Information-Computation Tradeoffs for Learning Margin Halfspaces with Random Classification Noise

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

Jelena Diakonikolas JELENA@CS.WISC.EDU
University of Wisconsin-Madison

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

Puqian Wang PWANG333@WISC.EDU
University of Wisconsin-Madison

Nikos Zarifis ZARIFIS@WISC.EDU
University of Wisconsin-Madison

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study the problem of PAC learning γ-margin halfspaces with Random Classification Noise. We establish an information-computation tradeoff suggesting an inherent gap between the sample complexity of the problem and the sample complexity of computationally efficient algorithms. Concretely, the sample complexity of the problem is $\widetilde{\Theta}(1/(\gamma^2\epsilon))$. We start by giving a simple efficient algorithm with sample complexity $\widetilde{O}(1/(\gamma^2\epsilon^2))$. Our main result is a lower bound for Statistical Query (SQ) algorithms and low-degree polynomial tests suggesting that the quadratic dependence on $1/\epsilon$ in the sample complexity is inherent for computationally efficient algorithms. Specifically, our results imply a lower bound of $\widetilde{\Omega}(1/(\gamma^{1/2}\epsilon^2))$ on the sample complexity of any efficient SQ learner or low-degree test.

Keywords: PAC Learning, Halfspaces, Margin, Random Classification Noise, SQ Model

1. Introduction

This work studies the efficient learnability of halfspaces with a margin in the presence of random label noise. Before we present our contributions, we provide the necessary background. A halfspace or Linear Threshold Function (LTF) is any Boolean-valued function $h : \mathbb{R}^d \to \{\pm 1\}$ of the form $h(\mathbf{x}) = \mathrm{sign}\,(\mathbf{w} \cdot \mathbf{x} - \theta)$, where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector and $\theta \in \mathbb{R}$ is the threshold. 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. The problem of learning halfspaces with a margin — i.e., under the assumption that no example lies too close to the separating hyperplane — is a textbook problem in machine learning, whose history goes back to the Perceptron algorithm of Rosenblatt (1958). Here we study the problem of PAC learning margin halfspaces in the presence of Random Classification Noise (RCN) (Angluin and Laird, 1988).

Before we describe the noisy setting (the focus of this work), we recall the basics in the realizable PAC model (Valiant, 1984) (i.e., when the labels are consistent with the target concept). We will henceforth assume that the threshold is $\theta = 0$, which is well-known to be no loss of generality. The setup is as follows: there is an unknown distribution $(\mathbf{x}, y) \sim D$

on $\mathbb{S}^{d-1} \times \{\pm 1\}$, where $\mathbb{S}^{d-1}$ is the unit sphere on $\mathbb{R}^d$, such that $y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$ for some $\mathbf{w}^* \in \mathbb{R}^d$ with $\|\mathbf{w}^*\|_2 = 1$. The margin assumption means that the marginal distribution of D on the examples $\mathbf{x}$, denoted by $D_{\mathbf{x}}$, puts no probability mass on points with distance less than $\gamma \in (0,1)$ from the separating hyperplane $\mathbf{w}^* \cdot \mathbf{x} = 0$; that is, we have that $\mathbf{Pr}_{\mathbf{x} \sim D_{\mathbf{x}}}[|\mathbf{w}^* \cdot \mathbf{x}| < \gamma] = 0$. The parameter γ is called the margin of the target halfspace.

In this context, the learning algorithm is given as input a desired accuracy $\epsilon > 0$ and a training set $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ of i.i.d. samples from D. The goal is to output a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ whose misclassification error $\mathrm{err}_{0-1}^D(h) := \mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y]$ is at most ϵ with high probability[1]. The aforementioned setting is well-understood. First, it is known that the sample complexity of the learning problem, ignoring computational considerations, is $\Theta(1/(\gamma^2\epsilon))$; see, e.g., Shalev-Shwartz and Ben-David (2014)[2]. Moreover, the Perceptron algorithm is a computationally efficient PAC learner achieving this sample complexity. (This follows, e.g., by combining the mistake bound of $O(1/\gamma^2)$ of the online Perceptron with the online-to-PAC conversion in Littlestone (1989).) That is, *in the realizable setting, there exists a computationally efficient learner for margin halfspaces achieving the optimal sample complexity* (within constant factors).

The high-level question that serves as the motivation for this work is the following:

Can we develop "similarly efficient" algorithms in the presence of label noise, and specifically in the (most) basic model of Random Classification Noise?

By the term "similarly efficient" above, we mean that we would like a polynomial-time algorithm with near-optimal sample complexity (up to logarithmic factors).

This problem appears innocuous and our initial efforts focused towards obtaining such an algorithm. After several failed attempts, we established an information-computation tradeoff strongly suggesting that such an algorithm does not exist. We next describe our setting in more detail.

Learning Margin Halfspaces with RCN The RCN model (Angluin and Laird, 1988) is the most basic model of random label noise. In this model, the label of each example is independently flipped with probability exactly η, where $0 < \eta < 1/2$ is a noise parameter. Since its introduction, RCN has been studied extensively in learning theory from both an information-theoretic and an algorithmic standpoint. One of the early fundamental results in this field was given by Kearns (1998), who showed that any Statistical Query (SQ) algorithm can be transformed into a PAC learning algorithm that is tolerant to RCN. This transformation preserves statistical and computational efficiency *within polynomial factors*.

We return to our problem of PAC learning margin halfspaces with RCN. The setup is very similar to the one above. The only difference is that the labels are now perturbed by RCN with noise rate η (see Definition 1). As a result, the optimal misclassification error is equal to η, and the goal is to find a hypothesis that with high probability satisfies $\mathrm{err}_{0-1}^D(h) \leq \eta + \epsilon$. A closely related objective would be to approximate the target halfspace, i.e., the function $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$, within any desired accuracy $\epsilon' > 0$. It is well-known (and easy to derive) that the two goals are essentially equivalent, up to rescaling the parameter

1. Throughout this introduction, we will take the failure probability to be a small constant, say 1/10.
2. We are implicitly assuming that $d = \Omega(1/\gamma^2)$; otherwise, a sample complexity bound of $\widetilde{O}(d/\epsilon)$ follows from standard VC-dimension arguments.

ϵ by a factor of $(1-2\eta)$. In this paper, we phrase our results for the misclassification error with respect to the observed labels.

In this context, the sample complexity of PAC learning margin halfspaces with RCN is equal to $\widetilde{\Theta}(1/((1-2\eta)\gamma^2\epsilon))$. This bound can be derived, e.g., from the work of Massart and Nedelec (2006). (That is, the sample complexity of the RCN learning problem is essentially the same as in the realizable case — assuming η is bounded from $1/2$ — within logarithmic factors.) On the algorithmic side, a number of works, starting with Bylander (1994), developed polynomial sample and time algorithms for this learning task. Specifically, Bylander (1994) developed a careful adaptation of the Perceptron algorithm for this purpose. Subsequently, Blum et al. (1997) pointed out that an SQ version of the Perceptron algorithm coupled with Kearns' reduction immediately implies the existence of an efficient RCN learner (see also Cohen (1997) for a closely related work). More recently, in a related context, Diakonikolas et al. (2019) pointed out that a simple convex surrogate loss can be used for this purpose (see also Diakonikolas et al. (2020) for a related setting).

The preceding paragraph might suggest that the RCN version of the problem is fully resolved. The catch is that *all known algorithms for the problem require sample complexities that are polynomially worse than the information-theoretic minimum.* Specifically, for all known polynomial-time algorithms, the dependence of the sample complexity on the inverse of the accuracy parameter ϵ is *at least quadratic* — while the information-theoretic minimum scales near-linearly with $1/\epsilon$. It is thus natural to ask whether a computationally efficient algorithm with (near-)optimal sample complexity exists. This leads us to the following question:

Is the existing gap between the sample complexity of known efficient algorithms and the information-theoretic sample complexity inherent?

In this paper, we resolve the above question in the affirmative for a broad class of algorithms — specifically, for all Statistical Query algorithms and low-degree polynomial tests.

1.1. Our Results

The following definition summarizes our setting.

Definition 1 (PAC Learning Margin Halfspaces with RCN) *Let D be a distribution over $\mathbb{S}^{d-1} \times \{\pm 1\}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$, and let $\mathbf{w}^* \in \mathbb{S}^{d-1}$. Let $\gamma \in (0,1)$ and $\eta \in (0, \frac{1}{2})$. For each sample $(\mathbf{x}, y) \sim D$, the following assumptions both hold:*

(A_1) The unit vector $\mathbf{w}^*$ satisfies the γ-margin condition, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[|\mathbf{w}^* \cdot \mathbf{x}| < \gamma] = 0$.

(A_2) For each point $\mathbf{x} \in \mathbb{S}^{d-1}$, the corresponding label y satisfies: with probability $1-\eta$, $y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$; otherwise, $y = -\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$.

Given i.i.d. samples from D, the goal of the learner is to output a hypothesis h that with high probability satisfies $\mathrm{err}^D_{0-1}(h) := \mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \eta + \epsilon$.

While our definition applies to homogeneous halfspaces, the case of general halfspaces, i.e., functions of the form $y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} - t)$, can easily be reduced to the homogeneous case by increasing the dimension by one, i.e., writing $y = \mathrm{sign}((\mathbf{w}^*, -t) \cdot (\mathbf{x}, 1))$. By rescaling these

vectors and letting $\mathbf{w}' := (\mathbf{w}^*, t)/\sqrt{1+t^2}$, and $\mathbf{x}' := (\mathbf{x}, 1)/\sqrt{2}$, we have a homogeneous halfspace $y = \mathrm{sign}(\mathbf{w}' \cdot \mathbf{x}')$ in $d+1$ dimensions with margin $\gamma/\sqrt{2(1+t^2)} \geq \gamma/2$ equivalent to our original problem. By the homogeneity assumption, it follows that the assumption that the examples lie on the unit sphere is also generic (up to scaling). We also recall that it can be assumed without loss of generality that the noise rate η is known to the algorithm (Angluin and Laird, 1988).

To the best of our knowledge, prior to our work, the best known sample-complexity upper bound of an efficient algorithm for our problem was $\widetilde{O}(1/(\gamma^4\epsilon^2))$ (Diakonikolas et al., 2019). Our first result is a computationally efficient learner with sample complexity $\widetilde{O}(1/(\gamma^2\epsilon^2))$.

Theorem 2 (Algorithmic Result) *There exists an algorithm that draws $N = \widetilde{O}(1/(\gamma^2\epsilon^2))$ samples, runs in time $\mathrm{poly}(N, d)$ and learns γ-margin halfspaces up to misclassification error $\eta + \epsilon$ with probability at least $9/10$.*

See Theorem 9 for a more detailed formal statement. While the above sample bound does not improve on the ϵ-dependence over prior algorithmic results, it does improve the dependence on the margin parameter γ quadratically — nearly matching the information-theoretic lower bound (within logarithmic factors). An independent and contemporaneous work by Kontonis et al. (2023) obtained a similar $O_\eta(1/(\gamma^2\epsilon^2))$ sample complexity result for learning γ-margin halfspaces with RCN in polynomial time, using a different algorithm and techniques.

Our second and main result is an information-computation tradeoff suggesting that the quadratic dependence in $1/\epsilon$ is inherent for polynomial-time algorithms. Formally, we establish such a tradeoff in the Statistical Query model and (via a known reduction) for low-degree polynomial tests.

Statistical Query (SQ) Model Before we state our main result, we recall the basics of the SQ model (Kearns, 1998). Instead of drawing samples from the input distribution, SQ algorithms are given query access to the distribution via the following oracle:

Definition 3 (STAT Oracle) *Let D be a distribution on $\mathbb{R}^d$. A statistical query is a bounded function $f : \mathbb{R}^d \to [-1, 1]$. For tolerance $\tau > 0$ of the statistical query, the $\mathrm{STAT}(\tau)$ oracle responds to the query f with a value v such that $|v - \mathbf{E}_{\mathbf{x}\sim D}[f(\mathbf{x})]| \leq \tau$.*

We note that other oracles have been considered in the literature, in particular VSTAT (Definition 10); our lower bound also holds with respect to these oracles.

An SQ lower bound for a learning problem Π is typically of the following form: any SQ algorithm for Π must either make at least q queries or it makes at least one query with small tolerance τ. When simulating a statistical query in the standard PAC model (by averaging i.i.d. samples to approximate expectations), the number of samples needed for a τ-accurate query can be as high as $\Omega(1/\tau^2)$. Thus, we can intuitively interpret an SQ lower bound as a tradeoff between runtime of $\Omega(q)$ and a sample complexity of $\Omega(1/\tau^2)$.

We are now ready to state our SQ lower bound:

Theorem 4 (SQ Lower Bound) *For any constant $c > 0$, any SQ algorithm that learns γ-margin halfspaces on the unit sphere in the presence of RCN with $\eta = 1/3$ to error $\eta + \epsilon$ requires at least $2^{(1/\gamma)^{\Omega(c)}}$ queries or makes at least one query with tolerance $O(\epsilon\gamma^{1/4-c})$.*

The reader is referred to Theorem 11 for a more detailed formal statement. The intuitive interpretation of our result is that any (sample simulation of an) SQ algorithm for the class of γ-margin halfspaces with RCN either draws at least $\Omega(1/(\gamma^{1/2-c}\epsilon^2))$ samples or requires at least $2^{(1/\gamma)^{\Omega(c)}}$ time. That is, for sufficiently small ϵ (namely, $\epsilon \leq \gamma^{3/2+c}$), the computational sample complexity of the problem (in the SQ model) is polynomially higher than its information-theoretic sample complexity. See Theorem 43 for the implications to low-degree polynomial tests.

Finally, we note that SQ lower bounds have been previously obtained for the (more challenging) problem of learning halfspaces with bounded (Massart) noise in a variety of regimes (Diakonikolas and Kane, 2022; Diakonikolas et al., 2022a; Nasser and Tiegel, 2022). Importantly, all these previous results make essential use of the bounded noise model and do not apply in the RCN setting.

1.2. Our Techniques

Upper Bound Our algorithmic approach is quite simple: we use projected subgradient descent applied to the leaky ReLU loss with parameter η, as was done in previous work (Diakonikolas et al., 2019). However, our analysis never explicitly makes a connection to minimizing the leaky ReLU loss; for our arguments, this loss is irrelevant. Instead, we make a novel connection between the (sub)gradient field of the leaky ReLU loss and the disagreement between how the vector $\mathbf{w}$ at which the subgradient is evaluated and an optimal vector $\mathbf{w}^*$ would classify points. Through this connection, we leverage the regret analysis of projected subgradient descent to obtain a novel regret bound on the disagreement probability $\mathbf{Pr}[\mathrm{sign}(\mathbf{w}_t \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})]$, where $\mathbf{w}_t$ are the iterates of the algorithm. The obtained bound decomposes into three terms: (i) the standard regret term, which is bounded by choosing the algorithm iteration count to be sufficiently high, (ii) an error term bounded by the subgradient norm of empirical leaky ReLU at $\mathbf{w}^*$, and (iii) an error term that corresponds to the uniform convergence error of the disagreement function $\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\}$. The latter two terms are then bounded choosing the sample size to be sufficiently high, yet bounded by $\widetilde{O}(1/(\epsilon^2\gamma^2))$.

