A Theoretical Framework for Learning from Quantum Data

Mohsen Heidari
Purdue University
mheidari@purdue.edu

Arun Padakandla
University of Tennessee
arunpr@utk.edu

Wojciech Szpankowski
Purdue University
szpan@purdue.edu

***Abstract*—Over decades traditional information theory of source and channel coding advances toward learning and effective extraction of information from data. We propose to go one step further and offer a theoretical foundation for learning classical patterns from *quantum data*. However, there are several roadblocks to lay the groundwork for such a generalization. First, classical data must be replaced by a density operator over a Hilbert space. Hence, deviated from problems such as *state tomography*, our samples are i.i.d density operators. The second challenge is even more profound since we must realize that our only interaction with a quantum state is through a measurement which – due to no-cloning quantum postulate – loses information after measuring it. With this in mind, we present a quantum counterpart of the well-known PAC framework. Based on that propose a quantum analogous of the ERM algorithm for learning measurement hypothesis classes. Then, we establish upper bounds on the quantum sample complexity quantum concept classes.**

I. Introduction

Over the past few decades, we have been mastering the ability to *learn* from data to perform many tasks such as classification, statistical inference, and pattern recognition. Recent achievements in quantum information processing to collect, store and process quantum systems endow us with a more powerful ability: learning from quantum data.

As research in quantum information theory suggests, fundamental concepts in classical settings admits multiple quantum counter-parts. For example, the task of communicating data over quantum channels leads to multiple notions of *capacity* [1]. The task of "learning" from "quantum data" is not an exception. Recently, researchers have been developing different learning frameworks [2]–[6].

From the perspective of quantum statistical learning theory, which is the view of this work, the learning models can be grouped into two main categories. The first category, refered to as *state tomography* or *state discrimination*, the objective is to find an approximate description of an unknown quantum state or distinguish it from another state using *measurements* on multiple copies of the state [7]–[9]. A survey on this topic is provided in [10]. An operational view of learning quantum states is introduced by [2]. Another related work in this line is [11] where the objective is to learn an unknown measurement E from samples of the form $\{(\rho_i, \mathrm{tr}\{E\rho_i\})\}_{i=1}^n$, where ρ_i's are independent and identically distributed (i.i.d.) random quantum states. Quantum state classification in this model then studied under various restrictions on the states (e.g., pure, mixed) [12], [13].

In the second group of works, which is refered to as the *quantum oracle model*, we measure identical copies of a superposition state to solve a classical learning problem [5], [14]. Learning using this method has been explored in several works such as [14]–[17] and analogous of the well-known *agnostic* PAC framework was introduced in [18].

The main departure point of this article from the mentioned models stems from the the fact that samples are not identical copies of each other; rather they are i.i.d. quantum states. Further, we are not required to learn the states rather we need only to learn a classical attribute to such states. That is we have an *ensemble* of quantum states, and associated to each state we have a classical attribute/label. Or alternatively, one can think of a quantum system that is measured by an unknown measurement (nature's measurement). We have access to the post-measurement states as well as the classical outcomes. The objective is to learn this measurement. Applications of this model has been studied in integrated quantum photonics [19]. That said, we propose a different model learning model for learning from quantum data. As a prototype, consider the following problem:

Suppose that there is a physical device randomly emitting a sequence of quantum states (e.g, photons), say $\rho_1, \rho_2, ...$ Associating to each state is a classical attribute $y_i \in \mathcal{Y}$, such as "red" or "blue" as its color. The probability distribution of the states and the underlying law governing their classical attribute are unknown. However, we know that the states belong to a family of parametrized quantum systems. We seek for a procedure that, given a number of training quantum states with their labels, learns the device's coloring/labeling law in order to predict the label of a new quantum state from this device.

Our problem formulation is motivated by the original/early questions that led to the theory of statistical learning. Suppose a computing device is provided with m training samples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : 1 \leq i \leq m$, can it learn the probabilistic/functional relationship between the label $y \in \mathcal{Y}$ and the features $x \in \mathcal{X}$. More specifically, under what conditions can an algorithm pick out a function from its library (hypothesis class) that best approximates the probabibilistic/functional relationship? The pursuit of an answer to this question led to the elegant theory of PAC learning, VC dimension, Radamacher complexity and such. As we describe in the sequel, our work formulates this very question in a quantum setup and we provide an initial set of our findings.

