Proper Agnostic Learning of Functions of Halfspaces

Sergei Tikhonov*
UT Austin
tikhonov@utexas.edu

Arsen Vasilyan*
Aarhus University
arsenvasilyan@gmail.com

Abstract

We study the problem of computationally efficient *proper* agnostic learning of multidimensional concept classes under the Gaussian distribution. In this setting, given i.i.d. labeled samples from an unknown distribution over $\mathbb{R}^d \times \{\pm 1\}$ whose marginal on $\mathbb{R}^d$ is Gaussian, the goal is to output a hypothesis from a target class $\mathcal{F}$ whose 0-1 loss is within ε of that of the best classifier in $\mathcal{F}$. We give the first efficient proper agnostic learning algorithm for arbitrary Boolean functions of K halfspaces under Gaussian marginals. Our algorithm runs in time $d^{O(K^2 \log(1/\varepsilon)/\varepsilon^2)} + (K/\varepsilon)^{O(K^3/\varepsilon^{2.5})}$. Prior to our work, the only known algorithm for $K \geq 2$ was brute-force search, with run-time exponential in d. Moreover, the dependence of our run-time on the dimension d matches that of the best known *improper* learning algorithm, namely $d^{\widetilde{O}(K^2/\varepsilon^2)}$. For the special case of a single halfspace ($K = 1$), the best previous run-time was $d^{O(1/\varepsilon^4)} + (1/\varepsilon)^{O(1/\varepsilon^6)}$. Our algorithm improves this to $d^{O(1/\varepsilon^2)} + (1/\varepsilon)^{O(1/\varepsilon^{2.5})}$. Once again, the dependence on d matches that of the best known improper algorithm, namely $d^{O(1/\varepsilon^2)}$. Furthermore, the dependence of our run-time on the dimension d is essentially optimal in the statistical query model.

1 Introduction

In this work, we focus on computationally efficient *proper agnostic learning* of multidimensional concepts such as halfspaces and Boolean functions of halfspaces. Proper agnostic learning is the fundamental task of *fitting an approximately optimal classifier* from a hypothesis class $\mathcal{F}$ to an i.i.d. labeled dataset S. More formally:

Definition 1.1 (Proper Agnostic Learning)**.** Let $\mathcal{F}$ be a hypothesis class. Given an i.i.d. sample set $S = \{(\mathbf{x}_i, y_i)\}$ drawn from a distribution D, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$, a proper agnostic learning algorithm[1] for $\mathcal{F}$ must produce a hypothesis $h \in \mathcal{F}$ satisfying

$$\Pr_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT}_{\mathcal{F}} + \varepsilon$$

with high probability, where $\mathrm{OPT}_{\mathcal{F}} := \inf_{h' \in \mathcal{F}} \Pr_{(\mathbf{x},y)\sim D}[h'(\mathbf{x}) \neq y]$.

Fitting a near-optimal classifier from a hypothesis class, when no classifier fits the data perfectly, is one of the most basic goals in machine learning. Yet it is well known that without any assumption on the distribution D, this task is NP-hard even for the elementary class of *halfspaces* (also known as *linear classifiers*) [GR06, FGKP06], and is therefore believed to require run-time exponential in the data dimension d. This contrasts sharply with the realizable setting, in which the data is linearly separable and efficient algorithms have been known for decades [Ros58]. We therefore focus on the

*Alphabetical order.

[1]Here, agnostic learning means achieving error at most $\mathrm{OPT}_{\mathcal{F}} + \varepsilon$. Some works also use "agnostic learning" or "semi-agnostic learning" for weaker guarantees, such as $C \cdot \mathrm{OPT}_{\mathcal{F}} + \varepsilon$ with $C > 1$.

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

distribution-specific setting, where an assumption is placed on the distribution of the datapoints $\{\mathbf{x}_i\}$. Specifically, we adopt the standard assumption that $\{\mathbf{x}_i\}$ are drawn from a Gaussian distribution (see e.g. [KKMS08, Vem10b, Vem10c, DKS18, ABL17, DKTZ22, DKTZ20, DKK+21]).

Our approach applies to the concept class of *Boolean functions of K halfspaces*, yielding the first efficient proper agnostic learning algorithm for this hypothesis class under the Gaussian distribution. Specifically, our algorithm runs in time $d^{O(K^2 \log(1/\varepsilon)/\varepsilon^2)} + (K/\varepsilon)^{O(K^3/\varepsilon^{2.5})}$ and returns a Boolean function of K halfspaces whose prediction error is within ε of the optimal error among all Boolean functions of K halfspaces, which we denote as OPT_K. Previously, for $K \geq 2$, the only available algorithm for proper learning of this concept class was the exponential-time algorithm based on brute-force search.

We emphasize that even for the most basic class of multi-dimensional concepts — the class of halfspaces (a.k.a. linear classifiers) — our work yields significant improvements over what was known previously. The prior state-of-the-art algorithm of [DKK+21] runs in time $d^{O(1/\varepsilon^4)} + (1/\varepsilon)^{O(1/\varepsilon^6)}$, where d is the data dimension and ε is the accuracy parameter. Our algorithm for this task has the improved run-time of $d^{O(1/\varepsilon^2)} + (1/\varepsilon)^{O(1/\varepsilon^{2.5})}$. In particular, the dependence of our run-time on the dimension d is likely to be optimal as it matches the state-of-the-art improper learning run-time $d^{O(1/\varepsilon^2)}$ [KKMS08, DKN10], as well as the run-time lower bound of $d^{\Omega(1/\varepsilon^2)}$ for statistical query algorithms [DKZ20, GGK20, DKPZ21b].

In addition to proper agnostic learning of halfspaces and Boolean functions of halfspaces, our approach extends to *intersections of halfspaces* — also yielding the first efficient proper agnostic learning algorithm under the Gaussian distribution for this hypothesis class. Overall, we summarize our run-times in Table 1.

Concept class	**This Work**	**Prior Best**
Boolean functions of K halfspaces	$d^{O(K^2 \log(1/\varepsilon)/\varepsilon^2)} + (K/\varepsilon)^{O(K^3/\varepsilon^{2.5})}$	$(1/\varepsilon)^{\mathrm{poly}(dK)}$ (brute force search)
Intersections of K halfspaces	$d^{O(\log K \, \log(1/\varepsilon)/\varepsilon^2)} + \left(\sqrt{\log K}/\varepsilon\right)^{O(K^2\sqrt{\log K}/\varepsilon^{2.5})}$	$(1/\varepsilon)^{\mathrm{poly}(dK)}$ (brute force search)
Halfspaces	$d^{O(1/\varepsilon^2)} + (1/\varepsilon)^{O(1/\varepsilon^{2.5})}$	$d^{O(1/\varepsilon^4)} + (1/\varepsilon)^{O(1/\varepsilon^6)}$ [DKK+21]

Table 1: Run-time comparison for proper agnostic learning under Gaussian marginals.

Comparison with improper learning algorithms. Our run-time for proper agnostic learning of functions of K halfspaces is close to matching the $d^{\tilde{O}(K^2/\varepsilon^2)}$ run-time of the state-of-the-art *improper* learning algorithm of [KKMS08, KOS08, PSW26]. That algorithm achieves error $\mathrm{OPT}_K + \varepsilon$, but the classifier it produces is a Polynomial Threshold Function (PTF), i.e. of the form $\mathrm{sign}(P(\mathbf{x}))$ where P is a low-degree multivariate polynomial[2].

Such improper learning algorithms have the obvious drawback of outputting a highly complex function—a PTF—whose accuracy is only guaranteed to match that of the best classifier in a much simpler hypothesis class (here, functions of K halfspaces). Moreover, the use of PTFs incurs overhead in storage, evaluation time, and adversarial robustness of the decision boundary [LV25b].

Due to these limitations, matching the run-time of improper learning algorithms with proper ones has been a major research direction in computational learning theory [DKK+21, LRV22, LV25a, PJMMN25, BLQT22]. We come close to achieving this for functions of halfspaces: the dependence of our run-time $d^{O(K^2 \log(1/\varepsilon)/\varepsilon^2)} + (K/\varepsilon)^{O(K^3/\varepsilon^{2.5})}$ on the dimension d matches that of the best known *improper* learning algorithm, $d^{O(K^2 \log(1/\varepsilon)/\varepsilon^2)}$ [KKMS08, KOS08, PSW26].

Comparison to computational hardness results. Based on statistical query lower bounds of [DKZ20, GGK20, DKPZ21b], it is believed that any agnostic learning algorithm for the class

[2]A learning algorithm for a hypothesis class $\mathcal{F}$ is called *improper* if the hypothesis it produces need not itself lie in $\mathcal{F}$.

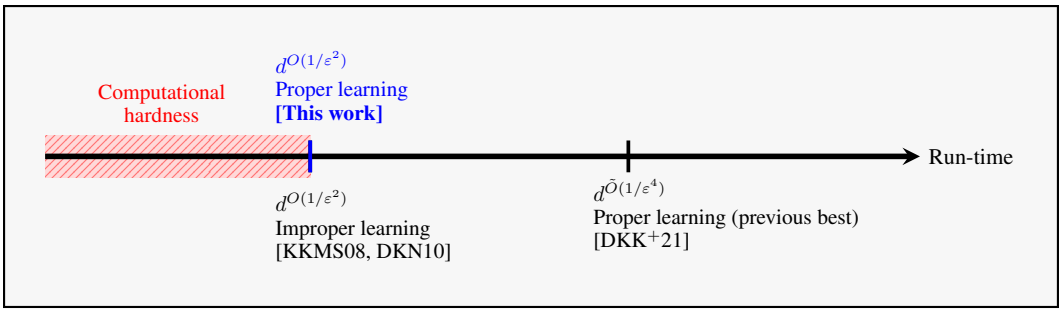

Figure 1: Agnostic proper learning of a single halfspace from Gaussian data in the *moderate-accuracy regime* ($\varepsilon \geq \frac{1}{\log(d)}$). Our algorithm improves over the previous best proper agnostic learning run-time and matches the best known improper learning run-time. The highlighted region indicates run-time lower bounds for statistical query algorithms [DKZ20, GGK20, DKPZ21b].

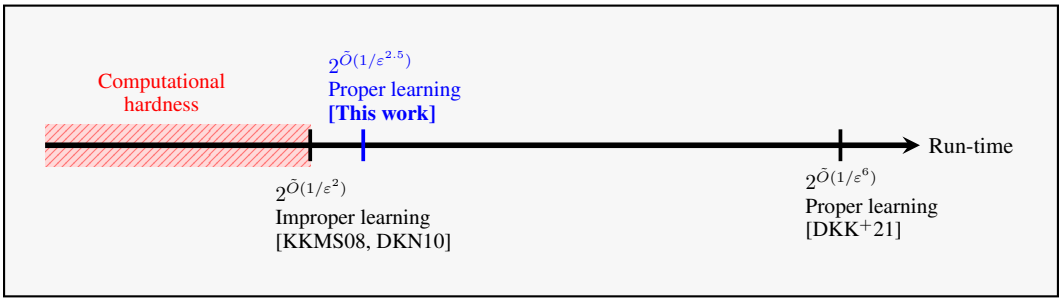

Figure 2: Agnostic proper learning of a single halfspace from Gaussian data in the *high-accuracy regime* ($\varepsilon \leq \frac{1}{\log^2(d)}$). Our algorithm improves over the previous best proper agnostic learning run-time and comes close to matching the best known improper learning run-time. The highlighted region indicates run-time lower bounds for statistical query algorithms [DKZ20, GGK20, DKPZ21b].

of halfspaces under the Gaussian distribution needs to run in time at least $d^{\Omega(1/\varepsilon^2)}$. This belief is further substantiated by run-time lower bounds based on cryptographic hardness assumptions [DKR23]. Therefore, the run-time of our algorithm for proper agnostic learning of a single halfspace is close to matching not only that of the state-of-the-art improper agnostic learning algorithm [KKMS08, DKN10], but also the run-time lower bounds for the agnostic learning of halfspaces under Gaussian data points.

Overall, see Figures 1 and 2 for a summary of how our run-time compares with the previous best run-time of [DKK+21] and computational intractability results for statistical query algorithms [DKZ20, GGK20, DKPZ21b].

- We see that in the moderate-accuracy regime, where $\varepsilon \geq \frac{1}{\log(d)}$, our run-time matches the run-time of the best improper learning algorithm [KKMS08, DKN10] and statistical query lower bounds (See Figure 1).

- In the high-accuracy regime, where $\varepsilon \leq \frac{1}{\log^2(d)}$, our run-time is $2^{\tilde{O}(1/\varepsilon^{2.5})}$, improving on the run-time of $2^{\tilde{O}(1/\varepsilon^6)}$ of [DKK+21] and coming close to matching the run-time $2^{\tilde{O}(1/\varepsilon^2)}$ of the improper learning algorithm [KKMS08, DKN10].

For Boolean functions of K halfspaces, our dimension-dependent exponent matches the SQ lower-bound exponent $\Omega(K^2/\varepsilon^2)$ up to a $\log(1/\varepsilon)$ factor, for $\varepsilon \geq d^{-c}$ with sufficiently small $c > 0$ [DKPZ21b, Theorem 4 and Proposition 21]. Likewise, for the class of intersections of K halfspaces, our run-time comes close to matching the state-of-the-art run-time lower bounds of $d^{\widetilde{\Omega}(\log K)}$ for statistical query due to [DKPZ21a, HSSVG22].

1.1 Our Contributions

Our main theorem is the first proper learning algorithm for Boolean functions of halfspaces under Gaussian marginals.

Theorem 1.2 (Proper learning of Boolean functions of halfspaces)**.** *There exists a proper agnostic learner for the class of arbitrary Boolean functions of halfspaces under Gaussian marginals with the following guarantee. For every $K \geq 1$, every $\varepsilon, \delta \in (0, 1/2)$, and every distribution D over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(0, \mathbf{I}_d)$, the learner draws $N = d^{O(K^2 \log(1/\varepsilon)/\varepsilon^2)} \log(1/\delta)$ samples, runs in time $d^{O(K^2 \log(1/\varepsilon)/\varepsilon^2)} \log(1/\delta) + (K/\varepsilon)^{O(K^3/\varepsilon^{2.5})} \log(1/\delta)$, and returns a hypothesis $h \in \mathcal{C}_K$ such that, with probability at least $1 - \delta$, $\mathrm{err}_D(h) \leq \mathrm{OPT}_K + \varepsilon$.*

For Boolean functions of K halfspaces, the improper learner by Klivans et al. [KOS08] runs in time $d^{O(K^2/\varepsilon^4)}$, while the result of Pesenti et al. [PSW26] improves this to $d^{\widetilde{O}(K^2/\varepsilon^2)}$. This matches the dimension dependence in the running time of our proper learner.

Moreover, in the case of a single halfspace ($K = 1$), we give an algorithm with a slightly faster time $d^{O(1/\varepsilon^2)}$ (Theorem A.13), matching the best known improper polynomial regression exponent under Gaussian marginals [KKMS08, DKN10]. See Appendix C for proper agnostic learning of intersections of halfspaces.

1.2 Algorithm overview

For the sake of simplicity, let us first focus on the problem of proper agnostic learning of a single halfspace. We discuss the more general case of Boolean functions of halfspaces in Section 1.3. Recall that the previous best run-time for this task was achieved by the algorithm of [DKK+21] that follows the outline below:

1. Compute a degree-$1/\varepsilon^4$ polynomial P that approximately minimizes the L_2 error, i.e. minimizing $\sqrt{\mathbb{E}_{\mathbf{x},y\sim D}[(y - P(\mathbf{x}))^2]}$.
2. Compute a matrix $\mathbf{M}(P) = \mathbb{E}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\top]$, and set V to be the span of eigenvectors of $\mathbf{M}(P)$ whose eigenvalues are sufficiently large.
3. Perform a brute-force search among all halfspaces with normals in V and select the most accurate halfspace in this collection (as measured on a validation set).

A key obstacle for improving the run-time of [DKK+21] is that the first step above is based on L_2 regression, and one **needs** degree-$1/\varepsilon^4$ polynomial to ε-approximate a halfspace in L_2 error. Yet, even writing down a degree-$1/\varepsilon^4$ polynomial requires $d^{\Omega(1/\varepsilon^4)}$ time and space.

Instead of using the L_2 loss, we use the following regularized convex objective

$$\mathcal{L}_{\mu,\nu}(P) := \mathbb{E}[\ell(yP(\mathbf{x}))] + \mu\, \mathbb{E}[P(\mathbf{x})^2] + \nu\, \mathrm{tr}\big(\mathbf{M}(P)^{1/2}\big), \tag{1}$$

where the loss term $\mathbb{E}[\ell(yP(\mathbf{x}))]$ is chosen so that its negative derivative[3] $\psi(u) := -\ell'(u)$ satisfies the following properties:

1. **Boundedness:** ψ is uniformly bounded.
2. **Lipschitzness:** ψ is Lipschitz.
3. **Label-flip symmetry:** $\psi(yP(\mathbf{x})) + \psi(-yP(\mathbf{x})) = 1$.

Boundedness of the negative derivative ψ makes sure the function ℓ does not grow too fast. At a high level, this allows us to argue the correctness of our algorithms using the existence of a polynomial that approximates halfspaces in L_1 norm (and satisfies a number of other technical conditions). This allows us to use a lower-degree polynomial because, unlike the case of the L_2 norm, approximating halfspaces under the Gaussian in L_1 norm (Lemma A.8) requires only degree $O(1/\varepsilon^2)$. Lipschitzness of ψ provides the needed smoothness; in particular, it allows us to control the change in the objective under small perturbations of the polynomial P, which is an important technique in our analysis (Lemma A.10). The label-flip symmetry gives a useful decomposition of

[3]The loss $\ell(u)$ is decreasing in the margin u, $\ell'(u) \leq 0$. We define $\psi(u) := -\ell'(u)$ so that ψ is nonnegative.

the loss derivative (Lemma 3.5): it separates the label-dependent term from the term depending only on the fitted polynomial P.

The logistic loss $\ell(u) := \log(1 + e^{-u})$ [Ber44, Cox58], whose negative derivative is $\psi(u) := -\ell'(u) = 1/(1 + e^u)$, satisfies the boundedness, Lipschitzness, and label-flip symmetry conditions. In this paper, we use the logistic loss, allowing us to work with a classical and well-studied convex surrogate.

The term $\mathbb{E}[P(\mathbf{x})^2]$ is included mainly for making the objective strongly convex; see Appendix F. This rules out the well-known pathology of unregularized logistic regression in separable or degenerate cases, where the infimum is approached only as the coefficient norm diverges [AA84].

