Learning *General* Halfspaces with *General* Massart Noise under the Gaussian Distribution

Ilias Diakonikolas* UW Madison ilias@cs.wisc.edu

Daniel M. Kane† UC San-Diego dakane@ucsd.edu

Vasilis Kontonis UW Madison kontonis@wisc.edu

Christos Tzamos UW Madison tzamos@wisc.edu

Nikos Zarifis‡ UW Madison zarifis@wisc.edu

April 14, 2022

Abstract

We study the problem of PAC learning halfspaces on $\mathbb{R}^d$ with Massart noise under the Gaussian distribution. In the Massart model, an adversary is allowed to flip the label of each point $\mathbf{x}$ with unknown probability $\eta(\mathbf{x}) \le \eta$, for some parameter $\eta \in [0, 1/2]$. The goal of the learner is to output a hypothesis with misclassification error of $\mathrm{OPT} + \epsilon$, where OPT is the error of the target halfspace. This problem had been previously studied under two assumptions: (i) the target halfspace is *homogeneous* (i.e., the separating hyperplane goes through the origin), and (ii) the parameter η is *strictly* smaller than $1/2$. Prior to this work, no nontrivial bounds were known for the general case when either of these assumptions is removed. Here we study the general problem and establish the following results:

- For $\eta < 1/2$, we give a learning algorithm for general halfspaces with sample and computational complexity $d^{O_\eta(\log(1/\gamma))}\mathrm{poly}(1/\epsilon)$, where $\gamma := \max\{\epsilon, \min\{\mathbf{Pr}[f(\mathbf{x}) = 1], \mathbf{Pr}[f(\mathbf{x}) = -1]\}\}$ is the "bias" of the target halfspace f. Prior efficient algorithms could only handle the special case of $\gamma = 1/2$. Interestingly, we also establish a qualitatively matching lower bound of $d^{\Omega(\log(1/\gamma))}$ on the complexity of any Statistical Query (SQ) algorithm of the problem.
- For $\eta = 1/2$, we give a learning algorithm for general halfspaces with sample and computational complexity $O_\epsilon(1)\, d^{O(\log(1/\epsilon))}$. This result is new even for the subclass of homogeneous halfspaces; prior algorithms for homogeneous Massart halfspaces provide vacuous guarantees for $\eta = 1/2$. We complement our upper bound with a nearly-matching SQ lower bound of $d^{\Omega(\log(1/\epsilon))}$, which holds even for the special case of homogeneous halfspaces.

Taken together, our results qualitatively characterize the complexity of learning general halfspaces with general Massart noise under Gaussian marginals. Our techniques rely on determining the existence (or non-existence) of low-degree polynomials whose expectations distinguish Massart halfspaces from random noise.

*Supported by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant.

†Supported by NSF Medium Award CCF-2107547, NSF Award CCF-1553288 (CAREER), and a Sloan Research Fellowship.

‡Supported in part by NSF Award CCF-1652862 (CAREER) and a DARPA Learning with Less Labels (LwLL) grant.

1 Introduction

This work focuses on the distribution-specific PAC learning of halfspaces in the presence of label noise. Before we describe our contributions, we provide the context for this work.

1.1 Background

A *halfspace* or Linear Threshold Function (LTF) is any Boolean-valued function $f : \mathbb{R}^d \mapsto \{\pm 1\}$ of the form $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} - t^*)$, for a vector $\mathbf{w}^* \in \mathbb{R}^d$ (known as the weight vector) and a scalar $t^* \in \mathbb{R}$ (known as the threshold). Halfspaces are a central class of Boolean functions in several areas of computer science, including complexity theory, learning theory, and optimization [Ros58, Nov62, MP68, Yao90, GHR92, FS97, Vap98, STC00, O'D14]. In this work, we focus on the algorithmic problem of learning halfspaces from labeled examples, arguably one of the most extensively studied and influential problems in machine learning.

The computational problem of PAC learning halfspaces is known to be efficiently solvable without noise (see, e.g., [MT94]) in the distribution-independent setting. The complexity of the problem in the presence of noisy data crucially depends on the noise model and the underlying distributional assumptions. In this work, we study the problem of distribution-specific PAC learning of halfspaces in the presence of Massart noise. Formally, we have the following definition.

Definition 1.1 (Distribution-specific PAC Learning with Massart Noise)**.** *Let $\mathcal{C}$ be a concept class of Boolean functions over $X = \mathbb{R}^d$, $\mathcal{F}$ be a* known family *of structured distributions on X, $0 \leq \eta \leq 1/2$, and $0 < \epsilon < 1$. Let f be an unknown target function in $\mathcal{C}$. A* noisy example oracle*, $\mathrm{EX}^{\mathrm{Mas}}(f, \mathcal{F}, \eta)$, works as follows: Each time $\mathrm{EX}^{\mathrm{Mas}}(f, \mathcal{F}, \eta)$ is invoked, it returns a labeled example $(\mathbf{x}, y)$, such that: (a) $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$, where $\mathcal{D}_{\mathbf{x}}$ is a fixed distribution in $\mathcal{F}$, and (b) $y = f(\mathbf{x})$ with probability $1 - \eta(\mathbf{x})$ and $y = -f(\mathbf{x})$ with probability $\eta(\mathbf{x})$, for an* unknown *function $\eta(\mathbf{x}) \leq \eta$. Let $\mathcal{D}$ denote the joint distribution on $(\mathbf{x}, y)$ generated by the above oracle. A learning algorithm is given i.i.d. samples from $\mathcal{D}$ and its goal is to output a hypothesis h such that with high probability it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \min_{c \in \mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[c(\mathbf{x}) \neq y]$.*

Remark 1.2. The noise rate parameter η in Definition 1.1 is allowed to be *equal* to $1/2$. This is consistent with the original definition of the Massart model in [MN06], and — as we argue in the subsequent discussion — is well-motivated in a number of practical applications. On the other hand, prior algorithmic work in the theoretical machine learning community imposed the crucial requirement that η is *strictly smaller* than $1/2$. This distinction turns out to be very significant and serves as one of the main motivations for the current work.

The Massart noise model in the above form was defined in [MN06]. A very similar noise model had been defined in the 80s by Sloan and Rivest [Slo88, Slo92, RS94, Slo96], and a related definition had been considered even earlier by Vapnik [Vap82]. The Massart model is a generalization of the Random Classification Noise (RCN) model [AL88] (where the flipping probability is uniform) and is a special case of the agnostic model (where the label noise is fully adversarial) [Hau92, KSS94].

The Massart model is a natural semi-random noise model that is more realistic than RCN. Specifically, label noise can reflect computational difficulty, ambiguity, or random factors. For example, a cursive "e" might be substantially more likely to be misclassified as "a" than an upper case Roman letter. Massart noise allows for such variations in misclassification rates without knowledge of which instances are more likely to be misclassified. That is, Massart noise-tolerant learners are less brittle than RCN tolerant learners. Agnostic learning is, of course, even more robust; unfortunately, agnostic learning is known to be computationally intractable in many settings of interest [GR06, FGKP06, Dan16, DKZ20, GGK20, DKPZ21].

1.2 Motivation for this Work

The algorithmic task of PAC learning with Massart noise is a classical problem in computational learning theory. In the distribution-independent setting, known efficient algorithms [DGT19, CKMY20] achieve error $\eta + \epsilon$ and we now know [DK20] that this error bound is close to best possible for efficient Statistical Query (SQ) algorithms [Kea98], even if $\mathrm{OPT} = \mathbf{E}[\eta(\mathbf{x})]$ is very small. This lower bound further motivates the study of the distribution-specific setting (the focus of the current work).

The work of [ABHU15] initiated the algorithmic study of learning Massart halfspaces under structured distributions. This work focused on the class of *homogeneous* halfspaces, i.e., functions of the form $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$, and gave a polynomial-time learning algorithm with $\mathrm{OPT} + \epsilon$ error under the uniform distribution on the unit sphere. The [ABHU15] algorithm succeeds when the parameter η is smaller than a sufficiently small constant ($\approx 10^{-6}$). A sequence of subsequent works [ABHZ16, YZ17, ZLC17, BZ17, MV19, DKTZ20a, ZSA20, ZL21] have given efficient learning algorithms for homogeneous Massart halfspaces that succeed for all $\eta < 1/2$ and under weaker distributional assumptions. The current state-of-the-art algorithms [DKTZ20a, ZSA20, ZL21] have sample and computational complexity $\mathrm{poly}(d, 1/\epsilon, 1/(1-2\eta))$ and succeed for all $\eta < 1/2$ under a class of distributions including isotropic log-concave distributions.

To summarize the preceding discussion, *known efficient algorithms for Massart halfspaces in the distribution-specific setting succeed under two crucial assumptions:*

(i) The target halfspace is homogeneous (i.e., has zero threshold), and

(ii) The upper bound parameter of the Massart noise satisfies $\eta < 1/2$.

Perhaps surprisingly, prior to the current work, no non-trivial bounds were known for the *general case* of this learning problem, where either of these two assumptions is removed. This represents a fundamental gap in our algorithmic understanding of learning halfspaces in the Massart model and serves as the main motivation of the current work.

In this work, we study the general version of this problem for the prototypical setting that the examples are drawn from the Gaussian distribution. As our main contribution, *we essentially characterize the complexity of the problem by giving the first efficient learning algorithms coupled with qualitatively matching SQ lower bounds.*

In the following paragraphs, we provide a more detailed technical motivation of the regimes we study followed by a detailed description of our results.

Massart Learning of *General* Halfspaces Suppose that the Massart noise rate η is a constant *strictly smaller* than $1/2$. Even for this "low-noise" regime, all previous efficient learning algorithms that achieve error $\mathrm{OPT} + \epsilon$ require that the unknown halfspace is *homogeneous*. Superficially, this might seem like an innocuous assumption. After all, it seems straightforward to reduce a general halfspace to a homogeneous one by adding an extra constant coordinate to every sample. Given this, one could use a learner for the homogeneous case on the modified instance. It turns out that this intuition is fundamentally flawed. While such a reduction is valid in the distribution-independent setting, it does not work in the distribution-specific setting because it alters the marginal distribution on the examples. In light of this state of affairs, it is natural to ask whether an efficient Massart learner exists for *general* halfspaces in the low-noise noise regime where previous algorithms succeeded.

Question 1.3. What is the complexity of learning *general* halfspaces in the *constant-bounded* Massart noise setting, i.e., when $\eta = 1/2 - c$ for some universal constant $c > 0$?

As we will show, the complexity of the problem in this regime is characterized by the "bias" γ of the target halfspace (see Definition 1.5) and (inherently) scales *quasi-polynomially* with $1/\gamma$ (Theorems 1.6 and 1.7). We note that previous algorithms only handle the special case of $\gamma = 1/2$.

Learning Halfspaces with *General* Massart Noise The original definition of the Massart noise model [MN06] allows the upper bound on the noise rate to be *equal* to $1/2$. On the other hand, all known Massart learning algorithms crucially rely on the assumption that $\eta < 1/2$. The latter assumption might have been motivated by the random classification noise model, where the $\eta = 1/2$ regime is not meaningful. In the Massart model, however, it may well be the case that $\eta(\mathbf{x}) = 1/2$, for a small probability subset of the domain and $\eta(\mathbf{x})$ is small otherwise, in which case the optimal misclassification error OPT will be similarly small.

Understanding the complexity of learning halfspaces with *general* Massart noise, i.e., in the regime where $\eta = 1/2$, is both of theoretical and practical significance. From the theoretical standpoint, the "high-noise" regime subsumes the well-studied Tsybakov model [MT99, Tsy04] and can be viewed as the strongest known non-adversarial noise model in the literature. Interestingly, the general Massart noise model has also been previously studied in the statistical learning theory literature under the name *benign noise*, see, e.g., [Han09, HY15]. In the "benign noise" model, the only assumption made about the label noise is that the *Bayes optimal classifier* lies in the target class. (It is an easy exercise, see Appendix B, that the benign noise model is equivalent to the general Massart model.) In addition to its theoretical interest, the general Massart model naturally arises in a number of practical applications. A concrete example is that of *human annotator noise* [CM84, KB09, BK09, KB10], where it has been shown [KB10] that human annotators (especially non-experts) often flip coins (corresponding to $\eta(\mathbf{x}) = 1/2$) when presented with a hard-to-classify example. The preceding discussion motivates the following question:

Question 1.4. What is the complexity of learning halfspaces with *general* Massart noise?

Interestingly, prior to this work, this question remained wide-open even for the special case of homogeneous halfspaces. Specifically, previous Massart learning algorithms (for homogeneous halfspaces) provide *vacuous guarantees* for $\eta = 1/2$, because they require sample complexity scaling polynomially with the parameter $1/\beta$, where $\beta := 1 - 2\eta$. It is worth noting that a dependence on β is not information-theoretically required for the problem. Specifically, $O(d/\epsilon^2)$ samples suffice to achieve error $\mathrm{OPT} + \epsilon$ (see, e.g., [MN06]), alas with an exponential time algorithm.

At a high-level, the β-dependence in previous approaches is due to the fact that previous algorithms solve the (harder) *parameter recovery* problem, i.e., they approximate the hidden weight vector $\mathbf{w}^*$ (e.g., within small angle). While the sample complexity of PAC learning (Definition 1.1) is independent of β, parameter learning requires at least $1/\beta$ samples. Consequently, a genuinely new approach is required to handle the $\beta = 0$ regime. That is, to answer Question 1.4, we need to understand to what extent it is possible to PAC learn without relying on parameter recovery.

As we will show, handling the $\beta = 0$ case comes at a cost. Our results (Theorems 1.9 and 1.10) establish an information-computation tradeoff for this problem (that persists even for homogeneous halfspaces) scaling *quasi-polynomially* with the parameter $1/\epsilon$.

1.3 Our Results

In this work, we answer Questions 1.3 and 1.4 by providing both efficient learning algorithms and nearly-matching lower bounds in the Statistical Query (SQ) model. Perhaps surprisingly, we show that the complexity of our learning problem scales quasi-polynomially with $1/\epsilon$, where ϵ is the excess error. A conceptual implication of our findings is that Questions 1.3 and 1.4, while seemingly orthogonal, are in fact intimately connected at a technical level.

Learning with Constant-Bounded Massart Noise Here we address the regime of learning general halfspaces with η-Massart noise, where $\eta \le 1/2 - c$ for some constant $c > 0$. It turns out that the complexity of the problem in this setting depends on the "bias" of the target halfspace.

Definition 1.5 ($(1-\gamma)$-biased Halfspace). *For $\gamma \in [0, 1/2]$, we say that the halfspace f is at most $(1-\gamma)$-biased (with respect to $\mathcal{D}$) if* $\min(\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) = +1], \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) = -1]) \ge \gamma$.

Note that homogeneous halfspaces correspond to the special case of $\gamma = 1/2$. Recall that prior literature gave $\mathrm{poly}(d/\epsilon)$-time learners for homogeneous Massart halfspaces with $\eta < 1/2$. Our first main result is a learning algorithm for general halfspaces whose complexity scales quasi-polynomially with $1/\gamma$. More specifically, we establish the following theorem (see also Theorem 4.1):

Theorem 1.6 (Learning General Massart Halfspaces with Constant-Bounded Noise). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal that satisfies the constant-bounded Massart noise condition with respect to an at most $(1-\gamma)$-biased halfspace. There exists an algorithm that draws $N = d^{O(\log(\min(1/\gamma, 1/\epsilon)))}\mathrm{poly}(1/\epsilon)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and computes a halfspace h such that with high probability it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \le \mathrm{OPT} + \epsilon$.*

Qualitatively, Theorem 1.6 yields a $\mathrm{poly}(d/\epsilon)$-time algorithm for halfspaces with bias bounded away from zero, specifically $\gamma \ge c$ where $c > 0$ is an absolute constant. Furthermore, for general halfspaces it yields a PTAS with quasi-polynomial dependence on $1/\epsilon$, i.e., with runtime $d^{O(\log(1/\epsilon))}$.

Perhaps surprisingly, we provide strong evidence that the upper bound of Theorem 1.6 is best possible for any value of γ. Specifically, we prove a matching lower bound in the Statistical Query (SQ) model of [Kea98]. We recall that SQ algorithms are a broad class of algorithms that are only allowed to query expectations of bounded functions of the distribution rather than directly access samples; see Section 6.1 for the definition and additional discussion. Formally, we prove (see also Theorem 6.11):

Theorem 1.7 (SQ Lower Bound for Massart Halfspaces with Constant-Bounded Noise). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal that satisfies the Massart noise condition with parameter $\eta = 1/2 - \Omega(1) < 1/2$ with respect to an at most $(1-\gamma)$-biased halfspace. For any $\gamma > \epsilon$, any SQ algorithm that for any such distribution $\mathcal{D}$ learns a hypothesis $h : \mathbb{R}^d \mapsto \{\pm 1\}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \le \mathrm{OPT} + \epsilon$, either requires queries with tolerance at most $d^{-\Omega(\log(1/\gamma))}$ or makes at least $2^{d^{\Omega(1)}}$ statistical queries.*

Informally, Theorem 1.7 shows that no SQ algorithm can learn the subclass of $(1-\gamma)$-biased halfspaces in the constant-bounded Massart noise model with sub-exponential in $d^{\Omega(1)}$ many queries, unless it uses queries of very small tolerance – that would require at least $d^{\Omega(\log(1/\gamma))}$ samples to simulate (as long as $\gamma > \epsilon$[1]). This "fine-grained" lower bound that can be viewed as an information-computation tradeoff for the problem (within the class of SQ algorithms) and matches the upper bound of Theorem 1.6. As a corollary, we obtain that learning general halfspaces with constant-bounded Massart noise has SQ complexity $d^{\Theta(\log(1/\epsilon))}$.

Theorems 1.6 and 1.7 together qualitatively characterize the complexity of learning general halfspaces in the constant-bounded Massart setting. We view this inherent quasi-polynomial dependence as rather surprising. Even though the class of general halfspaces has one additional parameter compared to the homogeneous case (the unknown threshold), the learning problem becomes harder and exhibits a quasi-polynomial dependence on the inverse of the bias parameter γ.

[1]Of course, if $\gamma \le \epsilon$, one of the two constant functions suffices.

Remark 1.8. While the algorithm of Theorem 1.6 is essentially optimal in the constant-bounded Massart regime, its running time has quasi-polynomial dependence on $1/\beta$, where $\beta := 1 - 2\eta$. As a result, the algorithm is not optimal (as far as we know) when the parameter η is *very* close to $1/2$, e.g., $\eta = 1/2 - 1/d$. Characterizing the complexity of the learning problem when η is very close to (but not equal to) $1/2$ is left as an open problem for future work.

Learning with *General* Massart Noise Our second main result essentially characterizes the complexity of learning halfspaces with general Massart noise, i.e., the $\eta = 1/2$ case. On the positive end, we develop an algorithm for this general setting with the following guarantees.

Theorem 1.9 (Learning General Halfspaces with General Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal and such that $\mathcal{D}$ satisfies the Massart noise condition for $\eta = 1/2$ with respect to a general target halfspace. There is an algorithm that draws $N = d^{O(\log(1/\epsilon))}$ samples from $\mathcal{D}$, runs in time $2^{\mathrm{poly}(1/\epsilon)}\mathrm{poly}(N, d)$, and computes a halfspace h such that with high probability $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$.*

We reiterate that no nontrivial algorithm was known in the general Massart model, even for homogeneous halfspaces (our algorithm works for general halfspaces). Qualitatively, Theorem 1.9 gives a PTAS with runtime $O_\epsilon(1)d^{O(\log(1/\epsilon))}$ for the general problem. Note that if $\epsilon \geq 1/\log^{\Omega(1)}(d)$, the runtime of our algorithm is $d^{O(\log(1/\epsilon))}$.

It is worth comparing Theorem 1.9 with the algorithm of Theorem 1.6, which handles the case where $\eta < c$ for some constant $c < 1/2$. The latter result yields a $\mathrm{poly}(d/\epsilon)$ time algorithm when the bias of the target halfspace is a positive constant. On the other hand, the algorithm of Theorem 1.9 has a quasi-polynomial dependence in $1/\epsilon$, even for the subclass of homogeneous halfspaces.

Perhaps surprisingly, we establish an SQ lower bound suggesting that this quasi-polynomial dependence is necessary in the general Massart model, even for homogeneous halfspaces.

Theorem 1.10 (SQ Lower Bound for Homogeneous Halfspaces with General Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal and such that $\mathcal{D}$ satisfies the Massart noise condition for $\eta = 1/2$ with respect to a homogeneous halfspace. Then any SQ algorithm that for any such distribution $\mathcal{D}$ outputs a hypothesis h with $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, either requires queries with tolerance at most $d^{-\Omega(\log(1/\epsilon))}$ or makes at least $2^{d^{\Omega(1)}}$ statistical queries.*

(For a more detailed statement, see Theorem 6.21.) Informally, Theorem 1.10 shows that no SQ algorithm can learn homogeneous halfspaces in the general Massart model with sub-exponential in $d^{\Omega(1)}$ many queries, unless it uses queries of very small tolerance – that would require at least $d^{\Omega(\log(1/\epsilon))}$ samples to simulate. Note that the sample complexity of the algorithm establishing Theorem 1.9 (which can be implemented in the SQ model) matches our SQ lower bound. Furthermore, this implies that the runtime of our algorithm is essentially optimal (within SQ algorithms) for all $\epsilon \geq 1/\log^{\Omega(1)}(d)$.

Remark 1.11. It is worth noting that Theorems 1.9 and 1.10 provably separate the SQ complexity of learning halfspaces with general Massart noise from the SQ complexity of the corresponding agnostic learning problem. For agnostically learning halfspaces under Gaussian marginals, the L_1-regression algorithm [KKMS08] has complexity $d^{O(1/\epsilon^2)}$ (and is known to be implementable in SQ). Moreover, a matching SQ lower bound of $d^{\Omega(1/\epsilon^2)}$ is known [DKPZ21]. That is, while agnostic learning requires runtime $d^{\mathrm{poly}(1/\epsilon)}$, Massart learning can be achieved in time $O_\epsilon(1)\, d^{O(\log(1/\epsilon))}$.

Remark 1.12. Interestingly, the Massart model remains meaningful even for the "very large" noise setting, where $\eta > 1/2$. For this extreme regime, it is not hard to show (see Appendix A) that this model becomes *equivalent* to the agnostic model.

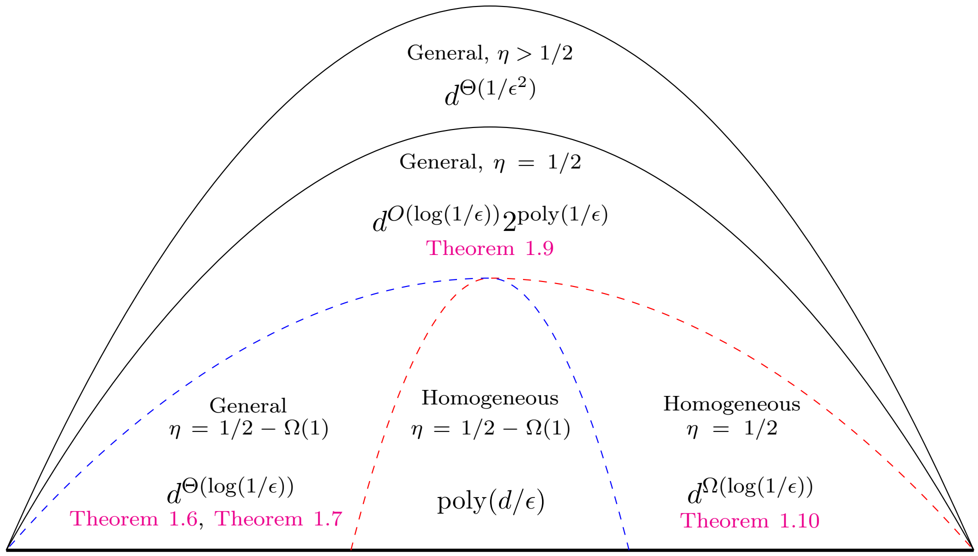


Figure 1: Overview of the (SQ) complexity of learning halfspaces with Massart noise. (1) For homogeneous halfspaces and $\eta < 1/2$, efficient algorithms were previously known under Gaussian or isotropic log-concave marginals. (2) For $\eta > 1/2$, the problem is equivalent to agnostic learning, for which tight upper and lower bounds were previously known. (3) The remaining regimes (general halfspaces and/or $\eta = 1/2$) are characterized in the current paper.

1.4 Brief Overview of Techniques

Leveraging Low-Degree Moments The unifying theme of both our upper and lower bound techniques is the use of low-degree moments. For our purposes, the low-degree moments of a distribution correspond to the values of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y]$ for low-degree polynomials p. It is clear that one can compute approximations of the (up to) k-degree moments of such a distribution with $\mathrm{poly}(d^k/\epsilon)$ samples and time. At a high-level, we are looking for a moment that will provide us with information about the correlation between $\mathbf{x}$ and y. In more detail, we would like to find polynomials p such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y] \neq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{x})]\, \mathbf{E}_{y\sim\mathcal{D}_y}[y]$, or equivalently find polynomials p with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{x})] = 0$ and $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y] \neq 0$. If no such polynomial exists, we can leverage this fact to prove SQ lower bounds; if such polynomials exist, we can hope to use them as a starting point for an algorithm.

Learning with Constant-Bounded Massart Noise: Theorem 1.6 At a high-level, our algorithm fits in the certificate-based framework developed in [DKTZ20b, DKK+20, DKK+21b]. (A similar framework was developed independently in [CKMY20] to properly learn Massart halfspaces in the distribution-free setting, matching the error of the [DGT19] algorithm.) The framework relies on the following fact: if the true halfspace is given by $f(\mathbf{x}) = \mathrm{sign}(\ell^*(\mathbf{x}))$, where $\ell^*(\mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x} - t^*$, then for any affine function $\ell(\mathbf{x})$ we have that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})yT(\mathbf{x})] \geq 0$ for all non-negative functions $T(\mathbf{x})$ if and only if $\ell = \ell^*$. We can think of this as an (infinite) linear program that can be used to solve for $\mathbf{w}^*$ and t^*. In order to solve this program, we need a separation oracle. In particular, for any hypothesis $\mathbf{w}$ and t that is too far from the truth, i.e., $\mathrm{sign}(\ell(\mathbf{x}))$ has error greater than $\mathrm{OPT} + \epsilon$, we need to be able to find an explicit non-negative function $T(\mathbf{x})$ such that the above constraint is violated. This essentially amounts to finding a function $T(\mathbf{x})$ that concentrates on the values of $\mathbf{x}$ for which our current hypothesis $\ell(\mathbf{x})$ is incorrect.

Our main new idea here is that for any sub-optimal halfspace guess h, there exists a certificate function of the form $T(\mathbf{x}) = \mathbb{1}_S(\mathbf{x})e^{\mathbf{v}\cdot\mathbf{x}}$ (*exponential-shift certificate*), where S is a thin strip around the current (known) halfspace h and $\mathbf{v}$ is an appropriate weight vector. However, finding a good weight vector $\mathbf{v}$ is a non-convex (hard) optimization problem in general. We simulate the behavior of the exponential-shift certificate by taking $T(\mathbf{x})$ to be the square of a low-degree polynomial $q(\mathbf{x})$ restricted to an appropriately chosen thin strip S. To construct our polynomial certificate, we essentially need to prove that for any at most $(1-\gamma)$-biased one-dimensional threshold function f, there exists a polynomial q of degree $O(\log(1/\gamma))$ such that $q^2(\mathbf{x})$ concentrates on the positive values of f; see Lemma 2.3. To prove the existence of such a polynomial, we establish that considering the $O(\log(1/\gamma))$-degree Taylor approximation of the exponential-shift $e^{\mathbf{v}\cdot\mathbf{x}}$ suffices; see Lemma 4.5. We can efficiently compute a polynomial certificate since, once we have a fixed strip S, we can compute the low-degree moments of y restricted to S. From there, finding an appropriate polynomial q amounts to finding a negative eigenvalue of a quadratic form. For more details, we refer the reader to Subsection 2.1 and Section 4.

Learning with *General* Massart Noise: Theorem 1.9 In the case of learning with *general* Massart noise ($\eta = 1/2$), the certificate approach fails, as it requires polynomials of large degree — polynomial in $1/\epsilon$ — to get concentration in the disagreement region. This is prohibitively large as it would result in a runtime of $d^{\mathrm{poly}(1/\epsilon)}$, which we can readily get by simply running the agnostic learner of [KKMS08]. The main idea to obtain an improved bound is to bypass the limitations of the certificate approach by relying on *correlation* properties rather than *concentration*. As we show, there is always some polynomial p of low degree, logarithmic in $1/\epsilon$, that correlates with the label y, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y] > 0$. Our algorithm exploits this weaker property and turns it into an iterative process that constantly improves the current guess.

Since low-degree polynomials are able to get non-trivial correlation, if we compute the low-order moment tensors of y, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x}^{\otimes k}y]$, we can use them to compute a low-dimensional subspace V onto which $\mathbf{w}^*$ has non-trivial ($\mathrm{poly}(\epsilon)$) projection; see Proposition 5.9 and Lemma 5.10. This means that picking a random element $\mathbf{v} \in V$ will, with reasonable probability (say $1/3$), have non-trivial, i.e., $\mathrm{poly}(\epsilon)$, correlation with $\mathbf{w}^*$. Our algorithm improves an initial guess $\mathbf{w}$ of the optimal weight vector as follows: by applying the above technique to a thin strip perpendicular to our current guess $\mathbf{w}$, we can – with some non-trivial probability – find a vector that correlates non-trivially with the projection of $\mathbf{w}^*$ on the orthogonal complement of $\mathbf{w}$. This, in turn, will allow us to compute a new guess $\mathbf{w}'$ with a slightly better correlation with $\mathbf{w}^*$; see Lemma 5.13. Repeating this process $\mathrm{poly}(1/\epsilon)$ times will produce a vector with at most $\mathrm{OPT} + \epsilon$ error; see Section 2.2. Each iteration of the above algorithm requires $d^{O(\log(1/\epsilon))}$ samples and time to compute the moments of order $O(\log(1/\epsilon))$. We then need to run the algorithm many times to find a trial in which we get lucky for $\mathrm{poly}(1/\epsilon)$ rounds in a row. This latter operation increases the complexity by a $2^{\mathrm{poly}(1/\epsilon)}$ factor.

The crucial structural result that we exploit is that for any ϵ-biased halfspace $f(\mathbf{x})$ we can construct a *mean-zero* polynomial p, which is a function of $\mathbf{w}^* \cdot \mathbf{x}$, and matches the sign of $f(\mathbf{x})$ everywhere, see (Informal) Proposition 2.8 and Subsection 5.1. We then show that, even when Massart noise with $\eta = 1/2$ is applied to f, this polynomial p will achieve non-trivial correlation with f, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y] > 0$. This implies that some low-order moment of our distribution must have a non-trivial component in the $\mathbf{w}^*$-direction; see Subsection 5.3. Flattening the moment tensors and performing SVD on the resulting matrices, we can efficiently construct a subspace (spanned by the top eigenvectors) onto which $\mathbf{w}^*$ has non-trivial projection.

SQ Lower Bounds: Theorems 1.7 and 1.10 The connection between low-degree moments and the correlation between $\mathbf{x}$ and y is even more evident when trying to establish SQ lower bounds. Using the framework introduced in [DKS17], we can show the following: if there exists a Massart halfspace whose degree-k moments have $\mathbf{x}$ independent of y, then distinguishing a random rotation of this distribution from one in which $\mathbf{x}$ is *truly* independent of y is roughly $d^{\Omega(k)}$-hard in the SQ model. Thus, proving SQ lower bounds amounts to finding Massart halfspace distributions that fool low-order moments in this way.

We note that the moment-matching and noise conditions amount to a system of linear inequalities in terms of the noise function $\eta(\mathbf{x})$. To construct our hard examples, we use linear programming (LP) duality to show the existence of solutions to this system. It should be noted that LP duality has been previously used to provide SQ lower bounds for agnostic learning of halfspaces, see, e.g., [DFT+15, DKPZ21]. In the current setting, however, we have to construct instances that satisfy the much more restrictive (constant-bounded) Massart noise assumption.

To achieve this, we show that it is possible to add constant-bounded Massart noise to a classifier $f(\mathbf{x})$ in order to cause it to fool polynomials of degree at most k if and only if there is no polynomial p of degree at most k with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[p(\mathbf{x})] = 0$, such that the sign of $p(\mathbf{x})$ agrees everywhere with the optimal halfspace $f(\mathbf{x})$. In order to prove that there exists no such low-degree, zero-mean, sign-matching polynomial, we establish a much more general result: the sign of any low-degree, zero-mean polynomial $p(\mathbf{x})$ cannot be too biased, and therefore $p(\mathbf{x})$ cannot match the sign of any heavily-biased function; see Lemma 6.18. In particular, if f is a halfspace with bias $(1-\gamma)$, this is true for all zero-mean polynomials of degree k up to (roughly) $\log(1/\gamma)$, implying the SQ lower bound of Theorem 1.7. We remark that our main structural lemma (Lemma 6.18) can readily be used to establish SQ lower bounds for Massart learning other geometric concept classes such as intersections of two homogeneous halfspaces.

It remains to provide a proof overview of Theorem 1.10 for learning homogeneous halfspaces with *general* Massart noise ($\eta = 1/2$). Interestingly, instead of adapting the methodology described in the preceding paragraphs, we prove Theorem 1.10 via a "reduction" to our SQ lower bound of Theorem 1.7, i.e., for learning (general) halfspaces with *constant-bounded* Massart noise. At a high-level, this is achieved by using the noise to "wash out" any information except when $\mathbf{x}$ lies in a thin strip not passing through the origin. Restricted to this strip, $f(\mathbf{x})$ is now a non-homogeneous halfspace, and our lower bound for general halfspaces applies.

1.5 Organization

The structure of this paper is as follows: In Section 2, we provide a detailed overview of our techniques. In Section 3, we introduce the required notation and preliminaries. In Section 4, we give our algorithm for learning general halfspaces with constant-bounded Massart noise, establishing Theorem 1.6. In Section 5, we give our algorithm for learning halfspaces under general Massart noise, establishing Theorem 1.9. Finally, in Section 6, we prove our SQ lower bounds, Theorem 1.7 and Theorem 1.10.

2 Detailed Technical Overview

2.1 Learning General Halfspaces with Constant-Bounded Massart Noise: Theorem 1.6

Our learning algorithm for this regime leverages the certificate framework developed in [DKTZ20b, DKK+20, DKK+21b]. This framework makes essential use of the Massart noise condition, i.e., the fact that $\eta(\mathbf{x}) \le 1/2$ for every $\mathbf{x} \in \mathbb{R}^d$. Using this fact, we have the following characterization of the affine function $\ell^*(\mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x} - t^*$ defining the target halfspace $f(\mathbf{x}) = \mathrm{sign}(\ell^*(\mathbf{x}))$. For every non-negative function $T(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}^+$, it holds that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell^*(\mathbf{x}) y \; T(\mathbf{x})] = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell^*(\mathbf{x}) \mathrm{sign}(\ell^*(\mathbf{x}))(1 - 2\eta(\mathbf{x})) \; T(\mathbf{x})] \ge 0 \,.$$

On the other hand, when $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \ne y] \ge \mathrm{OPT} + \epsilon$, there exists a non-negative function $T(\mathbf{x})$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x}) y \; T(\mathbf{x})] < 0$. We say that such a T is a "certifying" function (or simply a certificate) for the guess $\ell(\mathbf{x})$, because it proves that $\ell(\mathbf{x})$ is not optimal.

Definition 2.1 (Certificate Oracle)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal satisfying the Massart noise condition with respect to some halfspace. Fix $\epsilon, \delta \in (0, 1]$. Given any affine function $\ell(\mathbf{x})$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \ne y] \ge \mathrm{OPT} + \epsilon$, a ρ-certificate oracle returns a non-negative function $T : \mathbb{R}^d \mapsto \mathbb{R}_+$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x}) y \; T(\mathbf{x})] \le -\rho \|\ell(\mathbf{x})\|_2$.*

In [DKTZ20b, DKK+20, DKK+21b], it was shown that under Massart label noise, the problem of efficiently learning an optimal halfspace can be reduced to the problem of efficiently computing certifying functions. At a high-level, the certifying functions can be viewed either as separation oracles for a convex program (that can be solved via the ellipsoid method) or as loss functions in an online convex optimization problem (that can be solved via online gradient descent). In our setting, we need to adapt the proofs of [DKTZ20b] slightly to work for general halfspaces. For completeness, we provide the details of this reduction in Subsection 4.3.

Proposition 2.2. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal. Assume that $\mathcal{D}$ satisfies the η-Massart noise condition with respect to some halfspace. Fix $\epsilon, \delta \in (0, 1)$. Given a ρ-certificate oracle $\mathcal{O}$ with runtime $T_{\mathcal{O}}(\rho)$, there exists an algorithm that makes $M = \mathrm{poly}(\frac{d}{\epsilon\rho})$ calls to $\mathcal{O}$, draws $N = \mathrm{poly}(\frac{d}{\epsilon\rho}) \log(1/\delta)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(d, N, M) T_{\mathcal{O}}(\rho)$ time and computes a hypothesis h, such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \ne y] \le \mathrm{OPT} + \epsilon$, with probability $1 - \delta$.*

Given Proposition 2.2, it remains to construct an efficient certificate oracle. The first step is to restrict our search over some *parametric class* of non-negative functions.

Non-Continuous Certificates When we do not restrict our search to continuous functions, we can use certificates of the form $T(\mathbf{x}) = \mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \ne \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} - b)\}$, for some $\mathbf{v} \in \mathbb{R}^d, b \in \mathbb{R}$. It is not hard to see (Appendix C) that, for any $\eta \in [0, 1/2]$, taking $T(\mathbf{x}) = \mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \ne f(\mathbf{x})\}$, i.e., the indicator of the disagreement region of $\mathrm{sign}(\ell(\mathbf{x}))$ and $f(\mathbf{x})$, we obtain that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x}) y \; T(\mathbf{x})] \le -\Omega(\epsilon^2) \|\ell(\mathbf{x})\|_2$. Observe that, when we consider non-continuous certificates, there exist $\mathrm{poly}(\epsilon)$-certificates independent of both the noise level η and the bias of the halfspace γ. However, finding such certificates is computationally hard in general. In [DKK+20], the authors provided a polynomial-time oracle for non-continuous certificates, i.e., subsets of the disagreement region, under two crucial assumptions. Specifically, they assumed that (1) the target halfspace is homogeneous and (2) the distribution $\mathcal{D}$ satisfies the Tsybakov noise condition, which is significantly weaker than the 1/2-Massart noise assumption. In particular, the technique of [DKK+20] to isolate a subset of the disagreement region does not work when we relax either of the above assumptions. In what follows, we describe an efficient certificate oracle for general halfspaces under the assumption that $\eta = 1/2 - \Omega(1)$, i.e., in the constant-bounded Massart noise regime.

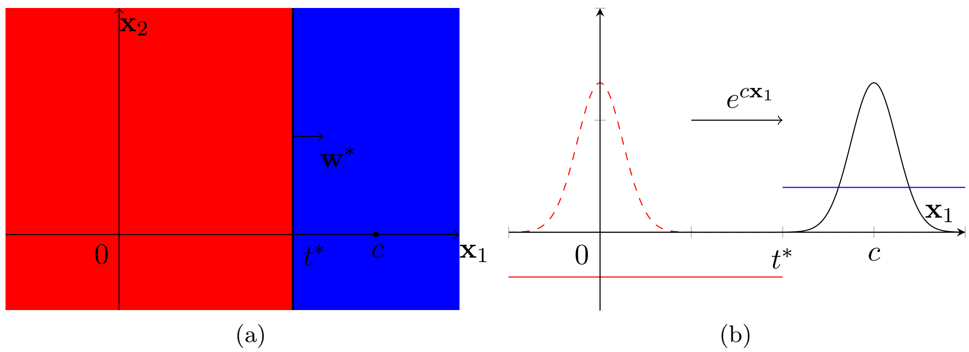


Figure 2: **(a)** An instance where $f(\mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x} - t^*$ is very biased and the constant hypothesis $\ell(\mathbf{x}) = -1$ agrees with $f(\mathbf{x})$ "almost everywhere", i.e., in the red region. A non-negative certifying function against the hypothesis -1 must put significantly more weight to the disagreement region (colored in blue). Notice that by Gaussian concentration we have that $t^* = O(\sqrt{\log(1/\gamma)})$.
(b) The certifying function $T(\mathbf{x}) = e^{c\ (\mathbf{w}^* \cdot \mathbf{x})}$ essentially moves the mean of the Gaussian from $\mathbf{0}$ to $c\mathbf{w}^*$. Choosing $c = \Theta(\sqrt{\log(1/\gamma)})$ implies that the mass of the blue region with respect to the shifted normal $\mathcal{N}(c\mathbf{w}^*, \mathbf{I})$ will be much larger than the mass of the red region, making Equation (1) true.

