Two new results about quantum exact learning

Srinivasan Arunachalam
Center for Theoretical Physics, MIT, USA
arunacha@mit.edu

Sourav Chakraborty
Indian Statistical Institute, Kolkata, India.
sourav@isical.ac.in

Troy Lee
Centre for Quantum Software and Information, School of Software, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia.
troyjlee@gmail.com

Manaswi Paraashar
Indian Statistical Institute, Kolkata, India.
manaswi.isi@gmail.com

Ronald de Wolf
QuSoft, CWI and University of Amsterdam, the Netherlands.
rdewolf@cwi.nl

Abstract

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). Our main tool is an improvement of Chang's lemma for 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-Gortler, SICOMP'04) by a $\log Q$-factor.

2012 ACM Subject Classification Hardware → Quantum computation; Theory of computation → Sample complexity and generalization bounds; Theory of computation → Boolean function learning

Keywords and phrases Quantum computing, Exact learning, Analysis of Boolean functions, Fourier sparse Boolean functions

Digital Object Identifier 10.4230/LIPIcs.ICALP.2019.11

Related Version A full version of the paper is available at `https://arxiv.org/abs/1810.00481`.

Funding *Srinivasan Arunachalam*: Work done when 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*.
Sourav Chakraborty: Work done while on sabbatical at CWI, supported by ERC QPROGRESS.
Troy Lee: 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.
Ronald de Wolf: Supported by ERC QPROGRESS, and QuantERA project QuantAlgo 680-91-034.

© Srinivasan Arunchalam, Sourav Chakraborty, Troy Lee, Manaswi Paraashar and Ronald de Wolf;
licensed under Creative Commons License CC-BY

46th International Colloquium on Automata, Languages, and Programming (ICALP 2019).
Editors: Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi;
Article No. 11; pp. 11:1–11:14

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

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 the Harrow-Hassidim-Lloyd linear-systems solver [19] to speed up learning algorithms for specific machine learning tasks (see [34, 29, 1, 9, 16] 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 [4] 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. *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:

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

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 [11] 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 [33]). 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 [8], is very

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

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 $\chi_S(x) = (-1)^{x \cdot S}$; 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 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.[2] Atıcı and Servedio [6] showed that $(\log n)$-juntas can be exactly learned 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 [24]). These cases (and others surveyed in [4]) show that uniform quantum examples (and in particular Fourier sampling) can be more useful than classical examples.[3]

In this paper we consider the concept class of n-bit Boolean functions 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, 23] 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 [10], Rudelson and Vershynin [27] and Candés and Tao [12].

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 combines the fact that a uniform quantum example allows us to Fourier-sample the target concept, with some Fourier analysis of k-Fourier-sparse functions. In particular, we significantly strengthen "Chang's lemma" for the special case of k-Fourier-sparse Boolean functions. This lemma upper bounds the dimension of the span of the large-weight part of the Fourier support of a Boolean function, and our Theorem 13 improves this bound almost quadratically for the special case of k-Fourier-sparse functions. Our learner has two phases. In the first phase, using Chang's lemma, we show that the span of the Fourier support of the target function can be learned from $O(k(\log k)^2)$ Fourier samples. 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 $O(rk \log k)$

[2] 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.

[3] This is not the case in Valiant's *PAC-learning* model [32] 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 [5].

classical examples. Since it is known that $r = O(\sqrt{k}\log k)$ [28], 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.

Since $r \geq \log k$, the second phase of our learner is always at least as expensive as the first phase. It might be possible to improve the upper bound to $O(k \cdot \text{polylog}(k))$ quantum examples, but that would require additional ideas to improve phase 2. We also prove a (non-matching) lower bound of $\Omega(k \log k)$ uniform quantum examples, using techniques from quantum information theory. We omitted some proofs due to space limitations; these may be found in [3].

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*.[4] 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 $\leq 1/3$.[5] Servedio and Gortler [30] 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

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

[4] 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.

[5] 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 [4, Section 4.1] for more.

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:

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

Note that this inequality is tight both for the class of linear functions and for the class of point functions.

Our proof combines the quantum adversary method [2, 7, 31] with an entropic argument to show that we can always find a query whose outcome (no matter whether it's 0 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 0s elsewhere. For $x \in \{0,1\}^n$, $i \in [n]$, let x^i be the input obtained by flipping the ith bit in x.

For $f : \{0,1\}^n \to \{-1,1\}$ 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 [26, 35] 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 $\sum_{i=1}^n S_i x_i$. 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

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

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 domain $\{-1,1\}$, 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 few structural results about Fourier coefficients and dimension.

