JMLR: Workshop and Conference Proceedings vol 23 (2012) 16.1–16.22 25th Annual Conference on Learning Theory

Computational Bounds on Statistical Query Learning

Vitaly Feldman VITALY@POST.HARVARD.EDU
IBM Almaden Research Center

Varun Kanade VKANADE@FAS.HARVARD.EDU
SEAS, Harvard University

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

We study the complexity of learning in Kearns' well-known *statistical query* (SQ) learning model (Kearns, 1993). A number of previous works have addressed the definition and estimation of the information-theoretic bounds on the SQ learning complexity, in other words, bounds on the query complexity. Here we give the first strictly computational upper and lower bounds on the complexity of several types of learning in the SQ model.

As it was already observed, the known characterization of distribution-specific SQ learning (Blum, et al. 1994) implies that for weak learning over a fixed distribution, the query complexity and computational complexity are essentially the same. In contrast, we show that for both distribution-specific and distribution-independent (strong) learning there exists a concept class of polynomial query complexity that is not efficiently learnable unless $\mathsf{RP} = \mathsf{NP}$. We then prove that our distribution-specific lower bound is essentially tight by showing that for every concept class C of polynomial query complexity there exists a polynomial time algorithm that given access to random points from any distribution D and an NP oracle, can SQ learn C over D.

We also consider a restriction of the SQ model, the correlational statistical query (CSQ) model (Bshouty and Feldman, 2001; Feldman, 2008) of learning which is closely-related to Valiant's model of evolvability (Valiant, 2007). We show a similar separation result for distribution-independent CSQ learning under a stronger assumption: there exists a concept class of polynomial CSQ query complexity which is not efficiently learnable unless every problem in $\mathsf{W}[\mathsf{P}]$ has a randomized fixed parameter tractable algorithm.

Keywords: statistical query learning, computational lower bounds, evolvability

1. Introduction

The statistical query learning model of Kearns (1998) is a natural restriction of the PAC learning model in which a learning algorithm is allowed to obtain estimates of statistical properties of the examples but cannot see the examples themselves. Formally, the learning algorithm is given access to STAT(f, D) – a *statistical query oracle* for the unknown target function f and distribution D over some domain X. A query to this oracle is a function of an example $\psi : X \times \{-1, 1\} \to \{-1, 1\}$. The oracle may respond to the query with any value v satisfying $|\mathbb{E}_{x \sim D}[\psi(x, f(x))] - v| \leq \tau$ where $\tau \in [0, 1]$ is the *tolerance* of the query. The learning algorithm is considered to be efficient if it runs in polynomial time, uses queries that can be evaluated in polynomial time and tolerance is lower-bounded by the inverse of a polynomial (all polynomials are in the standard learning problem parameters such as the dimension n and the inverse of the desired accuracy ϵ).

Kearns demonstrated that any learning algorithm that is based on statistical queries can be automatically converted to a learning algorithm robust to random classification noise of arbitrary rate smaller than the information-theoretic barrier of $1/2$ (Kearns, 1998). Most known learning algorithms and techniques can be converted to statistical query algorithms and hence the SQ model proved to be a powerful approach for the design of noise-tolerant learning algorithms (Kearns, 1998; Bylander, 1994; Blum et al., 1997; Dunagan and Vempala, 2004). In fact, since the introduction of the model virtually all[1] known noise-tolerant learning algorithms were obtained from SQ algorithms. The basic approach was also extended to deal with noise in several other learning scenarios (Decatur, 1993; Jackson et al., 1997; Bshouty and Feldman, 2002) and has also found applications in other areas including privacy-preserving learning and learning on multi-core systems (Bansal et al., 2002; Blum et al., 2005; Chu et al., 2006; Kasiviswanathan et al., 2008; Gupta et al., 2011).

Query complexity of SQ learning: Kearns has also demonstrated that there are information-theoretic impediments unique to SQ learning: parity functions require an exponential number of SQs to be learned Kearns (1998). Further, Blum et al. (1994) proved that the number of SQs required for weak learning (that is, one that gives a non-negligible advantage over the random guessing) of a concept class C over a fixed distribution D is characterized by a relatively simple combinatorial parameter of C called the *statistical query dimension* SQ-DIM(C, D). SQ-DIM(C, D) measures the maximum number of "nearly uncorrelated" (relative to distribution D) functions in C. These bounds for weak learning were strengthened and extended to other variants of statistical queries in several works (Bshouty and Feldman, 2002; Blum et al., 2003; Yang, 2005; Feldman, 2008). In addition to its use for learning, SQ-DIM was shown to be closely related to other measures of complexity, such as margin complexity and communication complexity (Simon, 2006; Sherstov, 2007; Feldman, 2009b; Kallweit and Simon, 2011).

More recently, Simon (2007) described an explicit[2] characterization of strong SQ learning with respect to a fixed distribution D. Simpler and stronger characterizations of strong SQ learning were subsequently derived by Feldman (2009b) and Szörényi (2009). Feldman's characterization is based on maximizing SQ-DIM of a concept class that can be obtained by subtracting a fixed function from each function of the given concept class. Szörényi characterization result is based on measuring the maximum number of functions in C whose pairwise correlations are nearly identical. It should be noted that all of these bounds characterize learning in the distribution-specific setting. Characterizing the query complexity of SQ learning in the distribution-independent setting is still an open problem.

Some of the notable applications of these characterizations are lower bounds for SQ learning of intersections-of-halfspaces by Klivans and Sherstov (2007), an upper-bound on the SQ dimension of halfspaces by Sherstov (2007) and a lower bound on strong SQ learning of depth-3 monotone formulas by Feldman et al. (2011).

Evolvability: We also address the existence of computational barriers in Valiant's recent model for evolvability (Valiant, 2009). In Valiant's model evolvability of a certain (presumably useful) functionality is cast as a problem of learning the desired functionality through a process in which, at each step, the most "fit" candidate function is chosen from a small pool of mutations of the current

1. A notable exception is the algorithm for learning parities of Blum et al. (2003) which is tolerant to random noise, albeit not in the same strong sense as the algorithms derived from SQs.
2. An earlier work has also considered this question but the characterization that was obtained is in terms of query-answering protocols that are essentially specifications of non-adaptive algorithms (Balcázar et al., 2007).

candidate. Limits on the number of steps and the amount of computation performed at each step are imposed to make this process naturally plausible. A class of functions C is considered evolvable if there exists a single representation scheme R and a mutation algorithm M on R that, when guided by such selection, guarantees convergence to the desired function for every function in C.

It has been observed by Valiant that all evolvable concept classes are also SQ learnable. Further, Feldman (2008) demonstrated that evolvability in Valiant's original[3] model is equivalent to learning by *correlational statistical queries* (CSQ). A CSQ is a restriction of SQ which allows only query functions of the form $\psi(x, \ell) \equiv g(x) \cdot \ell$, where $g(x)$ is a boolean function. In the context of SQ learning, CSQs were first defined by Bshouty and Feldman (2002), but an essentially equivalent form of queries was earlier defined by Ben-David et al. (1990) who introduced it as an instance of their Learning-By-Distances model. In both of these works it was observed that when the distribution is fixed CSQ learning is equivalent to SQ learning. On the other hand, in (Feldman, 2008, 2011) it was shown that distribution-independent CSQ learning is strictly weaker than SQ learning. This separation is based on information-theoretic lower bounds related to those known in the SQ model.

1.1. Overview of Our Results

While the query complexity of weak SQ learning is fairly well-studied, we are not aware of any prior work that deals with computational hardness specific to SQ learning. It would be natural to try to derive computational hardness of SQ learning using one of the numerous concept classes for which computational lower bounds in the more general PAC learning model are known (under a variety of cryptographic assumptions) (Valiant, 1984; Kearns and Valiant, 1994; Kharitonov, 1995). However, this does not work since for all the commonly-studied concept classes (such as polynomial-size constant-depth circuits, for example), the query complexity of the computationally hard-to-PAC-learn class is superpolynomial. In other words, for the concept classes we are aware of, the query complexity of SQ learning upper-bounds (up-to-polynomial factors) the computational complexity of PAC learning.

As it turns out, this situation is not incidental. Most of the known computational-hardness results prove hardness of weak learning over a fixed distribution. As we observe in Theorem 1, the characterization of weak distribution-specific SQ learning by Blum et al. (1994) implies that in this setting the query complexity of SQ learning and the (non-uniform) computational complexity of SQ learning are the same up to a polynomial factor.

Lower bounds: In Theorems 3 and 5 we show that, in contrast to Theorem 1, for both strong distribution-specific and distribution-independent SQ learning there exist concept classes of polynomial query complexity that are not efficiently SQ learnable unless $\mathsf{RP} = \mathsf{NP}$. As in PAC learning, weak and strong distribution-independent SQ learning are equivalent and therefore we do not treat weak distribution-independent SQ learning separately (Aslam and Decatur, 1998)). In Theorem 7 we prove an analogous separation of computational and information-theoretic complexity for distribution-independent CSQ learning, albeit under a stronger assumption: there exists a concept class of polynomial CSQ query complexity which is not efficiently learnable unless every problem in $\mathsf{W[P]}$ has a randomized fixed parameter tractable algorithm. In particular this would imply an

3. In subsequent work other loss functions were used to measure "fitness" of candidate mutations and the resulting models can be equivalent to all of SQ (Feldman, 2009a).

algorithm for the weighted circuit satisfiability problem[4] that runs in time $\mathrm{poly}(2^{\ell}, n)$. In the learning context this would imply an efficient, distribution-independent and proper learning algorithm for juntas (Alekhnovich et al., 2008). These are the first strictly computational lower bounds in all three settings. We remark that our hardness results for SQ learning are stronger than those available for (improper) PAC learning since the latter are based on cryptographic assumptions.