The Exponential Shift Certificate In order to get a handle on the optimization problem of finding a certificate, the first step is to consider smooth function classes. We will show how to construct a certificate function against the constant guess -1, when the true halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} - t^*)$ is $(1-\gamma)$-biased: the probability of the negative region is very large $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x}) = -1] = 1 - \gamma$; see Figure 2a. Finding certifying functions against the constant guess -1 captures many of the challenges of the general case. The idea is to find a continuous function $T(\mathbf{x})$ that puts more weight on the disagreement region of $f(\mathbf{x})$ (colored blue in Figure 2a) than the agreement region (colored red in Figure 2a). In particular, in order to find a certificate against the constant guess -1, we want to find some function $T(\mathbf{x})$ so that the following ratio is a sufficiently large constant (greater than $1/\beta$):

$$\frac{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[T(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) = +1\}]}{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[T(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) = -1\}]} \tag{1}$$

Our key idea is to use an exponential-shift certificate of the form $T(\mathbf{x}) = e^{\mathbf{v}\cdot\mathbf{x}}$. By multiplying the Gaussian density with the exponential function $T(\mathbf{x})$ we essentially shift the mean of the Gaussian from $\mathbf{0}$ to $\mathbf{v}$. Setting $\mathbf{v}$ to be a large multiple of $\mathbf{w}^*$, we can re-center the Gaussian of the ratio of Equation (1) to lie well within the disagreement (blue) region; see Figure 2b. In order to make the ratio of Equation (1) sufficiently large, it suffices to set $\mathbf{v} = c\ \mathbf{w}^*$ for $c = \Theta(\sqrt{\log(1/\gamma)})$; see also the proof of Lemma 4.5.

The Polynomial-Shift Certificate Even though we have shown that there exists an exponential-shift certificate, it is still not easy to compute: minimizing $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})ye^{\mathbf{v}\cdot\mathbf{x}}]$ over $\mathbf{v} \in \mathbb{R}^d$ is a non-convex objective that, in general, is hard to optimize. In order to circumvent this obstacle, we consider polynomial certificates of the form $q^2(\mathbf{x})$ for some low-degree polynomial q. Such certificates were also used in [DKTZ20b] in order to learn homogeneous halfspaces with Tsybakov noise. The advantage of having $T(\mathbf{x}) = q^2(\mathbf{x})$, for some low-degree polynomial q, is that we can efficiently compute certificates via semi-definite programming (SDP); see Subsection 4.2.

As with the exponential-shift certificate, we can provide a high-level overview of some of the ideas involved by trying to construct a certificate against the constant hypothesis -1, when the optimal halfspace is almost everywhere equal to -1, i.e., $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})=-1]=1-\gamma$. Our goal in this case is to find a polynomial that assigns much more weight to the positive region (colored blue in Figure 2a) than the negative region (colored red in Figure 2a). The main structural result that enables our algorithm is the following lemma, where we show that we can achieve the same effect of the exponential-shift certificate discussed above using a low-degree polynomial (see Lemma 4.5 for a more detailed statement and proof).

Lemma 2.3 (Low-Degree Polynomial Shift)**.** *Let $f(\mathbf{x})$ be an at most $(1-\gamma)$-biased halfspace. There exists an absolute constant $C\geq 1$ such that*

$$e^{k/C-C\log(1/\gamma)}\leq \max_{p:\deg(p)\leq k}\frac{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[p^2(\mathbf{x})\mathbb{1}\{f(\mathbf{x})=+1\}]}{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[p^2(\mathbf{x})\mathbb{1}\{f(\mathbf{x})=-1\}]}\leq e^{C\;k\log k-\log(1/\gamma)/C}\,.$$

The lower bound of Lemma 2.3 implies that there exists a polynomial certificate, namely a polynomial that makes Equation (1) true, of degree $O(\log(1/\gamma))$. The upper bound implies that degree $\Omega(\log(1/\gamma))$ is essentially necessary.

It is not hard to prove the upper bound using the anti-concentration of Gaussian polynomials; see Appendix F. To prove the lower bound of Lemma 2.3, we construct a low-degree polynomial using the Taylor approximation to the exponential-shift certificate that we defined previously. For $c=\Theta(\sqrt{\log(1/\gamma)})$, we consider the *square of the Taylor expansion* $S_k(cx)$ of the function e^{cx}. Using the Gaussian concentration and the fast convergence rate of the Taylor polynomial of e^x, in Claim 4.7, we show that for degree $k=\Theta(c^2)=\Theta(\log(1/\gamma))$, $S_k^2(c\,\mathbf{w}^*\cdot\mathbf{x})$ is very close to $e^{2c\;(\mathbf{w}^*\cdot\mathbf{x})}$ in the L_1 sense. Therefore, we can use $T(\mathbf{x})=S_k^2(c\,\mathbf{w}^*\cdot\mathbf{x})$ in the ratio of Equation (1) and obtain the same guarantees (up to constant factors) with the exponential shift $e^{2c\;(\mathbf{w}^*\cdot\mathbf{x})}$ that we discussed previously.

Our certificate against general (non-constant) hypotheses is the product of (the square of) a polynomial and a band around the current guess $\ell(\mathbf{x})$, i.e., $T(\mathbf{x})=\mathbb{1}\{r_1\leq\ell(\mathbf{x})\leq r_2\}q^2(\mathbf{x})$. Notice that, since the current guess $\ell(\mathbf{x})$ is known to the certificate algorithm, in order to find a band of the form $\mathbb{1}\{r_1\leq\ell(\mathbf{x})\leq r_2\}$, we simply need to perform a brute-force search over the two thresholds r_1,r_2. Our technical contribution here is the following proposition, showing the existence of low-degree polynomial certificates for general halfspaces with constant-bounded Massart noise (see Subsection 4.1 for a more detailed statement and proof).

Proposition 2.4 (Polynomial Certificate for Halfspaces with Constant-Bounded Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$ whose $\mathbf{x}$-marginal is the standard normal. Assume that $\mathcal{D}$ satisfies the constant-bounded Massart noise condition with respect to some at most $(1-\gamma)$-biased target halfspace $f(\mathbf{x})$. Let $\ell(\mathbf{x})$ be any linear function such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x}))\neq y]\geq\mathrm{OPT}+\epsilon$. There exist $r_1,r_2\in\mathbb{R}$ and polynomial $q(\mathbf{x})$ of degree $k=\Theta(\log(1/\gamma))$ with $\|q(\mathbf{x})\|_2=1$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(\mathbf{x})y\;\mathbb{1}\{r_1\leq\ell(\mathbf{x})\leq r_2\}q^2(\mathbf{x})\right]\leq-\epsilon^2\mathrm{poly}(\gamma)\;\|\ell(\mathbf{x})\|_2\,.$$

2.2 Learning Halfspaces with General Massart Noise: Theorem 1.9

For the purpose of this description, we will assume that the target halfspace $f(\mathbf{x})=\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})$ is homogeneous. This special case captures the key ideas of our algorithm; given such a result, the generalization to general halfspaces (under general Massart noise) is fairly straightforward. (This generalization is carried out in Appendix D.) To facilitate the intuition, we present the main ideas behind the algorithm of Theorem 1.9 in the following paragraphs. Our main technical contribution in this context is the construction of a low-degree sign-matching polynomial. This is presented at the end of this subsection and in full detail in Subsection 5.1.

At a high-level, our algorithm for learning halfspaces in the $\eta = 1/2$ regime consists of a random walk on the unit d-dimensional sphere, where we iteratively perform a random step in order to update our current guess $\mathbf{w}^{(t)}$, i.e.,

$$\mathbf{w}^{(t+1)} \leftarrow \frac{\mathbf{w}^{(t)} + \lambda \mathbf{v}}{\|\mathbf{w}^{(t)} + \lambda \mathbf{v}\|_2} . \tag{2}$$

Our goal is to find a way to sample the update vector $\mathbf{v}$, so that at every step there is some non-trivial probability that we make progress towards the optimal direction $\mathbf{w}^*$. In order to make progress towards $\mathbf{w}^*$, it suffices to have an update vector $\mathbf{v}$ that correlates with $\mathbf{w}^*$ and belongs in the orthogonal complement, $\mathbf{w}^\perp$, of $\mathbf{w}$. In what follows, we will denote by $(\mathbf{w}^*)^{\perp_\mathbf{w}}$ the normalized projection of $\mathbf{w}^*$ onto the subspace $\mathbf{w}^\perp$; see Figure 3a. Given such a vector $\mathbf{v}$, we can show that there exists a step size λ such that the update rule of Equation (2) moves $\mathbf{w}^{(t)}$ closer to $\mathbf{w}^*$ by a non-trivial amount with constant probability; see Lemma 5.13. Observe that a uniformly random unit direction in $\mathbb{R}^d$ has roughly $1/\sqrt{d}$ correlation with $(\mathbf{w}^*)^{\perp_\mathbf{w}}$ with constant probability. However, in order to hit $\mathbf{w}^*$, we need to perform roughly d consecutive successful updates (see Lemma 5.13), resulting in an algorithm with $2^{d^{\Omega(1)}}$ runtime.

The main algorithmic result of this section is the following proposition, which shows that we can efficiently sample an update vector $\mathbf{v}$ that improves the current guess with non-trivial probability.

Proposition 2.5 (Correlated Update Oracle)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta = 1/2$, with respect to a target halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Let $\mathbf{w} \in \mathbb{R}^d$ be a unit vector such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$, for some $\epsilon \in (0, 1]$. There exists an algorithm that draws $N = d^{O(\log(1/\epsilon))} \log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and with probability at least $1 - \delta$ returns a distribution $\mathcal{V}$ on $\mathbb{R}^d$ such that*

$$\mathop{\mathbf{Pr}}_{\mathbf{v}\sim\mathcal{V}} \left[(\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{v} \geq \mathrm{poly}(\epsilon) \right] \geq \frac{1}{3} .$$

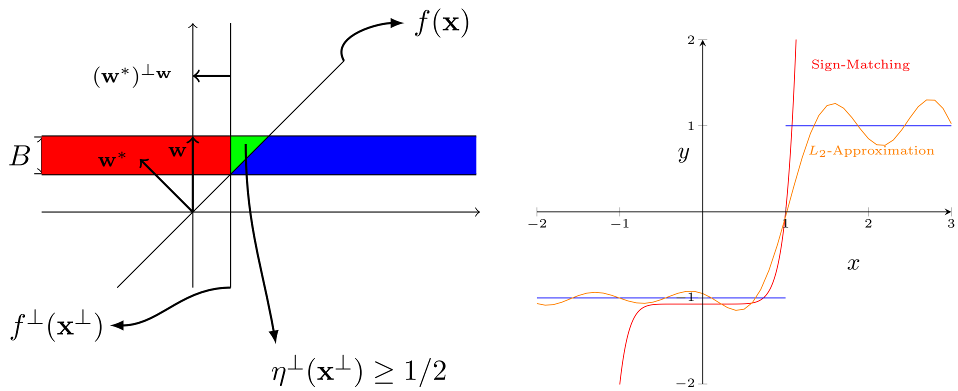
Moreover, $\mathcal{V}$ has description size $\mathrm{poly}(d/\epsilon)$ and can be sampled in $\mathrm{poly}(d/\epsilon)$ time.

(See Proposition 5.2 for a more detailed statement.) Our plan is to construct a subspace V of $\mathbb{R}^d$ such that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp_\mathbf{w}})\|_2 \geq \mathrm{poly}(\epsilon)$. To sample good update vectors $\mathbf{v}$, as claimed in Proposition 2.5, we can generate a random vector $\mathbf{v}$ on the unit sphere of V. However, we need to make sure that the dimension of V is sufficiently small, namely at most $\mathrm{poly}(1/\epsilon)$.

Improving the Constant Guess Let us assume for now that our current guess is $\mathbf{w} = \mathbf{0}$. Then, in order to make progress towards $\mathbf{w}^*$, one can simply use the degree-one Chow parameters of y, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]$. Observe that the degree-one Chow parameters have positive correlation with the optimal direction $\mathbf{w}^*$, since $(\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]) \cdot \mathbf{w}^* = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(1 - 2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\ (\mathbf{x} \cdot \mathbf{w}^*)] > 0$. Therefore, the degree-one Chow parameters are a good first update to the guess $\mathbf{w} = \mathbf{0}$.

Projecting onto $\mathbf{w}^\perp$: Subsection 5.2 In order to further improve a non-trivial guess $\mathbf{w}$, we need to find a good update direction $\mathbf{v}$ that correlates non-trivially with $(\mathbf{w}^*)^{\perp_\mathbf{w}}$. A natural attempt to do so would be to project $\mathbf{x}$ onto the orthogonal complement of the current guess, i.e., $\mathbf{w}^\perp$, and then compute the Chow parameters of the projected points. However, by doing so, the optimal classifier of the projected examples will no longer be a halfspace; see Figure 3a. In particular, the noise function $\eta^\perp(\mathbf{x}^\perp)$ after projection will be larger than $1/2$ for a large fraction of the points. To make the dataset nearly separable by a halfspace with normal vector $(\mathbf{w}^*)^{\perp_\mathbf{w}}$, we condition on a thin band and then project $\mathbf{x}$ onto $\mathbf{w}^\perp$. Doing so, apart from a small region close to the optimal classifier, $f^\perp$ (see Figure 3a), the noise function will be at most $1/2$. By making the band sufficiently thin, we can control the probability of this "high-noise" area. Notice that in the projected instance the optimal halfspace is no longer homogeneous. When the band (that we condition on) is far from the origin, the resulting optimal halfspace $f^\perp$ of the projected instance will be potentially very biased; see Figure 3a.

Assuming that our current halfspace $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})$ is at least ϵ suboptimal compared to $\mathbf{w}^*$, we show that there exists a thin band conditional on which the current hypothesis is roughly $\epsilon/\sqrt{\log(1/\epsilon)}$-suboptimal. Moreover, this band is not very far from the origin, which implies that the optimal halfspace $f^\perp$ conditional on the band will not be very biased: its threshold will be at most $O(\sqrt{\log(1/\epsilon)})$. It is worth noting that a similar orthogonal projection step onto $\mathbf{w}^\perp$ was used in [DKK+20] to learn homogeneous halfspaces with Tsybakov noise. The major difference between the setting of the current paper and [DKK+20] is that in the general Massart regime it is not possible to control the distance of the band from the origin. Specifically, it could be the case that $\eta(\mathbf{x}) = 1/2$ for all $\mathbf{x}$ close to $\mathbf{w} \cdot \mathbf{x} = 0$, forcing us to pick a band whose optimal halfspace $f^\perp$ in the subspace $\mathbf{w}^\perp$ actually has threshold $\Omega(\sqrt{\log(1/\epsilon)})$, see Figure 3a. In contrast, in [DKK+20], the "soft" Tsybakov noise condition allows for the band to be picked arbitrarily close to origin resulting in nearly homogeneous halfspaces $f^\perp$. For the details of this projection step, see Lemma 5.6. In what follows, we denote the distribution of the projected instance over $\mathbf{w}^\perp \times \{\pm 1\}$ by $\mathcal{D}^\perp$. We elaborate further on this step in Subsection 5.2.



(a) (b)

Figure 3: **(a)** After we condition on the band B, we project $\mathbf{x}$ to the subspace $(\mathbf{w}^*)^{\perp_\mathbf{w}}$ and nearly maintain the Massart noise property with respect to the biased halfspace $f^\perp(\mathbf{x}^\perp) = \mathrm{sign}((\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^\perp + b)$. In particular, it holds that $\eta^\perp(\mathbf{x}^\perp) \leq 1/2$ everywhere apart from a small area (green). Since the underlying distribution is the standard Gaussian, a band with large negative mass (blue) cannot be very far from the origin, and therefore we have that $|b| = O(\sqrt{\log(1/\epsilon)})$.
(b) The sign-matching polynomial that corresponds to the red curve does not need to closely approximate the threshold function. Its degree scales as $\Theta(b^2)$, which is at most $O(\log(1/\epsilon))$ for an ϵ-biased halfspace. On the other hand, to get an L_2 or L_1 approximation to error ϵ (orange curve), it is known that $\mathrm{poly}(1/\epsilon)$ degree is necessary.

Using the Low-Order Chow Tensors: Subsection 5.3 To obtain a good update vector $\mathbf{v}$, a natural approach is to use the degree-one Chow parameters. However, since the optimal halfspace

$f^\perp$ is biased, the degree-one Chow parameters are not guaranteed to correlate well with the direction of $f^\perp$. We thus need to look at higher-order Chow parameter tensors of $\mathcal{D}^\perp$. Recall that, to show Proposition 2.5, we want to construct a subspace V of $\mathbb{R}^d$ such that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp_\mathbf{w}})\|_2 \geq \mathrm{poly}(\epsilon)$. We now show how to find such a subspace using the low-order Chow tensors of $\mathcal{D}^\perp$.

Since we are working in Gaussian space, instead of considering the moment-tensor $\mathbf{x}^{\otimes m}$, we use the degree-m Hermite moment tensor $\mathbf{H}^m(\mathbf{x})$; this corresponds to replacing all the monomials of the tensor $\mathbf{x}^{\otimes m}$ by their corresponding Hermite monomials. (For tensor notation, we refer to Section 3.)

Definition 2.6 (Hermite Moment-Tensor)**.** *Let $\mathcal{H}$ be the linear operator that maps any d-variate monomial to the corresponding (normalized) d-variate polynomial in the Hermite basis, i.e., $\mathcal{H}(\mathbf{x}^\alpha) = h_\alpha(\mathbf{x})$. We define the degree-$m$ Hermite moment tensor as $(\mathbf{H}^m)_\alpha = \mathcal{H}((\mathbf{x}^{\otimes m})_\alpha)$.*

Using the Hermite moment-tensors, we also define the order-m Chow parameter tensors of a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{\pm 1\}$.

Definition 2.7 (Order-m Chow Tensor of $\mathcal{D}$)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal distribution. We define the order-m Chow tensor of $\mathcal{D}$ to be*

$$\mathbf{T}^m(\mathcal{D}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{H}^m(\mathbf{x})y] \,.$$

When it is clear from the context, we shall omit the distribution $\mathcal{D}$ and simply write $\mathbf{T}^m$.

Our high-level plan is to treat the above tensors as $d \times d^{m-1}$ matrices and perform SVD to find the top few left singular vectors, i.e., the singular vectors whose singular values are larger than some threshold. To show that the subspace V spanned by such eigenvectors contains a non-trivial part of $(\mathbf{w}^*)^{\perp_\mathbf{w}}$, we need to construct a *mean-zero polynomial* that only depends on the direction $(\mathbf{w}^*)^{\perp_\mathbf{w}}$ and correlates well with the label y (Lemma 5.10). This leads us to our main structural result.

The Sign-Matching Polynomial: Subsection 5.1 We show that there exists a mean-zero polynomial p that achieves non-trivial correlation with $(\mathbf{w}^*)^{\perp_\mathbf{w}}$, i.e., $\mathbf{E}_{(\mathbf{x}^\perp,y)\sim\mathcal{D}^\perp}[p(\mathbf{x})y] \geq \mathrm{poly}(\epsilon)$. Even though the noise of $\mathcal{D}^\perp$ is not exactly Massart – recall that there exists a small region where $\eta^\perp(\mathbf{x}^\perp) > 1/2$ – let us assume that $\eta^\perp(\mathbf{x}^\perp) \leq 1/2$ everywhere for simplicity. In this case, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x}^\perp,y)\sim\mathcal{D}^\perp}[p(\mathbf{x})y] = \mathop{\mathbf{E}}_{(\mathbf{x}^\perp,y)\sim\mathcal{D}^\perp_\mathbf{x}}\left[p(\mathbf{x})\mathrm{sign}((\mathbf{w}^*)^{\perp_\mathbf{w}}\cdot\mathbf{x}^\perp + b)(1-2\eta^\perp(\mathbf{x}^\perp))\right] \,.$$

Since $1-2\eta^\perp(\mathbf{x}^\perp) \geq 0$, in order to achieve non-trivial correlation it suffices to find p such that $p(\mathbf{x})$ matches the sign of $f^\perp(\mathbf{x}^\perp)$.

At this point, we made crucial use of the Massart noise condition; in particular, this is not possible in the agnostic model. In the agnostic model, to achieve non-trivial correlation, one needs to actually approximate the threshold function; see Figure 3b. More specifically, to guarantee positive correlation in the agnostic model, we need a polynomial whose L_1 error with $f(\mathbf{x})$ is $O(\epsilon)$. Unfortunately, this cannot be done for any polynomial with degree $o(1/\epsilon^2)$; see, e.g., Proposition 2.1 of [DKPZ21]. In the Massart noise setting, we show that we can construct a *zero-mean sign-matching* polynomial of degree only $\log(1/\epsilon)$ that achieves $\mathrm{poly}(\epsilon)$ correlation with y. We remark that the mean-zero condition, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{x})] = 0$ is crucial here. Non-zero mean polynomials, like the linear polynomial $(\mathbf{w}^*)^{\perp_\mathbf{w}}\cdot\mathbf{x}^\perp + b$, might give constant correlation, but do not reveal any information about the optimal direction. More concretely, we establish the following proposition; see Proposition 5.3 and Lemma 5.4 for the corresponding formal statements.

(Informal) Proposition 2.8 (Sign-Matching Polynomial). *Let $b \in \mathbb{R}$. There exists a zero mean, unit variance polynomial $p : \mathbb{R} \mapsto \mathbb{R}$ of degree $k = \Theta(b^2 + 1)$ such that the sign of p matches the sign of the threshold function* $\mathrm{sign}(z - b)$*, i.e.,* $\mathrm{sign}(p(z)) = \mathrm{sign}(z - b)$*, for all $z \in \mathbb{R}$.*

Notice that when we apply the above proposition, the threshold b of the corresponding halfspace $f^\perp$ will be at most $O(\sqrt{\log(1/\epsilon)})$ resulting in a polynomial of degree $O(\log(1/\epsilon))$; see also Figure 3a. In the case of homogeneous halfspaces, this happens because we use the orthogonal projection step; see Subsection 5.2. In the case of general halfspaces, we may have such thresholds to start with.

2.3 SQ Lower Bounds: Theorems 1.7 and 1.10

Learning General Halfspaces with Constant-Bounded Massart Noise: Theorem 1.7 Our SQ lower bounds make essential use of the "hidden-direction" framework developed in [DKS17]. Using this framework, we construct SQ lower bounds for learning halfspaces in high dimensions using a carefully constructed one-dimensional Massart noise instance. In particular, if we can construct a one-dimensional Massart noise distribution $\mathcal{D}$ on $\mathbb{R} \times \{\pm 1\}$ such that $\mathbf{E}_{(z,y)\sim\mathcal{D}}[z^k y] = \mathbf{E}_{z\sim\mathcal{D}_z}[z^i]\,\mathbf{E}_{y\sim\mathcal{D}_y}[y]$, for all $i \in [k]$, then we can obtain a family of $2^{d^{\Omega(1)}}$ distributions on $\mathbb{R}^d \times \{\pm 1\}$ whose pairwise correlation is $d^{-\Omega(k)}$. Using standard SQ lower bound arguments (see Lemma 6.6), having such a family of pairwise correlated distributions implies a $d^{\Omega(k)}$ SQ lower bound for learning the halfspace. (For a brief review on SQ lower bound machinery, we refer the reader to Section 6.1.)

Our main technical result is the following proposition (also see Proposition 6.13).

Proposition 2.9. *Fix $\eta \in (0, 1/2)$ such that $\eta = 1/2 - \Omega(1)$ and $\gamma \in (0, 1/2)$. There exists a distribution $\mathcal{D}$ on $(z, y) \in \mathbb{R} \times \{\pm 1\}$ whose z-marginal is the standard normal distribution with the following properties.*

- *$\mathcal{D}$ satisfies the η-Massart noise condition with respect to a halfspace $f(z)$ with $\mathbf{Pr}_{z\sim\mathcal{D}_z}[f(z) = +1] = \gamma$.*
- *There exists an absolute constant C such that for any integer $k \le C\log(1/\gamma)$, it holds $\mathbf{E}_{(z,y)\sim\mathcal{D}}[yz^k] = \mathbf{E}_{y\sim\mathcal{D}_y}[y]\,\mathbf{E}_{z\sim\mathcal{D}_z}[z^k]$.*

To prove the existence of the (constant-bounded) Massart noise instance of Proposition 2.9, we use (infinite-dimensional) LP duality, where our variable is the signal function $\beta(z) = 1 - 2\eta(z)$. We have the following pair of primal and dual linear programs. We denote $\mathcal{P}_k^0$ the linear space of zero-mean polynomials of degree at most k.

Primal	**Dual**
Find $\beta(z) \in L^\infty(\mathbb{R})$	Find $p(z) \in \mathcal{P}_k^0$
such that $\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)] = 0\ \forall p \in \mathcal{P}_k^0$	such that $\beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^+] > \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^-]$
$\mathop{\mathbf{Pr}}_{z\sim\mathcal{N}}[\beta \le \beta(z) \le 1] = 1$	

Using an infinite-dimensional variant of the theorem of the alternative for linear programming, to show that the primal problem is feasible, it suffices to show that the dual is infeasible. To show that the dual is infeasible, we prove a stronger statement: mean-zero polynomials of low-degree cannot match the sign of very biased Boolean functions. We prove the following (see Lemma 6.18 for the formal statement and proof).

Lemma 2.10. *Let $f : \mathbb{R} \mapsto \{\pm 1\}$ be any one-dimensional Boolean function, $\beta \in (0, 1)$ and $k \in \mathbb{Z}_+$. There exists a universal constant $C > 0$ such that if $\mathbf{Pr}_{z\sim\mathcal{N}}[f(z) = 1] \leq 2^{-Ck}(1 - \beta)$, then for any mean-zero polynomial of degree at most k it holds $\beta\,\mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^+] < \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^-]$.*

We remark that in the above lemma we do not require f to be a halfspace. Consequently, it is easy to adapt our argument to work for other Boolean concept classes that depend on a low-dimensional subspace V. For example, using the above lemma, we can obtain an SQ lower bound for learning intersections of 2 homogeneous halfspaces with constant-bounded Massart noise.

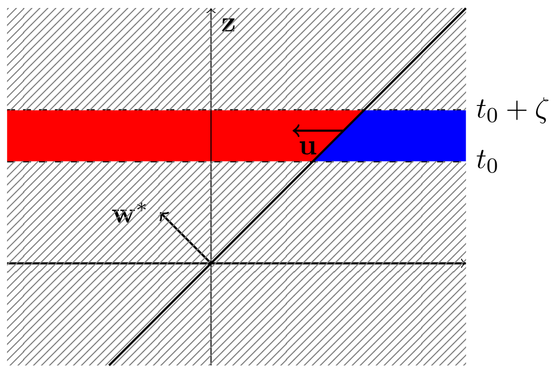


Figure 4: The "high-noise" ($\eta = 1/2$) Massart distribution $\mathcal{D}$ on $((\mathbf{x}, z), y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$ that we construct in our "reduction" from learning a biased halfspace with Massart noise. The $\mathbf{x}$-marginal of $\mathcal{D}$ is the standard normal distribution. Conditional on z, (x, y) has constant-bounded Massart noise when $z \in [t_0, t_0 + \zeta]$; this corresponds to the blue/red area. When $z \notin [t_0, t_0 + \zeta]$ we set y to be ± 1 with probability 1/2 independently of $\mathbf{x}$, i.e., $\eta((\mathbf{x}, z)) = 1/2$ in the gray area.

Learning Homogeneous Halfspaces with General Massart Noise: Theorem 1.10 Our main idea to show an SQ lower bound for learning homogeneous halfspaces with general Massart noise is to use the fact that by setting $\eta(\mathbf{x}) = 1/2$, we essentially remove all the useful signal from some parts of the space. Our "reduction" to learning halfspaces with constant-bounded Massart noise works as follows: We create a $(d + 1)$-dimensional high-noise instance, i.e., with $\eta = 1/2$, over $((\mathbf{x}, z), y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$ from many d-dimensional constant-bounded Massart noise instances. For every $(\mathbf{x}, z)$ outside of a thin slice (see Figure 4), we set $\eta(\mathbf{x}) = 1/2$. For $(\mathbf{x}, z)$ in the slice, we set the conditional distribution of $(\mathbf{x}, y)$ on z to satisfy the constant-bounded Massart noise condition with respect to some non-homogeneous optimal halfspace (denoted by $\mathbf{u}$ in Figure 4). We show that any hypothesis that achieves error $\mathrm{OPT} + \epsilon$ in the $(d + 1)$-dimensional "high-noise" instance will perform well (on average) on the d-dimensional constant-bounded Massart noise instances — a problem that we have already showed to be hard in the SQ model. We refer to Subsection 6.3 for the detailed proof.

3 Preliminaries

Basic Notation For $n \in \mathbb{Z}_+$, let $[n] := \{1, \ldots, n\}$. We 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 $\|\mathbf{x}\|_2 := (\sum_{i=1}^d \mathbf{x}_i^2)^{1/2}$ 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 slightly abuse notation and denote $\mathbf{e}_i$ 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 $\mathbf{x} \cdot \mathbf{v}$, 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 $\mathbf{A}$. For a square symmetric matrix $\mathbf{M}$, we say that $\mathbf{M}$ is positive semi-definite if only if all the eigenvalues of $\mathbf{M}$ are non-negative. For $m \in \mathbb{Z}_+$, we denote $\mathcal{S}^m$ the set of positive (symmetric) semi-definite matrices of dimension m. 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 $\mathbf{x} \notin A$.

We write $E \gtrsim F$ for two expressions E and F to denote that $E \geq c\,F$, where $c > 0$ is a sufficiently large universal constant (independent of the variables or parameters on which E and F depend). Similarly, we write $E \lesssim F$ to denote that $E \leq c\,F$, where $c > 0$ is a sufficiently small universal constant.

Probability Notation 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 $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, we denote $\phi_d(\cdot)$ the pdf of the d-dimensional Gaussian and we use the $\phi(\cdot)$ for the pdf of the standard normal. In this work we usually consider the standard normal, i.e., $\mu = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$, and therefore we denote it simply $\mathcal{N}$; the dimension will be always clear from the context. Moreover, we denote as $\mathcal{N}(\mu_1, \mu_2)$ the 2-dimensional Gaussian distribution with mean (μ_1, μ_2) and $\mathcal{N}(\mu)$ the 1-dimensional Gaussian distribution with mean μ. For $(\mathbf{x}, y)$ distributed according to $\mathcal{D}$, we denote $\mathcal{D}_\mathbf{x}$ to be the distribution of $\mathbf{x}$ and $\mathcal{D}_y$ to be the distribution of y. 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 $\mathbf{x}_\mathbf{v}$. For a set B and a distribution $\mathcal{D}$, we denote $\mathcal{D}_B$ to be the distribution D conditional on B. We define the standard L_p norms with respect to the Gaussian measure, i.e., $\|g(\mathbf{x})\|_p = (\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|g(\mathbf{x})|^p])^{1/p}$. To distinguish them from vector norms we shall write $\|g(\mathbf{x})\|_p$ instead of $\|g\|_p$ when it is not clear from the context. We also need the following fact providing upper and lower bounds on Gaussian tails and the Gaussian anticoncentation property for intervals.

Fact 3.1 (Gaussian Density Properties)**.** *Let $\mathcal{N}$ be the standard one-dimensional normal distribution. Then, the following properties hold:*

1. *For any $t > 0$, it holds $e^{-t^2}/4 \leq \mathbf{Pr}_{z\sim\mathcal{N}}[z > t] \leq e^{-t^2/2}/2$.*
2. *For any $a, b \in \mathbb{R}$ with $a \leq b$, it holds $\mathbf{Pr}_{z\sim\mathcal{N}}[a \leq z \leq b] \leq (b-a)/\sqrt{2\pi}$.*

Function Families (LTFs and Polynomials) We use $\mathcal{C}_V$ to denote 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 homogeneous 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 $\mathcal{P}_{k,d}$ the space of polynomials on $\mathbb{R}^d$ of degree at most k. We will use the following fact shows that for the probability that two halfspaces disagree can be upper bounded by the difference of their thresholds plus the angle of their corresponding normal vectors.

Fact 3.2 (see, e.g., Fact 3.5 of [DKK+21a])**.** *Let $\mathcal{N}$ be the standard normal distribution in $\mathbb{R}^d$. Let $\mathbf{v}, \mathbf{u}$ be unit vectors in $\mathbb{R}^d$ and $t_1, t_2 \in \mathbb{R}$. It holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[\mathrm{sign}(\mathbf{u} \cdot \mathbf{x} + t_1) \neq \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t_2)] \leq O(\theta(\mathbf{u}, \mathbf{v})) + O(|t_1 - t_2|)$.*

Multilinear Algebra Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ be a d-dimensional multi-index vector, where for all $i \in [d]$, α_i is non-negative integer. We denote $|\alpha| = \sum_{i=1}^d \alpha_i$ and for a d-dimensional vector $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d)$, we denote $\mathbf{w}^\alpha = \prod_{i=1}^d \mathbf{w}_i^{\alpha_i}$. An order-$k$ tensor $\mathbf{A}$ is an element of the k-fold tensor product of subspaces $\mathbf{A} \in \mathcal{V}_1 \otimes \ldots \otimes \mathcal{V}_k$. We will be exclusively working with subspaces of $\mathbb{R}^d$ so a tensor A can be represented by a sequence of coordinates, that is $A_{i_1,\ldots,i_k}$. The tensor product of order k tensor $\mathbf{A}$ and an order m tensor $\mathbf{B}$ is an order $k+m$ tensor defined as $(\mathbf{A} \otimes \mathbf{B})_{i_1,\ldots,i_k,j_1,\ldots,j_m} = \mathbf{A}_{i_1,\ldots,i_k} \mathbf{B}_{j_1,\ldots,j_m}$. Moreover, we denote by $\mathbf{A}^{\otimes k}$ the k-fold tensor product of $\mathbf{A}$ with itself. We define the dot product of two tensors (of the same order) to be $\mathbf{A} \cdot \mathbf{B} = \sum_\alpha \mathbf{A}_\alpha \mathbf{B}_\alpha$. We also denote the Frobenius norm of a tensor by $\|\mathbf{A}\|_F = \sqrt{\mathbf{A} \cdot \mathbf{A}}$. In this work, we will be using the (appropriately) normalized Hermite tensors (Definition 2.6). For these tensors we have the following fact.

Fact 3.3. *Let $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$ and let $h_m : \mathbb{R} \mapsto \mathbb{R}$ be the univariate normalized (with respect to the standard normal) degree-m Hermite polynomial and $\mathbf{H}^m$ be the Hermite moment tensor of Definition 2.6. It holds $h_m(\mathbf{x} \cdot \mathbf{v}) = \mathbf{H}^m \cdot \mathbf{v}^{\otimes m}$.*

4 A Certificate-Based Algorithm For Learning General Halfspaces with Constant-Bounded Massart Noise

In this section we show that, when the noise upper bound η is bounded away from $1/2$, we can learn at most $(1-\gamma)$-biased halfspaces with sample complexity and runtime $d^{O(\log(1/\gamma))}\mathrm{poly}(1/\epsilon)$. Our algorithmic result matches our SQ lower bound, given in Theorem 1.7. We first state the formal version of Theorem 1.6.

Theorem 4.1 (Learning $(1-\gamma)$-biased Halfspaces with Constant-Bounded Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard Gaussian. Assume that $\mathcal{D}$ satisfies the η-Massart noise condition with respect to some $(1-\gamma)$-biased optimal halfspace and $\beta = 1 - 2\eta$. Let $\epsilon, \delta \in (0, 1]$. There exists an algorithm that draws $N = d^{O(\log(1/(\gamma\beta)))}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and computes a halfspace $h \in \mathcal{C}$ such that with probability at least $1-\delta$ we have that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$.*

The main technical ingredient in our proof is Proposition 4.2 where we show that, given a linear function $\ell(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - t$ whose classification error is greater than $\mathrm{OPT} + \epsilon$, there exists a low-degree polynomial certifying function. We prove the existence of such a certificate in Subsection 4.1. In Subsection 4.2, we show that we can efficiently compute the low-degree polynomial certificate, and we bound the sample complexity and runtime of the corresponding SDP of the optimization problem. Finally, in Subsection 4.3, we show that given an efficient certificate oracle, we can use online gradient descent in order to learn the optimal halfspace, i.e., we provide the formal version and proof of Proposition 2.2.

4.1 The Low-Degree Polynomial Certificate

Here we show that given any ϵ suboptimal linear hypothesis $\ell(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - t$ there exists a low degree polynomial certifying function. We establish the following.

Proposition 4.2 (Certificate for General Halfspaces with Massart Noise: $\eta < 1/2$.)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the η-Massart noise condition with respect to some at most $(1-\gamma)$-biased optimal halfspace $f(\mathbf{x})$, where $\gamma \in (0, 1/2]$.*

Let $\ell(\mathbf{x})$ be any linear function such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x}))\neq y]\geq \mathrm{OPT}+\epsilon$. There exist $r_1,r_2\in\mathbb{R}$ and polynomial $q(\mathbf{x})=\sum_{|\alpha|\leq k}c_\alpha\mathbf{x}^\alpha$ of degree $k=\Theta(\log(\frac{1}{\gamma\beta}))$ and $\|q(\mathbf{x})\|_2=1$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y\ \mathbb{1}\{r_1\leq\ell(\mathbf{x})\leq r_2\}q^2(\mathbf{x})]\leq-\epsilon^4\mathrm{poly}(\beta\gamma)\|\ell(\mathbf{x})\|_2\,.$$

Moreover, the coefficients of q are bounded, specifically $\sum_{|\alpha|\leq k}|c_\alpha|\leq d^{O(k)}$.

To prove the proposition, we distinguish different cases for the hypothesis $h(\mathbf{x})$. First we assume that $\mathbf{w}=0$, i.e., the guess is not formally a halfspace but a constant function. This case may look easy to handle but captures many difficulties of designing certificates for biased halfspaces under Massart noise: when the constant guess is -1, we have to construct a polynomial that puts a large weight to the positive region of $f(\mathbf{x})$, i.e., $\{\mathbf{x}:f(\mathbf{x})=+1\}$. Since the probability of this region can be as small as γ, it may be far from the origin (i.e., $t^*=\Theta(\sqrt{\log(1/\gamma)})$); see Figure 2a. It can be seen that it is challenging to isolate such regions far from the origin and the required polynomial degree is roughly $\log(\frac{1}{\beta\gamma})$. The argument and the details for this case are given in Subsection 4.1.1. In Subsection 4.1.2, we show that, similarly to the case of constant hypotheses, we can construct polynomial certificates against very biased hypotheses, i.e., when $\ell(\mathbf{x})=\mathbf{w}\cdot\mathbf{x}-t$ and it holds $t/\|\mathbf{w}\|_2\geq C\sqrt{\log(1/(\beta\gamma))}$. Finally, in Subsection 4.1.3, we handle the remaining hypotheses, namely we provide certificates when $t/\|\mathbf{w}\|_2\leq C\sqrt{\log(1/(\beta\gamma))}$. Having Lemma 4.3, Lemma 4.9 and Lemma 4.13, we immediately obtain Proposition 4.2.

4.1.1 Certificate Against Constant Hypotheses

Here we show that we can construct a polynomial certificate that certifies the non-optimality of any constant hypothesis, i.e., $h(\mathbf{x})$ corresponds to the constant function $+1$ (or -1) for every $\mathbf{x}\in\mathbb{R}^d$. In this case, we do not require a "band" $\mathbb{1}\{r_1\leq\mathbf{w}\cdot\mathbf{x}\leq r_2\}$ as part of the certificate, and therefore we may set $r_1=-\infty$, $r_2=+\infty$ in Proposition 4.2. We prove the following lemma:

Lemma 4.3 (Certificate against Constant Hypotheses). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the η-Massart noise condition with respect to some (at least) $(1-\gamma)$-biased optimal halfspace. For every constant hypothesis $s\in\{-1,+1\}$, there exists a polynomial $q(\mathbf{x})=\sum_{|\alpha|\leq k}c_\alpha\mathbf{x}^\alpha$ of degree $k=\Theta(\log(\frac{1}{\beta\gamma}))$ with $\|q\|_2=1$, and sum of (absolute) coefficients $\sum_{|\alpha|\leq k}|c_\alpha|\leq d^{O(k)}$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[sy\ q^2(\mathbf{x})]\leq-\mathrm{poly}(\beta\gamma)\,.$$

Proof. Our plan is to pick the polynomial $q(\mathbf{x})$ so that it takes larger values to the side of the optimal halfspace $f(\mathbf{x})=\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}-t^*)$ that has the opposite sign of s. In the example of Figure 2a, in order to have a certificate for the constant hypothesis $s=-1$, we want to have a polynomial that takes much larger values in the blue area than the red area. We shall set $q(\mathbf{x})=p(\mathbf{w}^*\cdot\mathbf{x})$ for some one-dimensional polynomial $p:\mathbb{R}\to\mathbb{R}$. Since the certificate that we construct only depends on the direction of the optimal halfspace, it follows that we can project the points on the subspace spanned by $\mathbf{w}^*$. The following claim shows that the projection of a distribution with η-Massart noise onto a lower dimensional subspace that contains the direction of the optimal halfspace also satisfies the η-Massart noise condition with respect to the same optimal halfspace.

