Proceedings of Machine Learning Research vol 178:1–25, 2022 35th Annual Conference on Learning Theory

Near-Optimal Statistical Query Hardness of Learning Halfspaces with Massart Noise

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin, Madison

Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study the problem of PAC learning halfspaces with Massart noise. Given labeled samples (x, y) from a distribution D on $\mathbb{R}^d \times \{\pm 1\}$ such that the marginal D_x on the examples is arbitrary and the label y of example x is generated from the target halfspace corrupted by a Massart adversary with flipping probability $\eta(x) \leq \eta \leq 1/2$, the goal is to compute a hypothesis with small misclassification error. The best known $\mathrm{poly}(d, 1/\epsilon)$-time algorithms for this problem achieve error of $\eta + \epsilon$, which can be far from the optimal bound of $\mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \mathbf{E}_{x \sim D_x}[\eta(x)]$. While it is known that achieving $\mathrm{OPT} + o(1)$ error requires super-polynomial time in the Statistical Query model, a large gap remains between known upper and lower bounds.

In this work, we essentially characterize the efficient learnability of Massart halfspaces in the Statistical Query (SQ) model. Specifically, we show that no efficient SQ algorithm for learning Massart halfspaces on $\mathbb{R}^d$ can achieve error better than $\Omega(\eta)$, even if $\mathrm{OPT} = 2^{-\log^c(d)}$, for any universal constant $c \in (0, 1)$. Furthermore, when the noise upper bound η is close to $1/2$, our error lower bound becomes $\eta - o_\eta(1)$, where the $o_\eta(1)$ term goes to 0 when η approaches $1/2$. Our results provide strong evidence that known learning algorithms for Massart halfspaces are nearly best possible.

Keywords: Statistical Query Model, Halfspaces, Linear Threshold Functions, Massart Noise

1. Introduction

A halfspace, or Linear Threshold Function (LTF), is any function $f : \mathbb{R}^m \to \{\pm 1\}$ of the form $f(x) = \mathrm{sign}(w \cdot x - \theta)$, for some weight vector $w \in \mathbb{R}^m$ and threshold $\theta \in \mathbb{R}$. (The function $\mathrm{sign} : \mathbb{R} \to \{\pm 1\}$ is defined as $\mathrm{sign}(t) = 1$ if $t \geq 0$ and $\mathrm{sign}(t) = -1$ otherwise.) Halfspaces are a fundamental class of Boolean functions that have been extensively studied in computational complexity and learning theory over several decades (Minsky and Papert, 1968; Yao, 1990; Goldmann et al., 1992; Shawe-Taylor and Cristianini, 2000). The problem of learning an unknown halfspace is as old as the field of machine learning, starting with the Perceptron algorithm (Rosenblatt, 1958).

In the realizable PAC model (Valiant, 1984), i.e., when the labels are consistent with the target function, halfspaces are efficiently learnable via Linear Programming, see, e.g., Maass and Turan (1994). In the presence of noisy data, the complexity of learning halfspaces depends on the underlying noise model. Here we study the complexity of learning halfspaces with Massart noise. In the Massart (or bounded) noise model, the label of each example x is flipped independently with probability $\eta(x) \leq \eta$, for some parameter $\eta \leq 1/2$.

Definition 1 (PAC Learning with Massart Noise) *Let $\mathcal{C}$ be a class of Boolean-valued functions over $X = \mathbb{R}^m$, D_x be a fixed unknown distribution over X, and $0 \leq \eta \leq 1/2$. Let $f \in \mathcal{C}$ be the target concept. A* Massart oracle*, $\mathrm{EX}^{\mathrm{Mas}}(f, D_x, \eta)$, works as follows: Each time $\mathrm{EX}^{\mathrm{Mas}}(f, D_x, \eta)$ is invoked, it returns an example (x, y), where $x \sim D_x$, $y = f(x)$ with probability $1 - \eta(x)$ and $y = -f(x)$ with probability $\eta(x)$, for some* unknown *function $\eta(x) : X \to [0, 1/2]$ with $\eta(x) \leq \eta$, $x \in X$. Let D be the joint distribution on (x, y) generated by the Massart oracle. A PAC learning algorithm is given i.i.d. samples from D and outputs a hypothesis $h : X \to \{\pm 1\}$ such that with high probability the error $\mathbf{Pr}_{(x,y)\sim D}[h(x) \neq y]$ is as small as possible.*

A remark is in order. While the TCS community had only considered the case that the upper bound η on the Massart noise rate is *strictly smaller* than $1/2$, this is not an essential assumption in the model. In fact, the original definition of the Massart model (Massart and Nedelec, 2006) allows for $\eta = 1/2$. (Note that it is possible that $\eta = 1/2$ while OPT is much smaller.)

The Massart model is a natural semi-random input model formulated in Massart and Nedelec (2006). An equivalent model had been defined in the 80s (Sloan, 1988, 1992; Rivest and Sloan, 1994; Sloan, 1996) (under the name "malicious misclassification noise") and a similar definition had been proposed even earlier by Vapnik (1982). The *sample complexity* of learning halfspaces with Massart noise is well-understood. Specifically, halfspaces on $\mathbb{R}^m$ are learnable to error $\mathrm{OPT} + \epsilon$ in the Massart model with $O(m/\epsilon^2)$ samples. In sharp contrast, our understanding of the *algorithmic aspects* of PAC learning natural concept classes with Massart noise is startlingly poor and has remained a tantalizing open problem in computational learning theory since the 1980s.

Sloan (1988) defined the malicious misclassification noise model and asked whether there exists an efficient learning algorithm for Boolean disjunctions— a very special case of halfspaces — in this model. Cohen (1997) asked the same question for the general class of halfspaces. The problem remained open and was highlighted in A. Blum's FOCS 2003 tutorial (Blum, 2003). Surprisingly, until recently, it was not even known whether there exists an efficient algorithm that achieves misclassification error 49% for Massart halfspaces with noise rate upper bound of $\eta = 1\%$.

Diakonikolas et al. (2019) made the first algorithmic progress on this learning problem. Specifically, they gave a $\mathrm{poly}(m, 1/\epsilon)$-time learning algorithm for Massart halfspaces with error guarantee of $\eta + \epsilon$, where η is the upper bound on the Massart noise rate. This is an *absolute* error guarantee which cannot be improved in general — since it may well be the case that $\mathrm{OPT} = \eta$ (this in particular happens when $\eta(x) = \eta$ for all $x \in X$). Motivated by Diakonikolas et al. (2019), more recent work (Diakonikolas et al., 2021a) gave an efficient boosting algorithm, achieving error $\eta + \epsilon$ for any concept class, assuming the existence of a weak learner for the class.

The aforementioned error bound of $\eta + \epsilon$ can be very far from the information-theoretically optimum error of $\mathrm{OPT} + \epsilon$. Recall that $\mathrm{OPT} = \mathbf{E}_{x\sim D_x}[\eta(x)] \leq \eta$ and it could well be the case that $\mathrm{OPT} \ll \eta$. Follow-up work by Chen et al. (2020) showed that *exact* learning — specifically, obtaining error of $\mathrm{OPT} + o(1)$, when OPT is close to $1/2$ — requires super-polynomial time in the Statistical Query (SQ) model of Kearns (1998). The latter SQ lower bound is very fragile in the sense that it does not even rule out *any* constant factor approximation algorithm for the problem, i.e., a $\mathrm{poly}(m, 1/\epsilon)$-time learning algorithm with error $C \cdot \mathrm{OPT} + \epsilon$, for a universal constant $C > 1$. See Appendix A for a more detailed summary of prior and related work.

The aforementioned progress notwithstanding, a very large gap remains in our understanding of the efficient learnability of halfspaces in the presence of Massart noise.

Question 1.1 *Is there an efficient learning algorithm for Massart halfspaces achieving a* relative *error guarantee? Specifically, if* $\mathrm{OPT} \ll \eta$ *is it possible to achieve error significantly better than* η*? What is the best error (as a function of* OPT *and* η*) that can be achieved in polynomial time?*

We emphasize here that, throughout this work, we focus on *improper learning*, where the learning algorithm is allowed to output any polynomially evaluatable hypothesis.
In this paper, *we essentially resolve the efficient PAC learnability of Massart halfspaces in the SQ model.* Specifically, we prove a near-optimal super-polynomial SQ lower bound for this problem, which provides *strong evidence that known efficient algorithms are nearly best possible.*

Statistical Query (SQ) Model SQ algorithms are the class of algorithms that are only allowed to query expectations of bounded functions of the underlying distribution rather than directly access samples. The SQ model was introduced in Kearns (1998) in the context of supervised learning as a natural restriction of the PAC model (Valiant, 1984) and has been extensively studied in learning theory. A recent line of work (Feldman et al., 2013, 2015, 2017; Feldman, 2017) generalized the SQ framework for search problems over distributions. See Feldman (2016) for a survey.

The class of SQ algorithms is broad: a range of known algorithmic techniques in machine learning are known to be implementable in the SQ model. These include spectral techniques, moment and tensor methods, local search (e.g., Expectation Maximization), and many others (see, e.g., Chu et al. (2006); Feldman et al. (2013, 2017)). In the context of PAC learning classes of Boolean functions (the topic of this paper), with the exception of learning algorithms using Gaussian elimination (in particular for the concept class of parities, see, e.g., Blum et al. (2003)), all known algorithms with non-trivial performance guarantees are either SQ or are implementable using SQs.

1.1. Our Contributions

We show that any efficient SQ learning algorithm for Massart halfspaces on $\mathbb{R}^m$ cannot obtain error better than $\Omega(\eta)$, even if the optimal is as small as $\mathrm{OPT} = 2^{-\log^c(m)}$, for any constant $c \in (0,1)$.

Theorem 2 (Main Result) *For any universal constants* c, c' *with* $0 < c < 1$ *and* $0 < c' < 1 - c$*, the following holds. For any sufficiently large positive integer* m *and any* $0 < \eta < 1/2$*, there is no SQ algorithm that PAC learns the class of halfspaces in* $\mathbb{R}^m$ *with* η*-Massart noise to error better than* $\Omega(\eta)$ *using at most* $\exp(\log^{1+c}(m))$ *queries of accuracy no better than* $\exp(-\log^{1+c}(m))$*. This holds even if the optimal classifier has error* $\mathrm{OPT} = \exp(-\log^{c'}(m))$.

