Two new results about quantum exact learning

Srinivasan Arunachalam[1], Sourav Chakraborty[2], Troy Lee[3], Manaswi Paraashar[2], and Ronald de Wolf[4]

[1]IBM T. J. Watson Research Center

[2]Indian Statistical Institute, Kolkata, India

[3]Centre for Quantum Software and Information, University of Technology Sydney, Australia

[4]QuSoft, CWI and University of Amsterdam, the Netherlands

We present two new results about exact learning by quantum computers. First, we show how to exactly learn a k-Fourier-sparse n-bit Boolean function from $O(k^{1.5}(\log k)^2)$ uniform quantum examples for that function. This improves over the bound of $\widetilde{\Theta}(kn)$ uniformly random *classical* examples (Haviv and Regev, CCC'15). Additionally, we provide a possible direction to improve our $\widetilde{O}(k^{1.5})$ upper bound by proving an improvement of Chang's lemma for k-Fourier-sparse Boolean functions. Second, we show that if a concept class $\mathcal{C}$ can be exactly learned using Q quantum membership queries, then it can also be learned using $O\left(\frac{Q^2}{\log Q}\log|\mathcal{C}|\right)$ *classical* membership queries. This improves the previous-best simulation result (Servedio and Gortler, SICOMP'04) by a $\log Q$-factor.

1 Introduction

1.1 Quantum learning theory

Both quantum computing and machine learning are hot topics at the moment, and their intersection has been receiving growing attention in recent years as well. On the one hand there are particular approaches that use quantum algorithms like Grover search [18] and

Srinivasan Arunachalam: Work done while a Postdoc at Center for Theoretical Physics, MIT and PhD student at QuSoft, CWI, Amsterdam, the Netherlands. Supported by ERC Consolidator Grant 615307 QPROGRESS and MIT-IBM Watson AI Lab under the project *Machine Learning in Hilbert space*. Srinivasan.Arunachalam@ibm.com

Sourav Chakraborty: Work done while on sabbatical at CWI, supported by ERC Consolidator Grant 615307 QPROGRESS. sourav@isical.ac.in

Troy Lee: Partially supported by the Australian Research Council (Grant No: DP200100950). Part of this work was done while at the School for Physical and Mathematical Sciences, Nanyang Technological University and the Centre for Quantum Technologies, Singapore, supported by the Singapore National Research Foundation under NRF RF Award No. NRF-NRFF2013-13. troyjlee@gmail.com

Manaswi Paraashar: manaswi.isi@gmail.com

Ronald de Wolf: Partially supported by ERC Consolidator Grant 615307-QPROGRESS (which ended February 2019), and by the Dutch Research Council (NWO) through Gravitation-grant Quantum Software Consortium 024.003.037 and through QuantERA project QuantAlgo 680-91-034. rdewolf@cwi.nl

A conference version of this paper appeared in the proceedings of the 46th International Colloquium on Automata, Languages and Programming (ICALP 19), Leibniz International Proceedings in Informatics (LIPIcs) volume 132, pp.16:1-16:15, 2019.

the Harrow-Hassidim-Lloyd linear-systems solver [19] to speed up learning algorithms for specific machine learning tasks (see [1, 8, 16, 28, 33] for recent surveys of this line of work). On the other hand there have been a number of more general results about the sample and/or time complexity of learning various concept classes using a quantum computer (see [3] for a survey). This paper presents two new results in the latter line of work. In both cases the goal is to *exactly* learn an unknown target function with high probability; for the first result our access to the target function is through quantum examples for the function, and for the second result our access is through membership queries to the function.

1.2 Exact learning of sparse functions from uniform quantum examples

Let us first explain the setting of distribution-dependent learning from examples. Let $\mathcal{C}$ be a class of functions, a.k.a. a *concept class*. For concreteness assume they are ± 1-valued functions on a domain of size N; if $N = 2^n$, then the domain may be identified with $\{0,1\}^n$. Suppose $c \in \mathcal{C}$ is an unknown function (the *target* function or concept) that we want to learn. A learning algorithm is given *examples* of the form $(x, c(x))$, where x is distributed according to some probability distribution D on $[N]$. An (ε, δ)-learner for $\mathcal{C}$ w.r.t. D is an algorithm that, for every possible target concept $c \in \mathcal{C}$, produces a hypothesis $h : [N] \to \{-1, 1\}$ such that with probability at least $1 - \delta$ (over the randomness of the learner and the examples for the target concept c), h's generalization error is at most ε, i.e.,

$$\Pr_{x \sim D}[c(x) \neq h(x)] \leq \varepsilon,$$

where $x \sim D$ means x is sampled according to the distribution D. In other words, from D-distributed examples the learner has to construct a hypothesis that mostly agrees with the target concept *under the same* D.

In the early days of quantum computing, Bshouty and Jackson [10] generalized this learning setting by allowing coherent *quantum* examples. A quantum example for concept c w.r.t. distribution D, is the following $(\lceil \log N \rceil + 1)$-qubit state:

$$\sum_{x \in [N]} \sqrt{D(x)} |x, c(x)\rangle.$$

Clearly such a quantum example is at least as useful as a classical example, because measuring this state yields a pair $(x, c(x))$ where $x \sim D$. Bshouty and Jackson gave examples of concept classes that can be learned more efficiently from quantum examples than from classical random examples under specific D. In particular, they showed that the concept class of DNF-formulas can be learned in polynomial time from quantum examples under the *uniform* distribution, something we do not know how to do classically (the best classical upper bound is quasi-polynomial time [32]). The key to this improvement is the ability to obtain, from a uniform quantum example, a sample $S \sim \widehat{c}(S)^2$ distributed according to the squared *Fourier coefficients* of c.[1] This *Fourier sampling*, originally due to Bernstein and Vazirani [7], is very powerful. For example, if $\mathcal{C}$ is the class of $\mathbb{F}_2$-linear functions on $\{0,1\}^n$, then the unknown target concept c is a character function $\chi_S(x) = (-1)^{x \cdot S}$ [2]; its only non-zero Fourier coefficient is $\widehat{c}(S)$ hence one Fourier-sample gives us the unknown S with certainty. In contrast, learning linear functions from classical

[1] Parseval's identity implies $\sum_{S \in \{0,1\}^n} \widehat{c}(S)^2 = 1$, so this is indeed a probability distribution.

[2] The linear functions with domain $\{0,1\}^n$ and range $\{0,1\}$ are defined as $(S \cdot x) \bmod 2$, for $S \subseteq [n]$. The definition of linear functions we give here are for functions with range $\{-1,1\}$ rather than $\{0,1\}$.

uniform examples requires $\Theta(n)$ examples. Another example where Fourier sampling is proven powerful is in learning the class of ℓ-juntas on n bits.[3] Atıcı and Servedio [5] showed that $(\log n)$-juntas can be exactly learned by a quantum learner under the uniform distribution in time polynomial in n. Classically it is a long-standing open question if a similar result holds when the learner is given uniform classical examples (the best known algorithm runs in quasi-polynomial time [23]). These cases (and others surveyed in [3]) show that uniform quantum examples (and in particular Fourier sampling) can be more useful than classical examples.[4]

In this paper we consider the concept class of n-bit Boolean functions (with domain $\{0,1\}^n$ and range $\{-1,1\}$) that are *k-sparse in the Fourier domain*: $\widehat{c}(S) \neq 0$ for at most k different S's. This is a natural generalization of the above-mentioned case of learning linear functions, which corresponds to $k = 1$. It also generalizes the case of learning ℓ-juntas on n bits, which are functions of sparsity $k = 2^\ell$. Variants of the class of k-Fourier-sparse functions have been well-studied in the area of *sparse recovery*, where the goal is to recover a k-sparse vector $x \in \mathbb{R}^N$ given a low-dimensional linear sketch Ax for a so-called "measurement matrix" matrix $A \in \mathbb{R}^{m \times N}$. See [20, 22] for some upper bounds on the size of the measurement matrix that suffice for sparse recovery. Closer to the setting of this paper, there has also been extensive work on learning the concept class of n-bit *real-valued* functions that are k-sparse in the Fourier domain. In this direction Cheraghchi et al. [14] showed that $O(nk(\log k)^3)$ uniform examples suffice to learn this concept class, improving upon the works of Bourgain [9], Rudelson and Vershynin [26] and Candés and Tao [11].

In this paper we focus on *exactly* learning the target concept from uniform examples, with high success probability. So $D(x) = 1/2^n$ for all x, $\varepsilon = 0$, and $\delta = 1/3$. Haviv and Regev [21] showed that for classical learners $O(nk \log k)$ uniform examples suffice to learn k-Fourier-sparse functions, and $\Omega(nk)$ uniform examples are necessary. In Section 3 we study the number of uniform *quantum* examples needed to learn k-Fourier-sparse Boolean functions, and show that it is upper bounded by $O(k^{1.5}(\log k)^2)$. For $k \ll n^2$ this quantum bound is much better than the number of uniform examples used in the classical case. Proving the upper bound is done in two phases. In the first phase we use the fact that a uniform quantum example allows us to Fourier-sample the target concept and, with some Fourier analysis of k-Fourier-sparse functions, we learn the Fourier span using $O(rk)$ examples, where r is the Fourier dimension of the target concept (see Section 2 for the definition of Fourier dimension). In the second phase, we reduce the number of variables to the dimension r of the Fourier support, and then invoke the classical learner of Haviv and Regev to learn the target function from $O(rk \log k)$ classical examples. Since it is known that $r = O(\sqrt{k} \log k)$ [27], the two phases together imply that $O(k^{1.5}(\log k)^2)$ uniform quantum examples suffice to exactly learn the target with high probability. We also prove a (non-matching) lower bound of $\Omega(k \log k)$ uniform quantum examples, using techniques from quantum information theory.

We believe that the sample complexity for Phase 1 of our learning algorithm is actually $\tilde{O}(k)$. Towards that end, we propose a possible way to prove the sample complexity of our Phase 1 to $\tilde{O}(k)$. The first step in Phase 1 of our algorithm is to obtain an $S \neq 0^n$

[3] We say $f : \{0,1\}^n \to \{-1,1\}$ is an *ℓ-junta* if there exists a set $S \subseteq [n]$ of size $|S| \leq \ell$ such that f depends only on the variables whose indices are in S.