Claim 4.4 (Projections preserve Massart Noise). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$ satisfying the η-Massart noise condition with respect to some optimal halfspace $f(\mathbf{x}):\mathbb{R}^d\mapsto\{\pm1\}$. Let V be* *any subspace of* $\mathbb{R}^d$ *that contains the normal vector of the optimal halfspace* $f(\mathbf{x})$*. Then, for every function* $g : \mathbb{R}^d \mapsto \mathbb{R}$ *it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[g(\mathrm{proj}_V(\mathbf{x}))y] = \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_V}[g(\mathbf{v})\beta_V(\mathbf{v})f(\mathbf{v})] \,,$$

where $\beta_V(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ *satisfies* $\beta_V(\mathbf{x}) \in [1-2\eta, 1]$.

Proof. Assume that the optimal halfspace is $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} - t^*)$. From the fact that the distribution $\mathcal{D}$ satisfies the η-Massart noise condition, it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y|\mathbf{x}] = f(\mathbf{x})(1-\eta(\mathbf{x})) - f(\mathbf{x})\eta(\mathbf{x}) = f(\mathbf{x})(1-2\eta(\mathbf{x})) = f(\mathbf{x})\beta(\mathbf{x})$. Therefore, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[g(\mathrm{proj}_V(\mathbf{x}))y] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[g(\mathrm{proj}_V(\mathbf{x}))\beta(\mathbf{x})f(\mathbf{x})] \\
&= \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_V}\left[\mathop{\mathbf{E}}_{\mathbf{u}\sim(\mathcal{D}_\mathbf{x})_{V^\perp}}[g(\mathbf{v})\beta(\mathbf{u}+\mathbf{v})\mathrm{sign}(\mathbf{w}^*\cdot(\mathbf{u}+\mathbf{v})-t^*)\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_V}\left[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{v}-t^*)\; g(\mathbf{v}) \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_{V^\perp}}[\beta(\mathbf{u}+\mathbf{v})]\right],
\end{aligned}$$

where we used the fact that the subspace V contains the normal vector of the optimal halfspace $\mathbf{w}^*$. Observe that the "projected" noise function $\beta_V(\mathbf{v}) := \mathbf{E}_{\mathbf{u}\sim(\mathcal{D}_\mathbf{x})_{V^\perp}}[\beta(\mathbf{u}+\mathbf{v})]$ again satisfies the η-Massart noise condition, i.e., $\beta_V(\mathbf{v}) \in [1-2\eta, 1] = [\beta, 1]$. This concludes the proof of Claim 4.4. ☐

Since $q(\mathbf{x}) = p(\mathbf{w}^* \cdot \mathbf{x})$, we can use Claim 4.4 to project onto the subspace spanned by $\mathbf{w}^*$ and obtain that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[syq^2(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_{\mathbf{w}^*}}\left[s\;\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{v}-t^*)\; p^2(\mathbf{w}^*\cdot\mathbf{v})\; \beta_{\mathbf{w}^*}(\mathbf{v})\right], \tag{3}$$

where the "projected" noise function $\beta_{\mathbf{w}^*}(\mathbf{v}) \in [\beta, 1]$. For simplicity, in what follows, we will continue denoting β the projected noise function $\beta_{\mathbf{w}^*}$. Without loss of generality, we may assume that $\mathbf{w}^* = \mathbf{e}_1$ and in that case, using the fact that the projection of $\mathcal{D}_\mathbf{x}$ onto $\mathbf{w}^*$ is a one-dimensional standard normal distribution, we have that the expression of Equation (3) can be simplified as

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[syq^2(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[s\beta(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1-t^*)\; p^2(\mathbf{x}_1)]\,. \tag{4}$$

Moreover, to simplify the notation, assume that the constant guess is $s = -1$ (the case of $s = +1$ is similar) and that the halfspace puts γ mass on the positive side, i.e., $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f(\mathbf{x}) = +1] = \gamma$; see also Figure 2a. In that case, we want to construct a univariate polynomial $p(\mathbf{x}_1)$ so that $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[\beta(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1-t^*)\; p^2(\mathbf{x}_1)]$ is sufficiently positive. Observe that the worst case noise function $\beta(\mathbf{x}_1)$ is to set $\beta(\mathbf{x}_1) = \beta$ for all points $\mathbf{x}_1 \geq t^*$ and $\beta(\mathbf{x}_1) = 1$ for all $\mathbf{x}_1 < t^*$. In that case, we want to find a univariate polynomial p such that

$$\beta \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_1 \geq t^*\}p^2(\mathbf{x}_1)] \geq \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_1 \leq t^*\}p^2(\mathbf{x}_1)]\,. \tag{5}$$

To do that, we show our main structural result, i.e., that as long as the degree k is larger than $(t^*)^2$, there exists a polynomial that can make the above inequality true. In fact, we can make the ratio of $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_1 \geq t^*\}p^2(\mathbf{x}_1)] / \mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_1 \leq t^*\}p^2(\mathbf{x}_1)]$ grow exponentially fast with respect to the degree-k of the polynomial p. We prove the following lemma.

Lemma 4.5 (Polynomial Shift). *Let $t \in \mathbb{R}$ and $b \geq 1$. There exists a univariate polynomial $p(x) = \sum_{i=0}^{k} a_i x^i$ of degree $k = \Theta(t^2 + b^2)$, L_2 norm $\|p(x)\|_2 = 1$, and sum of (absolute) coefficients $\sum_{i=0}^{k} |a_i| = O(1)$, such that*

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \geq t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \leq t\}]} \geq e^{b^2} \quad \text{and} \quad \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \geq t\}] \geq e^{-O(k)}\,.$$

Proof. We first show that there exists a positive function that makes the ratio of Lemma 4.5 large. We shall consider the exponential function $x \mapsto e^{cx}$ for some constant $c \geq |t|$ to be determined later. We have that

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[e^{cx}\mathbb{1}\{x \geq t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[e^{cx}\mathbb{1}\{x \leq t\}]} = \frac{\mathbf{Pr}_{x\sim\mathcal{N}}[x \geq t - c]}{\mathbf{Pr}_{x\sim\mathcal{N}}[x \leq t - c]} \geq e^{(c-t)^2/2}\,, \tag{6}$$

where the last inequality follows from the fact that $c \geq t$ and standard lower bounds on the tail probability of the standard normal distribution. Therefore, for $c = \Theta(|t| + b)$ we immediately obtain the claimed bound. We now have to replace the function e^{cx} by the square of a polynomial p. To do so we use the Taylor expansion of the exponential function. Denote by $S_k(x)$ the degree-k Taylor expansion of e^x, i.e., $S_k(x) = \sum_{i=0}^{k} x^i/i!$. Using Taylor's remainder theorem and the fact that $e^x = \sum_{i=0}^{\infty} x^i/i!$ for all $x \in \mathbb{R}$, we obtain the following fact.

Fact 4.6 (Taylor Expansion of e^x). *Fix $R > 0$ and let $S_k(x)$ be the degree-k Taylor expansion of e^x.*

1. *For all $x \in [-R, R]$ we have that $|S_k(x) - e^x| \leq e^R \frac{R^{k+1}}{(k+1)!}$.*
2. *For all $x \in \mathbb{R}$ it holds that $|S_k(x)| \leq e^{|x|}$.*

Since we have to construct the square of a polynomial p we cannot simply use $S_k(x)$. However, we can consider the function e^{2cx} and use $S_k^2(cx)$ as our approximation. We first show that as long as k is at least $\Omega(c^2)$, we have that $S_k^2(cx)$ is a good approximation of e^{2cx} with respect to the L_1 norm.

Claim 4.7 (L_1 error of $S_k(x)$). *Assume that $k \geq 32c^2$. It holds*

$$\|S_k^2(cx) - e^{2cx}\|_1 \lesssim e^{-k/32}\,.$$

Proof. For every $x \in [-R, R]$ it holds that

$$|S_k^2(x) - (e^x)^2| = |S_k(x) - e^x||S_k(x) + e^x| \leq e^R \frac{R^{k+1}}{(k+1)!} \; 2e^R = 2e^{2R}\frac{R^{k+1}}{(k+1)!}\,, \tag{7}$$

where we used the second item of Fact 4.6, i.e, that $|S_k(x)| \leq e^R$. The above pointwise approximation guarantee together with the strong concentration properties of the Gaussian distribution (see, e.g., Fact 3.1) allow us to get a bound for the L_1-approximation error of $S_k^2(x)$. Using the error bound of Equation (7), we bound the L_1-approximation error as follows:

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(cx) - e^{2cx}|] &\leq \max_{|x|\leq R/c} |S_k^2(cx) - e^{2cx}| + \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(cx) + e^{2cx}|\; \mathbb{1}\{|x| \geq R/c\}] \\
&\leq 2e^{2R}\frac{R^{k+1}}{(k+1)!} + (\|S_k^2(cx)\|_2 + \|e^{2cx}\|_2)\sqrt{\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}}[|x| \geq R/c]} \\
&\leq 2e^{2R}\frac{R^{k+1}}{(k+1)!} + 2\sqrt{2}e^{4c^2-(R/c)^2/2}\,,
\end{aligned} \tag{8}$$

where for the last inequality we used the second item of Fact 4.6 to obtain that $\|S_k^2(x)\|_2 \le \|e^{2c|x|}\|_2$, the fact that $\|e^{2c|x|}\|_2 \le \sqrt{2}e^{4c^2}$, and the tail probability upper bound for the normal distribution, i.e., $\mathbf{Pr}_{x\sim\mathcal{N}}[|x| \ge (R/c)] \le e^{-(R/c)^2/2}$, see Fact 3.1. By choosing $k = 32mc^2$, for any $m \ge 1$, and $R = k/8$ the estimate of Equation (8) becomes

$$\|S_k^2(cx) - e^{2cx}\|_1 \lesssim e^{-mc^2} .$$

□

We are now ready to finish the proof of Lemma 4.5. Recall that to make Inequality (6) true, we chose $c = \Theta(b + |t|)$. Let $k = 32mc^2$, for some sufficiently large absolute constant $m > 0$. Given Inequality (6), we show that by replacing e^{2cx} with $S_k^2(cx)$, i.e., the Taylor expansion of e^{cx} squared, we get a similar formula for the lower-bound of the ratio. It suffices to show that

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \le t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[S_k^2(cx)\mathbb{1}\{x \le t\}]} \ge \frac{1}{2} \quad \text{and} \quad \frac{\mathbf{E}_{x\sim\mathcal{N}}[S_k^2(cx)\mathbb{1}\{x \ge t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \ge t\}]} \ge \frac{1}{2} . \tag{9}$$

We start from the first inequality of Equation (9). To prove it, it suffices to show that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(x) - e^{2cx}|\ \mathbb{1}\{x \le t\}] \le \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[e^{2cx}\ \mathbb{1}\{x \le t\}] . \tag{10}$$

From the L_1-bound of Claim 4.7, we have that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(x) - e^{2cx}|\ \mathbb{1}\{x \le t\}] \le \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(x) - e^{2cx}|] \lesssim e^{-mc^2} .$$

Moreover, to bound the $\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\ \mathbb{1}\{x \le t\}]$ from below, we use the lower bound for the tails of a Gaussian distribution (Fact 3.1). First, note that $\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\ \mathbb{1}\{x \le t\}] = e^{2c^2}\,\mathbf{Pr}_{x\sim\mathcal{N}}[x \le t - 2c]$. Therefore, we have that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[e^{2cx}\ \mathbb{1}\{x \le t\}] \ge \frac{1}{4}e^{2c^2}e^{-(t-2c)^2} \ge \frac{1}{4}e^{-7c^2} ,$$

where we used that $c \ge |t|$. Therefore, Inequality (10) is true for m being a large enough absolute constant.

For the second inequality of Equation (9), using the L_1-bound of Claim 4.7 we obtain that $\mathbf{E}_{x\sim\mathcal{N}}[S_k^2(cx)\mathbb{1}\{x \ge t\}] \ge \mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \ge t\}] - \|S_k^2(x) - e^{2cx}\|_1$, and therefore, taking m to be larger than an absolute constant we have that, there exists an absolute constant $C > 0$, such that

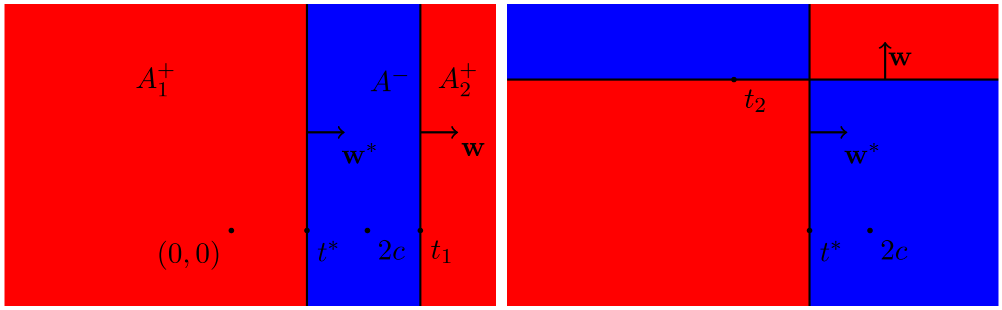
$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[S_k^2(cx)\mathbb{1}\{x \ge t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \ge t\}]} \ge 1 - \frac{Ce^{-mc^2}}{\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \ge t\}]} \ge \frac{1}{2} ,$$

where we used that $\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \ge t\}] = e^{2c^2}\,\mathbf{Pr}_{x\sim\mathcal{N}}[x \ge t - 2c] \ge e^{2c^2}/2$ since $c \ge t$.

We next show that we can normalize the polynomial $p(x) = S_k(cx)$ without making the expectation of p^2 over $x \ge t$, i.e., $\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \ge t\}]$, too small. From the second item of Fact 4.6 we obtain that $\|S_k^2(cx)\|_2 \le \|e^{2c|x|}\|_2 = e^{O(c^2)}$. The result follows from the second inequality of Equation (9) and the fact that $\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \ge t\}] \ge e^{2c^2}/2$.

Finally, we bound the coefficients of the polynomial $S_k^2(cx)$. We have $S_k^2(cx) = \left(\sum_{i=0}^k \frac{c^i}{i!}x^i\right)^2$. It holds that $\left(\sum_{i=0}^k \frac{c^i}{i!}\right)^2 = e^{2c}$. From the L_1 approximation guarantee of Claim 4.7 we obtain that $\|S_k^2(cx)\|_2 \ge \|S_k^2(cx)\|_1 \ge \|e^{2cx}\|_1 - 1 \ge e^{2c^2}/2$, since $c \ge 1$. We conclude that the sum of the absolute coefficients of $S_k^2(cx)/\|S_k^2(cx\|_2$ is at most 2, which implies that the sum of the absolute coefficients of $S_k(cx)$ is at most $\sqrt{2}$.

□



(a) (b)

Figure 5: **(a)** The guess $\mathbf{w} = \mathbf{e}_1 - t_1$ and t_1 is larger than $C\sqrt{\log(1/(\beta\gamma))}$ for some sufficiently large constant C. We can pick c such that $2c - t^*$ and $t_1 - 2c$ are both large constants making the exponential shift certificate $e^{2c\mathbf{x}_1}$ work, similarly to the constant case of Figure 2a.
(b) The guess $\mathbf{w} = \mathbf{e}_2 - t_2$ and t_1 is larger than $C\sqrt{\log(1/(\beta\gamma))}$ for some sufficiently large constant C. The region $\mathbf{x}_2 \geq t_2$ has probability at most $e^{-O(t_2^2)}$ and for $\mathbf{x}_2 \leq t_2$ we can treat $\mathbf{x}_2 - t_2$ as a constant negative guess.

To conclude the proof of Lemma 4.3, notice that using Lemma 4.5, we obtain that there exists a polynomial p of degree $(t^*)^2 + \log(1/\beta)$ such that Equation (5) is true. Moreover, from the same lemma, we obtain that $\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[p^2(\mathbf{x}_1)] = e^{-O((t^*)^2 + \log(1/\beta))}$. Since the halfspace is at least $(1-\gamma)$-biased, from standard upper bounds on the Gaussian tails, we obtain that $|t^*| \lesssim \sqrt{\log(1/\gamma)}$ obtaining the claimed bound for the degree of the certifying polynomial p. It remains to bound the coefficients of the multivariate polynomial $q(\mathbf{x}) = p(\mathbf{w}^* \cdot x)$. We are going to use the following fact.

Fact 4.8 (Lemma 3.6 of [DKTZ20b]). *Let $p(t) = \sum_{i=0}^{k} c_i t^i$ be a degree-k univariate polynomial. Given $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq 1$, define the multivariate polynomial $q(\mathbf{x}) = p(\mathbf{w} \cdot \mathbf{x}) = \sum_{S:|S| \leq k} C_S \mathbf{x}^S$. Then we have that $\sum_{S:|S| \leq k} C_S^2 \leq d^{2k} \sum_{i=0}^{k} c_i^2$.*

From Lemma 4.5 we know that the sum of the absolute coefficients of p is $O(1)$. Thus using Fact 4.8 we obtain that the sum of absolute coefficients of $p(\mathbf{w} \cdot \mathbf{x})$ is at most $d^{O(k)}$.

□

4.1.2 Certificate Against "Large Threshold" Halfspaces

We now deal with the case where the halfspace has a large threshold but is not the constant hypothesis. In particular, we assume that $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} - t)$ with $t/\|\mathbf{w}\| \geq C\sqrt{\log(1/(\beta\gamma))}$ for some sufficiently large absolute constant C. This case is, in fact, a generalization of the constant hypothesis case that corresponds to $\mathbf{w} = \mathbf{0}$. In this case we will show that roughly the same polynomial that we used for constant guesses in Subsection 4.1.1 can also be made to work for very biased guesses. We prove the next lemma.

Lemma 4.9 (Certificate against "Large Threshold" Hypotheses). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the η-Massart noise condition with respect to some (at most) $(1-\gamma)$-biased optimal halfspace. Define the linear function $\ell(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} - t)$ and assume that $t/\|\mathbf{w}\|_2 \geq C\sqrt{\log(1/(\beta\gamma))}$ for some sufficiently large absolute constant C. Then, there exists polynomial $q(\mathbf{x}) = \sum_{|\alpha| \leq k} c_\alpha \mathbf{x}^\alpha$ of degree $\Theta(\log(\frac{1}{\beta\gamma}))$, L_2 norm $\|q(\mathbf{x})\|_2 = 1$, and sum of* *absolute coefficients* $\sum_{|\alpha|\leq k}|c_\alpha| = d^{O(k)}$, *such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})yq^2(\mathbf{x})] \leq -\|\ell(\mathbf{x})\|_2 \text{ poly}(\beta\gamma)\,.$$

Proof. Using Claim 4.4, we can project the distribution $\mathcal{D}$ on the 2-dimensional subspace V spanned by $\mathbf{w}$ and $\mathbf{w}^*$. We have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})yp^2(\mathbf{w}^*\cdot\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}_V}[\ell(\mathbf{v})\beta_V(\mathbf{v})\text{sign}(\mathbf{w}^*\cdot\mathbf{v}-t^*)p^2(\mathbf{w}^*\cdot\mathbf{v})]\,. \tag{11}$$

In what follows, we again abuse notation and denote $\beta_V(\mathbf{v})$ simply as $\beta(\mathbf{v})$. By the spherical symmetry of the Gaussian distribution, we can, without loss of generality, assume that $\mathbf{w}^* = \mathbf{e}_1$ and $\mathbf{w} = \|\mathbf{w}\|_2(\cos\theta\mathbf{e}_1+\sin\theta\mathbf{e}_2)$. Moreover, we may assume that $t>0$ and $\theta\in[0,\pi/2]$ (see, e.g., Figure 5) since the other cases are similar. Our certificate will be the same polynomial as in the previous case of Lemma 4.3. In particular, we will again use the square of the Taylor approximation $S_k^2(c\mathbf{x}_1)$ of the exponential $e^{2c\mathbf{x}_1}$.

We decompose the problem of proving that this polynomial is a certificate for general halfspaces to the problem of showing that it is a certificate for halfspaces with large thresholds simultaneously in both orthogonal directions. In particular, using Equation (11), we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})yp^2(\mathbf{w}^*\cdot\mathbf{x})]\\
&\quad= \|\mathbf{w}\|_2\cos\theta \mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}\left[\left(\mathbf{x}_1-\frac{t}{2\|\mathbf{w}\|_2\cos\theta}\right)\beta(\mathbf{x}_1,\mathbf{x}_2)p^2(\mathbf{x}_1)\text{sign}(\mathbf{x}_1-t^*)\right]\\
&\quad+ \|\mathbf{w}\|_2\sin\theta \mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}\left[\left(\mathbf{x}_2-\frac{t}{2\|\mathbf{w}\|_2\sin\theta}\right)\beta(\mathbf{x}_1,\mathbf{x}_2)p^2(\mathbf{x}_1)\text{sign}(\mathbf{x}_1-t^*)\right]\,.
\end{aligned}\tag{12}$$

To simplify the notation, set $t_1 = t/(2\|\mathbf{w}\|_2\cos\theta)$ and $t_2 = t/(2\|\mathbf{w}\|_2\sin\theta)$ and notice that by the assumptions of Lemma 4.9, it holds that both t_1 and t_2 are larger than $C/2\sqrt{\log(1/(\beta\gamma))}$, see Figure 5. We will show that we can pick $c=\Theta(\log(1/(\beta\gamma))$ and $k=\Theta(c^2)$ so that the polynomial $S_k^2(c\mathbf{x}_1)$ simultaneously satisfies for $i=1,2$, the following inequality

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_i-t_i)\beta(\mathbf{x}_1,\mathbf{x}_2)\text{sign}(\mathbf{x}_1-t^*)p^2(\mathbf{x}_1)] \leq -t_i \text{ poly}(\beta\gamma)\,.$$

Since $t_i\geq 1$, it follows that we can replace the t_i poly$(\beta\gamma)$ above by $(2t_i+1)$poly$(\beta\gamma)$. Having these bounds and using Equation (12), we obtain

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})yp^2(\mathbf{w}^*\cdot\mathbf{x})] &= -(\|\mathbf{w}\|_2(\cos\theta+\sin\theta)+t) \text{ poly}(\gamma\beta)\\
&= -(\|\mathbf{w}\|_2+t) \text{ poly}(\gamma\beta) = -\|\ell(\mathbf{x})\|_2 \text{ poly}(\beta\gamma)\,,
\end{aligned}$$

where for the last inequality we used the fact that $\cos\theta+\sin\theta\geq 1$, for $\theta\in[0,\pi/2]$ and that $\sqrt{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\ell(\mathbf{x})^2]} = \sqrt{\|\mathbf{w}\|_2^2+t^2}$. We bound the first term in claim Claim 4.12 and the second in Claim 4.10.

We first bound the contribution of the direction that is orthogonal to the optimal, i.e., $\mathbf{e}_2$. At a high-level, we have that, since t_2 is a large multiple of $\sqrt{\log(1/(\beta\gamma))}$, "for most" values of $\mathbf{x}_2$, the quantity $\mathbf{x}_2-t_2$ will be negative. Therefore, $\mathbf{x}_2-t_2$ roughly corresponds to the constant guess -1, that we covered in Lemma 4.3.

Claim 4.10. *It holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_2-t_2)\beta(\mathbf{x}_1,\mathbf{x}_2)\text{sign}(\mathbf{x}_1-t^*)p^2(\mathbf{x}_1)] \leq -t_2 \text{ poly}(\beta\gamma)\,.$$

Proof. When $\mathbf{x}_2 - t_2 < 0$, we can treat $\mathbf{x}_2 - t_2$ as a constant negative guess, and use directly the bound obtained in the proof of Lemma 4.3, for $s = -1$, to make the term $\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[(-1)\beta(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \le -C'(\beta\gamma)^{\rho_1}$, for some absolute constants $\rho_1, C' > 0$. In particular, since in this case it holds $\mathbf{x}_2 - t_2 \le 0$, the "worst-case" noise function $\beta(\mathbf{x}_1, \mathbf{x}_2)$ is $\beta(\mathbf{x}_1, \mathbf{x}_2) = \beta$ for all $(\mathbf{x}_1, \mathbf{x}_2)$ such that $\mathbf{x}_1 \ge t^*$ and $\beta(\mathbf{x}_1, \mathbf{x}_2) = 1$ for $\mathbf{x}_1 \le t^*$, see Figure 5b. Notice that this worst case $\beta(\mathbf{x}_1, \mathbf{x}_2)$ does not depend on $\mathbf{x}_2$. Therefore, we obtain

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\mathbb{1}\{\mathbf{x}_2 \le t_2\}(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1, \mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \\
&\quad \le \mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_2 \le t_2\}(\mathbf{x}_2 - t_2)] \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[(-1)\beta(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \\
&\quad \le -\mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_2 \le t_2\}|\mathbf{x}_2 - t_2|] \;\; C'(\beta\gamma)^{\rho_1} \\
&\quad \le -C'/2 \; t_2 \;\; (\beta\gamma)^{\rho_1}, \qquad (13)
\end{aligned}$$

where the last inequality follows by our assumption that $t_2 \ge 1$.

On the other hand, the probability of the region $\{\mathbf{x}_2 \in \mathbb{R} : \mathbf{x}_2 - t_2 \ge 0\}$ is exponentially small which allows us to control the contribution of the region where $\mathbf{x}_2 \ge t_2$. To bound the expectation of $p^4(\mathbf{x}_1)$, we use the following lemma known as Bonami-Beckner inequality or simply Gaussian hypercontractivity.

Lemma 4.11 (Gaussian Hypercontractivity). *Let $p : \mathbb{R} \to \mathbb{R}$ be any polynomial of degree at most ℓ. Then, for every $q \ge 2$, it holds $\mathbf{E}_{x\sim\mathcal{N}}[|p(x)|^q] \le (q-1)^{q\ell/2}\Big(\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)]\Big)^{q/2}$.*

Using Cauchy-Schwarz and the fact that $\beta(\mathbf{x}) \le 1$, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\mathbb{1}\{\mathbf{x}_2 \ge t_2\}(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1, \mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \\
&\quad \le \sqrt{\mathop{\mathbf{Pr}}_{\mathbf{x}_2\sim\mathcal{N}}[\mathbf{x}_2 \ge t_2]} \left(\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[(\mathbf{x}_2 - t_2)^4]\right)^{1/4} \left(\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[p^8(\mathbf{x}_1)]\right)^{1/4}. \qquad (14)
\end{aligned}$$

Using Gaussian hypercontractivity, Lemma 4.11, we obtain that $(\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[p^8(\mathbf{x}_1)])^{1/4} = e^{O(k)}$. Moreover, from standard bounds of Gaussian tails, Fact 3.1, we have that $\mathbf{Pr}_{\mathbf{x}_2\sim\mathcal{N}}[\mathbf{x}_2 \ge t_2] \le e^{-t_2^2/2}$. Finally, we have that $(\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[(\mathbf{x}_2 - t_2)^4])^{1/4} = (3 + 6t_2^2 + t_2^4)^{1/4} \le 2t_2$, for all $t_2 \ge 1$. Combining the above bounds, we obtain that

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\mathbb{1}\{\mathbf{x}_2 \ge t_2\}(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1, \mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \le t_2 e^{-t_2^2/4} e^{O(k)}. \qquad (15)$$

Recall that $k = O(\log(1/(\beta\gamma)))$, therefore there exists an absolute constant $\rho_2 > 0$ such that $e^{O(k)} \le (1/(\beta\gamma))^{\rho_2}$. Therefore, combining eq. (13) and eq. (15), we get that

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1, \mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \le -Ct_2(\beta\gamma)^{\rho_1} + t_2 e^{-t_2^2/4}(\beta\gamma)^{\rho_2}.$$

Hence, because $t_2 \ge C\sqrt{\log(1/(\beta\gamma))}/2$, for C sufficiently large positive absolute constant, we have that $\mathbf{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1, \mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] = -t_2 \; \mathrm{poly}(\beta\gamma)$. ☐

We now bound the contribution of the direction parallel to the optimal vector, i.e., $\mathbf{e}_1$.

Claim 4.12. *It holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_1 - t_1)\beta(\mathbf{x}_1, \mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \le -t_1 \; \mathrm{poly}(\beta\gamma).$$

Proof. Notice that we can marginalize out the direction $\mathbf{e}_2$ in this case. The analysis is similar to that of Lemma 4.3. However, we now have to distinguish three different intervals for $\mathbf{x}_1$ inside which $(\mathbf{x}_1 - t_1)\mathrm{sign}(\mathbf{x}_1 - t^*)$ has the same sign. We will crucially use the fact that $|t^*| = O(\sqrt{\log(1/\gamma)})$, since the Gaussian puts at least γ mass on the positive side of the halfspace, and that the constant C is sufficiently large so that t_1 is at least a large constant multiple of t^*. We have the region $A^+ = A_1^+ \cup A_2^+$, where $A_1^+ = \{\mathbf{x}_1 : \mathbf{x}_1 \le t^*\}$ and $A_2^+ = \{\mathbf{x}_1 : \mathbf{x}_1 \ge t_1\}$. In A^+ it holds $(\mathbf{x}_1 - t_1)\mathrm{sign}(\mathbf{x}_1 - t^*) \ge 0$. We also define the disagreement region $A^- = \{\mathbf{x}_1 : t^* \le \mathbf{x}_1 \le t_1\}$, where $(\mathbf{x}_1 - t_1)\mathrm{sign}(\mathbf{x}_1 - t^*) \le 0$. The worst case noise function puts $\beta(\mathbf{x}_1, \mathbf{x}_2) = \beta$ when $\mathbf{x}_1 \in A^-$ and $\beta(\mathbf{x}_1, \mathbf{x}_2) = 1$ otherwise. Thus, in order to show that p is a certifying polynomial, we have to show that

$$\frac{\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[|\mathbf{x}_1 - t_1| p^2(\mathbf{x}_1) \mathbb{1}\{\mathbf{x}_1 \in A^-\}]}{\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[|\mathbf{x}_1 - t_1| p^2(\mathbf{x}_1) \mathbb{1}\{\mathbf{x}_1 \in A^+\}]} \ge \frac{1}{\beta} . \tag{16}$$

Similarly to the proof of Lemma 4.5 we will first use an exponential function to shift the mean of the Gaussian, i.e., use $e^{2c\mathbf{x}_1}$ instead of $p^2(\mathbf{x}_1)$ in the ratio of Inequality (16). Our goal is to show that there exists some $c \ge t^*$ such that the probability of the region A^- is much larger than the probability of A^+. Since $|t_1|$ is larger than some sufficiently large constant multiple of $\sqrt{\log(1/(\beta\gamma))}$ the interval $[t^*, t_1]$ will contain most of the mass of the shifted normal $\mathcal{N}(c, 1)$ for c some constant multiple of $\sqrt{\log(1/(\beta\gamma))}$, see Figure 2b.

We start by bounding from above the denominator of Inequality (16).

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}_1 \sim \mathcal{N}}[|\mathbf{x}_1 - t_1| e^{2c\mathbf{x}_1} \mathbb{1}\{\mathbf{x}_1 \in A^+\}] &\le \|\mathbf{x}_1 - t_1\|_2 \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}_1 \sim \mathcal{N}}[e^{4c\mathbf{x}_1} \mathbb{1}\{\mathbf{x}_1 \in A^+\}]} \\
&\le 2t_1 e^{4c^2} \left(\sqrt{\mathop{\mathbf{Pr}}_{\mathbf{x}_1 \sim \mathcal{N}}[\mathbf{x}_1 \le t^* - 4c]} + \sqrt{\mathop{\mathbf{Pr}}_{\mathbf{x}_1 \sim \mathcal{N}}[\mathbf{x}_1 \ge t_1 - 4c]} \right) \\
&\le 2t_1 e^{4c^2} (e^{-(t^* - 4c)^2/4} + e^{-(t_1 - 4c)^2)/4}) \\
&\le t_1 e^{4c^2} \beta/16 ,
\end{aligned} \tag{17}$$

where we used that $\|\mathbf{x}_1 - t_1\|_2 = (\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[(\mathbf{x}_1 - t_1)^2]^{1/2} = (1 + t_1^2)^{1/2} \le 2t_1$ for all $t_1 \ge 1$ and, for the last inequality we used the Gaussian tail upper bound, see Fact 3.1, and the fact that $c = \Theta(\sqrt{\log(1/(\beta\gamma))})$, i.e., we get the constant $1/16$ by appropriately choosing some $c = \Theta(\sqrt{\log(1/(\beta\gamma))})$. We next bound the numerator from below. We have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}_1 \sim \mathcal{N}}[|\mathbf{x}_1 - t_1| e^{2c\mathbf{x}_1} \mathbb{1}\{\mathbf{x}_1 \in A^-\}] &\ge \frac{t_1}{2} \mathop{\mathbf{E}}_{\mathbf{x}_1 \sim \mathcal{N}}[e^{2c\mathbf{x}_1} \mathbb{1}\{t^* \le \mathbf{x}_1 \le t_1/2\}] \\
&= \frac{t_1}{2} e^{4c^2} \mathop{\mathbf{Pr}}_{\mathbf{x}_1 \sim \mathcal{N}}[t^* \le \mathbf{x}_1 + 2c \le t_1/2] \ge \frac{t_1 e^{4c^2}}{4} ,
\end{aligned} \tag{18}$$

where for the last inequality we used the fact that $\mathbf{Pr}_{\mathbf{x}_1 \sim \mathcal{N}}[t^* \le \mathbf{x}_1 + 2c \le t_1/2]$ is at least $1/2$ since $t_1/2 - t^*$ is at least a large absolute constant and $c = \Theta(\sqrt{\log(1/(\beta\gamma))})$. In particular, both $c - t^*$ and $t_1/2 - c$ are at least large absolute constants. See Figure 5a. Thus, it holds that

$$\frac{\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[|\mathbf{x}_1 - t_1| \mathbb{1}\{\mathbf{x}_1 \in A^-\} e^{2c\mathbf{x}_1}]}{\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[|\mathbf{x}_1 - t_1| \mathbb{1}\{\mathbf{x}_1 \in A^+\} e^{2c\mathbf{x}_1}]} \ge \frac{4}{\beta} .$$

Similarly to the proof of Lemma 4.5, we now replace the exponential function e^{2cx} by the square of the Taylor expansion of e^{cx}, i.e., $S_k^2(cx)$. It suffices to bound the following ratios

$$\frac{\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[e^{2c\mathbf{x}_1} |\mathbf{x}_1 - t_1| \mathbb{1}\{\mathbf{x}_1 \in A^+\}]}{\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[(S_k(c\mathbf{x}_1))^2 |\mathbf{x}_1 - t_1| \mathbb{1}\{\mathbf{x}_1 \in A^+\}]} \ge \frac{1}{2} \quad \text{and} \quad \frac{\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[(S_k(c\mathbf{x}_1))^2 |\mathbf{x}_1 - t_1| \mathbb{1}\{\mathbf{x}_1 \in A^-\}]}{\mathbf{E}_{\mathbf{x}_1 \sim \mathcal{N}}[e^{2c\mathbf{x}_1} |\mathbf{x}_1 - t_1| \mathbb{1}\{\mathbf{x}_1 \in A^-\}]} \ge \frac{1}{2} . \tag{19}$$

We start by showing the first inequality of Equation (19). It suffices to show that

$$\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[|e^{2c\mathbf{x}_1}-S_k^2(c\mathbf{x}_1)||\mathbf{x}_1-t_1|\mathbb{1}\{\mathbf{x}_1\in A^+\}]\leq \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[e^{2c\mathbf{x}_1}|\mathbf{x}_1-t_1|\mathbb{1}\{\mathbf{x}_1\in A^+\}]\,.$$

Using the L_1 approximation error bound for the Taylor polynomial of Claim 4.7 and Hölder's inequality we can get an L_2 approximation guarantee. Assuming that k is some sufficiently large absolute constant multiple of c^2, i.e., $k=C'c^2$, it holds

$$\begin{aligned}\|e^{2c\mathbf{x}_1}-S_k^2(c\mathbf{x}_1)\|_2 &\leq \|e^{2c\mathbf{x}_1}-S_k^2(c\mathbf{x}_1)\|_1^{1/3}\|e^{2c\mathbf{x}_1}-S_k^2(c\mathbf{x}_1)\|_4^{2/3}\\ &\leq e^{-C'k}e^{O(c^2)}\leq e^{-\Omega(k)}\,,\end{aligned}\tag{20}$$

where the last inequality follows from the fact that $S_k^2(c\mathbf{x}_1)\leq e^{2c|\mathbf{x}_1|}$, $\|e^{2c|\mathbf{x}_1|}\|_4\leq e^{O(c^2)}$ and C' is sufficiently large absolute constant. We can now use Cauchy-Schwarz and the L_2 approximation guarantee of $S_k^2(c\mathbf{x}_1)$ to obtain

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[|e^{2c\mathbf{x}_1}-S_k^2(c\mathbf{x}_1)||\mathbf{x}_1-t_1|\mathbb{1}\{\mathbf{x}_1\in A^+\}]\\ &\quad\leq \|\mathbf{x}_1-t_1\|_2\|e^{2c\mathbf{x}_1}-S_k^2(c\mathbf{x}_1)\|_2\\ &\quad\leq 2t_1e^{-\Omega(k)}\,,\end{aligned}\tag{21}$$

where we used that $\|\mathbf{x}_1-t_1\|_2\leq 2t_1$ for $t_1\geq 1$. We have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[e^{2c\mathbf{x}_1}|\mathbf{x}_1-t_1|\mathbb{1}\{\mathbf{x}_1\in A^+\}]&\geq \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[e^{2c\mathbf{x}_1}|\mathbf{x}_1-t_1|\mathbb{1}\{\mathbf{x}_1\leq t^*\}]\\ &\geq (t_1-t^*)e^{-(2c-t^*)^2}/4\geq t_1e^{-4c^2}/8\,,\quad \triangleright\ t^*\leq c \text{ and } t^*\leq t_1/2\end{aligned}\tag{22}$$

where for the second inequality we used the Gaussian tail lower bounds of Fact 3.1. Combining Inequality (21) and Inequality (22), we obtain that for k larger than some constant multiple of c^2 the first inequality of Equation (19) holds. The proof of the second inequality of Equation (19) is similar. We obtain that Inequality (16) holds for the polynomial $S_k^2(c\mathbf{x}_1)$ for $k=\Theta(1/\log(1/(\beta\gamma)))$ and $c=\Theta(1/\sqrt{\log(1/(\beta\gamma))})$.

Finally, using Fact 4.6 we have that $\|S_k^2(c\mathbf{x}_1)\|_2\leq\|e^{2c|\mathbf{x}_1|}\|_2=e^{O(c^2)}$. Using the estimate of Equation (18) and the L_2 approximation guarantee of Equation (20) we obtain that

$$\frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[S_k^2(c\mathbf{x}_1)|\mathbf{x}_1-t_1|\mathbb{1}\{\mathbf{x}_1\in A^-\}]}{\|S_k^2(c\mathbf{x}_1)\|_2}=\frac{e^{-\Omega(c^2)}-e^{-\Omega(k)}}{e^{O(c^2)}}=e^{-O(c^2)}=\mathrm{poly}(\gamma\beta)\,,$$

where we used the fact that k is a sufficiently large constant multiple of c^2 and $c^2=\Theta(\log(1/(\beta\gamma)))$. We conclude that

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}\left[(\mathbf{x}_1-t_1)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1-t^*)\frac{S_k^2(c\mathbf{x}_1)}{\|S_k^2(c\mathbf{x}_1)\|_2}\right]\\ &\quad\leq \beta\frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[S_k^2(c\mathbf{x}_1)|\mathbf{x}_1-t_1|\mathbb{1}\{\mathbf{x}_1\in A^-\}]}{\|S_k^2(c\mathbf{x}_1)\|_2}-\frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[S_k^2(c\mathbf{x}_1)|\mathbf{x}_1-t_1|\mathbb{1}\{\mathbf{x}_1\in A^+\}]}{\|S_k^2(c\mathbf{x}_1)\|_2}\\ &\quad\leq -t_1\mathrm{poly}(\beta\gamma)\,.\end{aligned}$$

The proof of the upper bound on the coefficients of the polynomial $S_k^2(c\mathbf{x}_1)/\|S_k^2(c\mathbf{x}_1)\|_2$ is similar to that of the constant hypothesis case, see Lemma 4.3.