As our first contribution, we propose a quantum counter part of the probably approximately correct (PAC) learning framework as developed by [20], [21]. In our model the samples are pairs (ρ_i, y_i), where ρ_i's are density operators on a Hilbert space H_X and $y_i \in \mathcal{Y}$ are the classical labels. What we therefore seek is a measurement that will label a quantum state correctly. Hence, the predictors are measurements modeled as positive operator-valued measure (POVM). Analogous to the standard PAC, our quantum algorithm has a library of POVMs modeling the concept class of candidate predictors. By fixing a loss operator, we are lead to the analogous fundamental question of PAC learning: What is the quantum sample complexity for learning a measurement class?

To answer this question, we propose the quantum analogous of ERM algorithm using which we provide a bound on the quantum sample complexity. We will show that our model subsumes the classical PAC framework under some orthogonality condition. Further, our sample complexity bounds matches with classical ones. As a result, we conclude that the task of learning from quantum states is harder than classical. In other words, quantum sample complexity is not smaller than the classical sample complexity. We further show that the quantum sample complexity of a quantum concept class depends not only on its size but on a fundamental property called *compatibility* of the measurements in the class [22]. Such intrinsic quantum nature of the problem precludes a straightforward use of already developed complexity measures such as VC dimension, covering number and fat-shattering dimension [23], and Rademacher complexity from statistical learning theory [24].

As a careful reader will recognize, this learning framework hides several complexities. In what follows, we briefly highlight some its challenges and differences from previous models.

First, our only interaction with a quantum state is through a measurement. This necessitates the learning algorithm to be implemented via a *quantum measurement* with possible classical post-processing. Hence, abiding axioms of quantum mechanics, we can process the training samples only once, as they collapses after the measurement. This is a challenge; because, unlike the mentioned models, we do not have access to identical copies of the training samples. This difficulty is exacerbated as the *no-cloning* principle prohibits making new copies from the states at hand.

The second challenge arises from the *uncertainty principle*. Usually a learning algorithm needs to estimate multiple parameters via different measurements on the samples (e.g., empirical loss of different predictors). Ideally, we would like to combine these measurements and use one set of samples for all estimations. However, such measurements might not be *compatible* and hence, if we combine them the estimations' accuracy can drop significantly [22]. Motivated by the notion of *unbiased measurements* [25], [26], we propose *compatibility* covering in Section III.

Third, the training states are not completely distinguishable as they are not orthogonal. Hence, the amount of information we can extract from the samples is limited by the amount of their overlaps.

In this paper is organized as follows: In Section II, we formally describe the elements of our model and define a new quantum analogous of PAC. Then, in Section II-A we argue that classical learning is subsumed under this model. In section III we elaborate on the compatibility issue and propose our sample complexity bound. Lastly, in Section III-A we propose QERM to prove our results.

A. Preliminaries

Quantum states as usual are density operators, that are linear operators, self-adjoint, unit-trace and positive semi-definite. We denote by $\mathcal{D}(H)$ the set of all density operators on H. Any quantum measurement in this paper is modeled by a POVM. We denote a POVM as $\mathcal{M} := \{M_v, v \in \mathcal{V}\}$, where $\mathcal{V} \subset \mathbb{R}$ is the (finite) set of possible outcomes. Operators of the measurement satisfy the following conditions: $M_v = M_v^\dagger \geq 0$, $\sum_v M_v = I$, where I is the identity operator. For short-hand, we use $[n]$ to denote the set $\{1, 2, ..., n\}$ for any $n \in \mathbb{N}$.

II. The Proposed Quantum Learning Model

In this section, we formally propose our learning model. We discuss the differences between this model and the standard PAC framework. Also we show the classical learning framework is subsumed under our model.

