Quantum statistical query learning

Srinivasan Arunachalam* Alex B. Grilo† Henry Yuen‡

Abstract

We propose a learning model called the *quantum statistical learning (*QSQ*)* model, which extends the SQ learning model introduced by Kearns [Kea98] to the quantum setting. Our model can be also seen as a restriction of the quantum PAC learning model: here, the learner does not have direct access to quantum examples, but can only obtain estimates of measurement statistics on them. Theoretically, this model provides a simple yet expressive setting to explore the power of quantum examples in machine learning. From a practical perspective, since simpler operations are required, learning algorithms in the QSQ model are more feasible for implementation on near-term quantum devices.

We prove a number of results about the QSQ learning model. We first show that parity functions, $O(\log n)$-juntas and polynomial-sized DNF formulas are efficiently learnable in the QSQ model, in contrast to the classical setting where these problems are provably hard. This implies that many of the advantages of quantum PAC learning can be realized even in the more restricted quantum SQ learning model.

It is well-known that *weak statistical query dimension*, denoted by $\mathsf{WeakSQDIM}(\mathcal{C})$, characterizes the complexity of learning a concept class $\mathcal{C}$ in the classical SQ model. We show that $\log(\mathsf{WeakSQDIM}(\mathcal{C}))$ is a lower bound on the complexity of QSQ learning, and furthermore it is tight for certain concept classes $\mathcal{C}$. Additionally, we show that this quantity provides strong lower bounds for the small-bias quantum communication model under product distributions.

Finally, we introduce the notion of *private* quantum PAC learning, in which a quantum PAC learner is required to be *differentially private*. We show that learnability in the QSQ model implies learnability in the quantum private PAC model. Additionally, we show that in the private PAC learning setting, the classical and quantum sample complexities are equal, up to constant factors.

1 Introduction

The prospect of using quantum computers to perform machine learning has received much attention lately, given their potential to offer significant speedups for solving certain problems of practical relevance. There has been a flurry of proposed quantum algorithms for performing computations that are ubiquitous in machine learning, ranging from convex optimization, matrix completion, clustering, support vector machines [KP17, BKL+19, LMR13, RML14]. Due to the assumptions required by these quantum algorithms, the evidence for a quantum computational advantage in performing machine learning tasks is murky at best [Tan19, CGL+19]. It is therefore an active area of research to obtain evidence (even conditional) for quantum advantage in machine learning.

*IBM Research. Srinivasan.Arunachalam@ibm.com

†QuSoft and CWI, Amsterdam. alexg@cwi.com

‡University of Toronto. hyuen@cs.toronto.edu

Quantum learning theory has provided a theoretical framework to study the capabilities and limitations of quantum machine learning. Here, the focus is not only on the computational complexity of learning algorithms, but also on *information-theoretic* measures such as sample and query complexity. One of the first classical learning models that were generalized to the quantum setting was Valiant's Probably Approximately Correct (PAC) model of learning [Val84]. In the classical PAC model of learning, the goal is to learn a collection of Boolean functions, which is often referred to as a *concept class* $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$. The elements of a concept class are called *concepts*. In the PAC model of learning, there is an unknown distribution $D : \{0,1\}^n \to [0,1]$ and a learner is given *labelled examples* $\{x_i, c^*(x_i)\}_i$ where x_i is drawn from the distribution D and $c^* \in \mathcal{C}$ is the unknown *target concept*. The goal of a learner is the following: for every unknown D and c^*, use labelled examples to produce a hypothesis h that satisfies $\Pr_{x\sim D}[h(x) = c^*(x)] \geq 2/3$. The *quantum PAC model*, introduced by Bshouty and Jackson [BJ95], considers the extension of Valiant's PAC model where the learning algorithm is not given labelled examples $\{x_i, c(x_i)\}_i$, but instead is given copies of a *quantum example*

$$|\psi_{c^*}\rangle = \sum_{x\in\{0,1\}^n} \sqrt{D(x)}|x, c^*(x)\rangle.$$

which is a *superposition* of labeled examples. Observe that simply measuring $|\psi_{c^*}\rangle$ in the computational basis gives a classical labelled example. Quantum examples are well-motivated in quantum computing: they arise naturally in quantum query algorithms, and also have interesting complexity-theoretic applications [AT07].

In the distribution-independent PAC learning model, Arunachalam and de Wolf [AW18] showed that the sample complexity of quantum and classical PAC learning is the same. However, in the uniform distribution learning model (i.e., when we fix D to be the uniform distribution), quantum examples have been shown to be very powerful. In particular, given uniform quantum examples $\frac{1}{\sqrt{2^n}}\sum_x |x, c^*(x)\rangle$ a quantum learner can efficiently sample from the Fourier distribution $\{\widehat{c^*}(S)^2\}_S$, a tool that has been used to provide even exponential advantage over the known classical algorithms [BJ95, AS05, GKZ19, ACL+19].

In this paper, we further investigate the power of quantum examples in learning, by defining a restricted quantum learning model and studying its capabilities and limitations. We call it the *quantum statistical query (*QSQ*)* model , which extends the well-studied (classical) *statistical query* (SQ) learning model introduced by Kearns [Kea98]. In SQ learning, the learner constructs a hypothesis not by examining a sequence of labelled examples, but instead by adaptively querying an oracle to obtain *estimates* of statistical properties of the labelled examples. Though this model is weaker than PAC learning, it is rich enough to capture many known learning algorithms [BDMN05, Fel16, Fel17].

Quantum statistical query model. In the QSQ model, the learner – which is still a *classical* randomized algorithm – can query an oracle to obtain statistics of quantum examples to compute a hypothesis. Roughly speaking, these statistics correspond to the average value obtained if a quantum computer would repeatedly measure copies of quantum examples using a specified measurement M. More concretely, in quantum computing measurements are defined by Hermitian matrices called observables, and the statistics obtained by the learner consist of an estimate of the expectation value $\langle\psi_{c^*}|M|\psi_{c^*}\rangle$ for a chosen observable M. When M is diagonal, this reduces to the case of making classical SQ queries, and the power of QSQ appears when M corresponds to measurements of $|\psi_{c^*}\rangle$ in a non-classical basis.

We motivate the study of this model in several ways. Some concept classes appear to be learned more efficiently in the quantum PAC setting (at least in the *distribution-dependent* setting); a natural question is whether these efficiency gains come from the ability of the quantum learning algorithm to *directly* manipulate coherent superpositions of labeled data (i.e., quantum examples), or does the weaker quantum statistical query access suffice? Are there classical-quantum learning separations even in this weak statistical query model, where the learner can only access the data through measurement statistics of quantum examples?

Another motivation comes from the consideration that QSQ learners are more practically feasible than general quantum PAC learners. A general quantum PAC learning algorithm could perform complex entangling unitaries and measurements on many quantum examples simultaneously in order to extract joint statistics. However, this seems far beyond the capabilities of noisy, near-term quantum computers. In the QSQ learning model, the learner can only obtain statistics about individual quantum examples. In a practical implementation of these quantum learning algorithms, this would only require measuring a single quantum example at a time. One could imagine a scenario where classical learning algorithms can query a cloud-based quantum computer to solve a learning task; the QSQ model would lend itself naturally to this situation.

Our contributions. The first contribution of our paper is providing a definition for the quantum statistical query model. We then prove a number of results regarding this model.

1. We show that a query-efficient QSQ learner for a concept class $\mathcal{C}$ under a distribution D implies a sample-efficient quantum PAC learner for $\mathcal{C}$ under the same distribution, and furthermore implies a sample-efficient *noisy* quantum PAC learning of $\mathcal{C}$. This is exactly analogous to how classical SQ learning is a restriction of noisy PAC learning, which is itself a restriction of standard PAC learning.

2. We present three learning problems that can be solved efficiently in the QSQ model, but not in the classical SQ model. In particular, we show that it is possible to learn parity functions, juntas, and DNF formulas under the uniform distribution in polynomial time in the QSQ model; in contrast, the same problems are provably hard in the classical SQ model. Notice that for juntas and DNFs, no efficient classical learning algorithm is currently known even in the setting where the learner is given the classical samples.

3. We show that while the statistical query dimension characterizes the query complexity in the SQ model, its logarithm is a tight lower-bound to the query complexity of QSQ learning. We also show a connection between one-way communication complexity (under product distributions) with weak statistical query dimension. In particular, this connection allows us to prove non-trivial lower bounds on the communication complexity (under product distributions) even for *inverse exponential bias* in computing the function value.

4. Our final contribution in this paper is to define the notion of *privacy* in the quantum PAC learning model. We then lift the fundamental connection between classical statistical query learning and private PAC learning to the quantum setting and show that learnability in the quantum SQ model implies private quantum PAC learnability. Finally, we provide a combinatorial characterization of the sample complexity of private quantum PAC learning and using this characterization we show that the sample complexities of private classical and quantum PAC learning are equal, up to constant factors.

Acknowledgements. We thank Sasho Nikolov and Tanay Mehta for useful discussions. SA did part of this work at MIT and was supported by the MIT-IBM Watson AI Lab under the project *Machine Learning in Hilbert space*. HY was supported by NSERC Discovery Grant 2019-06636. Part of this work was done while AG and HY were visiting the Simons Institute for the Theory of Computing.

Organization. In Section 3 we introduce the quantum statistical query model. In Section 4 we present three concept classes that are efficiently learnable in the quantum statistical query model. We follow by showing a lower bound to the query complexity in the QSQ model and its relations to communication complexity in Section 5. Finally, in Section 6 we present connections between the QSQ model and quantum differential privacy.

2 Preliminaries

We let $[n] = \{1,\ldots,n\}$. For $s \in \{0,1\}^n$, define $\mathsf{supp}(s) = \{i \in [n] : s_i = 1\}$. For $S \subseteq [n]$, denote $S^c = [n]\backslash S$ be the complement of S.

Quantum computing. We briefly review the basic concepts in quantum computing. We define $|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $|1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ as the canonical basis for $\mathbb{C}^2$. A qubit $|\psi\rangle$ is a unit vector in $\mathbb{C}^2$, i.e., $\alpha|0\rangle+\beta|1\rangle$ for $\alpha,\beta \in \mathbb{C}$ that satisfy $|\alpha|^2+|\beta|^2 = 1$. Multi-qubit quantum states are obtained by taking tensor products of single-qubit states: an arbitrary n-qubit quantum state $|\psi\rangle \in \mathbb{C}^{2^n}$ is a unit vector in $\mathbb{C}^{2^n}$ and can be expressed as $|\psi\rangle = \sum_{x\in\{0,1\}^n} \alpha_x|x\rangle$ where $\alpha_x \in \mathbb{C}$ and $\sum_x |\alpha_x|^2 = 1$. We denote by $\langle\psi|$ as the conjugate transpose of the quantum state $|\psi\rangle$. On a quantum computer one is allowed to arbitrary quantum gates (or operations) that correspond to unitary matrices. One gate we use often is the Hadamard gate, defined as $\mathsf{H} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. An *observable* M is a Hermitian matrix, which encodes a measurement in quantum mechanics. The average measurement outcome of a state $|\psi\rangle$ using the observable O is given by the expectation value $\langle\psi|O|\psi\rangle$.

