over the past six decades (Minsky and Papert, 1968; Shawe-Taylor and Cristianini, 2000). The problem of learning halfspaces (in various models) is as old as the field of machine learning, starting with the Perceptron algorithm (Rosenblatt, 1958; Novikoff, 1962), and has been one of the most influential problems in the field with techniques such as SVMs (Vapnik, 1998) and AdaBoost (Freund and Schapire, 1997) coming out of this study.

Here we study the task of learning halfspaces in the *agnostic framework* (Haussler, 1992; Kearns et al., 1994), which models the phenomenon of learning from adversarially labeled data. While halfspaces are efficiently learnable in the presence of consistently labeled examples (see, e.g., Maass and Turan (1994)) — i.e., in Valiant's original PAC model (Valiant, 1984) — even *weak* agnostic learning is computationally hard without distributional assumptions (Guruswami and Raghavendra, 2006; Feldman et al., 2006; Daniely, 2016). To circumvent this computational intractability, a line of work has focused on the *distribution-specific* agnostic PAC model — where the learner has a priori information about the distribution on examples. In this setting, computationally efficient noise-tolerant learning algorithms are known (Kalai et al., 2008; Klivans et al., 2009; Awasthi et al., 2017; Daniely, 2015; Diakonikolas et al., 2018, 2020d) with various time-accuracy tradeoffs.

Definition 1 (Distribution-Specific Agnostic Learning) *Let $\mathcal{C}$ be a class of Boolean-valued functions on $\mathbb{R}^d$. Given i.i.d. labeled examples $(\mathbf{x}, y)$ from a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{\pm 1\}$, such that the marginal distribution $\mathcal{D}_{\mathbf{x}}$ is promised to lie in a known distribution family $\mathcal{F}$ and no assumptions are made on the labels, the goal of the learner is to output a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ with small misclassification error, $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \overset{\mathrm{def}}{=} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$, as compared to the optimal misclassification error, $\mathrm{OPT} \overset{\mathrm{def}}{=} \inf_{g\in\mathcal{C}} \mathrm{err}_{0-1}^{\mathcal{D}}(g)$, by any function in the class.*

Throughout this paper, we will focus on the natural and well-studied case that the underlying distribution on examples is the standard multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Some additional comments are in order on Definition 1. In *improper* learning, the only assumption about the hypothesis h is that it is *polynomially evaluable*. In other words, we assume that $h \in \mathcal{H}$, where $\mathcal{H}$ is a (potentially complex) class of polynomially evaluable functions. In contrast, in *proper* learning we have the additional requirement that the hypothesis h is proper, i.e., $h \in \mathcal{C}$. These notions of learning are essentially equivalent in terms of sample complexity, but not always equivalent in terms of computational complexity. In particular, there exist concept classes that are efficiently improperly learnable, while proper learning is computationally hard.

The classical L_1-polynomial regression algorithm of Kalai et al. (2008) agnostically learns halfspaces under the Gaussian distribution, within error $\mathrm{OPT} + \epsilon$, with sample complexity and runtime of $d^{\mathrm{poly}(1/\epsilon)}$. On the lower bound side, recent work has provided evidence that this complexity cannot be improved. Specifically, Diakonikolas et al. (2020c); Goel et al. (2020); Diakonikolas et al. (2021) obtained Statistical Query (SQ) lower bounds of $d^{\mathrm{poly}(1/\epsilon)}$ for this problem. That is, the complexity of this learning problem is well-understood.

The polynomial regression algorithm Kalai et al. (2008) is the only known agnostic learner for halfspaces and is inherently *improper*: instead of a halfspace, its output hypothesis is a degree-k polynomial threshold function (PTF), i.e., the sign of a degree-k polynomial, where $k = \mathrm{poly}(1/\epsilon)$. For the corresponding proper learning problem, prior to the present work, no non-trivial computational upper bound was known.

Importance of Proper Learning. While an improper hypothesis suffices for the purpose of prediction, an improper learner comes with some disadvantages. In our context, having such a complex output hypothesis requires spending $d^{\mathrm{poly}(1/\epsilon)}$ time for even evaluating the hypothesis on a single example. Moreover, storing the hypothesis function requires keeping track of the $d^{\mathrm{poly}(1/\epsilon)}$ coefficients defining the corresponding polynomial. In contrast, a proper hypothesis is easy to interpret and provides the most succinct representation. Specifically, a halfspace hypothesis would require only $O(d)$ time for evaluation and $O(d)$ storage space. Even though it is known that $d^{\mathrm{poly}(1/\epsilon)}$ time is required for identifying a good hypothesis during training, prior to this work, it was not clear whether one can learn a succinct hypothesis that is more efficient at test time.

The preceding discussion motivates the following natural question:

Is there an efficient proper *agnostic learner for halfspaces under Gaussian marginals?*

The main result of this paper (Theorem 2) is the first agnostic proper learner for this problem whose complexity qualitatively matches that of the known improper learner (Kalai et al., 2008).

Faster Runtime via Approximate Learning. In view of the known SQ lower bounds for our problem (Diakonikolas et al., 2020c; Goel et al., 2020; Diakonikolas et al., 2021), it is unlikely that the $d^{\mathrm{poly}(1/\epsilon)}$ runtime for agnostically learning halfspaces can be improved, even under the Gaussian distribution. A line of work (Klivans et al., 2009; Awasthi et al., 2017; Daniely, 2015; Diakonikolas et al., 2018, 2020d) has focused on obtaining faster learning algorithms with relaxed error guarantees. Specifically, Awasthi et al. (2017) gave the first $\mathrm{poly}(d/\epsilon)$ time *constant-factor* approximation algorithm – i.e., an algorithm with misclassification error of $C \cdot \mathrm{OPT} + \epsilon$, for some universal constant $C > 1$ – for *homogeneous* halfspaces under the Gaussian, and, more generally, under any isotropic log-concave distribution. More recently, Daniely (2015) obtained a polynomial time approximation scheme (PTAS), i.e., an algorithm with error $(1 + \gamma) \cdot \mathrm{OPT} + \epsilon$ and runtime $d^{\mathrm{poly}(1/\gamma)}/\mathrm{poly}(\epsilon)$, under the uniform distribution on the sphere (and, effectively, under the Gaussian distribution).

Interestingly, the constant factor approximation algorithm of Awasthi et al. (2017) is proper. On the other hand, the PTAS of Daniely (2015) is inherently improper, in part because it relies on the combination of the localization method Awasthi et al. (2017) and the (improper) polynomial regression algorithm Kalai et al. (2008). It is thus natural to ask the following question:

Is there a proper *PTAS for agnostically learning halfspaces under Gaussian marginals?*

As our second main contribution (Theorem 3), we give such a proper PTAS qualitatively matching the complexity of the known improper PTAS (Daniely, 2015).

1.2. Our Contributions

In this paper, we initiate a systematic algorithmic investigation of proper learning in the agnostic distribution-specific PAC model. Our main result is the first proper agnostic learner for the class of halfspaces under the Gaussian distribution, whose sample complexity and runtime qualitatively match the performance of the previously known improper algorithm.

Theorem 2 (Proper Agnostic Learning of Halfspaces) *Let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. There exists an algorithm that, given $\epsilon, \delta > 0$, and $N = d^{O(1/\epsilon^4)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ i.i.d. samples from $\mathcal{D}$, the algorithm runs in time $\mathrm{poly}(N) + (1/\epsilon)^{O(1/\epsilon^6)}\log(1/\delta)$, and computes a halfspace hypothesis h such that, with probability at least $1 - \delta$, it holds $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq \mathrm{OPT} + \epsilon$.*

Theorem 2 gives the first non-trivial agnostic proper learner for the class of halfspaces under natural distributional assumptions. The runtime of our algorithm is $d^{\mathrm{poly}(1/\epsilon)}$, which *qualitatively* matches the complexity of the improper polynomial regression algorithm (Kalai et al., 2008) and is known to be qualitatively best possible in the SQ model (Diakonikolas et al., 2020c; Goel et al., 2020; Diakonikolas et al., 2021).

The analysis of Kalai et al. (2008) established an upper bound of $d^{O(1/\epsilon^4)}$ on the complexity of polynomial regression for our setting. This bound was later improved to $d^{O(1/\epsilon^2)}$, using optimal bounds on the underlying polynomial approximations (Diakonikolas et al., 2010b). Designing a proper learner that *quantitatively* matches this upper bound is left as an interesting open question.

Our second main contribution is the first proper polynomial-time approximation scheme (PTAS) for the agnostic learning problem. In our context, a PTAS is an algorithm that, for any $\gamma, \epsilon > 0$, runs in time $d^{\mathrm{poly}(1/\gamma)}/\mathrm{poly}(\epsilon)$ and outputs a hypothesis h satisfying $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$. The parameter $\gamma > 0$ quantifies the approximation ratio of the algorithm. Prior work (Daniely, 2015) gave an improper PTAS for agnostically learning *homogeneous* halfspaces, i.e., halfspaces whose separating hyperplane goes through the origin. We give a proper algorithm for this problem.

Theorem 3 (Proper PTAS for Agnostically Learning Halfspaces) *Let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. There exists an algorithm that, given $\gamma, \epsilon, \delta > 0$ and $N = d^{\mathrm{poly}(1/\gamma)} \mathrm{poly}(1/\epsilon) \log(1/\delta)$ i.i.d. samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and computes a halfspace h such that, with probability $1 - \delta$, it holds $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq (1 + \gamma)\mathrm{OPT} + \epsilon$, where OPT is the optimal misclassification error of any homogeneous halfspace.*

Theorem 3 gives the first proper PTAS for agnostically learning homogeneous halfspaces under any natural distributional assumptions and qualitatively matches the complexity of the improper PTAS by Daniely (2015). We note that the homogeneity assumption is needed for technical reasons and is also required in the known improper learning algorithm. Obtaining a PTAS for agnostically learning arbitrary halfspaces remains an open problem (even for improper learners).

Remark 4 (Extension to Other Non-Linear Activations) While the focus of the current paper is on the class of halfspaces, our algorithmic techniques are sufficiently robust and naturally generalize to other activation functions, i.e., functions of the form $f(\mathbf{x}) = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is a well-behaved activation function. Specifically, in Appendix D, we use our methods to develop the first proper agnostic learner for ReLU regression (Diakonikolas et al., 2020a).

Broader Context This work is the starting point of the broader research direction of designing *proper* agnostic learners in the distribution-specific setting for various expressive classes of Boolean functions. Here we make a first step in this direction for the class of halfspaces under the Gaussian distribution. The polynomial regression algorithm Kalai et al. (2008) is an improper agnostic learner that has been showed to succeed for broader classes of geometric functions, including degree-d PTFs (Diakonikolas et al., 2010a; Kane, 2011; Diakonikolas et al., 2014; Harsha et al., 2014), intersections of halfspaces (Kalai et al., 2008; Klivans et al., 2008; Kane, 2014), and broader families of convex sets (Klivans et al., 2008). An ambitious research goal is to develop a general methodology that yields proper agnostic learners for these concept classes under natural and broad distributional assumptions, matching the performance of polynomial regression.

1.3. Overview of Techniques

In this section, we provide a detailed overview of our algorithmic and structural ideas that lead to our proper learners.

Proper Agnostic Learning Algorithm The main idea behind our proper learning algorithm is to start with a good improper hypothesis and compress it down to a halfspace, while maintaining the same error guarantees. Our algorithm starts by computing the low-degree polynomial P that best approximates the labels in L_2-norm (Lemma 7). We then take a two-step approach to identify a near-optimal halfspace. First, by identifying the high-influence directions of P, we construct a low-dimensional subspace of $\mathbb{R}^d$ and show that it contains the normal vector to a near-optimal halfspace (Proposition 5). Then, we exhaustively search over vectors in this subspace (through an appropriately fine cover) and output the one with minimum error.

The main technical challenge comes in identifying such a subspace that is large enough to contain a good proper hypothesis, but also small enough so that exhaustive searching is efficient. To identify this subspace, we consider an appropriate matrix (defined by the high-influence directions of the polynomial P) and take the subspace defined by its large eigenvectors (Proposition 5). Exploiting the concentration guarantees of polynomials under the Gaussian distribution, we show that the resulting subspace is small enough to enumerate over (Lemma 6).

In more detail, we first find the polynomial $P(\mathbf{x})$ of degree $k = O(1/\epsilon^4)$ that approximates the labels y in the L_2 sense, that is, minimizes $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2]$. We then consider the influence of the polynomial P along a direction $\mathbf{u}$, $\mathrm{Inf}_{\mathbf{u}}(P) = \mathbf{u}^T\mathbf{M}\mathbf{u}$, for the matrix $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$, as a measure for how much the polynomial P changes along the direction $\mathbf{u}$. The key observation is that, along low-influence directions, the polynomial remains essentially constant and, as we show, the optimal halfspace must also be essentially constant as well. This allows us to prune down these directions and focus on a subspace of lower-dimension. Our main structural result (Proposition 5) formalizes this intuition showing that the subspace V of eigenvectors whose eigenvalues are larger than $\Theta(\epsilon^2)$ contains a normal vector $\mathbf{w}_V$ that (together with an appropriate threshold) achieves error $\mathrm{OPT} + \epsilon$.

Finally, while Proposition 5 establishes that we can remove directions of low-influence, we need to argue that the number of relevant eigenvectors is sufficiently small to simplify the problem. As we show in Lemma 6, the dimension of the resulting subspace V is $O(1/\epsilon^6)$, and thus finding a good hypothesis in this subspace takes time independent of the original dimension d. The key ingredient in bounding the dimension of V is to use concentration of polynomials under the Gaussian distribution to argue that the Frobenius norm of $\mathbf{M}$ is bounded, and thus the number of eigenvectors with large eigenvalues is bounded.

Proper PTAS for Agnostic Learning Our algorithm for obtaining a proper PTAS works in the same framework as Daniely (2015), who gave a non-proper PTAS for homogeneous halfspaces by combining the algorithm of Awasthi et al. (2017) with the L_1-polynomial regression algorithm of Kalai et al. (2008).