Recall that the efficient algorithm of Diakonikolas et al. (2019) (which can be implemented in the SQ model) achieves error arbitrarily close to η. Moreover, it is easy to see that the Massart learning problem is computationally easy when $\mathrm{OPT} \ll 1/m$. As a result, the "inapproximability gap" of $\Omega(\eta)$ versus $2^{-\log^c(m)}$ established by Theorem 2 is essentially best possible (up to the universal constant in the $\Omega(\cdot)$). For a more detailed statement, see Theorem 9.

Remark 3 When the Massart noise rate upper bound η approaches $1/2$, we can replace the lower bound of $\Omega(\eta)$ appearing in Theorem 2 by the sharper lower bound of $\eta - o_\eta(1)$. Here the term $o_\eta(1)$ goes to 0 as η approaches $1/2$. See Theorem 33 for the statement in this regime.

It is worth comparing Theorem 2 to the hardness result of Daniely (2016) for PAC learning halfspaces in the *agnostic* model. Daniely's result is qualitatively similar to our Theorem 2 with two differences: (1) The lower bound in Daniely (2016) only applies against the (much more challenging) agnostic model. (2) In the agnostic setting, it is hard to learn halfspaces within error

significantly better than $1/2$, rather than error $\Omega(\eta)$ in the Massart setting. Theorem 9 proves an SQ lower bound for a much more benign semi-random noise model at the cost of allowing somewhat better error in polynomial time. We reiterate that error arbitrarily close to η is efficiently achievable for Massart noise (Diakonikolas et al., 2019), and therefore our hardness gap is nearly best possible.

1.2. Overview of Techniques

At a high level, our proof leverages the SQ lower bound framework developed in Diakonikolas et al. (2017). We stress that, while this framework is a key ingredient of our construction, employing it in our context requires new conceptual and technical ideas, as we explain in the proceeding discussion.

Roughly speaking, the prior work (Diakonikolas et al., 2017) established the following generic SQ-hardness result: Let A be a one-dimensional distribution that matches the first k moments with the standard Gaussian G and satisfying the additional technical condition that its chi-squared norm with G is not too large. Suppose we want to distinguish between the standard high-dimensional Gaussian $N(0, I)$ on $\mathbb{R}^m$ and a distribution $\mathbf{P}_v^A$ that is a copy of A in a random direction v and is a standard Gaussian in the orthogonal complement. Then any SQ algorithm for this hypothesis testing task requires super-polynomial complexity. Roughly speaking, any SQ algorithm distinguishing between the two cases requires either at least $m^{\Omega(k)}$ samples or at least $2^{m^{\Omega(1)}}$ time.

Here we require a generalization of the latter generic result that holds even if the one-dimensional distribution A *nearly* matches the first k moments with G. Furthermore, in contrast to the unsupervised estimation problem studied in Diakonikolas et al. (2017), in our context we require a generic statement establishing the SQ-hardness of a *binary classification* problem. Such a statement (Proposition 15) is not hard to derive from the techniques of Diakonikolas et al. (2017). In more detail, Proposition 15 shows the following: Let A and B be univariate distributions (approximately) matching their first k moments with G (and each having not too large chi-squared norm with respect to G) and let $p \in (0, 1)$. We consider the distribution on labeled samples $\mathbf{P}_v^{A,B,p}$ that returns a sample from $(\mathbf{P}_v^A, 1)$ with probability p and a sample from $(\mathbf{P}_v^B, -1)$ with probability $1 - p$. Given labeled examples from $\mathbf{P}_v^{A,B,p}$, for an unknown direction v, the goal is to output a Boolean-valued hypothesis with small misclassification error. Note that it is straightforward to obtain error $\min\{p, 1 - p\}$ (as one of the two constant functions achieves this). We show that obtaining slightly better error is hard in the SQ model.

To leverage the aforementioned result in our circumstances, we would like to establish the existence of a distribution (X, Y) on $\mathbb{R} \times \{\pm 1\}$ *that corresponds to a halfspace with Massart noise* such that both the distribution of X conditioned on $Y = 1$ (denoted by $(X \mid Y = 1)$) and the distribution of X conditioned on $Y = -1$ (denoted by $(X \mid Y = -1)$) approximately match their first k moments with the standard Gaussian. *Note that k here is a parameter that we would like to make as large as possible. In particular, to prove a super-polynomial SQ lower bound, we need to be able to make this parameter k* super-constant *(as a function of the ambient dimension).*

Naturally, a number of obstacles arise while trying to achieve this. *In particular, achieving the above goal directly is provably impossible* for the following reason. Any distribution X that even approximately matches a *constant* number of low-order moments with the standard Gaussian will satisfy $\mathbf{E}[f(X)] \approx \mathbf{E}[f(G)]$ for any halfspace (LTF) f. To see this fact, we can use the known statement (see, e.g., Diakonikolas et al. (2010)) that any halfspace f can be sandwiched between low-degree polynomials $f_+ \geq f \geq f_-$ with $\mathbf{E}[f_+(G) - f_-(G)]$ small. This structural result implies that if both conditional distributions $(X \mid Y = 1)$ and $(X \mid Y = -1)$ approximately match their

low-degree moments with G, then $\mathbf{E}[f(X)|Y=1]$ will necessarily be close to $\mathbf{E}[f(X)|Y=-1]$, which cannot hold in the presence of Massart noise.

In order to circumvent this obstacle, we will instead prove a super-polynomial SQ lower bound against learning degree-d polynomial threshold functions (PTFs) under the Gaussian distribution with Massart noise, for an appropriate (super-constant) value of the degree d. Since a degree-d PTF on the vector random variable $X \in \mathbb{R}^m$ is equivalent to an LTF on $X^{\otimes d}$ — a random variable in m^d dimensions — we will thus obtain an SQ lower bound for the original halfspace Massart learning problem. We note that a similar idea was used in Daniely (2016) to prove an SQ lower bound for the problem of learning halfspaces in the agnostic model.

The next challenge is, of course, to construct the required moment-matching distributions in one dimension. *Even for our reformulated PTF learning problem, it remains unclear whether this is even possible.* For example, let $f(x) = \mathrm{sign}(p(x))$ be a degree-d PTF. Then it will be the case that $\mathbf{E}[p(X)Y] = \mathbf{E}[p(X)f(X)(1-2\eta(X))] = \mathbf{E}[|p(X)|(1-2\eta(X))] > 0$. This holds despite the fact that $\mathbf{E}[p(X) \mid Y=1] \approx \mathbf{E}[p(X) \mid Y=-1] \approx \mathbf{E}[p(G)]$. If $\mathbf{E}[p(G)] > 0$, it will be the case that $\mathbf{E}[p(X) \mid Y=-1]$ will be positive, despite the fact that the conditional distribution of $X \mid Y=-1$ is almost entirely supported on the region where $p(X) < 0$. Our construction will thus need to take advantage of finding points where $|p(X)|$ is very large.

Fortunately for us, something of a miracle occurs here. Consider a discrete univariate Gaussian G_δ with spacing σ between its values. It is not hard to show that G_δ approximately matches moments with the standard Gaussian G to error $\exp(-\Omega(1/\sigma^2))$ (see Lemma 18). On the other hand, all but a tiny fraction of the probability mass of G_σ is supported on $d = \tilde{O}(1/\sigma)$ points. Unfortunately, a discrete Gaussian is not quite suitable for the conditional distributions in our construction, as its χ^2 inner product with respect to the standard Gaussian is infinite. We can fix this issue by replacing the single discrete Gaussian with an average of discrete Gaussians with different offsets. Doing so, we obtain a distribution that nearly-matches many moments with the standard Gaussian such that all but a small fraction of its mass is supported on a small number of intervals.

As a first attempt, we can let one of our conditional distributions be this average of "offset discrete Gaussians" described above, and the other be a similar average with different offsets. Thus, both conditional distributions nearly-match moments with the standard Gaussian and are approximately supported on a small number of (disjoint) intervals. This construction actually suffices to prove a lower bound for (the much more challenging) agnostic learning model. Unfortunately, for the Massart noise model, additional properties are needed. In particular, for a univariate PTF with Massart noise, it must be the case that except for points x in a small number of intervals, we have that $\mathbf{Pr}[Y=1 \mid X=x] > \mathbf{Pr}[Y=-1 \mid X=x]$; whereas in the above described construction we have to alternate infinitely many times between $Y=1$ being more likely and $Y=-1$ more likely.

To circumvent this issue, we need the following subtle modification of our construction. Let $G_{\sigma,\theta}$ be the discrete Gaussian supported on the points $n\sigma+\theta$, for $n \in \mathbb{Z}$ (Definition 17). Our previous (failed) construction involved taking an average of $G_{\sigma,\theta}$, for some *fixed* σ and θ varying in some range. Our modified construction will involve taking an average of $G_{\sigma,\theta}$, where both σ and θ *vary together*. The effect of this feature will be that instead of producing a distribution whose support is a set of evenly spaced intervals of the same size, the support of our distributions will instead consist of a set of evenly spaced intervals whose size grows with the distance from the origin. This means that for points x near 0, the support will essentially be a collection of small, disjoint intervals. But when x becomes large enough, these intervals will begin to overlap, causing all sufficiently large points x to be in our support. By changing the offsets used in defining the conditional distribution for $Y=1$ and the conditional distribution for $Y=-1$, we can ensure that

for points x with $|x|$ small that the supports of the two conditional distributions remain disjoint. This allows us to take the optimal error OPT to be very small. However, for larger values of $|x|$, the supports become the same. Finally, by adjusting the prior probabilities of $Y = 1$ and $Y = -1$, we can ensure that $\mathbf{Pr}[Y = 1 \mid X = x] > ((1-\eta)/(\eta))\,\mathbf{Pr}[Y = -1 \mid X = x]$ for all points x with $|x|$ sufficiently large. This suffices to show that the distribution corresponds to a Massart PTF.

2. Preliminaries

We will use the framework of Statistical Query (SQ) algorithms for problems over distributions introduced in Feldman et al. (2013). We start by defining a decision problem over distributions.