[4] This is not the case in Valiant's *PAC-learning* model [31] of distribution-independent learning. There we require the same learner to be an (ε, δ)-learner for $\mathcal{C}$ w.r.t. *every* possible distribution D. One can show in this model (and also in the broader model of *agnostic* learning) that the quantum and classical sample complexities are equal up to a constant factor [4].

such that $\widehat{c}(S) \neq 0$, where c is the k-Fourier-sparse target concept. It follows from Chang's lemma [13], a central result in additive combinatorics, that in expectation $O(k\sqrt{\log k}/\sqrt{r})$ Fourier-samples are sufficient to obtain one such S. In Section 3.3 we present an improvement of Chang's lemma for the case of k-Fourier-sparse Boolean functions. Using this improvement we can show that in expectation $O((k \log k)/r)$ Fourier-samples are sufficient to obtain an $S \neq \emptyset$ such that $\widehat{c}(S) \neq 0$. We conjecture (Conjecture 1) a generalization of our improvement of Chang's Lemma which, if true, would imply that Phase 1 of our algorithm can be done in $\tilde{O}(k)$ many expected number of samples. Our improvement of Chang's lemma and the techniques used therein might be of independent interest.

1.3 Exact learning from quantum membership queries

Our second result is in a model of active learning. The learner still wants to exactly learn an unknown target concept $c : [N] \to \{-1, 1\}$ from a known concept class $\mathcal{C}$, but now the learner can choose which points of the truth-table of the target it sees, rather than those points being chosen randomly. More precisely, the learner can query $c(x)$ for any x of its choice. This is called a *membership query*.[5] Quantum algorithms have the following query operation available:

$$O_c : |x, b\rangle \mapsto |x, b \cdot c(x)\rangle,$$

where $b \in \{-1, 1\}$. For some concept classes, quantum membership queries can be much more useful than classical. Consider again the class $\mathcal{C}$ of $\mathbb{F}_2$-linear functions on $\{0, 1\}^n$. Using one query to a uniform superposition over all x and doing a Hadamard transform, we can Fourier-sample and hence learn the target concept exactly. In contrast, $\Theta(n)$ classical membership queries are necessary and sufficient for classical learners. As another example, consider the concept class $\mathcal{C} = \{\delta_i \mid i \in [N]\}$ of the N point functions, where $\delta_i(x) = 1$ iff $i = x$. Elements from this class can be learned using $O(\sqrt{N})$ quantum membership queries by Grover's algorithm, while every classical algorithm needs to make $\Omega(N)$ membership queries.

For a given concept class $\mathcal{C}$ of ± 1-valued function on $[N]$, let $D(\mathcal{C})$ denote the minimal number of classical membership queries needed for learners that can exactly identify every $c \in \mathcal{C}$ with success probability 1 (such learners are deterministic without loss of generality). Let $R(\mathcal{C})$ and $Q(\mathcal{C})$ denote the minimal number of classical and quantum membership queries, respectively, needed for learners that can exactly identify every $c \in \mathcal{C}$ with error probability $\leq 1/3$.[6] Servedio and Gortler [29] showed that these quantum and classical measures cannot be too far apart. First, using an information-theoretic argument they showed

$$Q(\mathcal{C}) \geq \Omega\left(\frac{\log |\mathcal{C}|}{\log N}\right).$$

Intuitively, this holds because a learner recovers roughly $\log |\mathcal{C}|$ bits of information, while every quantum membership query can give at most $O(\log N)$ bits of information. Note that this is tight for the class of linear functions, where the left- and right-hand sides are both constant. Second, using the so-called hybrid method they showed

$$Q(\mathcal{C}) \geq \Omega(1/\sqrt{\gamma(\mathcal{C})}),$$

[5] Think of the set $\{x \mid c(x) = 1\}$ corresponding to the target concept: a membership query asks whether x is a member of this set or not.

[6] We can identify each concept with a string $c \in \{-1, 1\}^N$, and hence $\mathcal{C} \subseteq \{-1, 1\}^N$. The goal is to learn the unknown $c \in \mathcal{C}$ with high probability using few queries to the corresponding N-bit string. This setting is also sometimes called "oracle identification" in the literature; see [3, Section 4.1] for more references.

for some combinatorial parameter $\gamma(\mathcal{C})$ that we will not define here (but which is $1/N$ for the class $\mathcal{C}$ of point functions, hence this inequality is tight for that $\mathcal{C}$). They also noted the following upper bound:

$$D(\mathcal{C}) = O\left(\frac{\log|\mathcal{C}|}{\gamma(\mathcal{C})}\right).$$

Combining these three inequalities yields the following relation between $D(\mathcal{C})$ and $Q(\mathcal{C})$

$$D(\mathcal{C}) \leq O(Q(\mathcal{C})^2 \log|\mathcal{C}|) \leq O(Q(\mathcal{C})^3 \log N). \tag{1}$$

This shows that, up to a $\log N$-factor, quantum and classical membership query complexities of exact learning are polynomially close. While each of the three inequalities that together imply (1) can be individually tight (for different $\mathcal{C}$), this does not imply (1) itself is tight.

Note that Eq. (1) upper bounds the membership query complexity of *deterministic* classical learners. We are not aware of a stronger upper bound on *bounded-error* classical learners. However, in Section 4 we tighten that bound further by a $\log Q(\mathcal{C})$-factor:

$$R(\mathcal{C}) \leq O\left(\frac{Q(\mathcal{C})^2}{\log Q(\mathcal{C})}\log|\mathcal{C}|\right) \leq O\left(\frac{Q(\mathcal{C})^3}{\log Q(\mathcal{C})}\log N\right).$$

This inequality is tight both for the class of linear functions and the class of point functions.

Our proof combines the quantum adversary method [2, 6, 30] with an entropic argument to show that we can always find a query whose outcome (no matter whether it is 1 or -1) will shrink the concept class by a factor $\leq 1 - \frac{\log Q(\mathcal{C})}{Q(\mathcal{C})^2}$. While our improvement over the earlier bounds is not very large, we feel our usage of entropy to save a log-factor is new and may have applications elsewhere.

2 Preliminaries

Notation. Let $[n] = \{1, \ldots, n\}$. For an n-dimensional vector space, the standard basis vectors are $\{e_i \in \{0,1\}^n \mid i \in [n]\}$, where e_i is the vector with a 1 in the ith coordinate and zeros elsewhere. For $x \in \{0,1\}^n$ and $i \in [n]$, let x^i be the input obtained by flipping the ith bit in x.

For a Boolean function $f : \{0,1\}^n \to \{-1,1\}$ and $B \in \mathbb{F}_2^{n\times n}$, define $f \circ B : \{0,1\}^n \to \{-1,1\}$ as $(f \circ B)(x) := f(Bx)$, where the matrix-vector product Bx is over $\mathbb{F}_2$. Throughout this paper, the rank of a matrix $B \in \mathbb{F}_2^{n\times n}$ will be taken over $\mathbb{F}_2$. Let $B_1, \ldots, B_n$ be the columns of B.

Fourier analysis on the Boolean cube. We introduce the basics of Fourier analysis here, referring to [25, 34] for more. Define the inner product between functions $f, g : \{0,1\}^n \to \mathbb{R}$ as

$$\langle f, g\rangle = \mathbb{E}_{x\in\{0,1\}^n}[f(x)\cdot g(x)],$$

where the expectation is uniform over all $x \in \{0,1\}^n$. For $S \in \{0,1\}^n$, the character function corresponding to S is given by $\chi_S(x) := (-1)^{S\cdot x}$, where the dot product $S \cdot x$ is $\sum_{i=1}^n S_i x_i$. For every $j \in [n]$, we use the notation χ_j to denote the function $\chi_{\{j\}}$. Observe that the set of functions $\{\chi_S\}_{S\in\{0,1\}^n}$ forms an orthonormal basis for the space of real-valued functions over the Boolean cube. Hence every $f : \{0,1\}^n \to \mathbb{R}$ can be written uniquely as

$$f(x) = \sum_{S\in\{0,1\}^n} \widehat{f}(S)(-1)^{S\cdot x} \quad \text{for all } x \in \{0,1\}^n,$$

where $\widehat{f}(S) = \langle f, \chi_S \rangle = \mathbb{E}_x[f(x)\chi_S(x)]$ is called a *Fourier coefficient* of f. For $i \in [n]$, we write $\widehat{f}(e_i)$ as $\widehat{f}(i)$ for notational convenience.

Parseval's identity states that $\sum_{S\in\{0,1\}^n} \widehat{f}(S)^2 = \mathbb{E}_x[f(x)^2]$. If f has range $\{-1,1\}$, then Parseval gives $\sum_{S\in\{0,1\}^n} \widehat{f}(S)^2 = 1$, so $\{\widehat{f}(S)^2\}_{S\in\{0,1\}^n}$ forms a probability distribution. The *Fourier weight* of function f on $\mathcal{S} \subseteq \{0,1\}^n$ is defined as $\sum_{S\in\mathcal{S}} \widehat{f}(S)^2$.

For $f : \{0,1\}^n \to \mathbb{R}$, the *Fourier support* of f is $\mathrm{supp}(\widehat{f}) = \{S : \widehat{f}(S) \neq 0\}$. The *Fourier sparsity* of f is $|\mathrm{supp}(\widehat{f})|$. The *Fourier span* of f, denoted $\mathrm{Fspan}(f)$, is the span of $\mathrm{supp}(\widehat{f})$. The *Fourier dimension* of f, denoted $\mathrm{Fdim}(f)$, is the dimension of the Fourier span. We say f is *k-Fourier-sparse* if $|\mathrm{supp}(\widehat{f})| \leq k$.

We now state a number of known structural results about Fourier coefficients and dimension.

Theorem 1 ([27])**.** *The Fourier dimension of a k-Fourier-sparse $f : \{0,1\}^n \to \{-1,1\}$ is $O(\sqrt{k}\log k)$.*[7]

Lemma 1 ([17, Theorem 12])**.** *Let $k \geq 2$. The Fourier coefficients of a k-Fourier-sparse Boolean function $f : \{0,1\}^n \to \{-1,1\}$ are integer multiples of $2^{1-\lfloor \log k \rfloor}$.*

Definition 1. *Let $f : \{0,1\}^n \to \{-1,1\}$ and suppose $B \in \mathbb{F}_2^{n\times n}$ is invertible. Define f_B as*

$$f_B(x) = f((B^{-1})^{\mathsf{T}}x).$$

Lemma 2. *Let $f : \{0,1\}^n \to \mathbb{R}$ and suppose $B \in \mathbb{F}_2^{n\times n}$ is invertible. Then the Fourier coefficients of f_B are $\widehat{f_B}(Q) = \widehat{f}(BQ)$ for all $Q \in \{0,1\}^n$.*

Proof. Write out the Fourier expansion of f_B:

$$f_B(x) = f((B^{-1})^{\mathsf{T}}x) = \sum_{S\in\{0,1\}^n} \widehat{f}(S)(-1)^{(B^{-1}S)\cdot x} = \sum_{Q\in\{0,1\}^n} \widehat{f}(BQ)(-1)^{Q\cdot x},$$

where the second equality used $\langle S, (B^{-1})^{\mathsf{T}}x\rangle = \langle B^{-1}S, x\rangle$ and the last used the substitution $S = BQ$. □

The following lemma (Lemma 3) easily follows by applying Lemma 2 with an invertible linear map B that maps e_i to B_i, for every $i \in [r]$.