SQ Lower Bound To prove our SQ lower bound, we bound from below the SQ dimension of the problem; using standard results (Feldman et al., 2017) this implies the desired lower bound guarantees. Bounding the SQ dimension from below amounts to establishing the existence of a large set of distributions whose pairwise correlations are small. Inspired by the technique of Diakonikolas et al. (2017), we achieve this by selecting our distributions to be random rotations of a single distribution, each behaving in a standard way in all but one critical direction. To ensure the necessary margin property, we make the $\mathbf{x}$-marginals of our distribution uniform over the hypercube in $d \ll 1/\gamma$ dimensions — as opposed to Gaussian-like (as in Diakonikolas et al. (2017)).

The distributions we consider are quite simple. We define $f_{\mathbf{v}}(\mathbf{x}) = \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} - t)$, where $\mathbf{v}$ is a randomly chosen Boolean-valued vector and the threshold t is chosen so that the probability that $\mathbf{v} \cdot \mathbf{x} > t$ is of the order of ϵ. We then let $\mathbf{x}$ be the uniform distribution over the hypercube and let $y = f_{\mathbf{v}}(\mathbf{x})$ with probability $2/3$ and $-f_{\mathbf{v}}(\mathbf{x})$ otherwise. By picking many different vectors $\mathbf{v}$, we get many different LTFs. We claim that there exist many of these LTFs whose pairwise correlations (with respect to the distribution where

$\mathbf{x}$ is independent over the hypercube and y is independent of $\mathbf{x}$) are small, as long as the corresponding defining vectors $\mathbf{v}$ and $\mathbf{v}'$ have small inner product (see Lemma 13 and Lemma 14). Intuitively, this should hold because (i) both distributions are already ϵ-close to the base distribution, and (ii) when $\mathbf{u}$ and $\mathbf{v}$ are nearly orthogonal, $f_{\mathbf{v}}(\mathbf{x})$ and $f_{\mathbf{u}}(\mathbf{x})$ are nearly independent of each other.

To analyze this inner product, we use a Fourier analytic approach. First, we note that the sizes of the individual Fourier coefficients of $f_{\mathbf{u}}$ and $f_{\mathbf{v}}$ can be computed using Kravchuk polynomials (see Claim 18). This allows us to show that they do not have too much Fourier mass in low degrees. Second, we note that when taking the inner product of the degree-k parts of the Fourier transforms of $f_{\mathbf{u}}$ and $f_{\mathbf{v}}$, we will have large amounts of cancellation of terms, particularly when $|\mathbf{u} \cdot \mathbf{v}|$ is small or when k is large (see Claim 20 and Claim 21). The size of the remaining term after the cancellation can be written in terms of another Kravchuk polynomial, which we can bound. A careful analysis of all of the relevant terms gives us the necessary correlation bounds which imply our main result.

1.3. Notation

For $n \in \mathbb{Z}_+$, we use $[n]$ to denote the set $\{1, \ldots, n\}$. We use small boldface characters for vectors and capital bold characters for matrices. For $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_2 \coloneqq (\sum_{i=1}^d \mathbf{x}_i^2)^{1/2}$ denotes the ℓ_2-norm of $\mathbf{x}$. We use $\mathbf{x} \cdot \mathbf{y}$ for the inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\mathcal{B}_d = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$ to denote the unit centered Euclidean ball in $\mathbb{R}^d$; when the dimension is clear from the context, we omit it from the subscript. We use $\mathbb{1}\{A\}$ to denote the indicator function of A; equal to one if A is a true statement, and equal to zero otherwise.

We use $\mathbf{E}_{x \sim D}[x]$ for the expectation of the random variable x according to the distribution 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. For $(\mathbf{x}, y)$ distributed according to D, we denote by $D_{\mathbf{x}}$ the marginal distribution of $\mathbf{x}$ and by D_y the marginal distribution of y. We denote by $\mathcal{U}_d$ the uniform distribution over $\{\pm 1\}^d$.

2. Computationally Efficient Learning Algorithm

In this section, we give the algorithm establishing Theorem 2. We start by providing some intuition for our algorithm and its analysis. We then formally state the algorithm and bound its sample complexity and runtime. Due to space constraints, some of the technical details and proofs are deferred to Appendix A.

Leaky ReLU, its subgradient, and intuition. The Leaky ReLU loss function with parameter $\eta \in (0, 1/2)$ is defined by

$$\mathrm{LeakyReLU}_\eta(z) := (1-\eta) z \mathbb{1}\{z \geq 0\} + \eta z \mathbb{1}\{z < 0\} \,. \tag{1}$$

While the Leaky ReLU has been used as a convex surrogate for margin halfspace classification problems, this is not the core of our approach: we never argue about minimizing the expected leaky ReLU loss nor that its minimizer is a good classifier. Instead, we rely on the following vector-valued function $\mathbf{g}_\eta : \mathbb{R}^d \to \mathbb{R}^d$

$$\mathbf{g}_\eta(\mathbf{w}; \mathbf{x}, y) = \frac{1}{2}[(1-2\eta)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) - y]\mathbf{x} \,. \tag{2}$$

When $\mathbf{x}$ and y are clear from the context, we omit them and instead simply write $\mathbf{g}_\eta(\mathbf{w})$. The connection between $\mathbf{g}_\eta$ and LeakyReLU is that $\mathbf{g}_\eta(\mathbf{w})$ is a subgradient of LeakyReLU$(\mathbf{w})$; see, e.g., Diakonikolas et al. (2019, Lemma 2.1). However, this connection is not important for our analysis and we do not make any explicit use of the leaky ReLU function itself. What we do rely on is the following key observation.

Proposition 5 *For any* $\mathbf{w}, \bar{\mathbf{w}} \in \mathbb{R}^d$,

$$(\mathbf{g}_\eta(\mathbf{w}) - \mathbf{g}_\eta(\bar{\mathbf{w}})) \cdot (\mathbf{w} - \bar{\mathbf{w}}) = (1-2\eta)\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \mathrm{sign}(\bar{\mathbf{w}}\cdot\mathbf{x})\}(|\mathbf{w}\cdot\mathbf{x}| + |\bar{\mathbf{w}}\cdot\mathbf{x}|).$$

Proof By a direct calculation,

$$\begin{aligned}(\mathbf{g}_\eta(\mathbf{w}) - \mathbf{g}_\eta(\bar{\mathbf{w}})) \cdot (\mathbf{w} - \bar{\mathbf{w}}) &= \frac{1-2\eta}{2}\big((\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) - \mathrm{sign}(\bar{\mathbf{w}}\cdot\mathbf{x}))(\mathbf{w}\cdot\mathbf{x} - \bar{\mathbf{w}}\cdot\mathbf{x})\big) \\ &= \frac{1-2\eta}{2}\big(|\mathbf{w}\cdot\mathbf{x}| + |\bar{\mathbf{w}}\cdot\mathbf{x}| \\ &\qquad - (\mathbf{w}\cdot\mathbf{x})\,\mathrm{sign}(\bar{\mathbf{w}}\cdot\mathbf{x}) - (\bar{\mathbf{w}}\cdot\mathbf{x})\,\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\big).\end{aligned}$$

In the last expression, the term in the parentheses is zero when the signs of $\mathbf{w}\cdot\mathbf{x}$ and $\bar{\mathbf{w}}\cdot\mathbf{x}$ agree; otherwise it is equal to $2(|\mathbf{w}\cdot\mathbf{x}| + |\bar{\mathbf{w}}\cdot\mathbf{x}|)$, leading to the claimed identity. ■

In particular, recalling that $\mathbf{w}^*$ is the weight vector of the target halfspace (see Definition 1), Proposition 5 implies that

$$(\mathbf{g}_\eta(\mathbf{w}) - \mathbf{g}_\eta(\mathbf{w}^*)) \cdot (\mathbf{w} - \mathbf{w}^*) = (1-2\eta)\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\}(|\mathbf{w}\cdot\mathbf{x}| + |\mathbf{w}^*\cdot\mathbf{x}|). \quad (3)$$

In other words, the inner product $(\mathbf{g}_\eta(\mathbf{w}) - \mathbf{g}_\eta(\mathbf{w}^*)) \cdot (\mathbf{w} - \mathbf{w}^*)$ is proportional to the Boolean function $\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\}$ that indicates disagreement between $\mathbf{w}$ and $\mathbf{w}^*$ in how they classify points $\mathbf{x}$. In particular, if we argue that $\mathbf{E}_{\mathbf{x}\sim D_\mathbf{x}}[\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\}] = \mathbf{Pr}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})] \leq \bar{\epsilon}$ for some $\mathbf{w}$, then we can immediately conclude that the misclassification error of $\mathbf{w}$ is $\eta + (1-2\eta)\bar{\epsilon}$, due to Definition 1, Item (A_2). Thus, for $\bar{\epsilon} = \frac{\epsilon}{1-2\eta}$, the misclassification error is $\eta + \epsilon$. This is the approach that we take.

To carry out the analysis, given $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots \mathbf{x}^{(N)}$ drawn i.i.d. from $D_\mathbf{x}$, we use

$$\widehat{\mathbf{Pr}}_N(\mathbf{w}) := (1/N)\sum_{i=1}^{N} \mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}^{(i)}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}^{(i)})\} \quad (4)$$

to denote the empirical probability of disagreement between $\mathbf{w}$ and $\mathbf{w}^*$.

Projected subgradient descent and disagreement regret. Our Algorithm 1 is the simple projected subgradient descent, applied to the subgradient of the empirical leaky ReLU function, which given a sample $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$ drawn i.i.d. from D is defined by

$$\hat{\mathbf{g}}_N(\mathbf{w}) := (1/N)\sum_{i=1}^{N} \mathbf{g}_\eta(\mathbf{w}; \mathbf{x}^{(i)}, y^{(i)}). \quad (5)$$

To utilize standard regret bounds for (projected) subgradient descent, we first observe that $\mathbf{g}_\eta$ is bounded for all $\mathbf{w} \in \mathbb{R}^d$. As a consequence, $\hat{\mathbf{g}}_N$ admits the same upper bound.

Input: $\epsilon > 0$, $\gamma \in (0,1)$, $\eta \in (0,1/2)$, i.i.d. sample $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$ from D, any $\mathbf{w}_0 \in \mathcal{B}$
Let $T = \lceil \frac{16(1-\eta)^2}{\gamma^2\epsilon^2} - 1 \rceil$, $\mu = \frac{2}{(1-\eta)\sqrt{T+1}}$ **for** $t = 0 : T-1$ **do**
  $\hat{\mathbf{g}}_N(\mathbf{w}_t) = \frac{1}{2N}\sum_{i=1}^N \left((1-2\eta)\mathrm{sign}(\mathbf{w}_t \cdot \mathbf{x}^{(i)}) - y^{(i)}\right)\mathbf{x}^{(i)}$
  $\mathbf{w}_{t+1} = \mathrm{proj}_{\mathcal{B}}(\mathbf{w}_t - \mu\hat{\mathbf{g}}_N(\mathbf{w}_t))$, where $\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le 1\}$
end
return $\{\mathbf{w}_0, \ldots, \mathbf{w}_T\}$

Algorithm 1: PAC Learner for Margin Halfspaces with RCN

Claim 6 *Given any $\mathbf{w} \in \mathbb{R}^d$ and any $(\mathbf{x}, y) \in \mathbb{S}^{d-1} \times \{-1, 1\}$, $\|\mathbf{g}_\eta(\mathbf{w}; \mathbf{x}, y)\|_2 \le 1 - \eta$. As a consequence, given any set of points $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathbb{S}^{d-1} \times \{-1, 1\}$, $i \in [N]$, $\|\hat{\mathbf{g}}_N(\mathbf{w})\|_2 \le 1 - \eta$.*

Proof By the definition of $\mathbf{g}_\eta$ from Equation (2), we have

$$\|\mathbf{g}_\eta(\mathbf{w}; \mathbf{x}, y)\|_2 = (1/2)\left|(1-2\eta)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) - y\right| \|\mathbf{x}\|_2 \le (1/2)\left|(1-2\eta) + 1\right| = 1 - \eta,$$

where we have used that $\mathrm{sign}(\cdot) \in \{-1, 1\}$, $y \in \{-1, 1\}$ and $\|\mathbf{x}\|_2 = 1$. The bound on $\|\hat{\mathbf{g}}_N(\mathbf{w})\|_2$ follows immediately from this bound, by its definition and Jensen's inequality. ■

The following lemma provides what can be interpreted as a regret bound for the disagreement probability $\mathbf{Pr}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})]$. We refer to is as the "disagreement regret" and bound it using the regret analysis of projected subgradient descent, combined with Equation (3) and Definition 1, Item (A_1). Its proof is provided in Appendix A.

Lemma 7 *Consider Algorithm 1. There exists $t \in \{0, \ldots T\}$ such that*

$$\mathbf{Pr}[\mathrm{sign}(\mathbf{w}_t \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})] \le E_1 + E_2 + E_3,$$

where $E_1 = \frac{2(1-\eta)}{(1-2\eta)\gamma\sqrt{T+1}}$, $E_2 = \frac{2}{(1-2\eta)\gamma}\|\hat{\mathbf{g}}_N(\mathbf{w}^*)\|_2$, and $E_3 = \frac{1}{T+1}\sum_{t=0}^T \left[\mathbf{Pr}[\mathrm{sign}(\mathbf{w}_t \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})] - \widehat{\mathbf{Pr}}_N(\mathbf{w}_t)\right]$.

In Lemma 7, error E_1 is simply the empirical regret, which can be bounded by choosing the number of iterations T in Algorithm 1 to be sufficiently high. Errors E_2 and E_3 determine the sample complexity of our algorithm, and are dealt with in what follows.

Bounding the required number of samples. We now show that the errors E_2 and E_3 can be controlled by choosing a sufficiently large sample size N. We then combine everything we have shown so far to state our main result on upper bounds in Theorem 9.

Lemma 8 *Let E_2 and E_3 be defined as in Lemma 7. For any $\bar{\epsilon} > 0, \delta > 0$, if $N = \Omega(\frac{d}{\bar{\epsilon}} + \frac{\eta}{(1-2\eta)^2\bar{\epsilon}^2\gamma^2})\log(\frac{1}{\delta}))$, then with probability at least $1 - \delta$ we have $E_2 + E_3 \le \frac{\bar{\epsilon}}{2}$.*

We are now ready to state and prove our main upper bound result.

