Fat Shattering, Joint Measurability, and PAC Learnability of POVM Hypothesis Classes

Abram Magner[1*] and Arun Padakandla[2]

[1*]Department of Computer Science, University at Albany, State University of New York, 1400 Washington Ave, Albany, 12222, NY, United States of America.

[2]Eurecom, Nice, France.

*Corresponding author(s). E-mail(s): amagner@albany.edu;
Contributing authors: padakand@eurecom.fr;

Keywords: learning theory, POVM, PAC learning, joint measurability

We characterize learnability for quantum measurement classes by establishing matching necessary and sufficient conditions for their probably approximately correct (PAC) learnability, along with corresponding sample complexity bounds, in the setting where the learner is given access only to prepared quantum states. We first show that the empirical risk minimization (ERM) rule proposed in previous work is not universal, nor does uniform convergence of the empirical risk characterize learnability. Moreover, we show that VC dimension generalization bounds in previous work are in many cases infinite, even for measurement classes defined on a finite-dimensional Hilbert space, and even for learnable classes. To surmount the failure of the standard ERM to satisfy uniform convergence, we define a new learning rule – *denoised empirical risk minimization*. We show this to be a universal learning rule for both classical probabilistically observed concept classes and quantum measurement classes, and the condition for it to satisfy uniform convergence is finite fat shattering dimension of the class. The fat shattering dimension of a hypothesis class is a measure of complexity that intervenes in sample complexity bounds for regression in classical learning theory. We give sample complexity upper and lower bounds for learnability in terms of finite fat-shattering dimension and approximate finite partitionability into approximately jointly measurable subsets. We link fat shattering dimension with partitionability into approximately

jointly measurable subsets, leading to our matching conditions. We also show that every measurement class defined on a finite-dimensional Hilbert space is PAC learnable. We illustrate our results on several example POVM classes.

1 Introduction

Classical statistical learning theory formulates the broad problem of learning a relationship between two random quantities $X \in \mathcal{X}$ – known as features – and $Y \in \mathcal{Y}$ – class labels – as follows: the data are assumed to be generated from some *unknown* probability distribution $P_{X,Y}$, and a *learner* is given access to a dataset consisting of m independent and identically distributed samples (X_i, Y_i). The learner's task is to select a *hypothesis* from a fixed, known set Hyp of functions from $\mathcal{X}$ to $\mathcal{Y}$ (the *hypothesis/concept class*) that best approximates the joint distribution $P_{X,Y}$, *without knowledge of* $P_{X,Y}$ *itself.*

This work:

The intent of this work is to provide matching necessary and sufficient conditions for learnability in the following supervised quantum learning[1] scenario: there is an unknown joint probability distribution on prepared quantum states and classical labels. A hypothesis class consisting of quantum measurements is fixed and known to a learner. The learner is given access to a training dataset of these state-label pairs, but can only interact with the states by observing the classical outcomes of measuring them. It then outputs a hypothesis that is as close as possible to minimizing the expected *risk* over all hypotheses. This learning scenario was first posed in [1] as a quantum version of the classical PAC (Probably Approximately Correct) learning setting (see Appendix C) in which hypotheses are quantum measurements. This setting has extensive motivations ranging from building universal quantum state discriminators to classification of unknown quantum processes to classifying quantum phases of multipartite systems (see also [2, 3]). We discuss the classification of quantum many-body systems further in Appendix D. The setting was then further developed in [4]. In contrast with more well-established quantum learning frameworks [5], which deal with quantum algorithms for learning classical hypotheses (e.g., boolean functions $f : \{0,1\}^n \to \{0,1\}$) from superpositions of states corresponding to classical bit strings, our framework covers a distinct scenario in which input data consists of unknown quantum states, and the goal is to learn a measurement that predicts attributes (e.g., a class label) of those states.

More specifically, the authors of [1] formulated the quantum PAC learning framework that we study as follows: we fix a domain $\mathcal{X}$ consisting of quantum states, along with a codomain $\mathcal{Y}$. Analytically, quantum states are described by density matrices on a fixed Hilbert space $\mathcal{H}$ over the complex numbers $\mathbb{C}$. We take the codomain $\mathcal{Y} = \{0, 1\}$ for binary classification, but our results can be generalized further. A POVM hypothesis class Hyp is a set of *positive operator-valued measures* [6][2], which specify quantum measurements with outcomes in $\mathcal{Y}$. Additionally, we fix a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty]$. For binary classification, we take the misclassification loss

[1] For more background on classical statistical learning theory, see Appendix C.
[2] See the definitions from quantum mechanics, collected in the supplementary material.

$\ell(y_1, y_2) = I[y_1 \neq y_2]$, where $I[\cdot]$ is the indicator function. The learning process is as follows: an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ is fixed. To produce a single training example, $(X, Y) \sim \mathcal{D}$ is sampled, and then a quantum register is prepared in state X. **Here and throughout, a quantum register is a collection of qubits prepared in a state that is a density matrix in $\mathcal{H}$, which may be multidimensional**. The learner is given access to the quantum register and Y, and can only interact with the quantum register by measuring it and observing the outcome. This occurs independently m times to produce a training set of size m. The learner is then allowed to make arbitrarily many measurements of the given quantum registers and by an arbitrary procedure then producing a resulting POVM h from the class Hyp. We note that each measurement alters the state of the register according to the axioms of quantum mechanics (see Section B and specifically equation (B72)). The risk of a hypothesis is given by $R(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h[X], Y)]$, where $h[X]$ denotes a random variable whose distribution is that of the outcome of measuring a quantum register in state X with POVM h. Then the goal of the learner is to output a hypothesis with risk close enough to the minimal risk achieved by any hypothesis in the class. We define this setup formally in Definition 1 below.

The main problems of interest are similar to the ones asked in the classical PAC learning framework: perhaps the most immediate one is, what is a natural necessary and sufficient condition for PAC learnability of a POVM concept class? Is there a learning rule that is universal, in the sense that it is a PAC learning rule whenever the concept class is learnable? The present paper answers both of these questions. In the classical case with deterministic (function) concept classes, one of the fundamental results, which is sometimes called the *fundamental theorem of concept learning*, gives a necessary and sufficient condition for learnability of a concept class for binary classification under the misclassification loss: namely, learnability is equivalent to finiteness of the Vapnik-Chervonenkis (VC) dimension of the class [7]. The recent paper [4] gave one possible generalization of VC dimension to the quantum setting, resulting in a sufficient condition for learnability of POVM classes, along with a sample complexity[3] upper bound for one particular learning rule. However, it gave no necessary conditions and did not explore the tightness of the upper bound or the universality of the learning rule. The present paper finds that this sufficient condition is substantially weak and that the learning rule is very far from universal. We provide a new learning rule – *denoised empirical risk minimization (DERM)*, that we can show to be universal, along with matching necessary and sufficient conditions for learnability. See Section 1.2 for a fuller list of our contributions.

1.1 Prior work

The literature on statistical problems involving quantum states and measurements is quite broad. For example, a wealth of quantum state estimation problems have been posed [8–10], wherein the input is a sequence of multiple quantum registers, all prepared in a single unknown state. This set of works also includes works on

[3]The *sample complexity* of a learning rule is the minimum number of samples required to guarantee that with probability at least $1 - \delta$, the risk (i.e., expected loss) of the learned hypothesis is within ϵ of the minimum possible.

state tomography [11–19]. The task in such studies is to glean information about the single, unknown state – specifically, to *estimate* it. Estimation is *not* the same thing as learning, and so these are is in contrast with our work, in which the goal is more analogous to the classical supervised learning problem: i.e., our goal is to learn a statistical association between unknown quantum states sampled according to an unknown distribution and their classical labels. This statistical association need not reflect any intrinsic physical information about the states. We also point out that there are various works, such as [20, 21] that mix what is called PAC learning with quantum information, but these differ substantially from our setting: e.g., they assume a uniform distribution on the input, so they are not distribution-free; or they strongly constrain the input state to correspond to a bit string; or they output a boolean function instead of a POVM. There is also a large and expanding body of work in quantum machine learning in which hypothesis classes consist of specially structured POVMs – as recent examples, [22–26]. The focus in such works is different from that of the framework we study, since they aim to solve *classical* learning problems by suitably encoding classical input data as quantum states, then choosing a suitable measurement from the hypothesis class. In our case, the inputs $\mathcal{X}$ are intrinsically quantum and are not encodings of known classical inputs.

In the supplementary materials, we give a more extended discussion of prior works and how they differ from ours, including, in particular, how works on channel tomography are not applicable to solve our learning problem.

At first glance, the paper [27] has a more related goal to ours – producing an optimal POVM from training data. However, training samples consist of the density matrices encoding states, rather than quantum registers, as well as the probabilities of outcomes of measurements by an unknown POVM. In contrast, in the framework that we consider, the inputs to a learner are not analytical state descriptions; rather, they are quantum registers prepared in those states. Furthermore, we are given, not probabilities of outcomes, but the outcomes themselves. Finally, the statistical relationship between the state and the label in our case can be arbitrary, whereas in the cited paper, it is governed by a single unknown POVM.

Two recent papers are the most relevant to the present one and, indeed, are the sources of the framework that we study in this paper: [1, 4]. The paper [1] formulated the POVM class PAC learning framework, showed that finite-cardinality POVM classes are PAC learnable, and pointed out the usefulness of *joint measurability* in reducing sample complexity, resulting in the *Quantum Empirical Risk Minimization (QERM)* learning rule. The QERM rule is a generalization of the classical ERM, which is the cornerstone of classical statistical learning theory. Joint measurability is a property of a set S of quantum measurements that allows one to generate a sample from the distribution of outcomes of applying every measurement in S to a register prepared in an arbitrary state ρ. This is not possible for a generic set of measurements, since applying a measurement to ρ changes the state of the register. More formally, a set S of measurements is jointly measurable if there exists a single "root" measurement Π_{root} such that the outcome of measuring an arbitrary state ρ by any measurement Π in S can be simulated by first measuring ρ with Π_{root} and then post-processing the outcome with a classical channel α_{Π}. This allows for sample reuse across the set S.

The paper [4] studied the same setting, extending the sample complexity upper bounds for the QERM rule under the assumption that a partition of the hypothesis class into jointly measurable subsets is given, by formulating one possible generalization of the classical VC dimension of a probabilistically observed concept class. This implicitly showed that there exist PAC learnable POVM classes with infinite cardinality but left open the problem of giving necessary and sufficient conditions for a given class to be learnable. For example, no necessary conditions were given, in contrast with the present work. We will also show in this work that the upper bounds in that work are frequently vacuous.

In the course of proving our results, it will be convenient to define a PAC learning framework for what we call *probabilistically observed concept classes* (POCC), which we study as a technical tool for our quantum results. In this framework, each concept is a function from $\mathcal{X}$ to the set of probability distributions on $\mathcal{Y}$, and the task is, as usual to learn a risk-minimizing concept. However, on any sampled $x \in \mathcal{X}$ from the training set, for any concept h, the learner is only allowed to see a sample from the distribution $h(x)$. We will denote such samples by $h[x]$. This is in contrast with the theory of *probabilistic concepts* (p-concepts) introduced in [28]. There, concepts are similarly conditional distributions, but the learner is allowed to see the entire distribution $h(x)$.

A key concept in learning theory that arises in our necessary and sufficient conditions is the *fat shattering dimension* of a POVM class. This is our generalization of the classical learning-theoretic notion having the same name. Classically, the γ-fat shattering dimension is a natural number-valued measure of "complexity" of a hypothesis class Hyp consisting of either p-concepts or functions whose codomain is a continuous subset of the real numbers. It is a generalization of the VC dimension, which arises in binary classification, and it intervenes in a similar fashion to the VC dimension in generalization bounds (hence, sample complexity bounds) for empirical risk minimization for regression problems and p-concepts, as in [28]. Intuitively, for $\gamma \geq 0$, the γ-fat shattering dimension of a hypothesis class Hyp is the maximum cardinality m of any dataset S that is "shattered" by hypotheses in Hyp, in the sense that one can produce an arbitrary binary string with length m by first choosing *witness numbers* $r_1, ..., r_m \in [0, 1]$, then evaluating an appropriate hypothesis $h \in \text{Hyp}$ on S and rounding the outputs of h appropriately to 0 or 1. The rounding scheme depends on the parameter γ and the witness numbers: namely, if the output of h on the jth state is $p \geq r_j + \gamma$, then this rounds to 1; if the output is $p \leq r_j - \gamma$, then this rounds to 0. In full generality, this measure of complexity is not necessarily easy to evaluate for a given hypothesis class. However, by making certain geometric assumptions about the structure of the hypothesis classes that we consider, we are able to upper bound the fat shattering dimension of convex POVM classes to their geometry (specifically, their number of extreme points).

1.2 Our contributions

1. **Results on failure of ERM and uniform convergence:** We first show that the natural ERM learning rule proposed and studied in [1, 4] can fail for probabilistically observed concept classes that are PAC learnable. We probe this phenomenon further, showing that the empirical risk can fail to satisfy the uniform convergence

property for learnable hypothesis classes Hyp. That is, the supremal deviation of the empirical risk from expected value, where the supremum ranges over all elements of Hyp, does not converge to 0 as the number of samples tends to ∞. This implies that in the probabilistically observed and the quantum case, the QERM learning rule cannot be universal in the sense of being a PAC learning rule if and only if the class to which it is applied is learnable.

2. **Learnability of finite dimensional hypothesis classes:** We then show that every POVM class defined on a finite-dimensional Hilbert space is PAC learnable. This implies that the nontrivial *qualitative* question of learnability/non-learnability only occurs in the infinite-dimensional case. Furthermore, this implies that recovering classical learning theory from the POVM class framework requires mapping of classical classes to POVM classes over infinite-dimensional Hilbert spaces. This is an indication that infinite-dimensional Hilbert spaces are of fundamental interest for a complete quantum learning theory.
3. **Matching conditions for learnability of infinite-dimensional hypothesis classes:**

   We then turn to the problem of characterizing learning in the infinite-dimensional case. Motivated by our result on the failure of ERM, we define a new learning rule, which we call *denoised empirical risk minimization* (DERM). At the heart of this is a partition of the hypothesis class into *approximately jointly measurable* subsets. Intuitively, approximate joint measurability of a set S of POVMs allows us to reuse samples in the training set to evaluate the denoised empirical risk for every element of S.

   We show a sample complexity upper bound for DERM in terms of a suitably defined version of the fat shattering dimension and the *approximate joint measurability (JM) covering number* of the hypothesis class, which implies that finiteness of both of these quantities for a POVM class is a *sufficient condition* for learnability.

   We then exhibit a link between the JM covering number and the fat shattering dimension, thereby connecting the quantum concept with a learning theoretic complexity measure. Specifically, we show that the JM covering number lower bounds the fat shattering dimension, which implies that finite fat shattering dimension alone is a sufficient condition for learnability.