Fourier analysis. We now introduce the basics of Fourier analysis on the Boolean cube. For $S \in \{0,1\}^n$, we define the *character function* $\chi_S : \{0,1\}^n \to \{-1,1\}$ as $\chi_S(x) = (-1)^{S\cdot x}$ where $S \cdot x = \sum_i S_i \cdot x_i \pmod 2$. For $f : \{0,1\}^n \to \{-1,1\}$, the Fourier coefficients of f are

$$\widehat{f}(S) = \mathbb{E}_{x\in\{0,1\}^n}[f(x)\cdot\chi_S(x)] \quad \text{for every } S \in \{0,1\}^n,$$

where the expectation is taken with respect to the uniform distribution over $\{0,1\}^n$. Every function $f : \{0,1\}^n \to \mathbb{R}$ can be written uniquely as $f(x) = \sum_{S\in\{0,1\}^n} \widehat{f}(S)\chi_S(x)$. Parseval's identity states that $\sum_S \widehat{f}(S)^2 = \mathbb{E}[f(x)^2] = 1$. Hence, $\{\widehat{f}(S)^2\}_{S\in\{0,1\}^n}$ forms a probability distribution. For every $i \in [n]$, we define the ith influence as

$$\mathrm{Inf}_i(f) = \sum_{\substack{S\in\{0,1\}^n:\\ S_i=1}} \widehat{f}(S)^2.$$

2.1 PAC learning

Valiant [Val84] introduced the Probably Approximately Correct (PAC) model of learning, which gives a formalization of what "learning a function" means. In this learning model, a *concept class* $\mathcal{C}$ is a collection of Boolean functions $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$. The functions inside $\mathcal{C}$ are referred to as *concepts*. Let $D : \{0,1\}^n \to [0,1]$ be an *unknown distribution* over the Boolean cube. In the PAC model, a learner $\mathscr{A}$ is given many *labelled examples* $(x, c(x))$ where x is drawn from the distribution D and $c \in \mathcal{C}$ is the *unknown* target concept. The goal of an (ε,δ)-learner is the following: with probability at least $1-\delta$ (probability taken according to internal randomness of $\mathscr{A}$ and D), output a hypothesis $h : \{0,1\}^n \to \{0,1\}$ that satisfies $\Pr_{x\sim D}[h(x) = c(x)] \geq 1-\varepsilon$. The (ε,δ)-sample complexity of a learning algorithm $\mathscr{A}$ is the maximal number of labelled examples used, maximized over all $c \in \mathcal{C}$ and distributions $D : \{0,1\}^n \to [0,1]$. The (ε,δ)-sample complexity of learning $\mathcal{C}$ is the minimal sample complexity over all (ε,δ)-learners for $\mathcal{C}$.

We say $\mathscr{A}$ is a *uniform-(ε,δ) learner* for a concept class $\mathcal{C}$ if the distribution D is *fixed* to be the uniform distribution over $\{0,1\}^n$ and $\mathscr{A}$ learns $\mathcal{C}$ under the uniform distribution.

2.1.1 Quantum PAC learning.

The quantum PAC model of learning was introduced by [BJ95]. In this model a quantum learning algorithm has access to a quantum computer and *quantum examples* $\sum_x \sqrt{D(x)}|x, c(x)\rangle$, and the goal is still output a *classical* hypothesis h with the same requirements as in the classical setting. For every $\mathcal{C}$, the (ε,δ)-quantum PAC complexities are defined as the quantum analogues to the classical complexity measures. For more on these learning models, we refer the reader to [AW17] and the references therein.

Noisy quantum examples. Following the work of Grilo et al. [GKZ19], we define noisy quantum PAC learning. Here, a learner is provided with copies of a *noisy quantum example* for a concept $c \in \mathcal{C}$ and distribution D as a superposition of *noisy classical examples*.[1] Understanding the quantum and classical learnability of functions in the noisy setting is motivated by the connection to important problems in cryptography such as learning parity with noise [Pie12] and learning with errors problem [Reg09]. More concretely, a η-noisy quantum example for a concept c is given by

$$|\widehat{\psi}_c\rangle = \sum_{x\in\{0,1\}^n} \sqrt{D(x)}|x, c(x) \oplus b_x\rangle, \tag{1}$$

where each b_x is an i.i.d. variable which equals 0 with probability $1-\eta$ and 1 otherwise. Here again the goal of a learner is to learn a concept class $\mathcal{C}$ under all distributions D. The complexity of such learners is defined exactly as we defined it for quantum PAC learning, except that we also allow the sample complexity of an η-noisy PAC learner to depend on the factor $1/(1-2\eta)$.[2]

[1] In the classical setting, a noisy classical PAC learner obtains many $(x, c(x)+b_x)$ where x is sampled from D and b_x is an independent random variable which equals 0 with probability $1-\eta$ and 1 otherwise and using these noisy examples a learner needs to learn c.

[2] Note that when $\eta = 1/2$, a learner is obtaining uniformly random bits of information in which case we cannot hope to learn c.

3 Quantum statistical query learning

In this section, we introduce the model of quantum statistical query learning (QSQ). We start by briefly describing the *classical* SQ learning. Let $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{-1,1\}\}$ be a concept class. The goal of a statistical learning algorithm is to learn an unknown $c^* \in \mathcal{C}$ under an unknown distribution $D : \{0,1\}^n \to [0,1]$. A (classical) SQ learning algorithm has access to a *statistical query oracle* Stat which takes as input a *tolerance* parameter $\tau \geq 0$ and a function $\phi : \{0,1\}^n \times \{-1,1\} \to \{-1,1\}$ and returns a number α such that

$$\left|\alpha - \mathbb{E}_{x\sim D}[\phi(x, c^*(x))]\right| \leq \tau .$$

The SQ learning algorithm adaptively chooses a sequence $\{(\phi_i, \tau_i)\}$, and based on the responses of $\{\mathsf{Stat}(\phi_i, \tau_i)\}_i$, it outputs a hypothesis $h : \{0,1\}^n \to \{0,1\}$. We say that an SQ learning algorithm $\mathscr{A}$ ε-learns $\mathcal{C}$ with query complexity Q and tolerance τ if, for every $c^* \in \mathcal{C}$ and distribution D, $\mathscr{A}$ makes Q classical Stat queries with tolerance at least τ, and outputs a hypothesis h that is $1-\varepsilon$ close to c^* under D, i.e., $\Pr_{x\sim D}[h(x) = c^*(x)] \geq 1 - \varepsilon$.[3]

We extend this learning model to allow the algorithm to make *quantum statistical queries*.

Definition 3.1 *Let $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$ be a concept class and $D : \{0,1\}^n \to [0,1]$ be a distribution. A* quantum statistical query oracle $\mathsf{Qstat}(M, \tau)$ *for some $c^* \in \mathcal{C}$ receives as inputs a tolerance parameter $\tau \geq 0$ and an* observable $M \in (\mathbb{C}^2)^{\otimes n+1} \times (\mathbb{C}^2)^{\otimes n+1}$ *satisfying $\|M\| \leq 1$, and outputs a number α satisfying*

$$\left|\alpha - \langle \psi_{c^*} | M | \psi_{c^*} \rangle\right| \leq \tau,$$

where $|\psi_{c^*}\rangle = \sum_{x\in\{0,1\}^n} \sqrt{D(x)} |x, c^*(x)\rangle$.

Observe that Qstat generalizes the classical $\mathsf{Stat}(\phi, \tau)$: if we take the diagonal matrix

$$M = \sum_{z\in\{0,1\}^n} \phi(z, c(z)) |z, c(z)\rangle\langle z, c(z)|,$$

then $\mathsf{Qstat}(M, \tau)$ outputs a number $\alpha \in \mathbb{R}$ that is τ-close to $\mathbb{E}_{x\sim D}[\phi(x, c(x))]$, as in the classical case. Allowing M to be an arbitrary quantum observable lets the QSQ learning algorithm to acquire a broader range of statistics from the coherent superposition of labeled examples.

Definition 3.2 *Let $\mathcal{C}$ be a concept class and $D : \{0,1\}^n \to [0,1]$ be a distribution. We stay that $\mathcal{C}$ can be ε-learned in the quantum statistical query model with Q queries, if there is an algorithm $\mathscr{A}$ such that for every $c^* \in \mathcal{C}$, $\mathscr{A}$ makes at most Q* Qstat *queries and outputs a hypothesis h satisfying $\Pr_{x\sim D}[h(x) \neq c^*(x)] \leq \varepsilon$.*

We justify this model as follows. In the classical case, one can think of the input ϕ to the Stat oracle as a specification of a *statistic* about the distribution of examples $(x, c^*(x))$, and the output of the Stat oracle is an *estimation* of ϕ: one can imagine that the oracle receives i.i.d. labeled examples $(x, c^*(x))$ and empirically computes an estimate of ϕ, which is then forwarded to the learning algorithm. In the quantum setting, one can imagine the analogous situation where the

[3]Note that in the SQ model, there is no "δ"-parameter associated to a learner, i.e., we require a SQ learner to always output a hypothesis h that is ε-close to c under D.

oracle receives copies of the quantum example state $|\psi_{c^*}\rangle$, and performs a measurement indicated by the *observable* M on each copy and outputs an estimate of $\langle\psi_{c^*}|M|\psi_{c^*}\rangle$. We emphasize that the learning algorithm is still a *classical* randomized algorithm and only receives statistical estimates of measurements on quantum examples. Similar to the PAC setting, we are also interested in the sample and time complexity of learning concept classes in the quantum statistical model.

Definition 3.3 *Let $\mathcal{C}$ be a concept class and $D : \{0,1\}^n \to [0,1]$ be a distribution. We define $\mathsf{QSQ}_\varepsilon(c,D)$ as the minimal number of* Qstat *queries that a learner $\mathscr{A}$ needs to make to ε-learn c. We define the* statistical query complexity *of $\mathcal{C}$ as*

$$\mathsf{QSQ}_\varepsilon(\mathcal{C}) = \max_{c\in\mathcal{C}} \max_D \mathsf{QSQ}_\varepsilon(c,D).$$

We say that $\mathcal{C}$ can be ε-learned in polynomial time (polynomial with respect to the precision $1/\tau$, the error parameter $1/\varepsilon$ and the description size of $\mathcal{C}$) under the distribution D in the QSQ *model if there is a polynomial time algorithm $\mathscr{A}$ that ε-learns $\mathcal{C}$ under D. We say that $\mathcal{C}$ can be ε-learned in polynomial time if the learning algorithm $\mathscr{A}$ works for every distribution $D : \{0,1\}^n \to [0,1]$.*[4]

When the bias is not explicitly mentioned, $\mathsf{QSQ}(\mathcal{C})$ denotes the statistical query complexity of learning $\mathcal{C}$ with bias $\varepsilon = 1/3$.

Like in the classical case, our first observation is that if there exists an efficient quantum statistical learning algorithm using Qstat queries then there also exists a quantum learning algorithm in the standard and noisy PAC setting.