Definition 4 (Decision Problem over Distributions) *We denote by $\mathcal{B}(\mathcal{D}, D)$ the decision (or hypothesis testing) problem in which the input distribution D' is promised to satisfy either (a) $D' = D$ or (b) $D' \in \mathcal{D}$, and the goal of the algorithm is to distinguish between these two cases.*

Definition 5 (STAT **Oracle)** *For a tolerance parameter $\tau > 0$ and any bounded function $f : \mathbb{R}^n \to [-1, 1]$,* STAT$(\tau)$ *returns a value $v \in [\mathbf{E}_{x\sim D}[f(x)] - \tau, \mathbf{E}_{x\sim D}[f(x)] + \tau]$.*

Definition 6 (Pairwise Correlation) *The pairwise correlation of two distributions with pdfs $D_1, D_2 : \mathbb{R}^m \to \mathbb{R}_+$ with respect to a distribution with density $D : \mathbb{R}^m \to \mathbb{R}_+$, where the support of D contains the supports of D_1 and D_2, is defined as $\chi_D(D_1, D_2) \stackrel{\text{def}}{=} \int_{\mathbb{R}^m} D_1(x)D_2(x)/D(x)dx - 1$. We say that a set of s distributions $\mathcal{D} = \{D_1, \ldots, D_s\}$ over $\mathbb{R}^m$ is (γ, β)-correlated relative to a distribution D if $|\chi_D(D_i, D_j)| \leq \gamma$ for all $i \neq j$, and $|\chi_D(D_i, D_j)| \leq \beta$ for $i = j$.*

Definition 7 (SQ Dimension) *For $\beta, \gamma > 0$, a decision problem $\mathcal{B}(\mathcal{D}, D)$, where D is a fixed distribution and $\mathcal{D}$ is a family of distributions, let s be the maximum integer such that there exists a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ such that $\mathcal{D}_D$ is (γ, β)-correlated relative to D and $|\mathcal{D}_D| \geq s$. We define the* SQ dimension *with pairwise correlations (γ, β) of $\mathcal{B}$ to be s and denote it by* SD$(\mathcal{B}, \gamma, \beta)$.

Lemma 8 (Corollary 3.12 in Feldman et al. (2013)) *Let $\mathcal{B}(\mathcal{D}, D)$ be a decision problem, where D is the reference distribution and $\mathcal{D}$ is a class of distributions. For $\gamma, \beta > 0$, let $s =$* SD$(\mathcal{B}, \gamma, \beta)$. *For any $\gamma' > 0$, any SQ algorithm for $\mathcal{B}$ requires at least $s \cdot \gamma'/(\beta - \gamma)$ queries to* STAT$(\sqrt{\gamma + \gamma'})$.

3. SQ Hardness of Learning Halfspaces with Massart Noise

In this section, we establish the following result, which implies Theorem 2.

Theorem 9 (SQ Hardness of Learning Massart Halfspaces on $\mathbb{R}^M$) *Let* OPT > 0 *and $M \in \mathbb{Z}_+$ be such that $\log(M)/(\log\log(M))^3$ is at least a sufficiently large constant multiple of $\log(1/\mathrm{OPT})$. There exists a parameter $\tau \stackrel{\text{def}}{=} M^{-\Omega\left(\frac{\log(M)}{\log\log(M)^3}/\log(1/\mathrm{OPT})\right)}$ such that no SQ algorithm can learn the class of halfspaces on $\mathbb{R}^M$ in the presence of η-Massart noise, where* OPT $< \eta < 1/2$, *within error better than $\Omega(\eta)$ using at most $1/\tau$ queries of tolerance τ. This holds even if the optimal binary classifier has misclassification error at most* OPT.

As an immediate corollary of Theorem 9, by taking $\mathrm{OPT} = \exp(-\log^{c'}(M))$ for any fixed constant $c' \in (0,1)$, we obtain a super-polynomial SQ lower bound against learning a hypothesis with error better than $\Omega(\eta)$, even when error OPT is possible. Specifically, this setting of parameters immediately implies Theorem 2, since $\frac{\log(M)}{\log\log(M)^3}/\log(1/\mathrm{OPT}) = \frac{\log^{1-c'}(M)}{\log\log(M)^3} > \log^{c}(M)$, where $0 < c < 1 - c'$, and therefore $1/\tau \gg \exp(\log^{1+c}(M))$.

Remark 10 In addition to Theorem 9, we establish an alternative "inapproximability gap" of $1/2 - O(\sqrt{1/2 - \eta})$ versus $\exp(-\log^{c'}(M))$, which implies a sharper error lower bound when η approaches $1/2$. Specifically, for η close to $1/2$, we obtain an error lower bound of $\eta - o_\eta(1)$, even if $\mathrm{OPT} = \exp(-\log^{c'}(M))$. See Theorem 33 for the formal statement.

In Section 3.1, we review the SQ framework of Diakonikolas et al. (2017) with the necessary enhancements and modifications required for our supervised setting. In Section 3.2, we establish the existence of the one-dimensional distributions with the desired approximate moment-matching properties. In Section 3.3, we put everything together to complete the proof of Theorem 9. Finally, in Appendix F, we establish our sharper lower bounds for η close to $1/2$, proving Theorem 33.

3.1. Generic SQ Lower Bound Construction

We start with the following definition and moment-matching condition:

Definition 11 (High-Dimensional Hidden Direction Distribution) *For a distribution A on the real line with probability density function $A(x)$ and a unit vector $v \in \mathbb{R}^m$, consider the distribution over $\mathbb{R}^m$ with probability density function $\mathbf{P}_v^A(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right)/(2\pi)^{(m-1)/2}$. That is, $\mathbf{P}_v^A$ is the product distribution whose orthogonal projection onto the direction of v is A, and onto the subspace perpendicular to v is the standard $(m-1)$-dimensional normal distribution.*

Condition 12 *Let $k \in \mathbb{Z}_+$ and $\nu > 0$. The distribution A is such that (i) the first k moments of A agree with the first k moments of $N(0,1)$ up to error at most ν, and (ii) $\chi^2(A, N(0,1))$ is finite.*

Note that Condition 12-(ii) above implies that the distribution A has a pdf, which we will denote by $A(x)$. We will henceforth blur the distinction between a distribution and its pdf.

Our main result in this subsection makes essential use of the following key lemma:

Lemma 13 (Correlation Lemma) *Let $k \in \mathbb{Z}_+$. If the univariate distribution A satisfies Condition 12, then for all $v, v' \in \mathbb{R}^m$, with $|v \cdot v'|$ less than a sufficiently small constant, we have that $|\chi_{N(0,I)}(\mathbf{P}_v^A, \mathbf{P}_{v'}^A)| \leq |v \cdot v'|^{k+1}\chi^2(A, N(0,1)) + \nu^2$.*

This lemma is a technical generalization of Lemma 3.4 from Diakonikolas et al. (2017), which applied under *exact* moment matching assumptions. The proof is deferred to Appendix B.

We will establish an SQ lower bound for the following binary classification problem.

Definition 14 (Hidden Direction Binary Classification Problem) *Let A and B be distributions on $\mathbb{R}$ satisfying Condition 12 with parameters $k \in \mathbb{Z}_+$ and $\nu \in \mathbb{R}_+$, and let $p \in (0,1)$. For $m \in \mathbb{Z}_+$ and a unit vector $v \in \mathbb{R}^m$, define the distribution $\mathbf{P}_v^{A,B,p}$ on $\mathbb{R}^m \times \{\pm 1\}$ that returns a sample from $(\mathbf{P}_v^A, 1)$ with probability p and a sample from $(\mathbf{P}_v^B, -1)$ with probability $1-p$. The corresponding binary classification problem is the following: Given access to a distribution on labeled examples of the form $\mathbf{P}_v^{A,B,p}$, for a fixed but unknown unit vector v, output a hypothesis $h : \mathbb{R}^m \to \{\pm 1\}$ such that $\mathbf{Pr}_{(X,Y)\sim\mathbf{P}_v^{A,B,p}}[h(X) \neq Y]$ is (approximately) minimized.*

Note that it is straightforward to obtain misclassification error $\min\{p, 1-p\}$ (as one of the identically constant functions achieves this guarantee). We show that obtaining slightly better error is hard in the SQ model. The following result is the basis for our SQ lower bounds:

Proposition 15 (Generic SQ Lower Bound) *Consider the classification problem of Definition 14. Let $\tau g \nu^2 + 2^{-k}(\chi^2(A, N(0,1)) + \chi^2(B, N(0,1)))$. Then any SQ algorithm that, given access to a distribution $\mathbf{P}_v^{A,B,p}$ for an unknown $v \in \mathbb{R}^m$, outputs a hypothesis $h : \mathbb{R}^m \to \{\pm 1\}$ such that $\mathbf{Pr}_{(X,Y)\sim \mathbf{P}_v^{A,B,p}}[h(X) \neq Y] < \min(p, 1-p) - 4\sqrt{\tau}$ must either make queries of accuracy better than $2\sqrt{\tau}$ or must make at least $2^{\Omega(m)}\tau/(\chi^2(A, N(0,1)) + \chi^2(B, N(0,1)))$ statistical queries.*

The proof of Proposition 15 is deferred to Appendix C.

3.2. Construction of Univariate Moment-Matching Distributions

In this subsection, we give our univariate approximate moment-matching construction (Proposition 16), which is the key new ingredient to establish our desired SQ lower bound. The moment-matching construction of this subsection (along with its refinement for η close to $1/2$ presented in Section F) is the main technical contribution of this work.

We will use G for the measure of the univariate standard Gaussian distribution $N(0,1)$ and $g(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ for its probability density function.

We will construct two (non-negative, finite) measures $\mathcal{D}_+$ and $\mathcal{D}_-$ on this space with appropriate properties. The main technical result of this section is captured in the following proposition.

Proposition 16 *Let $0 < \epsilon < s < 1$ be real numbers such that s/ϵ is at least a sufficiently large universal constant. Let $0 < \eta < 1/2$. There exist measures $\mathcal{D}_+$ and $\mathcal{D}_-$ over $\mathbb{R}$ and a union J of $d = O(s/\epsilon)$ intervals on $\mathbb{R}$ such that:*

1. *(a) $\mathcal{D}_+ = 0$ on J, and (b) $\mathcal{D}_+/\mathcal{D}_- > (1-\eta)/\eta$ on $J^c := \mathbb{R} \setminus J$.*
2. *All but $\zeta = O(\eta s/\epsilon)\exp(-\Omega(s^4/\epsilon^2))$ of the measure of $\mathcal{D}_-$ lies in J.*
3. *For any $t \in \mathbb{N}$, the distributions $\mathcal{D}_+/\|\mathcal{D}_+\|_1$ and $\mathcal{D}_-/\|\mathcal{D}_-\|_1$ have their first t moments matching those of G within additive error at most $(t+1)!\exp(-\Omega(1/s^2))$.*
4. *(a) $\mathcal{D}_+$ is at most $O(s/\epsilon)$ G, and (b) $\mathcal{D}_-$ is at most $O(s\eta/\epsilon)$ G.*
5. *(a) $\|\mathcal{D}_+\|_1 = \Theta(1)$, and (b) $\|\mathcal{D}_-\|_1 = \Theta(\eta)$.*

Discussion Essentially, in our final construction, $\mathcal{D}_+$ will be proportional to the distribution of X conditioned on $Y = 1$ and $\mathcal{D}_-$ proportional to the distribution of X conditioned on $Y = -1$. Furthermore, the ratio of the probability of $Y = 1$ to the probability of $Y = -1$ will be equal to $\|\mathcal{D}_+\|_1/\|\mathcal{D}_-\|_1$. The Massart PTF that $f(X)$ is supposed to simulate will be -1 on $X \in J$ and 1 elsewhere (thus making it a degree-$2d$ PTF).

