New Bounds on Quantum Sample Complexity of Measurement Classes

Mohsen Heidari
Department of Computer Science
Indiana University
Email: mheidar@iu.edu

Wojciech Szpankowski
Department Computer Science
Purdue University
Email: szpan@purdue.edu

***Abstract*—This paper studies the sample complexity of classical inference from quantum states under a quantum counterpart of the well-known agnostic *probably approximately correct* (PAC) model. In this model, a learner is trained using n quantum states with classical labels. The learner's objective is to find a quantum measurement that predicts the label of the unseen samples with high accuracy. The model subsumes well-studied problems such as state discrimination, quantum property testing, and classical learning. The learner also presents standard models such as variational quantum algorithmss (VQAs) and quantum neural networks (QNNs). Recent works showed that sample complexity of learning a quantum concept class scales linearly with the size of the class; whereas for classical (finite) concept classes, the sample complexity scales with the logarithm of the concept class size. Therefore, there is a question about the gap between quantum and classical sample complexity. This work proposes a new approach to quantum learning that leads to significant improvements over quantum sample complexity bounds. For that a new quantum learning algorithm is proposed for which we derive upper bounds on the sample complexity of several quantum concept classes. Particularly, we prove that the quantum sample complexity of a quantum concept class $\mathcal{C}$ grows with $\mathcal{O}(\log|\mathcal{C}^*|)$, where $\mathcal{C}^*$ is defined as the set of extreme points of the convex closure of $\mathcal{C}$. Consequently, we prove that for finite quantum concept classes, in the worst case, the bound grows with $\mathcal{O}(\log|\mathcal{C}|)$ as $|\mathcal{C}^*| \leq |\mathcal{C}|$.**

I. Introduction

Quantum learning is one of the leading applications of quantum computing both for classical and quantum problems. While some models suggest quantum-enhancements of classical learning by mapping data into input quantum states [1]–[6], quantum computers (QCs) have a far greater capability to learn patterns from inherently quantum data. This is possible by directly operating on quantum states of physical systems (e.g., photons or states of matter) or their qubit representations [7]–[14]. Learning from quantum data has been studied extensively in recent literature in the context of diverse applications, including condensed matter for phase-of-matter detection [7], [15], ground-state search [8], [9], [16], entanglement detection [10], [11], [17]–[20], and theoretical chemistry [12]–[14], [21], [22].

The focus of this paper is on the fundamental limits of quantum learning, particularly the quantum sample complexity as a measure of the hardness of training a model or tuning a quantum device. In classical learning theory, limits of learning have been studied for decades under the well-known PAC framework [23], [24]. In this work, we study a quantum counterpart of this problem under the *quantum probably approximately correct* (QPAC) framework [25].

Several models for quantum learning have been introduced and studied [26]–[29]. A survey on this topic is provided in [30]. In *quantum state discrimination*, the objective is to distinguish an unknown quantum state ρ from another (known or unknown) state using *measurements* on multiple samples [27], [29], [31]–[36]. In another model, introduced by Bshouty and Jackson [28], one is interested in solving a classical PAC problem using a quantum oracle that outputs identical copies of an associated superposition state [28], [37]–[40].

QPAC is a recent framework that subsumes several models such as state discrimination, quantum property testing, quantum state classification and classical PAC. It consists of a set of n labeled quantum states $(\rho_i, y_i)_{i=1}^n$ as the training samples. The samples are randomly generated independent and identically distributed (i.i.d.) and according to an unknown but fixed probability distribution. There is no structural assumption about the states ρ_i, the labeling y_i, and the underlying distribution. Here, predictors are quantum measurements applied to the quantum states. Thus, one seeks a model training procedure for finding a quantum measurement to minimize the prediction loss. The objective is to obtain a loss that is close to the optimal loss within a library of predictors (a.k.a concept class). Quantum sample complexity is, then, the minimum number of samples to guarantee such requirement for all quantum states, labeling, and the underlying distributions.

Therefore, QPAC is a stronger requirement than the above models, as it is a distribution-free and state-free condition. It is also stronger than PAC, as PAC is only distribution-free. Moreover, QPAC abides by quantum mechanical laws such as no-cloning, state collapse, and measurement incompatibility. Such properties prohibit sample reuse and, thus, raise new challenges for learning in quantum settings. Moreover, quantum models are significantly richer than classical models. Hence, given the fragility of quantum samples, the strictness of QPAC, and the richness of quantum models, one expects that quantum sample complexity to be significant, if not exponentially, greater than the classical one.

This paper presents a new bound on quantum sample complexity which is rather surprising. We show that the quantum sample complexity could potentially be comparable

to the classical one. We introduce a novel algorithm called *quantum shadow risk minimization* (QSRM).

In classical learning *empirical risk minimization* (ERM) is a brute-force search in a given concept class $\mathcal{C}$ to minimize the empirical loss. With that ERM is an *agnostic* PAC learner as long as the training samples are a good representation of the true distribution which is guaranteed if the sample size is $n = O(\log|\mathcal{C}|)$ [41]. Extending this algorithm to quantum is not straightforward. One could naively propose the same technique to compute the empirical risk of each quantum predictor and chose the one with the minimum risk. However, the no-cloning and measurement incompatibility make this approach prohibitive. Essentially, the training samples will be distorted each time we measure the empirical risk of a predictor. If we use that naively, then we might need fresh samples for each predictor. That gives a sample complexity that grows with the size of the class $n = O(|\mathcal{C}|)$. There have been multiple attempts [25], [42]. [25] introduced an algorithm that measures the risk of compatible predictors together and hence obtained a better bound that grows with $O(\log|\mathcal{C}|)$ in fully compatible scenarios and grows $O(|\mathcal{C}|)$ in the worst case scenario with fully incompatible class. This is clearly larger than classical. Whether one can reduce it remains the question.