Similar to the PAC framework, our model consists of multiple components which are defined in the following. Let $\mathcal{X}$ be a finite set. The feature set is a collection of fixed density operators $\rho_x, x \in \mathcal{X}$, acting on a fixed Hilbert space H_X. The set of possible classical labels is a finite set $\mathcal{Y}$. For example, in binary classification of qubits H_X is a two dimensional Hilbert space and $\mathcal{Y} = \{0, 1\}$.

For compactness, we consider an auxiliary quantum register (pure state) for storing the classical labels. Let H_Y denote the Hilbert space of the labels created as $H_Y = \text{span}\{|y\rangle : y \in \mathcal{Y}\}$.With this notation, ρ_x together with its label y are represented by the bipartite quantum state $\rho_x \otimes |y\rangle\langle y|$. Hence, the feature-label set is given by $\{\rho_x \otimes |y\rangle\langle y| : x \in \mathcal{X}, y \in \mathcal{Y}\}$.

Consider an unknown, but fixed, probability distribution D on $\mathcal{X} \times \mathcal{Y}$. As the training set, we are given n i.i.d. samples $\rho_{x_i} \otimes |y_i\rangle\langle y_i|, i \in [n]$, where (x_i, y_i) are drawn from D. With this setup, the training samples are represented by the tensor product state $S_n = \bigotimes_{i=1}^n \left(\rho_i \otimes |y_i\rangle\langle y_i|\right)$. Further, the average density operator of each sample is $\rho_{XY} = \sum_{x,y} D(x, y)\rho_x \otimes |y\rangle\langle y|$.

We seek a procedure that given the training samples, construct a predictor for the task of classification (statistical inference). The predictor is given the only feature state ρ_x and is tasked to produce a label. Since the features are quantum states and the labels are classical, the predictors are quantum measurements.That said, a predictor is a POVM $\mathcal{M} := \{M_y : y \in \mathcal{Y}\}$ acting on the X-system only. To test a predictor $\mathcal{M}$, a new sample is drawn according to D. If $\rho_x \otimes |y\rangle\langle y|$ is the realization of the test sample, then without

revealing y, we measure ρ_x with $\mathcal{M}$. The outcome of this predictor is $\hat{y}$ with probability $\mathrm{tr}\{M_{\hat{y}}\rho_x\}, \hat{y} \in \mathcal{Y}$. Note that this is different from the classical settings, where the output of the predictor is a deterministic function of the samples. Since our labels are essentially stored in classical registers, we employ a conventional loss function to measure the accuracy of the predicted label. Thus, by $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto [0,1]$ we denote the (normalized) loss function. Therefore, the true risk of a predictor $\mathcal{M}$ with respect to the underlying sample's distribution D is

$$L_D(\mathcal{M}) \triangleq \sum_{(x,y,\hat{y})\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} D(x,y)\ \ell(y,\hat{y})\ \ \mathrm{tr}\{M_{\hat{y}}\rho_x\}.$$

The concept class in our model is a collection $\mathcal{C}$ of predictors and its minimum loss is denoted by $\mathsf{opt}_C \triangleq \inf_{\mathcal{M}\in\mathcal{C}} L_D(\mathcal{M})$. Before describing the rest of the model, let us present the following example.

Example 1. *Consider electrons with spin pointing in a direction, represented by a $3-$dim unit vector in the Bloch sphere. Let finite set $\mathcal{X} = \{(\theta_i,\phi_j) = (\frac{i\pi}{20}, \frac{j2\pi}{20}) : 0 \le i,j \le 19\}$ represent the possible spin axis directions. We have two labels in $\mathcal{Y} = \{blue, red\}$. Nature decides to label an electron 'blue' if the axis of its spin is orthonormal to a specific orthant. Otherwise the electron is labeled 'red'. For this she chooses a specific orthant $\mathcal{O}$. This establishes a relationship - $p_{Y|X}$ - between the elements $(x,y) \in \mathcal{X}\times\mathcal{Y}$. Going further, she chooses a distribution p_X, samples X wrt this distribution, endows an electron with the corresponding spin and hands only the electron to us. Our predictor is aware of $\mathcal{X}$, its association with the spin directions, i.e the mapping $x \to \rho_x$, and $\mathcal{Y}$. Oblivious to both the nature's decision and the orthant, but possessing the prepared electron, a predictor's task is to unravel the label. The predictor is a measurement with two outcomes 'blue' and 'red'. An optimal predictor will be able to distinguish whether the axis of an electron's spin is orthonormal to $\mathcal{O}$ or otherwise.*