Theorem 9 *Let D be a distribution on pairs $(\mathbf{x}, y) \in \mathbb{S}^{d-1} \times \{\pm 1\}$ as in Definition 1. Then, there is an algorithm (Algorithm 1) that for any given $\epsilon, \delta \in (0,1)$ uses $N = O(\frac{d(1-2\eta)}{\epsilon} + \frac{\eta}{\epsilon^2\gamma^2})\log(\frac{1}{\delta}))$ samples, runs in time $O(\frac{Nd}{\epsilon^2\gamma^2})$ and learns γ-margin halfspaces corrupted with η-RCN up to error $\eta + \epsilon$, with probability at least $1 - \delta$.*

Proof Applying Lemma 7 and Lemma 8, we have that for $N = O(\frac{d(1-2\eta)}{\epsilon} + \frac{\eta}{\epsilon^2\gamma^2})\log(\frac{1}{\delta}))$, with probability at least $1-\delta$, there exists $t \in \{0,\dots,T\}$ in Algorithm 1 such that

$$\mathbf{Pr}[\mathbb{1}\{\mathrm{sign}(\mathbf{w}_t \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\}] \leq E_1 + \frac{\epsilon}{2(1-2\eta)},$$

where $E_1 = \frac{2(1-\eta)}{(1-2\eta)\gamma\sqrt{T+1}}$. Hence, to ensure that $\mathbf{Pr}[\mathrm{sign}(\mathbf{w}_t \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})] \leq \frac{\epsilon}{1-2\eta}$, it suffices to choose $T \geq \frac{16(1-\eta)^2}{\gamma^2\epsilon^2} - 1$, which is what Algorithm 1 does. The bound on the runtime is then simply $O(TNd)$, as the complexity of each iteration is dominated by the computation of $\hat{\mathbf{g}}_N$, which takes $O(Nd)$ time. By Definition 1, Item (A_2), such a $\mathbf{w}_t$ misclassifies points $\mathbf{x}$ drawn from D with probability $\eta+\epsilon$.

What we have shown so far is that at least one of the vectors $\mathbf{w}_0,\dots,\mathbf{w}_T$ output by Algorithm 1 attains the target misclassification error $\eta+\epsilon$, but we have not specified which one. The appropriate vector can be determined by drawing a fresh sample $\{(\tilde{\mathbf{x}}^{(i)}, \tilde{y}^{(i)})\}_{i=1}^{N'}$ of size $N' = O(\log(\frac{T}{\epsilon})\log(\frac{1}{\delta})) = O(\log(\frac{1}{\epsilon\gamma})\log(\frac{1}{\delta}))$ and selecting the vector $\mathbf{w}_t \in \{\mathbf{w}_0,\dots,\mathbf{w}_T\}$ with minimum empirical misclassification error $\frac{1}{N'}\sum_{i=1}^{N'} \mathbb{1}\{\mathrm{sign}(\mathbf{w}_t \cdot \tilde{\mathbf{x}}^{(i)}) \neq \tilde{y}^{(i)}\}$. Clearly, this additional step does not negatively impact the sample complexity or the runtime stated in Theorem 9. The standard analysis for this part is provided in Appendix A. ■

Removing the Dependence on d from the Sample Complexity In Theorem 9, the sample complexity N depends on d via the term $\frac{d}{\epsilon}\log(\frac{1}{\delta})$, which comes from the VC dimension of $O(d)$ that appears when bounding the error term E_3. The dependence on d can be avoided and replaced by $1/\gamma^2$, using standard dimension-reduction; see Appendix A.

Low-Noise Regime When the noise parameter η is equal to zero (i.e., in the realizable setting), the sample complexity of the problem is $\Theta(\frac{1}{\gamma^2\epsilon})$ and is achievable via the classical Perceptron algorithm. Based on the result of Theorem 9 and with the dimension reduction discussed in the previous paragraph, we recover this optimal sample complexity with our algorithm not only for $\eta = 0$, but also whenever $\eta = O(\epsilon)$.

3. SQ Lower Bound For Learning Margin Halfspaces with RCN

In this section, we establish our SQ lower bound result (Theorem 4) and its associated implication for low-degree polynomial tests. In addition to the STAT oracle defined in the introduction, we also consider the VSTAT oracle, defined below.

Definition 10 (VSTAT Oracle) *Let D be a distribution on $\mathbb{R}^d$. A statistical query is a bounded function $f : \mathbb{R}^d \to [-1,1]$. For $t > 0$, the* VSTAT(t) *oracle responds to the query f with a value v such that $|v - \mathbf{E}_{\mathbf{x}\sim D}[f(\mathbf{x})]| \leq \tau$, where $\tau = \max\left(1/t, \sqrt{\mathbf{Var}_{\mathbf{x}\sim D}[f(\mathbf{x})]/t}\right)$.*

Our main SQ lower bound result is stated in the following theorem.

Theorem 11 (Main SQ Lower Bound) *Fix $c \in (0,1/2)$. Any SQ algorithm that learns the class of γ-margin halfspaces on $\mathbb{S}^{d-1}$ in the presence of RCN with $\eta = 1/3$ within misclassification error $\eta+\epsilon$ either requires queries of accuracy better than $O(\epsilon\gamma^{1/4-c/2})$, i.e., queries to* STAT($O(\epsilon\gamma^{1/4-c/2})$) *or* VSTAT($O(\gamma^{c-1/2}/\epsilon^2)$)*, or needs to make at least $2^{\Omega(\gamma^{-c})}$ statistical queries.*

3.1. Proof of Theorem 11

To prove the theorem, we construct a family of non-homogeneous margin halfspaces with RCN such that any SQ learner requires the desired complexity. This result can be translated to an SQ lower bound for homogeneous halfspaces with almost as good margin (see paragraph after Definition 1). The family of halfspaces that we construct is supported on $\{\pm 1\}^d$ and has margin $\gamma = \Omega(1/d)$. Note that we can straightforwardly extend this construction to a higher dimensional space (by setting the values of the new coordinates of points $\mathbf{x} \sim D_{\mathbf{x}}$ to zero). Hence, our construction directly implies a similar SQ lower bound for γ-margin halfspaces on the unit sphere $\mathbb{S}^{d-1}$ for any $d \gg 1/\gamma$.

Fix $\epsilon \in (0, 1/2)$. For $\mathbf{v} \in \{\pm 1\}^d$, let $f_{\mathbf{v}}(\mathbf{x}) = \mathbb{1}\{\mathbf{v} \cdot \mathbf{x} - t \geq 0\}$ and choose $t \in \mathbb{R}$ so that $\mathbf{Pr}[f_{\mathbf{v}}(\mathbf{x}) > 0] = 2\epsilon$. We define the distribution $D_{\mathbf{v}}$ over $\{\pm 1\}^d \times \{0, 1\}$ as follows. We choose the marginal distribution of $\mathbf{x}$, denoted by $(D_{\mathbf{v}})_{\mathbf{x}}$, to be the uniform distribution over the set $\{\pm 1\}^d$. For each $\mathbf{x}$, we couple the random variable y by setting $\mathbf{Pr}[y = f_{\mathbf{v}}(\mathbf{x})|\mathbf{x}] = 1 - \eta$ and $\mathbf{Pr}[y \neq f_{\mathbf{v}}(\mathbf{x})|\mathbf{x}] = \eta$. Let $A_{\mathbf{v}}$ be the conditional distribution of $D_{\mathbf{v}}$ given $y = 1$ and let $B_{\mathbf{v}}$ be the conditional distribution of $D_{\mathbf{v}}$ given $y = 0$. We denote by $A_{\mathbf{v}}(\mathbf{x})$ and $B_{\mathbf{v}}(\mathbf{x})$ the pmf of $A_{\mathbf{v}}$ and $B_{\mathbf{v}}$, respectively. Moreover, we denote by $\mathcal{U}_d(\mathbf{x})$ the pmf of $\mathcal{U}_d$. We first give a closed form expression for the pmf of $A_{\mathbf{v}}$ and $B_{\mathbf{v}}$. Its proof can be found in Appendix B.

Claim 12 *It holds* $A_{\mathbf{v}}(\mathbf{x}) = \frac{\eta + (1-2\eta) f_{\mathbf{v}}(\mathbf{x})}{\eta + (1-2\eta)\,\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]} \mathcal{U}_d(\mathbf{x})$ *and* $B_{\mathbf{v}}(\mathbf{x}) = \frac{1 - \eta - (1-2\eta) f_{\mathbf{v}}(\mathbf{x})}{1 - \eta - (1-2\eta)\,\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]} \mathcal{U}_d(\mathbf{x})$.

Fix $\mathbf{v}, \mathbf{u} \in \{\pm 1\}^d$. We associate each $\mathbf{v}$ and $\mathbf{u}$ to a distribution $D_{\mathbf{v}}$ and $D_{\mathbf{u}}$, constructed as above. The following lemma provides explicit bounds on the correlation between the distributions $D_{\mathbf{v}}$ and $D_{\mathbf{u}}$, and its proof can be found in Appendix B.

Recall that the pairwise correlation of two distributions with pmfs D_1, D_2 with respect to a distribution with pmf D is defined as $\chi_D(D_1, D_2) + 1 := \sum_{x \in \mathcal{X}} D_1(x) D_2(x)/D(x)$ (see Definition 30). We have the following lemma:

Lemma 13 *Let D_0 be a product distribution over $\mathcal{U}_d \times \{0, 1\}$, where $\mathbf{Pr}_{(\mathbf{x},y)\sim D_0}[y = 1] = \mathbf{Pr}_{(\mathbf{x},y)\sim D_{\mathbf{v}}}[y = 1]$. We have $\chi_{D_0}(D_{\mathbf{v}}, D_{\mathbf{u}}) \leq 2(1 - 2\eta)(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x}) f_{\mathbf{u}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]\,\mathbf{E}[f_{\mathbf{u}}(\mathbf{x})])$ and $\chi^2(D_{\mathbf{v}}, D_0) \leq (1 - 2\eta)(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]^2)$.*

To bound the correlation between $f_{\mathbf{v}}, f_{\mathbf{u}}$, we use the following key lemma whose proof can be found in Section 3.2.

Lemma 14 (Correlation Bound) *Let $\mathbf{v}, \mathbf{u} \in \{\pm 1\}^d$ and $f_{\mathbf{v}}(\mathbf{x}) = \mathbb{1}\{\mathbf{v} \cdot \mathbf{x} \geq 2t - d\}$. Choose t so that $\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})] = \epsilon$ for $\epsilon \in (0, 1)$. Assume that $|\mathbf{v} \cdot \mathbf{u}| \leq O(d/\mathrm{polylog}(d/\epsilon))$. Then there is an absolute constant $C > 0$ such that $\mathbf{E}[f_{\mathbf{v}}(\mathbf{x}) f_{\mathbf{u}}(\mathbf{x})] \leq C \log^2(d/\epsilon)\epsilon^2 |\mathbf{v} \cdot \mathbf{u}|/d + \epsilon^2$.*

The following fact states that there exists a large set of almost orthogonal vectors in $\{\pm 1\}^d$. Its proof can be found in Appendix B.

Fact 15 *Let $d \in \mathbb{Z}_+$. Let $0 < c < 1/2$. There exists a collection $\mathcal{S}$ of $2^{\Omega(d^c)}$ vectors in $\{\pm 1\}^d$, such that any pair $\mathbf{v}, \mathbf{u} \in \mathcal{S}$, with $\mathbf{v} \neq \mathbf{u}$, satisfies $|\mathbf{v} \cdot \mathbf{u}| < d^{1/2+c}$.*

By Lemma 14, we get that for any two vectors $\mathbf{v}, \mathbf{u} \in \{\pm 1\}^d$, we have that $\chi_{D_0}(D_{\mathbf{v}}, D_{\mathbf{u}}) \leq C \log^2(d/\epsilon)(1 - 2\eta)\epsilon^2 |\mathbf{v} \cdot \mathbf{u}|/d$ and $\chi^2(D_{\mathbf{v}}, D_0) \leq (1 - 2\eta)4\epsilon$ for some $C > 0$.

By Fact 15, for any $0 < c < 1/2$, there exists a set S of $2^{\Omega(d^c)}$ vectors such that for any two vectors $\mathbf{v}, \mathbf{u} \in S$, we have that $|\mathbf{v} \cdot \mathbf{u}|/d \leq d^{c-1/2}$. Denote by $\mathcal{D}$ the set containing the distributions $D_{\mathbf{v}}$ for each $\mathbf{v} \in S$ and let D_0 be a product distribution over $\mathcal{U}_d \times \{0,1\}$, where $\mathbf{Pr}_{(\mathbf{x},y)\sim D_0}[y = 1] = \eta + 2(1-2\eta)\epsilon$. By standard results (see Lemma 32), for the decision problem $\mathcal{B}(\mathcal{D}, D_0)$ of distinguishing between a distribution in $\mathcal{D}$ and the reference distribution D_0, the following holds: any SQ algorithm either requires a query of tolerance at most $O(\epsilon d^{c/2-1/4})$ or needs to make at least $2^{\Omega(d^c)}$ many queries.

It remains to reduce the testing (decision) problem above to the learning problem. This is standard, but we include it here for completeness. Suppose we have access to an algorithm $\mathcal{A}$ that solves the RCN problem with margin γ to excess error $\epsilon' > 0$. For the distributions in the set $\mathcal{D}$ of hard distributions, the margin γ is $1/(2d)$. We describe how algorithm $\mathcal{A}$ can be used to solve the testing problem $\mathcal{B}(\mathcal{D}, D_0)$. If the underlying distribution were $D_{\mathbf{v}}$ for some $\mathbf{v} \in \{\pm 1\}^d$, then algorithm $\mathcal{A}$ would produce a hypothesis h such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D_{\mathbf{v}}}[h(\mathbf{x}) \neq y] \leq \eta + \epsilon'$. If the underlying distribution were D_0 — i.e., the one with independent labels — then the best attainable error would be $\eta + 2(1-2\eta)\epsilon$ (achieved by the constant hypothesis $h(\mathbf{x}) \equiv 1$). Therefore, for $\eta = 1/3$ and $\epsilon' = \epsilon/4$, algorithm $\mathcal{A}$ solves the decision problem $\mathcal{B}(\mathcal{D}, D_0)$. This completes the proof of Theorem 11.

3.2. Proof of Lemma 14

We start with some definitions of the Fourier transform over the uniform distribution on the hypercube. For a subset $T \subseteq [d]$ and $\mathbf{x} \in \{\pm 1\}^d$, we denote $\chi_T(\mathbf{x}) = \prod_{i \in T} \mathbf{x}_i$. For a function f from $\{\pm 1\}^d$, let $\widehat{f}(T) = \mathbf{E}[f(\mathbf{x})\chi_T(\mathbf{x})]$. For a boolean function $f : \{\pm 1\}^d \mapsto \{0,1\}$, we can write f in the Fourier basis as follows, $f(\mathbf{x}) = \sum_{T \subseteq [d]} \mathbf{E}[f(\mathbf{z})\chi_T(\mathbf{z})]\chi_T(\mathbf{x}) = \sum_{T \subseteq [d]} \widehat{f}(T)\chi_T(\mathbf{x})$. Note that $\chi_T(\mathbf{x})$ is an orthonormal polynomial basis under the uniform distribution over $\{\pm 1\}^d$; this means that $\mathbf{E}[\chi_T(\mathbf{x})\chi_{T'}(\mathbf{x})] = \delta_{T,T'}$, where δ is the Kronecker delta. Further, for any two functions $f_1, f_2 : \{\pm 1\}^d \mapsto \{0,1\}$, we have that $\mathbf{E}[f_1(\mathbf{x})f_2(\mathbf{x})] = \sum_{T \subseteq [d]} \widehat{f_1}(T)\widehat{f_2}(T)$. We also define the normalized Kravchuk polynomials as follows.