Similarly to the algorithm of Daniely (2015), we start by learning a halfspace (with normal vector) $\mathbf{w}_0$ with error $O(\mathrm{OPT})$, using any of the known constant factor approximations as a black-box (Awasthi et al., 2017; Diakonikolas et al., 2018, 2020d), and then partition the space according to the distance to the halfspace $\mathbf{w}_0$. Daniely's algorithm (Daniely, 2015) is based on the observation that points far from the true halfspace are accurately classified by the halfspace $\mathbf{w}_0$. Thus, one can use the improper learner of Kalai et al. (2008) to classify nearby points.

A simple adaptation of this idea would be to replace the improper algorithm of Kalai et al. (2008) with our new proper algorithm for agnostically learning halfspaces. There are two main complications however. First, the guarantees of our proper algorithm crucially rely on having Gaussian marginals, and therefore we cannot readily apply it once we restrict our attention only to points around $\mathbf{w}_0$. We deal with this issue by using a "soft" localization technique introduced in Diakonikolas et al. (2018) to randomly partition points in two groups. In particular, we perform rejection sampling according to a judiciously chosen weight function such that the distribution conditional on acceptance is still a Gaussian, albeit with very small variance along the direction of $\mathbf{w}_0$, see Lemma 14. By running our proper algorithm, we can obtain a halfspace $\mathbf{w}_1$ that is near-optimal under the conditional distribution.

The second obstacle is that while we can obtain two halfspaces ($\mathbf{w}_0$ and $\mathbf{w}_1$) that each are near-optimal for their corresponding groups, combining them into a *single* halfspace that works well for the entire distribution is not immediate. We remark that this is not an issue for the improper approximation scheme of Daniely (2015), since an improper learner is allowed to output a different classifier for different subsets of $\mathbb{R}^d$. To handle this issue, we additionally show that the halfspace $\mathbf{w}_1$ we obtain after localization will in fact perform well overall. In more detail, we show that the halfspace $\mathbf{w}_1$ cannot have very large angle with $\mathbf{w}_0$ and also its bias is small, see Proposition 15. Given these closeness properties, we can then show that the halfspace $\mathbf{w}_1$ achieves the desired error guarantees over the entire distribution, see Lemma 16.

2. Preliminaries

We will use small boldface characters for vectors and capital bold characters for matrices. For $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, $\mathbf{x}_i$ denotes the i-th coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 \stackrel{\mathrm{def}}{=} (\sum_{i=1}^d \mathbf{x}_i^2)^{1/2}$ denotes the ℓ_2-norm of $\mathbf{x}$. We will use $\mathbf{x} \cdot \mathbf{y}$ for the inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\theta(\mathbf{x}, \mathbf{y})$ for the angle between $\mathbf{x}, \mathbf{y}$. We will use $\mathbb{1}_A$ to denote the characteristic function of the set A, i.e., $\mathbb{1}_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $\mathbb{1}_A(\mathbf{x}) = 0$ if $\mathbf{x} \notin A$.

Let $\mathbf{e}_i$ be the i-th standard basis vector in $\mathbb{R}^d$. For $\mathbf{x} \in \mathbb{R}^d$ and $V \subseteq \mathbb{R}^d$, $\mathbf{x}_V$ denotes the projection of $\mathbf{x}$ onto the subspace V. Note that in the special case where V is spanned from one unit vector $\mathbf{v}$, then we simply write $\mathbf{x}_\mathbf{v}$ to denote $\mathbf{v}\ (\mathbf{x} \cdot \mathbf{v})$, i.e., the projection of $\mathbf{x}$ onto $\mathbf{v}$. For a subspace $U \subset \mathbb{R}^d$, let $U^\perp$ be the orthogonal complement of U. For a vector $\mathbf{w} \in \mathbb{R}^d$, we use $\mathbf{w}^\perp$ to denote the subspace spanned by vectors orthogonal to $\mathbf{w}$, i.e., $\mathbf{w}^\perp = \{\mathbf{u} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{u} = 0\}$. For a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathrm{tr}(\mathbf{A})$ denotes the trace of the matrix $\mathbf{A}$.

We use $\mathbf{E}_{x \sim \mathcal{D}}[x]$ for the expectation of the random variable x according to the distribution $\mathcal{D}$ and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. For simplicity of notation, we may omit the distribution when it is clear from the context. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the d-dimensional Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. For $(\mathbf{x}, y)$ distributed according to $\mathcal{D}$, we denote $\mathcal{D}_\mathbf{x}$ to be the distribution of $\mathbf{x}$. For unit vector $\mathbf{v} \in \mathbb{R}^d$, we denote $\mathcal{D}_\mathbf{v}$ the distribution of $\mathbf{x}$ on the direction $\mathbf{v}$, i.e., the distribution of $\mathbf{x}_\mathbf{v}$.

We use $\mathcal{C}_V$ for the set of Linear Threshold Functions (LTFs) with normal vector contained in $V \subseteq \mathbb{R}^d$, i.e., $\mathcal{C}_V = \{\mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t) : \mathbf{v} \in V, \|\mathbf{v}\|_2 = 1, t \in \mathbb{R}\}$; when $V = \mathbb{R}^d$, we simply write $\mathcal{C}$. Moreover, we define $\mathcal{C}_0$ to be the set of unbiased LTFs, i.e., $\mathcal{C}_0 = \{\mathrm{sign}(\mathbf{v} \cdot \mathbf{x}) : \mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\|_2 = 1\}$. We denote by $\mathcal{P}_k$ the space of polynomials on $\mathbb{R}^d$ of degree at most k.

3. Proper Agnostic Learning Algorithm

In this section, we present our proper agnostic learning algorithm for halfspaces, establishing Theorem 2. The pseudocode of our algorithm is given in Algorithm 1.

3.1. Analysis of Algorithm 1: Proof of Theorem 2

The main structural result that allows us to prove Theorem 2 is the following proposition, establishing the following: Given a multivariate polynomial P of degree $\Theta(1/\epsilon^4)$ that correlates well with the labels, we can use its *high-influence directions* to construct a subspace that contains a near-optimal halfspace. Specifically, we show:

Proposition 5 *Let $C > 0$ be a sufficiently large universal constant. Fix any $\epsilon \in (0, 1]$ and set $k = C/\epsilon^4$. Let $P(\mathbf{x}) \in \mathcal{P}_k$ be a degree-k polynomial such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + O(\epsilon^3)$. Moreover, let $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\top]$ and V be*

Algorithm 1 Agnostic Proper Learning Halfspaces

```
procedure AGNOSTIC-PROPER-LEARNER($\epsilon, \delta, \mathcal{D}$)
Input: $\epsilon > 0$, $\delta > 0$ and sample access to distribution $\mathcal{D}$
Output: A hypothesis $h \in \mathcal{C}$ such as $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq \min_{f\in\mathcal{C}} \mathrm{err}_{0-1}^{\mathcal{D}}(f) + \epsilon$ with probability $1-\delta$.
    $k \leftarrow C/\epsilon^4$, $\eta \leftarrow \epsilon^2/C$.        ▷ $C$ is a sufficiently large constant
    Find $P(\mathbf{x})$ such $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2] + O(\epsilon^3)$.
    Let $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$.
    Let $V$ be the subspace spanned by the eigenvectors of $\mathbf{M}$ whose eigenvalues are at least $\eta$.
    Construct an $\epsilon$-cover $\mathcal{H}$ of LTF hypotheses with normal vectors in $V$        ▷ see Fact 21.
    Draw $\Theta(\frac{1}{\epsilon^2}\log(|\mathcal{H}|/\delta))$ i.i.d. samples from $\mathcal{D}$ and construct the empirical distribution $\widehat{\mathcal{D}}$.
    $h \leftarrow \mathrm{argmin}_{h'\in\mathcal{H}} \mathrm{err}_{0-1}^{\widehat{\mathcal{D}}}(h')$
    return $h$.
```

the subspace spanned by the eigenvectors of $\mathbf{M}$ *with eigenvalues larger than* η*, where* $\eta = \epsilon^2/C$*. Then, for any* $f \in \mathcal{C}$*, it holds* $\min_{\mathbf{v}\in V, t\in\mathbb{R}} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - \mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t))y] \leq \epsilon$.

The proof of Proposition 5 is the bulk of the technical work of this section and is deferred to Section 3.2. In the body of this subsection, we show how to use Proposition 5 to establish Theorem 2.

The next lemma bounds from above the dimension of the subspace spanned by the high-influence directions of a degree-k polynomial that minimizes the L_2-error with the labels y.

Lemma 6 *Fix* $\epsilon > 0$ *and let* $P(\mathbf{x})$ *be a degree-*k *polynomial, with* $k = O(1/\epsilon^4)$*, such that* $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2] + O(\epsilon^3)$*. Let* $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$ *and* V *be the subspace spanned by the eigenvectors of* $\mathbf{M}$ *with eigenvalues larger than* η*. Then the dimension of the subspace* V *is* $\dim(V) = O(k/\eta)$*.*

Proof Let P be a polynomial such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2] + O(\epsilon^3)$ and let $P^* = \mathrm{argmin}_{P'\in\mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2]$. First, we note that $\mathbf{E}_{(x,y)\sim\mathcal{D}}[(y-P^*(\mathbf{x}))^2] \leq \mathbf{E}_{(x,y)\sim\mathcal{D}}[(y-0)^2] = 1$. Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[P(\mathbf{x})^2] \leq 5$.

Let V denote the subspace spanned by the eigenvectors of $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$ with eigenvalues at least η. We will show that $m = \dim(V) = O(k/\eta)$. We can write

$$m\,\eta \leq \mathrm{tr}\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]\right) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathrm{tr}(\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top)\right] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|\nabla P(\mathbf{x})\|_2^2\right] . \quad (1)$$

It is sufficient to show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\|\nabla P(\mathbf{x})\|_2^2] = O(k)$. By writing $P(\mathbf{x})$ in the Hermite basis, from Fact 19, it holds

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|\nabla P(\mathbf{x})\|_2^2\right] = \sum_{\alpha\in\mathbb{N}^d, |\alpha|\leq k} |\alpha| c_\alpha^2 \leq k \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[P(\mathbf{x})^2] \leq 5k\,. \quad (2)$$

Combining Equations (1) and (2), we obtain that $m = O(k/\eta)$, and the proof is complete. ■

For the proof of Theorem 2, we require a standard result on L_2-polynomial regression required to compute the polynomial of Proposition 5. The proof can be found on Appendix B.3.

Lemma 7 (L_2-Polynomial Regression) *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Let $k \in \mathbb{Z}_+$ and $\epsilon, \delta > 0$. There is an algorithm that draws $N = (dk)^{O(k)} \log(1/\delta)/\epsilon^2$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and outputs a polynomial $P(\mathbf{x})$ of degree at most k such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[(f(\mathbf{x}) - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[(f(\mathbf{x}) - P'(\mathbf{x}))^2] + \epsilon$, with probability $1 - \delta$.*

By running the L_2-regression algorithm of the above lemma, we obtain a polynomial P matching the requirements of our dimension-reduction result (Proposition 5). To complete the proof of Theorem 2, we perform SVD on the influence matrix $\mathbf{M}$ (see Proposition 5), and then create a sufficiently fine cover of the low-dimensional subspace V. The details and the full proof of Theorem 2 can be found in Appendix B.2.

3.2. Proof of Proposition 5

Suppose for the sake of contradiction that there exists a halfspace $f \in \mathcal{C}$ such that for every halfspace $f' \in \mathcal{C}_V$, it holds

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f'(\mathbf{x}))y] \geq \epsilon \,. \tag{3}$$

Our plan is to use the above fact in order to contradict the (approximate) optimality of the polynomial $P(\mathbf{x})$. To achieve this, we need to construct a polynomial $P''(\mathbf{x})$ with error *strictly less than* $\min_{P' \in \mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2]$. In the following simple claim, we show that in order to construct such a polynomial $P''(\mathbf{x})$, one needs to find a polynomial $Q(\mathbf{x})$ of degree at most k that correlates well with the difference $y - P(\mathbf{x})$, its proof can be found on Appendix B.1.

Claim 8 *It suffices to show that there exists a polynomial $Q(\mathbf{x})$ of degree at most k with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q^2(\mathbf{x})] \leq 9$ that $(\epsilon/4)$-correlates with $(y - P(\mathbf{x}))$, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[Q(\mathbf{x})(y - P(\mathbf{x}))] \geq \epsilon/4$.*

We now construct such a polynomial $Q(\mathbf{x})$. We can write that $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) = \mathrm{sign}(\mathbf{w}_V \cdot \mathbf{x} + \mathbf{w}_{V^\perp} \cdot \mathbf{x} + t)$. Note that $\mathbf{w}_{V^\perp} \neq \mathbf{0}$, since otherwise we would have $f \in \mathcal{C}_V$. For simplicity, we denote $\xi = \mathbf{w}_{V^\perp}/\|\mathbf{w}_{V^\perp}\|_2$. Notice that the direction ξ has low influence, since $\xi \in V^\perp$. Recall that by $\mathcal{D}_\xi$ we denote the projection of $\mathcal{D}$ onto the (one-dimensional) subspace spanned by ξ. We define $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z} + \mathbf{x}_V)]$ to be a convex combination of halfspaces in $\mathcal{C}_V$. In particular, $f_V(\mathbf{x})$ is a smoothed version of the halfspace $\mathrm{sign}(\mathbf{w}_V \cdot \mathbf{x} + t)$ whose normal vector belongs in V. Our argument consists of two main claims. In Lemma 9, we show that the function $f(\mathbf{x}) - f_V(\mathbf{x})$ correlates non-trivially with $y - P(\mathbf{x})$. Then we show that we can approximate $f(\mathbf{x}) - f_V(\mathbf{x})$ with a low-degree polynomial $Q(\mathbf{x})$ that maintains non-trivial correlation with $y - P(\mathbf{x})$; see Lemma 11. We start with the first lemma.