Theorem 3.4 *Let $\mathcal{C}$ be a concept class. Let $\tau,\delta > 0$ and $\eta < \max\{1/2, 2\tau^2\}$. Suppose there exists an ε-*QSQ *algorithm that makes Q* Qstat *queries with tolerance at least τ. Then,*

1. *there exists a (ε,δ)-quantum PAC learner for $\mathcal{C}$ that uses $O(\tau^{-2} Q \log(Q/\delta))$ quantum examples.*
2. *there exists a η-noisy (ε,δ)-quantum PAC learner for $\mathcal{C}$ that uses $O((\tau - \sqrt{\eta/2})^{-2} Q \log(Q/\delta))$ many η-noisy quantum examples.*

Proof. Suppose there exists a quantum statistical algorithm that makes the Q Qstat queries $\{(M_1,\tau),\ldots,(M_Q,\tau)\}$ and the output of Qstat queries are $\{\alpha_1,\ldots,\alpha_Q\}$, where

$$\Big|\alpha_i - \langle\psi_c|M_i|\psi_c\rangle\Big| \le \tau,$$

and $|\psi_c\rangle = \sum_x \sqrt{D(x)}|x,c(x)\rangle$.[5] We now prove the first statement in the theorem. Consider a quantum PAC learner that does the following: for every $i \in [Q]$, a quantum learner obtains $T = \log(Q/\delta)/\tau^2$ copies of $|\psi_c\rangle$ and measures each of them according to M_i with outcomes $\{a_1^i,\ldots,a_T^i\} \in [-1,1]^T$. The quantum PAC learner simply passes $\beta_i = \frac{1}{T}\sum_{j=1}^T a_j^i$ to the QSQ learner. By a Chernoff bound, observe that

$$\Pr\Big[\Big|\beta_i - \langle\psi_c|M_i|\psi_c\rangle\Big| \le \tau\Big] \ge 1 - \frac{\delta}{Q} \qquad \text{for every } i \in [Q],$$

[4]Here, by polynomial-time algorithm, we mean the number of gates used in the quantum algorithm is polynomial in the relevant parameters.

[5]We remark that we consider non-adaptive queries for simplicity and that our argument works even if the QSQ learner makes Q *adaptive* Qstat queries.

where the probability is taken over the randomness in measurement. By the union bound over all Q, with probability at least $1-\delta$, the QSQ learner obtains $\{\beta_1,\ldots,\beta_Q\}$ which are the response Q Qstat queries up to precision τ. Hence the QSQ learner (and the quantum PAC learner) outputs a hypothesis h that satisfies $\Pr_{x\sim D}[h(x)=c(x)]\geq 2/3$. The total number of quantum examples used by quantum PAC learner is $Q\cdot\log(Q/\delta)/\tau^2$.

We now prove the second statement in the theorem. Consider the case where an η-noisy quantum PAC learner is given $T=(\tau-\sqrt{\eta/2})^{-2}\log(Q/\delta)$ noisy quantum examples $\bigotimes_{j\in[T]}|\widehat{\psi_j}\rangle$ for each query $i\in[Q]$, where each $|\widehat{\psi_j}\rangle$ is a fresh noisy example as described in Equation (1). The noisy PAC learner behaves like a standard PAC learner, for every $i\in[Q]$, the learner measures each of the T copies according to M_i, obtains $\{\alpha_1^i,\ldots,\alpha_T^i\}$ and passes $\beta_i'=\frac{1}{T}\sum_{j=1}^T a_j^i$ to the QSQ learner. Before we analyze the difference between β_i' and $\langle\widehat{\psi_c}|M_i|\widehat{\psi_c}\rangle$, we first observe that for every M_i satisfying $\|M_i\|\leq 1$ and for every $j\in[T]$, we have

$$\mathbb{E}\Big[\big|\langle\psi_c|M_i|\psi_c\rangle-\langle\widehat{\psi_j}|M_i|\widehat{\psi_j}\rangle\big|\Big]\leq\mathbb{E}\Big[\big\||\widehat{\psi_j}\rangle\langle\widehat{\psi_j}|-|\psi_c\rangle\langle\psi_c|\big\|\Big]=\mathbb{E}\Big[\sqrt{1-\langle\widehat{\psi_j}|\psi_c\rangle}\Big]=\sqrt{1-\sqrt{1-\eta}}\leq\sqrt{\eta/2},$$

where we use the definition of trace distance in the first inequality and $\sqrt{1-x}\leq x/2$ for $x\leq 1$ in the last inequality. Hence for every i,j, we have

$$\mu:=\mathbb{E}\Big[\langle\widehat{\psi_j}|M_i|\widehat{\psi_j}\rangle\Big]\in\Big[\langle\psi_c|M_i|\psi_c\rangle-\sqrt{\eta/2},\langle\psi_c|M_i|\psi_c\rangle+\sqrt{\eta/2}\Big],\tag{2}$$

and such value is *independent* of j. Using a Chernoff bound over the T noisy quantum examples, we have

$$\Pr\Big[\big|\beta_i'-\mu\big|\leq\tau-\sqrt{\eta/2}\Big]\geq 1-\frac{\delta}{Q},\tag{3}$$

where the probability is taken over the randomness in measurement. In particular, with probability $1-\frac{\delta}{Q}$ we have that

$$\Big|\beta_i'-\langle\psi_c|M_i|\psi_c\rangle\Big|\leq\Big|\beta_i'-\mu\Big|+\Big|\mu-\langle\psi_c|M_i|\psi_c\rangle\Big|\leq\tau-\sqrt{\eta/2}+\sqrt{\eta/2}=\tau,\tag{4}$$

where the first inequality used the triangle inequality and the last inequality used Equations (2) and (3). We now use the same argument as the PAC setting to argue that with probability at least $1-\delta$, a QSQ learner which obtains $\{\beta_1',\ldots,\beta_Q'\}$ will output h that satisfies $\Pr_{x\sim D}[h(x)=c(x)]\geq 2/3$, hence proving the theorem statement. The total number of quantum examples used by quantum PAC learner is $O((\tau-\sqrt{\eta/2})^{-2}Q\log(Q/\delta))$. □

4 Learning concept classes quantum efficiently

In this section, we show how to quantum-efficiently learn concept classes in the QSQ model that are provably hard to learn in the classical SQ model. Our key technical tool that will lead to such learning algorithms is a procedure to estimate the Fourier mass of a concept c on a subset of $\{0,1\}^n$ using a single Qstat.

Lemma 4.1 *Let $f:\{0,1\}^n\to\{-1,1\}$ and $|\psi_f\rangle=\frac{1}{\sqrt{2^n}}\sum_x|x,f(x)\rangle$. There is a procedure that on input $T\subseteq\{0,1\}^n$, outputs a τ-estimate of $\sum_{S\in T}\widehat{f}(S)^2$ using one* Qstat *query with tolerance τ.*

Proof. Let $M = \sum_{S\in T} |S\rangle\langle S|$. The observable used in Qstat is then

$$M' = \mathsf{H}^{\otimes(n+1)} \cdot \left(\mathbb{I}^{\otimes n} \otimes |1\rangle\langle 1|\right) \cdot M \cdot \left(\mathbb{I}^{\otimes n} \otimes |1\rangle\langle 1|\right) \cdot \mathsf{H}^{\otimes(n+1)}.$$

Operationally, M' corresponds to first apply the Fourier transform on $|\psi_f\rangle$, post-selecting on the last qubit being 1 and finally applying M to the first n qubits. In order to see the action of M' on $|\psi_f\rangle$, first observe that $\mathsf{H}^{\otimes(n+1)}|\psi_f\rangle$ yields

$$\frac{1}{\sqrt{2^n}} \sum_x |x, f(x)\rangle \to \frac{1}{2^n} \sum_{x,y} \sum_{b\in\{0,1\}} (-1)^{x\cdot y + b\cdot f(x)} |y, b\rangle.$$

Conditioned on the $(n+1)$-th qubit being 1, we have that the resulting quantum state is $|\psi'_f\rangle = \sum_Q \widehat{f}(Q)|Q\rangle$. Applying M to the resulting state gives us

$$\langle \psi'_f | M | \psi'_f \rangle = \sum_{\substack{R,Q\in\{0,1\}^n \\ S\in T}} \langle S|Q\rangle\langle S|R\rangle \widehat{f}(R)\widehat{f}(Q) = \sum_{S\in T} \widehat{f}(S)^2.$$

Therefore, one Qstat(M', τ) query results in a τ-approximation of $\sum_{S\in T} \widehat{f}(S)^2$. □

We now use this lemma to show efficient QSQ learners for parities, k-juntas and DNFs.

4.1 Learning Parities

We start by showing a polynomial time QSQ learner for parities. Classically, Kearns [Kea98] showed that $2^{\Omega(n)}$ Stat queries (with tolerance at least $2^{-\Omega(n)}$) are necessary to weakly learn parities under the uniform distribution.

Lemma 4.2 *The concept class $\mathcal{C} = \{\chi_s : \{0,1\}^n \to \{0,1\}\}_s$ of parities can be exactly learned with $O(n)$* Qstat *queries with tolerance at least $1/3$ under the uniform distribution.*

Proof. Let $c : \{0,1\}^n \to \{-1,1\}$ be a parity function defined as $c(x) = (-1)^{s\cdot x}$ for an unknown $s \in \{0,1\}^n$. Then it is not hard to see that $\mathrm{Inf}_i(c) = 1$ for all $i \in \mathrm{supp}(s)$ and $\mathrm{Inf}_i(c) = 0$ otherwise. Using this observation, the $O(n)$ query quantum algorithm is straightforward: for every $i \in [n]$, we use Lemma 4.1 to estimate the ith influence using one Qstat query with tolerance $\tau = 1/3$ (we let $T = \{S \subseteq \{0,1\}^n : s_i = 1\}$ in Lemma 4.1, in which case we have $\sum_{S\in T} \widehat{f}(S)^2 = \mathrm{Inf}_i(f)$). Suppose $\mathrm{Inf}_i(f) = 1$, then the outcome of the ith Qstat query is in the interval $[2/3, 4/3]$ and if $\mathrm{Inf}_i(f) = 0$, the outcome of the Qstat query is in the interval $[-1/3, 1/3]$. Given the outcomes of the queries, a quantum learning algorithm can easily learn $s \in \{0,1\}^n$, and hence c exactly. In order to understand why this algorithm can be implemented efficiently observe that for $T = \{S \subseteq \{0,1\}^n : s_i = 1\}$, the corresponding M we need to implement in Lemma 4.1 can be written as

$$M = \sum_{S\in T} |S\rangle\langle S| = |1\rangle\langle 1|_i \otimes \mathsf{H}^{\otimes(n-1)} \cdot \left(|0\rangle\langle 0|^{\otimes [n]\setminus\{i\}}\right) \cdot \mathsf{H}^{\otimes(n-1)}, \tag{5}$$

where the ith qubit is fixed to $|1\rangle\langle 1|$ and the remaining $n-1$ qubits (excluding the ith qubit) can be obtained by applying the Hadamard transform on the $n-1$ qubits. Since M can be implemented using poly(n) gates, one can learn parities quantum efficiently. □

4.2 Learning k-juntas

Using a similar idea to the parities problem, we show how to efficiently learn the class of $O(\log n)$-juntas under the uniform distribution in polynomial time in the QSQ model. The idea of this quantum learning algorithm is to first learn most of k influential variables of a junta using Lemma 4.1 and then approximates all the 2^k Fourier coefficients of the function using 2^k classical Stat queries with tolerance 2^{-k}.

Lemma 4.3 *Let $\mathcal{C}$ be the concept class of k-juntas. Then, $O(n+2^{O(k)})$ many* Qstat *queries with tolerance at least $O(\varepsilon \cdot 2^{-k/2})$ suffices to ε-learn $\mathcal{C}$ under the uniform distribution.*

Proof. Our algorithm is divided into two steps: first, we use $O(n)$ Qstat queries with tolerance $O(\varepsilon/k)$ to learn the variables $i \in [n]$ for which $\mathrm{Inf}_i(f) \geq \varepsilon/k$. Let T be the set of such variables. Next we use classical statistical queries to estimate all the Fourier coefficients $\widehat{f}(V)$ for every $V \subseteq T$ using at most $O(2^k)$ Qstat queries with tolerance $2^{-\Omega(k)}$.