Lemma 3. *Let $f : \{0,1\}^n \to \{-1,1\}$, and $B \in \mathbb{F}_2^{n\times n}$ be an invertible matrix such that the first r columns of B are a basis of the Fourier span of f, and $\widehat{f}(B_1), \ldots, \widehat{f}(B_r)$ are non-zero. Then*

1. *The Fourier span of $\widehat{f_B}$ is spanned by $\{e_1, \ldots, e_r\}$, i.e., f_B has only r influential variables.*
2. *For every $i \in [r]$, $\widehat{f_B}(i) \neq 0$.*

Here is the well-known fact, already mentioned in the introduction, that one can Fourier-sample from uniform quantum examples:

[7] Note that this theorem is optimal up to the logarithmic factor for the addressing function $\mathrm{Add}_m : \{0,1\}^{\log m+m} \to \{-1,1\}$ defined as $\mathrm{Add}_m(x,y) = 1 - 2y_x$ for all $x \in \{0,1\}^{\log m}$ and $y \in \{0,1\}^m$, i.e., the output of $\mathrm{Add}_m(x,y)$ is determined by the value y_x, where x is treated as the binary representation of a number in $\{0,\ldots,m-1\}$. For the Add_m function, the Fourier dimension is m and the Fourier sparsity is m^2.

Lemma 4. *Let $f : \{0,1\}^n \to \{-1,1\}$. There exists a procedure that uses one uniform quantum example and satisfies the following: with probability $1/2$ it outputs an S drawn from the distribution $\{\widehat{f}(S)^2\}_{S\in\{0,1\}^n}$, otherwise it rejects.*

Proof. Using a uniform quantum example $\frac{1}{\sqrt{2^n}}\sum_x |x, f(x)\rangle$, one can obtain $\frac{1}{\sqrt{2^n}}\sum_x f(x)|x\rangle$ with probability $1/2$: replace $f(x) \in \{-1,1\}$ by $(1-f(x))/2 \in \{0,1\}$ unitarily, apply the Hadamard transform to the last qubit and measure it. With probability $1/2$ we obtain the outcome 0, in which case our procedure rejects. Otherwise the remaining state is $\frac{1}{\sqrt{2^n}}\sum_x f(x)|x\rangle$. Apply Hadamard transforms to all n qubits to obtain $\sum_S \widehat{f}(S)|S\rangle$. Measuring this quantum state gives an S with probability $\widehat{f}(S)^2$. □

Information theory. We refer to [15] for a comprehensive introduction to classical information theory, and here just remind the reader of the basic definitions. A random variable $\mathbf{A}$ with probabilities $\Pr[\mathbf{A} = a] = p_a$ has *entropy* $H(\mathbf{A}) := -\sum_a p_a \log(p_a)$. For a pair of (possibly correlated) random variables $\mathbf{A}, \mathbf{B}$, the *conditional entropy* of $\mathbf{A}$ given $\mathbf{B}$, is $H(\mathbf{A} \mid \mathbf{B}) := H(\mathbf{A}, \mathbf{B}) - H(\mathbf{B})$. This equals $\mathbb{E}_{b\sim\mathbf{B}}[H(\mathbf{A} \mid \mathbf{B} = b)]$. The *mutual information* between $\mathbf{A}$ and $\mathbf{B}$ is $I(\mathbf{A} : \mathbf{B}) := H(\mathbf{A}) + H(\mathbf{B}) - H(\mathbf{A}, \mathbf{B}) = H(\mathbf{A}) - H(\mathbf{A} \mid \mathbf{B})$. The *binary entropy* $H(p)$ is the entropy of a bit with distribution $(p, 1-p)$. If ρ is a density matrix (i.e., a trace-1 positive semi-definite matrix), then its singular values form a probability distribution P, and the *von Neumann entropy* of ρ is $S(\rho) := H(P)$. We refer to [24, Part III] for a more extensive introduction to quantum information theory.

3 Exact learning of k-Fourier-sparse functions

In this section we consider exactly learning the concept class $\mathcal{C}$ of k-Fourier-sparse Boolean functions:

$$\mathcal{C} = \{f : \{0,1\}^n \to \{-1,1\} : |\mathrm{supp}(\widehat{f})| \leq k\}.$$

The goal is to exactly learn $c \in \mathcal{C}$ given *uniform examples* from c of the form $(x, c(x))$ where x is drawn from the uniform distribution on $\{0,1\}^n$. Haviv and Regev [21] considered learning this concept class and showed the following results.

Theorem 2 (Corollary 3.6 of [21])**.** *For every $n > 0$ and $k \leq 2^n$, the number of uniform examples that suffice to learn $\mathcal{C}$ with probability $1 - 2^{-\Omega(n \log k)}$ is $O(nk \log k)$.*

Theorem 3 (Theorem 3.7 of [21])**.** *For every $n > 0$ and $k \leq 2^n$, the number of uniform examples necessary to learn $\mathcal{C}$ with constant success probability is $\Omega(k(n - \log k))$.*

Our main results in this section are about the number of uniform *quantum* examples that are necessary and sufficient to exactly learn the class $\mathcal{C}$ of k-Fourier-sparse functions. A uniform quantum example for a concept $c \in \mathcal{C}$ is the quantum state

$$\frac{1}{\sqrt{2^n}} \sum_{x\in\{0,1\}^n} |x, c(x)\rangle.$$

Our first theorem of this section (Section 3.1) gives an upper bound on the number of uniform quantum examples that are sufficient to learn $\mathcal{C}$ by giving a learning algorithm.

Theorem 4. *For every $n > 0$ and $k \leq 2^n$, the number of uniform quantum examples that suffice to learn $\mathcal{C}$ with probability $\geq 2/3$ is $O(k^{1.5}(\log k)^2)$.*

The learning algorithm has two phases: Phase 1 is described in Section 3.1.1 and Phase 2 is discussed in Section 3.1.2.

In the theorem below (Section 3.2) we prove the following (non-matching) lower bound on the number of uniform quantum examples necessary to learn $\mathcal{C}$.

Theorem 5. *For every $n > 0$, constant $c \in (0,1)$ and $k \leq 2^{cn}$, the number of uniform quantum examples necessary to learn $\mathcal{C}$ with constant success probability is $\Omega(k \log k)$.*

In Section 3.3 we give a possible direction to prove an improved sample complexity for Phase 1 of our learning algorithm.

3.1 Upper bound on learning k-Fourier-sparse Boolean functions

We split our quantum learning algorithm into two phases. Suppose $c \in \mathcal{C}$ is the unknown concept, with Fourier dimension r. In the first phase the learner uses samples from the distribution $\{\widehat{c}(S)^2\}_{S \in \{0,1\}^n}$ to learn the Fourier span of c. In the second phase the learner uses uniform *classical* examples to learn c exactly, knowing its Fourier span. Phase 1 uses $O(rk)$ uniform quantum examples (for Fourier-sampling) and Phase 2 uses $O(rk \log k)$ uniform *classical* examples.

Theorem 6. *Let $k, r > 0$. There exists a quantum learner that exactly learns (with high probability) an unknown k-Fourier-sparse $c : \{0,1\}^n \to \{-1,1\}$ with Fourier dimension upper bounded by some known r, from $O(rk \log k)$ uniform quantum examples.*

The learner may not know the exact Fourier dimension r in advance, but Theorem 1 gives an upper bound $r = O(\sqrt{k} \log k)$, so our Theorem 4 follows immediately from Theorem 6.

Before we prove this Theorem 6, we first give a "trivial" algorithm for learning the Fourier support of Fourier-sparse functions quantumly. Gopalan et al. [17] showed that every k-Fourier-sparse Boolean function is "$2^{-\lceil \log k \rceil}$-granular", i.e., every Fourier coefficient of a k-Fourier-sparse Boolean function c is either 0 or an integer multiple of $2^{-\lceil \log k \rceil}$. Using this observation, if one is allowed to Fourier-sample from c, then each S with non-zero $\widehat{c}(S)$ will be observed with probability $\Omega(1/k^2)$, and using a coupon collector argument, we obtain the entire Fourier support using $O(k^2 \log k)$ many Fourier-samples. Our main contribution in Theorem 6 is to use the Fourier *dimension* in order to improve this trivial quantum algorithm. In particular observe that for functions with Fourier dimension $\log k$ (such as $(\log k)$-juntas), the theorem above scales as $O(k \log^2 k)$ which is better than the trivial algorithm by a factor of nearly k.

3.1.1 Phase 1: Learning the Fourier span

In this phase of the algorithm our goal is to learn the r-dimensional Fourier span of the k-Fourier-sparse target concept c, using $O(rk)$ Fourier-samples. The algorithm is very simple: Fourier-sample more and more S's and keep track of their span; stop when we reach dimension r. The key is the following technical lemma, which says that if our current span V' does not yet equal the full Fourier span V, then there is significant Fourier weight outside of V'. This implies that a small expected number of additional Fourier-samples will give us an $S \in V \setminus V'$, which will grow our current span. After r such grow-steps we have learned the full Fourier span.

Lemma 5. *Let $V \subseteq \{0,1\}^n$ be the r-dimensional Fourier span of k-Fourier-sparse function $c : \{0,1\}^n \to \{-1,1\}$, and $V' \subseteq V$ be a proper subspace. Then $\sum_{S \in V \setminus V'} \widehat{c}(S)^2 \geq 1/k$.*

Proof. Let us assume the worst case, which is that $\dim(V') = r - 1$. Because we can do an invertible linear transformation on c as in Lemma 2, we may assume without loss of generality that the one "missing" dimension corresponds to the variable x_r (i.e., $V = \mathrm{span}(V' \cup \{e_r\})$). Let g be the (not necessarily Boolean-valued) part of c with Fourier coefficients in V':

$$g(x) := \sum_{S \in V'} \widehat{c}(S)\chi_S(x).$$

Suppose, towards a contradiction, that the Fourier weight $W := \sum_{S \in V \setminus V'} \widehat{c}(S)^2$ is $< 1/k$. This implies that c and g have the same sign on every $x \in \{0,1\}^n$, as follows (using Cauchy-Schwarz):

$$|c(x) - g(x)| = \left| \sum_{S \in V \setminus V'} \widehat{c}(S)\chi_S(x) \right| \leq \sqrt{kW} < 1.$$

Since c depends on the variable x_r, there exists an $x \in \{0,1\}^n$ where x_r is influential, i.e., $c(x) \neq c(x^r)$. But g is independent of x_r, which implies $c(x) = \mathrm{sign}(g(x)) = \mathrm{sign}(g(x^r)) = c(x^r)$, a contradiction. Hence $W \geq 1/k$. □

We now conclude Phase 1 by presenting a quantum learning algorithm that learns the Fourier span of an unknown r-dimensional $c \in \mathcal{C}$, given uniform quantum examples for c.