Contributions: We propose a new approach called QSRM to measure the empirical risk of the predictors in the class that substantially improves the quantum sample complexity. Particularly, we prove that the quantum sample complexity of any measurement concept class $\mathcal{C}$ grows with $\mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{|\mathcal{C}^*|}{\delta}\right)$, where $\mathcal{C}^*$ is the set of extreme points of $\mathcal{C}$. Hence, in the worst case, the bound grows with $O(\log|\mathcal{C}|)$ if $\mathcal{C}$ is finite. Interestingly, the bound can be much lower than $\log|\mathcal{C}|$. This is surprising especially since sample duplication is prohibited and measurement incompatibility would lead to an exponentially larger sample complexity with standard methods. Such bound resembles those via Vapnik–Chervonenkis (VC) dimension in classical learning. However, whether $\log|\mathcal{C}^*|$ is the candidate for the quantum counterpart of the VC dimension remains to be seen.

II. Model Formulation

Notations: For shorthand, denote $[d]$ as $\{1, 2, ..., d\}$. For any $d \in \mathbb{N}$, let H_d be the Hilbert space of d-qubits with $\dim = 2^d$. The identity operator on H_d is denoted by I_d. As usual, a quantum state is defined as a *density operator*; that is a Hermitian, unit-trace, and non-negative linear operator. A quantum measurement $\mathcal{M}$ is a positive operator-valued measure (POVM) represented by a set of operators $\mathcal{M} := \{M_v, v \in \mathcal{V}\}$, where $\mathcal{V}$ is theset of possible outcomes, $M_v \geq 0$ for any $v \in \mathcal{V}$, and $\sum_{v\in\mathcal{V}} M_v = I_d$. For an operator A, denote $\|A\|_1 = \mathrm{tr}\{|A|\}$ as the trace norm, and $\|A\|_2 = \sqrt{\mathrm{tr}\{A^\dagger A\}}$ as the Hilbert–Schmidt norm.

A. Quantum Learning Model

Before presenting the main results, we formally define our quantum learning model. In this model [25], the objective is to distinguish between multiple groups of unknown quantum states without prior knowledge about the states. Available is only a training set of quantum states with a classical label determining their group index. We seek an agnostic procedure that given enough samples learns the labeling law. The model is defined more precisely as follows.

Let $\mathcal{Y}$ denote the labeling set and $\mathcal{H}$ be the underlying Hilbert space[1]. Each time, a sample $(|\phi\rangle_i, y_i), i \in [n]$ is randomly generated according to an unknown but fixed probability distribution D. A predictor is a quantum measurement $\mathcal{M} := \{M_{\hat{y}} : \hat{y} \in \mathcal{Y}\}$ that acts on the quantum states and outputs $\hat{y} \in \mathcal{Y}$ as the predicted label. Note that, unlike classical learning, the predicted label is random even for a fixed input. From Born's rule, $\hat{y}_i$ is generated randomly with probability $\langle\phi_i|M_{\hat{y}_i}|\phi_i\rangle$. The prediction loss is determined via a loss function $l : \mathcal{Y}\times\mathcal{Y} \to [0,\infty)$. The risk of a predictor $\mathcal{M}$ is computed by randomly generating a test sample $(|\phi\rangle_{test}, y_{test})$ according to D and measuring $|\phi\rangle_{test}$ with $\mathcal{M}$ to get $\hat{y}_{test}$. Hence, from Born's rule, the generalization (expected) loss is calculated as $L_D(\mathcal{M}) = \mathbb{E}[l(Y,\hat{Y})]$, where the expectation is taken over the sample's distribution D and the distribution of $\hat{Y}$.

Remark 1. *The generalization can be written compactly in terms of the density operators. Let ρ_y be a mixed state (density operator) representing the overall state of the system under label y. We can view ρ_y as the state averaged under the condition that the label is y. Then, the generalization loss of $\mathcal{M}$ is given by*

$$L_D(\mathcal{M}) = \sum_{y\in\mathcal{Y}}\sum_{\hat{y}\in\mathcal{Y}} D_Y(y)l(y,\hat{y})\,\mathrm{tr}\{M_{\hat{y}}\rho_y\}. \tag{1}$$

The problem in the binary case is simplified. The following is an example of this setting.

