STATISTICAL QUERY FRAMEWORK

Honam Wong
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
hnwong@seas.upenn.edu

ABSTRACT

This report presents a comprehensive analysis of the Statistical Query (SQ) learning framework, a restricted model of computation that captures the vast majority of modern learning algorithms. We systematically explore the information-theoretic limitations imposed by this model to establish unconditional lower bounds. We begin by formalizing the SQ and Correlational SQ (CSQ) models, providing a rigorous derivation of their equivalence in binary classification. We then introduce the machinery of Fourier analysis on the Boolean hypercube to derive lower bounds for parities and subsequently for Disjunctive Normal Forms (DNFs). Extending this analysis to the continuous domain, we detail the construction of orthogonal function families to prove super-polynomial lower bounds for learning one-layer neural networks under Gaussian distributions. We also discuss the nuances between different statistical oracle implementations (STAT vs. VSTAT vs. 1-STAT), and the broader implications of SQ framework.

1 INTRODUCTION

The central goal of computational learning theory is to characterize the resources: time, memory, and data required to learn concepts from examples. The standard framework for this analysis is the Probably Approximately Correct (PAC) model, which assumes the learner has access to independent and identically distributed (i.i.d.) samples from an unknown distribution. While the PAC model captures the information-theoretic barriers to learning, it does not inherently constrain the computational techniques a learner might employ.

However, a dichotomy exists in learning theory: some problems, such as learning parity functions, are tractable in the PAC model using specific algebraic techniques (e.g., Gaussian elimination) but appear intractable for the majority of robust, noise-tolerant algorithms used in practice, such as Gradient Descent. To formally study this phenomenon, Kearns (1998) introduced the Statistical Query (SQ) model.

In the SQ model, the learner is prohibited from accessing individual data points. Instead, the learner must interact with the environment via a "statistical oracle" that provides noisy estimates of population averages. This restriction is powerful enough to model almost all gradient-based, moment-matching, and local search algorithms while excluding brittle algebraic methods. Consequently, proving a lower bound in the SQ model serves as a strong impossibility result for a wide class of practical learning algorithms.

2 THE STATISTICAL QUERY MODEL

2.1 ORACLE DEFINITIONS

We consider a learning problem defined over an instance space $\mathcal{X}$ and a label space $\mathcal{Y}$. A concept class $\mathcal{C}$ is a set of functions $c : \mathcal{X} \to \mathcal{Y}$. We assume an unknown distribution $\mathcal{D}$ over $\mathcal{X}$ and an unknown target concept $c \in \mathcal{C}$.

Definition 1 (Statistical Query (STAT) Oracle)**.** *The learner has access to a STAT oracle for the distribution P over labeled examples (x, y). A query consists of a bounded function $\psi : \mathcal{X} \times \mathcal{Y} \to$*

$[-1, 1]$ *and a tolerance parameter* $\tau > 0$*. The oracle returns a value* v *such that:*

$$|v - \mathbb{E}_{(x,y)\sim P}[\psi(x, y)]| \leq \tau. \tag{1}$$

The tolerance τ is a critical parameter. Information-theoretically, simulating a query with tolerance τ requires $O(1/\tau^2)$ samples. Thus, an SQ lower bound showing that a problem requires tolerance $\tau < d^{-\omega(1)}$ implies that any statistical algorithm requires super-polynomial data.

2.2 Correlational Statistical Queries (CSQ)

A specific subclass of statistical queries, known as Correlational Statistical Queries (CSQ), is sufficient to analyze many problems, particularly those involving gradient-based updates.

Definition 2 (CSQ Oracle)**.** *A query is a correlational query if it is of the form* $\psi(x, y) = \phi(x) \cdot y$ *for some function* $\phi : \mathcal{X} \to \mathbb{R}$*. The learner requests an estimate of the correlation:*

$$\langle \phi, c \rangle_{\mathcal{D}} = \mathbb{E}_{x\sim\mathcal{D}}[\phi(x)c(x)]. \tag{2}$$

within tolerance τ*. We typically assume normalization* $\|\phi\| \leq 1$*.*

2.3 Equivalence of SQ and CSQ for Binary Classification

While the CSQ model appears restrictive, Szörényi (2009) and others have observed it is polynomially equivalent to the general SQ model for binary classification where $y \in \{\pm 1\}$. This equivalence is crucial because it allows us to focus our lower bound proofs on the simpler geometry of correlations (inner products).