We now provide an explanation of the properties established in Proposition 16. Property 1(a) says that Y will deterministically be -1 on J, while property 1(b) says that the ratio between $\mathcal{D}_+$ and $\mathcal{D}_-$ will be greater than $(1-\eta)/\eta$ on the complement of J. This implies that Y amounts to $f(X)$ with Massart noise at most η. Property 2 implies that Y only disagrees with the target PTF with probability roughly ζ, i.e., that the optimal misclassification value OPT will be less than ζ.

Property 3 says that $\mathcal{D}_+$ and $\mathcal{D}_-$, after rescaling, approximately match many moments with the standard Gaussian, which will be necessary in establishing our SQ lower bounds. Property 4 is necessary to show that $\mathcal{D}_+$ and $\mathcal{D}_-$ have relatively small chi-squared norms. Finally, Property5 is necessary to figure out how big the parameter p (i.e., $\Pr[Y = 1]$) should be (approximately).
Proof [of Proposition 16] We will use the following two-parameter family of discrete Gaussians.

Definition 17 (Discrete Gaussian) *For $\sigma \in \mathbb{R}_+$ and $\theta \in \mathbb{R}$, let $G_{\sigma,\theta}$ denote the measure of the "σ-spaced discrete Gaussian distribution". In particular, for each $n \in \mathbb{Z}$, $G_{\sigma,\theta}$ assigns mass $\sigma g(n\sigma+\theta)$ to the point $n\sigma + \theta$.*

Note that $G_{\sigma,\theta}$ is not a probability measure as its total measure is not equal to one. However, it is not difficult to show (see Lemma 18 below) that the measure of $G_{\sigma,\theta}$ is close to one for small $\sigma > 0$, hence can be intuitively thought of as a probability distribution.
Remark: Due to space limitations, the proofs of some intermediate lemmas are given in Appendix D.

The following lemma shows that the moments of $G_{\sigma,\theta}$ approximately match the moments of the standard Gaussian measure G.

Lemma 18 *For $t \in \mathbb{N}, \sigma \geq 0$, and $\theta \in \mathbb{R}$ we have that $\left|\mathbf{E}[G^t_{\sigma,\theta}] - \mathbf{E}[G^t]\right| = t!\, O(\sigma)^t \exp(-\Omega(1/\sigma^2))$.*

The proof of Lemma 18 proceeds by analyzing the Fourier transform of $G_{\sigma,\theta}$ and using the fact that the t^{th} moment of a measure is proportional to the t^{th} derivative of its Fourier transform at 0.

Note that Lemma 18 for $t = 0$ implies the total measure of $G_{\sigma,\theta}$ is $\exp(-\Omega(1/\sigma^2))$ close to one, i.e., for small $\sigma > 0$ $G_{\sigma,\theta}$ can be thought of as a probability distribution.

Definition of the Measures $\mathcal{D}_+$ and $\mathcal{D}_-$ We define our measures as mixtures of discrete Gaussian distributions. This will allow us to guarantee that they nearly match moments with the standard Gaussian. In particular, for a sufficiently large constant $C > 0$, we define:

$$\mathcal{D}_+ := C\,(s/\epsilon) \int_0^\epsilon \frac{1}{s+y}\, G_{s+y,y/2}\, dy\ , \text{ and } \mathcal{D}_- := \eta\,(s/\epsilon) \int_0^\epsilon \frac{1}{s+y}\, G_{s+y,(y+s)/2} dy\ . \quad (1)$$

We will require the following explicit formulas for $\mathcal{D}_+$ and $\mathcal{D}_-$.

Lemma 19 *For all $x \in \mathbb{R}$, we have that $\mathcal{D}_+(x) = C\, g(x)\,(s/\epsilon) \sum_{n\in\mathbb{Z}} \frac{\mathbf{1}\{x\in[ns,ns+(n+1/2)\epsilon]\}}{|n+1/2|}$, where by $\mathbf{1}\{x \in [ns, ns + (n + 1/2)\epsilon]\}$ we denote the indicator function of the event that x is between ns and $ns + (n + 1/2)\epsilon$, even in the case where $n < 0$ and $ns + (n + 1/2)\epsilon < ns$.*

Similarly, we have that $\mathcal{D}_-(x) = \eta\, g(x)\,(s/\epsilon) \sum_{n\in\mathbb{Z}} \frac{\mathbf{1}\{x\in[(n+1/2)s,(n+1/2)s+(n+1/2)\epsilon]\}}{|n+1/2|}$.

Intuition on Definition of $\mathcal{D}_+$ and $\mathcal{D}_-$ We now attempt to provide some intuition regarding the definition of the above measures. We start by noting that each of $\mathcal{D}_+(x)$ and $\mathcal{D}_-(x)$ will have size roughly $g(x)$ on its support. This can be seen to imply on the one hand that the chi-squared divergence of (the normalization of) $\mathcal{D}_\pm$ from the standard Gaussian G is not too large, and on the other hand that $\mathcal{D}_\pm$ roughly satisfy Gaussian concentration bounds.

The critical information to consider is the support of these distributions. Each of the two measures is supported on a union of intervals. Specifically, $\mathcal{D}_+$ is supported on intervals located at the point $n\,s$ of width $|n + 1/2|\epsilon$; and $\mathcal{D}_-$ is supported on intervals located at the point $(n + 1/2)s$ of width $|n + 1/2|\epsilon$. In the case where $\epsilon \ll s$, these intervals will be disjoint for small values of n

(roughly, for $|n| \ll s/\epsilon$). The factor of $C > 0$ difference in the definitions of the two measures will ensure that $\mathcal{D}_+ > \mathcal{D}_-(1-\eta)/\eta$ on their joint support; and once $|n|$ has exceeded a sufficiently large constant multiple of s/ϵ, the intervals will be wide enough that they overlap causing the support to be everything.

In other words, for $|x|$ less than a sufficiently small constant multiple of s^2/ϵ, $\mathcal{D}_+$ and $\mathcal{D}_-$ will be supported on $O(s/\epsilon)$ many intervals and will have disjoint supports. We define J to be the union of the $O(s/\epsilon)$ many intervals in the support of $\mathcal{D}_-$ that are not in the support of $\mathcal{D}_+$. With this definition, we will have that (1) $\mathcal{D}_+$ is equal to zero in J, and (2) $\mathcal{D}_+/\mathcal{D}_-$ is sufficiently large on J^c. Furthermore, since $\mathcal{D}_-$ only assigns mass to J^c for x with $|x| \gg s^2/\epsilon$, we can take $\zeta = \exp(-\Omega(s^4/\epsilon^2))$.

Given the above intuition, we begin the formal proof, starting with moment-matching.

Lemma 20 *For $t \in \mathbb{N}$, the distributions $\mathcal{D}_+/\|\mathcal{D}_+\|_1$ and $\mathcal{D}_-/\|\mathcal{D}_-\|_1$ match the first t moments with the standard Gaussian G to within additive error $t!\, O(s)^t \exp(-1/s^2)$.*

Proof This follows from Lemma 18 by noting that both of these distributions are mixtures of discrete Gaussians with $\sigma = \Theta(s)$. ■

Our next lemma provides approximations to the corresponding L_1 norms.

Lemma 21 *We have that $\|\mathcal{D}_+\|_1 = \Theta(1)$ and $\|D_-\|_1 = \Theta(\eta)$.*

For the rest of the proof, it will be important to analyze the intervals on which $\mathcal{D}_+$ and $\mathcal{D}_-$ are supported. To this end, we start by introducing the following notation.

Definition 22 *For $m \in \mathbb{Z}$, let I_+^m be the interval with endpoints ms and $ms + (m+1/2)\epsilon$, and let I_-^m be the interval with endpoints $(m+1/2)s$ and $(m+1/2)s + (m+1/2)\epsilon$. Additionally, for $x \in \mathbb{R}$, let $n_+(x)$ be the number of integers m such that $x \in I_+^m$, and $n_-(x)$ be the number of integers m such that $x \in I_-^m$.*

The following corollary is an easy consequence of the definition.

Corollary 23 *For $x \in \mathbb{R}$, we have that $x \in I_+^m$ only if $m = x/s + O((|x|+s)\epsilon/s)$. Similarly, $x \in I_-^m$ only if $m = x/s - 1/2 + O((|x|+s)\epsilon/s)$.*

Combining Corollary 23 with the formulas for $\mathcal{D}_+$ and $\mathcal{D}_-$ given in Lemma 19, we have that:

Corollary 24 *For all $x \in \mathbb{R}$ we have that*

$$\mathcal{D}_+(x) = \Theta(Cg(x)(s^2/\epsilon)n_+(x)/(|x|+s)) \text{ and } \mathcal{D}_-(x) = \Theta(\eta g(x)(s^2/\epsilon)n_-(x)/(|x|+s)) \,.$$

Proof This follows from the explicit formulas for $\mathcal{D}_+$ and $\mathcal{D}_-$ given in Lemma 19 along with Corollary 23, which implies that the denominators $|n+1/2|$ are $\Theta((|x|+s)/s)$. ■

We next need to approximate the size of $n_+(x)$ and $n_-(x)$. We have the following lemma.

Lemma 25 *For $x \in \mathbb{R}$, we have $n_+(x), n_-(x) = |x|\,(1/s - 1/(s+\epsilon)) + O(1) = |x|\,\Theta(\epsilon/s^2) + O(1)$.*

Proof We will prove the desired statement for $x \geq 0$ and $n_+(x)$. The other cases follow symmetrically. Note that $x \in I_+^m$ only for non-negative m. For such m, $x \in I_+^m$ if and only if $ms \leq x \leq m(s+\epsilon)+\epsilon/2$. This is the difference in the number of m's for which $ms \leq x$ and the number of m's for which $m(s+\epsilon)+\epsilon/2 < x$. The former is $|x|/s + O(1)$ and the latter is $|x|/(s+\epsilon) + O(1)$. The lemma follows. ■

Lemma 25 implies the following.