Learning Algorithm as a Quantum Measurement: A quantum learning algorithm is a process that with the training samples as the input, selects a predictor from the concept class.[1] This process is modeled as a quantum measurement on the joint space of all training the samples, i.e., $H_{XY}^{\otimes n}$. The outcome of this measurement is a classical number as the index of the selected predictor in the concept class.

Definition 1. *Let H_{XY} be the feature-label Hilbert space. Also let $\mathcal{C}$ be the concept class whose members are indexed by a set $\mathcal{J}$. Then, a (proper) quantum learning algorithm is a sequence of POVMs $\mathcal{A}_n := \{A_{n,j} : j \in \mathcal{J}\}, n \in \mathbb{N}$, acting on $H_{XY}^{\otimes n}$, the space of n samples, and with outcomes in $\mathcal{J}$.*

Unlike the classical settings, even if the samples are fixed, the algorithm's output is a random variable on $\mathcal{J}$. That said, we can write $M_J \in \mathcal{C}$ as the selected predictor with J being a random variable on $\mathcal{J}$.

[1]Our focus is on *proper* algorithms. Generally, we allow the selected predictor to be outside of the concept class.

With all the components described, we are ready to define the quantum version of PAC learnability.

Definition 2 (QPAC)**.** *Given a concept class $\mathcal{C}$, an algorithm $\mathcal{A}_n, n \in \mathbb{N}$ QPAC learns $\mathcal{C}$, if there exists a function $n_{\mathcal{C}} : (0,1)^2 \mapsto \mathbb{N}$ such that for every $\epsilon,\delta \in (0,1)$ and all $n \ge n_{\mathcal{C}}(\epsilon,\delta)$*

$$\sup_D \sum_{j\in\mathcal{J}} \mathrm{tr}\{A_{n,j}\rho_{XY}^{\otimes n}\}\mathbb{1}\{L_D(\mathcal{M}_j) > \mathsf{opt}_{\mathcal{C}} + \epsilon\} \le \delta,$$

where ρ_{XY} is the average density operator of the samples with respect to D and $\mathcal{M}_j \in \mathcal{C}$ is the jth predictor in the class.

Our goal is to characterize concept classes that are learnable and quantify their sample complexity. Before that, let us discuss the connection to the classical PAC.

A. Classical PAC learning is a special case

We argue that the proposed formulation subsumes the classical PAC learning framework.

Theorem 1. *For a classical PAC learning model with feature-label set $\mathcal{X}\times\mathcal{Y}$, hypothesis class $\mathcal{H}$, loss function $l : \mathcal{Y}\times\mathcal{Y} \mapsto [0,1]$, and algorithm A, there exist a corresponding element in the quantum learning model such that A is a PAC learning algorithm with respect to the classical model if and only if its quantum counter part is a QPAC learning algorithm under the quantum model.*

Proof idea:. We set $\rho_x = |x\rangle\langle x|, \forall x \in \mathcal{X}$, where $|x\rangle$'s are pure orthogonal states. As a result the feature-label density operators are $|x\rangle\langle x| \otimes |y\rangle\langle y|, x \in \mathcal{X}, y \in \mathcal{Y}$. As for the quantum hypothesis class, for any $f \in \mathcal{H}$ define the POVM $\mathcal{M}_f = \{M_y^f : y \in \mathcal{Y}\}$ where $M_y^f \triangleq \sum_{x:f(x)=y} |x\rangle\langle x|$. Then, our hypothesis class $\mathcal{C}$ is the collection of such POVMs $\mathcal{M}_f, f \in \mathcal{H}$. It is not difficult to see that the risk of any predictor $\mathcal{M}_f$ equals $L_D(\mathcal{M}_f) = \mathbb{E}_D[l(Y,f(X))]$ which is the classical risk of f. Further, since the states are completely distinguishable, one can show that any classical learning algorithm can be implemented by a quantum algorithm. As a result, of these arguments, we can show Definition 2 reduces to the standard PAC definition and that the classical sample complexity matches with quantum samples complexity. □