The basic technique for our separation is fairly standard. We create a concept class in which every concept contains an easy to learn information z and an information-theoretically hard to learn information $y = s(z)$ for some computationally hard to compute function $s(z)$. Knowing y is necessary for learning the concept class. A computationally unbounded learner can find y by computing $s(z)$. We then give a reduction from computing $s()$ to learning the concept class. The main difference between our constructions is the way we ensure that z is always learnable while completely hiding $s(z)$. It is easy to achieve this in the distribution-specific setting. We can simply use the uniform distribution and then split the domain in two parts: the first one used for encoding z and the second one for encoding $s(z)$ in an information-theoretically secure way (specifically, using a parity function defined by $s(z)$). Such straightforward approach does not work in a distribution-independent setting since the distribution can give weight 0 to the points encoding z. Our simple solution is based on restricting the encoding of $s(z)$ only to points of the domain that contain z as their prefix. This ensures that, relative to the unknown distribution, the target function is either close-to-constant or can be used to recover z. A somewhat similar technique was used by Servedio (2000) is his computational hardness results for attribute-efficient learning.

Designing the concept class and establishing the desired separation for distribution-independent CSQ learning (and thereby evolvability) is substantially more challenging, primarily since both learning and proving lower bounds in this model are technically more involved. Our computationally-unbounded learner is based on a strengthening of singleton (a class containing all functions that are positive on exactly one point of the domain) learning algorithm of Feldman (2009a) that actually recovers the point where the singleton is positive. Our computational lower bound is based on the recent lower bound for distribution-independent learning of conjunctions (Feldman, 2011). The lower bound is weaker than the lower bound for parities used for our SQ separations and therefore we obtain a hardness result based on fixed-parameter intractability of $\mathsf{W}[\mathsf{P}]$. Further our proof requires a special error correcting code for recovering a small set from a partial erasure of its elements. We design a simple code for this purpose using a Reed-Solomon code.

Upper bounds: In Theorem 9 we describe our learning result using an NP oracle. For any concept class C which has polynomial query complexity and $C \in \mathsf{P}$, we give an algorithm that, given access to random (unlabeled) points from any distribution D, SQ oracle and NP oracle our algorithm learns C. Here by $C \in \mathsf{P}$ we mean that the set of all binary strings representing concepts in C is a language in P. This is a natural assumption and is easy to verify for all the commonly-studied concept classes and representations. However, we note that it is not satisfied by the (artificial) examples we use for our lower bounds. As we show in Remarks 10 and 11 the examples are still learnable using an NP oracle and also that there exist concept classes in P with purely computational lower bounds under a stronger cryptographic assumption.

A variant of the SQ model which includes access to random points from the underlying distribution was also introduced and studied by Kearns (1998). Our lower bounds and the known characterizations of SQ complexity all apply to this variant. Hence for this variant of the model, our algo-

4. Given a boolean circuit, does it have a satisfying assignment of Hamming weight ℓ?

rithm provides an essentially matching upper bound for both distribution-specific and distribution-independent settings. For the more restricted setting (that is, without access to random points from D), we can still obtain this upper bound when the input distribution is fixed by using a single polynomial-size random sample from D as non-uniform advice to our algorithm.

One might again be tempted to apply the intuition from the PAC model where such an upper bound is well-known and trivial. However, the result is completely incomparable to an upper bound for PAC and relies on a recent result by Szörényi (2009). Szörényi gives an ingenious argument proving that the query complexity of every SQ learner for C that uses only queries from C which are consistent with all the previous oracle's answers is polynomially related to the query complexity of learning C. For comparison, we note that the best known algorithms for learning some of the concept classes in Angluin's exact learning model (Angluin, 1988) use a substantially more powerful Σ_3^p oracle (Bshouty et al., 1996) (assuming, of course, that the polynomial hierarchy does not collapse).

Organization: We describe notation and some preliminaries in Section 2. Section 3 describes in greater detail the various statistical query learning models we consider. Section 4 contains the lower bounds for distribution-specific SQ learning, distribution-independent SQ learning and distribution-independent CSQ learning. Section 5 gives the upper bounds on SQ learning.

2. Notation and Preliminaries

2.1. Notation

In this paper, binary strings are strings over $\{-1,1\}$ and boolean functions have range $\{-1,1\}$. Let $[k]$ denote the set $\{1,2,\ldots,k\}$ and for any k bit string z, let $S(z) = \{i \mid z_i = -1\} \subseteq [k]$. Define $\mathsf{MAJ}_z : \{-1,1\}^k \to \{-1,1\}$ to be the majority function over $S(z)$. Formally for $x \in \{-1,1\}^k$, $\mathsf{MAJ}_z(x) = -1$ if $|S(x) \cap S(z)|/|S(z)| \geq 1/2$ and $\mathsf{MAJ}_z(x) = 1$ otherwise. Define $\mathsf{PAR}_z : \{-1,1\}^k \to \{-1,1\}$ as the parity function over the bits in $S(z)$. Formally, for $x \in \{-1,1\}^k$, $\mathsf{PAR}_z(x) = -1$ if $|S(x) \cap S(z)|$ is odd and $\mathsf{PAR}_z(x) = 1$ if $|S(x) \cap S(z)|$ is even, i.e. $\mathsf{PAR}_z(x) = \prod_{i \in S(z)} x_i$. Define $\mathsf{OR}_z : \{-1,1\}^k \to \{-1,1\}$ as the OR function over the bits in $S(z)$. Formally, for $x \in \{-1,1\}^k$, $\mathsf{OR}_z(x) = -1$ if $|S(x) \cap S(z)| \geq 1$ and $\mathsf{OR}_z(x) = 1$ otherwise.

2.2. Fourier Analysis

Under the uniform distribution over $\{-1,1\}^n$, the set of parity functions, $\langle \mathsf{PAR}_z \rangle_{z \in \{-1,1\}^n}$ forms an orthonormal basis (Fourier basis) for real-valued functions defined over $\{-1,1\}^n$. For a function $f : \{-1,1\}^n \to \mathbb{R}$, $\hat{f}(S(z)) = \mathbb{E}_{x \sim U_n}[f(x)\mathsf{PAR}_z(x)]$ is the Fourier coefficient of f corresponding to the subset $S(z)$; here U_n denotes the uniform distribution over $\{-1,1\}^n$. Fourier analysis has been used extensively for learning with respect to the uniform distribution (for a survey see (Mansour, 1994)). In particular, Kushilevitz and Mansour (1993) showed that with blackbox access to a function f, all Fourier coefficients of f with magnitude at least θ, can be obtained in time $\mathrm{poly}(n, 1/\theta)$. We refer to this as the KM algorithm and use it frequently in our proofs.

3. Statistical Query Learning Models

In this section, we describe the various SQ models considered in this paper. Let X be the instance space and suppose that n characterizes the representation size of each element $x \in X$ (e. g. $X =$

$\{-1,1\}^n$ or $X = \mathbb{R}^n$). A concept class C over X is a subset of boolean functions defined over X, where each concept $c \in C$ is represented as a binary string; it is required that there exist an efficient Turing machine that outputs $c(x)$, given $c \in C$ and $x \in X$ as inputs (cf. Kearns and Vazirani (1994)). Let D be a distribution over X. In the SQ model (Kearns, 1998), the learning algorithm has access to a statistical query oracle, $\mathsf{STAT}(f, D)$, to which it can make a query of the form (ψ, τ), where $\psi : X \times \{-1,1\} \to [-1,1]$ is the query function and τ is the (inverse polynomial) tolerance. The oracle responds with a value v such that $|\mathbb{E}_D[\psi(x, f(x))] - v| \leq \tau$, where $f \in C$ is the target concept. The goal of the learning algorithm is to output a hypothesis, $h : X \to \{-1,1\}$, such that $\mathrm{err}_D(h, c) = \Pr_{x\sim D}[h(x) \neq c(x)] \leq \epsilon$.

In this paper, we will use the following characterization of the SQ model due to Bshouty and Feldman (2002) (see also Feldman, 2008): A statistical query $\psi : X \times \{-1,1\} \to [-1,1]$ is said to be *target-independent* if $\psi(x, b) \equiv \psi^{\mathrm{ti}}(x)$ for some function $\psi^{\mathrm{ti}} : X \to [-1,1]$. A statistical query is said to be *correlational* if $\psi(x, y) \equiv y\psi^{\mathrm{cor}}(x)$ for some function $\psi^{\mathrm{cor}} : X \to [-1,1]$. Bshouty and Feldman (2002) showed that any statistical query (ψ, τ) (in Kearns' model) can be replaced by two queries, one of which is target-independent and the other correlational, each with tolerance $\tau/2$. We denote an oracle that accepts only target-independent or correlational queries as $\mathsf{SQ}\text{-}\mathcal{O}(f, D)$.

Distribution-Specific SQ Learning: We first define the notion of distribution-specific SQ learning. It is required that the running time of the algorithm is polynomial in the parameters n and $1/\epsilon$ and also that the queries made by the algorithm are efficiently evaluatable[5] and use a tolerance parameter τ, that is lower-bounded by some inverse polynomial in n and $1/\epsilon$.

