PAC Learning of Quantum Measurement Classes: Sample Complexity Bounds and Universal Consistency

Arun Padakandla
University of Tennessee, Knoxville

Abram Magner
University at Albany, SUNY

Abstract

We formulate a quantum analogue of the fundamental classical PAC learning problem. As on a quantum computer, we model data to be encoded by modifying specific attributes - spin axis of an electron, plane of polarization of a photon - of sub-atomic particles. Any interaction, including reading off, extracting or learning from such data is via quantum measurements, thus leading us to a problem of PAC learning Quantum Measurement Classes. We propose and analyze the sample complexity of a new ERM algorithm that respects quantum non-commutativity. Our study entails that we define the VC dimension of Positive Operator Valued Measure(ments) (POVMs) concept classes. Our sample complexity bounds involve optimizing over partitions of jointly measurable classes. Finally, we identify universally consistent sequences of POVM classes. Technical components of this work include computations involving tensor products, trace and uniform convergence bounds.

1 INTRODUCTION

The blueprint of today's learning algorithms can be traced back to the foundational ideas developed in the early works of statistical learning theory (SLT). Questions such as : When is a concept class learnable? What parameters quantify its complexity regarding its learnability? How does the sample complexity depend on these parameters? were formulated four decades ago keeping in mind Turing's model of computation. 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 the above questions manifest in this new quantum computing framework? Do these lead to interesting formulations requiring new physical and mathematical ideas? These questions motivate the current work.

Learning from physical realizations of quantum states: In the classical model of computation, data stored in registers is modeled as elements taking values in sets, and computations or algorithms are modeled as functions or sequences of functions operating on these sets. In order to classify, say, an image stored in a collection of registers, a learning algorithm has to identify an optimal *function* from a concept class – a library of deterministic *functions*. Classically, the learning algorithm is able to observe the exact contents of registers arbitrarily many times.

In contrast, on a quantum computer, data is encoded on sub-atomic particles by modifying their attributes. For example, 8 bits can be stored on an electron by preparing the axis of its spin to be one among 256 outward radial directions. As we will discuss in more detail below, data stored in such a manner cannot be directly observed. Moreover, learning information about the value stored in the attributes of a particle necessarily results in a change in the value/attribute.

Our framework will focus on learning in a supervised setting, where the data is encoded, as described above, in attributes of sub-atomic particles – i.e., as a physical realization (*not* an analytical description) of a *quantum state*, and the associated labels, which we interpret as inferences about the data, are stored in classical registers. The combination of the quantum data with the classical label is referred to as a *classical-quantum* (CQ) state [Wilde, 2017, Sec. 4.3.4]. Any interaction with a sub-atomic particle is governed by the laws of quantum mechanics, and in particular, ascertaining its state or properties thereof is via a *measurement*. Therefore, the prediction of the label of a quantum

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

state must proceed via a measurement that is applied to the particle and results in a classical outcome that is interpreted as the predicted label.

The task of a quantum learning algorithm in this formulation, given a set of physical realizations of quantum states along with their associated labels, is to identify an "optimal" measurement from a fixed concept class, which now consists of a library of measurements. Measurements are formalized via the generic framework of *positive operator-valued measurements* (POVMs) [Holevo, 2019, Sec. 2.3]. A concept class thus consists of a set of POVMs.

In this work, we answer the above three questions in the context of POVM concept classes. We then initiate a study of universal consistency [Stephane Boucheron and Massart, 2013] and identify universally consistent sequences of POVM concept classes.

Challenges inherent in our setting: This natural, deceptively simple formulation hides the involved challenges. The axioms of quantum mechanics, that govern all our interaction with quantum data, manifests in three unique properties - non-commutativity, quantum indeterminism and entanglement. Non-commutativity implies that measuring a quantum state causes it to *collapse*. Post-measurement collapse results in a significant blowup of the sample complexity of the empirical risk minimization (ERM) rule, thus leading us to modify this foundational algorithm of SLT. Leveraging the elegant notion of joint measurability, also referred to as compatibility (Sec. 3.1) herein, we propose a modified ERM rule (Alg. 1 in Sec. 3.2) that prevents this blowup. Next, we quantify the effect of quantum indeterminism in the characterization of the VC dimension of a POVM concept class (Defn. 5). Our analysis (Sec. 4) reveals the functional dependence of the sample complexity (Thm. 1) on this measure of complexity. Our formulation generalizes (Rem. 1) the original PAC formulation [Vapnik and Chervonenkis, 1971, Vapnik and Chervonenkis, 1974, Vapnik, 1982, Valiant, 1984, Vapnik, 1995, Vapnik, 1998] and we recover all corresponding results (Rem. 4). We also generalize the results of [Heidari et al., 2021], which only deals with finite concept classes and a simpler notion of compatibility. In Sec. 5, we use optimal state discrimination to identify sequences of POVM concept classes that are universally consistent (Defn. 7).

Prior Work: Related prior work can be classified broadly into two sets. The first set, which includes this work, revolves around learning a quantum state [Arunachalam and de Wolf, 2017, Arunachalam et al., 2020], or its properties [Anshu et al., 2021], by measuring multiple identically prepared states. This set includes works on state tomography [O'Donnell and Wright, 2016, O'Donnell and Wright, 2017, Haah et al., 2016, Aaronson, 2006, Rehacek and Paris, 2004, Waseem et al., 2020, Altepeter et al., 2004, Aaronson, 2018, Aaronson et al., 2018], state discrimination [Bae and Kwek, 2015] and quantum property testing [Montanaro and de Wolf, 2018, Bubeck et al., 2020] that investigate the number of prepared states to accomplish the learning task. We highlight the work of Cheng et. al. [Cheng et al., 2016] that studies the related problem when the training samples form the collection $(\rho_i, \mathrm{tr}(M\rho_i)) : 1 \leq i \leq n$, where ρ_i denotes the ith quantum state, and the algorithm must find the unknown optimal POVM M. [1] See [Kearns and Schapire, 1990] for a closely related classical study. The related paper [Heidari et al., 2021] defines and studies a similar framework and learning rule to our own but only proves a sample complexity bound for *finite* POVM classes. Furthermore, it makes much stronger assumptions regarding the mutual compatibility of POVMs in the classes under study. The extension in the present work of statistical learning theory machinery to the quantum setting, necessary for proving finite upper bounds on the sample complexity for learning infinite-cardinality POVM classes, is a key contribution of our work.

The second set is comprised of works [Gortler and Servedio, 2001, Servedio, 2001, Lloyd et al., 2014, Schuld et al., 2014, Atici and Servedio, 2005, Anguita et al., 2003, Wiebe et al., 2012, Aïmeur et al., 2012] that explore the use of a quantum computer to speed up classical learning tasks. Surveys [Arunachalam and de Wolf, 2017, Khan and Robles-Kelly, 2020] provide a detailed account of works in this set.

Notation: For integer $n \in \mathbb{N}, [n] \triangleq \{1, \cdots, n\}$. We reserve $\mathcal{H}$ with appropriate subscripts to denote a finite-dimensional Hilbert space. The symbols $\mathcal{L}(\mathcal{H})$, $\mathcal{R}(\mathcal{H})$, $\mathcal{P}(\mathcal{H})$, $\mathcal{D}(\mathcal{H})$ denote the collection of linear, Hermitian, positive and density operators acting on $\mathcal{H}$ respectively. A POVM is a subcollection $\mathcal{M} \triangleq \{M_y \in \mathcal{P}(\mathcal{H}) : y \in \mathcal{Y}\}$ of indexed positive operators that sum to the identity I_H, i.e, $\sum_{y\in\mathcal{Y}} M_y = I_H$. For clarity, a POVM $\mathcal{M} = \{M_y \in \mathcal{P}(\mathcal{H}) : y \in \mathcal{Y}\}$ with outcomes in $\mathcal{Y}$ is often referred to as a $\mathcal{Y}-$POVM or $\mathcal{Y}-$POVM on $\mathcal{H}$. 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 let $\mathcal{M}^n \triangleq \{M_{y^n} \triangleq M_{y_1} \otimes \cdots \otimes M_{y_n} : y^n \in \mathcal{Y}^n\}$ with a slight abuse of notation. For a Hilbert space $\mathcal{H}_X$ and a

[1] Our work differs considerably since we are provided the actual labels, not the associated probabilities $\mathrm{tr}(M\rho_i) : 1 \leq i \leq n$. As discussed in [Kearns and Schapire, 1990], this makes the problem substantially more challenging.

set $\mathcal{Y}$, we let $\mathscr{M}_{X\rightarrow\mathcal{Y}}$ denote the set of all $\mathcal{Y}-$POVMs on $\mathcal{H}_X$. See [Holevo, 2019, Sec. 2.3] for its structure. We let $\mathscr{M}_{X^n\rightarrow\mathcal{Y}}$ denote the set of all $\mathcal{Y}-$POVMs on $\mathcal{H}_X^{\otimes n}$. We let $\mathscr{S}_{X\rightarrow\mathcal{Y}}$ denote the subcollection of sharp (projective) measurements on $\mathcal{H}_X$ with outcomes in $\mathcal{Y}$. We let $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ and $\mathcal{H}_Z\triangleq\mathcal{H}_X\otimes\mathcal{H}_Y$. 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.

2 MODEL FORMULATION, PROBLEM STATEMENT

Let $\mathcal{X}$ represent an (arbitrary) set of features, $\mathcal{Y}$ a finite set of labels and $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ the set of all distributions/densities/PMFs on $\mathcal{X}\times\mathcal{Y}$. In the classical model of computation, data and hence features are stored in registers, and operated on by functions. Therefore, random variables (RVs) $(X,Y)\in\mathcal{X}\times\mathcal{Y}$ with an unknown distribution p_{XY} 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 $l_p(f)\triangleq l_{p_{XY}}(f)\triangleq\mathbb{E}_{p_{XY}}\{l(f(X),Y)\}=\mathbb{E}_p\{l(f(X),Y)\}$ with respect to a loss function $l:\mathcal{Y}^2\rightarrow[A,B]$ bounded between $A,B\geq 0$. A learning algorithm's (LA's) task is to identify an optimal predictor – a *function* – from within a/its concept class $\mathcal{C}\subseteq\mathcal{F}_{\mathcal{X}\rightarrow Y}$ – a library of *functions*. We focus on supervised learning, wherein the LA is provided with n training samples modelled as n independent and identically distributed (IID) pairs $(X_i,Y_i)\in\mathcal{X}\times\mathcal{Y}:i\in[n]$ with unknown density p_{XY}. A $\mathcal{C}-$LA is therefore again a sequence of *functions* $A_n:(\mathcal{X}\times\mathcal{Y})^n\rightarrow\mathcal{C}:n\geq 1$, of which the ERM rule [Shalev-Shwartz and Ben-David, 2014, Sec. 2.2] is a canonical example.

Before we describe a quantum formulation, we briefly highlight two consequences of the classical model that are often taken for granted. Stored in registers and modeled as elements in a set, features can be duplicated (indiscriminately), thereby enabling one to *simultaneously* retain an "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). In other words, the outcome of evaluating a function on a feature is deterministic.