Corollary 26 *We have that $n_+(x) \geq 1$ for all $x \gg s^2/\epsilon$.*

Combining Lemma 25 with Corollary 24, we get the following.

Corollary 27 *For all $x \in \mathbb{R}$, we have that* $\mathcal{D}_+(x) = O\left(g(x)(s/\epsilon)\right), \quad \mathcal{D}_-(x) = O\left(g(x)\,\eta\,(s/\epsilon)\right)$.

A combination of Lemma 25 with Corollary 24 also implies that on the support of $\mathcal{D}_+$ the ratio $\mathcal{D}_+/\mathcal{D}_-$ is sufficiently large.

Corollary 28 *If $x \in \mathbb{R}$ is such that $\mathcal{D}_+(x) > 0$, then $\mathcal{D}_+(x)/\mathcal{D}_-(x) > (1-\eta)/\eta$.*

Proof If $\mathcal{D}_+(x) > 0$, then $n_+(x) > 0$. Lemma 25 implies that $n_-(x) = n_+(x) + O(1)$, and therefore $n_-(x)/n_+(x) = O(1)$. Combining this with Corollary 24, we have that $\mathcal{D}_+(x)/\mathcal{D}_-(x) = \Omega(C/\eta)$. For C a sufficiently large universal constant, this implies our result. ■

We also need to show that the intersection of the supports of $\mathcal{D}_+$ and $\mathcal{D}_-$ occurs only for $|x|$ sufficiently large. Specifically, we have the following lemma.

Lemma 29 *For $x \in \mathbb{R}$, it holds that $\min(n_+(x), n_-(x)) > 0$ only if $|x| = \Omega(s^2/\epsilon)$.*

Proof We have that $\min(n_+(x), n_-(x)) > 0$ only if there exist integers m and m' with $x \in I_+^m \cap I_-^{m'}$. By Corollary 23, it must be the case that $|m|, |m'| = O(|x|/s + 1)$. On the other hand, we have that I_+^m is an interval containing the point ms, and $I_-^{m'}$ is an interval containing the point $(m'+1/2)s$. These points must differ by at least $s/2$, and therefore the sum of the lengths of these intervals must be at least $s/2$. On the other hand, these intervals have length $|m+1/2|\epsilon$ and $|m'+1/2|\epsilon$ respectively. Thus, $\min(n_+(x), n_-(x)) > 0$ can only occur if $s/2 = O(|x|\epsilon/s + \epsilon)$, which implies that $x = \Omega(s^2/\epsilon)$, as desired. ■

We define J to be $J \stackrel{\text{def}}{=} \mathbb{R} \setminus \bigcup_{m \in \mathbb{Z}} I_+^m$. We claim that J is a union of $O(s/\epsilon)$ many intervals. Indeed, by Corollary 26, J is an interval $J_0 = [-O(s^2/\epsilon), O(s^2/\epsilon)]$ minus all of the intervals I_+^m that intersect J_0. By Corollary 23, I_+^m intersects J_0 only when $|m| = O(s/\epsilon)$. Thus, J is an interval minus a union of $O(s/\epsilon)$ other intervals. Thus, it is a union of $O(s/\epsilon)$ many intervals.

We can now directly verify the properties of Proposition 16. The definition of J implies that $n_+(x) = 0$ on J, which itself implies that $\mathcal{D}_+(x) = 0$ for $x \in J$. The latter fact combined with Corollary 28 imply Property 1. Lemma 29 implies that the intersection of J^c with the support of $\mathcal{D}_-$ consists only of points x with $|x| = \Omega(s^2/\epsilon)$. This fact and Corollary 27 imply Property 2 (by Gaussian concentration). Property 3 follows from Lemma 20. Property 4 follows from Corollary 27. Property 5 follows from Lemma 21. This completes the proof of Proposition 16. ■

3.3. Putting Everything Together: Proof of Theorem 9

Parameter Setting Recall the parameters in the theorem statement. We have that $\mathrm{OPT} > 0$ and $M \in \mathbb{Z}_+$ are such that $\log(M)/(\log\log(M))^3$ is at least a sufficiently large constant multiple of $\log(1/\mathrm{OPT})$. Moreover, we define a parameter τ which is set to $M^{-\Theta\left(\frac{\log(M)}{\log\log(M)^3}/\log(1/\mathrm{OPT})\right)}$, where the implied constant in the exponent is sufficiently small.

Let $C > 0$ be a sufficiently large universal constant. We define positive integers m and d as follows: $m = \lceil C\log(1/\tau)\rceil$ and $d = \lceil C\sqrt{\log(1/\mathrm{OPT})\log(1/\tau)\log\log(1/\tau)}\rceil$. Observe that

$$\binom{2d+m}{m} \leq m^{2d} = \exp(O(C\sqrt{\log(1/\mathrm{OPT})\log(1/\tau)(\log\log(1/\tau))^3})) \,. \tag{2}$$

We note that if $\log(1/\tau)$ is a sufficiently small constant multiple of $\frac{\log^2(M)}{(\log\log(M))^3\log(1/\mathrm{OPT})}$, then the RHS of (2) is less than M. Thus, by decreasing M if necessary, we can assume that $M = \binom{2d+m}{m}$. Consider the Veronese mapping, denote by $V_{2d} : \mathbb{R}^m \to \mathbb{R}^M$, such that the coordinate functions of V_{2d} are exactly the monomials in m variables of degree at most $2d$.

Hard Distributions We can now formally construct the family of high-dimensional distributions on labeled examples that (1) corresponds to Massart halfspaces, and (2) is SQ-hard to learn. We define univariate measures $\mathcal{D}_+$ and $\mathcal{D}_-$ on $\mathbb{R}$, as given by Proposition 16, with s and ϵ picked so that s^2/ϵ is a sufficiently large constant multiple of $\sqrt{\log(1/\mathrm{OPT})}$ and s/ϵ a sufficiently small constant multiple of d (for example, by taking $s = C^2\sqrt{\log(1/\mathrm{OPT})}/d = \Theta(1/\sqrt{\log(1/\tau)\log\log(1/\tau)})$ and $\epsilon = C^3\sqrt{\log(1/\mathrm{OPT})}/d^2$).

For a unit vector $v \in \mathbb{R}^m$, consider the distribution $\mathbf{P}_v^{\mathcal{D}_+,\mathcal{D}_-,p}$, as in Proposition 15, with $p = \|\mathcal{D}_+\|_1/(\|\mathcal{D}_+\|_1 + \|\mathcal{D}_-\|_1)$. By property 5 of Proposition 16, note that $\min(p, 1-p) = 1-p = \Theta(\eta)$. Our hard distribution is the distribution (X', Y') on $\mathbb{R}^M \times \{\pm 1\}$ obtained by drawing (X, Y) from $\mathbf{P}_v^{\mathcal{D}_+,\mathcal{D}_-,p}$ and letting $X' = V_{2d}(X)$ and $Y' = Y$.

We start by showing that this corresponds to a Massart halfspace (for the proof, see Appendix E).

Claim 30 *The distribution (X', Y') on $\mathbb{R}^M \times \{\pm 1\}$ is a Massart LTF distribution with optimal misclassification error* OPT *and Massart noise rate upper bound of η.*

We are now ready to complete the proof of our SQ lower bound. It is easy to see that finding a hypothesis that predicts Y' given X' is equivalent to finding a hypothesis for Y given X (since $Y = Y'$ and there is a known 1-1 mapping between X and X'). The pointwise bounds on $\mathcal{D}_+$ and $\mathcal{D}_-$, specifically properties 4 and 5 in Proposition 16, imply that $\chi^2(\mathcal{D}_\pm/\|\mathcal{D}_\pm\|_1\,, G) \leq O(s/\epsilon)^2 = \mathrm{polylog}(M)$. The parameter ν in Proposition 15 is $k!\exp(-\Omega(1/s^2)) = \exp(-\Omega(1/s^2))$ after taking k to be a sufficiently small constant multiple $\log(1/s)/s^2$.

Thus, by Proposition 15, in order to output a hypothesis with error smaller than $\min(p, 1-p) = \Theta(\eta)$, any SQ algorithm either needs queries with accuracy better than

$$\nu^2 + 2^{-k}(\chi^2(A, G) + \chi^2(B, G)) = \exp(-\Omega(\log(1/s)/s^2))\mathrm{polylog}(M) < \tau$$

or a number of queries more than $2^{\Omega(m)}\tau(\chi^2(A, G) + \chi^2(B, G)) > 1/\tau$. Therefore, Proposition 15 implies that it is impossible for an SQ algorithm to learn a hypothesis with error better than $\Theta(\eta)$ without either using queries of accuracy better than τ or making at least $1/\tau$ many queries. This completes the proof of Theorem 9.

4. Conclusions

This work gives a super-polynomial SQ lower bound with *near-optimal inapproximability gap* for the fundamental problem of (distribution-free) PAC learning Massart halfspaces. Our lower bound provides strong evidence that known algorithms for this problem are essentially best possible. An obvious open question is whether the constant factor in the $\Omega(\eta)$-term of our lower bound can be improved to $C = 1$ for all $\eta > 0$. Recall that we have shown such a bound for η close to $1/2$. Followup work (Nasser and Tiegel, 2022) showed that this is indeed possible with our techniques via a modification of our one-dimensional construction. This matches known algorithms *exactly*, specifically showing that the error of $\eta + \epsilon$ cannot be improved even for small values of $\eta > 0$. For a list of open questions coming out of our work, see Appendix G.

Acknowledgments

Ilias Diakonikolas was 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. Daniel M. Kane was supported by NSF Medium Award CCF-2107547, NSF Award CCF-1553288 (CAREER), a Sloan Research Fellowship, and a grant from CasperLabs.

References

D. Angluin and P. Laird. Learning from noisy examples. *Mach. Learn.*, 2(4):343–370, 1988.

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

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

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

A. Blum. Machine learning: My favorite results, directions, and open problems. In *44th Symposium on Foundations of Computer Science (FOCS 2003)*, pages 11–14, 2003.

A. Blum, A. M. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. In *37th Annual Symposium on Foundations of Computer Science, FOCS '96*, pages 330–338, 1996.

A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1997.

A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, 2003.