Lemma 9 *It holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(y - P(\mathbf{x}))] \geq \epsilon - 2\sqrt{\eta}$.*

Proof We have that $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z} + \mathbf{x}_{\xi^\perp})] = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[\mathrm{sign}(\mathbf{w}_V \cdot \mathbf{x}_V + \mathbf{w} \cdot \mathbf{z} + t)]$ and, since f_V is a convex combination of halfspaces in $\mathcal{C}_V$, from Equation (3), we see that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \geq \epsilon$. Thus, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(y - P(\mathbf{x}))] &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})] \\ &\geq \epsilon - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})] \,. \end{aligned} \tag{4}$$

To deal with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))P(\mathbf{x})]$, we first observe that for any function $g(\mathbf{x})$ depending only on the projection of $\mathbf{x}$ onto the subspace $\xi^{\perp}$, i.e., such that $g(\mathbf{x})=g(\mathbf{x}_{\xi^{\perp}})$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))g(\mathbf{x})]=\mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}_{\xi^{\perp}}}\left[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_{\xi}}[f(\mathbf{v}+\mathbf{z})-f_V(\mathbf{v})]\,g(\mathbf{v})\right]=0\,,$$

since for every $\mathbf{x}\in\mathbb{R}^d$ it holds $f_V(\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_{\xi}}[f(\mathbf{x}_{\xi^{\perp}}+\mathbf{z})]=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_{\xi}}[f(\mathbf{x}_V+\mathbf{z})]$. Unfortunately, we cannot directly do the above trick because $P(\mathbf{x})$ does not depend only on $\mathbf{x}_{\xi^{\perp}}$. However, since V contains the high-influence eigenvectors, it holds that P is almost a function of $\mathbf{x}_{\xi^{\perp}}$. In fact, we show that we can replace the polynomial P by a different polynomial of degree at most k that only depends on the projection of $\mathbf{x}$ on $\xi^{\perp}$. Similarly to the definition of the "smoothed" halfspace f_V, we define $R(\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_{\xi}}[P(\mathbf{x}_{\xi^{\perp}}+\mathbf{z})]$. We first prove that $R(\mathbf{x})$ is close to $P(\mathbf{x})$ in the L_2-sense.

Claim 10 *Let $R(\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_{\xi}}[P(\mathbf{x}_{\xi^{\perp}}+\mathbf{z})]$. It holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(P(\mathbf{x})-R(\mathbf{x}))^2]\leq\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\nabla P(\mathbf{x})\cdot\xi)^2]$.*

Proof We start by showing that without loss of generality we may assume that $\xi=\mathbf{e}_1$. Let $\mathbf{U}$ be an orthogonal matrix such that $\mathbf{U}\xi=\mathbf{e}_1$. Since $P(\mathbf{x})$ is a polynomial, we can apply the orthogonal transformation $\mathbf{U}$ to $\mathbf{x}$ and then use the Hermite basis to represent it, that is $P(\mathbf{x})=\sum_{\alpha\in\mathbb{N}^d}c_\alpha H_\alpha(\mathbf{U}^\top\mathbf{x})$. Our objective is equivalent to $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\nabla P(\mathbf{x})\cdot\xi)^2-(P(\mathbf{x})-R(\mathbf{x}))^2]\geq 0$. By the change of variables $\mathbf{x}\mapsto\mathbf{U}\mathbf{x}$, we have that $\mathbf{E}_{(\mathbf{U}\mathbf{x})\sim\mathcal{D}_{\mathbf{x}}}[(\nabla_{\mathbf{U}\mathbf{x}}P(\mathbf{U}\mathbf{x})\cdot(\mathbf{U}\xi))^2-(P(\mathbf{U}\mathbf{x})-R(\mathbf{U}\mathbf{x}))^2]\geq 0$, where we used the chain rule for the gradient. Observe that $R(\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_{\xi}}[P((\mathbf{I}-\xi\xi^\top)\mathbf{x}+\mathbf{z})]=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_{\xi}}\left[\sum_{\alpha\in\mathbb{N}^d}c_\alpha H_\alpha(\mathbf{U}^\top(\mathbf{I}-\xi\xi^\top)\mathbf{x}+\mathbf{U}^\top\mathbf{z})\right]$, therefore it holds that $R(\mathbf{U}\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_{\xi}}[\sum_{\alpha\in\mathbb{N}^d}c_\alpha H_\alpha(\mathbf{U}^\top(\mathbf{I}-\xi\xi^\top)\mathbf{U}\mathbf{x}+\mathbf{U}^\top\mathbf{z})]$. Moreover, $P(\mathbf{U}\mathbf{x})=\sum_{\alpha\in\mathbb{N}^d}c_\alpha H_\alpha(\mathbf{U}^\top\mathbf{U}\mathbf{x})$. Using the fact that $\mathbf{U}^\top\mathbf{U}=\mathbf{I}$ and $\mathbf{U}^\top\xi\xi^\top\mathbf{U}=\mathbf{e}_1\mathbf{e}_1^\top$, it follows that without loss of generality, we may assume that $\xi=\mathbf{e}_1$.

To keep notation simple, we write $P(\mathbf{x})=\sum_{\alpha\in\mathbb{N}^d}c_\alpha H_\alpha(\mathbf{x})$. Note that

$$P(\mathbf{x})-\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{D}_{\mathbf{e}_1}}[P(\mathbf{x}_{\xi^{\perp}}+\mathbf{x}_1)]=\sum_{\alpha\in\mathbb{N}^d}c_\alpha H_\alpha(\mathbf{x})-\sum_{\alpha\in\mathbb{N}^d}c_\alpha\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{D}_{\mathbf{x}_1}}[H_\alpha(\mathbf{x})]=\sum_{\alpha\in\mathcal{S}}c_\alpha H_\alpha(\mathbf{x})\ , \quad (5)$$

where $\mathcal{S}$ contains all the tuples for which the first index is non-zero, this follows from the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[H_\alpha(\mathbf{x})]=0$. Applying Parseval's identity yields

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(P(\mathbf{x})-\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{D}_{\mathbf{e}_1}}[P(\mathbf{x}_{\xi^{\perp}}+\mathbf{x}_1)])^2]=\sum_{\alpha\in\mathcal{S}}c_\alpha^2\,. \quad (6)$$

From Fact 19, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\nabla P(\mathbf{x})\cdot\mathbf{e}_1)^2]=\sum_{\alpha\in\mathbb{N}^d}\alpha_1 c_\alpha^2\geq\sum_{\alpha\in\mathcal{S}}c_\alpha^2\ , \quad (7)$$

where we used that $\alpha_1\geq 1$ on the set $\mathcal{S}$. Combining (6) and (7) completes the proof. ■

Adding and subtracting $R(\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_{\xi}}[P(\mathbf{x}_{\xi^{\perp}}+\mathbf{z})]$, we get

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))P(\mathbf{x})]=\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))(P(\mathbf{x})-R(\mathbf{x}_{\xi^{\perp}}))]+\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))R(\mathbf{x}_{\xi^{\perp}})]\,.$$

The second term is equal to zero, from the fact that $\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z}+\mathbf{x}_{\xi^\perp})-f_V(\mathbf{x}_{\xi^\perp})]=0$. From the Cauchy-Schwartz inequality, we get

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x}))(P(\mathbf{x})-R(\mathbf{x}_{\xi^\perp}))] &\le \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x}))^2]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(P(\mathbf{x})-R(\mathbf{x}_{\xi^\perp}))^2]}\\ &\le 2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(P(\mathbf{x})-R(\mathbf{x}_{\xi^\perp}))^2]}\le 2\sqrt{\eta}\;, \qquad (8)\end{aligned}$$

where we used Claim 10. Using Equation (4), we get that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-f_V(\mathbf{x}))(y-P(\mathbf{x}))]\ge \epsilon-2\sqrt{\eta}$. which completes the proof of Lemma 9. ■

Our final claim replaces $f-f_V$ by its polynomial approximation. By Hermite concentration arguments, we can show that we can use a polynomial $Q(\mathbf{x})$ of degree $O(1/\epsilon^4)$. We show:

Lemma 11 *There exists a polynomial $Q(\mathbf{x})$ of degree $O(1/\epsilon^4)$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q(\mathbf{x})(y-P(\mathbf{x}))]\ge \epsilon/2-2\sqrt{\eta}$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q^2(\mathbf{x})]\le 9$.*

Proof We will require the following result from Klivans et al. (2008) which bounds the Hermite concentration of LTFs.

Fact 12 (Theorem 15 of Klivans et al. (2008)) *Let $f\in\mathcal{C}$, and let S be the Hermite expansion up to degree k of f, i.e., $S(\mathbf{x})=\sum_{|\alpha|\le k}\hat{f}(\alpha)H_\alpha(\mathbf{x})$. Then $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(S(\mathbf{x})-f(\mathbf{x}))^2]=O(1/\sqrt{k})$.*

For any polynomial $Q(\mathbf{x})$, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q(\mathbf{x})(y-P(\mathbf{x}))] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x})+(f(\mathbf{x})-f_V(\mathbf{x}))-(f(\mathbf{x})-f_V(\mathbf{x})))(y-P(\mathbf{x}))]\\ &\ge \epsilon-2\sqrt{\eta}+\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x})-(f(\mathbf{x})-f_V(\mathbf{x})))(y-P(\mathbf{x}))]\;, \qquad (9)\end{aligned}$$

where we used Lemma 9. By choosing $Q(\mathbf{x})=S(\mathbf{x})-\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp}+\mathbf{z})]$, where we denote by $S(\mathbf{x})$ the Hermite expansion of f truncated up to degree k, $S(\mathbf{x})=\sum_{|\alpha|\le k}\hat{f}(\alpha)H_\alpha(\mathbf{x})$, we will show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x})-Q(\mathbf{x}))^2]=O(1/\sqrt{k})$. Using the elementary inequality $(a+b)^2\le 2a^2+2b^2$, we get that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x})-Q(\mathbf{x}))^2]\le 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-S(\mathbf{x}))^2]+2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f_V(\mathbf{x})-\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp})])^2]\;.$$

Moreover, by Jensen's inequality, it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f_V(\mathbf{x})-\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp}+\mathbf{z})])^2]\le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-S(\mathbf{x}))^2]=O(1/\sqrt{k})\;,$$

where in the last equality we used Fact 12. Note that from the reverse triangle inequality, it holds that

$$\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[Q^2(\mathbf{x})]}\le \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x}))^2]}+O(1/k^{1/4})\le 2+O(1/k^{1/4})\;. \qquad (10)$$

Choosing $k=O(1/\epsilon^4)$ and applying Cauchy-Schwartz to the Equation (9), we get

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[Q(\mathbf{x})(y-P(\mathbf{x}))] &\ge \epsilon-2\sqrt{\eta}-\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x})-(f(\mathbf{x})-f_V(\mathbf{x})))^2]\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2]}\\ &\ge \epsilon/2-2\sqrt{\eta}\;,\end{aligned}$$

where we used the fact that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \leq \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-0)^2] \leq 1$; the polynomial $P(\mathbf{x})$ is closer to y than the trivial polynomial 0. For this choice of k, Equation (10) gives $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[Q^2(\mathbf{x})] \leq 9$. This completes the proof of Lemma 11. ■

By choosing $\eta = \epsilon^2/64$, Lemma 11 contradicts our assumption that $P(\mathbf{x})$ is $O(\epsilon^3)$-close to the polynomial $P'(\mathbf{x})$ that minimizes the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2]$, see Claim 8. This completes the proof of Proposition 5.

4. Agnostic Proper PTAS for Homogeneous Halfspaces

In this section, we provide a proper PTAS for agnostically learning homogeneous halfspaces, thereby establishing Theorem 3. Concretely, let $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$ be an optimal halfspace, i.e., $\mathrm{OPT} = \min_{h\in\mathcal{C}_0} \mathrm{err}_{0-1}^{\mathcal{D}}(h) = \mathrm{err}_{0-1}^{\mathcal{D}}(f)$. For any $\gamma, \epsilon \in (0,1)$ our algorithm computes a halfspace $h(\mathbf{x})$ such that $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$. The pseudocode of our algorithm is given in Algorithm 2.

Algorithm 2 Agnostic Proper PTAS for Homogeneous Halfspaces

```
procedure AGNOSTIC-PROPER-PTAS($\gamma, \epsilon, \delta, \mathcal{D}$)        ▷ $C, C'$ are absolute constants
Input: $\gamma > 0$, $\epsilon, \delta > 0$, and distribution $\mathcal{D}$
Output: A hypothesis $h \in \mathcal{C}$ such that $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$ with probability $1-\delta$
    $\sigma \leftarrow C'\,\mathrm{OPT}/\gamma$
    Let $\mathbf{w}_0$ be the normal vector of the homogeneous halfspace computed using Lemma 13
    If $\epsilon > C\,\mathrm{OPT}$:
        return $h_0(\mathbf{x}) = \mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x})$
    Let $\mathcal{D}_A$ be the distribution $\mathcal{D}$ after applying rejection sampling with $\mathbf{w}_0$ and $\sigma$ (Lemma 14)
    Run Algorithm 1 on $\mathcal{D}_A$ with accuracy $\Theta(\gamma^2)$ and confidence $\delta$ to get $(\mathbf{w}, t)$
    return $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t)$
```

4.1. Analysis of Algorithm 2: Proof of Theorem 3

The following lemma provides us with an efficient algorithm that learns a halfspace within error $O(\mathrm{OPT})$ in polynomial time. This halfspace serves as the initialization of Algorithm 2: we will use it to perform localization around it.