Let c be a k-junta over the variables in $Q \subseteq [n]$ with size $|Q| = k$, i.e., $c(x) = f(x_Q)$ for some arbitrary $f : \{0,1\}^k \to \{0,1\}$. Then, it is not hard to see that for all $i \notin Q$, we have $\mathrm{Inf}_i(c) = 0$. Since the goal of a quantum statistical learner is to ε-learn c, it suffices to obtain the variables $i \in [n]$ whose influence $\mathrm{Inf}_i(f)$ is at least $\varepsilon/2k$. Our quantum learning algorithm proceeds as follows: for every $i \in [n]$, we use Lemma 4.1 to estimate $\mathrm{Inf}_i(f)$ upto precision $\varepsilon/(5k)$. Suppose the outcome of these queries is $\alpha_1, \ldots \alpha_n$, we let

$$T = \Big\{ i \in [n] : \alpha_i \geq \varepsilon/4k \Big\}.$$

First, notice that $T \subseteq Q$, since $\mathrm{Inf}_i(f) = 0$ implies that $\alpha_i \leq \varepsilon/5k$. Observe also that for every $i \in Q \setminus T$, we have $\mathrm{Inf}_i(f) < \varepsilon/(2k)$: in order to see this, suppose $\mathrm{Inf}_i(f) \geq \varepsilon/2k$ for some $i \in Q \setminus T$, then the Qstat query to estimate $\mathrm{Inf}_i(f)$ would produce an α_i such that

$$\alpha_i \geq \mathrm{Inf}_i(f) - \varepsilon/(5k) \geq \varepsilon/4k,$$

but this contradicts the fact that $i \notin T$. Hence T has captured all the variables with high influences. In particular

$$\sum_{i \in [n] \setminus T} \mathrm{Inf}_i(c) = \sum_{i \in Q \setminus T} \mathrm{Inf}_i(f) \leq k \cdot \frac{\varepsilon}{2k} = \varepsilon/2, \tag{6}$$

where the first equality used the fact that $\mathrm{Inf}_i(c) = 0$ for every $i \notin Q$ and the inequality used that there are at most k influential variables, hence $|Q| \leq k$

In the second phase, we ε-approximately learn the junta. In order to do this, we estimate the Fourier coefficients for all subsets of T. For every $V \subseteq T$, we make one Qstat query to approximate $\widehat{f}(V)$ upto error $\sqrt{\varepsilon/2} \cdot 2^{-k/2}$: for $V \subseteq [n]$, let $\phi(x,b) = b \cdot (-1)^{V \cdot x}$ for all $x \in \{0,1\}^n, b \in \{0,1\}$, hence $\mathbb{E}_x[\phi(x,c(x))] = \mathbb{E}_x\Big[c(x) \cdot (-1)^{V \cdot x}\Big] = \widehat{c}(V)$. Overall, it takes $2^{|V|} \leq 2^k$ many Qstat queries to estimate all Fourier coefficients $\{\widehat{c}(V) : V \subseteq T\}$. Once we obtain all these approximations $\{\alpha_V\}_{V \subseteq Q}$, we output the function

$$g(x) = \mathrm{sign}\Big(\underbrace{\sum_{V \subseteq T} \alpha_V \cdot \chi_V(x)}_{:=h(x)} \Big), \quad \text{for every } x \in \{0,1\}^n.$$

We now argue that g is in fact ε-close to c:

$$\begin{aligned}\Pr_{x\in\{0,1\}^n}[c(x)\neq g(x)] &= \mathbb{E}_x[c(x)\neq \mathrm{sign}(h(x))]\\ &\leq \mathbb{E}_x[|c(x)-h(x)|^2]\\ &= \sum_V (\widehat{h}(V)-\widehat{c}(V))^2 = \sum_{V\subseteq T}(\alpha_V-\widehat{c}(V))^2+\sum_{V\subseteq[n]\setminus T}\widehat{c}(V)^2 \qquad (7)\\ &\leq 2^k\cdot\frac{\varepsilon}{2^{k+1}}+\sum_{i\in[n]\setminus T}\mathrm{Inf}_i(c)\leq\varepsilon,\end{aligned}$$

where $[\cdot]$ is the indicator of an event, the second equality used Plancherel's identity to conclude $\mathbb{E}_x(c(x)-h(x))^2=\sum_V(\widehat{c}(V)-\widehat{h}(V))^2$, the second inequality follows by definition of $\mathrm{Inf}_i(c)=\sum_{S\subseteq[n]:S\ni i}\widehat{c}(S)^2$ and the fact that Qstat queries return α_V which are a $\varepsilon/2\cdot 2^{-k/2}$ approximation of $\widehat{c}(V)$, and finally the last inequality used Eq. (6). For the same reason as in Lemma 4.2, phase 1 can be performed quantum-efficiently (since the Ms can be expressed as Eq. (5) which takes $\mathrm{poly}(n)$ gates to implement) and phase 2 takes time polynomial in $n, 2^k$ since each ϕ can be computed in time $O(n)$ and we make 2^k many Stat queries. □

Notice that classically, every SQ learner for k-juntas needs to make $n^{\Omega(k)}$ Stat queries with tolerance at least $n^{-\Omega(k)}$, since this class contains at least $\binom{n}{k}$ distinct parity functions.

4.3 Learning Disjunctive Normal Forms (DNFs)

Finally, we give a polynomial time learning for learning $\mathrm{poly}(n)$-sized DNFs in the QSQ model. Classically we need $n^{\Omega(\log n)}$ classical Stat queries (with tolerance $n^{-\Omega(\log n)}$) to learn DNFs (since $\mathrm{poly}(n)$-sized DNFs contain $O(\log n)$-juntas which in turn contain at least $n^{O(\log n)}$ distinct parity functions).

The key step of the proof is to replace the membership queries in the well-known Goldreich-Levin (GL) algorithm [GL89, KM93] by quantum statistical queries. In particular, for a function $c:\{0,1\}^n\to\{-1,1\}$, our "quantum statistical" GL algorithm makes $\mathrm{poly}(n,1/\tau)$ Qstat queries with tolerance at least τ and returns a set $U=\{T_1,\ldots,T_\ell\}\subseteq[n]$ such that if $|\widehat{c}(T)|\geq\tau$, then $T\in U$, and if $T\in U$, we have $|\widehat{c}(T)|\geq\tau/2$. Using this subroutine, for an s-term DNF we can find all the Fourier coefficients which satisfy $|\widehat{c}(T)|\geq 1/s$ using $\mathrm{poly}(n,s)$ many Qstat queries. After this, one can use the classical algorithm for DNF learning by [Fel12] in order to approximate the s-term DNF. Overall our quantum statistical oracle uses $\mathrm{poly}(n)$ Qstat queries of tolerance $1/\mathrm{poly}(n)$ to learn $\mathrm{poly}(n)$-sized DNF formulas.

In order to prove the main lemma statement, we first argue that, in the classical Goldreich-Levin algorithm (GL algorithm) [GL89, KM93], we can replace *classical membership queries* by *quantum statistical queries*.

Theorem 4.4 (Goldreich-Levin theorem using Qstat queries) *Let $f:\{-1,1\}^n\to\{-1,1\}$, $\tau\in(0,1]$. There exists a $\mathrm{poly}(n,1/\tau)$-time quantum statistical learning algorithm that with high probability outputs $U=\{T_1,\ldots,T_\ell\}\subseteq[n]$ such that: (i) if $|\widehat{f}(T)|\geq\tau$, then $T\in U$; and (ii) if $T\in U$, then $|\widehat{f}(T)|\geq\tau/2$.*

We do not prove Theorem 4.4 here since it follows the classical GL algorithm almost exactly.[6] Instead, we only state the difference between the proofs of classical GL algorithm and Theorem 4.4.

[6] An interested reader is referred to Section 3.5 of [O'D14] for details.

In classical GL algorithm, one uses classical membership queries to perform the following task in time $\mathrm{poly}(n, 1/\varepsilon)$: let $Q \subseteq [n]$, for every $S \subseteq Q$, estimate $\sum_{V \subseteq Q^c} \widehat{c}(S \cup V)^2$ upto precision ε. Instead, in the proof of Theorem 4.4, we simply use Lemma 4.1 to estimate $\sum_{V \subseteq Q^c} \widehat{c}(S \cup V)^2$ using one quantum statistical query $\mathrm{Qstat}(c, \varepsilon)$ (we use Lemma 4.1 by setting $T = \{(S \cup V) : V \subseteq Q^c\}$ for a fixed S). The remaining part of GL algorithm, as well as the proof of Theorem 4.4, does not involve membership queries to c. Observe again that for a fixed S and $T = \{(S \cup V) : V \subseteq Q^c\}$, we can write M in Lemma 4.1 as

$$M = \sum_{R \in T} |R\rangle\langle R| = |1\rangle\langle 1|_S \otimes |0\rangle\langle 0|_{Q \setminus S} \otimes \mathsf{H}^{\otimes (n-|Q|)} \cdot \left(|0\rangle\langle 0|^{\otimes [n] \setminus Q} \right) \cdot \mathsf{H}^{\otimes (n-|Q|)},$$

wherein for all $i \in S$, we fix the ith qubit to $|1\rangle\langle 1|$, for $j \in Q \setminus S$ we set the jth qubit to be $|0\rangle\langle 0|$ and the remaining $n - |Q|$ qubits can be obtained by applying the Hadamard transform on the $n - |Q|$ qubits. Clearly such Ms can be implemented quantum efficiently using $\mathrm{poly}(n)$ gates. We now prove our main lemma using Theorem 4.4.

Lemma 4.5 *Let $\mathcal{C}$ be the concept class of* $\mathrm{poly}(n)$*-sized DNFs. Then there exists a* $\mathrm{poly}(n)$*-query* QSQ *algorithm that ε-learns $\mathcal{C}$ under the uniform distribution.*

Proof. Our quantum DNF learning algorithm follows the same ideas as the classical DNF learning by Feldman [Fel12]. We simply replace the classical membership queries in Feldman's algorithm by quantum statistical queries.[7] The only use of membership queries in Feldman's algorithm (in particular, Corollary 5.1 of [Fel12]) is to run GL algorithm to collect all the "large" Fourier coefficients of low Hamming weight: i.e., for an s-term DNF c, Feldman's learning algorithm uses membership queries to find all the Ss that satisfy $|\widehat{c}(S)| \geq \Omega(\varepsilon/s)$. In order to collect such Ss, we use GL (see Theorem 4.4) which makes $\mathrm{poly}(n, s/\varepsilon)$ quantum statistical queries to find all the heavy Fourier coefficients of c and discard those coefficients with large Hamming weight. The remaining part of the Feldman's algorithm in order to ε-learn c does not require membership queries to c and our quantum learner simply continues with Feldman's algorithm. The overall running time of Feldman's algorithm and our quantum learning algorithm is $\mathrm{poly}(n, s/\varepsilon) = \mathrm{poly}(n/\varepsilon)$ since we are concerned with $S = \mathrm{poly}(n)$-sized DNFs. ◻

5 Statistical query dimension

In a seminal work, Blumer et al. [BEHW89] showed that sample complexity of PAC learning is characterized by a combinatorial parameter called VC dimension (which was defined by Vapnik and Chervonenkis [VC71]). Similarly, Blum et al. [BFJ+94] introduced a combinatorial parameter called *statistical query dimension* that characterizes the sample complexity of weak statistical query learning.[8] Roughly the statistical query dimension for a concept class $\mathcal{C}$ and distribution D measures the maximum number of concepts in $\mathcal{C}$ that are nearly uncorrelated with respect to D. Let us define it more formally.