In our proposed quantum learning model, feature $x\in\mathcal{X}$ is encoded by modifying specific characteristics of a sub-atomic particle. Let $\rho_x\in\mathcal{D}(\mathcal{H}):x\in\mathcal{X}$ model the behaviour of a particle representing feature $x\in\mathcal{X}$. While $\mathcal{X}\times\mathcal{Y}$ indexes (all) possible feature-label pairs, the feature-label pair (x,y) is encoded in a physically realized quantum system modeled via density operators $\rho_{xy}\triangleq\rho_x\otimes|y\rangle\langle y|\in\mathcal{D}(\mathcal{H}_Z)$, where $\mathcal{H}_Z=\mathcal{H}_X\otimes\mathcal{H}_Y$, $\mathcal{H}_Y=\text{span}\{|y\rangle:y\in\mathcal{Y}\}$ with $\langle y|\hat{y}\rangle=\delta_{y\hat{y}}$ ensuring label distinguishability. Note that we have modeled labels, as before, to be stored in classical registers and have thus adopted the well-established notion of a *classical-quantum* (CQ) state [Wilde, 2017, Sec. 4.3.4] to model a feature-label pair.[2]

Labeled features are generated with respect to an unknown distribution p_{XY}. Specifically, the density operator of a feature-label pair is

$$\rho_{XY}\triangleq\int_{\mathcal{X}\times\mathcal{Y}}\rho_x\otimes|y\rangle\langle y|\,dp_{XY}\in\mathcal{D}(\mathcal{H}_Z) \qquad (1)$$

which reduces to $\rho_{XY}=\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}p_{XY}(x,y)\rho_x\otimes|y\rangle\langle y|$ if $|\mathcal{X}|<\infty$. A predictor is only provided the quantum system corresponding to the feature, i.e., the $X-$system. Mathematically, the density operator of the quantum state provided to a predictor is $\sum_{x\in\mathcal{X}}p_X(x)\rho_x$. We exemplify this below.

Ex. 1. *Consider electrons with spins. An electron can be prepared with the axis of its spin pointing in a direction, represented by a 3−dimensional unit vector. Let finite set* $\mathcal{X}=\{(\theta_i,\phi_j)=(\frac{i\pi}{8},\frac{j2\pi}{8}):0\leq i,j\leq 7\}$ *index unit vectors in the Bloch sphere representing the possible spin axis directions. We have two labels in* $\mathcal{Y}=\{blue,red\}$*. Nature decides to label an electron 'red' if the axis of its spin is orthonormal to a specific orthant. Otherwise the electron is labeled 'blue'. 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 *with respect to 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* $x\rightarrow\rho_x$*, and* $\mathcal{Y}$*. Oblivious to both the nature's decision and the orthant, but possessing the prepared electron, a predictor's task is to determine the label.*[3]

Previously stored in registers, features were labeled via functions. In contrast, on a quantum computing device, features encoded in quantum states can be labeled only through *measurements*. A predictor $\mathcal{M}=\{M_y\in\mathcal{P}(\mathcal{H}_X):y\in\mathcal{Y}\}$ is therefore a $\mathcal{Y}-$POVM on $\mathcal{H}_X$.

[2] A CQ state [Wilde, 2017, Sec. 4.3.4] models a scenario wherein some components of data is stored in classical registers, while others are encoded in quantum states that obey the axioms of quantum mechanics.

[3] For example, we may know that 000 is encoded as horizontal spin axis, 111 as vertical spin axis and so on. However, we are unaware of which spin axis is labeled as red or blue, i.e., we are unaware of the rule based on which a spin axis is labeled.

Since our labels are stored in registers, a conventional bounded loss function $l : \mathcal{Y} \times \mathcal{Y} \to [A, B]$ quantifies a predictor's performance. The true risk of predictor $\mathcal{M} = \{M_y : y \in \mathcal{Y}\}$ if the underlying distribution is p_{XY} is

$$l_p(\mathcal{M}) \triangleq \sum_{(x,y,\hat{y})\in\mathcal{Z}} p_{XY}(x,y)l(y,\hat{y}) \operatorname{tr}((\rho_x \otimes |y\rangle \langle y|)(M_{\hat{y}} \otimes I_Y)), \tag{2}$$

where $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}^2$. A concept class $\mathcal{C} \subseteq \mathscr{M}_{X\to Y}$ is a subcollection of *measurements*, i.e., *POVMs* with outcomes in $\mathcal{Y}$. Henceforth, a concept class, a POVM concept class and a measurement concept class all refer to the same object and are used interchangeably. For an underlying distribution p_{XY}, we let $l^*(p_{XY}, \mathcal{C}) \triangleq \inf_{\mathcal{M}\in\mathcal{C}} l_p(\mathcal{M})$ denote the risk of an optimal predictor in $\mathcal{C}$. The goal of a $\mathcal{C}-$LA is to choose a predictor $\mathcal{M} \in \mathcal{C}$ which is a POVM whose true risk $l_p(\mathcal{M})$ is close to $l^*(p_{XY}, \mathcal{C})$, while being oblivious to the underlying distribution p_{XY}. To accomplish this, a $\mathcal{C}-$LA is provided n training samples - labeled features - randomly chosen IID from p_{XY}. The density operator of the received training samples is therefore

$$\rho_{XY}^{\otimes n} \triangleq \int_{\mathcal{X}^n\times\mathcal{Y}^n} \rho_{x^n} \otimes |y^n\rangle \langle y^n| \, dp_{XY}^n \in \mathcal{D}(\mathcal{H}_Z^{\otimes n}) \tag{3}$$

where $\rho_{x^n} \triangleq \bigotimes_{i=1}^{n} \rho_{x_i}$, $|y^n\rangle \langle y^n| \triangleq \bigotimes_{i=1}^{n} |y_i\rangle\langle y_i|$.

Mathematically, any operation on a quantum state with classical outcomes is a measurement. A $\mathcal{C}-$LA must therefore measure the n training samples modeled via density operator $\rho_{XY}^{\otimes n}$ and output an index in $\mathcal{C}$.

Defn 1. *A POVM concept class for labeling features in $\mathcal{D}(\mathcal{H}_X)$ with labels in $\mathcal{Y}$ is a subcollection $\mathcal{C} \subseteq \mathscr{M}_{X\to Y}$ of POVMs. A $\mathcal{C}-$learning algorithm ($\mathcal{C}-$LA) is a sequence $A_n \triangleq \{A_{\mathcal{M}}^n \in \mathcal{P}(\mathcal{H}_Z^{\otimes n}) : \mathcal{M} \in \mathcal{C}\} \in \mathscr{M}_{Z^n\to\mathcal{C}} : n \geq 1$ of $\mathcal{C}-$POVMs on $\mathcal{H}_Z^{\otimes n}$.*

We elaborate on Def. 1 for clarity. A $\mathcal{C}-$LA is a sequence $A_n : n \geq 1$ of measurements. Each measurement in this sequence is a $\mathcal{C}-$POVM. When measurement A_n is performed on n training samples, the outcome obtained is the index of the chosen predictor in $\mathcal{C}$. Having clarified this, we now enquire how many samples does a $\mathcal{C}-$LA need to identify $\epsilon-$optimal predictor with confidence at least $1 - \delta$? We are thus led to the notion of PAC learnability of measurement spaces.

Defn 2. *An algorithm $A_n \triangleq \{A_{\mathcal{M}}^n \in \mathcal{P}(\mathcal{H}_Z^{\otimes n}) : \mathcal{M} \in \mathcal{C}\} \in \mathscr{M}_{Z^n\to\mathcal{C}} : n \geq 1$ (PAC) learns $\mathcal{C}$ if for every $\epsilon > 0$, $\delta > 0$, there exists a $N(\epsilon, \delta) \in \mathbb{N}$ such that for all $n \geq N(\epsilon, \delta)$ and every distribution p_{XY}, the probability of A_n choosing a predictor $\mathcal{M} \in \mathcal{C}$ for which the true risk $l_p(\mathcal{M}) > l^*(p_{XY}, \mathcal{C}) + \epsilon$, is at most δ. In this case, the function $N : [A, B] \times (0, 1) \to \mathbb{N}$ is the sample complexity of algorithm $A_n : n \geq 1$ in learning $\mathcal{C}$. We say that the $\mathcal{C}-$LA $A_n : n \geq 1$ efficiently PAC learns $\mathcal{C}$ if $N(\epsilon, \delta)$ is polynomial in $\frac{1}{\epsilon}, \frac{1}{\delta}$. A concept class $\mathcal{C} \subseteq \mathscr{M}_{X\to Y}$ is (efficiently) PAC learnable if there exists a $\mathcal{C}-$LA that (efficiently) PAC learns $\mathcal{C}$.*

Remark 1. *Suppose we are given a classical PAC learning problem with feature set $\mathcal{X}$, a function concept class $\mathcal{C} \subseteq \mathcal{F}_{\mathcal{X}\to\mathcal{Y}}$. We can recover this classical formulation via the following substitutions in the problem formulated herein. For this, choose $\mathcal{H}_X =$ span$\{|x\rangle : x \in \mathcal{X}\}$ with $\langle x|\hat{x}\rangle = \delta_{x\hat{x}}$ and $\rho_x = |x\rangle \langle x|$. We then encode the concept class of functions in terms of POVMs as $\mathcal{M}_y^f = \sum_{x:f(x)=y} |x\rangle \langle x|$ for each $f \in \mathcal{C}$.*

While the classical problem is well understood [Shalev-Shwartz and Ben-David, 2014, Devroye et al., 1996], the unique behaviour of quantum systems and complexity of the involved mathematical objects throws up challenges leaving the above problem, which we tackle here, unresolved.

The need to modify the ERM rule The following example illustrates the need to modify the classical ERM algorithm; we do so in Sec. 3.2 and analyze the resulting sample complexity in Sec. 4.

Ex. 2. *Suppose we are provided $n = 20$ training samples, i.e., 20 electrons with corresponding labels in registers, and our library $\mathcal{C}$ consists of 2 predictors $\mathcal{M}_1 = \{M_{1b}, M_{1r}\}, \mathcal{M}_2 = \{M_{2b}, M_{2r}\}$. An ERM rule would attempt to measure all 20 electrons with both measurements and choose that measurement, whose outcomes disagree least with the provided labels. However, every measurement alters the spin, rendering it unusable to perform the next measurement. Moreover, electron spins cannot be replicated (No Cloning Theorem [Wootters and Zurek, 1982]). This indicates that every training sample can be used to evaluate the empirical risk (ER) of just one measurement.*

Unlike in Remark 1, non-commutativity of quantum measurements and the No Cloning theorem [Wootters and Zurek, 1982] suggest that each training sample can be used to evaluate the ER of just one measurement. Moreover the outcome of measurements are random (Remark 2) resulting in random empirical risks. How should a $\mathcal{C}-$LA optimally utilize training samples to identify the best predictor from within $\mathcal{C}$? The discussion thus far, exemplified through Ex. 1 - 2, suggests that each training sample can yield the ER of just one measurement. This results in a blow-up of the sample complexity of any concept class. Our first simple step (Sec. 3.1) is to leverage *compatible* measurements [Holevo, 2019, Sec. 2.3.2] to 're-use' samples effectively. We build on this to derive an upper bound on the sample complexity. In contrast with the previous work [Heidari et al., 2021], our work necessitates a substantial extension to cover the ubiquitous case of infinite-cardinality concept classes and to generalize beyond compatible partitions defined in terms of commuting POVMs.

3 STOCHASTICALLY COMPATIBLE POVMS AND A NEW ERM RULE

While classical register contents maybe duplicated, thereby permitting simultaneous evaluations of all functions in $\mathcal{C} \subseteq \mathcal{F}_{X \to Y}$, the No Cloning theorem [Wootters and Zurek, 1982] precludes reproducing quantum states. Moreover, any measurement results in the *collapse* of a quantum state: performing POVM $\mathcal{M}_1 = \{M_{1y} \otimes I_Y : y \in \mathcal{Y}\}$ on each training sample results in the modified training samples described via $\tilde{\rho}_{XY}^{\otimes n}$, where $\tilde{\rho}_{XY} = \sum_{z \in \mathcal{Y}} \int \sqrt{M_{1z}} \rho_x \sqrt{M_{1z}} \otimes |y\rangle \langle y| \, dp_{XY}$. Performing subsequent POVMs on the collapsed training samples does not enable us to estimate the true risk of these predictors.