Lemma 13 (Awasthi et al. (2017); Diakonikolas et al. (2020d)) *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. There is an algorithm that draws $N = O((d/\epsilon^4)\log(1/\delta))$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and outputs a hypothesis $h \in \mathcal{C}$ such that, with probability at least $1-\delta$, we have $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq O(\mathrm{OPT}) + \epsilon$.*

The following lemma provides a "soft" localization procedure. Instead of performing rejection sampling inside a band around $\mathbf{w}_0$, i.e., $|\mathbf{x} \cdot \mathbf{w}_0| < \sigma$, we perform rejection sampling with weight $e^{-\mathbf{w}_0 \cdot \mathbf{x}(\sigma^{-2}-1)}$: samples that are far from the halfspace $\mathbf{w}_0$ are accepted with very small probability. Using this rejection sampling process, we get that the distribution conditional on acceptance is a normal distribution. This allows us to use our proper learning algorithm of Section 3.

Lemma 14 (Lemma 4.7 of Diakonikolas et al. (2018)) *Let $\mathbf{w}_0 \in \mathbb{R}^d$ be a unit vector and let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Fix $\sigma \in (0,1)$ and define the distribution*

$\mathcal{D}_A$ as follows: draw a sample $(\mathbf{x}, y)$ from $\mathcal{D}$ and accept it with probability $e^{-(\mathbf{w}_0\cdot\mathbf{x})^2(\sigma^{-2}-1)/2}$. $\mathcal{D}_A$ is the distribution of $(\mathbf{x}, y)$ conditional on acceptance. The $\mathbf{x}$-marginal of $\mathcal{D}_A$ is $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{I} - (1-\sigma^2)\mathbf{w}_0\mathbf{w}_0^\top$, and the probability that some point will be accepted is σ.

The main technical tool of this section is the following proposition. It shows that, if a halfspace performs reasonably well with respect to the "localized" distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ of Lemma 14, then it cannot be very biased or have very large angle with the initial guess $\mathbf{w}_0$. This allows us to prove that the halfspace that we find using the "localized" distribution will perform well over the initial Gaussian, $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Proposition 15 *Let $\mathbf{w}_0 \in \mathbb{R}^d$ be a unit vector and let $\alpha, \gamma \in (0, 1/4]$. Let $\mathcal{D}_A$ be defined as in Lemma 14, i.e., its $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{I} - (1-\sigma^2)\mathbf{w}_0\mathbf{w}_0^T$ for some $\sigma \in (0, \cos(\pi\alpha))$. Moreover, assume that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x}) \neq y] \leq \alpha/3$. There exists an algorithm that runs in time $d^{\mathrm{poly}(1/(\gamma\alpha))}\log(1/\delta)$ and with probability at least $1-\delta$ returns a halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} + t)$ such that $|t| = O(\sigma\alpha)$, $\theta(\mathbf{w}, \mathbf{w}_0) = O(\sigma\alpha)$. Moreover, it holds*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq y] \leq \min_{\bar{h}\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\bar{h}(\mathbf{x}) \neq y] + \alpha\gamma\,.$$

The proof of Proposition is quite technical and is deferred to Section 4.2. The following lemma is similar to the localization lemma (Lemma 2.1) given in Daniely (2015). We need to adapt it to work in our setting, where we use a soft localization procedure (see Lemma 14), as opposed to a hard one. Its proof can be found on Appendix C.1.

Lemma 16 (Gaussian Localization) *Let $R(\mathbf{x})$ be the event that the sample $\mathbf{x}$ is rejected from the rejection sampling procedure of Lemma 14 with vector $\mathbf{w}_0$ and $\sigma = \Theta\left(\frac{\mathrm{OPT}}{\alpha}\right)$. Let $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x})$, $h'(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} + t)$ be halfspaces with $t = O(\sigma\alpha)$ and $\theta(\mathbf{w}_0, \mathbf{w}) = O(\sigma\alpha)$. Then, $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), R(\mathbf{x})] = O(\alpha\,\mathrm{OPT})$.*

We are now ready to prove Theorem 3.

Proof [Proof of Theorem 3] Our analysis follows the cases of Algorithm 2. Initially, Algorithm 2 computes $h_0 = \mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x})$, using the algorithm of Lemma 13. From Lemma 13, we have that for this halfspace it holds, with probability at least $1-\delta/2$, that $\mathrm{err}_{0-1}^{\mathcal{D}}(h_0) = C\,\mathrm{OPT} + \epsilon$ for some absolute constant $C > 1$. The runtime of this step is $\mathrm{poly}(d, 1/\epsilon)\log(1/\delta)$. Therefore, when $\epsilon > 2C\,\mathrm{OPT}$, we directly get that $\mathrm{err}_{0-1}^{\mathcal{D}}(h_0) \leq \mathrm{OPT} + \epsilon$.

For the case when $\epsilon \leq 2C\,\mathrm{OPT}$, Algorithm 2 returns a halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)$, where $\mathbf{w} \in \mathbb{R}^d$ and $t \in \mathbb{R}$. We show that for this hypothesis h it holds $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$. Let $\mathcal{D}_A$ be the distribution conditional on acceptance (see Lemma 14) with parameters $\mathbf{w}_0$ (the normal vector of the halfspace h_0) and $\sigma = \Theta(\mathrm{OPT}/\alpha)$, for some parameter $\alpha = \Theta(\gamma)$ and $R(\mathbf{x})$ (resp. $A(\mathbf{x})$) be the event that the sample $\mathbf{x}$ is rejected (resp. accepted). The error of the hypothesis h is

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y, A(\mathbf{x})] + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y, R(\mathbf{x})]\,. \tag{11}$$

Let us first bound the $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y, A(\mathbf{x})] = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq y]\,\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[A(\mathbf{x})]$. From Proposition 15, we have that with sample complexity and runtime $d^{\mathrm{poly}(1/\gamma)}\log(1/\delta)$ we get that, with probability at least $1-\delta/2$, it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq y] \leq \min_{\bar{h}\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\bar{h}(\mathbf{x}) \neq y] + \alpha\gamma \leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[f(\mathbf{x}) \neq y] + \alpha\gamma\,,$$

and $|t| = O(\sigma\alpha)$, $\theta(\mathbf{w}, \mathbf{w}_0) = O(\sigma\alpha)$. From Lemma 14, it holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[A(\mathbf{x})] = \sigma$, thus

$$\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y, A(\mathbf{x})] \leq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[f(\mathbf{x}) \neq y, A(\mathbf{x})] + \alpha\gamma\sigma \; . \tag{12}$$

To bound $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y, R(\mathbf{x})]$, observe that $\theta(\mathbf{w}_0, \mathbf{w}^*) = O(\mathrm{OPT})$, because $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h_0(\mathbf{x}) \neq f(\mathbf{x})] = \theta(\mathbf{w}_0, \mathbf{w}^*)/\pi = O(\mathrm{OPT})$. Thus, with two applications of Lemma 16, we get $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq h_0(\mathbf{x}), R(\mathbf{x})] = O(\alpha\mathrm{OPT})$ and $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq h_0(\mathbf{x}), R(\mathbf{x})] = O(\alpha\mathrm{OPT})$. Using the triangle inequality, we get

$$\begin{aligned}\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y, R(\mathbf{x})] &\leq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y, R(\mathbf{x})] + \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq h_0(\mathbf{x}), R(\mathbf{x})] \\ &+ \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq h_0(\mathbf{x}), R(\mathbf{x})] = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y, R(\mathbf{x})] + O(\alpha\mathrm{OPT}) \; . \end{aligned} \tag{13}$$

Substituting Equations (12) and (13) into Equation (11), we get

$$\begin{aligned}\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] &= \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[f(\mathbf{x}) \neq y, A(\mathbf{x})] + \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y, R(\mathbf{x})] + O(\alpha\mathrm{OPT} + \alpha\gamma\sigma) \\ &= \mathrm{OPT} + O((\alpha+\gamma))\mathrm{OPT} = (1+O(\gamma))\mathrm{OPT} \; ,\end{aligned}$$

where we used the fact that $\alpha = \Theta(\gamma)$. Combining the above cases, we obtain that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq (1+O(\gamma))\mathrm{OPT} + \epsilon$. Combining the runtime of the above two steps, we obtain that the total runtime of our algorithm is $d^{\mathrm{poly}(1/\gamma)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$. ■

4.2. Proof of Proposition 15

We first make the $\mathbf{x}$-marginal isotropic by multiplying samples with $\mathbf{\Sigma}^{-1/2}$ and then use the proper learning algorithm of Theorem 2 with target accuracy $\epsilon = \alpha\gamma$. The sample complexity and runtime are thus $d^{\mathrm{poly}(1/(\alpha\gamma))}$. From the guarantee of Theorem 2, we immediately obtain that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq y] \leq \min_{\bar{h}\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\bar{h}(\mathbf{x}) \neq y] + \alpha\gamma/3$. It now remains to bound the bias t and the angle $\theta(\mathbf{w}, \mathbf{w}_0)$ of the returned halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)$. Using our assumption for the misclassification error of $\mathbf{w}_0$ with respect to $\mathcal{D}_A$, we obtain that $\min_{\bar{h}\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x}) \neq y] \leq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x}) \neq y] \leq \alpha$. From the triangle inequality, we obtain that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x})] \leq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq y] + \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x}) \neq y] \leq \alpha$. Therefore, we have

$$\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x})] = \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h(\mathbf{\Sigma}^{1/2}\mathbf{x}) \neq \mathrm{sign}((\mathbf{\Sigma}^{1/2}\mathbf{w}_0)\cdot\mathbf{x})] \, .$$

If the halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)$ has zero bias, i.e., $t = 0$, we have that by the spherical symmetry of the Gaussian distribution it holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h(\mathbf{\Sigma}^{1/2}\mathbf{x}) \neq \mathrm{sign}((\mathbf{\Sigma}^{1/2}\mathbf{w}_0)\cdot\mathbf{x})] = \theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0)/\pi$. Unfortunately, the same is not true when one of the halfspaces has non-zero bias. However, we can prove that the angle $\theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0)/\pi$ is still a *lower bound* on the probability of disagreement, i.e., $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h(\mathbf{\Sigma}^{1/2}\mathbf{x}) \neq \mathrm{sign}((\mathbf{\Sigma}^{1/2}\mathbf{w}_0)\cdot\mathbf{x})]$.

Formally, we prove the following claim showing that when one of the halfspaces is homogeneous, the probability mass of the disagreement region is at least a constant multiple of the angle between the normal vectors. The proof follows from the observation that we can always minimize the disagreement probability between a homogeneous and an arbitrary halfspace by centering the Gaussian exactly at their intersection point. We provide the detailed proof in Appendix C.

Claim 17 *For $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d, t \in \mathbb{R}$ define the halfspaces $h_0(\mathbf{x}) = \mathrm{sign}(\mathbf{u} \cdot \mathbf{x})$, $h_1(\mathbf{x}) = \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t)$. It holds $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[h_1(\mathbf{x}) \neq h_0(\mathbf{x})] \geq \theta(\mathbf{u}, \mathbf{v})/\pi$.*

Using Claim 17 and the fact that $\mathbf{Pr}_{(\mathbf{x},y) \sim \mathcal{D}_A}[h(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x})] \leq \alpha$ that we showed above, we have that $\theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0) \leq \pi\alpha$. We have

$$\cos(\theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0)) = \frac{\mathbf{w} \cdot (\mathbf{\Sigma}\mathbf{w}_0)}{\sqrt{\mathbf{w} \cdot (\mathbf{\Sigma}\mathbf{w})}\sqrt{\mathbf{w}_0 \cdot (\mathbf{\Sigma}\mathbf{w}_0)}} = \frac{\sigma\, \mathbf{w} \cdot \mathbf{w}_0}{\sqrt{1 - (1 - \sigma^2)(\mathbf{w} \cdot \mathbf{w}_0)^2}} .$$

Since $\theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0) \leq \pi\alpha$ and cosine is a decreasing function in $[0, \pi]$, we obtain that $\sigma\, \mathbf{w} \cdot \mathbf{w}_0 \geq \cos(\pi\alpha)\sqrt{1 - (1 - \sigma^2)(\mathbf{w} \cdot \mathbf{w}_0)^2}$. Solving this quadratic inequality with respect to $\mathbf{w} \cdot \mathbf{w}_0$, we obtain

$$\mathbf{w} \cdot \mathbf{w}_0 \geq \sqrt{\frac{1}{1 + \sigma^2(\frac{1}{\cos^2(\pi\alpha)} - 1)}} = \sqrt{\frac{1}{1 + \sigma^2 \tan^2(\pi\alpha)}} . \tag{14}$$

Using the inequality $\cos^{-1}(\sqrt{1/(1+x)}) \leq \sqrt{x}$ that holds for every $x \geq 0$, we obtain that the angle $\theta(\mathbf{w}, \mathbf{w}_0) \leq \sigma \tan(\pi\alpha)$. Using the elementary inequality $\tan(\pi x) \leq 4x$ that holds for all $x \in [0, 1/4]$, we can further simplify the bound for the angle to $\theta(\mathbf{w}, \mathbf{w}_0) \leq 4\sigma\alpha = O(\sigma\alpha)$.

We next bound the bias of the returned halfspace h. Now that we know that the angle between the vectors $\mathbf{w}_0, \mathbf{w}$ is small, we can use the following lower bound on the disagreement between two halfspaces to get that the bias cannot be too large. We provide the proof of the following claim in Appendix C.