Proposition 1. *Let* $\mathcal{Y} = \{\pm 1\}$*. Any statistical query* $\psi(x, y)$ *can be estimated using two queries: one that depends only on the marginal distribution of* $\mathcal{X}$ *and one correlational query.*

Proof. Let $\psi : \mathcal{X} \times \{\pm 1\} \to [-1, 1]$ be an arbitrary query. We can decompose the expectation $\mathbb{E}[\psi(x, y)]$ by conditioning on the value of y:

$$\mathbb{E}[\psi(x, y)] = \mathbb{E}\left[\mathbb{I}(y = 1)\psi(x, 1) + \mathbb{I}(y = -1)\psi(x, -1)\right] \tag{3}$$

$$= \mathbb{E}\left[\frac{1 + y}{2}\psi(x, 1) + \frac{1 - y}{2}\psi(x, -1)\right]. \tag{4}$$

Rearranging terms to group by y:

$$\mathbb{E}[\psi(x, y)] = \mathbb{E}\left[\frac{\psi(x, 1) + \psi(x, -1)}{2} + y \cdot \frac{\psi(x, 1) - \psi(x, -1)}{2}\right]. \tag{5}$$

We define two functions over the input space $\mathcal{X}$:

$$g(x) = \frac{\psi(x, 1) + \psi(x, -1)}{2}, \quad h(x) = \frac{\psi(x, 1) - \psi(x, -1)}{2}. \tag{6}$$

The expectation becomes:

$$\mathbb{E}[\psi(x, y)] = \mathbb{E}_{x\sim\mathcal{D}}[g(x)] + \mathbb{E}_{(x,y)\sim P}[h(x) \cdot y]. \tag{7}$$

The term $\mathbb{E}[g(x)]$ is a query independent of the labels. The learner can estimate this using unlabelled data or by treating it as a standard query where the label is ignored. The second term is explicitly a correlational query with query function $h(x)$. Since ψ is bounded in $[-1, 1]$, both g and h are also bounded. Thus, access to a CSQ oracle is sufficient to implement any SQ algorithm for binary classification. □

3 Boolean Function Learning and SQ Dimension

We now apply the SQ framework to Boolean functions, utilizing Fourier analysis on the hypercube to derive explicit lower bounds.

3.1 Preliminaries: Fourier Analysis on Boolean Functions

Let $f : \{\pm 1\}^d \to \mathbb{R}$ be a function on the Boolean hypercube. The set of parity functions $\{\chi_S\}_{S\subseteq[d]}$ forms an orthonormal basis for the space of such functions under the uniform distribution, where $\chi_S(x) = \prod_{i\in S} x_i$. Any function f can be uniquely expanded as:

$$f(x) = \sum_{S\subseteq[d]} \hat{f}(S)\chi_S(x). \tag{8}$$

where $\hat{f}(S) = \mathbb{E}[f(x)\chi_S(x)]$ are the Fourier coefficients. By Parseval's Theorem, $\sum_S \hat{f}(S)^2 = \mathbb{E}[f(x)^2]$.

4 Boolean Function Learning and SQ Dimension

We now apply the SQ framework to Boolean functions, utilizing Fourier analysis on the hypercube to derive explicit lower bounds.

4.1 Preliminaries: Fourier Analysis on Boolean Functions

Let $f : \{\pm 1\}^d \to \mathbb{R}$ be a function on the Boolean hypercube. The set of parity functions $\{\chi_S\}_{S\subseteq[d]}$ forms an **orthonormal basis** for the space of such functions under the uniform distribution, where $\chi_S(x) = \prod_{i\in S} x_i$. Any function f can be uniquely expanded as:

$$f(x) = \sum_{S\subseteq[d]} \hat{f}(S)\chi_S(x). \tag{9}$$

where $\hat{f}(S) = \mathbb{E}[f(x)\chi_S(x)]$ are the Fourier coefficients. By Parseval's Theorem, $\sum_S \hat{f}(S)^2 = \mathbb{E}[f(x)^2]$.

4.2 SQ Lower Bound for Learning Parity with Noise

The "Learning Parity with Noise" (LPN) problem is a canonical hard problem. The target concept is a parity function $c(x) = \chi_{S^*}(x)$ for some unknown S^*. Blum et al. (1994) provided the first lower bounds for this task in the SQ model.