3.1 Stochastically Compatible Measurements

The ERM rule we propose is based on the divide-and-conquer approach. While POVMs are generally incompatible, certain subsets are jointly measurable [Holevo, 2019, Sec. 2.1.2], leading us to the following.[4]

Defn 3. *A subcollection $\mathscr{B} \subseteq \mathscr{M}_{X \to \mathcal{Y}}$ of $\mathcal{Y}-$POVMs on $\mathcal{H}_X$ is compatible if there exists a finite outcome set $\mathcal{W}$, a $\mathcal{W}-$POVM $\mathcal{G} = \{G_w \in \mathcal{P}(\mathcal{H}_X) : w \in \mathcal{W}\}$ on $\mathcal{H}_X$ and a stochastic matrix $\alpha^{\mathcal{B}}_{\mathcal{Y}|\mathcal{W}} : \mathcal{W} \to \mathcal{Y}$ for each $\mathcal{B} = \{B_y \in \mathcal{P}(\mathcal{H}_X) : y \in \mathcal{Y}\} \in \mathscr{B}$ such that*

$$B_y = \sum_{w \in \mathcal{W}} G_w \alpha^{\mathcal{B}}_{\mathcal{Y}|\mathcal{W}}(y|w) \text{ for all } y \in \mathcal{Y} \text{ and all } \mathcal{B} \in \mathscr{B}. \quad (4)$$

In this case, we say $\mathscr{B}$ is compatible with fine-graining $(\mathcal{W}, \mathcal{G}, \alpha^{\mathscr{B}}_{\mathcal{Y}|\mathcal{W}}) \triangleq (\mathcal{W}, \mathcal{G}, \alpha^{\mathcal{B}}_{\mathcal{Y}|\mathcal{W}} : \mathcal{B} \in \mathscr{B})$.

In other words, a compatible collection of POVMs can be applied by first applying a *single* POVM to a quantum state, then applying a noisy classical channel to the output of this POVM. The classical channel is unique to each individual POVM in the compatible collection.

From (4), it is evident that to obtain the joint statistics of all POVMs in a compatible subcollection $\mathscr{B}$, one can perform POVM $\mathcal{G}$ and pass the outcome through the stochastic matrix $\prod_{\mathcal{B} \in \mathscr{B}} \alpha^{\mathcal{B}}_{\mathcal{Y}|\mathcal{W}}$. This suggests that we partition the POVM concept class $\mathcal{C}$ into compatible subcollections and 're-use' training samples amongst POVMs within a compatible subcollections. Before we state this, we identify an important example of a compatible subcollection.

Remark 2. *A subcollection $\mathscr{B} \subseteq \mathcal{M}_{X \to \mathcal{Y}}$ of $\mathcal{Y}-$POVMs on $\mathcal{H}_X$ are compatible if for any pair $\mathcal{M} = \{M_y : y \in \mathcal{Y}\} \in \mathscr{B}, \mathcal{L} = \{L_y : y \in \mathcal{Y}\} \in \mathscr{B}$, we have $M_y L_{\hat{y}} = L_{\hat{y}} M_y$ for all $(y, \hat{y}) \in \mathcal{Y} \times \mathcal{Y}$.*

While commutative POVMs are compatible, the converse is not necessarily true. In fact, the problem of characterizing compatible POVMs remains active [Jae et al., 2019], [Guerini and Terra Cunha, 2018], [Kunjwal, 2014] [Karthik et al., 2015]. Notwithstanding this, the notion of compatibility provides us with the appropriate construct to modify a naive ERM rule that resulted in blowup of sample complexity. We specify the new ERM rule in the following.

3.2 ERM Rule for Learning Quantum Measurement Concept Classes

The idea of the new ERM rule is to *partition* the POVM concept class $\mathcal{C}$ into compatible subcollections. The set of training samples is also partitioned analogously so that there is a $1:1$ correspondence between the two partitions. The number n_i of samples allotted to a given partition element with index i in the compatible partition is chosen in accordance with our sample complexity upper bound in Theorem 1. That is, we choose $n_i = \frac{8(B-A)\log(1/\delta) + 8B^2 VC(\mathscr{B}_i)}{\epsilon^2}$, where the VC dimension $VC(\cdot)$ is defined in subsequent discussion. Each subset of the training samples is employed to evaluate the ER of all POVMs in the corresponding compatible subset. Finally, the POVM with the least ER among all POVMs is the chosen predictor. The following definition enables us specify the proposed ERM rule precisely.

Defn 4. *Let $\mathcal{C} \subseteq \mathscr{M}_{X \to \mathcal{Y}}$. We say $(\mathscr{B}_i : i \in I)$ is a compatible partition of $\mathcal{C}$ if (i) $\mathscr{B}_i \subseteq \mathcal{C}$ is compatible for each $i \in I$, (ii) $\mathscr{B}_i \cap \mathscr{B}_{\hat{i}} = \phi$ whenever $\hat{i} \neq i$ and (iii) $\bigcup_{i \in I} \mathscr{B}_i = \mathcal{C}$. In this case, suppose $\mathscr{B}_i$ is compatible with fine-graining $(\mathcal{W}_i, \mathcal{G}_i, \alpha^i_{\mathcal{Y}|\mathcal{W}_i}) \triangleq (\mathcal{W}_i, \mathcal{G}_i, \alpha^{\mathscr{B}_i}_{\mathcal{Y}|\mathcal{W}_i})$ for each $i \in I$, we say $\mathscr{B}_I \triangleq (\mathscr{B}_i : i \in I)$ is a compatible partition of $\mathcal{C}$ with fine-graining $(\mathcal{W}_i, \mathcal{G}_i, \alpha^i_{\mathcal{Y}|\mathcal{W}_i}) : i \in I$. Furthermore, we let $\mathfrak{B}$ denote the set of all com-*

[4] Note that our notion of *compatible* is referred to as stochastically compatible by Holevo [Holevo, 2019, Sec. 2.1.2, Pg. 13]. *Compatible* in [Holevo, 2019, Sec. 2.1.2] corresponds to the case when the stochastic matrices are $0-1$ valued.

patible partitions of POVM concept class $\mathcal{C} \subseteq \mathscr{M}_{X \to \mathcal{Y}}$.

We give a finite example below.

Ex. 3. *Suppose $\mathcal{Y} = \{0,1\}$, $\mathcal{C} = \{\mathcal{M}_1, \cdots, \mathcal{M}_5\}$ contains 5 POVMs. Suppose $\mathscr{B}_1 = \{\mathcal{M}_1, \mathcal{M}_2\}$ and $\mathscr{B}_2 = \{\mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5\}$ is a compatible partition of $\mathcal{C}$ (Defn. 4). Specifically, suppose $\mathscr{B}_1$ is compatible with fine-graining $(\mathcal{W}_1, \mathcal{G}_1, \alpha^1_{\mathcal{Y}|\mathcal{W}_1}, \alpha^2_{\mathcal{Y}|\mathcal{W}_1})$ and $\mathscr{B}_2$ is compatible with fine graining $(\mathcal{W}_2, \mathcal{G}_2, \alpha^3_{\mathcal{Y}|\mathcal{W}_2}, \alpha^4_{\mathcal{Y}|\mathcal{W}_2}, \alpha^5_{\mathcal{Y}|\mathcal{W}_2})$. In order to obtain outcomes that are statistically indistinguishable from that of the outcomes of POVMs in $\mathscr{B}_1$, one can perform $\mathcal{W}_1-$POVM $\mathcal{G}_1$ on the quantum state and post-process the outcome $w_1 \in \mathcal{W}_1$ by passing it through the product stochastic matrix*

$$\alpha^1_{\mathcal{Y}|\mathcal{W}_1}(y_1, y_2|w_1) \triangleq \alpha^{\mathscr{B}_1}_{\mathcal{Y}|\mathcal{W}_1}(y_1, y_2|w_1) \quad (5)$$

$$\triangleq \alpha^{\mathcal{M}_1}_{\mathcal{Y}|\mathcal{W}_1}(y_1|w_1)\alpha^{\mathcal{M}_2}_{\mathcal{Y}|\mathcal{W}_1}(y_2|w_1) \quad (6)$$

where $(w_1, y_1, y_2) \in \mathcal{W}_1 \times \mathcal{Y}^2$.

Ex. 3 identifies a compatible partition with two elements, i.e., $I = \{1,2\}$ with $t = 2$. The product stochastic matrix for $i = 1$ is explicitly stated in the example. The outcomes $(w^{(i)}_j : 1 \leq j \leq n_i)$ are the n_i outcomes one obtains on performing POVM $\mathcal{G}_i$ on the n_i quantum states. Each of these is passed through the product stochastic matrix $\alpha^i_{\mathcal{Y}|\mathcal{W}_i} = \alpha^{\mathscr{B}_i}_{\mathcal{Y}|\mathcal{W}_i}$. We have denoted the resulting collection of outcomes $(\hat{y}^{\mathcal{B}}_j : \mathcal{B} \in \mathscr{B}_i) : 1 \leq j \leq n_i$ to distinguish it from the labels provided as part of the training data. These were denoted as y_1, y_2 in Ex. 3. The remaining steps facilitate identifying the ERM POVM in each partition followed by identifying the global ERM POVM that is returned.

Remark 3. *If $\mathscr{B} = \{\mathcal{M}_j : 1 \leq j \leq J\}$ consists of commuting $\mathcal{Y}-$POVMs $\mathcal{M}_j = \{M_{jy} \in \mathcal{P}(\mathcal{H}) : y \in \mathcal{Y}\} : 1 \leq j \leq J$, then $\mathscr{B}$ is compatible with fine-graining $(\mathcal{Y}^J, \prod_{j=1}^J \mathcal{M}_j, \mathbb{1}_{Y_j|\underline{Y}})$, where $\mathcal{Y}^J = \times_{j=1}^J \mathcal{Y}_j$ is the Cartesian product, $\prod_{j=1}^J \mathcal{M}_j \triangleq \{M_{1y_1} M_{2y_2} \cdots M_{Jy_J} \in \mathcal{P}(\mathcal{H}) : (y_1, \cdots, y_J) \in \mathcal{Y}^J\}$ is the product POVM and $\mathbb{1}_{Y_j|\underline{Y}}(y|y_1, \cdots, y_J) = \mathbb{1}_{\{y=y_j\}}$ is the co-ordinate function.*

Remark 4. *For a set $\mathcal{X}$, consider the Hilbert space $\mathcal{H}_X \triangleq$ span$\{|x\rangle : x \in \mathcal{X}\}$ with $\langle \hat{x}|x\rangle = \delta_{\hat{x}x}$. The set of all stochastic matrices on $\mathcal{X}$ (that subsumes functions) forms a commuting set of operators on $\mathcal{H}_X$. Hence the classical PAC learning problem reduces to a POVM concept class in which all operators commute. The ERM-Q therefore reduces to the classical ERM with one commuting subset of POVMs.*