[7] Alternatively, we could have used the weak-quantum learning for DNFs from [BJ95], followed by the statistical query *boosting* algorithm by [AD98].

[8] Here, "weak" refers to the fact that the output hypothesis h of the learner needs to weak-approximate the target concept c^* under the unknown distribution D, i.e., $\Pr_{x \sim D}[h(x) = c^*(x)] \geq 1/2 + 1/\mathrm{poly}(n)$.

Definition 5.1 *Let $\mathcal{C}$ be a concept class and D be a distribution. Then* $\mathsf{WeakSQDIM}(\mathcal{C}, D)$ *is defined as the largest d such that there exists $\{c_1, \ldots, c_d\} \subseteq \mathcal{C}$ such that $|\mathbb{E}_{x\sim D}[c_i(x) \cdot c_j(x)]| \leq \frac{1}{d}$ for every $i \neq j$. We define* $\mathsf{WeakSQDIM}(\mathcal{C}) = \max_D\{\mathsf{WeakSQDIM}(\mathcal{C}, D)\}$.

Using this combinatorial quantity, Blum et al. [BFJ+94] showed the following characterization.

Theorem 5.2 (Blum et al. [BFJ+94]) *Let $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$ be a concept class and D be a distribution. Suppose* $\mathsf{WeakSQDIM}(\mathcal{C}, D) = d$.

- *There exists a* SQ *algorithm for learning $\mathcal{C}$ under D, with error $\frac{1}{2} - \frac{1}{3d}$, that makes d* Stat *queries each with tolerance at least $1/(3d)$.*
- *If all* Stat *queries are made with tolerance $\geq d^{-1/3}$, then at least $d^{1/3}$ queries to the* Stat *oracle is necessary in order to weakly* SQ *learn $\mathcal{C}$.*

Subsequently, there have been many works [Szö09, Yan01, Fel17] that generalized and strengthened $\mathsf{WeakSQDIM}(\mathcal{C}, D)$ in order to characterize other variants of SQ learning model. We do not define these strengthened combinatorial parameters since we will not be using them.

We now show that that for every concept class $\mathcal{C}$, distribution D and tolerance $\tau > 0$, every learner needs to make $\log_{1/\tau}(\mathsf{WeakSQDIM}(\mathcal{C}, D))$ many Qstat queries with tolerance at least τ in order to learn $\mathcal{C}$ under D with error at most $1/2 - 1/d$.

Lemma 5.3 *Let $\tau > 0$. Let $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{-1,1\}\}$ and $D : \{0,1\}^n \to [0,1]$ be a distribution such that* $\mathsf{WeakSQDIM}(\mathcal{C}, D) = d$. *Then, every weak* QSQ *learning algorithm for $\mathcal{C}$ (with error at most $1/2 - 1/d$) needs to make $\Omega(\log_{1/\tau} d)$* Qstat *queries each of tolerance at least τ. Moreover, this lower bound is tight for the class of parity functions on n bits.*

Proof. In order to prove the lemma, we will use the following simple fact: suppose we place d points on the unit interval $[-1,1]$, then one can always find a 2τ-sized ball within the unit interval that covers τd points. In order to see this: suppose by contradiction, assume that *every* 2τ-sized ball in the interval $[-1,1]$ covers strictly lesser than τd points. Then place $1/\tau$ 2τ-sized balls to cover $[-1,1]$. By assumption, since each of the $1/\tau$ 2τ-sized covers $< \tau d$ points, the total number of points in the interval $[-1,1]$ is strictly lesser than d, which contradicts the original assumption that we placed d points in the interval $[-1,1]$.

Let $\mathcal{C}$ be a concept class and D be a distribution satisfying $\mathsf{WeakSQDIM}(\mathcal{C}, D) = d$. By definition, there are d concepts $\mathcal{C}' = \{c_1, \ldots, c_d\}$ such that for every $c_i \neq c_j \in \mathcal{C}'$, we have $\Big|\mathbb{E}_{x\sim D}[c_i(x) \cdot c_j(x)]\Big| \leq 1/d$. We now show that every QSQ learner for $\mathcal{C}$ with bias $1/d$ and tolerance at least 2τ, needs to make $\Omega(\log_{1/\tau} d)$ quantum statistical queries. The proof goes via an adversarial argument, i.e., we show how the replies of a Qstat oracle can enforce a QSQ learner to make $\Omega(\log_{1/\tau} d)$ queries to the Qstat oracle.

Suppose the first Qstat query made by the learner is specified by the operator M_1 and precision 2τ. The adversarial $\mathsf{Qstat}(M_1, 2\tau)$ oracle computes $A_1 = \big(\langle \psi_c | M_1 | \psi_c \rangle\big)_{c \in \mathcal{C}'}$, which contains d values. By the argument in the beginning of the proof, there exists a point x_1 such that at least τd points in A_1 lie within a 2τ-radius of x_1. Then, the $\mathsf{Qstat}(M_1, 2\tau)$ oracle responds with x_1. This narrows down the search space for the learner to at least τd candidate concepts $\mathcal{C}_1 \subseteq \mathcal{C}$. Suppose the next quantum statistical query of the learner is $(M_2, 2\tau)$, the $\mathsf{Qstat}(M_2, 2\tau)$ oracle computes the sequence $A_2 = \left(\langle\psi_c|M_2|\psi_c\rangle\right)_{c\in\mathcal{C}_1}$ with at least τd values and responds with x_2 such that there are at least $\tau^2 d$ points in A_2 around x_2. This process repeats for all the T Qstat queries $\{(M_i, 2\tau) : i \in [T]\}$ made by the QSQ learner.

Suppose $T < \log_{1/(2\tau)} d$ queries. Then after making T queries, there are at least two distinct concepts $c_1, c_2 \in \mathcal{C}'$ that satisfy

$$\Pr_{x\sim D}[c_1(x) \neq c_2(x)] \geq 1/2 - 1/2d \tag{8}$$

and these concepts are consistent with all the T quantum statistical queries queries made so far. Let h be the output of the quantum learner. Given Eq. (8), it must be the case that either $\Pr_{x\sim D}[c_1(x) \neq h(x)] \geq 1/2 - 1/4d$ or $\Pr_{x\sim D}[c_2(x) \neq h(x)] \geq 1/2 - 1/4d$, and we can choose, adversarially, the concept that maximizes such distance. Hence T, the number of queries made by an QSQ learner, needs to be at least $\log_{1/(2\tau)} d$, proving the lower bound.

We now show that this lower bound is tight. Let PARITY_n be the class of parity functions on n bits. For $c, c' \in \mathsf{PARITY}_n$ with $c \neq c'$, under the uniform distribution $\mathcal{U}_n$ we have that $\mathbb{E}_{x\sim\mathcal{U}_n}[c(x)c'(x)] = 0$. Since $|\mathsf{PARITY}_n| = 2^n$, we have $\mathsf{WeakSQDIM}(\mathsf{PARITY}, \mathcal{U}_n) = 2^n$. Along with Lemma 4.2, the lower bound above is tight for PARITY_n under the uniform distribution. □

5.1 Connections to communication complexity

We now present connections between the weak statistical query dimension and communication complexity. Several works [KNR99, JZ09, ANTV99] showed a surprising connection between communication complexity and learning theory: for every $F : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$, the classical and quantum one-way communication complexities of F (under product distributions) are characterized by the VC dimension of the concept class $\mathcal{C}_F = \{F_x : \{0,1\}^n \to \{0,1\} : F_x(y) = F(x,y)\}_{x\in\{0,1\}^n}$. We now prove that weak statistical query dimension of $\mathcal{C}_F$ also lower bounds the complexity in this communication model when ε asymptotically goes to zero.

We now define the model formally. Let $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$ and $\mu : \mathcal{C} \times \{0,1\}^n \to [0,1]$ be a *product distribution*. We consider the following task: (c, x) are picked from $\mathcal{C} \times \{0,1\}^n$ according to μ, and Alice is given as input $c \in \mathcal{C}$ and Bob is given $x \in \{0,1\}^n$. Alice and Bob share random bits and Alice is allowed to send classical bits to Bob, who needs to output $c(x)$ with probability $1/2+\gamma$. We let $\mathsf{R}^{\to,\times}_{1/2+\gamma}(c)$ be the *minimum* number of bits that Alice communicates to Bob, so that he can output $c(x)$ with probability at least $1/2 + \gamma$ (where the probability is taken over the randomness of Bob as well as the distribution μ). Let $\mathsf{R}^{\to,\times}_{1/2+\gamma}(\mathcal{C}) = \max_{c\in\mathcal{C}}\{\mathsf{R}^{\to,\times}_{1/2+\gamma}(c)\}$.

We show that quantum statistical query complexity is an upper bound on $\mathsf{R}^{\to,\times}_{1/2+\gamma}(\mathcal{C})$. The proof follows simply by observing that Alice can simulate the quantum statistical queries and sends the outputs to Bob, who runs the learning algorithm and obtains a hypothesis h that $(1/2+\gamma)$-correlates with the unknown $c \in \mathcal{C}$. Bob then outputs $h(x)$.

Lemma 5.4 *Let $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$ and $\gamma > 0$. Then $\mathsf{R}^{\to,\times}_{1/2+\gamma}(\mathcal{C}) \leq \mathsf{QSQ}_{1/2+\gamma}(\mathcal{C}) \cdot \log(1/\tau)$, where the* QSQ *learner for $\mathcal{C}$ makes queries with tolerance at least $\tau > 0$.*

Proof. Let $\mathsf{QSQ}_{1/2+\gamma}(\mathcal{C}) = d$. For every $c \in \mathcal{C}$ and distribution $D : \{0,1\}^n \to [0,1]$, there exists

$\{(M_i,\tau)\}_{i\in[d]}$ and a QSQ learning algorithm $\mathscr{A}$ such that: given $\alpha_1,\ldots,\alpha_d$ satisfying

$$\Big|\alpha_i - \langle\psi_c|M_i|\psi_c\rangle\Big| \leq \tau \qquad \text{for every } i\in[d], \tag{9}$$

where $|\psi_c\rangle = \sum_x \sqrt{D(x)}|x,c(x)\rangle$, $\mathscr{A}$ can output a hypothesis $h:\{0,1\}^n\to\{0,1\}$ satisfying $\Pr_{x\sim D}[h(x)=c(x)]\geq 1/2+\gamma$.

