Towards a Theory of Learning Measurement Classes

Mohsen Heidari
Purdue University
mheidari@purdue.edu

Arun Padakandla
University of Tennessee
arunpr@utk.edu

Wojciech Szpankowski
Purdue University
szpan@purdue.edu

Abstract—In current day computers, data is stored in registers - modeled as elements in a set - and operations performed on them are modeled as functions. The theory of statistical learning in the context of current day computers is a problem of function learning. In future quantum computing devices, we encode data onto qubits or quantum states. Operations on quantum states with classical outcomes are through measurements. We formulate a problem of learning from data encoded onto quantum states. Mathematically, this manifests as a problem of learning in measurement spaces. In addition to being a strict generalization of the conventional problem of PAC learning, the non-commutative of measurements throw new challenges. We take a first step at solving this problem leveraging the ideas of compatible measurements and thereby analyzing the ERM algorithm. Our exposition illustrates an elegant intertwining of physical and mathematical ideas involved in quantum information processing.

This is a preliminary preprint of our findings and an enlarged version with more results and proofs are in the publication process. While we commit to disseminating the latter version on arXiv at the earliest possible chance, the current version, intended to convey the ideas and flavor of the work, must not be distributed without the author’s permission.

I. INTRODUCTION

The blueprint of today’s learning algorithms can be traced back to the foundational ideas developed in the early years of statistical learning theory. Questions such as: When is a concept class learnable? What quantifies its complexity in regards to learnability? How does the sample complexity depend on this complexity? were formulated four decades ago keeping in mind Turing’s model of computation. Our answers to these questions continue to influence today’s learning algorithms. On an alternate front, considerable efforts are underway to design and build a fundamentally new computing device based on the axioms of quantum mechanics. How do these questions manifest in this new quantum computing framework? Do these lead to interesting formulations and ideas physically and mathematically? Motivated by this, we put forth a formulation and report an initial set of findings.

Learning theory is aimed at extracting information from data. Information, as we know, is quantified by the distribution of this data. The broad question in learning theory is therefore: How does one learn the underlying distribution, or some of its characteristics, from data? In our current-day computers, data is stored in classical registers and any processing of data are through algorithms. The mathematical theory underlying Turing’s model of computation is set theory. Data in registers are modeled via elements taking values in sets. Naturally, computations on data are mathematically modeled via functions operating on sets. The design of any efficient data processing algorithm or computation involves designing or identifying suitable/ appropriate functions.\(^1\)

Going back to learning theory, the task of learning distributions from data is therefore modeled in the Turing framework as a problem of identifying appropriate functions. Indeed, concept classes are subsets of function spaces and a learning algorithm is tasked to identify a near-optimal function. In summary, learning theory in the context of current-day computers is a function learning problem.

In quantum computing devices, data is encoded on subatomic particles by modifying its specific characteristics. For example, the axis of spin of an electron can be chosen to be one among \(2^8\) uniformly spaced radial vectors in a Bloch sphere. The behaviour of these sub-atomic particles are governed by the (unique) axioms of quantum mechanics. Moreover, our interaction with these quantum systems are through measurements. Reading out data written on subatomic particles entails designing measurements. How then, do we learn from data stored on quantum computing devices?

Let us now describe our formulation of learning from quantum data using the popular parlance in learning theory. Features are encoded into qubits, as specified earlier by preparing sub-atomic particles with specific characteristics. A predictor is a measurement performed on qubits, or more generally quantum states, whose outcomes are the labels. Since measurement outcomes are classical, our labels reside in classical registers. Modeling labels to be stored in classical registers, while being a first step, can also be justified since labels are, in some sense, our inferences. A concept class is a library of measurements to choose from. In this formulation, we consider supervised learning, wherein our learning algorithm (LA) has to pin down a measurement from this concept class. To do this, it is provided with \(n\) labeled features, where as stated, features are encoded in qubits and labels in classical registers - a natural setting for a CQ [1, Sec. 4.3.4] modeling.