Theorem 7. *Let $k, r > 0$. There exists a quantum learner that uses uniform quantum examples for an unknown k-Fourier-sparse $c : \{0,1\}^n \to \{-1,1\}$ with Fourier dimension r. After processing each new quantum example it outputs a subspace of the Fourier span of c. This sequence of subspaces is non-decreasing, and after an expected number of at most $2rk$ quantum examples, the output equals the Fourier span of c.*

This quantum learner can actually run forever, but if we know the Fourier dimension r of c, or an upper bound r on the actual Fourier dimension (e.g., by Theorem 1), then we can stop the learner after processing $6rk$ examples; now, by Markov's inequality, with probability $\geq 2/3$ the last subspace will be the Fourier span of c.

Proof. In order to learn the Fourier span of c, the quantum learner simply takes Fourier-samples until they span an r-dimensional space. Since we can generate a Fourier-sample from an expected number of 2 uniform quantum examples (by Lemma 4), the expected number of uniform quantum examples needed is at most twice the expected number of Fourier-samples. If our current sequence of Fourier-samples spans an r'-dimensional space V', with $r' < r$, then Lemma 5 implies that the next Fourier-sample has probability at least $1/k$ of yielding an $S \notin V'$. Hence an expected number of at most k Fourier-samples suffices to grow the dimension of V' by at least 1. Since we stop at dimension r, the overall expected number of Fourier-samples is at most $2rk$. □

3.1.2 Phase 2: Learning the function completely

In the above Phase 1, the quantum learner obtains the Fourier span of c, which we will denote by $\mathcal{T}$. Using this, the learner can restrict to the following concept class

$$\mathcal{C}' = \{c : \{0,1\}^n \to \{-1,1\} \mid c \text{ is } k\text{-Fourier-sparse with Fourier span } \mathcal{T}\}$$

Let $\dim(\mathcal{T}) = r$. Let $B \in \mathbb{F}_2^{n \times n}$ be an invertible matrix whose first r columns form a basis for $\mathcal{T}$. Consider $c_B = c \circ (B^{-1})^\mathsf{T}$ for $c \in \mathcal{C}'$. By Lemma 3 it follows that c_B depends on

only its first r bits, and we can write $c_B : \{0,1\}^r \to \{-1,1\}$. Hence the learner can apply the transformation $c \mapsto c \circ (B^{-1})^\mathsf{T}$ for every $c \in \mathcal{C}'$ and restrict to the concept class

$$\mathcal{C}'_r = \{c' : \{0,1\}^r \to \{-1,1\} \mid c' = c \circ (B^{-1})^\mathsf{T} \text{ for some } c \in \mathcal{C}' \text{ and invertible } B\}.$$

We now conclude Phase 2 of the algorithm by invoking the classical upper bound of Haviv-Regev (Theorem 2) which says that $O(rk \log k)$ uniform classical examples of the form $(z, c'(z)) \in \{0,1\}^{r+1}$ suffice to learn $\mathcal{C}'_r$. Although we assume our learning algorithm has access to uniform examples of the form $(x, c(x))$ for $x \in \{0,1\}^n$, the quantum learner knows B and hence can obtain a uniform example $(z, c'(z))$ for c' by letting z be the first r bits of $B^\mathsf{T}x$ and $c'(z) = c(x)$.

3.2 Lower bound on learning k-Fourier-sparse Boolean functions

In this section we show that $\Omega(k \log k)$ uniform quantum examples are necessary to learn the concept class of k-Fourier-sparse Boolean functions.

Theorem 8. *For every n, constant $c \in (0,1)$ and $k \leq 2^{cn}$, the number of uniform quantum examples necessary to learn the class of k-Fourier-sparse Boolean functions, with success probability $\geq 2/3$, is $\Omega(k \log k)$.*

Proof. Assume for simplicity that k is a power of 2, so $\log k$ is an integer. We prove the lower bound for the following concept class, which was also used for the classical lower bound of Haviv and Regev [21]: let $\mathcal{V}$ be the set of distinct subspaces in $\{0,1\}^n$ with dimension $n - \log k$ and

$$\mathcal{C} = \{c_V : \{0,1\}^n \to \{-1,1\} \mid c_V(x) = -1 \text{ iff } x \in V, \text{ where } V \in \mathcal{V}\}.$$

Note that every function in $\mathcal{C}$ has Fourier sparsity at most k, $|\mathcal{C}| = |\mathcal{V}|$, and each $c_V \in \mathcal{C}$ evaluates to 1 on a $(1 - 1/k)$-fraction of its domain.

We prove the lower bound for $\mathcal{C}$ using a three-step information-theoretic technique. A similar approach was used in proving classical and quantum PAC learning lower bounds in [4]. Let $\mathbf{A}$ be a random variable that is uniformly distributed over $\mathcal{C}$. Suppose $\mathbf{A} = c_V$, and let $\mathbf{B} = \mathbf{B}_1 \dots \mathbf{B}_T$ be T copies of the quantum example

$$|\psi_V\rangle = \frac{1}{\sqrt{2^n}} \sum_{x \in \{0,1\}^n} |x, c_V(x)\rangle$$

for c_V. The random variable $\mathbf{B}$ is a function of the random variable $\mathbf{A}$. The following upper and lower bounds on $I(\mathbf{A} : \mathbf{B})$ are similar to [4, proof of Theorem 12] and we omit the details of the first two steps here.

1. $I(\mathbf{A} : \mathbf{B}) \geq \Omega(\log |\mathcal{V}|)$ because $\mathbf{B}$ allows one to recover $\mathbf{A}$ with high probability.
2. $I(\mathbf{A} : \mathbf{B}) \leq T \cdot I(\mathbf{A} : \mathbf{B}_1)$ using a chain rule for mutual information.
3. $I(\mathbf{A} : \mathbf{B}_1) \leq O(n/k)$.
   Proof (of 3). Since $\mathbf{AB}$ is a classical-quantum state, we have

   $$I(\mathbf{A} : \mathbf{B}_1) = S(\mathbf{A}) + S(\mathbf{B}_1) - S(\mathbf{AB}_1) = S(\mathbf{B}_1),$$

   where the first equality is by definition and the second equality uses $S(\mathbf{A}) = \log |\mathcal{V}|$ since $\mathbf{A}$ is uniformly distributed over $\mathcal{C}$, and $S(\mathbf{AB}_1) = \log |\mathcal{V}|$ since the matrix

   $$\sigma = \frac{1}{|\mathcal{V}|} \sum_{V \in \mathcal{V}} |V\rangle\langle V| \otimes |\psi_V\rangle\langle\psi_V|$$

is block-diagonal with $|\mathcal{V}|$ rank-1 blocks on the diagonal. It thus suffices to bound the entropy of the (vector of singular values of the) reduced state of $\mathbf{B}_1$, which is

$$\rho = \frac{1}{|\mathcal{V}|} \sum_{V \in \mathcal{V}} |\psi_V\rangle\langle\psi_V|.$$

Let $\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{2^{n+1}-1} \geq 0$ be the singular values of ρ. Since ρ is a density matrix, these form a probability distribution. Now observe that $\sigma_0 \geq 1 - 1/k$ since the inner product between $\frac{1}{\sqrt{2^n}} \sum_{x \in \{0,1\}^n} |x, 1\rangle$ and every $|\psi_V\rangle$ is $1 - 1/k$. Let $\mathbf{N} \in \{0, 1, \ldots, 2^{n+1} - 1\}$ be a random variable with probabilities $\sigma_0, \sigma_1, \ldots, \sigma_{2^{n+1}-1}$, and $\mathbf{Z}$ an indicator for the event "$\mathbf{N} \neq 0$." Note that $\mathbf{Z} = 0$ with probability $\sigma_0 \geq 1 - 1/k$, and $H(\mathbf{N} \mid \mathbf{Z} = 0) = 0$. By a similar argument as in [4, Theorem 15], we have

$$\begin{aligned} S(\rho) = H(\mathbf{N}) &= H(\mathbf{N}, \mathbf{Z}) = H(\mathbf{Z}) + H(\mathbf{N} \mid \mathbf{Z}) \\ &= H(\sigma_0) + \sigma_0 \cdot H(\mathbf{N} \mid \mathbf{Z} = 0) + (1 - \sigma_0) \cdot H(\mathbf{N} \mid \mathbf{Z} = 1) \\ &\leq H\Big(\frac{1}{k}\Big) + \frac{n+1}{k} \leq O\Big(\frac{n + \log k}{k}\Big) \end{aligned}$$

using $H(\alpha) \leq O(\alpha \log(1/\alpha))$.

Combining these three steps implies $T = \Omega(k(\log |\mathcal{V}|)/n)$. It remains to lower bound $|\mathcal{V}|$.

Claim 1. *The number of distinct d-dimensional subspaces of $\mathbb{F}_2^n$ is at least $2^{\Omega((n-d)d)}$.*

Proof. We can specify a d-dimensional subspace by giving d linearly independent vectors in it. The number of distinct sequences of d linearly independent vectors is exactly $(2^n - 1)(2^n - 2)(2^n - 4) \cdots (2^n - 2^{d-1})$, because once we have the first t linearly independent vectors, with span $\mathcal{S}_t$, then there are $2^n - 2^t$ vectors that do not lie in $\mathcal{S}_t$.

However, we are double-counting certain subspaces in the argument above, since there will be multiple sequences of vectors yielding the same subspace. The number of sequences yielding a fixed d-dimensional subspace can be counted in a similar manner as above and we get $(2^d - 1)(2^d - 2)(2^d - 4) \cdots (2^d - 2^{d-1})$. So the total number of subspaces is

$$\frac{(2^n - 1)(2^n - 2) \cdots (2^n - 2^{d-1})}{(2^d - 1)(2^d - 2) \cdots (2^d - 2^{d-1})} \geq \frac{(2^n - 2^{d-1})^d}{(2^d - 1)^d} \geq 2^{\Omega((n-d)d)}.$$

□

Combining this claim (with $d = n - \log k$) and $T = \Omega(k(\log |\mathcal{V}|)/n)$ gives $T = \Omega(k \log k)$. □

3.3 A potential direction to prove an improved sample complexity for Phase 1

In this section we give a potential direction to prove that in expectation $\tilde{O}(k)$ Fourier-samples are sufficient for Phase 1 of our learning algorithm presented in Section 3.1.1. Recall Phase 1 of our learning algorithm. Given a k-Fourier-sparse function c, Phase 1 starts by finding an $S \in \mathrm{supp}(\widehat{c})$ such that $S \neq 0^n$. Lemma 5 implies that an expected number of $O(k)$ many Fourier-samples are sufficient to sample such an S. Chang's lemma, a central result in additive combinatorics, gives tighter bound on the expected number of samples for this step. Chang's lemma upper bounds the dimension of the span of the "large" Fourier coefficients.