Example 1. *As an example, consider a simple setting where there are only three types of states $|\psi_j\rangle, j = 0, 1, 2$ with label set $\mathcal{Y} = \{0, 1\}$. Each labeled sample is either of the four possibilities: $(|\psi_0\rangle, 0), (|\psi_1\rangle, 1), (|\psi_2\rangle, 0)$, and $(|\psi_2\rangle, 1)$ with probabilities $p_{0,0}, p_{1,1}, p_{2,0}$, and $p_{2,1}$, respectively. Hence, the label of $|\psi_2\rangle$ is probabilistic. Also consider the 0-1 loss $l(y, \hat{y}) := \mathbb{1}_{\{y\neq\hat{y}\}}$. Then, the generalization loss of a measurement $\mathcal{M} = \{M_0, M_1\}$ is given by $L_{0\text{-}1}(\mathcal{M}) = p_{0,0}\langle\psi_0|M_1|\psi_0\rangle + p_{1,1}\langle\psi_1|M_0|\psi_1\rangle + p_{2,0}\langle\psi_2|M_0|\psi_2\rangle + p_{2,1}\langle\psi_2|M_0|\psi_2\rangle$. Hence, the corresponding density operators are $\rho_0 = p_{0,0}|\psi_0\rangle\langle\psi_0| + p_{2,0}|\psi_2\rangle\langle\psi_2|$, and $\rho_0 = p_{1,1}|\psi_1\rangle\langle\psi_1| + p_{2,1}|\psi_2\rangle\langle\psi_2|$.*

The generalization loss is compared to the optimal value within a concept class which is a collection $\mathcal{C}$ of quantum predictor measurements. With this setup, a quantum learning algorithm processes the training samples and finds a predictor $\mathcal{M}$ which may or may not belong to $\mathcal{C}$. Let opt be the minimum loss among all the predictors in $\mathcal{C}$.

Definition 1 (QPAC)**.** *A quantum learning algorithm agnostically QPAC learns a measurement class $\mathcal{C}$ if there exists*

[1]For presentation simplicity, we assume $\mathcal{Y}$ is finite and $\mathcal{H}$ is finite-dimensional.

a function $n_{\mathcal{C}} : (0,1)^2 \mapsto \mathbb{N}$ *such that for every* $\epsilon, \delta \in [0, 1]$ *and given* $n > n_{\mathcal{C}}(\epsilon, \delta)$ *samples drawn i.i.d. according to any probability distributions* D *and any unknown states* $|\phi\rangle_i, i \in [n]$*, the algorithm outputs, with probability of at least* $(1-\delta)$*, a measurement whose loss is less than* $\mathsf{opt}_{\mathcal{C}} = \inf_{\mathcal{M}\in\mathcal{C}} L_D(\mathcal{M}) + \epsilon$.[2]

The quantum sample complexity of a concept class $\mathcal{C}$ is the minimum of $n_{\mathcal{C}}$ for which there exists a QPAC algorithm.

Note that the state discrimination problem is a special case in which samples are identical and are either of two *a priori* known states. Also, note that QPAC also subsumes classical PAC as classical samples can be embedded into pure and orthogonal states, and functions can be considered as a special form of quantum measurements. Therefore, QPAC is a stronger requirement than PAC and other methods. It is a *agnostic*, distribution-free and state-free condition; whereas PAC is only distribution-free. In addition, principles such as the no-cloning and state collapse after measurements, indicate that quantum samples are more fragile than classical ones.

B. Related Works on QERM

It is known that classical ERM PAC learns any (classical) finite concept class $\mathcal{C}$ with sample complexity that scales with $\mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{|\mathcal{C}|}{\delta}\right)$. In quantum settings, there have been various attempts in developing counterparts of ERM algorithm. Due to the no-cloning theorem, the straightforward quantum extension of ERM results in a sample complexity of $\mathcal{O}\left(\frac{|\mathcal{C}|}{\epsilon^2}\log\frac{1}{\delta}\right)$, see [25] for more details. This is problematic as the sample complexity grows linearly with the size of the concept class.

In [25], a new ERM-type algorithm is introduced to improve this bound. The new bound depends on the measurement *incompatibility* structure of the concept class. Incompatible measurements cannot be measured simultaneously (for more details see [43]). On one extreme, all the measurements in the concept class are mutually compatible; on another extreme, there is no pair of compatible measurements. Based on this, an improved bound on sample complexity is as follows.

Fact 1 ([25]). *Quantum sample complexity of any finite concept class* $\mathcal{C}$ *is upper bounded as*

$$n_{\mathcal{C}}(\epsilon,\delta) \leq \min_{\mathcal{C}_r \text{Comp. partition}} \sum_{r=1}^{m} \left\lceil \frac{8}{\epsilon^2}\log\frac{2m|\mathcal{C}_r|}{\delta} \right\rceil,$$

where the minimization is taken over all compatibility partitioning of $\mathcal{C}$*. This bound ranges from* $\mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{|\mathcal{C}|}{\delta}\right)$*, for fully compatible class, to* $\mathcal{O}\left(\frac{|\mathcal{C}|}{\epsilon^2}\log\frac{1}{\delta}\right)$ *for fully incompatible class.*

In [42], this result was extended to infinite concept classes through an ϵ-netting argument. In this paper, we propose a new quantum ERM that substantially improves the above bounds to one that, in the worst case, grows with $\mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{|\mathcal{C}|}{\delta}\right)$ even for fully incompatible concept classes.

[2]Naturally, we are interested in efficient learning with $n_{\mathcal{C}}$ being at most polynomial in ϵ, δ and $\dim(\mathcal{H})$.

III. Main Results

In this section, we present the main results of the paper on the sample complexity. We prove the upper bound by proposing a quantum shadow risk minimization (QSRM).