□

This completes the proof of Lemma 4.9. □

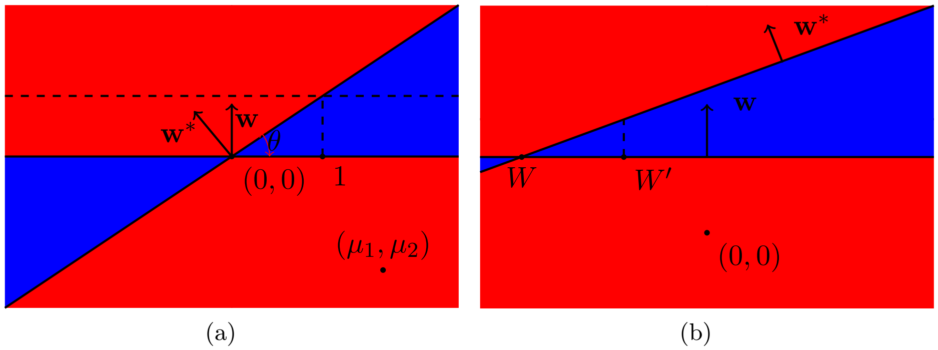


Figure 6: **(a)** Our certificate when the angle θ between the two halfspaces is large. Notice that we have changed coordinates so that the point $(0,0)$ is their crossing point. The mean of the Gaussian is now moved to the point $(\mu_1,\mu_2) = (-t/\tan(\theta) + t^*/\sin\theta, -t)$. Notice that since in this case we have assumed that $|t| \le C\sqrt{\log(1/(\beta\gamma))}$, the $\mathbf{e}_2$-coordinate of the mean of the Gaussian cannot be very far from the origin. In this case, we overestimate the agreement region between the two halfspaces (red area) considering to be $\{\mathbf{x}_1 \le 1\}$. In order to put more weight to the disagreement area, we again use the polynomial shift (see Lemma 4.5) in the direction $\mathbf{e}_1 = (\mathbf{w}^*)^{\perp_{\mathbf{w}}}$. Observe that this differs from the cases of Subsection 4.1.1 and Subsection 4.1.2, where the polynomial shift was along the direction of $\mathbf{w}^*$.
(b) The case where the angle θ between the two halfspaces is small, $\theta = O(\epsilon^2\gamma\beta)$. In that case, the crossing point W of the two halfspaces is very far from the origin, $|W| \ge \Omega(1/(\epsilon\gamma\beta))$. In this case, simply taking a band around $\ell(\mathbf{x})$ works as a certificate.

4.1.3 Certificate Against "Small Threshold" Halfspaces

Lemma 4.13 (Certificate against "Small Threshold" Hypotheses)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the η-Massart noise condition with respect to some at most $(1-\gamma)$-biased optimal halfspace. Define the linear function $\ell(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} - t)$ and assume that $t/\|\mathbf{w}\|_2 \le C\sqrt{\log(1/(\beta\gamma))}$ for some absolute constant $C > 0$. Moreover, assume that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \ge \mathrm{OPT} + \epsilon$. Then, there exists polynomial $q(\mathbf{x}) = \sum_{|\alpha|\le k} c_\alpha \mathbf{x}^\alpha$ of degree $\Theta(\log(\frac{1}{\beta}))$, norm $\|q(\mathbf{x})\|_2 = 1$, and sum of (absolute) coefficients $\sum_{|\alpha|\le k} |c_\alpha| \le d^{O(k)}$, and $r_1, r_2 \in \mathbb{R}$ with $|r_1 - r_2| \ge \mathrm{poly}(\epsilon\beta\gamma)$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x}) y \mathbb{1}\{r_1 \le \ell(\mathbf{x}) \le r_2\} q^2(\mathbf{x})] \le -\epsilon^4 \mathrm{poly}(\beta\gamma)\ \|\ell(\mathbf{x})\|_2 \,.$$

Proof. Denote $\ell^*(\mathbf{x}) = \mathbf{w}^*\cdot\mathbf{x} - t^*$ the optimal halfspace and denote $\ell(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} - t$ for some vectors $\mathbf{w}\in\mathbb{R}^d$ and some threshold $t \in \mathbb{R}$. Moreover, denote $\theta = \theta(\mathbf{w}^*, \mathbf{w})$ the angle between $\mathbf{w}^*$ and $\mathbf{w}$. We first observe that since $\mathrm{OPT} = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell^*(\mathbf{x}) \neq y]$ it holds that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] - \mathrm{OPT} = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq \mathrm{sign}(\ell^*(\mathbf{x}))\}\beta(\mathbf{x})] \,.$$

Thus, the disagreement probability between $\ell(\mathbf{x})$ and $\ell(\mathbf{x}^*)$ $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq \mathrm{sign}(\ell^*(\mathbf{x}))\}] \ge \epsilon$ from our assumption that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \ge \mathrm{OPT} + \epsilon$. From Fact 3.2 we know that

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_x}[\mathrm{sign}(\ell(\mathbf{x})) \neq \mathrm{sign}(\ell^*(\mathbf{x}))] \le O(\theta) + O(|t^* - t|) \,.$$

Thus, we either have that the angle of the two halfspaces is large $\theta = \Omega(\epsilon)$ or the difference of their thresholds $|t^* - t| = \Omega(\epsilon)$.

Since we can always write

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}&[(\mathbf{w}\cdot\mathbf{x}-t)y\ \mathbb{1}\{r_1\leq\ell(\mathbf{x})\leq r_2\}q^2(\mathbf{x})]\\&=\|\mathbf{w}\|_2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\cdot\mathbf{x}-\frac{t}{\|\mathbf{w}\|_2}\right)y\ \mathbb{1}\{r_1\leq\ell(\mathbf{x})\leq r_2\}q^2(\mathbf{x})\right],\end{aligned}$$

we will assume for simplicity that $\|\mathbf{w}\|_2=1$ and $|t|\leq C\sqrt{\log(1/(\beta\gamma))}$. We are going to construct a polynomial $q(\mathbf{x})$ that only depends on the subspace V spanned by $\mathbf{w},\mathbf{w}^*$. Therefore, as in the case of Lemma 4.3, we can project $\mathbf{x}$ to the subspace V spanned by $\mathbf{w},\mathbf{w}^*$ and preserve the η-Massart noise assumption, see Claim 4.4. Let $\theta=\theta(\mathbf{w},\mathbf{w}^*)$ be the angle between $\mathbf{w},\mathbf{w}^*$. Without loss of generality, we may assume that $\mathbf{w}=\mathbf{e}_2$ and $\mathbf{w}^*=-\sin\theta\mathbf{e}_1+\cos\theta\mathbf{e}_2$, see Figure 6a. Therefore, in this case we have that $\ell(\mathbf{x})=\mathbf{x}_2-t$ and $\ell^*(\mathbf{x})=-\sin\theta\mathbf{e}_1+\cos\theta\mathbf{e}_2-t^*$.

We first provide a certificate that works when the angle θ is non-trivial, that is when $\theta\geq C'\epsilon^2\gamma\beta$, for some small enough constant C' chosen appropriately. We change the coordinates so that the crossing point of the two linear functions ℓ,ℓ^* is the origin $(0,0)$. This will move the mean of the Gaussian to the point $(\mu_1,\mu_2)=(-t/\tan\theta+t^*/\sin\theta,-t)$, see Figure 6a.

Assume first that $C\epsilon^2\gamma\beta=\theta\leq\pi/2$ and $\mu_1\geq 0$, we show later that the other cases are symmetric. Set $r_1=0$, $r_2=\tan(\theta)=\Theta(\theta)$, and let $q(\mathbf{x})$ depend only on $\mathbf{x}_1$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}&[\ell(\mathbf{x})y\ \mathbb{1}\{r_1\leq\ell(\mathbf{x})\leq r_2\}q^2(\mathbf{x})]\\&=\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}(\mu_1,\mu_2)}[\mathbf{x}_2\mathrm{sign}(-\sin\theta\mathbf{x}_1+\cos\theta\mathbf{x}_2)\ \beta(\mathbf{x}_1,\mathbf{x}_2)\mathbb{1}\{0\leq\mathbf{x}_2\leq r_2\}q^2(\mathbf{x}_1)]\\&\leq\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}(\mu_1,\mu_2)}[\mathbf{x}_2\mathrm{sign}(1-\mathbf{x}_1)\ \beta(\mathbf{x}_1,\mathbf{x}_2)\ \mathbb{1}\{0\leq\mathbf{x}_2\leq r_2\}q^2(\mathbf{x}_1)]\,,\end{aligned}\tag{23}$$

where for the inequality we used the fact that $0\leq\mathbf{x}_2\leq r_2=\tan(\theta)$, and therefore it holds that if $-\sin\theta\mathbf{x}_1+\cos\theta\mathbf{x}_2\geq 0$ then it holds that $1-\mathbf{x}_1\geq 0$, for all $(\mathbf{x}_1,\mathbf{x}_2)$ such that $\mathbf{x}_2\in[0,r_2]$. Observe now that in order to maximize the quantity of Equation (23) the "worst-case" noise function $\beta(\mathbf{x}_1,\mathbf{x}_2)$ is equal to β for all points where the integral is negative, and $+1$ when the integral is positive, that is $\beta(\mathbf{x}_1,\mathbf{x}_2)=\beta\mathbb{1}\{\mathbf{x}_1\geq 1\}+\mathbb{1}\{\mathbf{x}_1<1\}$. Since this "worst-case" noise $\beta(\mathbf{x}_1,\mathbf{x}_2)$ is independent of $\mathbf{x}_2$, it follows that we can decompose the expectation of Equation (23), i.e.,

$$\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}(\mu_1)}[q^2(\mathbf{x}_1)\mathrm{sign}(1-\mathbf{x}_1)\beta(\mathbf{x}_1)]\mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}(\mu_2)}[\mathbf{x}_2\mathbb{1}\{0\leq\mathbf{x}_2\leq r_2\}]\,.$$

Since $|\mu_2|=|t|\leq C\sqrt{\log(1/(\beta\gamma))}$ from standards bounds of the Gaussian tail probability (Fact 3.1), we obtain

$$\mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}(\mu_2)}[\mathbf{x}_2\mathbb{1}\{0\leq\mathbf{x}_2\leq r_2\}]\geq\theta^2\mathrm{poly}(\beta\gamma)\,.$$

It remains to bound the term $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}(\mu_1)}[q^2(\mathbf{x}_1)\mathrm{sign}(1-\mathbf{x}_1)\beta(\mathbf{x}_1)]$. We observe that the "worst-case" value of $\mu_1\geq 0$ in order to maximize this expectation is $\mu_1=0$. Now, we can use the same argument as in the proof of Lemma 4.3; notice that in this case the threshold is 1 instead of $\sqrt{\log(1/\gamma)}$ and therefore, by picking $k=\Theta(\log(1/(\beta)))$ we have that there exists polynomial q such that $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}(\mu_1)}[q^2(\mathbf{x}_1)\mathrm{sign}(1-\mathbf{x}_1)\beta(\mathbf{x}_1)]\leq-\mathrm{poly}(\beta\gamma)$ and thus, we obtain the bound:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y\ \mathbb{1}\{r_1\leq\mathbf{w}\cdot\mathbf{x}\leq r_2\}q^2(\mathbf{x})]\leq-\theta^2\mathrm{poly}(\beta\gamma)\|\ell(\mathbf{x})\|_2\,.$$

In the case where $\theta\in[0,\pi/2]$ and $\mu_1<0$, we may pick $r_1=-\tan(\theta)$ and $r_2=0$ resulting in a completely symmetric case to the previous one. Finally, the case where $\theta\in[\pi/2,\pi]$ is easier than the previous two cases since the disagreement region between the two halfspaces is now a superset of the corresponding region in the previous cases.

We now handle the case where the angle between the two halfspaces is small, i.e., $\theta \le C'\epsilon^2\beta\gamma$. We know that the disagreement between two halfspaces is upper bounded by their angle, i.e., $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})] \le O(\theta) + O(|t-t^*|)$. Since $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})] \ge \epsilon$, we obtain that the two thresholds t, t^* cannot be very close, i.e., $|t-t^*| = \Omega(\epsilon)$. As in the previous case, we may assume that $\mathbf{w} = \mathbf{e}_2$ and $\mathbf{w}^* = -\sin\theta\mathbf{e}_1 + \cos\theta\mathbf{e}_2$. Recall that the intersection point of the two halfspaces has coordinates $(t/\tan\theta - t^*/\sin\theta, t)$. This means that when $\theta = O(\epsilon^2\gamma\beta)$ the intersection point of the two halfspaces is very far from the origin: its first coordinate is $|t/\tan\theta - t^*/\sin\theta| = \Omega(1/(\epsilon\gamma\beta))$, since by the triangle inequality, it follows

$$\begin{aligned}|t\cos\theta - t^*| &\ge |t-t^*|\cos\theta - |t^*||\cos\theta - 1| \\ &\ge |t-t^*|(1-\epsilon^4\gamma^2\beta^2) - O(\sqrt{\log(1/(\beta\gamma))})(\epsilon^4\gamma^2\beta^2 - 2\epsilon^8\gamma^4\beta^4) = \Omega(\epsilon)\,,\end{aligned}$$

where we used the inequality $1-x^2 \le \cos x \le 1-x^2/2+x^4$, for all $x \in [0,\pi/2]$ and the fact that $\epsilon \le 1/2$. Combining, the above with $\sin(\theta) = O(\epsilon^2\beta\gamma)$, we get that $|t/\tan\theta - t^*/\sin\theta| = \Omega(1/(\epsilon\gamma^2\beta^2))$. In this case, we do not require a polynomial for the certificate, therefore the certificate is going to be simply a band, see Figure 6b. Let W be the $\mathbf{e}_1$-coordinate of the crossing point of the optimal halfspace with $\ell(\mathbf{x})$, see Figure 6b. We have that $|W| = \Omega(1/(\epsilon\beta\gamma))$. Without loss of generality, we may assume that $W < 0$ along with the band $\epsilon^2/2 \le \ell(\mathbf{x}) \le \epsilon^2$ since the proof of the other case is similar: we just need to consider the band $-\epsilon^2 \le \ell(\mathbf{x}) \le -\epsilon^2/2$ instead, see Figure 6b. Moreover, denote W' the $\mathbf{e}_1$-coordinate of the point of $\ell^*(x) = \ell(\mathbf{x}) + \epsilon^2$, see Figure 6b. It holds $W' = W + \epsilon^2/\tan\theta$, and note that $|W'| = \Omega(1/(\epsilon\beta\gamma))$ since $|W + \epsilon^2/\tan\theta| \ge (|t\cos\theta - t^*| - |\epsilon^2|)/\sin\theta \ge \Omega(1/(\epsilon\beta\gamma))$. We have

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathrm{sign}(-\mathbf{x}_1\sin\theta + \mathbf{x}_2\cos\theta - t^*)\mathbb{1}\{\epsilon^2/2 \le \ell(\mathbf{x}) \le \epsilon^2\}\beta(\mathbf{x}_1,\mathbf{x}_2)] \\ &\qquad \le \mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathrm{sign}(-\mathbf{x}_1 + W + \epsilon^2/\tan\theta)\mathbb{1}\{\epsilon^2/2 \le \ell(\mathbf{x}) \le \epsilon^2\}\beta(\mathbf{x}_1,\mathbf{x}_2)]\,, \end{aligned} \tag{24}$$

where we overestimated the contribution of the agreement area (red region in Figure 6b) by the region $\mathbf{x}_1 \le W'$. Since W' is still very far from the origin the contribution of the region $\mathbf{x}_1 \le W'$ is going to be very small. To bound the quantity of Equation (24), we first bound from above the region where $\mathrm{sign}(-\mathbf{x}_1 + W')$ is positive:

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathbb{1}\{\epsilon^2/2 \le \ell(\mathbf{x}) \le \epsilon^2\}\beta(\mathbf{x})\mathbb{1}\{\mathbf{x}_1 \le W'\}] \le \mathop{\mathbf{Pr}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbf{x}_1 \le W'] \le e^{-1/(\epsilon\beta\gamma)^2}\,, \tag{25}$$

where we used the fact that $|W'| = \Omega(1/(\epsilon\gamma\beta))$. Next, we bound from below the region where $\mathrm{sign}(-\mathbf{x}_1 + W')$ is negative:

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathbb{1}\{\epsilon^2/2 \le \ell(\mathbf{x}) \le \epsilon^2\}\beta(\mathbf{x})\mathbb{1}\{\mathbf{x}_1 \ge W'\}] \\ &\qquad \ge \beta \mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}}[\ell(\mathbf{x}_2)\mathbb{1}\{\epsilon^2/2 \le \ell(\mathbf{x}_2) \le \epsilon^2\}] \mathop{\mathbf{Pr}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbf{x}_1 \ge W'] \\ &\qquad \ge \frac{\epsilon^2}{4}\beta \mathop{\mathbf{Pr}}_{\mathbf{x}_2\sim\mathcal{N}}[\epsilon^2/2 \le \ell(\mathbf{x}_2) \le \epsilon^2] \ge \epsilon^4\beta\mathrm{poly}(\beta\gamma)\,, \end{aligned} \tag{26}$$

where we used that $\mathbf{Pr}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbf{x}_1 \ge W'] \ge 1/2$ and that $\beta(\mathbf{x}) \ge \beta$. Using Equation (25) and Equation (26) to Equation (24), we get

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathrm{sign}(-\mathbf{x}_1\sin\theta + \mathbf{x}_2\cos\theta - t^*)\mathbb{1}\{\epsilon^2/2 \le \ell(\mathbf{x}) \le \epsilon^2\}\beta(\mathbf{x}_1,\mathbf{x}_2)] \le -\epsilon^4\mathrm{poly}(\beta\gamma)\,.$$

We combine the above cases to obtain the claimed bound. □

4.2 Efficiently Computing the Certificate via SDP

In this section, we show that we can efficiently compute our polynomial certificate given labeled examples from the target distribution. The following is the main proposition of this subsection, where we bound the number of samples and the runtime needed to compute the certificate given samples from the distribution $\mathcal{D}$. The proof is similar to that given in [DKTZ20b]; we adapt it to work for our certifying function and for general (as opposed to homogeneous) halfspaces.

Proposition 4.14 (Certificate Oracle). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the η-Massart noise condition with respect to some (at least) $(1-\gamma)$-biased optimal halfspace $f(\mathbf{x})$. Let $\ell(\mathbf{x})$ be any linear function such that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})] \geq \epsilon$, for $\epsilon \in (0,1)$. There exists an algorithm that draws $N = d^{O(\log(1/(\beta\gamma)))}\log(1/\delta)/\epsilon^2$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N,d)$, and with probability $1-\delta$ returns a positive function $T(\mathbf{x})$ with $\|T(\mathbf{x})\|_4 \leq 1$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x})\ell(\mathbf{x})y] \leq -\epsilon^4 d^{-O(\log(1/(\beta\gamma)))}\|\ell(\mathbf{x})\|_2 \,.$$

Proof. From Proposition 4.2, we know that we are looking for a certificate function $T(\mathbf{x})$ of the form $T(\mathbf{x}) = \mathbb{1}_B(\mathbf{x})q^2(\mathbf{x})$, where B is a band with respect $\ell(\mathbf{x})$, i.e., $B = \{r_1 \leq \ell(\mathbf{x}) \leq r_2\}$ for some $r_1, r_2 \in \mathbb{R}\cup\{\infty\}$ and $q(\mathbf{x})$ is a $k = \Theta(\log(1/\gamma\beta))$ degree polynomial. We illustrate how we can formulate the search of such function as an SDP. For the rest of this section, let $\mathbb{1}_B(\mathbf{x})$ be the indicator function of the region $B = \{\mathbf{x} : r_1 \leq \ell(\mathbf{x}) \leq r_2\}$, for some appropriate choices $r_1, r_2 \in \mathbb{R}\cup\{\infty\}$ and $\lambda = \epsilon^4 d^{-O(k)}\|\ell(\mathbf{x})\|_2$. Denote by $\mathbf{m}(\mathbf{x})$ the vector containing all monomials up to degree k, such that $\mathbf{m}_S(\mathbf{x}) := \mathbf{x}^S$, indexed by the multi-index S satisfying $|S| \leq k$. Recall that if $S = (s_1, \ldots, s_d)$, then $\mathbf{x}^S = \prod_{i=1}^d \mathbf{x}_i^{s_i}$. The dimension of $\mathbf{m}(\mathbf{x}) \in \mathbb{R}^m$ is $m = \binom{d+k}{k}$. Let $\mathbf{M} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbf{m}(\mathbf{x})\mathbf{m}(\mathbf{x})^\top \mathbb{1}_B(\mathbf{x})\ell(\mathbf{x})y\right]$, for any real matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$, we define the following function

$$\mathcal{L}_\ell(\mathbf{A}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbf{m}(\mathbf{x})^\top \mathbf{A}\ \mathbf{m}(\mathbf{x})\mathbb{1}_B(\mathbf{x})\ell(\mathbf{x})y\right] = \mathrm{tr}\left(\mathbf{AM}\right) \,. \tag{27}$$

Notice that $\mathcal{L}_\ell$ is linear in its variable $\mathbf{A}$. From Proposition 4.2, we know that if $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})] \geq \epsilon$, then there exists a (normalized) polynomial $q(\mathbf{x}) = \mathbf{m}(\mathbf{x})\cdot\mathbf{a}$, with $\|\mathbf{a}\|_1 = 1$ such that $\mathbf{E}_{(\mathbf{x},y)}[\ell(\mathbf{x})y\mathbb{1}_B(\mathbf{x})q(\mathbf{x})] \leq -\lambda$. Therefore, for $\mathbf{A} = \mathbf{a}\mathbf{a}^\top$, we have $q^2(\mathbf{x}) = \mathbf{m}(\mathbf{x})^\top\mathbf{A}\ \mathbf{m}(\mathbf{x})$ and hence, $\mathcal{L}_\ell(\mathbf{A}) \leq -\lambda$. It follows that there exists a positive semi-definite rank-1 matrix $\mathbf{A}$ such that $\mathcal{L}_\ell(\mathbf{A}) \leq -\lambda$. Moreover, $\|\mathbf{A}\|_F^2 = \left\|\mathbf{a}\mathbf{a}^T\right\|_F = \|\mathbf{a}\|_2^4 \leq 1$. Recall that, $\mathcal{S}^m$ is the set of (symmetric) positive semi-definite matrices of m-dimension. We formulate the following semi-definite program

$$\begin{aligned} \text{Find} \quad & \mathbf{A} \in \mathcal{S}^m \\ \text{s. t.} \quad & \mathrm{tr}(\mathbf{AM}) \leq -\lambda \\ & \|\mathbf{A}\|_F^2 \leq 1 \end{aligned} \tag{28}$$

Moreover, from Proposition 4.2, the SDP (28) is feasible if $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})] \geq \epsilon$. We define $\widetilde{\mathbf{M}} = \frac{1}{N}\sum_{i=1}^N \mathbf{m}(\mathbf{x}^{(i)})\mathbf{m}(\mathbf{x}^{(i)})^\top\mathbb{1}_B(\mathbf{x}^{(i)})y^{(i)}\ell(\mathbf{x}^{(i)})$, the empirical estimate of $\mathbf{M}$ using N samples from $\mathcal{D}$. Using the following fact, we bound the sample size required so that $\widetilde{\mathbf{M}}$ is sufficiently close to $\mathbf{M}$ which is similar to Lemma 3.8 of [DKTZ20b]. (See Appendix F for the proof).

Fact 4.15 (Estimation of $\mathbf{M}$). *Let $\Omega = \{\mathbf{A} \in \mathcal{S}^m : \|\mathbf{A}\|_F \leq 1\}$ and $\epsilon, \delta \in (0,1)$. Let $\ell(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + t$ with $|\ell(\mathbf{x})|^2 \leq C$ and $\widetilde{\mathbf{M}} = \frac{1}{N}\sum_{i=1}^N \mathbf{m}(\mathbf{x}^{(i)})\mathbf{m}(\mathbf{x}^{(i)})^\top\mathbb{1}_B(\mathbf{x}^{(i)})y^{(i)}\ell(\mathbf{x}^{(i)})$. There exists an algorithm* *that draws* $N=\frac{d^{O(\log\frac{1}{\gamma\beta})}}{C\epsilon^2}\log(1/\delta)$ *samples from* $\mathcal{D}$*, runs in* $\mathrm{poly}(N,d)$ *time and with probability at least* $1-\delta$ *outputs a matrix* $\widetilde{\mathbf{M}}$ *such that*

$$\mathbf{Pr}\left[\sup_{\mathbf{A}\in\Omega}\left|\mathrm{tr}(\mathbf{A}\widetilde{\mathbf{M}})-\mathrm{tr}(\mathbf{A}\mathbf{M})\right|\geq\epsilon\right]\leq 1-\delta\,.$$

Using Fact 4.15, we replace the matrix $\mathbf{M}$ in Equation (27) with the estimate $\widetilde{\mathbf{M}}$ and define the following "empirical" SDP

$$\begin{aligned}\text{Find}\quad & \mathbf{A}\in\mathcal{S}^m\\ \text{such that}\quad & \mathrm{tr}(\mathbf{A}\widetilde{\mathbf{M}})\leq-\lambda/2\\ & \|\mathbf{A}\|_F^2\leq 1\end{aligned}\tag{29}$$

From Fact 4.15, we obtain that with N samples we can get a matrix $\widetilde{\mathbf{M}}$ such that $|\mathrm{tr}(\mathbf{A}\widetilde{\mathbf{M}}-\mathrm{tr}(\mathbf{A}\mathbf{M})|\leq-\lambda/2$ with probability at least $1-\delta'$. From Proposition 4.2, we know that with the given bound for k and $\|\mathbf{A}\|_F$, there exists $\mathbf{A}^*$ such that

$$\mathrm{tr}(\mathbf{A}^*\mathbf{M})\leq-\lambda.$$

Therefore, the SDP (28) is feasible. Moreover, from Fact 4.15 we get that

$$\mathrm{tr}(\mathbf{A}^*\widetilde{\mathbf{M}})\leq-\lambda/2\,.$$

Thus, the SDP (29) is feasible. Since the dimension of the matrix $\mathbf{A}$ is smaller than the number of samples, we have that the runtime of the SDP is polynomial in the number of samples. Solving the SDP with tolerance $\lambda/4$, we obtain an almost feasible $\widetilde{\mathbf{A}}$, in the sense that $\mathrm{tr}(\widetilde{\mathbf{A}}\widetilde{\mathbf{M}})\leq-\lambda/4$. Using again the guarantee of Fact 4.15, we get that solving the SDP (29), we obtain a positive-semi definite matrix $\widetilde{\mathbf{A}}$ such that $\mathrm{tr}(\widetilde{\mathbf{A}}\mathbf{M})\leq-\lambda/4$. Moreover, we have that for the matrix $\mathbf{A}$ returned by our SDP it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{m}(\mathbf{x})^T\mathbf{A}\mathbf{m}(\mathbf{x}))^4]\leq\|\mathbf{A}\|_F^4\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{m}(\mathbf{x})^T\mathbf{m}(\mathbf{x}))^4]=d^{\log(1/(\beta\gamma))}\tag{30}$$

To complete the proof, we need to show how to guess the band B. For some large enough constant $C>0$, let $\mathcal{T}=\{\pm\lambda^2,\pm2\lambda^2,\ldots,C\sqrt{\log(1/\lambda)}\}$. Assume that for some $B=\{\mathbf{x}:r_1\leq\ell(\mathbf{x})\leq r_2\}$, with $r_1,r_2\in\mathbb{R}$ and some polynomial $q(\mathbf{x})$ with $\|q^2(\mathbf{x})\|_2\leq 1$, such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}_B(\mathbf{x})q^2(\mathbf{x})\ell(\mathbf{x})y\right]\leq-\lambda/4$. Then, there exists $\tilde{r}_1,\tilde{r}_2\in\mathcal{T}$, with $|r_1-\tilde{r}_1|\leq\lambda^2$ and $|r_2-\tilde{r}_2|\leq\lambda^2$, such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}\{\tilde{r}_1\leq\ell(\mathbf{x})\leq\tilde{r}_2\}q^2(\mathbf{x})\ell(\mathbf{x})y\right]\leq\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}_B(\mathbf{x})q^2(\mathbf{x})\ell(\mathbf{x})y\right]+2\lambda^2\leq-\lambda/4\,,$$

where we used Cauchy–Schwarz inequality and the Gaussian concentration. Thus, by setting $\delta'=\Theta(\delta\lambda^4/\log(1/\lambda))$ solving the SDP (29) for all the different choices of $r_1,r_2\in\mathcal{T}$, with $r_1\leq r_2$, we guarantee that the algorithm will return a polynomial $q(\mathbf{x})$ and some thresholds $\tilde{r}_1,\tilde{r}_2\in\mathcal{T}$, such that $T(\mathbf{x})=\mathbb{1}\{\tilde{r}_1\leq\ell(\mathbf{x})\leq\tilde{r}_2\}q^2(\mathbf{x})$ is a certifying function, i.e., it holds

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}\{\tilde{r}_1\leq\ell(\mathbf{x})\leq\tilde{r}_2\}q^2(\mathbf{x})\ell(\mathbf{x})y\right]\leq-\lambda/4\,,$$

with probability $1-\delta$. Finally, using Inequality (30) we obtain that the L_4 norm of $T(\mathbf{x})$ is bounded from above as $\|T(\mathbf{x})\|_4\leq d^{O(\log(1/(\beta\gamma)))}$. Thus, the function $T(\mathbf{x})/\|T(\mathbf{x})\|_4$ is an $\epsilon^4 d^{-O(\log(1/(\beta\gamma)))}$-certificate.

□

4.3 Learning a Near-Optimal Halfspace via Online Convex Optimization

In this subsection, we present a black-box approach to learn halfspaces with η-Massart Noise given a ρ-certifying oracle. A similar reduction for homogeneous halfspaces based on online-convex optimization was given in [DKK+20]. Here we adapt it so that it handles non-homogeneous halfspaces. Another difference is that in [DKK+20] the certificate function was bounded in the L_∞ sense. Here we have certificates bounded in the L_4 norm. The arguments however are similar, and we provide them here for completeness. Formally, we prove:

Proposition 4.16. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the η-Massart noise condition with respect to some halfspace. Fix $\epsilon, \delta \in (0,1)$. Given a ρ-certificate oracle $\mathcal{O}$ that returns certifying functions with bounded ℓ_4-norm. There exists an algorithm that makes $T = O(\frac{d\log(1/\epsilon)}{\rho^2\epsilon^4})$ calls to $\mathcal{O}$, draws $N = \tilde{O}(\frac{d}{\rho^2\epsilon^4}\log(1/\delta))$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(d, N, T)$ time and computes a hypothesis h, such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, with probability $1-\delta$.*

We will require the following standard regret bound from online convex optimization.

Lemma 4.17 (see, e.g., Theorem 3.1 of [Haz16])**.** *Let $\mathcal{V} \subseteq \mathbb{R}^n$ be a non-empty closed convex set with diameter K. Let $r_1, \ldots, r_T$ be a sequence of T convex functions $r_i : \mathcal{V} \mapsto \mathbb{R}$ differentiable in open sets containing $\mathcal{V}$, and let $G = \max_{i\in[T]} \|\nabla_{\mathbf{w}} r_i\|_2$. Pick any $\mathbf{w}^{(1)} \in \mathcal{V}$ and set $\eta_i = \frac{K}{G\sqrt{t}}$ for $i \in [T]$. Then, for all $\mathbf{u} \in \mathcal{V}$, we have that $\sum_{i=1}^{T}(r_i(\mathbf{w}^{(t)}) - r_i(\mathbf{u})) \leq \frac{3}{2}GK\sqrt{T}$.*

We show below that the optimal vector $\mathbf{w}^*$ and threshold t^* and our current candidate vector $\mathbf{w}^{(i)}$ and threshold $t^{(i)}$ have a separation in the value of r_i.

Lemma 4.18 (Error of r_i)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the η-Massart noise condition with respect to the optimal halfspace $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t^*)$. Let $\mathbf{w}^{(i)}$ with $\|\mathbf{w}^{(i)}\| \leq 1$ and $t^{(i)} \in \mathbb{R}$. Fix $\rho \in (0,1)$ and let $r_i(\mathbf{w}, t) = -\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(T^{(i)}(\mathbf{x}) + \rho)y(\mathbf{x}, 1)] \cdot (\mathbf{w}, t)$, where $T^{(i)}(\mathbf{x})$ is a non-negative function returned from a (2ρ)-certificate oracle, we have that*

$$r_i(\mathbf{w}^{(i)}, t^{(i)}) - r_i(\mathbf{w}^*, t^*) \geq \rho\epsilon^2/2 \,.$$

Proof. Let $\ell^{(i)}(\mathbf{x}) = \mathbf{w}^{(i)} \cdot \mathbf{x} + t^{(i)}$. Using the fact that $T^{(i)}(\mathbf{x}) \geq 0$, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T^{(i)}(\mathbf{x})(\mathbf{w}^* \cdot \mathbf{x} + t^*)y] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[T^{(i)}(\mathbf{x})|\mathbf{w}^* \cdot \mathbf{x} + t^*|\beta(\mathbf{x})] > 0 \,.$$

Therefore, we have that for every $i \in [T]$, it holds $r_i(\mathbf{w}^*, t^\star) \leq -\rho\,\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^* \cdot \mathbf{x} + t^*|\beta(\mathbf{x})]$. Let $I = \{\mathbf{x} : \mathbf{w}^* \cdot \mathbf{x} \in (-t^* - \epsilon/2, -t^* + \epsilon/2)\}$ and note that it should hold that $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y] = \mathrm{OPT} \leq 1/2 - \epsilon$, otherwise the (2ρ)-certifying oracle would not be able to return a function, and therefore $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})] \geq 2\epsilon$. We have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^* \cdot \mathbf{x} + t^*|\beta(\mathbf{x})] &\geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^* \cdot \mathbf{x} + t^*|\{\mathbf{x} \notin I\}\beta(\mathbf{x})] \geq \frac{\epsilon}{2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{\mathbf{x} \notin I\}\beta(\mathbf{x})] \\ &= \frac{\epsilon}{2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})] - \frac{\epsilon}{2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{\mathbf{x} \in I\}\beta(\mathbf{x})] \geq \epsilon^2/2 \,.\end{aligned}$$

where in the last inequality we used that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})] \geq 2\epsilon$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{\mathbf{x} \in I\}] \leq \epsilon$ from Fact 3.1.

It remains to bound from below $r_i(\mathbf{w}^{(i)}, t^{(i)})$. Using the fact that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T^{(i)}(\mathbf{x})(\mathbf{w}^{(i)} \cdot \mathbf{x} + t^{(i)})y] \leq -2\rho\|\ell^{(i)}(\mathbf{x})\|_2$, we have

$$\begin{aligned}r_i(\mathbf{w}^{(i)}, t^{(i)}) &= -\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(T^{(i)}(\mathbf{x}) + \rho)(\mathbf{w}^{(i)} \cdot \mathbf{x} + t^{(i)})y\right] \\ &\geq 2\rho\|\ell^{(i)}(\mathbf{x})\|_2 - \rho\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}^{(i)} \cdot \mathbf{x} + t^{(i)})y\right] \geq \rho\|\ell^{(i)}(\mathbf{x})\|_2 \geq 0 \,,\end{aligned}$$

Input:

1. $\epsilon, \delta > 0$.
2. A distribution $\mathcal{D}$ that satisfies the η-Massart Noise condition.
3. Access to a ρ-certificate oracle $\mathcal{O}$

Output: A vector $\mathbf{w}$ and threshold $t \in \mathbb{R}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t) \neq y] \leq \mathrm{OPT}+\epsilon$

Define: $T = C\ d\log(1/\epsilon)/(\rho\epsilon)^2$, $N = C\ d/\epsilon^2 \log(1/\epsilon)\log(T/\delta)$, for some large enough constant $C > 0$, $\mathcal{V} = \{(\mathbf{w}, t) \in \mathbb{R}^{d+1} : \|\mathbf{w}\|_2 \leq 1\}, |t| \leq 4\sqrt{\log(1/\epsilon)}\}$

1. $\mathbf{w}^{(0)} \leftarrow \mathbf{e}_1$, $t^{(0)} \leftarrow 0$
2. For $i \in [T]$ do
   (a) $\eta_i \leftarrow 1/(\sqrt{i} + \rho)$
   (b) If $(\mathbf{w}^{(i)}, t^{(i)}) = \mathbf{0}$ then $T^{(i)}(\mathbf{x}) = 0$.
   (c) Else let ANS $\leftarrow \mathcal{O}((\mathbf{w}^{(i)}, t^{(i)}))$.
   (d) If ANS = FAIL then **return** $(\mathbf{w}^{(i)}, t^{(i)})$ else $T^{(i)} \leftarrow$ ANS
   (e) Let $\nabla\hat{r}_i(\mathbf{w}, t)$ be an estimator of $-\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(T^{(i)}(\mathbf{x}) + \rho\right) y(\mathbf{x}, 1)\right]$ (Lemma 4.19)
   (f) $(\mathbf{w}^{(i+1)}, t^{(i+1)}) \leftarrow \mathrm{proj}_{\mathcal{V}}\left((\mathbf{w}^{(i)}, t^{(i)}) - \eta_i\nabla\hat{r}_i(\mathbf{w}^{(i)}, t^{(i)})\right)$
3. **return** $(\mathbf{w}^{(T+1)}, t^{(T+1)})$

Algorithm 1: Learning Halfspaces with η-Massart noise.

where we used the Cauchy-Schwarz inequality and the fact that $\mathbf{x}$ is standard normal. □

Since we do not have access to r_i precisely, we need a function $\hat{r}_i$, which is close to r_i with high probability. The following simple lemma gives us an efficient way to compute an approximation $\hat{r}_i$ of r_i.

Lemma 4.19 (Estimating the function r_i)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal and let $T^{(i)}(\mathbf{x})$ be a non-negative function returned by a (2ρ)-certificate oracle. Moreover, assume that $T^{(i)}(\mathbf{x})$ has bounded ℓ_4 norm, i.e., $\|T^{(i)}(\mathbf{x})\|_4 \leq 1$. Then after drawing $O(d\log(1/\epsilon)/\epsilon^2 \log(d/\delta))$ samples from $\mathcal{D}$, with probability at least $1-\delta$, we can compute an estimator $\hat{r}_i$ that satisfies the following conditions:*

- $\left\|\nabla\hat{r}_i(\mathbf{w}, t) - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(T^{(i)}(\mathbf{x}) + \rho)y(\mathbf{x}, 1)]\right\|_2 \leq \epsilon/\sqrt{\log(1/\epsilon)}$
- $\|\nabla\hat{r}_i(\mathbf{w}, t)\|_2 \leq 2\sqrt{d}$.

The proof of the lemma above can be found on Appendix F. We now proceed with the proof of Proposition 4.16.

Proof of Proposition 4.16. Note for the proof for simplicity, we assume that we have access to (2ρ)-certifying oracle, but the same argument works for $\rho' = \rho/2$ and access to ρ'-certifying oracle. Let

$\ell^{(i)}(\mathbf{x}) = \mathbf{w}^{(i)} \cdot \mathbf{x} + t^{(i)}$. We define $\mathcal{V} = \{(\mathbf{w}, t) \in \mathbb{R}^{d+1} : \|\mathbf{w}\|_2 \leq 1, |t| \leq 4\sqrt{\log(1/\epsilon)}\}$. Let T be the number of optimization steps that our algorithm runs. Assume, in order to reach a contradiction, that for all steps $i \in [T]$ it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell^{(i)}(\mathbf{x})) \neq y] \geq \mathrm{OPT} + \epsilon$. Let $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t^*)$. For each step i, let define $T^{(i)}(\mathbf{x})$ to be the non-negative function outputted by the (2ρ)-certifying oracle $\mathcal{O}$. Thus, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T^{(i)}(\mathbf{x}) y(\mathbf{w}^{(i)} \cdot \mathbf{x} + t^{(i)})] \leq -\rho \|\ell^{(i)}(\mathbf{x})\|_2 \,.$$

Let $\nabla \hat{r}_i(\mathbf{w}, t)$ be an estimator of $\nabla r_i(\mathbf{w}, t) = -\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\left(T^{(i)}(\mathbf{x}) + \rho\right) y(\mathbf{x}, 1)]$ such that with probability at least $1 - \frac{\delta}{T}$ it holds $\|\nabla \hat{r}_i(\mathbf{w}, t) - \nabla r_i(\mathbf{w}, t)\|_2 \leq \frac{1}{2}\rho\epsilon^2$ for all $(\mathbf{w}, t) \in \mathcal{V}$. Moreover, from Lemma 4.19, for $N = \tilde{O}(d/\log(T/\delta)/(\epsilon^4\rho^2))$ samples, we can achieve that.