Lemma 6 (Chang's lemma)**.** *Let $\alpha \in (0,1)$ and $\rho > 0$. For every $f : \{0,1\}^n \to \{-1,1\}$ that satisfies $\widehat{f}(0^n) = 1 - 2\alpha$, we have*

$$\dim(\mathrm{span}\{S : |\widehat{f}(S)| \geq \rho\alpha\}) \leq \frac{2\log(1/\alpha)}{\rho^2}. \tag{2}$$

Let us consider Chang's lemma for a k-Fourier-sparse Boolean function $c : \{0,1\}^n \to \{-1,1\}$ of Fourier dimension r and let $\rho \in (0,1]$. In particular, consider the case $\rho\alpha = 1/k$. In this case, since all elements of the Fourier support satisfy $|\widehat{c}(S)| \geq 1/k$ by Lemma 1, the left-hand side of Eq. (2) equals the Fourier dimension r of c. Thus Chang's lemma gives

$$r \leq 2\alpha^2 k^2 \log \rho k \leq 2\alpha^2 k^2 \log k,$$

which implies

$$\sum_{S \neq 0^n} \widehat{c}(S)^2 = \Omega\left(\frac{\sqrt{r}}{k\sqrt{\log k}}\right). \tag{3}$$

Thus an expected number of $O((k\sqrt{\log k})/\sqrt{r})$ many Fourier-samples are sufficient to obtain an $S \in \mathrm{supp}(\widehat{c})$ such that $S \neq 0^n$ in Phase 1. This is already an improvement from what Lemma 5 guaranteed.

In this section we give an improvement of Chang's lemma for k-Fourier-sparse Boolean functions:

Theorem 9. *Let $\alpha \in (0,1)$ and $k \geq 2$. For every k-Fourier-sparse $f : \{0,1\}^n \to \{-1,1\}$ that satisfies $\widehat{f}(0^n) = 1 - 2\alpha$ and $\mathrm{Fdim}(f) = r$, we have*

$$\widehat{f}(0^n) \leq 1 - \frac{r}{k \log k}.$$

We remark that in a follow-up paper [12], a subset of the authors gave a refinement of the theorem above.

Before giving a proof of Theorem 9, let us first discuss how this theorem improves the analysis of Phase 1 of our learning algorithm. Theorem 9 implies that for a k-Fourier-sparse Boolean function $c : \{0,1\}^n \to \{-1,1\}$ of Fourier dimension r,

$$\sum_{S : S \neq 0^n} \widehat{c}(S)^2 = \Omega(r/(k \log k)).$$

This is a better lower bound on the Fourier weight of c on the set $\{0,1\}^n \setminus \{0^n\}$ than that obtained from Chang's lemma (Equation 3). Thus an expected number of $O((k \log k)/r)$ many uniformly quantum samples is sufficient to obtain an $S \in \mathrm{supp}(\widehat{c})$ such that $S \neq 0^n$.

We suspect that Theorem 9 can in fact lead to an $\tilde{O}(k)$ learning algorithm for Phase 1. Towards that end we make the following conjecture which can be viewed as a generalization of Theorem 9.

Conjecture 1. *Let $n > 0$ and $1 \leq k \leq 2^n$. For every k-Fourier-sparse $f : \{0,1\}^n \to \{-1,1\}$ with Fourier span $\mathcal{V}$ and Fourier dimension r, the following holds: for every $r' > 0$ and $\mathcal{S} \subset \mathcal{V}$ satisfying $\dim(\mathrm{span}(\mathcal{S})) = r'$, we have*

$$\sum_{S \in \mathrm{span}(\mathcal{S})} \widehat{f}(S)^2 \leq 1 - \frac{r - r'}{k \log k}.$$

If the above conjecture is true then it would imply an $\tilde{O}(k)$ learning algorithm for Phase 1. Let $c : \{0,1\}^n \to \{-1,1\}$ be a k-Fourier-sparse function of Fourier dimension r. Assuming Conjecture 1 to be true we have

$$\sum_{S \notin \mathrm{span}(\mathcal{S})} \widehat{c}(S)^2 \geq \frac{r - r'}{k \log k}.$$

So the expected number of samples to increase the dimension by 1 is $\leq \frac{k \log k}{r - r'}$. Accordingly, the expected number of Fourier-samples needed to learn the whole Fourier span of f is at most

$$\sum_{i=1}^{r} \frac{k \log k}{i} \leq O(k \log k \log r),$$

where the final inequality used $\sum_{i=1}^{r} \frac{1}{i} = O(\log r)$. We now proceed to the proof of Theorem 9.

3.3.1 Proof of Theorem 9

We first define the following notation. For $U \subseteq [r]$, let $f^{(U)}$ be the function obtained by fixing the variables $\{x_i\}_{i \in U}$ in f to $x_i = (1 + \mathrm{sign}(\widehat{f}(i)))/2$ for all $i \in U$. Note that fixing variables cannot increase Fourier sparsity. For $i, j \in [r]$, define $f^{(i)} = f^{(\{i\})}$ and $f^{(ij)} = f^{(\{i,j\})}$. In this proof, for an invertible matrix $B \in \mathbb{F}_2^{n \times n}$, we will often treat its columns as a basis for the space $\mathbb{F}_2^n$. Recall $f_B(x) = f((B^{-1})^T x)$ from Definition 1. We let $f_B^{(i)}$ be the function obtained by fixing $x_i = (1 + \mathrm{sign}(\widehat{f}(i)))/2$ in the function f_B.

The core idea in the proof of the theorem is the following structural lemma, which says that there is a particular x_i that we can fix in the function f_B without decreasing the Fourier dimension very much.

Lemma 7. *For every k-Fourier-sparse Boolean function $f : \{0,1\}^n \to \{-1,1\}$ with $\mathrm{Fdim}(f) = r$, there exists an invertible matrix $B \in \mathbb{F}_2^{n \times n}$ and an index $i \in [r]$ such that $\mathrm{Fdim}(f_B^{(i)}) \geq r - \log k$ and $\widehat{f_B}(j) \neq 0$ for all $j \in [r]$.*

We defer the proof of the lemma to later and first conclude the proof of the theorem assuming the lemma. Consider the matrix B defined in Lemma 7. Using Lemma 3 it follows that f_B has only r influential variables, so we can write $f_B : \{0,1\}^r \to \{-1,1\}$, where $\widehat{f_B}(j) \neq 0$ for every $j \in [r]$. Also, $\widehat{f_B}(0^r) = \widehat{f}(0^n) = 1 - 2\alpha$. For convenience, we abuse notation and abbreviate $f = f_B$. It remains to show that for every $f : \{0,1\}^r \to \{-1,1\}$ with $\widehat{f}(j) \neq 0$ for all $j \in [r]$, we have $2\alpha = 1 - \widehat{f}(0^r) \geq r/(k \log k)$. We prove this by induction on r.

Base case. Let $r = 1$. Then $k = 2$ (since $r \geq \log k$ and $k \geq 2$ by assumption). Note that the only Boolean functions with Fourier dimension 1 and $|\mathrm{supp}(\widehat{f})| \leq 2$ are $\{\chi_j, -\chi_j\}$, where $\chi_j = (-1)^{x_j}$, for $j \in [n]$. In both these cases $1 - \widehat{f}(0^r) = 1$ and $r/(k \log k) = 1/2$ (although the Fourier sparsity of χ_j is 1, we are implicitly working with a concept class of 2-sparse Boolean functions, hence $k = 2$).

Induction hypothesis. Suppose that for all $p \in \{1, \ldots, r-1\}$ and k-Fourier-sparse Boolean function $g : \{0,1\}^p \to \{-1,1\}$ with $\mathrm{Fdim}(g) = p$ and $\widehat{g}(j) \neq 0$ for all $j \in [p]$, we have $1 - \widehat{g}(0^p) \geq p/(k \log k)$.

Induction step. Let $i \in [r]$ be the index from Lemma 7. Note that $f^{(i)}$ is still k-Fourier-sparse and $\widehat{f^{(i)}}(0^{r-1}) = 1 - 2\alpha + |\widehat{f}(i)|$. Since $|\widehat{f}(i)| \geq 1/k$ (by Lemma 1), we have

$$\widehat{f^{(i)}}(0^{r-1}) \geq 1 - 2\alpha + 1/k.$$

Since $r - \log k \leq \mathrm{Fdim}(f^{(i)}) \leq r-1$, we can use the induction hypothesis on the function $f^{(i)}$ to conclude that

$$2\alpha \geq 1 - \widehat{f^{(i)}}(0^{r-1}) + \frac{1}{k} \geq \frac{r - \log k}{k \log k} + \frac{1}{k} = \frac{r}{k \log k}.$$

This concludes the proof of the induction step and the theorem. We now prove Lemma 7.

Proof of Lemma 7. In order to construct B as in the lemma statement, we first make the following observation.

Observation 1. *For every Boolean function $f : \{0,1\}^n \to \{-1,1\}$ with $\mathrm{Fdim}(f) = r$, there exists an invertible $B \in \mathbb{F}_2^{n \times n}$ such that:*

1. *The Fourier coefficient $\widehat{f_B}(1)$ is non-zero.*
2. *There exists a $t \in [r]$ such that, for all $j \in \{2, \ldots, t\}$, we have $\mathrm{Fdim}(f_B^{(j)}) \leq r - t$.*
3. *The Fourier span of $f_B^{(1)}$ is spanned by $\{e_{t+1}, \ldots, e_r\}$.*
4. *For $\ell \in \{t+1, \ldots, r\}$, the Fourier coefficients $\widehat{f_B^{(1)}}(\ell)$ are non-zero.*

We defer the proof of this observation to the end. We proceed to prove the lemma assuming the observation. Note that Property 3 gives the following simple corollary:

Corollary 1. *$f_B^{(1)}$ is a function of $x_{t+1}, \ldots, x_r$ and independent of $x_2, \ldots, x_t$ (and hence $f_B^{(1)} = f_B^{(i1)} = f_B^{(1i)}$ for every $i \in \{2, \ldots, t\}$).*

We now show that not only $f_B^{(1)}$, but all the functions $f_B^{(2)}, \ldots, f_B^{(t)}$ are independent of $x_2, \ldots, x_t$.

