Learning Quantum Processes Without Input Control

Marco Fanizza,[1,*] Yihui Quek,[2,3,†] and Matteo Rosati[4,‡]

[1]*Física Teòrica: Informació i Fenòmens Quàntics, Departament de Física, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain*

[2]*Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Berlin 14195, Germany*

[3]*Massachusetts Institute of Technology, Cambridge, Massachusetts, USA*

[4]*Dipartimento di Ingegneria Civile, Informatica e delle Tecnologie Aeronautiche, Università Roma Tre, Via Vito Volterra 62, Rome I-00146, Italy*

We introduce a general statistical learning theory for processes that take as input a classical random variable and output a quantum state. Our setting is motivated by the practical situation in which one desires to learn a quantum process governed by classical parameters that are out of one's control. This framework is applicable, for example, to the study of astronomical phenomena, disordered systems and biological processes not controlled by the observer. We provide an algorithm for learning with high probability in this setting with a finite amount of samples, even if the concept class is infinite. To do this, we review and adapt existing algorithms for shadow tomography and hypothesis selection, and combine their guarantees with the uniform convergence on the data of the loss functions of interest. As a byproduct, we obtain sufficient conditions for performing shadow tomography of classical-quantum states with a number of copies, which depends on the dimension of the quantum register, but not on the dimension of the classical one. We give concrete examples of processes that can be learned in this manner, based on quantum circuits or physically motivated classes, such as systems governed by Hamiltonians with random perturbations or data-dependent phase shifts.

I. INTRODUCTION

The goal of science is to gain a better understanding of nature. Because "nature isn't classical, dammit" [1], the tools of quantum information processing have been central to this pursuit. Insights transposed from statistical learning theory to the quantum domain have established rigorous guarantees for learners that predict properties of quantum states [2–6], classify phases of matter [5], learn quantum channels [7], or approximate models of physical dynamics [8].

A feature shared by many of these works is that they require the quantum learner to have precise control of the unknown object—for instance, the ability to request multiple identical copies of the unknown state or run the unknown process on a string of well-chosen inputs. That is, the unknown process is treated as a black box, to be applied on inputs specially designed by the learner. However, this assumption is not always satisfied in practice. A scientist can typically *observe but not fully control* an unknown process of interest—think of an astronomer analyzing signals generated by rare celestial events, a biologist probing molecular mechanics induced by biochemical signals in a noisy environment, or a physicist characterizing systems obeying Hamiltonians subject to random perturbations. In this work, we show that quantum processes can be learned even without the strong assumption of input control.

That is (see Fig. 1), the learner receives as examples the input-output pairs

$$(x_1, \rho(x_1)), \ldots (x_n, \rho(x_n)), \tag{1}$$

where x_i are classical inputs and the $\rho(x_i)$ are the quantum states output by the unknown process with these inputs. Here the lack of input control is reflected in the fact that the x_i are not chosen by the learner but are samples from some distribution $\mathcal{D}$. Practically, the imperfect control could consist in the impossibility of choosing the input to a process or the inevitability of the input fluctuating throughout the various applications of the process. For instance, in

*Corresponding authors. marco.fanizza@uab.cat

†yihuiquek3.14@gmail.com
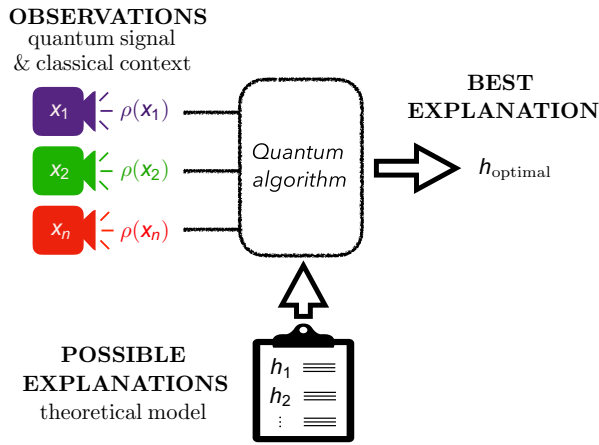
‡matteo.rosati@uniroma3.it

FIG. 1. Schematic depiction of the experimental setting for which we provide theoretical learning guarantees. A size-*n* training set comprising couples of a classical label *x* and a corresponding quantum state $\rho(x)$ is produced by sampling from random sources parametrized by $x_1, x_2, \ldots, x_n$. The learner tries to find a theoretical explanation to the observed classical-quantum data using a hypothesis class of operator-valued functions $h(x)$, possibly of infinite cardinality. We present quantum algorithms that are able to identify a near-optimal hypothesis h_{optimal} and obtain sample complexity guarantees in terms of covering numbers of the hypothesis class.

optical imaging of celestial, atmospheric, planetary, or biological events, $\rho(x)$ could describe the quantum state of an optical signal emitted or backscattered by an unknown object, while *x* is a set of environmental parameters influencing the state of the signal, e.g., temperature, optical depth, distance.

We model the situations described above as follows: the learner has access to a source that outputs classical-quantum states

$$\sum_{x \in \mathcal{X}} \mathcal{D}(x)|x\rangle\langle x| \otimes \rho(x),$$

$$\text{where } \rho : \mathcal{X} \to \mathcal{S} \text{ and } x \overset{\mathcal{D}}{\leftarrow} \mathcal{X}, \quad (2)$$

with $\mathcal{S}$ being the set of states of some Hilbert space $\mathcal{H}$ and ρ the unknown process. At each data collection, the input will be a product state $(|x_1\rangle\langle x_1| \otimes \rho(x_1)) \otimes \cdots \otimes (|x_n\rangle\langle x_n| \otimes \rho(x_n))$. This generalizes the setting of Refs. [7,9] on probably approximately correctly (PAC)-learning quantum channels.

Our goal is to learn a possible classical-quantum source from which the data are sampled, and we construct algorithms to do so that have bounded sample complexity. A key object in the analysis is a certain distinguishability measure of the class of candidate classical-quantum processes, which we identify, that can be bounded even if this class is infinite. As an application, we also improve the

shadow tomography procedure of Ref. [3] for an unknown classical-quantum state: a naive application of shadow tomography would not be guaranteed to work if the classical register is infinite dimensional, while we show that only the dimension of the quantum register matters.

Our theory also goes beyond previous works to encompass the *agnostic case*: the case when the unknown process is *not* included in the hypothesis set $\mathcal{C}$. Note that the agnostic case was mentioned by Ref. [7], who provided a lower bound on the sample complexity, and tackled in a specific case by Ref. [9]. This is as opposed to the *realizable* case, where the unknown process is guaranteed to be from the set. This is in keeping with the setting of learning nature: agnostic learning models the situation when we learn a natural process using as hypotheses only the limited models that are within the reach of our theoretical understanding.

Our work overcomes a key technical hurdle not tackled by previous approaches: without input control, our learner cannot obtain identical copies of the process sampled at well-chosen points $\{x_i\}$. This assumption is crucial in, for instance, Refs. [8] and [10]. Instead, for every sampled x_i, she gets only a single copy of $\rho(x_i)$, which in general differs from the other copies $\rho(x_j)$, $j \neq i$. Nevertheless, we design a measurement strategy that learns even without the luxury of identical copies. Ref. [7] hinted that a VC-dimension-like quantity for this setting might be impossible to define. Here instead we establish a fundamental prerequisite for the definition of a statistically meaningful dimension (analogous to the fundamental theorem for uniform convergence of Ref. [11]), providing sufficient conditions for learning concept classes of infinite cardinality with finite data: we introduce learning algorithms that succeed with high probability if a suitable *covering number* of the quantum concept class $\mathcal{C}$ grows sufficiently slowly with the sample size.

Our theory of learnability extends *beyond* the usual setting of learning quantum processes given by quantum circuits. In fact, our learning model also encompasses the following scenarios:

(a) We want to study how a small quantum system behaves in a variable environment. In this case, the classical random variable is a measurement of the status of the environment, for example, a measure of classical fields. The copy of the quantum state is a copy of the state of the system corresponding to the measured state of the environment. One can imagine applying this scenario to molecules or nanostructures. Notice that the border between environment and object is arbitrary, therefore in the classical random variable one could include the outcomes of some predetermined measurement on the object of interest. With the same idea in mind, the quantum state could be also not the original state of the

system but some postprocessing of it, for example, via a quantum sensor.

(b) We want to do imaging of a system, that is to associate to each point of the system a quantum state or channel. When the detector clicks, we receive as experimental data a pair comprising the position and the quantum state corresponding to that position. If our experimental setup uses spontaneous or stimulated emission, such that in a specific time interval we cannot guarantee that we can obtain an observation at a specific position, our model correctly represents the fact that we receive data from random positions and we cannot afford to receive multiple copies of the state corresponding to an arbitrary position.

(c) We are studying a class of stars with a combination of classical and quantum sensors: for example, we get classical electromagnetic signals from a star, and a quantum sensor collects information about gravitational waves. We would like to study correlations between the electromagnetic and gravitational waves, but since these events are rare and unique we cannot repeat them at will.

Some toy models for concept classes inspired by these scenarios are discussed as applications. In these cases, we can find bounds on the covering number, from which sample complexity bounds can be obtained.

A. Setting: learning quantum processes with random classical input

We now go into more detail about our learning setting, which is a natural quantum generalization of supervised learning and builds on that of Ref. [7]. Suppose there is an unknown function $\rho : \mathcal{X} \to \mathcal{S}$ to be learned, $\mathcal{X}$ possibly of infinite cardinality, and a distribution $\mathcal{D} : \mathcal{X} \to [0, 1]$. In fact, we will focus on processes ρ that map from a *classical* domain $\mathcal{X}$ to a *quantum* set $\mathcal{S} \subseteq \mathcal{L}(\mathcal{H})$, where $\mathcal{L}(\mathcal{H})$ is the set of linear operators on a finite-dimensional Hilbert space $\mathcal{H}$. When we want to keep track of the dimension d of a Hilbert space, we use the notation $\mathcal{H}^{(d)}$. Furthermore, we will always be interested in the case where the unknown process ρ outputs a *quantum state*, that is, all operators in $\mathcal{S}$ are positive semidefinite and have unit trace.

The learner receives as input samples the pairs $(x_i, \rho(x_i))_{i=1}^{n}$, where $x_i \overset{\mathcal{D}}{\leftarrow} \mathcal{X}$. Furthermore, the learner has a set of hypotheses, $\mathcal{C} = \{h : \mathcal{X} \to \mathcal{L}(\mathcal{H})\}$, and would like to use the smallest possible number n of samples to choose a candidate h from the class that accounts well for the observations. The accuracy of the learner's output h relative to the true function ρ will be measured by the *true risk* $R_\rho : \mathcal{C} \to \mathbb{R}$, defined via a loss function $L : \mathcal{L}(\mathcal{H}) \times$

$\mathcal{L}(\mathcal{H}) \to \mathbb{R}$:

$$R_\rho(h) := \mathbb{E}_{x \sim \mathcal{D}}[L(\rho(x), h(x))] \quad \text{(True risk)}. \quad (3)$$

Therefore, it is in the learner's interest to minimize the true risk, although she can only do so approximately, as detailed in the next section.

In what is known as the *realizable* setting of learning (studied by Ref. [7]), there is a promise that the unknown function comes from $\mathcal{C}$. We present results for this setting too, but go one step further. In learning nature, our scientific models are but approximate descriptions that correspond more closely to reality at some scales than others. Thus, we primarily treat the *agnostic* (or unrealizable) setting, in which no hypothesis in $\mathcal{C}$ is guaranteed to correspond exactly to the unknown function.

We will focus on concept classes $\mathcal{C}$ that output two types of quantum objects:

(a) **(Case 1) Quantum states:** $\mathcal{C}$ consists of hypotheses $h(x) = \sigma_h(x)$, which are state-valued functions. The true function $\rho(x)$ is also a state-valued function, and we use trace distance $L_s(\sigma_h, \rho) = d_{\text{tr}}(\sigma_h, \rho)$, a natural notion of distance between quantum states, as the loss function.

(b) **(Case 2) Quantum events or projectors:** $\mathcal{C}$ consists of hypotheses $h(x) = \Pi_h(x)$ which are projector-valued functions. Again, the true function $\rho(x)$ is a state-valued function and we will use as loss function the probability of not accepting the projector, i.e., $L_p(\Pi_h, \rho) := 1 - \text{Tr}[\rho \Pi_h]$.

We note that in case 2 we can switch out the projectors for general positive-operator valued measurement (POVM) elements, by a standard dilation argument with which we can represent them as projectors on a larger space. However, the dilation is not unique and what we will say in the following will depend on the dilation. Therefore, for simplicity, we will always speak only about projectors.

A quick note on the motivation for defining projector-valued concept classes. In the classical case, when the label y is not a deterministic function of x, one speaks of learning probabilistic concepts [12]. In Ref. [12] two possible approaches are considered. The first is to learn a deterministic concept, which maximizes the probability of correct prediction, the second is to learn the conditional probability distributions $p(y|x)$ on average. Learning projector-valued classes is a generalization of the first approach, and learning state-valued classes is a generalization of the second approach.

Moreover, estimating $L_p(\Pi_h, \rho)$ for every $h \in \mathcal{C}$ (as we do later) encompasses shadow tomography [2,3], a task where one is given a fixed state and a list of observables and has to output the expectation values of all observables in the list. Shadow tomography corresponds to the case

where $\rho(x)$ and $\Pi_h(x)$ are constant as functions of x. Most importantly, our algorithm for this risk estimation problem on projector-valued functions is a key part of our strategy to attack case 1.

B. Results

In this section, we state our main results, based on the algorithms summarized in Fig. 2 and obtained in Secs. III, IV, V. Recall that the goal is to learn an unknown function ρ and the output of our learning algorithm will be some hypothesis $h \in \mathcal{C}$. While ρ is not required to be in $\mathcal{C}$, our algorithms will always output some $h \in \mathcal{C}$ close to achieving the minimum *empirical risk*, i.e., the average

loss computed on the examples $(x_i, \rho(x_i))_{i=1}^n$:

$$\hat{R}_\rho(h) := \frac{1}{n} \sum_{i=1}^n L(\rho(x_i), h(x_i)) \quad \text{(Empirical risk).} \quad (4)$$

That is to say, we use minimizing the empirical risk as a proxy for minimizing the true risk. This principle, known as *empirical risk minimization* (ERM), originated in classical statistical learning theory [11]; the contribution of this paper is to adapt it to learning quantum-valued classes. In this paper, we distinguish between two types of tasks: the term *empirical risk minimization* refers to the task of outputting a *single hypothesis* from $\mathcal{C}$ that approximately

---

Input:

- hypothesis class $\mathcal{C} = \{h(x)\}_h$;
- loss function L;
- samples $(x_i, \rho(x_i))_{i=1}^n$ for unknown ρ, a state-valued function.

Goal: Empirical risk minimization,
i.e., for parameters $\eta, \epsilon > 0$, output $h^* \in \mathcal{C}$ such that

$$R(h^*) \leq \eta \min_{h \in \mathcal{C}} \hat{R}_\rho(h) + \epsilon.$$

Case 1 (States)

Hypotheses are $h(x) = \sigma_h(x)$,
state-valued functions.
Loss function: $L_s(\sigma_h, \rho) := d_{\text{tr}}(\sigma_h, \rho)$.

Quantum empirical risk estimation:
Input: $\bigotimes_{i=1}^n \rho_i$; m lists of projectors $\{\Pi_1^{(c)}, ..., \Pi_n^{(c)}\}_{c=1,...,m}$, so that $\mu_c = \frac{1}{n} \sum_{i=1}^n \text{Tr}[\rho_i \Pi_i^{(c)}]$. We will ultimately plug in projector-valued functions derived from state-valued functions, see the definition of A_{ij} below.
Output: $\hat{\mu}_c$ estimate of μ_c, $\forall i \in [m]$

+

Hypothesis selection:
Input: $\hat{\mu}_{ij}$ estimates of the risk; descriptions of the projector-valued functions $A_{ij}(s) := (\sigma_i(s) - \sigma_j(s))_+$. (Note: $d_{\text{tr}}(\sigma_i, \sigma_j) = \text{Tr}(\sigma_i A_{ij}) - \text{Tr}(\sigma_j A_{ij})$.)
Output: i^* s.t. minimizes $\max_{i<j} |\frac{1}{n} \sum_{s=1}^n \text{Tr}[\sigma_{i^*}(s) A_{ij}(s)] - \hat{\mu}_{ij}|$.

Case 2 (Events)

Hypotheses are $h(x) = \Pi_h(x)$,
projector-valued functions.
Loss function: $L_p(\Pi_h, \rho) := 1 - \text{Tr}[\rho \Pi_h]$.