A. Measuring the Empirical Loss

We start with defining the loss measurement for each predictor $\mathcal{M}$. Without loss of generality, assume $l : \mathcal{Y}\times\mathcal{Y} \mapsto [0, 1]$. Let $\mathcal{Z}$ be the image set of l. Since $\mathcal{Y}$ is a finite set, then so is $\mathcal{Z}$. With that, the loss observable for any predictor $\mathcal{M} := \{M_{\hat{y}} : \hat{y} \in \mathcal{Y}\}$ is given by $\mathcal{L}_M := \left\{L_z^M : z \in \mathcal{Z}\right\}$, where

$$L_z^M = \sum_{y,\hat{y}\in\mathcal{Y}} \mathbb{1}_{\{l(y,\hat{y})=z\}} M_{\hat{y}} \otimes |y\rangle\langle y|, \qquad \forall z \in \mathcal{Z}. \tag{2}$$

Therefore, the loss of $\mathcal{M}$ for predicting y from a given ρ_x is obtained by applying $\mathcal{L}_M$ on $\rho_x \otimes |y\rangle\langle y|$. The result is a random variable $Z = \ell(y, \hat{Y})$ taking values from $\mathcal{Z}$ as in (2). With this formulation the expected loss of $\mathcal{M}$ equals to $L_D(\mathcal{M}) = \langle\mathcal{L}_M\rangle_{\rho_{XY}}$.Now given a set of samples $\mathcal{S}_n = \{(|\phi_i\rangle, y_i) : i \in [n]\}$ the empirical expected loss of $\mathcal{M}$ over the samples is

$$L_{\mathcal{S}_n}(\mathcal{M}) \triangleq \frac{1}{n}\sum_i \langle\phi_i, y_i|\mathcal{L}_M|\phi_i, y_i\rangle,$$

where $z(i)$ is the outcome of the measurement $\mathcal{L}_M$ on the ith sample. Note that, unlike classical learning, even for fixed samples the empirical loss is not fixed.

B. Joint Estimation of Empirical Loss

The main challenge in the quantum setting is to measure the empirical loss for all the measurements in the class. This however is impossible due to the no-cloning and measurement incompatibility. The loss measurements $\mathcal{L}_M$ might be incompatible for different $\mathcal{M} \in \mathcal{C}$ — hence impossible to be measured simultaneously. In [25] it was proposed to partition $\mathcal{C}$ into mutually compatible subsets. With that approach bounds on the sample complexity were introduced. Unfortunately, in the worst-case scenario, the bounds could grow with $|\mathcal{C}|$. In what follows, we introduce a new approach that, in the worst-case, grows with $\log|\mathcal{C}|$.

Our approach is inspired by Shadow Tomography [44], [45] that is applied to *identical copies* of quantum states. In what follows we explain this procedure. We perform the following procedure for each sample $(\rho_i, y_i), i \in [n]$ and then explain our approach.

First, we generate a unitary operator U_i randomly and uniformly from the space of all unitary operators on the underlying Hilbert space $\mathcal{H}$. We rotate ρ_i by applying U_i resulting the state $U_i^\dagger\rho_i U_i$. Let $\{|j\rangle, j = 1, 2, ..., \dim_{\mathcal{H}}\}$ be a basis on $\mathcal{H}$. We measure the rotated state on this basis. From Born's rule the probability of getting the output j is $P_j = \langle j|U_i^\dagger\rho_i U_i|j\rangle$. Suppose the outcome for the ith sample is $j_i \in [\dim_{\mathcal{H}}]$. Then, the following state is prepared $\omega_i = U_i|j_i\rangle\langle j_i|U_i^\dagger$. As a result, the expected state ω_i over the

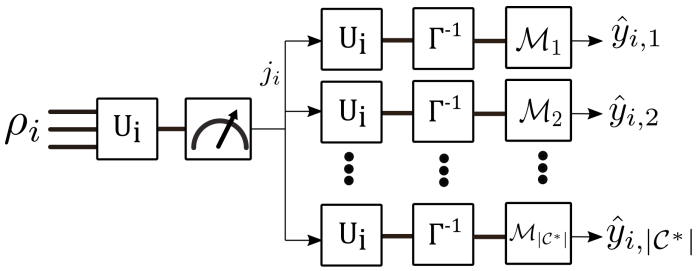

Fig. 1. The process for estimating the empirical loss of each measurement.

measurement randomness (P_j) and the choice of unitary U_i equals to $\Gamma[\rho_i]$, where Γ is a mapping defined as

$$\Gamma[O] := \mathbb{E}_U\Big[\sum_{j\in[\dim_{\mathcal{H}}]} \langle j|U^\dagger O U|j\rangle \;\; U\,|j\rangle\langle j|\,U^\dagger\Big], \qquad (3)$$

for any any operator O on $\mathcal{H}$. Observe that Γ is a linear mapping on $\mathcal{B}(\mathcal{H})$ with its inverse, denoted as Γ^{-1}, also a linear mapping. We apply Γ^{-1} on ω_i resulting in the following final state

$$\hat{\rho}_i := \Gamma^{-1}\big[U_i\,|j_i\rangle\langle j_i|\,U_i^\dagger\big]. \qquad (4)$$

This state is the classical "shadow" of ρ_i that is stored in a classical computer.