Definition 16 (Normalized Kravchuk Polynomials) *For $n, a, b \in \mathbb{Z}_+$ with $0 \leq a, b \leq n$, the normalized Kravchuk polynomial $\mathcal{K}(n, a, b)$ is defined by*

$$\mathcal{K}(n,a,b) := \frac{1}{\binom{n}{a}\binom{n}{b}} \sum_{A \subseteq [n], B \subseteq [n], |A|=a, |B|=b} (-1)^{|A \cap B|} .$$

One can think of the normalized Kravchuk polynomial $\mathcal{K}(n, a, b)$ as the expectation over the random subsets A, B of size a and b of -1 to the number of elements in the intersection of A and B. Note that for $n, a, b \in \mathbb{Z}_+$ $\mathcal{K}(n, a, b) = \mathcal{K}(n, b, a)$ and $|\mathcal{K}(n, a, b)| = |\mathcal{K}(n, a, n-b)|$. Furthermore, by definition, it also holds that $|\mathcal{K}(n, a, b)| \leq 1$. The proof of the following lemma can be found in Appendix B.

Lemma 17 *Let $d, m, k \in \mathbb{Z}$. Then the following hold:*

1. *For $k \leq d/2$, it holds $|\mathcal{K}(d, m, k)| \leq e^k 2^{3k}((kd^{-1})^{k/2} + (|d/2 - m|d^{-1})^k)$.*
2. *If $k \leq d/2$ and $|d/2 - m| \leq d/4$, then $|\mathcal{K}(d, m, k)| = \exp(-\Omega(k))$.*

For a vector $\mathbf{v} \in \{\pm 1\}^d$, we define the boolean function $f_{\mathbf{v}}(\mathbf{x}) = \mathbb{1}\{\mathbf{v} \cdot \mathbf{x} \geq t\}$. We first calculate the Fourier transform of $f_{\mathbf{v}}(\mathbf{x})$. The proof can be found in Appendix B.

Claim 18 (Fourier Coefficients) *Fix vector $\mathbf{v} \in \{\pm 1\}^d$ and let $f_{\mathbf{v}}(\mathbf{x}) = \mathbb{1}\{\mathbf{v} \cdot \mathbf{x} \geq t\}$. For $T \subseteq [d]$, we have that the Fourier coefficient of f at $\chi_T(\mathbf{x})$, i.e., $\widehat{f}(T)$, is given by*

$$\widehat{f}(T) = \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})\chi_T(\mathbf{x})] = \chi_T(\mathbf{v})(-1)^{|T|}2^{-d}\sum_{s=t}^{d}\binom{d}{s}\mathcal{K}(d,s,|T|) \, .$$

Proof of Lemma 14 Using Claim 18, we have that

$$\begin{aligned}\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})f_{\mathbf{u}}(\mathbf{x})] &= \sum_{T\subseteq[d]} \widehat{f}_{\mathbf{v}}(T)\widehat{f}_{\mathbf{u}}(T) = \sum_{k=0}^{d}\left(2^{-d}\sum_{s=t}^{d}\binom{d}{s}\mathcal{K}(d,k,s)\right)^2 \sum_{T\subseteq[d],|T|=k}\chi_T(\mathbf{v})\chi_T(\mathbf{u}) \\ &= \sum_{k=0}^{d}\binom{d}{k}\left(2^{-d}\sum_{s=t}^{d}\binom{d}{s}\mathcal{K}(d,k,s)\right)^2 \mathcal{K}(d,m,k) \, ,\end{aligned}$$

where m is the number of components for which $\mathbf{v}, \mathbf{u}$ agree. We proceed by bounding each term of this sum. To this end, we denote $R_k = \binom{d}{k}\left(2^{-d}\sum_{s=t}^{d}\binom{d}{s}\mathcal{K}(d,k,s)\right)^2 \mathcal{K}(d,k,m)$. First note that $R_0 = \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]^2$; to see this, observe that $R_0 = \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})\chi_{\emptyset}(\mathbf{x})]^2 = \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]^2$. Next, we bound R_d. We have the following claim, whose proof can be found in Appendix B.

Claim 19 *It holds that $|R_d| \leq 2^{-2d}\binom{d-1}{t-1}^2$.*

Therefore, using that $\binom{d-1}{t-1} = \binom{d-1}{d-t-2} = (t/d)\binom{d}{t}$, we get that $|R_d| \leq (t/d)^2(2^{-d}\binom{d}{t})^2 \leq \epsilon^2/d$. We next bound R_k for $k \in \{1, 2, \ldots, d-1\}$ (see Appendix B for the proof).

Claim 20 *Let $c > 0$ be a sufficiently large constant. We have that $\sum_{k=c\log(d/\epsilon)}^{d-c\log(d/\epsilon)} R_k \leq \epsilon^2/d$.*

Finally, the following claim bounds the small degree terms.

Claim 21 *Let $k' = c\log(d/\epsilon)$, where $c > 0$ is the absolute constant as in Claim 20. For $0 \leq k \leq k'$ or $d - k' \leq k \leq d$, we have $|R_k| \leq 4\epsilon^2 k'|\mathbf{v} \cdot \mathbf{u}|/d$.*

Proof We provide the proof for the case where $0 \leq k \leq k'$, as the other case is symmetric. Let $a = |d-2m|/d$ and note that $a = |\mathbf{v}\cdot\mathbf{u}|/d$. From Lemma 17, we have that $|\mathcal{K}(d,m,k)| \leq (a^k + (\log(d/\epsilon)/d)^{k/2})$. Thus, it follows that

$$\begin{aligned}|R_k| &= \binom{d}{k}\left(2^{-d}\sum_{s=t}^{d}\binom{d}{s}\mathcal{K}(d,k,s)\right)^2 \mathcal{K}(d,m,k) \\ &\leq 2^{4k}\binom{d}{k}\left(2^{-d}\sum_{s=t}^{d}\binom{d}{s}\mathcal{K}(d,k,s)\right)^2 (a^k + (\log(d/\epsilon)/d)^{k/2}) \\ &\leq 2^{4k}\binom{d}{k}\left(2^{-d}\sum_{\substack{s=t \\ |s-d/2|\leq c'\sqrt{dk\log(d/\epsilon)}}}^{d}\binom{d}{s}|\mathcal{K}(d,k,s)| + (\epsilon/d)^{2ck}\right)^2 (a^k + (\log(d/\epsilon)/d)^{k/2}) \, ,\end{aligned}$$

where we used that $|\mathcal{K}(d,k,s)| \leq 1$ and that $\sum_{i=k}^{d} \binom{d}{i} 2^{-d} \leq 2\exp(-(k-n/2)^2/n)$ from Hoeffding's inequality, hence $\sum_{s \geq d/2c'\sqrt{dk\log(d/\epsilon)}}^{d} \binom{d}{s} \leq (\epsilon/d)^{2ck}$. Futhermore, note that from Lemma 17, we have that $|\mathcal{K}(d,k,s)| \leq (2k\sqrt{\log(d/\epsilon)/d})^k$ for $|s-d/2| \leq c'\sqrt{dk\log(d/\epsilon)}$. Therefore, we have that

$$|R_k| \leq 2^{4k}\binom{d}{k}\left((2\sqrt{\log(d/\epsilon)/dk})^k \sum_{s=t}^{d}\binom{d}{s}2^{-d} + (\epsilon/d)^{2ck}\right)^2 (a^k + (\log(d/\epsilon)/d)^{k/2})$$
$$\leq 2^{4k}\epsilon^2\binom{d}{k}(2\sqrt{\log(d/\epsilon)/dk})^{2k}(a^k + (\log(d/\epsilon)/d)^{k/2}) \,,$$

where we used that by our choice of t it holds $\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})] = \epsilon$; therefore, $\sum_{s=t}^{d}\binom{d}{s}2^{-d} \leq \epsilon$. Using the fact that $\kappa \leq c\log(d/\epsilon)$ and that $\binom{d}{k} \leq d^k$, we get that $|R_k| \leq \epsilon^2(2k')^{C'k}(a^k + d^{-k/2})$. Therefore, if $a \leq C/\mathrm{poly}(k')$ for some sufficiently small absolute constant $C > 0$, we get that all the terms are bounded by the first term, i.e., we get that $|R_k| \leq 4\epsilon^2 k'\alpha$. ■

In summary, we have that $\sum_{k=0}^{d} R_k \leq C\log^2(d/\epsilon)\epsilon^2|\mathbf{v}\cdot\mathbf{u}|/d + \epsilon^2$, for some absolute constant $C > 0$. This completes the proof. ■

4. Conclusions

We studied the classical problem of learning margin halfspaces with Random Classification Noise. Our main finding is an information-computation tradeoff for SQ algorithms and low-degree polynomial tests. Specifically, our lower bounds suggest that efficient learners require sample complexity at least $\Omega(1/(\gamma^{1/2}\epsilon^2))$ (while $\widetilde{O}(1/(\gamma^2\epsilon))$ samples information-theoretically suffice). A number of interesting open questions remain. First, there is still a gap between $\widetilde{O}(1/(\gamma^2\epsilon^2))$ — the sample complexity of our algorithm — and the lower bound of $\Omega(1/(\gamma^{1/2}\epsilon^2))$. We believe that an SQ lower bound of $\Omega(1/(\gamma\epsilon^2))$ can be obtained with a more careful construction, but it is not clear what the optimal bound may be. Second, it would be interesting to obtain reduction-based computational hardness matching our SQ lower bound, along the lines of recent results (Gupte et al., 2022; Diakonikolas et al., 2022b, 2023).

Acknowledgments

Ilias Diakonikolas was supported by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), and a DARPA Learning with Less Labels (LwLL) grant. Jelena Diakonikolas was supported by NSF Award CCF-2007757 and by the U. S. Office of Naval Research under award number N00014-22-1-2348. Daniel Kane was supported by NSF Medium Award CCF-2107547 and NSF Award CCF-1553288 (CAREER). Puqian Wang was supported in part by NSF Award CCF-2007757. Nikos Zarifis was supported in part by NSF award 2023239, NSF Medium Award CCF-2107079, and a DARPA Learning with Less Labels (LwLL) grant.

References

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

J. Błasiok, P. Ivanov, Y. Jin, C. Lee, R. Servedio, and E. Viola. Fourier growth of structured f2-polynomials and applications. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

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.

M. Brennan, G. Bresler, S. Hopkins, J. Li, and T. Schramm. Statistical query algorithms and low-degree tests are almost equivalent. *arXiv preprint arXiv:2009.06107*, 2020.

T. Bylander. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory, COLT 1994*, pages 340–347, 1994.

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.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.

I. Diakonikolas and D. Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4258–4282. PMLR, 2022.

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. doi: 10.1109/FOCS.2017.16.

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.

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

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

I. Diakonikolas, D. M. Kane, P. Manurangsi, and L. Ren. Cryptographic hardness of learning halfspaces with massart noise. *CoRR*, abs/2207.14266, 2022b. doi: 10.48550/arXiv.2207.14266. URL https://doi.org/10.48550/arXiv.2207.14266.

I. Diakonikolas, D. M. Kane, and L. Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. *CoRR*, abs/2302.06512, 2023. doi: 10.48550/arXiv.2302.06512. URL https://doi.org/10.48550/arXiv.2302.06512.

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.

A Gupte, N. Vafa, and V. Vaikuntanathan. Continuous LWE is as hard as LWE & applications to learning gaussian mixtures. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022*, pages 1162–1173, 2022.

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

V. Kontonis, F. Iliopoulos, K. Trinh, C. Baykal, G. Menghani, and E. Vee. Slam: Student-label mixing for distillation with unlabeled examples. *arXiv preprint arXiv:2302.03806*, 2023.

N. Littlestone. From online to batch learning. In *Proceedings of the Second Annual Workshop on Computational Learning Theory*, pages 269–284, 1989.

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

R. Nasser and S. Tiegel. Optimal SQ lower bounds for learning halfspaces with massart noise. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1047–1074. PMLR, 2022. URL https://proceedings.mlr.press/v178/nasser22a.html.

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

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

S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

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.

Appendix A. Omitted Proofs from Section 2

A.1. Proof of Lemma 7

We restate and prove the following.

Lemma 7 *Consider Algorithm 1. There exists $t \in \{0, \dots T\}$ such that*

$$\mathbf{Pr}[\mathrm{sign}(\mathbf{w}_t \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})] \leq E_1 + E_2 + E_3,$$

where $E_1 = \frac{2(1-\eta)}{(1-2\eta)\gamma\sqrt{T+1}}$, $E_2 = \frac{2}{(1-2\eta)\gamma}\|\hat{\mathbf{g}}_N(\mathbf{w}^*)\|_2$, *and* $E_3 = \frac{1}{T+1}\sum_{t=0}^{T}\big[\mathbf{Pr}[\mathrm{sign}(\mathbf{w}_t \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})] - \widehat{\mathbf{Pr}}_N(\mathbf{w}_t)\big]$.

Proof We first argue, using standard regret analysis provided for completeness, that

$$\frac{1}{t+1}\sum_{s=0}^{t}\hat{\mathbf{g}}_N(\mathbf{w}_s)\cdot(\mathbf{w}_s - \mathbf{w}^*) \leq \frac{2}{\mu(t+1)} + \frac{\mu(1-\eta)^2}{2}, \tag{6}$$

where μ is the step size specified in Algorithm 1.