Theorem 1 (SQ Lower Bound for Parity)**.** *Any SQ algorithm for learning parity functions on d bits requires $2^{\Omega(d)}$ queries.*

Proof. Consider the noiseless case ($\eta = 0$). The target is $y = \chi_{S^*}(x)$. An SQ algorithm makes a query $\phi : \{\pm 1\}^d \to [-1, 1]$ and receives an estimate of $\mathbb{E}[\phi(x)y] = \mathbb{E}[\phi(x)\chi_{S^*}(x)]$. Note that this expectation is exactly the Fourier coefficient $\hat{\phi}(S^*)$.

The algorithm does not know S^*. It chooses a query ϕ hoping to find a non-zero correlation. However, for any query function ϕ, the "energy" is bounded by Parseval's theorem:

$$\sum_{S\subseteq[d]} \hat{\phi}(S)^2 = \mathbb{E}[\phi(x)^2] \leq 1. \tag{10}$$

Role of Pairwise Orthogonality of parity functions: This inequality is the core of the proof. Because $\langle\chi_S, \chi_T\rangle = 0$ for distinct S, T, the "energy" of the query ϕ is partitioned strictly among the different coefficients. Parseval's theorem ensures that the sum of squared correlations is bounded. If parity functions are not (nearly) mutally orthogonal, query ϕ can have high correlations with multiple basis, allowing a single query to eliminate much of the hypothesis space.

Due to this orthogonality, for any tolerance τ, the number of subsets S for which $|\hat{\phi}(S)| \geq \tau$ is at most $1/\tau^2$. Let $A_\phi = \{S : |\hat{\phi}(S)| \geq \tau\}$. Then $|A_\phi| \leq 1/\tau^2$.

If the algorithm makes q queries $\phi_1, \ldots, \phi_q$, the total number of candidate sets S that have a "significant" correlation with *any* of the queries is at most q/τ^2. The total number of possible parity

functions is 2^d. If $q/\tau^2 \ll 2^d$, then for a randomly chosen S^*, it is highly probable that S^* is not in the set of "discovered" high correlations. In this scenario, the oracle can simply return 0 for every query, which is a valid response within tolerance τ (since $|\hat{\phi}(S^*)| < \tau$). Receiving 0 for all queries provides no information to distinguish S^* from any other uncorrelated parity. Thus, the algorithm fails to learn unless q or $1/\tau$ is exponential in d. □

4.3 Statistical Query Dimension

The argument for Parity relies on the fact that parity functions are pairwise orthogonal. This leads to the definition of SQ Dimension, a general structural property that characterizes hardness.

Definition 3 (SQ Dimension)**.** *Let $\mathcal{C}$ be a class of functions and $\mathcal{D}$ an input distribution. The SQ dimension of $\mathcal{C}$ with respect to $\mathcal{D}$, denoted $SQ_{\mathcal{D}}(\mathcal{C})$, is the largest integer d such that there exist d functions $f_1, \ldots, f_d \in \mathcal{C}$ satisfying:*

$$|\langle f_i, f_j \rangle_{\mathcal{D}}| \leq \frac{1}{d} \quad \textit{for all } i \neq j. \tag{11}$$

For Parity, the functions are exactly orthogonal, so the dimension is $|\mathcal{C}| = 2^d$.

Szörényi (2009) established the following "General Recipe" for lower bounds:

Theorem 2. *If a concept class $\mathcal{C}$ has SQ dimension d_{SQ}, then any SQ algorithm for learning $\mathcal{C}$ requires at least $\Omega(d_{SQ}\tau^2)$ queries or a tolerance $\tau \leq d_{SQ}^{-1/3}$.*

Proof. We consider a class $\mathcal{F}$ with SQ dimension $\geq D$. By definition, there exist $f_1, \ldots, f_D \in \mathcal{F}$ such that for all $i \neq j$:

$$|\mathbb{E}_x[f_i(x)f_j(x)]| \leq \frac{1}{D}.$$

The core idea is to show that for any single query, the number of functions in the class that are "ruled out" (i.e., have a large correlation with the query) is very small. For most i, the trivial oracle response of 0 is accurate.