   We next show a sample complexity *lower* bound, again in terms of the fat shattering dimension. This implies that finite fat shattering dimension is a *necessary* condition for learnability. That is, **finiteness of the fat shattering dimension is necessary and sufficient for POVM classes**. This constitutes a fundamental theorem of concept learning for these classes. While the main topic of this paper is the quantum setting, the result for POCC classes may be of independent interest, as they are the natural formulation of the learning problem in cases where the value of the loss function depends on unobserved variables whose joint distribution with the input X is known.

   Our results improve substantially on prior work on this learning scenario, which did not prove any necessary conditions and which provided sample complexity upper bounds that are frequently infinite for finite-dimensional hypothesis classes that we can show to be learnable.

4. **Example POVM classes:** We then give examples of learnable and non-learnable POVM classes, including quantum neural networks (learnable) and an infinite-dimensional, nontrivally quantum example of a non-learnable class.

   All proofs are provided in the supplementary material.

2 Main results: learnability, uniform convergence, and ERM

2.1 Preliminaries

We first define the learning problems relevant to us. Definitions from quantum mechanics can be found in [6] and in the supplementary material.

Definition 1 (POVM concept class learning problem [1, 4]). *In the POVM concept/hypothesis class learning problem, we fix a set of possible input mixed states $\mathcal{X}$, which are density operators on a common Hilbert space $\mathcal{H}$, and a set of possible classical outputs $\mathcal{Y}$. We fix a* loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty]$.

We fix a POVM concept class Hyp*, which is simply a set of POVMs on $\mathcal{H}$ having $|\mathcal{Y}|$ outcomes. Informally, a* ***learning rule*** *$\mathcal{A}$ in this context takes as input a dataset $\{(\rho_j, Y_j)\}_{j=1}^m$ consisting of quantum registers in states $\rho_j \in \mathcal{X}$ and classical outputs $Y_j \in \mathcal{Y}$. This dataset is sampled from an unknown joint distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. The learning rule interacts with the ρ_j via quantum measurements (formally, POVMs). More precisely, all decisions made by the learning rule can only depend on any of the ρ_j via the classical labels Y_j and via outcomes of POVMs applied to those states or to states resulting from previous measurements. A learning rule is thus a* Markov decision process, *as detailed in Section F. Finally, the learning rule outputs a POVM $\Phi_* \in$ Hyp with the goal of minimizing $\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(\Phi_*[X], Y)]$, where $\Phi_*[X] \in \mathcal{Y}$ denotes the random outcome resulting from measuring X with ρ_*. We give a more formal definition of a learning rule in the supplementary materials.*

We say that a POVM learning rule $\mathcal{A}$ is (ϵ, δ)-probably approximately correct (PAC) for Hyp *if there exists a sample size $m = m(\epsilon, \delta)$ such that for all distributions $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ with $S \sim \mathcal{D}^m$, with probability at least $1 - \delta$, $\mathcal{A}(S)$ outputs a hypothesis $h \in$ Hyp satisfying*

$$R(h) - \inf_{h_* \in \mathrm{Hyp}} R(h_*) \leq \epsilon. \tag{1}$$

We then say that Hyp *is (ϵ, δ)-PAC learnable if there exists an (ϵ, δ)-PAC learning rule for* Hyp*. Finally, we say that* Hyp *is PAC learnable if it is (ϵ, δ)-PAC learnable for all $\epsilon > 0, \delta > 0$.*

The learning problem defined in Definition 1 is related to the problem of *probabilistically observed concept class learning*, which we introduce below.

Definition 2 (Probabilistically observed concept class learning problem). *In the probabilistically observed concept class (POCC) learning problem, $\mathcal{X}$ becomes an arbitrary set, and* Hyp *consists of functions $f : \mathcal{X} \to \Delta(\mathcal{Y})$, where $\Delta(S)$ denotes the set of probability distributions on a set S.*

When a hypothesis $h \in \mathrm{Hyp}$ is applied to an element $x \in \mathcal{X}$, the learning rule only observes a random sample $Z \sim h(x)$, not $h(x)$ itself. We denote a generic sample from $h(x)$ by $h[x]$.

Given this setting, the definition of PAC learning remains the same as before.

Remark 1 (Probabilistic versus probabilistically observed concept learning)**.** *We emphasize the important distinction between the probabilistic concepts (also called p-concepts) of [29] and the probabilistically observed concepts in the present paper: in the p-concept framework, the output probability distribution itself is observed, rather than just a sample from it. In our setting, in contrast, our learning rules are only allowed to see a sample from an unknown output probability distribution.*

2.1.1 Connecting POVM classes with POCCs

Here we describe the connection between the POVM and POCC frameworks. The POVM framework is more general than the POCC one: we first show how to translate the problem of learning a POCC class to one of learning a POVM class, along with translations of POCC learning rules to POVM learning rules.

Given a POCC learning problem with domain $\mathcal{X}$ and hypothesis class Hyp, the quantumization of this problem is formulated as follows: we introduce a Hilbert space $\mathcal{H}$ with dimension equal to $|\mathcal{X}|$ (which may be uncountably infinite), and we choose, arbitrarily, an orthonormal basis $B = \{e_x\}_{x \in \mathcal{X}}$. Each $x \in \mathcal{X}$ corresponds to a basis element $e_x \in \mathcal{H}$. The domain of the POVM learning problem is the basis B. Each hypothesis $h \in \mathrm{Hyp}$ bijectively maps to a corresponding POVM Π_h defined as follows: Π_h first measures in the basis B, uniquely identifying the input state e_x with probability 1, then postprocesses e_x through the classical channel corresponding to $h(x)$. (We note that a POVM may be constructed by measurement of a state with a POVM, then postprocessing the outcome through a classical channel.)

A POCC learning rule is a function of inputs x and samples from an arbitrary set of hypotheses $h[x]$. The analogous POVM learning rule is the same function as in the classical case, applied to the classical outcome of measurement in the basis B, along with results of passing this outcome through channels associated with hypotheses in Hyp.

Thus, a POCC learning rule can be translated to a quantum one with exactly the same error characteristics. The situation becomes more complicated when we generalize to truly quantum learning settings, because certain operations that are possible in the classical case are not possible in the quantum. In particular, in quantum settings, the hypothesis class consists of non-orthogonal states, which cannot be almost surely distinguished from one another. Thus, our learning rules cannot be functions of the inputs themselves, but instead can only be functions of outcomes of measurements applied to these inputs.

2.1.2 Jointly measurable sets of POVMs

We also need to recall the notions of a fine-graining of a POVM and a *jointly measurable* set of POVMs [30]. Intuitively, joint measurability will allow us to reuse samples to evaluate multiple hypotheses.

Definition 3 (Fine-graining of a POVM)**.** *Let* Π *be a POVM. We say that* Π *has a fine-graining* (Π', α)*, where* Π' *is a POVM and* α *is a classical channel, if the outcome of* Π *on any state has the same distribution as the outcome of* Π' *on the same state, then passed through the channel* α*. We call* Π' *the* root *of the fine-graining.*

We denote the set of fine-grainings of Π *by* $\mathfrak{F}(\Pi)$*.*

For a given fine-graining Φ*, we denote its root by* $\mathfrak{R}(\Phi)$*.*

Definition 4 (Jointly measurable set of POVMs)**.** *A set* S *of POVMs is said to be* jointly measurable *if there exists a POVM* Π_* *such that, for every* $\Pi \in S$*,* Π *has a fine-graining with root POVM equal to* Π_**. We then say that* Π_* *is a root POVM for* S*.*

The consequence of joint measurability of S is that one can obtain a sample outcome from measuring a state ρ with every element of S by first measuring ρ with a root POVM and then passing this outcome through each of the classical channels corresponding to the fine-grainings of the different POVMs in S.

2.2 Failure of uniform convergence and ERM for PAC learnable probabilistically observed hypothesis classes

Having dispensed with preliminaries, we next develop our first main results. The empirical risk minimization (ERM) rule is a cornerstone of statistical learning theory in the setting of deterministic concept classes. For a dataset $S = \{(X_i, Y_i)\}_{i=1}^m$, the empirical risk of a hypothesis $h \in \mathrm{Hyp}$ is given by

$$\hat{R}(h, S) = \frac{1}{m} \sum_{j=1}^{m} \ell(h[X_i], Y_i). \tag{2}$$

In the deterministic case, a hypothesis class Hyp being PAC learnable is logically equivalent to ERM being a PAC learning rule, which is logically equivalent to it satisfying the following uniform convergence property: for any hypothesis $h \in \mathrm{Hyp}$ and any data-generating distribution $\mathcal{D}$, $\mathrm{P}_{S \sim \mathcal{D}}[|R(h) - \hat{R}(h, S)| \geq \epsilon] \leq \delta$.

The ERM rule has been proposed for use as a subroutine in the quantum setting in prior work [1] and also adopted in the more recent work [4]. Both of these works give sample complexity upper bounds for this ERM rule. Our first main result is that uniform convergence and the ERM rule can fail for a POCC class Hyp, despite Hyp being PAC learnable. This is in stark contrast to the deterministic case. We will see, in our Theorem 2, that this has further implications for the quantum setting, and thus for the tightness of the bounds in [4]. Theorems 1 and 2 imply that our subsequent quantum learnability results in Sections 3 and 4 cannot simply build on prior work by either bounding VC dimension or by tighter analysis of the ERM rule – rather, we must propose *new* learning rules.

Theorem 1 (Failure of uniform convergence and ERM for POCC classes)**.** *There exists a POCC class* Hyp *that is PAC learnable but for which the ERM rule is not PAC and does not satisfy the uniform convergence property.*

Furthermore, there exists a POCC class $\widehat{\mathrm{Hyp}}$ *and a choice of* $\mathcal{X}, \mathcal{Y}$*, and* $\mathcal{D}$ *for which the uniform convergence property is not satisfied, but the ERM rule is PAC.*

2.3 Failure of uniform convergence for most finite-dimensional POVM classes

We next show that the situation regarding ERM is even worse in the quantum case. In particular, a consequence of what we show next is that the sample complexity upper bounds in [4] are infinite (i.e., vacuous) for a very large class of POVM classes that are learnable. To do so, we recall the definition of a deterministic POVM.

Definition 5 (Deterministic POVM)**.** *A POVM* $\Pi = \{\Pi_0, \Pi_1\}$ *is deterministic if either* $\Pi_0 = 0$ *or* $\Pi_1 = 0$*.*

That is, the outcome of a deterministic POVM is the same when used to measure any state.

We also define the L_1 operator norm for operators on a Hilbert space $\mathcal{H}$. For an operator $\Gamma : \mathcal{H} \to \mathcal{H}$, the L_1 operator norm is given by

$$\|\Gamma\|_{op,L_1} = \sup_{x \in \mathcal{H}} \frac{\|\Gamma x\|_1}{\|x\|_1}. \tag{3}$$

Any norm generates a topology, which allows us to talk about open and closed sets.

Theorem 2 (Failure of uniform convergence of ERM for most finite-dimensional POVM classes)**.** *Let* $\mathcal{X}$ *be a subset of a finite-dimensional Hilbert space* $\mathcal{H}$*. Consider an* L_1 *operator norm-closed POVM hypothesis class* Hyp *satisfying the following conditions:*

1. Hyp *is jointly measurable.*
2. Hyp *has infinite cardinality.*

Then exactly one of the following conclusions holds:

1. *Uniform convergence for ERM does not hold for* Hyp*, and ERM is not PAC.*
2. *The only points of accumulation of* Hyp *are deterministic POVMs.*

One can construct a Hyp satisfying the conditions listed in Theorem 2 by selecting a two-outcome root POVM Π along with the class of binary symmetric channels α_p with crossover probabilities $p \in [1/4, 3/4]$. When Hyp is the set of POVMs consisting of compositions of Π and α_p, it is immediate that Hyp is jointly measurable, has infinite cardinality, and is L_1 norm-closed. In this case, the first conclusion holds, while the second does not. An example in which Hyp satisfies all hypotheses and only the second conclusion is as follows: we select the root POVM $\Pi := (0, I)$, where I is the identity operator, and binary symmetric channels $\{\alpha_{p_j}\}_{j=1}^{\infty}$ with crossover probabilities $p_j \to 0$ monotonically as $j \to \infty$, along with α_0. Then we define Hyp to be the set of POVMs that are compositions of Π with the α_{p_j} and α_0.

Theorem 2 effectively says that an infinite-cardinality (but possibly finite-dimensional) POVM class can only enjoy the uniform convergence property for ERM if it "clusters" around deterministic measurements. The only deterministic measurements are the ones whose outcomes do not depend on the states being measured. This implies that ERM is not a useful learning rule for a rich enough set of POVM classes. Since ERM was a core subroutine of [4], this provides useful insight on prior work: in particular, in that work, a sample complexity upper bound for the ERM rule is given in the case where one can find a finite-cardinality jointly measurable partition. Our theorem above implies that this upper bound must be ∞ unless almost all of the hypotheses are close to deterministic (and, thus, independent of the input state). In our subsequent theorems, we will show that the upper bound of infinity is, in infinitely many cases, hopelessly loose as a bound on the minimum possible sample complexity of learning (irrespective of the learning rule). Crucially, in the above theorem, we note that ERM failing to be PAC for a hypothesis class Hyp does not imply that the class is not learnable.

3 Main results: Every finite-dimensional POVM class is learnable

In this section, we give a complete characterization of learnability of POVM classes in the case where $\mathcal{X}$ is a (**possibly infinite-cardinality**) subset of the set of density operators on a finite-dimensional Hilbert space $\mathcal{H}$. We call a POVM class defined on $\mathcal{X}$ a finite-dimensional POVM class. It turns out that *every finite-dimensional POVM class is learnable* – Theorem 3.

Theorem 3 (Every finite-dimensional POVM class is learnable)**.** *Let the span of the domain $\mathcal{X}$ be a finite-dimensional subspace of the space of density operators on a Hilbert space $\mathcal{H}$. Let* Hyp *be a POVM class all of whose POVMs are defined on $\mathcal{X}$.*

Then Hyp *is PAC learnable with the following sample complexity:*

$$n_{\mathrm{Hyp}}(\epsilon, \delta) \leq \frac{8N}{\epsilon^2} \log \frac{2N}{\delta}, \tag{4}$$

where N is the $\epsilon/4$-total variation covering number of Hyp*, which is finite.*

We note that in the worst case, the covering number in Theorem 3 can be exponential in the dimension of the Hilbert space. However, hypothesis classes of interest, where the POVMs have constrained structure, have a much smaller covering number. Additionally, in certain cases, one can take advantage of joint measurability in order to tighten this bound.

The above theorem provides infinitely many examples of POVM classes that are learnable. Furthermore, this class of examples includes ones such that the sample complexity upper bounds given in [4] were infinite. Therefore, this is a substantial improvement on the previous results.

Interestingly, the proof involves concocting a learning rule that uses ERM, but in a different way from prior work. This approach only works for the finite-dimensional case, necessitating yet another learning rule for our subsequent results.