▶ **Theorem 1** ([28]). *The Fourier dimension of a k-Fourier-sparse $f : \{0,1\}^n \to \{-1,1\}$ is $O(\sqrt{k}\log k)$.*

▶ **Lemma 2** ([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 3.** *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 4.** *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:

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

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

An easy consequence is the next lemma:

▶ **Lemma 5.** *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 $\mathrm{Fspan}(f)$, and $\widehat{f}(B_1), \ldots, \widehat{f}(B_r)$ are non-zero. Then the Fourier span of f_B is spanned by $\{e_1, \ldots, e_r\}$, i.e., f_B has only r influential variables. Additionally, 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:

▶ **Lemma 6.** *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.*

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 [25, 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 7** (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 8** (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

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

We prove the following two theorems here.

▶ **Theorem 9.** *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 $O(k^{1.5}(\log k)^2)$.*

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

▶ **Theorem 10.** *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 $\Omega(k \log k)$.*

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 $O(k(\log k)^2)$ uniform quantum examples (for Fourier-sampling) and phase 2 uses $O(rk \log k)$ uniform *classical* examples. Note that since $r \geq \log k$, phase 2 of our learner is always at least as expensive as phase 1.

▶ **Theorem 11.** *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 9 follows immediately from Theorem 11.

3.1.1 Phase 1: Learning the Fourier span

A crucial ingredient that we use in phase 1 of our quantum learning algorithm is an improvement of Chang's lemma [13, 22] for k-Fourier-sparse Boolean functions. The original lemma upper bounds the dimension of the span of the "large" Fourier coefficients as follows.

▶ **Lemma 12** (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 k-Fourier-sparse Boolean functions. In particular, consider the case $\rho\alpha = 1/k$. In that case, since all elements of the Fourier support satisfy $|\widehat{f}(S)| \geq 1/k$ by Lemma 2, the left-hand side of Eq. (2) equals the Fourier dimension r of f. Chang's lemma gives

$$r \leq 2\alpha^2 k^2 \log k.$$

We now improve this upper bound on r nearly quadratically:

▶ **Theorem 13.** *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*

$$r \leq 2\alpha k \log k.$$

For a proof of this theorem, see the full version of the paper. We now illustrate how this theorem improves over Lemma 12. First, observe that $\alpha \geq 1/k$ (by Lemma 2), so $\alpha k \leq \alpha^2 k^2$. Second, consider a Boolean function f which satisfies $\alpha = 1/k^{3/4}$. Then, Chang's lemma (with $\rho = 1/k^{1/4}$) upper bounds the Fourier dimension of f as $r \leq O(\sqrt{k}\log k)$, which already follows from Theorem 1. Our Theorem 13 gives the much better upper bound $r \leq O(k^{1/4}\log k)$ in this case.

Now that we have a better understanding of the Fourier dimension of k-Fourier-sparse Boolean functions, we would like to understand how many Fourier samples suffice to obtain the Fourier span of f (in fact this will be our quantum learning algorithm for phase 1). Since the $\leq k$ squared non-zero Fourier coefficients of a k-Fourier-sparse function are each at least $1/k^2$, it is easy to see that after $O(k^2 \log k)$ Fourier samples we are likely to have seen every element in the Fourier support, and hence know the full Fourier support as well. We will improve on this easy bound below. The main idea is to show that if the span of the Fourier samples seen at a certain point has some dimension $r' < r$, then there is significant Fourier weight on elements outside of this span, so after a few more Fourier samples we will have grown the span. We now state this formally and prove the lemma.

▶ **Lemma 14.** *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*

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

Proof. Let $B \in \mathbb{F}_2^{r\times r}$ be an invertible matrix such that the first $r' < r$ columns of B form a basis for $\mathrm{span}(\mathcal{S})$. By Lemma 5, f_B depends only on r bits, so we write $f_B : \{0,1\}^r \to \{-1,1\}$. Let $\mathcal{W} = \mathrm{span}\{e_1, \ldots, e_{r'}\} \subseteq \{0,1\}^r$. Then

$$\sum_{S\in \mathrm{span}(\mathcal{S})} \widehat{f}(S)^2 = \sum_{S\in\mathcal{W}} \widehat{f}_B(S)^2. \tag{3}$$