Claim 18 *Let $h_1(\mathbf{x}) = \mathrm{sign}(\mathbf{u} \cdot \mathbf{x} + t_1)$, $h_2(\mathbf{x}) = \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t_2)$ be two halfspaces. Let $r_1 = t_1/\left\|\mathbf{u}\mathbf{\Sigma}^{1/2}\right\|_2$, $r_2 = t_2/\left\|\mathbf{v}\mathbf{\Sigma}^{1/2}\right\|_2$. It holds $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \mathbf{Pr}_{r \sim \mathcal{N}(0,1)}[\min(r_1, r_2) \leq r \leq \max(r_1, r_2)]$.*

From Claim 18 and the fact that $\mathbf{Pr}_{(\mathbf{x},y) \sim \mathcal{D}_A}[h(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x})] \leq \alpha$, we obtain that $\mathbf{Pr}_{r \sim \mathcal{N}(0,1)}[0 \leq r \leq |t_1|/\left\|\mathbf{w}\mathbf{\Sigma}^{1/2}\right\|_2] \leq \alpha$. Using the anti-anti-concentration property of the univariate Gaussian distribution, i.e., that $\mathbf{Pr}_{r \sim \mathcal{N}(0,1)}[0 \leq r \leq t] \geq \min(t/2, 2/3)$ and the fact that $\alpha \leq 1/4$, we obtain that $|t_1|/\left\|\mathbf{w}\mathbf{\Sigma}^{1/2}\right\|_2 \leq \alpha$. From Equation (14), we obtain that

$$\left\|\mathbf{w}\mathbf{\Sigma}^{1/2}\right\|_2 = \sqrt{1 - (1 - \sigma^2)(\mathbf{w} \cdot \mathbf{w}_0)^2} \leq \sqrt{\sigma^2 \frac{1 + \tan^2(\pi a)}{1 + \sigma^2 \tan^2(\pi\alpha)}} \leq 2\sigma ,$$

using the fact that $\sigma < 1$ and $\alpha < 1/4$. Therefore, we conclude that $|t_1| \leq 2\sigma\alpha = O(\sigma\alpha)$. This concludes the proof of Proposition 15.

References

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.

A. Daniely. A PTAS for agnostically learning halfspaces. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 484–502, 2015.

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.

I. Diakonikolas, P. Harsha, A. Klivans, R. Meka, P. Raghavendra, R. A. Servedio, and L. Y. Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *STOC*, pages 533–542, 2010a.

I. Diakonikolas, D. M. Kane, and J. Nelson. Bounded independence fools degree-2 threshold functions. In *FOCS*, pages 11–20, 2010b.

I. Diakonikolas, P. Raghavendra, R. A. Servedio, and L. Y. Tan. Average sensitivity and noise sensitivity of polynomial threshold functions. *SIAM J. Comput.*, 43(1):231–253, 2014.

I. Diakonikolas, D. M. Kane, and A. Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1061–1073, 2018.

I. Diakonikolas, S. Goel, S. Karmalkar, A. R. Klivans, and M. Soltanolkotabi. Approximation schemes for ReLU regression. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 1452–1485. PMLR, 2020a.

I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis. Algorithms and SQ lower bounds for PAC learning one-hidden-layer ReLU networks. In *Conference on Learning Theory, COLT 2020*, pages 1514–1539. PMLR, 2020b.

I. Diakonikolas, D. M. Kane, and N. Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under gaussian marginals. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020c.

I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Non-convex SGD learns halfspaces with adversarial label noise. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020d.

I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021.

V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami. New results for learning noisy parities and halfspaces. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 563–574. IEEE Computer Society, 2006.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

S. Goel, A. Gollakota, and A. R. Klivans. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

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. Harsha, A. R. Klivans, and R. Meka. Bounding the sensitivity of polynomial threshold functions. *Theory of Computing*, 10:1–26, 2014.

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

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.

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

D. M. Kane. The average sensitivity of an intersection of half spaces. In *Symposium on Theory of Computing, STOC 2014*, pages 437–440, 2014.

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

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.

A. Klivans, P. Long, and R. Servedio. Learning Halfspaces with Malicious Noise. *Journal of Machine Learning Research*, 10:2715–2740, 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.

W. Maass and G. Turan. How fast can a threshold gate learn? In S. Hanson, G. Drastal, and R. Rivest, editors, *Computational Learning Theory and Natural Learning Systems*, pages 381–414. MIT Press, 1994.

M. Minsky and S. Papert. *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, MA, 1968.

A. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. ISBN 978-1-10-703832-5.

F. Rosenblatt. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

J. Shawe-Taylor and N. Cristianini. *An introduction to support vector machines*. Cambridge University Press, 2000.

G. Szegö. *Orthogonal Polynomials*. Number τ. 23 in American Mathematical Society colloquium publications. American Mathematical Society, 1967. ISBN 9780821889527. URL https://books.google.com/books?id=3hcW8HBh7gsC.

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

V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Appendix A. Hermite Polynomials

We are also going to use the Hermite polynomials that form an orthonormal system with respect to the Gaussian measure. We denote by $L^2(\mathbb{R}^d, \mathcal{N}(\mathbf{0}, \mathbf{I}))$ the vector space of all functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f^2(\mathbf{x})] < \infty$. The standard inner product for this space is $f \cdot g := \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x})g(\mathbf{x})]$. While usually one considers the probabilists' or physicists' Hermite polynomials, in this work we define the *normalized* Hermite polynomial of degree i to be $H_0(x) = 1, H_1(x) = x, H_2(x) = \frac{x^2-1}{\sqrt{2}}, \ldots, H_i(x) = \frac{He_i(x)}{\sqrt{i!}}, \ldots$ where by $He_i(x)$ we denote the probabilists' Hermite polynomial of degree i. These normalized Hermite polynomials form a complete orthonormal basis for the single-dimensional version of the inner product space defined above. To get an orthonormal basis for $L^2(\mathbb{R}^d, \mathcal{N}(\mathbf{0}, \mathbf{I}))$, we use a multi-index $\alpha \in \mathbb{N}^d$ to define the d-variate normalized Hermite polynomial as $H_\alpha(\mathbf{x}) = \prod_{i=1}^d H_{\alpha_i}(\mathbf{x}_i)$. The total degree of H_α is $|\alpha| := \sum_i \alpha_i$. Given a function $f \in L^2(\mathbb{R}^d, \mathcal{N}(\mathbf{0}, \mathbf{I}))$, we compute its Hermite coefficients as $\hat{f}(\alpha) = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x})H_\alpha(\mathbf{x})]$ and express it uniquely as $\sum_{\alpha\in\mathbb{N}^d} \hat{f}(\alpha)H_\alpha(\mathbf{x})$. For more details on the Gaussian space and Hermite analysis, we refer the reader to O'Donnell (2014). Most of the facts about Hermite polynomials that we use in this work are well-known properties and can be found, for example, in Szegö (1967). We are going to use the following simple fact about the gradient of Hermite polynomials; for a proof see, for example, Lemma 6 of Kontonis et al. (2019).

Fact 19 *Let* $P(\mathbf{x}) = \sum_{\substack{\alpha\in\mathbb{N}^d\\|\alpha|\le k}} c_\alpha H_\alpha(\mathbf{x})$ *be a k-degree Hermite polynomial. It holds* $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(\nabla P(\mathbf{x}) \cdot \mathbf{e}_i)^2] = \sum_{\substack{\alpha\in\mathbb{N}^d\\|\alpha|\le k}} \alpha_i c_\alpha^2$.

Appendix B. Omitted Proofs from Section 3

B.1. Details of proof of Proposition 5

We restate and prove the following claim:

Claim 20 *It suffices to show that there exists a polynomial $Q(\mathbf{x})$ of degree at most k with* $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q^2(\mathbf{x})] \le 9$ *that $(\epsilon/4)$-correlates with $(y - P(\mathbf{x}))$, i.e.,* $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[Q(\mathbf{x})(y - P(\mathbf{x}))] \ge \epsilon/4$.

Proof Given such a polynomial $Q(\mathbf{x})$, we consider the polynomial $P''(\mathbf{x}) = P(\mathbf{x}) + \zeta Q(\mathbf{x})$, for $\zeta = c\,\epsilon$ and c a sufficiently small constant. Observe that $P''(\mathbf{x})$ has degree at most k and decreases the value of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2]$ by at least $\Omega(\epsilon^2)$, which contradicts the optimality of $P(\mathbf{x})$, i.e., that $P(\mathbf{x})$ is $O(\epsilon^3)$-close to the polynomial that minimizes the L_2-error with y. ■

B.2. Proof of Theorem 2

We require the following standard fact showing the existence of a small ϵ-cover $\widetilde{V}$ of the set V, i.e., a set $\widetilde{V}$ such that for any $\mathbf{v} \in V$ there exists $\tilde{\mathbf{v}} \in \widetilde{V}$ such that $\|\mathbf{v} - \tilde{\mathbf{v}}\|_2 \leq \epsilon$.

Fact 21 (see, e.g., Corollary 4.2.13 of Vershynin (2018)) *For any $\epsilon > 0$, there exists an explicit ϵ-cover of the unit ball in $\mathbb{R}^k$, with respect to the ℓ_2-norm, of size $O(1/\epsilon)^k$.*

In order to create an effective discretization of the hypotheses, we need the following fact.

Fact 22 *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ be unit vectors and $t_1, t_2 \in \mathbb{R}$. Then the following holds:*

1. $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_1) - \mathrm{sign}(\mathbf{v}\cdot\mathbf{x} + t_1)|] = O(\|\mathbf{u} - \mathbf{v}\|_2)$,
2. $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_1) - \mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_2)|] = O(|t_1 - t_2|)$ *and,*
3. *if $|t| > \log(1/\epsilon)$ then, for any unit vector $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathrm{sign}(\mathbf{v}\cdot\mathbf{x} + t) - \mathrm{sign}(t)|] = O(\epsilon)$.*

Proof The first statement is proved in Diakonikolas et al. (2018) (Lemma 4.2). For the second statement, assuming without loss of generality that $t_2 \geq t_1 > 0$, we note that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_1) - \mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_2)|] = \frac{2}{\sqrt{2\pi}} \int_{t_1}^{t_2} e^{-t^2/2} \mathrm{d}t = \mathop{\mathbf{Pr}}_{t\sim\mathcal{N}(0,1)}[t_1 \leq t \leq t_2] \,.$$

Using the anti-concentration property of the one-dimensional Gaussian distribution, we have that $\mathbf{Pr}_{t\sim\mathcal{N}(0,1)}[t_1 \leq t \leq t_2] \leq O(t_2 - t_1)$, proving the claim.

For the third statement, note that if $t > \Omega(\sqrt{\log(1/\epsilon)})$, then by the concentration properties of the Gaussian, we have that:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathrm{sign}(\mathbf{v}\cdot\mathbf{x} + t) - \mathrm{sign}(t)|] \leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{v}\cdot\mathbf{x}| \geq |t|] = O(\epsilon) \,.$$

This completes the proof of Fact 22. ■

Proof [Proof of Theorem 2] We first show that there is a set $\mathcal{H}$ of size $(1/\epsilon)^{O(1/\epsilon^6)}$ which contains tuples $(\mathbf{u}, t)$ with $\mathbf{u} \in \mathbb{R}^d$ and $t \in \mathbb{R}$, such that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t) \neq y] \leq \inf_{f\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y] + \epsilon \,.$$

First note we can assume that $1/\epsilon^6 \leq d$, since otherwise one can directly do a brute-force search over an ϵ-cover of the d-dimensional unit ball: we do not need to perform our dimension-reduction process. The runtime to perform this brute-force search will be $(1/\epsilon)^{O(d)}\log(1/\delta)$ which, by the assumption that $1/\epsilon^6 > d$, is smaller than $(1/\epsilon)^{O(1/\epsilon^6)}\log(1/\delta)$.

Let $f \in \mathcal{C}$ be such that the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})y]$ is maximized and let $k = O(1/\epsilon^4)$. By an application of Lemma 7 for $N = (d/\epsilon)^{O(1/\epsilon^4)}\mathrm{poly}(1/\epsilon)\log(1/\delta) = d^{O(1/\epsilon^4)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$, it follows that there exists a degree $O(1/\epsilon^4)$ polynomial $P(\mathbf{x})$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \le \min_{P'\in\mathcal{P}_k}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2] + O(\epsilon^3)\ ,$$

with probability $1-\delta/2$. Applying Proposition 5 to the polynomial $P(\mathbf{x})$, we get that the subspace V spanned by the eigenvectors of the matrix $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$ with eigenvalues larger than $\eta = \Theta(1/\epsilon^2)$ contains a vector $\mathbf{v}\in V$, such that

$$\min_{t\in\mathbb{R}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t))y] \le \epsilon\ . \tag{15}$$

Moreover, by Lemma 6, the dimension of V is $O(1/\epsilon^6)$. Applying Fact 21, we get that there exists an ϵ-cover $\widetilde{V}$ of the set V with respect the ℓ_2-norm of size $(1/\epsilon)^{O(1/\epsilon^6)}$. We show that there is an effective way to discretize the set of biases. From Fact 22, it is clear that the set $\mathcal{T} = \{\pm\epsilon, \pm 2\epsilon, \ldots, \pm O(\sqrt{\log(1/\epsilon)})\}$ is an effective cover of the parameter t.