These objects place us in a familiar PAC learning framework. What measurement concept classes are learnable? What is the sample complexity of learning measurement classes? How do we design an analyze an ERM learning algorithm? We address these three questions in this article. Our modest first step establishes (Thm. 1) that finite measurement classes are learnable. More importantly, we propose an elegant technique of optimizing sample complexity. This involves analyzing

\(^1\)Probabilistic algorithms can also be viewed as functions in a larger domain.
a new variant of the empirical risk minimization (ERM) algorithm. In the sequel, we highlight the challenges involved, our ideas and indicate the significance of our contribution, by placing it in perspective.

The above (simplified) description hides the complexities of the problem. The transition from classical to quantum involves both mathematical sophistication - graduating from elements, sets and functions to density operators, Hilbert spaces and measurements - as well as physical challenges. The latter has direct implication on learning algorithms, and in particular the ERM rule. When stored in a classical register, a feature \( x \) can be duplicated indiscriminately, thereby enabling us to simultaneously retain a ‘original copy’ \( x \) and the outcomes \( (f(x) : f \in C) \) of evaluating it through all functions in concept class \( C \) (Note 1). Given \( n \) training samples \((x_i, y_i) : 1 \leq i \leq n\), an ERM rule on a classical computer can simultaneously compute all empirical losses \( \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i) : f \in C \) wrt any loss function \( l \). Every training samples contributes to refining our estimate of the loss of every predictor in our concept class. The axioms of quantum mechanics precludes this luxury when features are encoded in quantum states. Collapse of a quantum state following a measurement (Axiom 3) implies that after the first measurement, the feature component of the training sample is altered irrevocably. This lends the training sample unusable for assessing loss of subsequent measurements.

Does this imply that every predictor can refine our loss estimate of just one predictor in our concept class, thereby blowing up the sample complexity? In Sec. III-B, we leverage compatible measurements [7, Sec. 2.3.2] to ‘re-use’ samples and thereby optimize complexity. Here, we characterize ERM measurement operators for compatible classes and prove that the marginals of the outcome are faithful. In fact, this points at possible interesting problems in optimal partitioning of measurement concept classes into mutually compatible measurements.

This work is inspired by an elegant intertwining of physical (Exs. 1 - 4) and mathematical concepts aimed at designing optimal information processing algorithms. Mathematically, the problem we formulate is a generalization (Rem. 1) of the fundamental conventional problem of PAC learning, and we can therefore recover current known results through a complete solution of the same. Having said this, we do admit that the mathematical sophistication, alluded to above, restricts us to finite dimensional Hilbert spaces. Therefore, strictly speaking, we recover the conventional problem of PAC learning with a finite set of feature. However, this is only a technical limitation, not a conceptual one. Moreover, our Hilbert spaces can be of an arbitrary finite dimension, thereby recovering a conventional PAC learning problem with arbitrary finite feature sets. Secondly, our problem formulation represents a graduation from function space learning to measurement space learning. Thirdly, this must be viewed as a modest first step towards an interesting exploration of the science of information. Our focus in this article is an exposition of the ideas. Detailed proofs are provided in [2].

We conclude this section recognizing related work. Cheng. et al. [3] study the problem of learning one particular measurement when provided with the density operator and the outcome statistics. Aaronson [4] studies the problem of state tomography and prove that the number of measurements growing linearly in the dimension suffices for learning states for most practical purposes. The problem of using quantum computers for classical learning has evinced interest. Bshouty and Jackson [5] investigate the use of a quantum computer in PAC learning function classes. We refer to [6] for an excellent survey on these works.