Fact 2 ([45]). *$\hat{\rho}_i$ is an unbiased estimate of ρ_i, that is $\mathbb{E}_{U_i,J_i}[\hat{\rho}_i] = \rho_i$, where the expectation is taken over the rotation U and measurement randomness.*

Now, we compute the expected loss per each shadow as

$$\hat{L}_i(M) = \sum_{\hat{y}} l(y,\hat{y})\,\mathrm{tr}\{M_{\hat{y}\hat{\rho}_i}\}.$$

Next, we average it over all the shadows to find

$$L_{\hat{\mathcal{S}}_n}(\mathcal{M}) := \frac{1}{n}\sum_{i=1}^{n}\hat{L}_i(M), \qquad (5)$$

where here $\hat{\mathcal{S}}_n$ represents the set of all the shadows $\hat{\rho}_i$ with their labels y_i. Therefore, $L_{\hat{\mathcal{S}}_n}$ is different from $L_{\mathcal{S}_n}$ which is based on the true samples. This process is demonstrated in Figure 1.

C. Shadow Norm

Theorem 1 of [45] gives a bound on the complexity of the median of means estimator for predicting the expectation of m arbitrary observables. The bound scales with $\log m$ and the maximum *shadow norm* of the observables. In what follows, we prove a similar result but for the empirical average estimator.

Definition 2. *The shadow norm of any operator O on $\mathcal{H}$ is*

$$\|O\|_{\text{shadow}} := \max_{\sigma\in\mathcal{D}[H]}\Big(\sum_{j\in[\dim_{\mathcal{H}}]}\langle j|U^\dagger\sigma U|j\rangle\;\;\langle j|U\Gamma^{-1}[O]U^\dagger|j\rangle^2\Big)^{1/2}.$$

Shadow norm for Clifford groups is bounded as

$$\|O\|_{shadow} \le \sqrt{3\,\mathrm{tr}\{O^2\}}$$

For k-local measurements is bounded as

$$\|O\|_{shadow} \le 2^k\|O\|_\infty$$

Our first result is as follows.

Theorem 1. *Let $\hat{\rho}_i, i\in[n]$ be the classical shadows of n copies of a mixed state ρ, as in (4). Then, given $\delta\in(0,1)$ and m arbitrary observables $O_1,\cdots,O_m$, the empirical average $\hat{o}_j := \frac{1}{n}\sum_i \mathrm{tr}\{O_j\hat{\rho}_i\}$ satisfies the additive error ε given that*

$$n \ge \mathcal{O}\left(\frac{1}{\varepsilon^2}\log\Big(\frac{m}{\delta}\Big)\max_i\left\|O_i - \frac{\mathrm{tr}\{O_i\}}{dim_{\mathcal{H}}}I\right\|^2_{\text{shadow}}\right).$$

The the objective of QSRM is to find $\min_{\mathcal{M}} L_{\hat{\mathcal{S}}_n}$. However, the upper bound in the above theorem becomes loose when the concept class is large, for example when $|\mathcal{C}|$ is infinite. We argue that only the extreme points of $\mathcal{C}$ are relevant.

D. Extreme points of a concept class

Let $\bar{\mathcal{C}}$ denote the convex closure (envelope) of $\mathcal{C}$. When $\mathcal{C}$ is finite, then $\bar{\mathcal{C}}$ is the set of all POVMs that can be written as a convex combination of measurements in $\mathcal{C}$. More precisely, POVMs of the form $\bar{\mathcal{M}} = \{\bar{M}_{\hat{y}}, \hat{y}\in\mathcal{Y}\}$ such that

$$\bar{M}_{\hat{y}} = \sum_{j=1}^{k}\alpha_j M^j_{\hat{y}}, \quad \forall\hat{y}\in\mathcal{Y},$$

where each $\mathcal{M}^j = \left\{M^j_{\hat{y}}, \hat{y}\in\mathcal{Y}\right\}$ belongs to $\mathcal{C}$, and $\alpha_j\in[0,1]$ with $\sum_j\alpha_j = 1$. By definition $\bar{\mathcal{C}}$ is a convex set.

Definition 3. *Given any concept class $\mathcal{C}$, by $\mathcal{C}^*$ denote the extreme points of its convex closure.*

The following Algorithm 1 and its performance discussed in Theorem 2 are the main contribution of this paper.

Algorithm 1: QSRM

Input: $\mathcal{C}^*$ of the concept class and n training samples.
for $i = 1$ **to** n **do**
  Generate a unitary U_i randomly.
  Apply U_i on ρ_i as in Figure 1.
  Measure along $\{|j\rangle, j\in[\dim_{\mathcal{H}}]\}$ to get j_i.
  for $\ell = 1$ **to** $|\mathcal{C}^*|$ **do**
    Prepare the state $\hat{\rho}_i$ as in (4).
    Measure $\hat{\rho}_i$ by $\mathcal{M}_\ell$ to get $\hat{y}_{i,\ell}$.
    Calculate the incurred loss $z_{i,\ell} = l(y_i,\hat{y}_{i,\ell})$.
Compute the estimated empirical loss for each measurement as $\hat{L}(\mathcal{M}_\ell) = \frac{1}{n}\sum_i \hat{z}_{i,\ell}$.
return $\hat{\mathcal{M}}$ as the measurement with the minimum $\hat{L}$.

