A Near-optimal SQ Lower Bound for Smoothed Agnostic Learning of Boolean Halfspaces

Tim Sinen
University of Bonn
Bonn, Germany
sinen@uni-bonn.de

Abstract

We study the complexity of smoothed agnostic learning of halfspaces on $\{\pm 1\}^n$ under uniform marginals in the model of Kou and Meka [2025] where each input coordinate is independently flipped with probability $\sigma \in (0, 1/2)$. We show that L^1 polynomial regression achieves runtime and sample complexity $\tilde{O}(n^{O(\log(1/\varepsilon)/\sigma)})$, and prove a nearly matching Statistical Query complexity lower bound of $n^{\Omega(\log(1+\sigma/\varepsilon^2)/\sigma)}$ in the parameter range of Theorem B.10. This complements the recent work of Diakonikolas and Kane [2026], which established analogous bounds in the continuous setting under Gaussian marginals.

1 Introduction

Agnostic learning, introduced by Haussler [1992] and Kearns et al. [1994], is one of the most general frameworks for supervised classification. The learner receives labeled examples (X, Y) drawn i.i.d. from an arbitrary joint distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, where $\mathcal{X}$ is a feature space, and must output a hypothesis $h\colon \mathcal{X} \to \{\pm 1\}$ such that, with probability at least $1 - \delta$, the misclassification error satisfies $\Pr_{(X,Y)\sim\mathcal{D}}[h(X) \neq Y] \leq \mathrm{opt}_{\mathcal{C}} + \varepsilon$, where $\mathrm{opt}_{\mathcal{C}} = \inf_{f\in\mathcal{C}} \Pr_{(X,Y)\sim\mathcal{D}}[f(X) \neq Y]$ is the error of the best classifier in the target class $\mathcal{C}$. Importantly, no assumption is made on the relationship between labels and features. Typically, the computational complexity of a learning algorithm is measured in terms of its running time and sample complexity.

It is well known that without any distributional assumptions, even for the fundamental class of parities or halfspaces, agnostic learning is computationally intractable [Guruswami and Raghavendra, 2009, Feldman et al., 2009, Daniely, 2016]. Under the assumption of structured marginals, positive results are known. Under Gaussian or, more broadly, log-concave marginals, the L^1 polynomial regression algorithm of Kalai et al. [2008] provides the best known agnostic learners for halfspaces [Diakonikolas et al., 2009, 2010, 2021a], but the resulting complexity remains exponential in $1/\varepsilon$, with strong evidence that improving it would require either profound new techniques or would have consequences for other learning problems that are considered hard [Klivans and Kothari, 2014, Diakonikolas et al., 2021b, 2023, Tiegel, 2023].

One approach to bypass this hardness is the framework of *smoothed agnostic learning*, recently introduced by Chandrasekaran et al. [2024], which relaxes the benchmark against which the learner must compete. Informally, in this model with domain $\mathbb{R}^n$, inspired by the smoothed analysis paradigm of Spielman and Teng [2004], the learner's performance is measured relative to $\mathrm{opt}_{\mathcal{C},\sigma} = \inf_{f\in\mathcal{C}} \Pr_{(X,Y)\sim\mathcal{D}}[f(T_\sigma X) \neq Y]$ rather than $\mathrm{opt}_{\mathcal{C}}$, where $T_\sigma X$ denotes an additively perturbed version of X parameterized by $\sigma > 0$, effectively smoothing the instance space.

Kou and Meka [2025] ported this framework to the Boolean hypercube $\{\pm 1\}^n$, where additive perturbations are not well-defined, by replacing them with independent random bit flips of rate σ. We introduce their framework in more detail.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Definition 1.1 (Smoothed agnostic learning on the Boolean hypercube)**.** Fix $\varepsilon > 0, \sigma \in (0, 1/2)$ and $\delta \in (0, 1)$. Let $\mathcal{C} \subseteq \{\{\pm 1\}^n \to \{\pm 1\}\}$ be a set of Boolean concepts, and let $\mathcal{D}$ be a distribution over $\{\pm 1\}^n \times \{\pm 1\}$. We say that an algorithm A *learns* $\mathcal{C}$ in the σ-smoothed setting if, given i.i.d. samples from $\mathcal{D}$, it outputs a hypothesis $h : \{\pm 1\}^n \to \{\pm 1\}$ such that, with probability at least $1 - \delta$, $\Pr_{(X,Y)\sim\mathcal{D}}\big[h(X) \neq Y\big] \leq \mathrm{opt}_{\mathcal{C},\sigma} + \varepsilon$, where

$$\mathrm{opt}_{\mathcal{C},\sigma} := \inf_{f\in\mathcal{C}} \mathbb{E}_{Z\sim B(\sigma)^n}\left[\Pr_{(X,Y)\sim\mathcal{D}}\big[f(X \odot Z) \neq Y\big]\right].$$

Here, $Z \sim B(\sigma)^n$ means each coordinate of $Z \in \{\pm 1\}^n$ is independently -1 (a flip) with probability σ, and $x \odot z$ denotes the Hadamard product.

Note that in this definition, for $\sigma \to 0$ we recover the classical worst-case agnostic learning framework. For $\sigma > 0$, the idea is that the perturbation effectively limits the expressive power of the hypothesis class $\mathcal{C}$ that the learner must compete against, ideally leading to efficient learners with better sample complexity and runtime bounds than known in the worst case. Also, we do not require h to be contained in $\mathcal{C}$, so it is enough to provide an *improper* learner.

Boolean Halfspaces. A Boolean halfspace is a function $f\colon \{\pm 1\}^n \to \{\pm 1\}$ of the form $f(x) = \mathrm{sign}\,(\sum_{i=1}^n w_i x_i - \theta)$, where $w \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$. The majority function $\mathrm{Maj}_n(x) = \mathrm{sign}\,(\sum_{i=1}^n x_i)$ is the halfspace corresponding to uniform weights and zero threshold.

Both in the continuous setting with additive Gaussian perturbations and on the Boolean hypercube with bit-flip noise, halfspaces become efficiently learnable under a wide class of X-marginals $\mathcal{D}_X$, a phenomenon with no known analogue in the worst-case setting. In more detail, the current best sample-complexity upper bounds for smoothed agnostic learning of halfspaces under subexponential marginals are $n^{\tilde{O}(1/\sigma^2)\log(1/\varepsilon)}$ in the continuous setting [Koehler and Wu, 2026] and $\tilde{O}\big(n^{\mathrm{poly}(1/(\sigma\varepsilon))}\big)$ on the Boolean hypercube [Kou and Meka, 2025]. Notably, both bounds essentially rely on L^1 polynomial regression on the smoothed target.

This raises a natural and fundamental question:

Is there a way to improve upon the complexity of L^1 polynomial regression for learning halfspaces in the smoothed agnostic setting?

SQ Lower Bounds. A powerful framework for studying such questions is the Statistical Query (SQ) model of Kearns [1998], where, instead of receiving samples, a learner obtains approximate answers to statistical queries from an oracle.

Definition 1.2 (STAT oracle)**.** For a tolerance parameter $\tau > 0$, $\mathrm{STAT}(\tau)$ is the oracle that, given any query function $h : \mathcal{X} \to [-1, 1]$, returns a value

$$v \in [\mathbb{E}_{X\sim\mathcal{D}}[h(X)] - \tau,\, \mathbb{E}_{X\sim\mathcal{D}}[h(X)] + \tau]\,.$$

Lower bounds in this model are typically stated as a dichotomy: any SQ algorithm must either use $n^{\omega(1)}$ queries to $\mathrm{STAT}(\tau)$ or require a query of tolerance $\tau = n^{-\omega(1)}$. These bounds are of unconditional information-theoretical nature. Crucially, a great number of known algorithmic techniques in machine learning can be captured in the SQ model — including gradient descent, spectral methods, and most approaches to convex optimization (for a survey, see Reyzin [2020]). This makes SQ lower bounds important evidence of computational hardness.

In this work, we ask for SQ lower bounds in the case of smoothed agnostic learning of Boolean halfspaces.

We note that the open question is not whether in this setting a fully polynomial SQ learner exists, as this is already ruled out by worst-case hardness. Indeed, a union bound over the coordinates of Z gives $\mathrm{opt}_{\mathcal{C},\sigma} \leq \mathrm{opt}_{\mathcal{C}} + n\sigma$, so running a smoothed learner with $\sigma = \varepsilon/(2n)$ yields an ordinary agnostic learner with error $\mathrm{opt}_{\mathcal{C}} + \varepsilon$ on the same distribution. Such a learner would use $\mathrm{poly}(n, 1/\sigma, 1/\varepsilon)$ queries and inverse polynomial tolerance, which contradicts known SQ lower bounds for agnostic learning [cf. Kalai et al., 2008, Dachman-Soled et al., 2015]. This argument, however, breaks down for $\sigma = \Theta(1)$ and says nothing about the *fine-grained complexity*: is L^1 regression essentially optimal for smoothed agnostic learning on the Boolean hypercube as well?

Our contribution. We give a positive answer to the last question on the Boolean hypercube.

(i) We prove a Statistical Query lower bound for smoothed agnostic learning of Boolean halfspaces on $\{\pm 1\}^n$ under the uniform distribution, which falls within the strictly subexponential class studied by Kou and Meka [2025]. This matches the upper bound achievable by L^1 polynomial regression up to a $\log(1/\sigma)/\sigma$ summand in the exponent.

(ii) As the central technical ingredient, we establish an L^1 approximation lower bound for the smoothed majority function $T_{1-2\sigma}\mathrm{Maj}_n$ under the uniform measure on $\{\pm 1\}^n$. The proof constructs an explicit dual witness which combines a symmetric trigonometric kernel with weighted Laguerre integration. This construction may be of independent interest for other SQ lower bounds and approximation problems on the Boolean hypercube.

In more detail, we have the following results.

Theorem 1.3 (informal version of Theorem B.10)**.** *Fix a constant smoothing parameter $\sigma \in (0, 0.499]$ and let $1/\operatorname{poly}(n) \le \varepsilon \le \varepsilon_0$ for a sufficiently small absolute constant $\varepsilon_0 > 0$. Assume n is sufficiently large.*

Any SQ algorithm that, given access to an unknown distribution on $\{\pm 1\}^n \times \{\pm 1\}$ with uniform X-marginal, outputs with probability at least $2/3$ a hypothesis with excess error at most ε against the best halfspace in the σ-smoothed setting must either make a STAT query of tolerance at most $n^{-\Omega(\log(1+\sigma/\varepsilon^2)/\sigma)}$ or make $2^{\Omega(\sqrt{n})}$ queries.

Notice that in the regime $\varepsilon \le \sigma$, this gives an SQ lower bound of $n^{\Omega(\log(1/\varepsilon)/\sigma)}$. As $\sigma \to 0$, we obtain a lower bound of $n^{\Omega(1/\varepsilon^2)}$ whenever $\varepsilon = \Omega(n^{-1/4+\delta})$ for any $\delta > 0$. This recovers the optimal[1] SQ lower bound for ordinary agnostic learning of Boolean halfspaces in the worst-case setting under a uniform marginal in that particular regime. On the upper bound side, we have the following result, which establishes that the SQ lower bound in Theorem 1.3 is essentially tight.

Theorem 1.4. *Let $\mathcal{C}$ be any class of Boolean functions on $\{\pm 1\}^n$, and let $\varepsilon > 0$, $\sigma \in (0, 1/2)$ and $\delta \in (0, 1)$. There is an algorithm that, given samples from any distribution $\mathcal{D}$ on $\{\pm 1\}^n \times \{\pm 1\}$ with uniform X-marginal, has sample complexity and running time $\operatorname{poly}(n^{\log(1/\varepsilon)/\sigma}, \log(1/\delta))$ and outputs a Boolean hypothesis h with $\operatorname{err}(h) \le \operatorname{opt}_{\mathcal{C},\sigma} + \varepsilon$ with probability at least $1 - \delta$.*

Remark 1.5. Theorem 1.4 still holds if the uniform X-marginal is replaced by an arbitrary product distribution $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ on $\{\pm 1\}^n$.

2 Related Work

The SQ framework was extended to general computational problems by Feldman et al. [2017], who introduced the notion of statistical dimension and showed that lower bounds on it directly translate into query complexity lower bounds.