Claim 2. *For all $i \in \{2, \ldots, t\}$, $f_B^{(i)}$ is a function of $\{x_1, x_{t+1}, \ldots, x_r\}$ and independent of $x_2, \ldots, x_t$.*

Proof. Without loss of generality, let $i = 2$. By Observation 1 (property 4), the character functions $\chi_{t+1}, \ldots, \chi_r$ are present in the Fourier expansion of $f_B^{(1)}$. We have $f_B^{(21)} = f_B^{(1)}$ by Corollary 1. Hence, for every $\ell \in \{t+1, \ldots, r\}$, at least one of the characters χ_ℓ or $\chi_1\chi_\ell$ is present in the Fourier expansion of $f_B^{(2)}$. Let y_ℓ be χ_ℓ or $\chi_1\chi_\ell$ (depending on which character function is present in the Fourier expansion of $f_B^{(2)}$). Note that the $r - t$ character functions $y_{t+1}, \ldots, y_r$ are linearly independent. By Observation 1 (Property 2), we have $\mathrm{Fdim}(f_B^{(2)}) \leq r - t$, which implies $\mathrm{Fspan}(f_B^{(2)}) \subseteq \mathrm{span}\{y_{t+1}, \ldots, y_r\}$ and $f_B^{(2)}$ is independent of $\{x_2, \ldots, x_t\}$. The same argument shows that for every $i, k \in \{2, \ldots, t\}$, $f_B^{(i)}$ is independent of x_k. □

Claim 3. *There exists an assignment of $(x_1, x_{t+1}, \ldots, x_r)$ to $(a_1, a_{t+1} \ldots, a_r)$ in f_B such that the resulting function depends on* all *variables $x_2, \ldots, x_t$.*[8]

[8]Observe that in this assignment, we have $x_1 = (1 - \mathrm{sign}(\widehat{f}(1)))/2$. Otherwise, by assigning $x_1 = (1 + \mathrm{sign}(\widehat{f}(1)))/2$ in f_B, we would obtain the function $f_B^{(1)}$ which we know is independent of $\{x_2, \ldots, x_t\}$ by Corollary 1.

Proof. Before proving the claim we first make the following observation. Let us consider an assignment of $(x_1, x_{t+1}, \ldots, x_r) = z$ in f_B and assume that the resulting function $f_{B,z}$ is independent of x_i for some $i \in \{2, \ldots, t\}$. Let us assign $x_i = (1 + \text{sign}(\widehat{f_B}(i)))/2$ in $f_{B,z}$ and call the resulting function $f^{(i)}_{B,z}$. Firstly, $f^{(i)}_{B,z} = f_{B,z}$ since $f_{B,z}$ was independent of x_i. Secondly, observe that $f_{B,z} = f^{(i)}_{B,z}$ could have alternatively been obtained by *first* fixing $x_i = (1 + \text{sign}(\widehat{f}(i)))/2$ in f_B and then fixing $(x_1, x_{t+1} \ldots, x_r) = z$. In this case, by Claim 2, after fixing x_i in f_B, $f^{(i)}_B$ is independent of $x_2, \ldots, x_t$ and after fixing $(x_1, x_{t+1}, \ldots, x_r) = z$, $f_{B,z}$ is a constant. This in particular shows that if there exists a z such that $f_{B,z}$ is independent of x_i for some $i \in \{2, \ldots, t\}$, then $f_{B,z}$ is also independent of $x_2, \ldots, x_t$.

Towards a contradiction, suppose that for every assignment of $(x_1, x_{t+1}, \ldots, x_r) = z$ to f_B, the resulting function $f_{B,z}$ is independent of x_i, for some $i \in \{2, \ldots, t\}$. Then by the argument in the previous paragraph, for every assignment z, $f_{B,z}$ is also independent of x_k for *every* $k \in \{2, \ldots, t\}$. This, however, contradicts the fact that $x_2, \ldots, x_t$ had non-zero influence on f_B (since B was chosen such that $\widehat{f_B}(j) \neq 0$ for every $j \in [r]$ in Lemma 7). This implies the existence of an assignment $(x_1, x_{t+1}, \ldots, x_r) = (a_1, a_{t+1} \ldots, a_r)$, such that the resulting function depends on all the variables $x_2, \ldots, x_t$. □

We now argue that the assignment in Claim 3 results in a function which resembles the AND function on $x_2, \ldots, x_t$, and hence has Fourier sparsity 2^{t-1}.

Claim 4. *Consider the assignment $(x_1, x_{t+1}, \ldots, x_r) = (a_1, a_{t+1} \ldots, a_r)$ in f_B as in Claim 3, then the resulting function g equals (up to possible negations of input and output bits) the $(t-1)$-bit* AND *function.*

Proof. By Claim 3, g depends on all the variables $x_2, \ldots, x_t$. This dependence is such that if *any one* of the variables $\{x_i : i \in \{2, \ldots, t\}\}$ is set to $x_i = (1 + \text{sign}(\widehat{f_B}(i)))/2$, then by Claim 2 the resulting function $g^{(i)}$ is independent of $x_2, \ldots, x_t$. Hence, $g^{(i)}$ is some constant $b_i \in \{-1, 1\}$ for every $i \in \{2, \ldots, t\}$. Note that these b_is are all the same bit b, because first fixing x_i (which collapses g to the constant b_i) and then x_j gives the same function as first fixing x_j (which collapses g to b_j) and then x_i. Additionally, by assigning $x_i = (1 - \text{sign}(\widehat{f_B}(i)))/2$ for every $i \in \{2, \ldots, t\}$ in g, the resulting function must evaluate to $1 - b$ because g is non-constant (it depends on $x_2, \ldots, x_t$). Therefore g equals (up to possible negations of input and output bits) the $(t-1)$-bit AND function. □

We now conclude the proof of Lemma 7. Let $f : \{0,1\}^n \to \{-1, 1\}$ be such that $\text{Fdim}(f) = r$. Let B be as defined in Observation 1. Consider the assignment of $(x_{t+1}, \ldots, x_r) = (a_{t+1}, \ldots, a_r)$ to f_B as in Claim 4, and call the resulting function f'_B. From Claim 4, observe that by setting $x_1 = a_1$ in f'_B, the resulting function is $g(x_2, \ldots, x_t)$ and by setting $x_1 = 1 - a_1$ in f'_B, the resulting function is a constant. Hence f'_B can be written as

$$f'_B(x_1, \ldots, x_t, a_{t+1}, \ldots, a_r) = \frac{1 - (-1)^{x_1 + a_1}}{2} b_{a_1, a_{t+1}, \ldots, a_r} + \frac{1 + (-1)^{x_1 + a_1}}{2} g(x_2, \ldots, x_t), \tag{4}$$

where $b_{a_1, a_{t+1}, \ldots, a_r} \in \{-1, 1\}$ (note that it is independent of $x_2, \ldots, x_t$ by Corollary 1). Since g essentially equals the $(t-1)$-bit AND function (by Claim 4), g has Fourier sparsity 2^{t-1} and $\widehat{g}(0^{t-1}) = 1 - 2^{-t+2}$. Hence the Fourier sparsity of f'_B in Eq. (4) equals 2^t. Since f'_B was a restriction of f_B, the Fourier sparsity of f'_B is at most k, hence $t \leq \log k$. This implies $\text{Fdim}(f^{(1)}_B) = r - t \geq r - \log k$, concluding the proof. □

It remains to prove Observation 1, which we do now.

Proof of Observation 1. Let $D \in \mathbb{F}_2^{n\times n}$ be an invertible matrix that maximizes $\mathrm{Fdim}(f_D^{(1)})$ subject to the constraint $\widehat{f_D}(1) \neq 0$. Suppose $\mathrm{Fdim}(f_D^{(1)}) = r - t$. Let $d_1, \ldots, d_{r-t}$ be a basis of $\mathrm{Fspan}(f_D^{(1)})$ such that $\widehat{f_D^{(1)}}(d_i) \neq 0$ for all $i \in [r-t]$. We now construct an invertible $C \in \mathbb{F}_2^{n\times n}$ whose first r columns form a basis for $\mathrm{Fspan}(f_D)$, as follows: let $c_1 = e_1$, and for $i \in [r-t]$, fix $c_{t+i} = d_i$. Next, assign vectors $c_2, \ldots, c_t$ arbitrarily from $\mathrm{Fspan}(f_D)$, ensuring that $c_2, \ldots, c_t$ are linearly independent from $\{c_1, c_{t+1}, \ldots, c_r\}$. We then extend to a basis $\{c_1, \ldots, c_n\}$ arbitrarily. Define C as $C = [c_1, \ldots, c_n]$ (where the c_is are column vectors). Finally, define our desired matrix B as the product $B = DC$. We now verify the properties of B.

Property 1: Using Lemma 3 we have

$$\widehat{f_{DC}}(1) = \widehat{f_D}(Ce_1) = \widehat{f_D}(c_1) = \widehat{f_D}(1) \neq 0,$$

where the third equality used $c_1 = e_1$, and $\widehat{f_D}(1) \neq 0$ follows from the definition of D.

We next prove the following fact, which we use to verify the remaining three properties.

Fact 1. *Let C, D be invertible matrices as defined above. For every $i \in [t]$, let $(f_D^{(i)})_C$ be the function obtained after applying the invertible transformation C to $f_D^{(i)}$ and $(f_{DC})^{(i)}$ be the function obtained after fixing x_i to $(1 + \mathrm{sign}(\widehat{f_{DC}}(i)))/2$ in f_{DC}. Then $(f_{DC})^{(i)} = (f_D^{(i)})_C$.*

Property 2: Fact 1 implies that $\mathrm{Fdim}((f_{DC})^{(i)}) = \mathrm{Fdim}((f_D^{(i)})_C)$. Since C is invertible, $\mathrm{Fdim}((f_D^{(i)})_C) = \mathrm{Fdim}(f_D^{(i)})$. From the choice of D, observe that for all $i \in \{2, \ldots, t\}$,

$$\mathrm{Fdim}(f_B^{(i)}) = \mathrm{Fdim}(f_{DC}^{(i)}) = \mathrm{Fdim}((f_D^{(i)})_C) = \mathrm{Fdim}(f_D^{(i)}) \leq \mathrm{Fdim}(f_D^{(1)}) = r - t,$$

where the inequality follows by definition of D.

Property 3: Note that $\mathrm{Fspan}(f_D^{(1)})$ is contained in $\mathrm{span}\{d_1, \ldots, d_{r-t}\}$ by construction. By making the invertible transformation by C, observe that $\mathrm{Fspan}((f_D^{(1)})_C) \subseteq \mathrm{span}\{e_{t+1}, \ldots, e_r\}$ (since for all $i \in [r-t]$, we defined $c_{t+i} = d_i$). Property 3 follows because $(f_D^{(1)})_C = f_{DC}^{(1)} = f_B^{(1)}$ by Fact 1.