We emphasize that Theorem 3 *does not* contradict Theorem 2 – the former states that finite-dimensional POVM classes are learnable *by some learning rule* and makes no statement about ERM; the latter is a statement specifically about the failure of ERM as a learning rule.

4 Main results: Matching necessary and sufficient conditions for infinite-dimensional POVM class learnability

The result in Theorem 3 leaves open the questions of necessary and sufficient conditions for infinite-dimensional POVM classes and POCC classes. Furthermore, the results in Theorems 1 and 2 motivate a search for an alternative learning rule to ERM for both the POCC and POVM cases.

In Section 4.1, we present our learning rule – the *denoised ERM* – for POVM classes and how it specializes to the POCC case. We then show in Section 4.2 necessary and sufficient conditions for PAC learnability of POCC and POVM classes. Specifically, Theorem 4 gives distinct necessary and sufficient conditions. In Theorem 5, we show an inequality relating a quantity called the *approximate joint measurability covering number* to the *fat shattering dimension* of Hyp, which allows us to conclude with Corollary 1 that the conditions in Theorem 4 are matching.

4.1 Rescuing ERM: De-noised empirical risk

We now turn to the definition of our new learning rule, called *denoised empirical risk minimization.* We first define the denoised empirical risk, which is for a hypothesis in a set of jointly measurable POVMs.

Definition 6 (Denoised empirical risk of a hypothesis). *Let H be a jointly measurable set of POVMs with a fine-graining $(\Pi, \{\alpha_h\}_{h\in H})$.*

Let $h \in H$ be a hypothesis, and let $S = \{(X_j, Y_j)\}_{j=1}^m \in (\mathcal{X} \times \mathcal{Y})^m$ be a dataset. Let Z_j denote the random outcome of measurement of X_j with the POVM Π. We define the denoised empirical risk of h on input S *to be*

$$\mathrm{DER}(h, S) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}[\ell(h[X_j], Y_j) \mid \{(Z_j, Y_j)\}_{j=1}^m]. \tag{5}$$

Note that the denoised empirical risk is a random variable.

Remark 2. *It is* ***essential*** *to note in this definition* what *is being conditioned on. Specifically, one can imagine definitions of the empirical risk that either suffer from the same flaws as the ordinary empirical risk or, alternatively, cannot be computed by the learner because of the learner's inability to see the X_j. Intuitively, our definition "averages out" the randomness from the classical channels, which mitigates the drawbacks of the ordinary empirical risk.*

To state the learning rule, we also need a relaxation of the notion of a jointly measurable partition introduced in [4]. To state this, we also need to define the total variation distance between POVMs.

Definition 7 (Total variation distance between POVMs). *Let Π_1, Π_2 be two POVMs with common domain $\mathcal{X}$. We define the total variation distance between Π_1, Π_2 as follows:*

$$d_{TV}(\Pi_1, \Pi_2) = \sup_{x\in\mathcal{X}} d_{TV}(\mathrm{Out}(\Pi_1, x), \mathrm{Out}(\Pi_2, x)), \tag{6}$$

where $\mathrm{Out}(\Pi, x)$ *denotes the random outcome of the POVM* Π *on the mixed state* x.

With this in hand, we next define an *approximately jointly measurable class.*

Definition 8 (Approximately jointly measurable class)**.** *We say that a collection* S *of POVMs is* γ*-approximately jointly measurable (or just* γ*-jointly measurable) if there exists a POVM* Π_* *such that, for every POVM* $\Pi \in S$, Π *has a fine-graining with root* Π' *satisfying* $d_{TV}(\Pi_*, \Pi') \leq \gamma$, *where we recall the definition of the total variation distance between two POVMs in Definition 7. We call* Π_* *a* center *of* S.

Our proofs will exploit the fact that an approximately jointly measurable class S can be approximated by a jointly measurable one by choosing a center POVM Π_* of S and replacing each element of S with a POVM whose root is Π_*. This is formalized in the following definition.

Definition 9 (Joint measurability-smoothed class)**.** *Let* S *be a* γ*-jointly measurable POVM class, and let* Π_* *be a center for* S. *For a POVM* $h \in S$, *we define* $\mathbb{S}_{JM}(h, \Pi_*)$ *to be the following POVM:*

$$\mathbb{S}_{JM}(h, \Pi_*) = \underset{\Pi \,:\, \Pi_* \in \mathfrak{R}(\mathfrak{F}(\Pi))}{\arg\min} d_{TV}(\Pi, h). \tag{7}$$

We denote by $\mathbb{S}_{JM}(S, \Pi_*)$ *the following set of POVMs:*

$$\mathbb{S}_{JM}(S, \Pi_*) = \{\mathbb{S}_{JM}(h, \Pi_*) \mid h \in S\}. \tag{8}$$

We next define joint measurability notions for general hypothesis classes.

Definition 10 (Approximately jointly measurable partition)**.** *Let* Hyp *be a POVM class. Then a partition* $\mathcal{P}$ *of* Hyp *is a* γ*-approximately jointly measurable partition of* Hyp *if each partition element* P_j *is* γ*-approximately jointly measurable.*

Definition 11 (Approximate joint measurability covering numbers)**.** *Let* S *be a collection of POVMs. We say that a collection of subsets* $\{S_j\}_{j\in\Omega}$ *of* S *is a* γ*-joint measurability covering of* S *if* $\bigcup_{j\in\Omega} S_j = S$ *and every* S_j *is* γ*-jointly measurable.*

We define the γ*-joint measurability covering number* $\mathcal{N}_{JM}(\gamma, S)$ *to be the minimum cardinality of a* γ*-joint measurability covering of* S.

Having defined the denoised empirical risk of a hypothesis and the notion of an approximately jointly measurable partition of a hypothesis class, we present the denoised empirical risk minimization rule in Algorithm 1. This rule is parametrized by an approximately-jointly measurable partition with finite cardinality and a choice, for each partition element P_j, of a number of samples n_j, satisfying $\sum_{j=1}^{R} n_j = m$, where m is the size of the input training set. We will explain how to choose n_j based on the sample complexity bounds that we will derive. The input training set is partitioned into consecutive subsets $\hat{S}_1, ..., \hat{S}_R$, of cardinalities $n_1, ..., n_R$.

Remark 3 (The distinction between learning rules and algorithms)**.** *A distinction is made in statistical learning theory between a learning rule and an algorithm. In the classical case, ERM is a learning rule,* not *an algorithm, because it does not specify* how *to achieve the minimization. Achieving the minimization algorithmically requires more refined structural knowledge of the hypotheses. Indeed, for many hypothesis classes, ERM is a PAC learning rule, but it is not efficiently implementable by an algorithm.*

Algorithm 1: Denoised empirical risk minimization learning rule

Data: Training set $S = \{(X_j, Y_j)\}_{j=1}^m$; Approximately jointly measurable partition $\mathcal{P} = \{P_j\}_{j=1}^R$ of Hyp with centers $\Pi_*^{(j)}$; partitioned training set $\{\hat{S}_i\}_{i=1}^R$

Result: A hypothesis $h_* \in \text{Hyp}$ meant to have almost minimum risk

```
for j = 1 to R do
    // Process partition element j.
    Let P̂_j = S_JM(P_j, Π_*^(j));
    Compute ρ̂_j = argmin_{h∈P̂_j} DER(h, Ŝ_j) and R̂_j = DER(ρ̂_j, Ŝ_j);
end
Let j_* = argmin_{j∈{1,...,R}} R̂_j;
Let h_* be some hypothesis in Hyp such that S_JM(h_*, Π_*^(j)) = ρ̂_{j_*};
return h_*;
```

The study of efficient algorithmic implementation of learning rules is the subject of computational, rather than statistical, learning theory.

In our case, the DERM is similarly a learning rule, not an algorithm. We do not, and cannot, prescribe how to construct the jointly measurable partition, for instance. It is only important that the partition exists.

4.2 Necessary and sufficient conditions for PAC learnability

We next turn to necessary and sufficient conditions for PAC learnability of POVM and POCC classes. To do this, we need to define the fat shattering dimension of a class of POVMs [31]. We start by recalling the fat shattering dimension in the deterministic hypothesis class case.

Definition 12 (Fat-shattering dimension (classical case))**.** *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. We say that a dataset $(x_1, ..., x_m) \in \mathcal{X}^m$ is γ-fat-shattered by $\mathcal{F}$ if there exist* witness numbers $r_1, ..., r_m$ *such that, for any $(b_1, ..., b_m) \in \{0,1\}^m$, there exists some $f \in \mathcal{F}$ such that, simultaneously, $f(x_i) \leq r_i - \gamma$ if $b_i = 0$ and $f(x_i) \geq r_i + \gamma$ if $b_i = 1$.*

We define the γ-fat-shattering dimension of $\mathcal{F}$ to be the largest m for which there exists a dataset $X \in \mathcal{X}^m$ that is γ-fat-shattered by $\mathcal{F}$.

To define an analogous notion for POVM classes Hyp, we identify each POVM h with its induced function mapping from $\mathcal{X}$ to probability distributions on $\{0, 1\}$, which themselves can be identified with real numbers in $[0, 1]$. Thus, each hypothesis induces a function $f_h : \mathcal{X} \to [0, 1]$, and we define the γ-fat-shattering dimension of Hyp to be that of the induced deterministic function hypothesis class.

Theorem 4 (Necessary and sufficient conditions for PAC learnability of POVM classes)**.** *Let* Hyp *be a POVM class (not necessarily finite-dimensional or finite-cardinality). Then the following statements hold.*

1. ***Necessary condition:*** *If* Hyp *is PAC learnable, then for every $\gamma > 0$,* Hyp *has finite γ-fat-shattering dimension.*

2. ***Sufficient condition:*** *If there exists a finite $d > 0$ such that, for every small enough $\gamma > 0$,* Hyp *has γ-fat-shattering dimension $\leq d$ and, additionally, for every $\alpha > 0$, there exists a finite α-almost-jointly-measurable partition $\mathcal{P}$ of* Hyp*, then* Hyp *is PAC learnable by DERM, with sample complexity*

$$n_{\text{Hyp}}(\epsilon, \delta) \leq \inf_{\mathcal{P}} O(|\mathcal{P}| \frac{d + \log(1/\delta)}{\epsilon^2}), \tag{9}$$

where $\mathcal{P}$ ranges over all finite $\epsilon/8$-almost-jointly-measurable partitions of Hyp*.*

We next present a theorem linking fat-shattering dimension to approximate joint measurability.

Theorem 5 (Lower bound on fat-shattering dimension via approximate joint measurability covering number)**.** *Let* Hyp *be a POVM class with $k \leq \mathcal{N}_{JM}(\gamma, \text{Hyp})$ and with $d = \text{fat}_{\gamma/4}(\text{Hyp})$. Then for every $\gamma > 0$, if $m \geq \binom{k}{2}$, then*

$$k \leq 2 \cdot (m \cdot (2/\gamma + 1)^2)^{\lceil d \cdot \log(\frac{2em}{d\gamma}) \rceil}. \tag{10}$$

In particular, if $\mathcal{N}_{JM}(\gamma, \text{Hyp}) = \infty$, then $\text{fat}_{\gamma/4}(\text{Hyp}) = \infty$.

Theorem 5 immediately implies the following corollary of Theorem 4.

Corollary 1 (Fundamental theorem of concept learning for POVM classes)**.** *Let* Hyp *be a POVM class. Then* Hyp *is PAC learnable if and only if, for every $\gamma > 0$,* Hyp *has finite γ-fat-shattering dimension. Furthermore, DERM is a PAC learning rule for* Hyp*. That is, DERM is a universal learning rule.*

Specifically, this is because finite fat-shattering dimension implies finiteness of $\mathcal{N}_{JM}(\gamma, \text{Hyp})$ for all γ, rendering the finiteness of $\mathcal{N}_{JM}(\gamma, \text{Hyp})$ redundant as a sufficient condition for learnability.

Remark 4. *Corollary 1 subsumes Theorem 4.*

Corollary 1 constitutes a fundamental theorem of concept learning for POVM classes. As mentioned in the introduction, this implies the same for POCC classes, which arise whenever the loss function value depends on unobserved variables.

We note that a simple and natural condition for finite fat-shattering dimension of a POVM class is that the class be convex, with finitely many extreme points. This is the content of our next corollary.

Corollary 2 (Fat shattering dimension and geometry of Hyp)**.** *Suppose that* Hyp *is a convex set with $k < \infty$ extreme points $V := \{\Pi^{(j)}\}_{j=1}^k$. Then the γ-fat-shattering dimension of* Hyp *is $< \infty$. Thus,* Hyp *is PAC learnable.*

Proof. Theorem 1.5 of [32] states that the γ-fat-shattering dimension of the convex hull $\text{conv}(V)$ of a set V of functions whose range lies in $[0, 1]$ is upper bounded as follows:

$$\text{fat}_\gamma(\text{conv}(V)) \leq C \cdot \frac{\text{fat}_{\gamma/4}(V)}{\gamma^2} \log^2 \left(\frac{2\text{fat}_{\gamma/4}(V)}{\gamma} \right). \tag{11}$$

Since the $\gamma/4$-fat shattering dimension of a finite set of hypotheses (in this case, the extreme points V of the convex set Hyp) is finite for any γ, this implies that the γ-fat shattering dimension of the convex hull (namely, Hyp) is also finite. □

4.3 Examples and applications

Here we present example POVM classes to illustrate our results. We first present an application of our results to the learnability of variational quantum circuits (sometimes called quantum neural networks).

Definition 13 (Quantum neural network [2])**.** *Consider a Hilbert space $\mathcal{H}$ with dimension $d < \infty$. An ℓ-layer quantum neural network is parametrized by a sequence of ℓ unitary operators $U_1, U_2, ..., U_\ell$ and a measurement Π, all mapping $\mathcal{H}$ to $\mathcal{H}$, with Π having two classical outcomes. Each of the unitary operators U_j acts nontrivially on only a subset of qubits.*

Corollary 3 (Learnability of quantum neural networks)**.** *For any $d < \infty$, the class of quantum neural networks on a Hilbert space with dimension d is PAC learnable.*

Proof. This is an immediate consequence of Theorem 3 and the fact that d is stipulated to be finite. □

We give a more detailed result on learning variational quantum circuits in the supplementary material, specifically Appendix G. There, in Theorem 8, we give a concrete sample complexity bound for general variational quantum circuits.

As in classical learning theory, one advantage of combinatorial bounds on generalization error, such as those in terms of VC or fat-shattering dimension, is that they are *distribution-free*, meaning that they do not depend on the data generating distribution. This generality, of course, comes at a cost of tightness, as is well-known to be the case for neural networks. We have nonetheless included the above example to illustrate the application of our bounds to a concrete hypothesis class. We emphasize that our bounds can be applied beyond classes of quantum neural networks, and that our focus in this work is on generality, as is the case in the classical works using the fat-shattering dimension.