Fix any $t \in \{0, 1, \dots, T-1\}$. Recall that $\mathbf{w}_{t+1} = \mathrm{proj}_{\mathcal{B}}(\mathbf{w}_t - \mu\hat{\mathbf{g}}_N(\mathbf{w}_t))$ and $\mathbf{w}^* = \mathrm{proj}_{\mathcal{B}}(\mathbf{w}^*)$. Hence, by the nonexpansivity of the projection operator, we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}_t - \mathbf{w}^* - \mu\hat{\mathbf{g}}_N(\mathbf{w}_t)\|_2. \tag{7}$$

Further, expanding the square $\|\mathbf{w}_t - \mathbf{w}^* - \mu\hat{\mathbf{g}}_N(\mathbf{w}_t)\|_2^2$ and using Claim 6, we get

$$\begin{aligned}\|\mathbf{w}_t - \mathbf{w}^* - \mu\hat{\mathbf{g}}_N(\mathbf{w}_t)\|_2^2 &= \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \mu^2\|\hat{\mathbf{g}}_N(\mathbf{w}_t)\|_2^2 - 2\mu\hat{\mathbf{g}}_N(\mathbf{w}_t)\cdot(\mathbf{w}_t - \mathbf{w}^*)\\ &\leq \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \mu^2(1-\eta)^2 - 2\mu\hat{\mathbf{g}}_N(\mathbf{w}_t)\cdot(\mathbf{w}_t - \mathbf{w}^*).\end{aligned}$$

Hence, combining the last inequality with Equation (7), we get

$$\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \mu^2(1-\eta)^2 - 2\mu\hat{\mathbf{g}}_N(\mathbf{w}_t)\cdot(\mathbf{w}_t - \mathbf{w}^*). \tag{8}$$

To obtain Equation (6), it remains to rearrange Equation (8) and telescope. In particular, for T itearations and $\mu = \frac{2}{(1-\eta)\sqrt{T+1}}$, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\hat{\mathbf{g}}_N(\mathbf{w}_t)\cdot(\mathbf{w}_t - \mathbf{w}^*) \leq \frac{2(1-\eta)}{\sqrt{T+1}}. \tag{9}$$

Writing $\hat{\mathbf{g}}_N(\mathbf{w}_t)\cdot(\mathbf{w}_t - \mathbf{w}^*)$ as $\hat{\mathbf{g}}_N(\mathbf{w}_t)\cdot(\mathbf{w}_t - \mathbf{w}^*) = \hat{\mathbf{g}}_N(\mathbf{w}^*)\cdot(\mathbf{w}_t - \mathbf{w}^*) + (\hat{\mathbf{g}}_N(\mathbf{w}_t) - \hat{\mathbf{g}}_N(\mathbf{w}^*))\cdot(\mathbf{w}_t - \mathbf{w}^*)$ and rearranging Equation (9), we further get

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}(\hat{\mathbf{g}}_N(\mathbf{w}_t) - \hat{\mathbf{g}}_N(\mathbf{w}^*))\cdot(\mathbf{w}_t - \mathbf{w}^*) &\leq \frac{2(1-\eta)}{\sqrt{T+1}} + \hat{\mathbf{g}}_N(\mathbf{w}^*)\cdot\Big(\mathbf{w}^* - \frac{1}{T+1}\sum_{t=1}^{t}\mathbf{w}_t\Big)\\ &\leq \frac{2(1-\eta)}{\sqrt{T+1}} + 2\|\hat{\mathbf{g}}_N(\mathbf{w}^*)\|_2,\end{aligned} \tag{10}$$

where we used Cauchy-Schwarz inequality and $\mathbf{w}^*, \mathbf{w}_t \in \mathcal{B}$, for all $t \in \{0, \dots, T\}$.

Recall from Equation (3) that

$$(\mathbf{g}_\eta(\mathbf{w}) - \mathbf{g}_\eta(\mathbf{w}^*)) \cdot (\mathbf{w} - \mathbf{w}^*) = (1-2\eta)\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\}(|\mathbf{w}\cdot\mathbf{x}| + |\mathbf{w}^*\cdot\mathbf{x}|)$$
$$\geq (1-2\eta)\gamma\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\},$$

where the inequality holds by Definition 1, Item (A_1). Hence, by the definitions of $\hat{\mathbf{g}}_N$ and $\widehat{\mathbf{Pr}}_N$ (from Equation (5) and Equation (4)), we have

$$(\hat{\mathbf{g}}_N(\mathbf{w}_t) - \hat{\mathbf{g}}_N(\mathbf{w}^*)) \cdot (\mathbf{w}_t - \mathbf{w}^*) \geq (1-2\eta)\gamma\widehat{\mathbf{Pr}}_N(\mathbf{w}_t). \tag{11}$$

Combining Equation (10) and Equation (11), we then obtain the claimed regret bound, using simple algebraic manipulations. ■

A.2. Proof of Lemma 8

We restate the lemma and provide proof.

Lemma 8 *Let E_2 and E_3 be defined as in Lemma 7. For any $\bar{\epsilon} > 0, \delta > 0$, if $N = \Omega(\frac{d}{\bar{\epsilon}} + \frac{\eta}{(1-2\eta)^2\bar{\epsilon}^2\gamma^2})\log(\frac{1}{\delta}))$, then with probability at least $1-\delta$ we have $E_2 + E_3 \leq \frac{\bar{\epsilon}}{2}$.*

We first bound the error $E_2 = \frac{2}{(1-2\eta)\gamma}\|\hat{\mathbf{g}}_N(\mathbf{w}^*)\|_2$. Observe that $\mathbf{E}[\mathbf{g}_\eta(\mathbf{w}^*)] = 0$ and that, by Claim 6, $\|\mathbf{g}_\eta(\mathbf{w}^*)\|_2 \leq 1-\eta$ surely. We use the following Bennett-type inequality

Fact 22 ((Smale and Zhou, 2007), Lemma 1) *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \in \mathbb{R}^d$ be random variables such that for each $i \in [n]$ it holds $\|\mathbf{Z}_i\|_2 \leq M < \infty$ almost surely and let $\sigma^2 = \sum_{i=1}^n \mathbf{E}[\|\mathbf{Z}_i\|_2^2]$. Then, we have that for any $\epsilon > 0$,*

$$\mathbf{Pr}\left[\left\|\frac{1}{n}\sum_{i=1}^n(\mathbf{Z}_i - \mathbf{E}[\mathbf{Z}_i])\right\|_2 \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon}{2M}\log\left(1+\frac{nM\epsilon}{\sigma^2}\right)\right) .$$

Note that $\mathbf{E}[\|\mathbf{g}_\eta(\mathbf{w}^*)\|_2^2 = O(\eta)$ and using Fact 22, along with the inequality $\log(1+z) \geq z/2$, for $z \in (0,1)$ (note that σ^2 is at most nM), we get that for any $\hat{\epsilon}$ and $N \geq \Omega(\frac{\log(1/\delta)}{\hat{\epsilon}^2})$, with probability at least $1-\delta/2$, we have

$$\|\mathbf{E}[\mathbf{g}_\eta(\mathbf{w}^*)] - \hat{\mathbf{g}}_N(\mathbf{w}^*)\|_2 = \|\hat{\mathbf{g}}_N(\mathbf{w}^*)\|_2 \leq \hat{\epsilon} .$$

To complete bounding E_2, it remains to choose $\hat{\epsilon} = \frac{(1-2\eta)\gamma\bar{\epsilon}}{8}$.

To complete the proof and bound $E_3 = \frac{1}{T+1}\sum_{t=0}^T\left[\mathbf{Pr}[\mathrm{sign}(\mathbf{w}_t\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})] - \widehat{\mathbf{Pr}}_N(\mathbf{w}_t)\right]$, we use uniform convergence results for the function $\mathbf{w} \mapsto \mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\}$. Note that this boolean concept class is a subset of the class of the intersection of two halfspaces. The latter class has VC dimension $O(d)$. Thus, by the standard VC inequality combined with uniform convergence (see e.g., p. 31 of Devroye and Lugosi (2001)) we have that $N = \Omega(\frac{d}{\bar{\epsilon}}\log(1/\delta))$ samples suffice so that with probability $1-\delta/2$, we have $E_3 \leq \frac{\bar{\epsilon}}{4}$. ■

A.3. Testing to Find the Right Hypothesis

The following lemma justifies the claim made at the end of the proof of Theorem 9 and completes its proof. We use the following fact from Shalev-Shwartz and Ben-David (2014).

Fact 23 (Thereom 6.8 of Shalev-Shwartz and Ben-David (2014)) *Given a finite set of hypotheses $\mathcal{H}$, by drawing $N = O(\frac{\log(|\mathcal{H}|+\log(\frac{1}{\delta}))}{\epsilon^2})$ samples, it is guranteed that with probability at least $1-\delta$ it holds*

$$\min_{h\in\mathcal{H}}\Big\{\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{h(\mathbf{x}^{(i)})\neq y^{(i)}\}\Big\}\leq \min_{h\in\mathcal{H}}\mathbf{Pr}[h(\mathbf{x})\neq y]+\epsilon\,.$$

A.4. Dimension Reduction

We will use the following Johnson-Lindenstrauss lemma as our main technique to reduce the dimension of the space.

Lemma 24 (Johnson-Lindenstrauss) *Let β,ϵ be some positive constants. Let $A\in\mathbb{R}^{m\times d}$ be a random matrix with each entry A_{ij} sampled from* Uniform$\{-1/\sqrt{m},1/\sqrt{m}\}$ *where $m=O(\log(1/\beta)/\epsilon^2)$. Then, for any unit vector $\mathbf{u},\mathbf{v}\in\mathbb{R}^d$, it holds*

$$\mathop{\mathbf{Pr}}_{A}[|\mathbf{u}\cdot\mathbf{v}-(A\mathbf{u})\cdot(A\mathbf{v})|\geq\epsilon]\leq\beta.$$

Consequently, for any unit vector $\mathbf{u}\in\mathbb{R}^d$, we have $\mathbf{Pr}_A[|\|\mathbf{u}\|_2^2-\|A\mathbf{u}\|_2^2|\geq\epsilon]\leq\beta$.

Corollary 25 *Let β,β',γ be some positive constants such that $\beta\leq\beta'$ and $\gamma\in(0,1)$. Let $A\in\mathbb{R}^{m\times d}$ be a random matrix with each entry A_{ij} sampled uniformly at random from $\{-1/\sqrt{m},1/\sqrt{m}\}$ where $m=O(\log(1/\beta)/\gamma^2)$. Then, for $\mathbf{x}\sim D_{\mathbf{x}}$ and $\mathbf{w}^*\in\mathcal{B}$, with probability at least $1-\beta/\beta'$ it holds*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim D_{\mathbf{x}}}[|\mathbf{w}^*\cdot\mathbf{x}-(A\mathbf{w}^*)\cdot(A\mathbf{x})|\geq\gamma/2]\leq\beta'.$$

In addition, with probability at least $1-\beta/\beta'$ it holds $\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[|\|\mathbf{x}\|_2^2-\|A\mathbf{x}\|_2^2|\geq\gamma/2]\leq\beta'$.

Proof Lemma 24 indicates that since the matrix A is generated independent of the distribution $D_{\mathbf{x}}$, for any unit vector $\mathbf{x}\sim D_{\mathbf{x}}$ the norm of the transformed vector $A\mathbf{x}$ is close to 1 with constant probability. To be specific, we have:

$$\mathop{\mathbf{Pr}}_{A,\mathbf{x}\sim D_{\mathbf{x}}}[|\mathbf{w}^*\cdot\mathbf{x}-(A\mathbf{w}^*)\cdot(A\mathbf{x})|\geq\gamma/2]\leq\beta. \tag{12}$$

Now let $P(A)=\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[|\mathbf{w}^*\cdot\mathbf{x}-(A\mathbf{w}^*)\cdot(A\mathbf{x})|\geq\gamma/2]=\mathbf{E}_{\mathbf{x}\sim D_{\mathbf{x}}}[\mathbb{1}\{|\mathbf{w}^*\cdot\mathbf{x}-(A\mathbf{w}^*)\cdot(A\mathbf{x})|\geq\gamma/2\}|A]$ be a random variable determined by A. Note that $\mathbf{E}_A[P(A)]\leq\beta$ by Equation (12). Thus, applying Markov inequality to $P(A)$ we get

$$\mathop{\mathbf{Pr}}_{A}[P(A)\geq\beta']\leq\frac{\mathbf{E}_A[P(A)]}{\beta'}\leq\frac{\beta}{\beta'}.$$

Therefore, for any given matrix A sampled from the distribution $A_{ij} \sim \text{Uniform}\{\pm\frac{1}{\sqrt{m}}\}$ where $m = O(\log(1/\beta)/\gamma^2)$, we have with probability at least $1-\beta/\beta'$,

$$\Pr_{\mathbf{x}\sim D_{\mathbf{x}}}[|\mathbf{w}^* \cdot \mathbf{x} - (A\mathbf{w}^*)\cdot(A\mathbf{x})| \geq \gamma/2] \leq \beta'.$$

Following the same idea, we can also show that with probability at least $1-\beta/\beta'$ it holds $\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[|\|\mathbf{x}\|_2^2 - \|A\mathbf{x}\|_2^2| \geq \gamma/2] \leq \beta'$. ■

Input: $\epsilon > 0$, $\gamma \in (0,1)$, $\eta \in (0,1/2)$, $m > 0$, sample $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$ drawn i.i.d. from D, a random matrix $A \in \mathbb{R}^{m\times d}$ generated such that $A_{ij} = 1/\sqrt{m}$ w.p. $1/2$ and $A_{ij} = -1/\sqrt{m}$ w.p. $1/2$, any $\mathbf{w}_0 \in \mathcal{B}$

$T = \lceil (\frac{48(1-\eta)}{(1-2\eta)\gamma\epsilon})^2 - 1 \rceil$, $\mu = \frac{1}{(1-\eta)\sqrt{T+1}}$, $\bar{\mathbf{x}}^{(i)} = A\mathbf{x}^{(i)}$ for $i = 1, \cdots, N$.

for $t = 0 : T-1$ **do**

$\bar{\mathbf{g}}_N(\bar{\mathbf{w}}_t) = \frac{1}{2N}\sum_{i=1}^N ((1-2\eta)\text{sign}(\bar{\mathbf{w}}_t \cdot \bar{\mathbf{x}}^{(i)}) - y^{(i)})\bar{\mathbf{x}}^{(i)}$

$\bar{\mathbf{w}}_{t+1} = \text{proj}_{\|\bar{\mathbf{w}}\|_2 \leq 1}(\bar{\mathbf{w}}_t - \mu\bar{\mathbf{g}}_N(\bar{\mathbf{w}}_t))$

end

return $\{A^\top\bar{\mathbf{w}}_0, \ldots, A^\top\bar{\mathbf{w}}_T\}$

Algorithm 2: Dimension-Reduced Margin Halfspace Learner with RCN

Theorem 26 *Fix $\epsilon > 0, \gamma \in (0,1)$. Let the number of iterations be $T = O(\frac{(1-\eta)^2}{\gamma^2\epsilon^2})$ and set the stepsize $\mu = \frac{1}{(1-\eta)\sqrt{T+1}}$. Furthermore, let $A \in \mathbb{R}^{m\times d}$ be a matrix generated from the distribution described in Algorithm 2 with $m = O(\frac{\log((1-2\eta)/(\delta\epsilon))}{\gamma^2})$. Then running Algorithm 2 for T iterations with $N = \widetilde{\Omega}((\frac{\eta}{\gamma^2\epsilon^2} + \frac{(1-2\eta)}{\gamma^2\epsilon}\log(\frac{1-2\eta}{\delta\epsilon}))\log(\frac{1}{\delta}))$ i.i.d. samples drawn from distribution D, Algorithm 2 learns γ-margin halfspaces corrupted with η-RCN with error $\eta+\epsilon$ with probability at least $1-\delta$.*

Proof The goal of the proof is to show that the analysis of Algorithm 1 can be transformed to Algorithm 2 with minor modifications. For simplicity, let's denote $\bar{\mathbf{w}} = A\mathbf{w} \in \mathbb{R}^m$ and $\bar{\mathbf{x}} = A\mathbf{x} \in \mathbb{R}^m$ for any $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$. Similarly, we have $\bar{\mathbf{w}}^* = A\mathbf{w}^*$ and $\bar{\mathbf{x}}^{(i)} = A\mathbf{x}^{(i)}$.