Using the separation between the optimal hypothesis and the current one, i.e., Lemma 4.18, for every step $i \in [T]$, we have that $r_i(\mathbf{w}^{(i)}, t^{(i)}) - r_i(\mathbf{w}^*, t^*) \geq \frac{1}{2}\rho\epsilon^2$. Therefore, it holds

$$\hat{r}_i(\mathbf{w}^{(i)}, t^{(i)}) - \hat{r}_i(\mathbf{w}^*, t^*) \geq \frac{1}{4}\rho\epsilon^2 \,, \tag{31}$$

with probability at least $1 - \frac{\delta}{T}$. In order to apply the online gradient descent algorithm, i.e., Lemma 4.17, we need to define the parameters of the algorithm. First, the convex set we optimize over is $\mathcal{V}$, hence, the diameter of $\mathcal{V}$ is $K = O(\sqrt{\log(1/\epsilon)})$. Furthermore, from Lemma 4.19, we get that $\|\nabla \hat{r}_i(\mathbf{w}, t)\|_2 = O(\sqrt{d})$, and therefore, from Lemma 4.17, it holds

$$\frac{1}{T}\sum_{i=1}^{T}\left(\hat{r}_i(\mathbf{w}^{(i)}, t^{(i)}) - \hat{r}_i\left(\mathbf{w}^*, t^*\right)\right) \lesssim \frac{\sqrt{d} + \sqrt{\log(1/\epsilon)}}{\sqrt{T}} \,.$$

By the union bound, it follows that with probability at least $1 - \delta$, we have that

$$\frac{1}{4}\rho\epsilon^2 \leq \frac{1}{T}\sum_{i=1}^{T}\left(\hat{r}_i(\mathbf{w}^{(i)}, t^{(i)}) - \hat{r}_i(\mathbf{w}^*, t^*)\right) \lesssim \frac{\sqrt{d} + \sqrt{\log(1/\epsilon)}}{\sqrt{T}} \,,$$

which leads to a contradiction for $T = \Theta(\frac{d\log(1/\epsilon)}{\rho^2\epsilon^4})$.

Thus, either there exists $i \in [T]$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell^{(i)}(\mathbf{x})) \neq y] \leq \mathrm{OPT} + \epsilon$, which the algorithm returns, or the (2ρ)-certifying oracle $\mathcal{O}$ did not provide a correct certificate, which happens with probability at most δ. Moreover, the algorithm calls the certificate oracle T times and the number of samples needed are N. The runtime is the number of steps and multiplies by the number of samples $T\ N$. This completes the proof. □

Given Proposition 4.16, the proof of Theorem 4.1 follows by showing that the function returned by Proposition 4.14 is a ρ-certifying oracle, for an appropriate choice of ρ.

Proof of Theorem 4.1. In order to prove Theorem 4.1, we need to construct a ρ-certifying oracle. Using $N = O(d^{\log(1/(\gamma\beta))})\log(1/\delta)/\epsilon^2$ samples, Proposition 4.14 provides with probability $1 - \delta$ a $\rho = \epsilon^2\mathrm{poly}(\beta\gamma)$ certifying oracle in runtime $\mathrm{poly}(N, d)$. Therefore, from Proposition 4.16 the proof follows. The overall samples complexity is $O(N\log(T))$ and the runtime is $\mathrm{poly}(N, d)$. □

5 Learning Halfspaces with General Massart Noise

In this section, we present our algorithm for learning halfspaces with general Massart noise. The main result of this section is the following theorem.

Theorem 5.1 (Learning Halfspaces with General Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta = 1/2$ with respect to a target (possibly biased) halfspace $f \in \mathcal{C}$. Let $\epsilon, \delta \in (0, 1]$. There exists an algorithm that draws $N = d^{O(\log(1/\epsilon))} \log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d) 2^{\mathrm{poly}(1/\epsilon)}$, and computes a halfspace $h \in \mathcal{C}$ such that with probability at least $1 - \delta$ it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$.*

We remark that the algorithm of Theorem 5.1 works for any halfspace, regardless of whether it is biased or not. In the presentation that follows, we will focus on learning homogeneous halfspaces for the sake of simplicity. It is not hard to generalize the algorithm to work for arbitrary halfspaces. (We present the general algorithm for arbitrary halfspaces in Appendix D.) The proof of Theorem 5.1 consists of three main parts; see Section 2.2 for a roadmap of the proof. First, in Subsection 5.2 we show that by restricting on a thin slice and then projecting the distribution on the subspace $\mathbf{w}^{\perp}$, we obtain an instance where the optimal halfspace has again 1/2-Massart noise everywhere apart from a small region close to the halfspace, see Figure 3a. In Subsection 5.1, we present our main technical contribution, i.e., that there exists a low-degree, mean-zero, polynomial that achieves non-trivial correlation when $\eta(\mathbf{x}) \leq 1/2$ "almost everywhere". Finally, in Subsection 5.3 we show that the existence of such polynomials implies that by finding the left singular vectors of (a flattened version) of the Chow tensors of the distribution, we can construct a $\mathrm{poly}(1/\epsilon)$-dimensional subspace inside which the direction $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ has $\mathrm{poly}(\epsilon)$ projection. Recall that $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ is the normalized projection of $\mathbf{w}^*$ onto the orthogonal complement of $\mathbf{w}^{\perp}$:

$$(\mathbf{w}^*)^{\perp_{\mathbf{w}}} = \frac{\mathrm{proj}_{\mathbf{w}^{\perp}}(\mathbf{w}^*)}{\|\mathrm{proj}_{\mathbf{w}^{\perp}}(\mathbf{w}^*)\|_2} .$$

In Subsection 5.3, we combine everything together to prove Theorem 5.1.

We now state the main technical result of this section. (This is the formal version of Proposition 2.5.)

Proposition 5.2. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta = 1/2$, with respect to some target halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Let $\mathbf{w} \in \mathbb{R}^d$ be a unit vector such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$ and $\theta(\mathbf{w}, \mathbf{w}^*) \leq \pi - \epsilon$ for some $\epsilon \in (0, 1/2]$. There exists an algorithm that draws $N = d^{O(\log(1/\epsilon))} \log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and with probability at least $1 - \delta$ returns a basis of a subspace $V \subseteq \mathbf{w}^{\perp}$ such that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp_{\mathbf{w}}})\|_2 \geq \mathrm{poly}(\epsilon)$.*

Observe that given a subspace V with the above guarantees, we can sample a uniformly random direction on the unit sphere of V and obtain a unit update vector v such that $\mathbf{v} \in \mathbf{w}^{\perp}$ and $\mathbf{v} \cdot (\mathbf{w}^*)^{\perp_{\mathbf{w}}} \geq \mathrm{poly}(\epsilon)$, which given Lemma 5.13 improves the current guess $\mathbf{w}$; see Subsection 5.3 for the details.

5.1 The Sign-Matching Polynomial

Here we give an explicit construction of a mean-zero low-degree polynomial that achieves non-trivial correlation with the labels y. The assumptions on the distribution $\mathcal{D}$ over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ are that it has Gaussian $\mathbf{x}$-marginal and "almost-Massart" noise, i.e., that is $\eta(\mathbf{x})$ is greater than 1/2 only on a very small region close to the optimal halfspace. We prove the following:

Proposition 5.3 (Correlation via the Sign-Matching Polynomial)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal distribution. Let $f(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} - b)$ be such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(b)f(\mathbf{x}) > 0\}] \geq \zeta$ for some $\zeta \in (0, 1/2]$, and $\eta(\mathbf{x}) > 1/2$ only when $0 \leq \mathrm{sign}(b)(\mathbf{v}^* \cdot \mathbf{x} - b) \leq \xi$, where ξ is a sufficiently small constant multiple of ζ^3. There exists a univariate polynomial $p(z)$ of degree $\Theta(b^2 + 1)$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{v}^* \cdot \mathbf{x})] = 0$, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p^2(\mathbf{v}^* \cdot \mathbf{x})] = 1$, and $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[yp(\mathbf{v}^* \cdot \mathbf{x})] = \mathrm{poly}(\zeta)$.*

Proof. The main ingredient of the proof is the following lemma that shows the existence of a mean zero polynomial that matches the sign of the linear function $z - b$; see Figure 3b.

Lemma 5.4 (Sign-Matching Polynomial)**.** *Let $b \in \mathbb{R}$. There exists a zero mean and unit variance polynomial $p : \mathbb{R} \mapsto \mathbb{R}$ of degree $k = \Theta(b^2 + 1)$ such that*

- *The sign of p matches the sign of the threshold function $\mathrm{sign}(z - b)$, i.e., $\mathrm{sign}(p(z)) = \mathrm{sign}(z - b)$ all $z \in \mathbb{R}$,*
- *for any $\rho \in (0, 1)$, it holds $|p(b + \mathrm{sign}(b)\rho)| \geq c^k \mathrm{poly}(\rho)$, where $c > 0$ is a universal constant,*
- *$p(z)$ is increasing for all $|z| \geq |b|$.*

Proof. We first pick some odd integer k large enough such that $2|b| \leq \sqrt{k} \leq 4 \max(|b|, 1)$. Let $q(z) = z^{3k}, r(z) = z^{2k} - (2k - 1)!!$. We consider the polynomial

$$p(z) = q(z) - \frac{q(b)}{r(b)} r(z) \,.$$

It holds $p(b) = 0$ and $\mathbf{E}_{z\sim\mathcal{N}}[p(z)] = 0$, since $\mathbf{E}_{z\sim\mathcal{N}}[z^{3k}] = 0$, for odd k and $\mathbf{E}_{z\sim\mathcal{N}}[z^{2k}] = (2k - 1)!!$.

For the rest of the proof, we need the following simple estimate for double factorials. Sharper bounds can be obtained via Stirling's approximation; for our purposes the following rough bounds suffice.

Fact 5.5. *Let m be a positive integer. Then $(m/2)^m \leq (2m - 1)!! \leq (2m)^m$.*

We next show that $p(z)$ is increasing for all $|z| \geq |b|$. We compute the derivative of $p(z)$, i.e.,

$$p'(z) = 3kz^{3k-1} - 2k\frac{q(b)}{r(b)} z^{2k-1} = kz^{2k-1}\left(3z^k - 2\frac{q(b)}{r(b)}\right) \,, \tag{32}$$

we observe that if $|z^k| \geq \frac{2}{3}|q(b)/r(b)|$, then $p(z)$ is increasing. We show that this is the case for all $|z| \geq |b|$. Since $|z|^k$ is increasing in $|z|$ it suffices to verify the inequality for $|z| = |b|$. In that case we obtain the inequality $(2k - 1)!!/b^{2k} \geq 5/3$, which can be verified by using Fact 5.5 and our assumption that $|b| \leq \sqrt{k}/2$.

We now prove that $\mathrm{sign}(p(z)) = \mathrm{sign}(z - b)$. Note that $\mathrm{sign}(r(z)) < 0$ for any $z \in [-|b|, |b|]$ because $|b| \leq \sqrt{k}/2$ and thus, from Fact 5.5, it holds $b^{2k} - (2k - 1)!! \leq (k/4)^k - (k/2)^k \leq 0$. Without loss of generality, assume that $b > 0$. Assume that there exists $s \in \mathbb{R}$ with $s \neq b$ such that $p(s) = 0$. First, assume that $s \geq 0$, by eq. (32), since $r(b) < 0$ we get that $p(z)$ is increasing for $z \geq 0$, therefore the polynomial $p(z)$ has only one root at $s = b$, therefore we have a contradiction. For the case that $s < 0$, first observe that for any $s = 0$, it holds that $q(s)r(b) = q(b)r(s)$ and in particular $\mathrm{sign}(q(s)r(b)) = \mathrm{sign}(q(b)r(s))$. Thus, if $-b > s$, then it holds $\mathrm{sign}(q(b)r(s)) = -\mathrm{sign}(b)$ and $\mathrm{sign}(q(s)r(b)) = -\mathrm{sign}(s)$, therefore for any such solution it holds $\mathrm{sign}(s) = \mathrm{sign}(b)$ but $s < 0$ and $b > 0$, so we have a contradiction. Finally, for the case that $s \leq -b$, recall that for $|z| \geq |b|$, $p(z)$ is increasing, so for $z < -b$, we have that $p(z)$ is increasing. But $p(-b) < 0$ because there is no other root in the interval $[-b, b)$, thus $p(s) < 0$ which leads to a contradiction. Therefore, the only root of the polynomial is at $s = b$.

We next prove that $\mathbf{E}_{z\sim\mathcal{N}}[p^2(z)] = (O(k))^{3k}$. By applying twice the inequality $(a+b)^2 \leq 2a^2+2b^2$, we get

$$\mathbf{E}_{z\sim\mathcal{N}}[p^2(z)] \lesssim \mathbf{E}_{z\sim\mathcal{N}}[z^{6k}] + \frac{q^2(b)}{r^2(b)}\left(\mathbf{E}_{z\sim\mathcal{N}}[z^{4k}] + ((2k-1)!!)^2\right) \lesssim (6k-1)!! + \frac{q^2(b)}{r^2(b)}(4k-1)!!\,.$$

Using that $|b| \leq \sqrt{k}/2$ and Fact 5.5, we have that $|b^{2k} - (2k-1)!!| \geq (k/2)^k/2$. Hence, we get that $(q(b)/r(b))^2 \lesssim (2|b|^3/k)^{2k} \lesssim k^k$. Therefore, using Fact 5.5, we get that $\mathbf{E}_{z\sim\mathcal{N}}[p^2(z)] = O((6k)^{3k})$.

Without loss of generality, assume that $b > 0$. Observe that, for $\rho \in (0, 1)$, we have

$$p(b+\rho) = (b+\rho)^{3k} - \frac{b^{3k}((b+\rho)^{2k} - (2k-1)!!)}{b^{2k} - (2k-1)!!} = (b+\rho)^{3k} - b^{3k} - \frac{b^{3k}\sum_{m=0}^{2k-1}\binom{2k}{m}b^m\rho^{2k-m}}{b^{2k} - (2k-1)!!}\,.$$

Recall that using the fact that $|b| \leq \sqrt{k}/2$, it holds that $b^{2k} - (2k-1)!! < 0$, thus

$$p(b+\rho) \geq (b+\rho)^{3k} - b^{3k} = \sum_{m=0}^{3k-1}\binom{3k}{m}b^m\rho^{3k-m}\,.$$

From the equation above, we get that $p(b+\rho) \gtrsim b^{3k-1}\rho + \rho^{3k}$. Let $\tilde{p}(z)$ be the polynomial $p(z)$ normalized, i.e., $\tilde{p}(z) = p(z)/(\mathbf{E}_{z\sim\mathcal{N}}[p^2(z)])^{1/2}$. Notice that normalizing p does not affect its sign or monotonicity. It remains to prove that $\tilde{p}(b+\rho) = c^k\mathrm{poly}(\rho)$, where $c > 0$ is a small enough constant. If $b \geq 1$, recall that in that case it holds that $\sqrt{k} \leq 4b$, therefore it holds $\tilde{p}(b+\rho) \gtrsim b^{3k-1}\rho/(6k)^{3k/2} \gtrsim \rho/(24)^{2k}$. For the case where $0 < b < 1$, we have that $k = O(1)$ and similarly, we have $\tilde{p}(b+\rho) \gtrsim \rho^{2k}/(6k)^{3k/2} = \mathrm{poly}(\rho)$, hence $\tilde{p}(b+\rho) = c^k\mathrm{poly}(\rho)$, for some universal constant $c > 0$. □

It remains to show that the polynomial p of the Lemma 5.4, achieves non-trivial correlation with y, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[yp(\mathbf{v}^*\cdot\mathbf{x})] = \mathrm{poly}(\zeta)$. To do this we are going to use the monotonicity properties of p, the fact that it matches the sign of $f(\mathbf{x})$, and the fact that it is not very flat close to its unique root at b, see Item 2 of Lemma 5.4. We assume that $b \geq 0$ as the other case follows similarly. Recall that, from the assumptions of Proposition 5.3, there is an explicitly defined region where the noise is above $1/2$, i.e., $\eta(\mathbf{x}) > 1/2$. We denote this region I^ξ, see also Figure 7, that is

$$I^\xi = \{\mathbf{x} \in \mathbb{R}^d : \eta(\mathbf{x}) > 1/2\} \subseteq \{\mathbf{x} \in \mathbb{R}^d : b \leq \mathbf{v}^*\cdot\mathbf{x} \leq b+\xi\}\,.$$

Since $p(\mathbf{v}^*\cdot\mathbf{x})$ (p from Lemma 5.4) matches the sign of $\mathbf{v}^*\cdot\mathbf{x} - b$ we have

$$\begin{aligned}\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{v}^*\cdot\mathbf{x})y] &= \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{v}^*\cdot\mathbf{x})\mathrm{sign}(\mathbf{v}^*\cdot\mathbf{x} - b)\beta(\mathbf{x})(\mathbb{1}\{\mathbf{x} \notin I^\xi\} - \mathbb{1}\{\mathbf{x} \in I^\xi\})] \\ &\geq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|p(\mathbf{v}^*\cdot\mathbf{x})|\beta(\mathbf{x})\mathbb{1}\{\mathbf{x} \notin I^\xi\}] - \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|p(\mathbf{v}^*\cdot\mathbf{x})|\mathbb{1}\{\mathbf{x} \in I^\xi\}]\,. \end{aligned} \tag{33}$$

We first bound from below the correlation of the sign-matching polynomial outside the region I^ξ, i.e., the region where the Massart noise condition is true. From Lemma 5.4, we have that $p(z)$ is

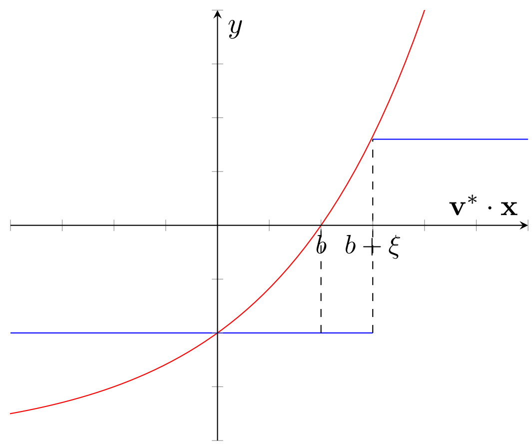


Figure 7: The sign-matching polynomial p corresponds to the red curve and the threshold function $f(\mathbf{x})$ to the blue. Observe that, in the region $I^\xi = \{b \le \mathbf{v}^* \cdot \mathbf{x} \le b+\xi\}$, the sign of the polynomial p does not agree with $f(\mathbf{x})$. This is the region where the Massart noise condition is (potentially) violated, as a result of the orthogonal projection step, Figure 3a. Even though the sign of p does not agree with $f(\mathbf{x})$ everywhere, we can take ξ to be small, i.e., ξ is a sufficiently small constant multiple of ζ^3, making the negative contribution of the region I_ξ small. We also use crucially the fact that p is monotone for $\mathbf{v}^* \cdot \mathbf{x} \ge b$ and not very flat in that region, see Item 2 of Lemma 5.4.

increasing for $z > b$, and therefore it holds

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|p(\mathbf{v}^* \cdot \mathbf{x})|\beta(\mathbf{x})\mathbb{1}\{\mathbf{x} \notin I^\xi\}] \\
&\quad\ge \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|p(\mathbf{v}^* \cdot \mathbf{x})|\beta(\mathbf{x})\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b+\zeta^2\}] \\
&\quad\ge |p(b+\zeta^2)| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b+\zeta^2\}] \\
&\quad= |p(b+\zeta^2)| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b\}] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{b \le \mathbf{v}^* \cdot \mathbf{x} \le b+\zeta^2\}] \right) , \qquad (34)
\end{aligned}$$

where for the first inequality we used the fact that $\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b+\xi\} \ge \mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b+\zeta^2\}$, because $\xi = \Theta(\zeta^3) \le \zeta^2$ from the assumptions of Proposition 5.3, and for the second, we used the monotonicity of p. We next bound from above the contribution of the sign-matching polynomial on the region where the Massart noise condition is violated, i.e., $\mathbf{x} \in I^\xi$. Moreover, using again the fact that $p(z)$ is increasing for $z > b$, we get

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|p(\mathbf{v}^* \cdot \mathbf{x})|]\mathbb{1}\{\mathbf{x} \in I^\xi\}] &\le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|p(\mathbf{v}^* \cdot \mathbf{x})|\mathbb{1}\{b \le \mathbf{v}^* \cdot \mathbf{x} \le b+\zeta^2\}] \\
&\le |p(b+\zeta^2)| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{b \le \mathbf{v}^* \cdot \mathbf{x} \le b+\zeta^2\}] . \qquad (35)
\end{aligned}$$

Substituting the bounds of eq. (34) and eq. (35) to eq. (33), we get that

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{v}^* \cdot \mathbf{x})y] &\ge |p(b+\zeta^2)| \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} \ge b\}] - 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{b \le \mathbf{v}^* \cdot \mathbf{x} \le b+\zeta^2\}] \right) \\
&\gtrsim |p(b+\zeta^2)|(\zeta - \zeta^2) \gtrsim |p(b+\zeta^2)|\zeta ,
\end{aligned}$$

where we used that from the assumptions of Proposition 5.3 we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(b)f(\mathbf{x}) > 0\}] \gtrsim \zeta$, and the anti-concentration property of the Gaussian distribution from Fact 3.1, i.e., $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{b < \mathbf{v}^* \cdot \mathbf{x} < b+\zeta^2\}] \le \zeta^2/(\sqrt{2\pi})$.

It remains to prove that $|p(b+\zeta^2)| = \mathrm{poly}(\zeta)$. From Lemma 5.4, we have that $|p(b+\zeta^2) = c^k\mathrm{poly}(\zeta)$, for some universal constant $c>0$. Note that from the assumption of Proposition 5.3 we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(b)\mathrm{sign}(\mathbf{v}^*\cdot\mathbf{x}-b)>0\}] \gtrsim \zeta$, hence, from Fact 3.1 we get $|b| \lesssim \sqrt{\log(1/\zeta)}$, therefore $c^k = \mathrm{poly}(\zeta)$, since $k=\Theta(b^2+1)$. Hence, we have $|p(b+\zeta^2)| = \mathrm{poly}(\zeta)$. This completes the proof of Proposition 5.3.

□

5.2 Projecting onto $\mathbf{w}^\perp$

In this subsection, we show that we can project $\mathcal{D}$ onto the subspace $(\mathbf{w}^*)^{\perp_\mathbf{w}}$ without violating the 1/2-Massart noise condition by a lot. At the same time we make sure that the "optimal halfspace" of the projected instance is not very biased. We show the following lemma.

Lemma 5.6 (Orthogonal Projection Inside a Band)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta=1/2$ with respect to $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})$. Let $\mathbf{w}$ be a unit vector such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\neq y] \geq \mathrm{OPT}+\epsilon$, and $\theta(\mathbf{w},\mathbf{w}^*) \leq \pi-\epsilon$, for some $\epsilon\in(0,1]$. Fix some $\rho>0$. For $t_1,t_2\in\mathbb{R}$, consider the band $B=\{t_1\leq \mathbf{w}\cdot\mathbf{x}\leq t_2\}$. Denote $\mathcal{D}^\perp = \mathcal{D}_B^{\perp_\mathbf{w}}$, i.e., $\mathcal{D}$ is the orthogonal projection onto $\mathbf{w}^\perp$ of the conditional distribution on B, and consider the halfspace $f^\perp:\mathbf{w}^\perp\mapsto\{\pm1\}$, with $f^\perp(\mathbf{x}) = \mathrm{sign}(\mathbf{x}\cdot(\mathbf{w}^*)^{\perp_\mathbf{w}}-b)$, for some threshold $b\in\mathbb{R}$. Moreover, define the noise function $\eta^\perp(\mathbf{x}) = \mathbf{Pr}_{(\mathbf{z},y)\sim\mathcal{D}^\perp}[y\neq f^\perp(\mathbf{z})|\mathbf{z}=\mathbf{x}]$. There exist $t_1,t_2\in\mathbb{R}$ multiples of ρ, with $|t_1-t_2|=\rho$, such that:*

1. $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}^\perp_\mathbf{x}}[(1-2\eta^\perp(\mathbf{x}))\mathbb{1}\{f^\perp(\mathbf{x})\mathrm{sign}(b)>0\}] \gtrsim \epsilon/\sqrt{\log(1/\epsilon)}-\rho/\epsilon$,
2. *if* $\eta^\perp(\mathbf{x})>1/2$ *then* $0<\mathrm{sign}(b)(\mathbf{x}\cdot(\mathbf{w}^*)^{\perp_\mathbf{w}}-b)\leq\rho/\epsilon$.

Remark 5.7. Observe that for Lemma 5.6 to give non-trivial guarantees we have to pick the size of the band, ρ, to be smaller than a sufficiently small constant multiple of $\epsilon/\sqrt{\log(1/\epsilon)}$. In fact, in what follows, we will be using bands of size $\rho=\mathrm{poly}(\epsilon)$. Moreover, the condition that $\theta(\mathbf{w},\mathbf{w}^*)\geq\pi-\epsilon$ is mostly technical. Having a vector $\mathbf{w}$ with $\theta(\mathbf{w},\mathbf{w}^*)\geq\pi-\epsilon$ we can always output $-\mathbf{w}$ and achieve classification error $O(\epsilon)$. See the proof of Theorem 5.1.

Proof. First, let us denote $\beta(\mathbf{x}) = 1-2\eta(\mathbf{x})$. Note that the assumption $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\neq y]\geq\mathrm{OPT}+\epsilon$ is equivalent to $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{f(\mathbf{x})\neq\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\}\beta(\mathbf{x})]\geq\epsilon$. We consider the region $B=\{t_1\leq\mathbf{w}\cdot\mathbf{x}\leq t_2\}$, for any $t_1,t_2\in\mathbb{R}$ with $|t_1-t_2|=\rho$. We denote by $\mathbf{x}^\perp$ the projection of $\mathbf{x}$ onto the subspace $\mathbf{w}^\perp$. Define the distribution $\mathcal{D}^\perp = \mathcal{D}_B^{\mathrm{proj}_{\mathbf{w}^\perp}}$, that is the distribution $\mathcal{D}$ conditioned on the set B and projected onto $\mathbf{w}^\perp$, the hypothesis $f^\perp(\mathbf{x}^\perp) = \mathrm{sign}(\mathbf{x}^\perp\cdot(\mathbf{w}^*)^{\perp_\mathbf{w}}+b)$ where $b\in\mathbb{R}$ is chosen appropriately below, and the noise function $\eta^\perp(\mathbf{x}^\perp) = \mathbf{Pr}_{(\mathbf{z},y)\sim\mathcal{D}^\perp}[y\neq f^\perp(\mathbf{z})|\mathbf{z}=\mathbf{x}^\perp]$, see Figure 3a.

Note that the distribution $\mathcal{D}^\perp$ does not satisfy the 1/2-Massart noise condition anymore. We first illustrate how the noise function changes. The orthogonal projection on $\mathbf{w}^\perp$ creates a region where the 1/2-Massart noise condition is violated, i.e., a region where $\eta^\perp(\mathbf{x}^\perp)$ may be more than 1/2, but we can control the probability that we get points inside this region, i.e., the green area of Figure 3a. More formally, we show that

$$\mathop{\mathbf{E}}_{(\mathbf{x}_\mathbf{w},\mathbf{x}^\perp)\sim(\mathcal{D}_B)_\mathbf{x}}\left[\beta(\mathbf{x})\mathbb{1}\{f(\mathbf{x}_\mathbf{w},\mathbf{x}^\perp)\neq f^\perp(\mathbf{x}^\perp)\}\right] \leq \mathop{\mathbf{Pr}}_{\mathbf{x}^\perp\sim\mathcal{D}^\perp_\mathbf{x}}[\eta^\perp(\mathbf{x}^\perp)>1/2] \lesssim \rho/\epsilon\,. \tag{36}$$

To show Inequality (36) notice that $\theta(\mathbf{w},\mathbf{w}^*)\gtrsim\epsilon$, otherwise we would have $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f(\mathbf{x})\neq\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})]\leq\epsilon$ from Fact 3.2. We can assume that $\mathbf{w}^* = \lambda_1\mathbf{w}+\lambda_2(\mathbf{w}^*)^{\perp_\mathbf{w}}$, where $\lambda_1=\cos\theta$ and $\lambda_2=\sin\theta$. Note that $\lambda_2\gtrsim\epsilon$, since $\theta\in[\Omega(\epsilon),\pi-\epsilon]$. Next we set $\mathbf{x}=(\mathbf{x}_\mathbf{w},\mathbf{x}^\perp)$, where $\mathbf{x}_\mathbf{w}=\mathbf{w}\cdot\mathbf{x}$. We show that the hypothesis $f^{\perp}(\mathbf{x}) = \mathrm{sign}((\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^{\perp} + t_1\lambda_1/\lambda_2)$ is almost as good as the $f(\mathbf{x})$ for the distribution $\mathcal{D}^{\perp}$. Note that in the statement of Lemma 5.6, we write $f^{\perp}(\mathbf{x}) = \mathrm{sign}((\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^{\perp} - b)$, where $b = -t_1\lambda_1/\lambda_2$.

Conditioned on $\mathbf{x} \in B$, i.e., $\mathbf{x}_\mathbf{w} = \mathbf{x} \cdot \mathbf{w} \in [t_1, t_1 + \rho]$, it holds that

$$\mathbf{w}^* \cdot \mathbf{x} = \lambda_1 \mathbf{x}_\mathbf{w} + \lambda_2 (\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^{\perp} = \lambda_1 t_1 + \lambda_2 (\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^{\perp} + s\rho\,,$$

for some $s \in [-1, 1]$ (recall that $|\lambda_1| \leq 1$). Notice that when $0 < \mathrm{sign}(-\lambda_1 t_1)(\lambda_1 t_1 + \lambda_2 (\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^{\perp}) < \rho$, $f^{\perp}(\mathbf{x}^{\perp})$ is not necessary equal to the sign of $(\mathbf{w}^* \cdot \mathbf{x})$ (recall that $\lambda_2 > 0$), and therefore we are inside the region that the 1/2-Massart noise condition is violated, otherwise the $f^{\perp}(\mathbf{x}^{\perp})$ always matches the sign of $\mathbf{w}^* \cdot \mathbf{x}$. This proves the second statement of Lemma 5.6.

To prove Inequality (36), we need to bound the probability of the event that $f^{\perp}(\mathbf{x}^{\perp})$ is not necessary equal to the sign of $(\mathbf{w}^* \cdot \mathbf{x})$, or equivalently if $\mathrm{sign}(-\lambda_1 t_1) > 0$ then $(\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^{\perp} \in [(-\lambda_1 t_1 - \rho)/\lambda_2, -\lambda_1 t_1/\lambda_2] =: I^{\rho}$, and otherwise $(\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^{\perp} \in [-\lambda_1 t_1/\lambda_2, -(\lambda_1 t_1 + \rho)/\lambda_2] =: I^{\rho}$. We have that

$$\Pr_{\mathbf{x}^{\perp} \sim \mathcal{D}^{\perp}_{\mathbf{x}}}[\eta^{\perp}(\mathbf{x}^{\perp}) > 1/2] \leq \Pr_{\mathbf{x}^{\perp} \sim \mathcal{D}^{\perp}_{\mathbf{x}}}\left[(\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^{\perp} \in I^{\rho}\right] = \frac{\Pr_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}\left[(\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x} \in I^{\rho}\right]}{\Pr_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbf{x} \in B]} \lesssim \rho/\lambda_2 \lesssim \rho/\epsilon\,,$$

where we used the anti-concentration property of the Gaussian distribution and the last inequality holds because we have that $\lambda_2 \gtrsim \epsilon$. This proves Inequality (36).

It remains to prove that there exist thresholds t_1, t_2 and a band $B = \{t_1 \leq \mathbf{w} \cdot \mathbf{x} \leq t_2\}$ with respect the t_1, t_2 such that $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}^{\perp}_{\mathbf{x}}}[(1 - 2\eta^{\perp}(\mathbf{x}))\mathbb{1}\{f^{\perp}(\mathbf{x}) \neq \mathrm{sign}(-b)\}] \gtrsim \frac{\epsilon}{\sqrt{\log(1/\epsilon)}}$, where $\mathcal{D}^{\perp}$ is the distribution $\mathcal{D}$ conditioned on the set B and projected onto $\mathbf{w}^{\perp}$. We first show the following claim

Claim 5.8. *Let* $\mathbf{v}$ *be a unit vector and let* S *be a subset of* $\mathbb{R}^d$ *such that* $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathbf{x} \in S\}\beta(\mathbf{x})] \geq \epsilon$, *for some* $\epsilon > 0$. *Fix* $\rho > 0$. *There exist* $t_1, t_2 \in \mathbb{R}$ *such that* t_1, t_2 *are multiples of* ρ, $|t_1 - t_2| = \rho$, *and*

$$\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathbf{x} \in S\}\mathbb{1}\{t_1 \leq \mathbf{v} \cdot \mathbf{x} \leq t_2\}\beta(\mathbf{x})] \gtrsim \frac{\epsilon\rho}{\sqrt{\log(1/\epsilon)}}\,.$$

Proof. Note that since $\beta(\mathbf{x})\mathbb{1}\{\mathbf{x} \in S\} \in [0, 1]$, we have $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathbf{x} \in S\}\mathbb{1}\{C\sqrt{\log(1/\epsilon)} \leq |\mathbf{v} \cdot \mathbf{x}|\}\beta(\mathbf{x})] \leq \mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbb{1}\{C\sqrt{\log(1/\epsilon)} \leq |\mathbf{v} \cdot \mathbf{x}|\}] \leq \epsilon^2$, where C is a large enough universal constant. To get this we used the upper bound on the Gaussian tails, Fact 3.1. Therefore, we can ignore the region far away from the origin and split the interval $|\mathbf{v} \cdot \mathbf{x}| \leq C\sqrt{\log(1/\epsilon)}$, into slices of length ρ. There must exist one slice inside which the expectation of $\beta(\mathbf{x})\mathbb{1}\{\mathbf{x} \in S\}$ is roughly $\epsilon\rho/\sqrt{\log(1/\epsilon)}$. Let $t_i = i\rho$ and $t_{-i} = -i\rho$, for $0 \leq i \leq C\sqrt{\log(1/\epsilon)}/\rho$, we define $B_i = \{t_i \leq \mathbf{v} \cdot \mathbf{x} \leq t_{i+1}\}$ and $B_{-i} = \{-t_{i+1} \leq \mathbf{v} \cdot \mathbf{x} \leq -t_i\}$. We have that

$$\sum_{|i| \leq C\sqrt{\log(1/\epsilon)}/\rho} \mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathbf{x} \in S\}\mathbb{1}\{\mathbf{x} \in B_i\}\beta(\mathbf{x})] \geq \epsilon - \epsilon^2 \geq \epsilon/2\,, \tag{37}$$

where we used that $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathbf{x} \in S\}\mathbb{1}\{C\sqrt{\log(1/\epsilon)} \leq |\mathbf{v} \cdot \mathbf{x}|\}\beta(\mathbf{x})] \leq \epsilon^2$. Using the fact that all the quantities we add in Equation (37) are positive, there exists B'_i, for some i', such that

$$\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathbf{x} \in S\}\mathbb{1}\{\mathbf{x} \in B_{i'}\}\beta(\mathbf{x})] \gtrsim \frac{\epsilon\rho}{\sqrt{\log(1/\epsilon)}}\,,$$

which completes the proof. □

An application of Claim 5.8 to the set $\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x})\}$, gives us that there exists a band $B = \{t_1 \leq \mathbf{w} \cdot \mathbf{x} \leq t_2\}$ with $|t_1 - t_2| = \rho$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\mathbb{1}\{f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x})\}\mathbb{1}\{\mathbf{x} \in B\}\beta(\mathbf{x})] \gtrsim \frac{\epsilon\rho}{\sqrt{\log(1/\epsilon)}} \,.$$

Moreover, note that for the distribution $\mathcal{D}_B$, that is the distribution $\mathcal{D}$ conditioned on B, it holds $\mathbf{E}_{\mathbf{x} \sim (\mathcal{D}_B)_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x})\}] \gtrsim \frac{\epsilon}{\sqrt{\log(1/\epsilon)}}$, where we used the Gaussian anti-concentration property, i.e., that $\mathbf{Pr}_{x \sim \mathcal{N}}[t_1 \leq x \leq t_2] \lesssim |t_2 - t_1|$.

We have that $f^\perp(\mathbf{x})$ agrees almost everywhere with $f(\mathbf{x})$ with respect to the distribution $\mathcal{D}_B$, i.e., we have that $\mathbf{E}_{(\mathbf{x}_\mathbf{w}, \mathbf{x}^\perp) \sim (\mathcal{D}_B)_\mathbf{x}}[\mathbb{1}\{f((\mathbf{x}_\mathbf{w}, \mathbf{x}^\perp)) \neq f^\perp(\mathbf{x}^\perp)\}] \leq \mathbf{Pr}_{\mathbf{x}^\perp \sim \mathcal{D}_\mathbf{x}^\perp}[\eta^\perp(\mathbf{x}^\perp) > 1/2] \lesssim \rho/\epsilon$ (Inequality (36)). Thus, using the triangle inequality, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x}_\mathbf{w}, \mathbf{x}^\perp) \sim (\mathcal{D}_B)_\mathbf{x}}\left[(1 - 2\eta^\perp(\mathbf{x}^\perp))\mathbb{1}\{\mathrm{sign}(\mathbf{x}_\mathbf{w}) \neq f^\perp(\mathbf{x}^\perp)\}\right] \gtrsim \frac{\epsilon}{\sqrt{\log(1/\epsilon)}} - \frac{\rho}{\epsilon} \,.$$

The proof concludes by noting that it holds $\mathrm{sign}(\mathbf{x}_\mathbf{w}) = \mathrm{sign}(-b)$, by the definition of $f^\perp(\mathbf{x})$. This completes the proof of Lemma 5.6. ☐

5.3 Using the Low-Order Chow Tensors

In this subsection, we use our structural result of Subsection 5.1 to construct the sampling oracle of Proposition 5.2 that provides us with good update vectors $\mathbf{v}$. To do so, we shall find a subspace V of $\mathbb{R}^d$ that contains non-trivial part of $\mathbf{v}^*$. In order to get a good update with non-trivial probability, we will sample a unit vector uniformly at random from V. The description of the corresponding algorithm is given in Algorithm 2.

Proposition 5.9. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Let $f(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} - b)$ be such that $\mathbf{Pr}_{(\mathbf{x},y) \sim \mathcal{D}}[y \neq f(\mathbf{x}) | \mathbf{x}] = \eta(\mathbf{x})$, $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(b)f(\mathbf{x}) > 0\}] \geq \zeta$ and $\eta(\mathbf{x}) > 1/2$ only when $\mathrm{sign}(b)(\mathbf{v}^* \cdot \mathbf{x} - b) \leq \xi$, where $\xi = \Theta(\zeta^3)$. There exists an algorithm with sample complexity and runtime $d^{O(\log(1/\zeta))} \log(1/\delta)$ that with probability at least $1 - \delta$ returns a basis of a subspace V of dimension $\mathrm{poly}(1/\zeta)$ such that $\|\mathrm{proj}_V(\mathbf{v}^*)\|_2 = \mathrm{poly}(\zeta)$.*

Proof. The following lemma shows that the existence of a mean-zero polynomial p that achieves non-trivial correlation with y implies that the subspace spanned by the top singular vectors of the Chow tensors of $\mathcal{D}$ will contain non-trivial part of $\mathbf{v}^*$. We remark that we do not rely on tensor SVD to obtain the singular vectors: in what follows we flatten the order-m Chow tensors and treat them as $d \times d^{m-1}$ matrices. We show the following lemma.

Lemma 5.10. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Let $p : \mathbb{R} \mapsto \mathbb{R}$ be a univariate, mean zero, unit variance polynomial of degree k such that for some unit vector $\mathbf{v}^* \in \mathbb{R}^d$ it holds $\mathbf{E}_{(\mathbf{x},y) \sim \mathcal{D}}[yp(\mathbf{v}^* \cdot \mathbf{x})] \geq \tau$ for some $\tau \in (0, 1]$. Let $\mathbf{T}'^m$ be an approximation of the order-m Chow-parameter tensor $\mathbf{T}^m$ of $\mathcal{D}$ such that $\|\mathbf{T}'^m - \mathbf{T}^m\|_F \leq \tau/(4\sqrt{k})$. Denote by V_m the subspace spanned by the left singular vectors of $(\mathbf{T}'^m)^\flat$ whose singular values are greater than $\tau/(4\sqrt{k})$. Moreover, denote by V the union of $V_1, \ldots, V_k$. Then we have that*

1. *$\dim(V) \leq 4k/\tau^2$, and*
2. *$\|\mathrm{proj}_V(\mathbf{v}^*)\|_2 \geq \tau/(4\sqrt{k})$.*

Input:

1. $\epsilon, \delta > 0$.
2. An empirical distribution $\widehat{\mathcal{D}}$ of the distribution $\mathcal{D}$ that satisfies the general Massart noise condition with respect to $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$.
3. A unit vector $\mathbf{v} \in \mathbb{R}^d$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$.

Output: A subspace V with $\dim(V) = \mathrm{poly}(1/\epsilon)$, and $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp \mathbf{w}})\|_2 \geq \mathrm{poly}(\epsilon)$.

Define: $k = C\log(1/\epsilon)$, $\sigma = \mathrm{poly}(\epsilon)/C$, $\rho = \epsilon^3/C$, $l = C\log(1/\epsilon)/\rho$, for $C > 0$ be a sufficiently large constant.