Generalized Threshold Search:
Input: $\otimes_{i=1}^n \rho_i$; a collection of pairs $(\{\Pi_i^{(c)}\}_{i \in [n]}, \theta_c)$, $c = 1, ..., m$;
Output: c s.t.
$\frac{1}{n} \sum_{i=1}^n \text{Tr}[\rho_i \Pi_i^{(c)}] \geq \theta_c$, if it exits

+

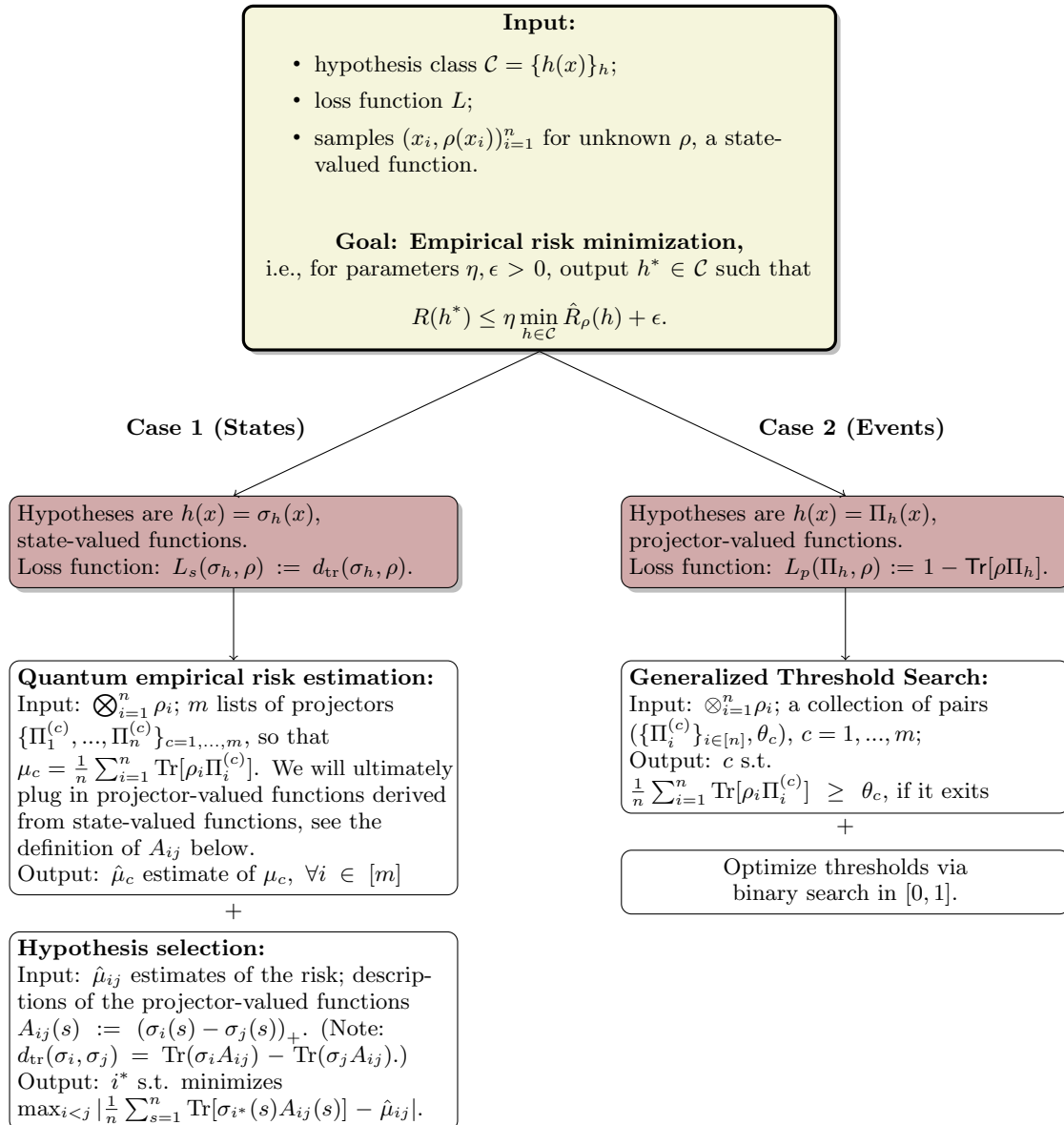Optimize thresholds via binary search in $[0, 1]$.

FIG. 2.   Summary of quantum empirical risk minimization algorithms.

minimizes the risk. Along the way, we will also develop an algorithm for *empirical risk estimation* (ERE), a term that refers to estimating the risk for *every hypothesis* in $\mathcal{C}$.

As formulated by Ref. [11], the success of ERM depends on a property of a concept class known as *uniform convergence* [13,14] of the empirical risks, which is controlled by a certain measure of the effective size of the concept class we define: the $\gamma_{1,q}$ covering number (see general treatments of covering numbers, e.g., Refs. [13,15]).

Definition 1 (Covering number). Let $\mathcal{G} \subseteq \mathcal{L}(\mathcal{H})^{\mathcal{X}}$ be a class of functions mapping to linear operators, and let $\epsilon \in (0, 1]$. The *covering number* of $\mathcal{G}$ is

$$\gamma_{1,q}(n, \epsilon, \mathcal{G}) := \max \left\{ N_{in}\left(\epsilon, \mathcal{G}, || \cdot ||_{1,q,\vec{x}}\right) | \vec{x} \in \mathcal{X}^n \right\}, \quad (5)$$

where for a pseudometric d, $N_{in}(\epsilon, \mathcal{G}, d)$ is the smallest cardinality of any internal ϵ-cover of $\mathcal{G}$ according to the pseudometric d. Here we have chosen as pseudometric the $||\cdot||_{1,q,\vec{x}}$ seminorm, which depends on the observed examples $\vec{x}$ as

$$\|g_1 - g_2\|_{1,q,\vec{x}} := \frac{1}{|\vec{x}|} \sum_{i=1}^{|\vec{x}|} ||g_1(x_i) - g_2(x_i)||_q, \quad (6)$$

where $||A||_q$ is a Schatten norm (see definitions in Sec. II D) and $|\vec{x}|$ is the length of $\vec{x}$, i.e., the sample size.

Intuitively, $\gamma_{1,q}$ describes the maximum number of hypotheses that can be pairwise distinguished on a dataset of size n with resolution ϵ, given information from the classical register only; in fact, the distinguishability between two hypotheses is measured by the average over classical outcomes of the appropriate distinguishability metric (operator norm for projectors and trace norm for states) for the corresponding values of the hypotheses.

Since the rate of convergence of the empirical risk to the true risk is controlled by the covering number $\gamma_{1,q}$ [11,13,15], it is possible to minimize the risk by optimizing over an ϵ-net of the concept class, which is finite-dimensional, rather than the class itself, which in general is infinite-dimensional. This is the basis of our technique. While the ERM principle is well established classically, the nontrivial part in the quantum case is to minimize the empirical risk $\hat{R}_\rho(h)$ based on the string of samples $(x_i, \rho(x_i))_{i=1}^n$. The main difficulties in this respect are the lack of identical copies of the input states and the fact that the observables associated to naive estimators of the empirical risks do not commute.

We now present our results establishing a quantum variant of ERM, for both types of concept classes considered.

1. Quantum empirical risk minimization and estimation

Let us first discuss projector-valued concept classes. Here, the naive strategy of measuring each ρ_{x_i} with $\Pi_h(x_i)$

for all h to estimate the empirical loss does not immediately work, since the projectors do not necessarily commute for different hypotheses. Nevertheless, we construct algorithms for both ERE and ERM on projector-valued concept classes, establishing the following theorems.

Theorem 1 (Quantum empirical risk minimization for projector-valued functions). Given access to a product state

$$\varrho = \rho_1 \otimes \cdots \otimes \rho_n \quad (7)$$

and a collection of lists of projectors $\{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}_{c=1,\ldots,m}$, with

$$\mu_c = \frac{1}{n} \sum_{i=1}^{n} \text{Tr}[\rho_i \Pi_i^{(c)}], \quad (8)$$

there is an algorithm which outputs c^* and $\hat{\mu}_{c^*}$ such that

$$\Pr(|\hat{\mu}_{c^*} - \max_{c \in [m]} \mu_c| \geq \epsilon \cup |\hat{\mu}_{c^*} - \mu_{c^*}| \geq \epsilon) \leq \delta \quad (9)$$

if n is large enough. In fact we can take

$$n = \tilde{O}\left(\frac{\log \frac{1}{\delta} \max(\log \frac{m}{\delta}, \log^2(em))}{\epsilon^2}\right). \quad (10)$$

Here, the notation $\tilde{O}$ hides logarithmic dependence on $\log m$, $\log \delta$, and ϵ. Note that the minimization in the caption of Theorem 1 refers to the minimization of the loss function, which is $1 - \mu_c$, whereas the theorem statement is about maximizing μ_c, which is equivalent.

Theorem 2 (Quantum empirical risk estimation for projector-valued functions). Given access to a product state

$$\varrho = \rho_1 \otimes \cdots \otimes \rho_n \quad (11)$$

and a collection of lists of projectors $\{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}_{c=1,\ldots,m}$, with

$$\mu_c = \frac{1}{n} \sum_{i=1}^{n} \text{Tr}[\rho_i \Pi_i^{(c)}], \quad (12)$$

there is an algorithm, which outputs estimates $\hat{\mu}_c, c \in [m]$ such that

$$\Pr(\exists c \in [m] : |\hat{\mu}_c - \mu_c| \geq \epsilon) \leq \delta \quad (13)$$

if n is large enough. In fact we can take

$$n = \tilde{O}\left(\frac{\log d \log \frac{1}{\delta} \max(\log \frac{m}{\delta}, \log^2(em))}{\epsilon^5}\right). \quad (14)$$

Here, the notation $\tilde{O}$ hides logarithmic dependence on $\log m, \log \delta, \log d$, and ϵ. Note that estimation, in comparison with minimization, shows a worse dependence on ϵ and an explicit dependence on d. However, it represents a key subroutine to realize the ERM algorithm in the case of state-valued concept classes, which is the subject of the next theorem.

Recall that, in this case, the loss function is different from the projector-valued case: it is the trace distance, rather than the overlap. This prevents us from immediately applying the techniques used for projector-valued classes, which only work to estimate expectation values.

Nevertheless, using a trick from *quantum hypothesis selection* [4], we can reduce the task of risk *minimization* for state-valued concept classes to estimating expectation values of Helstrom projectors constructed from the state-valued class. The result is the following theorem.

Theorem 3 (Quantum empirical risk minimization for state-valued functions). Let $\mathcal{C} = \{\sigma_i : [n] \to \mathcal{L}(\mathcal{H}^{(d)})\}_{i=1}^m$ be a class of state-valued functions and

$$\varrho = \rho_1 \otimes \cdots \otimes \rho_n. \tag{15}$$

There exists an algorithm, which given ϱ outputs i^* such that

$$\frac{1}{n} \sum_{s=1}^n d_{\text{tr}}(\sigma_{i^*}(s), \rho_s) \leq 3\eta + 4\epsilon, \tag{16}$$

where

$$\eta := \min_{i \in [m]} \frac{1}{n} \sum_{s=1}^n d_{\text{tr}}(\sigma_i(s), \rho_s), \tag{17}$$

with probability of error less than δ if n is large enough. In fact we can take

$$n = \tilde{O}\left(\frac{\log d \log \frac{1}{\delta} \max(\log \frac{m}{\delta}, \log^2(em))}{\epsilon^5}\right). \tag{18}$$

Here, the notation $\tilde{O}$ hides logarithmic dependence on $\log m, \log \delta, \log d$, and ϵ.

2. Learning via empirical risk minimization

Combining the theorems just stated and uniform convergence guarantees from statistical learning theory, we show the following sufficient conditions for learning.

Theorem 4 (Learning quantum processes via ERM). Suppose the concept class $\mathcal{C}$ consists of classical-quantum processes mapping to projectors or states and let $\epsilon > 0$ be the accuracy parameter. Furthermore, let $S = (x_i, \rho(x_i))_{i=1}^n$ be the training set, with $x_i \xleftarrow{\mathcal{D}} X$ and $\rho(\cdot)$ an unknown classical-quantum channel.

Then, the appropriate ERM algorithm of Theorems 1, 3, run on an ϵ-net of the concept class $\mathcal{C}$ (according to the appropriate pseudometric determined by $x_1, \ldots, x_n$), provide an *agnostic* learning algorithm $\mathcal{A} : \mathcal{X}^n \times \mathcal{L}(\mathcal{H}^{(d)})^{\otimes n} \to \mathcal{C}$. This algorithm outputs an hypothesis $\mathcal{A}(S)$ satisfying, for some fixed $\eta, \xi \geq 1$, and n large enough,

$$\Pr_S[R_\rho(\mathcal{A}(S)) - \eta \inf_{h \in \mathcal{C}} R_\rho(h) < \xi\epsilon] > 1 - \delta, \tag{19}$$

if

$$\lim_{n \to \infty} \frac{\log^2 \gamma_{1,q}(n, \epsilon, \mathcal{C})}{n} = 0, \qquad \forall \epsilon > 0. \tag{20}$$

In particular, this applies to risks defined via the loss functions L_p (in this case $\eta = 1$, $q = \infty$) and L_s (in this case $\eta = 3$, $q = 1$) for projector-valued and state-valued concept classes $\mathcal{C}$, respectively.

We remark that Theorem 4 applies even to agnostic learners, going beyond Ref. [7], which considered only the realizable setting. In fact our methods also imply that we can learn with an *infinite* concept class if $\lim_{n \to \infty} \log^2(\gamma_{1,q}(n, 0, \mathcal{C}))/n = 0$ [while in Ref. [7] show a sample complexity $O(\text{poly}(\log |\mathcal{C}|))$]. Moreover, in Appendix C we analyze the algorithm of Ref. [7] for learning pure states using a statistical learning theory approach, to show that $\lim_{n \to \infty} \log(\gamma_{1,1}(n, 0, \mathcal{C}))/n = 0$ suffices also with that algorithm.

As a consequence of proving Theorem 4 for the case when the concept class consists of projectors, we are also able to speed up shadow tomography of classical-quantum states, vis-a-vis Ref. [3]:

Theorem 5 (Shadow tomography of classical-quantum states). Suppose the concept class $\mathcal{C}$ consists of quantum processes mapping to projectors, $\epsilon > 0$ is the accuracy parameter, and the learner is given the same input as in the previous theorem. Suppose also that

$$\lim_{n \to \infty} \frac{\log^2 \gamma_{1,\infty}(n, \epsilon, \mathcal{C})}{n} = 0, \qquad \forall \epsilon > 0. \tag{21}$$

Then, the estimation algorithm of Theorem 2 run on an ϵ-net of the concept class $\mathcal{C}$ (according to the appropriate pseudometric determined by $x_1, \ldots, x_n$), provide an *agnostic* learning algorithm $\mathcal{A} : \mathcal{X}^n \times \mathcal{L}(\mathcal{H}^{(d)})^{\otimes n} \to [0, 1]^{|\mathcal{C}|}$, which output estimates of risks $\mu(h)$ for all concepts $h \in \mathcal{C}$, such that for some fixed $\xi \geq 0$,

$$\Pr_S[\forall h \in \mathcal{C}, \ |\mu(h) - R_\rho(h)| < \xi\epsilon] > 1 - \delta, \tag{22}$$

for n large enough.

In words, we can not only find the minimum possible risk in the concept class, but also simultaneously estimate risks for all concepts. In fact, when $\mathcal{C}$ is finite, algorithm $\mathcal{A}$ performs shadow tomography of classical-quantum states with a copy complexity of $\tilde{O}(\text{poly}(\log |\mathcal{C}|, \log d, 1/\epsilon))$, where d is the dimension of the quantum register only. By contrast, a naive application of shadow tomography [3] has a copy complexity scaling as the dimension of the full space, which may be infinite if $|\mathcal{X}| = \infty$. *Remark*: we mention that the algorithm of the aforementioned theorems performs ERM correctly with high probability on *any* possible data sequence, and not just on a subset of sequences which occur with high probability. If one is willing to accept that sometimes ERM can fail with non-negligible probability, learning becomes possible when $\lim_{n\to\infty}(\log \gamma_{1,\infty}(n,\epsilon,\mathcal{C}))/n = 0$, rather than $\lim_{n\to\infty}(\log^2 \gamma_{1,\infty}(n,\epsilon,\mathcal{C}))/n = 0$. We argue for this in Appendix D, by presenting a modification of the algorithms that perform ERM correctly only on a certain ϵ-net on a subset of the data, with high probability. We illustrate this fact only for risk minimization for a projector-valued class, but the same can be shown also for risk estimation for a projector-valued class, and for risk minimization for a state-valued class.

3. Examples of learnable classes

In Sec. VI, we evince the applicability of our results by giving examples of quantum processes covered by our theorems, for which we compute explicit upper bounds on the covering number $\gamma_{1,q}$ in terms of the dimensionality of the quantum systems and the *fat-shattering dimension* of an appropriate concept class $\mathcal{F}$. In all of the below examples, let $\mathcal{F}$ be a class of real-valued functions $g(x)$. We build our concept classes using finite-dimensional circuits and matrix functions that depend on the data via real functions $g(x)$ coming from concept classes $\mathcal{F}$. In fact, for the function classes with explicit data dependence that we consider, the covering number can slowly grow to infinity, but our theorems show they are still learnable.

(1) Quantum circuits in a particular architecture, possibly data dependent. Concept classes consisting of m-qubit quantum circuits chosen from a set S_m acting on arbitrary input states or projectors. That is, $\mathcal{C}_m = \{c_U(x) := U\rho(x)U^\dagger\}_{U \in S_m}$ where $\rho(x)$ is a process that prepares the input state to the circuit, or $\mathcal{C}_m = \{c_U(x) := U\Pi(x)U^\dagger\}_{U \in S_m}$, where $\Pi(x)$ is a projector. We give explicit upper bounds on the covering number for S_m being

(a) the class of one-dimensional local quantum circuits on m qubits of depth ℓ [16], constructed by applying ℓ 2-qubit nearest-neighbor gates on any pair of neighboring qubits,

(b) brickwork quantum circuits,

(c) the set of all unitaries,

(d) data-dependent circuits with any of the previous architectures, but modified by inserting in specific places in the sequence a number ℓ' of gates of the form $e^{iH_j g_{j'}(x)}$, $j' = 1, \ldots, \ell'$, $H_{j'}$ fixed.

(2) Gibbs states and low-energy projectors of a perturbed Hamiltonian. Concept classes given by a set of Gibbs states obtained by perturbing a Hamiltonian with a field-dependent term, but the specific dependence on the field is not known, namely $H_0 + g(x)V$. Similarly, concept class of projectors on low-energy eigenspace of $H_0 + g(x)V$.

(3) Phase shifts with position-dependent depth. Consider a spatially local channel that acts as a power of an unknown unitary channel at position x, according to some classical variable $g(x)$ at that position (for example, a thickness), which we probe with some position-dependent state $\rho(x)$. We can model this class as a state-valued [or projector-valued, when $\rho(x)$ are pure] function class: $\mathcal{C} = \{f \mid f(x) = U^{g(x)}\rho(x)(U^\dagger)^{g(x)} \; g \in \mathcal{G}, U = e^{iH} \in U(d)\}$.

C. Related work

There is a vast literature on how to learn quantum states and their properties. The exponential cost of full-state tomography motivated to consider the problem of "pretty-good tomography" [2], in which the goal of the learner is relaxed: she is content with obtaining a predictor function for *properties* of the state instead of a full description of it. A similar task, consisting of estimating expectation values of a fixed list of observables, is known by the name *shadow* tomography [3,4,17]. Even though we aim to learn processes and not quantum states, we have leveraged the tools developed in the shadow tomography literature for our work.

In this work, we delve into the considerably less studied realm of learning quantum *processes* that have a classical input variable and a quantum output. As mentioned, this model is motivated by the observation that many processes in nature leave a classical *imprint* of the conditions under which they occur, which the learner can detect with experimental equipment, but not observe. Our goal is to build a statistical learning theory for learning quantum processes that takes into account the unique structure of quantum measurements.

Several previous works have investigated similar directions. Ref. [7] first introduced the notion of probably approximately correct (PAC)-learning of quantum channels, which captures our notion of lack of input control—but solved the problem exclusively in the realizable setting. Their algorithms however have sample complexity scaling polynomially in $\log(|\mathcal{C}|)$, and thus cannot be applied when the concept class $\mathcal{C}$ is infinite. Our paper encompasses their setting (and improves their result for pure states) and solves the agnostic setting of channel

learning. We also define an extensive measure of the concept class—the covering number—which controls learnability and can be finite even when $\mathcal{C}$ is infinite. The work [9] addresses this aspect for the case of binary functions with quantum output (which are two states corresponding to 0 and 1, respectively), finding that Helstrom measurements are enough to get an upper bound on the sample complexity depending on the VC dimension of the associated classical function class.