Property 4: Using Fact 1, for every $\ell \in \{t+1, \ldots, r\}$, we have

$$\widehat{(f_B)^{(1)}}(\ell) = \widehat{(f_{DC})^{(1)}}(\ell) = \widehat{(f_D^{(1)})_C}(\ell) = \widehat{f_D^{(1)}}(c_\ell).$$

Since $c_\ell = d_{\ell-t}$, we have $\widehat{f_D^{(1)}}(c_\ell) = \widehat{f_D^{(1)}}(d_{\ell-t})$ and $\widehat{f_D^{(1)}}(d_1), \ldots, \widehat{f_D^{(1)}}(d_{r-t}) \neq 0$ by definition of d_i, hence the property follows.

Proof of Fact 1. Let $f_D = g$. We want to show that $(g^{(i)})_C = (g_C)^{(i)}$. For simplicity fix $i = 1$; the same proof works for every $i \in [t]$. Then,

$$(g^{(1)})(x) = \sum_{S\in\{0\}\times\{0,1\}^{n-1}} (\widehat{g}(S) + \widehat{g}(S \oplus e_1))\chi_S(x).$$

On transforming $g^{(1)}$ using the basis C we have:

$$(g^{(1)})_C(x) = \sum_{S\in\{0\}\times\{0,1\}^{n-1}} (\widehat{g}(CS) + \widehat{g}(C(S \oplus e_1))\chi_S(x). \tag{5}$$

Consider the function g_C. The Fourier expansion of g_C is $g_C(y) = \sum_{S\in\{0,1\}^n} \widehat{g}(CS)\chi_S(y)$ and the Fourier expansion of the $(g_C)^{(1)}$ can be written as

$$g_C^{(1)}(y) = \sum_{S\in\{0\}\times\{0,1\}^{n-1}} (\widehat{g}(CS) + \widehat{g}(CS \oplus Ce_1))\chi_S(y). \tag{6}$$

Using Eq. (5), (6), we conclude that $(g^{(1)})_C = (g_C)^{(1)}$, concluding the proof of the fact. □

This concludes the proof of the observation. □

This concludes the proof of the theorem.

4 Quantum vs classical membership queries

In this section we assume we can access the target function using membership queries rather than examples. Our goal is to simulate quantum exact learners for a concept class $\mathcal{C}$ by classical exact learners, without using many more membership queries. A key tool here will be the ("nonnegative" or "positive-weights") adversary method. This was introduced by Ambainis [2]; here we will use the formulation of Barnum et al. [6], which is called the "spectral adversary" in the survey [30].

Let $\mathcal{C} \subseteq \{0,1\}^N$ be a set of strings. If $N = 2^n$ then we may view such a string $c \in \mathcal{C}$ as (the truth-table of) an n-bit Boolean function, but in this section we do not need the additional structure of functions on the Boolean cube and may consider any positive integer N. Suppose we want to identify an unknown $c \in \mathcal{C}$ with success probability at least $2/3$ (i.e., we want to compute the identity function on $\mathcal{C}$). The required number of quantum queries to c can be lower bounded as follows. Let Γ be a $|\mathcal{C}| \times |\mathcal{C}|$ matrix with real, nonnegative entries and 0s on the diagonal (called an "adversary matrix"). Let D_i denote the $|\mathcal{C}| \times |\mathcal{C}|$ 0/1-matrix whose (c,c')-entry is $[c_i \neq c'_i]$.[9] Then it is known that at least (a constant factor times) $\| \Gamma \| / \max_{i\in[N]} \| \Gamma \circ D_i \|$ quantum queries are needed, where $\| \cdot \|$ denotes operator norm (largest singular value) and '$\circ$' denotes entrywise product of matrices. Let

$$\mathrm{ADV}(\mathcal{C}) = \max_{\Gamma \geq 0} \frac{\| \Gamma \|}{\max_{i\in[N]} \| \Gamma \circ D_i \|}$$

denote the best-possible lower bound on $Q(\mathcal{C})$ that can be achieved this way.

The key to our classical simulation is the next lemma. It shows that if $Q(\mathcal{C})$ (and hence $\mathrm{ADV}(\mathcal{C})$) is small, then there is a query that splits the concept class in a "mildly balanced" way.

Lemma 8. *Let $\mathcal{C} \subseteq \{0,1\}^N$ be a concept class and*

$$\mathrm{ADV}(\mathcal{C}) = \max_{\Gamma \geq 0} \frac{\| \Gamma \|}{\max_{i\in[N]} \| \Gamma \circ D_i \|}$$

be the nonnegative adversary bound for the exact learning problem corresponding to $\mathcal{C}$. Let μ be a distribution on $\mathcal{C}$ such that $\max_{c\in\mathcal{C}} \mu(c) \leq 5/6$. Then there exists an $i \in [N]$ such that

$$\min(\mu(C_i = 0), \mu(C_i = 1)) \geq \frac{1}{36\mathrm{ADV}(\mathcal{C})^2}.$$

[9] The bracket-notation $[P]$ denotes the truth-value of proposition P.

Proof. Define unit vector $v \in \mathbb{R}_+^{|\mathcal{C}|}$ by $v_c = \sqrt{\mu(c)}$, and adversary matrix

$$\Gamma = vv^* - \mathrm{diag}(\mu),$$

where $\mathrm{diag}(\mu)$ is the diagonal matrix that has the entries of μ on its diagonal. This Γ is a nonnegative matrix with 0 diagonal (and hence a valid adversary matrix for the exact learning problem), and $\| \Gamma \| \geq \| vv^* \| - \| \mathrm{diag}(\mu) \| \geq 1 - 5/6 = 1/6$. Abbreviate $A = \mathrm{ADV}(\mathcal{C})$. By definition of A, we have for this particular Γ

$$A \geq \frac{\| \Gamma \|}{\max_i \| \Gamma \circ D_i \|} \geq \frac{1}{6 \max_i \| \Gamma \circ D_i \|},$$

hence there exists an $i \in [N]$ such that $\| \Gamma \circ D_i \| \geq \frac{1}{6A}$. We can write $v = \begin{pmatrix} v_0 \\ v_1 \end{pmatrix}$ where the entries of v_0 are the ones corresponding to Cs where $C_i = 0$, and the entries of v_1 are the ones where $C_i = 1$. Then

$$\Gamma = \begin{pmatrix} v_0 v_0^* & v_0 v_1^* \\ v_1 v_0^* & v_1 v_1^* \end{pmatrix} - \mathrm{diag}(\mu) \quad \text{and} \quad \Gamma \circ D_i = \begin{pmatrix} 0 & v_0 v_1^* \\ v_1 v_0^* & 0 \end{pmatrix}.$$

It is easy to see that $\| \Gamma \circ D_i \| = \| v_0 \| \cdot \| v_1 \|$. Hence

$$\frac{1}{36A^2} \leq \| \Gamma \circ D_i \|^2 = \| v_0 \|^2 \| v_1 \|^2 = \mu(C_i = 0)\mu(C_i = 1) \leq \min(\mu(C_i = 0), \mu(C_i = 1)),$$

where the last inequality used $\max(\mu(C_i = 0), \mu(C_i = 1)) \leq 1$. □

Note that if we query the index i given by this lemma and remove from $\mathcal{C}$ the strings that are inconsistent with the query outcome, then we reduce the size of $\mathcal{C}$ by a factor $\leq 1 - \Omega(1/\mathrm{ADV}(\mathcal{C})^2)$. Repeating this $O(\mathrm{ADV}(\mathcal{C})^2 \log |\mathcal{C}|)$ times would reduce the size of $\mathcal{C}$ to 1, completing the learning task. However, we will see below that analyzing the same approach in terms of entropy gives a somewhat better upper bound on the number of queries.

Theorem 10. *Let $\mathcal{C} \subseteq \{0,1\}^N$ be a concept class and*

$$\mathrm{ADV}(\mathcal{C}) = \max_{\Gamma \geq 0} \frac{\| \Gamma \|}{\max_{i \in [N]} \| \Gamma \circ D_i \|}$$

be the nonnegative adversary bound for the exact learning problem corresponding to $\mathcal{C}$. Then there exists a classical learner for $\mathcal{C}$ using $O\left(\frac{\mathrm{ADV}(\mathcal{C})^2}{\log \mathrm{ADV}(\mathcal{C})} \log |\mathcal{C}|\right)$ membership queries that identifies the target concept with probability $\geq 2/3$.

Proof. Fix an arbitrary distribution μ on $\mathcal{C}$. We will construct a deterministic classical learner for $\mathcal{C}$ with success probability $\geq 2/3$ under μ. Since we can do this for every μ, the "Yao principle" [35] then implies the existence of a randomized learner that has success probability $\geq 2/3$ for every $c \in \mathcal{C}$.

Consider the following algorithm, whose input is an N-bit random variable $C \sim \mu$:

1. Choose an i that maximizes $H(C_i)$ and query that i.[10]

[10] Querying this i will give a fairly "balanced" reduction of the size of $\mathcal{C}$ irrespective of the outcome of the query. If there are several maximizing is, then choose the smallest i to make the algorithm deterministic.

2. Update $\mathcal{C}$ and μ by restricting to the concepts that are consistent with the query outcome.

3. Goto 1.

The queried indices are themselves random variables, and we denote them by $I_1, I_2, \ldots$. We can think of t steps of this algorithm as generating a binary tree of depth t, where the different paths correspond to the different queries made and their binary outcomes.

Let P_t be the probability that, after t queries, our algorithm has reduced μ to a distribution that has weight $\geq 5/6$ on one particular c:

$$P_t = \sum_{\substack{i_1,\ldots,i_t \in [N] \\ b \in \{0,1\}^t}} \Pr[I_1 = i_1, \ldots, I_t = i_t, C_{i_1} \ldots C_{i_t} = b] \cdot \left[\exists c \in \mathcal{C} \text{ s.t. } \mu(c \mid C_{i_1} \ldots C_{i_t} = b) \geq \frac{5}{6} \right].$$

Because restricting μ to a subset $\mathcal{C}' \subseteq \mathcal{C}$ cannot decrease probabilities of individual $c \in \mathcal{C}'$, this probability P_t is non-decreasing in t. Because N queries give us the target concept completely, we have $P_N = 1$. Let T be the smallest integer t for which $P_t \geq 5/6$. We will run our algorithm for T queries, and then output the c with highest probability under the restricted version of μ we now have. With μ-probability at least $5/6$, that c will have probability at least $5/6$ (under μ conditioned on the query-results). The overall error probability under μ is therefore $\leq 1/6 + 1/6 = 1/3$.