Let the inner product be defined as $\langle f, g \rangle := \mathbb{E}_{x\sim q}[f(x)g(x)]$. Fix a CSQ query defined by $\mathbb{E}[y\psi(x)]$. We define the sets of indices for which the correlation is significant:

$$A^+ := \{i \in [D] : \langle f_i, \psi \rangle \geq \tau\},$$
$$A^- := \{i \in [D] : \langle f_i, \psi \rangle \leq -\tau\}.$$

Our goal is to show that $|A^+|$ and $|A^-|$ are small. To do this, we examine the "rotation function" quantity:

$$Z = \left\langle \psi, \sum_{i \in A^+} f_i \right\rangle^2.$$

We derive both an upper bound and a lower bound for Z.

1. Upper Bound via Cauchy-Schwarz: Using the Cauchy-Schwarz inequality and the assumption that the query has unit norm ($\|\psi\| \leq 1$):

$$\begin{aligned} Z &\leq \|\psi\|^2 \left\| \sum_{i \in A^+} f_i \right\|^2 \\ &\leq 1 \cdot \sum_{i,j \in A^+} \langle f_i, f_j \rangle \\ &= \sum_{i \in A^+} \|f_i\|^2 + \sum_{i \neq j \in A^+} \langle f_i, f_j \rangle. \end{aligned}$$

Assuming the functions are normalized such that $\|f_i\|^2 \leq 1$, and using the SQ dimension property that $|\langle f_i, f_j \rangle| \leq 1/D$:

$$\begin{aligned} Z &\leq \sum_{i \in A^+} 1 + \frac{1}{D}|A^+|(|A^+| - 1) \\ &\leq |A^+| + \frac{|A^+|^2}{D}. \end{aligned}$$

2. Lower Bound via Definition of A^+: By the definition of the set A^+, for every $i \in A^+$, we have $\langle f_i, \psi \rangle \geq \tau$. Therefore:

$$\left\langle \psi, \sum_{i \in A^+} f_i \right\rangle = \sum_{i \in A^+} \langle \psi, f_i \rangle \geq |A^+|\tau.$$

Squaring this gives the lower bound for Z:

$$Z \geq \tau^2 |A^+|^2.$$

3. Combining the Bounds: Combining the upper and lower bounds for Z, we have:

$$\tau^2 |A^+|^2 \leq Z \leq \frac{|A^+|^2}{D} + |A^+|.$$

Dividing through by $|A^+|$ (assuming $|A^+| > 0$) yields:

$$\tau^2 |A^+| \leq \frac{|A^+|}{D} + 1 \implies |A^+| \left(\tau^2 - \frac{1}{D} \right) \leq 1.$$

Solving for $|A^+|$:

$$|A^+| \leq \frac{D}{D\tau^2 - 1}.$$

Assuming D is large enough such that $1/D$ is negligible compared to τ^2, this simplifies to:

$$|A^+| \leq O\left(\frac{1}{\tau^2} \right).$$

By a symmetric argument, we also have $|A^-| \leq O(1/\tau^2)$. This implies that a single query can have a correlation of magnitude at least τ with at most $O(1/\tau^2)$ functions in the set. For all other functions, the correlation is within the tolerance τ of 0. Thus, an adversary can simply answer 0 for every query, and this answer will be valid for the vast majority of the function class. There are D candidate functions in total, to narrow down the hypothesis space to a single function, the algorithm therefore requires $\Omega(D\tau^2)$ queries. □

This theorem essentially states that if a class contains a large number of nearly orthogonal functions, finding the correct one is akin to finding a needle in a haystack; no query can correlate with more than a few functions at once.

4.4 Hardness of Learning DNF

We apply this framework to Disjunctive Normal Form (DNF) formulas. A DNF is a disjunction of conjunctions. While polynomial-time algorithms exist for learning DNF given membership queries (e.g., the KM algorithm), the problem is believed to be hard from random examples alone.

The hardness in the SQ model is derived from the fact that DNF formulas can approximate parity functions.

1. A parity function on k variables can be expressed as a DNF formula of size 2^{k-1}.
2. If we allow DNF formulas of size polynomial in d (say d^c), we can represent parities of size $k = c \log d$.
3. The class of polynomial-size DNFs therefore contains the class of $O(\log d)$-size parities.

Since parities are orthogonal, the SQ dimension of this subclass is the number of subsets of size $\log d$, which is $d^{O(\log d)}$. This is super-polynomial. Therefore, Blum et al. (1994) showed that learning polynomial-size DNF is hard for SQ algorithms because they cannot efficiently isolate the high-degree correlations embedded within the DNF structure.