1. For $i \in [l]$, set $B_i = \{i\rho \leq \mathbf{w} \cdot \mathbf{x} \leq (i+1)\rho\}$ and $B_{-i} = \{-(i+1)\rho \leq \mathbf{w} \cdot \mathbf{x} \leq -i\rho\}$.
2. For $i \in \{-l, \dots, l\}$, repeat
   (a) Let $\widehat{\mathcal{D}}_i^{\perp}$ be the distribution $\widehat{\mathcal{D}}$ conditioned on B_i and projected onto $\mathbf{w}^{\perp}$.
   (b) For any $m \in [k]$, calculate the m-order Chow tensor $\mathbf{T}^m(\widehat{\mathcal{D}}_i^{\perp})$, see Definition 2.7.
   (c) Flatten each $\mathbf{T}^m(\widehat{\mathcal{D}}_i^{\perp})$ for $m \in [k]$, to a $d \times d^{m-1}$ matrix $\mathbf{M}^m = (\mathbf{T}^m(\widehat{\mathcal{D}}_i^{\perp}))^{\flat}$, see Section 3.
   (d) Let V_i^m be the set of the left singular vectors of $\mathbf{M}^m$, for all $m \in [k]$, whose singular values have absolute value at least σ.
3. **Return:** A basis of $V = \bigcup_{i=1}^{\ell} \bigcup_{m=1}^{k} V_i^m$.

Algorithm 2: Creating a Random Oracle with Good Correlation, see Proposition 5.2.

Proof. We note that $\|\mathbf{T}^m\|_F = \sup_{p\in\mathcal{H}_k} \mathbf{E}[yp(\mathbf{x})/\sqrt{\mathbf{E}[p^2(\mathbf{x})]}] \leq \sup_{p\in\mathcal{H}_k} \mathbf{E}[p^2(\mathbf{x})]^{1/2}/\mathbf{E}[p^2(\mathbf{x})]^{1/2} = 1$, where $\mathcal{H}_k$ is the set of polynomials formed from a linear combination of k-degree Hermite polynomials. Therefore, by the assumptions of Lemma 5.10 and the triangle inequality we obtain that $\|\mathbf{T}'^m\|_F \leq \|\mathbf{T}^m\|_F + 1 \leq 2$.

Recall that by $(\mathbf{T}'^m)^{\flat}$ we denote the $d \times d^{m-1}$ flattening of $\mathbf{T}'^m$, see Section 3. To simplify notation write $\mathbf{M}^m = (\mathbf{T}'^m)^{\flat} \in R^{d\times d^{m-1}}$. We have that V_m is the span of the left singular vectors of $\mathbf{M}^m$ with singular value at least τ and V be the union of all the V_m. We first show that the dimension of V is not very large.

Claim 5.11. *It holds that* $\dim(V) \leq 4k/\tau^2$.

Proof. We have that $\|\mathbf{M}^m\|_F = \|\mathbf{T}'^m\|_F \leq 2$. Therefore, since $\|\mathbf{M}^m\|_F$ is equal to the sum of the squares of the singular values $\sigma_1, \dots, \sigma_d$ of $\mathbf{M}^m$ we obtain that the number of singular values σ_i with $|\sigma_i| \geq \tau$ is

$$\frac{1}{\tau^2}\sum_{i=1}^{d} \mathbb{1}\{\sigma_i^2 \geq \tau^2\}\tau^2 \leq \frac{1}{\tau^2}\sum_{i=1}^{d}\sigma_i^2 = \frac{\|\mathbf{M}^m\|_F^2}{\tau^2} \leq \frac{4}{\tau^2}\,.$$

We can do the same calculation for all $\mathbf{T}^m$ to conclude that the dimension of the union of all subspaces is at most $4k/\tau^2$. □

Next, we show that V contains a non-trivial part of the optimal direction $\mathbf{v}^*$, i.e., it holds that $\|\mathrm{proj}_V(\mathbf{v}^*)\|_2 \geq \tau/(4\sqrt{k})$. We first note that since the univariate polynomial p is mean-zero, it can be written as a linear combination of the univariate Hermite polynomials $h_i(z)$ of degree greater than or equal to 1. Write $p(z) = \sum_{i=1}^{k} c_i h_i(z)$. We note that since $\mathbf{E}_{z\sim\mathcal{N}}[p(z)^2] = 1$, we have that $\sum_{i=1}^{k} c_i^2 = 1$. Moreover, since the order-i Chow tensor is defined as $\mathbf{T}^i = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{H}^i(\mathbf{x})y]$, it holds

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[yp(\mathbf{v}^*\cdot\mathbf{x})] = \sum_{i=1}^{k} c_i \mathbf{T}^i \cdot (\mathbf{v}^*)^{\otimes i} \geq \tau \,.$$

Therefore, by Cauchy-Schwarz, there exists $j \in \{1,\ldots,k\}$ such that $\mathbf{T}^j \cdot (\mathbf{v}^*)^{\otimes j} \geq \tau/\sqrt{k}$. Write $\mathbf{v}^*$ as $\mathbf{v}+\mathbf{u}$ where $\mathbf{v}\in V$ and $\mathbf{u}\in V^{\perp}$. Moreover, denote $\mathbf{r}\in\mathbb{R}^{d^{j-1}}$ to be the flattening of the tensor $\mathbf{v}^{\otimes(j-1)}$. We note that

$$\begin{aligned}\mathbf{T}^j \cdot (\mathbf{v}^*)^{\otimes j} &\leq \|\mathbf{T}^j - \mathbf{T}'^j\|_F + |\mathbf{T}'^j \cdot (\mathbf{v}\otimes(\mathbf{v}^*)^{\otimes j-1})| + |\mathbf{T}'^j \cdot (\mathbf{u}\otimes(\mathbf{v}^*)^{\otimes j-1})| \\ &\leq \|\mathbf{T}^j - \mathbf{T}'^j\|_F + |\mathbf{v}^\top \mathbf{M}^j \mathbf{r}| + |\mathbf{u}^\top \mathbf{M}^j \mathbf{r}| \\ &\leq 2\frac{\tau}{4\sqrt{k}} + \|\mathbf{v}\|_2\|\mathbf{M}^j\|_F \leq \frac{\tau}{2\sqrt{k}} + 2\|\mathbf{v}\|_2 \,,\end{aligned}$$

where we used that $\mathbf{u}$ belongs on the subspace spanned from the left singular vectors of $\mathbf{M}^j$ whose singular values are less than $\tau/(4\sqrt{k})$, and therefore it holds $|\mathbf{u}^\top\mathbf{M}^j\mathbf{v}| \leq \tau/(4\sqrt{k})$. Thus, we have $\|\mathbf{v}\|_2 \geq \frac{\tau}{4\sqrt{k}}$ and therefore, $\|\mathrm{proj}_V(\mathbf{v}^*)\|_2 \geq \tau/(4\sqrt{k})$.

□

The proof of Proposition 5.9 follows from an application of Proposition 5.3 and Lemma 5.10 with the appropriate parameters, which we state below. To conclude the proof, we need the following lemma which bounds the number of samples needed to estimate the order-m Chow parameters. Its proof can be found in Appendix E.

Lemma 5.12. *Fix $m\in\mathbb{Z}_+$ and $\epsilon,\delta\in(0,1)$. Let $\mathcal{D}$ be a distribution in $\mathbb{R}^d\times\{\pm1\}$ with standard normal $\mathbf{x}$-marginals. There is an algorithm that with $N = d^{O(m)}\log(1/\delta)/\epsilon^2$ samples and $\mathrm{poly}(d,N)$ runtime, outputs an approximation $\mathbf{T}'^m$ of the order-m Chow-parameter tensor $\mathbf{T}^m$ of $\mathcal{D}$ such that with probability $1-\delta$, it holds $\|\mathbf{T}'^m - \mathbf{T}^m\|_F \leq \epsilon$.*

Recall that from our assumptions $|b| \leq \sqrt{\log(1/\zeta)}$. Moreover, from Proposition 5.3, we have that for $k = O(\log(1/\zeta))$, there exists a degree-k polynomial p, such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[yp(\mathbf{v}^*\cdot\mathbf{x})] \geq \mathrm{poly}(\zeta)$. Thus, in order to apply Lemma 5.10, we need to estimate the first order-k Chow tensors. Hence, the sample complexity is $N = d^{O(\log(1/\zeta))}\mathrm{poly}(1/\zeta)\log(1/\delta)$ and the runtime is $\mathrm{poly}(N,d)$. This completes the proof of Proposition 5.9.

□

Combining Lemma 5.6 and Proposition 5.9, we can prove Proposition 5.2.

Proof of Proposition 5.2. From the assumption that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\neq y] \geq \mathrm{OPT}+\epsilon$, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\neq f(\mathbf{x})\}] \geq \epsilon$. To prove this, note that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\neq y] = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\neq f(\mathbf{x})\}] + \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})]$ and $\mathrm{OPT} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})]$, by combining with $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\neq y] \geq \mathrm{OPT}+\epsilon$, it follows that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\neq f(\mathbf{x})\}] \geq \epsilon \,.$$

Let $\widehat{\mathcal{D}}_N$ be the empirical distribution of $\mathcal{D}$ using $N = d^{\Theta(\log(1/\epsilon))}\log(1/\delta)$ samples. Let $\rho = C\epsilon^3$, for some small enough constant $C > 0$, and from Lemma 5.6, we get that there exists a set $B = \{t_1 \leq \mathbf{v}\cdot\mathbf{x} \leq t_2\}$, with $|t_1 - t_2| = \rho$, for which we can apply the algorithm of Proposition 5.9 to the distribution $\mathcal{D}_B^{\perp\mathbf{w}}$ ($\mathcal{D}_B^{\perp\mathbf{w}}$ is the distribution $\widehat{\mathcal{D}}_N$ conditioned on B and projected into $\mathbf{w}^\perp$), and get a vector space V with the following properties: (i) $V \subseteq \mathbf{w}^\perp$, (ii) the dimension of V is $\mathrm{poly}(1/\epsilon)$, and (iii) $\|\mathrm{proj}_V(\mathbf{w}^*)\|_2 = \mathrm{poly}(\epsilon\rho) = \mathrm{poly}(\epsilon)$ with probability $1-\delta$, if $N \geq d^{O(\log(1/\epsilon))}\log(1/\delta)$. Moreover, the algorithm runs in $\mathrm{poly}(N, d)$ runtime.

The problem, we need to overcome, is that we do not know the set B, so we are going to apply the algorithm of Proposition 5.9 to any possible set and take the union of the outputs. Let $l = \Theta(\log(1/\epsilon)/\rho)$ and let $t_i = i\rho$ and $t_{-i} = -i\rho$, for $0 \leq i \leq l$. Furthermore, we define $B_i = \{t_i \leq \mathbf{v}\cdot\mathbf{x} \leq t_{i+1}\}$ and $B_{-i} = \{-t_{i+1} \leq \mathbf{v}\cdot\mathbf{x} \leq -t_i\}$ (similarly to Claim 5.8). We then apply the algorithm of Proposition 5.9 to each of the distributions $\mathcal{D}_{B_i}^{\perp\mathbf{w}}$ and get vector spaces V_i. Moreover, the algorithm of Proposition 5.9 returns vector spaces V_i such that $V_i \subseteq \mathbf{w}^\perp$ and $\dim(V_i) = \mathrm{poly}(1/\epsilon)$. Finally, from Lemma 5.6, we know that there exists an index $j \in \{-l, \ldots, l\}$ such that applying the algorithm Proposition 5.9 on the distribution $\mathcal{D}_{B_j}^{\perp\mathbf{w}}$ gives a vector space V_j, with the additional property that $\|\mathrm{proj}_{V_j}((\mathbf{w}^*)^{\perp\mathbf{w}})\|_2 = \mathrm{poly}(\epsilon)$, with probability $1-\delta$. Thus, we let V to be the union of all subspaces V_i, i.e., $V = \bigcup_{i=-l}^{l} V_i$. It holds that $\dim(V) \leq \sum_{i=1}^{l}\dim(V_i) = \mathrm{poly}(1/\epsilon)$. Moreover, for any $\mathbf{v} \in V$, it holds $\mathbf{v}\cdot\mathbf{w} = 0$, this is because $V_i \subseteq \mathbf{w}^\perp$ and hence $V \subseteq \mathbf{w}^\perp$. Finally, we have that $\|\mathrm{proj}_{V_j}(\mathbf{w}^*)\|_2 = \mathrm{poly}(\epsilon)$ and because $V_j \subseteq V$ it holds that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp\mathbf{w}})\|_2 \geq \mathrm{poly}(\epsilon)$ with probability $1-\delta$. □

We now show that by finding a vector $\mathbf{u}$ that correlates well with $(\mathbf{v}^*)^{\perp\mathbf{v}}$, we can update our current guess vector $\mathbf{v}$ and get one with increased correlation with $\mathbf{v}^*$. Its proof can be found on Appendix E.

Lemma 5.13 (Correlation Improvement). *Fix unit vectors $\mathbf{v}^*, \mathbf{v} \in \mathbb{R}^d$. Let $\mathbf{u} \in \mathbb{R}^d$ such that $\mathbf{u}\cdot\mathbf{v}^* \geq c$, $\mathbf{u}\cdot\mathbf{v} = 0$, and $\|\mathbf{u}\|_2 \leq 1$, with $c > 0$. Then, for $\mathbf{v}' = \frac{\mathbf{v}+\lambda\mathbf{u}}{\|\mathbf{v}+\lambda\mathbf{u}\|_2}$, with $\lambda = c/2$, we have that $\mathbf{v}'\cdot\mathbf{v}^* \geq \mathbf{v}\cdot\mathbf{v}^* + \lambda^2/2$.*

We need the following standard fact that bounds from below the correlation of any vector with a random one.

Fact 5.14 (see, e.g., Remark 3.2.5 of [Ver18]). *Let $\mathbf{v}$ be a unit vector in $\mathbb{R}^d$. For a random unit vector $\mathbf{u} \in \mathbb{R}^d$, with at least $1/3$ probability, it holds $\mathbf{v}\cdot\mathbf{u} \gtrsim 1/\sqrt{d}$.*

We are now ready to prove the Theorem 5.1.

Proof of Theorem 5.1. First, let $\widehat{\mathcal{D}}_N$ be the empirical distribution of $\mathcal{D}$ using $N \in \mathbb{Z}_+$ samples. Let $\mathbf{w}^{(i)}$ be the current guess. If $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(i)}\cdot\mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$ and $\theta(\mathbf{w}^{(i)}, \mathbf{w}^* \leq \pi - \epsilon)$, then from Proposition 5.2 with $N_1 = d^{O(\log(1/\epsilon))}\log(1/\delta_1)$ samples from $\widehat{\mathcal{D}}$ and $\mathrm{poly}(N_1, d)$ time, we compute a subspace V such that $\|\mathrm{proj}_V(\mathbf{w}^*)\|_2 \geq \mathrm{poly}(\epsilon)$ and from Fact 5.14, we get a random a unit vector $\mathbf{v} \in V$ such that $\mathbf{v}\cdot\mathbf{w}^* = \mathrm{poly}(\epsilon)$ and $\mathbf{v}\cdot\mathbf{w}^{(i)} = 0$, with probability $(1-\delta_1)/3$. We call this event $\mathcal{E}_i$. By conditioning on the event $\mathcal{E}_i$; from Lemma 5.13, after we update our current hypothesis vector $\mathbf{w}^{(i)}$ with $\mathbf{v}$, we get a unit vector $\mathbf{w}^{(i+1)}$ such that $\mathbf{w}^{(i+1)}\cdot\mathbf{w}^* \geq \mathbf{w}^{(i+1)}\cdot\mathbf{w}^* + \mathrm{poly}(\epsilon)$.

After running the update step k times and conditioning on the events $\mathcal{E}_1, \ldots, \mathcal{E}_k$, then $\mathbf{w}^{(k)}\cdot\mathbf{w}^* \geq \mathbf{w}^{(1)}\cdot\mathbf{w}^* + k\,\mathrm{poly}(\epsilon)$; therefore, for $k = \mathrm{poly}(1/\epsilon)$, we get that the vector $\mathbf{w}^{(k)}$ that is competitive with the optimal hypothesis, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(k)}\cdot\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$ or $\theta(\mathbf{w}^{(i)}, \mathbf{w}^* \in (\pi - \epsilon, \pi)$ which means that $-\mathbf{w}^{(k)}$ is competitive with the optimal hypothesis, see Fact 3.2. The probability that all the events $\mathcal{E}_1, \ldots, \mathcal{E}_k$ hold simultaneously is at least $(1 - k\delta_1) + (1/3)^k$, and thus by choosing

Input:

1. $\epsilon, \delta > 0$.
2. Sample access to $\mathcal{D}$ with standard normal $\mathbf{x}$-marginal which satisfies the general Massart noise condition with respect to the hypothesis $f(\mathbf{x})$.

Output: A hypothesis $h(\mathbf{x})$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT}+\epsilon$, with probability $1-\delta$.

Define: $L = \{\}$, $N = d^{C\log(1/\epsilon)}\log(1/\delta)$, $\lambda = \mathrm{poly}(\epsilon)/C$, $T = 2^{C\mathrm{poly}(1/\epsilon)}$, for $C > 0$ a sufficiently large constant.

1. Initialize $\mathbf{w} = \mathbf{e}_1$.
2. Draw N samples and compute to be the empirical distribution $\widehat{D}$.
3. Repeat T times
   (a) Using $\mathbf{w}$ on Algorithm 2, generate vector space V.
   (b) Pick a random unit vector $\mathbf{v} \in V$.
   (c) Update current hypothesis $\mathbf{w}$ by $\mathbf{w} \leftarrow \frac{\mathbf{w}+\lambda\mathbf{v}}{\|\mathbf{w}+\lambda\mathbf{v}\|_2}$.
   (d) Update list of vectors $L \leftarrow L \cup \{\mathbf{w}\}$.
4. **Return:** $h(\mathbf{x}) = \mathrm{sign}(\hat{\mathbf{w}} \cdot \mathbf{x})$, where $\hat{\mathbf{w}} \in S$ and minimizes the error with respect y, i.e.,

$$\hat{\mathbf{w}} \leftarrow \operatorname*{argmin}_{\mathbf{w}\in L} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y] .$$

Algorithm 3: The Biased Random Walk for Learning Halfspaces with General Massart Noise, see Theorem 5.1.

$\delta_1 \leq 1/(3k)$, the probability of success is at least $\delta_2 = (1/3)^k$. By running the algorithm above $M = \log(1/\delta)/\delta_2$ times, we get a list of $2M$ vectors, that list contains all the $\mathbf{w}^{(i)}$ that generated in every step of the algorithm and the $-\mathbf{w}^{(i)}$. By applying Hoeffding's inequality, we get that the list L of $2M$ vectors contains a unit vector $\mathbf{w}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(k)}\cdot\mathbf{x}) \neq y] \leq \mathrm{OPT}+\epsilon$, with probability $1-\delta/2$. Finally, to evaluate all the vectors from the list, we need a few samples, from the distribution $\mathcal{D}$ to obtain the best among them, i.e., the one that minimizes the zero-one loss.

The size of the list of candidates is at most $M \leq 2^{\mathrm{poly}(1/\epsilon)}\log(1/\delta)$. Therefore, from Hoeffding's inequality, it follows that $O(\mathrm{poly}(1/\epsilon)\log(1/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most ϵ with probability at least $1-\delta/2$. Thus, with $N = d^{\log(1/\epsilon)}\log(1/\delta)$ samples and $\mathrm{poly}(N, d, 2^{\mathrm{poly}(1/\epsilon)})$ runtime, we get a hypothesis $\hat{\mathbf{w}}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\hat{\mathbf{w}}\cdot\mathbf{x}) \neq y] \leq \mathrm{OPT}+\epsilon$ with probability $1-\delta$. This completes the proof. □

6 Statistical Query Lower Bounds for Learning Massart Halfspaces

6.1 Background on SQ Lower Bounds

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

Definition 6.1. *Let $\mathcal{D}$ be a distribution on labeled examples supported on $X \times \{-1, 1\}$, for some domain X. A statistical query is a function $q : X \times \{-1, 1\} \to [-1, 1]$. We define $\mathrm{STAT}(\tau)$ to be the oracle that given any such query $q(\cdot, \cdot)$ outputs a value v such that $|v - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[q(\mathbf{x}, y)]| \le \tau$, where $\tau > 0$ is the tolerance parameter of the query.*

The SQ model was introduced by Kearns [Kea98] in the context of supervised learning as a natural restriction of the PAC model [Val84]. Subsequently, the SQ model has been extensively studied in a plethora of contexts (see, e.g., [Fel16] and references therein). The class of SQ algorithms is rather broad and captures a range of known supervised learning algorithms. More broadly, several known algorithmic techniques in machine learning are known to be implementable using SQs. These include spectral techniques, moment and tensor methods, local search (e.g., Expectation Maximization), and many others (see, e.g., [FGR+17, FGV17]). Recent work [BBH+21] has shown a near-equivalence between the SQ model and low-degree polynomial tests

Statistical Query Dimension To bound the complexity of SQ learning a concept class $\mathcal{C}$, we use the SQ framework for problems over distributions [FGR+17].

Definition 6.2 (Decision Problem over Distributions)**.** *Let $\mathcal{D}$ be a fixed distribution and $\mathfrak{D}$ be a family of distributions. We denote by $\mathcal{B}(\mathfrak{D}, \mathcal{D})$ the decision (or hypothesis testing) problem in which the input distribution $\mathcal{D}'$ is promised to satisfy either (a) $\mathcal{D}' = \mathcal{D}$ or (b) $\mathcal{D}' \in \mathfrak{D}$, and the goal is to distinguish between the two cases.*

Definition 6.3 (Pairwise Correlation)**.** *The pairwise correlation of two distributions with probability density functions $\mathcal{D}_1, \mathcal{D}_2 : \mathbb{R}^n \to \mathbb{R}_+$ with respect to a distribution with density $\mathcal{D} : \mathbb{R}^n \to \mathbb{R}_+$, where the support of $\mathcal{D}$ contains the supports of $\mathcal{D}_1$ and $\mathcal{D}_2$, is defined as $\chi_{\mathcal{D}}(\mathcal{D}_1, \mathcal{D}_2) := \int_{\mathbb{R}^n} \mathcal{D}_1(\mathbf{x})\mathcal{D}_2(\mathbf{x})/D(\mathbf{x})\,\mathrm{d}\mathbf{x} - 1$.*

Definition 6.4. *We say that a set of s distributions $\mathfrak{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_s\}$ over $\mathbb{R}^n$ is (γ, β)-correlated relative to a distribution $\mathcal{D}$ if $|\chi_{\mathcal{D}}(\mathcal{D}_i, \mathcal{D}_j)| \le \gamma$ for all $i \neq j$, and $|\chi_{\mathcal{D}}(\mathcal{D}_i, \mathcal{D}_i)| \le \beta$ for all i.*

Definition 6.5 (Statistical Query Dimension)**.** *For $\beta, \gamma > 0$ and a decision problem $\mathcal{B}(\mathfrak{D}, \mathcal{D})$, where $\mathcal{D}$ is a fixed distribution and $\mathfrak{D}$ is a family of distributions, let s be the maximum integer such that there exists a finite set of distributions $\mathfrak{D}_{\mathcal{D}} \subseteq \mathfrak{D}$ such that $\mathfrak{D}_{\mathcal{D}}$ is (γ, β)-correlated relative to $\mathcal{D}$ and $|\mathfrak{D}_{\mathcal{D}}| \ge s$. The* Statistical Query dimension *with pairwise correlations (γ, β) of $\mathcal{B}$ is defined to be s, and denoted by $\mathrm{SD}(\mathcal{B}, \gamma, \beta)$.*

Lemma 6.6 (Corollary 3.12 of [FGR+17])**.** *Let $\mathcal{B}(\mathfrak{D}, \mathcal{D})$ be a decision problem, where $\mathcal{D}$ is the reference distribution and $\mathfrak{D}$ is a class of distributions. For $\gamma, \beta > 0$, let $s = \mathrm{SD}(\mathcal{B}, \gamma, \beta)$. For any $\gamma' > 0$, any SQ algorithm for $\mathcal{B}$ requires queries of tolerance at most $\sqrt{\gamma + \gamma'}$ or makes at least $s\gamma'/(\beta - \gamma)$ queries.*

We next introduce some definitions related to the hidden-direction proof machinery that we use. We start with the following definition:

Definition 6.7 (High-Dimensional Hidden Direction Distribution)**.** *For a distribution A on the real line with probability density function $A(z)$ and a unit vector $\mathbf{v} \in \mathbb{R}^d$, consider the distribution over $\mathbb{R}^d$ with probability density function*

$$\mathbf{P}^A_{\mathbf{v}}(\mathbf{x}) = A(\mathbf{v} \cdot \mathbf{x}) \exp\left(-\|\mathbf{x} - (\mathbf{v} \cdot \mathbf{x})\mathbf{v}\|_2^2/2\right)/(2\pi)^{(d-1)/2} .$$

That is, $\mathbf{P}_{\mathbf{v}}$ is the product distribution whose orthogonal projection onto the direction of $\mathbf{v}$ is A, and onto the subspace perpendicular to $\mathbf{v}$ is the standard $(d-1)$-dimensional normal distribution.

Since we will be using mixtures of two hidden direction distributions we introduce the following notation.

Definition 6.8 (Mixture of Hidden Direction Distributions)**.** *Let A and B be distributions on $\mathbb{R}$. For $d \in \mathbb{Z}_+$ and a unit vector $\mathbf{v} \in \mathbb{R}^d$, define the distribution $\mathbf{P}^{A,B,p}_{\mathbf{v}}$ on $\mathbb{R}^d \times \{\pm 1\}$ that returns a sample from $(\mathbf{P}^A_{\mathbf{v}}, 1)$ with probability p and a sample from $(\mathbf{P}^B_{\mathbf{v}}, -1)$ with probability $1-p$.*

We will also use the following fact showing that there exists an exponentially large set of d-dimensional unit vectors all of which have small correlation. Moreover, one can show that a set of random vectors on the unit sphere will satisfy this property with nontrivial probability.

Fact 6.9 (Lemma 3.7 of [DKS17])**.** *For any constant $0 < c < 1/2$ there exists a set S of $2^{\Omega(d^c)}$ unit vectors in $\mathbb{R}^d$ such that any pair $\mathbf{u}, \mathbf{v} \in S$, with $\mathbf{u} \neq \mathbf{v}$, satisfies $|\mathbf{u} \cdot \mathbf{v}| \lesssim d^{c-1/2}$.*

The following fact shows that given a one-dimensional marginal A that matches m moments with the standard normal, the correlation between two hidden direction distributions with directions $\mathbf{v}$ and $\mathbf{u}$ is bounded above roughly by the $(m+1)$-th power of the correlation of their corresponding directions. Therefore, using Fact 6.9, we obtain that there exists an exponentially large (in the dimension d) set of distributions with pairwise correlation roughly d^{-m}.

Fact 6.10 (Lemma 3.4 from [DKS17])**.** *Let $m \in \mathbb{Z}_+$. If the univariate distribution A over $\mathbb{R}$ agrees with the first m moments of $\mathcal{N}(0,1)$, then for all $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$, we have that*

$$|\chi_{\mathcal{N}(\mathbf{0},\mathbf{I})}(\mathbf{P}^A_{\mathbf{v}}, \mathbf{P}^A_{\mathbf{u}})| \leq |\mathbf{v} \cdot \mathbf{u}|^{m+1} \chi^2(A, \mathcal{N}(0,1)) .$$

6.2 SQ Lower Bound for Learning Halfspaces with Constant-Bounded Massart Noise

The problem of learning homogeneous halfspaces with Massart Noise under the Gaussian distribution is by now well understood. All previous algorithms fit in the SQ framework and show that the SQ complexity of learning halfspaces with Massart noise is polynomial in the dimension d, the accuracy ϵ, and the noise rate η. In this section, we show that if the optimal halfspace f is γ-biased, i.e., $\mathbf{Pr}[f(\mathbf{x}) = +1] = \gamma$, the SQ complexity of learning f is quasi-polynomial in the bias γ, that is $d^{\Omega(\log(1/\gamma))}$ SQ queries are required. We prove the following theorem.

Theorem 6.11. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal that satisfies the η-Massart noise condition with parameter $\eta \in (0, 1/2)$ with respect to some unknown $(1-\gamma)$-biased optimal halfspace $f(\mathbf{x})$, for some $\gamma > 0$ less than a sufficiently small constant. Any SQ algorithm that, for any such distribution $\mathcal{D}$, learns a hypothesis $h : \mathbb{R}^d \mapsto \{\pm 1\}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$ for $\epsilon \leq \gamma(1-2\eta)$, either requires queries with tolerance at most $d^{-\Omega(\log(\eta/\gamma))}$ or makes at least $2^{d^{\Omega(1)}} d^{-\log(\eta/\gamma)}$ queries.*

Remark 6.12. We remark that when the bias γ is less than ϵ one can output the constant guess -1 and obtain error at most ϵ. Therefore, reasonable values for γ are $\Omega(\epsilon)$. In the extreme case where $\gamma = \Theta(\epsilon)$, Theorem 6.11 implies a quasi-polynomial SQ lower bound in the accuracy parameter ϵ. However, our result is fine-grained with respect to γ: we show that this quasi-polynomial dependency on the bias is required across the whole regime of γ.

The main structural result of the proof of Theorem 6.11 is the proposition that follows. We prove that we can construct a distribution $\mathcal{D}$ over labeled pairs $(z, y) \in \mathbb{R} \times \{\pm 1\}$ that satisfies the η-Massart noise assumption with respect to some (biased) halfspace f and whose low-order moments match the moments of the product distribution of the marginals $\mathcal{D}_z$ and $\mathcal{D}_y$. In other words, we show that we can construct and instance $\mathcal{D}$ whose z-marginal is a one dimensional Gaussian distribution and y is uncorrelated with z^k for any k less than a sufficiently small multiple of $\log(\eta/\gamma)$.

Proposition 6.13. *Fix $\eta \in (0, 1/2)$ and $\gamma > 0$ less than a sufficiently small constant. There exists a distribution $\mathcal{D}$ on $(z, y) \in \mathbb{R} \times \{\pm 1\}$ whose z-marginal is the standard normal distribution with the following properties.*

- *$\mathcal{D}$ satisfies the Massart noise condition with parameter η with respect to a halfspace $f(z)$ with $\mathbf{Pr}_{z \sim \mathcal{D}_z}[f(z) = 1] = \gamma$.*
- *For any integer $k \leq C \log(\eta/\gamma)$ where $C > 0$ is a sufficiently small constant, it holds $\mathbf{E}_{(z,y)\sim\mathcal{D}}[yz^k] = \mathbf{E}_{y\sim\mathcal{D}_y}[y]\,\mathbf{E}_{z\sim\mathcal{D}_z}[z^k]$.*

Proof. Our goal is to construct a distribution $\mathcal{D}$ on $\mathbb{R} \times \{\pm 1\}$ with Gaussian z-marginal satisfying the η-Massart noise condition. Recall that, for any such distribution, we denote $\beta = 1 - 2\eta$ and $\beta(z) = 1 - 2\eta(z)$. It holds

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}}[y|x = z] = -f(z)\eta(z) + f(z)(1 - \eta(z)) = f(z)(1 - 2\eta(z)) = f(z)\beta(z)\,.$$

Therefore, we need to prove that there exists a "noise" function $\beta(z) : \mathbb{R} \mapsto [\beta, 1]$ such that for every zero mean polynomial $p(z)$ of degree at most k, it holds $\mathbf{E}_{(z,y)\sim\mathcal{D}}[p(z)y] = \mathbf{E}_{z\sim\mathcal{N}}[p(z)\beta(z)f(z)] = 0$. Recall that by $\mathcal{P}_{k,d}$ we denote the space of polynomials of degree at most k over $\mathbb{R}^d$. In what follows, we will be using polynomials on the subspace of univariate mean-zero polynomials (with respect to the Gaussian measure) which we denote by

$$\mathcal{P}_k^0 \coloneqq \{p \in \mathcal{P}_{k,1} : \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[p(z)] = 0\}\,.$$

Using duality, we first show that such a noise function $\beta(z)$ exists when there exists no mean-zero polynomial p such that the expectation of $(f(z)p(z))^+$ is $1/\beta$ times larger than the expectation of $(f(z)p(z))^-$, where $z^+ \coloneqq \max(0, z)$ and $z^- \coloneqq |\min(0, z)|$. We prove the following lemma.

Lemma 6.14 (Moment-Matching Duality)**.** *Let $f : \mathbb{R} \mapsto \{\pm 1\}$ be any one-dimensional Boolean function. Assume that for any polynomial $p \in \mathcal{P}_k^0$, with $p \neq 0$, it holds that*

$$\beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^+] < \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^-]\,.$$

Then, there exists a function $\beta(z) : \mathbb{R} \mapsto \mathbb{R}$ such that $\mathbf{Pr}_{z\sim\mathcal{N}}[\beta \leq \beta(z) \leq 1] = 1$ and for every polynomial $p \in \mathcal{P}_k^0$ it holds that $\mathbf{E}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)] = 0$.

Remark 6.15. Even though we used the assumption that f is a one-dimensional function, this statement is true for any m-dimensional function as long as f is biased enough. Therefore, using similar techniques as in [DKPZ21], we can embed this m-dimensional subspace to d dimensions and get lower bounds for learning more general classes. In fact, for $m = 2$, one can show an SQ lower bound for intersections of two homogeneous halfspaces under constant-bounded Massart noise.

Proof. Treating the function $\beta(z)$ as an (infinite dimensional) variable we can formulate the following feasibility linear program.

$$\begin{aligned} \text{Find} \quad & \beta(z) \in L^{\infty}(\mathbb{R}) \\ \text{such that} \quad & \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)] = 0 \qquad \forall p \in \mathcal{P}_k^0 \\ & \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}}[\beta \le \beta(z) \le 1] = 1 \end{aligned} \tag{38}$$

We claim that the LP (38) is equivalent to the following LP:

$$\begin{aligned} \text{Find} \quad & \beta(z) \in L^{\infty}(\mathbb{R}) \\ \text{such that} \quad & \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)] = 0 && \forall p \in \mathcal{P}_k^0 \\ & \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\beta(z)h(z)] \le \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[h(z)] && \forall h \in L^1_+(\mathbb{R}) \\ \beta & \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[T(z)] \le \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\beta(z)T(z)] && \forall T \in L^1_+(\mathbb{R}) \end{aligned} \tag{39}$$

Recall that $L^1(\mathbb{R})$ are all the functions that have bounded L_1-norm, and we denote $L^1_+(\mathbb{R})$ to be the positive functions in $L^1(\mathbb{R})$.

Claim 6.16. *The LP (38) is equivalent to the LP (39).*

Proof. We now show the equivalence between the two formulations. We claim that the second constraint of LP (38) is equivalent with the second and the third constraints of LP (39). This follows by introducing the "dual variables" $h : \mathbb{R} \to \mathbb{R}$ and $T : \mathbb{R} \to \mathbb{R}$. First, we so that any valid solution $b(z)$ of LP (38) is also a valid solution for the LP (39). For any valid solution $b(z)$ of LP (38), it should hold that $b(z) \le 1$ and $b(z) \ge \beta$ with probability 1, thus for $h(z) \in L^1_+(\mathbb{R})$ it should hold that $b(z)h(z) \le h(z)$ and by taking expectation in both sides, we get the third inequality, similarly we get the fourth inequality from $b(z) \ge \beta$. To prove that any valid solution $b(z)$ of LP (39) is also a valid solution for the LP (38), first we assume, in order to reach a contradiction, that there is a set A with non-zero probability that $b(z) > 1$. Then, by taking $h(z) = \mathbb{1}\{z \in A\}$, the third inequality becomes

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[h(z)] < \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\beta(z)h(z)] \le \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[h(z)] \,,$$

which is a contradiction. For the case, we assume, to reach a contradiction, that there is a set B with non-zero probability that $b(z) < \beta$. Similarly, we have $T(z) = \mathbb{1}\{z \in B\}$ and get that $\beta \, \mathbf{E}_{z\sim\mathcal{N}}[T(z)] \le \mathbf{E}_{z\sim\mathcal{N}}[\beta(z)T(z)] < \beta \, \mathbf{E}_{z\sim\mathcal{N}}[T(z)]$, which is again a contradiction. ☐

At this point, we would like to use "LP duality" to argue that LP (39) is feasible if and only if its "dual LP" is infeasible. While such a statement turns out to be true, it requires some care to prove since we are dealing with infinite LPs (both in number of variables and constraints). More formally, we have that LP (39) is feasible if and only if there is no conical combination that yields the contradicting inequality $1 \leq 0$. We define the "dual LP" to be the following:

$$\begin{aligned} \text{Find} \quad & h \in L_+^1(\mathbb{R}), T \in L_+^1(\mathbb{R}), p \in \mathcal{P}_k^0 \\ \text{such that} \quad & f(z)p(z) + h(z) - T(z) = 0 \qquad \forall z \in \mathbb{R} \\ & \beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[T(z)] - \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[h(z)] > 0 \end{aligned} \tag{40}$$

The following lemma states that the sufficiently conditions so that LP (39) is feasible. Its proof can be found on Appendix G.

Lemma 6.17. *If there is no polynomial $p \in \mathcal{P}_k^0$ such that $\beta \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^+] > \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^-]$ then, the LP (39) is feasible if only if LP (40) is infeasible.*

Note that if LP (40) is feasible for some functions (h, T, p), then it should hold that $h(z) = (f(z)p(z))^-$ and $T(z) = (f(z)p(z))^+$. Therefore, we claim that the LP (40) is equivalent to the following LP.

$$\begin{aligned} \text{Find} \quad & p \in \mathcal{P}_k^0 \\ \text{such that} \quad & \beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^+] > \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^-] \end{aligned} \tag{41}$$

Observe that if LP (41) is infeasible then from Lemma 6.17, we get that LP (39) is feasible, therefore the proof of Lemma 6.14 follows. ☐

Lemma 6.18. *Let $f : \mathbb{R} \mapsto \{\pm 1\}$ be any one-dimensional Boolean function, $\beta \in (0, 1)$ and $k \in \mathbb{Z}_+$. If $\mathbf{Pr}_{z\sim\mathcal{N}}[f(z) = 1] \leq 2^{-Ck}(1-\beta)$ for some sufficiently large constant $C > 0$, then for any polynomial $p \in \mathcal{P}_k^0$, it holds $\beta \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^+] \leq \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^-]$.*

Proof. First, notice that, since $\mathbf{E}_{z\sim\mathcal{N}}[p(z)] = 0$, it holds that $\mathbf{E}_{z\sim\mathcal{N}}[p(z)^+] = \mathbf{E}_{z\sim\mathcal{N}}[p(z)^-]$. Moreover, we have that for any $z \in \mathbb{R}$ we either have $f(z) = +1$ or $f(z) = -1$ and therefore, it holds that $f(z)^+ + f(z)^- = 1$. Thus, $\mathbf{E}_{z\sim\mathcal{N}}[p(z)^+ f(z)^+] + \mathbf{E}_{z\sim\mathcal{N}}[p(z)^+ f(z)^-] = \mathbf{E}_{z\sim\mathcal{N}}[p(z)^- f(z)^+] + \mathbf{E}_{z\sim\mathcal{N}}[p(z)^- f(z)^-]$. We have that

$$\begin{aligned} \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^+] &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+ p(z)^+] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^- p(z)^-] , \\ \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^-] &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+ p(z)^-] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^- p(z)^+] . \end{aligned}$$

Therefore, LP (41) implies the following inequality

$$\begin{aligned} 2\beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+ p(z)^+] &> (1+\beta) \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+ p(z)^-] + (1-\beta) \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^- p(z)^+] \\ &\geq (1-\beta) \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^- p(z)^+] , \end{aligned} \tag{42}$$

where we used the fact that $\mathbf{E}_{z\sim\mathcal{N}}[f(z)^+ p(z)^-] \geq 0$. Notice that $\mathbf{E}_{z\sim\mathcal{N}}[|p(z)|] = 2\,\mathbf{E}_{z\sim\mathcal{N}}[p(z)^+] = 2\,\mathbf{E}_{z\sim\mathcal{N}}[p(z)^-]$, because $p(z)$ is a zero mean polynomial. Moreover, using $f(z)^- = 1 - f(z)^+$, it holds $\mathbf{E}_{z\sim\mathcal{N}}[f(z)^- p(z)^+] = \mathbf{E}_{z\sim\mathcal{N}}[|p(z)|]/2 - \mathbf{E}_{z\sim\mathcal{N}}[f(z)^+ p(z)^+]$. Thus, by substituting this in the Equation (42), we get

$$2\frac{1+\beta}{1-\beta} \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+ p(z)^+] > \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[|p(z)|] ,$$

or $4\,\mathbf{E}_{z\sim\mathcal{N}}[f(z)^+p(z)^+]/(1-\beta) > \mathbf{E}_{z\sim\mathcal{N}}[|p(z)|]$, where we used that $\beta<1$. From Cauchy–Schwarz inequality, we have

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+p(z)^+] \le \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+|p(z)|] \le (\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+])^{1/2}(\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[|p(z)|])^{1/2}\,.$$

Using Cauchy–Schwarz, it holds that $(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^2])^{1/2} \le (\mathbf{E}_{z\sim\mathcal{N}}[|p(z)|])^{1/3}(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^4])^{1/6}$ and from Lemma 4.11, we have that $(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^4])^{1/4} \le 3^{k/2}(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^2])^{1/2}$. Putting everything together, we have that $(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^2])^{1/2} \le 2^{Ck}\,\mathbf{E}_{z\sim\mathcal{N}}[|p(z)|]$, for C some large enough positive constant, thus

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[|p(z)|] \le \frac{4}{1-\beta}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+p(z)^+] \le \frac{4}{1-\beta}(\mathbf{Pr}[f(z)=1])^{1/2}2^{Ck}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[|p(z)|]\,,$$

which is a contradiction if $\mathbf{Pr}[f(z)=1] \le 2^{-Ck}(1-\beta)$ for large enough positive constant C. Thus, the LP (41) is infeasible and thus, the LP (38) is feasible. □

Using Lemma 6.14 and Lemma 6.18, we get that if $\gamma = \mathbf{Pr}_{z\sim\mathcal{N}}[f(z)=1] \le 2^{-Ck}(1-\beta)$ for some large enough constant $C>0$, then there exists a function $\beta:\mathbb{R}\mapsto\mathbb{R}$ such that for every polynomial $p\in\mathcal{P}_k^0$ it holds $\mathbf{E}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)]=0$, thus, for any $m\le k$, we have

$$\mathop{\mathbf{E}}_{(z,y)\sim\mathcal{D}}[yz^m] = \mathop{\mathbf{E}}_{(z,y)\sim\mathcal{D}}[y(z^m - \mathop{\mathbf{E}}_{z\sim\mathcal{D}_z}[z^m])] + \mathop{\mathbf{E}}_{(z,y)\sim\mathcal{D}_y}[y]\mathop{\mathbf{E}}_{z\sim\mathcal{D}_z}[z^m] = \mathop{\mathbf{E}}_{(z,y)\sim\mathcal{D}_y}[y]\mathop{\mathbf{E}}_{z\sim\mathcal{D}_z}[z^m]\,.$$

Moreover, using the fact that $\beta = 1-2\eta$, we get that the degree k is less than a sufficiently small multiply of $\log(\eta/\gamma)$, which completes the proof of Proposition 6.13.