Non-learnable POVM classes:

We next turn to examples of POVM classes that are *not* learnable. Theorem 3 implies that we must consider input state sets spanning an infinite-dimensional Hilbert space. One can easily concoct an unlearnable class from a classical hypothesis class with infinite VC dimension. Our next example goes further than this to provide intuition about legitimately quantum POVM classes that are unlearnable.

Example 1 (A POVM class that is not learnable)**.** *Consider a domain $\mathcal{X}$ whose span is infinite-dimensional in some Hilbert space H with a sequence of orthonormal vectors $\{\rho_j\}_{j=1}^{\infty}$ and a sequence of numbers $\{\beta_j\}_{j=1}^{\infty}$, with $\beta_j \in (1/2+\gamma, 1)$. We construct a sequence of hypotheses as follows (since each hypothesis has two outcomes, we only need to specify for each hypothesis the operator corresponding to the 0 outcome): letting*

$b = (b_1, b_2, ...) \in \{0,1\}^\infty$,

$$\Pi_0^{(b)} = \sum_{j=1}^{\infty} \beta_j^{1-b_j} (1-\beta_j)^{b_j} |\rho_j\rangle \langle\rho_j| . \tag{12}$$

Finally, let Hyp *be the class of all such POVMs:* Hyp $= \{\Pi^{(b)}\}_{b \in \{0,1\}^\infty}$*. We claim that* Hyp *is not PAC learnable. To show this, by Theorem 4, it is sufficient to show that* Hyp *has infinite γ-fat shattering dimension. We show this in the following sequence of steps:*

1. *We exhibit a candidate sequence $x_1, x_2, ...$ of elements of $\mathcal{X}$ that we will show to be fat-shattered by hypotheses in* Hyp*. Specifically, we take $x_j = |\rho_j\rangle \langle\rho_j|$.*
2. *We exhibit a* witness number $r = 1/2$.
3. *We exhibit, for each bit string $b = b_1, b_2, ... \in \{0,1\}^\infty$, a hypothesis $h \in$ Hyp such that when $b_j = 1$, $\mathrm{P}[h[x_j] = 1] \geq r + \gamma$ and when $b_j = 0$, $\mathrm{P}[h[x_j] = 0] \leq r - \gamma$, for every j. Specifically, we take $h = \Pi^{(\mathbf{1}-b)}$. We have, when $b_j = 0$,*

$$\mathrm{P}[h[x_j] = 0] = \mathrm{Tr}\Big(x_j \Pi_0^{(\mathbf{1}-b)}\Big) \tag{13}$$

$$= \langle\rho_j| \Pi_0^{(\mathbf{1}-b)} |\rho_j\rangle = \beta_j^{b_j} (1-\beta_j)^{1-b_j}, \tag{14}$$

so that $\mathrm{P}[h[x_j] = 0] = 1 - \beta_j \leq 1/2 - \gamma$, by assumption. Similarly, when $b_j = 1$, $\mathrm{P}[h[x_j] = 1] = 1 - \beta_j^{1-b_j} (1-\beta_j)^{b_j} \geq 1 - (1/2 - \gamma) = 1/2 + \gamma$.

Thus, Hyp *has infinite γ-fat-shattering dimension, which implies that it is not PAC learnable.*

We note that the results of prior papers were incapable of showing that *any* POVM class is unlearnable.

5 Conclusion

We have provided matching necessary and sufficient conditions for learnability of POVM hypothesis classes in terms of their fat-shattering dimension. To do so, we connected the learning-theoretic notion of fat-shattering dimension with the quantum concept of approximate joint measurability covering. The proof of our sufficient condition came via the introduction of a new universal learning rule, the de-noised empirical risk minimization rule. Additionally, we showed that all finite-dimensional POVM classes are learnable, and we provided quantitative sample complexity bounds for some example hypothesis classes.

There are various possible extensions of our work: for instance, a characterization of the fat-shattering dimension of a hypothesis class in terms of its Hilbert space geometry would be of interest. Additionally, our learning rule only makes *separable* measurements. In quantum hypothesis testing, where the goal is to distinguish between two *known* states with minimal error probability from m copies of one of them, block measurements have a provable advantage in terms of sample complexity. It would be interesting to understand whether this phenomenon holds in the learning setting.

6 Acknowledgments

This research is funded by NSF CCF grant number #2212327.

References

[1] Heidari, M., Padakandla, A., Szpankowski, W.: A theoretical framework for learning from quantum data. In: 2021 IEEE International Symposium on Information Theory (ISIT), pp. 1469–1474 (2021). https://doi.org/10.1109/ISIT45174.2021.9517721

[2] Heidari, M., Grama, A., Szpankowski, W.: Toward Physically Realizable Quantum Neural Networks. arXiv (2022). https://doi.org/10.48550/ARXIV.2203.12092 . https://arxiv.org/abs/2203.12092

[3] Wu, Y., Wu, B., Wang, J., Yuan, X.: Quantum phase recognition via quantum kernel methods. Quantum **7**, 981 (2023) https://doi.org/10.22331/q-2023-04-17-981

[4] Padakandla, A., Magner, A.: Pac learning of quantum measurement classes : Sample complexity bounds and universal consistency. In: Camps-Valls, G., Ruiz, F.J.R., Valera, I. (eds.) Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 151, pp. 11305–11319. PMLR, Virtual (2022). https://proceedings.mlr.press/v151/padakandla22a.html

[5] Arunachalam, S., De Wolf, R.: Optimal quantum sample complexity of learning algorithms. J. Mach. Learn. Res. **19**(1), 2879–2878 (2018)

[6] Wilde, M.: Quantum Information Theory. Cambridge University Press, Cambridge, UK (2017). https://arxiv.org/abs/1106.1445

[7] Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning - From Theory to Algorithms. Cambridge University Press, Cambridge, UK (2014)

[8] Arunachalam, S., De Wolf, R.: Guest column: A survey of quantum learning theory. SIGACT News **48**(2), 41–67 (2017) https://doi.org/10.1145/3106700.3106710

[9] Arunachalam, S., Grilo, A.B., Yuen, H.: Quantum statistical query learning. arXiv preprint arXiv:2002.08240 (2020) arXiv:2002.08240 [quant-ph]

[10] Anshu, A., Arunachalam, S., Kuwahara, T., Soleimanifar, M.: Sample-efficient learning of interacting quantum systems. Nature Physics, 1745–2481 (2021) https://doi.org/10.1038/s41567-021-01232-0

[11] O'Donnell, R., Wright, J.: Efficient quantum tomography. In: Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing. STOC '16, pp. 899–912. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2897518.2897544 . https://doi.org/10.1145/2897518.2897544

[12] O'Donnell, R., Wright, J.: Efficient quantum tomography ii. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing. STOC 2017, pp. 962–974. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3055399.3055454 . https://doi.org/10.1145/3055399.3055454

[13] Haah, J., Harrow, A.W., Ji, Z., Wu, X., Yu, N.: Sample-optimal tomography of quantum states. In: Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing. STOC '16, pp. 913–925. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2897518.2897585 . https://doi.org/10.1145/2897518.2897585

[14] Aaronson, S.: The learnability of quantum states. Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences **463** (2006) https://doi.org/10.1098/rspa.2007.0113

[15] Rehacek, J., Paris, M.: Quantum State Estimation. Lecture notes in physics. Springer, Berlin (2004). https://doi.org/10.1007/b98673 . https://cds.cern.ch/record/1391369

[16] Waseem, M.H., Faizan-e-Ilahi, Anwar, M.S.: Quantum state tomography. In: Quantum Mechanics in the Single Photon Laboratory. 2053-2563, pp. 6–1620. IOP Publishing, Bristol, UK (2020). https://doi.org/10.1088/978-0-7503-3063-3ch6 . http://dx.doi.org/10.1088/978-0-7503-3063-3ch6

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

[18] Aaronson, S.: Shadow tomography of quantum states. In: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. STOC 2018, pp. 325–338. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3188745.3188802 . https://doi.org/10.1145/3188745.3188802

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

[20] Bshouty, N.H., Jackson, J.C.: Learning dnf over the uniform distribution using a quantum example oracle. SIAM Journal on Computing **28**(3), 1136–1153 (1998)

[21] Atici, A., Servedio, R.A.: Improved bounds on quantum learning algorithms. Quantum Information Processing **4**(5), 355–386 (2005) https://doi.org/10.1007/s11128-005-0001-2

[22] Roget, M., Molfetta, G.D., Kadri, H.: Quantum perceptron revisited: Computational-statistical tradeoffs. In: The 38th Conference on Uncertainty in Artificial Intelligence (2022). https://openreview.net/forum?id=BuOcoU8jqxq

[23] Caro, M.C., Huang, H.-Y., Cerezo, M., Sharma, K., Sornborger, A., Cincio, L., Coles, P.J.: Generalization in quantum machine learning from few training data. Nature Communications **13**(1), 4919 (2022) https://doi.org/10.1038/s41467-022-32550-3

[24] Rosati, M.: A learning theory for quantum photonic processors and beyond (2023)

[25] Fanizza, M., Quek, Y., Rosati, M.: Learning quantum processes without input control (2024)

[26] Caro, M., Gur, T., Rouzé, C., Stilck França, D., Subramanian, S.: Information-theoretic generalization bounds for learning from quantum data (2023)

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

[28] Kearns, M.J., Schapire, R.E., Sellie, L.M.: Toward efficient agnostic learning. Machine Learning **17**(2-3), 115–141 (1994) https://doi.org/10.1007/bf00993468

[29] Kearns, M.J., Schapire, R.E.: Efficient distribution-free learning of probabilistic concepts. In: Proceedings of the 31st Annual Symposium on Foundations of Computer Science. SFCS '90, pp. 382–3911. IEEE Computer Society, USA (1990). https://doi.org/10.1109/FSCS.1990.89557 . https://doi.org/10.1109/FSCS.1990.89557

[30] Jae, J., Baek, K., Ryu, J., Lee, J.: Necessary and sufficient condition for joint measurability. Physical Review A **100**(3) (2019) https://doi.org/10.1103/physreva.100.032113

[31] Vovk, V., Papadopoulos, H., Gammerman, A.: Measures of Complexity: Festschrift for Alexey Chervonenkis. Springer, New York, NY (2015). https://doi.org/10.1007/978-3-319-21852-6

[32] Mendelson, S.: On the size of convex hulls of small sets. J. Mach. Learn. Res. **2**, 1–18 (2002) https://doi.org/10.1162/153244302760185225

[33] Cong, I., Choi, S., Lukin, M.: Quantum convolutional neural networks. Nature Physics **15**, 1–6 (2019) https://doi.org/10.1038/s41567-019-0648-8

[34] Szarek, S.J.: Metric Entropy of Homogeneous Spaces. arXiv (1997). https://doi.

org/10.48550/ARXIV.MATH/9701213 . https://arxiv.org/abs/math/9701213

Appendix A Supplementary material: proofs

A.1 Proof of Theorem 1

Consider $\mathcal{X} = \{0,1\}^2, \mathcal{Y} = \{0,1\}$, so that each $x \in \mathcal{X}$ can be written as (x_1, x_2). Furthermore, consider the following hypothesis classes $\widehat{\text{Hyp}}$ and Hyp: fix some small enough $\alpha > 0$, and let $\widehat{\text{Hyp}}$ consist of probabilistically observable concepts that, on input $x \in \mathcal{X}$, pass x_1 through a binary symmetric channel with crossover probability $0.1 + z$, where $z \in [-\alpha, \alpha]$. We then define Hyp to consist of $\widehat{\text{Hyp}}$, with one additional hypothesis: h_*, which, on input x, outputs a Bernoulli(1/2) random variable with probability 0.99 and outputs x_2 with probability 0.01.

One can think of $\widehat{\text{Hyp}}$ as a relaxation of the very simple deterministic hypothesis class Hyp' consisting of a single hypothesis: $f(x) = x_1$. Of course, Hyp' is agnostic-PAC learnable.

The theorem statement consists of the following claims:

1. The uniform convergence property for ERM fails to hold for the hypothesis class $\widehat{\text{Hyp}}$, even for distributions on $\mathcal{X} \times \mathcal{Y}$ for which ERM is PAC.
2. The hypothesis class Hyp is PAC learnable, but there exist distributions for which, simultaneously, the uniform convergence property fails to hold for ERM and ERM is not PAC.

Proof of claim 1: To show that the class $\widehat{\text{Hyp}}$ does not satisfy the uniform convergence property, we will exhibit a data-generating distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ for which, with non-negligible probability, there exist hypotheses in $\widehat{\text{Hyp}}$ whose empirical risks are bounded away from their true risks. In particular, let us consider a uniform distribution on $\mathcal{X}$, and a target function $f(x) = x_2$. Call the resulting joint distribution $\mathcal{D}$. Note that the expected risk $\mathbb{E}[R(h, f)] = 1/2$ for all $h \in \widehat{\text{Hyp}}$. In particular, this implies that we can trivially find a hypothesis whose expected risk is arbitrarily close to the minimum possible with probability exactly 1, so this distribution does not pose any fundamental difficulties from a learning perspective. Next, we show that for every m, with probability 1, there exists some hypothesis $h \in \widehat{Hyp}$ whose empirical risk on a dataset of length m is bounded away from its expected risk. Fix m samples $S = (x^{(1)}, ..., x^{(m)})$ drawn iid from $\mathcal{D}$. We claim that with probability 1, the set of *outputs* of all hypotheses in $\widehat{\text{Hyp}}$ on input S has cardinality 2^m, so that it is the set of bit strings of length m. To prove this, we upper bound the probability of the negation of this event: let B be the event that there exists some bit string that is *not* the output of any hypothesis in $\widehat{\text{Hyp}}$:

$$\mathrm{P}[B] = \mathrm{P}[\bigcup y \in \{0,1\}^m \bigcap_{h \in \widehat{\text{Hyp}}} [h(S) \neq y]] \leq \sum_{y \in \{0,1\}^m} \mathrm{P}[\bigcap_{h \in \widehat{\text{Hyp}}} [h(S) \neq y]]. \tag{A1}$$

Each term of the remaining sum can be computed by conditioning on the value of S. That is,

$$\mathrm{P}[\bigcap_{h\in\widehat{\mathrm{Hyp}}}[h(S)\neq y]] = \mathbb{E}[\mathrm{P}[\bigcap_{h\in\widehat{\mathrm{Hyp}}}[h(S)\neq y] \mid S]] = \mathbb{E}[\prod_{h\in\widehat{\mathrm{Hyp}}}\mathrm{P}[h(S)\neq y \mid S]] \quad \text{(A2)}$$

$$= \frac{1}{2^{2m}}\sum_{\hat{S}\in\{0,1\}^{2\times m}}\prod_{h\in\widehat{\mathrm{Hyp}}}\mathrm{P}[h(S)\neq y \mid S=\hat{S}]. \quad \text{(A3)}$$

Now, note that $\mathrm{P}[h(S)\neq y \mid S=\hat{S}] \leq c < 1$, by our choice of hypothesis class. This implies that

$$\mathrm{P}[\bigcap_{h\in\widehat{\mathrm{Hyp}}}[h(S)\neq y]] = \frac{1}{2^{2m}}\cdot 2^{2m}\prod_{h\in\widehat{\mathrm{Hyp}}}c \leq \prod_{h\in\widehat{\mathrm{Hyp}}}c = 0, \quad \text{(A4)}$$

where the last equality is by the fact that $\widehat{\mathrm{Hyp}}$ has infinite cardinality. This implies that

$$\mathrm{P}[B] \leq \sum_{y\in\{0,1\}^m} 0 = 0. \quad \text{(A5)}$$

Thus, with probability 1, the set of outputs of all hypotheses in $\widehat{\mathrm{Hyp}}$ in input S has cardinality 2^m. Now, this means that with probability 1, there exists a hypothesis h with exactly 0 misclassification error on S, so that its empirical risk is 0, while its expected risk is 1/2.