Definition 1 (Distribution-Specific SQ Learning) *Let X be the instance space (with representation size n), D a distribution over X and C a concept class over X. We say that C is* distribution-specific *SQ learnable with respect to distribution D, if there exists a randomized algorithm $\mathcal{A}$ that for every $\epsilon, \delta > 0$, every target concept $f \in C$, with access to oracle $\mathsf{SQ}\text{-}\mathcal{O}(f, D)$, outputs with probability at least $1-\delta$, a hypothesis h such that $\mathrm{err}_D(h, f) \leq \epsilon$. Furthermore, the running time of the algorithm must be polynomial in n and $1/\epsilon$ and $1/\delta$ and the queries made to the oracle and the output hypothesis must be polynomially evaluatable and have a tolerance τ that is lower-bounded by an inverse polynomial in n, $1/\epsilon$.*

To distinguish the computational complexity and information-theoretic (or statistical) complexity of SQ learning, we use the notion of *query complexity*. The query complexity for learning concept class C is the minimum number of queries required by any (possibly unbounded) algorithm to learn C in the SQ model. It is worthwhile to point that when defining query complexity, it is not required that the queries be efficiently evaluatable and need not even have a small representation.

Definition 2 (Distribution-Specific SQ Query Complexity) *Let X be the instance space (with representation size n), D a distribution over X and C a concept class over X. We say that the* query complexity *of learning C under distribution D to accuracy ϵ, is bounded by q if there exists a (possibly computationally unbounded) algorithm that for every concept $f \in C$, makes at most q queries to oracle $\mathsf{SQ}\text{-}\mathcal{O}(f, D)$ outputs a hypothesis h, such that $\mathrm{err}_D(h, f) \leq \epsilon$. The tolerance τ for the queries must be lower-bounded by an inverse polynomial in n and $1/\epsilon$.*

5. By this we mean that given a description of a query function ψ and it's input x, there exists an efficient Turing machine that outputs $\psi(x)$.

Distribution-Specific Weak SQ Learning: In the case of weak learning, the learning algorithm is required only to output a hypothesis whose error is at most $1/2 - \gamma$, where $1/\gamma$ is bounded by a polynomial in n. The definitions of distribution-specific weak SQ learning and distribution-specific weak query complexity are identical to the definitions above (except for the requirement on the error of the output hypothesis).

Distribution-Independent SQ Learning: In the case of distribution-independent SQ learning, the same learning algorithm is required to output an accurate hypothesis for all distributions. The definition of distribution-independent SQ learning and distribution-independent query complexity can be made as in the case of distribution-specific learning with this additional requirement. In the case of distribution-independent SQ learning, weak and strong learning are equivalent (cf. Aslam and Decatur (1998)), hence we do not consider the distribution-independent weak learning model. Formal definitions are provided in Appendix A

Correlational Statistical Query Learning: The correlational statistical query (CSQ) learning model was introduced by Feldman (2008) and he showed that this model is equivalent to Valiant's evolution model. In the CSQ model, the learner is only allowed to make statistical queries that are *correlational*. Let (ψ, τ) be a query, where $\psi : X \to [-1, 1]$ is the query function and τ is the (inverse polynomial) tolerance factor. A CSQ oracle, $\mathsf{CSQ}\text{-}\mathcal{O}(f, D)$, responds with a value v such that $|E_{x\sim D}[\psi(x)f(x)] - v| \leq \tau$, where $f \in C$ is the target concept.

Distribution-specific CSQ learning and distribution-specific SQ learning are essentially equivalent, as long as the learning algorithm has a random sample from the distribution as non-uniform advice[6]. Thus, we do not consider the case of distribution-specific CSQ learning separately. Also, in the case of CSQ learning, it does not make sense to consider models which have access to random unlabeled examples, since this would make it the same as SQ learning. The notion of distribution-independent CSQ learning and query complexity are same as in the SQ case, except that the learning algorithm only has access to a $\mathsf{CSQ}\text{-}\mathcal{O}(f, D)$ oracle. Formal definitions are provided in Appendix A

4. Computational Lower Bounds

In this section, we first state for completeness the result that for weak distribution-specific SQ/CSQ learning, the query complexity and computational complexity is essentially the same. This fact was observed by Blum et al. (1994) and follows easily from their characterization of SQ learning. Next, we show that for distribution-specific (strong) SQ learning, distribution-independent SQ learning and distribution-independent (strong) CSQ learning, the query complexity is significantly different from the computational complexity. In the case of distribution-specific (strong) SQ learning and distribution-independent SQ learning, we show that there exists a concept class that has polynomial query complexity, but cannot be efficiently learned in the respective SQ model unless $\mathsf{RP} = \mathsf{NP}$. In the case of distribution-independent CSQ learning, the separation is based on a stronger assumption: we show that there is a concept class C with polynomial query complexity (of CSQ learning), but cannot be learned efficiently unless every problem in $\mathsf{W[P]}$ has a randomized fixed parameter tractable algorithm.

6. See (Feldman, 2008) for more details.

4.1. Weak Distribution-Specific SQ/CSQ Learning

Blum et al. (1994) showed that weak distribution-specific SQ learnability of a concept class is characterized by a combinatorial parameter SQ-DIM(C, D), which characterizes the number of nearly uncorrelated concepts in C and observed that this implies the equivalence of query complexity and computational complexity in this model of learning. A short proof sketch of the following theorem is provided in Appendix B for completeness.

Theorem 1 *If a concept class C is weakly SQ learnable over a distribution D then there exists a polynomial-size circuit that weakly SQ learns C over a distribution D.*

4.2. Strong Distribution-Specific SQ/CSQ Learning

Let $\phi \in \{-1,1\}^m$ denote a 3-CNF formula over n variables (encoded as a string). Suppose ϕ is satisfiable and let $\zeta(\phi)$ denote the lexicographically first satisfying assignment of ϕ. Throughout this section $b \in \{-1,1\}$, $x \in \{-1,1\}^m$ and $x' \in \{-1,1\}^n$ and let bxx' denote the $m+n+1$ bit string obtained by concatenating b, x and x'. For $\phi \in \{-1,1\}^m$, where ϕ is a satisfiable 3-CNF formula, define the function $f_{\phi,y} : \{-1,1\}^{m+n+1} \to \{-1,1\}$ as follows:

$$f_{\phi,y}(bxx') = \begin{cases} \mathsf{MAJ}_\phi(x) & \text{if } b = 1 \\ \mathsf{PAR}_y(x') & \text{if } b = -1 \end{cases}$$

In other words, $f_{\phi,y}$ is a function that over one half of the domain is the majority function, MAJ_ϕ, and over the other half of the domain is the parity function, PAR_y. Note that the function $f_{\phi,y}$ is efficiently computable given the representation (ϕ, y). Define C_1 to be the following concept class:

$$C_1 = \{f_{\phi,\zeta(\phi)} \mid \phi \text{ is satisfiable}\}.$$

Theorems 2 and 3 show that the query complexity of C_1 is polynomial, but unless $\mathsf{RP} = \mathsf{NP}$ there is no polynomial time SQ algorithm for learning C_1. The proofs are provided in Appendix B. To prove that the query complexity is polynomial, the key idea is that the learning algorithm only needs to (proper) learn majorities in the SQ model, which is easy. The learning algorithm can recover ϕ and solve for $\zeta(\phi)$ (possibly using unbounded computation). Thus, $f_{\phi,\zeta(\phi)}$ can be exactly SQ learned using only polynomially many queries. On the other hand, we show that an efficient SQ learning algorithm for C_1 can be used to recover a satisfying assignment of the 3-CNF formula ϕ. The key point to note here is that parities are essentially invisible to statistical queries and hence the only way to learn C_1 is to obtain $\zeta(\phi)$ using ϕ, which is not possible unless $\mathsf{RP} = \mathsf{NP}$.

Theorem 2 *The query complexity of SQ learning C_1 with respect to the uniform distribution U is at most m.*

Theorem 3 *C_1 is not* efficiently *SQ learnable under the uniform distribution unless* $\mathsf{RP} = \mathsf{NP}$.

4.3. Strong Distribution-Independent SQ Learning

In this section, we consider the distribution-independent SQ learning model. As in the case of distribution-specific SQ/CSQ learning, we construct a concept class, C_2, such that C_2 is distribution-independently SQ learnable, but not *efficiently* distribution-independently SQ-learnable unless $\mathsf{RP} = \mathsf{NP}$.

Using the notation from Section 4.2 define $g_{\phi,y}$ as:

$$g_{\phi,y}(xx') = \begin{cases} PAR_y(x') & x = \phi \\ 1 & \text{otherwise} \end{cases}$$

Thus, $g_{\phi,y}$ is the function that equals $\mathsf{PAR}_y(x')$ on the part of the domain that has ϕ as the prefix and is the constant function 1 otherwise. Define the concept class C_2 as follows:

$$C_2 = \{g_{\phi,\zeta(\phi)} \mid \phi \text{ is satisfiable}\}.$$

First, we show that the distribution-independent query complexity of SQ learning C_2 is bounded by a polynomial in n. The key idea is that either the constant function 1 is an accurate predictor (if the distribution has almost no mass on points that have ϕ as a prefix), or else it is possible to recover the 3-CNF formula ϕ using statistical queries, and then (using possibly unbounded computation) the assignment $\zeta(\phi)$ can be obtained to learn $g_{\phi,\zeta(\phi)}$ exactly. On the contrary, we show that C_2 cannot be efficiently learned in the distribution-independent SQ model unless $\mathsf{RP} = \mathsf{NP}$. As in the previous case, we show that an efficient SQ algorithm for learning C_2 can be used to find a satisfying assignment to any 3-CNF formula ϕ, if it exists. The proofs of Theorems 4 and 5 are provided in Appendix B.

Theorem 4 *The distribution-independent query complexity of SQ learning C_2 is at most $2m + 1$.*

Theorem 5 *C_2 is not* efficiently *distribution-independently SQ learnable unless* $\mathsf{RP} = \mathsf{NP}$.