The term $\mathrm{tr}(\mathbf{M}(P)^{1/2})$ controls the spectrum of $\mathbf{M}(P) = \mathbb{E}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$. Since the final search subspace V is defined by the large-eigenvalue directions of $\mathbf{M}(P)$, this regularizer helps keep the search space low-dimensional; see Lemma 3.10. It also has a natural convex-analytic interpretation: by Lemma E.2, $\mathrm{tr}(\mathbf{M}(P)^{1/2})$ is the nuclear norm (Definition E.1) of the Hermite coefficient matrix of ∇P.

The trace of $\mathbf{M}(P)^{1/2}$ can be viewed as a convex surrogate for controlling the rank of $\mathbf{M}(P)$, in the same way that the lasso is used as a convex surrogate for sparsity [Tib96, Don06].

Algorithm 1 Agnostic Proper Learner for Halfspaces

procedure AGNOSTIC-PROPER-LEARNER(ε, δ, D)
  Input: $\varepsilon > 0$, $\delta > 0$ and sample access to a distribution D on $\mathbb{R}^d \times \{\pm 1\}$
  Output: A halfspace h such that $\mathrm{err}_D(h) \le \mathrm{OPT} + \varepsilon$ with probability at least $1 - \delta$
  Set $k \leftarrow C_0\varepsilon^{-2}$ and $\eta \leftarrow \varepsilon^2/C_0$ ▷ C_0 is a sufficiently large absolute const
  Set $\mu \leftarrow 1/128$ and $\nu \leftarrow c_\nu\varepsilon^{3/2}$ ▷ c_ν is a sufficiently small absolute const
  Find $P \in \mathcal{P}_k$ such that $\mathcal{L}_{\mu,\nu}(P) \le \inf_{P' \in \mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + O(\varepsilon^3)$ using Proposition 3.1
  Let $\mathbf{M}(P) = \mathbb{E}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$
  Let V be the span of the eigenvectors of $\mathbf{M}(P)$ whose eigenvalues are at least η
  Construct a ε-cover $\mathcal{H}$ of halfspaces with normals in V using Fact 3.11
  Draw $N_2 = O\big(\log(|\mathcal{H}|/\delta)/\varepsilon^2\big)$ i.i.d. samples from D and let $\widehat{D}$ be the empirical distribution
  $h \leftarrow \arg\min_{h' \in \mathcal{H}} \mathrm{err}_{\widehat{D}}(h')$
  return h

The algorithm computes a polynomial P, a matrix $\mathbf{M}(P)$, the subspace V spanned by eigenvectors of $\mathbf{M}(P)$ with eigenvalues at least η, and a finite cover of halfspaces with normals in V. The function f, the missing direction $\boldsymbol{\xi}$, the averaged function f_V, the approximating polynomial S, and the perturbation Q are introduced only in the analysis.

1.3 Extension to functions of halfspaces

We show how to extend the approach above from a single halfspace to arbitrary Boolean functions of several halfspaces. The core extension is based on the same principle as in the halfspace case: a polynomial is used as a device for identifying the low-dimensional structure on which a near-optimal classifier depends. Once this structure is recovered, we discard the polynomial and perform a search inside the original concept class.

The main additional difficulty is that the optimal classifier may now combine several halfspaces through an arbitrary Boolean rule. Thus, the decision boundary need not be close to a single hyperplane. Nevertheless, such classifiers still depend on the input only through the normals of the underlying halfspaces. Consequently, any near-optimal function of halfspaces varies meaningfully only along a low-dimensional subspace generated by these normals.

Our regularized polynomial objective (1) is designed to recover such subspace. As before, we optimize a convex surrogate loss over low-degree polynomials (Proposition 3.1) and then examine the gradient matrix $\mathbf{M}(P) = \mathbb{E}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$ of the computed polynomial. The eigenvectors with large eigenvalues identify directions along which P varies most on average and help us find a low-dimensional subspace. The spectral regularizer prevents the polynomial from spreading its variation across too many directions, so the large-eigenvalue eigenspace captures the directions that

are relevant for the near-optimal function of K halfspaces. The structural part of the argument is formalized in Theorem 3.9, which shows that this large-eigenvalue subspace contains the directions needed to approximate a near-optimal function of halfspaces.

The main technical difference from the case of proper agnostic learning a single halfspace is in the analysis — we can no longer average directions one at a time. Instead, we need to average over an entire subspace of low-influence directions. This is where the subspace Gaussian Poincaré inequality (Lemma I.2) is used. This smoothing step is formalized in Lemma 3.6 and is the key ingredient that lets the dimension-reduction argument extend from a single halfspace to Boolean functions of halfspaces.

Once this subspace is identified, we construct a ε-cover of functions of K halfspaces restricted to this subspace and search for the best candidate in the cover. The covering argument, stated in Fact 3.11, shows that this cover contains a proper hypothesis whose error is close to the best achievable error among all Boolean functions of K halfspaces. Finally, empirical risk minimization over the cover selects such a hypothesis with high probability. Lemma B.6 and Fact 3.11 together show how to perform this optimization step efficiently. Combining these ingredients gives the proper agnostic learning guarantee in Theorem 1.2.

Algorithm 2 Agnostic Proper Learner for Boolean functions of Halfspaces

procedure AGNOSTIC-PROPER-LEARNER-BOOLEAN-FUNCTIONS($K, \varepsilon, \delta, D$)
  Input: $K \geq 2$, $\varepsilon > 0$, $\delta > 0$ and sample access to a distribution D on $\mathbb{R}^d \times \{\pm 1\}$
  Output: A hypothesis $h \in \mathcal{C}_K$ such that $\mathrm{err}_D(h) \leq \mathrm{OPT}_K + \varepsilon$ with probability at least $1 - \delta$
  Set $k \leftarrow C_0 K^2 \log(1/\varepsilon)/\varepsilon^2$, $\eta \leftarrow \varepsilon^2/(C_0 K)$ ▷ C_0 is a sufficiently large absolute const
  Set $\mu \leftarrow 1/128$, $\nu \leftarrow c_\nu \varepsilon^{3/2}/K^{3/2}$ ▷ c_ν is a sufficiently small absolute const
  Find $P \in \mathcal{P}_k$ such that $\mathcal{L}_{\mu,\nu}(P) \leq \inf_{P' \in \mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + O(\varepsilon^3)$ using Proposition 3.1
  Let $\mathbf{M}(P) = \mathbb{E}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\top]$
  Let $\mathbf{V}$ be the span of the eigenvectors of $M(P)$ whose eigenvalues are at least η
  Construct a $c\varepsilon/K$-cover $\mathcal{H}$ of halfspaces with normals in $\mathbf{V}$ using Fact 3.11
  Draw $N_2 = O\big((K \log|\mathcal{H}| + 2^K + \log(1/\delta))/\varepsilon^2\big)$ i.i.d. samples from D and let $\widehat{D}$ be the empirical distribution
  For each tuple $(h_1, \ldots, h_K) \in \mathcal{H}^K$, compute the empirically optimal Boolean map $B : \{\pm 1\}^K \to \{\pm 1\}$ using Lemma B.6
  Return $h \leftarrow \arg\min_{\substack{h_1,\ldots,h_K \in \mathcal{H} \\ B:\{\pm 1\}^K \to \{\pm 1\}}} \mathrm{err}_{\widehat{D}}\big(\mathbf{x} \mapsto B(h_1(\mathbf{x}), \ldots, h_K(\mathbf{x}))\big)$

The algorithm computes a polynomial P, a matrix $\mathbf{M}(P)$, and the subspace V spanned by eigenvectors of $\mathbf{M}(P)$ with eigenvalues at least η. It then searches over Boolean functions of K halfspaces using a finite cover of halfspaces with normals in V. The function f, the subspace W, the averaged function f_V, the approximating polynomial S, and the perturbation Q are introduced only in the analysis.

1.4 Related Work

Proper Learning. Improperness is a well-known limitation of the agnostic-learning-from-approximation paradigm of [KKMS08], and there has been significant interest in designing proper agnostic learning algorithms matching its run-time. Despite this interest, few such algorithms are known: in addition to this work and the already-discussed [DKK+21], algorithms exist for proper agnostic learning of monotone functions [LRV22, LV25a] and convex functions [PJMMN25]. The algorithm of [MR02] can also match the run-time of [KKMS08] for proper agnostic learning of decision trees from samples. Our work adds intersections of halfspaces and Boolean functions of halfspaces to this list, while also improving the run-time for proper agnostic learning of halfspaces.

In the *weakly agnostic* setting, where the goal is a classifier with error $O(\mathrm{OPT}) + \varepsilon$ rather than $\mathrm{OPT} + \varepsilon$, the state-of-the-art algorithms [ABL17, DKS18, DKTZ22, DKTZ20] for learning halfspaces under the Gaussian distribution are already proper. These weakly-agnostic results are specific to halfspaces and do not extend to intersections or functions of halfspaces.

In the *realizable*[4] setting — where labels are described exactly by a function in the class — *partially proper* algorithms are known for learning intersections of halfspaces over the Gaussian distribution [Vem10b, Vem10c, Vem10a, DKS18]: when learning an intersection of k halfspaces, they return an intersection of $k' > k$ halfspaces. Prior to our work, no efficient proper agnostic learning algorithm achieving error $\mathrm{OPT} + \varepsilon$ was known for intersections of $k \geq 2$ halfspaces. For Boolean functions of halfspaces, even partially proper efficient algorithms were unknown.

Additionally, in the realizable setting, the recent work of [APS26] gives an improper learning algorithm for intersections of halfspaces under arbitrary data distributions with run-time $2^{\sqrt{d}(\log d)^{O(k)}}\mathrm{poly}(1/\varepsilon)$.

So far we have discussed only proper learning algorithms with access to i.i.d. labeled samples. Proper agnostic learning has also been studied with queries; see [EH89, BLQT22] for work on decision trees and [DKK+24] for work on halfspaces and RELUs.

Broader observations and connections. A common approach to learning multidimensional concepts is to identify their underlying low-dimensional structure. In particular, [DIKZ25] developed an iterative subspace approximation framework for learning multi-index models, including intersections of halfspaces. Although their algorithm is improper, we observe that their algorithm can be adapted to properly learn intersections of halfspaces by performing ERM over a finite cover of intersections in the learned subspace. However, their error guarantee is $K\widetilde{O}(\mathrm{OPT}) + \varepsilon$ in time $\mathrm{poly}(d)2^{\mathrm{poly}(K/\varepsilon)}$. Our work achieves proper learning with the sharper $\mathrm{OPT} + \varepsilon$ guarantee and extends to arbitrary Boolean functions of K halfspaces. The exponent of d in our run-time, however, depends on K and ε, which is believed to be unavoidable due to lower bounds for SQ algorithms.

Convex surrogates. Convex surrogate losses have long been central in statistics and machine learning, beginning with logistic models [Ber44, Cox58] and later support vector machines and boosting [CV95, FHT00]. Although the 0-1 loss is hard to optimize directly due to discontinuity and nonconvexity, convex surrogates are more than computational relaxations: they yield a margin-based formulation in which excess surrogate risk can be related to excess classification risk [BJM06].

Regularization. Regularization was developed to address instability in statistical estimation, particularly for shrinkage and ill-posed problems [HK70, TA77]. Norm penalties control function-class size, and thus margin, complexity, and generalization [CV95, BM02, SS02]. Different regularizers correspond to different structure: L_2 controls norm, L_1 promotes sparsity, and nuclear-norm regularization captures low-rank structure [Tib96, SRJ04, RFP10].

2 Notation

We work with distributions D over pairs $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$, where $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$. The label y may depend on $\mathbf{x}$ arbitrarily. For a classifier h, write $\mathrm{err}_D(h) := \Pr_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y]$.

For the pointwise loss we use ℓ, and for the corresponding population objective we use $\mathcal{L}$. With regularization, we write $\mathcal{L}_{\mu,\nu}$, where $\mu > 0$ and $\nu \geq 0$. We reserve λ for eigenvalues, and $\|\cdot\|_*$ denotes the nuclear norm.

Let $\mathcal{P}_k$ be the space of polynomials on $\mathbb{R}^d$ of degree at most k. Let $\mathcal{C}$ be the class of halfspaces $\mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle + t)$ with $\|\mathbf{w}\|_2 = 1$, together with the two constant classifiers. If $V \subseteq \mathbb{R}^d$ is a subspace, $\mathcal{C}_V$ denotes the subclass whose normal vectors lie in V, again including the two constant classifiers. All halfspace covers implicitly include both constant classifiers, which form the entire cover when $V = \{0\}$. For $K \geq 1$, let $\mathcal{C}_K$ be the class of arbitrary Boolean functions of K halfspaces, $\mathcal{C}_K := \{\mathbf{x} \mapsto B(f_1(\mathbf{x}), \ldots, f_K(\mathbf{x})) : f_1, \ldots, f_K \in \mathcal{C},\ B : \{-1,+1\}^K \to \{-1,+1\}\}$. Similarly, $\mathcal{C}_{K,V}$ denotes the subclass in which all halfspace normal vectors lie in V. We write $\mathrm{OPT} := \inf_{f\in\mathcal{C}} \mathrm{err}_D(f)$, $\mathrm{OPT}_K := \inf_{f\in\mathcal{C}_K} \mathrm{err}_D(f)$, and more generally $\mathrm{OPT}_{\mathcal{F}} := \inf_{f\in\mathcal{F}} \mathrm{err}_D(f)$.

For a subspace $V \subseteq \mathbb{R}^d$, write $V^\perp$ for its orthogonal complement and $\mathbf{x}_V$ for the orthogonal projection of $\mathbf{x}$ onto V. We use $O(\cdot)$, $\Theta(\cdot)$, and $o(\cdot)$ with absolute hidden constants unless stated otherwise, and write $\mathrm{poly}(\cdot)$ for a polynomial in its arguments.

[4]The work of [DKS18] also handles *nasty noise*, with error polynomial in the noise rate.

3 Proper Learning of Boolean Functions of Halfspaces

In this section, we analyze Algorithm 2 and show that with high probability it gives a Boolean function of K halfspaces whose prediction error is ε-close to the optimal error OPT_K among all such functions.

We denote by $\mathcal{C}_K$ the class of arbitrary Boolean functions of K halfspaces

$$\mathcal{C}_K := \Big\{ \mathbf{x} \mapsto B(f_1(\mathbf{x}), \dots, f_K(\mathbf{x})) : f_1, \dots, f_K \in \mathcal{C},\, B : \{-1,+1\}^K \to \{-1,+1\} \Big\}.$$

If $V \subseteq \mathbb{R}^d$ is a subspace, we write $\mathcal{C}_{K,V}$ for the corresponding subclass in which all halfspace normal vectors lie in V.

We first compute a near-minimizer of the regularized logistic-loss objective over low-degree polynomials.

Proposition 3.1 (Regularized logistic-loss polynomial regression)**.** *Fix $\mu > 0$ to be an absolute constant and let $\nu \geq 0$. Let D be a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(0, \mathbf{I}_d)$. Let $k \in \mathbb{Z}_+$ and $\varepsilon, \delta > 0$. There is an algorithm that draws $N = (dk)^{O(k)} \frac{\log(1/\delta)}{\varepsilon^2}$ samples from D, runs in time $\mathrm{poly}(N, d)$, and outputs a polynomial $P \in \mathcal{P}_k$ such that $\mathcal{L}_{\mu,\nu}(P) \leq \inf_{P' \in \mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + \varepsilon$ with probability at least $1 - \delta$.*

A proof can be found in Appendix F.

Klivans et al. [KOS08] showed that Gaussian surface area is a useful instrument for learning broad families of geometric concepts. We recall the definition.

Definition 3.2 (Gaussian surface area)**.** Let γ_d denote the standard Gaussian measure on $\mathbb{R}^d$. For a measurable set $A \subseteq \mathbb{R}^d$ and $\delta > 0$, define $A_\delta := \{\mathbf{x} \in \mathbb{R}^d : \mathrm{dist}(\mathbf{x}, A) \leq \delta\}$. The Gaussian surface area of A is

$$\mathrm{GSA}(A) := \liminf_{\delta \to 0^+} \frac{\gamma_d(A_\delta \setminus A)}{\delta}.$$

For Boolean functions of halfspaces, we use the following Gaussian surface area bound.

Fact 3.3 (GSA bound for Boolean functions of K halfspaces [KOS08, Fact 17])**.** *There exists an absolute constant $C > 0$ such that for every $K \geq 1$ and every $f \in \mathcal{C}_K$,*

$$\mathrm{GSA}(f) \leq O(K).$$

One can view Gaussian surface area as a quantitative measure of the regularity of the decision boundary, which we use to obtain the required low-degree approximation. For a single halfspace, the one-dimensional FT-mollification argument of [DKN10] gives a sharper approximation bound (Appendix A). For multidimensional concepts, such as Boolean functions of halfspaces, the decision boundary is no longer determined by a single normal direction. We therefore use a Hermite-concentration argument based on the Ornstein–Uhlenbeck operator and Gaussian surface area, which applies to general decision boundaries.

Theorem 3.4 (Gaussian L_1 approximation for Boolean functions of halfspaces)**.** *There exist absolute constants $C_{\mathrm{BH}}, C'_{\mathrm{BH}} > 0$ with the following property. For every integer $K \geq 1$, every $\tau \in (0, 1/10)$, and every $f \in \mathcal{C}_K$, there exists a polynomial $S : \mathbb{R}^d \to \mathbb{R}$ of degree at most $C_{\mathrm{BH}} \frac{K^2 \log(1/\tau)}{\tau^2}$ such that*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)} [\,|f(\mathbf{x}) - S(\mathbf{x})|\,] \leq \tau,$$

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)} \|\nabla S(\mathbf{x})\|^2 \leq C'_{\mathrm{BH}} \frac{K^2}{\tau},$$

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)} [S(\mathbf{x})^2] \leq 1.$$

A proof can be found in Appendix H. Compared with the one-dimensional halfspace approximation Theorem A.2, this argument incurs an additional logarithmic factor $\log(1/\tau)$ in the polynomial degree.

The lemma below shows how the label-flip symmetry can be leveraged for decomposing the loss into a label-dependent and a model-dependent component.

Lemma 3.5. *Assume $\psi : \mathbb{R} \to \mathbb{R}$ satisfies $\psi(t)+\psi(-t) = 1$ for all $t \in \mathbb{R}$, and set $\phi(u) := \frac{1}{2} - \psi(u)$. Then ϕ is odd and, for all $y \in \{\pm 1\}$ and $u \in \mathbb{R}$,*

$$y\,\psi(yu) = \frac{1}{2}\,y - \phi(u). \tag{2}$$

If ψ is L-Lipschitz, then so is ϕ.

Proof. The symmetry gives $\phi(-u) = \frac{1}{2} - \psi(-u) = \frac{1}{2} - (1 - \psi(u)) = -\phi(u)$, so ϕ is odd. Hence, since $y \in \{\pm 1\}$, $y\,\psi(yu) = y\left(\frac{1}{2} - \phi(yu)\right) = \frac{1}{2}\,y - y\phi(yu) = \frac{1}{2}\,y - \phi(u)$. Finally, $|\phi(u) - \phi(v)| = |\psi(u) - \psi(v)| \le L|u - v|$. □

We also use the following lemma.