Ref. [8] also considers agnostic learning for polynomial-time quantum processes. However, the paper models each process as an experiment over which the experimenter has full control. Accordingly, the learner runs the process on identical copies of the input in order to implement the hypothesis selection procedure (Sec. F.3.b). On the other hand, we are interested in the setting where the experimenter must contend with nonidentical copies of the input. Additionally, while the model in Ref. [8] applies to channels with bounded input and output dimensions, where one can always find an ϵ-net directly for the concept class (by its compactness), in our case instead we are interested in learning correlations between a classical and a quantum random variable, with the classical variable living possibly in an unbounded space. In this case, one cannot always define a finite ϵ-net on the sources, but we show that learning can be still guaranteed in terms of ϵ-nets of the datasets formed by pairs of classical and quantum variables.

Ref. [18] also works in a PAC setting for learning quantum channels acting on classical inputs drawn from a distribution, but their notion of learning is to find a good predictor function (mapping to real values) for the expected value of a fixed set of observables on the output of an unknown quantum process. The focus of their work is whether quantum machine-learning algorithms can have a large advantage over classical machine learning for various notions of prediction error. By contrast, our notion of learning is different and is not about predicting observables accurately. Furthermore, our work provides constructive quantum algorithms that achieve the sample complexity bounds we mention, while no such algorithms are presented in this other work.

Ref. [19] also considered the problem of learning classical-quantum processes, providing generalization bounds in terms of entropic quantities. However, their setting is closer to a discrimination problem: the task is to find, among a set of possible projectors, one that best approximates joint distributions of classical data x and a set of labels c, when measured on quantum states $\rho(x)$ representing embedding of x. Our setting is different, as their projectors do not depend on the specific value of x; moreover, their bounds on the generalization error depend on the source and not on the concept class, capturing a different aspect.

Other relevant works on learning quantum processes are Ref. [20], which deals with estimating local properties of the output of an unknown circuit, with guarantees on the average error for certain distribution of input states, Ref. [21], which shows how to efficiently learn the matrix elements of a channel in the Pauli basis, as a generalization of shadow tomography, and Refs. [22–24], which study how to estimate and test properties of an unknown channel with or without quantum memory.

Finally, the possibility of having generalization bounds for various loss functions evaluated on quantum circuits, using Rademacher complexities and covering numbers, has been considered extensively, especially with quantum machine learning applications in mind [25–32]. Relations between different combinatorial dimensions that cater to different modes of learning quantum states have also been explored in Ref. [33]. Our bounds on the covering numbers of processes described by quantum circuits are technically analogous to these existent results, which are devised for a different scenario with full control over the inputs. Indeed, such generalization bounds are valid if the sample complexity is defined as the number of unique data-point pairs $(x_i, \rho(x_i))$ seen by the learner, without restrictions on the number of times the pair $(x_i, \rho(x_i))$ can be actually produced (assuming that one is able to minimize the average of the loss function in some efficient way for every finite dataset). This analysis is not generically applicable in our case, since we do not have full control of the source and a good model for explaining the sample data cannot be in general inferred from finite samples. Obviously, since the covering numbers we consider are not constructed directly from the loss function but are dependent on the input data, they are generally worse in performance than those obtained for the cases with controlled input.

D. Learning algorithms: technical overview

In this section we explain the ideas behind our algorithms and their proofs (see also Fig. 2).

1. Threshold search for general product states (summary of Sec. III)

The key technical tool we establish is a generalization of the *Threshold Search* algorithm in Ref. [4], which takes as input many identical copies of a state σ, a collection of pairs consisting of a projector and a threshold $(\Pi^{(c)}, \theta_c)$, $c = 1, \ldots, m$ and reports if there is one projector whose average exceeds the threshold.

In our generalization, the algorithm takes as input only *a single state* ϱ—a product (of possibly *nonidentical*) states, $\varrho = \otimes_{i=1}^{n} \rho_i$—and a collection of pairs, each comprising a list of projectors and a threshold $(\{\Pi_i^{(c)}\}_{i\in[n]}, \theta_c)$; as output, the algorithm reports if there is one projector list whose average on the sample exceeds the threshold: $1/n \sum_{i=1}^{n} \text{Tr}[\rho_i \Pi_i^{(c)}] \geq \theta_c - \epsilon$. Lemma 1 gives the detailed formulation of this result. In the following we outline its main technical steps. For each list of projectors $\{\Pi_i^{(c)}\}_{j\in[n]}$,

one can construct the collective projectors that accept at least k times, denoted as $\{E_k^{(c)}\}_{k=0}^n$,

$$E_k^{(c)} := \sum_{\vec{v} \in \{0,1\}^n: ||\vec{v}||_1 = k} \bigotimes_{i=1}^n (\Pi_i^{(c)} + v_i(\mathbf{1} - 2\Pi_i^{(c)})). \quad (23)$$

When the projective measurement $\{E_k^{(c)}\}_k$ is executed on ϱ, the random variable associated to the outcome k, denoted as $\mathsf{T}^{(c)}$, is a sum of Bernoulli variables, $\mathsf{T}^{(c)} = \sum_{i=1}^n \mathsf{T}_i^{(c)}$, where $\mathsf{T}_i^{(c)} \sim (\mathrm{Tr}[\rho_i \Pi_i^{(c)}], 1 - \mathrm{Tr}[\rho_i \Pi_i^{(c)}])$. Via standard Chernoff bound arguments, to distinguish between

(a) $1/n \sum_{i=1}^n \mathrm{Tr}[\rho_i \Pi_i^{(c)}] \geq \theta_c$ and
(b) $1/n \sum_{i=1}^n \mathrm{Tr}[\rho_i \Pi_i^{(c)}] \leq \theta_c - \epsilon$

for a given c, it is sufficient to measure the binary event $\sum_{t > n(\theta_c - \epsilon/3)} E_t^{(c)}$, obtaining the correct answer with high probability with $n = O(1/\epsilon^2)$. However, after one of these measurements is applied ϱ can change significantly, so that the method of checking each threshold by measuring the threshold's accompanying projective measurement will not work naively. Therefore, we construct a more clever measurement that is "gentler": each threshold is associated with one such measurement. One performs the list of measurements sequentially on the same state and the measurement disturbs the state in a controlled way when it rejects. We now explain how such measurements are constructed in more detail.

From the projectors $\{E_k^{(c)}\}_k$, one can construct the events B_c:

$$B_c := \sum_{t=1}^n \Pr(\mathsf{X} + t > \theta n) E_t^{(c)}, \quad (24)$$

where X is an exponential random variable, $p_\lambda(\mathsf{X} = x) := \lambda e^{-\lambda x}$, for some $0 < \lambda < 1$. When B_c is measured on ϱ, the probability of accepting is $\Pr(\mathsf{X} + \mathsf{T}^{(c)} > \theta n)$. The exponential random variable has the role of smoothing the event B_c, so that the probability of accepting is still exponentially suppressed in n when $1/n \sum_{i=1}^n \mathrm{Tr}[\rho_i \Pi_i^{(c)}] \geq \theta_c - \epsilon$, but at the same time, when B_c rejects, the state ϱ does not change a lot. More explicitly, we can show in Theorem 9 that for $\lambda = 1/D\sqrt{n}$ and some constants $D > 0$:

$$\mathrm{Tr}[\varrho B_c] \leq e^{\left(-\frac{\sqrt{n}}{D}\left(\theta - \frac{1}{n}\sum_{i=1}^n \mathrm{Tr}[\rho_i \Pi_i^{(c)}]\right)\right) + e/(2D^2)}, \quad (25)$$

and, if $\mathrm{Tr}[\varrho B_c] < 1/4$, with

$$\varrho|_{\sqrt{\mathbf{1} - B_c}} := \frac{\sqrt{\mathbf{1} - B_c}\varrho\sqrt{\mathbf{1} - B_c}}{\mathrm{Tr}[\varrho(\mathbf{1} - B_c)]}$$

being the postmeasurement state conditioned on rejecting B_c, for some constant $C > 0$:

$$d_{\mathrm{tr}}\left(\varrho, \varrho|_{\sqrt{\mathbf{1} - B_c}}\right) \leq C\mathrm{Tr}[\varrho B_c]. \quad (26)$$

Note that the last equation is a consequence of the gentle measurement lemma for the fidelity (e.g., Proposition 2.2 in Ref. [4]), together with a bound on the χ^2 divergence of the distributions of $\mathsf{X} + \mathsf{T}^{(c)}$ and $\mathsf{X} + \mathsf{T}^{(c)}$ conditioned on $\mathsf{X} + \mathsf{T}^{(c)} \leq \theta n$.

Based on the previous definitions, the algorithm (Algorithm 1) for threshold search consists in measuring the events B_c, $c = 1, \ldots, m$ in sequence until a first acceptance c^* occurs. Then c^* is declared as the index of the list of projectors with expectation value above threshold. Lemma 1 shows that $n = O(\log^2(em)/\epsilon^2)$ is a sufficient condition for Algorithm 1 to succeed with probability larger than 0.03. The proof is based on a quantum union bound for sequential measurement proved in Ref. [4].

In fact, our contribution is to note that the proofs of Ref. [4] can be adapted to the case of nonidentical states and projectors, using the fact that the concentration properties of the Poisson binomial distribution are sufficient to reproduce the same argument of Ref. [4] with some appropriate adjustments. We do not make additional comments here on the proof, and we redirect to Ref. [4] for additional explanations on the idea behind the algorithm and remarks on its connections with adaptive data analysis.

2. Quantum empirical risk minimization (summary of Sec. IV)

Projector-valued concept classes (Theorem 1 and Theorem 2)—The main idea that lets us perform ERM for projector-valued functions is to observe that, in this case, the empirical risk is the average of expectation values of a list of projectors. Hence, we can directly apply the techniques of Sec. I D 1 to check if empirical risks are above or below a threshold. Furthermore, for our purposes, it is sufficient to consider concept classes of finite cardinality. Thus, in the following we assume that our concept class can be described as m lists of projectors $\{\Pi_i^{(c)}\}_{i \in [n]}, c \in [m]$. Our Algorithm 2 for ERM works as follows.

(I) *Sampling-without-replacement step:* given an input $\varrho = \otimes_{i=1}^n \rho_i$ a sufficiently large number of random batches $\varrho_s = \otimes_{k=1}^l \rho_{s,k}$ of size l, obtained by sampling without replacement from the list $i = 1, \ldots, n$. In the same way, one selects corresponding lists of projectors $\{\Pi_{s,k}^{(c)}\}_{k \in [l]}$ from the initial list $\{\Pi_j^{(c)}\}_{j \in [n]}$, for all c.

Via an argument based on concentration inequalities for sampling without replacement, we show that this procedure has the effect that the empirical risk on each batch ϱ_s, i.e., $1/l \sum_{j=1}^l \mathrm{Tr}[\rho_{s,j} \Pi_{s,j}^{(c)}]$, is ϵ-close to that on the full training set ϱ. The precise relation between the batch size l, the number of batches, the batch-risk approximation error ϵ, and the probability of error of this procedure is given in Lemma 3.

(II) *Binary search for optimal threshold:* the empirical risks take values in $[0, 1]$, therefore we start with a candidate minimum empirical risk 0.5, and use Algorithm 1 on the first batch to check if there is a candidate below this threshold. If we find one, we guess a new minimum empirical risk at 0.25, otherwise we guess 0.75 (these are approximate values, see the technical treatment for the details). We iterate this procedure via binary search, using a new batch for each new threshold, and selecting the upper or lower halves of the candidate interval, until we get to approximate the value of the empirical risk with precision ϵ: we will terminate with a number of step $O(\log(1/\epsilon))$.

Since Algorithm 2 is guaranteed to work only with a constant probability, we may need to repeat it a certain number of times to ensure we are able to get a candidate below threshold with high probability, if there is one. Moreover, we also need to check that the candidate risk is indeed below threshold, which is done by a further measurement of the corresponding empirical average.

Crucially, since in all of these steps we cannot reuse batches ϱ_s, the empirical risks must be close between different batches, which is guaranteed by the sampling-without-replacement step. By a union bound on the probability of the several types of error, we can obtain our guarantee for ERM, Theorem 1.

A similar approach can be used to obtain Theorem 2, viewing ERE as a generalization of *shadow tomography*. Shadow tomography [3] consists of the following task: given n copies of an unknown quantum state σ, and a list of projectors $\Pi_1, \ldots \Pi_m$, output approximations of $\mathrm{Tr}(\rho \Pi_c) \ \forall c \in [m]$. The Threshold Search algorithm of Ref. [4] was indeed used to obtain the state-of-the-art sample complexity for shadow tomography.

The idea is to keep a list of candidate intervals for the true values of $\mathrm{Tr}[\sigma \Pi_c]$ for each c, and use their extremes as thresholds in the Threshold Search algorithm, to be run with projectors $\Pi_1, \ldots \Pi_m$ as well as $\mathbf{1} - \Pi_1, \ldots \mathbf{1} - \Pi_m$. If one of the true values of $\mathrm{Tr}[\sigma \Pi_c]$ is far from the candidate values, one such c^* with this property will be found after a certain number of attempts, with high probability. If this happens, the list of candidates for the true expectation values is updated. Given a specific way to compute the candidate intervals, the process is guaranteed to terminate with intervals of size $O(\epsilon)$, containing the true values, after $\tilde{O}(\log d/\epsilon^3)$ rounds [3].

Our Algorithm 3 generalizes the IID case with two main modifications, as follows.

(I) *Looking for bad estimates in the general product states case:* with our generalization of Threshold Search for general product states and our sampling-without-replacement step, we can follow the same

scheme for ERE: given a list of candidate values for $1/n \sum_{i=1}^{n} \mathrm{Tr}[\rho_i \Pi_i^{(c)}]$, search for c^* such that the true value of $1/n \sum_{i=1}^{n} \mathrm{Tr}[\rho_i \Pi_i^{(c)}]$ is ϵ-far from the candidate.

(II) *Update rule for the general product states' case:* we generalize Aaronson's update procedure. We stress that in both Aaronson's and our generalization this step can be carried out by a classical computer, even if the computation involves quantum states.

In the original work, the estimates $\mu_{t,c}$ for the projector Π_c at update step t are obtained starting from q copies of a maximally mixed state, that is from the state $\rho_0^* := (\mathbf{1}/d)^{\otimes q}$. At each step, the estimates $\mu_{t,c}$ are the expectation of the empirical average of Π_c on ρ_t^*, which is also the predicted fraction of acceptances when Π_c is measured on each subsystem $\mathcal{H}_i$ in the tensor product $\bigotimes_{i=1}^{q} \mathcal{H}_i$. If at step $t = 1, \ldots, T$ some c^* is detected as associated to a prediction smaller (larger) than the true value, ρ_t^* is obtained postselecting ρ_{t-1}^* on the event for which measuring Π_c on each subsystem accepts a fraction of times slightly larger (smaller) than what previously predicted. Clearly, this makes the estimates progressively more accurate.

For the general product states case, the main change is that we start from the following classical-quantum state at step $t = 0$:

$$\rho_0^* := \left(\frac{1}{n} \sum_{i=1}^{n} |i\rangle\langle i| \otimes \frac{\mathbf{1}}{d} \right)^{\otimes q}, \qquad (27)$$

where $|i\rangle\langle i|$ are orthonormal projectors of an auxiliary system. At each step ρ_t^* is used as a guess for the classical-quantum state

$$\sigma = \left(\frac{1}{n} \sum_{i=1}^{n} |i\rangle\langle i| \otimes \rho_i \right)^{\otimes q}, \qquad (28)$$

in the following sense. Note that the quantities $1 - R_{\varrho_s}(c)$ can be expressed as expectation values of projectors

$$\Pi^{(c)} := \sum_{i=1}^{n} |i\rangle\langle i| \otimes \Pi_i^{(c)} \qquad (29)$$

on σ, $1 - R_{\varrho_s}(c) = \mathrm{Tr}[\Pi^{(c)} \sigma]$. The empirical average of $\Pi^{(c)}$ on $\mathcal{H}^{\otimes q}$ is $\overline{\Pi^{(c)}} := 1/q(\Pi^{(c)} \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \Pi^{(c)} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \mathbf{1} \otimes \cdots \otimes \Pi^{(c)})$. It follows that $\mathrm{Tr}[\Pi^{(c)} \sigma] = \mathrm{Tr}[\overline{\Pi^{(c)}} \sigma^{\otimes q}]$. ρ_t^* is a guess for $\sigma^{\otimes q}$ in the sense that, at each step t, the estimates of the risks are $\mu_{c,t} = \mathrm{Tr}[\overline{\Pi^{(c)}} \rho_t^*]$. The algorithm now runs exactly as

Aaronson's. The key point to conclude is that

$$\rho_0^* = \rho_0^{\otimes q} = \frac{1}{d^q}\sigma^{\otimes q} + \left(1 - \frac{1}{d^q}\right)\omega, \qquad (30)$$

for some positive semidefinite and trace-1 ω. Therefore, with probability at least $1/d^q$, ρ_0^* behaves as σ^q, which would accept all the postselection updates with high probability. On the other hand, if some estimate is incorrect at time t, one can show that ρ_t^* should reject with high probability. These two facts give rise to contradiction unless the number of necessary rounds is $\tilde{O}(\log d/\epsilon^3)$.

The proof of Theorem 2 is then obtained through a union bound over all possible sources of errors of Algorithm 3. We also note that, for projector-valued functions, the estimation of the empirical risks is similar to the *diverse-state setting* for shadow tomography considered in Ref. [17], where the state is a general product state but the projectors do not change with the subsystems.

State-valued concept classes (Theorem 3)—Here the loss function used is different from the projector-valued case: it is trace distance, and not overlap. This prevents us from immediately recycling the technique used previously, which works only to estimate expectation values.

Also in this case, it will suffice for the moment to consider finite cardinality classes, described by lists of states $\{\sigma_c(s)\}_{s\in[n]}$, $c \in [m]$. Equivalently, we can encode this information into a collection of classical-quantum states $\sigma_c = 1/n \sum_{k=1}^{n} |s\rangle\langle s| \otimes \sigma_c(s)$, while the information about the input $\varrho = \otimes_{s=1}^{n} \rho(s)$ is encoded in the classical-quantum state $\sigma = 1/n \sum_{k=1}^{n} |s\rangle\langle s| \otimes \rho(s)$. Then, the empirical risk for ϱ and the hypothesis $\sigma_c(s)$ is simply $d_{\text{tr}}(\sigma, \sigma_c)$.