Remark 5. *While the ERM-Q algorithm 're-uses' samples to obtain ER of POVMS within a compatible subset, it makes no attempt to 're-use' samples across subsets of a compatible partition. Using $\{\rho_x : x \in \mathcal{X}\}$ and the obtained outcomes one can, in theory, characterize all possible collapsed states and 're-use' samples across partitions to glean partial information on the ER of incompatible POVMs. Since POVMs obfuscate phase information of the collapsed state, we have not attempted to explore this direction in this first step. Our sample complexity of ERM-Q is therefore only an upper bound.*

Algorithm 1: ERM-Quantum (ERM-Q) Algorithm

```
Input: POVM Concept class $\mathcal{C}$ and $n$ training samples with $(y_1, \cdots, y_n)$ denoting the labels.
Output: Index of the selected predictor in $\mathcal{C}$
Identify a compatible partition $\mathscr{B}_I \triangleq (\mathscr{B}_i : i \in I)$ of $\mathcal{C}$ with fine graining $(\mathcal{W}_i, \mathcal{G}_i, \alpha^i_{\mathcal{Y}|\mathcal{W}_i}) : i \in I$ and let $t = |I|$. Henceforth, let $\alpha^i_{\mathcal{Y}|\mathcal{W}_i} = \prod_{\mathcal{B} \in \mathscr{B}_i} \alpha^{\mathcal{B}}_{Y|\mathcal{W}_i}$ be the product of the stochastic matrices.
Partition $n$ training samples into $t = |I|$ subsets with $i-$th subset having $n_i$ training samples. Let $(y^{(i)}_j : 1 \leq j \leq n_i)$ denote the training sample labels of samples provided in $i-$th subset. $n_i$ is determined in relation to $\mathtt{VC}(\mathscr{B}_i)$ appearing in (9)
for $i = 1$ to $t$ do
    Perform $\mathcal{W}_i-$POVM $\mathcal{G}_i$ on each of the $n_i$ samples. For $j = 1, \cdots, n_i$, let $w^{(i)}_j \in \mathcal{W}_i$ denote the outcome of the POVM on the $j-$th quantum state.
    Postprocess $(w^{(i)}_j : 1 \leq j \leq n_i)$ by passing each component through the stochastic matrix $\alpha^i_{\mathcal{Y}|\mathcal{W}_i}$. For POVM $\mathcal{B} \in \mathscr{B}_i \subseteq \mathcal{C}$, let $\hat{y}^{\mathcal{B}}_j$ denote the post-processed outcome corresponding to sample $j$.
    Compute ER $\mathtt{ER}(\mathcal{B}, \mathscr{B}_i) \triangleq \frac{1}{n_i}\sum_{j=1}^{n_i} l(\hat{y}^{\mathcal{B}}_j, y^{(i)}_j)$ of POVM $\mathcal{B} \in \mathscr{B}_i \subseteq \mathcal{C}$.
    Let $\mathcal{B}^*_i = \arg\min_{\mathcal{B} \in \mathscr{B}_i} \mathtt{ER}(\mathcal{B}, \mathscr{B}_i)$.
return Index $\arg\min_{i \in I} \mathcal{B}^*_i$ that globally minimizes ER.
```

4 AN UPPER BOUND ON THE SAMPLE COMPLEXITY OF POVM CONCEPT CLASSES

We derive an upper bound on the sample complexity of ERM-Q algorithm. As noted before, in contrast to the classical PAC, there are two points of departure - non-commutativity and random outcomes. The former led us to partitioning $\mathcal{C}$. The latter forces us to deal with randomness in outcome labels. We provide several steps of the proof in Sec. 4.2. To aid clarity and readability, probability of events are expressed as expectations of corresponding indicator RVs, which is further expressed as a summation. E.g., if X, Y, Z has a distribution $p_{XY}p_{Z|X}$, then $P(l(Y,Z) > a)$ is written as $\sum_{x,y,z} p_{XY}(x,y)p_{Z|X}(z|x)\mathbb{1}_{\{l(y,z)>a\}}$. This is done only to permit an uncluttered integration of quantum and classical probability. We emphasize that none of the arguments rely on summations confined to finite ranges or suprema being maxima.

4.1 VC Dimension of Compatible POVMs and Sample Complexity of ERM-Q

The divide-and-conquer approach of ERM-Q is naturally reflected in its sample complexity. Below, we define a notion of VC dimension of a compatible concept class and use it to upper bound the sample complexity of ERM-Q.

Defn 5. *Let $\mathscr{B} = \{\mathcal{B}_k : k \in \mathcal{K}\}$ be a compatible partition with fine-graining $(\mathcal{W}, \mathcal{G} \triangleq \mathcal{G}_{\mathcal{W}}, \alpha_k \triangleq \alpha^k_{\mathcal{Y}|\mathcal{W}} : k \in \mathcal{K})$ where $\mathcal{G} = \{G_w \in \mathcal{P}(\mathcal{H}) : w \in \mathcal{W}\}$. Let*

$$\beta_k(y_k|x) \triangleq \sum_{w \in \mathcal{W}} \operatorname{tr}(\rho_x G_w)\alpha_k(y_k|w) \tag{7}$$

and

$$\beta_{\mathcal{K}}(y_k : k \in \mathcal{K}|x) = \sum_{w \in \mathcal{W}} \operatorname{tr}(\rho_x G_w) \prod_{k \in \mathcal{K}} \alpha_k(y_k|w). \tag{8}$$

For $x \in \mathcal{X}$, we say $(y_k : k \in \mathcal{K})$ is reachable from x if $\beta_{\mathcal{K}}(y_k : k \in \mathcal{K}|x) > 0$. Let $r \in \mathbb{N}$ and let $\underline{\underline{y}} \triangleq (\underline{y}_1, \cdots, \underline{y}_r) \in \mathcal{Y}^{|\mathcal{K}|} \times \cdots \times \mathcal{Y}^{|\mathcal{K}|}$ where $\underline{y}_i = (y_{ki} : k \in \mathcal{K})$ for $i \in [r]$. We say $\underline{\underline{y}}$ is reachable from $\underline{x} = (x_1, \cdots, x_r) \in \mathcal{X}^r$ if $\prod_{i=1}^r \beta_{\mathcal{K}}(\underline{y}_i|x_i) > 0$. Let $\theta_r(\mathscr{B}, \underline{x}) \triangleq \{\underline{\underline{y}} \in \mathcal{Y}^{|\mathcal{K}|r} : \underline{\underline{y}} \text{ is reachable from } \underline{x}\}$. We say $\mathcal{B}$ shatters $\underline{x} \in \mathcal{X}^r$ if $|\theta_r(\mathscr{B}, \underline{x})| = |\mathcal{Y}|^r$. The VC dimension of $\mathscr{B}$, denoted $\mathsf{VC}(\mathscr{B})$, is the maximal $d \in \mathbb{N}$ for which there exists a set $\underline{x} \in \mathcal{X}^d$ that is shattered by $\mathscr{B}$. Finally, we let $\mathcal{S}_r(\mathscr{B}) \triangleq \max_{\underline{x} \in \mathcal{X}^r} |\theta_r(\mathscr{B}, \underline{x})|$.

This leads to the following theorem.

Theorem 1. *The sample complexity $N : [A, B] \times (0, 1) \to \mathbb{N}$ of the ERM-Q algorithm in learning $\mathcal{C}$ is bounded by*

$$N(\epsilon, \delta) \leq \tag{9}$$

$$\min_{(\mathscr{B}_1, \cdots, \mathscr{B}_{|I|}) \in \mathfrak{B}} \frac{8((B - A)\log\left(\frac{|I|}{\delta}\right) + B^2 \sum_{i=1}^{|I|} \mathsf{VC}(\mathscr{B}_i)))}{\epsilon^2}$$

where the minimum is over the collection $\mathfrak{B}$ of all compatible partitions of $\mathcal{C}$ as defined in Defn. 4.

Remark 6. *From the definition of reachability and $\theta_r(\mathscr{B}, \underline{x})$, it is evident that the VC dimension of POVM classes is, owing to its indeterminsim, quite high. Indeed, any set of outcome labels that get a non-zero probability of occurrence, no matter how small, is reachable. This increases the sample complexity of POVM classes. Quantum indeterminism is unfortunately unavoidable. Additionally, our bound is finite only when a finite-cardinality compatible partition of the hypothesis class exists. This condition naturally holds in certain use cases: e.g., where an experimenter has access to finitely many "root" measurement devices (POVMs), and they form an infinite-cardinality concept class by passing the results of these POVMs through infinitely many classical channels. In this scenario, the natural partition is immediate. It is likely that our results can be further extended to the case where we only require a looser version of compatibility among POVMs in a partition element. Since the space of POVMs of a given dimension is compact, finitude of the partition may then be guaranteed. However, the problem of actually specifying a partition for which our bound is minimized can likely only be solved under further assumptions on the hypothesis class.*

Our provided sample complexity bound is a nontrivial extension of the techniques in the classical learning theory case to the quantum framework, which is necessary for showing finite sample complexity bounds for infinite-cardinality learnable hypothesis classes. Our upper bound is likely not tight for a wide variety of POVM classes. Thus, it is of interest to further improve our upper bound and to provide sample complexity lower *bounds that match upper bounds. Ultimately, the goal is to give necessary and sufficient conditions for learnability of POVM classes, and the present work is a first step in this direction.*

4.2 Proof of Theorem 1 : Outline

To indicate the general direction of the proof, we identify here two main terms T_1 and T_2 that need to be bounded on the above in order to derive sample complexity. Upper bounds on the two main terms are derived in the supplementary material.

Preliminaries and Notation: Let $\mathscr{B}_1, \cdots, \mathscr{B}_{|I|}$ be a compatible partition of $\mathcal{C}$. Throughout, we focus on proving the uniform convergence property [Shalev-Shwartz and Ben-David, 2014, Defn. 4.3] of one partition element $\mathscr{B}_1$, henceforth denoted $\mathscr{B}$. Let $\mathscr{B} = \{\mathcal{M}_1, \cdots, \mathcal{M}_K\} \subseteq \mathcal{C}$ be compatible, with fine graining $(\mathcal{W}, \mathcal{G}_{\mathcal{W}}, \alpha^k_{Y|W} : 1 \leq k \leq K)$. Henceforth, we let $\alpha_k = \alpha^k_{Y|W} : k \in [K]$ and $\mathcal{G} = \mathcal{G}_{\mathcal{W}}$. To prevent overloading, we let $\mathcal{Z} = \mathcal{Y}$ and use z, Z_k, z_k to refer to the outcome of the POVMs in $\mathscr{B}$ on the $X-$system of the training sample. Recall that ERM-Q performs POVM $\mathcal{G}^n = \{G_{w_1} \otimes \cdots \otimes G_{w_n} : (w_1, \cdots, w_n) \in \mathcal{W}^n\}$ on $\rho_X^{\otimes n}$, where $\rho_X = \int \rho_x dp_X = \sum_{x \in \mathcal{X}} p_X(x)\rho_x$ with p_X unknown. Each of the n outcomes $W_i : i \in [n]$ (RVs) is passed through stochastic matrix $(\alpha_{[K]}(\underline{z}|w) = \alpha_{[K]}(z_1, \cdots, z_K|w) : (\underline{z}, w) = (z_1, \cdots, z_K, w) \in \mathcal{Z}^K \times \mathcal{W})$, where[5]

$$\alpha_{[K]}(z_1, \cdots, z_K|w) \triangleq \prod_{k=1}^{K} \alpha_k(z_k|w) \tag{10}$$