Theorem 2. *Suppose ℓ is a bounded loss function and $\mathcal{C}$ is a measurement class with finite extreme points. Then, QSRM (Algorithm 1) agnostically QPAC learns $\mathcal{C}$ with quantum sample complexity bounded as*

$$n_{\mathcal{C}}(\epsilon,\delta) = \mathcal{O}\left(\frac{V_{\mathcal{C}^*}}{\epsilon^2}\log\frac{|\mathcal{C}^*|}{\delta}\right),$$

where $\mathcal{C}^*$ is the set of extreme points as in Definition 3 and

$$V_{\mathcal{C}^*} := \max_{\mathcal{M}\in\mathcal{C}^*} \max_{y} \left\| \mathcal{L}_M(y) - \frac{\mathrm{tr}\{\mathcal{L}_M(y)\}}{dim_{\mathcal{H}}} I \right\|^2_{\text{shadow}},$$

where the shadow norm is as in Definition 2.

Note that $|\mathcal{C}^*| \leq |\mathcal{C}|$. Even when $\mathcal{C}$ is infinite $\mathcal{C}^*$ can be finite. Hence an interesting distinction compared to the classical is observed. When $\mathcal{C}$ is finite, then the bound in theorem scales at most with $\mathcal{O}\Big(\frac{1}{\epsilon^2}\log\frac{|\mathcal{C}|}{\delta}\Big)$. Interestingly, in other scenarios where $|\mathcal{C}^*| \ll |\mathcal{C}|$ the bound is significantly lower. However, even though the Hilbert space is finite-dimensional and $\mathcal{Y}$ is finite, there could be infinitely many extreme points in $\bar{\mathcal{C}}$. In that case, an ϵ-netting argument will give a bound on the sample complexity.

IV. Proof of the main results

A. Proof of Theorem 1

The proof of the theorem follows from a concentration of measures for bounded variances. We use the following result from [46, Theorem 8.2].

Lemma 1 (Method of Bounded Variances)**.** *Let $X_1, ..., X_n$ be a set of random variables and let $Y_n = f(X_1, ..., X_n)$, where f a function such that $\mathbb{E}[Y_n] < \infty$. Let $D_i = \mathbb{E}[Y_n|\mathbf{X}^i] - \mathbb{E}[Y_n|\mathbf{X}^{i-1}]$ and $|D_i| \leq c_i$ for some constants $c_i > 0$. Also let, $V = \sum_{i=1}^n \sup_{\mathbf{x}^{i-1}} \mathsf{var}(D_i|\mathbf{x}^{i-1})$. Then,*

$$\mathbb{P}\Big\{|Y_n - \mathbb{E}[Y_n]| > \epsilon\Big\} \leq 2\exp\Big\{-\frac{\epsilon^2}{4V}\Big\},$$

where $\epsilon \leq \frac{2V}{\max_i c_i}$.

Corollary 1. *For any O, and small enough $\epsilon > 0$, the shadow empirical loss in (5) satisfies*

$$\mathbb{P}\Big\{|\hat{o} - \mathrm{tr}\{O\rho\}| > \epsilon\Big\} \leq 2\exp\Big\{\frac{-n\epsilon^2}{4\max_i \mathsf{var}(\mathrm{tr}\{O\hat{\rho}_i\})}\Big\}.$$

Proof. The corollary is proved from Lemma 1 with $f = \hat{o} := \frac{1}{n}\sum_i \mathrm{tr}\{O\hat{\rho}_i\}$ implying that

$$D_i = \frac{1}{n}(\mathrm{tr}\{O\hat{\rho}_i\} - \mathbb{E}[\mathrm{tr}\{O\hat{\rho}_i\}]).$$

Since, $\hat{\rho}_i$'s are mutually independent and identically distributed then

$$V = \frac{1}{n^2}\sum_i \mathsf{var}(\mathrm{tr}\{O\hat{\rho}_i\}) \leq \frac{1}{n}\max_i \mathsf{var}(\mathrm{tr}\{O\hat{\rho}_i\}),$$

which gives the desired statement. □

Applying this result for $O = O_i$ in the theorem and a union bound give $\max_j |\hat{o}_j - \mathrm{tr}\{O_j\rho\}| \leq \varepsilon$ with probability greater than

$$2m\exp\Big\{\frac{-n\epsilon^2}{4\max_{i,j}\mathsf{var}(\mathrm{tr}\{O_j\hat{\rho}_i\})}\Big\}. \tag{6}$$

From [45, Lemma 1], the variance terms are upper bounded by the shadow norm as follows

$$\max_j \left\| O_j - \frac{\mathrm{tr}\{O_j\}}{\dim_{\mathcal{H}}} I \right\|^2_{shadow}.$$

Therefore, by equating the bound on the probability to δ gives the desired result as stated in the theorem.

B. Proof of Theorem 2.

We start with the following lemma:

Lemma 2. *Let $\bar{\mathcal{C}}$ be the convex closure of $\mathcal{C}$ and $\mathcal{C}^*$ be the set of all extreme points of $\bar{\mathcal{C}}$. Then, $\mathsf{opt}_{\mathcal{C}} = \mathsf{opt}_{\bar{\mathcal{C}}} = \inf_{\mathcal{M}\in\mathcal{C}^*} L_D(\mathcal{M})$.*

Proof. Note that $L_D(\mathcal{M})$ is linear, and hence convex, in $\mathcal{M}$. This is due to the linearity of the trace and the definition of the loss given in (1). As a result, given that $\mathcal{C} \subseteq \bar{\mathcal{C}}$ and that $\mathsf{opt}_{\bar{\mathcal{C}}} = \inf_{\mathcal{M}\in\bar{\mathcal{C}}} L_D(\mathcal{M})$ we find that $\mathsf{opt}_{\bar{\mathcal{C}}} = \mathsf{opt}_{\mathcal{C}}$. Moreover, since the above expression is a convex optimization, then the optimal values occur at the extreme points of $\bar{\mathcal{C}}$. Hence the proof is complete. □

This result implies that QPAC learning of $\mathcal{C}$ is reduced to its extreme points $\mathcal{C}^*$.