Thus, **the uniform convergence property does not hold for** $\widehat{\mathrm{Hyp}}$, and we have demonstrated this with a distribution on which empirical risk minimization trivially outputs a good hypothesis. This implies that uniform convergence is not necessary for ERM to be PAC. This completes the proof of Claim 1.

Proof of Claim 2: To show that there exist data-generating distributions for which ERM is not PAC for Hyp and the uniform convergence property fails to hold, we will consider the same input distribution and target function as in the proof of Claim 1: then the expected risk of h_* is as follows:

$$\mathbb{E}[R(h_*, f)] = \mathrm{P}[h_*(x)\neq f(x)] = \mathrm{P}[h_*(x)\neq x_2] = 0.99\cdot 1/2. \quad \text{(A6)}$$

That is, h_* is the unique hypothesis in Hyp with minimum expected risk for this distribution. By the analysis of $\widehat{\mathrm{Hyp}}$ above, though, ERM fails to return h_* on this hypothesis class asymptotically almost surely as the number of samples tends to ∞, because, with probability 1 over the choice of samples, there exists some *other* hypothesis that has empirical risk 0. Thus, ERM fails in this case – it returns a hypothesis with expected risk strictly bounded away from the minimum possible with probability 1.

To complete the proof of the claim, and, hence, the proof of Theorem 1, we need to show that Hyp is PAC learnable. This can be done in a variety of ways. We consider the learning rule that works as follows: on input $S = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, we perform the following operations:

1. Estimate the joint distribution $\mathcal{D}$ empirically:

$$\hat{p}(x, y) = \frac{|\{j \ : \ (x_j, y_j) = (x, y)\}|}{m}. \tag{A7}$$

2. Using the estimate $\hat{p}$ of $\mathcal{D}$, choose a hypothesis that minimizes the expected risk, where the expectation is computed according to $\hat{p}$ instead of $\mathcal{D}$. I.e., output a hypothesis $\hat{h} \in \mathrm{Hyp}$ such that

$$\hat{h} = \arg\min_{h \in \mathrm{Hyp}} \mathbb{E}_{(X,Y)\sim\hat{p}}[\ell(h[X], Y)]. \tag{A8}$$

To show that this learning rule is PAC, we note that by the strong law of large numbers and the fact that $|\mathcal{X} \times \mathcal{Y}|$ is finite, m can be chosen sufficiently large so that the total variation distance between $\hat{p}$ and $\mathcal{D}$ is less than ϵ with probability arbitrarily close to 1. This immediately implies that, for any $h \in \mathrm{Hyp}$, simultaneously,

$$|\mathbb{E}_{(X,Y)\sim\hat{p}}[\ell(h[X], Y)] - R(h)| < \epsilon. \tag{A9}$$

This implies the desired property. With the proofs of Claims 1 and 2 completed, the proof of Theorem 1 is complete.

A.2 Proof of Theorem 2

We first show that Hyp is compact in the topology generated by the L_1 operator norm on operators on $\mathcal{H}$ (here we note that each hypothesis in Hyp is given by $(\Pi_0, I - \Pi_0)$, so that Hyp may be equated with the set of operators Π_0 yielding the 0 outcome). Because the set of operators on $\mathcal{H}$ is finite-dimensional (since $\mathcal{H}$ itself was assumed to be so), all norms on them are equivalent, in the sense that they generate the same topology. Furthermore, compactness is equivalent to Hyp being closed and bounded. We have assumed that Hyp is closed. To show that it is bounded, we note that for any $(\Pi_0, I - \Pi_1) \in \mathrm{Hyp}$, we can write Π_0 as a convex combination of orthogonal projections:

$$\Pi_0 = \sum_{j=1}^{\dim(\mathcal{H})} a_j \, |v_j\rangle\langle v_j| \,, \tag{A10}$$

where $a_j \in [0, 1]$. Then a loose bound on the L_1 operator norm of Π_0 is given by

$$||\Pi_0||_{op,L_1} \leq \sum_{j=1}^{\dim(\mathcal{H})} a_j \leq \dim(\mathcal{H}). \tag{A11}$$

This implies boundedness of Hyp, which implies that Hyp is compact in the topology generated by the L_1 operator norm.

Compactness, in turn, implies that Hyp must have at least one accumulation point. Let us assume that Hyp has an accumulation point Π_* that is non-deterministic. Then we may use the argument from the proof of Theorem 1 to show that uniform convergence fails and ERM is not PAC. In particular, there exists an open neighborhood of Π_* containing infinitely many non-deterministic elements in Hyp. This was the key property used in Theorem 1, which allows us to conclude that ERM is not PAC for Hyp, and it does not satisfy the uniform convergence property.

A.3 Proof of Theorem 3

The proof of this theorem consists of the following steps:

1. We introduce a statistical distance between POVMs – the total variation distance d_{TV} between them. This distance has the property that any two hypotheses within distance γ of each other have expected risks within γ of each other. Throughout, we choose $\gamma = \epsilon/4$.
2. We show that for every γ, the d_{TV} γ-covering number of a finite-dimensional POVM class is finite: i.e., it can be covered by finitely many d_{TV} balls of radius less than γ.
3. Using the finiteness of covering numbers, we define the smoothing of the hypothesis class by a given γ-covering, which is a hypothesis class consisting of the centers of the balls in the covering. This class is necessarily finite-cardinality.
4. By previous results in the literature [1], the smoothed class is agnostically PAC learnable because it is finite-cardinality. The output of a $(\epsilon/4, \delta)$-PAC learning rule on this hypothesis class has true risk within $\epsilon/2$ of the minimum possible within the smoothed class. This minimum has, by our result on the total variation metric, a true risk that is within $\epsilon/2$ of the infimum of possible true risks in the original hypothesis class. Thus, the hypothesis returned by the learning rule on the smoothed class has true risk within ϵ of the infimum for the original class, with probability at least $1 - \delta$.

We now give the details of the above steps.

Step 1: Defining d_{TV} between POVMs

The definition is given in Definition 7.

We next state and prove the lemma connecting d_{TV} with the expected risks of the two hypotheses.

Lemma 1 (Connecting d_{TV} with expected risks)**.** *Let* $\Pi_1, \Pi_2 \in \mathrm{Hyp}$. *Then*

$$|R(\Pi_1) - R(\Pi_2)| \leq 2 d_{TV}(\Pi_1, \Pi_2). \tag{A12}$$

Proof. We have

$$\begin{aligned} &|R(\Pi_1) - R(\Pi_2)| && \text{(A13)} \\ &= |\mathrm{P}_{(X,Y)\sim\mathcal{D}}[\mathrm{Out}(\Pi_1, X) \neq Y] - \mathrm{P}_{(X,Y)\sim\mathcal{D}}[\mathrm{Out}(\Pi_2, X) \neq Y]| && \text{(A14)} \\ &= \mathbb{E}[|\mathrm{P}[\mathrm{Out}(\Pi_1, X) \neq Y \mid X, Y] - \mathrm{P}[\mathrm{Out}(\Pi_2, X) \neq Y \mid X, Y]|] && \text{(A15)} \end{aligned}$$

$$= \mathbb{E}[\sum_{b=0}^{1} |\mathrm{P}[\mathrm{Out}(\Pi_1, X) = b \mid X] - \mathrm{P}[\mathrm{Out}(\Pi_2, X) = b \mid X]| \cdot \mathrm{P}[Y = b]] \tag{A16}$$

$$\leq \mathbb{E}_X[\sum_{b=0}^{1} |\mathrm{P}[\mathrm{Out}(\Pi_1, X) = b \mid X] - \mathrm{P}[\mathrm{Out}(\Pi_2, X) = b \mid X]|] \tag{A17}$$

$$\leq \sup_{x \in \mathcal{X}} \sum_{b=0}^{1} |\mathrm{P}[\mathrm{Out}(\Pi_1, X) = b \mid X] - \mathrm{P}[\mathrm{Out}(\Pi_2, X) = b \mid X] \tag{A18}$$

$$= 2d_{TV}(\Pi_1, \Pi_2) \tag{A19}$$

The conditioning on Y disappears in equation (A16) because the event that $\mathrm{Out}(\Pi_j, X) = b$ is independent of Y given X. The first inequality is by upper bounding $\mathrm{P}[Y = b]$ by 1. □

Step 2: Finiteness of the d_{TV} covering numbers of Hyp

To show that the d_{TV} covering numbers of Hyp are finite, it is sufficient to show an upper bound on the corresponding packing numbers. This is the content of the next lemma, Lemma 2.

Lemma 2 (Finiteness of the total variation packing)**.** *Let* Hyp *be a finite-dimensional POVM class with outcomes in* $\{0, 1\}$*. For every* $\gamma > 0$*, the* γ*-*d_{TV}*-packing number of* Hyp *is finite.*

Proof. Since the Hilbert space $\mathcal{H}$ on which the operators in Hyp are defined is finite-dimensional, we take an orthonormal basis $|v_1\rangle, ..., |v_d\rangle$ for $\mathcal{H}$. We can uniquely encode each $h \in$ Hyp by its outcome-0 operator, which, in turn, can be represented by its $d \times d$ matrix over $\mathbb{C}$ with respect to the chosen basis. Thus, we can identify Hyp with a bounded subset $S \subset \mathbb{C}^{d \times d}$ (since we already know, from the proof of Theorem 2, that finite dimensionality of the POVM class (even if it is not closed) is sufficient to conclude boundedness of Hyp in the L_1 operator norm). Bounded subsets of finite-dimensional Euclidean spaces are known to have finite packing and covering numbers in every norm. More specifically, letting $\|\cdot\|$ be an arbitrary norm on $\mathbb{C}^{d \times d}$, since all norms on $\mathbb{C}^{d \times d}$ are topologically equivalent, S may be covered by a single sufficiently large $\|\cdot\|$-ball B, with diameter $\mathrm{diam}(S, \|\cdot\|)$. Then a covering of this ball is a covering of the set S. This implies that the $\|\cdot\|$-covering numbers of S are upper bounded by the $\|\cdot\|$-covering numbers of B. It is known that, for any $\epsilon > 0$, the ϵ-covering number of a ball in any norm is upper bounded by

$$\left(1 + \frac{2\mathrm{diam}(S, \|\cdot\|)}{\epsilon}\right)^{d^2}, \tag{A20}$$

which is finite since S is bounded. This implies finite d_{TV}-packing and covering numbers for Hyp.

□

Step 3: Constructing the TV-smoothed version of a hypothesis class

We next define the following *TV-smoothed* hypothesis class.

Definition 14 (TV-Smoothed version of a POVM class)**.** *Let* Hyp *be a finite-dimensional POVM class with outcomes* 0 *and* 1*. Let* $\mathcal{P}$ *be a finite* γ*-TV covering of* Hyp *by balls with centers* $\{\Pi^{(j)}\}_{j=1}^{|\mathcal{P}|}$*. We define the* γ*-smoothed version of* Hyp *to be simply the set of centers. We denote this by* $\mathbb{S}_{TV}(\mathrm{Hyp}, \mathcal{P})$*.*

Step 4: Learnability of the TV-smoothed class implies learnability of the original class

For any γ, the γ-TV-smoothing of a hypothesis class Hyp has the property that its infimum expected loss is less than γ away from that of the infimum expected loss of Hyp. This is the content of the next lemma, Lemma 3.

Lemma 3 (Infimal expected loss of $\mathbb{S}_{TV}(\mathrm{Hyp}, \mathcal{P})$)**.** *Let* $\widehat{\mathrm{Hyp}} = \mathbb{S}_{TV}(\mathrm{Hyp}, \mathcal{P})$*, where* $\mathcal{P}$ *is a finite* γ*-TV covering of* Hyp*. Then*

$$\inf_{h \in \mathrm{Hyp}} R(h) \leq \inf_{h \in \widehat{\mathrm{Hyp}}} R(h) \leq \inf_{h \in \mathrm{Hyp}} R(h) + 2\gamma. \tag{A21}$$

Proof. For any hypothesis $h_* \in \widehat{\mathrm{Hyp}}$ and h in the element of $\mathcal{P}$ corresponding to h_*, we have that $d_{TV}(h_*, h) \leq \gamma$. This implies, by Lemma 1, that

$$|R(h_*) - R(h)| \leq 2\gamma, \tag{A22}$$

which implies the desired result. □

Lemma 3 implies that if $\widehat{\mathrm{Hyp}}$ is $(\epsilon/2, \delta)$-PAC learnable, then with probability at least $1 - \delta$, we can find a hypothesis $h_* \in \widehat{\mathrm{Hyp}} \subseteq \mathrm{Hyp}$ such that $R(h_*) \leq \inf_{h \in \widehat{\mathrm{Hyp}}} R(h) + \epsilon/2 \leq \inf_{h \in \mathrm{Hyp}} R(h) + \epsilon/2 + 2\gamma = \epsilon$ (recalling that we set $\gamma = \epsilon/4$), which implies that Hyp is (ϵ, δ)-PAC learnable. In fact, since $\widehat{\mathrm{Hyp}}$ is finite, it is PAC learnable for every $\epsilon, \delta \to 0$, by the results in [1]. This completes the proof of learnability. To provide a more specific sample complexity bound, we recall the following result of [1].