for $(z_1, \cdots, z_K, w) \in \mathcal{Z}^K \times \mathcal{W}$. Provided with quantum states corresponding to features $(x_1, \cdots, x_n)$, it can be verified that the K random labels $\underline{Z}_i \triangleq (Z_{1i}, Z_{2i}, \cdots, Z_{Ki})$ output by the ERM-Q corresponding to state ρ_{x_i} have distribution

$$\begin{aligned} \beta_{[K]}(\underline{z}_i|x_i) &\triangleq \beta_{[K]}(z_{1i}, \cdots, z_{Ki}|x_i) \\ &\triangleq \sum_{w \in \mathcal{W}} \mathrm{tr}(\rho_{x_i} G_w)\alpha_{[K]}(z_{1i}, \cdots, z_{Ki}|w) \end{aligned} \tag{11}$$

where $(w, \underline{z}_i) \in \mathcal{W} \times \mathcal{Z}^K$. Moreover, the n vectors $\underline{\underline{Z}} \triangleq (\underline{Z}_i : i \in [n])$ are mutually independent and we let $\beta^n_{[K]}(\underline{\underline{z}}|\underline{x}) \triangleq \prod_{i=1}^n \beta_{[K]}(\underline{z}_i|x_i)$ specify the distribution of the outcome label RVs $\underline{\underline{Z}}$. Moreover, from (10), (11), the marginal of Z_{ki} conditioned on feature x_i is $\beta_k(z_{ki}|x_i) \triangleq \sum_{w \in \mathcal{W}} \mathrm{tr}(\rho_{x_i} G_w)\alpha_k(z_{ki}|w)$. Using all of this, it can be verified that the uniform convergence property [Shalev-Shwartz and Ben-David, 2014, Defn. 4.3] demands that we characterize $N(\epsilon, \delta) \in \mathbb{N}$ for which

$$\begin{aligned} &\sup_{p_{XY}} \sum_{\underline{x} \in \mathcal{X}^n} \sum_{\underline{y} \in \mathcal{Y}^n} \sum_{\underline{z}_1 \in \mathcal{Z}^K} \\ &\quad \cdots \sum_{\underline{z}_n \in \mathcal{Z}^K} p^n_{XY}(\underline{x}, \underline{y})\beta^n_{[K]}(\underline{\underline{z}}|\underline{x})\mathbb{1}_{\{\phi(\underline{x},\underline{y},\underline{\underline{z}})>\epsilon\}} \leq \delta, \end{aligned} \tag{12}$$

where $\phi(\cdot)$ is the maximal deviation of the empirical risk from the true risk, over all partition element indices k. Since $\mathbb{1}_{\{\phi(\underline{x},\underline{y},\underline{\underline{z}})>\epsilon\}} \leq \mathbb{1}_{\{\phi(\underline{x},\underline{y},\underline{\underline{z}})-\mathbb{E}\{\phi(\underline{X},\underline{Y},\underline{\underline{Z}})\}>\frac{\epsilon}{2}\}} + \mathbb{1}_{\{\mathbb{E}\{\phi(\underline{X},\underline{Y},\underline{\underline{Z}})\}>\frac{\epsilon}{2}\}}$, it suffices to derive conditions under which $T_1 \leq \delta$, where we define

$$\begin{aligned} &T_1 \triangleq \\ &\sum_{\underline{x},\underline{y},\underline{\underline{z}}} p^n_{XY}(\underline{x}, \underline{y})\beta^n_{[K]}(\underline{\underline{z}}|\underline{x})\mathbb{1}_{\{\phi(\underline{x},\underline{y},\underline{\underline{z}})-\mathbb{E}\{\phi(\underline{X},\underline{Y},\underline{\underline{Z}})\}>\frac{\epsilon}{2}\}} \end{aligned} \tag{13}$$

and $T_2 \triangleq \mathbb{E}\{\phi(\underline{X}, \underline{Y}, \underline{\underline{Z}})\} \overset{(ii)}{\leq} \frac{\epsilon}{2}$, where we have suppressed the summation ranges, and let $\underline{\underline{z}} = (\underline{z}_1, \cdots, \underline{z}_n) \in \mathcal{Z}^{KN}$.

Outline of Key Steps: The crux of the proof lies in deriving upper bounds on T_1 and T_2. The main tool in bounding T_1 is the Bounded Difference Inequality (BDI) [Stephane Boucheron and Massart, 2013, Thm. 6.2]. The bound on T_2 is derived using the VC inequality and Massart's lemma Specifically, the ghost sample technique followed by the symmetrization technique leads us to the Rademacher complexity of the POVM concept class. Finally, using Massart's lemma, we derive a bound on T_2. A complete proof is given in the supplementary material.

5 UNIVERSAL CONSISTENCY OF POVM CLASSES

We now identify specific sequences of POVM concept classes that are *universally consistent* [Devroye et al., 1996, Chap. 6]. We identify natural universally consistent sequences of POVM concept classes. We also discuss techniques for identifying other universally consistent sequences. Ours being the first in this theoretical line of study, we do not stress on the challenges of designing such classes. We begin by defining universal consistency and introduce a sequence of POVM concept classes of interest.

Defn 6. *Consider a distribution $p_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, a collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$ of density operators, the density operator of the feature-label quantum state $\rho_{XY} \triangleq \int_{\mathcal{X}\times\mathcal{Y}} \rho_x \otimes |y\rangle\langle y| \, dp_{XY} \in \mathcal{D}(\mathcal{H}_Z)$ and a loss function $l : \mathcal{Y}^2 \to [A, B]$. We let $l_p^* \triangleq \inf_{\mathcal{M} \in \mathscr{M}_{X\to Y}} l_p(\mathcal{M})$ denote the Bayes' risk.[6] A sequence $\mathcal{C}_k \subseteq \mathscr{M}_{X\to Y} : k \geq 1$ of POVM concept classes is universally consistent if $\lim_{k\to\infty} l_p(\mathcal{C}_k) = l_p^*$ for every $p_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, where for any $\mathcal{B} \subseteq \mathscr{M}_{X\to Y}$, we define $l_p(\mathcal{B}) \triangleq \inf_{\mathcal{M} \in \mathcal{B}} l_p(\mathcal{M})$.*

Universally consistent POVM concept classes enables us to squeeze the approximation error[7] to 0. Specifically, suppose a sequence $\mathcal{M}_k \in \mathscr{M}_{X\to Y} : k \geq 1$ is universally consistent, and if we can find a sequence $k(n) \in \mathbb{N}$ for $n \in \mathbb{N}$ such that $\lim_{n\to\infty} k(n) = \infty$, then the loss of this sequence $\lim_{n\to\infty} l_p(\mathcal{M}_{k(n)}) = l_p^*$ achieves the Bayes' risk. It is therefore of interest to find a sequence $k(n) : n \geq 1$ that grows slowly enough such that the estimation error also goes to zero. Next, we design a sequence of POVM concept classes that are universally consistent.

[5] The reader may recall Ex. 3

[6] $l_p(\mathcal{M})$ is defined in (2). We have done away with the subscripts XY in p_{XY} while denoting l_p^* and $l_p(\mathcal{M})$ to reduce notational clutter.

[7] See [Shalev-Shwartz and Ben-David, 2014, Sec. 5.2] for definitions of approximation and estimation errors.

5.1 Optimal state discrimination yields universal consistency

We recall that the learning algorithm is ignorant of the collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$ of density operators and is only provided the prepared quantum states and the corresponding labels. Our approach at designing universally consistent POVM concept classes is based on optimal state discrimination. Suppose the collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$ was perfectly distinguishable via a measurement $\mathcal{G} \in \mathscr{M}_{\mathcal{H}_X \to \mathcal{X}}$. Here $\mathscr{M}_{\mathcal{H}_X \to \mathcal{X}}$ denotes the set of all $\mathcal{X}-$POVMs on $\mathcal{H}_X$ and $\mathcal{G} = \{G_x : x \in \mathcal{X}\}$ has $|\mathcal{X}|$ positive operators that sum to the identity on $\mathcal{H}_X$ that satisfy $\mathrm{tr}(G_x \rho_{\hat{x}}) = \delta_{x\hat{x}}$. Then any universally consistent sequence of function concept classes can be adapted to form a universally consistent sequence of POVM concept classes. Our approach when the collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$ is not perfectly distinguishable is to employ an optimal state discriminator $\mathcal{G} \in \mathscr{M}_{\mathcal{H}_X \to \mathcal{X}}$ for the collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$.

To find the optimal state discriminator $\mathcal{G} \in \mathscr{M}_{\mathcal{H}_X \to \mathcal{X}}$ for the collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$, we turn to [Holevo, 2019, Sec. 2.4], wherein it is proven that the problem of identifying optimal state discriminator is one of maximizing an affine function over a convex set of observables. We next state the necessary facts.

Fact 1. *Given a Hilbert space $\mathcal{H}_X$ of finite dimension and a collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$ of density operators, there exists an optimal state discriminator $\mathcal{G} = \{G_x \in \mathcal{P}(\mathcal{H}) : x \in \mathcal{X}\} \in \mathscr{M}_{\mathcal{H}_X \to \mathcal{X}}$ such that, the probability of error in identification is minimized. Moreover, identifying an optimal state discriminator $\mathcal{G}$ is equivalent to maximizing an affine function over a convex set of observables.*

Fact 1 is stated in [Holevo, 2019, Sec. 2.4, Thm. 2.22]. Additionally, [Holevo, 2019, Ex. 2.27] proves that optimal quantum state discrimination is facilitated via unsharp observables - a phenomenon not observed in classical discrimination. While an analytical characterization of an optimal state discriminator $\mathcal{G}$ might be available only for specific structured collections $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$ of density operators, and not in general, the fact that $\mathcal{G}$ can be identified computationally efficiently via convex programming leads us to the use of this tool in our identification of a universally consistent sequence of POVM concept classes.

The second fact gives a universally consistent sequence of *function* concept classes.

Fact 2. *Let $\mathcal{X}$ be an arbitrary domain set with a distance metric $d(\cdot,\cdot)$ and $\mathcal{Y}$ be a finite set of labels. Let $l : \mathcal{Y}^2 \to [A, B]$ be a loss function and $\mathcal{F}_{\mathcal{X}\to\mathcal{Y}} \triangleq \{f : \mathcal{X} \to \mathcal{Y}\}$ be the set of all functions with domain $\mathcal{X}$ and range $\mathcal{Y}$. For a distribution $p_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ a function $f \in \mathcal{F}_{\mathcal{X}\to\mathcal{Y}}$ and any subcollection $\mathcal{H} \subseteq \mathcal{F}$, we let*

$$l_p(f) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} l(f(x), y) dp_{XY}(x, y), \quad l_p(\mathcal{H}) \triangleq \inf_{f\in\mathcal{H}} l_p(f), \tag{14}$$

and let $l_p^* \triangleq \inf_{f\in\mathcal{F}} l_p(f)$ denote the Bayes' risk. For $k \in \mathbb{N}$, let $\{A_{k,j} \subseteq \mathcal{X} : j \in \mathbb{N}\}$ be a partition of $\mathcal{X}$ such that $\lim_{k\to\infty} \sup_{j\in\mathbb{N}} diam(A_{k,j}) = 0$, where $diam(B) = sup_{c,d\in B} d(c, d)$ for $B \subseteq \mathcal{X}$ denotes the diameter of a set $B \subseteq \mathcal{X}$. For $k \in \mathbb{N}$, let

$$\mathcal{H}_k \triangleq \{h \in \mathcal{F}_{\mathcal{X}\to\mathcal{Y}} : h(a) = h(b) \tag{15}$$

$$\iff \exists j \in \mathbb{N}, a, b \in A_{k,j}\}. \tag{16}$$

Then, $\mathcal{H}_k : k \in \mathbb{N}$ is a universally consistent sequence of function concept classes, i.e., $\lim_{k\to\infty} l_p(\mathcal{H}_k) = l_p^*$ for each distribution $p \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

Fact 2 is a direct consequence of [Stephane Boucheron and Massart, 2013, Thm. 6.1]. We are now set to define a sequence of POVM concept classes that are universally consistent.