4.4. Strong Distribution-Independent CSQ Learning

Showing a separation between the computational complexity and query complexity of distribution-independent CSQ learning is significantly more involved. The separation in this case is based on a stronger assumption: $\mathsf{W[P]}$ does not have randomized fixed parameter tractable algorithms. A *fixed parameter tractable algorithm* for a decision problem (x, k) is allowed to take running time $f(k)p(|x|)$ where p is a polynomial and f is an arbitrary function. A complete problem for $\mathsf{W[P]}$ is weighted circuit satisfiabiliy, i.e. given a circuit ϕ and parameter k, does there exist a satisfying assignment of Hamming weight k? It is widely believed that $\mathsf{W[P]}$ does not have randomized fixed-parameter tractable algorithms and such an algorithm would also imply a subexponential time algorithm for circuit satisfiability (cf. Downey and Fellows (1995)).

The construction relies on Feldman's recent result (Feldman, 2011), where he shows that the class of disjunctions cannot be learned (for information theoretic reasons) in the distribution-independent CSQ model. The class of disjunctions on the other hand is weakly learnable in the distribution-independent CSQ model (Feldman, 2008). Unlike in the case of distribution-independent SQ model, this fact is required[7] because any algorithm that only uses *correlational* statistical queries can only get information about the distribution by first finding some function that is (at least weakly) correlated with the target function under that distribution.

Let $\phi \in \{-1, 1\}^m$ denote a circuit (represented as a string) with n input variables. For some parameter ℓ, let $\zeta(\phi)$ denote the lexicographicaly first satisfying assignment of Hamming weight ℓ. Let $n' = 3\ell n$ and let $\mathsf{Enc} : \{-1, 1\}^n \to \{-1, 1\}^{n'}$ be an encoding such that for any string

7. Note that the class of parities is not weakly learnable in the SQ model.

$s \in \{-1,1\}^n$ with ℓ "-1" bits, $\mathsf{Enc}(s) \in \{-1,1\}^{n'}$ has 3ℓ "-1" bits. Furthermore, recovering any ℓ of these 3ℓ "-1" bits of $\mathsf{Enc}(s)$ allows us to reconstruct s. Such encodings can be constructed using Reed-Solomon codes and are defined in Appendix C. We will explain shortly the necessity for these codes for our construction. Let $\xi(\phi) = \mathsf{Enc}(\zeta(\phi))$. Let $y \in \{-1,1\}^{n'}$ (recall that $n' = 3\ell n$), let $x \in \{-1,1\}^m$, $x' \in \{-1,1\}^{n'}$ and define $c_{\phi,y} : \{-1,1\}^{m+n'} \to \{-1,1\}$ as follows:

$$c_{\phi,y}(xx') = \begin{cases} \mathsf{OR}_y(x') & \text{if } x = \phi \\ 1 & \text{otherwise} \end{cases}$$

Define the concept class.

$$C_3 = \{c_{\phi,\xi(\phi)} \mid \phi \text{ has a satisfying assignment of Hamming weight at most } \ell\}.$$

Theorem 6 shows that the query complexity of CSQ learning C_3 is polynomial. This can be proved using the fact that OR is weakly learnable and by modifying Feldman's singleton learning algorithm (Feldman, 2009a). This enables us to recover ϕ and the lexicographically first satisfying assignment of ϕ can be easily constructed (using unbounded computation). On the other hand, Theorem 7 shows that an efficient CSQ algorithm for learning C_3, implies a $\mathrm{poly}(2^\ell, n)$ time for the weighted-circuit-SAT problem (given (ϕ, ℓ), does there exist a satisfying assignment for ϕ of Hamming weight ℓ?). The reduction requires us to set the accuracy of the learning algorithm to $O(2^{-\ell})$ and also allows us to only recover one-third of the bits of the hidden OR. For this reason we need to use an OR that uses $\xi(\phi) = \mathsf{Enc}(\zeta(\phi))$ rather than $\zeta(\phi)$. Recovering a third of the bits of $\xi(\phi)$ is enough to reconstruct $\zeta(\phi)$.

Theorem 6 *The distribution-independent CSQ query complexity of C_3 is at most* $\mathrm{poly}(n, 1/\epsilon)$

Theorem 7 *C_3 is not* efficiently *distribution-independently CSQ learnable, unless there exists a randomized algorithm that determines whether or not a given circuit ϕ, has a satisfying assignment of Hamming weight at most ℓ in time* $\mathrm{poly}(2^\ell, n)$.

5. Computational Upper Bounds

In this section, we consider the following question: how much computational power is sufficient for learning in these models, given that the query complexity is polynomial?

In the setting where the learning algorithm has access to i.i.d. unlabeled examples from the underlying distribution, we show that an NP-oracle suffices for learning. We show that if the query complexity for a class C is polynomial, then there exists a polynomial-time algorithm that with access to random unlabeled examples from the distribution and with access to an NP-oracle learns C. We use Szörényi's characterization of SQ learning, where he shows that any algorithm that makes consistent queries from the class C, learns C. We require an additional natural condition, $C \in \mathsf{P}$, i.e. given c as a bit string, there is a polynomial time algorithm that determines whether or not c is a valid representation of a concept in C.

Definition 3 (Consistent Learner) *(Szörényi, 2009) Let $\langle(\phi_i, \tau_i)\rangle_{i \geq i}$ be the queries made by an SQ learning algorithm $\mathcal{A}$ and let $\langle v_i \rangle_{i \geq 1}$ be the reponses of the SQ oracle. Algorithm $\mathcal{A}$ is said to be* consistent *if for every $j < i$, $|\mathbb{E}_D[\phi_j(x)\phi_i(x)] - v_j| \leq \tau_j$.*

Szörényi (2009) proved the following result.

Theorem 8 *(Szörényi, 2009) Let q be the query complexity of SQ learning concept class C with respect to distribution D, then there exists τ, such that $1/\tau$ is bounded by $\mathrm{poly}(q, n, 1/\epsilon)$ and any* consistent *algorithm that makes queries of the form (c, τ), where $c \in C$, eventually makes a query of the form (c', τ), where $\mathrm{err}(c') \leq \epsilon/2$. The total number of queries made by the algorithm is at most $\mathrm{poly}(q, n, 1/\epsilon)$.*

As a corollary of this result, we can show that an NP-oracle suffices for statistical query learning, when the learning algorithm also has access to unlabeled examples from the underlying distribution. The key idea is that it is possible to find queries from C that are consistent with the previous query responses by using a large enough sample and with access to an NP-oracle. The following theorem follows easily from Theorem 8.

Theorem 9 *Let $q(C, D, \epsilon)$ be the query complexity of SQ learning concept class $C \in \mathsf{P}$ with respect to distribution D to accuracy ϵ. Then there exists an algorithm that for every target function $f \in C$, for every distribution D, with access to random examples from distribution D, oracle $\mathsf{CSQ}\text{-}\mathcal{O}(f, D)$[8] and an NP-oracle, outputs $c' \in C$, such that $\mathrm{err}_D(c', f) \leq \epsilon$. The running time of the algorithm is $\mathrm{poly}(q(C, D, \epsilon), n, 1/\epsilon)$.*

Proof Let C be a concept class and assume that every $c \in C$ has a representation that uses at most $s(n, 1/\epsilon)$ bits for some polynomial s. Let $m = (16s(n, \epsilon^{-1})/\tau(n, \epsilon^{-1}))^2 \log(1/\delta)$. Then a random sample of size m from distribution D satisfies the following,

$$\forall c_1, c_2 \in C, |\mathbb{E}_D[c_1(x)c_2(x)] - \frac{1}{m}\sum_{k=1}^{m} c_1(x_k)c_2(x_k)| \leq \tau/4$$

Now, consider the following algorithm. First make any query $(c_1, \tau/4)$ and receive response v_1. Note that v_1 is also a valid query response for the query (c_1, τ). Given queries $(c_1, \tau/4), \ldots, (c_{i-1}, \tau/4)$ with responses $v_1, \ldots, v_{i-1}$. Find c_i such that for every $j < i$ it holds simultaneously that,

$$\left| \frac{1}{m}\sum_{k=1}^{m} c_i(x_k)c_j(x_k) - v_j \right| \leq \tau/2 \tag{1}$$

Now it is easy to see that such a c_i exists because the true target concept f satisfies this. It is also easy to see that such a c_i can be identified easily using an NP-oracle, since c_i has a polynomial-size representation (thus obtaining c_i one bit at a time), and so the fact that $c_i \in C$ and the relations (1) can be verified easily in polynomial time.

An algorithm that makes queries $(c_1, \tau), (c_2, \tau), \ldots$ and receives responses $v_1, v_2, \ldots$ is *consistent*. Hence there will be some $t = \mathrm{poly}(q(C, D, \epsilon)n, 1/\epsilon)$ such that $\mathrm{err}_D(c_t, f) \leq \epsilon/2$. ■

Remark 10 *We note that the concept classes C_1, C_2, and C_3 defined respectively in Sections 4.2, 4.3, and 4.4 are actually not recognized by a polynomial time Turing machine. This is because given*

8. Since we are assuming that our algorithm has access to i.i.d. random examples from the distribution an oracle that only responds to correlational statistical queries is sufficient.

a string of the form $(\phi, \zeta(\phi))$, it is not possible to verify that $\zeta(\phi)$ is indeed the lexicographically first satisfying assignment to ϕ unless $\mathsf{P} = \mathsf{NP}$. We note however that even then these classes can be learned with access to an NP-oracle because $C_1, C_2, C_3 \in \mathsf{P}^{\mathsf{NP}}$, i.e. with access to an NP-oracle, the lexicographically first satisfying assignments can be constructed (one bit at a time).