Note that in the setting of the above result, beyond possible computational advantages, the quantum learning does not benefit statistically. Hence, in this case the quantum sample complexity matches with the classical one. However, this might not be the case when the hypothesis class is classical, but ρ_x's are not orthogonal. Similarly in quantum source coding, when the states are not orthogonal, we get advantage in compression rates [27], [28].

III. Quantum PAC Learning Results

In this section, we present our main results which is a bound on quantum sample complexity. As discussed in the introduction, our bounds depend on the *compatibility* structure of the predictors in the concept class. To present our results, we need to elaborate on the notion of *compatibility*. The predictors in this paper are assumed to be *sharp* measurements. Thus,

from Theorem 2.13 of [22] the definition of compatibility is reduced to the following.

Definition 3. *A collection of sharp measurements $\mathcal{M}^j = \left\{M_y^j : y \in \mathcal{Y}\right\}, j = 1, 2, ..., k$, are compatible if their operators mutually commute, that is $[M_y^j, M_{\tilde{y}}^\ell] = 0$ for all $j, \ell \in [k]$ and all $y, \tilde{y} \in \mathcal{Y}$.*

Consequently, if $\mathcal{C}$ is a compatible concept class, then there exists a basis on which all the predictors are diagonalized.

If $\mathcal{C}$ is a general concept class. Then, we group its members into compatible subclasses.

Definition 4. *Given a collection of observables $\mathcal{C}$, a compatibility partitioning is a family of distinct subsets $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_m$ of $\mathcal{C}$ such that $\mathcal{C} = \bigcup_r \mathcal{C}_r$ and that the observables inside each $\mathcal{C}_r$ are compatible internally with each other.*

Note that there always exists a compatibility partitioning as the single element subsets of $\mathcal{C}$ form a valid covering. Further, note that the compatibility structure is an inherent property of the concept class which is independent of the samples.

Now with the above definitions, we are ready to present our main result in the following theorem.

Theorem 2. *Let $\mathcal{C}$ be a finite hypothesis class and $\Delta \leq I_d$ be the loss operator. Then, $\mathcal{C}$ is agnostic PAC learnable with sample complexity bounded from above as*

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

where $\mathcal{C}_r \subseteq \mathcal{C}$ form a compatibility covering of $\mathcal{C}$ and the minimization is taken over all such coverings.

The proof of the theorem is provided in the next subsection.

Remark 1. *If $\mathcal{C}$ is a compatible comcept class, then the sample complexity bound in Theorem 2 simplifies to $\left\lceil \frac{2}{\epsilon^2} \log \frac{|\mathcal{C}|}{\delta} \right\rceil$.*

A. QERM algorithm and the Proof of the main result

We prove Theorem 2 by proposing our QERM algorithm. As in the classical ERM, our algorithm is implemented by measuring the empirical loss for each predictor of the class. This is done by applying an appropriately designed quantum measurement on each sample. Then, we collect all measurement outcomes and process it classically to obtain the empirical losses. In what follows, we describe this process. Further, we propose a concentration analyses for quantum measurements.

We start with the measurement process for computing the empirical loss of only one predictor. Let $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto [0, 1]$ be the loss function and $\mathcal{Z}$ be the image set of ℓ. Given any predictor $\mathcal{M} := \{M_{\hat{y}} : \hat{y} \in \mathcal{Y}\}$, the corresponding loss observable is given by $\mathcal{L}_M := \left\{L_z^M : z \in \mathcal{Z}\right\}$, where