Defn 7. *Let $\mathcal{X}$ be a domain set, $\mathcal{Y}$ be a set of labels, $\mathcal{H}_X$ be a Hilbert space, $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$ be density operators of the quantum states encoding the features, $\mathcal{G} = \{G_x : x \in \mathcal{X}\} \in \mathscr{M}_{\mathcal{H}_X \to \mathcal{Y}}$ be a optimal state discriminator for the collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$ as defined in Fact 1 and $\mathcal{H}_k \subseteq \mathcal{F}_{\mathcal{X}\to\mathcal{Y}} : k \geq 1$ be a universally consistent sequence of function concept classes as stated in Fact 2. For $k \in \mathbb{N}$, $h \in \mathcal{H}_k$, we let*

$$\mathcal{M}_h^k = \{M_{h,y}^k = \sum_{x\in\mathcal{X}:h(x)=y} G_x \quad : \quad y \in \mathcal{Y}\} \tag{17}$$

and $\mathscr{M}_k \triangleq \{\mathcal{M}_h^k : h \in \mathcal{H}_k\}$.

It is straightforward to verify $\mathcal{M}_h^k \in \mathscr{M}_{X\to Y}$ and hence $\mathscr{M}_k \subseteq \mathscr{M}_{X\to Y}$ is a POVM concept class. Indeed $\sum_{y\in\mathcal{Y}} M_{h,y}^k = \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}:h(x)=y} G_x = \sum_{x\in\mathcal{X}} G_x = I_{\mathcal{H}_X}$, the identity on $\mathcal{H}_X$. We now assert that $\mathcal{M}_h^k : k \geq 1$ is a sequence of POVM concept classes that are universally consistent.

Theorem 2. *The sequence $\mathscr{M}_k : k \geq 1$ of POVM concept classes is universally consistent.*

6 CONCLUSION, FUTURE WORK

We studied the problem of learning from quantum data, entailing a graduation from function classes to measurement concept classes. Next, the challenges posed by quantum effects forced us to modify the foundational algorithm (ERM) of statistical learning theory and unravel its sample complexity. We conclude by identifying the related combinatorial optimization problem of identifying an optimal stochastic compatible partition of the concept class as a problem of independent interest.

References

[Aaronson, 2006] Aaronson, S. (2006). The learnability of quantum states. *Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences*, 463.

[Aaronson, 2018] Aaronson, S. (2018). Shadow tomography of quantum states. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 325–338, New York, NY, USA. Association for Computing Machinery.

[Aaronson et al., 2018] Aaronson, S., Chen, X., Hazan, E., and Kale, S. (2018). Online learning of quantum states. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8976–8986, Red Hook, NY, USA. Curran Associates Inc.

[Altepeter et al., 2004] Altepeter, J., James, D., and Kwiat, P. (2004). Qubit quantum state tomography. *Lecture Notes in Physics*, 649:113–145.

[Anguita et al., 2003] Anguita, D., Ridella, S., Rivieccio, F., and Zunino, R. (2003). Quantum optimization for training support vector machines. *Neural Networks*, 16(5):763–770. Advances in Neural Networks Research: IJCNN '03.

[Anshu et al., 2021] Anshu, A., Arunachalam, S., Kuwahara, T., and Soleimanifar, M. (2021). Sample-efficient learning of interacting quantum systems. *Nature Physics*, pages 1745–2481.

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

[Arunachalam et al., 2020] Arunachalam, S., Grilo, A. B., and Yuen, H. (2020). Quantum statistical query learning.

[Atici and Servedio, 2005] Atici, A. and Servedio, R. A. (2005). Improved bounds on quantum learning algorithms. *Quantum Information Processing*, 4(5):355–386.

[Aïmeur et al., 2012] Aïmeur, E., Brassard, G., and Gambs, S. (2012). Quantum speed-up for unsupervised learning. *Machine Learning*, 90(2):261–287.

[Bae and Kwek, 2015] Bae, J. and Kwek, L.-C. (2015). Quantum state discrimination and its applications. *Journal of Physics A: Mathematical and Theoretical*, 48(8):083001.

[Bubeck et al., 2020] Bubeck, S., Chen, S., and Li, J. (2020). Entanglement is necessary for optimal quantum property testing.

[Cheng et al., 2016] Cheng, H.-C., Hsieh, M.-H., and Yeh, P.-C. (2016). The learnability of unknown quantum measurements. *Quantum Info. Comput.*, 16(7–8):615–656.

[Devroye et al., 1996] Devroye, L., Gy orfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer.

[Gortler and Servedio, 2001] Gortler, S. J. and Servedio, R. A. (2001). Quantum versus classical learnability. In *Proceedings of Computational Complexity. Sixteenth Annual IEEE Conference*, page 0138, Los Alamitos, CA, USA. IEEE Computer Society.

[Guerini and Terra Cunha, 2018] Guerini, L. and Terra Cunha, M. (2018). Uniqueness of the joint measurement and the structure of the set of compatible quantum measurements. *Journal of Mathematical Physics*, 59(4):042106.

[Haah et al., 2016] Haah, J., Harrow, A. W., Ji, Z., Wu, X., and Yu, N. (2016). Sample-optimal tomography of quantum states. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 913–925, New York, NY, USA. Association for Computing Machinery.

[Heidari et al., 2021] Heidari, M., Padakandla, A., and Szpankowski, W. (2021). A theoretical framework for learning from quantum data. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1469–1474.

[Holevo, 2019] Holevo, A. S. (2019). *Quantum Systems, Channels, Information.* De Gruyter, 2 edition.

[Jae et al., 2019] Jae, J., Baek, K., Ryu, J., and Lee, J. (2019). Necessary and sufficient condition for joint measurability. *Phys. Rev. A*, 100:032113.

[Karthik et al., 2015] Karthik, H. S., Devi, A. R. U., and Rajagopal, A. K. (2015). Unsharp measurements and joint measurability. *Current Science*, 109(11):2061–2068.

[Kearns and Schapire, 1990] Kearns, M. J. and Schapire, R. E. (1990). Efficient distribution-free learning of probabilistic concepts. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, SFCS '90, page 382–391 vol.1, USA. IEEE Computer Society.

[Khan and Robles-Kelly, 2020] Khan, T. M. and Robles-Kelly, A. (2020). Machine learning: Quantum vs classical. *IEEE Access*, 8:219275–219294.

[Kunjwal, 2014] Kunjwal, R. (2014). A note on the joint measurability of povms and its implications for contextuality.

[Lloyd et al., 2014] Lloyd, S., Mohseni, M., and Rebentrost, P. (2014). Quantum principal component analysis. *Nature Physics*, 10(9):631–633.

[Montanaro and de Wolf, 2018] Montanaro, A. and de Wolf, R. (2018). A survey of quantum property testing.

[O'Donnell and Wright, 2016] O'Donnell, R. and Wright, J. (2016). Efficient quantum tomography. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 899–912, New York, NY, USA. Association for Computing Machinery.

[O'Donnell and Wright, 2017] O'Donnell, R. and Wright, J. (2017). Efficient quantum tomography ii. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 962–974, New York, NY, USA. Association for Computing Machinery.

[Rehacek and Paris, 2004] Rehacek, J. and Paris, M. (2004). *Quantum state estimation.* Lecture notes in physics. Springer, Berlin.

[Schuld et al., 2014] Schuld, M., Sinayskiy, I., and Petruccione, F. (2014). Quantum computing for pattern classification. In Pham, D.-N. and Park, S.-B., editors, *PRICAI 2014: Trends in Artificial Intelligence*, pages 208–220, Cham. Springer International Publishing.

[Servedio, 2001] Servedio, R. A. (2001). Separating quantum and classical learning. In *Proceedings of the 28th International Colloquium on Automata, Languages and Programming,*, ICALP '01, page 1065–1080, Berlin, Heidelberg. Springer-Verlag.

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

[Stephane Boucheron and Massart, 2013] Stephane Boucheron, G. L. and Massart, P. (2013). *Concentration Inequalities.* Oxford University Press, Oxford, United Kingdom.

[Valiant, 1984] Valiant, L. G. (1984). A theory of the learnable. *Commun. ACM*, 27(11):1134–1142.

[Vapnik, 1982] Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics).* Springer-Verlag, Berlin, Heidelberg.

[Vapnik and Chervonenkis, 1974] Vapnik, V. and Chervonenkis, A. (1974). *Theory of Pattern Recognition [in Russian].* Nauka, Moscow. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979).

[Vapnik, 1995] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag, Berlin, Heidelberg.

[Vapnik, 1998] Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley-Interscience.

[Vapnik and Chervonenkis, 1971] Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

[Waseem et al., 2020] Waseem, M. H., e Ilahi, F., and Anwar, M. S. (2020). Quantum state tomography. In *Quantum Mechanics in the Single Photon Laboratory*, 2053-2563, pages 6–1 to 6–20. IOP Publishing.

[Wiebe et al., 2012] Wiebe, N., Braun, D., and Lloyd, S. (2012). Quantum algorithm for data fitting. *Phys. Rev. Lett.*, 109:050505.

[Wilde, 2017] Wilde, M. (2017). *Quantum information theory.* Cambridge University Press, Cambridge, UK.

[Wootters and Zurek, 1982] Wootters, W. K. and Zurek, W. H. (1982). A single quantum cannot be cloned. *Nature*, 299(5886):802–803.

Supplementary Material: PAC Learning of Quantum Measurement Classes: Sample Complexity Bounds and Universal Consistency

A A COMPLETE PROOF OF THEOREM 1

In addition to deriving upper bounds on T_1 and T_2, we fill in missing steps and thereby provide a complete proof of Thm. 1. We refer the reader to Sec. 4.2 to recall the notation and tools introduced therein. The reader may note that we had stated in Sec. 4.2 the sufficiency of the uniform convergence property [Shalev-Shwartz and Ben-David, 2014, Definition 4.3]. Here, we provide a proof of this sufficiency tailored to the current scenario involving compatible partitions.