The proof follows from Theorem 1 and Lemma 2. For any fixed y, and $\mathcal{M}$ let $\mathcal{L}_M(y)$ be the expected loss when the label is y. We apply Theorem 1 with $O_j = \mathcal{L}_M(y)$ for $\mathcal{M} \in \mathcal{C}^*$. Let

$$V_{\mathcal{C}^*} := \max_{\mathcal{M}\in\mathcal{C}^*} \max_{y} \left\| \mathcal{L}_M(y) - \frac{\mathrm{tr}\{\mathcal{L}_M(y)\}}{\dim_{\mathcal{H}}} I \right\|^2_{shadow}.$$

Then, the theorem gives the following sample complexity bound: $n = \mathcal{O}\Big(\frac{V_{\mathcal{C}^*}}{\epsilon^2}\log\Big(\frac{|\mathcal{C}^*|}{\delta}\Big)\Big)$.

Now, let $\widehat{\mathcal{M}}$ and $\mathcal{M}^*$ be the measurements minimizing the $L_{\hat{\mathcal{S}}_n}$ and L_D, respectively. Then, with probability $(1-\delta)$ the following chain of inequalities holds:

$$L_D(\widehat{\mathcal{M}}) \leq \hat{L}(\widehat{\mathcal{M}}) + \frac{\epsilon}{2} \leq \hat{L}(\mathcal{M}^*) + \frac{\epsilon}{2} \leq L_D(\mathcal{M}^*) + \epsilon.$$

The left-hand side is the loss of the selected predictor by QSRM (Algorithm 1), and the right-hand side equals $\mathsf{opt}_{\mathcal{C}} + \epsilon$ and hence the proof is complete.

Conclusion

This paper studies the learning of quantum measurement classes. It introduces a novel quantum algorithm called QSRM for learning quantum concept classes. Using this algorithm, a new upper bound on the quantum sample complexity is derived. It is shown that the quantum sample complexity grows at most with the logarithm of the size of the extreme points of the convex closure of the concept class. This is a significant improvement over prior results. The approach is based on a novel method to estimate the empirical loss of the concept class via creating random shadows of the training samples. With that QSRM algorithm can perform risk minimization while abiding to no-cloning, state collapse, and measurement incompatibility.

Acknowledgment

This work was partially supported by the NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, CCF-2006440, and CCF-2211423.

References

[1] V. Giovannetti, S. Lloyd, and L. Maccone, "Quantum random access memory," *Physical Review Letters*, vol. 100, no. 16, apr 2008.

[2] D. K. Park, F. Petruccione, and J.-K. K. Rhee, "Circuit-based quantum random access memory for classical data," *Scientific Reports*, vol. 9, no. 1, mar 2019.

[3] P. Rebentrost, M. Mohseni, and S. Lloyd, "Quantum support vector machine for big data classification," *Physical Review Letters*, vol. 113, no. 13, sep 2014.

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

[5] ——, "Quantum principal component analysis," vol. 10, no. 9, pp. 631–633, jul 2014.

[6] M. Heidari, A. Y. Grama, and W. Szpankowski, "Toward physically realizable quantum neural networks," *accepted and to be appear in Association for the Advancement of Articial Intelligence (AAAI)*, 2022.

[7] J. Carrasquilla and R. G. Melko, "Machine learning phases of matter," vol. 13, no. 5, pp. 431–434, feb 2017.

[8] G. Carleo and M. Troyer, "Solving the quantum many-body problem with artificial neural networks," vol. 355, no. 6325, pp. 602–606, feb 2017.

[9] M. Broughton, G. Verdon, T. McCourt, A. J. Martinez, J. H. Yoo, S. V. Isakov, P. Massey, R. Halavati, M. Y. Niu, A. Zlokapa, E. Peters, O. Lockwood, A. Skolik, S. Jerbi, V. Dunjko, M. Leib, M. Streif, D. V. Dollen, H. Chen, S. Cao, R. Wiersema, H.-Y. Huang, J. R. McClean, R. Babbush, S. Boixo, D. Bacon, A. K. Ho, H. Neven, and M. Mohseni, "Tensorflow quantum: A software framework for quantum machine learning," *arXiv:2003.02989*, Mar. 2020.

[10] F. V. Massoli, L. Vadicamo, G. Amato, and F. Falchi, "A leap among entanglement and neural networks: A quantum survey," *arXiv:2107.03313*, Jul. 2021.

[11] S. Lu, S. Huang, K. Li, J. Li, J. Chen, D. Lu, Z. Ji, Y. Shen, D. Zhou, and B. Zeng, "Separability-entanglement classifier via machine learning," *Physical Review A*, vol. 98, no. 1, p. 012315, 2018.

[12] I. Kassal, J. D. Whitfield, A. Perdomo-Ortiz, M.-H. Yung, and A. Aspuru-Guzik, "Simulating chemistry using quantum computers," *Annual Review of Physical Chemistry*, vol. 62, no. 1, pp. 185–207, may 2011.

[13] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, "Quantum computational chemistry," *Reviews of Modern Physics*, vol. 92, no. 1, p. 015003, mar 2020.