The key idea is again inspired by Ref. [4], which used it for a task named *quantum hypothesis selection*. They used the following observations:

(a) The trace distance between pairs of states σ_i, σ_j can be written as the difference of two expectation values via Helstrom's theorem. Indeed, defining $A_{ij}(s) := (\sigma_i - \sigma_j)_+$, where $(\cdot)_+$ is the projector on the positive part of the argument, we have that

$$d_{\text{tr}}(\sigma_i, \sigma_j) = \text{Tr}(\sigma_i A_{ij}) - \text{Tr}(\sigma_j A_{ij}). \qquad (31)$$

(b) If a state ρ is close to σ_k, so are the expectation values. A good guess for ρ is then the state σ_{k^*} such that

$$k^* = \underset{k\in[m]}{\text{argmin}}\,\max_{i<j} |\text{Tr}[\rho A_{ij}] - \text{Tr}[\sigma_k A_{ij}]|. \qquad (32)$$

Indeed, suppose that i^* is such the true minimizer of $d_{\text{tr}}(\sigma_{i^*}, \rho) \le \eta$. Then we have that

$$\max_{i<j} |\text{Tr}[\rho A_{ij}] - \text{Tr}[\sigma_{k^*}A_{ij}]| \qquad (33)$$

$$\le \max_{i<j} |\text{Tr}[\rho A_{ij}] - \text{Tr}[\sigma_{i^*}A_{ij}]| \le \eta \qquad (34)$$

and via triangle inequalities

$$d_{\text{tr}}(\sigma_{k^*}, \rho) \le d_{\text{tr}}(\sigma_{k^*}, \sigma_{i^*}) + d_{\text{tr}}(\sigma_{i^*}, \rho) \qquad (35)$$

$$= |\text{Tr}(\sigma_{k^*}A_{k^*i^*}) - \text{Tr}(\sigma_{i^*}A_{k^*i^*})| + \eta \qquad (36)$$

$$\le |\text{Tr}(\sigma_{k^*}A_{k^*i^*}) - \text{Tr}(\rho A_{k^*i^*})|$$
$$+ |\text{Tr}(\rho A_{k^*i^*}) - \text{Tr}(\sigma_{i^*}A_{k^*i^*})| + \eta \le 3\eta. \qquad (37)$$

(c) Finally, if $\text{Tr}[\rho A_{ij}]$ are known only with precision ϵ, the above procedure is robust, and allows k^* to be found such that $d_{\text{tr}}(\rho, \sigma_{k^*}) \le 3\eta + 2\epsilon$ from approximate estimation of the expectation values of the Helstrom projectors.

Our crucial observation is that for the classical-quantum states σ_c, the Helstrom projectors are $A_{ij} := \sum_{s\in[n]} |x\rangle\langle x| \otimes A_{ij}(s)$, with $A_{ij}(s) := (\sigma_i(s) - \sigma_j(s))_+$. In turn, their expectation values can be estimated via our ERE algorithm for the lists of projectors $\{A_{ij}(s)\}_{s\in[n]}$, $i < j$, given ϱ as input. This reduces our procedure for learning with state-valued functions to a postprocessing of Algorithm 3 for ERE for projector-valued functions and proves Theorem 3.

3. Statistical learning for classical-quantum processes (summary of Sec. V)

Building on the ERM algorithms just reviewed, we are able to give sufficient conditions for the minimization of the true risk in both cases of projector- and state-valued functions. The algorithm works as follows.

(a) From the knowledge of the classical variables x_i for all i, we can construct an ϵ-net of the function class using the appropriate pseudometric: for two projector-valued functions $\Pi^{(c)}(x)$ and $\Pi^{(c')}(x)$, their pseudodistance on the data is $1/n \sum_{i=1}^{n} ||\Pi^{(c)}(x_i) - \Pi^{(c')}(x_i)||$, while for two state-valued functions $\sigma_c(x)$ and $\sigma_{c'}(x)$, their pseudodistance on the data is $1/n \sum_{i=1}^{n} d_{\text{tr}}(\sigma_c(x_i), \sigma_{c'}(x_i))$. The cardinality of the ϵ-net will be bounded by the $\gamma_{1,\infty}$ covering number in the case of projectors and by the $\gamma_{1,1}$ covering number in the case of states (see Definition 1).

(b) We then run ERM on $\varrho = \otimes_{i=1}^{n} \rho(x_i)$ using as the concept class the ϵ-net found at the previous step.

Let us observe that our ERM algorithms are guaranteed to work if the covering numbers $\gamma_{1,q}$ grow slowly with n, as they take the place of m in Theorems 1, 2, and 3. On the other hand, via classical statistical learning theory, uniform

convergence of the estimated empirical risk to the true risk is controlled by the covering numbers of the loss function, which depend on the unknown states $\rho(x)$.

Nevertheless, we are able to show that these covering numbers can be bounded by $\gamma_{1,q}$, which does not depend on $\rho(x)$ and is in principle computable by the learner. Therefore, we get a sufficient condition, independent of the data, for learning in terms of the growth of $\gamma_{1,q}$ with n, proving Theorem 4. Similarly, Theorem 5 can be obtained by running the ERE algorithm of Theorem 2 and checking that uniform convergence is also satisfied when $\gamma_{1,\infty}$ grows slowly with n.

E. Outlook

Our goal is to learn nature. In the past few years, quantum information processing tools have been fundamental to building a theory of learning states produced by quantum circuits. However, the circuit picture may not be the most natural one to describe all quantum mechanical processes. Our work represents a first incursion into the territory of developing a statistical learning theory for physical processes beyond quantum circuits. Our intention is to build a general theory of what we can learn via quantum processing, going beyond the terra firma of quantum circuits and what they can model, reaching characterization protocols in quantum mechanics that may be best expressed outside the circuit model, for instance, metrology, sensing, calibration, and verification, and eventually building a unified foundation for designing physical experiments.

II. PRELIMINARIES

A. Notation

We will consider random variables X valued in the set $\mathcal{X}$, denoting as $\mathcal{D}(x) = \mathrm{Pr}_{x \sim \mathcal{D}}(\mathsf{X} = x)$ the probability that the random variable takes value $x \in \mathcal{X}$; accordingly, $\mathrm{Pr}_{x \sim \mathcal{D}}(E)$ is the probability of an event $E \subseteq \mathcal{X}$. We will denote the set $\{1, \ldots, n\}$ as $[n]$. We denote the n-fold cartesian product of $\mathcal{X}$ as $\mathcal{X}^n$, elements in it as vectors $\vec{x} \in \mathcal{X}^n$, and we use the notation $|\vec{x}| := n$ to refer to the length of the vector. $\mathcal{D}^n$ denotes the probability distribution of $\vec{x}$. We denote a Hilbert space of dimension d by $\mathcal{H}^{(d)}$. We further denote by $\mathcal{L}(\mathcal{H}^{(d)})$ the set of linear operators on $\mathcal{H}^{(d)}$, and by $D(\mathcal{H}^{(d)}) \subseteq \mathcal{L}(\mathcal{H}^{(d)})$ the set of density matrices, that is, the subset of $\mathcal{L}(\mathcal{H}^{(d)})$, which is positive semidefinite and has unit trace. These matrices describe quantum states. For brevity, the Hilbert space of n qubits is denoted $\mathcal{H}_n$.

An arbitrary valid quantum operation on quantum states can be expressed as a quantum channel Φ, i.e., a completely positive trace-preserving map, and we denote the output of a channel applied to ρ as $\Phi[\rho]$. A special case of quantum channels are unitary channels, defined from a *unitary matrix* U (meaning $UU^\dagger = U^\dagger U = \mathbf{1}$). An application

of a unitary U to the state ρ results in the quantum state $U\rho U^\dagger$. Any quantum channel in finite dimension can be expressed in Kraus representation as $\Phi[\rho] = \sum_{i=1}^m K_i \rho K_i^\dagger$ for a suitable finite set of operators $\{K_i\}$.

In order to extract *classical* information out of a quantum state, one can perform a POVM, which is specified by a set of m positive semidefinite matrices $\{E_1, \ldots, E_m\}$ satisfying $\sum_i E_i = \mathbf{1}$. A measurement on a state ρ using such a POVM returns a classical outcome $i \in [m]$ with probability $\mathrm{Tr}[E_i \rho]$. For any operator $0 \le E \le \mathbf{1}$, we can define a two-outcome POVM $\{E, \mathbf{1} - E\}$, which we say that implements the measurement of the *event* E, and the probability associated to E is denoted as in the classical case as $\mathrm{Pr}(E) = \mathbb{E}_\rho[E]$. We will similarly use the notation $\mathbb{E}_\rho[A] := \mathrm{Tr}[\rho A]$ to express expectation values of operators A. We will denote the standard deviation of a classical random variable T as $\mathrm{stddev}[\mathsf{T}] := \sqrt{\mathbb{E}[(\mathsf{T} - \mathbb{E}[\mathsf{T}])^2]}$, where the expectation value is with respect to the probability distribution of T.

We will focus on *classical-quantum* states, i.e., states that can be written as $\rho = \sum_{x \in \mathcal{X}} \mathcal{D}(x) |x\rangle \langle x| \otimes \rho(x)$, where $\rho(x)$ are states of $\mathcal{H}^{(d)}$ and $\mathcal{D}$ is a probability distribution on $\mathcal{X}$, $\{|x\rangle\}_{x \in \mathcal{X}}$ are an orthonormal basis of a Hilbert space that we denote $\mathcal{H}_\mathcal{X}$. Note that for any operator $A \in \mathcal{L}(\mathcal{H}_\mathcal{X} \otimes \mathcal{H}^{(d)})$ and a classical-quantum state $\rho \in D(\mathcal{H}_\mathcal{X} \otimes \mathcal{H}^{(d)})$, $\mathbb{E}_\rho[A] = \sum_{x \in \mathcal{X}} \mathcal{D}(x) \mathrm{Tr}[A(x)\rho(x)] = \sum_{x \in \mathcal{X}} \mathcal{D}(x) \mathbb{E}_{\rho(x)}[A(x)]$, where $A(x) := \langle x|A|x\rangle \in \mathcal{L}(\mathcal{H}^{(d)})$. For a set of n operators on a Hilbert space $\mathcal{H}$, $A = \{A_1, \ldots, A_n\}$, we denote as $\overline{A}$ the operator on $\mathcal{H}^{\otimes n}$ defined as $\overline{A} := A_1 \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes A_2 \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \mathbf{1} \otimes \cdots \otimes A_n$.

1. Distances between probability distributions, quantum states, and quantum channels

We consider the following distances, defined for probability distributions on discrete supports. Let P, Q be two probability measures on the same support $\mathcal{X}$.

(a) **Total variation distance:**

$$TV(P, Q) := \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)|$$

$$= \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|. \quad (38)$$

(b) **Chi-squared distance:**

$$d_{\chi^2}(P, Q) := \sum_{x \in \mathcal{X}} Q(x) \left(1 - \frac{P(x)}{Q(x)}\right)^2. \quad (39)$$

(c) **Bhattacharya Coefficient:**

$$BC(P, Q) := \sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)}. \quad (40)$$

We will also need notions of continuity for matrices. First of all, we introduce some matrix norms. Let $M \in \mathbb{C}^{d \times d}$, then:

(a) The **trace norm** of M is

$$\|M\|_1 := \mathrm{Tr}\left[\sqrt{M^\dagger M}\right] = \sum_{i=1}^{\mathrm{rank}(M)} \sigma_i(M), \quad (41)$$

where $\sigma_i(M)$ are the singular values of M.

(b) For $p \in [1, \infty)$, the **Schatten p-norm** of M is

$$\|M\|_p := [\mathrm{Tr}(\sqrt{M^\dagger M}^p)]^{\frac{1}{p}}, \quad (42)$$

and $\|M\|_\infty = \lim_{p \to \infty} \|M\|_p$.

(c) The **spectral norm** is the maximum singular value of M:

$$\|M\| := \max_{i \in [\mathrm{rank}(M)]} \sigma_i(M), \quad (43)$$

which coincides with $\|M\|_\infty$. This is also the operator norm induced by the 2-norm for vectors:

$$\|M\| = \max_{\|x\|_2 = 1, x \in \mathbb{C}^n} \|Mx\|_2. \quad (44)$$

We will also use the following fact:

Fact 1. For any two vectors $u, v \in \mathbb{C}^n$ (not necessarily normalized),

$$\|uv^\dagger\|_1 = \|u\|_2 \|v\|_2, \quad (45)$$

which can be verified by applying definitions.

We will need, in particular, some measures of distance between quantum states and channels, defined as follows. For ρ, σ two quantum states in $D(\mathcal{H}^{(d)})$:

(a) **Trace distance:**

$$d_{\mathrm{tr}}(\rho, \sigma) := \frac{1}{2}\|\rho - \sigma\|_1 = \frac{1}{2}\mathrm{Tr}\left[\sqrt{(\rho - \sigma)^\dagger(\rho - \sigma)}\right]. \quad (46)$$

Note that in analogy to Eq.(38), this is half of the trace norm of the difference between ρ and σ.

(b) **Fidelity:**

$$F(\rho, \sigma) := \|\sqrt{\rho}\sqrt{\sigma}\|_1 = \mathrm{Tr}\left[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right]. \quad (47)$$

(c) **Bures distance:**

$$d_{\mathrm{Bures}}(\rho, \sigma)^2 := 2(1 - F(\rho, \sigma)). \quad (48)$$

B. Sequential measurements

The algorithms we will develop make extensive use of sequential measurements, which require us to specify what is the state of the quantum system conditioned on the outcome of a measurement. We stick to the usual convention, which states that if the outcome of a POVM $\{E_i\}_{i=1}^r$ is i, the state after the measurement is $\rho_{|\sqrt{E_i}} := \sqrt{E_i}\rho\sqrt{E_i}/\mathrm{Tr}[E_i\rho]$. The following facts will be useful [34].

(a) [4,35] Let P be the probability distribution on $[r]$ determined by the measurement $\mathcal{M} = \{\Pi_1, \ldots, \Pi_r\}$ on ρ, and let Q instead be the distribution determined by $\mathcal{M}$ on $\rho_{|\sqrt{A}}$, where $A = \sum_{i=1}^r a_i \Pi_i$ for some $a_i > 0$. Then

$$F(\rho, \rho_{|\sqrt{A}}) = \frac{\mathbb{E}_\rho[\sqrt{A}]}{\sqrt{\mathbb{E}_\rho[A]}} = BC(P, Q). \quad (49)$$

(b) [36,37] For each $i = 1, \ldots, m$, let $\Pi_0^{(i)}$ be projectors and $\Pi_1^{(i)} = \mathbf{1} - \Pi_0^{(i)}$. The probability p of getting always outcome 1 for sequential measurements $\{\{\Pi_0^{(1)}, \Pi_1^{(1)}\}, \ldots, \{\Pi_0^{(m)}, \Pi_1^{(m)}\}\}$ on the state ρ satisfies

$$p = \mathbb{E}_\rho\left[(\Pi_1^{(1)} \cdots \Pi_1^{(m)})(\Pi_1^{(1)} \cdots \Pi_1^{(m)})^\dagger\right]$$

$$\geq 1 - 2\sqrt{\sum_{i=1}^m \mathbb{E}_\rho[\Pi_0^{(i)}]}. \quad (50)$$

(c) Lemma 2.5 in Ref. [4]: Let $E_1, \ldots, E_m$ be events and consider a sequence of two-outcome POVMs implementing them, on the state ρ, so that the state post-selected on the events occurring is $\rho_{|\sqrt{E_{i-1}} \cdots \sqrt{E_1}} = \frac{\sqrt{E_{i-1}} \cdots \sqrt{E_1}\rho\sqrt{E_1} \cdots \sqrt{E_{i-1}}}{\mathrm{Tr}[\sqrt{E_{i-1}} \cdots \sqrt{E_1}\rho\sqrt{E_1} \cdots \sqrt{E_{i-1}}]}$ Let $q_1 = \mathbb{E}_\rho[E_1]$ and $q_i = \mathbb{E}_{\rho_{|\sqrt{E_{i-1}} \cdots \sqrt{E_1}}}[E_i]$, for $i > 1$, while $p_i = \mathbb{E}_\rho[E_i] > 0$ for all $i \in [m]$. We have that denoting $p_{[k]} = \prod_{i \in [k]} p_i$, $q_{[k]} = \prod_{i \in [k]} q_i$,

$$|p_{[m]} - q_{[m]}| \leq 2 \times \sum_{i=1}^{m-1} p_{[i]} d_{\mathrm{tr}}(\rho_{|\sqrt{E_i}}, \rho). \quad (51)$$

(d) Lemma 2.6 in Ref. [4]: with the same notation as the previous point, let $p_0 = 1$, $\rho_0 = \rho$, and $\rho_i = \rho_{i-1|\sqrt{E_i}}$ for all $i \in [m]$.
For all $t \in [m]$, let s_t denote the probability of observing outcomes $E_1, \ldots, E_{t-1}, I - E_t$. It holds that

$$1 \leq \sqrt{q_m}F(\rho, \rho_m) + \sum_{i=1}^m \sqrt{s_i}\sqrt{1 - p_i}. \quad (52)$$

Moreover, given a subset $\mathcal{B} \in [m]$ such that $1 - p_i \leq (\eta/m)^2$ for any $i \in \mathcal{B}$, it follows that (see proof of Lemma 4.2 in Ref. [4]):

$$1 - \sqrt{q_m}F(\rho, \rho_m) - \eta \leq \sqrt{\sum_{i \in [m]/\mathcal{B}} s_i} \sqrt{\sum_{i \in [m]/\mathcal{B}} (1 - p_i)}. \tag{53}$$

(e) Fix any POVM $\{E_1, \ldots, E_n\}$, a classical random variable X with values in $\mathbb{R}$, and $\theta \in \mathbb{R}$. Furthermore, consider the classical random variable T, defined on a fixed quantum state ρ, which has the distribution $\Pr(\mathsf{T} = t) = \mathbb{E}_\rho[E_t]$.
The quantum event

$$B = \sum_{t=1}^{n} \Pr(\mathsf{X} + t > \theta)E_t \tag{54}$$

is such that when measured on a state ρ, it is accepted with probability $\mathbb{E}_\rho[B] = \text{Tr}[\rho B] = \Pr(\mathsf{X} + \mathsf{T} > \theta)$. Furthermore, by Eq. (49), B has the property that

$$F(\rho, \rho_{|\sqrt{1-B}}) = BC(\mathsf{T}, (\mathsf{T}|\mathsf{T} + \mathsf{X} \leq \theta)). \tag{55}$$

C. Naive expectation estimation

A fundamental result on the concentration of sums of random variables is the following multiplicative Chernoff bound, with $\mathsf{X}_1 \ldots \mathsf{X}_n$ independent random variables and $0 \leq \epsilon \leq 1$:

$$\Pr\left(\sum_{i=1}^{n} \mathsf{X}_i \geq (1 + \epsilon) \sum_{i=1}^{n} \mathbb{E}[\mathsf{X}_i]\right) \leq e^{-\epsilon^2 \sum_{i=1}^{n} \mathbb{E}[\mathsf{X}_i]/3} \tag{56}$$

$$\Pr\left(\sum_{i=1}^{n} \mathsf{X}_i \leq (1 - \epsilon) \sum_{i=1}^{n} \mathbb{E}[\mathsf{X}_i]\right) \leq e^{-\epsilon^2 \sum_{i=1}^{n} \mathbb{E}[\mathsf{X}_i]/2}. \tag{57}$$

A consequence of this bound is the following proposition, related to what has been called *naive expectation estimation* by Ref. [4], but with IID observables. We retain the name:

Proposition 1 (Naive expectation estimation). The random variable X obtained by measuring the observable $\overline{\Pi} := \Pi_1 \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \Pi_2 \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes \Pi_n$ on $\varrho = \rho_1 \otimes \cdots \otimes \rho_n$ satisfies, for $\epsilon < 1$,

$$\Pr_\varrho[|\mathsf{X} - \mathbb{E}_\varrho[\overline{\Pi}]| \geq n\epsilon] \leq 2e^{-n\epsilon^2/3}. \tag{58}$$

Note that $\mathbb{E}_\varrho[\overline{\Pi}] = \sum_{s=1}^{n} \mathbb{E}_{\rho_s}[\Pi_s] := n(1 - R(h))$. This implies that, for given θ and ϵ, and $n \geq \frac{27}{\epsilon^2} \log \frac{2}{\delta}$ this measurement allows us to distinguish the two cases

(a) $|R(h) - \theta| \leq \epsilon$,
(b) $|R(h) - \theta| \geq 2\epsilon$,

with probability at least $1 - \delta$.