It remains to show that the set $\mathcal{H}$ is an effective cover, where $\mathcal{H} = \widetilde{V}\times\mathcal{T}$. We show that there exists a set of parameters $(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}$ which define a halfspace that correlates with the labels as well as the function f. Fix the parameters $(\mathbf{v},t)$ which minimize the Equation (15). Indeed, we have

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-\mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))y]\\ &\quad= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t))y] + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t)-\mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))y]\ .\end{aligned} \tag{16}$$

We claim that there exists $(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t)-\mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))y] = O(\epsilon)\ .$$

If $|t| > \sqrt{\log(1/\epsilon)}$, then from Fact 22, the constant hypothesis gets $O(\epsilon)$ error, so we need to check the case $|t| \le \sqrt{\log(1/\epsilon)}$. Applying Fact 22, we get that

$$\min_{(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t)-\mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t})|] = \min_{(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}} O(\|\mathbf{v}-\tilde{\mathbf{v}}\|_2+|t-\tilde{t}|) = O(\epsilon)\ . \tag{17}$$

Thus, substituting Equation (17) to Equation (16), we get

$$\min_{(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-\mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))y] \le \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t))y] + O(\epsilon) = O(\epsilon)\ , \tag{18}$$

where in the last equality we used Equation (15). Using the fact that for a boolean function $g(\mathbf{x})$ it holds $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[g(\mathbf{x})y] = 1 - 2\,\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[g(\mathbf{x})\neq y]$, we get

$$\min_{(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\tilde{\mathbf{u}}\cdot\mathbf{x}+\tilde{t})\neq y] \le \inf_{f\in\mathcal{C}}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})\neq y] + O(\epsilon)\ .$$

To complete the proof, we show that Step 10 of Algorithm 1 outputs a hypothesis close to the minimizer inside $\mathcal{H}$. From Hoeffding's inequality, it follows that $O(\frac{1}{\epsilon^2}\log(\mathcal{H}/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most ϵ with probability at least $1-\delta/2$. To bound the runtime of the algorithm, we note that L_2-regression has runtime $d^{O(1/\epsilon^4)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ and exhaustive search over an ϵ-cover takes time $(1/\epsilon)^{O(1/\epsilon^6)}\log(1/\delta)$. Thus, the total runtime of our algorithm in the case where $1/\epsilon^6 \le d$ is

$$\left(d^{O(1/\epsilon^4)} + (1/\epsilon)^{O(1/\epsilon^6)}\right)\log(1/\delta)\,.$$

This completes the proof of Theorem 2. ■

B.3. Proof of Lemma 7

We restate and prove the following lemma:

Lemma 23 (ℓ_2-Polynomial Regression) *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0},\mathbf{I})$. Let $k \in \mathbb{Z}_+$ and $\epsilon,\delta > 0$. There is an algorithm that draws $N = (dk)^{O(k)}\log(1/\delta)/\epsilon^2$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N,d)$, and outputs a polynomial $P(\mathbf{x})$ of degree at most k such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[(f(\mathbf{x})-P(\mathbf{x}))^2] \le \min_{P'\in\mathcal{P}_k}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[(f(\mathbf{x})-P'(\mathbf{x}))^2]+\epsilon$, with probability $1-\delta$.*

Proof Let S denote the empirical distribution of $\mathcal{D}$ with $N = (d/\epsilon)^{O(k)}$ samples. Recall that for any such $P(\mathbf{x})$, it holds that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[P^2(\mathbf{x})] \le 5$ (see Lemma 6). Writing $P(\mathbf{x})$ in the Hermite basis, $P(\mathbf{x}) = \sum_{\alpha\in\mathbb{N}^d} c_\alpha H_\alpha(\mathbf{x})$, it holds that $\sum_{\alpha\in\mathbb{N}^d} c_\alpha^2 = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[P^2(\mathbf{x})]$. The one-dimensional Hermite polynomials of k-degree are $H_k(z) = \sum_{m=0}^{\lfloor k/2\rfloor}\frac{(-1)^m z^{k-2m}}{m!(n-2m)!2^m}$. Thus, each monomial has coefficient absolute bounded by 2^k. Therefore, the maximum coefficient of a multidimensional Hermite polynomial $H_a(\mathbf{x})$ is $2^{|a|}$, thus the maximum coefficient of the polynomial $P(\mathbf{x})$ is $O(2^k)$. Let us now prove that for any degree-k polynomial $P(\mathbf{x})$ with coefficients bounded by $C = 2^{O(k)}$, we have

$$\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[P(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[P(\mathbf{x})y]\right| \le \epsilon\,,$$

with high constant probability. Write $P(\mathbf{x}) = \sum a_i m_i(\mathbf{x})$, where the summation ranges over all monomials m_i with degree less than k along with their coefficients a_i. We have

$$\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[P(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[P(\mathbf{x})y]\right| \le \sum |a_i|\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[m_i(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[m_i(\mathbf{x})y]\right|\,. \tag{19}$$

Using Markov's inequality, we have

$$\begin{aligned}\mathbf{Pr}\left[\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[m_i(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[m_i(\mathbf{x})y]\right| \ge \epsilon/(d^kC)\right] &\le \frac{C^2d^{2k}}{N\epsilon^2}\mathbf{Var}[m_i(\mathbf{x})y] \\ &\le \frac{C^2d^{2k}}{N\epsilon^2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[m_i^2(\mathbf{x})y^2] \\ &\le \frac{C^2d^{2k}}{N\epsilon^2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|\mathbf{x}\|_2^{2i}] = O\left(\frac{C^2i^id^{2k}}{N\epsilon^2}\right)\,.\end{aligned}$$

By using the fact that $N = (d\,k)^{O(k)}/\epsilon^2$ and applying above to the Equation (19), we have

$$\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[P(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[P(\mathbf{x})y]\right| \le C\sum\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[m_i(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[m_i(\mathbf{x})y]\right| \le \epsilon \;,$$

with high probability. Next, we need to bound the difference $\left|\mathbf{E}_{(\mathbf{x},y)\sim S}[P^2(\mathbf{x})] - \mathbf{E}_{(\mathbf{x},y)\sim \mathcal{D}}[P^2(\mathbf{x})]\right|$. This can be done by applying the same procedure as before and noting that the highest coefficient is at most C^2 and the degree is $2k$. Thus, for any k-degree polynomial P with high probability, we have

$$\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(P(\mathbf{x}) - y)^2] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[(P(\mathbf{x}) - y)^2]\right| \le \epsilon \;, \tag{20}$$

where we used the fact that $\mathbf{E}_{(\mathbf{x},y)\sim S}[y^2] = \mathbf{E}_{(\mathbf{x},y)\sim \mathcal{D}}[y^2]$. By solving a convex program, we can find a polynomial P such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(y - P(\mathbf{x}))^2] \le \min_{P'\in\mathcal{P}_k} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(y - P'(\mathbf{x}))^2] + \epsilon \;,$$

Note that if $P''(\mathbf{x}) = \mathrm{argmin}_{P'\in\mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim \mathcal{D}}[(y - P'(\mathbf{x}))^2]$, then

$$\min_{P'\in\mathcal{P}_k} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(y - P'(\mathbf{x}))^2] \le \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(y - P''(\mathbf{x}))^2] \le \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[(y - P''(\mathbf{x}))^2] \;,$$

where we used Equation (20). Thus, we have proved that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[(y - P(\mathbf{x}))^2] \le \min_{P'\in\mathcal{P}_k} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[(y - P'(\mathbf{x}))^2] + \epsilon \;,$$

with high constant probability. Using basic boosting procedures (see, e.g., exercise 1, chapter 13 of Shalev-Shwartz and Ben-David (2014)), we can boost the probability of success to $1-\delta$, with $N' = N\log(1/\delta) = (dk)^{O(k)}\log(1/\delta)/\epsilon^2$. ■

Appendix C. Omitted Proofs from Section 4

We restate and prove the following claims.

Claim 24 *For vectors* $\mathbf{v},\mathbf{u} \in \mathbb{R}^d, t \in \mathbb{R}$ *define the halfspaces* $h_0(\mathbf{x}) = \mathrm{sign}(\mathbf{u}\cdot\mathbf{x})$*,* $h_1(\mathbf{x}) = \mathrm{sign}(\mathbf{v}\cdot\mathbf{x} + t)$*. It holds* $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{x}) \neq h_0(\mathbf{x})] \ge \theta(\mathbf{u},\mathbf{v})/\pi$*.*

Proof Denote $\theta = \theta(\mathbf{u},\mathbf{v})$ and first assume that $\theta \in [0,\pi/2)$. Since the Gaussian distribution is invariant under rotations, for simplicity we may assume that $\mathbf{u},\mathbf{v}$ span $\mathbb{R}^2$. Morover, assume that two halfspaces intersect at the origin $(0,0)$ (if they do not intersect then the claimed lower bound on the disagreement $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{x}) \neq h_0(\mathbf{x})]$ is trivially true as their angle is 0). Moreover, assume that $\mathbf{u} = \mathbf{e}_1$ and that the Gaussian is centered at some point $(z,0)$, i.e., a point that lies on the $\mathbf{x}_1$-axis. This follows from the fact that $h_0(\mathbf{x}) = \mathrm{sign}(\mathbf{u}\cdot\mathbf{x})$ is homogeneous. After we change coordinates, the halfspace h_1 is also homogeneous, with normal vector $\mathbf{v} = (-\sin\theta,\cos\theta)$. We will show that the disagreement between the two halfspaces is *minimized* where $z = 0$, i.e., when the Gaussian is centered on the intersection point of the two halfspaces. Using the above, we obtain that

$$\begin{aligned}&\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{x})\neq h_0(\mathbf{x})]\\&\qquad=\int_{-\infty}^{0}\int_{\mathbf{x}_1\tan\theta}^{0}e^{-((\mathbf{x}_1-z)^2/2-\mathbf{x}_2^2/2}\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{x}_1+\int_{0}^{\infty}\int_{0}^{\mathbf{x}_1\tan\theta}e^{-((\mathbf{x}_1-z)^2/2-\mathbf{x}_2^2/2}\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{x}_1:=q(z)\end{aligned}$$

We will show that the function q is minimized for $z=0$. Taking the derivative with respect to z, we obtain

$$q'(z)=\int_{-\infty}^{0}\int_{\mathbf{x}_1\tan\theta}^{0}(\mathbf{x}_1-z)e^{-((\mathbf{x}_1-z)^2/2-\mathbf{x}_2^2/2}\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{x}_1+\int_{0}^{\infty}\int_{0}^{\mathbf{x}_1\tan\theta}(\mathbf{x}_1-z)e^{-((\mathbf{x}_1-z)^2/2-\mathbf{x}_2^2/2}\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{x}_1\,.$$

Observe that $q'(-z)=-q'(z)$, i.e, $q'(z)$ is an odd function with $q'(0)=0$. Thus, it can only be minimized at 0. We have that $q''(0)>0$ and therefore $z=0$ is the global minimizer of $q(z)$. The case $\theta\in[\pi/2,\pi]$ can be shown similarly. ∎

Claim 25 *Let* $h_1(\mathbf{x})=\mathrm{sign}(\mathbf{u}\cdot\mathbf{x}+t_1)$, $h_2(\mathbf{x})=\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t_2)$ *be two halfspaces. Let* $r_1=t_1/\left\|\mathbf{u}\mathbf{\Sigma}^{1/2}\right\|_2$, $r_2=t_2/\left\|\mathbf{v}\mathbf{\Sigma}^{1/2}\right\|_2$*. It holds*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}[h_1(\mathbf{x})\neq h_2(\mathbf{x})]\geq\mathop{\mathbf{Pr}}_{r\sim\mathcal{N}(0,1)}[\min(r_1,r_2)\leq r\leq\max(r_1,r_2)]\,.$$

Proof We first observe that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}[h_1(\mathbf{x})\neq h_2(\mathbf{x})]&=\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma}\mathbf{x})\neq h_2(\mathbf{\Sigma}\mathbf{x})]\\&\geq\left|\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma}\mathbf{x})\neq 0]-\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_2(\mathbf{\Sigma}\mathbf{x})\neq 0]\right|\,,\end{aligned}$$

where in the last step we used triangle inequality. Moreover, using that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma}\mathbf{x})\neq 0]=\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma}\mathbf{x})]$, we have

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}[h_1(\mathbf{x})\neq h_2(\mathbf{x})]\geq\left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma}\mathbf{x})]-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_2(\mathbf{\Sigma}\mathbf{x})]\right|\,.$$

Note that $h_1(\mathbf{\Sigma}\mathbf{x})=\mathrm{sign}(\mathbf{u}\cdot\mathbf{\Sigma}\mathbf{x}/\left\|\mathbf{u}\mathbf{\Sigma}^{1/2}\right\|_2+r_1)$ and $h_2(\mathbf{\Sigma}\mathbf{x})=\mathrm{sign}(\mathbf{v}\cdot\mathbf{\Sigma}\mathbf{x}/\left\|\mathbf{v}\mathbf{\Sigma}^{1/2}\right\|_2+r_2)$, thus

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}[h_1(\mathbf{x})\neq h_2(\mathbf{x})]&\geq\left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma}\mathbf{x})]-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_2(\mathbf{\Sigma}\mathbf{x})]\right|\\&=\left|\mathop{\mathbf{Pr}}_{r\sim\mathcal{N}(0,1)}[r\leq r_1]-\mathop{\mathbf{Pr}}_{r\sim\mathcal{N}(0,1)}[r\leq r_2]\right|\,,\end{aligned}$$

which completes the proof. ∎

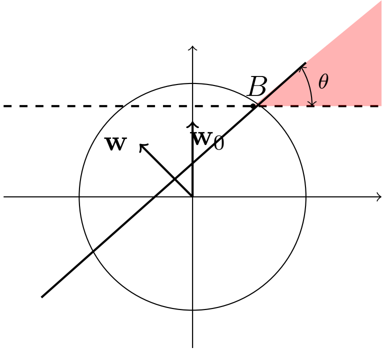

Figure 1: Localization technique, Lemma 16

C.1. Proof of Lemma 16

We restate and prove the following lemma.