Theorem 6 ([1], Theorem 2)**.** *Any finite POVM class* Hyp *is agnostic* (ϵ, δ)*-PAC-learnable with sample complexity bounded by*

$$n_{\mathrm{Hyp}}(\epsilon, \delta) = \min_{\mathcal{P} = (\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{|\mathcal{P}|})} \sum_{r=1}^{|\mathcal{P}|} \frac{8}{\epsilon^2} \log \frac{2|\mathcal{P}||\mathcal{P}_r|}{\delta}, \tag{A23}$$

where $\mathcal{P}$ *ranges over all possible joint measurability partitions of* Hyp*.*

In our case, we will apply this bound to $\mathbb{S}_{TV}(\mathrm{Hyp}, \mathcal{P})$, and hence to Hyp itself. In general, it can be difficult to find a minimal jointly measurable partition, and so we state the following worst-case bound, which we get from the trivial joint measurability partition whose elements are all singletons. We let N be the γ-TV-covering number of Hyp.

$$n_{\mathrm{Hyp}}(\epsilon,\delta) \le n_{\mathbb{S}(\mathrm{Hyp},\mathcal{P})}(\epsilon,\delta) \le \sum_{r=1}^{N} \frac{8}{\epsilon^2}\log\frac{2N}{\delta} = \frac{8N}{\epsilon^2}\log\frac{2N}{\delta}. \tag{A24}$$

This completes the proof of Theorem 3.

A.4 Proof of Theorem 4

A.4.1 Proof that learnable implies finite fat-shattering dimension

We first prove the necessary condition. The chain of logic is as follows:

1. We first prove that if Hyp is PAC learnable, then the corresponding POCC class is PAC learnable. We do this by a reduction.
2. We then show that if a POCC class is PAC learnable, then the corresponding p-concept class is PAC learnable, again by a reduction.
3. It is known that a p-concept class is PAC learnable if and only if, for every γ, its γ-fat shattering dimension is finite. Since all fat-shattering dimensions are the same in our chain of reductions, this implies that the γ-fat shattering dimension of Hyp must be finite.

Step 1: POVM learnability implies POCC class learnability

We recall that a POVM class induces in a natural way a corresponding POCC class. Namely, every POVM induces a unique conditional distribution on outcomes when used to measure a given state. Thus, each POVM induces a function from states to probability distributions on outcomes, which define probabilistically observed concepts.

Lemma 4. *Suppose that* Hyp *is a PAC learnable POVM class. Then it is PAC learnable as a POCC class.*

Proof. Let A be an (ϵ,δ)-PAC learning rule for Hyp. We may use exactly the same learning rule in the POCC framework, and it yields the same guarantees. □

Step 2: POCC class learnability implies p-concept class learnability

We recall that every POCC class has an associated p-concept class, trivially.

Lemma 5. *Suppose that* Hyp *is a PAC learnable POCC class. Then it is PAC learnable as a p-concept class.*

Proof. Let Hyp be a POCC class, and let $\widehat{\mathrm{Hyp}}$ denote the corresponding p-concept class. Note that these have exactly the same γ-fat-shattering dimension, for every γ. We will show that if Hyp is (ϵ,δ)-PAC learnable, then so is $\widehat{\mathrm{Hyp}}$, with the L_1 risk.

Let A be a learning rule for Hyp. Then it is also a learning rule for $\widehat{\mathrm{Hyp}}$, and the risks of all hypotheses in Hyp are the same as those in $\widehat{\mathrm{Hyp}}$, as we show next. In particular, the misclassification risk for $h \in \mathrm{Hyp}$ is given by

$$R_{\mathrm{Hyp}}(h) = \mathrm{P}[h[X] \neq Y] \tag{A25}$$

$$= \mathrm{P}[h[X] = 1, Y = 0] + \mathrm{P}[h[X] = 0, Y = 1] \tag{A26}$$
$$= \mathbb{E}[\mathrm{P}[h[X] = 1, Y = 0 \mid X] + \mathrm{P}[h[X] = 0, Y = 1 \mid X]] \tag{A27}$$
$$= \mathbb{E}[\mathrm{P}[h[X] = 1 \mid X]\mathrm{P}[Y = 0 \mid X] + \mathrm{P}[h[X] = 0 \mid X]\mathrm{P}[Y = 1 \mid X]] \tag{A28}$$
$$= \mathbb{E}[h(X)(1 - \mathrm{P}[Y = 1 \mid X]) + (1 - h(X))\mathrm{P}[Y = 1 \mid X]] \tag{A29}$$
$$= \mathbb{E}[h(X) + \mathrm{P}[Y = 1 \mid X] - 2\mathrm{P}[Y = 1 \mid X]h(X)]. \tag{A30}$$

Meanwhile, the L_1 risk for $h \in \widehat{\mathrm{Hyp}}$ is

$$R_{\widehat{\mathrm{Hyp}}}(h) = \mathbb{E}[|h(X) - Y|] \tag{A31}$$
$$= \mathbb{E}[\mathbb{E}[|h(X) - Y| \mid X]] \tag{A32}$$
$$= \mathbb{E}[\mathrm{P}[Y = 1 \mid X](1 - h(X)) + (1 - \mathrm{P}[Y = 1 \mid X])h(X)] \tag{A33}$$
$$= \mathbb{E}[h(X) + \mathrm{P}[Y = 1 \mid X] - 2\mathrm{P}[Y = 1 \mid X]h(X)] \tag{A34}$$
$$= R_{\mathrm{Hyp}}(h). \tag{A35}$$

This implies that Hyp being PAC learnable implies that $\widehat{\mathrm{Hyp}}$ is PAC learnable with the same parameters. □

Step 3: p-concept class learnability only if fat shattering dimension is finite

It is shown in [28] that p-concept classes are learnable with respect to the L_1 risk only if they have finite γ-fat-shattering dimension for every $\gamma > 0$. By steps 1 and 2 of our proof, this implies that if a POVM class Hyp is PAC learnable, then it has finite γ-fat-shattering dimension for every γ, which completes the proof of the necessary condition of the theorem.

A.4.2 Proof that finite fat-shattering dimension and finite partitionability implies learnable

We now show that the stated sufficient conditions imply learnability of a POVM class. The proof outline when Hyp is jointly measurable, in which case all that is needed is finite fat shattering dimension, is as follows.

1. In the preliminaries, we show that the expected value of the denoised empirical risk of a POVM is its expected risk (Lemma 6 below). We also define an appropriate generalization of the Rademacher complexity of Hyp.
2. We define $\Phi_{\mathrm{Hyp}}(S)$, for a training set S, to be the supremum, over all hypotheses $h \in \mathrm{Hyp}$, of the deviation of the denoised empirical risk of h from the expected risk. We use McDiarmid's inequality to show that $\Phi_{\mathrm{Hyp}}(S)$ is well-concentrated around its mean.
3. We upper bound the mean of $\Phi_{\mathrm{Hyp}}(S)$ in terms of the Rademacher complexity of Hyp. It is in this step that we use Lemma 6.
4. We apply known bounds on the Rademacher complexity in terms of covering numbers and then a known bound on the covering numbers in terms of the fat shattering

dimension. This implies an upper bound on the sample complexity of PAC learning Hyp using DERM, which implies learnability for jointly measurable hypothesis classes with finite fat shattering dimension.

To establish that the conditions given in the theorem statement – namely, that Hyp has finite fat-shattering dimension and that for every $\alpha > 0$, there exists a finite α-approximately jointly measurable partition $\mathcal{P}$ of Hyp – are sufficient, we reason as follows:

1. We note that the joint measurability smoothing of a hypothesis class is a jointly measurable class. Thus, inside the loop of DERM, we are computing the denoised empirical risk $\hat{R}_j$ of a jointly measurable class and the denoised empirical risk minimizer $\hat{\rho}_j$ of that class.
2. By the reasoning for the jointly measurable class case earlier in this proof, with probability at least $1 - \delta/|\mathcal{P}|$, $\hat{R}_j$ is within $\epsilon/2$ of the minimal true risk inside the smoothed partition element $\hat{P}_j$, provided that $|\hat{S}|_j$ is chosen to be sufficiently large as a function of ϵ, δ, and the number of partition elements $|\mathcal{P}|$. Taking a union bound over all partition elements ensures that with probability at least $1 - \delta$, this holds for *every* partition element.
3. We show that the γ-joint measurability smoothing of each partition element P_j results in hypotheses whose risks are at most 2γ away from the risks of the corresponding unsmoothed hypotheses. This is the content of Lemma 11. Taking $\gamma = \epsilon/4$, we get that the resulting hypothesis h_* returned by the learning rule is at most $2\gamma + \epsilon/2 = \epsilon$ away from the infimal risk of the hypothesis class Hyp, with probability at least $1 - \delta$.

Step 1: Preliminaries for upper bounding sample complexity for jointly measurable Hyp

Our first lemma says that the expected value of the denoised empirical risk of a POVM hypothesis is the true risk.

Lemma 6 (Expected value of the denoised empirical risk)**.** *Let* Hyp *be a POVM class, and let* $h \in$ Hyp. *We have the following identity:*

$$\mathbb{E}[\mathrm{DER}(h, S)] = R(h). \tag{A36}$$

Proof. This is a result of the tower property of conditional expectation:

$$\mathbb{E}[\mathrm{DER}(h, S)] = \mathbb{E}[\mathbb{E}[\frac{1}{m}\sum_{j=1}^{m} \ell(h[X_j], Y_j) \mid Y, Z]] = \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}[\mathbb{E}[\ell(h[X_j], Y_j) \mid Y_j, Z]] \tag{A37}$$

$$= \frac{1}{m}\sum_{j=1}^{m} R(h) = R(h). \tag{A38}$$

More precisely, the first equality uses the definition of $\mathrm{DER}(h, S)$. The second equality uses linearity of expectation and the independence of (X_j, Y_j, Z_j) on (X_i, Y_i, Z_i) for

$j \neq i$. The third equality is where we use the tower property of conditional expectation: in general, for random variables A, B on a common probability space, $\mathbb{E}[\mathbb{E}[A \mid B]] = \mathbb{E}[A]$. The third equality also invokes the definition of the risk $R(h)$. □

We next define an analogue of the Rademacher complexity [7] of a POVM class.

Definition 15 (Rademacher complexity of a jointly measurable POVM class)**.** *Let* Hyp *be a POVM class consisting of jointly measurable POVMs with root POVM ρ, and let $\mathcal{D}$ be some distribution on $\mathcal{X} \times \mathcal{Y}$.*

We define the mth Rademacher complexity (with respect to the specified fine graining of which ρ is the root) of Hyp *to be*

$$\mathfrak{R}_m(\mathrm{Hyp}) = \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{h\in\mathrm{Hyp}} \sum_{j=1}^{m} \sigma_j \zeta(S_j)] \tag{A39}$$

where $S \sim \mathcal{D}^m$, S_j denotes (X_j, Y_j), and $\zeta(S_j)$ denotes the random variable $\mathbb{E}[\ell(h[X_j], Y_j) \mid Y_j, Z_j]$, and Z_j is the random outcome of measuring X_j with ρ.

Step 2: Defining $\Phi_{\mathrm{Hyp}}(S)$ – the point generalization gap, and applying McDiarmid

We define the following function, called the *point generalization gap* for DERM of Hyp on the dataset $S \in (\mathcal{X} \times \mathcal{Y})^m$:

$$\Phi_H(S) = \Phi_{\mathrm{Hyp}}(S) = \sup_{h\in\mathrm{Hyp}} (\mathrm{DER}(h, S) - R(h)). \tag{A40}$$

Next, we apply McDiarmid's inequality to show that $\Phi(S)$ is close to its expected value with high probability. To do this, we need to upper bound the maximum possible value of $|\Phi(S)-\Phi(\hat{S})|$, where $\hat{S}$ differs from S in exactly one coordinate (say, coordinate j). We have

$$\begin{aligned}
&|\Phi(S) - \Phi(\hat{S})| && \text{(A41)}\\
&= |\sup_{h\in\mathrm{Hyp}} (\mathrm{DER}(h, S) - R(h)) - \sup_{h\in\mathrm{Hyp}} (\mathrm{DER}(h, \hat{S}) - R(h))| && \text{(A42)}\\
&\leq |\sup_{h\in\mathrm{Hyp}} (\mathrm{DER}(h, S) - \mathrm{DER}(h, \hat{S}))| && \text{(A43)}\\
&= \frac{1}{m}| \sup_{h\in\mathrm{Hyp}} \sum_{k\neq j} \mathbb{E}[\ell(h[x_k], y_k) \mid (z_k, y_k)] - \mathbb{E}[\ell(h[x_k], y_k) \mid (z_k, y_k)] && \text{(A44)}\\
&\quad + \mathbb{E}[\ell(h[x_j], y_j) \mid (z_j, y_j)] - \mathbb{E}[\ell(h[\hat{x}_j], \hat{y}_j) \mid (\hat{z}_j, \hat{y}_j)]| && \text{(A45)}\\
&= \frac{1}{m}| \sup_{h\in\mathrm{Hyp}} |\mathbb{E}[\ell(h[x_j], y_j) \mid (z_j, y_j)] - \mathbb{E}[\ell(h[\hat{x}_j], \hat{y}_j) \mid (\hat{z}_j, \hat{y}_j)]| && \text{(A46)}\\
&\leq 1/m. && \text{(A47)}
\end{aligned}$$

The first inequality is because the difference between suprema is less than or equal to the supremum of the difference. The final inequality is because the loss is bounded between 0 and 1. This is essentially exactly the same as in the classical case.

Applying McDiarmid's inequality, we then get the following intermediate bound.

Lemma 7 (Application of McDiarmid's inequality to $\Phi(S)$)**.** *For every $\gamma > 0$, we have*

$$\mathrm{P}[|\Phi(S) - \mathbb{E}_S[\Phi(S)]| \geq \gamma] \leq \exp\left(-2m\gamma^2\right). \tag{A48}$$

Since we want this probability to be at most δ, we choose γ such that

$$e^{-2m\gamma^2} = \delta, \tag{A49}$$

which implies

$$\log\delta = -2m\gamma^2 \implies \gamma = \sqrt{\frac{\log(1/\delta)}{2m}}. \tag{A50}$$

That is, with probability at least $1-\delta$, we have

$$\Phi(S) \leq \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\log(1/\delta)}{2m}}. \tag{A51}$$

Step 3: Upper bounding $\mathbb{E}[\Phi(S)]$ via Rademacher complexity

Next, we need an upper bound on the expected value of $\Phi(S)$ in terms of the Rademacher complexity defined in Definition 15.

Lemma 8 (Rademacher complexity upper bound)**.** *We have the following:*

$$\mathbb{E}[\Phi_{\mathrm{Hyp}}(S)] \leq 2\mathfrak{R}(\mathrm{Hyp}). \tag{A52}$$

Proof. We follow the pattern as in classical statistical learning theory: we first rewrite $R(h)$ in $\mathbb{E}[\Phi(S)]$ as the expected value of an empirical expectation.

$$\begin{aligned}\mathbb{E}_S[\Phi(S)] &= \mathbb{E}[\sup_{h\in\mathrm{Hyp}}(\mathrm{DER}(h,S) - R(h))] &\text{(A53)}\\ &= \mathbb{E}[\sup_{h\in\mathrm{Hyp}}(\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\ell(h[X_j],Y_j) \mid Y_j, Z_j] - R(h))], &\text{(A54)}\end{aligned}$$

where we recall that Z_j is the random outcome of measuring X_j with the root measurement ρ.