5 HARDNESS OF LEARNING NEURAL NETWORKS

We now extend the SQ dimension arguments to the continuous domain to analyze modern deep learning architectures. Specifically, we focus on the problem of learning a one-hidden-layer neural network with Gaussian inputs.

5.1 PROBLEM SETUP AND SIGN-SYMMETRY

We consider a neural network function $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \psi\left(\sum_{i=1}^{m} a_i \phi(w_i \cdot x)\right). \tag{12}$$

where $x \sim \mathcal{N}(0, I_d)$, ϕ is the activation function (e.g., ReLU), and ψ is an outer link function. We interpret this in the probabilistic concept setting where $y \in \{\pm 1\}$ and $\mathbb{E}[y|x] = f(x)$.

To prove a lower bound, we utilize the construction from Goel et al. (2020), which relies on the property of **sign-symmetry**, which is weaker than usual symmetry.

Assumption 1. *The input distribution $\mathcal{D}$ is sign-symmetric if for any sign vector $z \in \{\pm 1\}^d$, the distribution of $x \circ z$ (element-wise product) is identical to $\mathcal{D}$. The standard Gaussian $\mathcal{N}(0, I_d)$ satisfies this.*

5.2 CONSTRUCTION OF ORTHOGONAL FAMILIES

We define a family of functions indexed by subsets $S \subset [d]$ of size k. Let $\chi(w)$ be the parity of vector $w \in \{\pm 1\}^k$.

$$f_S(x) = \psi\left(\sum_{w \in \{\pm 1\}^k} \chi(w) \phi\left(\frac{w \cdot x_S}{\sqrt{k}}\right)\right). \tag{13}$$

Here, x_S is the projection of x onto the coordinates in S. We assume ψ is an odd function. And we usually assume $k = \Theta(\log m)$ and $m = \text{poly}(d)$ such that the neural network is poly(d)-sized.

Theorem 3. *For any distinct sets S, T of size k, $\langle f_S, f_T \rangle_{\mathcal{N}(0,I)} = 0$.*

Proof. The proof exploits sign-symmetry. First, we establish how f_S transforms under sign flips. Let $z \in \{\pm 1\}^d$.

$$f_S(x \circ z) = \psi\left(\sum_{w} \chi(w) \phi\left(\frac{w \cdot (x \circ z)_S}{\sqrt{k}}\right)\right) \tag{14}$$

$$= \psi\left(\sum_{w} \chi(w) \phi\left(\frac{(w \circ z_S) \cdot x_S}{\sqrt{k}}\right)\right). \tag{15}$$

Let $w' = w \circ z_S$. Since w sums over the hypercube, w' also sums over the hypercube. Note that $\chi(w) = \chi(w' \circ z_S) = \chi(w')\chi(z_S)$.

$$f_S(x \circ z) = \psi\left(\sum_{w'} \chi(w')\chi(z_S) \phi\left(\frac{w' \cdot x_S}{\sqrt{k}}\right)\right) \tag{16}$$

$$= \psi\left(\chi(z_S)\left[\sum_{w'} \chi(w') \phi\left(\frac{w' \cdot x_S}{\sqrt{k}}\right)\right]\right). \tag{17}$$

Since ψ is an odd function ($\psi(-u) = -\psi(u)$) and $\chi(z_S) \in \{\pm 1\}$, we can pull the sign out:

$$f_S(x \circ z) = \chi(z_S) f_S(x) \tag{18}$$

Now, consider the inner product under the distribution $\mathcal{D}$:

$$\langle f_S, f_T \rangle_{\mathcal{D}} = \mathbb{E}_{x \sim \mathcal{D}}[f_S(x) f_T(x)] \tag{19}$$

$$= \mathbb{E}_{z \sim U(\{\pm 1\}^d)} \mathbb{E}_{x \sim \mathcal{D}}[f_S(x \circ z) f_T(x \circ z)] \quad \text{(By sign symmetry)} \tag{20}$$

$$= \mathbb{E}_{x \sim \mathcal{D}}\left[f_S(x) f_T(x) \mathbb{E}_z[\chi(z_S)\chi(z_T)]\right]. \tag{21}$$

The inner term $\mathbb{E}_z[\chi(z_S)\chi(z_T)] = \mathbb{E}_z[\chi(z_{S\Delta T})]$. Since $S \neq T$, the symmetric difference $S\Delta T$ is non-empty. The expectation of a non-trivial parity over uniform random signs is exactly 0. Thus, the entire inner product is 0. □

5.3 Implications for Deep Learning

The size of this orthogonal family corresponds to the number of subsets of size k, which is $\binom{d}{k} \approx d^k$. By the SQ dimension theorem, any SQ algorithm requires $d^{\Omega(k)}$ queries.