□

Proposition 6.19 (SQ Complexity of Hypothesis Testing)**.** *Fix $\eta\in(0,1/2)$ and $\gamma>0$ less than a sufficiently small constant. There exist:*

- *a family of distributions $\mathfrak{D}$ such that every $\mathcal{D}\in\mathfrak{D}$ is a distribution over $(\mathbf{x},y)\in\mathbb{R}^d\times\{\pm1\}$, its $\mathbf{x}$-marginal is the standard normal distribution, and $\mathcal{D}$ satisfies the η-Massart noise condition with respect to some γ-biased halfspace, and*
- *a reference distribution $\mathcal{R}$ over $(\mathbf{x},y)\in\mathbb{R}^d\times\{\pm1\}$, whose $\mathbf{x}$-marginal is the standard normal and y is independent of $\mathbf{x}$,*

such that any SQ algorithm that decides whether the input distribution belongs to $\mathfrak{D}$ or is equal to the reference $\mathcal{R}$ either requires queries with tolerance at most $d^{-\Omega(\log(\eta/\gamma))}/\sqrt{\gamma}$ or makes at least $2^{d^{\Omega(1)}}d^{-\log(\eta/\gamma)}$ queries.

Proposition 6.19 follows from the lemma below showing that we can construct a family of distributions with small pairwise correlation

Lemma 6.20 (Correlated Family of Distributions)**.** *Fix $t\in\mathbb{R}$ such that $|t|$ is larger than some absolute constant. Define $\gamma(t) := \mathbf{Pr}_{z\sim\mathcal{N}}[z\ge|t|]$. There exist:*

- *a set $\mathfrak{D}_t$ of $2^{d^{\Omega(1)}}$ distributions on $\mathbb{R}^d\times\{\pm1\}$, such that every $\mathcal{D}\in\mathfrak{D}_t$ is a distribution over $(\mathbf{x},y)\in\mathbb{R}^d\times\{\pm1\}$, its $\mathbf{x}$-marginal is the standard normal distribution, and $\mathcal{D}$ satisfies the η-Massart condition with $\eta\in(0,\frac{1}{2})$ with respect to a halfspace $f(\mathbf{x})=\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t)$ where $\mathbf{v}$ is a unit vector in $\mathbb{R}^d$. Moreover, all $\mathcal{D}\in\mathfrak{D}_t$ have the same y-marginal,*

- *a reference distribution $\mathcal{R}_t$ in $\mathbb{R}^d \times \{\pm 1\}$, where for $(\mathbf{x}, y) \sim \mathcal{R}_t$ we have that $\mathbf{x}$ is distributed according to the standard normal $\mathcal{N}(\mathbf{0}, \mathbf{I})$, y is independent of $\mathbf{x}$, and the distribution of y is equal with the y-marginal of any $\mathcal{D} \in \mathfrak{D}_t$.*

Moreover, the set $\mathfrak{D}_t$ is $(d^{-\Omega(\log(\eta/\gamma(t)))}/\gamma(t), 4/\gamma(t))$-correlated with respect to the reference distribution $\mathcal{R}_t$.

Proof. From Proposition 6.13, we know that for every t, with $|t|$ larger than a sufficiently large constant, there exists a distribution $\mathcal{D}_t$ on $\mathbb{R} \times \{\pm 1\}$ whose z-marginal is a standard normal, $\mathcal{D}$ satisfies the η-Massart noise distribution with respect to a γ-biased halfspace $f : \mathbb{R} \mapsto \{\pm 1\}$, with $f(z) = \mathrm{sign}(z + t)$, and for every $k = \Theta(\log(\eta/\gamma))$ it holds that $\mathbf{E}_{(z,y)\sim\mathcal{D}_t}[yz^k] = \mathbf{E}_{y\sim\mathcal{D}_{ty}}[y]\,\mathbf{E}_{z\sim\mathcal{N}}[z^k]$. Let $\beta_t(z)$ be the noise function corresponding to $\mathcal{D}_t$ and $\phi(z)$ be the density function of the single dimensional standard normal distribution. We define the following densities on $\mathbb{R}$: $A_t(z) = (1 + \beta(z)f(z))\phi(z)/(1+c)$ and $B_t(z) = (1-\beta(z)f(z))\phi(z)/(1-c)$ where $c = \mathbf{E}_{z\sim\mathcal{N}}[\beta(z)f(z)] = \mathbf{E}_{y\sim\mathcal{D}_{ty}}[y]$. It holds that

$$\mathop{\mathbf{E}}_{z\sim A_t}[z^k] = \frac{\int z^k\phi(z) + z^k\beta(z)f(z)\mathrm{d}z}{1+c} = \frac{\mathbf{E}_{z\sim\mathcal{N}}[z^k] + \mathbf{E}_{(z,y)\sim\mathcal{D}_t}[z^k y]}{1+c}$$
$$= \frac{\mathbf{E}_{z\sim\mathcal{N}}[z^k] + \mathbf{E}_{z\sim\mathcal{D}_{tz}}[z^k]\,\mathbf{E}_{y\sim\mathcal{D}_{ty}}[y]}{1+\mathbf{E}_{y\sim\mathcal{D}_{ty}}[y]} = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[z^k]\,.$$

Similarly, we have that $\mathbf{E}_{z\sim B_t}[z^k] = \mathbf{E}_{z\sim\mathcal{N}}[z^k]$. Therefore, the distributions A_t and B_t match the first $\Theta(\log(\eta/\gamma))$ moments with $\mathcal{N}$.

Moreover, we have that

$$\chi^2(A_t, \mathcal{N}) = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\frac{1+\beta(z)f(z)}{1+c}\right)^2 - 1\right] \le \frac{4}{(1+c)^2}\ ,$$

and

$$\chi^2(B_t, \mathcal{N}) = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\frac{1-\beta(z)f(z)}{1+c}\right)^2 - 1\right] \le \frac{4}{(1-c)^2}\ ,$$

which means that $\max((1+c)\chi^2(A_t,\mathcal{N}), (1-c)\chi^2(B_t,\mathcal{N})) \lesssim 1/\gamma$. Let S be as in Fact 6.9. Choose $p = (1+c)/2$ and consider the mixture distributions $\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}$, for $\mathbf{v} \in S$. We set the hard family of distributions $\mathfrak{D}_t = \{\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p} : \mathbf{v} \in S\}$. First, we show that $\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}$ corresponds to a distribution that satisfies the η-Massart noise condition. We have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}}[y|\mathbf{x}] = \frac{(1+c)}{2}(1 + f(\mathbf{x}\cdot\mathbf{v})\beta(\mathbf{x}\cdot\mathbf{v}))\frac{\phi(\mathbf{x})}{1+c} - \frac{(1-c)}{2}(1 - f(\mathbf{x}\cdot\mathbf{v})\beta(\mathbf{x}\cdot\mathbf{v}))\frac{\phi(\mathbf{x})}{1-c}$$
$$= f(\mathbf{x}\cdot\mathbf{v})\beta(\mathbf{x}\cdot\mathbf{v})\ ,$$

thus, indeed the distribution $\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}$ satisfies the η-Massart noise condition, because from Proposition 6.13, we know that $\beta = 1 - 2\eta \le \beta(z) \le 1$ almost surely for all $z \in \mathbb{R}$.

Let $\mathcal{R}_t$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that if $(\mathbf{x}, y) \sim \mathcal{R}_t$ then $\mathbf{x} \sim \mathcal{N}_d$, y is independent of $\mathbf{x}$, and $y = 1$ with probability p and $y = -1$ otherwise. We need to show that for $\mathbf{u}, \mathbf{v} \in S$ we have that $|\chi_{\mathcal{R}_t}(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}, \mathbf{P}_{\mathbf{u}}^{A_t,B_t,p})|$ is small. Since $\mathcal{R}_t, \mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}$, and $\mathbf{P}_{\mathbf{u}}^{A_t,B_t,p}$ all assign $y = 1$ with probability p, we have that

$$\chi_{\mathcal{R}_t}(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}, \mathbf{P}_{\mathbf{u}}^{A_t,B_t,p}) = p\ \chi_{\mathcal{R}_t|y=1}\big((\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p} \mid y=1), (\mathbf{P}_{\mathbf{u}}^{A_t,B_t,p} \mid y=1)\big) +$$
$$(1-p)\ \chi_{\mathcal{R}_t|y=-1}\big((\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p} \mid y=-1), (\mathbf{P}_{\mathbf{u}}^{A_t,B_t,p} \mid y=-1)\big)$$
$$= p\ \chi_{\mathcal{N}_d}(\mathbf{P}_{\mathbf{v}}^{A_t}, \mathbf{P}_{\mathbf{u}}^{A_t}) + (1-p)\ \chi_{\mathcal{N}_d}(\mathbf{P}_{\mathbf{v}}^{B_t}, \mathbf{P}_{\mathbf{u}}^{B_t}).$$

By Fact 6.10, it follows that

$$\begin{aligned}\chi_{\mathcal{R}_t}(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p},\mathbf{P}_{\mathbf{u}}^{A_t,B_t,p}) &\le d^{-\Omega(\log(\eta/\gamma))}(p\,\chi^2(A_t,N(0,1))+(1-p)\chi^2(B_t,N(0,1)))\\ &= d^{-\Omega(\log(\eta/\gamma))}/\gamma\,.\end{aligned}$$

A similar computation shows that

$$\chi_{\mathcal{R}_t}(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p},\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}) = \chi^2(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p},\mathcal{R}_t) \le p\,\chi^2(A_t,N(0,1))+(1-p)\chi^2(B_t,N(0,1) \le 4/\gamma\,.$$

Thus, the set $\mathfrak{D}_t$ is $(d^{-\Omega(\log(\eta/\gamma))}/\gamma, 4/\gamma)$-correlated with respect the reference distribution $\mathcal{R}_t$. □

Proof of Proposition 6.19. Fix a $\gamma > 0$ less than a sufficiently small constant, and let $t \in \mathbb{R}$ such that $\mathbf{Pr}_{z\sim\mathcal{N}}[z \ge |t|] = \gamma$ Moreover, because $|\mathfrak{D}_t| = 2^{d^{\Omega(1)}}$ and the set $\mathfrak{D}_t$ is $(d^{-\Omega(\log(\eta/\gamma))}/\gamma, 4/\gamma)$-correlated with respect $\mathcal{R}_t$ we have that $\mathrm{SD}(\mathcal{B}, d^{-\Omega(\log(\eta/\gamma))}/\gamma, 4/\gamma) = 2^{d^{\Omega(1)}}$, thus an application of Lemma 6.6 completes the proof. □

Proof of Theorem 6.11. Let $\mathfrak{D}$ and $\mathcal{R}$ be as in Proposition 6.19. Note that from the construction, for any $\mathcal{D} \in \mathfrak{D}$, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{R}}[y = i] = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[y = i] = p_i$, for $i \in \{\pm 1\}$ Let $\mathcal{A}$ an algorithm that outputs a hypothesis h with respect a distribution $\mathcal{D}$ that satisfies the η-Massart Noise condition such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \le \mathrm{OPT} + \epsilon$, where OPT is the error achieved by the best classifier. Moreover, from for any classifier h' it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{R}}[h'(\mathbf{x}) \neq y] \ge \min_{i\in\{\pm1\}} p_i$. Thus, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{R}}[h'(\mathbf{x}) \neq y] - \mathrm{OPT} \ge \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\beta(\mathbf{x})f(\mathbf{x})] \ge (1-2\eta)\gamma$. Thus, algorithm $\mathcal{A}$ for $\epsilon \le \gamma\beta$ would solve the decision problem $\mathcal{B}(\mathfrak{D},\mathcal{R})$ (with one additional query $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})y]$ up to accuracy $\gamma\beta/2$) and from Proposition 6.19 the result follows. □

6.3 SQ Lower Bound for Learning Homogeneous Halfspaces with General Massart Noise

Learning homogeneous halfspaces under η-Massart noise for any constant $\eta < 1/2$ is known to be solvable with polynomially many statistical queries [DKTZ20a]. When $\eta = 1/2$, we show that any SQ algorithm for homogeneous Massart halfspaces requires $d^{\Omega(\log(1/\epsilon))}$ queries. For comparison, we recall that the SQ complexity of learning halfspaces under adversarial noise is $d^{\Omega(1/\epsilon^2)}$ [DKPZ21].

Our formal result is the following theorem.

Theorem 6.21 (1/2-Massart Noise SQ-Lower Bound). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Massart noise condition for $\eta = 1/2$, for homogeneous halfspaces, where the $\mathbf{x}$-marginal is distributed according to the standard normal. Any SQ algorithm that, for any such distribution $\mathcal{D}$, finds a hypothesis h such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \le \mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $d^{-\Omega(\log(1/\epsilon))}$ or makes at least $2^{d^{\Omega(1)}}$ queries.*

To prove the theorem above, we are going to use a "reduction" to the problem of learning general halfspaces with constant-bounded Massart noise.

Proposition 6.22 (SQ Complexity of Hypothesis Testing). *Let $\epsilon > 0$ less than a sufficiently small constant. There exist:*

- *a family of distributions $\mathfrak{D}_B$ such that every $\mathcal{D} \in \mathfrak{D}$ is a distribution over $(\mathbf{x}', y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$, its $\mathbf{x}'$-marginal is the standard normal distribution, and $\mathcal{D}$ satisfies the Massart noise condition for $\eta = 1/2$, with respect to some unbiased halfspace, and*
- *a reference distribution $\mathcal{R}$ over $((\mathbf{x}, z), y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$, whose $(\mathbf{x}, z)$-marginal is the standard normal and y depends only on z,*

such that any SQ algorithm that decides whether the input distribution belongs to $\mathfrak{D}$ or is equal to the reference $\mathcal{R}$ either requires queries with tolerance at most $d^{-\Omega(\log(1/\epsilon))}$ or makes at least $2^{d^{\Omega(1)}} d^{-\log(1/\epsilon)}$ queries.

Proof. Proposition 6.22 follows from the lemma below showing that we can construct a family of distributions with small pairwise correlation

Lemma 6.23 (Correlated Family of Distributions)**.** *Let $\epsilon > 0$ less than a sufficiently small constant. There exist:*

- *a set $\mathfrak{D}_B$ of $2^{d^{\Omega(1)}}$ distributions on $\mathbb{R}^{d+1} \times \{\pm 1\}$, such that every $\mathcal{D} \in \mathfrak{D}_B$ is a distribution over $(\mathbf{x}', y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$, its $\mathbf{x}'$-marginal is the standard normal distribution, and $\mathcal{D}$ satisfies the Massart noise condition for $\eta = 1/2$ with respect to a halfspace $f(\mathbf{x}') = \mathrm{sign}((\mathbf{v}, 1) \cdot \mathbf{x}')$ where $\mathbf{v}$ is a unit vector in $\mathbb{R}^d$. Moreover, all $\mathcal{D} \in \mathfrak{D}_B$ have the same y-marginal.*
- *A reference distribution $\mathcal{R}$ in $\mathbb{R}^{d+1} \times \{\pm 1\}$, where for $((\mathbf{x}, z), y) \sim \mathcal{R}$ we have that $\mathbf{x}'$ is a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, y depends only on z, and for any function $h : \mathbb{R}^{d+1} \mapsto \{\pm 1\}$ and any $\mathcal{D} \in \mathfrak{D}_B$, it holds that $\mathbf{Pr}_{(\mathbf{x}', y) \sim \mathcal{R}}[h(\mathbf{x}') \neq y] - \min_{f \in \mathcal{C}} \mathbf{Pr}_{(\mathbf{x}', y) \sim \mathcal{D}}[f(\mathbf{x}') \neq y] \geq 2\epsilon$.*

Moreover, the set $\mathfrak{D}_B$ is $(d^{-\Omega(\log(1/\epsilon))}, 16\epsilon)$-correlated with respect to the reference distribution $\mathcal{R}$.

Proof. Fix $t_0 > 0$ and $\zeta > 0$ such that $\mathbf{Pr}_{z \sim \mathcal{N}}[z \geq t_0] = \gamma$ for $\gamma = 4\sqrt{\epsilon}$ and $\mathbf{Pr}_{z \sim \mathcal{N}}[t_0 + \zeta > z \geq t_0] = \gamma/2$. We are going to construct a new set of distributions $\mathfrak{D}_B$ as follows:

1. For $\eta = 1/4$ and some threshold t denote $(\mathfrak{D}_t, \mathcal{R}_t)$ be the family of distributions and their corresponding reference distributions from Lemma 6.20. Recall that the distributions in the family $\mathfrak{D}_t$ are indexed by a set of unit vectors S, i.e., $\mathfrak{D}_t = \{\mathcal{D}_{\mathbf{u},t} : \mathbf{u} \in S\}$. Every $\mathcal{D}_{\mathbf{u},t} \in \mathfrak{D}_t$ is a distribution in $\mathbb{R}^d$ that satisfies the η-Massart noise condition with $\eta = 1/4$ and with respect to the halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{u} \cdot \mathbf{x} + t)$.

2. Fix some direction $\mathbf{u} \in S$. We define a new distribution $\mathcal{D}'_{\mathbf{u}}$ on $((\mathbf{x}, z), y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$ where the "extra" coordinate z is drawn from the standard normal distribution $\mathcal{N}(0, 1)$. When z falls inside some thin interval $[t_0, t_0 + \zeta]$ we sample $(\mathbf{x}, y)$ from the 1/4-Massart noise distribution $\mathcal{D}_{\mathbf{u},z}$. This corresponds to the blue/red region in Figure 4. When z falls outside $[t_0, t_0 + \zeta]$, we draw $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and set y to be ± 1 with probability 1/2, independently of $\mathbf{x}$. This "high-noise" area corresponds to the gray area of Figure 4. More formally, we define

$$\mathcal{D}'_{\mathbf{u}}((\mathbf{x}, z), y) = \begin{cases} \mathcal{D}'_{\mathbf{u},z}(\mathbf{x}, y)\phi(z) & \text{if } z \in [t_0, t_0 + \zeta] \\ \frac{1}{2}\phi_{d+1}((\mathbf{x}, z)) & \text{otherwise} . \end{cases}$$

   Let $\mathfrak{D}_B$ be the set of these distributions.

3. We define a reference distribution $\mathcal{R}$ on $((\mathbf{x}, z), y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$ similarly. The "extra" coordinate z is again drawn from the standard normal distribution $\mathcal{N}(0, 1)$. When z falls inside some thin interval $[t_0, t_0 + \zeta]$, i.e., the blue/red region of Figure 4 we sample $(\mathbf{x}, y)$ from the reference distribution $\mathcal{R}_t$. When z falls outside $[t_0, t_0 + \zeta]$, i.e., in the gray area of Figure 4, we draw $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and set y to be ± 1 with probability 1/2, independently of $\mathbf{x}$. We have

$$\mathcal{R}((\mathbf{x}, z), y) = \begin{cases} \mathcal{R}_z(\mathbf{x}, y)\phi(z) & \text{if } z \in [t_0, t_0 + \zeta] \\ \frac{1}{2}\phi_{d+1}((\mathbf{x}, z)) & \text{otherwise} . \end{cases}$$

Our first step is to prove that any "hidden-direction" $(d+1)$-dimensional distribution $\mathcal{D}'_{\mathbf{u}}$ that we create out of the Massart-noise instances $\mathcal{D}_{\mathbf{u},t}$ satisfies the $1/2$-Massart noise condition with a homogeneous optimal halfspace. We show the following claim.

Claim 6.24. *For the distribution $\mathcal{D}'_{\mathbf{u}}$, the optimal hypothesis is $f'(\mathbf{x}') = \mathrm{sign}(\mathbf{x}' \cdot (\mathbf{u}, 1))$, thus $\mathcal{D}'_{\mathbf{u}}$ satisfies the Massart noise condition for $\eta = 1/2$ with respect to a homogeneous $(d+1)$-dimensional halfspace.*

Proof. In what A be the event that $z \in [t_0, t_0 + \zeta]$. Assume in order to reach to a contradiction that $f'(\mathbf{x}')$ is not the optimal hypothesis. Let $h(\mathbf{x}')$ to be the optimal hypothesis for the distribution $\mathcal{D}_{\mathbf{u}'}$. First, observe that for any hypothesis $h'(\mathbf{x}')$ it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[h'(\mathbf{x}') \neq y, A^c] = \mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[A^c]/2 \; .$$

Thus, we need only to consider the error inside the A. For the function $f'(\mathbf{x}')$, we have that

$$\begin{aligned} \mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[f'(\mathbf{x}') \neq y, A] &= \int_{[t_0,t_0+\zeta]} \mathop{\mathbf{Pr}}_{((\mathbf{x},t),y)\sim\mathcal{D}'_{\mathbf{u}}}[f'((\mathbf{x},t)) \neq y | t = z]\phi(z)\mathrm{d}z \\ &= \int_{[t_0,t_0+\zeta]} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathbf{u},z}}[\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + z) \neq y]\phi(z)\mathrm{d}z \; . \end{aligned} \tag{43}$$

Similarly, for any other hypothesis $h'(\mathbf{x}')$, we have that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[h'(\mathbf{x}') \neq y, A] = \int_{[t_0,t_0+\zeta]} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathbf{u},z}}[h'((\mathbf{x},z)) \neq y]\phi(z)\mathrm{d}z \; . \tag{44}$$

Moreover, for the $h'(\mathbf{x}')$ it should hold that $\mathbf{Pr}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[f'(\mathbf{x}') \neq y, A] - \mathbf{Pr}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[h(\mathbf{x}') \neq y, A] > 0$, thus combining with eq. (43) and eq. (44), we have that there exists $z \in [t_0, t_0 + \zeta]$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathbf{u},z}}[\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + z) \neq y] > \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathbf{u},z}}[h'((\mathbf{x}, z)) \neq y]$, which is a contradiction, because for each $\mathcal{D}_{\mathbf{u},t}$ the optimal classifier is $f(\mathbf{x}) = \mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t)$. □

We next have to show that the family of "hidden-direction" distributions $\mathfrak{D}_B$ is pairwise correlated. We have the following claim.

Claim 6.25. *For every $\mathcal{D} \in \mathfrak{D}_B$ it holds $\chi_{\mathcal{R}}(\mathcal{D}, \mathcal{D}) \leq 4$ and for every distinct $\mathcal{D}_1, \mathcal{D}_2 \in \mathfrak{D}_B$ it holds $\chi_{\mathcal{R}}(\mathcal{D}_1, \mathcal{D}_2) \leq d^{-\Omega(\log(1/\gamma))}$. In other words, the family $\mathfrak{D}_B$ is $(d^{-\Omega(\log(1/\gamma))}, 4)$-correlated with respect to the reference distribution $\mathcal{R}$.*

Proof. Note that by the construction of the distributions $\mathcal{D}_1, \mathcal{D}_2$, for a point $\mathbf{x}' = (\mathbf{x}, z) \in \mathbb{R}^{d+1}$, such that $z \notin [t_0, t_0 + \zeta]$ the conditional distributions $\mathcal{D}_1 \mid z$, $\mathcal{D}_2 \mid z$ and $\mathcal{R} \mid z$ coincide. Therefore, since $\mathcal{D}_1, \mathcal{D}_2, \mathcal{R}$ give the same distribution on z, for $z \notin [t_0, t_0 + \zeta]$ it holds $\chi_{\mathcal{R}|z}((\mathcal{D}_1 \mid z), (\mathcal{D}_2 \mid z)) = 0$, for $z \notin [t_0, t_0 + \zeta]$. Thus, to compute the correlation of $\mathcal{D}_1, \mathcal{D}_2$ it suffices to consider $z \in [t_0, t_0 + \zeta]$, i.e.,

$$\begin{aligned} \chi_{\mathcal{R}}(\mathcal{D}_1, \mathcal{D}_2) &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\chi_{\mathcal{R}_z}((\mathcal{D}_1 \mid z), (\mathcal{D}_2 \mid z))] \\ &= \int_{t_0}^{t_0+\zeta} \chi_{\mathcal{R}_z}((\mathcal{D}_1 \mid z), (\mathcal{D}_2 \mid z))\phi(z)\mathrm{d}z \; , \end{aligned}$$

where $(\mathcal{D}_1 \mid z)$ (resp. $(\mathcal{D}_2 \mid z)$) is the pdf $\mathcal{D}_1$ (resp. $\mathcal{D}_2$) conditioned on z. From Lemma 6.20, it holds that

$$\chi_{\mathcal{R}}(\mathcal{D}_1, \mathcal{D}_2) \leq \frac{d^{-\Omega(\log(1/\gamma))}}{\gamma} \int_{t_0}^{t_0+\zeta} \phi(z)\mathrm{d}z = d^{-\Omega(\log(1/\gamma))} \; ,$$

where we used that $\mathbf{Pr}_{z\sim\mathcal{N}}[t_0 + \zeta > z \geq t_0] = \gamma/2$. The second part follows similarly. □

Moreover, if r is the best hypothesis for $\mathcal{R}$ and f for $\mathcal{D}'_{\mathbf{u}}$, then from Lemma 6.20, we have

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{R}}[r(\mathbf{x}') \neq y] - \mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[f(\mathbf{x}') \neq y] \geq \frac{1}{3}\gamma \int_{t_0}^{t_0+\zeta} \phi(z)\mathrm{d}z \geq \frac{\gamma^2}{6} ,$$

where in the last inequality we used that t_0, ζ are chosen such that $\mathbf{Pr}_{z\sim\mathcal{N}}[t_0 + \zeta > z \geq t_0] = \gamma/2$, by substituting $\gamma = 4\sqrt{\epsilon}$ the result follows. □

To prove Proposition 6.22, from Lemma 6.23, we have that the set $\mathfrak{D}_B$ is $(d^{-\Omega(\log(1/\epsilon))}, 16)$-correlated with respect $\mathcal{R}$. Thus, we have that $\mathrm{SD}(\mathcal{B}, d^{-\Omega(\log(1/\epsilon))}, 16) = 2^{d^{\Omega(1)}}$ and an application of Lemma 6.6 completes the proof. □

Proof of Theorem 6.21. Fix $\epsilon > 0$ less than a sufficiently small constant. Let $\mathfrak{D}_B$ and $\mathcal{R}$ as in Lemma 6.23. Let $\mathcal{A}$ be an algorithm that given $\epsilon' > 0$ and $\mathcal{D}$ that satisfies the 1/2-Massart Noise Condition with respect the halfspace $f'(\mathbf{x}')$ computes a hypothesis h such that

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

We show that $\mathcal{A}$ can solve the Decision Problem $\mathcal{B}(\mathfrak{D}_B, \mathcal{R})$. Let $\mathcal{D} \in \mathfrak{D}_B$ with optimal halfspace $f(\mathbf{x})$, then from Lemma 6.23, we have that for any hypothesis h it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{R}}[h(\mathbf{x}') \neq y] - \mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}}[f(\mathbf{x}') \neq y] \geq 2\epsilon .$$

The algorithm $\mathcal{A}$ for $\epsilon' \leq 2\epsilon$ would solve the decision problem $\mathcal{B}(\mathfrak{D}_B, \mathcal{R})$ (with one additional query $\mathbf{E}_{(\mathbf{x}',y)\sim\mathcal{D}}[h(\mathbf{x}')y]$ up to accuracy ϵ), thus from Proposition 6.22, we get our result. □

References

[ABHU15] P. Awasthi, M. F. Balcan, N. Haghtalab, and R. Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 167–190, 2015.

[ABHZ16] P. Awasthi, M. F. Balcan, N. Haghtalab, and H. Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pages 152–192, 2016.

[AL88] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[BBH+21] M. Brennan, G. Bresler, S. B. Hopkins, J. Li, and T. Schramm. Statistical query algorithms and low-degree tests are almost equivalent. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021.

[BK09] E. Beigman and B. B. Klebanov. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 280–287, 2009.

[BZ17] M.-F. Balcan and H. Zhang. Sample and computationally efficient learning algorithms under s-concave distributions. In *Advances in Neural Information Processing Systems*, pages 4796–4805, 2017.

[CKMY20] S. Chen, F. Koehler, A. Moitra, and M. Yau. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[CM84] R. S. Chhikara and J. McKeon. Linear discriminant analysis with misallocation in training samples. *Journal of the American Statistical Association*, 79(388):899–906, 1984.

[CW01] A. Carbery and J. Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in R^n. *Mathematical Research Letters*, 8(3):233–248, 2001.

[Dan16] 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.

[DFT+15] D. Dachman-Soled, V. Feldman, L.-Y. Tan, A. Wan, and K. Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In Piotr Indyk, editor, *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 498–511. SIAM, 2015.

[DGT19] I. Diakonikolas, T. Gouleakis, and C. Tzamos. Distribution-independent pac learning of halfspaces with massart noise. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4751–4762. Curran Associates, Inc., 2019.

[DK20] I. Diakonikolas and D. M. Kane. Hardness of learning halfspaces with massart noise. *CoRR*, abs/2012.09720, 2020.

[DKK+20] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. A polynomial time algorithm for learning halfspaces with tsybakov noise. *arXiv*, 2020.

[DKK+21a] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021.

[DKK+21b] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Efficiently learning halfspaces with tsybakov noise. *STOC*, 2021.

[DKPZ21] 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.

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, 2017.

[DKTZ20a] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 1486–1513. PMLR, 2020.

[DKTZ20b] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with tsybakov noise. *arXiv*, 2020.

[DKZ20] 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*, 2020.

[Fan68] K. Fan. On infinite systems of linear inequalities. *Journal of Mathematical Analysis and Applications*, 21(3):475 – 478, 1968.

[Fel16] V. Feldman. Statistical query learning. In *Encyclopedia of Algorithms*, pages 2090–2095. 2016.

[FGKP06] 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.

[FGR+17] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM*, 64(2):8:1–8:37, 2017.

[FGV17] V. Feldman, C. Guzman, and S. S. Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In Philip N. Klein, editor, *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017*, pages 1265–1277. SIAM, 2017.

[FS97] 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.

[GGK20] 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.

[GHR92] M. Goldmann, J. Håstad, and A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.

[GR06] 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.

[Han09] S. Hanneke. Theoretical foundations of active learning. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA MACHINE LEARNING DEPT, 2009.

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[Haz16] E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[HY15] S. Hanneke and L. Yang. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16:3487–3602, 2015.

[KB09] B. B. Klebanov and E. Beigman. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503, 2009.

[KB10] B. B. Klebanov and E. Beigman. Some empirical evidence for annotation noise in a benchmarked dataset. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 438–446, 2010.

[Kea98] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[KKMS08] 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.

[KSS94] M. Kearns, R. Schapire, and L. Sellie. Toward Efficient Agnostic Learning. *Machine Learning*, 17(2/3):115–141, 1994.

[MN06] P. Massart and E. Nedelec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 10 2006.

[MP68] M. Minsky and S. Papert. *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, MA, 1968.

[MT94] 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.

[MT99] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 12 1999.

[MV19] O. Mangoubi and N. K. Vishnoi. Nonconvex sampling with the metropolis-adjusted langevin algorithm. In *Conference on Learning Theory, COLT 2019*, pages 2259–2293, 2019.

[Nov62] A. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

[O'D14] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[Ros58] F. Rosenblatt. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.

[RS94] R. Rivest and R. Sloan. A formal model of hierarchical concept learning. *Information and Computation*, 114(1):88–114, 1994.

[Slo88] R. H. Sloan. Types of noise in data for concept learning. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, COLT '88, pages 91–96, San Francisco, CA, USA, 1988. Morgan Kaufmann Publishers Inc.

[Slo92] R. H. Sloan. Corrigendum to types of noise in data for concept learning. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992*, page 450, 1992.

[Slo96] R. H. Sloan. *Pac Learning, Noise, and Geometry*, pages 21–41. Birkhäuser Boston, Boston, MA, 1996.

[SSBD14] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[STC00] J. Shawe-Taylor and N. Cristianini. *An introduction to support vector machines*. Cambridge University Press, 2000.

[Tsy04] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[Val84] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap82] V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics*. Springer-Verlag, Berlin, Heidelberg, 1982.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

[Ver18] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[Yao90] A. Yao. On ACC and threshold circuits. In *Proceedings of the Thirty-First Annual Symposium on Foundations of Computer Science*, pages 619–627, 1990.

[YZ17] S. Yan and C. Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 1056–1066, 2017.

[ZL21] C. Zhang and Y. Li. Improved algorithms for efficient active learning halfspaces with massart and tsybakov noise. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021.

[ZLC17] Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 1980–2022, 2017.

[ZSA20] C. Zhang, J. Shen, and P. Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

A "Massart" Noise with $\eta > 1/2$ is Equivalent to Agnostic Learning

Lemma A.1 (Learning with $\eta > 1/2$)**.** *Let $\mathcal{C}$ be a class of Boolean functions on $\mathbb{R}^d$ and let $\mathcal{F}$ be a class of distributions over $\mathbb{R}^d$. Fix $\eta \in (1/2, 1]$ and let $\mathcal{A}$ be a learning algorithm that given m samples from a distribution $\mathcal{D}'$ with η-semi-random noise, learns a hypothesis $h : \mathbb{R}^d \mapsto \{\pm 1\}$ such that*

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

Then $\mathcal{A}$ can learn $\mathcal{C}$ in the agnostic PAC learning model.

Proof. Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d \times \{\pm 1\}$. Since $\eta > 1/2$, we can create a distribution $\mathcal{D}'$ with η-semi-random noise as follows:

1. Draw $(\mathbf{x}, y) \sim \mathcal{D}$.
2. With probability $2\eta - 1$ return $(\mathbf{x}, y)$ and with probability $2(1-\eta)$ return $(\mathbf{x}, \hat{y})$ where $\hat{y} \in \{\pm 1\}$ is uniformly random and independent of $\mathbf{x}$.

For any $h \in \mathcal{C}$ (in fact for any classifier in general), it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}'}[h(\mathbf{x}) = y|\mathbf{x}] \geq 2(1-\eta)/2 \geq 1-\eta$ and thus we have that $\eta(\mathbf{x}) \leq \eta$. The classifier f of Definition 1.1 can therefore be any classifier in $\mathcal{C}$. For every classifier h it holds

$$\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}'}[h(\mathbf{x}) \neq y] = (2\eta - 1) \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] + 2(\eta - 1)(1/2)\,. \tag{45}$$

We can therefore, use $\mathcal{A}$ on the samples from $\mathcal{D}'$ and obtain a classifier h such that

$$\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}'}[h(\mathbf{x}) = y] \leq \min_{c\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}'}[c(\mathbf{x}) = y] + \epsilon\,.$$

Using Equation (45) we obtain that for the same classifier h it holds

$$\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) = y] \leq \min_{c\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[c(\mathbf{x}) = y] + \epsilon\,,$$

and therefore $\mathcal{A}$ can learn the class $\mathcal{C}$ with respect to any distribution $\mathcal{D}$, i.e., in the agnostic PAC learning setting. □

B Benign Noise is Equivalent to 1/2-Massart Noise

Fact B.1. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$. $\mathcal{D}$ satisfies the 1/2-Massart noise condition with respect to a halfspace $f(\mathbf{x})$ if only if the distribution $\mathcal{D}$ satisfies the Benign noise condition with respect to a halfspace $f(\mathbf{x})$.*

Proof. We prove each direction separately.

Claim B.2. *If $\mathcal{D}$ satisfies the Massart noise condition with $\eta = 1/2$ with respect to a halfspace $f(\mathbf{x})$, then $\mathcal{D}$ satisfies the Benign noise condition with respect to the halfspace $f(\mathbf{x})$.*

Proof. Assume in order to reach in contradiction that the optimal classifier is not f and is h the optimal one. Therefore, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] < \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$. It holds that

$$\begin{aligned}\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] &= \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1-\eta(\mathbf{x}))] + \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) = f(\mathbf{x})\}\eta(\mathbf{x})] \\ &= \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1-2\eta(\mathbf{x}))] + \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})]\,.\end{aligned}$$

Note that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})]$, therefore, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1 - 2\eta(\mathbf{x}))] < 0 ,$$

which means that there is a point (a mass with non-zero measure) with $\eta(\mathbf{x}) > 1/2$, which leads to a contradiction. Therefore, if $\eta(\mathbf{x}) \leq 1/2$ then the optimal classifier is $f(\mathbf{x})$.

□

Next, we prove the other direction.

Claim B.3. *If $\mathcal{D}$ satisfies the Benign noise condition with respect to a halfspace $f(\mathbf{x})$, then $\mathcal{D}$ satisfies the $1/2$-Massart noise condition with respect to the halfspace $f(\mathbf{x})$.*

Proof. It suffices to prove that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\eta(\mathbf{x}) > 1/2\}] = 0$. Because $f(\mathbf{x})$ is the optimal classifier, that means that other classifier $h(\mathbf{x})$ gets more error, therefore it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1 - 2\eta(\mathbf{x}))] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})] .$$

Because $f(\mathbf{x})$ is the optimal classifier, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \geq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y]$, therefore

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1 - 2\eta(\mathbf{x}))] \geq 0 .$$

Assume that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\eta(\mathbf{x}) > 1/2\}] = a > 0$, therefore we can set $h(\mathbf{x}) = f(\mathbf{x})$ when $\eta(\mathbf{x}) \leq 1/2$ and $h(\mathbf{x}) \neq f(\mathbf{x})$ otherwise. Hence, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1 - 2\eta(\mathbf{x}))] = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(1 - 2\eta(\mathbf{x}))\mathbb{1}\{\eta(\mathbf{x}) > 1/2\}] < 0$, which is a contradiction. Therefore, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\eta(\mathbf{x}) > 1/2\}] = 0$, that means that the measure of points that $\eta(\mathbf{x}) > 1/2$ is 0, hence, it satisfies the 1/2-Massart noise condition. □

□

C A Non-Continuous Certificate

Non-Continuous Certificates. When we do not restrict our search to continuous functions in order to find a certificate we can search over functions of the form $T(x) = \mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} - b)\}$ for some $\mathbf{v} \in \mathbb{R}^d, b \in \mathbb{R}$.