Lemma 26 (Gaussian Localization) *Let $R(\mathbf{x})$ be the event that the sample $\mathbf{x}$ is rejected from the rejection sampling procedure of Lemma 14 with vector $\mathbf{w}_0$ and $\sigma = \Theta\left(\frac{\mathrm{OPT}}{\alpha}\right)$. Let $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x})$, $h'(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t)$ be halfspaces with $t = O(\sigma\alpha)$ and $\theta(\mathbf{w}_0, \mathbf{w}) = O(\sigma\alpha)$. Then, $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), R(\mathbf{x})] = O(\alpha\,\mathrm{OPT})$.*

Proof Let $\theta = \theta(\mathbf{w}_0, \mathbf{w})$ and fix $r = \Theta(1/\alpha^{1/3})\max(1, t/\sin\theta)$. We define $B = r\sin\theta(\sqrt{1 - t^2/r^2} - t/r\frac{\cos\theta}{\sin\theta})$. Observe that in order for $B > 0$ we need $r \geq t/\sin\theta$, which is true by our assumptions. Also note that $B = \Theta(r\sin\theta)$. We have that

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), R(\mathbf{x})] = \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| \geq B, R(\mathbf{x})] + \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| < B, R(\mathbf{x})]\,.$$

We first bound from above the term $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| < B, R(\mathbf{x})]$. It holds that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| < B, R(\mathbf{x})] &\leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}_0 \cdot \mathbf{x}| < B, R(\mathbf{x})] \\ &= \mathrm{erf}(B/\sqrt{2}) - \sigma\,\mathrm{erf}(B/(\sigma\sqrt{2})) \\ &= O(B^3/\sigma^2) = O(\alpha\,\mathrm{OPT})\,, \qquad (21)\end{aligned}$$

where erf is the error function and in the last equality we used the error of the Taylor expansion of degree-2. In order to bound the second term, we define V to be the subspace spanned by the vectors $\mathbf{w}_0, \mathbf{w}$. Let $\mathbf{x}', x''$ be the solutions of the system of equations $\{\mathbf{w} \cdot \mathbf{x} + t = 0, \|\mathbf{x}_V\|_2 = r^2\}$; observe that $\min(|\mathbf{x}' \cdot \mathbf{w}_0|, |\mathbf{x}'' \cdot \mathbf{w}_0|) = B$, thus, if $\|\mathbf{x}_V\|_2 \leq r$ and $|\mathbf{w}_0 \cdot \mathbf{x}| \geq B$, then $h(\mathbf{x}) = h'(\mathbf{x})$ (see Figure 1), thus

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| \geq B, R(\mathbf{x})] &\leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| \geq B, R(\mathbf{x})] \\ &\leq C\theta e^{-r} = O(\alpha\,\mathrm{OPT})\,. \qquad (22)\end{aligned}$$

Combining Equations (21), (22), we get $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), R(\mathbf{x})] = O(\alpha\,\mathrm{OPT})$. ■

Appendix D. Agnostic Proper Learning of ReLus

In this section, we use our techniques to develop a proper agnostic learning algorithm that handles ReLU activations. We work in the standard L_2-regression setting, i.e., we want to find a weight vector $\mathbf{w}$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\max(0,\mathbf{w}\cdot\mathbf{x}))^2] \leq \min_{f\in\mathcal{C}^\rho}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-f(\mathbf{x}))^2]+\epsilon\;,$$

where by $\mathcal{C}^\rho$ we denote the class of ReLU activations, i.e., $\mathcal{C}^\rho=\{\mathbf{x}\mapsto\max(0,\mathbf{w}\cdot\mathbf{x}+t):\|\mathbf{w}\|_2\leq 1,\mathbf{w}\in\mathbb{R}^d,t\in\mathbb{R}^d\}$. Moreover, we are going to use $\mathcal{C}^\rho_0=\{\mathbf{x}\mapsto\max(0,\mathbf{w}\cdot\mathbf{x}):\|\mathbf{w}\|_2\leq 1,\mathbf{w}\in\mathbb{R}^d\}$ and $\rho(\mathbf{x})$ to denote the ReLU activation function. Finally, observe that for $\mathbf{w}\in\mathbb{R}^d$ with $\|\mathbf{w}\|_2\leq 1$ it holds $\max(0,\mathbf{w}\cdot\mathbf{x})=\|\mathbf{w}\|_2\max(0,\mathbf{w}\cdot\mathbf{x}/\left\|\mathbf{w}\right\|_2)$. To keep the presentation simple we are going to assume, similarly to Diakonikolas et al. (2020a) that the observed labels y are bounded in $[-1,1]$. For the rest of the section, we assume that for the labels y, it holds $|y|<1$.

The pseudocode of our algorithm is given in Algorithm 3.

Algorithm 3 Agnostic Proper Algorithm for ReLU Regression

```
procedure AGNOSTIC-LEARNER($\epsilon, \delta, \mathcal{D}$)
Input: $\epsilon > 0$, $\delta > 0$ and distribution $\mathcal{D}$
Output: A hypothesis $h \in \mathcal{C}$ such as $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(h(\mathbf{x})-y)^2] \leq \min_{f\in\mathcal{C}^\rho}\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-y)^2]+\epsilon$ with probability $1-\delta$.

    $k \leftarrow C/\epsilon^{4/3}$, $\eta \leftarrow \epsilon^2/C$.
    Find $P(\mathbf{x})$ such $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k}\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2]+O(\epsilon^3)$.
    Let $\mathbf{M}=\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$.
    Let $V$ be the subspace spanned by the eigenvectors of $\mathbf{M}$ whose eigenvalues are at least $\eta$.
    Construct an $\epsilon$-cover $\mathcal{H}$ of hypotheses with normal vectors in $V$        $\triangleright$ see Fact 21.
    Draw $\Theta(\frac{1}{\epsilon^2}\log(|\mathcal{H}|/\delta))$ i.i.d. samples from $\mathcal{D}$ and construct the empirical distribution $\widehat{\mathcal{D}}$.
    $h \leftarrow \mathrm{argmin}_{h'\in\mathcal{H}}\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[(h'(\mathbf{x})-y)^2]$
    return $h$.
```

Theorem 27 *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\mathbb{R}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0},\mathbf{I})$. Algorithm 3 draws $N=d^{O(1/\epsilon^{4/3})}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ samples from $\mathcal{D}$, runs in time $(d^{O(1/\epsilon^{4/3})}+(1/\epsilon)^{O(1/\epsilon^{10/3})})\log(1/\delta)$, and computes a hypothesis $h\in\mathcal{C}^\rho$ such that, with probability at least $1-\delta$, we have that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-h(\mathbf{x}))^2] \leq \min_{f\in\mathcal{C}^\rho_0}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-f(\mathbf{x}))^2]+\epsilon\;.$$

The main structural result of this section is the following proposition showing that we can perform dimension reduction by looking at high-influence directions of a low-degree polynomial.

Proposition 28 *Let C be a sufficiently large constant, fix $\epsilon>0$, $k=C/\epsilon^{4/3}$. Let $P(\mathbf{x})\in\mathcal{P}_k$ be a polynomial such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2]+O(\epsilon^3)\,.$$

Moreover, let $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$ and V be the subspace spanned by the eigenvectors of $\mathbf{M}$ with eigenvalues larger than η, where $\eta = \epsilon^2/C$. Then, for any function $f \in \mathcal{C}_0^\rho$, it holds

$$\min_{\mathbf{v}\in V, \|\mathbf{v}\|_2\leq 1, t\in\mathbb{R}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v}\cdot\mathbf{x}+t)-y)^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-y)^2]+\epsilon\,.$$

The proof of Proposition 28 is similar to the proof of Proposition 5. We provide the details below for completeness.

Proof Suppose for the sake of contradiction that there exists a hypothesis $f \in \mathcal{C}_0^\rho$ such that for every hypothesis $f' \in \mathcal{C}_V^\rho$, it holds

$$\min_{\mathbf{v}\in V, \|\mathbf{v}\|_2\leq 1, t\in\mathbb{R}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v}\cdot\mathbf{x}+t)-y)^2] > \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-y)^2]+\epsilon\,. \tag{23}$$

Equivalently, from the above equation, we have that for every $\mathbf{v} \in V$ with $\|\mathbf{v}\|_2 \leq 1$ and $t \in \mathbb{R}$:

$$2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-\rho(\mathbf{v}\cdot\mathbf{x}+t))y] > \epsilon + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f^2(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\rho^2(\mathbf{v}\cdot\mathbf{x}+t)]\,. \tag{24}$$

Claim 29 *It suffices to show that there exists some polynomial $Q(\mathbf{x})$ of degree at most k, with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q^2(\mathbf{x})] \leq 4$, that $(\epsilon/4)$-correlates with $(y - P(\mathbf{x}))$, i.e.,*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[Q(\mathbf{x})(y-P(\mathbf{x}))] \geq \epsilon/4\,.$$

Proof We have that the polynomial $P(\mathbf{x})+\zeta Q(\mathbf{x})$, for $\zeta = \Theta(\epsilon)$, has degree at most k and decreases the value of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2]$ by at least $\Omega(\epsilon^2)$, which contradicts the optimality of $P(\mathbf{x})$, i.e., that $P(\mathbf{x})$, $O(\epsilon^3)$-close to the polynomial that minimizes the L_2 error with y. ■

We now construct such a polynomial $Q(\mathbf{x})$. We have $f(\mathbf{x}) = \rho(\mathbf{w}\cdot\mathbf{x}) = \rho(\mathbf{w}_V\cdot\mathbf{x}+\mathbf{w}_{V^\perp}\cdot\mathbf{x})$, for some $0 < a \leq 1$. It holds that $\mathbf{w}_{V^\perp} \neq \mathbf{0}$ since otherwise we would have that $f \in \mathcal{C}_V^\rho$. For simplicity, we denote $\xi = \mathbf{w}_{V^\perp}/\left\|\mathbf{w}_{V^\perp}\right\|_2$. Notice that the direction ξ is of low influence since $\xi \in V^\perp$. Recall, that by $\mathcal{D}_\xi$ we denote the projection of $\mathcal{D}$ onto the (one-dimensional) subspace spanned by ξ. We define $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z}+\mathbf{x}_V)]$: a convex combination of hypotheses in $\mathcal{C}_V^\rho$. In particular, $f_V(\mathbf{x})$ is a smoothed version of the hypothesis $\rho(\mathbf{w}_V\cdot\mathbf{x}+t)$, whose normal vector belongs in V. We first observe that by (24) $f_V(\mathbf{x})$ cannot correlate too well with y:

$$\begin{aligned}2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-f_V(\mathbf{x}))y] &= 2\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-\rho(\mathbf{w}\cdot\mathbf{x}_V+\mathbf{w}\cdot\mathbf{z}))y]]\\ &\geq \epsilon+\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f^2(\mathbf{x})]-\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\rho^2(\mathbf{w}\cdot\mathbf{x}_V+\mathbf{w}\cdot\mathbf{z})]]=\epsilon\,,\end{aligned} \tag{25}$$

where the last equality follows by the fact that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f^2(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{D}_{\xi^\perp}}\left[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[f^2(\mathbf{u}+\mathbf{z})]\right] = \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}\left[\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{D}_{\xi^\perp}}[\rho^2(\mathbf{u}\cdot\mathbf{w}_V+\mathbf{w}\cdot\mathbf{z})]\right] = \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}\left[\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{D}_\mathbf{x}}[\rho^2(\mathbf{u}_V\cdot\mathbf{w}+\mathbf{w}\cdot\mathbf{z})]\right].$$

Our argument consists two main claims. We first show that the function $f(\mathbf{x}) - f_V(\mathbf{x})$ correlates non-trivially with $y - P(\mathbf{x})$. Then we show that we can approximate $f(\mathbf{x}) - f_V(\mathbf{x})$ by a low degree polynomial $Q(\mathbf{x})$ that maintains non-trivial correlation with $y - P(\mathbf{x})$, see Claim 31. We start by proving the first claim.

Claim 30 *It holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))(y-P(\mathbf{x}))]\geq\epsilon/2-\sqrt{2\eta}\,.$$

Proof We have $f_V(\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z}+\mathbf{x}_{\xi^\perp})]=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[\rho(\mathbf{w}_V\cdot\mathbf{x}_V+\mathbf{w}\cdot\mathbf{z})]$ and, since f_V is a convex combination of hypothesis in $\mathcal{C}_V^\rho$, from Equation (23), we see that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-f_V(\mathbf{x}))y]\geq\epsilon$. Thus, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-f_V(\mathbf{x}))(y-P(\mathbf{x}))]&=\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-f_V(\mathbf{x}))y]-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))P(\mathbf{x})]\\&\geq\epsilon/2-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))P(\mathbf{x})]\,.\end{aligned}\tag{26}$$

To deal with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))P(\mathbf{x})]$, we first observe that for any function $g(\mathbf{x})$ depending only on the projection of $\mathbf{x}$ onto the subspace $\xi^\perp$, i.e., it holds $g(\mathbf{x})=g(\mathbf{x}_{\xi^\perp})$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))g(\mathbf{x})]=\mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}_{\xi^\perp}}\left[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{v}+\mathbf{z})-f_V(\mathbf{v})]\,g(\mathbf{v})\right]=0\,,$$

since for every $\mathbf{x}\in\mathbb{R}^d$, it holds $f_V(\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{x}_{\xi^\perp}+\mathbf{z})]=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{x}_V+\mathbf{z})]$. Unfortunately, this is not true since $P(\mathbf{x})$ is not only a function of $\mathbf{x}_{\xi^\perp}$. However, since V contains the high influence eigenvectors it holds that P is almost a function of $\mathbf{x}_{\xi^\perp}$. In fact, we show that we can replace the polynomial P by a different polynomial of degree at most k that only depends on the projection of $\mathbf{x}$ on $\xi^\perp$. Similarly to the definition of the "smoothed" hypothesis f_V, we define $R(\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[P(\mathbf{x}_{\xi^\perp}+\mathbf{z})]$. We first prove that $R(\mathbf{x})$ is close to $P(\mathbf{x})$ in the L_2 sense.