We rewrite $R(h)$ as

$$R(h) = \mathbb{E}_{\hat{S}}[\mathrm{DER}(h,\hat{S})], \tag{A55}$$

where $\hat{S}$ is independent and equal in distribution to S. Plugging this into (A54),

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_{S,\hat{S}}[\sup_{h\in\mathrm{Hyp}}(\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\ell(h[X_j],Y_j) \mid Y_j, Z_j] - \mathbb{E}[\ell(h[\hat{X}_j],\hat{X}_j) \mid \hat{Y}_j, \hat{Z}_j])]. \tag{A56}$$

The remaining steps can be carried out *exactly* as in the classical case, and we will end up with an upper bound by the Rademacher complexity of the set of denoised empirical risks of hypotheses in the class. Let $\zeta(S_j)$ denote the random variable $\mathbb{E}[\ell(h[X_j], Y_j) \mid Y_j, Z_j]$. Then the above can be more succinctly written as

$$\mathbb{E}_{S,\hat{S}}[\sup_{h\in\mathrm{Hyp}}(\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\ell(h[X_j], Y_j) \mid Y_j, Z_j] - \mathbb{E}[\ell(h[\hat{X}_j], \hat{Y}_j) \mid \hat{Y}_j, \hat{Z}_j])] \tag{A57}$$

$$= \frac{1}{m}\mathbb{E}_{S,\hat{S}}[\sup_{h\in\mathrm{Hyp}}\sum_{j=1}^{m}\zeta(S_j) - \zeta(\hat{S}_j)]. \tag{A58}$$

Introducing Rademacher random variables $\sigma_j \in \{-1, 1\}$, for $j \in [m]$, this is equal to

$$\frac{1}{m}\mathbb{E}_{S,\hat{S}}[\sup_{h\in\mathrm{Hyp}}\sum_{j=1}^{m}\zeta(S_j) - \zeta(\hat{S}_j)] = \frac{1}{m}\mathbb{E}_{S,\hat{S},\sigma}[\sup_{h\in\mathrm{Hyp}}\sum_{j=1}^{m}\sigma_j\zeta(S_j) - \sigma_j\zeta(\hat{S}_j)]. \tag{A59}$$

Note that this is because if two random variables $\zeta(S_j)$ and $\zeta(\hat{S}_j)$ are independent and identically distributed, their difference is equal in distribution to $\sigma_j \cdot (\zeta(S_j) - \zeta(\hat{S}_j))$. Finally, this is less than or equal to

$$\frac{1}{m}\mathbb{E}_{S,\hat{S},\sigma}[\sup_{h\in\mathrm{Hyp}}\sum_{j=1}^{m}\sigma_j\zeta(S_j) - \sigma_j\zeta(\hat{S}_j)] \leq \frac{2}{m}\mathbb{E}_{S,\hat{S},\sigma}[\sup_{h\in\mathrm{Hyp}}\sum_{j=1}^{m}\sigma_j\zeta(S_j)] = 2\mathfrak{R}_m(\mathrm{Hyp}). \tag{A60}$$

□

Step 4: Upper bounding the Rademacher complexity via covering numbers and fat shattering dimension

By Step 3, if we want to show that a hypothesis class is learnable, then we should show that the Rademacher complexity above is $o(1)$ as $m \to \infty$. We will show that the *fat-shattering dimension* is the relevant quantity, and that it is sufficient for this quantity to be finite. Lemmas 9 and 10 below establish the connection between the Rademacher complexity and the fat shattering dimension.

Definition 16 (Covering numbers $\mathcal{N}(\alpha, \mathcal{F}, m)$)**.** *Let $\mathcal{F}$ be a metric space with metric d. We say that a subset $\hat{\mathcal{F}} \subseteq \mathcal{F}$ is an α-covering of $\mathcal{F}$ with respect to d if for every $f \in \mathcal{F}$, there exists $g \in \hat{\mathcal{F}}$ such that $d(f, g) \leq \alpha$. The α-covering number of $\mathcal{F}$ with respect to d is defined to be the minimum cardinality of any α-covering of $\mathcal{F}$ and is denoted by $\mathcal{N}_d(\alpha, \mathcal{F})$. In place of d in the subscript, we may put a norm, in which case the relevant metric is the one induced by that norm.*

For a class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$, we define the following norm with respect to a set $S = (x_1, ..., x_m) \in \mathcal{X}^m$:

$$\|f - g\|_{\infty,S} = \sup_{j\in\{1,2,...,m\}} |f(x_j) - g(x_j)|. \tag{A61}$$

The covering numbers of $\mathcal{F}$ with respect to this norm on the set S are well-defined. We then define the covering number $\mathcal{N}(\alpha, \mathcal{F}, m) = \sup_{S \in \mathcal{X}^m} \mathcal{N}_{\|\cdot\|_{\infty,S}}(\alpha, \mathcal{F}, S)$.

Lemma 9 (Upper bound on Rademacher complexity via covering numbers [31]). *Let $\mathcal{F}$ be a function class. Then*

$$\mathfrak{R}_m(\mathcal{F}) \leq \inf_{\alpha} \left(\alpha + \sqrt{\frac{2\log(\mathcal{N}(\alpha, \mathcal{F}, m))}{m}} \right). \tag{A62}$$

Lemma 10 (Upper bound on covering numbers via fat-shattering dimension [31]). *Let $\mathcal{F}$ be a class of functions with range $[0,1]$. Then the following upper bound on the covering numbers holds: let $\alpha \geq 0$ and $d = \mathrm{fat}_{\alpha/4}(\mathcal{F})$. Then*

$$\mathcal{N}(\alpha, \mathcal{F}, m) \leq 2 \cdot \left(m \cdot (2/\alpha + 1)^2\right)^{\lceil d \log\left(\frac{2em}{d\alpha}\right)\rceil} \tag{A63}$$

Plugging the upper bound in Lemma 10 into the one in Lemma 9, and then plugging that into Lemma 8 and, finally, using Lemma 7 yields the following statement: with probability at least $1-\delta$,

$$\Phi(S) \leq 2\mathfrak{R}_m(\mathrm{Hyp}) \leq 2 \inf_{\alpha} \left(\alpha + \sqrt{\frac{2\log 2 + 2\lceil d \log\left(\frac{2em}{d\alpha}\right)\rceil \log(m(2/\alpha+1)^2)}{m}} \right), \tag{A64}$$

which can be upper bounded by setting $\alpha = \Theta(\frac{\log m}{\sqrt{m}})$. This results in the following bound:

$$\mathbb{E}_S[\Phi(S)] \leq O(\frac{\sqrt{d}\log m}{\sqrt{m}}) \implies \Phi(S) \leq O(\frac{\sqrt{d}\log m + \sqrt{\log(1/\delta)}}{\sqrt{m}}) \tag{A65}$$

This directly translates to a finite upper bound on the number of samples $n_{\mathrm{Hyp}}(\epsilon, \delta)$ required to achieve a risk within ϵ of the infimum, provided that the fat shattering dimension d is finite. Specifically,

$$n_{\mathrm{Hyp}}(\epsilon, \delta) \leq \inf_{\mathcal{P}} O(|\mathcal{P}|\frac{d + \log(1/\delta)}{\epsilon^2}), \tag{A66}$$

where $\mathcal{P}$ ranges over all finite $\epsilon/8$-almost-jointly-measurable partitions of Hyp.

This completes the proof of sufficiency of finite fat shattering dimension for learnability in the case where Hyp is a jointly measurable class.

Completing the proof of the theorem

It remains to prove the following lemma, which is the remaining detail for completing the argument in the case where Hyp can be partitioned into finitely many approximately jointly measurable classes:

Lemma 11 (Risk of joint measurability-smoothed hypothesis classes)**.** *Let* $\hat{\text{Hyp}} = \mathbb{S}_{JM}(\text{Hyp}, \mathcal{P})$*, where* $\mathcal{P}$ *is a finite* γ*-joint measurability smoothing of* Hyp*. Then*

$$\inf_{h \in \text{Hyp}} R(h) - 2\gamma \leq \inf_{h \in \hat{\text{Hyp}}} R(h) \leq \inf_{h \in \text{Hyp}} R(h) + 2\gamma. \tag{A67}$$

Proof. It suffices to prove the analogous chain of inequalities for any single hypothesis h and its smoothed version $\hat{h}$. For this, we use the data processing inequality for total variation distance: let (Π, α) and $(\hat{\Pi}, \alpha)$ be respective fine grainings of h and $\hat{h}$, noting that by definition of the smoothing operation that the two classical channels are the same. Then for any $x \in \mathcal{X}$,

$$d_{TV}(\alpha \circ \text{Out}(\Pi, x), \alpha \circ \text{Out}(\hat{\Pi}, x)) \leq d_{TV}(\text{Out}(\Pi, x), \text{Out}(\hat{\Pi}, x)) \leq \gamma. \tag{A68}$$

This implies, by Lemma 1, that $|R(h) - R(\hat{h})| \leq 2\gamma$, which implies the stated result. ☐

This completes the proof of the sufficient condition part of Theorem 4.

A.5 Proof of Theorem 5

We start by defining the packing numbers of a set S with respect to a metric.

Definition 17 (Packing numbers)**.** *Let d be a metric on a set S, and let $\gamma \geq 0$. We say that a subset $A \subseteq S$ is a γ-d-packing of S if for all $a, b \in A$, we have $d(a, b) > \gamma$.*

Then we define the γ-d-packing number of S to be the maximum cardinality of any γ-d-packing of S. We denote this number by $\mathcal{M}(\gamma, S, d)$.

Before stating and proving the main lemmas used in the proof of the theorem, we recall that a maximal γ-d-packing of a set S is also a γ-d-covering of S: otherwise, if there is some $x \in S$ that is not within γ of any packing element, then we could enlarge the packing by including x. So the cardinality of this packing is an upper bound on the γ-d-covering number of S, so that $\mathcal{N}(\gamma, S, d) \leq \mathcal{M}(\gamma, S, d)$.

Lemma 12 (Relating d_{TV} covering and packing numbers to $\mathcal{N}_{JM}(\gamma, S)$)**.** *Let S be a collection of POVMs We have*

$$\mathcal{N}_{JM}(\gamma, S) \leq \mathcal{N}(\gamma, S, d_{TV}) \leq \mathcal{M}(\gamma, S, d_{TV}). \tag{A69}$$

Proof. The second inequality is well-known, so we focus on the first. Let $\hat{S}$ be a γ-covering of S. We will show that $\hat{S}$ is also a γ-jm covering of S, which immediately implies the inequality.

Let Π be the center of one element of $\hat{S}$. We claim that the closed d_{TV} ball $B_{TV}(\Pi, \gamma)$ centered at Π with radius γ is γ-jointly measurable. We choose the root POVM to be Π itself. Now, for every $\Pi' \in B_{TV}(\Pi, \gamma)$, we consider the trivial fine-graining (Π', Id), where Id is the identity channel. Trivially, $d_{TV}(\Pi, \Pi') \leq \gamma$. This completes the proof of the claim. ☐

The next lemma relates the d_{TV}-packing number of a set of POVMs to the m-sample d_{TV}-covering number, which we recall is upper bounded by a function of the fat-shattering dimension.

Lemma 13 (Relating the d_{TV}-packing number to the m-sample d_{TV}-covering number)**.** *Let* Hyp *be a class of POVMs. Suppose that k is some number satisfying $k \leq \mathcal{M}(\gamma, \mathrm{Hyp}, d_{TV})$. If $m \geq \binom{\mathcal{M}(\gamma,\mathrm{Hyp},d_{TV})}{2}$, then*

$$\mathcal{M}(\gamma, \mathrm{Hyp}, d_{TV}) \leq \mathcal{N}(\gamma, \mathrm{Hyp}, m). \tag{A70}$$

Proof. To lower bound the m-sample covering number of Hyp by some number k, we must exhibit a set S of m points $\{\hat{\rho}_j\}_{j=1}^m$ such that $\mathcal{N}(\gamma, \mathrm{Hyp}, S) \geq k$.

An important consequence of Lemma 12 is that if there exists a γ-jm partition of Hyp with cardinality at least k, then there exist hypotheses $h_1, ..., h_k \in \mathrm{Hyp}$ such that for every $i \neq j \in [k]$, we have $d_{TV}(h_i, h_j) \geq \gamma$. This is a direct consequence of the packing number bound in the lemma.

The fact that $d_{TV}(h_i, h_j) \geq \gamma$ means that there exists a state $\rho_{i,j} \in \mathcal{X}$ such that $d_{TV}(h_i(\rho_{i,j}), h_j(\rho_{i,j})) \geq \gamma$. Thus, we choose our set S to be $\{\rho_{i,j}\}_{i \neq j \in [k]}$, along with an arbitrary collection of $m - \binom{k}{2}$ other states. It is then easily checked that the covering number $\mathcal{N}(\gamma, \mathrm{Hyp}, S) \geq k$, and we can set $k = \mathcal{M}(\gamma, \mathrm{Hyp}, d_{TV})$. This completes the proof. □

The proof of the theorem is then a direct application of Lemmas 12, 13, and 10.

Appendix B Definitions from quantum mechanics

We give below a brief introduction to relevant definitions and notation from quantum information. This is meant only to highlight the bare minimum necessary concepts for this paper. The reader is encouraged to consult [6] for more extensive discussions of quantum information. We note that the reader does not need any physics background at all, and the required mathematics is not beyond the training of most learning theorists.

To describe quantum states, we fix a Hilbert space $\mathcal{H}$ over the complex numbers $\mathbb{C}$. A *pure state* is a unit vector in $\mathcal{H}$, which, in the bra-ket notation of quantum mechanics, is denoted by $|v\rangle$. The dual space to $\mathcal{H}$ is the vector space of linear functionals $\langle v| : \mathcal{H} \to \mathbb{C}$, where $\langle v|w\rangle$ is defined to be the inner product of $|v\rangle$ with $|w\rangle$. It is frequently convenient to identify pure states $|v\rangle$ with their outer product forms $|v\rangle\langle v|$, which are operators from $\mathcal{H} \to \mathcal{H}$. The reason for this is the outer product forms fit nicely into the density matrix formalism, which we discuss next.

Mixed states (which we just call states in this paper) are more general: they are convex combinations of pure states and are also called *density matrices*. They capture statistics resulting from drawing pure states from a probability distribution. However, it should be noted that a single density matrix can arise from multiple distinct convex combinations of pure states.

A quantum measurement is specified by a positive operator-valued measure (POVM), defined as follows.

Definition 18 (POVM)**.** *A POVM with k outcomes, defined on a Hilbert space $\mathcal{H}$, is a k-tuple $\Pi = (\Pi_1, ..., \Pi_k)$ of positive semidefinite Hermitian operators on $\mathcal{H}$ that sum to the identity operator.*

We note that each operator Π_j, by virtue of being positive semidefinite Hermitian, has a decomposition as $\Pi_j = M_j^* M_j$, where M_j^* denotes the adjoint operator.