Remark 11 *Under stronger cryptographic assumptions, we can construct classes $C_1', C_2', C_3' \in \mathsf{P}$ that are also not efficiently learnable in the respective statistical query models. The functions constructed can be of the form $(s(z), z)$, where $s(z)$ is easy to find information and s is a one-way permutation (that is cannot be inverted efficiently). An additional implication of such constructions is average-case computational hardness: learning is hard for most functions in $C_1'/C_2'/C_3'$.*

References

M. Alekhnovich, M. Braverman, V. Feldman, A. Klivans, and T. Pitassi. The complexity of properly learning simple classes. *Journal of Computer and System Sciences*, 74(1):16–34, 2008.

D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

J. Aslam and S. Decatur. General bounds on statistical query learning and pac learning with noise via hypothesis boosting. *Information and Computation*, 141(2):85–118, 1998.

J. Balcázar, J. Castro, D. Guijarro, J. Köbler, and W. Lindner. A general dimension for query learning. *Journal of Computer and System Sciences*, 73(6):924–940, 2007.

N. Bansal, A. Blum, and S. Chawla. Correlation clustering. In *Proceedings of FOCS*, pages 238–247, 2002.

S. Ben-David, A. Itai, and E. Kushilevitz. Learning by distances. In *Proceedings of COLT*, pages 232–245, 1990.

A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of STOC*, pages 253–262, 1994.

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

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

A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *Proceedings of PODS*, pages 128–138, 2005.

N. Bshouty and V. Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002. ISSN 1533-7928.

N. Bshouty, R. Cleve, R. Gavaldà, S. Kannan, and C. Tamon. Oracles and queries that are sufficient for exact learning. *J. Comput. Syst. Sci.*, 52(3):421–433, 1996.

T. Bylander. Learning linear threshold functions in the presence of classification noise. In *Proceedings of COLT*, pages 340–347, 1994.

C. Chu, S. Kim, Y. Lin, Y. Yu, G. Bradski, A. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *Proceedings of NIPS*, pages 281–288, 2006.

S. Decatur. Statistical queries and faulty pac oracles. In *Proceedings of the Sixth Workshop on Computational Learning Theory*, pages 262–268, 1993.

Rod G. Downey and Michael R. Fellows. Fixed-parameter tractability and completeness i: Basic results. *SIAM J. Comput.*, 24:873–921, August 1995. ISSN 0097-5397.

J. Dunagan and S. Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *Proceedings of STOC*, pages 315–320, 2004.

V. Feldman. Evolvability from learning algorithms. In *Proceedings of STOC*, pages 619–628, 2008.

V. Feldman. Robustness of evolvability. In *Proceedings of COLT*, pages 277–292, 2009a.

V. Feldman. A complete characterization of statistical query learning with applications to evolvability. In *Proceedings of FOCS*, pages 375–384, 2009b.

V. Feldman. Distribution-independent evolvability of linear threshold functions. *Journal of Machine Learning Research - COLT Proceedings*, 19:253–272, 2011.

V. Feldman, H. Lee, and R. Servedio. Lower bounds and hardness amplification for learning shallow monotone formulas. *Journal of Machine Learning Research - COLT Proceedings*, 19:273–292, 2011.

A. Gupta, M. Hardt, A. Roth, and J. Ullman. Privately releasing conjunctions and the statistical query barrier. In *STOC*, pages 803–812, 2011.

J. Jackson, E. Shamir, and C. Shwartzman. Learning with queries corrupted by classification noise. In *Proceedings of the Fifth Israel Symposium on the Theory of Computing Systems*, pages 45–53, 1997.

M. Kallweit and H. Simon. A close look to margin complexity and related parameters. *Journal of Machine Learning Research - COLT Proceedings*, 19:437–456, 2011.

S. Kasiviswanathan, H. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? In *Proceedings of FOCS*, pages 531–540, 2008.

M. Kearns and L. Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM*, 41(1):67–95, 1994.

M. Kearns and U. Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, MA, 1994.

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

M. Kharitonov. Cryptographic lower bounds for learnability of boolean functions on the uniform distribution. *Journal of Computer and System Sciences*, 50:600–610, 1995.

A. Klivans and A. Sherstov. Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning*, 69(2-3):97–114, 2007.

E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, 1993.

Y. Mansour. Learning boolean functions via the fourier transform. In V. P. Roychodhury, K. Y. Siu, and A. Orlitsky, editors, *Theoretical Advances in Neural Computation and Learning*, pages 391–424. Kluwer, 1994.

R. O'Donnell. *Computational Applications of Noise Sensitivity*. PhD thesis, MIT, 2003. URL citeseer.ist.psu.edu/630676.html.

R. Servedio. Computational sample complexity and attribute-efficient learning. *Journal of Computer and System Sciences*, 60(1):161–178, 2000.

A. A. Sherstov. Halfspace matrices. In *Proceedings of Conference on Computational Complexity*, pages 83–95, 2007.

H. Simon. Spectral norm in learning theory: Some selected topics. In *Proceedings of Algorithmic Learning Theory*, pages 13–27, 2006.

H. Simon. A characterization of strong learnability in the statistical query model. In *Proceedings of Symposium on Theoretical Aspects of Computer Science*, pages 393–404, 2007.

B. Szörényi. Characterizing statistical query learning:simplified notions and proofs. In *Proceedings of ALT*, pages 186–200, 2009.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

L. G. Valiant. Evolvability. *Journal of the ACM*, 56(1):3.1–3.21, 2009. Earlier version in ECCC, 2006.

Ke Yang. New lower bounds for statistical query learning. *Journal of Computer and System Sciences*, 70(4):485–509, 2005.

Appendix A. Models

We give formal definitions omitted in Section 3.

Distribution-Independent SQ Learning:

Definition 4 (Distribution-Independent SQ Learning) *Let X be the instance space (with representation size n) and C a concept class over X. We say that C is* distribution-independently *SQ learnable if there exists a randomized algorithm $\mathcal{A}$ that for every $\epsilon, \delta > 0$, every target concept $f \in C$ and for every distribution D over X, with access to oracle $\mathsf{SQ}\text{-}\mathcal{O}(f, D)$, outputs with probability at least $1 - \delta$, a hypothesis h such that $\mathrm{err}_D(h, f) \leq \epsilon$. Furthermore, the running time of the algorithm must be polynomial in n and $1/\epsilon$ and $1/\delta$ and the queries made to the oracle and the output hypothesis must be polynomially evaluatable and have a tolerance τ that is lower-bounded by an inverse polynomial in n, $1/\epsilon$.*

Definition 5 (Distribution-Independent SQ Query Complexity) *Let X be the instance space (with representation size n) and C a concept class over X. We say that the distribution-independent* query complexity *of learning C to accuracy ϵ is bounded by q, if there exists a (possibly computationally unbounded) algorithm that for every concept $f \in C$, every distribution D over X makes at most q queries to oracle $\mathsf{SQ}\text{-}\mathcal{O}(f, D)$ outputs a hypothesis h, such that $\mathrm{err}_D(h, f) \leq \epsilon$. The tolerance τ for the queries must be lower-bounded by an inverse polynomial in n and $1/\epsilon$.*

Distribution-Independent CSQ Learning:

Definition 6 (Distribution-Independent CSQ Learning) *Let X be the instance space (with representation size n) and C a concept class over X. We say that C is* distribution-independently *CSQ learnable if there exists a randomized algorithm $\mathcal{A}$ that for every $\epsilon, \delta > 0$, every target concept $f \in C$ and for every distribution D over X, with access to oracle $\mathsf{CSQ}\text{-}\mathcal{O}(f, D)$, outputs with probability at least $1 - \delta$, a hypothesis h such that $\mathrm{err}_D(h, f) \leq \epsilon$. Furthermore, the running time of the algorithm must be polynomial in n, $1/\epsilon$ and $1/\delta$ and the queries made to the oracle and the output hypothesis must be polynomially evaluatable and have a tolerance τ, that is lower-bounded by a polynomial in n, $1/\epsilon$.*

As in the previous cases, one can define the *distribution-independent query complexity* of CSQ learning. This captures the information-theoretic complexity of CSQ learning.

Definition 7 (Distribution-Independent CSQ Query Complexity) *Let X be the instance space (with representation size n) and C a concept class over X. We say that the distribution-independent* query complexity *of learning C to accuracy ϵ is bounded by q, if there exists a (possibly computationally unbounded) algorithm that for every concept $f \in C$, every distribution D over X makes at most q queries to oracle $\mathsf{CSQ}\text{-}\mathcal{O}(f, D)$ outputs a hypothesis h, such that $\mathrm{err}_D(h, f) \leq \epsilon$. The tolerance τ for the queries must be lower-bounded by an inverse polynomial in n and $1/\epsilon$.*

Appendix B. Omitted Proofs

B.1. Weak Distribution-Specific SQ/CSQ Learning

Proof [of Theorem 1] SQ-DIM(C, D, γ) is the size of the largest subset $S \subseteq C$, such that for every $c_1, c_2 \in S$, $\mathrm{err}_D(c_1, c_2) = \Pr_{x\sim D}[c_1(x) \neq c_2(x)] \geq 1/2 - \gamma$. A simple weak-learning algorithm just tries every concept from S[9], and at least one of them will have error less than $1/2 - \gamma$ with the target function f (since S is the largest such subset, if this weren't the case adding the target function f to S would give a larger subset with the same property). Blum et al. (1994) showed that SQ-DIM(C, D, γ) is polynomially related to the query complexity of weak SQ learning C under D to accuracy $1/2 - \gamma$. ■