In particular, it has been remarkably successful for understanding the limits of distribution-specific agnostic learning. In their seminal work, Dachman-Soled et al. [2015] introduced the notion of *approximate resilience*, equivalently L^1-inapproximability by low-degree polynomials, and showed that it essentially characterizes the SQ complexity of agnostic learning under the uniform distribution on $\{\pm 1\}^n$. This gives a general framework for deriving SQ lower bounds for learning Boolean concepts, and it is the most relevant precursor for our witness construction. The approximate resilience framework was subsequently extended by Hsu et al. [2022] to obtain near-optimal SQ lower bounds for agnostically learning intersections of halfspaces under Gaussian marginals.

Under Gaussian marginals, the main technique is the moment-matching paradigm of Diakonikolas et al. [2017], in its supervised form [Diakonikolas et al., 2020, 2021b]. There, one constructs a pair of one-dimensional distributions, one for each label, whose first m moments are the same as those of a reference distribution while being far apart in total variation distance, and then embeds this structure along hidden directions in high dimensions. It has since yielded strong SQ lower bounds across a range of learning problems [Diakonikolas et al., 2020, 2021b, Diakonikolas and Kane, 2022]. In particular, Diakonikolas et al. [2021b] proved that L^1 polynomial regression is essentially

[1] E.g., the optimal exponent $1/\varepsilon^2$ is achievable via the sparse parity reduction which we also discuss below. We also expect that the SQ lower bound of Diakonikolas et al. [2021b], which is stronger in terms of query complexity, can be transferred to the Boolean hypercube under uniform marginals via the block embedding we also discuss in the following.

optimal among SQ algorithms for agnostically learning halfspaces under Gaussian marginals, and more recently, Diakonikolas and Kane [2026] established the first SQ lower bound for smoothed agnostic learning in the continuous subgaussian setting, as we already briefly discussed above. This provides evidence that even with smoothing, L^1 regression cannot be substantially improved, which is a negative answer to our question in the continuous setting. Our high-level proof idea follows this paradigm, but the construction of Diakonikolas and Kane [2026] does not directly transfer to the Boolean hypercube, as we discuss in Section 4.

The connections between polynomial approximation and hardness in learning theory are not coincidental: at the heart of all such lower bounds lies the question of how well a degree-m polynomial can approximate a target function in a weighted L^p norm. This theme appears across several areas. Most relevantly, Bun and Thaler [2013] and Bun et al. [2018] developed systematic dual-polynomial methods for approximate-degree lower bounds, whose central technical task — bounding the L^1 norm of a dual witness constructed from orthogonal polynomials — is structurally analogous to our normalization problem of bounding the L^∞ norm of the sine-Laguerre witness (Lemma B.5).

Weaker and transferred lower bounds. A weaker bound already follows from sparse parities, in the spirit of Kalai et al. [2008] and Chandrasekaran et al. [2024, Theorem E.7]. For odd k and $|S| = k$, draw x uniformly at random and assign it label $\chi_S(x)$. One of the halfspaces $\pm\mathrm{Maj}_S$ then has smoothed correlation $\rho^k|\widehat{\mathrm{Maj}_S}(S)| = \Theta(\rho^k/\sqrt{k})$ with the labels, so any smoothed agnostic learner with $\varepsilon \le c\rho^k/\sqrt{k}$ must output a hypothesis with correlation to χ_S of $\Omega(\rho^k/\sqrt{k})$. With $\Lambda := \log(1+\sigma/\varepsilon^2)/\sigma$ and $k = \Theta(\Lambda)$, SQ hardness of learning parities [Blum et al., 1994] gives $Q/\tau^2 \ge n^{\Omega(\Lambda)}$, which is the same exponent as in Theorem B.10. However, since there are only $\binom{n}{k} = n^{O(\Lambda)}$ candidate parities, they can be learned with $n^{O(\Lambda)}$ queries of constant tolerance, so this family cannot yield the much stronger query-complexity bound of Theorem B.10, which requires exponentially many nearly uncorrelated distributions.

Alternatively, Gaussian SQ lower bounds can be transferred to the hypercube by coupling each Gaussian coordinate to a block of uniform bits and approximating Gaussian halfspaces by halfspaces on the block sums using Wasserstein bounds, e.g. [Rio, 2009]. We expect that applying this to Diakonikolas and Kane [2026] gives an alternative route to a bound of similar form. Such a transfer, however, inherits the exponent of Diakonikolas and Kane [2026], which is of order $\log(1/\varepsilon) + 1/(\sigma+\varepsilon^2)$ after translating Gaussian to Boolean smoothing.[2] Up to constants, the exponent of Theorem B.10 is at least as large, and exceeds it by a factor of order $\min\{\log(1/\varepsilon), 1/\sigma\}$ whenever $\varepsilon \le \sigma$. To our knowledge, such a multiplicative dependence is not known in the Gaussian setting. Moreover, the transfer does not yield a bounded approximation-resilient witness.

3 Preliminaries

We briefly review the most relevant definitions, facts and notation needed for a high-level overview of our techniques.

Notation. For $n \ge 1$, we set $[n] = \{1, \dots, n\}$. For a measure μ and $p \ge 1$, we write $\|f\|_{L^p(\mu)}$ for the usual L^p norm, $\|f\|_{L^\infty(\mu)}$ for the essential supremum, and $\langle f, g\rangle_{L^2(\mu)} = \int fg\, d\mu$. When μ is clear from context we write $\|f\|_p$ and $\langle f, g\rangle$. Probability distributions are denoted by $\mathcal{D}$, with subscripts and decorations (e.g. $\mathcal{D}_u$, $\mathcal{D}_0$) for indexed families; we write $\mathcal{D}_X$ for the X-marginal of $\mathcal{D}$, and use script $\mathscr{D}$ for classes of distributions. We write U_n for the uniform probability measure on the Boolean hypercube $\{\pm1\}^n$, so that $\|f\|_{L^p(U_n)} = \mathbb{E}_{X\sim U_n}\left[|f(X)|^p\right]^{1/p}$. Finally, $\mathcal{H}_n$ denotes the class of halfspaces on $\{\pm1\}^n$, and for $m \ge 0$, $\mathcal{P}_{\le m}$ is the space of multilinear polynomials on $\{\pm1\}^n$ of degree at most m.

Fourier–Walsh Analysis. Over the Boolean hypercube $\{\pm1\}^n$, for $S \subseteq [n]$, we set $\chi_S(x) = \prod_{i\in S} x_i$. The family $\{\chi_S\}_{S\subseteq[n]}$ is an orthonormal basis of $L^2(U_n)$, that is, every $f : \{\pm1\}^n \to \mathbb{R}$

[2] In Diakonikolas and Kane [2026], the exponent is $\log(1/\varepsilon) + (s+\varepsilon)^{-2}$ for additive Gaussian noise of standard deviation s. Flipping each bit with probability σ acts on a block sum like Gaussian noise of standard deviation $\Theta(\sqrt{\sigma})$, so $s = \Theta(\sqrt{\sigma})$.

admits a unique expansion

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S(x)$$

where $\hat{f}(S) = \mathbb{E}_{X \sim U_n}[f(X)\chi_S(X)]$. The degree of the expansion is $\deg(f) = \max\{|S| \mid \hat{f}(S) \neq 0\}$. Parseval's identity gives $\|f\|_2^2 = \sum_S \hat{f}(S)^2$. A Boolean function $f : \{\pm 1\}^n \to \mathbb{R}$ is called *symmetric* if $f(x_{\pi(1)}, \ldots, x_{\pi(n)}) = f(x_1, \ldots, x_n)$ for every permutation $\pi \in S_n$. For such functions, it is convenient to organize the Fourier basis by level. The *symmetric Fourier mode* of level k is

$$\Psi_{k,n}(x) := \binom{n}{k}^{-1/2} \sum_{|S|=k} \chi_S(x), \quad 0 \leq k \leq n$$

and the family $\{\Psi_{k,n}\}_{k=0}^n$ is orthonormal in $L^2(U_n)$.

Noise operator. For $\rho \in [0,1]$, the noise operator is $T_\rho f(x) := \mathbb{E}_Z[f(x \odot Z)]$ with $Z \sim B((1-\rho)/2)^n$, i.e., each coordinate of x is flipped independently with probability $(1-\rho)/2$. Equivalently, in the Fourier–Walsh basis, $T_\rho f = \sum_S \rho^{|S|} \hat{f}(S)\chi_S$. In particular, σ-smoothing as in Definition 1.1 corresponds to T_ρ with $\rho = 1 - 2\sigma$.

SDA Framework. In following, let $\mathcal{D}$ be a distribution over an input space $\mathcal{X}$.

Definition 3.1 (VSTAT Oracle, Feldman et al., 2017, Definition 2.3)**.** For a sample size parameter $t > 0$, $\mathrm{VSTAT}(t)$ is the oracle that, given any query function $h : \mathcal{X} \to [0,1]$, returns a value $v \in [p - \tau, p + \tau]$, where $p = \mathbb{E}_{X \sim \mathcal{D}}[h(X)]$ and

$$\tau = \max\left\{\frac{1}{t}, \sqrt{\frac{p(1-p)}{t}}\right\}.$$

We observe that $\mathrm{VSTAT}(t)$ always returns the expectation within $1/\sqrt{t}$, so it is at least as powerful as $\mathrm{STAT}(1/\sqrt{t})$ and at most as powerful as $\mathrm{STAT}(1/t)$. This allows us to convert VSTAT lower bounds to lower bounds for the STAT oracle, after rescaling the query answer.

For two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{X}$, the *pairwise correlation* relative to $\mathcal{D}$ is

$$\chi_{\mathcal{D}}(\mathcal{D}_1, \mathcal{D}_2) = \left|\left\langle \frac{\mathcal{D}_1}{\mathcal{D}} - 1, \frac{\mathcal{D}_2}{\mathcal{D}} - 1 \right\rangle_{\mathcal{D}}\right|.$$

When $\mathcal{D}_1 = \mathcal{D}_2$, this is equivalent to the χ^2-divergence $\chi^2(\mathcal{D}_1 \| \mathcal{D})$. The *average correlation* of a non-empty set of distributions $\mathscr{D}'$ relative to $\mathcal{D}$ is

$$\gamma(\mathscr{D}', \mathcal{D}) = \frac{1}{|\mathscr{D}'|^2} \sum_{\mathcal{D}_1, \mathcal{D}_2 \in \mathscr{D}'} \chi_{\mathcal{D}}(\mathcal{D}_1, \mathcal{D}_2).$$

Definition 3.2 (Statistical dimension of a set of distributions, Feldman et al., 2017, Definition 2.5)**.** For $\bar{\gamma} > 0$, a set of distributions $\mathscr{D}$ over $\mathcal{X}$, and a reference distribution $\mathcal{D}$ over $\mathcal{X}$, the *statistical dimension of $\mathscr{D}$ relative to $\mathcal{D}$ with average correlation $\bar{\gamma}$*, denoted $\mathrm{SDA}(\mathscr{D}, \mathcal{D}, \bar{\gamma})$, is the largest d such that $\gamma(\mathscr{D}', \mathcal{D}) \leq \bar{\gamma}$ for any subset $\mathscr{D}' \subseteq \mathscr{D}$ with $|\mathscr{D}'| \geq |\mathscr{D}|/d$.

Intuitively, $\mathrm{SDA}(\mathscr{D}, \mathcal{D}, \bar{\gamma})$ corresponds to the size of a "distinguishable" family relative to $\mathcal{D}$. A larger value of d indicates greater difficulty for statistical queries to distinguish members of $\mathscr{D}$ from $\mathcal{D}$.

Definition 3.3 (Search problem over distributions, Feldman et al., 2017, Definition 2.6)**.** A *search problem* P over a set of solutions F and a class of distributions $\mathscr{D}$ over $\mathcal{X}$ is a map $P : \mathscr{D} \to 2^F$ that assigns to each $\mathcal{D} \in \mathscr{D}$ a non-empty set $P_{\mathcal{D}} \subseteq F$ of *valid solutions*. An algorithm *solves* P if, given access to an unknown $\mathcal{D} \in \mathscr{D}$, it outputs some $f \in P_{\mathcal{D}}$. For $f \in F$, we write $P_f := \{\mathcal{D} \in \mathscr{D} \mid f \in P_{\mathcal{D}}\}$ for the set of distributions for which f is a valid solution.