M. Brennan, G. Bresler, S. B. Hopkins, J. Li, and T.Schramm. Statistical query algorithms and low-degree tests are almost equivalent. *CoRR*, abs/2009.06107, 2020. URL https://arxiv.org/abs/2009.06107. Conference version appeared in COLT'21.

N. Bshouty and V. Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002.

S. Chen, F. Koehler, A. Moitra, and M. Yau. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. *CoRR*, abs/2006.04787, 2020. URL https://arxiv.org/abs/2006.04787.

C.-T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. Bradski, A. Y. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 281–288, Cambridge, MA, USA, 2006. MIT Press.

E. Cohen. Learning noisy perceptrons by a perceptron in polynomial time. In *Proceedings of the Thirty-Eighth Symposium on Foundations of Computer Science*, pages 514–521, 1997.

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

I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. *SIAM J. on Comput.*, 39(8):3441–3462, 2010.

I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 73–84, 2017. Full version at http://arxiv.org/abs/1611.03473.

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

I. Diakonikolas, T. Gouleakis, and C. Tzamos. Distribution-independent PAC learning of halfspaces with massart noise. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 4751–4762, 2019.

I. Diakonikolas, D. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. A polynomial time algorithm for learning halfspaces with tsybakov noise. *CoRR*, abs/2010.01705, 2020a. URL https://arxiv.org/abs/2010.01705.

I. Diakonikolas, D. M. Kane, and N. Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and relus under gaussian marginals. *CoRR*, abs/2006.16200, 2020b. URL https://arxiv.org/abs/2006.16200. In NeurIPS'20.

I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with massart noise under structured distributions. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 1486–1513. PMLR, 2020c.

I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with tsybakov noise. *CoRR*, abs/2006.06467, 2020d. URL https://arxiv.org/abs/2006.06467.

I. Diakonikolas, R. Impagliazzo, D. M. Kane, R. Lei, J. Sorrell, and C. Tzamos. Boosting in the presence of massart noise. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1585–1644. PMLR, 2021a. URL http://proceedings.mlr.press/v134/diakonikolas21d.html.

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

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*, 2021c.

V. Feldman. Evolvability from learning algorithms. In Cynthia Dwork, editor, *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, 2008*, pages 619–628. ACM, 2008.

V. Feldman. Distribution-independent evolvability of linear threshold functions. In Sham M. Kakade and Ulrike von Luxburg, editors, *COLT 2011 - The 24th Annual Conference on Learning Theory, June 9-11, 2011*, volume 19 of *JMLR Proceedings*, pages 253–272. JMLR.org, 2011.

V. Feldman. Hardness of proper learning (1988; pitt, valiant). In *Encyclopedia of Algorithms*. 2015.

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

V. Feldman. A general characterization of the statistical query complexity. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 785–830. PMLR, 2017.

V. Feldman, P. Gopalan, S. Khot, and A. Ponnuswami. New results for learning noisy parities and halfspaces. In *Proc. FOCS*, pages 563–576, 2006.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of STOC'13*, pages 655–664, 2013. Full version in Journal of the ACM, 2017.

V. Feldman, W. Perkins, and S. Vempala. On the complexity of random satisfiability problems with planted solutions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC, 2015*, pages 77–86, 2015.

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.

S. Goel, A. Gollakota, and A. R. Klivans. Statistical-query lower bounds via functional gradients. *CoRR*, abs/2006.15812, 2020. URL https://arxiv.org/abs/2006.15812. In NeurIPS'20.

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

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.

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

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

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

A. R. Klivans and P. Kothari. Embedding hard learning problems into gaussian space. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2014*, pages 793–809, 2014.

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.

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

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

R. Nasser and S. Tiegel. Optimal sq lower bounds for learning halfspaces with massart noise. *CoRR*, abs/2201.09818, 2022. URL https://arxiv.org/abs/2201.09818.

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

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

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

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.

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.

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

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

V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics*. Springer-Verlag, Berlin, Heidelberg, 1982. ISBN 0387907335.

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.

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

C. Zhang, J. Shen, and P. Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. *coRR*, abs/2002.04840, 2020. In NeurIPS'20.

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.

Appendix

Appendix A. Additional Related Work

Here we summarize the most relevant literature on learning halfspaces in related noise models. Massart noise lies in between Random Classification noise and the Agnostic model.

Random Classification Noise Random Classification Noise (RCN) (Angluin and Laird, 1988) is the special case of Massart noise where each label is flipped with probability *exactly* $\eta < 1/2$. Halfspaces are known to be efficiently learnable *to optimal accuracy* in the (distribution-independent) PAC model with RCN (Blum et al., 1996, 1997). In fact, it is well-known that any SQ learning algorithm (Kearns, 1998) can be transformed to an RCN noise tolerant learning algorithm — a fact that inherently fails in the presence of Massart noise. Roughly speaking, the ability of the Massart adversary to choose *whether* to flip a given label and, if so, with what probability, makes the algorithmic problem in this model significantly more challenging.

Agnostic Learning The agnostic model (Haussler, 1992; Kearns et al., 1994) is the strongest noise model in the literature, where an adversary is allowed to adversarially corrupt an arbitrary $\mathrm{OPT} < 1/2$ fraction of the labels. In the distribution-independent setting, even *weak* agnostic PAC learning of halfspaces (i.e., obtaining a hypothesis with non-trivial accuracy) is known to be intractable. A long line of work (see, e.g., Guruswami and Raghavendra (2006); Feldman et al. (2006)) has established NP-hardness of weak agnostic *proper* learning. (See Feldman (2015) for a survey on

hardness of proper learning results.) More recently, Daniely (2016) gave super-polynomial lower bounds for *improper* learning, under certain average-case complexity assumptions, and simultaneously established SQ lower bounds for the problem. Concretely, Daniely (2016) showed that no polynomial-time SQ algorithm for agnostically learning halfspaces on $\mathbb{R}^m$ can compute a hypothesis with error $1/2 - 1/m^c$, for some constant $c > 0$, even for instances with optimal error $\mathrm{OPT} = 2^{-\log^{1-\nu}(m)}$, for some constant $\nu \in (0, 1/2)$.

Finally, it is worth noting that learning to *optimal* accuracy in the agnostic model is known to be computationally hard even in the distribution-specific PAC model, and in particular under the Gaussian distribution (Klivans and Kothari, 2014; Goel et al., 2020; Diakonikolas et al., 2020b, 2021c). However, these distribution-specific hardness results are very fragile and do not preclude efficient constant factor approximations. In fact, efficient constant factor approximate learners are known for the Gaussian and other well-behaved distributions (see, e.g., Awasthi et al. (2017); Diakonikolas et al. (2018)).

Prior SQ Lower Bound for Massart Halfspaces Chen et al. (2020) showed an SQ lower bound of $m^{\Omega(\log(1/\epsilon))}$ for learning halfspaces with Massart to error $\mathrm{OPT} + \epsilon$, when OPT is close to $1/2$. Specifically, Chen et al. (2020) observed a connection between SQ learning with Massart noise and the *Correlational Statistical Query (CSQ)* model, a restriction of the SQ model defined in Bshouty and Feldman (2002) (see also Feldman (2008, 2011)). Given this observation, Chen et al. (2020) deduced their SQ lower bound by applying *as a black-box* a previously known CSQ lower bound by Feldman (2011). This approach is inherently limited to exact learning. Establishing lower bounds for approximate learning requires new ideas.

Distribution-Specific Learning We note that $\mathrm{poly}(m, 1/\epsilon)$ time learning algorithms for homogeneous Massart halfspaces with optimal error guarantees have been developed when the marginal distribution on examples is well-behaved (Awasthi et al., 2015, 2016; Zhang et al., 2017; Yan and Zhang, 2017; Zhang et al., 2020; Diakonikolas et al., 2020c,d,a, 2021b). The hardness result obtained in this paper provides additional motivation for such distributional assumptions. As follows from our inapproximability result, without some niceness assumption on the distribution of examples, obtaining even extremely weak relative approximations to the optimal error is hard.

Appendix B. Proof of Lemma 13

Let θ be the angle between v and v'. By making an orthogonal change of variables, we can reduce to the case where $v = (1, 0, \ldots, 0)$ and $v' = (\cos(\theta), \sin(\theta), 0, 0, \ldots, 0)$. Then by definition we have that $\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'}) + 1$ is

$$\int_{\mathbb{R}^m} \left(\frac{A(x_1)A(\cos(\theta)x_1 + \sin(\theta)x_2)g(x_2)g(\sin(\theta)x_1 - \cos(\theta)x_2)}{g(x_1)g(x_2)} \right) g(x_3)\cdots g(x_m)dx_1 \cdots dx_m \, .$$

Noting that the integral over $x_3, \ldots, x_m$ separates out, we are left with

$$\int_{\mathbb{R}^2} \left(\frac{A(x)A(\cos(\theta)x + \sin(\theta)y)g(y)g(\sin(\theta)x - \cos(\theta)y)}{g(x)g(y)} \right) dxdy \, .$$

Integrating over y gives

$$\int \frac{A(x)}{g(x)} \left(\int A(\cos(\theta)x + \sin(\theta)y)g(\sin(\theta)x - \cos(\theta)y)dy \right) dx = \int \frac{A(x)U_{\cos(\theta)}A(x)}{g(x)} dx \, ,$$

where U_t is the Ornstein-Uhlenbeck operator. We will simplify our computations by expressing the various quantities in terms of the eigenbasis for this operator.

In particular, let $h_n(x) = He_n(x)/\sqrt{n!}$ where $He_n(x)$ is the probabilist's Hermite polynomial. We note the following basic facts about them:

1. $\int_{\mathbb{R}} h_i(x)h_j(x)g(x)dx = \delta_{i,j}$.
2. $U_t(h_n(x)g(x)) = t^n h_n(x)g(x)$.