A. Preliminaries

We supplement notation adopted in [1] with the following. For \( n \in \mathbb{N} \), \( [n] := \{1, \cdots, n\} \). We let \( \mathcal{L}(\mathcal{H}), \mathcal{D}(\mathcal{H}) \) denote the collections of linear, positive and density operators acting on Hilbert space \( \mathcal{H} \). All our Hilbert spaces are assumed to be finite dimensional. For clarity, a POVM \( \mathcal{M} = \{M_y : y \in \mathcal{Y}\} \) with outcomes in \( \mathcal{Y} \) is either referred to as a \( \mathcal{Y} \)–POVM or \( \mathcal{Y} \)–POVM on \( \mathcal{H}_X \). We let \( \mathcal{S}_{\mathcal{X} \rightarrow \mathcal{Y}} \) denote the subcollection of sharp (projective) measurements on \( \mathcal{H}_X \) with outcomes in \( \mathcal{Y} \). In this work, we are required to perform the same measurement on multiple quantum states. For a \( \mathcal{Y} \)–POVM \( \mathcal{M} = \{M_y : y \in \mathcal{Y}\} \), we therefore define \( \mathcal{M}^n := \{M_{y^n} : y^n \in \mathcal{Y}^n\} \) with a slight abuse of notation. For a Hilbert space \( \mathcal{H}_X \) and a set \( \mathcal{Y} \), we let \( \mathcal{M}_{\mathcal{X} \rightarrow \mathcal{Y}} \) denote the set of all \( \mathcal{Y} \)–POVMs on \( \mathcal{H}_X \). See [2, Sec. 2.3] for its structure. \( \mathcal{M}_{\mathcal{X} \rightarrow \mathcal{Y}} \) denotes the set of all \( \mathcal{Y} \)–POVMs on \( \mathcal{H}_X^{\otimes n} \). We let \( \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \) and \( \mathcal{H}_Z := \mathcal{H}_X \otimes \mathcal{H}_Y \). We abbreviate random variables, probability mass function, empirical risk as RVs, PMF, ER respectively. We use PMF and distribution interchangeably. We write \( X \sim p_X \) if RV \( X \) is distributed with PMF \( p_X \).

II. MODEL FORMULATION AND PROBLEM STATEMENT

Let \( \mathcal{X} \) denote a feature set and \( \mathcal{Y} \) a finite set of labels. Feature-label pairs occurring in nature are modeled as generated IID wrt an unknown distribution \( p_{X|Y} \). A predictor, provided with only the feature, predicts its label. Its performance is measured by comparing the predicted and true labels via a loss function. A learning algorithm’s (LA’s) task is to identify an optimal predictor from within its concept class - a library of predictors. To accomplish this, it is provided \( n \) training samples, i.e. \( n \) labeled features, generated IID from the same unknown distribution \( p_{X|Y} \).

Our formulation of this abstract problem of prediction is based on Turing’s model of computation. In this model, data is stored in registers and computation is modeled via evaluation through functions. We are thus led to a concrete formulation - our familiar PAC framework for learning from classical data. Random variables \((X, Y) \in \mathcal{X} \times \mathcal{Y} \) with an unknown distribution \( p_{X|Y} \) model feature-label pairs occurring in nature. A predictor - a function \( f \in \mathcal{F}_{\mathcal{X} \rightarrow \mathcal{Y}} \) - is...
only provided X. Its performance is measured through its true risk \( l_p(f) := E_D \{ l(f(X), Y) \} \) wrt a loss function \( l : \mathcal{Y}^2 \rightarrow [0, \infty) \). A learning algorithm’s (LA’s) task is to identify an optimal predictor from within a / its concept class \( \mathcal{C} \subseteq \mathcal{F}_{\mathcal{X} \rightarrow \mathcal{Y}} \) - a library of functions. To accomplish this, it is provided with \( n \) training samples - \( n \) elements in \( \mathcal{X} \times \mathcal{Y} \). The design of a \( \mathcal{C}-\text{LA} \) entails designing a sequence of functions \( A_n : (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{C} : n \geq 1 \).