Let us decompose f_B as follows: $f_B(x_1,\ldots,x_r) = g(x_1,\ldots,x_{r'}) + g'(x_1,\ldots,x_r)$, where

$$g(y) = \sum_{T\in\{0,1\}^{r'}} \widehat{f}_B(T,0^{r-r'})\chi_T(y,0^{r-r'}) \quad \text{for every } y\in\{0,1\}^{r'}, \tag{4}$$

and

$$g'(x) = \sum_{S\notin\mathcal{W}} \widehat{f}_B(S)\chi_S(x) \quad \text{for every } x\in\{0,1\}^r.$$

Now by Parseval's identity we have

$$\mathbb{E}_{y\in\{0,1\}^{r'}}[g(y)^2] = \sum_{T\in\{0,1\}^{r'}} \widehat{g}(T)^2 = \sum_{S\in\mathcal{W}} \widehat{f}_B(S)^2, \tag{5}$$

where the second equality used Eq. (4). Combining Eq. (5) with an averaging argument, there exists an assignment of $a = (a_1,\ldots,a_{r'}) \in \{0,1\}^{r'}$ to $(y_1,\ldots,y_{r'})$ such that

$$g(a_1,\ldots,a_{r'})^2 \geq \sum_{S\in\mathcal{W}} \widehat{f}_B(S)^2, \tag{6}$$

Consider the function h defined as

$$h(z_1,\ldots,z_{r-r'}) = f_B(a_1,\ldots,a_{r'},z_1,\ldots,z_{r-r'}) \quad \text{for every } z_1,\ldots,z_{r-r'}\in\{0,1\}. \tag{7}$$

Note that h has Fourier sparsity at most the Fourier sparsity of f_B, hence at most k. Also, the Fourier dimension of h is at most $r-r'$. Finally note that

$$\begin{aligned}
\widehat{h}(0^{r-r'}) &= \mathbb{E}_{z\in\{0,1\}^{r-r'}}[h(z)] \\
&= \mathbb{E}_{z\in\{0,1\}^{r-r'}}[f_B(a,z)] && \text{(by Eq. (7))} \\
&= \mathbb{E}_{z\in\{0,1\}^{r-r'}}\Big[\sum_{S_1\in\{0,1\}^{r'}}\sum_{S_2\in\{0,1\}^{r-r'}} \widehat{f}_B(S_1,S_2)\chi_{S_1}(a)\chi_{S_2}(z)\Big] \\
& && \text{(Fourier expansion of } f_B) \\
&= \sum_{S_1\in\{0,1\}^{r'}} \widehat{f}_B(S_1,0^{r-r'})\chi_{S_1}(a,0^{r-r'}) && \text{(using } \mathbb{E}_{z\in\{0,1\}^{r-r'}}\chi_S(z) = \delta_{S,0^{r-r'}}) \\
&= g(a_1,\ldots,a_{r'}) && \text{(by definition of } g \text{ in Eq. (4))} \\
&\geq \Big(\sum_{S\in\mathcal{W}} \widehat{f}_B(S)^2\Big)^{1/2}. && \text{(by Eq. (6))}
\end{aligned}$$

Using Theorem 13 for the function h, it follows that $\widehat{h}(0^{r-r'}) \leq 1 - (r-r')/(k\log k)$, which in particular implies

$$\sum_{S\in\mathrm{span}(\mathcal{S})} \widehat{f}(S)^2 = \sum_{S\in\mathcal{W}} \widehat{f}_B(S)^2 \leq \widehat{h}(0^{r-r'})^2 \leq 1 - \frac{r-r'}{k\log k},$$

where the first equality used Eq. (3). ◀

▶ **Theorem 15.** *For every k-Fourier-sparse Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ with Fourier dimension r, its Fourier span can be learned using an expected number of $O(k \log k \log r)$ quantum examples.*

Proof. We only use the quantum examples for Fourier sampling; an expected number of two quantum examples suffices to get one Fourier sample. At any point of time let $\mathcal{S}$ be the set of samples we have received. Let the dimension of the span of $\mathcal{S}$ be r'. Now if we receive a new sample S such that $S \notin \text{span}(\mathcal{S})$, then the dimension of the samples we have seen increases by 1. By Lemma 14

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

So the expected number of samples to increase the dimension by 1 is $\leq \frac{k \log k}{r - r'}$. Hence, 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 $\sum_{i=1}^{r} \frac{1}{i} = O(\log r)$. ◀

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 of B form a basis for $\mathcal{T}$. Consider $c_B = c \circ (B^{-1})^\mathsf{T}$ for $c \in \mathcal{C}'$. By Lemma 5 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 7) 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

We show that $\Omega(k \log k)$ uniform quantum examples are necessary to learn the concept class of k-Fourier-sparse Boolean functions. See the full version of the paper for the proof.