Definition 3.4 (Statistical dimension of a search problem, Feldman et al., 2017, Definition 3.1)**.** Let P be a search problem over a set of solutions F and a class of distributions $\mathscr{D}$ over $\mathcal{X}$. For $\bar{\gamma} > 0$ and $\eta \in (0,1)$, the *statistical dimension with average correlation $\bar{\gamma}$ and solution set bound η*, denoted

$\mathrm{SDA}(P, \bar{\gamma}, \eta)$, is the largest d such that there exists a reference distribution $\mathcal{D}$ over $\mathcal{X}$ and a finite set $\mathscr{D}_{\mathcal{D}} \subseteq \mathscr{D}$ with the following property: for every solution $f \in F$, the set $\mathscr{D}_f = \mathscr{D}_{\mathcal{D}} \setminus P_f$ satisfies $|\mathscr{D}_f| \geq (1-\eta) \cdot |\mathscr{D}_{\mathcal{D}}|$ and $\mathrm{SDA}(\mathscr{D}_f, \mathcal{D}, \bar{\gamma}) \geq d$.

Note that the witnessing family $\mathscr{D}_{\mathcal{D}}$ is chosen so that no fixed solution $f \in F$ solves more than an η-fraction of its distributions, and the remaining unsolved distributions $\mathscr{D}_f$ still have statistical dimension at least d relative to $\mathcal{D}$. Thus, a large value of $\mathrm{SDA}(P, \bar{\gamma}, \eta)$ means that every candidate solution fails on a large subfamily that still has high statistical dimension. This implies a lower bound on any SQ algorithm by the following theorem.

Theorem 3.5 (Feldman et al., 2017, Theorem 3.2)**.** *Let P be a search problem over a set of solutions F and a class of distributions $\mathscr{D}$ over $\mathcal{X}$. For $\bar{\gamma} > 0$ and $\eta \in (0,1)$, let $d = \mathrm{SDA}(P, \bar{\gamma}, \eta)$. Any randomized SQ algorithm that solves P with probability $\alpha > \eta$ requires at least*

$$\frac{\alpha - \eta}{1 - \eta} \cdot d$$

calls to $\mathrm{VSTAT}(1/(3\bar{\gamma}))$.

The L^1 regression algorithm. The algorithmic baseline for agnostic learning under structured marginals is the L^1 polynomial regression algorithm of Kalai et al. [2008]. Given N samples (x_i, y_i) from a distribution $\mathcal{D}$ over $\{\pm 1\}^n \times \{\pm 1\}$, the algorithm solves the linear program

$$\min_{p \in \mathcal{P}_{\leq d}} \frac{1}{N} \sum_{i=1}^{N} |p(x_i) - y_i|$$

over polynomials of degree at most d, and outputs $h(x) = \mathrm{sign}(p(x) - t)$ for a suitable threshold t. If every target f in a concept class $\mathcal{C}$ can be approximated by a polynomial p_f of degree d such that

$$\|f - p_f\|_{L^1(\mathcal{D}_X)} \leq \varepsilon,$$

then the algorithm achieves error $\mathrm{opt}_{\mathcal{C}} + O(\varepsilon)$ with probability at least $1 - \delta$ with runtime and sample complexity of $\mathrm{poly}(n^d, \log(1/\delta))$.

4 Technical Overview

In this section, we give an overview of the structure of our proof for Theorem 1.3 and a proof sketch of Theorem 1.4. Throughout, we set $\rho := 1 - 2\sigma$, and for a distribution $\mathcal{D}$ over $\{\pm 1\}^n \times \{\pm 1\}$ we write $\mathrm{opt}_{\sigma,\mathcal{D}} := \mathrm{opt}_{\mathcal{H}_n,\sigma}$ for the smoothed optimum over Boolean halfspaces (cf. Definition 1.1).

Also, we write $\mathrm{err}_{\mathcal{D}}(h) = \Pr_{(X,Y)\sim\mathcal{D}}[h(X) \neq Y]$ for the true error of a hypothesis h. When the distribution $\mathcal{D}$ is clear from context, we drop the $\mathcal{D}$ in these definitions.

At a high level, the SQ lower bound has two main ingredients. First, we establish an L^1 approximation lower bound for the smoothed majority $T_\rho \mathrm{Maj}_n$. This is done by constructing an explicit dual witness which is bounded in L^∞, orthogonal to all low-degree polynomials, and still has nontrivial correlation with $T_\rho \mathrm{Maj}_n$. Second, we embed this one-dimensional hard instance into many different directions $u \in \{\pm 1\}^n$, hidden through a large family of labeled distributions $\{\mathcal{D}_u\}_{u \in \mathcal{U}}$. That is, for every $u \in \mathcal{U}$ the labels and the smoothed halfspace $T_\rho \mathrm{Maj}_n(u \odot x)$, whose normal vector is parallel to u, have detectable correlation under $\mathcal{D}_u$. Learning the halfspace therefore effectively corresponds to finding the hidden direction. We then exploit the standard *many-vs-one* trick used by Feldman et al. [2017]: distributions encoding different directions have very small pairwise correlation relative to a reference distribution $\mathcal{D}_0$, so no SQ learner can efficiently distinguish them from this reference, which makes finding the hidden direction hard. The statistical-dimension framework of Feldman et al. [2017] then converts the precise correlation bounds into an SQ lower bound.

Note that the restriction to odd n in our argument is purely for convenience and does not affect the asymptotic bounds; the case of even n reduces to odd $n - 1$ by ignoring one coordinate.

Technical comparison with Diakonikolas and Kane [2026]. In the continuous Gaussian setting, Diakonikolas and Kane [2026] embed a one-dimensional hard instance into a random direction

of $\mathbb{R}^n$ via the Non-Gaussian Component Analysis (NGCA) framework, and establish a lower bound via an implicit Hermite witness for the smoothed sign function, combined with concentration arguments to pass from L^2 to L^1. On the hypercube, a discrete analogue of the embedding step via hidden juntas was developed by Diakonikolas et al. [2022b] and used by Diakonikolas et al. [2022a] to obtain SQ lower bounds for Boolean halfspaces with Massart noise. Our embedding via the sign flips $x \mapsto u \odot x$ is an alternative whose instances are symmetric functions of $u \odot x$ depending on all n coordinates. This avoids the polynomial dimension overhead of the junta construction and yields a slightly better trade-off between the size of the family and the pairwise correlations. The main obstacle is the witness, as the symmetric Fourier modes that diagonalize the noise operator T_ρ are not uniformly bounded, so the Hermite-based argument has no direct counterpart. We instead construct an explicit witness within the class of bounded functions.

4.1 Construction of the Sine-Laguerre Dual Witness

The foundation of the hardness result is an L^1 approximation lower bound for the smoothed majority $T_\rho \mathrm{Maj}_n$ against polynomials of degree $\leq m$, that is, we want to lower bound

$$\inf_{\deg(p)\leq m} \|T_\rho \mathrm{Maj}_n - p\|_{L^1(U_n)}.$$

A *dual witness* is a function $\psi : \{\pm 1\}^n \to [-1, 1]$ with $\psi \perp \mathcal{P}_{\leq m}$. For any such ψ, weak duality gives

$$\|T_\rho \mathrm{Maj}_n - p\|_{L^1(U_n)} \;\geq\; \langle T_\rho \mathrm{Maj}_n - p, \psi\rangle \;=\; \langle T_\rho \mathrm{Maj}_n, \psi\rangle$$

for every $p \in \mathcal{P}_{\leq m}$, so it suffices to construct a witness ψ with large enough correlation with the smoothed majority.

Construction (informal). The smoothed majority is a symmetric odd function; it is natural to ask this of the witness as well. Moreover, every Fourier coefficient of the majority of level $2r + 1$ has sign $(-1)^r$ [cf. O'Donnell, 2014, Chapter 5], so we need to match this oscillatory behavior in order to obtain a satisfyingly large correlation. A natural first candidate is the sine kernel $\sin(\theta \sum_{i=1}^n x_i)$ for some scaling $\theta \in \mathbb{R}$, having Fourier coefficient $(-1)^r (\sin\theta)^{2r+1} (\cos\theta)^{n-2r-1}$ at level $2r + 1$, which indeed gives the desired oscillation; the function is odd, symmetric and bounded by 1 by definition. The only thing missing is the required orthogonality to lower levels. The trick lies in the observation that by the reparameterization $\theta = \arctan\sqrt{y/n}$, the Fourier coefficient becomes $(-1)^r n^{-r-1/2} y^{r+1/2} (1 + y/n)^{-n/2}$, so for vanishing on the levels $1 \leq 2r + 1 \leq m$, it is sufficient to find a weight that cancels the moments $y^{1/2}, y^{3/2}, \ldots, y^{k-1/2}$ with respect to $(1 + y/n)^{-n/2}\, dy$, where $k := \lfloor (m + 1)/2 \rfloor$. This is accomplished by the weight $(-1)^k L_k^{(1/2)}(y)(1 + y/n)^{n/2} e^{-y}$, using Laguerre orthogonality with respect to $y^{1/2} e^{-y}\, dy$. Averaging the resulting functions for indices k and $k + 1$ and normalizing gives the *sine-Laguerre witness* $\psi_{k,n}$ (cf. Definition B.1). The particular choice of the Laguerre polynomial (in fact, other polynomial families are conceivable for establishing the orthogonality as well) is motivated by the observation that under the substitution $y = s^2$ the averaged Laguerre kernel becomes a weighted Hermite function, whose Fourier transform is uniformly bounded. This becomes pivotal in verifying that the weighting does not blow up the L^∞ norm of the witness, so that renormalization preserves correlation with $T_\rho \mathrm{Maj}_n$, i.e., we show that the witness resulting from the averaging with the Laguerre weight has norm $O(1)$ in L^∞ (cf. Lemma B.5).

Using the explicit Fourier representations for both our witness and the majority, we obtain the following lower bound.

Lemma 4.1 (informal version of Lemma B.3)**.** *Fix $c > 0$ and $\eta \in (0, 1/2)$. For every $\rho \in (0, 1)$, sufficiently large odd n and integer $1 \leq m \leq cn^{1/2-\eta}$, it holds*

$$\inf_{\deg p\leq m} \|T_\rho \mathrm{Maj}_n - p\|_{L^1(U_n)} \geq \Omega_{c,\eta}(\rho^{2m+1}/\sqrt{m}).$$

It is sharp in the sense that we cannot hope for a better dependence on the exponent of m in the denominator, see the discussion in the last section.

4.2 Reduction to a Search Problem

Next, we embed ψ into a hidden-direction family $\{\mathcal{D}_u\}_{u\in\mathcal{U}}$. We show that any smoothed agnostic learner with small excess error solves the resulting hidden-direction search problem for recovering the direction $u \in \mathcal{U}$ when the SQ learner has access to an unknown $\mathcal{D}_u$ of this family. For this, it is important to recognize that over the Boolean hypercube, the smoothed error of a halfspace g can be written in terms of its smoothed correlation with the labels as

$$\mathbb{E}_{Z\sim B(\sigma)^n} \Pr_{(X,Y)\sim\mathcal{D}}[g(X\odot Z)\neq Y] = \frac{1}{2}\left(1-\mathbb{E}_{(X,Y)\sim\mathcal{D}}[Y\,T_\rho g(X)]\right). \tag{S}$$

Lemma 4.2 (Hidden-direction family, informal version of Lemma B.8)**.** *Fix $c>0$ and $\eta\in(0,1/2)$. Let n be odd and large enough, $1\le m\le cn^{1/2-\eta}$ be an integer, and $k=\lfloor(m+1)/2\rfloor$. Let $\psi=\psi_{k,n}:\{\pm1\}^n\to[-1,1]$ be the sine-Laguerre witness from Lemma 4.1. For $u\in\{\pm1\}^n$, define the u-rotated witness $\psi^{(u)}(x):=\psi(u\odot x)$ and the planted distribution*

$$\mathcal{D}_0(x,y):=2^{-(n+1)},\qquad \mathcal{D}_u(x,y):=2^{-(n+1)}\big(1+y\,\psi^{(u)}(x)\big)$$

for $x\in\{\pm1\}^n$ and $y\in\{\pm1\}$. Then there exists $\mathcal{U}\subseteq\{\pm1\}^n$ of size $|\mathcal{U}|\ge\exp(\Omega(\sqrt{n}))$ with $|\langle u,v\rangle|\le n^{3/4}$ for all distinct $u,v\in\mathcal{U}$, such that:

(i) For every $h:\{\pm1\}^n\to[-1,1]$ and every $u\in\mathcal{U}$,

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}_u}[Yh(X)] \;=\; \langle h,\psi^{(u)}\rangle_{U_n}.$$

(ii) For all distinct $u,v\in\mathcal{U}$, $\chi_{\mathcal{D}_0}(\mathcal{D}_u,\mathcal{D}_v)=|\langle\psi^{(u)},\psi^{(v)}\rangle_{U_n}|\le n^{-\Omega(m)}$.