We first show that as a consequence of Lemma 24 and Corollary 25, with a large probability that the A generated in Algorithm 2 is a "good matrix" in the sense that for most of the points in D, $A\mathbf{x}$ will not be far away from $\mathbf{x}$. Formally, we have the following claim.

Claim 27 *Fix some constants $\gamma, \bar{\epsilon}, \delta > 0, N > 1$ and let $m = O(\frac{\log(1/\beta)}{\gamma^2})$ where $\beta = \frac{\bar{\epsilon}\delta}{20N}$. For any A generated in Algorithm 2, denote $\mathcal{E}_A = \{\mathbf{x} \in \mathcal{S}^{d-1} : |\mathbf{w}^*\cdot\mathbf{x} - \bar{\mathbf{w}}^*\cdot\bar{\mathbf{x}}| \leq \gamma/2,\ |\|\mathbf{x}\|_2^2 - \|\bar{\mathbf{x}}\|_2^2| \leq \gamma/2\}$ and let $\mathcal{E}$ be the set of A such that $\bar{\mathbf{w}}^*$ is close to $\mathbf{w}^*$ and moreover, for any $\mathbf{x} \sim D_{\mathbf{x}}$, $\mathbf{x} \in \mathcal{E}_A$ with high probability, i.e., $\mathcal{E} = \{A \in \mathbb{R}^{m\times d} : \mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[\mathbf{x} \in \mathcal{E}_A] \geq 1 - \bar{\epsilon}/(2N),\ |\|\mathbf{w}^*\|_2^2 - \|\bar{\mathbf{w}}^*\|_2^2| \leq \gamma/2\}$. Then,*

1. *$\mathcal{E}$ happens with probability at least $1 - \frac{2}{5}\delta$;*
2. *If $\mathcal{E}$ happens, then for any N i.i.d. samples $\{\mathbf{x}^{(i)}\}_{i=1}^N$, it holds $\mathbf{Pr}[\mathbf{x}^{(i)} \in \mathcal{E}_A, \forall i \in [N]] \geq 1 - \frac{\bar{\epsilon}}{2}$.*

Proof According to Lemma 24, we know that $\mathbf{Pr}[|\|\mathbf{w}^*\|_2^2 - \|\bar{\mathbf{w}}^*\|_2^2| \le \gamma/2] \ge 1-\beta = 1-\bar{\epsilon}\delta/(20N)$. Furthermore, recall that in Corollary 25 (with a union bound) we showed with probability at least $1-4N\beta/\bar{\epsilon} = 1-\delta/5$, $\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[\mathbf{x}\in\mathcal{E}_A] \ge 1-\bar{\epsilon}/(2N)$, therefore the first claim follows from a union bound on these 2 events.

Now conditioned on $\mathcal{E}$. Given any N samples $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$, we know that $\mathbf{Pr}[\mathbf{x}^{(i)} \in \mathcal{E}_A] \ge 1-\frac{\bar{\epsilon}}{2N}$, hence applying union bound we get $\mathbf{Pr}[\mathbf{x}^{(i)}\in\mathcal{E}_A, \forall i\in[N]\,] \ge 1-\frac{\bar{\epsilon}}{2}$. ■

For the analysis below, we will condition on the event $A\in\mathcal{E}$ which occurs with probability at least $1-\frac{2}{5}\delta$. We begin with showing that under such fixed A, the norm of $\|\bar{\mathbf{g}}_N(\bar{\mathbf{w}})\|_2$ can be bounded by $2(1-\eta)$.

Claim 28 *Given N samples $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$ and suppose that $\mathbf{x}^{(i)}\in\mathcal{E}_A, \forall i=1,\cdots,N$, we have*

$$\sup_{\bar{\mathbf{w}}\in\mathbb{R}^m} \|\bar{\mathbf{g}}_N(\bar{\mathbf{w}})\|_2 \le 2(1-\eta).$$

Proof Since $\mathbf{x}^{(i)}\in\mathcal{E}_A$, we have $\|\bar{\mathbf{x}}^{(i)}\|_2 \le \|\mathbf{x}^{(i)}\|_2 + \gamma/2$. Hence, it holds

$$\begin{aligned}\sup_{\bar{\mathbf{w}}\in\mathbb{R}^m} \|\bar{\mathbf{g}}_N(\bar{\mathbf{w}})\|_2 &= \sup_{\bar{\mathbf{w}}\in\mathbb{R}^m}\left\{\frac{1}{2N}\Big\|\sum_{i=1}^N((1-2\eta)\mathrm{sign}(\bar{\mathbf{w}}\cdot\bar{\mathbf{x}}^{(i)}) - y^{(i)})\bar{\mathbf{x}}^{(i)}\Big\|_2\right\}\\ &\le (1-\eta)\|\bar{\mathbf{x}}^{(i)}\|_2 \le (1-\eta)(\|\mathbf{x}^{(i)}\|_2+\gamma/2) \le 2(1-\eta),\end{aligned}$$

where in the last inequality we used the fact that $\|\mathbf{x}^{(i)}\|_2 = 1$ and $\gamma\in(0,1)$. ■

We then study the difference between $\|\bar{\mathbf{w}}_{t+1}-\bar{\mathbf{w}}^*\|_2$ and $\|\bar{\mathbf{w}}_t-\bar{\mathbf{w}}^*\|_2$, which is almost analogous to the analysis we have seen in Lemma 7 with the only differences being that: (i) with probability at least $1-\bar{\epsilon}/2$ we have $\mathbf{x}^{(i)}\in\mathcal{E}_A$ for $i\in[N]$, hence $\|\bar{\mathbf{g}}_N(\bar{\mathbf{w}})\|_2 \le 2(1-\eta)$ for every $\bar{\mathbf{w}}_t$ and $\bar{\mathbf{w}}^*$ as shown in Claim 27 and Claim 28; (ii) since $|\mathbf{w}^*\cdot\mathbf{x}^{(i)} - \bar{\mathbf{w}}^*\cdot\bar{\mathbf{x}}^{(i)}| \le \gamma/2$ for all $i\in[N]$, it holds $\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}) = \mathrm{sign}(\bar{\mathbf{w}}^*\cdot\bar{\mathbf{x}}^{(i)})$; (iii) $\|\bar{\mathbf{w}}^*\|_2 \le 2$ since we have $|\|\mathbf{w}^*\|_2^2 - \|\bar{\mathbf{w}}^*\|_2^2| \le \gamma/2$ conditioning on the event $\mathcal{E}$.

Now we further condition on the event that $\mathbf{x}^{(i)}\in\mathcal{E}_A$ for $i\in[N]$ and denote the distribution of $D_{\mathbf{x}}$ restricted on $\mathcal{E}_A$ as $D_{\mathbf{x}}(\mathcal{E}_A)$. Then, choosing $\mu = \frac{1}{(1-\eta)\sqrt{T+1}}$ and following the same steps as in Lemma 7, we have $\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}(\mathcal{E}_A)}[\mathrm{sign}((A^\top\bar{\mathbf{w}}_t)\cdot\mathbf{x}) \ne \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})] \le E_1' + E_2' + E_3'$, where $E_1' = \frac{8(1-\eta)}{(1-2\eta)\gamma\sqrt{T+1}}$, $E_2' = \frac{6}{(1-2\eta)\gamma}\|\bar{\mathbf{g}}_N(\bar{\mathbf{w}}^*)\|_2$ and $E_3' = \frac{1}{T+1}\sum_{t=0}^T\{\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}(\mathcal{E}_A)}[\mathrm{sign}((A^\top\bar{\mathbf{w}}_t)\cdot\mathbf{x}) \ne \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})] - \widehat{\mathbf{Pr}}_N(A^\top\bar{\mathbf{w}})\}$.

We show that our choice of N and T suffices to make $\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[\mathrm{sign}((A^\top\bar{\mathbf{w}}_t)\cdot\mathbf{x}) \ne \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})] \le \bar{\epsilon}$. First, $T = \left(\frac{48(1-\eta)}{(1-2\eta)\gamma\bar{\epsilon}}\right)^2$ renders $E_1' \le \bar{\epsilon}/6$. Next, observe that $\mathbf{E}_{\mathbf{x}\sim D_{\mathbf{x}}(\mathcal{E}_A)}[\mathbf{g}_\eta(\bar{\mathbf{w}}^*;\bar{\mathbf{x}},y)] = 0$ and recall that $\|\mathbf{g}_\eta(\bar{\mathbf{w}}^*;\bar{\mathbf{x}}^{(i)},y^{(i)})\|_2 \le 2(1-\eta)$, thus by Fact 22 we know $N = \Omega\big(\log(1/\delta)/\bar{\epsilon}^2\big)$ suffices to make $E_2' \le \bar{\epsilon}/6$ with probability $1-\delta/10$. Finally, since linear threshold function class $\bar{\mathbf{w}}\mapsto\mathbb{1}\{\mathrm{sign}((A^\top\bar{\mathbf{w}})\cdot\mathbf{x}) \ne \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\}$ has VC dimension $m+1$, therefore by standard VC dimension arguments choosing $N \ge \Omega(\frac{m}{\bar{\epsilon}}\log(1/\delta))$ we are guaranteed to have $E_3' \le \frac{\bar{\epsilon}}{6}$ with probability $1-\delta/10$. Recall that $\mathbf{Pr}[\mathbf{x}\in\mathcal{E}_A] \ge 1-\bar{\epsilon}/2$, hence under our choice of m, N, T, under the condition of event $\mathcal{E}$ and $\mathbf{x}^{(i)}\in\mathcal{E}_A$ for $i\in[N]$, we have with probability at least $1-\delta/5$,

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim D_{\mathbf{x}}}[\mathrm{sign}((A^\top\bar{\mathbf{w}}_t)\cdot\mathbf{x}) \ne \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})] \le \mathop{\mathbf{Pr}}_{\mathbf{x}\in D_{\mathbf{x}}(\mathcal{E}_A)}[\mathrm{sign}((A^\top\bar{\mathbf{w}}_t)\cdot\mathbf{x}) \ne \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})]\left(1-\frac{\bar{\epsilon}}{2}\right)+\frac{\bar{\epsilon}}{2} \le \bar{\epsilon}.$$

Finally, applying a union bound on all of these 3 events, we know that with probability at least $1 - \frac{2}{5}\delta - \frac{\bar{\epsilon}}{2} - \frac{1}{5}\delta \geq 1 - \delta$, $\mathbf{Pr}_{\mathbf{x}\sim D_{\mathbf{x}}}[\mathrm{sign}((A^\top \bar{\mathbf{w}}_t)\cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})] \leq \bar{\epsilon}$. By substituting $\bar{\epsilon}$ with $\epsilon/(1-2\eta)$, we get a learner with error $\eta + \epsilon$ and the proof is complete. ■

Appendix B. Omitted Proofs from Section 3

B.1. Additional Background on SQ Model

We will use the framework of Statistical Query (SQ) algorithms for problems over distributions (Feldman et al., 2017). We require the following standard definition.

Definition 29 (Decision/Testing Problem over Distributions) *Let D be a distribution and $\mathcal{D}$ be a family of distributions over $\mathbb{R}^d$. 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.*

To define the SQ dimension, we need the following definition.

Definition 30 (Pairwise Correlation) *The pairwise correlation of two distributions with probability mass functions (pmfs) $D_1, D_2 : \mathcal{X} \to \mathbb{R}_+$ with respect to a distribution with pmf $D : \mathcal{X} \to \mathbb{R}_+$, where the support of D contains the supports of D_1 and D_2, is defined as $\chi_D(D_1, D_2) + 1 \coloneqq \sum_{x\in\mathcal{X}} D_1(x)D_2(x)/D(x)$. We say that a collection of s distributions $\mathcal{D} = \{D_1, \ldots, D_s\}$ over $\mathcal{X}$ 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 $i = j$.*

The following notion of dimension effectively characterizes the difficulty of the decision problem.

Definition 31 (SQ Dimension) *For $\gamma, \beta > 0$, a decision problem $\mathcal{B}(\mathcal{D}, D)$, where D is fixed and $\mathcal{D}$ is a family of distributions over $\mathcal{X}$, let s be the maximum integer such that there exists $\mathcal{D}_D \subseteq D$ such that D_D is (γ, β)-correlated relative to D and $|D_D| \geq s$. We define the* Statistical Query dimension *with pairwise correlations (γ, β) of $\mathcal{B}$ to be s and denote it by $\mathrm{SD}(\mathcal{B}, \gamma, \beta)$.*

The connection between SQ dimension and lower bounds is captured by the following lemma.

Lemma 32 ((Feldman et al., 2017)) *Let $\mathcal{B}(D, D)$ be a decision problem, where D is the reference distribution and D is a class of distributions over $\mathcal{X}$. For $\gamma, \beta > 0$, let $s = \mathrm{SD}(\mathcal{B}, \gamma, \beta)$. Any SQ algorithm that solves $\mathcal{B}$ with probability at least $2/3$ requires at least $s \cdot \gamma/\beta$ queries to the $\mathrm{STAT}(\sqrt{2\gamma})$ or $\mathrm{VSTAT}(1/\gamma)$ oracles.*

B.2. Proof of Fact 15

We restate and prove the following fact.

Fact 15 *Let $d \in \mathbb{Z}_+$. Let $0 < c < 1/2$. There exists a collection $\mathcal{S}$ of $2^{\Omega(d^c)}$ vectors in $\{\pm 1\}^d$, such that any pair $\mathbf{v}, \mathbf{u} \in \mathcal{S}$, with $\mathbf{v} \neq \mathbf{u}$, satisfies $|\mathbf{v}\cdot\mathbf{u}| < d^{1/2+c}$.*

Proof Sample two vectors $\mathbf{v}, \mathbf{u}$ at random, i.e., $\mathbf{v}, \mathbf{u} \sim \mathcal{U}_d$. Note that $\mathbf{v} \cdot \mathbf{u}$ is a sum of Rademacher random variables. We use the following concentration inequality:

Fact 33 *Let $z_1, \ldots, z_n$ be Rademacher random variables. Then, for any $t > 0$, it holds*

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{n} z_i\right| \geq t\sqrt{n}\right] \leq 2\exp\left(-\frac{t^2}{2}\right).$$

Using Fact 33, we get that for $t = d^c$ for some $0 < c < 1/2$, we get that

$$\mathbf{Pr}\left[|\mathbf{v} \cdot \mathbf{u}| \geq d^{1/2+c}\right] \leq 2\exp\left(-\frac{d^{2c}}{2}\right).$$

From union bound we get that there exists $2^{\Omega(d^c)}$ such vectors. ■

B.3. Proof of Claim 12

We restate and prove the following claim.