Corollary 1. *Any SQ algorithm for learning one-hidden-layer neural networks (specifically the class $\mathcal{C}_{orth}(n, k)$) requires $d^{\Omega(k)}$ queries or tolerance $d^{-\Omega(k)}$.*

By setting $k = \log d$, we obtain a super-polynomial lower bound $d^{\Omega(\log d)}$. This result applies to Gradient Descent trained from random initialization. It suggests that learning even simple networks is computationally hard without specific structural priors or initialization strategies that bypass the SQ framework.

6 Oracle Variants: STAT vs. VSTAT vs. 1-STAT

In previous sections, we have discussed using STAT oracle to establish SQ lower bound, in the literature Feldman et al. (2017), there also exists variants of oracles, for example, with variance-dependent tolerance.

The **STAT(τ)** oracle returns a value v such that $|v - \mathbb{E}[\psi]| \leq \tau$. The tolerance τ is fixed regardless of the query. This models adversarial noise where the error magnitude is independent of the query's inherent variance.

6.1 The VSTAT Oracle

The **VSTAT(t)** oracle captures the intuition that empirical averages converge faster for random variables with lower variance. For a query $\psi : \mathcal{X} \to [0, 1]$ with mean $p = \mathbb{E}[\psi]$, VSTAT(t) returns a value v satisfying:

$$|v - p| \leq \max\left(\frac{1}{t}, \sqrt{\frac{p(1-p)}{t}}\right). \tag{22}$$

The term $\sqrt{p(1-p)/t}$ corresponds to the standard deviation of an average of t Bernoulli trials.

- **Strength:** VSTAT is strictly stronger than STAT($1/\sqrt{t}$) because it provides much tighter bounds for biased queries (where $p \approx 0$ or $p \approx 1$).
- **Adversarial Nature:** Despite modeling sample capacity t, VSTAT is often defined as an adversarial oracle within the specific variance-based bounds, rather than a purely stochastic one.

Proving hardness in the VSTAT model is therefore a stronger result than in the STAT model, as it rules out algorithms that might exploit the lower noise on low-variance queries.

6.2 The 1-STAT Oracle

The **1-STAT** oracle (often referred to as the one-sample oracle) represents the most fundamental level of access, corresponding to drawing a single fresh sample from the distribution. For a query $\psi : \mathcal{X} \to [0, 1]$, 1-STAT returns a random variable v defined by:

$$v = \psi(x), \quad \text{where } x \sim \mathcal{D}. \tag{23}$$

Unlike STAT and VSTAT, which return estimates of the expectation $\mathbb{E}[\psi]$, 1-STAT returns a specific realization of the query function on a random sample.

- **Strength:** 1-STAT is the weakest oracle in the hierarchy. While it provides an unbiased estimator of the mean, it has maximal variance. To simulate an oracle call of VSTAT(t) or STAT($1/\sqrt{t}$), an algorithm must average the results of $O(t)$ independent calls to 1-STAT.

- **Stochastic Nature:** Unlike STAT and VSTAT, which are often modeled as adversarial (returning *any* value within the error tolerance), 1-STAT is purely stochastic. It directly models standard learning algorithms (like Stochastic Gradient Descent) that process data one sample at a time.

Any lower bound established for the STAT or VSTAT oracles immediately implies a sample complexity lower bound for the 1-STAT oracle (and thus for general PAC learning), often converting the query complexity q into a sample complexity of roughly $q \cdot t$.

7 SEARCH PROBLEM

While the previous sections focused on supervised learning (classifying labeled data), the Statistical Query framework provides equally powerful machinery for analyzing unsupervised learning and high-dimensional search problems. In these settings, the learner is given access to samples from a distribution D and must either estimate D (density estimation) or distinguish D from a reference noise distribution D_0 (hypothesis testing).

The hardness in these problems typically arises from the geometry of high-dimensional space: when a signal is "hidden" in a random direction or a random subset of coordinates, statistical queries—which essentially compute projections or correlations—fail to distinguish the signal from noise unless the signal is overwhelmingly strong.