Lemma 4.2 makes the link between smoothed correlation with the labels and the search problem framework of the SDA bound precise: Let P be the problem whose instances are the distributions $\mathcal{D}_u$, $u\in\mathcal{U}$, and whose valid solutions are hypotheses h satisfying $\mathbb{E}_{\mathcal{D}_u}[Yh(X)]\ge\varepsilon$. Then Lemma 4.2(i) tells us that h solves $\mathcal{D}_u$ if and only if $\langle h,\psi^{(u)}\rangle_{U_n}\ge\varepsilon$. For this search problem, the usual SQ framework gives the following lower bound.

Theorem 4.3 (Hidden-direction SQ lower bound, informal version of Theorem B.9)**.** *Let $\{\mathcal{D}_u\}_{u\in\mathcal{U}}$ be the distribution family from Lemma 4.2, and let γ_n be as in Lemma B.8. Assume $\varepsilon>0$, $\varepsilon^2\ge2\gamma_n$, and that every $\mathcal{D}_u$ admits a valid solution. Then any randomized SQ algorithm which, given access to an unknown $\mathcal{D}_u$, outputs with probability at least $2/3$ a hypothesis $h:\{\pm1\}^n\to[-1,1]$ satisfying*

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}_u}[Yh(X)]\ge\varepsilon$$

must either make $2^{\Omega(\sqrt{n})}$ STAT queries, or use tolerance $\tau\le n^{-\Omega(m)}$.

Note that Theorem 4.3 applies to improper learners which return values in $[-1,1]$; Theorem 1.3 likewise extends to such learners if its error guarantee $\mathrm{err}_{\mathcal{D}}(h)\le\mathrm{opt}_{\sigma,\mathcal{D}}+\varepsilon$ is replaced by the correlation guarantee $\mathbb{E}_{\mathcal{D}}[Yh(X)]\ge2\varepsilon$.

Theorem 4.3 follows from two facts. First, since the $\psi^{(u)}$ are nearly orthogonal by Lemma 4.2(ii), a Cauchy–Schwarz argument shows that every hypothesis solves at most $O(1/\varepsilon^2)$ of the distributions $\{\mathcal{D}_u\}_{u\in\mathcal{U}}$.

Second, every sufficiently large subfamily has average pairwise correlation at most $O(\bar\gamma)$, where $\bar\gamma=n^{-\Theta(m)}$ is the decay rate from Lemma 4.2(ii). In particular, the distributions not solved by a fixed hypothesis still form a family of high statistical dimension, i.e., for $\eta=o(1)$, the SDA framework of Feldman et al. [2017] gives

$$\mathrm{SDA}(P,\,O(\bar\gamma),\eta)\;\gtrsim\;|\mathcal{U}|\bar\gamma,$$

so since $|\mathcal{U}|\ge\exp(\Omega(\sqrt{n}))$ and $\bar\gamma\ge n^{-O(m)}$, this is $2^{\Omega(\sqrt{n})}$ in the regime $m\log n=o(\sqrt{n})$. The matching tolerance bound $\tau\le n^{-\Omega(m)}$ follows from Theorem 3.5 and converting its VSTAT bound to STAT.

We now have all the ingredients to sketch the proof of our main result, which is a corollary of Theorem 4.3.

Proof sketch of Theorem 1.3. By Lemma 4.1, a short calculation shows that we can choose $m = \Theta(\log(1+\sigma/\varepsilon^2)/\sigma)$ such that $\kappa_m := \langle T_\rho \mathrm{Maj}_n, \psi_{k,n}\rangle_{U_n} \geq 4\varepsilon$, where $k = \lfloor (m+1)/2 \rfloor$. The choice in Theorem B.10 also ensures $\gamma_n \leq 2\varepsilon^2$.

The theorem follows by relating smoothed agnostic learning to the hidden-direction search problem. Each distribution $\mathcal{D}_u$ has a uniform X-marginal on $\{\pm 1\}^n$; a lower bound against the family $\{\mathcal{D}_u\}_{u\in\mathcal{U}}$ therefore implies the same lower bound for smoothed agnostic learning under a uniform X-marginal. For $u \in \mathcal{U}$, we write $g_u(x) := \mathrm{sign}(\langle u, x\rangle)$.

Since g_u is a halfspace, (S) gives $\mathrm{opt}_{\sigma,\mathcal{D}_u} \leq \frac{1}{2}\big(1 - \mathbb{E}_{\mathcal{D}_u}[Y\, T_\rho g_u(X)]\big)$. Hence any Boolean hypothesis h with $\mathrm{err}_{\mathcal{D}_u}(h) \leq \mathrm{opt}_{\sigma,\mathcal{D}_u} + \varepsilon$ satisfies

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_u}[Yh(X)] &= 1 - 2\,\mathrm{err}_{\mathcal{D}_u}(h) \\
&\geq \mathbb{E}_{\mathcal{D}_u}[Y\, T_\rho g_u(X)] - 2\varepsilon \\
&= \langle T_\rho g_u, \psi^{(u)}\rangle_{U_n} - 2\varepsilon \\
&= \langle T_\rho \mathrm{Maj}_n, \psi\rangle_{U_n} - 2\varepsilon \\
&= \kappa_m - 2\varepsilon \\
&\geq 2\varepsilon,
\end{aligned}$$

where the second equality is Lemma 4.2(i) applied to $T_\rho g_u$, and the third is the rotation symmetry $x \mapsto u \odot x$.

Thus, the above learner solves the hidden-direction search problem P from Theorem 4.3, in particular with correlation threshold 2ε. By Theorem 4.3, any such SQ algorithm must either use a query with tolerance $\tau \leq n^{-\Omega(m)}$ or make $2^{\Omega(\sqrt{n})}$ queries to STAT. Substituting the above choice of m gives the claimed SQ lower bound. □

4.3 Proof Sketch of Theorem 1.4

Finally, we give a proof sketch of Theorem 1.4, given by a direct application of the L^1 regression algorithm of Kalai et al. [2008].

Proof sketch. For any $f \in \mathcal{C}$, let p_d be the Fourier truncation of $T_\rho f$ at degree d. Parseval's identity and the Cauchy–Schwarz inequality bound the tail by $\|T_\rho f - p_d\|_1 \leq \rho^{d+1}$. With $d = O(\log(1/\varepsilon)/\sigma)$ this is $O(\varepsilon)$, after which the L^1 regression algorithm of Kalai et al. [2008] outputs a Boolean hypothesis h with $\mathrm{err}(h) \leq \mathrm{opt}_{\mathcal{C},\sigma} + \varepsilon$ via a standard thresholding step. □

5 Conclusion and Future Directions

We studied the complexity of smoothed agnostic learning of Boolean halfspaces for distributions with uniform X-marginal. On the algorithmic side, the standard L^1 polynomial regression method gives a learner with sample complexity and running time $\mathrm{poly}\big(n^{\log(1/\varepsilon)/\sigma}, \log(1/\delta)\big)$. Our main result shows that this dependence is essentially unavoidable for SQ algorithms in the parameter range of Theorem B.10: Any SQ learner must either use tolerance $\tau \leq n^{-\Omega(\log(1+\sigma/\varepsilon^2)/\sigma)}$ or make $2^{\Omega(\sqrt{n})}$ queries.

Several questions remain open. First, on the lower bound side, it would be interesting to close the remaining gaps in Theorem B.10 compared to the upper bound. For example, a logarithmic gap in the exponent remains when $\sigma = \Theta(\varepsilon^2)$. Closing it via our approach would require a lower bound in Lemma 4.1 that decays more slowly in m than $\rho^{2m+1}/\sqrt{m}$ for fixed $\rho < 1$. An obvious strengthening to $\inf_{\deg(p)\leq m} \|T_\rho \mathrm{Maj}_n - p\|_{L^1(U_n)} \geq \rho^m$ fails for $\rho \to 1$, since every halfspace on $\{\pm 1\}^n$ has a polynomial L^1 approximation of degree at most m with error $\tilde{O}(1/\sqrt{m})$, uniformly in n [Diakonikolas et al., 2009, 2010].

Second, one could ask whether stronger SQ lower bounds hold for more specific Boolean marginals beyond the uniform distribution, or if it is possible to improve the upper bounds of Kou and Meka [2025] for general subexponential marginals, where the best known dependence $n^{\mathrm{poly}(1/(\sigma\varepsilon))}$ remains far from the $n^{O(\log(1/\varepsilon)/\sigma)}$ upper bound of the uniform case.

As mentioned in the introduction, we believe that the idea behind the construction of the sine-Laguerre witness might be applicable to obtain SQ lower bounds for other smoothed settings, e.g. intersections of halfspaces or ReLUs.

Additionally, motivated by the worst-case agnostic literature on approximate learners [Daniely, 2015, Awasthi et al., 2017], it is natural to ask whether smoothed agnostic learners with relaxed guarantees of the form $O(\mathrm{opt}_\sigma) + \varepsilon$ or $\mathrm{opt}_\sigma^{O(1)} + \varepsilon$ admit faster algorithms.

Finally, an important direction is to understand whether non-SQ algorithms can bypass the SQ barrier, or whether L^1 polynomial regression is optimal here in a stronger computational sense. This question is fundamental to agnostic learning more broadly, well beyond the smoothed setting.

Acknowledgments

The author thanks Heiko Röglin for helpful discussions and feedback on early drafts of this work as well as the anonymous reviewers for helping to improve the quality of the presentation of the results. This work was supported by the Lamarr Institute for Machine Learning and Artificial Intelligence.

References

Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM*, 63(6):50:1–50:27, 2017.

Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the 26th Annual ACM Symposium on Theory of Computing (STOC)*, pages 253–262, 1994.

Mark Bun and Justin Thaler. Dual lower bounds for approximate degree and Markov–Bernstein inequalities. In *Proceedings of the 40th International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 7965 of *Lecture Notes in Computer Science*, pages 303–314, 2013.

Mark Bun, Robin Kothari, and Justin Thaler. The polynomial method strikes back: Tight quantum query bounds via dual polynomials. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing (STOC)*, pages 297–310, 2018.

Gautam Chandrasekaran, Adam R. Klivans, Vasilis Kontonis, Raghu Meka, and Konstantinos Stavropoulos. Smoothed analysis for learning concepts with low intrinsic dimension. In *Proceedings of the 37th Conference on Learning Theory (COLT)*, volume 247 of *Proceedings of Machine Learning Research*, pages 876–922, 2024.

Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 498–511, 2015.

Amit Daniely. A PTAS for agnostically learning halfspaces. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pages 484–502, 2015.

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

Ilias Diakonikolas and Daniel M. Kane. Near-optimal statistical query hardness of learning halfspaces with Massart noise. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pages 4258–4282, 2022.

Ilias Diakonikolas and Daniel M. Kane. Statistical query lower bounds for smoothed agnostic learning, 2026. Preprint.

Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 171–180, 2009.

Ilias Diakonikolas, Daniel M. Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2010.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, 2017.

Ilias Diakonikolas, Daniel M. Kane, and Nikos Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under Gaussian marginals. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostic proper learning of halfspaces under Gaussian marginals. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pages 1522–1551, 2021a.

Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under Gaussian marginals in the SQ model. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pages 1552–1584, 2021b.

Ilias Diakonikolas, Daniel M. Kane, Lisheng Ren, and Yuxin Sun. SQ lower bounds for learning single neurons with Massart noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24006–24018, 2022a.

Ilias Diakonikolas, Daniel M. Kane, and Yuxin Sun. Optimal SQ lower bounds for robustly learning discrete product distributions and Ising models. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pages 3936–3978, 2022b.