▶ **Theorem 16.** *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 $\Omega(k \log k)$.*

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. [7], which is called the "spectral adversary" in the survey [31].

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 $[c_i \neq c'_i]$.[6] 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 17.** *For $N \geq 1$, let $\mathcal{C} \subseteq \{0,1\}^N$ be a concept class and suppose $\mathrm{ADV}(\mathcal{C}) = \max_{\Gamma \geq 0} \| \Gamma \| / \max_{i \in [N]} \| \Gamma \circ D_i \|$ is the nonnegative adversary bound for the exact learning problem corresponding to $\mathcal{C}$. Let μ be a distribution on $\mathcal{C}$ such that $\max_{c \in \mathcal{C}} \mu(c) \leq 5/6$, and let $\mathbf{C}$ be a random variable distributed according to μ. Then there exists an $i \in [N]$ such that*

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

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

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

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

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 $\mathbf{C}_i = 1$. 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

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

where the last inequality used $\max(\mu(\mathbf{C}_i = 0), \mu(\mathbf{C}_i = 1)) \leq 1$. ◀

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 18.** *For $N \geq 1$, let $\mathcal{C} \subseteq \{0,1\}^N$ be a concept class and suppose $\mathrm{ADV}(\mathcal{C}) = \max_{\Gamma \geq 0} \| \Gamma \| / \max_{i \in [N]} \| \Gamma \circ D_i \|$ is the nonnegative adversary bound for the exact learning problem corresponding to $\mathcal{C}$. Then there exists a classical learner for the concept class $\mathcal{C}$ 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" [36] 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 $\mathbf{C} \sim \mu$:

1. Choose an i that maximizes $H(\mathbf{C}_i)$ and query that i.[7]
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 $\mathbf{I}_1, \mathbf{I}_2, \ldots$. 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 by the algorithm 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:

$$P_t = \sum_{i_1,\ldots,i_t \in [N],\ b \in \{0,1\}^t} \Pr[\mathbf{I}_1 = i_1, \ldots, \mathbf{I}_t = i_t, \mathbf{C}_{i_1} \ldots \mathbf{C}_{i_t} = b] \cdot [\exists c \in \mathcal{C} \text{ s.t. } \mu(c \mid \mathbf{C}_{i_1} \ldots \mathbf{C}_{i_t} = b) \geq 5/6].$$

[7] 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.

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:

$$\begin{aligned} E_t &= H(\mathbf{C} \mid \mathbf{C}_{\mathbf{I}_1}, \ldots, \mathbf{C}_{\mathbf{I}_t}) \\ &= \sum_{i_1,\ldots,i_t \in [N],\ b \in \{0,1\}^t} \Pr[\mathbf{I}_1 = i_1, \ldots, \mathbf{I}_t = i_t, \mathbf{C}_{i_1} \ldots \mathbf{C}_{i_t} = b] \cdot H(\mathbf{C} \mid \mathbf{C}_{i_1} \ldots \mathbf{C}_{i_t} = b). \end{aligned}$$

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

Let $i_1, ..., i_t$ and b be such that there is no c in $\mathcal{C}$ s.t. $\mu(c \mid \mathbf{C}_{i_1} \ldots \mathbf{C}_{i_t} = b) \geq 5/6$ (note that the μ-probability of getting such an $i_1, ..., i_t$ and b, is $1 - P_t$). 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 17 to μ', there is an $i_{t+1} \in [N]$ with $p := \min(\mu'(\mathbf{C}_{i_{t+1}} = 0), \mu'(\mathbf{C}_{i_{t+1}} = 1)) \geq \frac{1}{36A'^2} \geq \frac{1}{36A^2}$. Note that $H(p) \geq \Omega(\log(A)/A^2)$. Hence

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

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 $E_0 = H(\mathbf{C}) \leq \log|\mathcal{C}|$ 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 18 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 $A = \Theta(\sqrt{N})$, $|\mathcal{C}| = N$, 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: $R(\mathcal{C}) = O(Q(\mathcal{C})^2 + Q(\mathcal{C}) \log N)$.