Consider the product distribution $\mu=\mu_1\times\mu_2$ where $\mu_1:\mathcal{C}\to[0,1]$ and $\mu_2:\{0,1\}^n\to[0,1]$. Suppose Alice receives $c\in\mathcal{C}$ according the distribution μ_1 and Bob obtains $x\in\{0,1\}^n$ from distribution μ_2. Bob now runs the quantum statistical query protocol for the distribution $D=\mu_2$. In order to run the protocol, Alice sends $\alpha_1,\ldots,\alpha_d$ to Bob where αs are defined in Eq. (9) for the state $|\psi_c\rangle=\sum_x\sqrt{\mu_2(x)}|x,c(x)\rangle$ (note that the distribution μ is known both to Alice and Bob explicitly). The total cost of sending α_i is at most $\log(1/\tau)$. Bob uses these αs and obtains a hypothesis h that satisfies $\Pr_{x\sim\mu_2}[h(x)=c(x)]\geq 1/2+\gamma$. Bob then outputs $h(x)$. By the promise of the QSQ algorithm, for every $c\in\mathcal{C}$, we have

$$\Pr_{x\sim\mu_2}[h(x)=c(x)]\geq\frac{1}{2}+\gamma.$$

In particular, this implies $\Pr_{(c,x)\sim\mu}[h(x)=c(x)]\geq\frac{1}{2}+\gamma$. Hence, we have $\mathsf{R}^{\rightarrow,\times}_{1/2+\gamma}(\mathcal{C})\leq d\log(1/\tau)$, thereby proving the lemma statement. □

Similar to $\mathsf{R}^{\rightarrow,\times}_{1/2+\gamma}(\mathcal{C})$, we can define $\mathsf{Q}^{\rightarrow,\times}_{1/2+\gamma}(\mathcal{C})$ as the *quantum communication complexity* of computing $\mathcal{C}$ under product distributions, wherein Alice is allowed to send *quantum bits* to Bob. We observe that $\mathsf{WeakSQDIM}(\mathcal{C})$ can be used to lower bound $\mathsf{Q}^{\rightarrow,\times}_{1/2+\gamma}(\mathcal{C})$.

Lemma 5.5 *Let $\mathcal{C}\subseteq\{c:\{0,1\}^n\to\{0,1\}\}$. For every $\gamma>0$, we have*

$$\Omega(\log(\gamma\cdot\sqrt{\mathsf{WeakSQDIM}(\mathcal{C})}))\leq\mathsf{Q}^{\rightarrow,\times}_{1/2+\gamma}(\mathcal{C}). \tag{10}$$

Proof. The main technical tool in the proof is a combinatorial quantity called *discrepancy*. which we do not define here, but we use its connections to WeakSQDIM and communication complexity. Sherstov [She08] showed that for any $\mathcal{C}$ and $F_{\mathcal{C}}:\mathcal{C}\times\{0,1\}^n\to\{0,1\}$ such that $F(c,x)=c(x)$, we have that

$$\sqrt{\frac{1}{2}\mathsf{WeakSQDIM}(\mathcal{C}_F)}\leq\frac{1}{\mathsf{disc}^{\times}(F)}\leq 8\mathsf{WeakSQDIM}(\mathcal{C}_F)^2.$$

Our lemma statement follows from the result of [Kla07], who showed that for any function $F:X\times Y\to[0,1]$, distribution μ and $\gamma>0$, we have $\mathsf{Q}^{\mu}_{1/2+\gamma}(F)\geq\log_2\left(\frac{\gamma}{\mathsf{disc}_{\mu}(F)}\right)$. □

Prior to our work, Kremer *et al.* [KNR99] and Ambainis *et al* [ANTV99] related the VC dimension and communication complexity by showing that for every concept class $\mathcal{C}$ we have

$$(1-\mathsf{H}_2(\varepsilon))\cdot\mathsf{VC}(\mathcal{C})\leq\mathsf{R}^{\rightarrow,\times}_{\varepsilon}(\mathcal{C})\leq\mathsf{VC}(\mathcal{C}), \tag{11}$$

where $H_2(\cdot)$ is the binary entropy function. In particular for constant ε, they showed the characterization $\mathsf{R}^{\rightarrow,\times}_{\varepsilon}(\mathcal{C})=\Theta(\mathsf{VC}(\mathcal{C}))$. A priori it might seem that the lower bound of $\Omega(\log\mathsf{WeakSQDIM}(\mathcal{C}))$ in Lemma 5.5 is exponentially worse than the upper bound that we get in Lemma 5.4 and might not be useful in comparison to the lower bound in Eq. (11). However note that for every $\mathcal{C}$, the best lower bound that Eq. (11) can yield is $\mathsf{VC}(\mathcal{C}) \leq \log|\mathcal{C}|$, and one can obtain a similar lower bound using our Lemma 5.5 since $\mathsf{WeakSQDIM}(\mathcal{C})$ could be as large as $|\mathcal{C}|$. In fact we show that in the small-error regime, our lower bound can be *exponentially* better than what can get from Eq. (11).

Suppose $\varepsilon = 1/2 + \gamma$ for some $\gamma \ll 1/2$ in Eq. (11). The lower bound scales then as

$$(1 - \mathsf{H}_2(\varepsilon)) \cdot \mathsf{VC}(\mathcal{C}) = \left(1 - \mathsf{H}_2\left(\frac{1}{2} + \gamma\right)\right) \cdot \mathsf{VC}(\mathcal{C}) = \Theta(\gamma^2 \cdot \mathsf{VC}(\mathcal{C})),$$

where we used the Taylor series expansion of $\mathsf{H}_2(\cdot)$ to conclude $\mathsf{H}_2(1/2 + \gamma) = \Theta(\gamma^2)$ for $\gamma \ll 1/2$. Let $\mathcal{C} = \mathsf{PARITY}_n$ and $\gamma = \mathsf{WeakSQDIM}(\mathcal{C})^{-1/3} = 2^{-n/3}$, then Eq. (11) gives us the trivial

$$\mathsf{R}^{\rightarrow,\times}_{\frac{1}{2}+\gamma}(\mathcal{C}) \geq \frac{\mathsf{VC}(\mathcal{C})}{\mathsf{WeakSQDIM}(\mathcal{C})^3} = \Omega(n \cdot 2^{-2n/3}),$$

however Eq. (10) gives us a stronger bound of $\mathsf{Q}^{\rightarrow,\times}_{\frac{1}{2}+\gamma}(\mathcal{C}) \geq \Omega\left(\log\left(\mathsf{WeakSQDIM}(\mathcal{C})^{\frac{1}{6}}\right)\right) = \Omega(n)$. Notice that this allows us to give non-trivial lower bounds on the communication complexity even for *inverse exponential bias*.

6 Quantum learning in a differential private setting

In this section we describe the connections between the QSQ model and private learning. We start with a brief overview of classical differential privacy.

6.1 Differential privacy

Differential privacy is an important framework that provides a mathematical model for the notion of privacy of individuals on database queries [Dwo06, DN04, DMNS16, BDMN05]. More concretely, an algorithm $\mathscr{A}$ is said to be α-differentially private if for any two neighbor databases[9] X and X', where two databases are neighbors if they differ in a single position, and for every subset $\mathscr{F}$ of the possible outcomes of $\mathscr{A}$ we have

$$\Pr\left[\mathscr{A}(X) \in \mathscr{F}\right] \leq e^{\alpha} \Pr\left[\mathscr{A}(X') \in \mathscr{F}\right].$$

Given the success of differential privacy (in theory and practice), this notion was extended also to learning algorithms by Kasiviswanathan *et al.* [KLN+08]. In this setting, we extend the requirements of standard PAC learning to require the learning algorithm to be differentially-private. Classically, it is well known that that if a concept class can be learned in statistical query model, it can be private PAC learnable and this connection has provided many consequences (which we do not discuss here, and refer the interested reader to [KLN+08, BBKN14, Vad17] for more on differential privacy and its applications).

6.1.1 Laplacian mechanism

The Laplacian mechanism is a technique used often to ensure that the output of a classical algorithm is differentially-private. The mechanism works as follows: suppose we want to compute

[9]We can see a database X as a string in Σ^n, for some alphabet Σ.

function $f : [0,1]^T \to [0,1]$ whose input variables have small influences, then the Laplacian mechanism first computes f on an input $(x_1,\ldots,x_T)$, then adds noise from the Laplace distribution to $f(x_1,\ldots,x_T)$ and outputs the resulting value.

Definition 6.1 (Laplacian mechanism) *Let $T \geq 1$, $f : [0,1]^T \to [0,1]$ and $a_1,\ldots,a_T \in [0,1]$. The* Laplacian mechanism *for computing f, first computes $a' = f(a_1,\ldots,a_T)$ and outputs $a' + x$ where x is drawn from the Laplacian distribution $D : \mathbb{R} \to [0,1]$ with parameter αn defined as*

$$D_{\alpha\cdot n}(x) = \frac{\alpha n}{2} e^{-|x|\cdot\alpha n},$$

where $|x|$ is the absolute value of x.

The output of the Laplacian mechanism can be shown to compute f in a differentially private manner. In particular, it is well-known that it can be used to compute the average of numbers (i.e., given $a_1,\ldots,a_T \in \mathbb{R}$, compute $a' = f(a_1,\ldots,a_T) = \frac{1}{T}\sum_{i=1}^{T} a_i$) privately.

Lemma 6.2 *The Laplacian mechanism for computing the average of T numbers $\{a_1,\ldots,a_T\}$ with Laplacian parameter $\alpha \cdot T$ is α-differentially private. Moreover, there exists some universal constant $C > 0$ such that with probability at least $1-\delta$ the output v of the Laplacian mechanism satisfies:*

$$\left|v - \frac{1}{T}\sum_{i=1}^{T} a_i\right| \leq C \cdot \frac{1}{\alpha T}\log\left(\frac{1}{\delta}\right).$$

For a proof of this lemma and additionally applications of the Laplacian mechanism in differential privacy, we refer the reader to [DR14, Section 3.3].

6.2 Private quantum PAC learning

Given the success of differential privacy, its quantum analogue was recently proposed by Aaronson and Rothblum [AR19], which we define now.

Definition 6.3 *Two product states $|\Phi\rangle = |\phi_1\rangle \otimes \cdots \otimes |\phi_n\rangle$. and $|\Psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$ are neighbors* if *there exists at most one $i \in [n]$ such that $|\phi_i\rangle \neq |\psi_i\rangle$. A quantum algorithm $\mathscr{A}$ is α-differential private on some subset S of product states if for all states $|\Phi\rangle, |\Psi\rangle \in S$ that are neighbors and every subset $\mathscr{F}$ of the possible outputs of $\mathscr{A}$ we have that*[10]

$$\Pr[\mathscr{A}(|\Psi\rangle) \in \mathscr{F}] \leq e^{\alpha} \Pr[\mathscr{A}(|\Phi\rangle) \in \mathscr{F}].$$

Inspired by this definition, we now define private learning a concept class.

Definition 6.4 *Let $\mathcal{C}$ be a concept class. We say a $\mathscr{A}$ is a* $(\alpha,\varepsilon,\delta)$-differentially private quantum PAC *learning algorithm for $\mathcal{C}$ with sample complexity T if $i)$ $\mathscr{A}$ is α-differentially private and $ii)$ for every distribution $D : \{0,1\}^n \to [0,1]$, $\mathscr{A}$ uses T copies of $|\psi_c\rangle = \sum_x \sqrt{D(x)}|x, c(x)\rangle$, and with probability at least $1-\delta$ outputs h such that $\Pr_{x\sim D}[h(x) \neq c(x)] \leq \varepsilon$.*

[10]Following [BNS13], we define the stronger notion of privacy where the probabilities are close not only for every possible output, but also for every *subset* of outputs.