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

By applying $\mathcal{L}_M$ on a given sample $\rho_x \otimes |y\rangle\langle y|$ we obtained the loss value of $\mathcal{M}$ for predicting that sample. Note that, unlike the classical settings, when the predictor and the samples are fixed the loss value is still a random variable. In that case, the "conditional" expectation of Z for a fixed sample is given by $\langle \mathcal{L}_M \rangle_{\rho_x \otimes |y\rangle\langle y|}$, where $\langle \cdot \rangle$ is the *expectation value* of an observable in a quantum state. Hence, the overall expectation of Z equals $\mathbb{E}[Z] = \langle \mathcal{L}_M \rangle_{\rho_{XY}}$, where ρ_{XY} is the average density operator of the sample. Hence, it is not difficult to see that the true risk of a predictor $\mathcal{M}$ equals to

$$L_D(\mathcal{M}) = \langle \mathcal{L}_M \rangle_{\rho_{XY}} = \mathbb{E}[Z].$$

We compute an empirical loss of $\mathcal{M}$ by applying $\mathcal{L}_M$ on each sample. Let $z(i)$ be the realization of the loss value measured on the ith sample. Then, the empirical loss is given by

$$L_{\hat{D}}(\mathcal{M}) \triangleq \frac{1}{n} \sum_i z(i).$$

Next, we provide a quantum sample complexity analysis. For that, we present a quantum analogous of Chernoff-Hoeffding inequality.

Lemma 1. *Let $\rho_i, i \in [n]$ be i.i.d. random density operators on a finite dimensional Hilbert space H. Let $\bar{\rho} = \mathbb{E}[\rho_i]$ be their average density operator. Let $\mathcal{M}$ be a (discrete) observable on H with outcomes bounded by the interval $[a, b]$, where a, $b \in \mathbb{R}$. If V_i is the outcome of $\mathcal{M}$ for measuring ρ_i, then for any $t \geq 0$*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} V_i - \langle \mathcal{M} \rangle_{\bar{\rho}}\right| \geq t\right\} \leq 2\exp\left\{-\frac{nt^2}{2(b-a)^2}\right\},$$

where $\langle \mathcal{M} \rangle_{\bar{\rho}}$ is the expectation value of $\mathcal{M}$ in state $\bar{\rho}$.

The proof is omitted as it is a direct consequence of Theorem A.19 in [29].

We apply Lemma 1 where the measurement is $\mathcal{L}_M$ and the random states are our i.i.d. samples with $\bar{\rho} = \rho_{XY}$ as the average density operator. Hence, by an appropriate choice of t, given $\delta \in [0, 1]$, with probability $(1 - \delta)$ the following inequality holds

$$|L_{\hat{D}}(\mathcal{M}) - L_D(\mathcal{M})| \leq \sqrt{\frac{\log(2/\delta)}{2n}}.$$

As a next step, we would like to measure the empirical loss for all the predictors in the given hypothesis class. However, this is not straightforward as in the classical setting. Because, after measuring the empirical loss of one predictor, the quantum state of the samples collapses and we might not be able to "re-use" the samples to measure the loss of another predictor. Further, the no-cloning principle prohibits creating multiple copies of the training samples.

That said, a naive strategy is to partition the training samples into several batches, one for each predictor $\mathcal{M} \in \mathcal{C}$. For this strategy it is easy to verify that

$$\sup_{\mathcal{M} \in \mathcal{C}} |L_{\hat{D}}(\mathcal{M}) - L_D(\mathcal{M})| \leq \sqrt{\frac{|\mathcal{C}|}{2n} \log \frac{2}{\delta}}.$$

Therefore, the sample complexity blows up with the square of the size of the hypothesis class.

We improve upon this bound by leveraging from the compatibility notion.

QERM for Compatible Classes: Suppose the predictors in the hypothesis class $\mathcal{C}$ are compatible. Let index the elements of $\mathcal{C}$ by $\mathcal{J} = \{1, 2, ..., |\mathcal{C}|\}$. For each measurement $\mathcal{M}$, we have the loss observable $\mathcal{L}_M$ with operators as in (1). Since $\mathcal{M} \in \mathcal{C}$ are compatible, then so are $\mathcal{L}_M$. Hence, we create the POVM $\mathcal{L}^{\mathcal{C}}_{\text{QERM}} := \left\{L_{\mathbf{z}} : \mathbf{z} \in \mathcal{Z}^{|\mathcal{C}|}\right\}$, with operators

$$\mathcal{L}^{\mathcal{C}}_{\text{QERM}} := \left\{L_{\mathbf{z}} = \prod_{j \in \mathcal{J}_{\mathcal{C}}} L^{M_j}_{z_j} : \mathbf{z} \in \mathcal{Z}^{|\mathcal{C}|}\right\}, \tag{2}$$

where $\left\{L^{M_j}_z : z \in \mathcal{Z}\right\}$ are the operators of the $\mathcal{L}_{M_j}$.