B.2. Strong Distribution-Specific SQ Learning

Proof [of Theorem 2] Let $f_{\phi,\zeta(\phi)} \in C_1$ be the target function. We show how to obtain ϕ using statistical queries. For $i \in \{1, \ldots, m\}$, define the function $\psi_i : \{-1, 1\}^{m+n+1} \times \{-1, 1\} \to [-1, 1]$ as follows:

$$\psi_i(bxx', y) = \begin{cases} 0 & \text{if } b = -1 \\ x_i y & \text{if } b = 1 \end{cases}$$

Then, observe that $\mathbb{E}_{U_{m+n+1}}[\psi(bxy, f_{\phi,\zeta(\phi)}(bxy))] = (1/2)\mathbb{E}_{U_m}[x_i \mathsf{MAJ}_\phi(x)]$. It is well known (see for example O'Donnell (2003)) that if $\phi_i = -1$ (i.e. the i^{th} bit is part of the majority function) then $\mathbb{E}_{U_m}[x_i \mathsf{MAJ}_\phi(x)] = \Omega(1/\sqrt{m})$ and 0 otherwise. Hence, by setting $\tau = \Theta(1/\sqrt{m})$, the query (ψ_i, τ) reveals the bit ϕ_i. Thus, using at most $m = O(n^3)$ statistical queries, we obtain ϕ. Now, it is easy to obtain (possibly using unbounded computation) the value $\zeta(\phi)$ and thus obtain the function, $f_{\phi,\zeta(\phi)}$. ■

Proof [of Theorem 3] Suppose to the contrary that $\mathcal{A}$ is a (possibly randomized) algorithm that learns C_1 to error at most 0.1 (in fact, to any value noticeably lower than $1/4$) in polynomial time. We show that using $\mathcal{A}$ it is possible (with high probability) to find a satisfying assignment to any 3-CNF formula ϕ, if one exists. Thus, failure to find a satisfying assignment implies that ϕ is unsatisfiable.

Let ϕ be a 3-CNF instance. Suppose ϕ is a satisfiable, so that $f_{\phi,\zeta(\phi)} \in C_1$; we show that in this case a solution to ϕ can be obtained with high probability. Suppose $\mathcal{A}$ makes q statistical queries each with tolerance τ to learn C_1. We show that we can simulate any statistical query (ψ, τ) with respect to $f_{\phi,\zeta(\phi)}$ efficiently. The queries made by $\mathcal{A}$ to the oracle SQ-$\mathcal{O}$ may be target-independent or correlational. Below, we consider the two cases:

1. Let $(\psi^{\mathrm{ti}}, \tau)$ be a *target-independent* query; we need to return an (additive) τ-approximation to the value $\mathbb{E}_{U_{m+n+1}}[\psi(bxx')]$. This is easily achieved by drawing a sample of size $\tilde{O}(1/\tau^2)$ from the uniform distribution and returning the empirical estimate.

2. Let ψ^{cor} be the *correlational* query. In this case, we need to return an (additive) τ-approximation to the value $\mathbb{E}_{U_{m+n+1}}[\psi^{\mathrm{cor}}(bxx')f_{\phi,\zeta(\phi)}(bxx')]$. For $b \in \{-1, 1\}$, define $\psi_b^{\mathrm{cor}}(xx') \equiv \psi^{\mathrm{cor}}(bxx')$.

9. Note that this is a non-uniform algorithm, since the set S needs to be given as advice to the algorithm

Then,

$$\mathbb{E}_{U_{m+n+1}}[\psi^{\text{cor}}(bxx')f_{\phi,\zeta(\phi)}(bxx')] = \frac{1}{2}\mathbb{E}_{U_{m+n}}[\psi_1^{\text{cor}}(xx')\mathsf{MAJ}_\phi(x)] + \frac{1}{2}\mathbb{E}_{U_{m+n}}[\psi_{-1}^{\text{cor}}(xx')\mathsf{PAR}_{\zeta(\phi)}(x')]$$

It suffices to find τ-approximations to both the terms in the above expression. To obtain a τ-approximate estimate of $\mathbb{E}_{U_{m+n}}[\psi_1^{\text{cor}}(xx')\mathsf{MAJ}_\phi(x)]$, as in the earlier case, we can draw a sample of size $\tilde{O}(1/\tau^2)$ from U_{m+n} and return the empirical estimate (since we can efficiently compute the functions ψ_1^{cor} and MAJ_ϕ).

We show that either 0 is a τ-approximation to $\mathbb{E}_{U_{m+n}}[\psi_{-1}^{\text{cor}}(xx')\mathsf{PAR}_{\zeta(\phi)}(x')]$ or we find a satisfying assignment to ϕ using Fourier analysis. Observe that $\mathbb{E}_{U_{m+n}}[\psi_{-1}^{\text{cor}}(xx')\mathsf{PAR}_{\zeta(\phi)}(x')]$ is simply the Fourier coefficient of ψ_{-1}^{cor} corresponding to $\zeta(\phi)$ (or actually the set $S(\zeta(\phi)) = \{m+i \mid \zeta(\phi)_i = -1\} \subseteq [m+n]$). We know that all Fourier coefficients of ψ_{-1}^{cor} of magnitude larger than $\tau/2$ can be estimated to accuracy $\tau/4$ using the KM algorithm in time $\text{poly}(n, 1/\tau)$ (see Section 2.2 or Kushilevitz and Mansour (1993)). Furthermore, the number of such coefficients is polynomial in n, $1/\tau$. We check whether any such coefficient (interpretted as a string of length n) is a satisfying assignment of ϕ. If we find an assignment, we are done; if not we know that the coefficient $|\widehat{\psi_{-1}^{\text{cor}}}(S(\zeta(\phi)))| \leq \tau$, since $\zeta(\phi)$ is a solution to ϕ and we would have identified it as such, had it been in the list of heavy coefficients. Thus, 0 is an (additive) τ-approximate estimate to the term $\mathbb{E}_{U_{m+n}}[\psi_{-1}^{\text{cor}}(xx')\mathsf{PAR}_{\zeta(\phi)}(x')]$.

Thus, we have shown that we can either find a satisfying assignment to ϕ or simulate the SQ-$\mathcal{O}$ oracle response satisfactorily to all queries made by algorithm $\mathcal{A}$. In the latter case, the algorithm outputs h such that $\text{err}_{U_{m+n+1}}(h, f_{\phi,\zeta(\phi)}) \leq 0.1$, i.e. $\mathbb{E}_{U_{m+n+1}}[h(bxx')f_{\phi,\zeta(\phi)}(bxx')] \geq 4/5$. Let $h_b(xx') \equiv h(bxx')$, then

$$\mathbb{E}_{U_{m+n+1}}[h(bxx')f_{\phi,\zeta(\phi)}(bxx')] = \frac{1}{2}\mathbb{E}_{U_{m+n}}[h_1(xx')\mathsf{MAJ}_\phi(x)] + \frac{1}{2}\mathbb{E}_{U_{m+n}}[h_{-1}(xx')\mathsf{PAR}_{\zeta(\phi)}(x')]$$

The above equation implies that $\mathbb{E}_{U_{m+n}}[h_{-1}(xx')\mathsf{PAR}_{\zeta(\phi)}(x')] = \hat{h}_{-1}(S(\zeta(\phi))) \geq 3/5$, where $\hat{h}_{-1}(S(\zeta(\phi)))$ is the Fourier coefficient of h_{-1} corresponding to the set $S(\zeta(\phi))$. Thus identifying all large coefficients of h_{-1}, by the KM algorithm, and checking whether any of the coefficients (when interpreted as a string of length n) satisfies ϕ, a satisfying assignment of ϕ is obtained (since $\zeta(\phi)$ has a large Fourier coefficient).

Thus, if ϕ is satisfiable, using $\mathcal{A}$ it is possible to find, with high probability, a satisfying assignment to ϕ. If we fail to find the satisfying assignment, then ϕ is unsatisfiable. Hence, an algorithm to efficiently SQ learn C_1 does not exist unless RP = NP. ■

B.3. Strong Distribution-Independent SQ Learning

Proof [of Theorem 4] Let $g_{\phi,\zeta(\phi)}$ be the target function, D the target distribution and let $\epsilon > 0$ be the target error rate. We first test if the hypothesis, the constant 1 function, is ϵ-accurate. This can be tested using a single correlational statistical query $(1, \epsilon/4)$. If the value returned is at least $1 - 3\epsilon/4$, then $\mathbb{E}_D[g_{\phi,\zeta(\phi)}(xx')] \geq 1 - \epsilon$, i.e. the constant 1 hypothesis is ϵ-accurate. If not, we know that $g_{\phi,\zeta(\phi)}$ is -1 on at least $\epsilon/4$ fraction of the domain (under the target distribution D).

Now, suppose that $\Pr_D[g_{\phi,\zeta(\phi)}(xx') = -1] \geq \epsilon/4$. For $i = 1, \ldots, m$, define $\psi_i : \{-1,1\}^{m+n} \to [-1,1]$ as the following function: $\psi_i(xx') = 1$, if $x_i = 1$ and $\psi_i(xx') = 0$ otherwise.

Consider the following expectation,

$$\mathbb{E}_D[\psi_i(xx') - g_{\phi,\zeta(\phi)}(xx')\psi_i(xx')]$$

If $\phi_i = -1$ (i.e. the i^{th} bit of the representation of the 3-CNF formula ϕ is -1), then $g_{\phi,\zeta(\phi)}(xx') = 1$ for all points where $\psi_i(xx') \neq 0$. This is because $g_{\phi,\zeta(\phi)}$ is the constant 1 function on points which do not have ϕ as a prefix, and if $\psi_i(xx') \neq 0$, then $x_i = 1 \neq \phi_i$. Thus, for all points $\psi_i(xx') = g_{\phi,\zeta(\phi)}(xx')\psi_i(xx')$ and hence the value of the above expecation is exactly 0.