In classical literature it is well-known that if a concept class is learnable in the SQ model, then it can also be learned privately in the PAC learning model. We now show that this implication also holds true in the quantum case. The proof follows similarly to Theorem 3.4: we use $O(\tau^{-2})$ quantum examples to simulate a Qstat oracle with tolerance τ and then, to ensure privacy the of each query, we use the well-known Laplacian mechanism (see Section 6.1.1) in the simulation of Qstat by quantum examples.

Theorem 6.5 *Let $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$. If there exists a learning algorithm that ε-learns $\mathcal{C}$ using d* Qstat *queries with tolerance at least τ, then the quantum sample complexity of $(\alpha, \varepsilon, \delta)$-private quantum PAC learning $\mathcal{C}$ is $O\Big(\Big(\frac{d}{\tau^2} + \frac{d}{\varepsilon\tau}\Big) \cdot \log\Big(\frac{d}{\beta}\Big)\Big)$.*

Proof. The proof here is similar to the proof of Theorem 3.4 where we showed quantum SQ learnability implies quantum PAC learnability. Suppose $\mathsf{QSQ}(\mathcal{C}) = d$. For every $c \in \mathcal{C}$ and distribution $D : \{0,1\}^n \to [0,1]$: suppose the QSQ learner $\mathscr{A}$ makes the queries $\{(M_i, \tau)\}_{i\in[d]}$ and obtains $\alpha_1, \ldots, \alpha_d$ satisfying

$$\Big|\alpha_i - \langle \psi_c | M_i | \psi_c \rangle\Big| \leq \tau \qquad \text{for every } i \in [d],$$

for $|\psi_c\rangle = \sum_x \sqrt{D(x)}|x, c(x)\rangle$, then $\mathscr{A}$ outputs a hypothesis h satisfying $\Pr_{x\sim D}[h(x) \neq c(x)] \leq \eta$. Let

$$Q = C\alpha^{-1} \cdot \Big(\frac{1}{\tau^2} + \frac{2}{\tau}\Big) \cdot \log\Big(\frac{2d}{\delta}\Big)$$

where C is the constant defined in Theorem 6.2. Consider a quantum PAC learner that for every $i \in [d]$, the learner obtains Q many (fresh) quantum examples $|\psi_c\rangle$, which are measured according to the observable M_i with outcomes $a_1^i, \ldots, a_Q^i$. The learner then applies the Laplacian mechanism LM to compute the average of $a_1^i, \ldots, a_Q^i$: first compute $b^i = \sum_{j=1}^{Q} a_j^i$ and then apply to b^i the Laplacian noise with parameter $\alpha \cdot Q$, resulting in $\tilde{b}^i$ (see Definition 6.1). The quantum PAC learner feeds the QSQ learner with $\{\widetilde{b}^1, \ldots, \widetilde{b}^d\}$, and outputs the hypothesis h provided by the QSQ learner. The sample complexity of this PAC learner is $O\Big(d\alpha^{-1} \cdot \Big(\frac{1}{\tau^2} + \frac{2}{\tau}\Big) \cdot \log\Big(\frac{d}{\delta}\Big)\Big)$.

We first analyze the correctness of our quantum PAC learner. Similar to the proof of Theorem 3.4, observe that Q is large enough to ensure that, with probability at least $1 - \delta/(2d)$, we have $|b^i - \langle \psi_c | M_i | \psi_c \rangle| \leq \tau/2$ for every i. Next, by Lemma 6.2, with probability at least $1 - \delta/(2d)$, we have

$$\Big|\tilde{b}^i - b^i\Big| \leq C \cdot \frac{1}{\alpha Q} \cdot \log\Big(\frac{2d}{\delta}\Big) \leq \frac{\tau}{2},$$

where the last inequality used the definition of Q. The difference between the quantum SQ query response and $\tilde{b}^i$ can be bounded using the triangle inequality by

$$|\langle \psi_c | M_i | \psi_c \rangle - \tilde{b}^i| \leq |\langle \psi_c | M_i | \psi_c \rangle - b^i| + |b^i - \tilde{b}^i| \leq \tau.$$

Moreover, by a union bound we have: with probability at least $1 - \delta$, the quantum PAC learner answers all d QSQ queries with error at most τ. Hence, with probability $\geq 1 - \delta$, the output h of the quantum QSQ learner (and hence the quantum PAC learner) satisfies $\Pr_{x\sim D}[h(x) \neq c(x)] \leq \varepsilon$.

We now analyze the privacy of our quantum PAC learner. For that, let us analyze the privacy for computing $\tilde{b}^i$ for some fixed i. Let $\mathscr{Q}$ be the procedure that computes $\tilde{b}^i$ from $|\psi_c\rangle^{\otimes Q}$. It follows that

$$\Pr[\mathscr{Q}\big(|\psi_c\rangle^{\otimes Q}\big) = y] = \Pr_{a_1^i,\ldots,a_Q^i}[\mathsf{LM}\big(\{a_1^i,\ldots,a_Q^i\}\big) = y]$$
$$\leq e^{\alpha}\cdot \Pr_{a_1^i,\ldots,a_Q^i}[\mathsf{LM}\big(\{a_1^i,\ldots,a_{Q-1}^i,w\}\big) = y] = e^{\alpha}\cdot\Pr[\mathscr{Q}\big(|\psi_c\rangle^{\otimes Q-1}\otimes|\phi\rangle\big) = y],$$

where we use that the Laplacian mechanism with our parameters is α-differential private and we assume for simplicity that the (possibly) different entry is the last state in the tensor product.

Notice that the quantum PAC learning algorithm $\mathscr{A}$ receives as input $|\psi_c\rangle^{\otimes Qd}$, and runs $\mathscr{Q}$ d-times in parallel and runs a procedure $\mathscr{S}$ that computes the hypothesis basis on the classical statistics. Let assume again for simplicity that the neighbor $|\Phi\rangle$ of $|\psi_c\rangle^{\otimes Qd}$ has its different entry in the last position. In this case, for any subset of outputs $\mathscr{F}$, we have that

$$\begin{aligned}
&\Pr[\mathscr{A}\big(|\psi_c\rangle^{\otimes Qd}\big)\in\mathscr{F}]\\
&=\Pr[\mathscr{S}(y_1,\ldots,y_d)\in\mathscr{F}]\Pr[\mathscr{Q}\big(|\psi_c\rangle^{\otimes Q}\big)=y_1]\cdots\Pr[\mathscr{Q}\big(|\psi_c\rangle^{\otimes Q}\big)=y_d]\\
&\leq e^{\alpha}\Pr[\mathscr{S}(y_1,\ldots,y_d)\in\mathscr{F}]\Pr[\mathscr{Q}\big(|\psi_c\rangle^{\otimes Q}\big)=y_1]\cdots\Pr[\mathscr{Q}\big(|\psi_c\rangle^{\otimes Q-1}\otimes|\phi\rangle\big)=y_d]\\
&=\Pr[\mathscr{A}\big(|\Phi\rangle\big)\in\mathscr{F}],
\end{aligned}$$

showing that $\mathscr{A}$ is also α-private. □

An immediate corollary of this theorem along with the results in Section 4 is the following (which was not known before).

Corollary 6.6 *Parities, k-juntas and DNFs can be* privately *quantum PAC learned under the uniform distribution.*

6.3 Representation dimensions and private quantum PAC learning

It is well-known that the sample complexity of classical and quantum PAC learning is characterized by VC dimension [BEHW89, Han16, AW18]. Classically, in the private setting, a series of results [BNS13, BBKN14, FX15] showed that the *representational dimension* of the concept class $\mathcal{C}$ ($\mathsf{PRDIM}(\mathcal{C})$) characterizes the sample complexity of *private* PAC learning. Here, we show that $\mathsf{PRDIM}(\mathcal{C})$ also characterizes the sample complexity of private *quantum* PAC learning $\mathcal{C}$.

In order to define the representation dimension of a concept class $\mathcal{C}\subseteq\{c:\{0,1\}^n\to\{0,1\}\}$, we first define the *probabilistic representation* of $\mathcal{C}$ and its *probabilistic representational dimension*.

Definition 6.7 (Representation of concept classes) *A hypothesis class $\mathcal{H}\subseteq\{h:\{0,1\}^n\to\{0,1\}\}$ is an* ε-representation *of $\mathcal{C}$ if for every $c\in\mathcal{C}$ and distribution $D:\{0,1\}^n\to[0,1]$, there exists $h\in\mathcal{H}$ such that $\Pr_{x\sim D}[h(x)\neq c(x)]\leq\varepsilon$.*

Similarly, let $P:[r]\to[0,1]$ be a distribution and $\mathscr{H}=\{\mathcal{H}_1,\ldots,\mathcal{H}_r\}$ be a collection of hypothesis classes. We say $(P,\mathscr{H})$ is an (ε,δ)-probabilistic representation *of $\mathcal{C}$, if for every $c\in\mathcal{C}$ and distribution $D:\{0,1\}^n\to[0,1]$, we have*

$$\Pr_{i\sim P}[\exists h\in\mathcal{H}_i \text{ s.t. } \Pr_{x\sim D}[h(x)\neq c(x)]\leq\varepsilon]\geq 1-\delta.$$

Define $\mathsf{size}(\mathscr{H})=\max\{\log|\mathcal{H}_i|:\mathcal{H}_i\in\mathscr{H}\}$

We are now ready to define the probabilistic representational dimension of a concept class.

Definition 6.8 (Representational dimension [BBKN14, BNS13]) *Let $\mathcal{C} \subseteq \{c : \{0,1\}^n \to \{0,1\}\}$ be a concept class. The (ε, δ) probabilistic representational dimension of $\mathcal{C}$, $\mathsf{PRDIM}(\mathcal{C})$ is defined as*

$$\min\Big\{\mathsf{size}(\mathscr{H}) : \textit{ there exists } (P, \mathscr{H}) \textit{ that } (\varepsilon, \delta) - \textit{probabilistically represents } \mathcal{C}\Big\},$$

We now show that for every concept class $\mathcal{C}$, the quantum sample complexity of *private* PAC learning $\mathcal{C}$ is characterized by the representation dimension of a concept class. Since $\mathsf{PRDIM}(\mathcal{C})$ is an upper-bound to the *classical* sample complexity of private PAC learning (which in its turn is an upper bound to the quantum sample complexity[11]), we only need to show that $\mathsf{PRDIM}(\mathcal{C})$ is also a *lower bound* on the quantum sample complexity of quantum private PAC learning. Together with the corresponding classical characterization [BBKN14, BNS13] (which inspires our proof), our result implies that quantum and classical sample complexities of private PAC learning are equal, up to constant factors.