Before we describe a quantum formulation, we briefly highlight two consequences of the Turing model (that we take for granted). Stored in registers and modeled as elements in a set, features can be duplicated (indiscriminately), thereby enabling one to simultaneously retain a ‘original copy’ \( x \) and the outcomes \( (f(x) : f \in \mathcal{C}) \) of evaluating it through any arbitrary collection of functions (Note 1). Secondly, if one ignores circuit reliability issues, no uncertainty is involved in evaluating features through functions (Note 2).

On a quantum computing device, data is stored in qubits, or more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states. The behaviour of the latter is modeled via density operators. While more generally encoded onto quantum states.
Moreover, electron spins

\[ C \subseteq \mathcal{M}_{X \rightarrow Y} \] is (PAC) learnable if there exists a C–LA that learns C.

**Remark 1.** We are given a classical PAC learning problem with feature set X, concept class C \( \subseteq \mathcal{F}_{X \rightarrow Y} \). We can recover this classical formulation via the following substitutions in the problem formulated herein. Let chooses 

\[ \mathcal{H}_X = \text{span} \{ |x \rangle : x \in X \} \] with \( |x \rangle = \delta_{xx} \) and \( \rho_x = |x \rangle \langle x | \).

Next, let 

\[ M_Y^n = \sum_{x \in X} f(x) = y \langle x | y \rangle \] for each \( f \in C \).

The reader may verify that this recovers the classical learning PAC problem. Details are worked out in [2].

The goal of our investigation is to characterize concept classes C \( \subseteq \mathcal{M}_{X \rightarrow Y} \) that are learnable, quantify their complexity in regards to learnability and thereby derive their sample complexity. In the context of classical data, this study is now well understood. However, the unique behaviour of quantum systems and complexity of the involved mathematical objects throws up challenges leaving the above problem unresolved. In this article, we present our modest first step. We prove finite concept classes are learnable. More importantly, we leverage an elegant notion of compatible measurements to optimize the sample complexity. Our running example provides a good

setting to illustrate the challenges and our ideas.

**Ex. 4.** Suppose we are provided \( n = 20 \) labeled
electrons\(^8\) and our library C consists of 2 predictors \( M_1 = \{ M_{1b}, M_{1r} \}, M_2 = \{ M_{2b}, M_{2r} \} \). An ERM rule\(^9\) would attempt to measure all \( n = 20 \) electrons with both measurements and choose that measurement, whose outcomes disagree least with the provided labels.

However, a measurement alters the spin, lending it unusable to perform the next measurement. Moreover, electron spins cannot be replicated (No Cloning Theorem). In other words, every training sample can be used to evaluate the empirical risk (ER) of just one measurement.

Contrasting with Note 1, we recognize that noncommutativity of quantum measurements and the No Cloning theorem suggest that each training sample can be used to evaluate the ER of just one measurement. Moreover the outcome of measurements are random (Note 2) resulting in random empirical risks. How should a C–LA optimally utilize training samples to identify the `best' predictor from within C? The discussion thus far, exemplified through Ex. 1 - 4, suggests that each training sample can yield the ER of just one measurement. As an informed reader will note, this results in a blow-up of the sample complexity of any concept class. Our first simple step (Sec. III-B) is to leverage compatible measurements [7, Sec. 2.3.2] to `re-use' samples effectively. We build on this
to derive an upper bound on the sample complexity of finite concept classes (Thm. 1), thereby proving their learnability. As we noted, this is a first step in what could be an interesting exploration.

**III. ERM Rule and Compatible Measurements**

**A. Quantum State Collapse and Inefficiency of Standard ERM**

In order to characterize learnability, it is natural to analyze an ERM algorithm. It is instructive to briefly revisit the classical PAC learning scenario. For a predictor \( f \in C \), let RVs \( Y, \hat{Y}_f \) be jointly distributed with PMF

\[ p_{Y\hat{Y}_f}(y, \hat{y}) := \sum_{x \in X} p_{XY}(x, y) \mathbb{1}(y = f(x)) : (y, \hat{y}) \in \mathcal{Y}^2. \] (1)