Ilias Diakonikolas, Daniel M. Kane, and Lisheng Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and ReLU regression under Gaussian marginals. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 7922–7938, 2023.

Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S. Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM*, 64(2):8:1–8:37, 2017.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

Daniel Hsu, Clayton Sanford, Rocco A. Servedio, and Emmanouil V. Vlatakis-Gkaragkounis. Near-optimal statistical query lower bounds for agnostically learning intersections of halfspaces with Gaussian marginals. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pages 283–312, 2022.

Adam T. Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

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

Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2–3):115–141, 1994.

Adam R. Klivans and Pravesh Kothari. Embedding hard learning problems into Gaussian space. In *Proceedings of Approximation, Randomization, and Combinatorial Optimization (APPROX/RANDOM)*, volume 28 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 793–809, 2014.

Frederic Koehler and Beining Wu. Constructive approximation under Carleman's condition, with applications to smoothed analysis. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing (STOC)*, 2026.

Roelof Koekoek, Peter A. Lesky, and René F. Swarttouw. *Hypergeometric Orthogonal Polynomials and Their q-Analogues*. Springer Monographs in Mathematics. Springer, 2010.

Yiwen Kou and Raghu Meka. Smoothed agnostic learning of halfspaces over the hypercube. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

Lev Reyzin. Statistical queries and statistical algorithms: Foundations and applications, 2020. Preprint.

Emmanuel Rio. Upper bounds for minimal distances in the central limit theorem. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45(3):802–817, 2009.

Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004.

Gábor Szegő. *Orthogonal Polynomials*, volume 23 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, 4th edition, 1975.

Stefan Tiegel. Hardness of agnostically learning halfspaces from worst-case lattice problems. In *Proceedings of the 36th Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pages 3029–3064, 2023.

Haiyong Wang and Zhimin Zhang. Superconvergence points of Hermite spectral interpolation. *Journal of Scientific Computing*, 109(15), 2026.

A Some Facts about Orthogonal Polynomials

This appendix collects some standard identities for the Hermite, Laguerre, and Krawtchouk families used in the deferred proofs of the paper. Many of them can be found in the literature, e.g. [Szegő, 1975, Koekoek et al., 2010].

Throughout, we write

$$\widehat{f}(\xi) := (2\pi)^{-1/2} \int_{\mathbb{R}} f(x)\, e^{-ix\xi}\, dx \tag{A.1}$$

for the unitary Fourier transform of $f \in L^2(\mathbb{R})$.

A.1 Hermite Polynomials and Hermite Functions

The *physicist's* Hermite polynomials are given by the Rodrigues formula as

$$H_k(x) := (-1)^k e^{x^2} \frac{d^k}{dx^k} e^{-x^2}, \qquad k \geq 0. \tag{A.2}$$

The associated normalized Hermite functions

$$\Phi_k(x) := \left(2^k k! \sqrt{\pi}\right)^{-1/2} H_k(x)\, e^{-x^2/2} \tag{A.3}$$

form an orthonormal basis of $L^2(\mathbb{R})$, satisfy the Hermite ODE

$$-\Phi_k''(x) + x^2 \Phi_k(x) \;=\; (2k+1)\, \Phi_k(x) \tag{A.4}$$

and are eigenfunctions of the (unitary) Fourier transform,

$$\widehat{\Phi_k}(\xi) \;=\; (-i)^k\, \Phi_k(\xi). \tag{A.5}$$

Additionally, they obey the ladder identities

$$x\, \Phi_k \;=\; \sqrt{\tfrac{k+1}{2}}\, \Phi_{k+1} + \sqrt{\tfrac{k}{2}}\, \Phi_{k-1} \tag{A.6}$$

and

$$\Phi_k' \;=\; \sqrt{\tfrac{k}{2}}\, \Phi_{k-1} - \sqrt{\tfrac{k+1}{2}}\, \Phi_{k+1} \tag{A.7}$$

with the convention $\Phi_{-1} \equiv 0$.

Finally, we need the following basic Hermite estimates.

Lemma A.1. *There exists an absolute constant $C > 0$ such that for all $n \geq 0$,*

$$\|\Phi_n'\|_{L^\infty(\mathbb{R})} \leq C\, (n+1)^{1/4} \tag{A.8}$$

and

$$\|\Phi_n\|_{L^1(\mathbb{R})} \leq C\, (n+1)^{1/4}. \tag{A.9}$$

Proof. The bound (A.8) is the case $k = 1$ of [Wang and Zhang, 2026, Lemma 3.2]. For (A.9), the Cauchy–Schwarz inequality yields

$$\begin{aligned} \|\Phi_n\|_{L^1(\mathbb{R})} &\leq \left(\int_{\mathbb{R}} \frac{dx}{x^2+n+1}\right)^{1/2} \left((n+1)\|\Phi_n\|_2^2 + \|x\Phi_n\|_2^2\right)^{1/2} \\ &\overset{(A.6)}{\leq} \left(\frac{\pi}{(n+1)^{1/2}}\right)^{1/2} \left(2(n+1)\right)^{1/2} = \sqrt{2\pi}\, (n+1)^{1/4}, \end{aligned}$$

where we used orthonormality. □

A.2 Laguerre Polynomials

For $\alpha > -1$, the generalized Laguerre polynomials $\{L_m^{(\alpha)}\}_{m\ge 0}$ are orthogonal with respect to the weight $x^\alpha e^{-x}$ on $[0,\infty)$ and are given by

$$L_m^{(\alpha)}(x) \;:=\; \sum_{j=0}^{m} (-1)^j \binom{m+\alpha}{m-j} \frac{x^j}{j!}, \tag{A.11}$$

where $\binom{a}{k} = a(a-1)\cdots(a-k+1)/k!$ for $a \in \mathbb{R}$, $k \in \mathbb{N}_0$. Moreover, the moment identity

$$\int_0^\infty x^{\alpha'-1}\, e^{-x}\, L_m^{(\alpha)}(x)\, dx \;=\; \binom{\alpha-\alpha'+m}{m} \Gamma(\alpha') \tag{A.12}$$

holds for $\alpha > -1$ and $\operatorname{Re}\,\alpha' > 0$, where Γ denotes the Gamma function. From this, a short computation shows

$$\int_0^\infty L_m^{(\alpha)}(x)\, x^{r+\alpha}\, e^{-x}\, dx \;=\; (-1)^m \binom{r}{m} \Gamma(r+\alpha+1) \tag{A.13}$$

for any integers $m, r \ge 0$, where we use the convention $\binom{r}{m} = 0$ when $r < m$.

For $m \ge 0$, the Hermite and Laguerre polynomials are linked through the identity

$$H_{2m}(x) \;=\; (-1)^m\, 2^{2m}\, m!\, L_m^{(-1/2)}(x^2). \tag{A.14}$$

Combining (A.14) with the definition of the normalized Hermite function Φ_n, for $n \ge 1$, the packet

$$A_n(x) \;:=\; (-1)^n\, x\, L_n^{(-1/2)}(x^2)\, e^{-x^2/2} \tag{A.15}$$

admits the factorization

$$A_n(x) \;=\; d_n\, x\, \Phi_{2n}(x), \qquad d_n \;:=\; \frac{\left(2^{2n}(2n)!\sqrt{\pi}\right)^{1/2}}{2^{2n}\, n!}. \tag{A.16}$$

Squaring the definition of d_n gives

$$d_n^2 \;=\; \sqrt{\pi}\, \frac{\binom{2n}{n}}{4^n} \;=\; \Theta(n^{-1/2}) \tag{A.17}$$

where we used Stirling's formula $\binom{2n}{n} \sim 4^n/\sqrt{\pi n}$.

This yields the following further estimates.

Lemma A.2. *There is an absolute constant $C > 0$ and, for every $j \ge 0$, a constant $C_j > 0$ such that for all $M \ge 1$,*

$$\|\widehat{A_M}\|_\infty \le C, \tag{A.18}$$

$$\|s^j A_M\|_1 + \|A_M^{(j)}\|_1 \le C_j\, M^{(j+1)/2}. \tag{A.19}$$

Proof. By (A.16)–(A.17), $A_M(s) = d_M\, s\, \Phi_{2M}(s)$ with $|d_M| \lesssim M^{-1/4}$. Using (A.5) and $\widehat{sf} = i\widehat{f}'$, we obtain

$$\|\widehat{A_M}\|_\infty = |d_M|\|\Phi_{2M}'\|_\infty \overset{(A.8)}{\lesssim} M^{-1/4} M^{1/4} = 1,$$

which is (A.18).

For (A.19), the ladder identities (A.6)–(A.7) show that each multiplication by s or differentiation shifts the Hermite index by one, and all coefficients are in $O_j(\sqrt{M})$ throughout the first $j+1$ steps. Hence

$$\|s^j A_M\|_1 + \|A_M^{(j)}\|_1 \lesssim_j |d_M| M^{(j+1)/2} \max_{\substack{\ell\ge 0 \\ |\ell-2M|\le j+1}} \|\Phi_\ell\|_1$$

$$\overset{(A.9)}{\lesssim_j} M^{-1/4} M^{(j+1)/2} M^{1/4} = M^{(j+1)/2},$$

as required. □

A.3 Krawtchouk Polynomials

For $0 \le k \le n$ and $r \in \{-1, -1+2/n, \dots, 1-2/n, 1\}$, the normalized binary Krawtchouk polynomial of degree k is

$$\mathcal{K}_k^{(n)}(r) \;:=\; \binom{n}{k}^{-1} \sum_{b=0}^{k} (-1)^b \binom{n(1-r)/2}{b} \binom{n(1+r)/2}{k-b}. \tag{A.20}$$

We use the convention $\binom{a}{b} = 0$ whenever $b < 0$ or $b > a$ for nonnegative integers a, b. Writing $d = n(1-r)/2 \in \{0, 1, \dots, n\}$ for the Hamming distance, this is

$$\mathcal{K}_k^{(n)}(r) \;=\; \binom{n}{k}^{-1} K_k(d; n), \qquad K_k(d; n) \;:=\; \sum_{b=0}^{k} (-1)^b \binom{d}{b} \binom{n-d}{k-b} \tag{A.21}$$

where $K_k(\,\cdot\,; n)$ is the unnormalized Krawtchouk polynomial.

The Krawtchouk polynomials satisfy the three-term recurrence in the degree variable

$$(n-2d)\, K_k(d; n) \;=\; (k+1)\, K_{k+1}(d; n) \;+\; (n-k+1)\, K_{k-1}(d; n) \tag{A.22}$$

with the convention $K_{-1} \equiv K_{n+1} \equiv \mathcal{K}_{-1}^{(n)} \equiv \mathcal{K}_{n+1}^{(n)} \equiv 0$. Using $n - 2d = nr$ and $(k+1)\binom{n}{k+1} = (n-k)\binom{n}{k}$, $(n-k+1)\binom{n}{k-1} = k\binom{n}{k}$, this is equivalent to the normalized recurrence

$$nr\, \mathcal{K}_k^{(n)}(r) \;=\; (n-k)\, \mathcal{K}_{k+1}^{(n)}(r) \;+\; k\, \mathcal{K}_{k-1}^{(n)}(r). \tag{A.23}$$

For $0 \le \ell \le n$, $x, y \in \{\pm 1\}^n$, and $D := |\{i \mid x_i \ne y_i\}|$, we have

$$\binom{n}{\ell}^{-1} \sum_{|S|=\ell} \chi_S(x)\, \chi_S(y) \;=\; \frac{K_\ell(D; n)}{\binom{n}{\ell}} \;=\; \mathcal{K}_\ell^{(n)}\!\left(\frac{\langle x, y\rangle}{n}\right). \tag{A.24}$$

This follows from the observation that if b of the ℓ selected coordinates lie in the disagreement set, then $\chi_S(x)\chi_S(y) = (-1)^b$, and summing over b gives the Krawtchouk polynomial.

Using $\chi_S(w \odot x) = \chi_S(w)\chi_S(x)$ for $w \in \{\pm 1\}^n$, expanding the product, and applying (A.24) gives the addition formula

$$\mathbb{E}_{x \sim U_n}[\Psi_{\ell,n}(u \odot x)\, \Psi_{\ell',n}(v \odot x)] \;=\; \mathcal{K}_\ell^{(n)}\!\left(\frac{\langle u, v\rangle}{n}\right) \mathbf{1}_{\{\ell = \ell'\}} \tag{A.25}$$

for every $u, v \in \{\pm 1\}^n$.