Theorem 6.9 *If there exists an $(\alpha, \varepsilon, \delta)$-quantum private PAC learner for a concept class $\mathcal{C}$ with sample complexity T, then the (ε, β)-probabilistic dimension* $\mathsf{PRDIM}(\mathcal{C}) = O(T\alpha + +\log\log 1/\beta)$.[12]

Proof. Let $\mathscr{A}$ be a $(\alpha, \varepsilon, 1/2)$-quantum private learning algorithm for $\mathcal{C}$ using a hypothesis class $\mathcal{F}$, with sample complexity T. Fix $c \in \mathcal{C}$ and distribution D and define $\mathcal{F}' \subseteq \mathcal{F}$ as $\mathcal{F}' = \{h \in \mathcal{F} : \Pr_{x\sim D}[c(x) \neq h(x)] \leq \varepsilon\}$. By the "$\delta$-learning promise" of $\mathscr{A}$, we know

$$\Pr\Big[\mathscr{A}\Big(|\psi_c\rangle^{\otimes T}\Big) \in \mathcal{F}'\Big] \geq 1 - \delta, \tag{12}$$

where the probability is taken with respect to the randomness of $\mathscr{A}$. Let $|\psi_{\mathbf{0}}\rangle = \sum_x \sqrt{D(x)}|x, 0\rangle$. The α-quantum differential privacy of $\mathscr{A}$ implies that

$$\begin{aligned}\Pr[\mathscr{A}\Big(|\psi_{\mathbf{0}}\rangle^{\otimes T}\Big) \in \mathcal{F}'] &\geq e^{-\alpha} \cdot \Pr[\mathscr{A}\Big(|\psi_{\mathbf{0}}\rangle^{\otimes T-1} \otimes |\psi_c\rangle\Big) \in \mathcal{F}'] \\ &\geq e^{-2\alpha} \Pr[\mathscr{A}\Big(|\psi_{\mathbf{0}}\rangle^{\otimes T-2} \otimes |\psi_c\rangle^{\otimes 2}\Big) \in \mathcal{F}'] \geq \cdots \geq e^{-T\alpha} \cdot \Pr[\mathscr{A}\Big(|\psi_c\rangle^{\otimes T}\Big) \in \mathcal{F}']\end{aligned}$$

which is at least $(1-\delta)e^{-T\alpha}$ using Eq. (12). In particular, we have that $\Pr\Big[\mathscr{A}\Big(|\psi_{\mathbf{0}}\rangle^{\otimes T}\Big) \notin \mathcal{F}'\Big] \leq 1 - (1-\delta)e^{-T\alpha}$. Suppose, we run $\mathscr{A}$ $K = \ln(1/\beta) \cdot e^{T\alpha}/(1-\delta)$ many times on input $|\psi_{\mathbf{0}}\rangle^{\otimes T}$, and let $\mathcal{H}$ be the set of the outcomes of $\mathscr{A}$ on each execution. The probability that $\mathcal{H}$ does not contain an ε-good hypothesis is at most

$$\Big(1 - (1-\delta) \cdot e^{-T\alpha}\Big)^K \leq \exp(-K(1-\delta)e^{-T\alpha}) \leq \beta,$$

using $(1-x)^t \leq e^{-xt}$ in the first inequality and the definition of K in the second inequality. Let $\widetilde{\mathcal{F}} \subseteq \mathcal{F}$ be the set of hypothesis that have a non-zero probability of being output when $\mathscr{A}$ is given

[11] In particular, this inequality holds because the following algorithm is a *private quantum learner*: suppose a quantum learner obtains T quantum examples, measures each quantum example in the computational basis and then runs the classical private learning algorithm on the T classical examples. This quantum algorithm satisfies the conditions of quantum differential privacy, because a neighboring quantum state that is provided to the quantum learner will result in neighboring classical examples and by assumption we know that the classical learner is differentially private.

[12] One can further prune this bound to get the ε dependence in the upper bound on $\mathsf{PRDIM}(\mathcal{C})$ by using ideas in [BNS13, Lemma 3.16], we omit it here.

the input $|\psi_{\mathbf{0}}\rangle^{\otimes T}$. Let also $\mathscr{H} = \left\{\mathcal{H} \subseteq \widetilde{\mathcal{F}} : |\mathcal{H}| \le \ln(1/\beta)\cdot e^{T\alpha}/(1-\delta)\right\}$ and P be the uniform distribution over all $\mathcal{H} \in \mathscr{H}$. Then $(P, \mathscr{H})$ is an (ε, β)-probabilistic representation for the concept class $\mathcal{C}$ and it follows that

$$\mathsf{PRDIM}(\mathcal{C}) \le \max_{\mathcal{H} \in \mathsf{supp}(\mathscr{H})} \{\ln |\mathcal{H}|\} \le O(T\alpha + \log\log 1/\beta),$$

which proves the theorem statement. □

7 Open questions and future work.

We now conclude with a few open questions. In our definition of QSQ, we have a classical randomized learner and one could possibly consider a general definition of QSQ model wherein the algorithm can make *quantum superposition queries* to the oracle, or ask the oracle to perform joint, entangling measurements on multiple copies of $|\psi_{c^*}\rangle$.

Secondly, Bun and Zhandry [BZ16] showed that classical PAC learning is strictly more powerful than its private version under cryptographic assumptions. We leave understanding if such a separation also works in a (post-) quantum scenario as an open question.

References

[ACL+19] S. Arunachalam, S. Chakraborty, T. Lee, M. Paraashar, and R. de Wolf. Two new results about quantum exact learning. In *46th International Colloquium on Automata, Languages, and Programming, ICALP*, pages 16:1–16:15, 2019. arXiv:1810.00481. 2

[AD98] J. A. Aslam and S. E. Decatur. General bounds on statistical query learning and PAC learning with noise via hypothesis boosting. *Information and Computation*, 141(2):85–118, 1998. Earlier in FOCS'93. 12

[ANTV99] A. Ambainis, A. Nayak, A. Ta-Shma, and U. V. Vazirani. Dense quantum coding and a lower bound for 1-way quantum automata. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, pages 376–383, 1999. 14, 15

[AR19] S. Aaronson and G. N. Rothblum. Gentle measurement of quantum states and differential privacy. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pages 322–333, 2019. 17

[AS05] A. Atıcı and R. Servedio. Improved bounds on quantum learning algorithms. *Quantum Information Processing*, 4(5):355–386, 2005. quant-ph/0411140. 2

[AT07] D. Aharonov and A. Ta-Shma. Adiabatic quantum state generation. *SIAM Journal on Computing*, 37(1):47–82, 2007. 2

[AW17] 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

[AW18] S. Arunachalam and R. de Wolf. Optimal quantum sample complexity of learning algorithms. *Journal of Machine Learning Research*, 19(71):1–36, 2018. Earlier version in CCC'17. arXiv:1607.00932. 2, 19

[BBKN14] A. Beimel, H. Brenner, S. P. Kasiviswanathan, and K. Nissim. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94(3):401–437, 2014. 16, 19, 20

[BDMN05] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 128–138, 2005. 2, 16

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989. Earlier in FOCS'88. 12, 19

[BFJ+94] A. Blum, M. L. Furst, J. C. Jackson, M. J. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, pages 253–262, 1994. 12, 13

[BJ95] N. H. Bshouty and J. C. Jackson. Learning DNF over the uniform distribution using a quantum example oracle. In *Proceedings of the Eigth Annual Conference on Computational Learning Theory, COLT*, pages 118–127, 1995. 2, 5, 12

[BKL+19] F. G. S. L. Brandão, A. Kalev, T. Li, C. Yen-Yu Lin, K. M. Svore, and X. Wu. Quantum SDP solvers: Large speed-ups, optimality, and applications to quantum learning. In *46th International Colloquium on Automata, Languages, and Programming, ICALP*, pages 27:1–27:14, 2019. 1

[BNS13] A. Beimel, K. Nissim, and U. Stemmer. Characterizing the sample complexity of private learners. In *Innovations in Theoretical Computer Science, ITCS*, pages 97–110, 2013. 17, 19, 20

[BZ16] M. Bun and M. Zhandry. Order-revealing encryption and the hardness of private learning. In *Theory of Cryptography - 13th International Conference, TCC 2016-A*, pages 176–206, 2016. 21

[CGL+19] N-H. Chia, A. Gilyén, T. Li, H-H. Lin, E. Tang, and C. Wang. Sampling-based sublinear low-rank matrix arithmetic framework for dequantizing quantum machine learning, 2019. arXiv:1910.06151. 1

[DMNS16] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality,*, 7(3):17–51, 2016. 16

[DN04] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In *Advances in Cryptology - CRYPTO 2004, 24th Annual International Cryptology Conference*, pages 528–544, 2004. 16

[DR14] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. 17

[Dwo06] C. Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP*, pages 1–12, 2006. 16

[Fel12] V. Feldman. Learning DNF expressions from Fourier spectrum. In *COLT 2012 - The 25th Annual Conference on Learning Theory*, pages 17.1–17.19, 2012. 11, 12

[Fel16] V. Feldman. Statistical query learning. In *Encyclopedia of Algorithms*, pages 2090–2095. 2016. 2

[Fel17] V. Feldman. A general characterization of the statistical query complexity. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 785–830, 2017. 2, 13

[FX15] V. Feldman and D. Xiao. Sample complexity bounds on differentially private learning via communication complexity. *SIAM J. Comput.*, 44(6):1740–1764, 2015. Earlier in COLT'14. 19

[GKZ19] A. B. Grilo, I. Kerenidis, and T. Zijlstra. Learning with Errors is easy with quantum samples. *Physical Review Letters A*, 99:032314, 2019. arXiv: 1702.08255. 2, 5

[GL89] O. Goldreich and L. A. Levin. A hard-core predicate for all one-way functions. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing*, pages 25–32, 1989. 11

[Han16] S. Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17:38:1–38:15, 2016. 19

[JZ09] R. Jain and S. Zhang. New bounds on classical and quantum one-way communication complexity. *Theoretical Computer Science*, 410(26):2463–2477, 2009. 14

[Kea98] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998. Earlier version in STOC'03. 1, 2, 9

[Kla07] H. Klauck. Lower bounds for quantum communication complexity. *SIAM Journal on Computing*, 37(1):20–46, 2007. 15

[KLN+08] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. D. Smith. What can we learn privately? In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 531–540, 2008. 16

[KM93] E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, 1993. 11

[KNR99] I. Kremer, N. Nisan, and D. Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999. Earlier in STOC'95. 14, 15

[KP17] I. Kerenidis and A. Prakash. Quantum recommendation systems. In *8th Innovations in Theoretical Computer Science Conference, ITCS*, pages 49:1–49:21, 2017. 1

[LMR13] S. Lloyd, M. Mohseni, and P. Rebentrost. Quantum algorithms for supervised and unsupervised machine learning, 2013. arXiv:1307.0411. 1

[O'D14] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. 11

[Pie12] K. Pietrzak. Cryptography from learning parity with noise. In *SOFSEM 2012: Theory and Practice of Computer Science - 38th Conference on Current Trends in Theory and Practice of Computer Science*, 2012. 5

[Reg09] O. Regev. On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM*, 56(6):34:1–34:40, 2009. Earlier in STOC'05. 5

[RML14] P. Rebentrost, M. Mohseni, and S. Lloyd. Quantum support vector machine for big data classification. *Physical Review Letters*, 113:130503, 2014. 1

[She08] A. A. Sherstov. Halfspace matrices. *Computational Complexity*, 17(2):149–178, 2008. Earlier in CCC'07. 15

[Szö09] B. Szörényi. Characterizing statistical query learning: simplified notions and proofs. In *International Conference on Algorithmic Learning Theory*, pages 186–200. Springer, 2009. 13

[Tan19] E. Tang. A quantum-inspired classical algorithm for recommendation systems. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pages 217–228, 2019. 1

[Vad17] S. P. Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. 2017. 16

[Val84] L. Valiant. A theory of the learnable. *Journal of the ACM*, 27(11):11341142, 1984. 2, 5

[VC71] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. English translation of 1968 Russian paper in *Dokl. Akad. Nauk.* 181(4). 12

[Yan01] K. Yang. On learning correlated boolean functions using statistical queries. In *Algorithmic Learning Theory, 12th International Conference, ALT 2001, Washington, DC, USA, November 25-28, 2001, Proceedings*, pages 59–76, 2001. 13