We can now write $A(x)$ in this basis as

$$A(x) = \sum_{n=0}^{\infty} a_n h_n(x)g(x) \ .$$

From this, we obtain that

$$\begin{aligned}\chi^2(A, N(0,1)) &= \int_{\mathbb{R}} \left(\sum_{n=0}^{\infty} a_n h_n(x)g(x)\right)^2 / g(x)dx \\ &= \int_{\mathbb{R}} \sum_{n,m=0}^{\infty} a_n a_m h_n(x)h_m(x)g(x)dx \\ &= \sum_{n=0}^{\infty} a_n^2 \ .\end{aligned}$$

Furthermore, we have that

$$\int_{\mathbb{R}} h_s(x)A(x)dx = \int_{\mathbb{R}} \sum_{n=0}^{\infty} a_n h_s(x)h_n(x)g(x)dx = a_s \ .$$

For $1 \le s \le k$, we have that

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

We therefore have that

$$a_s = \sum_{t=0}^{\lfloor s/2 \rfloor} \left(\frac{\sqrt{s!}(-1)^t x^{s-2t}}{2^t t!(s-2t)!}\right) \mathbf{E}[A^{s-2t}] \ .$$

Note that the above is close to

$$\sum_{t=0}^{\lfloor s/2 \rfloor} \left(\frac{\sqrt{s!}(-1)^t x^{s-2t}}{2^t t!(s-2t)!}\right) \mathbf{E}[G^{s-2t}] = \mathbf{E}[h_s(G)] = 0 \ .$$

In particular, the difference between the two quantities is at most

$$\nu \sum_{t=0}^{\lfloor s/2 \rfloor} \left(\frac{\sqrt{s!}}{2^t t!(s-2t)!}\right) \ .$$

It is easy to see that the denominator is minimized when $t = s/2 - O(\sqrt{s})$. From this it follows that this sum is $2^{O(s)}\,\nu$. Therefore, we have that $a_s = 2^{O(s)}\,\nu$, for $1 \leq s \leq k$. Furthermore, $a_0 = \int A(x)dx = 1$. Thus, we have that

$$\begin{aligned}
\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'}) + 1 &= \int_{\mathbb{R}} \frac{A(x) U_{v\cdot v'} A(x)}{g(x)} dx \\
&= \int_{\mathbb{R}} \left(\sum_{n=0}^{\infty} a_n h_n(x) g(x)\right) \left(\sum_{n'=0}^{\infty} a'_n (v\cdot v')^{n'} h'_n(x) g(x)\right) / g(x) dx \\
&= \int_{\mathbb{R}} \sum_{n,n'=0}^{\infty} a_n a'_n (v\cdot v')^{n'} h_n(x) h'_n(x) g(x) dx \\
&= \sum_{n=0}^{\infty} a_n^2 (v\cdot v')^n \\
&= 1 + \sum_{n=1}^{k} a_n^2 (v\cdot v')^n + \sum_{n=k+1}^{\infty} a_n^2 (v\cdot v')^n \,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
|\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'})| &\leq O(\nu^2) \sum_{n=1}^{k} 2^{O(n)} |v\cdot v'|^n + |v\cdot v'|^{k+1} \sum_{n=0}^{\infty} a_n^2 \\
&\leq \nu^2 + |v\cdot v'|^{k+1} \chi^2(A, N(0,1)) \,.
\end{aligned}$$

This completes our proof.

Appendix C. Proof of Proposition 15

We will use the following standard fact:

Fact 31 *For any constant $c > 0$ there exists a set S of $2^{\Omega_c(m)}$ unit vectors in $\mathbb{R}^m$ such that any pair $u, v \in S$, with $u \neq v$, satisfies $|u \cdot v| < c$.*

In fact, an appropriate size set of random unit vectors satisfies the above statement with high probability. We note that Diakonikolas et al. (2017) made use of a similar statement, albeit with different parameters.

The proof proceeds as follows: We start by defining a related hypothesis testing problem $\mathcal{H}$ and show that $\mathcal{H}$ efficiently reduces to our learning (search) problem. We then leverage Lemma 13 and Fact 31 to prove an SQ lower bound for $\mathcal{H}$, which in turns implies an SQ lower bound for our learning task.

Let S be a set of $2^{\Omega(m)}$ unit vectors in $\mathbb{R}^m$ whose pairwise inner products are at most a sufficiently small universal constant c. (In fact, any constant $c < 1/2$ suffices.) By Fact 31, such a set is guaranteed to exist. Given S, our hypothesis testing problem is defined as follows.

Definition 32 (Hidden Direction Hypothesis Testing Problem) *In the context of Definition 14, the testing problem $\mathcal{H}$ is the task of distinguishing between: (i) the distribution $\mathbf{P}_v^{A,B,p}$, for v randomly chosen from S, and (ii) the distribution G' on $\mathbb{R}^m \times \{\pm 1\}$, where for $(X,Y) \sim G'$ we have*

that X is a standard Gaussian $G \sim N(0, I)$, and Y is independently 1 with probability p and -1 with probability $1-p$.

We claim that $\mathcal{H}$ efficiently reduces to our learning task. In more detail, any SQ algorithm that computes a hypothesis h satisfying $\mathbf{Pr}_{(X,Y)\sim \mathbf{P}_v^{A,B,p}}[h(X) \neq Y] < \min(p, 1-p) - 4\sqrt{\tau}$ can be used as a black-box to distinguish between $\mathbf{P}_v^{A,B,p}$, for v randomly chosen from S, and G'. Indeed, suppose we have such a hypothesis h. Then, with one additional query to estimate the $\mathbf{Pr}[h(X) \neq Y]$, we can distinguish between $\mathbf{P}_v^{A,B,p}$, for v randomly chosen from S, and G' for the following reason: For any function h, we have that $\mathbf{Pr}_{(X,Y)\sim G'}[h(X) \neq Y] \geq \min(p, 1-p)$.

It remains to prove that solving the hypothesis testing problem $\mathcal{H}$ is impossible for an SQ algorithm with the desired parameters. We will show this using Lemma 8.

More specifically, we need to show that for $u, v \in S$ we have that $|\chi_{G'}(\mathbf{P}_v^{A,B,p}, \mathbf{P}_u^{A,B,p})|$ is small. Since $G', \mathbf{P}_v^{A,B,p}$, and $\mathbf{P}_u^{A,B,p}$ all assign $Y = 1$ with probability p, it is not hard to see that

$$\begin{aligned}\chi_{G'}(\mathbf{P}_v^{A,B,p}, \mathbf{P}_u^{A,B,p}) &= p\, \chi_{(G'|Y=1)}\left((\mathbf{P}_v^{A,B,p} \mid Y=1), (\mathbf{P}_u^{A,B,p} \mid Y=1)\right) + \\ &\quad (1-p)\, \chi_{(G'|Y=-1)}\left((\mathbf{P}_v^{A,B,p} \mid Y=-1), (\mathbf{P}_u^{A,B,p} \mid Y=-1)\right) \\ &= p\, \chi_G(\mathbf{P}_v^A, \mathbf{P}_u^A) + (1-p)\, \chi_G(\mathbf{P}_v^B, \mathbf{P}_u^B)\,.\end{aligned}$$

By Lemma 13, it follows that

$$\chi_{G'}(\mathbf{P}_v^{A,B,p}, \mathbf{P}_u^{A,B,p}) \leq \nu^2 + 2^{-k}(\chi^2(A, N(0,1)) + \chi^2(B, N(0,1))) = \tau\,.$$

A similar computation shows that

$$\chi_{G'}(\mathbf{P}_v^{A,B,p}, \mathbf{P}_v^{A,B,p}) = \chi^2(\mathbf{P}_v^{A,B,p}, G') \leq \chi^2(A, N(0,1)) + \chi^2(B, N(0,1))\,.$$

An application of Lemma 8 for $\gamma = \gamma' = \tau$ and $\beta = \chi^2(A, N(0,1)) + \chi^2(B, N(0,1))$ completes the proof.

Appendix D. Proof of Proposition 16

In this section, we prove the lemmas and claims that are used in the proof of Proposition 16.

D.1. Proof of Lemma 18

We consider the Fourier transform of $G_{\sigma,\theta}$. Note that $G_{\sigma,\theta}$ is the pointwise product of G with a mesh of delta-functions. Therefore, its Fourier transform is the convolution of their Fourier transforms. The Fourier transform of G is $\sqrt{2\pi}G$. The Fourier transform of the net of delta-functions $f(\xi) = \sum_{n\in\mathbb{Z}} \delta(\xi - n/\sigma)e^{2\pi i\theta\xi}$. Thus, we have that the Fourier transform of $G_{\sigma,\theta}$ at ξ is

$$\sum_{n\in\mathbb{Z}} \sqrt{2\pi} g(\xi + n/\sigma) e^{-2\pi i n\theta/\sigma}\,.$$

The t^{th} moment of a pseudodistribution is proportional to the value of the t^{th} derivative of its Fourier transform at $\xi = 0$. For G, this is $\sqrt{2\pi}g^{(t)}(0)$. For $G_{\sigma,\theta}$, it is equal to this term plus

$$\sum_{n\in\mathbb{Z}, n\neq 0} \sqrt{2\pi} g^{(t)}(n/\sigma) e^{-2\pi i n\theta/\sigma}\,.$$

Computing the derivative of g using Cauchy's integral formula (integrating around a circle of radius $1/(2\sigma)$ centered at n/σ), we find that

$$|g^{(t)}(n/\sigma)| = t! O(\sigma)^t \exp(-\Omega(n/\sigma)^2).$$

Taking a sum over n yields our result.

D.2. Proof of Lemma 19

To prove the lemma, we unravel the definition of the discrete Gaussian to find that:

$$\begin{aligned}
\mathcal{D}_+(x) &= Cs/\epsilon \int_0^\epsilon \frac{\sum_{n\in\mathbb{Z}}(s+y)g(n(s+y)+y/2)\delta(x-(n(s+y)+y/2))}{s+y}dy \\
&= Cs/\epsilon \sum_{n\in\mathbb{Z}} \int_0^\epsilon g(n(s+y)+y/2)\delta(x-(n(s+y)+y/2))dy \\
&= Cs/\epsilon \sum_{n\in\mathbb{Z}} \int_0^\epsilon g((n+1/2)y+ns)\delta(x-((n+1/2)y+ns))dy \\
&= Cg(x)s/\epsilon \sum_{n\in\mathbb{Z}} \frac{\mathbf{1}\{x\in[ns, ns+(n+1/2)\epsilon]\}}{|n+1/2|} \, .
\end{aligned}$$

The calculation for $\mathcal{D}_-(x)$ is similar.

D.3. Proof of Lemma 21

Applying Lemma 18 with $t = 0$, we get that $\|G_{s+y,y/2}\|_1 = \Theta(1)$. Thus, working from the definition, we find that

$$\begin{aligned}
\|\mathcal{D}_+\|_1 &= C\,(s/\epsilon) \int_0^\epsilon \frac{\Theta(1)}{s+y}dy \\
&= C\,(s/\epsilon) \int_0^\epsilon \Theta(1/s)dy \\
&= \Theta(1) \, .
\end{aligned}$$

The proof for $\mathcal{D}_-$ follows similarly.