B Omitted Proofs

In this appendix, we provide the deferred proofs of the theorems and lemmas from the main section.

B.1 Proof of Lemma 4.1

For $x \in \{\pm 1\}^n$, let $S(x) := \sum_{i=1}^n x_i$, and for $y \ge 0$, let $\theta_n(y) := \arctan\sqrt{y/n}$ and $\phi_{y,n}(x) := \sin(\theta_n(y) S(x))$. By construction, $\phi_{y,n}$ is symmetric, odd and bounded in absolute value by 1.

For odd n and $\theta := \theta_n(y)$, we have $\phi_{y,n} = \operatorname{Im} e^{i\theta \sum_{j=1}^n x_j} = \operatorname{Im} \prod_{j=1}^n (\cos\theta + i x_j \sin\theta)$. Expanding the product and plugging in the substitution gives

$$\begin{aligned}
\phi_{y,n}(x) &= (\cos\theta)^n \sum_{\substack{d \le n \\ d \text{ odd}}} (-1)^{(d-1)/2} (\tan\theta)^d \binom{n}{d}^{1/2} \Psi_{d,n}(x) \\
&= \left(1 + \frac{y}{n}\right)^{-n/2} \sum_{\substack{d \le n \\ d \text{ odd}}} (-1)^{(d-1)/2} \left(\frac{y}{n}\right)^{d/2} \binom{n}{d}^{1/2} \Psi_{d,n}(x) \\
&= \left(1 + \frac{y}{n}\right)^{-n/2} \sum_{r=0}^{(n-1)/2} (-1)^r \binom{n}{2r+1}^{1/2} n^{-r-\frac{1}{2}}\, y^{r+\frac{1}{2}}\, \Psi_{2r+1,n}(x). \qquad (\phi)
\end{aligned}$$

Weighting $\phi_{y,n}$ against $L(y)\,(1+y/n)^{n/2}e^{-y}\,dy$ therefore multiplies the level-$(2r+1)$ coefficient by $\int_0^\infty L(y)\,y^{r+1/2}e^{-y}\,dy$, which vanishes for all small r when L is a Laguerre polynomial, cf. (A.13). The average of two consecutive witnesses turns out to give better asymptotics than a single witness.

Definition B.1 (Sine–Laguerre witness)**.** Let $n \geq 1$ be odd, and let $\phi_{y,n}$ be as in (ϕ). For $\ell \geq 1$, the *ℓ-th unnormalized Laguerre witness* is

$$\Omega_{\ell,n}(x) := (-1)^\ell \int_0^\infty L_\ell^{(1/2)}(y)\left(1+\frac{y}{n}\right)^{n/2} e^{-y}\,\phi_{y,n}(x)\,dy,$$

and for $j \geq 1$, the *two-term averaged sine–Laguerre witness* is $\widetilde{\Omega}_{j,n} := \frac{1}{2}(\Omega_{j,n}+\Omega_{j+1,n})$. Whenever $\widetilde{\Omega}_{j,n} \not\equiv 0$, the *normalized sine–Laguerre witness* is

$$\psi_{j,n} := \frac{\widetilde{\Omega}_{j,n}}{\|\widetilde{\Omega}_{j,n}\|_\infty}.$$

Lemma B.2. *Let $n \geq 1$ be odd, $m \geq 1$, and set $k := \lfloor (m+1)/2 \rfloor$. Assume $k \leq (n-1)/2$. Then the normalized sine–Laguerre witness $\psi_{k,n}$ is a valid dual witness, i.e., $\|\psi_{k,n}\|_\infty \leq 1$ and $\psi_{k,n} \perp \mathcal{P}_{\leq m}$.*

Proof. The first part holds by definition. By the parameter shift $L_{k+1}^{(1/2)} - L_k^{(1/2)} = L_{k+1}^{(-1/2)}$, the kernel of $\widetilde{\Omega}_{k,n}$ can be written as $\frac{1}{2}(-1)^{k+1}L_{k+1}^{(-1/2)}(y)$. Moreover, by (A.13),

$$\int_0^\infty L_{k+1}^{(-1/2)}(y)\,y^{r+1/2}e^{-y}\,dy = (-1)^{k+1}\binom{r+1}{k+1}\Gamma\big(r+\tfrac{3}{2}\big),$$

which vanishes for $r < k$. Inserting (ϕ) thus gives

$$\begin{aligned}\widetilde{\Omega}_{k,n} &= \frac{(-1)^{k+1}}{2}\int_0^\infty L_{k+1}^{(-1/2)}(y)\left(1+\frac{y}{n}\right)^{n/2} e^{-y}\,\phi_{y,n}\,dy \\ &= \frac{1}{2}\sum_{r=k}^{(n-1)/2}(-1)^r\binom{r+1}{k+1}\Gamma\big(r+\tfrac{3}{2}\big)\binom{n}{2r+1}^{1/2} n^{-r-\frac{1}{2}}\,\Psi_{2r+1,n}. \qquad \text{(E)}\end{aligned}$$

For $k \leq (n-1)/2$, we readily see $\widetilde{\Omega}_{k,n} \not\equiv 0$ as well as $\widetilde{\Omega}_{k,n} \perp \mathcal{P}_{\leq 2k} \supseteq \mathcal{P}_{\leq m}$. □

Lemma B.3. *Fix $0 < \rho < 1$. For every $c > 0$ and $0 < \eta < 1/2$, there exist $n_0(c,\eta)$ and $K_{c,\eta} > 0$ such that, for all odd $n = 2N+1 \geq n_0(c,\eta)$ and all integers $1 \leq m \leq cn^{1/2-\eta}$,*

$$\inf_{\deg p \leq m} \|T_\rho\,\mathrm{Maj}_n - p\|_{L^1(U_n)} \geq K_{c,\eta}\frac{\rho^{2m+1}}{\sqrt{m}}.$$

Remark B.4. The assumption $m \leq cn^{1/2-\eta}$ simplifies the asymptotics here but is plausibly not essential for the lower bound; without it, however, the proof of Lemma B.5 would fail.

Proof of Lemma B.3. In Lemma B.5 we show $\|\widetilde{\Omega}_{k,n}\|_\infty = O_{c,\eta}(1)$, so by weak duality and the previous lemma it suffices to lower-bound $\langle T_\rho\,\mathrm{Maj}_n, \widetilde{\Omega}_{k,n}\rangle$. By O'Donnell [2014, Thm. 5.19], the Fourier coefficients of majority vanish on sets of even size, and for $|S| = j$ odd,

$$\widehat{\mathrm{Maj}_n}(S) = (-1)^{(j-1)/2}\frac{\binom{N}{(j-1)/2}}{\binom{n-1}{j-1}}\cdot\frac{2}{2^n}\binom{n-1}{N}.$$

Combining this with (E), in the correlation $\langle T_\rho\,\mathrm{Maj}_n, \widetilde{\Omega}_{k,n}\rangle$ all summands per non-vanishing level are nonnegative, so we may keep only those with index $k \leq r \leq 2k-1$. Since $2k-1 \leq m \leq N$

and $m^2 \le n/2$ for large n, after some simplification of the terms, it holds

$$\begin{aligned}\langle T_\rho \operatorname{Maj}_n, \widetilde{\Omega}_{k,n}\rangle &\ge \underbrace{\frac{\binom{2N}{N}\sqrt{n}}{4^N}}_{=\Theta(1)} \frac{\rho}{4} \sum_{r=k}^{2k-1} \rho^{2r} \underbrace{\frac{N^{\underline{r}}}{n^r}}_{\ge 2^{-r-1}} \underbrace{\frac{\Gamma(r+\frac{1}{2})}{\Gamma(r+1)}}_{\gtrsim m^{-1/2}} \binom{r+1}{k+1} \\ &\gtrsim_{c,\eta} \frac{\rho}{\sqrt{m}} \sum_{r=k}^{2k-1} \rho^{2r}\, 2^{-r} \binom{r+1}{k+1} \\ &\ge \frac{\rho^{2m+1}}{\sqrt{m}} \sum_{r=k}^{2k-1} 2^{-r} \binom{r}{k},\end{aligned}$$

where $N^{\underline{r}} := \prod_{j=0}^{r-1}(N-j)$.

Since $2^{-r-1}\binom{r}{k}$ is the probability that the $(k+1)$-st head of a fair coin appears at toss $r+1$, the last sum is twice the probability of at least $k+1$ heads in $2k$ tosses, that is, $1 - 4^{-k}\binom{2k}{k} \ge \frac{1}{2}$. Hence $\langle T_\rho \operatorname{Maj}_n, \widetilde{\Omega}_{k,n}\rangle \gtrsim \rho^{2m+1}/\sqrt{m}$. □

Lemma B.5. *For every $c > 0$ and $0 < \eta < 1/2$ there is a constant $C_{c,\eta} > 0$ such that for all odd n and all integers $1 \le m \le cn^{1/2-\eta}$, with $k = \lfloor (m+1)/2 \rfloor$, it holds $\|\widetilde{\Omega}_{k,n}\|_\infty \le C_{c,\eta}$.*

Proof. Fix $x \in \{\pm 1\}^n$ and put $\zeta := S(x)/\sqrt{n}$; recall that $k+1 \le 2cn^{1/2-\eta}$. Let A_{k+1} be as in (A.15), and set $w(s) := e^{-s^2/2}(1+s^2/n)^{n/2}$ and $\nu(s) := \sqrt{n}\arctan(s/\sqrt{n})$. Changing variables $y = s^2$ in (E) gives

$$\begin{aligned}\widetilde{\Omega}_{k,n}(x) &= \frac{(-1)^{k+1}}{2} \int_0^\infty L_{k+1}^{(-1/2)}(y)\, e^{-y} \Big(1 + \frac{y}{n}\Big)^{n/2} \sin\big(\theta_n(y) S(x)\big)\, dy \\ &= \int_0^\infty A_{k+1}(s)\, w(s) \sin(\zeta \nu(s))\, ds = \frac{1}{2i} \underbrace{\int_{\mathbb{R}} A_{k+1}(s)\, w(s)\, e^{i\zeta\nu(s)}\, ds}_{=:I(\zeta)},\end{aligned}$$

as A_{k+1} and ν are odd and w is even. Since $w(s) \to 1$ and $\nu(s) \to s$ as $n \to \infty$, the integral $I(\zeta)$ is essentially $\sqrt{2\pi}\,\widehat{A_{k+1}}(-\zeta)$, the Fourier transform of a Hermite function, which is uniformly bounded by (A.18). In the following, making this heuristic rigorous yields $|I(\zeta)| \lesssim_{c,\eta} 1$, which establishes the lemma.