D. Growth functions and covering numbers: measuring the effective size of a concept class

In classical statistical learning theory, conditions that guarantee the success of ERM to learn an unknown function in the concept class $\mathcal{C}$ are given in terms of effective measures of the size of $\mathcal{C}$. We will be interested in two such measures, the growth function and the covering number.

For a finite sequence $x_1, \ldots, x_n$ and a set of functions $\mathcal{C} = \{f_\alpha : \mathcal{X} \to \mathcal{Y}\}$, we can identify a set of equivalence classes given by grouping together the functions that give the same outcomes on the inputs $x_1 \cdots, x_n$. Naturally, if $\mathcal{Y}$ is finite, the cardinality of this set is finite. However, even if $\mathcal{Y}$ is infinite, $\mathcal{C}$ could still be such that there are only finitely many equivalence classes. Indeed, for a fixed function class, the maximum number of equivalence classes induced by a sample of length n is called the *growth function $G(n)$*. A formal definition of the growth function is as follows [13,14]:

Definition 2 (Growth function). Let $\mathcal{C} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a class of functions with target space $\mathcal{Y}$. For every subset $\Xi \subseteq \mathcal{X}$ define the **restriction of $\mathcal{C}$** to Ξ as

$$\mathcal{C}|_\Xi := \{f \in \mathcal{Y}^\Xi \mid \exists F \in \mathcal{C} \; \forall x \in \Xi : f(x) = F(x)\}. \tag{59}$$

The growth function G assigned to $\mathcal{C}$ is then defined for all $n \in \mathbb{N}$ as

$$G(n) := \max_{\Xi \subseteq \mathcal{X}: |\Xi| \leq n} |\mathcal{C}|_\Xi|.$$

By a slight abuse of notation, we will also write $\mathcal{C}|_{\vec{x}}$ for a concept class restricted to a domain given by the set of values appearing in $\vec{x}$. We will also want to define ϵ-nets over classes of functions and will be interested in their size.

Definition 3 (Covering number N_{in}). For a set $\mathcal{M}$ and a pseudometric [38] $d : \mathcal{M} \times \mathcal{M} \to [0, c]$, let $(\mathcal{M}, d)$ be a pseudometric space, let the sets $A, B \subseteq \mathcal{M}$ and fix $\epsilon > 0$.

The set $A \subseteq B$ is an internal ϵ-net of B if $\forall b \in B, \exists a \in A : d(a, b) \leq \epsilon$.

The ϵ-covering number of B, denoted by

$$N_{in}(\epsilon, B, d), \tag{60}$$

is the smallest cardinality of any internal ϵ-net of B.

The first pseudometric that will be of interest to us is the one built on the $\|\cdot\|_{p,\vec{x}}$ seminorm.

Definition 4 ($\|\cdot\|_{p,\vec{x}}$ seminorms). For any set $\mathcal{X}, \vec{x} \in \mathcal{X}^n$ and any function class $\mathcal{G} \subseteq [0,c]^{\mathcal{X}}$ define the $\|\cdot\|_{p,\vec{x}}$ seminorm on the linear span of $\mathcal{G}$ for $p \in [1,\infty)$ as

$$\|g\|_{p,\vec{x}} := \left(\frac{1}{n} \sum_{i=1}^{n} |g(x_i)|^p \right)^{1/p}, \qquad (61)$$

and note that $\|g\|_{\infty,\vec{x}} := \max_{i \in [n]} |g(x_i)|$. The corresponding pseudometric is then given by $d : (g_1, g_2) \mapsto \|g_1 - g_2\|_{p,\vec{x}}$.

Definition 5 (Loss-function covering numbers). Let $\mathcal{G} \subseteq [0,c]^{\mathcal{X}}$ be a class of real-valued functions. For positive integer p, the pth loss function covering number is

$$\Gamma_p(n, \epsilon, \mathcal{G}) := \max \left\{ N_{in}\left(\epsilon, \mathcal{G}, \|\cdot\|_{p,\vec{x}}\right) \mid \vec{x} \in \mathcal{X}^n \right\}. \quad (62)$$

We make three remarks:

(1) For a finite target space $\mathcal{Y} = 1, 2, \dots |\mathcal{Y}|$, and for any $\vec{x} \in \mathcal{X}^n$, observe that $\|g - f\|_{\infty,\vec{x}} < \epsilon < 1$ iff $g|_{\vec{x}} = f|_{\vec{x}}$. This follows from the definition of spectral norm. Thus

$$N_{in}\left(\epsilon, \mathcal{F}, \|\cdot\|_{\infty,\vec{x}}\right) = |\mathcal{F}|_{\vec{x}}| \qquad (63)$$

and hence

$$\Gamma_\infty(n, \epsilon, \mathcal{F}) = G(n). \qquad (64)$$

(2) The name of the above covering number comes from the fact that we will often be interested in choosing $\mathcal{G}$ to be the *induced loss function class*, defined below:

Definition 6 (Induced loss function class). For any function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ and loss function $L : \mathcal{Y} \times \mathcal{Y} \to [0,c]$ define

$$\mathcal{G}_{\mathcal{F},L} := \{g : \mathcal{X} \times \mathcal{Y} \to [0,c]$$
$$\mid \exists h \in \mathcal{F} : g(x,y) = L(y, h(x))\}. \qquad (65)$$

On a related note, observe that, when the loss function is $L(y,x) = |x - y|$, for any $h_1, h_2 \in \mathcal{F}$, their pseudodistance $\|h_1 - h_2\|_{1,\vec{x}}$ upper bounds the difference between their empirical risks, $\hat{R}(h_1) - \hat{R}(h_2)$. This fact will be crucial in the next few paragraphs.

We will also need to define covering numbers for classes mapping to operators instead of real intervals (in the following, $\mathcal{H}$ denotes some fixed Hilbert space). For this case we will use a different seminorm, defined for operator-valued functions:

Definition 7 ($\|\cdot\|_{p,q,\vec{x}}$ seminorms). For any set $\mathcal{X}, \vec{x} \in \mathcal{X}^n$ and any function class $\mathcal{C} \subseteq \{h : \mathcal{X} \to M\}$, where $M \subseteq \mathcal{L}(\mathcal{H})$, we define the $\|\cdot\|_{p,q,\vec{x}}$ seminorm on the linear span of $\mathcal{C}$ for $p \in [1,\infty), q \in [1,\infty)$ as

$$\|g\|_{p,q,\vec{x}} := \left(\frac{1}{n} \sum_{i=1}^{n} \|g(x_i)\|_q^p \right)^{1/p}, \qquad (66)$$

where $\|\cdot\|_q$ is a Schatten q norm, and note that $\|g\|_{\infty,q,\vec{x}} := \max_{i \in [n]} \|g(x_i)\|_q$. The corresponding pseudometric is then given by $d : (g_1, g_2) \mapsto \|g_1 - g_2\|_{p,q,\vec{x}}$.

Accordingly, we will define the following covering number:

Definition 8 (Operator-class covering numbers). Let $\mathcal{C} \subseteq \mathcal{L}(\mathcal{H})^{\mathcal{X}}$ be a class of operator-valued functions. For positive integers p, q, the (p,q)th operator-class covering number is

$$\gamma_{p,q}(n, \epsilon, \mathcal{C}) := \max \left\{ N_{in}\left(\epsilon, \mathcal{C}, \|\cdot\|_{p,q,\vec{x}}\right) \mid \vec{x} \in \mathcal{X}^n \right\}. \quad (67)$$

For the concept classes of interest $\mathcal{C} : \mathcal{C} \subseteq \{h : \mathcal{X} \to M\}$ where $M \subseteq \mathcal{L}(\mathcal{H})$, we elect to use the following natural operator-class covering numbers, which we can also relate to the corresponding loss-function covering numbers:

(1) **(Case 1: Quantum states)** In the case of M being quantum states, we can take $p = 1, q = 1$. The pseudodistance between two concepts becomes twice the average trace distance between the states, $\|g_1 - g_2\|_{1,1,\vec{x}} = \sum_{i=1}^{n} 1/n \|g_1(x_i) - g_2(x_i)\|_1$. Note that the empirical risk (defined via L_s) evaluated on the input source and the concepts g_1 and g_2 satisfies

$$|\hat{R}_s(g_1) - \hat{R}_s(g_2)|$$

$$= \frac{1}{n} \left| \sum_{i=1}^{n} L_s(\rho(x_i), g_1(x_i)) - \sum_{i=1}^{n} L_s(\rho(x_i), g_2(x_i)) \right| \qquad (68)$$

$$= \frac{1}{2n} \left| \sum_{i=1}^{n} \left(\|g_1(x_i) - \rho(x_i)\|_1 - \|g_2(x_i) - \rho(x_i)\|_1 \right) \right| \qquad (69)$$

$$\leq \frac{1}{2} \|g_1 - g_2\|_{1,1,\vec{x}}, \qquad (70)$$

where the inequality follows from the triangle inequality. This means that if $N_{in}(\mathcal{C}, \epsilon, \|\cdot\|_{1,1,\vec{x}})$ is an ϵ-net for $\mathcal{C}$, then one can construct an ϵ-net for $\mathcal{G}_{\mathcal{C},L_s}$, simply by associating every point $c \in$

$N_{in}(\mathcal{C}, \epsilon, ||\cdot||_{p,q,\vec{x}})$ with a point $g_c = L_s \circ c$, which is in $\mathcal{G}_{\mathcal{C},L_s}$. This implies in turn that

$$\Gamma_1(n,\epsilon,\mathcal{G}_{\mathcal{C},L_s}) \le \gamma_{1,1}(n,2\epsilon,\mathcal{C}) \le \gamma_{1,1}(n,\epsilon,\mathcal{C}). \tag{71}$$

(2) **(Case 2: Projectors)** In the case of M being projectors, we can take $p=1, q=\infty$. The pseudodistance between two concepts becomes the average operator norm distance between the projectors, $||g_1 - g_2||_{1,\infty,\vec{x}} = \sum_{i=1}^{n} 1/n ||g_1(x_i) - g_2(x_i)||_\infty$. Note that the empirical risk (defined via L_p) evaluated on the input source and the concepts g_1 and g_2 satisfies

$$|\hat{R}_p(g_1) - \hat{R}_p(g_2)| \tag{72}$$

$$= \left| \sum_{i=1}^{n} \frac{1}{n} L_p(\rho(x),g_1(x)) - \sum_{i=1}^{n} \frac{1}{n} L_p(\rho(x),g_2(x)) \right| \tag{73}$$

$$\le ||g_1 - g_2||_{1,\infty,\vec{x}}. \tag{74}$$

Similarly to the above, we can conclude that

$$\Gamma_1(n,\epsilon,\mathcal{G}_{\mathcal{C},L_p}) \le \gamma_{1,\infty}(n,\epsilon,\mathcal{C}). \tag{75}$$

Note that both properties above are a generalization of the decreasing property of covering numbers for real-valued function classes under composition with Lipschitz functions [13]. We need operator-class covering numbers since we can construct ϵ-nets with respect to the $||\cdot||_{p,q,\vec{x}}$ seminorm, but not with respect to the loss function directly, since the value of the loss cannot be obtained immediately from the data.

E. Statistical learning theorems

The principle of empirical risk minimization dictates that, once a loss function L with values in the interval $[0,c]$ has been fixed, in order to find a function in the concept class that minimizes the true risk relative to the unknown concept f,

$$R(h) := \mathbb{E}_{x\sim\mathcal{D}}[L(f(x),h(x))] \quad \text{(True risk)}, \tag{76}$$

it suffices with high probability to find one that minimizes the empirical risk over the given sample:

$$\hat{R}(h) := \frac{1}{n}\sum_{i=1}^{n} L(f(x_i),h(x_i)) \quad \text{(Empirical risk)}. \tag{77}$$

In the classical case, the utility of this principle comes from the fact that the true risk cannot be estimated directly from the sample, but the empirical risk can. Crucially, the

quality of the approximation and the rate of the above-mentioned convergence depend on the concept class under study.

In particular, *discrete-output* concept classes are very well-understood. In this special case, the rate of convergence of the empirical risk to the true risk is quantified by growth functions. Namely, a well-known fact (see, for instance, Refs. [11,13,14]) states that given a sample $S = (x_i, f(x_i))_{i=1}^{n}$, with probability at least $(1-\delta)$ with respect to repeated sampling of training data of size n we have

$$\forall h \in \mathcal{F}: |R(h) - \hat{R}(h)| \le c\sqrt{\frac{8\ln\left(G(2n)\frac{4}{\delta}\right)}{n}}. \tag{78}$$

However, in our setting, we would like to be able to characterize channels that map to a continuous set of states. Since the restrictions of such concept classes to a finite sample are potentially infinite, the above statement is not useful. Nevertheless, a crucial observation is that, even for infinite concept classes $\mathcal{F}$, we can obtain a PAC bound for learning $\mathcal{F}$ via the covering number of a class $\mathcal{G}$ related to $\mathcal{F}$. $\mathcal{G}$ is induced by composing the chosen loss function with the functions in $\mathcal{F}$, as earlier stated in Definition 6.

The learning theorem is then as follows:

Theorem 6 (PAC bound via uniform covering numbers (reported as in Ref. [14], also equivalent to Theorem 17.1 in Ref. [13]). For any concept class $\mathcal{F}$ and loss function L with values in $[0,c]$, define $\mathcal{G}_{\mathcal{F},L}$ as in Definition 6. For any $\epsilon > 0$ and any probability measure $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ it holds

$$\Pr_{S\sim\mathcal{D}^n}[\exists h \in \mathcal{F}: |R(h) - \hat{R}(h)| \ge \epsilon] \tag{79}$$

$$\le 4\Gamma_1(2n,\epsilon/8,\mathcal{G}_{\mathcal{F},L})e^{-\frac{n\epsilon^2}{32c^2}}, \tag{80}$$

where $S = ((x_i, y_i))_{i=1}^{n}$ is the training sample and $(x_i, y_i) \sim \mathcal{D}$.

That is to say, the empirical risk $\hat{R}$ converges to the true risk R depending on the speed of growth of Γ_1.

As noted above, for the loss functions we consider the covering number $\Gamma_1(n,\epsilon,\mathcal{G}_{\mathcal{F},L})$ can be bounded by the corresponding operator-class covering number $\gamma_{1,q}(n,\epsilon,\mathcal{C})$, which effectively controls our learning algorithms.

III. THRESHOLD SEARCH FOR NONIDENTICAL STATES

Recall the discussion in the introduction about the difficulties encountered in naively generalizing the strategy of classical ERM to our quantum setting: they stem from not having an arbitrary number of identical copies of a quantum state $\rho(x_i)$ and from not being able to naively estimate all the empirical risks at the same time. In this section,

we nevertheless introduce a tool for performing our "quantum" version of ERM: an adaptation of *quantum threshold search* [3,4]. The gist of the algorithm is to use identical copies of some state of interest ρ, to pick an (observable, threshold) pair from amongst a set of such pairs, with the property that the observable exceeds the threshold on ρ—or, reports that no such pair is available. We define it more formally:

Problem 1 (Quantum threshold search [3,4]). Given as input

(a) Parameters $\epsilon, \delta > 0$
(b) Access to *identical* copies of a state $\rho \in D(\mathcal{H}^{(d)})$
(c) A collection of known projectors $\{\Pi_c\}_{c=1}^m$ where $\Pi_c \in \mathcal{L}(\mathcal{H}^{(d)})$.
(d) A collection of known thresholds $\{\theta_c\}_{c=1}^m$ where $\theta_c \in [0,1]$.

Output either

(a) $\text{Tr}[\rho \Pi_c] > \theta_c - \epsilon$ for some particular c,
(b) $\text{Tr}[\rho \Pi_c] \leq \theta_c$ for all c,

with probability of a correct statement at least $1 - \delta$.

This section is devoted to presenting an algorithm for threshold search that relaxes the need for access to identical copies of ρ, solving the same task on *general product states*. Intuitively, this is necessary for our setting because the learner, lacking control over the input to the process, receives as examples the sequence $(x_i, \rho(x_i))_{i=1}^n$ where $\rho(x_i)$ are not identical to each other. Thus, in contrast to the measurement in Ref. [4], which works on $\rho^{\otimes n}$, our key tool will be a measurement on the product state $\rho(x_1) \otimes \ldots \otimes \rho(x_n)$. The measurement reports if a threshold has been exceeded. We will eventually use this measurement to perform ERM for our setting. We define the properties of this measurement in Lemma 1. Our general proof strategy follows closely that of Ref. [4], adapting when needed.