It remains to upper bound T. To this end, define the following "energy function" in terms of conditional entropy:

$$\begin{aligned} E_t &= H(C \mid C_{I_1}, \ldots, C_{I_t}) \\ &= \sum_{\substack{i_1,\ldots,i_t \in [N] \\ b \in \{0,1\}^t}} \Pr[I_1 = i_1, \ldots, I_t = i_t, C_{i_1} \ldots C_{i_t} = b] \cdot H(C \mid C_{i_1} \ldots C_{i_t} = b). \end{aligned}$$

Because conditioning on a random variable cannot increase entropy, E_t is non-increasing in t. We will show below that as long as $P_t < 5/6$, the energy shrinks significantly with each new query.

Let $C_{i_1} \ldots C_{i_t} = b$ be such that there is no $c \in \mathcal{C}$ s.t. $\mu(c \mid C_{i_1} \ldots C_{i_t} = b) \geq 5/6$ (note that this event happens in our algorithm with μ-probability $1 - P_t$). Let μ' be μ restricted to the class $\mathcal{C}'$ of concepts c where $c_{i_1} \ldots c_{i_t} = b$. The nonnegative adversary bound for this restricted concept class is $A' = \mathrm{ADV}(\mathcal{C}') \leq \mathrm{ADV}(\mathcal{C}) = A$. Applying Lemma 8 to μ', there is an $i_{t+1} \in [N]$ with $p := \min(\mu'(C_{i_{t+1}} = 0), \mu'(C_{i_{t+1}} = 1)) \geq \frac{1}{36A'^2} \geq \frac{1}{36A^2}$. Note that $H(p) \geq \Omega(\log(A)/A^2)$. Hence

$$\begin{aligned} H(C \mid C_{i_1} \ldots C_{i_t} = b) - H(C \mid C_{i_1} \ldots C_{i_t} = b, C_{i_{t+1}}) &= H(C_{i_{t+1}} \mid C_{i_1} \ldots C_{i_t} = b) \\ &\geq \Omega(\log(A)/A^2). \end{aligned}$$

This implies $E_t - E_{t+1} \geq (1 - P_t) \cdot \Omega(\log(A)/A^2)$. In particular, as long as $P_t < 5/6$, the $(t+1)$st query shrinks E_t by at least $\frac{1}{6}\Omega(\log(A)/A^2) = \Omega(\log(A)/A^2)$. Since $E_0 = H(C) \leq \log|\mathcal{C}|$ and E_t cannot shrink below 0, there can be at most $O\left(\frac{A^2}{\log A} \log|\mathcal{C}|\right)$ queries before P_t grows to $\geq 5/6$. □

Since $\mathrm{ADV}(\mathcal{C})$ lower bounds $Q(\mathcal{C})$, Theorem 10 implies the bound

$$R(\mathcal{C}) \leq O\left(\frac{Q(\mathcal{C})^2}{\log Q(\mathcal{C})} \log |\mathcal{C}|\right)$$

claimed in our introduction. Note that this bound is tight up to a constant factor for the class of N-bit point functions, where $Q(\mathcal{C}) = \Theta(\sqrt{N})$, $|\mathcal{C}| = N$, and $R(\mathcal{C}) = \Theta(N)$ classical queries are necessary and sufficient.

5 Future work

Neither of our two results is tight. As directions for future work, let us state two conjectures, one for each model:

- k-Fourier-sparse functions can be learned from $O(k \cdot \mathrm{polylog}(k))$ uniform quantum examples.
- For all concept classes $\mathcal{C}$ of Boolean-valued functions on a domain of size N we have: $R(\mathcal{C}) = O(Q(\mathcal{C})^2 + Q(\mathcal{C}) \log N)$.

Acknowledgements. We thank Swagato Sanyal for pointing out an error in a previous version of this paper.

References

[1] J. Adcock, E. Allen, M. Day, S. Frick, J. Hinchliff, M. Johnson, S. Morley-Short, S. Pallister, A. Price, and S. Stanisic. Advances in quantum machine learning, 2015. URL https://arxiv.org/abs/1512.02900.

[2] A. Ambainis. Quantum lower bounds by quantum arguments. *Journal of Computer and System Sciences*, 64(4):750–767, 2002. DOI: 10.1006/jcss.2002.1826. Earlier version in STOC'00.

[3] S. Arunachalam and R. de Wolf. Guest column: A survey of quantum learning theory. *SIGACT News*, 48(2):41–67, 2017. DOI: 10.1145/3106700.3106710. arXiv:1701.06806.

[4] S. Arunachalam and R. de Wolf. Optimal quantum sample complexity of learning algorithms. *Journal of Machine Learning Research*, 19, 2018. URL http://jmlr.org/papers/v19/18-195.html. Earlier version in CCC'17.

[5] A. Atıcı and R. Servedio. Quantum algorithms for learning and testing juntas. *Quantum Information Processing*, 6(5):323–348, 2009. DOI: 10.1007/s11128-007-0061-6.

[6] H. Barnum, M. Saks, and M. Szegedy. Quantum query complexity and semi-definite programming. In *Proceedings of 18th IEEE Conference on Computational Complexity*, pages 179–193, 2003. DOI: 10.1109/CCC.2003.1214419.

[7] E. Bernstein and U. Vazirani. Quantum complexity theory. *SIAM Journal on Computing*, 26(5):1411–1473, 1997. DOI: 10.1137/S0097539796300921. Earlier version in STOC'93.

[8] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd. Quantum machine learning. *Nature*, 549(7671), 2017. DOI: 10.1038/nature23474.

[9] J. Bourgain. An improved estimate in the restricted isometry problem. In *Geometric Aspects of Functional Analysis*, volume 2116 of *Lecture Notes in Mathematics*, pages 65–70. Springer, 2014. DOI: 10.1007/978-3-319-09477-9_5.

[10] N. H. Bshouty and J. C. Jackson. Learning DNF over the uniform distribution using a quantum example oracle. *SIAM Journal on Computing*, 28(3):1136—-1153, 1999. DOI: 10.1145/225298.225312. Earlier version in COLT'95.

[11] E. J. Candés and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12): 5406–5425, 2006. DOI: 10.1109/TIT.2006.885507.

[12] Sourav Chakraborty, Nikhil S Mande, Rajat Mittal, Tulasimohan Molli, Manaswi Paraashar, and Swagato Sanyal. Tight Chang's-lemma-type bounds for Boolean functions, 2020. URL `https://arxiv.org/abs/2012.02335`.

[13] M. C. Chang. A polynomial bound in Freiman's theorem. *Duke Mathematics Journal*, 113(3):399–419, 2002. DOI: 10.1215/S0012-7094-02-11331-3.

[14] M. Cheraghchi, V. Guruswami, and A. Velingker. Restricted isometry of Fourier matrices and list decodability of random linear codes. *SIAM Journal on Computing*, 42(5):1888–1914, 2013. DOI: 10.1137/120896773.

[15] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991. DOI: 10.1002/047174882X.

[16] V. Dunjko and H. J. Briegel. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics*, 81(7):074001, 2018. DOI: doi:10.1088/1361-6633/aab406.

[17] P. Gopalan, R. O'Donnell, R. A. Servedio, A. Shpilka, and K. Wimmer. Testing Fourier dimensionality and sparsity. *SIAM Journal on Computing*, 40(4):1075–1100, 2011. DOI: 10.1137/100785429. Earlier version in ICALP'09.

[18] L. K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of 28th ACM STOC*, pages 212–219, 1996. DOI: 10.1145/237814.237866.

[19] A. Harrow, A. Hassidim, and S. Lloyd. Quantum algorithm for solving linear systems of equations. *Physical Review Letters*, 103(15):150502, 2009. DOI: 10.1103/PhysRevLett.103.150502.

[20] H. Hassanieh, P. Indyk, D. Katabi, and E. Price. Nearly optimal sparse Fourier transform. In *Proceedings of 44th ACM STOC*, pages 563–578, 2012. DOI: 10.1145/2213977.2214029.

[21] I. Haviv and O. Regev. The list-decoding size of Fourier-sparse Boolean functions. *ACM Transactions on Computation Theory*, 8(3):10:1–10:14, 2016. DOI: 10.1145/2898439. Earlier version in CCC'15.

[22] P. Indyk and M. Kapralov. Sample-optimal Fourier sampling in any constant dimension. In *Proceedings of 55th IEEE FOCS*, pages 514–523, 2014. DOI: 10.1109/FOCS.2014.61.

[23] E. Mossel, R. O'Donnell, and R. Servedio. Learning functions of k relevant variables. *Journal of Computer and System Sciences*, 69(3):421–434, 2004. DOI: 10.1016/j.jcss.2004.04.002. Earlier version in STOC'03.

[24] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010. DOI: 10.1017/CBO9780511976667.

[25] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. DOI: 10.1017/CBO9781139814782.

[26] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008. DOI: 10.1002/cpa.20227.

[27] Swagato Sanyal. Fourier sparsity and dimension. volume 15, pages 1–13. Theory of Computing, 2019. DOI: 10.4086/toc.2019.v015a011.

[28] M. Schuld, I. Sinayskiy, and F. Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015. DOI: 10.1080/00107514.2014.964942.

[29] R. Servedio and S. Gortler. Equivalences and separations between quantum and classical learnability. *SIAM Journal on Computing*, 33(5):1067–1092, 2004. DOI: 10.1137/S0097539704412910. Combines earlier papers from ICALP'01 and CCC'01.

[30] R. Špalek and M. Szegedy. All quantum adversary methods are equivalent. In *Proceedings of 32nd ICALP*, volume 3580 of *Lecture Notes in Computer Science*, pages 1299–1311, 2005. DOI: 10.1007/11523468_105.

[31] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. DOI: 10.1145/1968.1972.

[32] K. A. Verbeurgt. Learning DNF under the uniform distribution in quasi-polynomial time. In *Proceedings of 3rd Annual Workshop on Computational Learning Theory (COLT'90)*, pages 314–326, 1990. URL `https://dl.acm.org/doi/10.5555/92571.92659`.

[33] P. Wittek. *Quantum Machine Learning: What Quantum Computing Means to Data Mining*. Elsevier, 2014. DOI: 10.1016/C2013-0-19170-2.

[34] R. de Wolf. A brief introduction to Fourier analysis on the Boolean cube. *Theory of Computing*, 2008. DOI: 10.4086/toc.gs.2008.001. ToC Library, Graduate Surveys 1.

[35] A. C-C. Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of 18th IEEE FOCS*, pages 222–227, 1977. DOI: 10.1109/SFCS.1977.24.