Observe that \( l_p(f) = \mathbb{E}_{Y\hat{Y}_f}[l(Y, \hat{Y}_f)] \). Since \( (X_i, Y_i) \sim \prod_{i=1}^n p_{XY}, \) \( (Y_i, \hat{Y}_f) \) are IID \( p_{XY}, \) \( Y_i \) for all predictors \( f \in C \), and hence \( \frac{1}{n} \sum_{i=1}^n l(Y_i, \hat{Y}_f) \) converges to \( l_p(f) \) simultaneously for \( f \in C \), albeit at different rates.\(^10\) This guarantees that the empirical risks are good representations of the true risk for all predictors \( f \in C \), which is the foundational basis of optimality of the ERM rule.

An analogous analysis of the quantum setting illustrates mathematically the challenge alluded to in Ex. 4. For any measurement/predictor \( M = \{ M_{\tilde{y}} : \tilde{y} \in \mathcal{Y} \}, \) let 

\[ \mathcal{M} := \{ M_{\tilde{y}} \otimes \hat{y} : (\tilde{y}, \hat{y}) \in \mathcal{Y} \times \mathcal{Y} \} \] be a measurement designed to be performed on a training sample whose outcome includes the provided label. Observe that measuring a fresh unmeasured\(^11\) random training sample with density operator \( p_{XY} \in \mathcal{D}(\mathcal{H}_Z) \) with \( \mathcal{M} \) yields outcome RVs \( Y, \hat{Y}_M \) with distribution

\[ p_{Y\hat{Y}_M}(y, \hat{y}) := \sum_{x \in X} p_{XY}(x, y) \mathbb{1}(x = f(y)) : (y, \hat{y}) \in \mathcal{Y}^2. \] (2)

Observe that \( l_p(M) = \mathbb{E}_{Y\hat{Y}_M}[l(Y, \hat{Y}_f)] \). Suppose we choose \( \mathcal{M}_1 = \{ M_{y} : y \in \mathcal{Y} \} \) as our first measurement and perform the same on the \( n \) training samples, then the effective measurement on the \( n \) training samples is \( \mathcal{M}^n_1 = \{ M_{y^n} \otimes \hat{y}^n \} : (\hat{y}^n, y^n) \in \mathcal{Y} \times \mathcal{Y}^n \) and the \( n \) outcome RVs \( (Y_i, \hat{Y}_M, i) \) are IID \( p_{Y\hat{Y}_M} \). Hence, \( \mathbb{E} \frac{1}{n} \sum_{i=1}^n l(Y_i, \hat{Y}_M, i) \) converges to \( l_p(M_1) \) for the first measurement. However, after the first measurement is performed, the state of the training samples is altered.

The behaviour of the \( n \) collapsed training samples after measurement \( M_1 \) is performed is modeled as \( \hat{\rho}_{XY} \), where\(^12\)

\[ \hat{\rho}_{XY} := \sum_{x \in X} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Y}} p_{XY}(x, y) \sqrt{M_{1z} \rho_x \sqrt{M_{1z} \otimes |y \rangle \langle y |}}. \]

Suppose we next choose \( \mathcal{M}_2 = \{ M_{2y} : y \in \mathcal{Y} \} \) as our second measurement and perform the same on the \( n \) collapsed

\(^8\)The labels corresponding to the electrons are stored in classical registers.

\(^9\)In Sec. III, we characterize an algorithm \( E_{n} = (E^n_{\mathcal{M}} \in \mathcal{P}(\mathcal{H}_{Z}^\otimes n)) : \mathcal{M} \in \mathcal{C} \) corresponding to the well known ERM rule by specifying \( E^n_{\mathcal{M}} \). For now, the reader is encouraged to focus on the intuitive idea.

\(^10\)We have not made this mathematically precise in terms of mode etc. Our emphasis here is to convey the ideas.

\(^11\)The import of these adjectives will be clear in due course.