Lemma 3.6 (Subspace Gaussian Poincaré)**.** *Let $W \subseteq \mathbb{R}^d$ be a subspace of dimension r, and let $\{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_r\}$ be an orthonormal basis of W. Write $\mathbf{x} = \mathbf{u} + \mathbf{z}$ with $\mathbf{u} \in W^\perp$ and $\mathbf{z} = \sum_{i=1}^r z_i \boldsymbol{\xi}_i$, where $\mathbf{u}$ and $z_1, \dots, z_r$ are independent and $z_1, \dots, z_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then for every weakly differentiable $G : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}\left[\Big(G(\mathbf{x}) - \mathbb{E}[G(\mathbf{x}) \mid \mathbf{u}]\Big)^2\right] \le \sum_{i=1}^r \mathbb{E}\left[(\partial_{\boldsymbol{\xi}_i} G(\mathbf{x}))^2\right].$$

A proof can be found in Appendix I. We will use this concentration inequality to show that averaging over a low-influence direction changes the relevant quantities by only a small amount.

3.1 Reducing to a Low-Dimensional Search Space

The first step in our analysis is to define the appropriate averaging operator. Fix $f = B(f_1, \dots, f_K) \in \mathcal{C}_K$, where $f_j(\mathbf{x}) = \operatorname{sign}(\langle \mathbf{w}_j, \mathbf{x}\rangle + t_j)$. Let $V \subseteq \mathbb{R}^d$ be a subspace and let $W := \operatorname{span}\{(\mathbf{w}_1)_{V^\perp}, \dots, (\mathbf{w}_K)_{V^\perp}\} \subseteq V^\perp$. Let $\{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_r\}$ be an orthonormal basis of W, where $r = \dim(W) \le K$. Decompose $\mathbf{x} = \mathbf{x}_{W^\perp} + \mathbf{x}_W$, where $\mathbf{x}_{W^\perp}$ and $\mathbf{x}_W$ are the projections of $\mathbf{x}$ onto $W^\perp$ and W, respectively. For $\mathbf{z} = \sum_{i=1}^r z_i \boldsymbol{\xi}_i$ with $z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ independent of $\mathbf{x}_{W^\perp}$, we have a resampling identity

$$\mathbf{x} = \mathbf{x}_{W^\perp} + \mathbf{x}_W \overset{\text{d}}{=} \mathbf{x}_{W^\perp} + \mathbf{z}. \tag{3}$$

This is analogous to the one-dimensional resampling identity used for a single halfspace in (6), except that now $\mathbf{z}$ is a Gaussian vector lying in the subspace W.

Definition 3.7 (Averaging a Boolean function of halfspaces)**.** For each $\mathbf{z} \in W$, define the shifted Boolean function of halfspaces $h_{\mathbf{z}}(\mathbf{x}) := f(\mathbf{x}_{W^\perp} + \mathbf{z})$. Based on the resampling identity (3), define

$$f_V(\mathbf{x}) := \mathop{\mathbb{E}}_{\mathbf{z}}[h_{\mathbf{z}}(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{z}}[f(\mathbf{x}_{W^\perp} + \mathbf{z})]. \tag{4}$$

If $W = \{0\}$, define $f_V(\mathbf{x}) := f(\mathbf{x})$.

For each fixed $\mathbf{z} \in W$, the shifted function $h_{\mathbf{z}}$ belongs to $\mathcal{C}_{K,V}$. Indeed, for each j, set $t_{j,\mathbf{z}} := t_j + \langle (\mathbf{w}_j)_{V^\perp}, \mathbf{z}\rangle$. Since $(\mathbf{w}_j)_{V^\perp} \in W$ and $(\mathbf{w}_j)_V \perp W$, we have

$$f_j(\mathbf{x}_{W^\perp} + \mathbf{z}) = \operatorname{sign}(\langle (\mathbf{w}_j)_V, \mathbf{x}\rangle + t_{j,\mathbf{z}}).$$

Thus each shifted halfspace has normal vector in V. Applying the same Boolean map B gives $h_{\mathbf{z}} \in \mathcal{C}_{K,V}$ for every fixed $\mathbf{z}$, and therefore $f_V \in \operatorname{conv}(\mathcal{C}_{K,V})$.

Remark 3.8. If $(\mathbf{w}_j)_V = 0$, it is a constant function, which we regard as a degenerate halfspace.

Theorem 3.9 (Low-dimensional reduction for Boolean functions of halfspaces)**.** *There exists an absolute constant $C > 0$ and an absolute constant $c_\nu > 0$ such that the following holds. Fix $K \ge 1$ and $\varepsilon \in (0, 1/10)$, and set $k := CK^2 \log(1/\varepsilon)/\varepsilon^2$, $\mu := 1/128$, $\eta := \varepsilon^2/(1024K)$, and $\nu := c_\nu \varepsilon^{3/2}/K^{3/2}$. Let $P \in \mathcal{P}_k$ satisfy $\mathcal{L}_{\mu,\nu}(P) \le \inf_{P' \in \mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + O(\varepsilon^3)$, where the hidden constant is sufficiently small. Let V be the span of the eigenvectors of $\mathbf{M}(P)$ whose eigenvalues are at least η. Then for every $f \in \mathcal{C}_K$, the averaged function f_V from Definition 3.7 satisfies*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \le \varepsilon. \tag{5}$$

Consequently, there exists $h \in \mathcal{C}_{K,V}$ such that

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - h(\mathbf{x}))y] \le \varepsilon.$$

The idea of the proof is to show that averaging the Boolean function over the missing directions does not lose much correlation with the labels. The quantity $f - f_V$ measures the contribution of the directions outside V that is removed by averaging. Thus, proving that $\mathbb{E}[(f - f_V)y]$ is small means that these missing directions do not carry significant label correlation.

The proof, given in Appendix B, has four main steps. First, Lemma B.2 lower bounds the correlation of $f - f_V$ with the weighted labels $y\psi(yP)$ in terms of its correlation with the labels. Second, Lemma B.3 approximates $f - f_V$ by a low-degree polynomial Q. Third, Lemma B.4 shows that the regularization cost of this perturbation is small. Finally, Lemma B.5 uses approximate optimality of P to upper bound the possible gain from such perturbations.

3.2 Finite Search in the Reduced Subspace for Boolean Functions of Halfspaces

In this section, we show that the existence guarantee from Section 3.1 can be converted into a proper learner for arbitrary Boolean functions of halfspaces.

Lemma 3.10 (Dimension bound)**.** *Let V be the span of the eigenvectors of $\mathbf{M} = \mathbf{M}(P)$ with eigenvalues at least $\eta > 0$, and suppose that P is an $O(\varepsilon^3)$-approximate minimizer of $\mathcal{L}_{\mu,\nu}$. Then*

$$\dim(V) \le \frac{\mathrm{tr}(\mathbf{M}^{1/2})}{\sqrt{\eta}} = O\left(\frac{1}{\nu\sqrt{\eta}}\right).$$

To turn the existence of a low-dimensional subspace into a proper learner, we use the same discretization facts as in [DKK+21, Section 3.1]; see also [Ver18, Corollary 4.2.13] for the covering bound. In particular, we construct a standard ε-net $\widetilde{V} \subseteq \{\mathbf{v} \in V : \|\mathbf{v}\|_2 = 1\}$ for the unit sphere of the low-dimensional subspace V, together with a discrete set $\widetilde{T}$ of thresholds $t \in \mathbb{R}$. This procedure results in the ε-cover of candidate halfspaces

$$\mathcal{H} := \{\mathrm{sign}(\langle \mathbf{v}, \mathbf{x}\rangle + t) : \mathbf{v} \in \widetilde{V},\ t \in \widetilde{T}\}.$$

Fact 3.11 (Covering and threshold discretization)**.** *Let $V \subseteq \mathbb{R}^d$ be a subspace with $r = \dim(V)$. For every $\varepsilon \in (0, 1/2)$ and every halfspace $h(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{v}, \mathbf{x}\rangle + t)$ with $\mathbf{v} \in V$ and $\|\mathbf{v}\|_2 = 1$, there exists a halfspace $\widetilde{h}$ in a ε-cover $\mathcal{H}$ of size $(1/\varepsilon)^{O(r)}$ satisfying*

$$\Pr_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[h(\mathbf{x}) \ne \widetilde{h}(\mathbf{x})] \le O(\varepsilon).$$

A brute-force search over all Boolean maps $B : \{-1, +1\}^K \to \{-1, +1\}$ requires considering 2^{2^K} candidates, which is doubly exponential in K. Instead, we use the standard hyperplane-arrangement trick, formalized in Lemma B.6: after the fixed halfspaces partition the sample into cells, the optimal Boolean map is obtained by optimizing independently on each cell.

With these tools in place, we can prove the main learning guarantee, Theorem 1.2. The proof is given in Appendix B.

Acknowledgments and Disclosure of Funding

Arsen Vasilyan was supported by the NSF AI Institute for Foundations of Machine Learning (IFML). His work on this paper was completed while he was at UT Austin.

References

[AA84] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[ABL17] Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63(6):1–27, 2017.

[APS26] Josh Alman, Shyamal Patel, and Rocco A. Servedio. Learning functions of halfspaces. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing (STOC)*, 2026. To appear.

[Ber44] Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365, 1944.

[BJM06] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BLQT22] Guy Blanc, Jane Lange, Mingda Qiao, and Li-Yang Tan. Properly learning decision trees in almost polynomial time. *Journal of the ACM*, 69(6):1–19, 2022.

[BM02] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[Cox58] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B*, 20(2):215–242, 1958.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[DIKZ25] Ilias Diakonikolas, Giannis Iakovidis, Daniel M. Kane, and Nikos Zarifis. Robust learning of multi-index models via iterative subspace approximation. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2231–2239. IEEE, 2025.

[DKK+21] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1522–1551, 2021.

[DKK+24] Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostically learning multi-index models with queries. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1931–1952. IEEE, 2024.

[DKN10] I. Diakonikolas, D. M. Kane, and J. Nelson. Bounded independence fools degree-2 threshold functions. In *Proceedings of the 51st IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2010.

[DKPZ21a] I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals, 2021. arXiv:2102.04401.

[DKPZ21b] 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 Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1552–1584. PMLR, 2021.

[DKR23] Ilias Diakonikolas, Daniel 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*, volume 202, pages 7922–7938. PMLR, 23–29 Jul 2023.

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073, 2018.

[DKTZ20] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Non-convex sgd learns halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 33:18540–18549, 2020.

[DKTZ22] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with adversarial label noise via online gradient descent. In *International Conference on Machine Learning*, pages 5118–5141. PMLR, 2022.

[DKZ20] Ilias Diakonikolas, Daniel Kane, and Nikos Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under Gaussian marginals. *Advances in Neural Information Processing Systems*, 33:13586–13596, 2020.

[Don06] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[EH89] Andrzej Ehrenfeucht and David Haussler. Learning decision trees from random examples. *Information and Computation*, 82(3):231–246, 1989.

[FGKP06] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society, 2006.

[FHT00] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[GGK20] Surbhi Goel, Aravind Gollakota, and Adam Klivans. Statistical-query lower bounds via functional gradients. *Advances in Neural Information Processing Systems*, 33:2147–2158, 2020.

[GR06] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. In *Proceedings of the 47th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 543–552. IEEE Computer Society, 2006.

[HK70] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[HSSVG22] 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. PMLR, 2022.

[KKMS08] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KOS08] A. R. Klivans, R. O'Donnell, and R. A. Servedio. Learning geometric concepts via gaussian surface area. In *49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 541–550, 2008.

[LRV22] Jane Lange, Ronitt Rubinfeld, and Arsen Vasilyan. Properly learning monotone functions via local correction. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 75–86. IEEE, 2022.

[LV25a] Jane Lange and Arsen Vasilyan. Agnostic proper learning of monotone functions: beyond the black-box correction barrier. *SIAM Journal on Computing*, pages FOCS23–1, 2025.

[LV25b] Jane Lange and Arsen Vasilyan. Robust learning of halfspaces under log-concave marginals. In *Advances in Neural Information Processing Systems*, 2025.

[MR02] Dinesh P. Mehta and Vijay Raghavan. Decision tree approximations of boolean functions. *Theoretical Computer Science*, 270(1–2):609–623, 2002.

[Naz03] F. Nazarov. On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a gaussian measure. In *Geometric Aspects of Functional Analysis*, volume 1807 of *Lecture Notes in Mathematics*, pages 169–187. Springer, 2003.

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

[PJMMN25] Renato Ferreira Pinto Jr, Cassandra Marcussen, Elchanan Mossel, and Shivam Nadimpalli. Learning and testing convex functions. *arXiv preprint arXiv:2511.11498*, 2025.

[PSW26] L. Pesenti, L. Slot, and M. Wiedmer. Agnostic learning in (almost) optimal time via gaussian surface area, 2026. arXiv:2603.06027.

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[Riv74] T. J. Rivlin. *The Chebyshev Polynomials*. Wiley, 1974.

[Ros58] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[SRJ04] Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, pages 1329–1336, 2004.

[SS02] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, 2002.

[Sze67] Gábor Szegő. *Orthogonal Polynomials*, volume 23 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, 1967.

[TA77] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. Winston, 1977.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[Vem10a] Santosh S. Vempala. Corrigendum: A random sampling algorithm for learning an intersection of halfspaces. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, page 123. IEEE Computer Society, 2010.

[Vem10b] Santosh S. Vempala. Learning convex concepts from gaussian distributions with PCA. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 124–130. IEEE Computer Society, 2010.

[Vem10c] Santosh S. Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *Journal of the ACM*, 57(6):32:1–32:14, November 2010.

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

A Proper learning of Halfspaces

We will also use the following standard facts.

To compute a polynomial satisfying the required near-optimality condition for this objective, we use the following uniform convergence result.

Proposition A.1 (Regularized logistic-loss polynomial regression)**.** *Fix $\mu > 0$ to be an absolute constant and let $\nu \geq 0$. Let D be a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(0, \mathbf{I}_d)$. Let $k \in \mathbb{Z}_+$ and $\varepsilon, \delta > 0$. There is an algorithm that draws $N = (dk)^{O(k)} \frac{\log(1/\delta)}{\varepsilon^2}$ samples from D, runs in time $\mathrm{poly}(N, d)$, and outputs a polynomial $P \in \mathcal{P}_k$ such that $\mathcal{L}_{\mu,\nu}(P) \leq \inf_{P' \in \mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + \varepsilon$ with probability at least $1 - \delta$.*

A proof can be found in Appendix F.

Theorem A.2 (Gaussian L_1 approximation)**.** *There exists an absolute constant C_{HS} with the following property. For every $\tau \in (0, 1/10)$ and every shift $b \in \mathbb{R}$, there exists a univariate polynomial $S : \mathbb{R} \to \mathbb{R}$ of degree at most $C_{\mathrm{HS}}\tau^{-2}$ such that*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[\,|S(x) - \mathrm{sign}(x+b)|\,] \le \tau,$$
$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[S'(x)^2] \le C\tau^{-1},$$
$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[S(x)^2] \le 5.$$

A proof is included in Appendix G. It is a one-dimensional FT-mollification construction [DKN10]; the only additional point is that, besides L^1-approximation, we also require the control of L^2 and gradient norms.

The next lemma is the conditional version of Gaussian Poincaré inequality.

Lemma A.3 (Conditional Gaussian Poincaré)**.** *Let $\boldsymbol{\xi} \in \mathbb{R}^d$ be a unit vector and write $\mathbf{x} = \mathbf{u} + z\boldsymbol{\xi}$ with $\mathbf{u} \in \boldsymbol{\xi}^\perp$ and $z \sim \mathcal{N}(0,1)$ independent. For every weakly differentiable $G : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}\left[\Big(G(\mathbf{x}) - \mathbb{E}[G(\mathbf{x}) \mid \mathbf{u}]\Big)^2\right] \le \mathbb{E}\left[(\partial_{\boldsymbol{\xi}} G(\mathbf{x}))^2\right].$$

A proof can be found in Appendix I.

A.1 Reducing to a Low-Dimensional Search Space

Let $V \subseteq \mathbb{R}^d$ be a subspace and let $f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle + t)$ be a halfspace with $\|\mathbf{w}\|_2 = 1$. Decompose $\mathbf{w} = \mathbf{w}_V + \mathbf{w}_{V^\perp}$, where $\mathbf{w}_V$ is in V and $\mathbf{w}_{V^\perp}$ is orthogonal to V. If $\mathbf{w}_{V^\perp} \neq 0$, let $\boldsymbol{\xi} := \frac{\mathbf{w}_{V^\perp}}{\|\mathbf{w}_{V^\perp}\|_2} \in V^\perp$ be the normalized missing direction. Decompose $\mathbf{x} = \mathbf{x}_{\boldsymbol{\xi}^\perp} + \mathbf{x}_{\boldsymbol{\xi}}$, where $\mathbf{x}_{\boldsymbol{\xi}^\perp}$ and $\mathbf{x}_{\boldsymbol{\xi}}$ are the projections of $\mathbf{x}$ onto $\boldsymbol{\xi}^\perp$ and $\mathrm{span}\{\boldsymbol{\xi}\}$, respectively. For $z \sim \mathcal{N}(0,1)$ independent of $\mathbf{x}_{\boldsymbol{\xi}^\perp}$, we have a resampling identity

$$\mathbf{x} = \mathbf{x}_{\boldsymbol{\xi}^\perp} + \mathbf{x}_{\boldsymbol{\xi}} \overset{\mathrm{d}}{=} \mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi}. \tag{6}$$

Definition A.4 (Averaging a halfspace)**.** For each $z \in \mathbb{R}$, define the shifted halfspace $h_z(\mathbf{x}) := f(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi})$. Based on the resampling identity (6), define

$$f_V(\mathbf{x}) := \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[h_z(\mathbf{x})] = \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[f(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi})]. \tag{7}$$

If $\mathbf{w}_{V^\perp} = 0$, define $f_V(\mathbf{x}) := f(\mathbf{x})$.

For each fixed $z \in \mathbb{R}$, the shifted function h_z belongs to $\mathcal{C}_V$. Indeed, set $t_z := t + z\langle \mathbf{w}, \boldsymbol{\xi}\rangle$. Since $\boldsymbol{\xi} \in V^\perp$ and $\mathbf{w}_{V^\perp}$ is parallel to $\boldsymbol{\xi}$, we have

$$h_z(\mathbf{x}) = f(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi}) = \mathrm{sign}(\langle \mathbf{w}_V, \mathbf{x}\rangle + t_z). \tag{8}$$

Thus each shifted halfspace has normal vector in V. Therefore $h_z \in \mathcal{C}_V$ for every fixed z, and consequently $f_V \in \mathrm{conv}(\mathcal{C}_V)$.