On the other hand, if $\phi_i = 1$, then whenever $g_{\phi,\zeta(\phi)}(xx') = -1$, $\psi_i(xx') = 1$. When $g_{\phi,\zeta(\phi)}(xx') = 1$, $\psi_i(xx') - g_{\phi,\zeta(\phi)}(xx')\psi_i(xx') = 0$. Recall that $\Pr_D[g_{\phi,\zeta(\phi)}(xx') = -1] \geq \epsilon/4$, thus the above expectation is at least $\epsilon/2$.

As $\mathbb{E}_D[\psi_i(xx') - g_{\phi,\zeta(\phi)}(xx')\psi_i(xx')] = \mathbb{E}_D[\psi_i(xx')] - \mathbb{E}_D[g_{\phi,\zeta(\phi)}(xx')\psi_i(xx')]$, an $\epsilon/8$ accurate estimate to the above expectation can be obtained by making a target independent query $(\psi_i, \epsilon/16)$ and a correlational query $(\psi_i, \epsilon/16)$. Thus, by looking at the query responses it is possible to determine whether $\phi_i = 1$ or $\phi_i = -1$.

Using $2m$ queries, each bit of ϕ can be determined, and then $\zeta(\phi)$ can be obtained, if necessary by brute force, to output $g_{\phi,\zeta(\phi)}$. ■

Proof [of Theorem 5] Suppose to the contrary and let $\mathcal{A}$ be a (possibly randomized) algorithm that efficiently learns C_2 in the distribution-independent SQ model. We show that if ϕ is a satisfiable 3-CNF formula then, using $\mathcal{A}$, a satisfying assignment can be constructed with high probability.

Let ϕ be a 3-CNF formula that is satisfiable, so that $g_{\phi,\zeta(\phi)} \in C_2$. Let D_2 be the distribution defined as follows: $D_2(xx') = 2^{-n}$ if $x = \phi$, $D_2(xx') = 0$ otherwise; thus, D_2 is the uniform distribution on strings of the form $\phi x'$.

Let $g_{\phi,\zeta(\phi)}$ be the target concept from C_2 and D_2 the target distribution. Suppose $\epsilon \leq 1/4$. We run $\mathcal{A}$ to learn $g_{\phi,\zeta(\phi)}$. We need to show that we can simulate the queries made by $\mathcal{A}$ to the oracle $\mathsf{SQ}\text{-}\mathcal{O}(g_{\phi,\zeta(\phi)}, D_2)$.

As in the proof of Theorem 3, response to a target-independent query can be simulated by drawing a sample from D_2 of size $\tilde{O}(1/\tau^2)$ and returning the empirical estimate. In the case of correlational queries also, the main idea is similar to that used in the proof of Theorem 3. Let $(\psi^{\text{cor}}, \tau)$ be a correlational query, define $\psi^{\text{cor}}_\phi : \{-1,1\}^n \to [-1,1]$ to be the function $\psi^{\text{cor}}_\phi(x') = \psi^{\text{cor}}(\phi x')$. Thus, $\mathbb{E}_{D_2}[\psi^{\text{cor}}(xx')g_{\phi,\zeta(\phi)}(xx')] = \mathbb{E}_{U_n}[\psi^{\text{cor}}_\phi(x')\mathsf{PAR}_{\zeta(\phi)}(x')]$. This is just the Fourier coefficient of ψ^{cor}_ϕ on the subset $S(\zeta(\phi))$. Thus, we obtain all large (of magnitude greater than $\tau/2$) Fourier coefficients of ψ^{cor}_ϕ and check whether any of them (i.e. their string representations of length n) are a satisfying assignment to ϕ. If not, then 0 is valid (τ-approximate) answer to the query $(\psi^{\text{cor}}, \tau)$.

Thus, we can simulate access to the $\mathsf{SQ}\text{-}\mathcal{O}(g_{\phi,\zeta(\phi)}, D_2)$ oracle to $\mathcal{A}$ or else we find a satisfying assignment to ϕ. Suppose we don't find a satisfying assignment to ϕ and $\mathcal{A}$ runs to completion, then for the output hypothesis, h, $\text{err}_{D_2}(h, g_{\phi,\zeta(\phi)}) \leq 1/4$ or equivalently, $\mathbb{E}_{D_2}[h(xx')g_{\phi,\zeta(\phi)}(xx')] \geq 1/2$. Again define $h_\phi(x') = h(\phi x')$, so that $\mathbb{E}_{U_n}[h_\phi(x')\mathsf{PAR}_{\zeta(\phi)}(x')] \geq 1/2$. Thus, looking at the heavy Fourier coefficients of h_ϕ reveals a satisfying assignment to ϕ. The above algorithm works correctly with high probability.

If we are unable to find a satisfying assignment of ϕ, then we report ϕ as being unsatisfiable. Thus, we get a randomized polytime algorithm for 3-CNF. ■

Appendix C. Strong Distribution-Independent CSQ Learning: Lower Bounds

In this section, we provide details omitted from Section 4.4. We first show that the class of disjunctions is weakly learnable in the CSQ model (distribution-independently).

C.1. Weak Distribution-independent CSQ Learning Disjunctions

Let $\mathsf{DISJ}_n = \{\mathsf{OR}_z \mid z \in \{-1,1\}^n\}$ be the class of disjunctions over n variables. Let $x \in \{-1,1\}^n$ and let $x_1, x_2, \ldots, x_n$ be the input bits. Let $\mathcal{W} = \{-1, x_1, x_2, \ldots, x_n\}$ be a set of $n+1$ functions, where -1 is the constant function that is -1 everywhere, and x_i is the function $w(x) = x_i$. The following simple lemma shows that for every $z \in \{-1,1\}^n$ and every distribution D over $\{-1,1\}^n$, there exists $w \in \mathcal{W}$ such that $\mathbb{E}_D[\mathsf{OR}_z(x)w(x)] \geq 1/(2n)$. Thus, this implies that the class DISJ_n is efficiently weakly distribution-independently CSQ learnable. Feldman (2008) gives a proof of this lemma, but we include a proof for completeness.

Lemma 12 *For every $\mathsf{OR}_z \in \mathsf{DISJ}_n$ and every distribution D over $\{-1,1\}^n$, there exists $w \in \mathcal{W}$ such that $\mathbb{E}_D[\mathsf{OR}_z(x)w(x)] \geq 1/(2n)$.*

Proof For a string $z \in \{-1,1\}^n$, recall that $S(z) = \{i \mid z_i = -1\}$. Let $\beta_z(x) = \sum_{i \in S(z)} x_i - |S(z)| + 1$. Then observe that $\mathsf{OR}_z(x) = \mathrm{sign}(\beta_z(x))$, since $\mathsf{OR}_z(x) = 1$ if all x_i such that $i \in S(z)$ are 1, in which case $\beta_z(x) = \sum_{i \in S(z)} x_i - |S(z)| + 1 = 1$, otherwise $\beta_z(x) = \sum_{i \in S(z)} x_i - |S(z)| + 1$ is at most -1.

Note that $\beta_z(x) = \sum_{i \in S(z)} x_i - |S(z)| + 1$ is always an odd integer, and hence $|\beta_z(x)| \geq 1$ for all x. Thus, for all x, $\beta_z(x)\,\mathrm{sign}(\beta_z(x)) \geq 1$.

Then for any distribution D over $\{-1,1\}^n$ we have,

$$\mathbb{E}_{x \sim D}[\beta_z(x)\mathsf{OR}_z(x)] = \mathbb{E}_{x \sim D}[\beta_z(x)\,\mathrm{sign}(\beta_z(x))] \geq 1$$

Hence, either $\mathbb{E}_D[(-1)\cdot\mathsf{OR}_z(x)] \geq 1/(2(|S(z)|-1))$ or there exists $i \in S(z)$ such that $\mathbb{E}_D[x_i\mathsf{OR}_z(x)] \geq 1/(2|S(z)|)$. ■

C.2. Encoding Sparse Strings

We give here a simple implementation of the encoding of sparse strings described in Section 4.4. Let s be a string of length n that contains at most ℓ "-1"-bits. We want to encode s as a string, $\mathsf{Enc}(s)$, of length $3\ell n$ that has at most 3ℓ "-1"-bits such that identifying any ℓ of the 3ℓ positions that have "-1" suffice to recover the string s. For a string s' of length $3\ell n$, let $\mathsf{Dec}(s')$ denote the (unique) string of length n, such that $\mathsf{Enc}(\mathsf{Dec}(s')) = s'$, or the null string if no such string exists.

For simplicity, let n be a power of 2, say $n = 2^k$. Given $s \in \{-1,1\}^n$ with at most ℓ "-1" bits, do the following: Identify the set $S = \{i \mid s_i = -1\}$; notice that $|S| = \ell$. We use the Reed-Solomon code to encode the elements of S using a set T, $|T| = 3\ell$ such that identifying any subset of T of size ℓ allows us to recover S. This is done by interpretting $i \in S$ as elements of the field $\mathbb{F}_{2^k}$ and constructing a polymial of degree $\ell - 1$, using elements of S as the coefficients. The set T contains an evaluation of this polynomial at 3ℓ different points in $\mathbb{F}_{2^k}$. Clearly, identifying any ℓ elements

of T is enough to perform interpolation and hence obtain S. Now, we can encode T using a string of length $3\ell n$, with at most 3ℓ, "-1" bits as follows: Let $T = \{t_1, \ldots, t_{3\ell}\}$ and consider the string $s' = \mathsf{Enc}(s)$ as 3ℓ blocks of length n. In the i^{th} block, only the t_i^{th} bit is -1 and the rest are all 1. Notice, that although t_i are technically elements of $\mathbb{F}_{2^k}$, they can be interpretted as integers less than n. Thus identifying the positions of any ℓ, "-1" bits of s' allows for decoding and recovering s. Denote by $\mathsf{Dec}(s')$ the string s, if s' is any string that has at least ℓ "-1" bits and must have been a (corrupted) version of $\mathsf{Enc}(s)$.