A.1 Uniform Convergence Property Suffices

As stated in Sec. 4.2, let $\mathscr{B}_1, \cdots, \mathscr{B}_{|I|}$ be a compatible partition of $\mathcal{C}$. Suppose $\mathscr{B}_j = \{\mathcal{M}_{j1}, \cdots, \mathcal{M}_{jK}\}$ is compatible with fine graining $(\mathcal{W}_j, \mathcal{G}_j = \mathcal{G}_{\mathcal{W}_j}, \alpha_k^j = \alpha_{\mathcal{Y}|\mathcal{W}_j}^{j,k})$ where $\mathcal{G}_j = \{\mathcal{G}_{jw} \in \mathcal{P}(\mathcal{H}) : w \in \mathcal{W}_j\} \in \mathscr{M}_{X \to \mathcal{W}_j}$.[8] Suppose n_j of the n training samples are used to evaluate the ER of POVMs in $\mathscr{B}_j$. Let $\mathcal{Z} = \mathcal{Y}$. The ERM-Q algorithm generates RVs $Z_{ki}^j \in \mathcal{Z} : 1 \leq k \leq K, \leq i \leq n_j$ with distribution

$$\left(\sum_{w \in \mathcal{W}_j} \operatorname{tr}(\mathcal{G}_{jw}\rho_{x_i})\alpha_k^j(z|w) : z \in \mathcal{Z} = \mathcal{Y} \right).$$

For $j = 1, \cdots, |I|$, ERM-Q identifies

$$k_j^* \triangleq \underset{1 \leq k \leq K}{\arg\min} \frac{1}{n_j} \sum_{i=1}^{n_j} l(y_i, Z_{ki}^j)$$

the best POVM in compatible partition element $\mathscr{B}_j$ and declares the index of $\arg\min_{1 \leq j \leq |I|} k_j^*$ as the chosen predictor. Suppose, for every $j = 1, \cdots, |I|$ and every $1 \leq k \leq K$, we have

$$\Bigg| \frac{1}{n_j} \sum_{i=1}^{n_j} l(y_i, Z_{ki}^j) \tag{18}$$

$$- \sum_{(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}} p_{XY}(x, y) \tag{19}$$

$$\cdot \sum_{w \in \mathcal{W}_j} \operatorname{tr}(\mathcal{G}_{jw}\rho_x)\alpha_k^j(z|w)l(y, z) \Bigg| \tag{20}$$

$$\leq \frac{\epsilon}{4} \tag{21}$$

where the the second term within the modulus above is $l_p(\mathcal{M}_{j,k})$, whenever $(x_i, y_i) : 1 \leq i \leq n_j$ are drawn according to p_{XY}, then by the definition of the infimum, we can choose a POVM whose true risk is within $\frac{\epsilon}{4}$ of the infimum and thereby guarantee that the true risk of the index $\arg\min_{1 \leq j \leq |I|} k_j^*$ is indeed within $\frac{\epsilon}{2}$ of the true risk of $l^*(p_{XY}, \mathcal{C})$. We have thus argued that proving that each compatible partition $\mathscr{B}_1, \cdots, \mathscr{B}_{|I|}$ possessing the uniform convergence property is a sufficient property for proving PAC learnability of the ERM-Q algorithm. In the following two sections we derive bounds on T_1 and T_2 defined in (13).

A.2 An upper bound on the first Term T_1 in (13)

Recall

$$T_1 \triangleq \tag{22}$$

$$\sum_{\underline{x}, \underline{y}, \underline{\underline{z}}} p_{XY}^n(\underline{x}, \underline{y}) \beta_{[K]}^n(\underline{\underline{z}}|\underline{x}) \mathbb{1}_{\{\phi(\underline{x}, \underline{y}, \underline{\underline{z}}) - \mathbb{E}\{\phi(\underline{X}, \underline{Y}, \underline{\underline{Z}})\} > \frac{\epsilon}{2}\}} \cdot \tag{23}$$

Henceforth, we include a subscript, focus on the one-sided term and let

$$\phi_{p\beta}^+(\underline{x}, \underline{y}, \underline{\underline{z}}) \tag{24}$$

$$\triangleq \sup_{1 \leq k \leq K} \sum_{i=1}^{n} \frac{l(y_i, z_{ki})}{n} - \sum_{\substack{(a,b,c) \in \\ \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}} l(b, c) p_{XY}(a, b) \beta_k(c|a). \tag{25}$$

[8] We could just choose $K = \max_{1 \leq j \leq |I|} |\mathscr{B}_j|$ and populate the smaller compatible partitions with trivial (Identity) POVMs.

With this, the first term T_1 is

$$T_1 \tag{26}$$

$$= P \circ \beta\left(\phi^{+}_{p\beta}(\underline{X},\underline{Y},\underline{\underline{Z}}) - \mathbb{E}\{\phi^{+}_{p\beta}(\underline{X},\underline{Y},\underline{\underline{Z}})\} > \frac{\epsilon}{4}\right). \tag{27}$$

In order to upper bound T_1, we use the Bounded Difference Inequality stated in Theorem 3. Choose $\mathcal{B} = \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}^K$ and $B_i = (X_i, Y_i, \underline{Z}_i) = (X_i, Y_i, Z_{1i}, \cdots, Z_{Ki})$ and $f(\underline{x}, \underline{y}, \underline{\underline{z}}) = \phi_{p\beta}(\underline{x}, \underline{y}, \underline{\underline{z}})$. Let $\underline{x}, \hat{\underline{x}} \in \mathcal{X}^n$, $\underline{y}, \hat{\underline{y}} \in \mathcal{Y}^n$ and $\underline{\underline{z}}, \hat{\underline{\underline{z}}} \in \mathcal{Z}^{Kn}$, be such that $x_t = \hat{x}_t$, $y_t = \hat{y}_t$ and $\underline{z}_t = \hat{\underline{z}}_t$ for $t \in [n] \setminus i$. For such a choice, we have

$$\phi^{+}_{p\beta}(\underline{x}, \underline{y}, \underline{\underline{z}}) - \phi^{+}_{p\beta}(\hat{\underline{x}}, \hat{\underline{y}}, \hat{\underline{\underline{z}}}) \tag{28}$$

$$= \sup_{1 \le k \le K} \sum_{i=1}^{n} \frac{l(y_i, z_{ki})}{n} - \sum_{\substack{(a,b,c) \in \\ \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}} l(b,c) p_{XY}(a,b) \beta_k(c|a)$$

$$- \sup_{1 \le k \le K} \sum_{i=1}^{n} \frac{l(\hat{y}_i, \hat{z}_{ki})}{n} - \sum_{\substack{(a,b,c) \in \\ \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}} l(b,c) p_{XY}(a,b) \beta_k(c|a)$$

$$\le \sup_{1 \le k \le K} \tag{29}$$

$$\left\{ \sum_{i=1}^{n} \frac{l(y_i, z_{ki})}{n} - \sum_{\substack{(a,b,c) \in \\ \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}} l(b,c) p_{XY}(a,b) \beta_k(c|a) \right.$$

$$\left. - \sum_{i=1}^{n} \frac{l(\hat{y}_i, \hat{z}_{ki})}{n} - \sum_{\substack{(a,b,c) \in \\ \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}} l(b,c) p_{XY}(a,b) \beta_k(c|a) \right\} \tag{30}$$

$$= \sup_{1 \le k \le K} \left\{ \sum_{i=1}^{n} \frac{l(y_i, z_{ki})}{n} - \sum_{i=1}^{n} \frac{l(\hat{y}_i, \hat{z}_{ki})}{n} \right\} \le \frac{B-A}{n} \tag{31}$$

where (30) follows from $\sup_a f_1(a) - \sup_a f_2(a) \le \sup_a (f_1(a) - f_2(a))$ and (31) follows from the bounds on the loss function $l(\cdot,\cdot)$. Similarly, we can prove $\phi^{+}_{p\beta}(\hat{\underline{x}}, \hat{\underline{y}}, \hat{\underline{\underline{z}}}) - \phi^{+}_{p\beta}(\underline{x}, \underline{y}, \underline{\underline{z}}) \le \frac{B-A}{n}$. This proves that $\phi^{+}_{p\beta}(\underline{x}, \underline{y}, \underline{\underline{z}})$ possesses the bounded difference property and we have $P \circ \beta(\phi^{+}_{p\beta}(\underline{X},\underline{Y},\underline{\underline{Z}}) - \mathbb{E}\{\phi^{+}_{p\beta}(\underline{X},\underline{Y},\underline{\underline{Z}})\} \ge \frac{\epsilon}{2}) \le \exp\left\{-\frac{n\epsilon^2}{2(B-A)^2}\right\}$. Analogously defining

$$\phi^{-}_{p\beta}(\underline{x}, \underline{y}, \underline{\underline{z}}) \tag{32}$$

$$\triangleq \sup_{1 \le k \le K} \sum_{\substack{(a,b,c) \in \\ \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}} l(b,c) p_{XY}(a,b) \beta_k(c|a) - \sum_{i=1}^{n} \frac{l(y_i, z_{ki})}{n}$$

and following a similar sequence of steps, one can prove $P \circ \beta(\phi^{-}_{p\beta}(\underline{X},\underline{Y},\underline{\underline{Z}}) - \mathbb{E}\{\phi^{-}_{p\beta}(\underline{X},\underline{Y},\underline{\underline{Z}}) \ge \frac{\epsilon}{2}\}) \le \exp\left\{-\frac{n\epsilon^2}{2(B-A)^2}\right\}$ and we can conclude $P \circ \beta(\phi_{p\beta}(\underline{X},\underline{Y},\underline{\underline{Z}}) - \mathbb{E}\{\phi_{p\beta}(\underline{X},\underline{Y},\underline{\underline{Z}})\} \ge \frac{\epsilon}{2}) \le 2\exp\left\{-\frac{n\epsilon^2}{4(B-A)^2}\right\}$.

A.3 Bounding the Second Term T_2 in (13) via Rademacher Complexity

We recall $T_2 \triangleq \mathbb{E}\{\phi(\underline{X},\underline{Y},\underline{\underline{Z}})\}$, where

$$\phi(x_1, \cdots, x_n, y_1, \cdots, y_n, \underline{z}_1, \cdot\cdot\cdot, \underline{z}_n) \triangleq \tag{33}$$

$$\sup_{1 \le k \le K} \tag{34}$$

$$\left| \sum_{i=1}^{n} \frac{l(y_i, z_{ki})}{n} - \sum_{\substack{(a,b,c) \in \\ \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}} l(b,c) p_{XY}(a,b) \beta_k(c|a) \right|. \tag{35}$$

In bounding $\mathbb{E}\{\phi(\underline{X},\underline{Y},\underline{\underline{Z}})\}$, we employ the ghost sample+symmetrization trick [Devroye et al., 1996, Thm. 12.4]. Observe that

$$\mathbb{E}\{\phi(\underline{X},\underline{Y},\underline{\underline{Z}})\} \tag{36}$$

$$= \sum_{\underline{x},\, \underline{y},\, \underline{\underline{z}}} p^n_{XY}(\underline{x}, \underline{y}) \beta^n_{[K]}(\underline{\underline{z}}|\underline{x}) \tag{37}$$

$$\sup_{k \in [K]} \left| \sum_{i=1}^{n} \frac{l(y_i, z_{ki})}{n} - \sum_{\substack{(a,b,c) \in \\ \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}} l(b,c) p_{XY}(a,b) \beta_k(c|a) \right|$$

$$= \sum_{\underline{x},\, \underline{y},\, \underline{\underline{z}}} p^n_{XY}(\underline{x}, \underline{y}) \beta^n_{[K]}(\underline{\underline{z}}|\underline{x}) \tag{38}$$

$$\sup_{k \in [K]} \Big| \sum_{i=1}^{n} \frac{l(y_i, z_{ki})}{n} \tag{39}$$

$$- \sum_{\underline{a},\, \underline{b},\, \underline{\underline{c}}} p^n_{XY}(\underline{a}, \underline{b}) \beta^n_{[K]}(\underline{\underline{c}}|\underline{a}) \left[\sum_{j=1}^{n} \frac{l(b_j, c_{kj})}{n} \right] \Big| \tag{40}$$

$$\le \sum_{\underline{x},\, \underline{y},\, \underline{\underline{z}}} \sum_{\underline{a},\, \underline{b},\, \underline{\underline{c}}} p^n_{XY}(\underline{a}, \underline{b}) \beta^n_{[K]}(\underline{\underline{c}}|\underline{a}) p^n_{XY}(\underline{x}, \underline{y}) \beta^n_{[K]}(\underline{\underline{z}}|\underline{x}) \tag{41}$$