Remark A.5. If $\mathbf{w}_V = 0$, it is a constant function, which we regard as a degenerate halfspace.

Using Definition A.4 and the definition of the regularizer logistic loss (1), we can construct a subspace that contains a near-optimal halfspace.

Theorem A.6 (Existence of a Low-Dimensional Subspace for Halfspaces)**.** *There exists an absolute constant $C > 0$ and $c_\nu > 0$ such that the following holds. Fix $\varepsilon \in (0, 1/10)$, set $k := C/\varepsilon^2$, $\mu := 1/128$, $\eta := \varepsilon^2/1024$, and $\nu := c_\nu \varepsilon^{3/2}$. Let $P \in \mathcal{P}_k$ satisfy $\mathcal{L}_{\mu,\nu}(P) \le \inf_{P'\in\mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + O(\varepsilon^3)$, where the hidden constant is sufficiently small. Let V be the span of the eigenvectors of $\mathbf{M}(P)$ whose eigenvalues are at least η. Then for every halfspace $f \in \mathcal{C}$, the averaged function f_V from Definition A.4 satisfies*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \le \varepsilon. \tag{9}$$

Consequently, there exists $h \in \mathcal{C}_V$ such that

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - h(\mathbf{x}))y] \le \varepsilon.$$

Proof. Fix a halfspace $f(\mathbf{x}) = \operatorname{sign}(\langle \mathbf{w}, \mathbf{x}\rangle + t)$. If $\mathbf{w}_{V^\perp} = 0$, then $f_V(\mathbf{x}) = f(\mathbf{x})$, and the claim is immediate. Hence assume $\mathbf{w}_{V^\perp} \neq 0$ and let $\boldsymbol{\xi} = \mathbf{w}_{V^\perp}/\|\mathbf{w}_{V^\perp}\|_2$.

The next lemma controls the correlation between $f - f_V$ and the negative derivative of the logistic loss, $\psi(yP(\mathbf{x}))$. We remind the reader that the function ψ is defined in (1).

Lemma A.7 (Correlation with the logistic gradient)**.** *For the averaged function f_V from Definition A.4,*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))\, y\, \psi(yP(\mathbf{x}))] \geq \frac{1}{2} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \frac{1}{2}\sqrt{\eta}. \tag{10}$$

Proof. Using the label-flip property (2),

$$\begin{aligned}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))\, y\, \psi(yP(\mathbf{x}))] = \frac{1}{2} &\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \\ &- \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))\phi(P(\mathbf{x}))].\end{aligned} \tag{11}$$

We claim that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))\phi(P(\mathbf{x}))] \leq \frac{1}{2}\sqrt{\eta}. \tag{12}$$

Let $G(\mathbf{x}) := \phi(P(\mathbf{x}))$ and $G_V(\mathbf{x}) := \mathbb{E}_{z\sim\mathcal{N}(0,1)}[G(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi})]$. Since $G_V(\mathbf{x})$ depends only on $\mathbf{x}_{\boldsymbol{\xi}^\perp}$,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))G_V(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[G_V(\mathbf{x}) \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[f(\mathbf{x}) - f_V(\mathbf{x}) \mid \mathbf{x}_{\boldsymbol{\xi}^\perp}]\right] = 0,$$

where the last equality follows from the definition of f_V. Hence $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))\phi(P(\mathbf{x}))] = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))(G(\mathbf{x}) - G_V(\mathbf{x}))]$. By Lemma A.3, $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(G(\mathbf{x}) - G_V(\mathbf{x}))^2] \leq \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(\partial_{\boldsymbol{\xi}} G(\mathbf{x}))^2]$. Since $|\phi'| \leq 1/4$,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(\partial_{\boldsymbol{\xi}} G(\mathbf{x}))^2] \leq \frac{1}{16} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(\partial_{\boldsymbol{\xi}} P(\mathbf{x}))^2] = \frac{1}{16}\boldsymbol{\xi}^\top \mathbf{M}(P)\boldsymbol{\xi} \leq \frac{\eta}{16},$$

because $\boldsymbol{\xi} \in V^\perp$ and all eigenvalues of $\mathbf{M}(P)$ on $V^\perp$ are smaller than η. Finally, $|f(\mathbf{x}) - f_V(\mathbf{x})| \leq 2$, so the Cauchy–Schwarz inequality gives

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))(G(\mathbf{x}) - G_V(\mathbf{x}))] \leq \|f - f_V\|_2 \|G - G_V\|_2 \leq 2 \cdot \frac{\sqrt{\eta}}{4} = \frac{1}{2}\sqrt{\eta}.$$

This proves (12). Combining (11) and (12) gives (10). □

Lemma A.8 (Low-degree approximation of $f - f_V$)**.** *Fix $\tau \in (0, 1/10)$. Suppose $k \geq C_{\mathrm{HS}}\tau^{-2}$. Then there exists $Q \in \mathcal{P}_k$ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2] \leq 5$ and*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x}))|] \leq 2\tau. \tag{13}$$

Proof. By Theorem A.2, there exists a polynomial S of degree at most $C_{\mathrm{HS}}\tau^{-2}$ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S(\mathbf{x}) - f(\mathbf{x})|] \leq \tau$ and $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[S(\mathbf{x})^2] \leq 5$. Since $k \geq C_{\mathrm{HS}}\tau^{-2}$, we have $S \in \mathcal{P}_k$. Define $S_V(\mathbf{x}) := \mathbb{E}_{z\sim\mathcal{N}(0,1)}[S(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi})]$ and set $Q(\mathbf{x}) := S(\mathbf{x}) - S_V(\mathbf{x})$. Gaussian averaging in the $\boldsymbol{\xi}$-direction preserves degree, so $Q \in \mathcal{P}_k$. Moreover,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[\operatorname{Var}(S(\mathbf{x}) \mid \mathbf{x}_{\boldsymbol{\xi}^\perp})] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[S(\mathbf{x})^2] \leq 5.$$

Also, $Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x})) = (S(\mathbf{x}) - f(\mathbf{x})) - (S_V(\mathbf{x}) - f_V(\mathbf{x}))$. Hence

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x}))|] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S(\mathbf{x}) - f(\mathbf{x})|] + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S_V(\mathbf{x}) - f_V(\mathbf{x})|].$$

By Jensen's inequality,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S_V(\mathbf{x}) - f_V(\mathbf{x})|] &\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}[|S(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi}) - f(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi})|]\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S(\mathbf{x}) - f(\mathbf{x})|],\end{aligned}$$

where the last equality uses the Gaussian resampling identity $\mathbf{x} \overset{d}{=} \mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi}$. Therefore $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x}))|] \leq 2\tau$. □

Lemma A.9 (Control of the regularization terms)**.** *Let $Q = S - S_V$ be the polynomial constructed in Lemma A.8. Then*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P(\mathbf{x})Q(\mathbf{x})] \le \sqrt{5\eta}. \tag{14}$$

$$\mathrm{tr}(\mathbf{M}(Q)^{1/2}) \le O(\tau^{-1/2}). \tag{15}$$

Proof. Recall that $Q(\mathbf{x}) = S(\mathbf{x}) - S_V(\mathbf{x})$, where $S_V(\mathbf{x}) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[S(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi})]$. Hence $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x}) \mid \mathbf{x}_{\boldsymbol{\xi}^\perp}] = 0$. Let $P_V(\mathbf{x}) := \mathbb{E}_{z\sim\mathcal{N}(0,1)}[P(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi})]$. Since $P_V(\mathbf{x})$ depends only on $\mathbf{x}_{\boldsymbol{\xi}^\perp}$, we have $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P_V(\mathbf{x})Q(\mathbf{x})] = 0$, and therefore $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P(\mathbf{x})Q(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(P(\mathbf{x}) - P_V(\mathbf{x}))Q(\mathbf{x})]$. By Lemma A.3,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(P(\mathbf{x}) - P_V(\mathbf{x}))^2] \le \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(\partial_{\boldsymbol{\xi}} P(\mathbf{x}))^2] \le \eta. \tag{16}$$

Using (16) and $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2] \le 5$ gives

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P(\mathbf{x})Q(\mathbf{x})] \le \Big(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\big[(P(\mathbf{x}) - P_V(\mathbf{x}))^2\big] \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\big[Q(\mathbf{x})^2\big]\Big)^{1/2} \le \sqrt{5\eta}.$$

This proves (14).

We now prove (15). By Theorem A.2, S also satisfies $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S(\mathbf{x})\|^2 \le O(\tau^{-1})$. Since $S_V(\mathbf{x}) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[S(\mathbf{x}_{\boldsymbol{\xi}^\perp} + z\boldsymbol{\xi})]$, Jensen's inequality gives $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S_V(\mathbf{x})\|^2 \le \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S(\mathbf{x})\|^2$. Thus, using the elementary inequality $\|a-b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we get that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla Q(\mathbf{x})\|^2 \le 2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S(\mathbf{x})\|^2 + 2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S_V(\mathbf{x})\|^2 \le O(\tau^{-1}).$$

Moreover, S depends only on the direction $\mathbf{w}$, while S_V depends only on the projected direction $\mathbf{w}_V$. Hence $Q = S - S_V$ depends only on $\mathrm{span}\{\mathbf{w}, \mathbf{w}_V\}$, and therefore $\mathrm{rank}(\mathbf{M}(Q)) \le 2$. By Lemma E.3,

$$\mathrm{tr}(\mathbf{M}(Q)^{1/2}) \le \sqrt{\mathrm{rank}(\mathbf{M}(Q))\mathrm{tr}(\mathbf{M}(Q))} = \sqrt{\mathrm{rank}(\mathbf{M}(Q)) \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla Q(\mathbf{x})\|^2} \le O(\tau^{-1/2}).$$

This proves (15). □

For a polynomial $Q \in \mathcal{P}_k$, write

$$\mathsf{Gain}_P(Q) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[Q(\mathbf{x})\, y\, \psi(yP(\mathbf{x}))] - 2\mu \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P(\mathbf{x})Q(\mathbf{x})] - \nu\, \mathrm{tr}\big(\mathbf{M}(Q)^{1/2}\big). \tag{17}$$

Intuitively, $\mathsf{Gain}_P(Q)$ lower-bounds the first-order decrease in the regularized loss $\mathcal{L}_{\mu,\nu}$ along Q. If averaging f to f_V lost too much label correlation, our constructed Q would have large positive gain, contradicting approximate optimality of P.

We next combine the previous estimates to lower bound the gain of the polynomial Q constructed in Lemma A.8. By (13), Lemma A.7, and the bound $|y\psi(yP(\mathbf{x}))| \le 1$,

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[Q(\mathbf{x})\, y\, \psi(yP(\mathbf{x}))] \ge \frac{1}{2} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \frac{1}{2}\sqrt{\eta} - 2\tau. \tag{18}$$

Combining (18), (14), and (15) with the definition of $\mathsf{Gain}_P(Q)$ in (17) gives

$$\mathsf{Gain}_P(Q) \ge \frac{1}{2} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \left(\frac{1}{2} + 2\mu\sqrt{5}\right)\sqrt{\eta} - 2\tau - O(\nu\tau^{-1/2}). \tag{19}$$

Lemma A.10 (Approximate optimality implies no large gain)**.** *Let $P \in \mathcal{P}_k$ satisfy $\mathcal{L}_{\mu,\nu}(P) \le \inf_{P'\in\mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + \alpha$. If $Q \in \mathcal{P}_k$ and $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2] \le 5$, then*

$$\mathsf{Gain}_P(Q) \le O(\sqrt{\alpha}).$$

Proof. If $\mathsf{Gain}_P(Q) \le 0$, the desired upper bound holds immediately because $\mathsf{Gain}_P(Q) \le 0 \le C\sqrt{\alpha}$ for any $C > 0$. Otherwise, $\mathsf{Gain}_P(Q) > 0$, which implies $\mathbb{E}[Q(\mathbf{x})^2] > 0$, and we consider $\zeta > 0$.

By the standard second-order Taylor bound, since $\ell'' \le 1/4$, for any a, b,

$$\ell(b) \le \ell(a) + \ell'(a)(b-a) + \frac{1}{8}(b-a)^2. \tag{20}$$

Apply (20) pointwise with $a = yP(\mathbf{x})$ and $b = y(P(\mathbf{x}) + \zeta Q(\mathbf{x}))$ and take expectation. Since $\ell'(u) = -\psi(u)$, we get

$$\begin{aligned}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[\ell(y(P(\mathbf{x}) + \zeta Q(\mathbf{x})))] \le \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[\ell(yP(\mathbf{x}))] \\ - \zeta \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[Q(\mathbf{x}) y \psi(yP(\mathbf{x}))] + \frac{1}{8}\zeta^2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2].\end{aligned}$$

Adding the regularization terms and using $\mathrm{tr}(M(P+\zeta Q)^{1/2}) \le \mathrm{tr}(\mathbf{M}(P)^{1/2}) + \zeta \mathrm{tr}(\mathbf{M}(Q)^{1/2})$, we get

$$\mathcal{L}_{\mu,\nu}(P + \zeta Q) \le \mathcal{L}_{\mu,\nu}(P) - \zeta \mathsf{Gain}_P(Q) + \left(\mu + \frac{1}{8}\right)\zeta^2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2].$$

Choosing $\zeta = \frac{\mathsf{Gain}_P(Q)}{2\left(\mu+\frac{1}{8}\right)\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2]}$ gives

$$\mathcal{L}_{\mu,\nu}(P + \zeta Q) \le \mathcal{L}_{\mu,\nu}(P) - \frac{\mathsf{Gain}_P(Q)^2}{4\left(\mu+\frac{1}{8}\right)\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2]}.$$

Since P is α-approximately optimal, the decrease is at most α. Using $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2] \le 5$, we get

$$\mathsf{Gain}_P(Q) \le 2\sqrt{5\left(\mu + \frac{1}{8}\right)\alpha}.$$

This proves the lemma. □

We now finish the proof of (9). Choose $\tau := \varepsilon/32$. By the choice of the sufficiently large constant C in $k = C/\varepsilon^2$, the condition $k \ge C_{\mathrm{HS}}\tau^{-2}$ holds. Apply Lemma A.10 with $\alpha = O(\varepsilon^3)$. By taking the hidden constant in the optimization accuracy sufficiently small, as allowed in the theorem statement, we get

$$\mathsf{Gain}_P(Q) \le \frac{\varepsilon}{32}.$$

On the other hand, (19) gives

$$\mathsf{Gain}_P(Q) \ge \frac{1}{2}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \left(\frac{1}{2} + 2\mu\sqrt{5}\right)\sqrt{\eta} - 2\tau - O(\nu\tau^{-1/2}).$$

Since $\eta = \varepsilon^2/1024$, $\mu = 1/128$, and $\nu = c_\nu\varepsilon^{3/2}$ with $c_\nu > 0$ sufficiently small, we get

$$\frac{1}{2}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \le \frac{\varepsilon}{32} + \left(\frac{1}{2} + 2\mu\sqrt{5}\right)\sqrt{\eta} + 2\tau + O(\nu\tau^{-1/2}) \le \frac{\varepsilon}{2}.$$

Therefore

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \le \varepsilon,$$

proving (9).

It remains to extract a single proper halfspace in V. If $\mathbf{w}_{V^\perp} = 0$, then $f \in \mathcal{C}_V$ and the claim is trivial. Otherwise, by Definition A.4, $f_V(\mathbf{x}) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[h_z(\mathbf{x})]$ with $h_z \in \mathcal{C}_V$ for every z. Hence $\mathbb{E}_{z\sim\mathcal{N}(0,1)}\mathbb{E}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - h_z(\mathbf{x}))y] = \mathbb{E}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \le \varepsilon$. Therefore there exists some z such that $\mathbb{E}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - h_z(\mathbf{x}))y] \le \varepsilon$. Since $h_z \in \mathcal{C}_V$, taking $h = h_z$ proves the second claim. □

A.2 Finite Search in the Reduced Subspace for Halfspaces

We now show that the low-dimensional subspace V from Theorem A.6 has dimension independent of d. This allows us to replace the search over $\mathbb{R}^d$ by a finite search over a net in V.

Lemma A.11 (Dimension bound)**.** *Let V be the span of the eigenvectors of $\mathbf{M} = \mathbf{M}(P)$ with eigenvalues at least $\eta > 0$, and suppose that P is an $O(\varepsilon^3)$-approximate minimizer of $\mathcal{L}_{\mu,\nu}$. Then*

$$\dim(V) \le \frac{\mathrm{tr}(\mathbf{M}^{1/2})}{\sqrt{\eta}} = O\left(\frac{1}{\nu\sqrt{\eta}}\right).$$

Proof. Let $\lambda_1, \dots, \lambda_r$ be the eigenvalues of $\mathbf{M}$ corresponding to V, where $r = \dim(V)$. Since $\lambda_i \ge \eta$,

$$\dim(V)\sqrt{\eta} = r\sqrt{\eta} \le \sum_{i=1}^{r} \sqrt{\lambda_i} \le \mathrm{tr}(\mathbf{M}^{1/2}).$$

Also, since $0 \in \mathcal{P}_k$ and the regularization terms are nonnegative,

$$\nu \mathrm{tr}(\mathbf{M}^{1/2}) \le \mathcal{L}_{\mu,\nu}(P) \le \mathcal{L}_{\mu,\nu}(0) + O(\varepsilon^3) = \log 2 + O(\varepsilon^3) = O(1).$$

Thus $\mathrm{tr}(\mathbf{M}^{1/2}) = O(1/\nu)$, and the claim follows. □

To turn the existence of a low-dimensional subspace into a proper learner, we use the same discretization facts as in [DKK+21, Section 3.1]; see also [Ver18, Corollary 4.2.13] for the covering bound. In particular, we construct a standard ε-net $\widetilde{V} \subseteq \{\mathbf{v} \in V : \|\mathbf{v}\|_2 = 1\}$ for the unit sphere of the low-dimensional subspace V, together with a discrete set $\widetilde{T}$ of thresholds $t \in \mathbb{R}$. This procedure results in the ε-cover of candidate halfspaces

$$\mathcal{H} := \{\mathrm{sign}(\langle \mathbf{v}, \mathbf{x}\rangle + t) : \mathbf{v} \in \widetilde{V},\ t \in \widetilde{T}\}.$$

Fact A.12 (Covering and threshold discretization)**.** *Let $V \subseteq \mathbb{R}^d$ be a subspace with $r = \dim(V)$. For every $\varepsilon \in (0, 1/2)$ and every halfspace $h(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{v}, \mathbf{x}\rangle + t)$ with $\mathbf{v} \in V$ and $\|\mathbf{v}\|_2 = 1$, there exists a halfspace $\widetilde{h}$ in a ε-cover $\mathcal{H}$ of size $(1/\varepsilon)^{O(r)}$ satisfying*

$$\Pr_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[h(\mathbf{x}) \ne \widetilde{h}(\mathbf{x})] \le O(\varepsilon).$$

We can now prove Theorem A.13.