C.3. Omitted Proofs

Proof [of Theorem 6] Let $c_{\phi,\xi(\phi)}$ be the target concept from C_3 and let $\epsilon > 0$ be the target error rate. We first test the hypothesis that is constant 1 everywhere. This can be tested using the correlational query $(1, \epsilon/4)$, where 1 is the constant 1 function. If the query response is greater than $1 - 3\epsilon/4$, then $\mathbb{E}_D[c_{\phi,\xi(\phi)}(xx')] \geq 1 - \epsilon$ and hence the constant 1 function is an ϵ-accurate hypothesis and we are done. Otherwise, $\Pr_D[c_{\phi,\xi(\phi)}(xx') = -1] \geq \epsilon/4$.

Let ϕ be the encoding of the 3-CNF-SAT formula corresponding to the target function $c_{\phi,\xi(\phi)}$. Let D_1 be the marginal distribution over the first m bits of the target distribution D. Suppose $h : \{-1,1\}^m \to \{-1,1\}$ is a function satisfying the two properties: (i) $h(\phi) = 1$, and (ii) $\Pr_{D_1}[h(x) = 1 \wedge x \neq \phi] \leq \epsilon/(100n)$.

Let D_2 be the distribution D conditioned on the first m bits being ϕ, i.e. $D_2(x') = D(\phi x')/D_1(\phi)$. Let $\mathcal{W}$ be as in Lemma 12 and let $w \in \mathcal{W}$ be such that $\mathbb{E}_{D_2}[\mathsf{OR}_{\xi(\phi)}(x')w(x')] \geq 1/(2n)$. Then, define the function $h^w(xx') = w(x')$ if $h(x) = 1$ and $h^w(xx') = 0$ otherwise. Note that,

$$\begin{aligned}\mathbb{E}_D[h^w(xx')c_{\phi,\xi(\phi)}(xx')] &\geq \Pr_{D_1}[x = \phi]\mathbb{E}_{D_2}[\mathsf{OR}_{\xi(\phi)}(x')w(x')] - \Pr_{D_1}[h^w(xx') = 1 \wedge x \neq \phi] \\ &\geq \frac{\epsilon}{4} \cdot \frac{1}{2n} - \frac{\epsilon}{100n}\end{aligned}$$

Now define $h_i^w : \{-1,1\}^{m+n} \to [-1,1]$ to be the function, where $h_i^w(xx') = w(x')$ if $h(x) = 1$ and $x_i = 1$, and $h_i^w(xx') = 0$ otherwise. Now note that if $\phi_i = 1$, $\mathbb{E}_D[h_i^w(xx)c_{\phi,\xi(\phi)}(xx')] \geq \epsilon/(8n) - \epsilon/(100n)$, as in the previous case. On the other hand if $\phi_i = -1$, then $\mathbb{E}_D[h_i^w(xx')c_{\phi,\xi(\phi)}(xx')] \leq \epsilon/(100n)$.

This gap between the expectations in the two cases is large enough that the response to the correlational statistical query $(h_i^w, \epsilon/(100n))$ distinguishes the case when $\phi = 1$ and $\phi = -1$. Thus m such correlational queries can be used to exactly determine ϕ and then (possibly using unbounded computation) $\xi(\phi)$ may be determined to identify $c_{\phi,\xi(\phi)}$.

Now, suppose we did not know that h satisfied the properties (i) and (ii), mentioned above. We could still carry out the operations described above to come up with a candidate $\tilde{\phi}$ and guess $c_{\tilde{\phi},\xi(\tilde{\phi})}$ to be the target concept. We can then simply make the correlational query $(c_{\tilde{\phi},\xi(\tilde{\phi})}, \epsilon/4)$ to check whether $c_{\tilde{\phi},\xi(\tilde{\phi})}$ is an ϵ-accurate hypothesis. Note that if h did indeed satisfy the properties (i) and (ii), then $\tilde{\phi} = \phi$.

The last part required to complete the proof is to show that it is easy to construct a random hypothesis h that satisfies properties (i) and (ii) with non-negligible (inverse polynomial) probability. Then, several such hypotheses may be generated and each tested until the right one (or one that is good enough) is found. But, this is exactly what Feldman's algorithm for CSQ learning singletons does Feldman (2009a). ■

Proof [of Theorem 7] Suppose that there exists an efficient algorithm, $\mathcal{A}$, that distribution-independently CSQ learns C_3. Let ϕ be a circuit formula. We show that if ϕ has a satisfying assignment of Hamming weight at most ℓ, then using $\mathcal{A}$ we can find such a solution, with high probability.

Let $z = \xi(\phi)$, $S(z) = \{i \mid z_i = -1\}$ and suppose that $|S(z)| = k$, where $k \leq 3\ell$. Note that $\zeta(\phi)$ has Hamming weight at most ℓ. Then, the function OR_z can be expressed as the following polynomial.

$$\begin{aligned}\mathsf{OR}_z(x') &= -1 + 2 \prod_{i \in S(z)} \frac{1 + x'_i}{2} \\ &= -1 + 2^{-k+1} \sum_{T \subseteq S(z)} \chi_T(x')\end{aligned}$$

where $\chi_T(x')$ is the parity function over T. Let t_z be the polynomial,

$$t_z(x') = -1 + 2^{-k+1} + 2^{-k+1} \sum_{\substack{T \subseteq S(z) \\ |T| > k/3}} \chi_T(x')$$

Define D_z to be the distribution where $D_z(x') = |t_z(x')|/(\sum_{x'} |t_z(x')|)$. Feldman (2011) showed that $\mathrm{sign}(t_z(x')) = \mathsf{OR}_z(x')$, and hence for all x', $D_z(x')\mathsf{OR}_z(x') = U_{n'}(x')t_z(x')$, where $U_{n'}$ is the uniform distribution over n' bits.

Define D to be the distribution over $\{-1,1\}^{m+n'}$, where $D(xx') = D_z(x')$ if $x = \phi$ and $D(xx') = 0$ if $x \neq \phi$. Now, we run algorithm $\mathcal{A}$ to learn C_3 to accuracy $\epsilon = 2^{-k-2}$, where D is the target distribution and $c_{\phi,\xi(\phi)}$ is the target concept. We need to show that we can simulate oracle CSQ-$\mathcal{O}$ for any query (ψ, τ). Let $\psi_\phi : \{-1,1\}^{n'} \to [-1,1]$ be the function where $\psi_\phi(x') = \psi(\phi x')$.

Note that,

$$\mathbb{E}_D[\psi(xx')c_{\phi,\xi(\phi)}(xx')] = \mathbb{E}_{D_z}[\psi_\phi(x')\mathsf{OR}_z(x')] = \mathbb{E}_{U_{n'}}[\psi_\phi(x')t_z(x')]$$

Then observe that,

$$\mathbb{E}_{U_{n'}}[\psi_\phi(x')t_z(x')] = (-1 + 2^{-k+1})\widehat{\psi_\phi}(\emptyset) + 2^{-k+1} \sum_{\substack{T \subseteq S(z) \\ |T| > k/3}} \widehat{\psi_\phi}(T)$$

Note that the only Fourier coefficients of ψ_ϕ that matter are those corresponding to the empty set and sets $T \subseteq S(z)$ such that $|T| \geq k/3$. There are at most 2^k subsets of $S(z)$. Using the KM algorithm, we can identify in time polynomial in $2^k, n, 1/\tau$, all Fourier coefficients of ψ_ϕ whose magnitude is at least $\tau/2^k$. Now if there exists a subset $T \subseteq S(z)$ such that $|\widehat{\psi_\phi}| \geq \tau/2^{-k}$ and $|T| > k/3$, then it will be in the list of coefficients obtained above. But note that T can be converted into a string of length n', say $\sigma(T)$, such that $\mathsf{Dec}(\sigma(T)) = \zeta(\phi)$ which is a satisfying assignment of ϕ. Thus, for each heavy (magnitude $\geq \tau/2^k$) Fourier coefficient of ψ_ϕ, we check if we get a satisfying assignment to ϕ. If not, then 0 is a valid answer (τ-approximate) to the query (ψ, τ).

The algorithm, $\mathcal{A}$, outputs a hypothesis h. Let $h_\phi(x') = h(\phi x')$. Note that

$$\begin{aligned}\mathbb{E}_D[h(xx')c_{\phi,\xi(\phi)}(xx')] &= \mathbb{E}_{U_{n'}}[h_\phi(x')t_z(x')] \\ &= (-1+2^{-k+1})\widehat{h_\phi}(\emptyset) + 2^{-k+1}\sum_{\substack{T\subseteq S(z)\\|T|>k/3}}\widehat{h_\phi}(T) \geq 1-2\epsilon = 1-2^{-k-1}\,.\end{aligned}$$

This means that

$$\sum_{\substack{T\subseteq S(z)\\|T|>k/3}}\widehat{h_\phi}(T) \geq 1/2.$$

Thus, as for the queries, identifying and decoding all large (magnitude $\geq 0.1/2^k$) Fourier coefficients of h_ϕ reveals a satisfying assignment of ϕ of Hamming weight at most ℓ. ■