Using $\log(1+z) \le z - z^2/(2(1+z))$ for $z \ge 0$, we have

$$w(n^{1/4}t) \le \exp\Big(-\frac{t^4}{4(1+t^2/\sqrt{n})}\Big) \le e^{(1-t^2)/4}. \tag{W}$$

For $|\zeta| \le n^{1/4}$, from the preceding bound, we obtain

$$\big\|w\, e^{i\zeta(\nu-s)}\big\|_2^2 = \|w\|_2^2 = n^{1/4} \int_{\mathbb{R}} w(n^{1/4}t)^2\, dt \le n^{1/4} \int_{\mathbb{R}} e^{(1-t^2)/2}\, dt \lesssim n^{1/4},$$

and similarly $\big\|\big(w\, e^{i\zeta(\nu-s)}\big)'\big\|_2^2 \lesssim n^{-1/4}$. By Young's and the Cauchy–Schwarz inequality, it follows that

$$\begin{aligned}|I(\zeta)| &= \Big|\int_{\mathbb{R}} \widehat{A_{k+1}}(\xi)\, \mathcal{F}\big[w\, e^{i\zeta(\nu-s)}\big](-\xi-\zeta)\, d\xi\Big| \le \|\widehat{A_{k+1}}\|_\infty \big\|\mathcal{F}\big[w\, e^{i\zeta(\nu-s)}\big]\big\|_1 \\ &\overset{(A.18)}{\lesssim} \Big(\big\|w\, e^{i\zeta(\nu-s)}\big\|_2 \big\|\big(w\, e^{i\zeta(\nu-s)}\big)'\big\|_2\Big)^{1/2} \lesssim 1.\end{aligned}$$

For $|\zeta| > n^{1/4}$, we use the principle of non-stationary phase, as $\nu'(s) = (1+s^2/n)^{-1} > 0$ for all $s \in \mathbb{R}$. Defining $Lf := (f/\nu')' = \big((1+s^2/n)f\big)'$, L maps the Schwartz class $\mathcal{S}(\mathbb{R})$ into itself. Further, as $wA_{k+1} \in \mathcal{S}(\mathbb{R})$, integrating $I(\zeta)$ by parts repeatedly introduces no boundary terms. Moreover,

$w'/w = -s^3/(n+s^2)$ gives the conjugation identity $L(wG) = w\left[(1+s^2/n)\,\partial_s + (2s-s^3)/n\right]G$, so substituting $s = n^{1/4}t$ and inducting on $J \geq 1$ yields

$$|I(\zeta)| = \Big|\Big(\frac{i}{\zeta}\Big)^J \int_{\mathbb{R}} L^J(wA_{k+1})\, e^{i\zeta\nu}\, ds\Big| \overset{(W)}{\lesssim_J} n^{-J/4} \sum_{j=0}^{J} n^{(j-J)/4} \|A_{k+1}^{(j)}\|_1$$
$$\overset{(A.19)}{\lesssim_J} n^{-J/4}(k+1)^{(J+1)/2} \lesssim_{c,J} n^{1/4-\eta(J+1)/2}.$$

Taking $J = \lceil 1/(2\eta)\rceil$ makes the exponent negative, and combining the bounds for $|\zeta| \leq n^{1/4}$ and $|\zeta| > n^{1/4}$ finishes the proof. □

B.2 Proof of Lemma 4.2

Throughout this subsection, $\psi := \psi_{k,n} = \sum_{d=0}^{n} b_{d,n}\Psi_{d,n}$ and $\psi^{(u)}(x) := \psi(u \odot x)$ for $u \in \{\pm 1\}^n$. Since ψ is odd and orthogonal to $\mathcal{P}_{\leq 2k}$, we have $b_{d,n} = 0$ unless d is odd and $d \geq 2k+1$; moreover $\sum_d b_{d,n}^2 = \|\psi\|_2^2 \leq 1$. We need two auxiliary facts, namely a geometric decay of the coefficients $b_{d,n}$ and a simple bound on Krawtchouk polynomials.

Lemma B.6. *Let $k \geq 1$ and $n \geq 8k+11$. Then for every odd d with $8k+9 \leq d \leq n-2$,*

$$\left|\frac{b_{d+2,n}}{b_{d,n}}\right| \leq \frac{2}{3}.$$

Proof. From (E), we get

$$b_{d,n} = (-1)^{(d-1)/2}\frac{1}{2\|\widetilde{\Omega}_{k,n}\|_\infty}\binom{n}{d}^{1/2} n^{-d/2}\binom{(d+1)/2}{k+1}\Gamma\big(\tfrac{d}{2}+1\big)$$

for odd $d \geq 2k+1$. A direct computation shows

$$\frac{|b_{d+2,n}|}{|b_{d,n}|} = \frac{d+3}{d+1-2k}\cdot\frac{1}{2}\sqrt{\frac{d+2}{d+1}}\cdot\sqrt{\Big(1-\frac{d}{n}\Big)\Big(1-\frac{d+1}{n}\Big)}$$
$$\leq \frac{8k+12}{6k+10}\cdot\frac{1}{2}\sqrt{\frac{8k+11}{8k+10}} \leq \frac{2}{3}$$

for $d \geq 8k+9$. □

Lemma B.7. *Let $n \geq 4$ and let D be an integer with $1 \leq D \leq \sqrt{n}$. Then for every $0 \leq j \leq D$ and every $r \in \{-1, -1+2/n, \ldots, 1\}$,*

$$|\mathcal{K}_j^{(n)}(r)| \leq 2\left(|r| + \sqrt{D/n}\right)^j.$$

Proof. Let $F_j := \frac{n^{\underline{j}}}{n^j}\mathcal{K}_j^{(n)}(r)$, where $n^{\underline{j}} := n(n-1)\cdots(n-j+1)$, so that $F_0 = 1$, $F_1 = r$ and, by (A.23),

$$F_j = rF_{j-1} - \frac{(j-1)(n-j+2)}{n^2}F_{j-2} \qquad (j \geq 2),$$

where the last coefficient is at most D/n for $j \leq D$. With $a := |r| + \sqrt{D/n}$ we have $|r|a + D/n \leq a^2$, and induction gives $|F_j| \leq |r|a^{j-1} + \frac{D}{n}a^{j-2} \leq a^j$. Finally, $\frac{n^{\underline{j}}}{n^j} = \prod_{i<j}(1-\frac{i}{n}) \geq 1 - \frac{j^2}{2n} \geq \frac{1}{2}$ as $j \leq \sqrt{n}$, hence $|\mathcal{K}_j^{(n)}(r)| \leq 2a^j$. □

Lemma B.8 (Hidden-direction family)**.** *Fix $0 < \rho < 1$, $c_0 > 0$, and $0 < \eta < 1/2$. There exist $n_0 = n_0(c_0, \eta)$ and $K_{c_0,\eta} > 0$ such that for every odd $n \geq n_0$ and every integer $1 \leq m \leq c_0 n^{1/2-\eta}$, with $k := \lfloor (m+1)/2 \rfloor$, the following holds for the distributions*

$$\mathcal{D}_u(x,y) := 2^{-(n+1)}\big(1 + y\,\psi^{(u)}(x)\big), \quad \mathcal{D}_0(x,y) := 2^{-(n+1)}$$

on $\{\pm 1\}^n \times \{\pm 1\}$, where $u \in \{\pm 1\}^n$.

(i) For every $u \in \{\pm 1\}^n$, the smoothed rotated majority $f_u(x) := T_\rho \operatorname{sign}(\langle u, x\rangle)$ satisfies

$$\mathbb{E}_{\mathcal{D}_u}[Y f_u(X)] \geq \frac{K_{c_0,\eta}\, \rho^{2m+1}}{\sqrt{m}}.$$

(ii) For all $u \neq v$ in $\{\pm 1\}^n$ with $|\langle u, v\rangle| \leq n^{3/4}$,

$$\chi_{\mathcal{D}_0}(\mathcal{D}_u, \mathcal{D}_v) \leq \gamma_n := 4\left(2n^{-1/4}\right)^{2k+1},$$

and

$$n^{-O(k)} \;\leq\; \gamma_n \;\leq\; 4\, n^{-(2k+1)/8}. \tag{γ}$$

(iii) There exists $\mathcal{U} \subseteq \{\pm 1\}^n$ with $|\mathcal{U}| = 2^{\Omega(\sqrt{n})}$ such that $|\langle u, v\rangle| \leq n^{3/4}$ for all distinct $u, v \in \mathcal{U}$.

Proof. Since $\|\psi\|_\infty \leq 1$, each $\mathcal{D}_u$ is a probability distribution. Take $n_0 = n_0(c_0, \eta)$ large enough so that Lemmas B.3 and B.6 and the estimates below apply uniformly over all allowed m. This is possible since $k \log n \leq c_0 n^{1/2-\eta} \log n = o(\sqrt{n})$.

Let $R_u := \mathcal{D}_u/\mathcal{D}_0 - 1$, so that $R_u(x, y) = y\, \psi^{(u)}(x)$ and $\|R_u\|_{L^2(\mathcal{D}_0)} \leq 1$.

(i) Under $\mathcal{D}_u$ we have $X \sim U_n$. Both f_u and $\mathbb{E}[Y \mid X = x] = \psi^{(u)}(x)$ are the result of applying the sign flip $x \mapsto u \odot x$ to $T_\rho \operatorname{Maj}_n$ and ψ, since T_ρ commutes with the flip. As the X-marginal U_n is preserved under the rotation, so is the correlation $\mathbb{E}_{\mathcal{D}_u}[Y f_u(X)] = \langle T_\rho \operatorname{Maj}_n, \psi\rangle_{U_n}$, which allows us to use the bound from the proof of Lemma B.3 together with Lemma B.5.

(ii) By assumption, we have $|\langle u, v\rangle|/n \leq n^{-1/4}$, so

$$\begin{aligned}
\left|\langle R_u, R_v\rangle_{L^2(\mathcal{D}_0)}\right| = \left|\sum_{d\geq 1} \langle \psi_d^{(u)}, \psi_d^{(v)}\rangle\right| &\overset{\text{(A.25)}}{\leq} \sum_{\substack{d\geq 1\\ d \text{ odd}}} b_{d,n}^2\, |\mathcal{K}_d^{(n)}(\langle u, v\rangle/n)| \\
&\overset{\text{(Lem. B.7)}}{\leq} 2\left(2n^{-1/4}\right)^{2k+1} + \sum_{\substack{d>\sqrt{n}\\ d \text{ odd}}} b_{d,n}^2 \\
&\overset{\text{(Lem. B.6)}}{\leq} 2\left(2n^{-1/4}\right)^{2k+1} + e^{-\Omega(\sqrt{n})} \\
&\leq\; 4\left(2n^{-1/4}\right)^{2k+1}
\end{aligned}$$

for n large enough, where $\psi_d := b_{d,n}\Psi_{d,n}$. The upper bound in ($\gamma$) follows from $2n^{-1/4} \leq n^{-1/8}$ for $n \geq 256$; the lower bound is immediate.

(iii) Let $s := \lfloor \exp(0.2\sqrt{n})\rfloor$; if $s \leq 1$ there is nothing to show. Otherwise, sample $u^{(1)}, \ldots, u^{(s)}$ independently and uniformly from $\{\pm 1\}^n$. For $a \neq b$, $\langle u^{(a)}, u^{(b)}\rangle$ is a sum of n independent Rademacher variables, so by Hoeffding's inequality and a union bound,

$$\Pr\left(\exists a \neq b :\ |\langle u^{(a)}, u^{(b)}\rangle| > n^{3/4}\right) \leq s^2 e^{-\sqrt{n}/2} \leq e^{(2/5-1/2)\sqrt{n}} < 1.$$

Hence some realization has all pairwise inner products with absolute value at most $n^{3/4}$; in particular, the vectors are distinct. □

B.3 Proof of Theorems 4.3 and 1.3

Theorem B.9 (Hidden-direction SQ lower bound)**.** *For every $c_0 > 0$ and $\eta_0 \in (0, 1/2)$, there is $n_0 = n_0(c_0, \eta_0)$ such that the following holds. Let $n \geq n_0$ be odd, $1 \leq m \leq c_0 n^{1/2-\eta_0}$ and $k := \lfloor (m+1)/2\rfloor$, and let $\mathcal{U}$ and γ_n be as in Lemma B.8. For $\varepsilon > 0$ with $\varepsilon^2 \geq 2\gamma_n$, let P be the search problem on $\{\mathcal{D}_u\}_{u\in\mathcal{U}}$ whose valid solutions for $\mathcal{D}_u$ are the hypotheses $h : \{\pm 1\}^n \to [-1, 1]$ with $\mathbb{E}_{\mathcal{D}_u}[Y h(X)] \geq \varepsilon$, and assume that every $\mathcal{D}_u$ admits a valid solution.*

Then any randomized SQ algorithm that solves P with probability at least $2/3$ on every $\mathcal{D}_u$ needs at least

$$\tfrac{2}{3}\,|\mathcal{U}|\gamma_n - 1 \;=\; 2^{\Omega(\sqrt{n})}$$

queries to $\mathrm{VSTAT}(1/(6\gamma_n))$. *In particular, it must either make* $2^{\Omega(\sqrt{n})}$ *STAT queries or use a query of tolerance*

$$\tau < \sqrt{6\gamma_n} \le 2\sqrt{6}\, n^{-(2k+1)/16}.$$

Proof. Write $\gamma := \gamma_n$. By Lemma B.8(ii), every nonempty $B \subseteq \mathcal{U}$ satisfies

$$\gamma(B, \mathcal{D}_0) \le \gamma + \frac{1}{|B|}. \tag{1}$$

We first show that any fixed hypothesis is valid for at most $1/\gamma$ distributions. Fix $h : \{\pm 1\}^n \to [-1, 1]$, and let $S_h := \{u \in \mathcal{U} \mid h \in P_{\mathcal{D}_u}\}$. There is nothing to prove if $S_h = \emptyset$. Otherwise, under $\mathcal{D}_0$, the label is uniform and independent of X, so $\mathbb{E}_{\mathcal{D}_0}[yh(x)] = 0$ and $\|yh(x)\|_{\mathcal{D}_0} \le 1$. For every $u \in S_h$, a change of measure gives $\langle yh, R_u\rangle_{\mathcal{D}_0} = \mathbb{E}_{\mathcal{D}_u}[Yh(X)] \ge \varepsilon$. Averaging over S_h, applying the Cauchy–Schwarz inequality, and using (1), we obtain

$$\varepsilon^2 \le \Big\| \frac{1}{|S_h|} \sum_{u \in S_h} R_u \Big\|_{\mathcal{D}_0}^2 \le \gamma(S_h, \mathcal{D}_0) \le \gamma + \frac{1}{|S_h|}.$$

Since $\varepsilon^2 \ge 2\gamma$, it follows that $|S_h| \le 1/\gamma$.