Theorem A.13 (Proper learning of halfspaces)**.** *There exists a proper agnostic learner for halfspaces under Gaussian marginals with the following guarantee. For every $\varepsilon, \delta \in (0, 1/2)$ and every distribution D over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(0, \mathbf{I}_d)$, the learner draws $N = d^{O(1/\varepsilon^2)} \log(1/\delta)$ examples, runs in time $d^{O(1/\varepsilon^2)} \log(1/\delta) + (1/\varepsilon)^{O(1/\varepsilon^{2.5})} \log(1/\delta)$, and returns a halfspace h such that, with probability at least $1 - \delta$, $\mathrm{err}_D(h) \le \mathrm{OPT} + \varepsilon$.*

Proof of Theorem A.13. Set $k = \Theta(\varepsilon^{-2})$, $\eta = \Theta(\varepsilon^2)$, $\mu := 1/128$, and $\nu = \Theta(\varepsilon^{3/2})$. By Proposition A.1, using

$$N_1 = d^{O(k)} \mathrm{poly}(1/\varepsilon) \log(1/\delta) = d^{O(1/\varepsilon^2)} \mathrm{poly}(1/\varepsilon) \log(1/\delta)$$

samples we compute a degree-k polynomial P such that $\mathcal{L}_{\mu,\nu}(P) \le \inf_{P' \in \mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + O(\varepsilon^3)$ with probability at least $1 - \delta/3$.

Form the influence matrix $\mathbf{M}(P) = \mathbb{E}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\top]$, let V be the span of the eigenvectors with eigenvalues at least η, and let $r = \dim(V)$. By Lemma A.11,

$$r = O\left(\frac{1}{\nu\sqrt{\eta}}\right) = O(\varepsilon^{-2.5}).$$

Fix any halfspace $f \in \mathcal{C}$ with $\mathrm{err}_D(f) \le \mathrm{OPT} + \varepsilon$. By Theorem A.6, there exists $h_V \in \mathcal{C}_V$ such that

$$\mathbb{E}[(f - h_V)y] \le \varepsilon,$$

hence

$$\mathrm{err}_D(h_V) \le \mathrm{err}_D(f) + \varepsilon/2 \le \mathrm{OPT} + 3\varepsilon/2.$$

Applying Fact A.12, we obtain a ε-cover $\mathcal{H}$ of halfspaces of size $|\mathcal{H}| = (1/\varepsilon)^{O(r)} = (1/\varepsilon)^{O(1/\varepsilon^{2.5})}$ containing a hypothesis $\widetilde{h}$ with

$$\mathrm{err}_D(\widetilde{h}) \le \mathrm{err}_D(h_V) + O(\varepsilon) \le \mathrm{OPT} + O(\varepsilon).$$

Finally, let $h = \arg\min_{g\in\mathcal{H}} \widehat{\mathrm{err}}(g)$, where $\widehat{\mathrm{err}}$ is computed on an independent validation sample of size $N_2 = O\Big(\frac{\log|\mathcal{H}|+\log(1/\delta)}{\varepsilon^2}\Big)$. By Hoeffding's inequality and a union bound over the ε-cover $\mathcal{H}$, with probability at least $1-\delta/3$,

$$\mathrm{err}_D(h) \le \min_{g\in\mathcal{H}} \mathrm{err}_D(g) + \varepsilon \le \mathrm{OPT} + O(\varepsilon).$$

Rescaling ε by a sufficiently small absolute constant gives the stated $\mathrm{OPT}+\varepsilon$ guarantee.

Combining this with the success event of Proposition A.1 completes the proof.

The running time is the sum of the time needed to optimize over the degree-k Hermite coefficients, compute the eigenspace of $\mathbf{M}(P)$, and search over the ε-cover $\mathcal{H}$. The first term is $d^{O(1/\varepsilon^2)}\log(1/\delta)$ by Proposition A.1, and the second is $(1/\varepsilon)^{O(1/\varepsilon^{2.5})}\log(1/\delta)$. □

B Proper Learning of Boolean Functions of Halfspaces

In this appendix, we prove the low-dimensional subspace existence and the resulting learning guarantee for Boolean functions of halfspaces.

Theorem B.1 (Low-dimensional reduction for Boolean functions of halfspaces)**.** *There exists an absolute constant $C>0$ and an absolute constant $c_\nu>0$ such that the following holds. Fix $K\ge 1$ and $\varepsilon\in(0,1/10)$, and set $k := C\frac{K^2\log(1/\varepsilon)}{\varepsilon^2}$, $\mu := 1/128$, $\eta := \frac{\varepsilon^2}{1024K}$, and $\nu := c_\nu\frac{\varepsilon^{3/2}}{K^{3/2}}$. Let $P\in\mathcal{P}_k$ satisfy $\mathcal{L}_{\mu,\nu}(P) \le \inf_{P'\in\mathcal{P}_k}\mathcal{L}_{\mu,\nu}(P') + O(\varepsilon^3)$, where the hidden constant is sufficiently small. Let V be the span of the eigenvectors of $\mathbf{M}(P)$ whose eigenvalues are at least η. Then for every $f\in\mathcal{C}_K$, the averaged function f_V from Definition 3.7 satisfies*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \le \varepsilon. \tag{21}$$

Consequently, there exists $h\in\mathcal{C}_{K,V}$ such that

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - h(\mathbf{x}))y] \le \varepsilon.$$

Proof. Fix $f\in\mathcal{C}_K$. If $W=\{0\}$, then $f_V = f$, and (21) is immediate. Hence assume $W\neq\{0\}$.

The next lemma controls the correlation between the averaging residual $f - f_V$ and the logistic gradient, i.e. the gradient of the logistic loss at the margin yP.

Lemma B.2 (Correlation with the logistic gradient)**.** *For the averaged function f_V from Definition 3.7,*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))\, y\, \psi(yP(\mathbf{x}))] \ge \frac{1}{2}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \frac{1}{2}\sqrt{K\eta}. \tag{22}$$

Proof. Using Lemma 3.5,

$$\begin{aligned}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))\, y\, \psi(yP(\mathbf{x}))] = \frac{1}{2}&\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \\ &- \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))\phi(P(\mathbf{x}))].\end{aligned} \tag{23}$$

We claim that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))\phi(P(\mathbf{x}))] \le \frac{1}{2}\sqrt{K\eta}. \tag{24}$$

Let $G(\mathbf{x}) := \phi(P(\mathbf{x}))$ and $G_V(\mathbf{x}) := \mathbb{E}_{\mathbf{z}}[G(\mathbf{x}_{W^\perp} + \mathbf{z})]$. Since $G_V(\mathbf{x})$ depends only on $\mathbf{x}_{W^\perp}$,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))G_V(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[G_V(\mathbf{x}) \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[f(\mathbf{x}) - f_V(\mathbf{x}) \mid \mathbf{x}_{W^\perp}]\right] = 0,$$

where the last equality follows from the definition of f_V. Hence

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))\phi(P(\mathbf{x}))] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))(G(\mathbf{x}) - G_V(\mathbf{x}))].$$

By Lemma 3.6,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(G(\mathbf{x}) - G_V(\mathbf{x}))^2] \le \sum_{i=1}^{r} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(\partial_{\boldsymbol{\xi}_i} G(\mathbf{x}))^2].$$

Since $|\phi'| \le 1/4$,

$$\sum_{i=1}^{r} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(\partial_{\boldsymbol{\xi}_i} G(\mathbf{x}))^2] \le \frac{1}{16}\sum_{i=1}^{r} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(\partial_{\boldsymbol{\xi}_i} P(\mathbf{x}))^2] = \frac{1}{16}\sum_{i=1}^{r} \boldsymbol{\xi}_i^\top \mathbf{M}(P)\boldsymbol{\xi}_i \le \frac{r\eta}{16} \le \frac{K\eta}{16},$$

because every $\boldsymbol{\xi}_i \in W \subseteq V^\perp$, and all eigenvalues of $\mathbf{M}(P)$ on $V^\perp$ are smaller than η. Finally, $|f(\mathbf{x}) - f_V(\mathbf{x})| \le 2$, so the Cauchy–Schwarz inequality gives

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(f(\mathbf{x}) - f_V(\mathbf{x}))(G(\mathbf{x}) - G_V(\mathbf{x}))] \le \|f - f_V\|_2 \|G - G_V\|_2 \le 2 \cdot \frac{\sqrt{K\eta}}{4} = \frac{1}{2}\sqrt{K\eta}.$$

This proves (24). Combining (23) and (24) gives (22). □

Lemma B.3 (Low-degree approximation of $f - f_V$)**.** *Fix $\tau \in (0, 1/10)$. Suppose $k \ge C_{\mathrm{BH}} \frac{K^2 \log(1/\tau)}{\tau^2}$. Then there exists $Q \in \mathcal{P}_k$ such that*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \left[|Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x}))|\right] \le 2\tau. \tag{25}$$

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2] \le 1 \tag{26}$$

Proof. By Corollary H.5, there exists a polynomial S of degree at most $C_{\mathrm{BH}} \frac{K^2 \log(1/\tau)}{\tau^2}$ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S(\mathbf{x}) - f(\mathbf{x})|] \le \tau$ and $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[S(\mathbf{x})^2] \le 1$. Since $k \ge C_{\mathrm{BH}} \frac{K^2 \log(1/\tau)}{\tau^2}$, we have $S \in \mathcal{P}_k$. Define $S_V(\mathbf{x}) := \mathbb{E}_{\mathbf{z}}[S(\mathbf{x}_{W^\perp} + \mathbf{z})]$ and set $Q(\mathbf{x}) := S(\mathbf{x}) - S_V(\mathbf{x})$. Gaussian averaging in the W-directions preserves degree, so $Q \in \mathcal{P}_k$. Moreover,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[\mathrm{Var}(S(\mathbf{x}) \mid \mathbf{x}_{W^\perp})] \le \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[S(\mathbf{x})^2] \le 1.$$

Also, $Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x})) = (S(\mathbf{x}) - f(\mathbf{x})) - (S_V(\mathbf{x}) - f_V(\mathbf{x}))$. Hence

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \left[|Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x}))|\right] \le \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S(\mathbf{x}) - f(\mathbf{x})|] + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S_V(\mathbf{x}) - f_V(\mathbf{x})|].$$

By Jensen's inequality,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S_V(\mathbf{x}) - f_V(\mathbf{x})|] &\le \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\mathop{\mathbb{E}}_{\mathbf{z}}[|S(\mathbf{x}_{W^\perp} + \mathbf{z}) - f(\mathbf{x}_{W^\perp} + \mathbf{z})|]\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[|S(\mathbf{x}) - f(\mathbf{x})|],\end{aligned}$$

where the last equality uses the Gaussian resampling identity $\mathbf{x} \overset{d}{=} \mathbf{x}_{W^\perp} + \mathbf{z}$. Therefore (25) follows. □

Lemma B.4 (Control of the regularization terms)**.** *Let $Q = S - S_V$ be the polynomial constructed in Lemma B.3. Then*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P(\mathbf{x})Q(\mathbf{x})] \le \sqrt{K\eta}. \tag{27}$$

$$\mathrm{tr}(\mathbf{M}(Q)^{1/2}) \le O\left(\frac{K^{3/2}}{\sqrt{\tau}}\right). \tag{28}$$

Proof. Recall that $Q(\mathbf{x}) = S(\mathbf{x}) - S_V(\mathbf{x})$, where $S_V(\mathbf{x}) = \mathbb{E}_{\mathbf{z}}[S(\mathbf{x}_{W^\perp} + \mathbf{z})]$. Hence $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x}) \mid \mathbf{x}_{W^\perp}] = 0$. Let $P_V(\mathbf{x}) := \mathbb{E}_{\mathbf{z}}[P(\mathbf{x}_{W^\perp} + \mathbf{z})]$. Since $P_V(\mathbf{x})$ depends only on $\mathbf{x}_{W^\perp}$, we have $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P_V(\mathbf{x})Q(\mathbf{x})] = 0$, and therefore

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P(\mathbf{x})Q(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(P(\mathbf{x}) - P_V(\mathbf{x}))Q(\mathbf{x})].$$

By Lemma 3.6,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(P(\mathbf{x}) - P_V(\mathbf{x}))^2] \le \sum_{i=1}^{r} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(\partial_{\boldsymbol{\xi}_i} P(\mathbf{x}))^2] = \sum_{i=1}^{r} \boldsymbol{\xi}_i^\top \mathbf{M}(P)\boldsymbol{\xi}_i \le r\eta \le K\eta.$$

Using (26) and the Cauchy–Schwarz inequality gives

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P(\mathbf{x})Q(\mathbf{x})] \le \sqrt{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[(P(\mathbf{x}) - P_V(\mathbf{x}))^2] \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2]} \le \sqrt{K\eta}.$$

This proves (27).

We now prove (28). By Theorem 3.4, S also satisfies $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S(\mathbf{x})\|^2 \le O(K^2/\tau)$. Since $S_V(\mathbf{x}) = \mathbb{E}_{\mathbf{z}}[S(\mathbf{x}_{W^\perp} + \mathbf{z})]$, Jensen's inequality gives

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S_V(\mathbf{x})\|^2 \le \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S(\mathbf{x})\|^2.$$

Thus, using the elementary inequality $\|\mathbf{a} - \mathbf{b}\|^2 \le 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla Q(\mathbf{x})\|^2 \le 2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S(\mathbf{x})\|^2 + 2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla S_V(\mathbf{x})\|^2 \le O(K^2/\tau).$$

Moreover, by Corollary H.5, S depends only on the span of the normals $\mathbf{w}_1, \dots, \mathbf{w}_K$, while S_V depends only on the span of the projected normals $(\mathbf{w}_1)_V, \dots, (\mathbf{w}_K)_V$. Hence $Q = S - S_V$ depends only on a subspace of dimension at most $2K$, and therefore $\operatorname{rank}(\mathbf{M}(Q)) \le 2K$. By Lemma E.3,

$$\operatorname{tr}(\mathbf{M}(Q)^{1/2}) \le \sqrt{\operatorname{rank}(\mathbf{M}(Q))\operatorname{tr}(\mathbf{M}(Q))} = \sqrt{\operatorname{rank}(\mathbf{M}(Q)) \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla Q(\mathbf{x})\|^2} \le O\left(\frac{K^{3/2}}{\sqrt{\tau}}\right).$$

This proves (28). □

For a polynomial $Q \in \mathcal{P}_k$, write

$$\mathsf{Gain}_P(Q) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[Q(\mathbf{x})\, y\, \psi(yP(\mathbf{x}))] - 2\mu \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P(\mathbf{x})Q(\mathbf{x})] - \nu \operatorname{tr}\big(\mathbf{M}(Q)^{1/2}\big). \tag{29}$$

Intuitively, $\mathsf{Gain}_P(Q)$ lower-bounds the first-order decrease in the regularized loss $\mathcal{L}_{\mu,\nu}$ along Q. If averaging f to f_V lost too much label correlation, our constructed Q would have large positive gain, contradicting approximate optimality of P.

Lemma B.5 (Approximate optimality implies no large gain)**.** *Let $P \in \mathcal{P}_k$ satisfy $\mathcal{L}_{\mu,\nu}(P) \le \inf_{P'\in\mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + \alpha$. If $Q \in \mathcal{P}_k$ and $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2] \le 5$, then*

$$\mathsf{Gain}_P(Q) \le O(\sqrt{\alpha}).$$

Proof. The proof is identical to the proof of Lemma A.10. □

We next combine the previous estimates to lower bound the gain of the polynomial Q constructed in Lemma B.3. By Lemma B.2, (25), and the bound $|y\psi(yP(\mathbf{x}))| \le 1$,

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[Q(\mathbf{x})\, y\, \psi(yP(\mathbf{x}))] \ge \frac{1}{2} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \frac{1}{2}\sqrt{K\eta} - 2\tau. \tag{30}$$

Combining (30), (27), and (28) with the definition of $\mathsf{Gain}_P(Q)$ in (29) gives

$$\mathsf{Gain}_P(Q) \ge \frac{1}{2} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \left(\frac{1}{2} + 2\mu\right)\sqrt{K\eta} - 2\tau - O\left(\nu\frac{K^{3/2}}{\sqrt{\tau}}\right). \tag{31}$$

We now finish the proof of (21). Choose $\tau := \varepsilon/32$. By the choice of the sufficiently large constant C in $k = CK^2\log(1/\varepsilon)/\varepsilon^2$, the condition $k \geq C_{\mathrm{BH}}\frac{K^2\log(1/\tau)}{\tau^2}$ holds. Apply Lemma B.5 with $\alpha = O(\varepsilon^3)$. Since $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[Q(\mathbf{x})^2] \leq 1 \leq 5$, and by choosing the hidden constant in the optimization accuracy sufficiently small as allowed in the premise of the theorem, we get

$$\mathsf{Gain}_P(Q) \leq \frac{\varepsilon}{32}.$$

Combining $\mathsf{Gain}_P(Q) \leq \varepsilon/32$ with (31), and using $\eta = \varepsilon^2/(1024K)$, $\mu = 1/128$, $\tau = \varepsilon/32$, and $\nu = c_\nu\varepsilon^{3/2}/K^{3/2}$ with $c_\nu > 0$ sufficiently small, we get

$$\frac{1}{2}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \leq \frac{\varepsilon}{32} + \left(\frac{1}{2} + 2\mu\right)\sqrt{K\eta} + 2\tau + O\left(\nu\frac{K^{3/2}}{\sqrt{\tau}}\right) \leq \frac{\varepsilon}{2}.$$

Therefore

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \leq \varepsilon,$$

proving (21).

It remains to extract a single proper Boolean function of halfspaces in V. If $W = \{0\}$, then $f \in \mathcal{C}_{K,V}$ and the claim is trivial. Otherwise, by Definition 3.7, $f_V(\mathbf{x}) = \mathbb{E}_{\mathbf{z}}[h_{\mathbf{z}}(\mathbf{x})]$ with $h_{\mathbf{z}} \in \mathcal{C}_{K,V}$ for every $\mathbf{z}$. Hence

$$\mathop{\mathbb{E}}_{\mathbf{z}}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - h_{\mathbf{z}}(\mathbf{x}))y] = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \leq \varepsilon.$$

Therefore there exists some $\mathbf{z}$ such that $\mathbb{E}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - h_{\mathbf{z}}(\mathbf{x}))y] \leq \varepsilon$. Since $h_{\mathbf{z}} \in \mathcal{C}_{K,V}$, taking $h = h_{\mathbf{z}}$ proves the consequent. □

We next show how to optimize the Boolean map once the halfspaces are fixed.