\(^12\)Our notation for \( \hat{\rho}_{XY} \) does not reflect its dependence on \( M_1 \), which it should. This choice has been made to reduce clutter.
training samples, then the \( n \) outcome RVs \((Y_i, \hat{Y}_{M_1,M_2,i}) : 1 \leq i \leq n\) are IID \( p_{Y\hat{Y}_{M_1,M_2}}(y, \hat{y}) = \sum_{x \in X} \sum_{z \in \mathcal{Z}} p_{XY}(x, y) \text{tr}(\sqrt{M_x} \rho_x \sqrt{M_z} M_z^2) : (y, \hat{y}) \in \mathcal{Y}^2\). Hence, for a non-trivial measurement \( M_1 \), the second \( \text{ER} \frac{1}{n} \sum_{i=1}^n l(Y_i, \hat{Y}_{M_1,M_2,i}) \) does not converge to \( l_p(M_2) \) for the second measurement in general.

**Remark 2.** From the above discussion, we conclude that if one employs the standard ERM algorithm to learn measurement classes, the collapse of quantum states implies that ER of the loss of the subsequent measurements are not guaranteed to provide an estimate of the subsequent measurement losses. Alternatively stated, a standard ERM algorithm is guaranteed in a blow of sample complexity. Each measurement requires its own set of training samples, which cannot be used for refining our estimate of the true risk of other measurements.

**B. Compatible Measurements**

How do we optimize our ERM algorithm? From our discussion in Sec. III-A, we recognize two distinct challenges. Firstly, outcomes are random, while function evaluations were deterministic (Note 2). This can be observed by the replacement of the indicator \( \phi \) in (1) with the trace in (2). Secondly, the collapse of our provided features post-measurement. While the second difficulty is a truly quantum phenomena, the first must be observed in learning of probabilistic concepts [7] - concept classes where functions are replaced by conditional distributions. The fact that sample complexity does not blow up in learning probabilistic concepts [7] suggests an approach.

Probabilistic concepts embed into measurement concept classes as one compatible [?, Sec. 2.3.2] set of measurements. Specifically, conditional distributions can be identified a specific set of measurements whose operators commute. The unique feature of quantum is complementarity - the presence of non-commutative measurements. We can therefore partition our measurement concept class into subsets wherein measurements in each subset have all commuting operators. We define the same.

**Defn 3.** A subcollection \( \mathcal{C}_B \subseteq \mathcal{M}_{X \to \mathcal{Y}} \times \mathcal{Y}-\text{POVMs} \) on \( \mathcal{H}_X \) are mutually compatible if for any pair \( \mathcal{M} = \{M_y : y \in \mathcal{M}\} \), \( \mathcal{L} = \{L_y : y \in \mathcal{Y}\} \in \mathcal{C}_B \), we have \( M_y L_y = L_y M_y \) for all \((y, \hat{y}) \in \mathcal{Y} \times \mathcal{Y}\).

**Remark 3.** We should note that our notion of mutual compatibility is more general than that in [?, Sec. 2.1]. In fact, our notion is just mutually commuting. Our choice of ‘compatible’ in place of commuting is only for aesthetic reasons. Another option is to restrict all our measurements to be projective or sharp [?, Defn 2.10]. In that case, mutually compatible is equivalent to mutually commutative. However, such a restriction would exclude classical probabilistic concepts. Our choice balances aesthetics and generality.

How do compatible measurements enable us ‘re-use’ samples effectively, or in other words, overcome the state collapse problem we discussed? Consider a subclass \( \mathcal{C}_B \subseteq \mathcal{C} \) that is mutually compatible. The idea is to fine-grain [?] these measurements to obtain a high resolution of operators. Specifically, for a collection \( \mathcal{C}_B \subseteq \mathcal{C} \), we can identify a sharp measurement \( \pi_B := \{\pi_b : b \in \mathcal{B}\} \subseteq \mathcal{X}_{\mathcal{Y}} \) such that for every \( \mathcal{M} \in \mathcal{C}_B \), there exists a stochastic matrix \( \alpha_{\mathcal{M}}^Y(y|b) : (y, b) \in \mathcal{Y} \times \mathcal{B} \) such that \( \pi_y = \sum_{b \in \mathcal{B}} \alpha_{\mathcal{M}}(y|b) \pi_b \). In order to obtain the ER of each measurement in \( \mathcal{C}_B \), we can just perform measurement \( \pi_B \) on each training sample and pass the outcome of the measurement through each of the stochastic matrices \( \alpha_{\mathcal{M}}^Y : \mathcal{M} \in \mathcal{C}_B \).