Measurement of a mixed state ρ by a POVM Π works as follows: it produces an *outcome* $\mathrm{Out}(\Pi, \rho)$ in $\{1, ..., k\}$, which is observed by the measurer and a post-measurement state ρ', which is not. The outcome is drawn from the following distribution:

$$\mathrm{P}[\mathrm{Out}(\Pi, \rho) = j] = \mathrm{Tr}\{\rho \Pi_j\}, \tag{B71}$$

where $\mathrm{Tr}\{\cdot\}$ denotes the trace.

The post-measurement state ρ' is dependent on $\mathrm{Out}(\Pi, \rho)$. If the outcome is j, then ρ' is given by

$$\rho' = \frac{M_j \rho M_j^*}{\mathrm{Tr}\{\Pi_j \rho\}}. \tag{B72}$$

These measurement rules are collectively called the *Born rule*. One can thus think of a POVM as a particular type of stochastic map from density matrices to ordered pairs whose first component is an outcome index and whose second component is a post-measurement density matrix. The particulars of the Born rule become important when one tries to define specific hypothesis classes and study their learning-theoretic measures of complexity (e.g., fat-shattering dimension). It is also worth emphasizing a few phenomena that differentiate the quantum learning setting from the classical case:

- Unknown states cannot be copied. That is, there is no general procedure that takes as input a register prepared in some state ρ and produces two registers, both in state ρ. Thus, for example, a learner cannot make a "backup copy" of a state in the training set.
- States cannot be directly observed by a learner. The only thing that can be observed is the outcome index of measurement of a state.

Appendix C Aspects of learning theory for those only familiar with quantum information

Here we describe the basics of classical statistical learning theory for an audience that may not be familiar with it. Our goal is to avoid common confusions, such as the distinction between state estimation and learning. This distinction is important in Section E.

In classical statistical learning theory, supervised learning is formulated as follows: a domain $\mathcal{X}$ and a co-domain $\mathcal{Y}$ (which we think of as the label set in a classification problem) are fixed and known to the learner. There is an unknown joint distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. A known hypothesis class Hyp consisting of deterministic functions $h : \mathcal{X} \to \mathcal{Y}$ is fixed. These hypotheses are meant to approximate the statistical association between inputs $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$. To measure the quality of approximation, a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y}$ is fixed, and the loss of a hypothesis on a pair in $\mathcal{X} \times \mathcal{Y}$ is defined by $\ell(h, x, y) = \ell(h(x), y)$. The *risk* of a hypothesis is its expected loss on a pair $(X, Y) \sim \mathcal{D}$: $R(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h, X, Y)]$.

The learner sees a training set consisting of independent samples from $\mathcal{D}$, and the goal of the learner is to choose a hypothesis h from Hyp with risk as close as possible to the worst risk of any hypothesis in the class. Formally, a hypothesis class is (ϵ, δ)-PAC learnable if there exists a learning rule (i.e., a function from datasets to Hyp) A and a number of samples $m(\epsilon, \delta)$ such that, for *every* distribution $\mathcal{D}$, A outputs a hypothesis h such that with probability at least $1 - \delta$,

$$R(h) \leq \inf_{h_* \in \mathrm{Hyp}} R(h_*) + \epsilon. \tag{C73}$$

Note that Hyp may be uncountably infinite, and so the infimum may not be achievable. The number of samples required for the risk bound (C73) to hold with probability $\geq 1 - \delta$ is the *sample complexity* of learning the hypothesis class.

We emphasize a few things about this:

- The above framework is *distribution-free*, in the sense that the number of samples and the learner must not depend a priori on any assumption about the form that $\mathcal{D}$ takes. However, the learner is assumed to have full knowledge of $\mathcal{X}$, $\mathcal{Y}$, $\ell(\cdot, \cdot, \cdot)$, and Hyp.
- Statistical learning theory does not deal with computational efficiency, as learning rules are not algorithms. Indeed, a hypothesis class may be PAC learnable but not efficiently so.
- The goal is to choose a hypothesis that captures the statistical association between X and Y as well as possible *compared to any other hypothesis in the class.* This is a distinct approach from estimating the distribution $\mathcal{D}$ of the data. The reason that the theory is formulated this way is that the problem of selecting a hypothesis from a well-designed class Hyp can have dramatically smaller sample complexity than that of estimating $\mathcal{D}$. This is a key difference between PAC learning and estimation.

Further philosophical grounding for statistical learning theory can be found in any of a number of textbooks on the subject (e.g., [7]).

C.1 Empirical risk minimization

The quintessential learning rule in classical learning theory is *empirical risk minimization.* Given a dataset $S = ((X_j, Y_j))_{j=1}^m$ and a hypothesis $h \in \mathrm{Hyp}$, the empirical risk of h is given by

$$\hat{R}(h, S) = \frac{1}{m} \sum_{j=1}^{m} \ell(h(X_j), Y_j). \tag{C74}$$

Then the empirical risk minimization (ERM) learning rule outputs the following:

$$h_* = \arg\min_{h \in \mathrm{Hyp}} \hat{R}(h, S). \tag{C75}$$

This is a central learning rule in the classical theory, as explained in Section C.2.

C.2 The fundamental theorem of concept learning

One of the fundamental results in classical statistical learning theory is the *fundamental theorem of concept learning*, sometimes called the fundamental theorem of PAC learning or of statistical learning (see [7], Theorem 6.7). It gives matching necessary and sufficient conditions for a hypothesis class to be learnable, under certain assumptions on the codomain $\mathcal{Y}$ and the loss function. Specifically, there is a combinatorial notion of complexity of the hypothesis class, known as the Vapnik-Chervonenkis (VC) dimension of Hyp. The fundamental theorem of concept learning relates the VC dimension of Hyp to its learnability. We summarize it below.

Theorem 7 (Fundamental theorem of concept learning [7]). *Let* Hyp *be a hypothesis class of functions from a domain* $\mathcal{X}$ *to* $\{0, 1\}$*, and let the loss function be the misclassification loss. Then the following are equivalent:*

1. *The ERM rule is a successful PAC learner for* Hyp*.*
2. Hyp *is PAC learnable.*
3. Hyp *has finite VC dimension.*

Appendix D Further discussion of motivating applications

Here we spell out a class of motivating applications in which the learning scenario that we study arises. While these exist in the literature, others may be possible.

Classifying quantum many-body systems:

We consider scenarios in which a quantum many-body system has a state with n qubits, and we would like to learn to classify such systems into one of a finite number of categories. This arises in the problem of *quantum phase recognition*, which has been studied in the literature [3, 33].

To perform quantum phase recognition, one could perform state tomography to estimate the system state ρ, then classify it using a classical machine learning architecture, but this involves estimation of an *exponential* number of parameters as a function of n, followed by manipulation of exponentially large feature vectors. Learning a quantum measurement hypothesis, which takes quantum state inputs, avoids this exponential bottleneck. This motivates the formulation of the learning problem studied in our paper. Indeed, avoiding the exponential bottleneck is precisely the reason given for the development of quantum machine learning architectures with quantum inputs in the work [33], which proposes quantum convolutional neural networks for quantum phase recognition, among other applications (such as learning quantum error correction schemes via supervised learning on quantum states).

We furthermore note that study of the basic classification scenario is a theoretical stepping stone toward more advanced learning scenarios, such as sequential and transfer learning in a quantum setting.

Appendix E Further discussion of prior work and relationship to tomography problems

Here we contrast our work with works on state and channel/process tomography. Our main messages are as follows:

1. It is not obvious how channel or state tomography can be used to construct a learning rule;
2. In particular, none of our results seem to follow from PAC learning results for state or channel tomography.

Our motivation for emphasizing these points is that readers sometimes confuse PAC frameworks for estimation/tomography with PAC learning, despite the fact that these are distinct problems.

We start by defining the basic versions of both state and channel tomography.

In state tomography, the problem is as follows: given m copies of an unknown mixed state ρ, produce an estimate $\hat{\rho}$ such that with probability $> 1-\delta$, $\|\hat{\rho}-\rho\|_F^2 \leq \epsilon$, where $\|\cdot\|_F$ is the Frobenius norm.

In channel tomography, one is given access to a quantum channel (formalized by a completely positive trace-preserving (CPTP) map) Φ, and one is allowed to query it m times by preparing input states, passing them through the channel, and measuring the output. The goal is to produce an estimate $\hat{\Phi}$ that is within ϵ of Φ in some metric, with probability $> 1-\delta$.

In both state and channel tomography, it is known that in the finite-dimensional case, only finitely many samples are needed. One may be tempted to try to use a solution to either problem to perform PAC learning with respect to a hypothesis class Hyp. We give a few examples to show the flaws in such approaches.

1. One can imagine viewing the POVM that we want to learn as a quantum channel and using channel tomography to estimate it. There are multiple problems with this approach: the result of channel tomography need not be a POVM in the hypothesis class. More seriously, channel tomography requires that we be allowed to prepare registers in arbitrary states (which we know) and feed them into the channel. Such power is not given to the learner in the PAC learning setting: in fact, input states are drawn from an unknown distribution, and they are not known to us.
2. One might suppose that if we give a little bit more power to channel tomography, then it might become relevant to PAC learning. In particular, suppose that, by any method whatsoever, we could produce a POVM Π (not necessarily in the hypothesis class!) such that $R(\Pi) \leq \epsilon + \inf_{\Pi_*} R(\Pi_*)$, where the infimum is taken over all POVMs. A learning rule that uses this ability must still produce an h_* that lies in the hypothesis class Hyp and is within ϵ of $\inf_{h\in\mathcal{H}} R(h)$ with probability at least $1-\delta$. One natural idea would be to choose $h_* = \arg\min_{h\in\mathrm{Hyp}} d_{TV}(h,\Pi)$. But this does not solve the problem: there exist learning scenarios in which there are multiple, well-separated Π_* that achieve nearly the minimum possible risk (which is not necessarily 0, since the distribution on input states may place positive probability on states that are close together) over all possible POVMs (not just the ones in the hypothesis class). In this case, channel tomography may output a Π that is far

from every hypothesis in the class, while there may exist a hypothesis $h \in \mathcal{H}$ that is very close to some *other* POVM Π_* with nearly minimal risk. This would violate the agnostic PAC learning condition.

3. One might instead think to use state tomography. In particular, the Choi-Jamiołkowski isomorphism result states that, given a CPTP map, if we input a suitably defined maximally mixed state in a larger space and have access to the extended CPTP map that acts via the identity on the environment, then the resulting state completely characterizes the CPTP map. We should note that the learner cannot construct this output state, because the learner cannot provide arbitrary inputs to the CPTP map (the inputs are decided strictly by the data-generating distribution $\mathcal{D}$, not the learner, as is the situation in classical statistical learning theory). Thus, quantum state tomography **cannot** be brought to bear to recover this state, and so we cannot even estimate it, let alone attack the learning problem.

Appendix F A formalization of quantum learning rules

There is some mathematical subtlety in defining a learning rule in the quantum setting. In particular, what it means, informally, for a learning rule to only be able to interact with a quantum register by measurement may be clear, but the formalization in terms of mathematical objects is less straightforward. For completeness, we give such a formalization in this section, in terms of Markov decision processes.

We define the following Markov decision process for a given dataset $S = ((X_1, Y_1), ..., (X_m, Y_m))$: the *state* Z_0 is initialized to $\otimes_{j=1}^m X_j$. At any timestep t, the set of possible actions consists of POVMs operating on the state Z_t, producing an observable outcome ω_t via the Born rule. The state Z_{t+1} is then derived from S_j again via the Born rule.

A POVM learning rule specifies a policy for this MDP, where, at each timestep $t \geq 0$, the action A_t at time t is conditionally independent of Z_j for any j, given the outcomes $\omega_0, ..., \omega_{t-1}$. Finally, the learning rule specifies a conditional distribution from outcome sequences to hypotheses $h \in \mathrm{Hyp}$.

Appendix G More details on sample complexity bounds for variational quantum circuits

Here we give more details for our sample complexity bounds for hypothesis classes corresponding to variational quantum circuits.

In great generality, one can define the following hypothesis class.

Definition 19 (General variational quantum circuit). *Let $\mathcal{H}$ be a Hilbert space with dimension d. We fix a POVM Π on $\mathcal{H}$ with two outcomes. We define the hypothesis class $\mathbb{G}$ to consist of hypotheses parametrized by an arbitrary unitary operator U from $\mathcal{H}$ to $\mathcal{H}$ that apply U to the input state ρ, then measure the resulting state with Π.*

We have the following theorem giving the sample complexity of $\mathbb{G}$.

Theorem 8 (Sample complexity bound for $\mathbb{G}$)**.** *The class $\mathbb{G}$ is (ϵ, δ)-PAC learnable with sample complexity*

$$n_{\mathbb{G}}(\epsilon, \delta) \leq O(\frac{8N}{\epsilon^2} \log \frac{2N}{\delta}) \leq (C/\epsilon)^{d+2} \log(1/\delta). \tag{G76}$$

Proof. By Theorem 3, it is sufficient to upper bound the $\epsilon/4$-TV-covering number of $\mathbb{G}$. We first note that for any $\gamma > 0$, the γ-TV covering number of $\mathbb{G}$ is upper bounded by the γ-L_1-operator norm covering number of the set of $d \times d$ unitary matrices, which in turn is upper bounded by the γ-L_1 norm covering number of the set of $d \times d$ unitary matrices, viewed as vectors.

The γ-L_1-operator norm covering number of the set of $d \times d$ unitary matrices is

$$\leq d(C/\gamma)^d, \tag{G77}$$

for some positive constant C, by upper bounding by the γ-L_2-operator norm covering number and using Theorem 7 of [34].

This implies that $N \leq d(C/\gamma)^d$, and setting $\gamma = \epsilon/4$, we get

$$N \leq d(C/\epsilon)^d, \tag{G78}$$

resulting in the bound

$$n(\epsilon, \delta) \leq O(\frac{d(C/\epsilon)^d}{\epsilon^2} \log \frac{2d(C/\epsilon)^d}{\delta}). \tag{G79}$$

Upper bounding factors that are polynomial in d and ϵ by $(C/\epsilon)^d$, this can be simplified to

$$n(\epsilon, \delta) \leq (C/\epsilon)^{d+2} \log(1/\delta), \tag{G80}$$

which completes the proof. □

This sample complexity scales exponentially with the dimension – as one might expect, since the hypotheses may be thought of as applying an arbitrary circuit to the input, then measuring by a fixed POVM, and so such a hypothesis class is not used in practice. Our purpose in spelling out this example is to illustrate the type of analysis that one would undertake in applying our results to a specific hypothesis class. Instead, the approach taken in works on variational quantum circuits is to constrain the form of the unitary operator. Specifically, [2] constrains U to be *bandlimited* in terms of its Fourier spectrum, which reduces the number of trainable parameters and, hence, the TV-covering number of the hypothesis space to a polynomial value in the dimension.