Claim 12 *It holds $A_{\mathbf{v}}(\mathbf{x}) = \frac{\eta+(1-2\eta)f_{\mathbf{v}}(\mathbf{x})}{\eta+(1-2\eta)\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]}\mathcal{U}_d(\mathbf{x})$ and $B_{\mathbf{v}}(\mathbf{x}) = \frac{1-\eta-(1-2\eta)f_{\mathbf{v}}(\mathbf{x})}{1-\eta-(1-2\eta)\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]}\mathcal{U}_d(\mathbf{x})$.*

Proof Denote $g_{D_{\mathbf{v}}}$ the pmf of $D_{\mathbf{v}}$. We show the claim only for the distribution $A_{\mathbf{v}}$ as $B_{\mathbf{v}}$ follows similarly. Note that $A_{\mathbf{v}}(\mathbf{x}) = \frac{g_{D_{\mathbf{v}}}(\mathbf{x},y=1)}{\mathbf{Pr}[y=1]}$. By construction, we have that $\mathbf{Pr}[y = 1] = \eta + (1-2\eta)2\epsilon$ and $g_{D_{\mathbf{v}}}(\mathbf{x}, y = 1) = (\eta\mathbb{1}\{f_{\mathbf{v}}(\mathbf{x}) = 0\} + (1-\eta)\mathbb{1}\{f_{\mathbf{v}}(\mathbf{x}) > 0\})\mathcal{U}_d(\mathbf{x}) = (\eta + (1-2\eta)f_{\mathbf{v}}(\mathbf{x}))\mathcal{U}_d(\mathbf{x})$. Therefore, $A_{\mathbf{v}}(\mathbf{x}) = \frac{\eta+(1-2\eta)f_{\mathbf{v}}(\mathbf{x})}{\eta+(1-2\eta)2\epsilon}\mathcal{U}_d(\mathbf{x})$. Similarly, we show that $B(\mathbf{x}) = \frac{1-\eta-(1-2\eta)f_{\mathbf{v}}(\mathbf{x})}{1-\eta-(1-2\eta)2\epsilon}\mathcal{U}_d(\mathbf{x})$. ■

B.4. Proof of Lemma 13

We restate and prove the following.

Lemma 13 *Let D_0 be a product distribution over $\mathcal{U}_d \times \{0,1\}$, where $\mathbf{Pr}_{(\mathbf{x},y)\sim D_0}[y = 1] = \mathbf{Pr}_{(\mathbf{x},y)\sim D_{\mathbf{v}}}[y = 1]$. We have $\chi_{D_0}(D_{\mathbf{v}}, D_{\mathbf{u}}) \leq 2(1-2\eta)(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})f_{\mathbf{u}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]\,\mathbf{E}[f_{\mathbf{u}}(\mathbf{x})])$ and $\chi^2(D_{\mathbf{v}}, D_0) \leq (1-2\eta)(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]^2)$.*

Proof Denote $\kappa_1 = 1/\,\mathbf{Pr}_{(\mathbf{x},y)\sim D_{\mathbf{v}}}[y = 1]$ and $\kappa_0 = 1/\,\mathbf{Pr}_{(\mathbf{x},y)\sim D_{\mathbf{v}}}[y = 0]$. We have that

$$\begin{aligned}\chi_{D_0}(D_{\mathbf{v}}, D_{\mathbf{u}}) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D_{\mathbf{v}}}[y = 1]\chi_{\mathcal{U}_d}(A_{\mathbf{v}}, A_{\mathbf{u}}) + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D_{\mathbf{v}}}[y = 0]\chi_{\mathcal{U}_d}(B_{\mathbf{v}}, B_{\mathbf{u}}) \\ &= \kappa_1^{-1}\chi_{\mathcal{U}_d}(A_{\mathbf{v}}, A_{\mathbf{u}}) + \kappa_0^{-1}\chi_{\mathcal{U}_d}(B_{\mathbf{v}}, B_{\mathbf{u}})\,.\end{aligned}$$

We now bound each term in the above expression.

Claim 34 *We have $\chi_{\mathcal{U}_d}(A_{\mathbf{v}}, A_{\mathbf{u}}) \leq (1-2\eta)\kappa_1^2(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})f_{\mathbf{u}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]\,\mathbf{E}[f_{\mathbf{u}}(\mathbf{x})])$ and $\chi_{\mathcal{U}_d}(B_{\mathbf{v}}, B_{\mathbf{u}}) \leq (1-2\eta)\kappa_0^2(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})f_{\mathbf{u}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]\,\mathbf{E}[f_{\mathbf{u}}(\mathbf{x})])$.*

Proof We first bound $\chi_{\mathcal{U}_d}(A_{\mathbf{v}}, A_{\mathbf{u}})$ as the other follows similarly. We have that

$$\begin{aligned}\chi_{\mathcal{U}_d}(A_{\mathbf{v}}, A_{\mathbf{u}}) &= \sum_{\mathbf{x}\in\{\pm 1\}^d} \frac{(A_{\mathbf{v}}(\mathbf{x}) - \mathcal{U}_d(\mathbf{x}))(A_{\mathbf{u}}(\mathbf{x}) - \mathcal{U}_d(\mathbf{x}))}{\mathcal{U}_d(\mathbf{x})} \\ &= \frac{1-2\eta}{(\eta + (1-2\eta)\,\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})])^2} \sum_{\mathbf{x}\in\{\pm 1\}^d} (f_{\mathbf{v}}(\mathbf{x}) - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})])(f_{\mathbf{u}}(\mathbf{x}) - \mathbf{E}[f_{\mathbf{u}}(\mathbf{x})])\mathcal{U}_d(\mathbf{x}) \\ &= \frac{1-2\eta}{(\eta + (1-2\eta)\,\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})])^2} \left(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x}) f_{\mathbf{u}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]\,\mathbf{E}[f_{\mathbf{u}}(\mathbf{x})]\right) .\end{aligned}$$

Working in a similar way, we also get that $\chi_{\mathcal{U}_d}(B_{\mathbf{v}}, B_{\mathbf{u}}) = (1-2\eta)\kappa_0^2(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})f_{\mathbf{u}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]\,\mathbf{E}[f_{\mathbf{u}}(\mathbf{x})])$. ■

Using Claim 34, we get that

$$\chi_{D_0}(D_{\mathbf{v}}, D_{\mathbf{u}}) \le (1-2\eta)(\kappa_1 + \kappa_0)\,\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})f_{\mathbf{u}}(\mathbf{x})] \le 2(1-2\eta)\,\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})f_{\mathbf{u}}(\mathbf{x})] .$$

It remains to bound $\chi^2(D_{\mathbf{v}}, D_0)$. We show the following:

Claim 35 *Let $\kappa = (1-2\eta)/(\eta + (1-2\eta)2\,\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]^2)^2$. It holds that*

$$\chi^2(D_{\mathbf{v}}, D_0) \le (1-2\eta)\,\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]^2 .$$

Proof Let $\kappa = (1-2\eta)/(\eta + (1-2\eta)2\,\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]^2)^2$ We have that

$$\chi^2(D_{\mathbf{v}}, D_0) = \kappa_1^{-1}\chi^2(\mathcal{U}_d, A_{\mathbf{v}}) + \kappa_0^{-1}\chi^2(\mathcal{U}_d, B_{\mathbf{v}}) .$$

$$\begin{aligned}\chi^2(\mathcal{U}_d, A_{\mathbf{v}}) = \sum_{\mathbf{x}\in\{\pm 1\}^d} \frac{(A_{\mathbf{v}}(\mathbf{x}) - \mathcal{U}_d(\mathbf{x}))^2}{\mathcal{U}_d(\mathbf{x})} &= (1-2\eta)\kappa_1^2 \sum_{\mathbf{x}\in\{\pm 1\}^d} (f_{\mathbf{v}}(\mathbf{x}) - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})])^2\mathcal{U}_d(\mathbf{x}) \\ &= (1-2\eta)\kappa_1^2(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]^2) .\end{aligned}$$

Similarly, we show that $\chi^2(\mathcal{U}_d, B_{\mathbf{v}}) = (1-2\eta)\kappa_0^2(\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})] - \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})]^2)$. Combining, we get the result. ■

■

B.5. Proof of Lemma 17

The following is a more detailed version of Lemma 17.

Lemma 36 *Let $d, m, k \in \mathbb{Z}$. Then, the following hold*

1. *$|\mathcal{K}(d, m, k)| \le 1$ for any $k \in \mathbb{Z}$.*
2. *For $k \le d/2$, it holds*

$$|\mathcal{K}(d, m, k)| \le e^k 2^{3k} \left(\left(\frac{k}{d}\right)^{k/2} + \left(\frac{|d/2 - m|}{d}\right)^k \right) .$$

3. *If $k \le d/2$ and $|d/2 - m| \le d/4$, then $|\mathcal{K}(d, m, k)| = \exp(-\Omega(k))$.*

Proof The first part follows from the fact that $\mathcal{K}(d, m, k)$ is the expectation of a random variable with support in $[-1, 1]$. For the next claims, we use the following fact.

Fact 37 (Claim 22 of Błasiok et al. (2021)) *Let $d, m, k \in \mathbb{Z}$, then*

$$|\mathcal{K}(d, m, k)| \le \frac{e^k 2^{3k}}{\binom{d}{k}} \left(\left(\frac{d}{k}\right)^{k/2} + \left(\frac{|d/2 - m|}{k}\right)^k \right) .$$

Using the inequality $\binom{d}{k} \ge (d/k)^k$, we have that

$$|\mathcal{K}(d, m, k)| \le e^k 2^{3k} \left(\left(\frac{k}{d}\right)^{k/2} + \left(\frac{|d/2 - m|}{d}\right)^k \right) .$$

If $k \le d/12$ and $|d/2 - m| \le d/4$, then we have that $|\mathcal{K}(d, m, k)| \le 2\exp(-0.2k)$. We provide a proof for the final part, i.e., the case where $d/12 \le k \le d/2$. Denote $Y_{A,B} = (-1)^{A \cap B}$. The sum we want to bound is equal to $\mathbf{E}_{A,B}[Y_{A,B}]$. Denote $A' = \{1, 3, \ldots, 2m - 1\}$ and $s_i^B = |B \cap \{2i - 1, 2i\}|$ for $i = 1, \ldots, m$. Note that the $\mathbf{E}_{A,B}[Y_{A,B} \mid A = A', s_1^B, \ldots, s_m^B] = 0$, if we condition that any $s_i^B = 1$. This holds because if $s_i^B = 1$ for some i, then we can swap which $2i$ and $2i - 1$ is in B to create B' and hence $(-1)^{A \cap B} + (-1)^{A \cap B'} = 0$.

It suffices to show that if we choose B at random, i.e., B is a uniform subset of $[d]$ of size k, then with probability at most $\exp(-\Omega(k))$ we are in the case where no s_i^B is equal to 1. To show that, we create a new random variable B' and we sample B' as follows: we let the size of B' be equal to $\mathrm{Bin}(d, k/d)$, which is equivalent to sampling each element of $[d]$ with probability k/d. Now each $s_i^{B'}$ is independent of each other. The probability that each $s_i^{B'}$ is equal to one is $2k/d(1 - k/d) = \Omega(k/d)$. Therefore, the probability that no $s_i^{B'}$ is one is at most

$$(1 - \Omega(k/d))^m \le \exp(-\Omega(-km/d)) = \exp(-\Omega(k)) ,$$

where we used that $|d/2 - m| \le d/4$. Therefore, $|\mathbf{E}[Y_{A,B'}|A]| \le \exp(-\Omega(k))$. It remains to relate the expectation with respect B' to the expectation of B. Note that according to the sampling rule, there is an $\Omega(1/\sqrt{k})$ probability of generating a uniform subset of size k, but the probability that Y is non-zero is at most $\exp(-\Omega(k))$. Therefore, we have that

$$|\mathbf{E}[Y_{A,B'}|A]| = |\mathbf{E}[Y_{A,B'}|A, |B'| = k]| \le \exp(-\Omega(k))\sqrt{k} = \exp(-\Omega(k)) ,$$

where we used that k is large enough by assumption. Therefore, $|\mathbf{E}[Y_{A,B}|A]| \le \exp(-\Omega(k))$ and the total expectation is at most $\exp(-\Omega(k))$.

■

B.6. Proof of Claim 18

We restate and prove the following.

Claim 18 (Fourier Coefficients) *Fix vector $\mathbf{v} \in \{\pm 1\}^d$ and let $f_{\mathbf{v}}(\mathbf{x}) = \mathbb{1}\{\mathbf{v} \cdot \mathbf{x} \geq t\}$. For $T \subseteq [d]$, we have that the Fourier coefficient of f at $\chi_T(\mathbf{x})$, i.e., $\widehat{f}(T)$, is given by*

$$\widehat{f}(T) = \mathbf{E}[f_{\mathbf{v}}(\mathbf{x})\chi_T(\mathbf{x})] = \chi_T(\mathbf{v})(-1)^{|T|}2^{-d}\sum_{s=t}^{d}\binom{d}{s}\mathcal{K}(d,s,|T|) \, .$$

Proof We have that

$$\begin{aligned}
\mathbf{E}[f_{\mathbf{v}}(\mathbf{x})\chi_T(\mathbf{x})] &= 2^{-d}\sum_{\mathbf{x}\in\{\pm 1\}^d} f_{\mathbf{v}}(\mathbf{x})\chi_T(\mathbf{x}) \\
&= 2^{-d}\sum_{s=t}^{d}\sum_{A\subseteq[d],|A|=s}\prod_{i\in T\cap A}\mathbf{v}_i\prod_{i\in T\cap\bar{A}}(-\mathbf{v}_i) \\
&= \chi_T(\mathbf{v})(-1)^{|T|}2^{-d}\sum_{s=t}^{d}\sum_{A\subseteq[d],|A|=s}(-1)^{|T\cap A|} \\
&= \chi_T(\mathbf{v})(-1)^{|T|}2^{-d}\sum_{s=t}^{d}\sum_{A\subseteq[d],|A|=s}\binom{d}{|T|}^{-1}\sum_{B\subseteq[d],|B|=|T|}(-1)^{|A\cap T|} \\
&= \chi_T(\mathbf{v})(-1)^{|T|}2^{-d}\sum_{s=t}^{d}\binom{d}{|s|}\mathcal{K}(d,s,|T|) \, ,
\end{aligned}$$

where in the first equality we changed the summation so that s is the number of $\mathbf{x}_i$ that agree with $\mathbf{v}_i$, and we sum from t as if $\mathbf{v}$ and $\mathbf{x}$ agree in more than t coordinates, then the indicator is positive. In the third inequality, we used the fact that, due to the symmetry, the sum only depends on the size of $|T|$; hence, we sum over all subsets with size $|T|$ and divide by the number of subsets with size $|T|$. ■

B.7. Proof of Claim 19

We restate and prove the following:

Claim 19 *It holds that $|R_d| \leq 2^{-2d}\binom{d-1}{t-1}^2$.*

Proof We have that

$$\mathcal{K}(d,d,s) = \frac{1}{\binom{d}{s}}\sum_{A\subseteq[d],|A|=s}(-1)^{|A\cap[d]|} = (-1)^s \, ,$$

hence, $R_d = (2^{-d}\sum_{s=t}^{d}(-1)^s\binom{d}{s})^2\mathcal{K}(d,d,m)$. Therefore, $|R_d| \leq 2^{-2d}(\sum_{s=t}^{d}(-1)^s\binom{d}{s})^2$. Using the two identities about binomial sums, i.e., that $\sum_{s=0}^{t}(-1)^s\binom{d}{s} = (-1)^t\binom{d-1}{t}$ and $\sum_{s=0}^{d}(-1)^s\binom{d}{s} = 0$, we have that $|R_d| \leq 2^{-2d}\binom{d-1}{t-1}^2$. ■

B.8. Proof of Claim 20

We restate and prove the following claim.