Lemma B.6 (Empirical optimization over the Boolean map)**.** *Fix a sample S and halfspaces $h_1, \ldots, h_K$. For each $b \in \{\pm1\}^K$, let*

$$S_b := \{(\mathbf{x}, y) \in S : (h_1(\mathbf{x}), \ldots, h_K(\mathbf{x})) = b\},$$

and let N_b^+ and N_b^- be the numbers of positive and negative labels in S_b. Then

$$\min_{g:\{\pm1\}^K\to\{\pm1\}} \mathrm{err}_S(g(h_1, \ldots, h_K)) = \frac{1}{|S|}\sum_{b\in\{\pm1\}^K} \min\{N_b^+, N_b^-\}.$$

The minimizer is obtained by assigning each cell its majority label, and can be computed in time $O(K|S| + 2^K)$.

Proof. For fixed $h_1, \ldots, h_K$, the cells S_b partition the sample, and $g(h_1, \ldots, h_K)$ is constant on each cell. Hence the empirical error separates over the cells. On cell S_b, the optimal choice of $g(b)$ is the majority label, contributing $\min\{N_b^+, N_b^-\}/|S|$ to the error. Summing over all $b \in \{\pm1\}^K$ gives the claim. Computing the cell counts requires evaluating the K halfspaces on the sample and then assigning a majority label to each of 2^K cells. □

Theorem B.7 (Proper learning of Boolean functions of halfspaces)**.** *There exists a proper agnostic learner for the class of arbitrary Boolean functions of K halfspaces under Gaussian marginals with the following guarantee. For every $K \geq 1$, every $\varepsilon, \delta \in (0, 1/2)$, and every distribution D over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(0, \mathbf{I}_d)$, the learner draws $N = d^{O(K^2\log(1/\varepsilon)/\varepsilon^2)}\log(1/\delta)$ samples, runs in time $d^{O(K^2\log(1/\varepsilon)/\varepsilon^2)}\log(1/\delta) + (K/\varepsilon)^{O(K^3/\varepsilon^{2.5})}\log(1/\delta)$, and returns a hypothesis $h \in \mathcal{C}_K$ such that, with probability at least $1 - \delta$, $\mathrm{err}_D(h) \leq \mathrm{OPT}_K + \varepsilon$.*

Proof. Set $k = \Theta\left(\frac{K^2\log(1/\varepsilon)}{\varepsilon^2}\right)$, $\eta = \Theta\left(\frac{\varepsilon^2}{K}\right)$, and $\nu = \Theta\left(\frac{\varepsilon^{3/2}}{K^{3/2}}\right)$.

By Proposition 3.1, using $N_1 = d^{O(K^2\log(1/\varepsilon)/\varepsilon^2)}\log(1/\delta)$ samples, we compute a polynomial $P \in \mathcal{P}_k$ such that $\mathcal{L}_{\mu,\nu}(P) \leq \inf_{P'\in\mathcal{P}_k}\mathcal{L}_{\mu,\nu}(P') + O(\varepsilon^3)$ with probability at least $1 - \delta/3$.

Form the influence matrix $\mathbf{M}(P) = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$, let V be the span of the eigenvectors with eigenvalues at least η, and let $r = \dim(V)$. By Lemma 3.10,

$$r = O\left(\frac{1}{\nu\sqrt{\eta}}\right) = O\left(\frac{K^2}{\varepsilon^{2.5}}\right).$$

Fix any $f \in \mathcal{C}_K$ with $\mathrm{err}_D(f) \le \mathrm{OPT}_K + \varepsilon$. By Theorem B.1, there exists $h_V \in \mathcal{C}_{K,V}$ such that

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - h_V(\mathbf{x}))y] \le \varepsilon.$$

Therefore

$$\mathrm{err}_D(h_V) \le \mathrm{err}_D(f) + \frac{\varepsilon}{2} \le \mathrm{OPT}_K + \frac{3\varepsilon}{2}.$$

Let $\mathcal{H} \subseteq \mathcal{C}_V$ be a halfspace cover at accuracy $c\varepsilon/K$, for a sufficiently small absolute constant $c > 0$. By Fact 3.11,

$$|\mathcal{H}| = (K/\varepsilon)^{O(r)}.$$

Define the finite class

$$\mathcal{G} := \left\{\mathbf{x} \mapsto g(h_1(\mathbf{x}), \dots, h_K(\mathbf{x})) : h_1, \dots, h_K \in \mathcal{H},\ g : \{\pm1\}^K \to \{\pm1\}\right\}.$$

We claim that $\mathcal{G}$ contains a hypothesis with error at most $\mathrm{OPT}_K + O(\varepsilon)$.

Indeed, write $h_V(\mathbf{x}) = B_0(g_1(\mathbf{x}), \dots, g_K(\mathbf{x}))$ for some $g_1, \dots, g_K \in \mathcal{C}_V$ and some Boolean map $B_0 : \{-1,+1\}^K \to \{-1,+1\}$. For each i, choose $\widetilde{g}_i \in \mathcal{H}$ such that

$$\Pr_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[g_i(\mathbf{x}) \ne \widetilde{g}_i(\mathbf{x})] \le c\varepsilon/K.$$

Then, by the union bound,

$$\Pr_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[B_0(g_1(\mathbf{x}), \dots, g_K(\mathbf{x})) \ne B_0(\widetilde{g}_1(\mathbf{x}), \dots, \widetilde{g}_K(\mathbf{x}))] \le \sum_{i=1}^{K} \Pr[g_i(\mathbf{x}) \ne \widetilde{g}_i(\mathbf{x})] \le c\varepsilon.$$

Thus

$$\mathrm{err}_D\big(\mathbf{x} \mapsto B_0(\widetilde{g}_1(\mathbf{x}), \dots, \widetilde{g}_K(\mathbf{x}))\big) \le \mathrm{err}_D(h_V) + c\varepsilon \le \mathrm{OPT}_K + O(\varepsilon).$$

Hence

$$\min_{h\in\mathcal{G}} \mathrm{err}_D(h) \le \mathrm{OPT}_K + O(\varepsilon).$$

It remains to select a good hypothesis from $\mathcal{G}$. Since $|\mathcal{G}| \le |\mathcal{H}|^K 2^{2^K}$ and $|\mathcal{H}| = (K/\varepsilon)^{O(r)}$, we have

$$\log|\mathcal{G}| \le O(Kr\log(K/\varepsilon) + 2^K).$$

With $r = O(K^2/\varepsilon^{2.5})$, draw an independent validation sample of size

$$N_2 = O\left(\frac{Kr\log(K/\varepsilon) + 2^K + \log(1/\delta)}{\varepsilon^2}\right) = O\left(\frac{K^3\log(K/\varepsilon)}{\varepsilon^{4.5}} + \frac{2^K + \log(1/\delta)}{\varepsilon^2}\right).$$

By Hoeffding's inequality and a union bound over $\mathcal{G}$, with probability at least $1 - \delta/3$, empirical risk minimization over $\mathcal{G}$ returns a hypothesis h satisfying

$$\mathrm{err}_D(h) \le \min_{h'\in\mathcal{G}} \mathrm{err}_D(h') + \varepsilon/2 \le \mathrm{OPT}_K + O(\varepsilon).$$

Rescaling ε by a sufficiently small absolute constant gives the stated $\mathrm{OPT}_K + \varepsilon$ guarantee.

Algorithmically, we do not enumerate all 2^{2^K} Boolean maps. We enumerate the tuples $(h_1, \dots, h_K) \in \mathcal{H}^K$, and for each fixed tuple compute the empirically optimal Boolean map using Lemma B.6. The cost per tuple is $O(KN_2 + 2^K)$, so the total search time is bounded by

$$|\mathcal{H}|^K \cdot \mathrm{poly}(2^K, K, 1/\varepsilon, \log(1/\delta)).$$

Since $|\mathcal{H}|^K = (K/\varepsilon)^{O(Kr)}$ and $r = O(K^2/\varepsilon^{2.5})$, the factor 2^K is absorbed into

$$(K/\varepsilon)^{O(Kr)} = (K/\varepsilon)^{O(K^3/\varepsilon^{2.5})}.$$

Combining this with the time for computing P and V gives the claimed running time. □

C Proper Learning of Intersections of Halfspaces

Perhaps the most important special case of a Boolean function of K halfspaces is an intersection of halfspaces, obtained by taking their conjunction. We denote this class by $\mathcal{C}_K^{\cap}$, and write $\mathcal{C}_{K,V}^{\cap}$ for the subclass in which all halfspace normal vectors lie in V. We also write

$$\mathrm{OPT}_K^{\cap} := \inf_{f \in \mathcal{C}_K^{\cap}} \mathrm{err}_D(f).$$

Although the previous section already applies to arbitrary Boolean functions of halfspaces, intersections have additional geometric structure. In particular, Nazarov's Gaussian surface area bound gives a sharper approximation theorem, and this leads to a better dependence on K.

Fact C.1 (GSA bound for intersections of halfspaces [Naz03])**.** *There exists an absolute constant $C > 0$ such that for every $K \geq 2$ and every $f \in \mathcal{C}_K^{\cap}$,*

$$\mathrm{GSA}(f) \leq O(\sqrt{\log K}).$$

This improves the general bound $O(K)$ for arbitrary Boolean functions of K halfspaces to $O(\sqrt{\log K})$ for intersections. Using the same Hermite-concentration argument as in Theorem 3.4, we obtain the following sharper approximation theorem.

Theorem C.2 (Gaussian L_1 approximation for intersections of halfspaces)**.** *There exist absolute constants $C_{\mathrm{IHS}}, C'_{\mathrm{IHS}} > 0$ with the following property. For every integer $K \geq 2$, every $\tau \in (0, 1/10)$, and every $f \in \mathcal{C}_K^{\cap}$, there exists a polynomial $S : \mathbb{R}^d \to \mathbb{R}$ of degree at most $C_{\mathrm{IHS}} \frac{\log K \, \log(1/\tau)}{\tau^2}$ such that*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)} [\, |f(\mathbf{x}) - S(\mathbf{x})| \,] \leq \tau,$$

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)} \|\nabla S(\mathbf{x})\|^2 \leq C'_{\mathrm{IHS}} \frac{\log K}{\tau},$$

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)} [S(\mathbf{x})^2] \leq 1.$$

Proof. The proof is identical to that of Theorem 3.4, except that the Gaussian surface area estimate from Fact 3.3 is replaced by Nazarov's sharper bound from Fact C.1. Thus the factor K^2 in the degree and gradient bounds is replaced by $\log K$. □

We now state the corresponding low-dimensional subspace theorem. Its proof is the same as the proof of Theorem B.1, with the approximation theorem above replacing Theorem 3.4.

Theorem C.3 (Existence of a Low-Dimensional Subspace for intersections of halfspaces)**.** *There exists an absolute constant $C > 0$ and an absolute constant $c_\nu > 0$ such that the following holds. Fix $K \geq 2$ and $\varepsilon \in (0, 1/10)$, and set $k := C \frac{\log K \, \log(1/\varepsilon)}{\varepsilon^2}$, $\mu := 1/128$, $\eta := \frac{\varepsilon^2}{1024K}$, and $\nu := c_\nu \frac{\varepsilon^{3/2}}{\sqrt{K \log K}}$. Let $P \in \mathcal{P}_k$ satisfy $\mathcal{L}_{\mu,\nu}(P) \leq \inf_{P' \in \mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + O(\varepsilon^3)$, where the hidden constant is sufficiently small. Let V be the span of the eigenvectors of $\mathbf{M}(P)$ whose eigenvalues are at least η. Then for every $f \in \mathcal{C}_K^{\cap}$, the averaged function f_V from Definition 3.7 satisfies*

$$\mathop{\mathbb{E}}_{(\mathbf{x},y) \sim D} [(f(\mathbf{x}) - f_V(\mathbf{x}))y] \leq \varepsilon.$$

Consequently, there exists $h \in \mathcal{C}_{K,V}^{\cap}$ such that

$$\mathop{\mathbb{E}}_{(\mathbf{x},y) \sim D} [(f(\mathbf{x}) - h(\mathbf{x}))y] \leq \varepsilon.$$

Proof. The proof is identical to that of Theorem B.1, using Corollary H.4 in place of Corollary H.5. The only changes are $k = O\Big(\frac{\log K \, \log(1/\varepsilon)}{\varepsilon^2}\Big)$ and $\mathrm{tr}(\mathbf{M}(Q)^{1/2}) \leq O\Big(\sqrt{\frac{K \log K}{\tau}}\Big)$, instead of $O(K^{3/2}/\sqrt{\tau})$. Accordingly, choosing $\nu = c_\nu \varepsilon^{3/2}/\sqrt{K \log K}$ with $c_\nu > 0$ sufficiently small makes the regularization term in the gain bound at most a small constant multiple of ε. □

We now turn the structural theorem into a proper learner. Since the Boolean map is fixed to be conjunction, there is no search over arbitrary Boolean maps.

Theorem C.4 (Proper learning of intersections of halfspaces)**.** *There exists a proper agnostic learner for $\mathcal{C}_K^{\cap}$ under Gaussian marginals with the following guarantee. For every $K \geq 2$, every $\varepsilon, \delta \in (0, 1/2)$, and every distribution D over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(0, \mathbf{I}_d)$, the learner draws*

$$N = d^{O(\log K \, \log(1/\varepsilon)/\varepsilon^2)} \log(1/\delta)$$

samples, runs in time

$$d^{O(\log K \, \log(1/\varepsilon)/\varepsilon^2)} \log(1/\delta) + \left(\frac{\sqrt{\log K}}{\varepsilon}\right)^{O(K^2\sqrt{\log K}/\varepsilon^{2.5})} \log(1/\delta),$$

and returns a hypothesis $h \in \mathcal{C}_K^{\cap}$ such that, with probability at least $1 - \delta$,

$$\mathrm{err}_D(h) \leq \mathrm{OPT}_K^{\cap} + \varepsilon.$$

Proof. The proof is identical to the proof of Theorem B.7, replacing Theorem B.1 by Theorem C.3 and replacing the finite class $\mathcal{G}$ by the cover of intersections inside V. With $\nu = \Theta(\varepsilon^{3/2}/\sqrt{K \log K})$ and $\eta = \Theta(\varepsilon^2/K)$, Lemma 3.10 gives

$$\dim(V) = O\left(\frac{1}{\nu\sqrt{\eta}}\right) = O\left(\frac{K\sqrt{\log K}}{\varepsilon^{2.5}}\right).$$

After truncating thresholds at $T = \Theta(\sqrt{\log(K/\varepsilon)})$, discretize the thresholds and unit normal vectors at $\Theta(\varepsilon/\sqrt{\log K \, \log(K/\varepsilon)})$. Gaussian tail bounds control the resulting perturbations of the linear forms, while Nazarov's bound [Naz03] controls the Gaussian mass of the corresponding margin strip, yielding an ε-cover of size

$$\left(\frac{\sqrt{\log K}}{\varepsilon}\right)^{O(K \dim(V))} = \left(\frac{\sqrt{\log K}}{\varepsilon}\right)^{O(K^2\sqrt{\log K}/\varepsilon^{2.5})}.$$

Empirical risk minimization over this cover, using an independent validation sample and a union bound, returns a hypothesis with error at most $\mathrm{OPT}_K^{\cap} + \varepsilon$. The running time is the sum of the time needed to compute P and V, and the time needed to search over the cover, which gives the claimed bound. □

D Hermite Polynomials

For a comprehensive discussion of Hermite polynomials, we refer the reader to Section 11.2 of [O'D14] and Chapter 5 of [Sze67].

Definition D.1 (Normalized Hermite basis)**.** For $j \in \mathbb{N}$, let $H_j : \mathbb{R} \to \mathbb{R}$ denote the normalized univariate Hermite polynomial

$$H_j(t) = \frac{(-1)^j}{\sqrt{j!}} \exp(t^2/2) \frac{d^j}{dt^j} \exp(-t^2/2).$$

For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, define $|\alpha| := \sum_{i=1}^d \alpha_i$ and

$$H_\alpha(\mathbf{x}) := \prod_{i=1}^d H_{\alpha_i}(\mathbf{x}_i).$$

We refer to $\{H_\alpha\}_{\alpha \in \mathbb{N}^d}$ as the normalized multivariate Hermite basis.

Fact D.2 (Orthonormality and Parseval)**.** *The basis $\{H_\alpha\}_{\alpha \in \mathbb{N}^d}$ is orthonormal in $L_2(\mathcal{N}(0, \mathbf{I}_d))$:*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)} [H_\alpha(\mathbf{x}) H_\beta(\mathbf{x})] = \mathbf{1}\{\alpha = \beta\}.$$

Consequently, every $P \in \mathcal{P}_k$ has a unique expansion[5]

$$P(\mathbf{x}) = \sum_{|\alpha| \leq k} c_\alpha H_\alpha(\mathbf{x}) = \langle \mathbf{c}, \Phi(\mathbf{x}) \rangle$$

[5] We use the coefficient representation or the feature-map representation depending on convenience.

Parseval's identity gives

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[P(\mathbf{x})^2] = \sum_{|\alpha|\le k} c_\alpha^2.$$

Fact D.3 (Derivative of Hermite Polynomial)**.** *For every multi-index $\alpha \in \mathbb{N}^d$ and every $i \in [d]$,*

$$\partial_i H_\alpha(\mathbf{x}) = \begin{cases} \sqrt{\alpha_i}\, H_{\alpha-\mathbf{e}_i}(\mathbf{x}), & \alpha_i \ge 1, \\ 0, & \alpha_i = 0. \end{cases}$$

Consequently, if $P(\mathbf{x}) = \sum_{|\alpha|\le k} c_\alpha H_\alpha(\mathbf{x})$, then

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)} \|\nabla P(\mathbf{x})\|^2 = \sum_{|\alpha|\le k} |\alpha| c_\alpha^2.$$

Lemma D.4 (Hermite expansion of the gradient)**.** *Let $P(\mathbf{x}) = \sum_{|\alpha|\le k} c_\alpha H_\alpha(\mathbf{x})$. Then*

$$\nabla P(\mathbf{x}) = \sum_{|\beta|\le k-1} \mathbf{a}_\beta(P) H_\beta(\mathbf{x}),$$

where

$$\mathbf{a}_\beta(P) = \left(\sqrt{\beta_1+1}\, c_{\beta+\mathbf{e}_1}, \ldots, \sqrt{\beta_d+1}\, c_{\beta+\mathbf{e}_d}\right)^\top.$$

Proof. For every $i \in [d]$, differentiating the normalized Hermite polynomial coordinatewise gives

$$\partial_i H_\alpha(\mathbf{x}) = \begin{cases} \sqrt{\alpha_i}\, H_{\alpha-\mathbf{e}_i}(\mathbf{x}), & \alpha_i \ge 1, \\ 0, & \alpha_i = 0 \end{cases}$$

Therefore

$$\partial_i P(\mathbf{x}) = \sum_{\substack{|\alpha|\le k \\ \alpha_i \ge 1}} \sqrt{\alpha_i}\, c_\alpha H_{\alpha-\mathbf{e}_i}(\mathbf{x})$$

Reindexing with $\beta = \alpha - \mathbf{e}_i$ and defining $a_{i,\beta}(P) := \sqrt{\beta_i+1}\, c_{\beta+\mathbf{e}_i}$, we get

$$\partial_i P(\mathbf{x}) = \sum_{|\beta|\le k-1} \sqrt{\beta_i+1}\, c_{\beta+\mathbf{e}_i} H_\beta(\mathbf{x}) = \sum_{|\beta|\le k-1} a_{i,\beta}(P) H_\beta(\mathbf{x})$$

Thus the coefficient of H_β in the Hermite expansion of ∇P is precisely

$$\mathbf{a}_\beta(P) = \left(\sqrt{\beta_1+1}\, c_{\beta+\mathbf{e}_1}, \ldots, \sqrt{\beta_d+1}\, c_{\beta+\mathbf{e}_d}\right)^\top$$

□

E The Nuclear Norm

In this appendix we collect the main facts about the regularizer

$$\mathrm{tr}\big(\mathbf{M}(P)^{1/2}\big) = \mathrm{tr}((\mathbb{E}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top])^{1/2}).$$

In particular, we show that it can naturally be viewed as the nuclear norm of the coefficient matrix of ∇P in the Hermite basis.