For sufficiently large n, set $\beta := \frac{1}{|\mathcal{U}|\gamma}$, and $d := |\mathcal{U}|\gamma - 1 = (1-\beta)|\mathcal{U}|\gamma$. For every hypothesis h, the subfamily $\mathcal{D}_h := \{\mathcal{D}_u \mid u \in \mathcal{U} \setminus S_h\}$ has size at least $|\mathcal{U}| - 1/\gamma = (1-\beta)|\mathcal{U}|$. Moreover,

$$\frac{|\mathcal{D}_h|}{d} \ge \frac{|\mathcal{U}| - 1/\gamma}{|\mathcal{U}|\gamma - 1} = \frac{1}{\gamma}.$$

Thus every $\mathcal{A} \subseteq \mathcal{D}_h$ with $|\mathcal{A}| \ge |\mathcal{D}_h|/d$ has average correlation at most 2γ, by (1). By the definition of SDA,

$$\mathrm{SDA}(P, 2\gamma, \beta) \ge d.$$

The problem P is total by assumption. Hence Theorem 3.5 gives the lower bound

$$\frac{2/3 - \beta}{1 - \beta}\, d = \frac{2}{3}|\mathcal{U}|\gamma - 1 \overset{\text{Lem. B.8(iii), (}\gamma\text{)}}{=} \exp\big(\Omega(\sqrt{n}) - O(k \log n)\big) = 2^{\Omega(\sqrt{n})},$$

on the number of queries to $\mathrm{VSTAT}(1/(6\gamma))$ for any SQ algorithm that solves P with probability at least $2/3$, where the last step uses $k = O(n^{1/2-\eta_0})$. This also justifies the choice $\beta < 2/3$ for sufficiently large n. The conversion from VSTAT to STAT is standard. □

Theorem B.10. *Let* $c_0 > 0$ *and* $\eta_0 \in (0, 1/2)$. *There exist constants* $a_0, n_0 > 0$, *depending only on* c_0 *and* η_0, *such that the following holds. Let* $\sigma \in (0, 0.499]$, $\varepsilon \in (0, 1/4]$ *and let* $n \ge n_0$ *be odd. Set*

$$m := \left\lfloor a_0 \frac{\log(1 + \sigma/\varepsilon^2)}{\sigma} \right\rfloor$$

and assume $1 \le m \le c_0 n^{1/2-\eta_0}$.[3] *Let* $\mathcal{U}$ *and* $\{\mathcal{D}_u\}_{u \in \mathcal{U}}$ *be as in Lemma B.8 for this* m. *Then any randomized SQ algorithm that, given access to an unknown* $\mathcal{D}_u$ *with* $u \in \mathcal{U}$, *outputs with probability at least* $2/3$ *a hypothesis* $h : \{\pm 1\}^n \to \{\pm 1\}$ *with* $\mathrm{err}_{\mathcal{D}_u}(h) \le \mathrm{opt}_{\sigma, \mathcal{D}_u} + \varepsilon$ *must either make* $2^{\Omega(\sqrt{n})}$ *STAT queries or use a query of tolerance*

$$\tau < 2\sqrt{6}\, n^{-m/16} = n^{-\Omega\left(\log(1+\sigma/\varepsilon^2)/\sigma\right)}.$$

Proof. We apply Theorem B.9 to the search problem P at threshold 2ε. To this end, we show that $T_\rho g_u$ is a valid solution, $\gamma_n \le 2\varepsilon^2$, and that every ε-optimal hypothesis solves P.

Choose $a_0 > 0$ sufficiently small and let $\rho := 1 - 2\sigma$. Since $-\log \rho = O(\sigma)$ and $m \ge 1$, the definition of m gives

$$\varepsilon^2 \le \frac{\sigma}{e^{\sigma m/a_0} - 1} \le \frac{a_0}{m} e^{-\sigma m/(2a_0)} \le \frac{a_0 \rho^{4m+2}}{m},$$

[3] For ε below a sufficiently small absolute constant, the restriction $m \ge 1$ can be dropped: if $m = 0$, then $\log(1 + \sigma/\varepsilon^2)/\sigma < 1/a_0 = O(1)$, and a similar family of hard instances (e.g., rotated majority functions along nearly orthogonal directions) yields the same SDA lower bound. Since the upper bound on ε is in this case only scaled by a constant, we ignore it for simplicity.

where the middle inequality uses $e^s - 1 \geq se^{s/2}$ for $s \geq 0$. Thus Lemma B.8(i) gives $\mathbb{E}_{\mathcal{D}_u}[YT_\rho g_u(X)] \geq 4\varepsilon$ for the halfspace $g_u(x) := \operatorname{sign}\langle u, x\rangle$.

To bound γ_n, note that by concavity of $s \mapsto \log(1 + s/\varepsilon^2)$ we have

$$\log \frac{1}{\varepsilon} \leq \frac{\log(1 + \sigma/\varepsilon^2)}{2\sigma} < \frac{m+1}{2a_0} \leq \frac{m}{a_0}.$$

Hence (γ) implies $\gamma_n \leq 4n^{-m/8} \leq 2\varepsilon^2$ for all sufficiently large n.

The search problem P at threshold 2ε is therefore total, with valid solutions $T_\rho g_u$. Moreover, every Boolean hypothesis h with $\operatorname{err}_{\mathcal{D}_u}(h) \leq \operatorname{opt}_{\sigma,\mathcal{D}_u} + \varepsilon$ solves P, since

$$\mathbb{E}_{\mathcal{D}_u}[Yh(X)] = 1 - 2\operatorname{err}_{\mathcal{D}_u}(h) \geq \mathbb{E}_{\mathcal{D}_u}[YT_\rho g_u(X)] - 2\varepsilon \geq 2\varepsilon.$$

Thus an SQ learner for the smoothed agnostic problem provides a solution to P with the same success probability and no additional queries. Applying Theorem B.9 at threshold 2ε shows that any such learner with success probability at least $2/3$ must either make $2^{\Omega(\sqrt{n})}$ STAT queries or use a query of tolerance

$$\tau < \sqrt{6\gamma_n} = O(n^{-m/16}) = n^{-\Omega(\log(1+\sigma/\varepsilon^2)/\sigma)},$$

where the last step follows from the definition of m and $m \geq 1$. □

We prove the more general version of Theorem 1.4 for product distributions.

Theorem B.11 (Upper bound for product marginals)**.** *Let $\mathcal{C}$ be any class of Boolean functions on $\{\pm 1\}^n$, and let $\varepsilon > 0$, $\sigma \in (0, 1/2)$ and $\delta \in (0, 1)$. There is an algorithm that, given samples from any distribution $\mathcal{D}$ on $\{\pm 1\}^n \times \{\pm 1\}$ whose X-marginal μ is a product distribution, has sample complexity and running time $\operatorname{poly}(n^{\log(1/\varepsilon)/\sigma}, \log(1/\delta))$ and outputs a Boolean hypothesis h with $\operatorname{err}(h) \leq \operatorname{opt}_{\mathcal{C},\sigma} + \varepsilon$ with probability at least $1 - \delta$.*

Proof. We use the L^1 polynomial regression algorithm [Kalai et al., 2008, Theorem 5]. For this, it suffices to show that for some $d = O(\log(1/\varepsilon)/\sigma)$, every $f \in \mathcal{C}$ admits a polynomial $p_f \in \mathcal{P}_{\leq d}$ with $\|T_\rho f - p_f\|_{L^1(\mu)} < \varepsilon/2$, where $\rho := 1 - 2\sigma$. We show this by taking p_f as the truncation of $T_\rho f$ at degree d in the μ-biased Fourier basis and bounding the $L^2(\mu)$ error.

Let $m_i := \mathbb{E}_\mu[X_i]$. Coordinates with $|m_i| = 1$ can be dropped after averaging f over their noisy values $m_i Z_i$, where $Z \sim B(\sigma)^n$. This preserves $T_\rho f$ on the support of μ and yields a function in $[-1, 1]$, which suffices below. Relabeling the remaining coordinates and denoting the resulting function again by f, assume $|m_i| < 1$ for all i. The law ν of $X \odot Z$ is again a product distribution, with means ρm_i. Let $\{\phi_S\}$ and $\{\psi_S\}$ be the biased Fourier bases of $L^2(\mu)$ and $L^2(\nu)$ [O'Donnell, 2014, Chapter 8], and write $f = \sum_{S\subseteq[n]} \tilde{f}(S)\psi_S$. Since T_ρ acts coordinatewise and preserves means, $T_\rho \psi_S = \lambda_S \phi_S$ with $\lambda_S := \prod_{i\in S} \lambda_i$, where

$$\lambda_i = \operatorname{Corr}(X_i, X_i Z_i) = \rho\sqrt{\frac{1 - m_i^2}{1 - \rho^2 m_i^2}} \leq \rho.$$

Hence $T_\rho f = \sum_{S\subseteq[n]} \lambda_S \tilde{f}(S)\phi_S$. Setting $p_f := \sum_{|S|\leq d} \lambda_S \tilde{f}(S)\phi_S$, we obtain

$$\|T_\rho f - p_f\|^2_{L^2(\mu)} = \sum_{|S|>d} \lambda_S^2 \tilde{f}(S)^2 \leq \rho^{2(d+1)} \sum_{S\subseteq[n]} \tilde{f}(S)^2 = \rho^{2(d+1)} \|f\|^2_{L^2(\nu)} \leq \rho^{2(d+1)}.$$

It holds $\rho = 1 - 2\sigma \leq e^{-2\sigma}$, so taking d at least a sufficiently large constant multiple of $\log(1/\varepsilon)/\sigma$ gives $\|T_\rho f - p_f\|_{L^1(\mu)} \leq \|T_\rho f - p_f\|_{L^2(\mu)} < \varepsilon/2$. As $\mathcal{P}_{\leq d}$ has dimension $n^{O(d)}$, the algorithm runs in $\operatorname{poly}(n^{\log(1/\varepsilon)/\sigma}, \log(1/\delta))$ time and sample complexity. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the SQ lower bound (Theorem B.10) and the matching upper bound (Theorem 1.4), with the precise scope (uniform marginal on the Boolean cube) and the technical contribution (sine-Laguerre witness, Lemma B.3) clearly identified.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion of this work discusses the remaining logarithmic gap between upper and lower bounds, the restriction to the uniform marginal, the remaining gap on the upper bound side in the subexponential setting of Kou and Meka [2025], and the open question of whether non-SQ algorithms can bypass the barrier.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems and lemmas in this work are numbered and cross-referenced. Informal statements in the main text (Theorems 1.3, 1.4, Lemmas 4.1, 4.2, Theorem 4.3) are paired with formal versions in the appendix (Corollary B.10, Theorem 1.4, Lemmas B.3, B.8, Theorem B.9) with complete proofs. All assumptions are stated explicitly in the formal theorem statements.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This work does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: This work does not include experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This work does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This work does not involve any human subjects, datasets, or deployed systems.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no direct path to any specific societal applications from the theoretical results presented in this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release any models, datasets or other artifacts that pose the risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This work does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This work does not introduce any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This work does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.