Now, adding and subtracting $R(\mathbf{x})=\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[P(\mathbf{x}_{\xi^\perp}+\mathbf{z})]$, we get

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))P(\mathbf{x})]=\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))(P(\mathbf{x})-R(\mathbf{x}_{\xi^\perp}))]+\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))R(\mathbf{x}_{\xi^\perp})]\,.$$

The second term equals to zero, from the fact that $\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z}+\mathbf{x}_{\xi^\perp})-f_V(\mathbf{x}_{\xi^\perp})]=0$. Using Cauchy-Schwarz inequality, we get

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))(P(\mathbf{x})-R(\mathbf{x}_{\xi^\perp}))]&\leq\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x})-f_V(\mathbf{x}))^2]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(P(\mathbf{x})-R(\mathbf{x}_{\xi^\perp}))^2]}\\&\leq\sqrt{2}\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(P(\mathbf{x})-R(\mathbf{x}_{\xi^\perp}))^2]}\leq\sqrt{2\eta}\,,\end{aligned}\tag{27}$$

where we used Claim 10. Using Equation (26), we get that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-f_V(\mathbf{x}))(y-P(\mathbf{x}))]\geq\epsilon/2-\sqrt{2\eta}\,,$$

which completes the proof of Claim 30. ■

Claim 31 *There exists a polynomial $Q(\mathbf{x})$ of degree $O(1/\epsilon^{4/3})$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[Q(\mathbf{x})(y-P(\mathbf{x}))]\geq\epsilon/4-\sqrt{2\eta}$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[Q^2(\mathbf{x})]\leq 4$.*

Proof For any polynomial $Q(\mathbf{x})$, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q(\mathbf{x})(y-P(\mathbf{x}))] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x})+(f(\mathbf{x})-f_V(\mathbf{x}))-(f(\mathbf{x})-f_V(\mathbf{x})))(y-P(\mathbf{x}))]\\ &\geq \epsilon/2-2\sqrt{\eta}+\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x})-(f(\mathbf{x})-f_V(\mathbf{x})))(y-P(\mathbf{x}))]\ ,\end{aligned} \tag{28}$$

where we used Claim 30. By choosing $Q(\mathbf{x}) = S(\mathbf{x}) - \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp}+\xi)]$, where we denote by $S(\mathbf{x})$ the Hermite expansion of f truncated up to degree k, $S(\mathbf{x}) = \sum_{|\alpha|\leq k}\hat{f}(\alpha)H_\alpha(\mathbf{x})$, we show that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x})-Q(\mathbf{x}))^2] \leq \epsilon^2\ .$$

We need the following fact:

Fact 32 (Goel et al. (2020)) *Let $f\in\mathcal{C}_0^\rho$, and let S be the Hermite expansion up to k-degree of f, i.e., $S(\mathbf{x}) = \sum_{|\alpha|\leq k}\hat{f}(\alpha)H_\alpha(\mathbf{x})$. Then $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(S(\mathbf{x})-f(\mathbf{x}))^2] = O(k^{-3/2})$.*

Using the inequality $(a+b)^2\leq 2a^2+2b^2$, we get that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x})-Q(\mathbf{x}))^2] \leq 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-S(\mathbf{x}))^2]+2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f_V(\mathbf{x})-\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp})])^2]\ .$$

Moreover, from Jensen's inequality, it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f_V(\mathbf{x})-\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp}+\xi)])^2] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-S(\mathbf{x}))^2] = O\left(k^{-3/2}\right)\ ,$$

where in the last equality we used the Fact 32. Choose $k=\Theta(1/\epsilon^{4/3})$. Applying Cauchy-Schwartz to the Equation (28), we get

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[Q(\mathbf{x})(y-P(\mathbf{x}))] &\geq \epsilon/2-\sqrt{2\eta}-\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x})-(f(\mathbf{x})-f_V(\mathbf{x})))^2]\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2]}\\ &\geq \epsilon/4-\sqrt{2\eta}\ ,\end{aligned}$$

where we used the fact that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2]\leq 2$. Note that from the reverse triangle inequality it holds that

$$\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[Q^2(\mathbf{x})]} \leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x}))^2]}+\epsilon \leq \sqrt{2}+\epsilon\ . \tag{29}$$

Equation (29) gives $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[Q^2(\mathbf{x})]\leq 4$. This completes the proof of Claim 31, which completes the proof of Claim 31.

■

By choosing $\eta=\Theta(\epsilon^2)$, Claim 31 contradicts our assumption that $P(\mathbf{x})$ is $O(\epsilon^3)$-close to the polynomial $P'(\mathbf{x})$ that minimizes the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2]$. This completes the proof. ■

The next lemma bounds the dimension of the subspace spanned by the high-influence directions of a polynomial that minimizes the L_2 error with the labels y.

Before we proceed to the proof of Theorem 27, we need an algorithm that calculates an approximate minimal polynomial for the Proposition 28.

Lemma 33 (L_2-Polynomial Regression) *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \mathbb{R}$ whose $\mathbf{x}$-marginal is the standard normal and whose labels are bounded by 1. Moreover, let $k \in \mathbb{Z}_+$, and $\epsilon, \delta > 0$. There is an algorithm that draws $N = (dk)^{O(k)} \log(1/\delta)/\epsilon^2$ samples, runs in time $\mathrm{poly}(N, d)$, and outputs a polynomial $P(\mathbf{x})$ of degree at most k such that with probability $1 - \delta$ it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + \epsilon$.*

The proof of this lemma is nearly identical to the proof of Lemma 7.

We need the following simple fact for ReLUs. An essentially identical fact was shown in Diakonikolas et al. (2020b) Equation (2) for the zero threshold case. We provide the proof here for completeness.

Fact 34 *Let $f_1(\mathbf{x}) = \rho(\mathbf{v} \cdot \mathbf{x} + T)$ and $f_2(\mathbf{x}) = \rho(\mathbf{u} \cdot \mathbf{x} + T)$, for $T \in \mathbb{R}$ and $\mathbf{v}, \mathbf{u}$ unit vectors in $\mathbb{R}^d$. Then $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f_1(\mathbf{x}) - f_2(\mathbf{x}))^2] = O(\|\mathbf{v} - \mathbf{u}\|_2^2)$.*

Proof The proof relies on the following fact.

Fact 35 (Correlated Differences, Lemma 6 of Kontonis et al. (2019)) *Let $r(\mathbf{x}) \in L_2(\mathbb{R}^d, \mathcal{N}^d)$ be differentiable almost everywhere and let*

$$D_\rho = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{I} & \rho\mathbf{I} \\ \rho\mathbf{I} & \mathbf{I} \end{pmatrix}\right).$$

We call ρ-correlated a pair of random variables $(\mathbf{x}, \mathbf{y}) \sim D_\rho$. It holds

$$\frac{1}{2} \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{z})\sim D_\rho}[(r(\mathbf{x}) - r(\mathbf{z}))^2] \leq (1 - \rho) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|\nabla r(\mathbf{x})\|_2^2\right].$$

Using this fact for $\rho = \mathbf{v} \cdot \mathbf{u}$, and using the approximation $(1 - \mathbf{v} \cdot \mathbf{u}) = \|\mathbf{u} - \mathbf{v}\|_2^2$ the result follows. ■

We also need the following fact about the biases of ReLUs.

Fact 36 *Let $f_1(\mathbf{x}) = \rho(\mathbf{v} \cdot \mathbf{x} - T)$ and $f_2(\mathbf{x}) = \rho(\mathbf{v} \cdot \mathbf{x} - T')$ with $T' \geq T$. Then $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f_1(\mathbf{x}) - f_2(\mathbf{x}))^2] = O((T - T')^2 + 2T'(T' - T)e^{-T^2/2})$.*

Proof Without loss of generality, we can assume that $\mathbf{v} = \mathbf{e}_1$. The result follows by noting that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f_1(\mathbf{x}) - f_2(\mathbf{x}))^2] \leq \int_{\mathbf{x}_1 \geq T} (f_1(\mathbf{x}) - f_2(\mathbf{x}))^2 \phi(\mathbf{x}) \mathrm{d}\mathbf{x} + \int_{T'}^{T} f_2(\mathbf{x}) \phi(\mathbf{x}) \mathrm{d}\mathbf{x}$. ■

We can now prove the main theorem of this section.

Proof [Proof of Theorem 27] Let $f \in \mathcal{C}_0^\rho$ such that the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})y]$ is maximized. Using Lemma 33 on the labels y, with $N = d^{O(1/\epsilon^{4/3})} \mathrm{poly}(1/\epsilon) \log(1/\delta)$ samples, we get an $k = O(1/\epsilon^{4/3})$-degree polynomial $P(\mathbf{x})$ and it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + \epsilon^3,$$

with probability $1 - \delta/2$. Applying Proposition 28 to the polynomial $P(\mathbf{x})$, we get that subspace V spanned by the eigenvectors of the matrix $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\top]$ with eigenvalues larger than $\eta = \Omega(1/\epsilon^2)$ contains a vector $\mathbf{v} \in V$, so that

$$\min_{\mathbf{v} \in V, \|\mathbf{v}\|_2 \leq 1, t \in \mathbb{R}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v} \cdot \mathbf{x} + t) - y)^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - y)^2] + \epsilon. \tag{30}$$

Moreover, from Lemma 6, the dimension of V is $O(1/\epsilon)^{10/3}$. Thus, applying Fact 21, we get that there exists a set $\tilde{V}$ which is an ϵ-cover of the set V with respect the ℓ_2-norm of size $(1/\epsilon)^{O(1/\epsilon^{10/3})}$. We will use the set $\mathcal{T} = \{\epsilon/A, 2\epsilon/A, \dots, 1\}$, where A is a large enough constant, and show that it is a good cover of the parameter a which is used as the guess of the norm of the vector $\mathbf{v}$. Finally, we need an effective cover for the biases t. Observe that from Fact 36, we need step-size $s = \epsilon^2/\sqrt{\log(1/\epsilon)}$ and the maximum negative value is $-\Theta(\sqrt{\log(1/\epsilon)})$. (If the value was larger, then the zero function would correlate as well.) We also need to bound the maximum positive value. We claim that the maximum positive value is some universal constant C'. This is because the error scales with the norm of the function that is returned by the algorithm and because we are trying to be competitive against the unbiased ReLU, the norm of the ReLU that is returned cannot be more than $2(\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\rho^2(\mathbf{w}\cdot\mathbf{x})] + 4\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y^2]) \leq C'$, for some large enough constant C'. Thus, the set of biases is $\mathcal{T}' = \{-C\sqrt{\log(1/\epsilon)}A/\epsilon, \dots, 0, s, 2s, \dots, CA/\epsilon'\}$, where we multiply with the minimal value of the guess of the norm. This is because if $\|\mathbf{v}\|_2 = \alpha$, then we have that $\rho(\mathbf{v}\cdot\mathbf{x}+t) = \alpha\rho(\mathbf{v}\cdot\mathbf{x}/\|\mathbf{v}\|_2 + t/\alpha)$.

We show that the set $\mathcal{H}$ is an effective cover, where $\mathcal{H} = \widetilde{V} \times \mathcal{T} \times \mathcal{T}'$. We show that there exist a set of parameters $(\tilde{\mathbf{v}}, \tilde{a}, \tilde{t}) \in \mathcal{H}$ which define a ReLU that correlates with the labels as well as the function f. Fix the parameters $\mathbf{v}, \alpha, t$ which minimize the Equation (30). Indeed, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[a\rho(\mathbf{v}\cdot\mathbf{x}+t) - \tilde{a}\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))^2]^{1/2} &\leq a \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v}\cdot\mathbf{x}+t) - \rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+t))^2]^{1/2} \\ &+ a \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}) - \rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+t))^2]^{1/2} + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t})^2]^{1/2}|(a-\tilde{a})| \;.\end{aligned} \tag{31}$$

Applying Facts 34 and 36, we get that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[a\rho(\mathbf{v}\cdot\mathbf{x}+t) - \tilde{a}\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))^2] \leq O(\epsilon) \;.$$

Thus, from the triangle inequality, we get that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\tilde{a}\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t})-y)^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v}\cdot\mathbf{x}+t)-y)^2] + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[a\rho(\mathbf{v}\cdot\mathbf{x}+t)-\tilde{a}\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-y)^2]+O($$

To complete the proof, it remains to show that Step 11 outputs a hypothesis close to the minimizer inside $\mathcal{H}$. We need the following claim:

Claim 37 *Let $h \in \mathcal{C}^\rho$ and let $\widehat{\mathcal{D}}$ be the empirical distribution with $N = O((1/\epsilon^2)\log(1/\delta))$ samples. Then, with probability $1-\delta$, it holds*

$$|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[(y-h(\mathbf{x}))^2] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-h(\mathbf{x}))^2]| \leq \epsilon \;.$$

Proof We need first to prove that with probability $1-\delta$ it holds:

$$|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[yh(\mathbf{x})] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[yh(\mathbf{x})]| \leq \epsilon \;.$$

Using Markov's inequality, we have

$$\begin{aligned}\mathbf{Pr}[|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[h(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})y]| \geq \epsilon] &\leq \frac{1}{N\epsilon^2}\mathbf{Var}[h(\mathbf{x})y] \\ &\leq \frac{1}{N\epsilon^2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[h^2(\mathbf{x})y^2] \\ &\leq O\left(\frac{1}{N\epsilon^2}\right) ,\end{aligned}$$

where we used the fact that our functions are bounded in L_2-norm. With the same procedure we bound the difference $|\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[h^2(\mathbf{x})] - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[h^2(\mathbf{x})]| \leq \epsilon$. By using the fact that $N = O(1/\epsilon^2)$, we get our result for constant probability. By applying a standard probability amplification technique, we can boost the confidence to $1-\delta$ with $N' = O(N\log(1/\delta))$ samples. ■

Therefore, from Claim 37, it follows that $O(\frac{1}{\epsilon^2}\log(\mathcal{H}/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most ϵ with probability at least $1-\delta/2$.

To bound the runtime of the algorithm, we note that L_2-regression has runtime $d^{O(1/\epsilon^{4/3})}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ and the exhaustive search over an ϵ-cover takes time $(1/\epsilon)^{O(1/\epsilon^{10/3})}\log(1/\delta)$ time. The total runtime of our algorithm in the case where $1/\epsilon^{10/3} \leq d$ is

$$\left(d^{O(1/\epsilon^{4/3})} + (1/\epsilon)^{O(1/\epsilon^{10/3})}\right)\log(1/\delta)\,.$$

In the case where $1/\epsilon^{10/3} > d$, one can directly do a brute-force search over an ϵ-cover of the d-dimensional unit ball: we do not need to perform our dimension-reduction process and the runtime is bounded above by the previous case. ■