Definition E.1 (Nuclear norm)**.** Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have singular values $\sigma_1, \ldots, \sigma_r$, where $r = \mathrm{rank}(\mathbf{A})$. Let $\lambda_1, \ldots, \lambda_r$ be the nonzero eigenvalues of $\mathbf{A}^\top\mathbf{A}$ (equivalently, of $\mathbf{A}\mathbf{A}^\top$). The nuclear norm of $\mathbf{A}$ is defined by

$$\|\mathbf{A}\|_* := \sum_{i=1}^r \sigma_i = \sum_{i=1}^r \sqrt{\lambda_i}.$$

Equivalently,

$$\|\mathbf{A}\|_* = \mathrm{tr}\left((\mathbf{A}^\top\mathbf{A})^{1/2}\right) = \mathrm{tr}\left((\mathbf{A}\mathbf{A}^\top)^{1/2}\right). \tag{32}$$

The reason we are interested in this regularizer is that the properties of Hermite polynomials allow us to rewrite $\operatorname{tr}(\mathbf{M}(P)^{1/2})$ as the nuclear norm of a matrix constructed from the Hermite coefficients of the vector-valued polynomial ∇P.

Lemma E.2 (Nuclear norm representation of the spectral regularizer)**.** *Let $\mathbf{a}_\beta(P) \in \mathbb{R}^d$ be the Hermite gradient coefficients from Lemma D.4. Define $\mathbf{A}(P)$ by collecting the vectors $\mathbf{a}_\beta(P)$ as columns:*

$$\mathbf{A}(P) := \big(\mathbf{a}_\beta(P)\big)_{|\beta|\le k-1}$$

Then $\mathbf{M}(P) = \mathbf{A}(P)\mathbf{A}(P)^\top$ and consequently

$$\operatorname{tr}(\mathbf{M}(P)^{1/2}) = \operatorname{tr}\big((\mathbf{A}(P)\mathbf{A}(P)^\top)^{1/2}\big) = \|\mathbf{A}(P)\|_*$$

Proof. By definition, $\mathbf{M}(P) = \mathbb{E}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]$. By Lemma D.4,

$$\mathbf{M}(P) = \mathbb{E}\left[\left(\sum_\beta \mathbf{a}_\beta(P) H_\beta(\mathbf{x})\right)\left(\sum_\gamma \mathbf{a}_\gamma(P) H_\gamma(\mathbf{x})\right)^\top\right]$$

Using orthonormality of the Hermite basis, Fact D.2, this becomes

$$\mathbf{M}(P) = \sum_{|\beta|\le k-1} \mathbf{a}_\beta(P)\mathbf{a}_\beta(P)^\top = \mathbf{A}(P)\mathbf{A}(P)^\top$$

Finally, by the definition of nuclear norm (32),

$$\operatorname{tr}(\mathbf{M}(P)^{1/2}) = \operatorname{tr}\big((\mathbf{A}(P)\mathbf{A}(P)^\top)^{1/2}\big) = \|\mathbf{A}(P)\|_*$$

□

As a result, $P \mapsto \operatorname{tr}(\mathbf{M}(P)^{1/2})$ is convex on $\mathcal{P}_k$, since $P \mapsto \mathbf{A}(P)$ is linear and the nuclear norm is convex. Therefore, $\operatorname{tr}(\mathbf{M}(P)^{1/2})$ can be used as a regularizer in our convex optimization problem.

Lemma E.3 (Rank-trace bound)**.** *For every polynomial $P \in \mathcal{P}_k$,*

$$\operatorname{tr}(\mathbf{M}(P)^{1/2}) \le \sqrt{\operatorname{rank}(\mathbf{M}(P))\operatorname{tr}(\mathbf{M}(P))} = \sqrt{\operatorname{rank}(\mathbf{M}(P))\ \mathbb{E}[\|\nabla P(\mathbf{x})\|_2^2]}.$$

Proof. Let $\lambda_1, \dots, \lambda_r$ be the nonzero eigenvalues of $\mathbf{M}(P)$, where $r := \operatorname{rank}(\mathbf{M}(P))$. Since $\mathbf{M}(P) \succeq 0$, all λ_i are nonnegative. Therefore

$$\operatorname{tr}(\mathbf{M}(P)^{1/2}) = \sum_{i=1}^r \sqrt{\lambda_i}.$$

By Cauchy–Schwarz inequality,

$$\left(\sum_{i=1}^r \sqrt{\lambda_i}\right)^2 \le \left(\sum_{i=1}^r 1\right)\left(\sum_{i=1}^r \lambda_i\right) = r\operatorname{tr}(\mathbf{M}(P)) = \operatorname{rank}(\mathbf{M}(P))\operatorname{tr}(\mathbf{M}(P)).$$

Taking square roots gives

$$\operatorname{tr}(\mathbf{M}(P)^{1/2}) \le \sqrt{\operatorname{rank}(\mathbf{M}(P))\operatorname{tr}(\mathbf{M}(P))}.$$

For the second part, using linearity and cyclic property of trace,

$$\operatorname{tr}(\mathbf{M}(P)) = \operatorname{tr}(\mathbb{E}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\top]) = \mathbb{E}[\nabla P(\mathbf{x})^\top \nabla P(\mathbf{x})] = \mathbb{E}[\|\nabla P(\mathbf{x})\|_2^2].$$

□

F Regularized Logistic-Loss Polynomial Regression

We restate and prove the following theorem. The proof is a uniform convergence result for a Lipschitz loss over a bounded finite-dimensional class. The main technical difficulty is that, in the Hermite basis, the feature map $\Phi(\mathbf{x})$ is unbounded under the Gaussian distribution. To address this, we use a standard truncation argument.

Theorem F.1 (Regularized logistic-loss polynomial regression)**.** *Fix $\mu > 0$ to be an absolute constant and let $\nu \geq 0$. Let D be a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(0, \mathbf{I}_d)$. Let $k \in \mathbb{Z}_+$ and $\varepsilon, \delta > 0$. There is an algorithm that draws $N = (dk)^{O(k)} \frac{\log(1/\delta)}{\varepsilon^2}$ samples from D, runs in time $\mathrm{poly}(N, d)$, and outputs a polynomial $P \in \mathcal{P}_k$ such that $\mathcal{L}_{\mu,\nu}(P) \leq \inf_{P' \in \mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + \varepsilon$ with probability at least $1 - \delta$.*

Proof. Let $\Phi(\mathbf{x}) := (H_\alpha(\mathbf{x})) \in \mathbb{R}^m$ be the normalized Hermite feature map with $m := |\{\alpha \in \mathbb{N}^d : |\alpha| \leq k\}| = \binom{d+k}{k}$. Every $P \in \mathcal{P}_k$ has a unique coefficient vector $\mathbf{c} \in \mathbb{R}^m$ such that

$$P_{\mathbf{c}}(\mathbf{x}) = \langle \mathbf{c}, \Phi(\mathbf{x}) \rangle = \sum_{|\alpha| \leq k} c_\alpha H_\alpha(\mathbf{x})$$

Based on orthogonality of the Hermite basis (Fact D.2),

$$\mathbb{E}[\Phi(\mathbf{x})\Phi(\mathbf{x})^\top] = \mathbf{I}_m, \qquad \mathbb{E}\,\|\Phi(\mathbf{x})\|_2^2 = m, \qquad \mathbb{E}[P_{\mathbf{c}}(\mathbf{x})^2] = \|\mathbf{c}\|_2^2.$$

By the nuclear norm representation of the trace in Lemma E.2, where $\mathbf{A}(\mathbf{c})$ is a fixed linear function of $\mathbf{c}$,

$$\mathrm{tr}(\mathbf{M}(P_{\mathbf{c}})^{1/2}) = \|\mathbf{A}(\mathbf{c})\|_*,$$

Thus, our loss (1) in Hermite basis

$$\mathcal{L}_{\mu,\nu}(\mathbf{c}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D} \ell(y\langle \mathbf{c}, \Phi(\mathbf{x})\rangle) + \mu\|\mathbf{c}\|_2^2 + \nu\|\mathbf{A}(\mathbf{c})\|_*.$$

Let $\ell_0 := \ell(0) = \log 2$ and let $\mathcal{B}_R$ be the Euclidean ball of radius $R := \sqrt{\ell_0/\mu}$:

$$\mathcal{B}_R := \{\mathbf{c} \in \mathbb{R}^m : \|\mathbf{c}\|_2 \leq R\}.$$

Since $\ell \geq 0$ and $\|\mathbf{A}(\mathbf{c})\|_* \geq 0$, we have $\mathcal{L}_{\mu,\nu}(\mathbf{c}) \geq \mu\|\mathbf{c}\|_2^2$, while $\mathcal{L}_{\mu,\nu}(0) = \ell_0$. Hence every population minimizer lies in $\mathcal{B}_R$. Since μ is an absolute constant, $R = O(1)$.

The technical difficulty is that the Hermite feature map $\Phi(\mathbf{x})$ is unbounded under the Gaussian distribution. Let $\Lambda := \frac{8Rm}{\varepsilon}$,

$$G_\Lambda(\mathbf{x}) := \mathbf{1}\{\|\Phi(\mathbf{x})\|_2 \leq \Lambda\}.$$

For $\mathbf{c} \in \mathcal{B}_R$, define

$$\ell_{\Lambda,\mathbf{c}}(\mathbf{x}, y) := G_\Lambda(\mathbf{x})\ell(y\langle \mathbf{c}, \Phi(\mathbf{x})\rangle) + (1 - G_\Lambda(\mathbf{x}))\ell_0.$$

Let the population truncated objective be

$$\mathcal{L}^\Lambda_{\mu,\nu}(\mathbf{c}) := \mathbb{E}[\ell_{\Lambda,\mathbf{c}}(\mathbf{x}, y)] + \mu\|\mathbf{c}\|_2^2 + \nu\|\mathbf{A}(\mathbf{c})\|_*,$$

and define the empirical truncated objective

$$\widehat{\mathcal{L}}^\Lambda_{\mu,\nu}(\mathbf{c}) := \frac{1}{N}\sum_{i=1}^N \ell_{\Lambda,\mathbf{c}}(\mathbf{x}_i, y_i) + \mu\|\mathbf{c}\|_2^2 + \nu\|\mathbf{A}(\mathbf{c})\|_*.$$

The algorithm computes $\widehat{\mathbf{c}} \in \mathcal{B}_R$ such that

$$\widehat{\mathcal{L}}^\Lambda_{\mu,\nu}(\widehat{\mathbf{c}}) \leq \inf_{\mathbf{c}\in\mathcal{B}_R} \widehat{\mathcal{L}}^\Lambda_{\mu,\nu}(\mathbf{c}) + \varepsilon/4, \tag{33}$$

and outputs $\widehat{P} = P_{\widehat{\mathbf{c}}}$.

We first compare the truncated and untruncated population objectives. Since logistic loss ℓ is 1-Lipschitz, for every $\mathbf{c} \in \mathcal{B}_R$,

$$\begin{aligned}
|\mathbb{E}\,\ell(y\langle \mathbf{c}, \Phi(\mathbf{x})\rangle) - \mathbb{E}\,\ell_{\Lambda,\mathbf{c}}(\mathbf{x}, y)| &\leq \mathbb{E}\left[|\langle \mathbf{c}, \Phi(\mathbf{x})\rangle| \mathbf{1}\{\|\Phi(\mathbf{x})\|_2 > \Lambda\}\right] \\
&\leq R\,\mathbb{E}\left[\|\Phi(\mathbf{x})\|_2 \mathbf{1}\{\|\Phi(\mathbf{x})\|_2 > \Lambda\}\right] \\
&\leq R\frac{\mathbb{E}\,\|\Phi(\mathbf{x})\|_2^2}{\Lambda} = \frac{Rm}{\Lambda} = \frac{\varepsilon}{8}.
\end{aligned}$$

The first inequality follows from the 1-Lipschitzness of the logistic loss and the definition of the truncated loss; the second follows from the Cauchy–Schwarz inequality and $\|\mathbf{c}\|_2 \le R$; the third uses Markov-type tail bound.

The regularizers are the same in $\mathcal{L}_{\mu,\nu}$ and $\mathcal{L}^{\Lambda}_{\mu,\nu}$, so

$$\sup_{\mathbf{c}\in\mathcal{B}_R} \left|\mathcal{L}_{\mu,\nu}(\mathbf{c}) - \mathcal{L}^{\Lambda}_{\mu,\nu}(\mathbf{c})\right| \le \frac{\varepsilon}{8}. \tag{34}$$

We next control the empirical error for the truncated loss class. Define $h_{\mathbf{c}}(\mathbf{x}, y) := \ell_{\Lambda,\mathbf{c}}(\mathbf{x}, y) - \ell_0$. For $\mathbf{c} \in \mathcal{B}_R$, we have $|h_{\mathbf{c}}(\mathbf{x}, y)| \le R\Lambda$,

$$\mathbb{E}[h_{\mathbf{c}}(\mathbf{x}, y)^2] \le R^2.$$

Moreover, by symmetrization ([BLM13, Lemma 11.4]) and the contraction principle ([BLM13, Theorem 11.6]),

$$\mathbb{E} \sup_{\mathbf{c}\in\mathcal{B}_R} \left| \frac{1}{N}\sum_{i=1}^{N} h_{\mathbf{c}}(\mathbf{x}_i, y_i) - \mathbb{E}\, h_{\mathbf{c}}(\mathbf{x}, y) \right| \le CR\sqrt{\frac{m}{N}}.$$

Applying Bousquet's version of Talagrand's inequality for bounded empirical processes, see [BLM13, Theorem 12.5], gives that with probability at least $1 - \delta$,

$$\sup_{\mathbf{c}\in\mathcal{B}_R} \left| \frac{1}{N}\sum_{i=1}^{N} \ell_{\Lambda,\mathbf{c}}(\mathbf{x}_i, y_i) - \mathbb{E}\, \ell_{\Lambda,\mathbf{c}}(\mathbf{x}, y) \right| \le C\left(R\sqrt{\frac{m}{N}} + R\sqrt{\frac{\log(e/\delta)}{N}} + R\Lambda\frac{\log(e/\delta)}{N} \right).$$

Since $\Lambda = 8Rm/\varepsilon$ and $R = O(1)$, the right-hand side is at most $\varepsilon/8$ provided

$$N \ge C_\mu m \frac{\log(e/\delta)}{\varepsilon^2}.$$

As $m = \binom{d+k}{k} \le (dk)^{O(k)}$, this is implied by

$$N = (dk)^{O(k)} \frac{\log(e/\delta)}{\varepsilon^2}.$$

Consequently, with probability at least $1 - \delta$,

$$\sup_{\mathbf{c}\in\mathcal{B}_R} \left|\widehat{\mathcal{L}}^{\Lambda}_{\mu,\nu}(\mathbf{c}) - \mathcal{L}^{\Lambda}_{\mu,\nu}(\mathbf{c})\right| \le \frac{\varepsilon}{8}. \tag{35}$$

Let $\mathbf{c}^\star \in \mathcal{B}_R$ be a minimizer of $\mathcal{L}_{\mu,\nu}$ over $\mathbb{R}^m$. On the event above,

$$\mathcal{L}_{\mu,\nu}(\widehat{\mathbf{c}}) \overset{(34)}{\le} \mathcal{L}^{\Lambda}_{\mu,\nu}(\widehat{\mathbf{c}}) + \frac{\varepsilon}{8} \overset{(35)}{\le} \widehat{\mathcal{L}}^{\Lambda}_{\mu,\nu}(\widehat{\mathbf{c}}) + \frac{\varepsilon}{4} \overset{(33)}{\le} \widehat{\mathcal{L}}^{\Lambda}_{\mu,\nu}(\mathbf{c}^\star) + \frac{\varepsilon}{2} \overset{(35)}{\le} \mathcal{L}^{\Lambda}_{\mu,\nu}(\mathbf{c}^\star) + \frac{3\varepsilon}{4} \le \mathcal{L}_{\mu,\nu}(\mathbf{c}^\star) + \varepsilon.$$

Therefore

$$\mathcal{L}_{\mu,\nu}(\widehat{P}) \le \inf_{P'\in\mathcal{P}_k} \mathcal{L}_{\mu,\nu}(P') + \varepsilon.$$

Finally, $\widehat{\mathcal{L}}^{\Lambda}_{\mu,\nu}$ is a finite-dimensional convex objective over $\mathcal{B}_R$: the loss term is convex, $\mu\|\mathbf{c}\|_2^2$ is convex, and $\|\mathbf{A}(\mathbf{c})\|_*$ is convex because $\mathbf{A}$ is linear and the nuclear norm is convex (Lemma E.2). The dimension is $m \le (dk)^{O(k)}$, so an $\varepsilon/4$-approximate minimizer can be found in time polynomial in the number of samples and the feature dimension, which is written as $\mathrm{poly}(N, d)$ with the dependence on k absorbed into $(dk)^{O(k)}$. □

G FT-Mollification

The univariate Gaussian L_1 approximation of a threshold by a degree-$O(\tau^{-2})$ polynomial is a standard consequence of the FT-mollification argument of [DKN10]. We record that the same construction also gives the L_2 and gradient bounds needed in the main text.