7.1 STATISTICAL DIMENSION FOR DISTRIBUTION FAMILIES

To analyze these problems, we generalize the notion of Statistical Dimension from Boolean functions to probability distributions. The bounds are based on inner products between functions of the form $(D'(x) - D(x))/D(x)$ where D' and D are distributions over $\mathcal{X}$. Specifically, for any real-valued function f, the difference in expectations under D' and D can be expressed as an inner product:

$$\mathbb{E}_{x\sim D'}[f(x)] - \mathbb{E}_{x\sim D}[f(x)] = \left\langle \frac{D' - D}{D}, f \right\rangle_D . \tag{24}$$

where the inner product $\langle f, g\rangle_D := \mathbb{E}_{x\sim D}[f(x)g(x)]$.

Based on this, we define the pairwise correlation of two distributions D_1 and D_2 relative to a reference distribution D (often denoted D_0) as:

$$\chi_D(D_1, D_2) = \left| \left\langle \frac{D_1}{D} - 1, \frac{D_2}{D} - 1 \right\rangle_D \right| \tag{25}$$

When $D_1 = D_2$, this quantity is known as the $\chi^2(D_1, D)$ distance, which widely appears in hypothesis testing and information theory literature.

A key notion for our statistical dimension is the average correlation of a set of distributions $\mathcal{D}'$ relative to a distribution D. We denote it by $\rho(\mathcal{D}', D)$ and define it as follows:

$$\rho(\mathcal{D}', D) := \frac{1}{|\mathcal{D}'|^2} \sum_{D_1, D_2 \in \mathcal{D}'} \chi_D(D_1, D_2) = \frac{1}{|\mathcal{D}'|^2} \sum_{D_1, D_2 \in \mathcal{D}'} \left| \left\langle \frac{D_1}{D} - 1, \frac{D_2}{D} - 1 \right\rangle_D \right| \tag{26}$$

This formulation allows us to bound the performance of any SQ algorithm attempting to distinguish a distribution in $\mathcal{D}'$ from the reference D, as the average correlation dictates the maximum information any single query can reveal about the true distribution.

7.2 CASE STUDY: APPROXIMATE MAX-XOR-SAT

To demonstrate the versatility of our techniques beyond standard learning and estimation, we consider the **MAX-XOR-SAT** problem. This is a "warm-up" search problem where the goal is to find an assignment that satisfies the maximum number of constraints.

7.2.1 Problem Definition

Let n be the number of boolean variables. We denote the set of all possible XOR clauses on n variables by $\mathcal{C} = \{0,1\}^n$. For a clause $c \in \mathcal{C}$, if the i-th bit $c_i = 1$, the variable x_i is included in the XOR sum. Otherwise, it is not. A clause c is satisfied by an assignment $a \in \{0,1\}^n$ if the inner product modulo 2 is 1: $\langle a, c \rangle \equiv 1 \pmod 2$.

The search problem is defined as follows:

Definition 4 (ϵ-approximate MAX-XOR-SAT)**.** *Let $\mathcal{D}$ be an unknown distribution over the set of clauses $\mathcal{C}$. For any assignment a, let $f_{\mathcal{D}}(a) = \mathbb{E}_{c\sim\mathcal{D}}[a \cdot c \pmod 2]$ be the fraction of clauses satisfied by a (or more precisely, the correlation). Let $M_{\mathcal{D}} = \max_a f_{\mathcal{D}}(a)$. The goal is to find an assignment a such that $f_{\mathcal{D}}(a) \geq M_{\mathcal{D}} - \epsilon$, given SQ access to $\mathcal{D}$.*

This problem is computationally hard in the worst case (NP-hard to approximate within $1/2 - \delta$). We provide strong evidence that it is also hard in the distributional SQ setting.

7.2.2 Statistical Dimension of MAX-XOR-SAT

To prove a lower bound, we formalize the search problem as a list of distributional problems $\mathcal{Z}$. For each possible assignment $a \in \{0,1\}^n$, there exists a "hard" distribution D_a where a is the unique solution.