We compute the empirical loss of all predictors in $\mathcal{C}$ by applying $\mathcal{L}^{\mathcal{C}}_{\text{QERM}}$ on each sample. Let $\mathbf{z}(i)$ be the outcome of $\mathcal{L}^{\mathcal{C}}_{\text{QERM}}$ when measuring the ith sample. By $z_j(i)$ denote the jth coordinate of the vector $\mathbf{z}(i)$. Then, the empirical loss of the jth predictor in $\mathcal{C}$ is given by

$$\mathcal{L}_{\hat{D}}(\mathcal{M}_j) = \frac{1}{n}\sum_{i=1}^{n} z_j(i). \tag{3}$$

Hence, we can simultaneously measure the empirical loss of all the predictors without the need for partitioning the training samples. We then establish the following result on the accuracy of the empirical loss.

Lemma 2. *Let $\mathcal{C}$ be a finite hypothesis class consisting of compatible predictors. Let $\mathcal{L}_{\hat{D}}(\mathcal{M}_j)$ be the empirical loss of the jth predictor of $\mathcal{C}$ as in (3). Then, for $\delta \in [0, 1]$, with probability at least $(1 - \delta)$, the following inequality holds*

$$\max_{\mathcal{M} \in \mathcal{C}} |L_{\hat{D}}(\mathcal{M}) - L_D(\mathcal{M})| \leq \sqrt{\frac{1}{2n} \log \frac{2|\mathcal{C}|}{\delta}}.$$

As a result, we expect that the sample complexity increases at most logarithmic with the size of the hypothesis class. Hence, we get a significant improvement over the naive strategy.

QERM for General Classes: Now we extend our approach for a general hypothesis class $\mathcal{C}$. The idea is to partition $\mathcal{C}$ into compatible subclasses as in Definition 4.

Class partitioning: Based on Definition 3, we can check if two measurements are compatible by checking whether their operators commute. Hence, with an exhaustive search one can find all possible ways of partitioning $\mathcal{C}$ into compatible subclasses. Note that the compatibility depends only on $\mathcal{C}$ and is independent of the samples. Hence, the partitioning can be done once as a pre-processing step.

Sample partitioning: With a partitioning, observables inside each subclass can be measured simultaneously. However, each compatible class must be supplied with an exclusive set of training samples. This is because measurements belonging to different subclasses may not be compatible. In other words, the n training samples have to be partitioned into multiple subsets, one for each subclass. The sample subsets are allowed to have different sizes. Let n_j be the size of the jth subset corresponding to jth subclass.

We repeat the process described in the previous part on each subclass with its sample subset. For that we create measurements $\mathcal{L}^{\mathcal{C}_r}_{QERM}$ as in (2) and compute the empirical loss of the predictors inside each subclass. We will show how to chose the batch sizes and the best partitioning of C.

With this approach, we formally propose the QERM algorithm as presented in Algorithm 1 and establish our theorem.

Algorithm 1: QERM

Input: Concept class $\mathcal{C}$ and n training samples.
Output: Index of the selected predictor in $\mathcal{C}$

1 Partition $\mathcal{C}$ into a set of compatible subclasses $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_m$.
2 Partition the samples into m bathes, one for each subclass.
3 **for** $r = 1$ **to** m **do**
4 Construct $\mathcal{L}^{\mathcal{C}_r}_{QERM}$ as in (2) and apply it on each sample in the rth batch.
5 Let $\mathbf{z}^r(i)$ be the vector outcome on the ith sample of batch r.
6 Compute $\bar{z}^r_j = \frac{1}{n_r}\sum_i \mathbf{z}^r_j(i)$, as the empirical loss of the jth predictor in $\mathcal{C}_r$.
7 **return** $\arg\min_{r,j} \bar{\mathbf{z}}^r_j$ as the index of the selected predictor.