Proposition G.1 ([DKN10])**.** *There exists an absolute constant $C_{\mathrm{HS}} > 0$ such that for every $\tau \in (0, 1/10)$ and every $b \in \mathbb{R}$, the FT-mollification construction of [DKN10] yields a polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree at most $C_{\mathrm{HS}}\tau^{-2}$ satisfying*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)} \big[\, |p(x) - \mathrm{sign}(x + b)| \,\big] \le \tau.$$

Fix $\tau \in (0, 1/10)$ and $b \in \mathbb{R}$. Let g be the FT-mollification of the sign function at scale $\Theta(\tau)$. The function g is monotone non-decreasing, bounded in $[-1, 1]$, and satisfies, for every integer $j \geq 0$,

$$\sup_{x \in \mathbb{R}} |g^{(j)}(x + b)| \leq (C/\tau)^j, \qquad \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,1)}[(g'(x + b))^2] \leq C\tau^{-1}.$$

The second bound follows from $g' \geq 0$, $\|g'\|_\infty \leq C/\tau$, $\int_{\mathbb{R}} g'(u)\, du \leq 2$, and the bounded Gaussian density. The construction also gives

$$\mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,1)}\big[|g(x + b) - \operatorname{sign}(x + b)|\big] \leq \tau/2.$$

Choose $k = \Theta(\tau^{-2})$ with a sufficiently large implicit constant, and let p be the degree-k Taylor polynomial at $x = 0$ of the function $x \mapsto g(x + b)$:

$$p(x) := \sum_{j=0}^{k} \frac{g^{(j)}(b)}{j!} x^j.$$

This polynomial satisfies the L_1 guarantee in Proposition G.1.

Fact G.2 (Chebyshev growth bound [Riv74, Theorem 2.20])**.** *Let $R, M > 0$, and let q be a polynomial of degree at most k. If*

$$|q(x)| \leq M \qquad \forall |x| \leq R,$$

then for every $|x| \geq R$,

$$|q(x)| \leq M \left(\frac{2|x|}{R}\right)^k.$$

Lemma G.3. *Let p be the polynomial defined above, where $k = \Theta(\tau^{-2})$ is chosen with a sufficiently large constant. Then, uniformly in b,*

$$\mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,1)}[p(x)^2] \leq 5,$$

$$\mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,1)}[(p'(x))^2] \leq C\tau^{-1}.$$

Proof. Choose a universal constant $\alpha \geq 4$ and set $R := \alpha\sqrt{k}$. By Taylor's theorem and the derivative bounds for g,

$$|g(x + b) - p(x)| \leq \left(\frac{C|x|}{\tau k}\right)^{k+1},$$

$$|g'(x + b) - p'(x)| \leq C\tau^{-1} \left(\frac{C|x|}{\tau k}\right)^k.$$

If $k \geq C\alpha^2\tau^{-2}$, then for all $|x| \leq R$,

$$|p(x)| \leq 2, \qquad |g'(x + b) - p'(x)| \leq C\tau^{-1}4^{-k}.$$

Hence

$$\mathbb{E}[p(x)^2 \mathbf{1}\{|x| \leq R\}] \leq 4,$$

and, using $\mathbb{E}[(g'(x + b))^2] \leq C\tau^{-1}$,

$$\mathbb{E}[(p'(x))^2 \mathbf{1}\{|x| \leq R\}] \leq C\tau^{-1}.$$

By Fact G.2, for all $|x| \geq R$,

$$|p(x)| \leq 2 \left(\frac{2|x|}{R}\right)^k.$$

Therefore, using $\mathbb{E}[|x|^{2k}] = (2k - 1)!! \leq (2k)^k$,

$$\mathbb{E}[p(x)^2] \leq 4 + 4 \left(\frac{2}{R}\right)^{2k} \mathbb{E}[|x|^{2k}] \leq 4 + 4 \left(\frac{8}{\alpha^2}\right)^k \leq 5.$$

Similarly, since $|p'(x)| \le C\tau^{-1}$ on $[-R, R]$, applying Fact G.2 to p' gives

$$|p'(x)| \le C\tau^{-1}\left(\frac{2|x|}{R}\right)^{k-1} \qquad \forall |x| \ge R.$$

Using $\mathbb{E}[|x|^{2m}] \le (2m)^m$ with $m = k-1$, we obtain

$$\mathbb{E}[(p'(x))^2 \mathbf{1}\{|x| > R\}] \le C\tau^{-2}\left(\frac{8}{\alpha^2}\right)^{k-1} \le C\tau^{-1},$$

for sufficiently large absolute constants in the choices of α and $k = \Theta(\tau^{-2})$. Combining the estimates on $|x| \le R$ and $|x| > R$ gives

$$\mathbb{E}[(p'(x))^2] \le C\tau^{-1}.$$

□

H Multidimensional OU

For a comprehensive discussion of Hermite polynomials, we refer the reader to Section 11.2 of [O'D14] and Chapter 5 of [Sze67]. We first prove a general approximation theorem in terms of Gaussian surface area, and then specialize it to intersections of halfspaces and Boolean functions of halfspaces.

Theorem H.1 (Gaussian L_1 approximation from Gaussian surface area)**.** *There exists an absolute constant $C > 0$ such that the following holds. Let $A \ge 1$ and let $f : \mathbb{R}^d \to \{\pm 1\}$ satisfy $\mathrm{GSA}(f) \le A$. For every $\tau \in (0, 1/10)$, there exists a polynomial $S : \mathbb{R}^d \to \mathbb{R}$ of degree at most $C\frac{A^2 \log(1/\tau)}{\tau^2}$ such that, for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$,*

$$\mathbb{E}[|f(\mathbf{x}) - S(\mathbf{x})|] \le \tau, \qquad \mathbb{E}[S(\mathbf{x})^2] \le 1, \qquad \mathbb{E}\,\|\nabla S(\mathbf{x})\|^2 \le C\frac{A^2}{\tau}.$$

Proof. Write the Hermite expansion $f = \sum_\alpha c_\alpha H_\alpha$. For $\rho \in (0, 1)$, let T_ρ denote the Ornstein–Uhlenbeck operator, so that for ρ-correlated standard Gaussians $\mathbf{X}, \mathbf{Y}$,

$$T_\rho f(\mathbf{x}) := \sum_\alpha \rho^{|\alpha|} c_\alpha H_\alpha(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}].$$

Since f is $\{\pm 1\}$-valued,

$$\mathbb{E}[|f(\mathbf{X}) - T_\rho f(\mathbf{X})|] \le \mathbb{E}[|f(\mathbf{X}) - f(\mathbf{Y})|] = 2\Pr[f(\mathbf{X}) \ne f(\mathbf{Y})].$$

By the Gaussian surface area noise-sensitivity bound [KOS08, Corollary 14],

$$\Pr[f(\mathbf{X}) \ne f(\mathbf{Y})] \le C_0\, A\sqrt{1-\rho}.$$

Choose $1 - \rho = c\tau^2/A^2$ for a sufficiently small absolute constant $c > 0$. Then

$$\mathbb{E}[|f - T_\rho f|] \le \frac{\tau}{2}.$$

Now truncate the Hermite expansion of $T_\rho f$ at degree m:

$$S(\mathbf{x}) := \sum_{|\alpha| \le m} \rho^{|\alpha|} c_\alpha H_\alpha(\mathbf{x}).$$

Since f is $\{\pm 1\}$-valued, Parseval's identity Fact D.2 gives

$$\mathbb{E}[S(\mathbf{x})^2] = \sum_{|\alpha| \le m} \rho^{2|\alpha|} c_\alpha^2 \le \sum_\alpha c_\alpha^2 = \mathbb{E}[f(\mathbf{x})^2] = 1.$$

Also,

$$\|T_\rho f - S\|_2^2 = \sum_{|\alpha| > m} \rho^{2|\alpha|} c_\alpha^2 \le \rho^{2m} \sum_\alpha c_\alpha^2 \le \rho^{2m}.$$

Hence

$$\|T_\rho f - S\|_1 \le \|T_\rho f - S\|_2 \le \rho^m.$$

Choose $m \ge \frac{\log(2/\tau)}{\log(1/\rho)}$. Since $\log(1/\rho) \ge 1-\rho$ for $\rho \in (0,1)$, it is enough to take

$$m = O\left(\frac{\log(1/\tau)}{1-\rho}\right) = O\left(\frac{A^2 \log(1/\tau)}{\tau^2}\right).$$

Then $\|T_\rho f - S\|_1 \le \tau/2$, and therefore

$$\mathbb{E}[|f - S|] \le \mathbb{E}[|f - T_\rho f|] + \mathbb{E}[|T_\rho f - S|] \le \tau.$$

It remains to prove the gradient bound. Let $a_n := \sum_{|\alpha|=n} c_\alpha^2$ denote the Hermite weight of f on level n. By Fact D.3,

$$\mathbb{E}\,\|\nabla S(\mathbf{x})\|^2 = \sum_{n=1}^{m} n\rho^{2n} a_n \le \sum_{n\ge 1} n\rho^{2n} a_n.$$

For every $n \ge 1$,

$$n\rho^{2n} \le \frac{1-\rho^n}{1-\rho}.$$

Indeed, $(1-\rho^n)/(1-\rho) = 1 + \rho + \cdots + \rho^{n-1} \ge n\rho^{n-1} \ge n\rho^{2n}$. Therefore

$$\mathbb{E}\,\|\nabla S(\mathbf{x})\|^2 \le \frac{1}{1-\rho} \sum_{n\ge 1} (1-\rho^n) a_n.$$

Since $\mathbb{E}[f(\mathbf{X})f(\mathbf{Y})] = \sum_{n\ge 0} \rho^n a_n$ and $\mathbb{E}[f(\mathbf{X})^2] = 1$, we have

$$\sum_{n\ge 1} (1-\rho^n) a_n = 1 - \mathbb{E}[f(\mathbf{X})f(\mathbf{Y})] = 2\Pr[f(\mathbf{X}) \ne f(\mathbf{Y})].$$

Using the same Gaussian surface area noise-sensitivity bound,

$$\sum_{n\ge 1} (1-\rho^n) a_n \le C\, A \sqrt{1-\rho}.$$

Thus

$$\mathbb{E}\,\|\nabla S(\mathbf{x})\|^2 \le C \frac{A}{\sqrt{1-\rho}}.$$

Substituting $1-\rho = \Theta(\tau^2/A^2)$ gives

$$\mathbb{E}\,\|\nabla S(\mathbf{x})\|^2 \le C \frac{A^2}{\tau}.$$

This completes the proof. □

We now record the Gaussian surface area bounds used in the main text.

Fact H.2 (Gaussian surface area bound for intersections of halfspaces [Naz03])**.** *There exists an absolute constant $C > 0$ such that for every $K \ge 2$ and every $f \in \mathcal{C}_K^{\cap}$,*

$$\mathrm{GSA}(f) \le O(\sqrt{\log K}).$$

Fact H.3 (Gaussian surface area bound for Boolean functions of K halfspaces [KOS08, Fact 17])**.** *There exists an absolute constant $C > 0$ such that for every $K \ge 1$ and every $f \in \mathcal{C}_K$,*

$$\mathrm{GSA}(f) \le O(K).$$

The following two corollaries are the approximation inputs used for intersections and Boolean functions of halfspaces.

Corollary H.4 (Gaussian L_1 approximation for intersections of halfspaces)**.** *There exist absolute constants $C_{\mathrm{IHS}}, C'_{\mathrm{IHS}} > 0$ with the following property. For every integer $K \geq 2$, every $\tau \in (0, 1/10)$, and every $f \in \mathcal{C}_K^{\cap}$, there exists a polynomial $S : \mathbb{R}^d \to \mathbb{R}$ of degree at most $C_{\mathrm{IHS}} \frac{\log K \, \log(1/\tau)}{\tau^2}$ such that, for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$,*

$$\mathbb{E}[|f(\mathbf{x}) - S(\mathbf{x})|] \leq \tau, \qquad \mathbb{E}[S(\mathbf{x})^2] \leq 1, \qquad \mathbb{E}\,\|\nabla S(\mathbf{x})\|^2 \leq C'_{\mathrm{IHS}} \frac{\log K}{\tau}.$$

Proof. By Fact C.1, we may apply Theorem H.1 with $A = C_1\sqrt{\log K}$, for a sufficiently large absolute constant C_1. This gives degree $O\Big(\frac{\log K \, \log(1/\tau)}{\tau^2}\Big)$ and gradient bound

$$\mathbb{E}\,\|\nabla S(\mathbf{x})\|^2 \leq O\left(\frac{\log K}{\tau}\right).$$

The L_1 and L_2 bounds are those of Theorem H.1. □

Corollary H.5 (Gaussian L_1 approximation for Boolean functions of halfspaces)**.** *There exist absolute constants $C_{\mathrm{BH}}, C'_{\mathrm{BH}} > 0$ with the following property. For every integer $K \geq 1$, every $\tau \in (0, 1/10)$, and every $f \in \mathcal{C}_K$, there exists a polynomial $S : \mathbb{R}^d \to \mathbb{R}$ of degree at most $C_{\mathrm{BH}} \frac{K^2 \log(1/\tau)}{\tau^2}$ such that, for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$,*

$$\mathbb{E}[|f(\mathbf{x}) - S(\mathbf{x})|] \leq \tau, \qquad \mathbb{E}[S(\mathbf{x})^2] \leq 1, \qquad \mathbb{E}\,\|\nabla S(\mathbf{x})\|^2 \leq C'_{\mathrm{BH}} \frac{K^2}{\tau}.$$

Moreover, if $f = B(f_1, \ldots, f_K)$ where $f_j(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}_j, \mathbf{x}\rangle + t_j)$, then S depends only on the span of the normal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$.

Proof. By Fact 3.3, we may apply Theorem H.1 with $A = C_1 K$, for a sufficiently large absolute constant C_1. This gives degree $O\Big(\frac{K^2 \log(1/\tau)}{\tau^2}\Big)$ and gradient bound

$$\mathbb{E}\,\|\nabla S(\mathbf{x})\|^2 \leq O\left(\frac{K^2}{\tau}\right).$$

The L_1 and L_2 bounds are those of Theorem H.1. If $f = B(f_1, \ldots, f_K)$, then f depends only on the span of the normal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$; the Ornstein–Uhlenbeck operator and Hermite truncation preserve this dependence. □

I Gaussian Poincaré inequality

We restate and prove the following lemmas.

Lemma I.1 (Conditional Gaussian Poincaré)**.** *Let $\boldsymbol{\xi} \in \mathbb{R}^d$ be a unit vector and write $\mathbf{x} = \mathbf{u} + z\boldsymbol{\xi}$, where $\mathbf{u} \in \boldsymbol{\xi}^{\perp}$ and $z \sim \mathcal{N}(0,1)$ are independent. For every weakly differentiable $G : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}\left[\Big(G(\mathbf{x}) - \mathbb{E}[G(\mathbf{x}) \mid \mathbf{u}]\Big)^2\right] \leq \mathbb{E}\left[(\partial_{\boldsymbol{\xi}} G(\mathbf{x}))^2\right].$$

Proof. Condition on $\mathbf{u} \in \boldsymbol{\xi}^{\perp}$ and set $g_{\mathbf{u}}(z) := G(\mathbf{u} + z\boldsymbol{\xi})$. Then $\mathrm{Var}(G(\mathbf{x}) \mid \mathbf{u}) = \mathrm{Var}(g_{\mathbf{u}}(z))$. By the Gaussian Poincaré inequality [BLM13, Theorem 3.20],

$$\mathrm{Var}(g_{\mathbf{u}}(z)) \leq \mathbb{E}[(g'_{\mathbf{u}}(z))^2 \mid \mathbf{u}].$$

Since $g'_{\mathbf{u}}(z) = \partial_{\boldsymbol{\xi}} G(\mathbf{u} + z\boldsymbol{\xi})$, we get

$$\mathrm{Var}(G(\mathbf{x}) \mid \mathbf{u}) \leq \mathbb{E}[(\partial_{\boldsymbol{\xi}} G(\mathbf{x}))^2 \mid \mathbf{u}].$$

Taking expectation over $\mathbf{u}$ proves the claim. □

Lemma I.2 (Subspace Gaussian Poincaré)**.** *Let $W \subseteq \mathbb{R}^d$ be a subspace of dimension r, and let $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r\}$ be an orthonormal basis of W. Write $\mathbf{x} = \mathbf{u} + \mathbf{z}$, where $\mathbf{u} \in W^{\perp}$, $\mathbf{z} = \sum_{i=1}^r z_i \boldsymbol{\xi}_i$, and $z_1, \ldots, z_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are independent of $\mathbf{u}$. Then for every weakly differentiable $G : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}\left[\Big(G(\mathbf{x}) - \mathbb{E}[G(\mathbf{x}) \mid \mathbf{u}]\Big)^2\right] \leq \sum_{i=1}^{r} \mathbb{E}\left[(\partial_{\boldsymbol{\xi}_i} G(\mathbf{x}))^2\right].$$

Proof. Condition on $\mathbf{u} \in W^\perp$ and set $g_{\mathbf{u}}(z_1, \dots, z_r) := G(\mathbf{u} + \sum_{i=1}^r z_i \boldsymbol{\xi}_i)$. Then $\mathrm{Var}(G(\mathbf{x}) \mid \mathbf{u}) = \mathrm{Var}(g_{\mathbf{u}}(z_1, \dots, z_r))$. By the Gaussian Poincaré inequality [BLM13, Theorem 3.20],

$$\mathrm{Var}(g_{\mathbf{u}}) \le \sum_{i=1}^r \mathbb{E}\left[\left(\frac{\partial g_{\mathbf{u}}}{\partial z_i} \right)^2 \middle| \mathbf{u} \right].$$

Since $\frac{\partial g_{\mathbf{u}}}{\partial z_i} = \partial_{\boldsymbol{\xi}_i} G(\mathbf{x})$, we get

$$\mathrm{Var}(G(\mathbf{x}) \mid \mathbf{u}) \le \sum_{i=1}^r \mathbb{E}[(\partial_{\boldsymbol{\xi}_i} G(\mathbf{x}))^2 \mid \mathbf{u}].$$

Taking expectation over $\mathbf{u}$ proves the claim. □

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 claims made are summarized in the abstract. Assumptions are summarized and brief comparison with prior work is included.

   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 paper clearly indicates that our work only applies in the well-studied setting of Gaussian data. The run-time bounds are clearly stated and it is clearly explained to which hypothesis classes our algorithms apply to.

   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 assumptions and new proofs are given in the paper core and the appendix. They are also properly referenced. When known propositions from the literature are used, this is stated explicitly and the relevant literature is cited.

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: The paper 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: paper does not include experiments requiring code.

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: the paper 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: We reviewed the NeurIPS Code of Ethics. We believe that our work conforms to the NeurIPS Code of Ethics.

   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: The contribution is theoretical and concerns the run-time required for learning some basic hypothesis classes. As such, we foresee no immediate positive or negative societal impact for our 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: the paper contains no data or models.

    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: the paper 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: the paper does not release 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: the paper does not involve crowdsourcing or 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: the paper does not involve crowdsourcing or 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: the core method development in this research 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.