[14] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, S. Sim, L. Veis, and A. Aspuru-Guzik, "Quantum chemistry in the age of quantum computing," *Chemical Reviews*, vol. 119, no. 19, pp. 10 856–10 915, aug 2019.

[15] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, "Machine learning quantum phases of matter beyond the fermion sign problem," vol. 7, no. 1, aug 2017.

[16] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," vol. 549, no. 7671, pp. 195–202, sep 2017.

[17] Y.-C. Ma and M.-H. Yung, "Transforming bell's inequalities into state classifiers with machine learning," *npj Quantum Information*, vol. 4, no. 1, jul 2018.

[18] B. C. Hiesmayr, "Free versus bound entanglement, a NP-hard problem tackled by machine learning," *Scientific Reports*, vol. 11, no. 1, oct 2021.

[19] C. Chen, C. Ren, H. Lin, and H. Lu, "Entanglement structure detection via machine learning," *Quantum Science and Technology*, 2021.

[20] D.-L. Deng, X. Li, and S. D. Sarma, "Quantum entanglement in neural network states," *Physical Review X*, vol. 7, no. 2, p. 021021, 2017.

[21] C. Hempel, C. Maier, J. Romero, J. McClean, T. Monz, H. Shen, P. Jurcevic, B. P. Lanyon, P. Love, R. Babbush, A. Aspuru-Guzik, R. Blatt, and C. F. Roos, "Quantum chemistry calculations on a trapped-ion quantum simulator," *Physical Review X*, vol. 8, no. 3, p. 031022, jul 2018.

[22] B. Bauer, S. Bravyi, M. Motta, and G. K.-L. Chan, "Quantum algorithms for quantum chemistry and quantum materials science," *Chemical Reviews*, vol. 120, no. 22, pp. 12 685–12 717, oct 2020.

[23] M. J. Kearns, R. E. Schapire, and L. M. Sellie, "Toward efficient agnostic learning," *Machine Learning*, vol. 17, no. 2-3, pp. 115–141, 1994.

[24] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, nov 1984.

[25] M. Heidari, A. Padakandla, and W. Szpankowski, "A theoretical framework for learning from quantum data," in *IEEE International Symposium on Information Theory (ISIT)*, 2021.

[26] S. Aaronson, "The learnability of quantum states," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 463, no. 2088, pp. 3089–3114, sep 2007.

[27] S. M. Barnett and S. Croke, "Quantum state discrimination," *Advances in Optics and Photonics*, vol. 1, no. 2, p. 238, feb 2009.

[28] N. H. Bshouty and J. C. Jackson, "Learning dnf over the uniform distribution using a quantum example oracle," *SIAM Journal on Computing*, vol. 28, no. 3, pp. 1136–1153, 1998.

[29] C. Badescu, R. O'Donnell, and J. Wright, "Quantum state certification," in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, jun 2019.

[30] A. Montanaro and R. de Wolf, "A survey of quantum property testing," *Theory of Computing*, vol. 1, no. 1, pp. 1–81, 2016.

[31] R. O'Donnell and J. Wright, "Efficient quantum tomography," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, jun 2016.

[32] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu, "Sample-optimal tomography of quantum states," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, jun 2016.

[33] S. Bubeck, S. Chen, and J. Li, "Entanglement is necessary for optimal quantum property testing," in *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, nov 2020.

[34] S. Gambs, "Quantum classification," *0809.0444 [quant-ph]*, Sep. 2008.

[35] M. Guta and W. Kotlowski, "Quantum learning: asymptotically optimal classification of qubit states," *New Journal of Physics*, vol. 12, no. 12, p. 123032, dec 2010.

[36] H.-C. Cheng, A. Winter, and N. Yu, "Discrimination of quantum states under locality constraints in the many-copy setting," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, jul 2021.

[37] S. Arunachalam and R. de Wolf, "A survey of quantum learning theory," *arXiv:1701.06806*, 2017.

[38] V. Kanade, A. Rocchetto, and S. Severini, "Learning dnfs under product distributions via μ-biased quantum fourier sampling," *arXiv:1802.05690v3*, 2019.

[39] E. Bernstein and U. Vazirani, "Quantum complexity theory," *SIAM Journal on Computing*, vol. 26, no. 5, pp. 1411–1473, oct 1997.

[40] R. A. Servedio and S. J. Gortler, "Equivalences and separations between quantum and classical learnability," *SIAM J. Comput.*, vol. 33, no. 5, p. 1067–1092, May 2004.

[41] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge University Press, 2014.

[42] A. Padakandla and A. Magner, "Pac learning of quantum measurement classes : Sample complexity bounds and universal consistency," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Eds., vol. 151. PMLR, 28–30 Mar 2022, pp. 11 305–11 319. [Online]. Available: https://proceedings.mlr.press/v151/padakandla22a.html

[43] A. S. Holevo, *Quantum Systems, Channels, Information*. DE GRUYTER, jan 2012.

[44] S. Aaronson, "Shadow tomography of quantum states," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 325–338. [Online]. Available: https://doi.org/10.1145/3188745.3188802

[45] H.-Y. Huang, R. Kueng, and J. Preskill, "Predicting many properties of a quantum system from very few measurements," *Nature Physics 16, 1050–1057 (2020)*, Feb. 2020.

[46] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, Jun. 2009.