Theorem 4. *For any $\delta > 0$, any SQ algorithm requires at least $2^{n/3} - 1$ queries to a STAT($2^{-n/3}$) oracle to solve $(1/2 - \delta)$-approximate MAX-XOR-SAT.*

Proof. We construct a hypothesis class of distributions $\mathcal{Z} = \{D_a\}_{a\in\{0,1\}^n}$ with large statistical dimension. Let the reference distribution D be the uniform distribution over all clauses $\mathcal{C}$. For each assignment a, we define the distribution D_a such that it is supported only on clauses satisfied by a. Specifically:

$$D_a(c) = \begin{cases} 2/|\mathcal{C}| & \text{if } a \cdot c = 1 \\ 0 & \text{if } a \cdot c = 0 \end{cases} \tag{27}$$

Under D_a, the assignment a satisfies 100% of clauses, while any other assignment $b \neq a$ satisfies exactly 50% (due to the orthogonality of parity functions). Thus, finding the optimal assignment is equivalent to distinguishing D_a from the uniform distribution.

We compute the pairwise correlation between D_a and D_b relative to the uniform distribution D:

$$\left\langle \frac{D_a}{D} - 1, \frac{D_b}{D} - 1 \right\rangle_D = \begin{cases} 0 & \text{for } a \neq b \\ 1 & \text{for } a = b \end{cases} \tag{28}$$

The calculation follows from the fact that $\frac{D_a(c)}{D(c)} - 1$ is simply the parity character $(-1)^{a\cdot c}$. Since parity characters are orthogonal over the uniform distribution, the pairwise correlation for distinct a, b is exactly 0.

Since the pairwise correlation is 0 for all 2^n pairs, the statistical dimension is $2^n - 1$. Applying the general SQ lower bound recipe, we obtain the exponential lower bound on query complexity. □

8 Broader Implications: A Unifying Principle of Hardness

Beyond the specific case studies of Parity, DNF, and Neural Networks, the Statistical Query framework provides a unifying explanation for the computational hardness of a vast landscape of high-dimensional learning problems. The underlying principle across these varied domains is consistent: hardness is established by constructing a large family of **mutually nearly orthogonal functions** (in supervised learning) or **uncorrelated distributions** (in unsupervised learning), thereby maximizing the lower bound of statistical dimension.

For the **Planted Clique** problem, this orthogonality manifests in the geometry of random graph subsets. Because random cliques share very few vertices, their corresponding indicator functions (or induced distributions) have negligible pairwise correlation, forcing any statistical algorithm to search an exponentially large space of size $\binom{n}{k}$ Feldman et al. (2017). Similarly, for learning **Gaussian**

Mixture Models (GMMs), hardness is established via *moment-matching* constructions. By designing a mixture that matches the first m moments of a standard Gaussian, the distribution becomes statistically orthogonal to any low-degree polynomial query, effectively "hiding" the non-Gaussian signal in higher-order moments and enforcing a complexity of $d^{\Omega(k)}$ Diakonikolas et al. (2017).

Indeed, SQ lower bounds have been applied in modern deep learning theory research to understand the power and limitations of feature learning. A line of deep learning literature characterizes data as single-index and multi-index models, where the function depends only on a few coordinates Bruna & Hsu (2025). **Single-Index Models** are of the form $y = \phi(\langle w^*, x \rangle)$. Here, the SQ complexity is dictated by the "information exponent", the smallest degree s for which the target function's Hermite coefficient is non-zero. If the first $s - 1$ coefficients vanish, the target is orthogonal to all degree $s - 1$ polynomials, rendering standard gradient-based methods ineffective without $d^{\Theta(s)}$ samples or queries Abbe et al. (2023).

9 Conclusion

The Statistical Query framework serves as a rigorous bridge between information theory and computational complexity. By focusing on algorithms that rely on statistical aggregates, we can prove unconditional lower bounds that apply to the vast majority of techniques used in machine learning today. We have shown that the geometric structure of the concept class, specifically its orthogonality or SQ dimension, is the primary determinant of learnability. Problems with high SQ dimension, such as Parity, DNF, and Neural Networks, present exponential barriers to statistical learning, necessitating either exponentially large datasets or algorithmic techniques that bypass the SQ paradigm entirely. Finally, we have discussed its implications in searching problems, other classical problems, and problems in the frontier of deep learning theory.

References

Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *arXiv preprint arXiv:2302.11055*, 2023.

Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pp. 253–262, 1994.

Joan Bruna and Daniel Hsu. Survey on algorithms for multi-index models, 2025. URL `https://arxiv.org/abs/2504.05426`.

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

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

Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pp. 3587–3596. PMLR, 2020.

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

Balázs Szörényi. Characterizing statistical query learning: simplified notions and proofs. In *International Conference on Algorithmic Learning Theory*, pp. 186–200. Springer, 2009.