Lemma 1 (Quantum threshold search on nonidentical states). Given as input

(a) Parameter $1 > \epsilon > 0$.
(b) Access to *a single* product state $\varrho := \rho_1 \otimes \ldots \otimes \rho_n \in (D(\mathcal{H}^{(d)}))^{\otimes n}$, which is a product of generally *nonidentical* qudit states.
(c) A collection of lists of known projectors

$$\{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}_{c=1}^m, \tag{81}$$

where each $\Pi_i^{(c)} \in \mathcal{L}(\mathcal{H}^{(d)})$.
(d) A collection of known thresholds $\{\theta_c\}_{c=1}^m$ where $\theta_c \in [0,1]$.

If the projectors and thresholds obey the promise that

$$\frac{1}{n}\sum_{i=1}^n \text{Tr}[\Pi_i^{(c)}\rho_i] > \theta_c \tag{82}$$

for at least one i, there is an algorithm such that

(a) At each step $c = 1, \ldots, m$ it performs a two-outcome measurement $\{B_c, \mathbf{1} - B_c\}$ based on the projectors $\{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}$.
(b) If the measurement accepts B_c, then the algorithm halts and outputs c, otherwise it passes to $c + 1$.

This algorithm is such that if

$$(\log m + C_2)^2 < C_1 n \epsilon^2 \tag{83}$$

for appropriate constants $C_1, C_2 > 0$, it outputs c such that

$$\frac{1}{n}\sum_{i=1}^n \text{Tr}[\Pi_i^{(c)}\rho_i] \geq \theta_c - \epsilon \tag{84}$$

with probability at least 0.03.

In order to get this improvement, we need to generalize the measurement constructed in Ref. [4], which in a certain sense is gentler than the projectors, which can check if the expectation is above or below threshold from Proposition 1. Roughly speaking, gentle means that the state does not change much after the measurement, with high probability. Implementations of product measurements that are gentle on product states had already been obtained in Ref. [3], and then reconsidered by Ref. [4] with a simpler analysis on identical states, which gave an improvement in the sample complexity of shadow tomography [4]. In these implementations, a parameter λ can be increased in order to obtain a gentler version of a projective measurement, sacrificing the amount of information that the measurement can reveal. We go one step further and generalize the stronger statements of Ref. [4] so that they apply to our setting, which involves products of nonidentical states.

To do this we need to study the behavior, under perturbation, of a random variable which is the sum of nonidentical Bernoulli random variables:

$$\mathsf{T} = \sum_{i=1}^n \mathsf{T}_i \quad \text{where } \mathsf{T}_i \sim (p_i, 1 - p_i). \tag{85}$$

The distribution of T is also known as the *Poisson binomial distribution* and we denote this as $T \sim PB(p_1, \ldots p_n)$.

We also need to consider exponential random variables X, which have a density $p_\lambda(x) := \lambda e^{-\lambda x}$, for some $\lambda > 0$, and satisfy $\mathbb{E}[\mathsf{X}] = 1/\lambda$.

We first present an entirely classical theorem. This theorem says that a measurement that checks for the event that $T + X$ exceeds a certain threshold is gentle when it rejects, in the sense that it only *gently* perturbs the distribution of T. This is a generalization of Theorem 1.2 of Ref. [4], which shows that adding exponential noise to a *binomial* random variable allows for gentle measurements.

Theorem 7 (Gentle classical measurement on Poisson binomial distribution). Let $T \sim PB(p_1, \ldots, p_n)$, and write $q_i = 1 - p_i$. Assume that X is an independent exponential random variable with mean $1/\lambda$ at least stddev$[T] = \sqrt{\sum_{i=1}^{n} p_i q_i}$ (and also at least 1). Let $\mathfrak{B}$ be the event that $T + X > \theta n$, and assume that $\Pr[\mathfrak{B}] < 1/4$. Then

$$d_{\chi^2}((T \mid \overline{\mathfrak{B}}), T) \leq C \left(\Pr[\mathfrak{B}] \cdot \frac{\mathbf{stddev}[T]}{\mathbb{E}[X]} \right)^2,$$

for a sufficiently large constant $C > 0$.

Proof. See Appendix A. The proof of Theorem 7 requires modification of some key details of the original proof. The main observation is that the crucial properties of the binomial distribution used in the proof, i.e., the form of the generating function and the probability of exceeding the expectation, also hold, with some caveats, for the Poisson binomial distribution. ∎

Now let us turn to the quantum problem. The reason we must consider Poisson binomial random variables is that they describe the probability distribution of outcomes when measuring a sum of (possibly nonidentical) local projectors on a product quantum state. The link to the learning setting is that the local projectors are exactly the ones in the set $\{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}$, which describe the cth hypothesis under consideration, in a way we will make formal in the next section.

In the agnostic learning setting, the learner receives the string of classical-quantum examples $(x_1, \rho(x_1)) \otimes \ldots \otimes (x_n, \rho(x_n))$. The quantum part of these examples are n nonidentical quantum states. We assume there is at least one hypothesis whose empirical risk on the examples goes below some threshold; this is captured by the promise in Lemma 1, that there is at least one set of projectors $\{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \text{Tr}[\Pi_i^{(c)} \rho_i] > \theta_c. \qquad (86)$$

In order to perform ERM, therefore, it suffices to find some c satisfying such a guarantee. The next theorem (Theorem 8) shows that the learner can perform a gentle *quantum* measurement, in the sense we next define, on the product state, in order to find such c with high probability. This is

also the measurement referred to in Lemma 1. It is sequential and adaptive, where the learner is presented with a projectors-threshold pair $(\{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}, \theta_c)$ at each step, and responds by making an appropriate measurement on her state (which is reused for multiple measurements) depending on past measurement outcomes. The cth measurement depends on the presented projectors-threshold pair and the outcomes of the first $c - 1$ measurements.

Theorem 8 should be viewed as a quantum counterpart to the classical Theorem 7, as it introduces the gentle quantum measurement that is at the core of our learning algorithm. Here we show that to every list of projectors $\{\Pi_1, \ldots, \Pi_n\}$, one can associate a gentle quantum observable B that, similarly, only perturbs a quantum state by a small amount, if a measurement of B rejects. In Theorem 7, the gentleness was in the sense that $d_{\chi^2}((T|\overline{\mathfrak{B}}), T)$ was bounded by $\Pr(\mathfrak{B})^2$. In the quantum Theorem 8, the gentleness is in the sense that the fidelity between the pre- and post-rejecting-measurement quantum state is exactly $BC((T|\overline{\mathfrak{B}}), T)$: this is analogous to Lemma 3.4 of Ref. [4]. The two classical distance measures on probability distributions that we have mentioned are related via the inequality

$$1 - BC(P, Q) \leq d_{\chi^2}(P, Q) \qquad (87)$$

for probability distributions P, Q.

Theorem 8 (Gentle quantum measurements on nonidentical product states). Fix $0 \leq \theta \leq 1$. Let X be any classical random variable taking values in $[0, \infty)$. For any list of projectors $\{\Pi_1, \ldots, \Pi_n\}$, $\Pi_i \in \mathcal{L}(\mathcal{H}^{(d)})$, there exists a quantum event $B \in L((\mathcal{H}^{(d)})^{\otimes n})$ such that when B is measured against a product state ϱ, i.e.,

$$\varrho := \rho_1 \otimes \ldots \otimes \rho_n, \qquad (88)$$

we have

$$\mathbb{E}_\varrho[B] = \Pr(T + X > \theta n), \qquad (89)$$

where $T \sim PB(p_1, \ldots, p_n)$ and $p_i = \mathbb{E}_{\rho_i}[\Pi_i]$. Furthermore,

$$F(\varrho, \varrho|_{\sqrt{1-B}}) = BC((T|T + X \leq \theta n), T). \qquad (90)$$

Proof. Observe that we may write the distribution of T explicitly as

$$\Pr[T = t] = \sum_{\vec{v} \in \{0,1\}^n : ||\vec{v}||_1 = t} \prod_{i=1}^{n} p_i^{v_i} (1 - p_i)^{(1-v_i)}. \qquad (91)$$

Then consider the projectors $\{E_k\}_{k=0}^{n}$,

$$E_k := \sum_{\vec{v} \in \{0,1\}^n : ||\vec{v}||_1 = k} \otimes_{i=1}^{n} (\Pi_i + v_i(\mathbf{1} - 2\Pi_i)). \qquad (92)$$

It is easy to see that the event B

$$B := \sum_{t=1}^{n} \Pr(\mathsf{X} + t > \theta n) E_t \tag{93}$$

fulfills Eq. (89). The "furthermore" part of the statement of the lemma follows immediately from Eq. (55). ∎

Theorem 8 applies to X being any classical random variable. Now we specialize to X being an exponential random variable. This specification allows us to obtain a bound on the Bures distance between the state before and after the rejecting measurement, together with a bound on the probability of B. For the latter, at variance with Ref. [4], we employ here a second-order Lagrange remainder that allows us to bypass the reduction to a fixed threshold, thus simplifying the overall proof procedure.

Theorem 9 (Gentle quantum measurements on nonidentical product states via exponential noise). Let $\varrho := \rho_1 \otimes \ldots \otimes \rho_n \in L((\mathcal{H}^{(d)})^{\otimes n})$ and consider a list of projectors $\{\Pi_1, \ldots, \Pi_n\}$, $\Pi_i \in \mathcal{L}(\mathcal{H}^{(d)})$. Let $1 > \lambda > 0$, and let $\theta \in [0, 1]$ be an arbitrary threshold. Let $\mathsf{T} \sim PB(p_1, \ldots, p_n)$ where $p_i = \mathrm{Tr}[\rho_i \Pi_i]$, and let the exponential random variable X be as in Theorem 7.

Then there exists a quantum event $B \in (\mathbb{C}^{d \times d})^{\otimes n}$ such that

$$\mathbb{E}_\varrho[B] = \Pr(\mathsf{T} + \mathsf{X} > \theta n), \tag{94}$$

and

$$d_{\mathrm{Bures}}\left(\varrho, \varrho|_{\sqrt{1-B}}\right) \lesssim \mathbb{E}_\varrho[B] \cdot \frac{\mathrm{stddev}[\mathsf{T}]}{\mathbb{E}[\mathsf{X}]}. \tag{95}$$

Moreover,

$$\mathbb{E}_\varrho[B] \leq \exp\left(-n\lambda\left(\theta - \frac{1}{n}\sum_{i=0}^{n} p_i - e\frac{\lambda}{2}\right)\right). \tag{96}$$

Proof. Equation (95) is an easy consequence of Theorems 7, 8 and Eq. (87).

Moreover, Eq. (96) follows from

$$\mathbb{E}_\varrho[B] := \Pr[\mathsf{T} + \mathsf{X} > \theta n] \tag{97}$$

$$= \sum_{t=0}^{n} \Pr(\mathsf{T} = t) \Pr(\mathsf{X} > \theta n - t) \tag{98}$$

$$\leq \mathbb{E}[\exp(-\lambda(\theta n - \mathsf{T}))] \tag{99}$$

$$= \exp(-\lambda\theta n)\mathbb{E}[\exp(\lambda\mathsf{T})] \tag{100}$$

$$= \exp(-\lambda\theta n)\prod_{i=0}^{n}(1 - p_i + p_i e^\lambda) \tag{101}$$

$$\leq \exp(-\lambda\theta n)\left(\frac{1}{n}\sum_{i=0}^{n}(1 + p_i(e^\lambda - 1))\right)^n \tag{102}$$

$$\leq \exp(-\lambda\theta n)\left(1 + \frac{\lambda}{n}\sum_{i=0}^{n}p_i + e\lambda^2/2\right)^n \tag{103}$$

$$\leq \exp(-\lambda\theta n)\exp\left(\lambda\sum_{i=0}^{n}p_i + en\lambda^2/2\right) \tag{104}$$

$$= \exp\left(-n\lambda\left(\theta - \frac{1}{n}\sum_{i=0}^{n}p_i - e\lambda/2\right)\right) \tag{105}$$

as desired. In the first inequality, we used that $\Pr[\mathsf{X} > t] \leq \exp(-\lambda t)$; in the fourth equality we used the MGF of a Poisson binomial random variable; in the second inequality we used the AM-GM inequality; in the third inequality we used that, from the second-order Lagrange remainder:

$$e^\lambda = 1 + \lambda + \lambda^2 \frac{e^{t^*}}{2} \leq 1 + \lambda + e\lambda^2/2,$$

with $t^* \in [0, \lambda]$. Finally, in the fourth inequality, we used that $1 + x \leq e^x$ for $x \in \mathbb{R}$. ∎

With these ingredients, we can prove Lemma 1. The algorithm that solves quantum threshold search is as follows:

ALGORITHM 1. Quantum threshold search on nonidentical states (ThresholdSearch)

Parameters: $\epsilon, \delta, D > 0$.
Promise: Of the pairs $\{(\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}, \theta_c)\}_{c=1}^{m}$, there is some $c \in [m]$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}[\Pi_i^{(c)}\rho_i] > \theta_c \tag{106}$$

Initialize: $\varrho^{(0)} \leftarrow \rho_1 \otimes \ldots \otimes \rho_n$.

1: **for** $c = 1, \ldots, m$ **do**
2:    Upon being presented with the pair

$$(\{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}, \theta_c) , \tag{107}$$

    measure the two-outcome POVM $\{B_c, \overline{B}_c := \mathbf{1} - B_c\}$, constructed as in Theorem 8 with threshold $\theta_c - \epsilon$ and with $1/\lambda = D\sqrt{n}$, on $\varrho^{(c-1)}$.
3:    If the measurement accepts, output c and break. If the measurement rejects, let $\varrho^{(c)}$ denote the post-measurement quantum state.
4: **end for**
5:    If none of the measurements were accepted, output "pass on all".

In Lemma 1, we claimed that Algorithm 1 is such that if

$$(\log m + C_2)^2 < C_1 n\epsilon^2, \tag{108}$$

for an appropriate constants C_1 and C_2, (and when $\epsilon < 1$), with probability at least 0.03, the algorithm halts and outputs a projector exceeding the threshold, i.e., it halts on i such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}[\Pi_i^{(c)}\rho_i] \geq \theta_c - \epsilon. \tag{109}$$

We now prove this. We will adopt the following notation, which mirrors the notation in Ref. [4]. For $c = 1, \ldots, m$, let

(1) $\mathsf{T}^{(c)} = \sum_{i=1}^{n} \mathsf{T}_i^{(c)}$, with $\mathsf{T}_i^{(c)} \sim (\mathbb{E}_{\rho_i}[\Pi_i^{(c)}], 1 - \mathbb{E}_{\rho_i}[\Pi_i^{(c)}])$ Bernoulli variables, implying $\mathsf{T}^{(c)} \sim PB(\mathbb{E}_{\rho_1}[\Pi_1^{(c)}], \ldots, \mathbb{E}_{\rho_n}[\Pi_n^{(c)}])$.
(2) $p_c = \mathbb{E}_{\varrho}[B_c]$;
(3) $\varrho^{(0)} = \rho_1 \otimes \ldots \otimes \rho_n$ and $\varrho^{(c)}$ be the quantum state after the cth measurement, *conditioned* on the event $\overline{B}_j$ occurring for all $1 \leq j \leq c$;
(4) $r_c = \mathbb{E}_{\varrho^{(c-1)}}[\overline{B}_c]$ be the probability that the event $\overline{B}_i$ occurs assuming all the events $\overline{B}_j$ with $1 \leq j \leq c - 1$ occurred;
(5) $q_c = r_1 \cdots r_c$ be the probability that *all* of the events $\overline{B}_j$ with $1 \leq j \leq c$ occur;
(6) $s_c = q_{c-1} \cdot \mathbb{E}_{\varrho_{c-1}}[B_c]$ be the probability of observing outcomes $\overline{B}_1 \cdots \overline{B}_{c-1} B_c$.

As in Ref. [4], our proof works by bounding the probabilities of two bad events. We now state what these events are and give an intuition for why their probabilities should be bounded, before we formalize the intuition.

(1) The algorithm outputs a *false negative*: it passes on all projector-threshold pairs, even though there was one that fulfilled the promise.
(2) The algorithm outputs a *false positive*: it outputs c that actually does not fulfil the promise, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}[\Pi_i^{(c)}\rho_i] \leq \theta_c - \epsilon. \tag{110}$$

Intuition for why these two probabilities are bounded: Suppose hypothetically that the algorithm had the luxury to measure the given projector on R *fresh* copies of $\rho_1 \otimes \ldots \otimes \rho_n$ each time. Then the promise of Lemma 1, and Chernoff's bound, ensure that with $R = O(1/\epsilon^2)$-many copies, one could identify some projector fulfilling (109) with high probability, so that neither false negative nor false positive would come to pass.

However, in Algorithm 1, we do not use R fresh copies on each of the m iterations of the For loop; instead, we make successive measurements on a single copy. Then the damage lemma [Eq. (51)], together with the properties of the specially constructed B_c in Theorem 8, ensures that the usage of a "damaged" copy does not affect the probability of acceptance on a new measurement too much, relative to using R fresh copies. The caveat is that "low damage" is only guaranteed if all of the previous measurements have rejected; care must then be taken to account for this.

Proof of Lemma 1. **Controlling the false negative probability**—the promise of Lemma 1 is that there is some $c \in [m]$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}[\Pi_i^{(c)}\rho_i] > \theta_c. \tag{111}$$

For this particular c, let us bound the probability of acceptance of the corresponding measurement B_c (which we previously denoted p_c).

To do so, let us recall that Theorem 9 guarantees that the B_c's are constructed in such a way that their measurement outcome statistics follow from that of a classical Poisson binomial random variable, $\mathsf{T}^{(c)} := \sum_{i=1}^{n} \mathsf{T}_i^{(c)}$. This is a sum of independent and not identically distributed binary random variables. Their concentration is captured by the multiplicative Chernoff bound in Eqs. (56) and (57). Applying this to the sum $\mathsf{T}^{(c)} := \sum_{i=1}^{n} \mathsf{T}_i^{(c)}$, which has expectation $\mathbb{E}[\mathsf{T}^{(c)}] = n\overline{p}_c := \sum_{i=1}^{n} \mathrm{Tr}[\rho_i \Pi_i^{(c)}]$, we have, by the promise, $\theta_c \leq \overline{p}_c$ and for any non-negative random variable X:

$$p_i := \mathbb{E}_{\varrho}[B_c] = \Pr(\mathsf{T}^{(c)} + \mathsf{X} > (\theta_c - \epsilon)n) \tag{112}$$

$$\geq \Pr(\mathsf{T}^{(c)} > (\theta_c - \epsilon)n) \geq \Pr(\mathsf{T}^{(c)} > (\overline{p}_c - \epsilon)n) \tag{113}$$

$$\geq 1 - e^{-n\epsilon^2/(2\overline{p}_i)} \geq 1 - e^{-1/4} \tag{114}$$

for $n = \Omega(1/\epsilon^2)$, where the second equality on the first line is by Eq. (94), and the first inequality on the third line follows from Eq. (57).