As a last step in the proof Theorem 2, we analyze the sample complexity and find an upper bound on $n(\delta, \epsilon)$. The argument follows from standard steps.

We apply Lemma 2 on each subclass $\mathcal{C}_r$ with the rth sample batch with n_r samples. Set $n_j = \lceil \frac{2}{\epsilon^2} \log \frac{2|\mathcal{C}_r|}{\delta} \rceil$. As a result, with probability $(1 - \delta)$

$$\max_{\mathcal{M} \in \mathcal{C}_r} |L_{\hat{D}}(\mathcal{M}) - L_D(\mathcal{M})| \leq \frac{\epsilon}{2}.$$

Hence, from the union bound, with probability at $(1 - (1 - \delta)^m) \approx 1 - m\delta$ we have that

$$\max_{1 \leq r \leq m} \max_{\mathcal{M} \in \mathcal{C}_r} |L_{\hat{D}}(\mathcal{M}) - L_D(\mathcal{M})| \leq \frac{\epsilon}{2}.$$

Let $\widehat{\mathcal{M}}$ and $\mathcal{M}^*$ be the predictors minimizing the empirical loss and the true loss, respectively. Then,

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

The left-hand side is the loss of the selected predictor by QERM and the righ-hand side equals $\mathsf{opt} + \epsilon$. Hence, the proof is complete by replacing δ with δ/m.

ACKNOWLEDGEMENT

This work was supported in part by NSF Center on Science of Information Grants CCF-0939370 and NSF Grants CCF-1524312, CCF-2006440, CCF-2007238, and Google Research Award.

References

[1] M. Wilde, *Quantum information theory*. Cambridge, UK: Cambridge University Press, 2013.

[2] 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.

[3] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, "Efficient quantum state tomography," *Nature Communications*, vol. 1, no. 1, dec 2010.

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

[5] 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.

[6] C. Bădescu, 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.

[7] 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.

[8] ——, "Efficient quantum tomography II," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, jun 2017.

[9] 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.

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

[11] H.-C. Cheng, M.-H. Hsieh, and P.-C. Yeh, "The learnability of unknown quantum measurements," *QIC, Vol. 16, No. 7-8, 0615-0656 (2016)*, Jan. 2015.

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

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

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

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

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

[17] 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. [Online]. Available: https://doi.org/10.1137/S0097539704412910

[18] S. Arunachalam and R. De Wolf, "Optimal quantum sample complexity of learning algorithms," *J. Mach. Learn. Res.*, vol. 19, no. 1, p. 2879–2878, Jan. 2018.

[19] Z. A. Kudyshev, S. I. Bogdanov, T. Isacsson, A. V. Kildishev, A. Boltasseva, and V. M. Shalaev, "Rapid classification of quantum sources enabled by machine learning," *Advanced Quantum Technologies*, vol. 3, no. 10, p. 2000067, sep 2020.

[20] 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.

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

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

[23] M. J. Kearns and R. E. Schapire, "Efficient distribution-free learning of probabilistic concepts," in *Colt Proceedings 1990*. Elsevier, 1990, p. 389.

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

[25] J. Lawrence, Č. Brukner, and A. Zeilinger, "Mutually unbiased binary observable sets onNqubits," *Physical Review A*, vol. 65, no. 3, feb 2002.

[26] S. Bandyopadhyay, P. O. Boykin, V. Roychowdhury, and F. Vatan, "A new proof for the existence of mutually unbiased bases," *quant-ph/0103162*, 2001.

[27] N. Datta, M.-H. Hsieh, and M. M. Wilde, "Quantum rate distortion, reverse shannon theorems, and source-channel separation," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 615–630, jan 2013.

[28] B. Schumacher, "Quantum coding," *Phys. Rev. A*, vol. 51, pp. 2738–2747, Apr 1995. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.51.2738

[29] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 569–579, mar 2002.