We use this in the following section to prove our results.

**IV. Finite Measurement Classes are Learnable**

We state our main results in this section. In regards to proofs, we only provide certain elements. Complete proofs are provided in [2].

We begin with the idea of partitioning concept classes into mutually compatible subsets and formalize the above stated fine-graining.

**Defn 4.** A collection \( \{\mathcal{C}_i : i \in \mathcal{I}\} \) of subsets of \( \mathcal{C} \) is a mutually compatible partition if (i) \( \mathcal{C}_i \subseteq \mathcal{I} \) is a partition, i.e., \( \mathcal{C}_i \cap \mathcal{C}_j = \phi \) if \( i \neq j \), \( \cup_i \mathcal{C}_i = \mathcal{C} \) and (ii) each \( \mathcal{C}_i \) is mutually compatible. We let \( \mathcal{X}_C \) denote the collection of all mutually compatible partitions of \( \mathcal{C} \). For a mutually compatible subcollection \( \mathcal{C}_B \subseteq \mathcal{C} \) of measurements, we let \( \{\pi_B : b \in \mathcal{B}\} \subseteq \mathcal{X}_{\mathcal{Y}} \) be a sharp measurement such that for every \( \mathcal{M} \in \mathcal{C}_B \), there exists a stochastic matrix \( \alpha_{\mathcal{M}}^Y(y|b) : (y, b) \in \mathcal{Y} \times \mathcal{B} \) such that \( \pi_y = \sum_{b \in \mathcal{B}} \alpha_{\mathcal{M}}^Y(y|b) \pi_b \).

Each mutually compatible partition provides us with a way to partition our concept class and re-use training samples within each subset of measurements. We provide an ERM algorithm and follow it by a result on the sample complexity.

**Algorithm 1:** Quantum ERM Algorithm

**Input:** Concept class \( \mathcal{C} \) and \( n \) training samples.

**Output:** Index of the selected predictor in \( \mathcal{C} \)

1. Identify a mutually compatible partition \( \mathcal{C}_B \subseteq \mathcal{C} \)
2. Partition \( n \) training samples into \( t \) subsets whose sizes \( n_1, \ldots, n_t \) are such that \( \frac{n_i}{n} \) is proportional to \( \left| \mathcal{C}_i \right| \).
3. For \( i = 1 \) to \( t \) do
   4. Measure the \( n_i \) training samples with measurement \( \{\pi_B : b_i \in \mathcal{B}_i\} \).
   5. Pass the outcomes through stochastic matrices \( \alpha_{\mathcal{M}}^Y \) for each \( \mathcal{M} \in \mathcal{C}_B \).
6. Return Among all the measurements in all of the mutually compatible subsets, choose the measurement with the least empirical risk.

13 These stochastic matrices can either be coupled are independent. It is only the marginals that matter.
Theorem 1. Every finite measurement class $C$ is (PAC) learnable with sample complexity at most

$$N(\epsilon, \delta) \leq \inf_{(C_B, i \in I) \in \mathcal{D}_C} \sum_{i \in I} \left\lceil \frac{\log \left( \frac{|C_B|}{\delta} \right)}{\epsilon^2} \right\rceil,$$

where the infimum is over the collection $\mathcal{D}_C$ of all mutually compatible partitions of $C$.

We refer the reader to [2] for a complete proof.

REFERENCES