Since we have now established that there is some i such that $1 - p_i \leq e^{-1/4}$, so there must also be some minimum $t \in [m]$ such that $(1 - p_1) \cdots (1 - p_t) \leq e^{-1/4}$. If $t = 1$, then $r_1 = 1 - p_1 \leq e^{-1/4} \leq 4/5$, that is $p_1 \geq 1/5$.

Otherwise, $t > 1$. Since t is minimal, $(1 - p_1) \cdots (1 - p_{t-1}) > e^{-1/4}$. Taking logs on both sides, this implies that $\sum_{c=1}^{t-1} p_c \leq -\log \prod_{c=1}^{t-1}(1 - p_c) \leq 1/4$ [where we have used the inequality $x < -\log(1 - x)$]. If $t > 1$, by Eq. (51), Theorem 7 and the fact that $d_{\mathrm{tr}}(\rho, \sigma) \leq d_{\mathrm{Bures}}(\rho, \sigma)$ [by the standard Fuchs–Van de Graaf

inequality [35]], we have

$$|(1 - p_1) \cdots (1 - p_t) - q_t| \tag{115}$$

$$\leq 2 \sum_{c=1}^{t-1} d_{\mathrm{tr}}(\varrho_{|\sqrt{1-B_c}}, \varrho) \tag{116}$$

$$\lesssim 2 \sum_{c=1}^{t-1} \mathbb{E}_\varrho[B_c] \cdot \frac{\mathrm{stddev}[\mathsf{T}^{(c)}]}{\mathbb{E}[\mathsf{X}]} \tag{}$$

$$\leq \left(\sum_{c=1}^{t-1} p_c\right) \cdot \frac{\sqrt{n}}{\mathbb{E}[\mathsf{X}]} \leq \frac{1}{4} \cdot \frac{\sqrt{n}}{\mathbb{E}[\mathsf{X}]}, \tag{117}$$

and also

$$|(1 - p_1) \cdots (1 - p_{t-1}) - q_{t-1}| \lesssim \frac{1}{4} \times \frac{\sqrt{n}}{\mathbb{E}[\mathsf{X}]}. \tag{118}$$

Recall that X is an exponential random variable with $\mathbb{E}[\mathsf{X}] = D\sqrt{n}$, meaning that choosing D large enough we have $q_t \leq 4/5$ and $q_{t-1} \geq 3/4$. The former means that the algorithm will output some $c \leq t$ with probability larger than $1/5$.

Controlling the false positive probability—now we restrict our attention to the first t projectors presented to the algorithm, since we have established in the previous section that the algorithm terminates with more than some constant probability after at most t iterations of the For loop. We now need to show that the algorithm, with sufficiently high probability, does not output a c with $\mu_c = 1/n \sum_{i=1}^n \mathrm{Tr}[\Pi_i^{(c)} \rho_i] \leq \theta_c - \epsilon$. Let us denote the set of such c as $\mathcal{B}$ (for Bad):

$$\mathcal{B} = \left\{ c \leq t \,\middle|\, \mu_c \leq \theta_c - \epsilon \right\}.$$

By Eq. (96) in Theorem 9, whenever $\mu_c \leq \theta_c - \epsilon$ (i.e., for all $c \in \mathcal{B}$), we have that

$$p_c \leq \exp\left(-n\lambda\left(\theta_c - \mu_c - e\lambda/2\right)\right)$$
$$\leq \exp\left(-D^{-1}\sqrt{n}\epsilon + eD^{-2}/2\right), \tag{119}$$

where we used that $\mathbb{E}[\mathsf{X}] = 1/\lambda = D\sqrt{n}$ for some $D > 0$. Therefore $p_c \leq (\eta/m)^2 < 1/5$ if $(2\log(m/\eta) + eD^{-2}/2)^2 \leq D^{-1}n\epsilon^2$, and $\eta \leq 0.01$.

The argument is now identical to Ref. [4], proof of Lemma 4.2. There, using Eq. (53), $F(\varrho, \varrho_t) < 1$, and $\sum_{c \in [t]/\mathcal{B}} p_c < 1/4$, $q_t < 4/5$, one obtains

$$\frac{1}{2}\sqrt{\sum_{c \in [t]/\mathcal{B}} s_c} \geq 0.99 - \sqrt{4/5} \tag{120}$$

$$\rightarrow \sum_{c \in [t]/\mathcal{B}} s_c \geq 4 \cdot (0.99 - \sqrt{4/5})^2 \geq 0.03. \tag{121}$$

Since $\sum_{c \in [t]/\mathcal{B}} s_c$ is the probability that the algorithm returns an index $c \in [t]$ with $\mu_c \geq \theta_c - \epsilon$, it follows that the algorithm is correct with probability at least 0.03. ∎

IV. QUANTUM EMPIRICAL RISK MINIMIZATION

In this section we present our main algorithm to perform ERM for both projector- and state-valued classes of quantum processes.

A. Empirical risk minimization for projector-valued functions

We first show how to estimate the empirical risk of a fixed number of projector-valued functions on a given product state. The main theorem we prove is the following.

Theorem 10 (Quantum empirical risk minimization—Theorem 1, refined). Given access to a product state

$$\varrho = \rho_1 \otimes \cdots \otimes \rho_n \tag{122}$$

and a collection of lists of projectors $\{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}_{c=1,\ldots,m}$, with

$$\mu_c = \frac{1}{n} \sum_{i=1}^n \mathrm{Tr}[\rho_i \Pi_i^{(c)}], \tag{123}$$

(i.e., μ_c is $1 - R_\varrho(c)$, where the second term is the empirical risk of concept c on the entire product state) there is an algorithm which outputs c^* together with an estimate $\hat{\mu}_{c^*}$ of μ_{c^*} such that

$$\Pr(|\hat{\mu}_{c^*} - \max_{c \in [m]} \mu_c| \geq \epsilon \cup |\hat{\mu}_{c^*} - \mu_{c^*}| \geq \epsilon) \leq \delta, \tag{124}$$

if n is large enough. In fact we can take

$$n = \frac{1}{\epsilon^2} \log\frac{1}{\delta} \log^2\frac{1}{\epsilon}$$
$$\times O\left(\max\left(\log\left(\frac{m}{\delta}\log\frac{1}{\delta}\log^2\frac{1}{\epsilon}\right), (\log m + C_1)^2\right)\right) \tag{125}$$

for some $C_1 > 0$.

Let us now convey the intuition behind the claimed algorithm (which is Algorithm 2). Observe that the task of finding the concept that attains the maximum overlap, $\max_{c \in [m]} \mu_c$, can be accomplished by binary search over the interval $[0, 1]$—namely, we search for the largest threshold such that there is at least one risk above it, out of the m possible ones. The remainder of this section gives the

details of how to do so. The key idea is to divide the given product state ϱ into *blocks* of $l = n/2Tk$ states, that is

$$\varrho = \varrho_1 \otimes \varrho_2 \cdots \otimes \varrho_{2Tk} \quad \text{where} \quad \varrho_s = \rho_{s,1} \otimes \cdots \otimes \rho_{s,l},$$
(126)

and take the block size l large enough so that the average risk on each block concentrates towards μ_c for each c. This gives us sufficient confidence to apply our Algorithm 1 (ThresholdSearch) on each block to check if there exists some concept that exceeds the current candidate threshold. Depending on the results of this check, we adjust the candidate threshold accordingly. The guarantees of this Algorithm (Algorithm 2) are given in Lemma 2. Subsequently, in Lemma 3, we derive the block size l we need to ensure the desired concentration.

Lemma 2 (Quantum empirical risk minimization given large product states). Given access to $2Tk$ blocks of states as in Eq. (126) and a collection of lists of projectors $\{\Pi_{s,j}^{(c)}\}_{c=1,\ldots,m,s=1,\ldots,2Tk,j=1,\ldots,l}$, suppose that the following conditions hold:

(1) For appropriate constants $C_1, C_2 > 0$

$$(\log m + C_2)^2 < C_1 l \epsilon^2.$$
(127)

(2) At the same time

$$l > \frac{\log(Tk/\delta)}{\epsilon^2},$$
(128)

for large enough $T = O(\log \frac{1}{\epsilon})$ and $k = O\left(\log \frac{1}{\delta} \log \frac{1}{\epsilon}\right)$.

(3) The numbers $0 \leq \mu_c \leq 1$ are approximations of the expected value of $\{\Pi_{s,j}^{(c)}\}_{s,j}$ with respect to the blocks of states for all $c \in [m]$, i.e., they satisfy

$$\left| \frac{1}{l} \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - \mu_c \right| \leq \epsilon/4, \ \forall c \in [m], s \in [2Tk].$$
(129)

Then there is an algorithm that outputs c^* and $\hat{\mu}_{c^*}$ such that

$$\Pr(|\hat{\mu}_{c^*} - \max_{c \in [m]} \mu_c| \geq 6\epsilon \cup |\hat{\mu}_{c^*} - \mu_{c^*}| \geq 6\epsilon) \leq \delta. \quad (130)$$

Here we note that "given large product states" in the title of the lemma refers to the requirement that the $n = 2Tkl$ is large enough to guarantee that Eq. (127) holds.

Proof. We use Algorithm 2, and the following notation:

(a) $h_s^{(c)} = \{\Pi_{s,1}^{(c)}, \ldots, \Pi_{s,n}^{(c)}\}$,

(b) $\overline{h_s^{(c)}} = \Pi_{s,1}^{(c)} \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \Pi_{s,2}^{(c)} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \mathbf{1} \otimes \cdots \otimes \Pi_{s,n}^{(c)}$,

(c) $R_{\varrho_s}(c) = 1 - \frac{1}{n}\mathbb{E}_{\varrho_s}[\overline{h_{c,s}}]$,

(d) $\mathsf{X}_{c,s}$ is the random variable obtained by measuring $\overline{h_{c,s}}$ on ϱ_s.

Algorithm in words: Algorithm 2 runs binary search to find an interval containing $\max_c \mu_c$, starting with the candidate interval $[0, 1]$, determining whether the desired value lies in the upper or lower half, and then updating the candidate interval to the relevant half and recursing. To determine in which half of the candidate interval the desired value lies, the algorithm uses up two blocks of samples, ϱ_{2s-1} and ϱ_{2s}: on the first block, we run ThresholdSearch, which also outputs a concept that exceeds the current candidate value θ_c. On the second block, we run a check, that confirms that this concept indeed exceeds the threshold

---

ALGORITHM 2. Learning projector-valued functions

Input: $2Tk$ product states $\{\varrho_s\}_{s=1,\ldots 2Tk}$ and $2Tkm$ sets of projectors $\{\{h_s^{(c)}\}_{c=1,\ldots,m}\}_{s=1,\ldots,2Tk}$, with $h_s^{(c)} = \{\Pi_{s,1}^{(c)}, \ldots, \Pi_{s,n}^{(c)}\}$.

Parameters: $\epsilon, \delta, k > 0$.
Initialize: $\theta = 1/2, low = 0, high = 1, failures = 0, s = 0$.

1: **while** $high - low \geq 6\epsilon$ **do**
2:     **if** $failures < k$ **then**
3:         $s+ = 1$
4:         **Threshold search**: Apply ThresholdSearch (Algorithm 1) on the set of projectors-threshold pairs $\{(h_{2s}^{(c)}, \theta - \epsilon)\}_{c=1}^m$ with the parameter $\epsilon/4$, to the product state ϱ_{2s-1}.
5:         **if Threshold search** doesn't output a concept **then**:
6:             $failures+ = 1$
7:         **else Threshold search** outputs concept c
8:             **Check**: measure $\overline{h_{2s}^{(c)}}$ on ϱ_{2s} and check if $\mathsf{X}_{c,2s} \geq n(\theta - 7/4\epsilon)$.
9:             **if Check** outputs 'yes' **then**:
10:                $low \leftarrow \theta - 2\epsilon, high \leftarrow high, \theta \leftarrow \frac{1}{2}(high + low)$ ▷ Update interval to upper half
11:                $failures \leftarrow 0$.
12:             **else Check** outputs 'no' and
13:                $failures+ = 1$
14:         **end if**
15:     **end if**
16:     **else** there have been k consecutive failures, so
17:         $low \leftarrow low, high \leftarrow \theta, \theta \leftarrow \frac{1}{2}(high + low)$ ▷ Update interval to lower half
18:         $failures \leftarrow 0$
19:     **end if**
20: **end while**
21: Output θ and the last selected concept, if there is one, otherwise pick it randomly.

θ_c (as ThresholdSearch only succeeds with probability 0.03). We declare a *failure* if by the end of this process, we have not received a "yes" from both algorithms; after k consecutive failures, we conclude that there was actually no concept exceeding the candidate threshold, and move on by decreasing the candidate threshold (line 17). Conversely, if at any point we receive a "yes" from both algorithms, we conclude that the candidate threshold was too low and increase it (line 10).

Error analysis: Let us now analyze the error in this algorithm. The algorithm will err if either line 17 or line 10 updates the candidate threshold wrongly. We are interested in the probability of either of these two events happening for a fixed candidate threshold; by a union bound, we will then multiply this probability by the number of candidate thresholds that are examined, which is $O(\log(1/\epsilon))$—as the interval does not get updated once it becomes smaller than 6ϵ, and every update approximately halves the size of the interval (see lines 10 and 17).

We thus define the following four error probabilities and their associated events:

(a) p_{FPTS}: ThresholdSearch outputs a false positive, i.e., it outputs that there is a concept above threshold when there is none.

(b) p_{FNTS}: ThresholdSearch outputs a false negative, i.e., it outputs "no concept above threshold" when, in fact, there was one.

(c) p_{FPC}: Check outputs a false positive, i.e., $X_{c,2s} \geq l(\theta - 7/4\epsilon)$ when in fact $\mu_c < \theta - 2\epsilon$.

(d) p_{FNC}: Check outputs a false negative, i.e., $X_{c,2s} < l(\theta - 7/4\epsilon)$ when in fact $\mu_c \geq \theta$.

Note that, by Lemma 1, it holds $p_{\text{FNTS}} < 0.97$, whereas p_{FNC} and p_{FPC} can be made exponentially small in the block size l by the multiplicative Chernoff bound (Proposition 1).

If line 10 updates wrongly, it can only be because ThresholdSearch outputs a concept even though there is no concept that exceeds the threshold *and* Check outputs 'yes' on that wrong concept. The probability of this happening for a given interval is at most $k\, p_{\text{FPTS}}\, p_{\text{FPC}}$, as there are k rounds where this could potentially happen and any such event triggers an update of the interval.

If line 17 updates wrongly, there must have been k consecutive failures when there actually was a concept above threshold. Each such failure is caused by one of the following events:

(a) Threshold search executes correctly at some point but Check falsely outputs no. The probability that this happens in k rounds is at most p_{FNC}. To see this, call τ_i the event that the first false negative check occurs at time i. $\{\tau_i\}$ are mutually exclusive events, and $\Pr(\tau_i) \leq (p_{\text{FNTS}})^{i-1}(1 -$

$p_{\text{FNTS}})p_{\text{FNC}}$. The probability that there is a false negative at some time i is $\sum_{i=1}^{k}\Pr(\tau_i) \leq p_{\text{FNC}}$.

(b) Threshold search wrongly outputs that no concept was above threshold (probability p_{FNTS}) for k consecutive times. The probability of these events is bounded by 0.97^k, as explained above.

Summing over the T rounds, the probability of error is then upper bounded as

$$p_{\text{err}} \leq T(0.97^k + kp_{\text{FPC}}(l) + p_{\text{FNC}}(l)). \tag{131}$$

With the choices made for the parameters of the algorithm, we obtain the guarantee given in the lemma statement. ∎

Next, we control the probability that the (possibly non-identical) product states give an accurate approximation to the expectation values of the projectors associated with the possible concepts, i.e., the probability that Eq. (129) is satisfied. The main technical ingredient is the following lemma, which shows, essentially, that Höffding's inequality is effective even for proving the concentration of the empirical mean of samples without replacement from a finite population.

Lemma 3. Let $\mathcal{Y} = \{1, \ldots, n\}$, $n \geq 3Kl$, and let the K sets of indices $\{\{X_{lk+i}\}_{i=1}^{l}\}_{k=0}^{K-1}$ be random samples drawn without replacement from $\mathcal{Y}$. Furthermore, consider m different finite populations of n numbers $\{\{x_{c,j}\}_{j=1}^{n}\}_{c=1}^{m}$, with $0 \leq x_{c,j} \leq 1$. From each population $\{x_{c,j}\}_{j=1}^{n}$ obtain K subsets of size l, as $\{\{x_{c,X_{lk+j}}\}_{j=1}^{l}\}_{k=0}^{K-1}$. Then it holds

$$\Pr\left(\max_{\substack{c=1,\ldots,m \\ k=0,\ldots,K-1}} \left|\frac{1}{l}\sum_{j=1}^{l} x_{c,X_{lk+j}} - \mu_c\right| \geq \epsilon\right) \tag{132}$$

$$\leq 2Kme^{-2l\epsilon^2/4}, \tag{133}$$

where $\mu_c = 1/n\sum_{j=1}^{n} x_{c,j}$.

The proof is given in the Appendix B (see the equivalent Theorem 19). With these two lemmas in hand, we are now able to prove that Algorithm 2 works for finding the empirical risk minimizer amongst projector-valued functions.

Proof of Theorem 10. Let $n \geq 6Tkl$ be the number of observations of the output of the unknown process (where we remind readers that each observation corresponds to a quantum state). That is, we set $K = 2Tk$ in Lemma 3. Sampling without replacement $2Tk$ lists of length l from $[n]$, let

the sth list define the product state

$$\varrho_s = \rho_{s,1} \otimes \cdots \otimes \rho_{s,l}, \qquad s = 1, \ldots, 2Tk \qquad (134)$$

and for each concept $c \in [m]$, the set of projectors

$$\{\Pi_{s,1}^{(c)}, \ldots \Pi_{s,l}^{(c)}\}, \qquad s = 1, \ldots, 2Tk. \qquad (135)$$

Now by identifying $x_{c,j} = \text{Tr}[\Pi_j^{(c)} \rho_j]$ in Lemma 3, we obtain the desired concentration

$$\left| \frac{1}{l} \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - \mu_c \right| \le \epsilon/4, \ \forall c \in [m], s \in [2Tk],$$
$$(136)$$

with probability $p_{\text{err}}^{(1)} \le 4Tkme^{-2l\epsilon^2/64}$,

If Eq. (136) is true, then condition Eq. (129) is satisfied and we can apply Algorithm 2 to obtain a good estimate with probability of error $p_{\text{err}}^{(2)} \le T(0.97^k + (k+1)e^{-l\epsilon^2/72})$.