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, 9 Dec 2015. arXiv:1512.02900.
2. A. Ambainis. Quantum lower bounds by quantum arguments. *Journal of Computer and System Sciences*, 64(4):750–767, 2002. Earlier version in STOC'00. quant-ph/0002066.
3. S. Arunachalam, S. Chakraborty, T. Lee, M. Paraashar, and R. de Wolf. Two new results about quantum exact learning. arXiv:1810.00481.
4. S. Arunachalam and R. de Wolf. Guest column: A survey of quantum learning theory. *SIGACT News*, 48(2):41–67, 2017. arXiv:1701.06806.
5. S. Arunachalam and R. de Wolf. Optimal quantum sample complexity of learning algorithms. *Journal of Machine Learning Research*, 19, 2018. Earlier version in CCC'17. arXiv:1607.00932.
6. A. Atıcı and R. Servedio. Quantum algorithms for learning and testing juntas. *Quantum Information Processing*, 6(5):323–348, 2009. arXiv:0707.3479.
7. 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.
8. E. Bernstein and U. Vazirani. Quantum complexity theory. *SIAM Journal on Computing*, 26(5):1411–1473, 1997. Earlier version in STOC'93.
9. J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd. Quantum machine learning. *Nature*, 549(7671), 2017. arXiv:1611.09347.
10. 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, 2014.
11. 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. Earlier version in COLT'95.
12. 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.
13. M. C. Chang. A polynomial bound in Freimans theorem. *Duke Mathematics Journal*, 113(3):399–419, 2002.
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.
15. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 1991.
16. V. Dunjko and H. Briegel. Machine learning & artificial intelligence in the quantum domain, 8 Sep 2017. arXiv:1709.02779.
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. 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. quant-ph/9605043.
19. A. Harrow, A. Hassidim, and S. Lloyd. Quantum algorithm for solving linear systems of equations. *Physical Review Letters*, 103(15):150502, 2009. arXiv:0811.3171.

403 **20** H. Hassanieh, P. Indyk, D. Katabi, and E. Price. Nearly optimal sparse Fourier transform. In
404 *Proceedings of 44th ACM STOC*, pages 563–578, 2012.

405 **21** I. Haviv and O. Regev. The list-decoding size of Fourier-sparse Boolean functions. *ACM*
406 *Transactions on Computation Theory*, 8(3):10:1–10:14, 2016. Earlier version in CCC'15.
407 arXiv:1504.01649.

408 **22** R. Impagliazzo, C. Moore, and A. Russell. An entropic proof of Chang's inequality. *SIAM*
409 *Journal of Discrete Mathematics*, 28(1):173–176, 2014. arXiv:1205.0263.

410 **23** P. Indyk and M. Kapralov. Sample-optimal Fourier sampling in any constant dimension. In
411 *Proceedings of 55th IEEE FOCS*, pages 514–523, 2014.

412 **24** E. Mossel, R. O'Donnell, and R. Servedio. Learning functions of k relevant variables. *Journal*
413 *of Computer and System Sciences*, 69(3):421–434, 2004. Earlier version in STOC'03.

414 **25** M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. 2000.

415 **26** R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

416 **27** M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measure-
417 ments. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

418 **28** S. Sanyal. Near-optimal upper bound on Fourier dimension of Boolean functions in terms of
419 Fourier sparsity. In *Proceedings of 42nd ICALP*, pages 1035–1045, 2015.

420 **29** M. Schuld, I. Sinayskiy, and F. Petruccione. An introduction to quantum machine learning.
421 *Contemporary Physics*, 56(2):172–185, 2015. arXiv:1409.3097.

422 **30** R. Servedio and S. Gortler. Equivalences and separations between quantum and classical
423 learnability. *SIAM Journal on Computing*, 33(5):1067–1092, 2004. Combines earlier papers
424 from ICALP'01 and CCC'01. quant-ph/0007036.

425 **31** R. Špalek and M. Szegedy. All quantum adversary methods are equivalent. In *Proceedings*
426 *of 32nd ICALP*, volume 3580 of *Lecture Notes in Computer Science*, pages 1299–1311, 2005.
427 quant-ph/0409116.

428 **32** L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134—-1142, 1984.

429 **33** K. A. Verbeurgt. Learning DNF under the uniform distribution in quasi-polynomial time. In
430 *Proceedings of 3rd Annual Workshop on Computational Learning Theory (COLT'90)*, pages
431 314–326, 1990.

432 **34** P. Wittek. *Quantum Machine Learning: What Quantum Computing Means to Data Mining*.
433 Elsevier, 2014.

434 **35** R. de Wolf. A brief introduction to Fourier analysis on the Boolean cube. *Theory of Computing*,
435 2008. ToC Library, Graduate Surveys 1.

436 **36** A. C-C. Yao. Probabilistic computations: Toward a unified measure of complexity. In
437 *Proceedings of 18th IEEE FOCS*, pages 222–227, 1977.