Fact C.1. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the η-Massart noise condition with respect to the optimal halfspace $f(\mathbf{x})$. Then, for any linear function $\ell(\mathbf{x})$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \geq \mathrm{OPT} + \epsilon$, let $T(\mathbf{x}) = \mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})\}$, then we have that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y \ T(\mathbf{x})] \leq -\Omega(\epsilon^2)\|\ell(\mathbf{x})\|_2 .$$

Proof. Using the η-Massart noise condition, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y \ T(\mathbf{x})] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\ell(\mathbf{x})f(\mathbf{x})\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})\}] \\ &= -\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\ell(\mathbf{x})|\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})\}] .\end{aligned}$$

From Lemma F.1, we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\ell(\mathbf{x})| \geq \|\ell(\mathbf{x})\|_2\epsilon/C] \geq 1-\epsilon/2$, for some absolute constant $C>0$. From the assumptions, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x}))\neq f(\mathbf{x})\}] \geq \epsilon$, therefore it holds

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{|\ell(\mathbf{x})| \geq \|\ell(\mathbf{x})\|_2\epsilon/c\}\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x}))\neq f(\mathbf{x})\}] \geq \epsilon/2\;,$$

and we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\ell(\mathbf{x})|\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x}))\neq f(\mathbf{x})\}] \\
&\qquad\geq (\|\ell(\mathbf{x})\|_2\epsilon/C)\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{|\ell(\mathbf{x})| \geq \|\ell(\mathbf{x})\|_2\epsilon/c\}\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x}))\neq f(\mathbf{x})\}] \\
&\qquad\gtrsim \|\ell(\mathbf{x})\|_2\epsilon^2/C\;.
\end{aligned}$$

And this completes the proof. □

D Learning General Halfspaces with General Massart Noise

In this section, we provide the algorithm of learning biased halfspaces with general Massart Noise.

Theorem D.1 (Learning General Halfspaces with General Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta=1/2$ with respect to some optimal (possibly biased) halfspace $f\in\mathcal{C}$. Let $\epsilon,\delta\in(0,1]$. There exists an algorithm that draws $N=d^{O(\log(1/\epsilon))}\log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N,d)2^{\mathrm{poly}(1/\epsilon)}$, and computes a halfspace $h\in\mathcal{C}$ such that with probability at least $1-\delta$,*

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

The proposition below is similar to Proposition 5.2.

Proposition D.2. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta=1/2$, with respect to some optimal halfspace $f(\mathbf{x})=\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t^*)$. Let $\mathbf{w}\in\mathbb{R}^d$ be a unit vector and $t\in\mathbb{R}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq y]\geq \mathrm{OPT}+\epsilon$ and $\theta(\mathbf{w}),\mathbf{w}^*)\leq\pi-\epsilon$, for some $\epsilon\in(0,1]$. There exists an algorithm that draws $N=d^{O(\log(1/\epsilon))}\log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N,d)$ and, with probability at least $1-\delta$ returns a basis of a subspace $V\subseteq\mathbf{w}^\perp$ such that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp_\mathbf{w}})\|_2=\mathrm{poly}(\epsilon)$.*

The proof of Proposition D.2 is the same as Proposition 5.2 with the only difference be that, instead of Lemma 5.6 we use the lemma below.

Lemma D.3. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition with $\eta=1/2$ with respect to $f(\mathbf{x})=\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t)$ with $t\in\mathbb{R}$. Fix $\epsilon>0$ and $\rho>0$ such that $\rho\lesssim\epsilon$. Let $\mathbf{w}$ be a unit vector such that for any $t'\in\mathbb{R}$, it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{f(\mathbf{x})\neq\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t')\}]\geq\epsilon$. For $t_1,t_2\in\mathbb{R}$ consider the band $B=\{t_1\leq\mathbf{w}\cdot\mathbf{x}\leq t_2\}$. Denote $\mathcal{D}^\perp=\mathcal{D}_B^{\perp_\mathbf{w}}$, i.e., $\mathcal{D}$ is the orthogonal projection onto $\mathbf{w}^\perp$ of the conditional distribution on B, and consider the halfspace $f^\perp:\mathbf{w}^\perp\mapsto\{\pm1\}$, with $f^\perp(\mathbf{x})=\mathrm{sign}(\mathbf{x}\cdot(\mathbf{w}^*)^{\perp_\mathbf{w}}-b)$, for some threshold $b\in\mathbb{R}$. Moreover, define the noise function*

$$\eta^\perp(\mathbf{x})=\mathop{\mathbf{Pr}}_{(\mathbf{z},y)\sim\mathcal{D}^\perp}[y\neq f^\perp(\mathbf{z})|\mathbf{z}=\mathbf{x}]\,.$$

There exist $t_1,t_2\in\mathbb{R}$ multiples of ρ, with $|t_1-t_2|=\rho$, such that:

- $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}^{\perp}_{\mathbf{x}}}[(1-2\eta^{\perp}(\mathbf{x}))\mathbb{1}\{f^{\perp}(\mathbf{x})\mathrm{sign}(b)>0\}]\gtrsim \epsilon/\sqrt{\log(1/\epsilon)}-\rho/\epsilon$,
- *if* $\eta^{\perp}(\mathbf{x})>1/2$*, then* $0<\mathrm{sign}(b)(\mathbf{x}\cdot(\mathbf{w}^*)^{\perp_{\mathbf{w}}}-b)\le\rho/\epsilon$.

Proof. First, we consider the region $B=\{t_1\le\mathbf{w}\cdot\mathbf{x}\le t_2\}$, for any $t_1,t_2\in\mathbb{R}$ with $|t_1-t_2|=\rho$. We denote by $\mathbf{x}^{\perp}$ the projection of $\mathbf{x}$ onto the subspace $\mathbf{w}^{\perp}$. Define the distribution $\mathcal{D}^{\perp}=\mathcal{D}_B^{\mathrm{proj}_{\mathbf{w}^{\perp}}}$, that is the distribution $\mathcal{D}$ conditioned on the set B and projected onto $\mathbf{w}^{\perp}$, the hypothesis $f^{\perp}(\mathbf{x}^{\perp})=\mathrm{sign}(\mathbf{x}^{\perp}\cdot(\mathbf{w}^*)^{\perp_{\mathbf{w}}}-b)$ where $b\in\mathbb{R}$ is chosen appropriately below, and the noise function $\eta^{\perp}(\mathbf{x}^{\perp})=\mathbf{Pr}_{(\mathbf{z},y)\sim\mathcal{D}^{\perp}}[y\neq f^{\perp}(\mathbf{z})|\mathbf{z}=\mathbf{x}^{\perp}]$.

Note that the distribution $\mathcal{D}^{\perp}$ does not satisfy the 1/2-Massart noise condition anymore. We first illustrate how the noise function changes. The orthogonal projection on $\mathbf{w}^{\perp}$ creates a region where the Massart condition is violated, i.e., a region where $\eta^{\perp}(\mathbf{x}^{\perp})\ge 1/2$, but we can control the probability that we get points inside this region. More formally, we show that$\mathbf{Pr}_{(\mathbf{x}^{\perp},y)\sim\mathcal{D}^{\perp}}[\eta^{\perp}(\mathbf{x}^{\perp})\ge 1/2]\lesssim\rho/\epsilon$.

To show that first notice that $\theta(\mathbf{v},\mathbf{w}^*)\ge\epsilon$, otherwise we would have $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[f(\mathbf{x})\neq\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})]\le\epsilon$. We can assume that $\mathbf{w}^*=\lambda_1\mathbf{w}+\lambda_2(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$, where $\lambda_1=\cos\theta$ and $\lambda_2=\sin\theta$. Next we set $\mathbf{x}=(\mathbf{x}_{\mathbf{w}},\mathbf{x}^{\perp})$, where $\mathbf{x}_{\mathbf{w}}=\mathbf{w}\cdot\mathbf{x}$. We show that the hypothesis $f^{\perp}(\mathbf{x})=\mathrm{sign}((\mathbf{w}^*)^{\perp_{\mathbf{w}}}\cdot\mathbf{x}^{\perp}+(t+t_1\lambda_1)/\lambda_2)$ is almost as good as the $f(\mathbf{x})$ for the distribution $\mathcal{D}^{\perp}$.

Conditioned on $\mathbf{x}\in B$, i.e., $\mathbf{x}_{\mathbf{w}}\in[t_1,t_1+\rho]$, it holds that

$$\mathbf{w}^*\cdot\mathbf{x}=\lambda_1\mathbf{x}_{\mathbf{w}}+\lambda_2(\mathbf{w}^*)^{\perp_{\mathbf{w}}}\cdot\mathbf{x}^{\perp}=\lambda_1 t_1+\lambda_2(\mathbf{w}^*)^{\perp_{\mathbf{w}}}\cdot\mathbf{x}^{\perp}+s\rho\,,$$

for some $s\in[-1,1]$ (recall that $|\lambda_1|\le 1$). Let $b=-(t+\lambda_1 t_1)/\lambda_2$. Notice that when $0\le\mathrm{sign}(b)((\mathbf{w}^*)^{\perp_{\mathbf{w}}}\cdot\mathbf{x}^{\perp}-b)<\rho/\lambda_2$, $f^{\perp}(\mathbf{x}^{\perp})$ is not equal to the sign of $(\mathbf{w}^*\cdot\mathbf{x}+t)$ (recall that $\lambda_2>0$), and therefore we are inside the region that Massart noise is violated. Thus, we need to bound the probability of this event I^{ξ}. We have that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}^{\perp},y)\sim\mathcal{D}^{\perp}}[\eta^{\perp}(\mathbf{x}^{\perp})>1/2]=\frac{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{w}^*)^{\perp_{\mathbf{w}}}\cdot\mathbf{x}\in I^{\xi}\right]}{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x}\in B]}\lesssim\rho/\lambda_2\lesssim\rho/\epsilon\,,$$

where we used the anti-concentration property of the Gaussian distribution and the last inequality holds because we have that $\lambda_2\gtrsim\epsilon$.

It remains to prove that there is a choice t_1,t_2 and a band $B=\{t_1\le\mathbf{w}\cdot\mathbf{x}\le t_2\}$ with respect the t_1,t_2 such that $\mathbf{E}_{\mathbf{x}\sim(\mathcal{D}_B^{\perp})_{\mathbf{x}}}[(1-2\eta^{\perp}(\mathbf{x}))\mathbb{1}\{f(\mathbf{x})^{\perp}\neq\mathrm{sign}(-b)\}]\gtrsim\frac{\epsilon}{\sqrt{\log(1/\epsilon)}}$, where $\mathcal{D}_B^{\perp}$ is the distribution $\mathcal{D}$ conditioned on the set B and projected onto $\mathbf{w}^{\perp}$. We first show the following claim Let $t_i=i\rho$ and $t_{-i}=-i\rho$, for $0\le i\le C\log(1/\epsilon)/\rho$ where $C>0$ is a large enough constant. We define $B_i=\{t_i\le\mathbf{v}\cdot\mathbf{x}\le t_{i+1}\}$ and $B_{-i}=\{-t_{i+1}\le\mathbf{v}\cdot\mathbf{x}\le -t_i\}$. For each B_i, we define the distributions $\mathcal{D}_{B_i}^{\perp}$, the hypothesis $f_i^{\perp}(\mathbf{x}^{\perp})=\mathrm{sign}(\mathbf{x}^{\perp}\cdot(\mathbf{w}^*)^{\perp_{\mathbf{w}}}-b_i)$ and the noise functions $\eta_i^{\perp}(\mathbf{x}^{\perp})$. We remind that from the assumptions, we have that for any $t'\in\mathbb{R}$, it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\beta(\mathbf{x})\mathbb{1}\{f(\mathbf{x})\neq\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t')\}]\ge\epsilon$. Choose $t'=t/\sin\theta$, where $\theta=\theta(\mathbf{w},\mathbf{w}^*)$ and an application of Claim 5.8 to the set $\{\mathbf{x}\in\mathbb{R}^d: f(\mathbf{x})\neq\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t')\}$, gives us that there exists an index i' such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{f(\mathbf{x})\neq\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t')\}\mathbb{1}\{\mathbf{x}\in B_{i'}\}\beta(\mathbf{x})]\gtrsim\frac{\epsilon\rho}{\sqrt{\log(1/\epsilon)}}\,.$$

Moreover, note that for the distribution $\mathcal{D}_{B_{i'}}$, that is the distribution $\mathcal{D}$ conditioned on B, it holds $\mathbf{E}_{\mathbf{x}\sim(\mathcal{D}_{B_{i'}})_{\mathbf{x}}}[\beta(\mathbf{x})\mathbb{1}\{f(\mathbf{x})\neq\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t')\}]\gtrsim\frac{\epsilon}{\sqrt{\log(1/\epsilon)}}$, where we used the Gaussian concentration. We have $f_{i'}^{\perp}(\mathbf{x})$ agrees almost everywhere with $f(\mathbf{x})$ with respect the distribution $\mathcal{D}_{B_{i'}}$, i.e., we have that $\mathbf{E}_{(\mathbf{x}_{\mathbf{w}},\mathbf{x}^{\perp})\sim(\mathcal{D}_{B_{i'}})_i}[\mathbb{1}\{f((\mathbf{x}_{\mathbf{w}},\mathbf{x}^{\perp}))\neq f^{\perp}(\mathbf{x}^{\perp})\}]\lesssim\rho/\epsilon$. Thus using the triangle inequality, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x}_{\mathbf{w}},\mathbf{x}^{\perp})\sim(\mathcal{D}_{B_{i'}})_{\mathbf{x}}}\left[(1-2\eta^{\perp}(\mathbf{x}^{\perp}))\mathbb{1}\{\mathrm{sign}(\mathbf{x}_{\mathbf{w}}+t')\neq f^{\perp}(\mathbf{x}^{\perp})\}\right]\gtrsim\frac{\epsilon}{\sqrt{\log(1/\epsilon)}}-\rho/\epsilon\,.$$

The proof concludes by noting that in each $B_{i'}$, it holds $\mathrm{sign}(\mathbf{x_w} + t') = \mathrm{sign}(-b_i)$ from construction of $f^{\perp}(\mathbf{x})$. This completes the proof of Lemma D.3.

□

We are now ready to prove the Theorem D.1. Note that this is the same proof as Theorem 5.1 the only difference is that we need to test the different thresholds.

Proof of Theorem D.1. First, let $\widehat{\mathcal{D}}_N$ be the empirical distribution of $\mathcal{D}$ using $N \in \mathbb{Z}_+$ samples. Let $\mathbf{w}^{(i)}$ be the current guess. If $\min_{t\in\mathbb{R}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t) \neq y] \geq \mathrm{OPT} + \epsilon$ and $\theta(\mathbf{w}^{(i)}, \mathbf{w}^a st) \leq \pi - \epsilon$, then from Proposition D.2 with $N_1 = d^{O(\log(1/\epsilon))} \log(1/\delta_1)$ samples from $\widehat{\mathcal{D}}$ and $\mathrm{poly}(N_1, d)$ time, we compute a subspace V such that $\|\mathrm{proj}_V(\mathbf{w}^*)\|_2 \geq \mathrm{poly}(\epsilon)$ and from Fact 5.14, we get a random a unit vector $\mathbf{v} \in V$ such that $\mathbf{v} \cdot \mathbf{w}^* = \mathrm{poly}(\epsilon)$ and $\mathbf{v} \cdot \mathbf{w}^{(i)} = 0$, with probability $(1 - \delta_1)/3$. We call this event $\mathcal{E}_i$. By conditioning on the event $\mathcal{E}_i$; from Lemma 5.13, after we update our current hypothesis vector $\mathbf{w}^{(i)}$ with $\mathbf{v}$, we get a unit vector $\mathbf{w}^{(i+1)}$ such that $\mathbf{w}^{(i+1)} \cdot \mathbf{w}^* \geq \mathbf{w}^{(i+1)} \cdot \mathbf{w}^* + \mathrm{poly}(\epsilon)$.

After running the update step k times and conditioning on the events $\mathcal{E}_1, \dots, \mathcal{E}_k$, then $\mathbf{w}^{(k)} \cdot \mathbf{w}^* \geq \mathbf{w}^{(1)} \cdot \mathbf{w}^* + k\,\mathrm{poly}(\epsilon)$; therefore, for $k = \mathrm{poly}(1/\epsilon)$, we get that the vector $\mathbf{w}^{(k)}$ that is competitive with the optimal hypothesis, i.e., $\min_{t\in\mathbb{R}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(k)} \cdot \mathbf{x} + t) \neq y] \leq \mathrm{OPT} + \epsilon$. The probability that all the events $\mathcal{E}_1, \dots, \mathcal{E}_k$ hold simultaneously is at least $(1 - k\delta_1) + (1/3)^k$, and thus by choosing $\delta_1 \leq 1/(3k)$, the probability of success is at least $\delta_2 = (1/3)^k$. By running the algorithm above $M = \log(1/\delta)/\delta_2$ times and along with an application of Hoeffding's inequality, we get a list L of M vectors such that contains all the unit vectors $\mathbf{w}$ the algorithm calculated along with $-\mathbf{w}$, therefore L contains a vector such $\mathbf{w}$ that such that $\min_{t\in\mathbb{R}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] \leq \mathrm{OPT} + \epsilon$, with probability $1 - \delta/2$. Moreover, from Fact 3.2 we get that if we set $\mathcal{T} = \{\pm\epsilon^2, \pm 2\epsilon^2, \dots, \pm 4\sqrt{\log(1/\epsilon)}\}$ we have that $\min_{t\in\mathcal{T}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] \leq \min_{t\in\mathbb{R}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) \neq y] + \epsilon^2$. Therefore, we construct the list $\mathcal{H} = \{(\mathbf{w}, t) : \mathbf{w} \in L, t \in \mathcal{T}\}$. Finally, to evaluate all the vectors from the list, we need a few samples, from the distribution $\mathcal{D}$ to obtain the best among them, i.e., the one that minimizes the zero-one loss.

The size of the list of candidates is at most $M \leq 2^{\mathrm{poly}(1/\epsilon)} \log(1/\delta)$. Therefore, from Hoeffding's inequality, it follows that $O(\mathrm{poly}(1/\epsilon)\log(1/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most ϵ with probability at least $1 - \delta/2$. Thus, with $N = d^{\log(1/\epsilon)} \log(1/\delta)$ samples and $\mathrm{poly}(N, d, 2^{\mathrm{poly}(1/\epsilon)})$ runtime, we get a hypothesis $(\hat{\mathbf{w}}, \hat{t})$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\hat{\mathbf{w}} \cdot \mathbf{x} + \hat{t}) \neq y] \leq \mathrm{OPT} + \epsilon$ with probability $1 - \delta$. This completes the proof. □

E Omitted Proofs from Section 5

We restate and prove Lemma 5.12.

Lemma E.1. *Fix $m \in \mathbb{Z}_+$ and $\epsilon, \delta \in (0,1)$. Let $\mathcal{D}$ be a distribution in $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginals. There is an algorithm that with $N = d^{O(m)} \log(1/\delta)/\epsilon^2$ samples and $\mathrm{poly}(d, N)$ runtime, outputs an approximation $\mathbf{T}'^m$ of the order-m Chow-parameter tensor $\mathbf{T}^m$ of $\mathcal{D}$ such that with probability $1 - \delta$, it holds*

$$\|\mathbf{T}'^m - \mathbf{T}^m\|_F \leq \epsilon \,.$$

Proof.

Fact E.2. *Fix $m \in \mathbb{Z}_+$ and $\epsilon, \delta \in (0,1)$. Let $\mathcal{D}$ be a distribution in $\mathbb{R}^d \times \{\pm 1\}$ and let $N = d^{O(m)} \log(1/\delta)/\epsilon^2$. Let α be a multi-index satisfying $|\alpha| \leq m$. Then there is an algorithm that with N samples and runtime $\mathrm{poly}(d, N)$, computes with probability $1 - \delta$ estimates $\hat{h}_\alpha$ such that*

$$\Big|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y h_\alpha(\mathbf{x})] - \hat{h}_\alpha\Big| \leq \epsilon \,,$$

for all $|\alpha| \leq m$.

Proof. Let $\widehat{\mathcal{D}}_N$ be the empirical distribution of $\mathcal{D}$ with N samples. Using Markov's inequality, we have

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

where we used the fact that for the Hermite polynomials it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_d}[h_\alpha^2(\mathbf{x})] = 1$. Moreover, all possible choices of α are at most d^m. Hence, using the fact that $N = O(d^m/\epsilon^2)$ and applying a union bound for all different h_α, we get that with constant probability it holds

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

for all $|\alpha| \leq m$. By applying a standard probability amplification technique (see, e.g., Ex. 1, Chapter 13 of [SSBD14]), we can boost the confidence to $1-\delta$ with $N' = O(N\log(1/\delta))$ samples. ☐

In order to estimate up to order-m Chow tensors, with accuracy ϵ, we need to learn each order-m Chow parameters up to accuracy $\epsilon^2 d^{-m}$, therefore we need to $d^{O(m)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ samples. ☐

We restate and prove Lemma 5.13.

Lemma E.3. *For unit vectors* $\mathbf{v}^*, \mathbf{v} \in \mathbb{R}^d$*, let* $\mathbf{u} \in \mathbb{R}^d$ *such that* $\mathbf{u}\cdot\mathbf{v}^* \geq c$, $\mathbf{u}\cdot\mathbf{v} = 0$*, and* $\|\mathbf{u}\|_2 \leq 1$*, with* $c > 0$*. Then, for* $\mathbf{v}' = \frac{\mathbf{v}+\lambda\mathbf{u}}{\|\mathbf{v}+\lambda\mathbf{u}\|_2}$*, with* $\lambda = c/2$*, we have that* $\mathbf{v}'\cdot\mathbf{v}^* \geq \mathbf{v}\cdot\mathbf{v}^* + \lambda^2/2$.

Proof. We will show that $\mathbf{v}'\cdot\mathbf{v}^* = \cos\theta' \geq \cos\theta + \lambda^2/2$, where $\cos\theta = \mathbf{v}\cdot\mathbf{v}^*$. We have that

$$\|\mathbf{v}+\lambda\mathbf{u}\|_2 = \sqrt{1+\lambda^2\|\mathbf{u}\|_2^2 + 2\lambda\mathbf{u}\cdot\mathbf{v}} \leq 1 + \lambda^2\|\mathbf{u}\|_2^2 , \tag{46}$$

where we used that $\sqrt{1+a} \leq 1 + a/2$. Using the update rule, we have

$$\mathbf{v}'\cdot\mathbf{v}^* = \mathbf{v}'\cdot(\mathbf{v}^*)^{\perp\mathbf{v}}\sin\theta + \mathbf{v}'\cdot\mathbf{v}\cos\theta = \frac{\lambda\mathbf{u}\cdot(\mathbf{v}^*)^{\perp\mathbf{v}}}{\|\mathbf{v}+\lambda\mathbf{u}\|_2}\sin\theta + \frac{\mathbf{v}+\lambda\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{v}+\lambda\mathbf{u}\|_2}\cos\theta .$$

Now using Equation (46), we get

$$\mathbf{v}'\cdot\mathbf{v}^* \geq \frac{\lambda\mathbf{u}\cdot(\mathbf{v}^*)^{\perp\mathbf{v}}}{1+\lambda^2\|\mathbf{u}\|_2^2}\sin\theta + \frac{\cos\theta}{1+\lambda^2\|\mathbf{u}\|_2^2} = \cos\theta + \frac{\lambda\mathbf{u}\cdot(\mathbf{v}^*)^{\perp\mathbf{v}}}{1+\lambda^2\|\mathbf{u}\|_2^2}\sin\theta + \frac{-\lambda^2\|\mathbf{u}\|_2^2\cos\theta}{1+\lambda^2\|\mathbf{u}\|_2^2} .$$

Then, using that $\mathbf{u}\cdot\mathbf{v}^* = \mathbf{u}\cdot(\mathbf{v}^*)^{\perp\mathbf{v}}\sin\theta$, we have that $\mathbf{u}\cdot(\mathbf{v}^*)^{\perp\mathbf{v}} \geq \frac{c}{\sin\theta}$, thus

$$\mathbf{v}'\cdot\mathbf{v}^* \geq \cos\theta + \frac{\lambda c - \lambda^2\|\mathbf{u}\|_2^2}{1+\lambda^2\|\mathbf{u}\|_2^2} \geq \cos\theta + \frac{\lambda c - \lambda^2}{1+\lambda^2\|\mathbf{u}\|_2^2} = \cos\theta + \frac{1}{4}\frac{c^2}{1+\lambda^2\|\mathbf{u}\|_2^2} ,$$

where in the first inequality we used that $\|\mathbf{u}\|_2 \leq 1$ and in the second that for $\lambda = c/2$ it holds $c - \lambda \geq c/2$. Finally, we have that

$$\cos\theta' = \mathbf{v}'\cdot\mathbf{v}^* \geq \cos\theta + \frac{1}{4}\frac{c^2}{1+\lambda^2\|\mathbf{u}\|_2^2} \geq \cos\theta + \frac{1}{8}c^2 = \cos\theta + \frac{1}{2}\lambda^2 .$$

This completes the proof. ☐

F Omitted Proofs from Section 4

Lemma F.1 (Theorem 8 of [CW01]). *Let $p : \mathbb{R}^d \mapsto R$ be a polynomial of degree at most n. Then there is an absolute constant $C > 0$ such that for any $0 < q < \infty$ and $t \geq 0$, it holds* $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}_d}[|p(\mathbf{x})| \leq \gamma] \leq Cq\gamma^{1/n}(\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_d}[|p(\mathbf{x})|^{q/n}])^{-1/q}$.

Lemma F.2. *Let $t \geq 0$. There exists an absolute constant $C' \geq 1$ such that for any univariate polynomial of degree k it holds*

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \geq t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \leq t\}]} \leq e^{C'k\log k - t^2/C'} .$$

Proof. We start by bounding from above the $\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \geq t\}]$. Using the Cauchy-Schwarz inequality, we get

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \geq t\}] \leq (\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)])^{1/2}(\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}}[x \geq t])^{1/2} \lesssim (\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)])^{1/2}e^{-t^2/4} . \tag{47}$$

In order to bound the $\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \leq t\}]$ from below, we are going to use Lemma F.1. By setting $q = 4k, n = k$ and $\gamma = (\frac{1}{12Ck})^k(\mathbf{E}_{x\sim\mathcal{N}}[p^4(x)])^{1/4}$, on Lemma F.1. We get that

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}}\left[|p(x)| \leq \left(\frac{1}{12Ck}\right)^k \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)]\right)^{1/4}\right] \leq 1/3 .$$

Therefore, by squaring we get

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}}\left[p^2(x) \geq \left(\frac{1}{12Ck}\right)^{2k} \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)]\right)^{1/2}\right] \geq 2/3 . \tag{48}$$

Furthermore, using the assumption that $t \geq 0$, we have $\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \leq t\}] \geq \mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \leq 0\}]$, hence, combing the with Equation (48), we get

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \leq t\}] \gtrsim \left(\frac{1}{12Ck}\right)^{2k} \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)]\right)^{1/2} . \tag{49}$$

Combining eq. (47) and eq. (49), we get

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \geq t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \leq t\}]} \lesssim (12Ck)^{2k}e^{-t^2/4} \leq e^{C'k\log k - t^2/C'} ,$$

for some $C' \geq 1$ absolute constant. $\square$

We prove below Fact 4.15, we restate it for convenience.

Fact F.3 (Estimation of $\mathbf{M}$). *Let $\Omega = \{\mathbf{A} \in \mathcal{S}^m : \|\mathbf{A}\|_F \leq 1\}$ and $\epsilon, \delta \in (0,1)$. Let $\ell(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + t$ with $|\ell(\mathbf{x})|^2 \leq C$ and $\widetilde{\mathbf{M}} = \frac{1}{N}\sum_{i=1}^N \mathbf{m}(\mathbf{x}^{(i)})\mathbf{m}(\mathbf{x}^{(i)})^\top \mathbb{1}_B(\mathbf{x}^{(i)})y^{(i)}\ell(\mathbf{x}^{(i)})$. There exists an algorithm that draws $N = \frac{d^{O(\log\frac{1}{\gamma\beta})}}{C\epsilon^2}\log(1/\delta)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(N,d)$ time and with probability at least $1-\delta$ outputs a matrix $\widetilde{\mathbf{M}}$ such that*

$$\mathbf{Pr}\left[\sup_{\mathbf{A}\in\Omega}\left|\mathrm{tr}(\mathbf{A}\widetilde{\mathbf{M}}) - \mathrm{tr}(\mathbf{A}\mathbf{M})\right| \geq \epsilon\right] \leq 1-\delta .$$

Proof. Using the Cauchy-Schwarz inequality, we get

$$\operatorname{tr}\left(\mathbf{A}(\mathbf{M}-\widetilde{\mathbf{M}})\right) \leq \|\mathbf{A}\|_F \left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F .$$

Therefore, it suffices to bound the probability that $\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F \geq \epsilon$. From Markov's inequality, we have

$$\mathbf{Pr}\left[\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F \geq \epsilon\right] \leq \frac{1}{\epsilon^2}\mathbf{E}\left[\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F^2\right] . \tag{50}$$

Using multi-indices S_1, S_2 that correspond to the monomials $\mathbf{x}^{S_1}, \mathbf{x}^{S_2}$ (as indices of the matrix $\mathbf{M}$), we have

$$\mathbf{E}\left[\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F^2\right] = \sum_{S_1,S_2:|S_1|,|S_2|\leq k} (\mathbf{M}_{S_1,S_2}-\widetilde{\mathbf{M}}_{S_1,S_2})^2 = \sum_{S_1,S_2:|S_1|,|S_2|\leq k} \mathbf{Var}[\widetilde{\mathbf{M}}_{S_1,S_2}] .$$

Using the fact that the samples $(\mathbf{x}^{(i)}, y^{(i)})$ are independent, we can bound from above the variance of each entry (S_1, S_2) of $\widetilde{\mathbf{M}}$

$$\begin{aligned}\mathbf{Var}[\widetilde{\mathbf{M}}_{S_1,S_2}] &\leq \frac{1}{N}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbf{x}^{2(S_1+S_2)}\left(\mathbb{1}_B(\mathbf{x})\ell(\mathbf{x})y\right)^2\right] \leq \frac{2}{N}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}^{2(S_1+S_2)}(\|\mathbf{x}\|_2^2+t^2)\right]\\ &\leq \frac{2C}{N}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\|\mathbf{x}\|_2^2)^{|S_1+S_2|+1}\right] .\end{aligned}$$

For every $n \geq 1$, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\|\mathbf{x}\|_2^2)^n\right] = d^{O(n)}$. Using the above bound for the variance and summing over all pairs S_1, S_2 with $|S_1|, |S_2| \leq k$, we obtain

$$\mathbf{E}\left[\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F^2\right] = \frac{2C}{N}d^{O(k)} \tag{51}$$

Combining Equation (50) and Equation (51) we obtain that with $N = d^{O(m)}/(C\epsilon^2)$ samples we can estimate $\mathbf{M}$ within the target accuracy with probability at least $3/4$. To amplify the probability to $1-\delta$, we can simply use the above empirical estimate ℓ times to obtain estimates $\widetilde{\mathbf{M}}^{(1)}, \ldots, \widetilde{\mathbf{M}}^{(\ell)}$ and keep the coordinate-wise median as our final estimate. It follows that $O(\log(m/\delta))$ repetitions suffice to guarantee confidence probability at least $1-\delta$.

□

We restate and prove Lemma 4.19.

Lemma F.4 (Estimating the function r_i). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal and let $T^{(i)}(\mathbf{x})$ be a non-negative function returned by a (2ρ)-certificate oracle. Moreover, assume that $T^{(i)}(\mathbf{x})$ has bounded ℓ_4 norm, i.e., $\|T^{(i)}(\mathbf{x})\|_4 \leq 1$. Then after drawing $O(d\log(1/\epsilon)/\epsilon^2\log(d/\delta))$ samples from $\mathcal{D}$, with probability at least $1-\delta$, we can compute an estimator $\hat{r}_i$ that satisfies the following conditions:*

- $\left\|\nabla\hat{r}_i(\mathbf{w},t) - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(T^{(i)}(\mathbf{x})+\rho)y(\mathbf{x},1)]\right\|_2 \leq \epsilon/\sqrt{\log(1/\epsilon)}$
- $\|\nabla\hat{r}_i(\mathbf{w},t)\|_2 \leq 2\sqrt{d}$.

Proof. Let $\mathcal{D}_N$ be the empirical distribution of $\mathcal{D}$ with N samples. It suffices to find N, such that with probability $1-\delta$ the

$$\left\| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}_N}[(T(\mathbf{x})+\rho)y\mathbf{x}] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)y\mathbf{x}] \right\|_2 \le \epsilon\,.$$

Let $\widehat{\mathbf{R}} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}_N}[(T(\mathbf{x})+\rho)y\mathbf{x}]$ and $\mathbf{R} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)y\mathbf{x}]$. From Markov's inequality, we have that

$$\mathbf{Pr}\left[\left\|\widehat{\mathbf{R}}-\mathbf{R}\right\|_2 \ge \epsilon\right] \le \frac{1}{\epsilon^2}\mathbf{E}[\|\widehat{\mathbf{R}}-\mathbf{R}\|_2^2] \le \frac{1}{N\epsilon^2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|(T(\mathbf{x})+\rho)y\mathbf{x}\|_2^2]\,.$$

From Cauchy-Schwarz, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\|(T(\mathbf{x})+\rho)\mathbf{x}\|_2^2] \le (\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(T(\mathbf{x})+\rho)^4])^{1/2}(\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\|\mathbf{x}\|_2^4])^{1/2}$. From the fact that $\|T\|_4 \le 1$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(T(\mathbf{x})+\rho)^4] \le 4(1+\rho^4)$. Moreover, $(\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\|\mathbf{x}\|_2^4])^{1/2} \lesssim d$, hence

$$\mathbf{Pr}\left[\left\|\widehat{\mathbf{R}}-\mathbf{R}\right\|_2 \ge \epsilon\right] \lesssim \frac{4d}{N\epsilon^2}\,.$$

For $N' = O(\frac{d}{\epsilon^2})$ samples, we get that the above holds with constant probability. Like in Fact 4.15, we can amplify the probability to $1-\delta$, using $N = N'\log(d/\delta)$ samples.

Therefore, let $\widetilde{\mathbf{R}}$ be the estimator of $\mathbf{R}$, we have

$$\left\|\widetilde{\mathbf{R}}\right\|_2 \le \|\mathbf{R}\|_2 + \epsilon \le 4\sqrt{(1+\rho^2)}\sqrt{d} + \epsilon \le 8\sqrt{d}\,. \tag{52}$$

Finally, it remains to show that we can compute an estimator $\widetilde{\mathbf{M}} = (\widetilde{\mathbf{R}}, \tilde{m})$ such that for any unit vector $\mathbf{v}\in\mathbb{R}^d$ and any $|t| \le 4\sqrt{\log(1/\epsilon)}$, we have with probability $1-\delta$ that

$$|\widetilde{\mathbf{R}}\cdot\mathbf{v} + \tilde{m}t - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)y(\mathbf{x}\cdot\mathbf{v}+t)]| \le \epsilon\,.$$

From the above, for any unit vector $\mathbf{v}\in\mathbb{R}^d$, we have with probability $1-\delta$ that $|\widetilde{\mathbf{R}}\cdot\mathbf{v} - \mathbf{R}\cdot\mathbf{v}| \le \|\widetilde{\mathbf{R}}-\mathbf{R}\|_2 \le \epsilon$. Therefore, it remains to estimate $\widetilde{m}$. Let $\mathcal{D}_{N'}$ be the empirical distribution of $\mathcal{D}$ with N' samples. From Markov's inequality we have for any $\epsilon' > 0$ that

$$\mathbf{Pr}[|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}_{N'}}[(T(\mathbf{x})+\rho)yt] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)yt]| \ge \epsilon'] \le \frac{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)^2]t^2}{N'\epsilon^2} \le \frac{4t^2}{N'\epsilon^2}\,.$$

Therefore, using $N' = O(\log(1/\epsilon)/\epsilon^2)$ we get an estimator with constant probability and hence by using the boosting technique as before, we get that with overall $N'\log(1/\delta)$ samples, we can boost the probability to $1-\delta$. □

G Omitted Proofs from Section 6

In this section, we prove Lemma 6.17. We restate the lemma for convenience.

Lemma G.1. *If there is no polynomial $p\in\mathcal{P}_k^0$ such that $\beta\|(pf)^+\|_1 \ge \|(pf)^-\|_1$ then, the LP (39) is feasible if only if LP (40) is infeasible.*

Proof. First we introduce some notation. We use $(\tilde{h}, c)$ for the inequality $\mathbf{E}_{z\sim\mathcal{N}}[\beta(z)\tilde{h}(z)] + c \le 0$, where $\tilde{h}\in L^1(\mathbb{R})$ and $c\in\mathbb{R}$. Moreover, let $\mathcal{S}$ be the set that contains all such tuples that describe the target system. For the set $\mathcal{S}$, the closed convex cone over $L^1(\mathbb{R})\times R$ is the smallest closed set $\mathcal{S}_+$ satisfying, if $A\in\mathcal{S}_+$ and $B\in\mathcal{S}_+$ then $A+B\in\mathcal{S}_+$ and, if $A\in\mathcal{S}_+$ then $\lambda A\in\mathcal{S}_+$ for all $\lambda\ge 0$. Note that the $\mathcal{S}_+$ contains the same feasible solutions as $\mathcal{S}$. In order to prove this, we need the following functional analysis result from [Fan68].

Fact G.2 (Theorem 1 of [Fan68]). *If $\mathcal{X}$ is a locally convex, real separated vector space then, a linear system described by $\mathcal{S}$ is feasible (i.e., there exists a $g \in \mathcal{X}^*$) if and only if $(0,1) \notin \mathcal{S}_+$.*

Our LP is defined by the following inequalities: $(pf, 0)$ for $p \in \mathcal{P}_k^0$, $(h, -\|h\|_1)$ for all $h \in L_+^1(\mathbb{R})$, $(-T, \beta\|T\|_1)$ for all $T \in L_+^1(\mathbb{R})$. By taking the convex cone defined from the above inequalities, we get the following set

$$\mathcal{S}'_+ = \{(pf + h - T, -\|h\|_1 + \beta\|T\|_1) : p \in \mathcal{P}_k^0, h \in L_+^1(\mathbb{R}), T \in L_+^1(\mathbb{R})\} .$$

We have the following lemma.

Lemma G.3. *If there is no polynomial $p \in \mathcal{P}_k^0$ such that $\beta\|(pf)^+\|_1 \geq \|(pf)^-\|_1$ then, the LP described by $\mathcal{S}$ is feasible if only if $(0,1) \notin \mathcal{S}'_+$.*

Proof. This proof follows from Fact G.2 by showing that $\mathcal{S}'_+$ is closed, i.e., $\mathcal{S}'_+ = \mathcal{S}_+$. Let $F = fp + h - T$ and $\lambda \geq \|h\|_1 - \beta\|T\|_1$, therefore the above LP is equivalent to $\mathbf{E}_{z\sim\mathcal{N}}[\beta(z)F(z)] \leq \lambda$. Let $(p_n, h_n, T_n, \lambda_n)$ be a sequence that satisfies our constraints, i.e., $(fp_n + h_n - T_n, \lambda_n) \in S'_+$, with $F_n = fp_n + h_n - T_n$ converging with respect the ℓ_1-norm to F_0 and λ_n converging to λ_0.

Note that $\lambda_n \geq \|(F_n - (fp_n))^+\|_1 - \beta\|(F_n - (fp_n))^-\|_1$. This holds because h_n and T_n are positive functions and $F_n - fp_n = h_n - T_n$, therefore $h_n = (F_n - (fp_n))^+$ and $T_n = (F_n - (fp_n))^-$. We have that

$$\|(F_n - (fp_n))^+\|_1 - \beta\|(F_n - (fp_n))^-\|_1 \geq -2\|F_n\|_1 + \|(fp_n)^-\|_1 - \beta\|(fp_n)^+\|_1 .$$

Note that from our assumption we have that for any polynomial $p \in \mathcal{P}_k^0$, it holds $\|(pf)^-\|_1 - \beta\|(pf)^+\|_1 > 0$. Because this is homogeneous with respect p, we can assume that $\|p\|_1 = 1$. Therefore, using the fact that the set $\|p\|_1 = 1$ and $\mathcal{P}_k^0$ is compact, we have that all the limits are inside $\mathcal{P}_k^0$, thus there exists a $c > 0$, such that $\|(pf)^-\|_1 - \beta\|(pf)^+\|_1 > c$. Therefore, it holds $\|(pf)^-\|_1 - \beta\|(pf)^+\|_1 > c\|p\|_1$.

From the above, we have that $\lambda_n + 2\|F_n\|_1 \geq c\|p_n\|_1$, and because $F_n \to F_0$ and $\lambda_n \to \lambda_0$, that means that $\|p_n\|_1$ is bounded. Since, an L^1 ball in $\mathcal{P}_k^0$ is compact, there is a subsequence such that $p_n \to p_0$. By setting $h_0 = (F_0 - (fp_0))^+$ and $T_0 = (F_0 - (fp_0))^-$, we find appropriate p_0, h_0, T_0 that give the appropriate limit points. Therefore, the set $\mathcal{S}'_+$ is closed and the lemma follows from Fact G.2. ☐

The proof of Lemma 6.17, follows from Lemma G.3, by noting that $(0,1) \notin \mathcal{S}'_+$ is equivalent to the infeasibility of LP (40). ☐