Appendix E. Proof of Claim 30

For a unit vector $v \in \mathbb{R}^m$, let $g_v : \mathbb{R}^m \to \{\pm 1\}$ be defined as $g_v(x) = -1$ if and only if $v \cdot x \in J$, where J is the union of intervals in the construction of Proposition 16. Note that g_v is a degree-$2d$ PTF on $\mathbb{R}^m$, since g_v is a $(2d+1)$-piecewise constant function of $v \cdot x$. Therefore, there exists some LTF $L : \mathbb{R}^M \to \{\pm 1\}$ such that $g_v(x) = L(V_{2d}(x))$ for all $x \in \mathbb{R}^m$.

Note that our hard distribution returns (X', Y') with $Y' = L(X')$, unless it picked a sample corresponding to a sample of $\mathcal{D}_-$ coming from J^c, which happens with probability at most $\zeta <$ OPT. Additionally, suppose that our distribution returned a sample with $X' = V_{2d}(X)$, for some $X \in \mathbb{R}^m$. By construction, conditioned on this event, we have that $Y' = 1$ with probability proportional to $\mathcal{D}_+(v \cdot X)$, and $Y' = -1$ with probability proportional to $\mathcal{D}_-(v \cdot X)$. We note that if

$L(V_{2d}(X)) = 1$, then $v \cdot X \notin J$; so, by Proposition 16 property 1(b), this ratio is at least $1 - \eta : \eta$. On the other hand, if $L(V_{2d}(X)) = -1$, then $v \cdot X \in J$, so $\mathcal{D}_+(v \cdot X) = 0$. This implies that the pointwise probability of error $\eta(X')$ is at most η, completing the proof of the claim.

Appendix F. Obtaining Optimal Error: The Case of Large η

In this section, we refine the construction of the previous subsections to obtain a sharp lower bound of $\eta - o_\eta(1)$, when η is close to $1/2$. Here the term $o_\eta(1)$ goes to zero when η approaches $1/2$. Specifically, we show:

Theorem 33 (Sharp SQ Hardness of Massart Halfspaces for Large η) *Let $\mathrm{OPT} > 0$ and $M \in \mathbb{Z}_+$ be such that $\log(M)/(\log\log(M))^3$ is at least a sufficiently large constant multiple of $\log(1/\mathrm{OPT})$. Let $c > 0$ be any parameter such that $c \gg \sqrt{1/2 - \eta}$. There exists a parameter*

$$\tau \stackrel{\mathrm{def}}{=} M^{-\Omega_c\left(\frac{\log(M)}{\log\log(M)^3}/\log(1/\mathrm{OPT})\right)}$$

such that no SQ algorithm can learn the class of halfspaces on $\mathbb{R}^M$ in the presence of η-Massart noise, where $\mathrm{OPT} < \eta \leq 1/2$, within error better than $1/2 - c$ using at most $1/\tau$ queries of tolerance τ. This holds even if the optimal classifier has misclassification error at most OPT.

Conceptually, Theorem 33 provides evidence that *even the constant factor* (of 1) in the error guarantee (of $\eta+\epsilon$) achieved by the Massart learner of Diakonikolas et al. (2019) cannot be improved in general.

The proof of Theorem 33 proceeds along the same lines as the proof of Theorem 9. The main difference is in the choice of the one-dimensional moment-matching distributions. For this, we use a construction that is qualitatively similar (though somewhat more sophisticated) to that used in the proof of Section 3.2.

Specifically, for some carefully chosen parameter $C > 0$ (to be determined), we define the positive measures:

$$\mathcal{D}_+ := C\,(s/\epsilon) \int_0^\epsilon \frac{G_{s+y,y/2}}{s+y} dy \;,$$

and

$$\mathcal{D}_- := (s/\epsilon) \int_0^\epsilon \frac{G_{s+y,(y+s)/2}}{s+y} dy \;.$$

As a refinement of Corollary 24, we obtain the following.

Corollary 34 *For all $x \in \mathbb{R}$, we have that*

$$\mathcal{D}_+(x) = Cg(x)(s^2/\epsilon)n_+(x)/(|x|+1)(1 + O(\epsilon/(|x|+1)))$$

and

$$\mathcal{D}_-(x) = g(x)(s^2/\epsilon)n_-(x)/(|x|+1)(1 + O(\epsilon/(|x|+1))) \;.$$

Proof This follows from the explicit formulas for $\mathcal{D}_+$ and $\mathcal{D}_-$ (Lemma 19) along with the fact that for $x \in I_m^{\pm}$, $1/|m|$ and $1/|m+1/2|$ are $s/(|x|+1)(1 + O(\epsilon/(|x|+1)))$. ■

Using the above corollary, we obtain the following.

Corollary 35 *For all $x \in \mathbb{R}$, we have that $\mathcal{D}_+(x)/\mathcal{D}_-(x) = C\,q(x)(1+O(\epsilon/(|x|+1)))$, where $q(x) = n_+(x)/n_-(x)$ is a rational number with numerator and denominator at most $O(|x|/s+1)$.*

We want to guarantee that for all $x \in \mathbb{R}$ it holds that $\mathcal{D}_+(x)/\mathcal{D}_-(x) \notin [\eta/(1-\eta), (1-\eta)/\eta]$. We note that this condition automatically holds for $|x|$ less than a sufficiently small constant multiple of s^2/ϵ, as in this range we have that $\min(n_+(x), n_-(x)) = 0$. For points x outside this range, we have that $\mathcal{D}_+(x)/\mathcal{D}_-(x) = C\,q(x)(1+O(\epsilon/s)^2)$. Furthermore, since $|n_+(x) - n_-(x)| \leq 1$, the latter implies that in this range $\mathcal{D}_+(x)/\mathcal{D}_-(x)$ is always one of:

- $C(1+O(\epsilon/s)^2)$,
- $C(1+1/m)(1+O(\epsilon/s)^2)$, for some integer m,
- $C(1-1/m)(1+O(\epsilon/s)^2)$, for some integer m.

We will arrange that this quantity is always in the appropriate range by picking the parameter C, so that for some well chosen m_0 we have that

$$C(1-1/m_0)(1+O(\epsilon/s)^2) \leq \eta/(1-\eta), \text{ and } C(1-1/(m_0+1))(1-O(\epsilon/s)^2) \geq (1-\eta)/\eta\ .$$

If the above holds, it is easy to see that $\mathcal{D}_+(x)/\mathcal{D}_-(x)$ will never be in the range $[\eta/(1-\eta), (1-\eta)/\eta]$ for any value of x. In order to arrange this, we set C to satisfy

$$C(1-1/m_0)(1+O(\epsilon/s)^2) = \eta/(1-\eta)\ .$$

In order for the second condition to hold, it must be the case that

$$\left(\frac{1-1/(m_0+1)}{1-1/m_0}\right)(1-O(\epsilon/s)^2) > ((1-\eta)/\eta)^2 = 1+O(1/2-\eta)\ .$$

For the latter to be true, it must hold that $1/m_0^2$ is at least a sufficiently large constant multiple of $(1/2-\eta)+(\epsilon/s)^2$, or that m_0 is at most a sufficiently small constant multiple of $\min(s/\epsilon, \sqrt{1/2-\eta})$.

In particular, if we take m_0 to be at most a sufficiently small constant multiple of $1/\sqrt{1/2-\eta}$ and ensure that ϵ/s is sufficiently small, this construction can be made to work with $C = 1+1/m_0$.

We then let J be the set of points $x \in \mathbb{R}$ for which $\mathcal{D}_-(x) > \mathcal{D}_+(x)$. It is easy to see that $J = \{x : m_0 \geq n_-(x) > n_+(x)\}$, and from this it can be seen that J is a union of $O(m_0 s/\epsilon)$ intervals. As before, $\mathcal{D}_+$ and $\mathcal{D}_-$ approximately match many moments with a Gaussian and the mass of $\mathcal{D}_+$ on J and $\mathcal{D}_-$ on J^c are both supported on points x such that $|x| \geq \Omega(s^2/\epsilon)$, and thus have mass $\exp(-\Omega(s^4/\epsilon^2))$.

Furthermore, we have that $\mathcal{D}_-(x)/\mathcal{D}_+(x) > (1-\eta)/\eta$ for $x \in J$ and $\mathcal{D}_+(x)/\mathcal{D}_-(x) > (1-\eta)/\eta$ for $x \in J^c$. Therefore, the appropriate hidden-direction distribution is a degree-$O(m_0 s/\epsilon)$ PTF with at most η Massart noise.

Finally, it is not hard to see that $\|\mathcal{D}_+\|_1/\|\mathcal{D}_-\|_1 = 1+1/m_0$. Therefore, by following the arguments of Section 3.3 mutatis-mutandis, it follows that for any constant $\eta < 1/2$ it is SQ-hard to learn an LTF with η-Massart noise to error better than $1/2 - c$ for any $c \gg \sqrt{1/2-\eta}$, even when OPT is almost polynomially small in the dimension. This completes the proof of Theorem 33.

Appendix G. Conclusions and Future Work

This work gives a super-polynomial Statistical Query (SQ) lower bound with *near-optimal inapproximability gap* for the fundamental problem of (distribution-free) PAC learning Massart halfspaces. Our lower bound provides strong evidence that known algorithms for this problem are essentially best possible. An obvious open question is whether the constant factor in the $\Omega(\eta)$-term of our lower bound can be improved to the value $C = 1$ for all $\eta > 0$. Recall that we have shown such a bound for η close to $1/2$. Followup work (Nasser and Tiegel, 2022) showed that this is indeed possible with our techniques via a modification of our one-dimensional construction. This matches known algorithms *exactly*, specifically showing that the error of $\eta + \epsilon$ cannot be improved even for small values of $\eta > 0$.

Interestingly, SQ lower bounds are the *only known* evidence of hardness for our Massart halfspace learning problem. Via a recent reduction (Brennan et al., 2020), our SQ lower bound implies a similar low-degree polynomial testing lower bound for the problem. An interesting open question is to prove similar hardness results against families of convex programming relaxations (obtained, e.g., via the Sum-of-Squares framework). Such lower bounds would likely depend on the underlying optimization formulation of the learning problem.

A related question is whether one can establish reduction-based computational hardness for learning halfspaces in the presence of Massart noise. Daniely (2016) gave such a reduction for the (much more challenging) problem of agnostically learning halfspaces, starting from the problem of strongly refuting random XOR formulas. It currently remains unclear whether the latter problem is an appropriate starting point for proving hardness in the Massart model. That said, obtaining reduction-based hardness for learning Massart halfspaces is left as an interesting open problem.