$$\sup_{k \in [K]} \Big| \sum_{i=1}^{n} \frac{l(y_i, z_{ki})}{n} - \sum_{j=1}^{n} \frac{l(b_j, c_{kj})}{n} \Big| \tag{42}$$

$$= \sum_{\substack{(\sigma_1, \cdots \sigma_n) \\ \in \{-1,+1\}^n}} \sum_{\substack{\underline{x},\, \underline{y},\, \underline{\underline{z}} \\ \underline{a},\, \underline{b},\, \underline{\underline{c}}}} \tag{43}$$

$$\frac{p^n_{XY}(\underline{a}, \underline{b}) \beta^n_{[K]}(\underline{\underline{c}}|\underline{a}) p^n_{XY}(\underline{x}, \underline{y}) \beta^n_{[K]}(\underline{\underline{z}}|\underline{x})}{2^n} \tag{44}$$

$$\sup_{k \in [K]} \Big| \sum_{i=1}^{n} \frac{\sigma_i l(y_i, z_{ki})}{n} - \sum_{j=1}^{n} \frac{\sigma_j l(b_j, c_{kj})}{n} \Big|$$

$$\le 2\mathfrak{R}_n(\mathscr{B}). \tag{45}$$

where

$$\mathfrak{R}_n(\mathscr{B}) \triangleq \tag{46}$$

$$\sum_{\substack{\sigma\\ \in\{\pm 1\}^n}} \sum_{\underline{x},\ \underline{y},\ \underline{\underline{z}}} \frac{p_{XY}^n(\underline{x},\underline{y})\beta_{[K]}^n(\underline{\underline{z}}|\underline{x})}{2^n}\cdot \tag{47}$$

$$\left[\sup_{k\in[K]} \frac{1}{n}\sum_{i=1}^{n} \sigma_i l(y_i, z_{ki})\right], \tag{48}$$

where (38) follows by introducing the ghost sample to the second term [Devroye et al., 1996, Step 1 in Proof of Thm. 12.7], (41) follows from the chain $\sup|\mathbb{E}\{\cdot\}| \leq \sup\mathbb{E}\{|\cdot|\} \leq \mathbb{E}\{\sup|\cdot|\}$ of inequalities, the next step follows by symmetrization by random signs [Devroye et al., 1996, Step 2 in Proof of Thm. 12.4], finally resulting in the familiar average Radamacher complexity. While these steps are fairly standard, we notice the crucial difference (45), where the averaging includes the POVM randomness $\beta_{[K]}(\cdot|\cdot)$.

A.4 Bounding $\mathfrak{R}_n(\mathscr{B})$ using Massart's Lemma

Once again, we follow a well established sequence of steps in statistical learning theory to bound $\mathfrak{R}_n(\mathscr{B})$. Recognizing $l(y,z) \leq B$, implying $||(l(y_i,z_i) : i \in [n])||_2 \leq B\sqrt{n}$ and the cardinality of the set $\{(l(y_i,z_{ki}) : i \in [n]) : k \in [K]\}$ is at most $|\theta_n(\mathscr{B},(x_1,\cdots,x_n))|$, we have

$$\mathfrak{R}_n(\mathscr{B}) \tag{49}$$

$$= \mathbb{E}_{\underline{X}\ \underline{Y}\ \underline{\underline{Z}}}\left\{\mathbb{E}_{\sigma}\left[\sup_{k\in[K]}\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(Y_i, Z_{ki})\right]\right\} \tag{50}$$

$$\leq \mathbb{E}_{\underline{X}\ \underline{Y}\ \underline{\underline{Z}}}\left\{\frac{B\sqrt{2n\log|\theta_n(\mathscr{B},\underline{X})|}}{n}\right\} \tag{51}$$

$$\overset{(i)}{\leq} \frac{B\sqrt{2n\mathbb{E}\{\log|\theta_n(\mathscr{B},\underline{X})|\}}}{n} \overset{(ii)}{\leq} \sqrt{\frac{2B^2\log\mathcal{S}_r(\mathscr{B})}{n}} \tag{52}$$

where the inequality in (49) follows from Massart's lemma provided in Theorem 4, (52(i)) follows from Jensen's inequality applied to the concave $\sqrt{\cdot}-$function, (52(ii)).

A.5 Collating Bounds

We now collate bounds from the first term and (52) and substitute in (13). Recollecting our compatible partition $\mathscr{B} = \{\mathcal{M}_k : 1 \leq k \leq K\}$, definitions (10) through (33), our uniform convergence bound for $\mathscr{B}$ is

$$\sup_{p_{XY}} \sum_{\underline{x}\in\mathcal{X}^n} \sum_{\underline{y}\in\mathcal{Y}^n} \sum_{\underline{z}_1\in\mathcal{Z}^K} \tag{53}$$

$$\cdots \sum_{\underline{z}_n\in\mathcal{Z}^K} p_{XY}^n(\underline{x},\underline{y})\beta_{[K]}^n(\underline{\underline{z}}|\underline{x})\mathbb{1}_{\{\phi(\underline{x},\underline{y},\underline{\underline{z}})>\eta(\epsilon,\delta)\}} \tag{54}$$

$$\leq \delta \tag{55}$$

where

$$\eta(\epsilon,\delta) \triangleq \sqrt{\frac{2(B-A)\log\left(\frac{1}{\delta}\right) + 8B^2\log\mathcal{S}_r(\mathscr{B})}{n}}. \tag{56}$$

A.6 Union Bound on the Compatible Partition

Having derived uniform convergence for one compatible subset $\mathscr{B}$ of POVMs, we employ a union bound for the compatible partition $\mathscr{B}_1,\cdots,\mathscr{B}_{|I|}$. Squeezing δ to $\frac{\delta}{|I|}$, we obtain the sample complexity stated in the theorem statement.

B PROOF OF THEOREM 2 : SKETCH AND OUTLINE

We first the idea of the proof before fleshing out the details. Let us first understand the case when $\rho_x = |x\rangle$ for $x \in \mathcal{X}$ with $\langle x|\hat{x}\rangle = \delta_{x\hat{x}}$. In essence, all the quantum states are distinguishable, reducing this to the classical problem. For a given $p_{XY} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, let $R(x \to y) \triangleq \sum_{\hat{y}\in\mathcal{Y}} p_{Y|X}(\hat{y}|x)l(y,\hat{y})$ denote the risk, i.e., the conditional expectation of the loss, of assigning label y to x. It can be shown that the Bayes' rule for this setting is given by the function $f^* : \mathcal{X} \to \mathcal{Y}$ defined as $f^*(x) = \arg\min_{y\in\mathcal{Y}} R(x \to y)$. Essentially, f^* is choosing the label that minimizes the expected loss for every x. This implies that $l_p^* = l_p(f^*)$. It is straightforward to recognize that, in this simplified case the sequence $\mathscr{M}_k : k \geq 1$ defined in (17) is indeed the sequence $\mathcal{H}_k : k \geq 1$ as defined in (15). Specifically, $\mathscr{M}_k = \mathcal{H}_k$ for $k \in \mathbb{N}$. The proof that $\mathscr{M}_k : k \geq 1$ forms a universally consistent sequence of POVM concept classes is now a direct consequence of [Stephane Boucheron and Massart, 2013, Thm. 6.1]. We alert the reader here that our claim of universal consistency implies that for *every* distribution $p_{XY} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, we have $\lim_{k\to\infty} l_p(\mathscr{M}_k) = l_p^*$.

What if the quantum states are not distinguishable, i.e., the operators $\rho_x : x \in \mathcal{X}$ have overlapping support spaces? We leverage the approach stated in the above to provide a sketch of the arguments. In this submission, we provide only a sketch of the arguments and follow it up with all the details fleshed out in a subsequent version available publicly. Having said that, we should mention that the sketch provided here has sufficient details for a reader informed in SLT and quantum theory. For a distribution p_{XY} and a collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$, we let $\mathcal{M}^p = \{M^p_y : y \in \mathcal{Y}\} \in \mathscr{M}_{X \to Y}$ denote a $\mathcal{Y}-$POVM for which

$$\sum_{\hat{y} \in \mathcal{Y}} \operatorname{tr}\big(M^p_y \rho_x\big) p_{Y|X}(\hat{y}|x) l(y, \hat{y}) \tag{57}$$

$$\leq \sum_{\hat{y} \in \mathcal{Y}} \operatorname{tr}\Big(M^p_{\tilde{y}} \rho_x\Big) p_{Y|X}(\hat{y}|x) l(\tilde{y}, \hat{y}) \tag{58}$$

for every

$$(x, y, \hat{y}) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}. \tag{59}$$

The proof is built on two facts that can be proved based on the arguments presented in proof of [Stephane Boucheron and Massart, 2013, Thm. 6.1] and the properties of a POVM. Firstly, (57) is a necessary and sufficient condition for a POVM $\mathcal{M}^p = \{M^p_y : y \in \mathcal{Y}\}$ to achieve the Bayes' loss for the distribution p_{XY} and a collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$. Secondly, given any $p_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and any collection $(\rho_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X})$ of density operators, we can identify a sequence $\mathcal{G}_k \in \mathscr{M}_k : k \geq 1$ that satisfy the above the condition in (57) in the limit. This is the approach we take to prove that the sequence $\mathscr{M}_k : k \geq 1$ of POVM concept classes is universally consistent.

C TOOLS AND PREREQUISITES

C.1 Bounded Difference Inequality

The following can be found in [Stephane Boucheron and Massart, 2013, Section 6.1].

Defn 8. *A function $f : \mathcal{B}^n \rightarrow \mathbb{R}$ has bounded difference property if for some non-negative constants $c_1, \cdots, c_n$, we have*

$$\sup_{b_1, \cdots, b_n, \hat{b}_i} |f(b_1, \cdots, b_n) - f(b_1, \cdots, b_{i-1}, \hat{b}_i, b_{i+1}, \cdots, b_n)| \tag{60}$$

$$\leq c_i \tag{61}$$

for $1 \leq i \leq n$.

Theorem 3. *Suppose $f : \mathcal{B}^n \rightarrow \mathbb{R}$ possesses the bounded difference property with constants $c_1, \cdots, c_n$ and let $v = \frac{1}{4}(c_1^2 + \cdots + c_n^2)$. Suppose $B_1, \cdots, B_n$ are independent and $A = f(B_1, \cdots, B_n)$, then*

$$P(A - \mathbb{E}\{A\} > t) \leq \exp\left\{-\frac{t^2}{2v}\right\}, \tag{62}$$

$$P(A - \mathbb{E}\{A\} < -t) \leq \exp\left\{-\frac{t^2}{2v}\right\}$$

C.2 Massart's Lemma

The facts presented below and a proof of the same can be found in [Shalev-Shwartz and Ben-David, 2014, Chapter 26].

Theorem 4. *Let $A = \{\underline{a}_1, \cdots, \underline{a}_r\} \subseteq \mathbb{R}^m$ with $\underline{a}_i = (a_{1i}, \cdots, a_{mi})$. Define $\underline{b} = \frac{1}{r}\sum_{i=1}^r \underline{a}_i$ and let $B_1, \cdots, B_m \in \{-1.+1\}$ be independent and uniformly distributed. Then*

$$\mathbb{E}_{B_1, \cdots, B_m}\left\{\sup_{1 \leq i \leq r} \frac{1}{m} \sum_{j=1}^m B_j a_{ji}\right\} \tag{63}$$

$$\leq \max_{1 \leq i \leq r} ||\underline{b} - \underline{a}_i||_2 \frac{\sqrt{2\log(r)}}{m}. \tag{64}$$