Claim 20 *Let $c > 0$ be a sufficiently large constant. We have that $\sum_{k=c\log(d/\epsilon)}^{d-c\log(d/\epsilon)} R_k \leq \epsilon^2/d$.*

Proof Note that, $\sum_{k=0}^{d} \binom{d}{k} \left(2^{-d}\sum_{s=t}^{d}\binom{d}{s}\mathcal{K}(d,k,s)\right)^2 = \mathbf{E}[f_{\mathbf{v}}^2(\mathbf{x})] = \epsilon$. From Lemma 17, we get that $|\mathcal{K}(d,m,k)| \leq \exp(-ck)$, where $c > 0$ is an absolute constant. Therefore, if $d/2 \geq k \geq c\log(d/\epsilon)$ we have that $|\mathcal{K}(d,m,k)| \leq \epsilon/d$. Furthermore, using the fact that $|\mathcal{K}(d,m,k)|$ is symmetric with center $d/2$, we get that if $d/2 \leq k \leq d - c\log(d/\epsilon)$, then we also have that $|\mathcal{K}(d,m,k)| \leq \epsilon/d$. Therefore, we have that

$$\sum_{k=c\log(d/\epsilon)}^{d/2-c\log(d/\epsilon)} R_k \leq \sum_{k=c\log(d/\epsilon)}^{d/2-c\log(d/\epsilon)} |R_k| \leq (\epsilon/d) \sum_{k=c\log(d/\epsilon)}^{d/2-c\log(d/\epsilon)} \binom{d}{k}\left(2^{-d}\sum_{s=t}^{d}\binom{d}{s}\mathcal{K}(d,k,s)\right)^2 \leq \epsilon^2/d\,.$$

This completes the proof. ■

Appendix C. Lower Bound for Low-Degree Polynomial Testing

We begin by formally defining a hypothesis problem.

Definition 38 (Hypothesis testing) *Let a distribution D_0 and a set $\mathcal{S} = \{D_u\}_{u\in S}$ of distributions on $\mathbb{R}^d$. Let μ be a prior distribution on the indices S of that family. We are given access (via i.i.d. samples or oracle) to an* underlying *distribution where one of the two is true:*

- *H_0: The underlying distribution is D_0.*
- *H_1: First u is drawn from μ and then the underlying distribution is set to be D_u.*

We say that a (randomized) algorithm solves the hypothesis testing problem if it succeeds with non-trivial probability (i.e., greater than 0.9).

Definition 39 *Let D_0 be the joint distribution D_0 over the pair $(\mathbf{x}, y) \in \{\pm 1\}^d \times \{0,1\}$ where $\mathbf{x} \sim \mathcal{U}_d$ and $y \sim D_0(y)$ independently of $\mathbf{x}$. Let D_v be the joint distribution over pairs $(\mathbf{x}, y) \in \{\pm 1\}^d \times \{0,1\}$ where the marginal on y is again $D_0(y)$ but the conditional distribution $E_{\mathbf{v}}(\mathbf{x}|1)$ is of the form $A_{\mathbf{v}}$ (as in Theorem 11) and the conditional distribution $E_{\mathbf{v}}(\mathbf{x}|0)$ is of the form $B_{\mathbf{v}}$. Define $\mathcal{S} = \{E_{\mathbf{v}}\}_{\mathbf{v}\in S}$ for S being the set of d-dimensional nearly orthogonal vectors from Fact 15 and let the hypothesis testing problem be distinguishing between D_0 vs. $\mathcal{S}$ with prior μ being the uniform distribution on S.*

We need the following variant of the statistical dimension from Brennan et al. (2020), which is closely related to the hypothesis testing problems considered in this section. Since this is a slightly different definition from the statistical dimension (SD) used so far, we will assign the distinct notation (SDA) for it.

Notation For $f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ and a distribution D, we define the inner product $\langle f, g \rangle_D = \mathbf{E}_{X \sim D}[f(X)g(X)]$ and the norm $\|f\|_D = \sqrt{\langle f, f \rangle_D}$.

Definition 40 (Statistical Dimension) *For the hypothesis testing problem of Definition 38, we define the* statistical dimension $\mathrm{SDA}(\mathcal{S}, \mu, n)$ *as follows:*

$$\mathrm{SDA}(\mathcal{S}, \mu, n) = \max \left\{ q \in \mathbb{N} : \mathop{\mathbf{E}}_{u,v \sim \mu} [|\langle \bar{D}_u, \bar{D}_v \rangle_{D_0} - 1| \mid E] \leq \frac{1}{n} \text{ for all events } E \text{ s.t. } \mathop{\mathbf{Pr}}_{u,v \sim \mu}[E] \geq \frac{1}{q^2} \right\} .$$

We will omit writing μ when it is clear from the context.

The following lemma translates the (γ, β)-correlation of $\mathcal{S}$ to a lower bound for the statistical dimension of the hypothesis testing problem. The proof is very similar to that of Corollary 8.28 of Brennan et al. (2020) but it is given below for completeness.

Lemma 41 *Let $0 < c < 1/2$ and $d, m \in \mathbb{Z}_+$. Consider the hypothesis testing problem of Definition 39. Then, for any $q \geq 1$,*

$$\mathrm{SDA}\left(\mathcal{D}, \left(\frac{\epsilon^{-1}\Omega(d)^{1/2-c}}{(1-2\eta)\epsilon(q^2/2^{\Omega(d^{c/2})}+1)}\right)\right) \geq q \, .$$

Proof The first part is to calculate the correlation of the set $\mathcal{S}$. By Theorem 11, we know that the set $\mathcal{S}$ is (γ, β)-correlated with $\gamma = (1-2\eta)\epsilon^2\Omega(d)^{c-1/2}$ and $\beta = 4(1-2\eta)\epsilon$.

We next calculate the SDA according to Definition 40. We denote by $\bar{E}_{\mathbf{v}}$ the ratios of the density of $E_{\mathbf{v}}$ to the density of R. Note that the quantity $\langle \bar{E}_{\mathbf{u}}, \bar{E}_{\mathbf{v}} \rangle - 1$ used there is equal to $\langle \bar{E}_{\mathbf{u}} - 1, \bar{E}_{\mathbf{v}} - 1 \rangle$. Let E be an event that has $\mathbf{Pr}_{\mathbf{u},\mathbf{v} \sim \mu}[E] \geq 1/q^2$. For d sufficiently large we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{u,v \sim \mu}[|\langle \bar{E}_{\mathbf{u}}, \bar{E}_{\mathbf{v}} \rangle - 1| E] &\leq \min\left(1, \frac{1}{|\mathcal{S}| \mathbf{Pr}[E]}\right)\beta + \max\left(0, 1 - \frac{1}{|\mathcal{S}| \mathbf{Pr}[E]}\right)\gamma \\
&\leq (1-2\eta)\epsilon\left(\frac{q^2}{2^{\Omega(d^c)}} + \frac{\epsilon}{\Omega(d)^{1/2-c}}\right) = (1-2\eta)\epsilon\left(\frac{\epsilon^{-1}\Omega(d)^{1/2-c}}{q^2/2^{\Omega(d^{c/2})}+1}\right)^{-1} ,
\end{aligned}$$

where the first inequality uses that $\mathbf{Pr}[\mathbf{u} = \mathbf{v}|E] = \mathbf{Pr}[\mathbf{u} = \mathbf{v}, E]/\mathbf{Pr}[E]$ and bounds the numerator in two different ways: $\mathbf{Pr}[\mathbf{u} = \mathbf{v}, E]/\mathbf{Pr}[E] \leq \mathbf{Pr}[\mathbf{u} = \mathbf{v}]/\mathbf{Pr}[E] = 1/(|\mathcal{S}|\mathbf{Pr}[E])$ and $\mathbf{Pr}[\mathbf{u} = \mathbf{v}, E]/\mathbf{Pr}[E] \leq \mathbf{Pr}[E]/\mathbf{Pr}[E] = 1$. ■

C.1. Preliminaries: Low-Degree Method

We begin by recording the necessary notation, definitions, and facts. This section mostly follows Brennan et al. (2020).

Low-Degree Polynomials A function $f : \mathbb{R}^a \to \mathbb{R}^b$ is a polynomial of degree at most k if it can be written in the form

$$f(x) = (f_1(x), f_2(x), \ldots, f_b(x)) \, ,$$

where each $f_i : \mathbb{R}^a \to \mathbb{R}$ is a polynomial of degree at most k. We allow polynomials to have random coefficients as long as they are independent of the input x. When considering *list-decodable estimation* problems, an algorithm in this model of computation is a polynomial $f : \mathbb{R}^{d_1 \times n} \to \mathbb{R}^{d_2 \times \ell}$, where d_1 is the dimension of each sample, n is the number of samples, d_2 is the dimension of the output hypotheses, and ℓ is the number of hypotheses returned. On the other hand, Brennan et al. (2020) focuses on *binary hypothesis testing* problems defined in Definition 38.

A degree-k polynomial test for Definition 38 is a degree-k polynomial $f : \mathbb{R}^{d \times n} \to \mathbb{R}$ and a threshold $t \in \mathbb{R}$. The corresponding algorithm consists of evaluating f on the input $x_1, \ldots, x_n$ and returning H_0 if and only if $f(x_1, \ldots, x_n) > t$.

Definition 42 (n-sample ϵ-good distinguisher) *We say that the polynomial $p : \mathbb{R}^{d \times n} \mapsto \mathbb{R}$ is an n-sample ϵ-distinguisher for the hypothesis testing problem in Definition 38 if*

$$|\mathop{\mathbf{E}}_{X \sim D_0^{\otimes n}}[p(X)] - \mathop{\mathbf{E}}_{u \sim \mu} \mathop{\mathbf{E}}_{X \sim D_u^{\otimes n}}[p(X)]| \geq \epsilon \sqrt{\mathop{\mathbf{Var}}_{X \sim D_0^{\otimes n}}[p(X)]} \, .$$

We call ϵ the advantage *of the distinguisher.*

Let $\mathcal{C}$ be the linear space of polynomials with a degree at most k. The best possible advantage is given by the *low-degree likelihood ratio*

$$\max_{\substack{p \in \mathcal{C} \\ \mathbf{E}_{X \sim D_0^{\otimes n}}[p^2(X)] \leq 1}} |\mathop{\mathbf{E}}_{u \sim \mu} \mathop{\mathbf{E}}_{X \sim D_u^{\otimes n}}[p(X)] - \mathop{\mathbf{E}}_{X \sim D_0^{\otimes n}}[p(X)]| = \left\| \mathop{\mathbf{E}}_{u \sim \mu} \left[(\bar{D}_u^{\otimes n})^{\leq k} \right] - 1 \right\|_{D_0^{\otimes n}} ,$$

where we denote $\bar{D}_u = D_u / D_0$ and the notation $f^{\leq k}$ denotes the orthogonal projection of f to $\mathcal{C}$.

Another notation we will use regarding a finer notion of degrees is the following: We say that the polynomial $f(x_1, \ldots, x_n) : \mathbb{R}^{d \times n} \to \mathbb{R}$ has *samplewise degree* (r, k) if it is a polynomial, where each monomial uses at most k different samples from $x_1, \ldots, x_n$ and uses degree at most r for each of them. In analogy to what was stated for the best degree-k distinguisher, the best distinguisher of samplewise degree (r, k)-achieves advantage $\left\| \mathbf{E}_{u \sim \mu}[(\bar{D}_u^{\otimes n})^{\leq r,k}] - 1 \right\|_{D_0^{\otimes n}}$ the notation $f^{\leq r,k}$ now means the orthogonal projection of f to the space of all samplewise degree-(r, k) polynomials with unit norm.

C.2. Hardness of Hypothesis Testing Against Low-Degree Polynomials

We restate and prove the following.

Theorem 43 *Let $0 < c < 1/2$. Consider the hypothesis testing problem of Definition 39. For $d \in \mathbb{Z}_+$ with d larger than an absolute constant, any $n \leq \Omega(d)^{1/2-c}/(\epsilon^2(1-2\eta))$ and any even integer $k < d^{c/4}$, we have that*

$$\left\| \mathop{\mathbf{E}}_{\mathbf{v} \sim \mu} \left[(\bar{E}_{\mathbf{v}}^{\otimes n})^{\leq \infty, \Omega(k)} \right] - 1 \right\|_{D_0^{\otimes n}}^2 \leq 1 \, .$$

Proof In Brennan et al. (2020), the following relation between SDA and low-degree likelihood ratio is established.

Theorem 44 (Theorem 4.1 of Brennan et al. (2020)) *Let $\mathcal{D}$ be a hypothesis testing problem on $\mathbb{R}^d$ with respect to null hypothesis D_0. Let $n, k \in \mathbb{N}$ with k even. Suppose that for all $0 \le n' \le n$, $\mathrm{SDA}(\mathcal{S}, n') \ge 100^k (n/n')^k$. Then, for all r, $\left\| \mathbf{E}_{u \sim \mu} \left[(\bar{D}_u^{\otimes n})^{\le r, \Omega(k)} \right] - 1 \right\|_{D_0^{\otimes n}}^2 \le 1$.*

In Lemma 41 we set $n = \Omega(d)^{1/2-c}/(\epsilon^2(1-2\eta))$ and $q = \sqrt{2^{\Omega(d^{c/2})}(n/n')}$. Then, $\mathrm{SDA}(\mathcal{S}, n') \ge \sqrt{2^{\Omega(d^{c/2})}(n/n')} \ge (100n/n')^k$ for $k < d^{c/4}$ and then we apply the theorem above.

■