We can make $p_{\text{err}}^{(1)} + p_{\text{err}}^{(2)} \le \delta$ with the choice of T, k as in Lemma 2 and $l = O(\max(\log(Tkm/\delta)/\epsilon^2), (\log m + C_2)^2/\epsilon^2))$. ∎

B. Empirical risk estimation for projector-valued functions

In this section we will show that if the size of a product state is sufficiently large with respect to the logarithm of the local dimension, we cannot only identify the concept yielding the highest minimum risk and estimate that risk, but in fact estimate the risks of *all* of the concepts at the same time. The key idea is built on the philosophy of shadow tomography: we find a function that predicts the expectation values

$$\mu_c = \frac{1}{n} \sum_{i=1}^{n} \text{Tr}[\rho_i \Pi_i^{(c)}], \qquad (137)$$

where c runs over the possible different concepts, given access to the product state

$$\varrho = \rho_1 \otimes \cdots \otimes \rho_n. \qquad (138)$$

In the case of lists of identical projectors ($\Pi_i^{(c)}$ independent of i), the task went under the name of *diverse-state setting* for shadow tomography in Ref. [17], where it was noted that the algorithm proposed there also works in this setting. Here we show that the improvements given by Ref. [4] carry through in this setting with the appropriate generalizations. We also make explicit an appropriate modification of the procedure of Ref. [3] to *update* the state guess, where a reference quantum state (possibly stored on a classical computer as a matrix) is used to compute the expectation values of the projectors, and updated to agree with the data from the experiment. Here we start with m lists of projectors

$$h_c = \{\Pi_1^{(c)}, \ldots, \Pi_n^{(c)}\}, \quad c \in [m] \qquad (139)$$

and an unknown product state

$$\varrho = \rho_1 \otimes \cdots \otimes \rho_n. \qquad (140)$$

As in the previous section we use the notation

(a) $h_{c,s} = \{\Pi_{s,1}^{(c)}, \ldots, \Pi_{s,l}^{(c)}\}, \quad \overline{h_{c,s}} = \Pi_{s,1}^{(c)} \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1}$
$+ \mathbf{1} \otimes \Pi_{s,2}^{(c)} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \mathbf{1} \otimes \cdots \otimes \Pi_{s,l}^{(c)}$
(b) $1 - R_{\varrho_s}(c) = \frac{1}{l}\mathbb{E}_{\varrho_s}[\overline{h_{c,s}}]$
(c) $\mathsf{X}_{h_{c,s}}$ is the random variable obtained by measuring $\overline{h_{c,s}}$ on ϱ_s.

Note that the quantities $1 - R_\varrho(c)$ can be also expressed as expectation values of projectors

$$\Pi^{(c)} := \sum_{i=1}^{n} |i\rangle\langle i| \otimes \Pi_i^{(c)} \qquad (141)$$

on the classical-quantum state

$$\sigma = \frac{1}{n} \sum_{i=1}^{n} |i\rangle\langle i| \otimes \rho_i. \qquad (142)$$

The algorithm works by keeping track of a classical estimate of σ that is updated sequentially. The estimate is initialized at time $t = 0$ as

$$\rho_0^* := \left(\frac{1}{n} \sum_{i=1}^{n} |i\rangle\langle i| \otimes \frac{\mathbf{1}}{d} \right)^{\otimes q}, \qquad (143)$$

and at each time t, ρ_t is obtained by picking one system of the q at random and computing the marginal of ρ_t^* in that system. The form of the state to update is the main difference between what we propose and the strategy in Ref. [3]. We give the proof in its entirety for convenience.

At each time step t, the algorithm is also provided with some $c \in [m]$ for which the current best estimate ρ_t makes poor predictions relative to σ, that is,

$$\mu_{c,t} = \mathbb{E}_{\rho_t}[\Pi^{(c)}] = \text{Tr}[\rho_t \sum_i |i\rangle\langle i| \otimes \Pi_i^{(c)}] \qquad (144)$$

is ϵ-far from

$$\mu_c = \mathbb{E}_\sigma[\Pi^{(c)}] = \sum_i \frac{1}{n} \text{Tr}[\Pi_i^{(c)} \rho_i]. \qquad (145)$$

(Later, in Lemma 5, we will explain how to find such a c.) The updating procedure then makes use of $\Pi^{(c)}$ to update

ρ_t^* to ρ_{t+1}^*. In doing so we can guarantee that at most a certain number T of updates will be required, in the following sense:

Lemma 4 (Update). There exists a sequential procedure to update a classical estimate for some unknown quantum state σ, that initializes the estimate at time $t = 0$ as ρ_0^* given in Eq. (143), and at time t, takes as input the current estimate ρ_t^* and some $c \in \mathcal{C}$ such that either

   (1) $\mu_c - \mathbb{E}_{\rho_t}[\Pi^{(c)}] \geq \epsilon$, or
   (2) $\mu_c - \mathbb{E}_{\rho_t}[\Pi^{(c)}] \leq -\epsilon$,

and outputs an updated estimate, ρ_{t+1}^*, such that after

$$T = O\left(\frac{\log d}{\epsilon^3}\left(\log \log d + \log \frac{1}{\epsilon}\right)\right)$$

updates, the estimate ρ_T fulfils

$$|\mu_c - \mathbb{E}_{\rho_T}[\Pi^{(c)}]| \leq \epsilon, \qquad \forall c \in [m]. \tag{146}$$

Proof. First of all, we construct the following projectors

$$\Pi^{(c)}(l) = \sum_{\vec{v} \in \{0,1\}^q : ||\vec{v}||_1 = l} \bigotimes_{i=1}^{n} (\Pi^{(c)} + v_i(\mathbf{1} - 2\Pi^{(c)})). \tag{147}$$

In other words, $\Pi^{(c)}(l)$ is the sum of all events that accept the projector $\Pi^{(c)}$ at *exactly* l points in $[q]$ and reject $\Pi^{(c)}$ at the remaining $q - l$ points in $[q]$. Finally, we define

$$\Pi^{(c),-}(r) = \sum_{l \leq r} \Pi^{(c)}(l), \tag{148}$$

$$\Pi^{(c),+}(r) = \sum_{l \geq r} \Pi^{(c)}(l). \tag{149}$$

In words, $\Pi^{(c),+}(r)$ is the sum of all events that reject $\Pi^{(c)}$ in *at least* r points in $[q]$ and reject $\Pi^{(c)}$ in all remaining points in $[q]$; $\Pi^{(c),-}(r)$ is the sum of all events that reject $\Pi^{(c)}$ in *at most* r points in $[q]$ and accept $\Pi^{(c)}$ at all remaining points in $[q]$. We now describe the update procedure. In case 1, we update ρ_t^* to ρ_{t+1}^* by postselecting on the event $F_t^+(c) = \Pi^{(c),+}((\hat{\mu}_{c,t} + \epsilon/2)q)$. In case 2, we then update ρ_t to ρ_{t+1} postselecting on the event $F_t^-(c) = \Pi^{(c),-}((\hat{\mu}_{c,t} - \epsilon/2)q)$. Define F_t as the appropriate accepting event at time t, i.e., $F_t := F_t^+$ in case 1 and $F_t := F_t^-$ in case 2.

We will now upper bound the probability that the first t postselection steps all succeed, following an argument of Ref. [3]. This is

$$p_t = \mathrm{Tr}[F_0 \rho_0^*] \ldots \mathrm{Tr}[F_{t-1} \rho_{t-1}^*], \tag{150}$$

and we will be able to upper bound it by $p_t \leq (1 - \epsilon)^{\Omega(t)}$. Indeed, we can show this by considering the random variable

$$\mathsf{X} = \text{number of acceptances resulting from measuring}$$

$$\left\{\Pi^{(c)}, \mathbf{1} - \Pi^{(c)}\right\} \text{ on } \rho_t^*. \tag{151}$$

Applying Markov's inequality, we can then see that, by writing $\overline{\Pi^{(c)}} := \Pi^{(c)} \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \Pi^{(c)} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \mathbf{1} \otimes \cdots \otimes \Pi^{(c)})$,

$$\mathbb{E}[\mathsf{X}] = \mathrm{Tr}[\rho_t^* \overline{\Pi^{(c)}}] = q \mathbb{E}_{\rho_t}[\Pi^{(c)}], \tag{152}$$

and, in case 1,

$$\Pr[\mathsf{X} \geq (\hat{\mu}_{c,t} + \epsilon/2)q] = \mathrm{Tr}[F_t^+(c)\rho_t^*] \tag{153}$$

$$\leq \frac{\hat{\mu}_{c,t}q}{(\hat{\mu}_{c,t} + \epsilon/2)q} \leq 1 - \Omega(\epsilon) \tag{154}$$

while an analogous calculation yields that in case 2,

$$\mathrm{Tr}[F_t^-(c)\rho_t^*] \leq 1 - \Omega(\epsilon). \tag{155}$$

This implies $p_t \leq (1 - \epsilon)^{\Omega(t)}$.

Now, we lower bound p_t. Hypothetically, suppose that at time t we were to apply the measurement F_t (which depends on which case we are in) to $\sigma^{\otimes q} = \left(\sum_{i=1}^{n} 1/n |i\rangle\langle i| \otimes \rho_i\right)^{\otimes q}$. By the promise that we are either in case 1 or 2, and by Chernoff's bound, at each step the measurements reject with probability $1 - \mathrm{Tr}[F_t \sigma^{\otimes q}] \leq e^{-\Omega(q\epsilon^2)}$. Applying the measurements in sequence, $F_t \ldots F_0$ also always accepts with high probability by the quantum union bound (e.g., Ref. [37]). In particular, the probability of accepting on every step, until step t, is

$$q_t \geq 1 - O\left(\sqrt{t} e^{-\Omega(q\epsilon^2)}\right). \tag{156}$$

But then, for every ρ_i, $i \in [n]$, it is possible to write $\mathbf{1}/d = 1/(d)\rho_i + (1 - 1/d)\omega_i$ for some state ω_i. Then, recalling that $\rho_0^* = (\sum_{i=1}^{n} 1/n |i\rangle\langle i| \otimes \mathbf{1}/d)^{\otimes q}$

$$\rho_0^* = \rho_0^{\otimes q} = \frac{1}{d^q} \sigma^{\otimes q} + \left(1 - \frac{1}{d^q}\right)\omega, \tag{157}$$

for some positive semidefinite and trace-1 ω.

This decomposition (specifically the fact that ω is PSD) elucidates the following lower bound:

$$p_t = \mathrm{Tr}[\rho_0^* F_0] \ldots \mathrm{Tr}[\rho_{t-1}^* F_{t-1}] \tag{158}$$

$$\geq \frac{1}{d^q}\left(1 - O\left(\sqrt{t} e^{-\Omega(q\epsilon^2)}\right)\right). \tag{159}$$

By taking $q = C/\epsilon^2 (\log \log d + \log(1/\epsilon))$ we have $p_t \geq 1/d^q (1 - \sqrt{t}\epsilon^2/\log d)$.

Putting this together with the upper bound $p_t \leq (1 - \epsilon)^{\Omega(t)}$, we need $T = O\left(\log d/\epsilon^3 \left(\log\log d + \log(1/\epsilon)\right)\right)$, where T is the number of updates after which we can estimate all the expectation values at the desired precision. ∎

To obtain c satisfying the promise of Lemma 4, we can use the following lemma, which is a simpler variant of Lemma 2.

Lemma 5. Given access to $2k$ product states

$$\varrho_s = \rho_{s,1} \otimes \cdots \otimes \rho_{s,l}, \qquad s = 1, \ldots, 2k, \qquad (160)$$

and a collection of lists of projectors $\{\Pi_{s,j}^{(c)}\}_{c=1,\ldots,m, s=1,\ldots,2k, j=1,\ldots,l}$ and numbers $0 \leq \mu_c \leq 1$, $c \in [m]$ such that

$$\left| \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - \mu_c \right| \leq \epsilon/4, \, \forall c \in [m], s \in [2k],$$

$$(161)$$

and numbers $\{\lambda_c\}_{c=1,\ldots,m}$. Then if we are guaranteed that

$$(\log m + C_2)^2 < C_1 l \epsilon^2, \qquad (162)$$

for appropriate constants C_1, C_2, and at the same time

$$l > \frac{\log(k/\delta)}{\epsilon^2}, \qquad (163)$$

for a large enough $k = O(\log\frac{1}{\delta}\log\frac{1}{\epsilon})$, there is an algorithm that, with probability larger than δ, if there exists some

c such that $|\mu_c - \lambda_c| \geq 2\epsilon$, either the algorithm declares failure or it outputs a c^* such that $|\mu_{c^*} - \lambda_{c^*}| \geq \epsilon/2$.

Proof. Use ThresholdSearch (Lemma 1) on the states ϱ_s, s odd with

(a) the list of projectors $\{\Pi_{s,1}^{(c)}, \Pi_{s,2}^{(c)}, \ldots, \Pi_{s,l}^{(c)}\}$ and threshold $\lambda_c + 7/4\epsilon$,

(b) together with the list of projectors $\{\mathbf{1} - \Pi_{s,1}^{(c)}, \mathbf{1} - \Pi_{s,2}^{(c)}, \ldots, \mathbf{1} - \Pi_{s,l}^{(c)}\}$ and threshold $1 - \lambda_c - 7/4\epsilon$,

$c \in [m]$, and precision parameter $\epsilon/4$. For each odd s, if c is output from the search, we measure

$$\Pi_{s+1,1}^{(c)} \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \Pi_{s+1,2}^{(c)} \otimes \cdots \otimes \mathbf{1}$$
$$+ \mathbf{1} \otimes \mathbf{1} \otimes \cdots \otimes \Pi_{s+1,l}^{(c)} \qquad (164)$$

on ϱ_{s+1}, with a resulting random variable $\mathsf{X}_{c,s+1}$.

If $|(\mathsf{X}_{c,s+1})/l - \lambda_c| > \epsilon$, we declare the check passed and output c. If all the checks are not passed, we declare failure.

The analysis is as follows: If there exists c such that $|\mu_c - \lambda_c| \geq 2\epsilon$, then by Eq. (127) we have $|1/l \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - \lambda_c| > 7/4\epsilon$, therefore either $1/l \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] \geq \lambda_c + 7/4\epsilon$, or $1/l \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\mathbf{1} - \Pi_{s,j}^{(c)}] \leq 1 - \lambda_c - 7/4\epsilon$. Since the promise of Threshold Search is fulfilled either for $\{\Pi_{s,j}^{(c)}\}_{j \in [l]}$ or $\{\mathbf{1} - \Pi_{s,j}^{(c)}\}_{j \in [l]}$, each time ThresholdSearch is performed, with probability larger than 0.03 it outputs a concept c such that $|1/l \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - \lambda_c| > 6/4\epsilon$. In this case, by Chernoff bound, Eq. (58) we have

$$\Pr(|\mathsf{X}_{c,s+1} - l\lambda_c| \leq l\epsilon) = \Pr(|\mathsf{X}_{c,s+1} - \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] + \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - l\lambda_c| \leq l\epsilon)$$

$$\leq \Pr\left(\left||\mathsf{X}_{c,s+1} - \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}]| - |\sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - l\lambda_c|\right| \leq l\epsilon\right) \leq \Pr(|\mathsf{X}_{c,s+1} - \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}]| \geq l\epsilon/2) \leq 2e^{-l\frac{\epsilon^2}{6}}.$$

$$(165)$$

On the other hand, if a concept is selected with $|\mu_c - \lambda_c| \leq \epsilon/2$, we have

$$\Pr(|\mathsf{X}_{c,s+1} - l\lambda_c| \geq l\epsilon) = \Pr(|\mathsf{X}_{c,s+1} - \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] + \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - l\lambda_c| \geq l\epsilon)$$

$$\leq \Pr(|\mathsf{X}_{c,s+1} - \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}]| + |\sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - l\lambda_c| \geq l\epsilon) \leq \Pr(|\mathsf{X}_{c,s+1} - \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}]|) \geq l\epsilon/2) \leq 2e^{-l\frac{\epsilon^2}{6}}.$$

$$(166)$$

ALGORITHM 3.   Empirical risk estimation

Input: Product states

$$\varrho_s = \rho_{s,1} \otimes \dots \otimes \rho_{s,2lk}, \qquad s \in [T] \;, \qquad (167)$$

Parameters: T, q.

1: Initialize on a classical computer, the classical estimate

$$\rho_0^* := \left(\sum_{s=1}^{n} |s\rangle\langle s| \otimes \frac{1}{d} \right)^{\otimes q} . \qquad (168)$$

2: **for** $t = 1, \dots, T$ **do**
3: $\quad c \leftarrow \mathsf{ThresholdSearch}$ on the state ϱ_s with projectors and parameters as described in the proof of Lemma 5. If $\mathsf{ThresholdSearch}$ declares failure then Break.
4: $\quad \rho_{t+1}^* \leftarrow \mathsf{Update}(\rho_t^*, c)$.
5: **end for**
6: Output estimates $\mu_{c,T} = \mathbb{E}_{\rho_T}[\Pi^{(c)}], \forall c \in \mathcal{C}$.

By an argument identical to the analysis of the error probability in Lemma 2, the probability of error is bounded as $p_{\mathrm{err}} \le 0.97^k + 2(k+1)e^{-l\frac{\epsilon^2}{6}}$, and the choice of k and l in the lemma statement makes it less than δ. ∎

Thus, our algorithm for ERE simply starts from an estimate dependent on the empirical distribution of measurement outcomes of the classical register, and then interleaves the $\mathsf{ThresholdSearch}$ and Update subroutines to progressively update this estimate. This is summed up in the following algorithm:

We can now prove the following.

Theorem 11 (Quantum empirical risk estimation for projector-valued functions (Theorem 2, refined)). Given access to a product state

$$\varrho = \rho_1 \otimes \dots \otimes \rho_n \qquad (169)$$

and a collection of lists of projectors $\{\Pi_1^{(c)}, \dots, \Pi_n^{(c)}\}_{c=1,\dots,m}$, with

$$\mu_c = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Tr}[\rho_i \Pi_i^{(c)}], \qquad (170)$$

there is an algorithm which outputs estimates $\hat{\mu}_c$ such that

$$\Pr(|\hat{\mu}_c - \mu_c| \ge 2\epsilon) \le \delta \qquad \forall c \in [m] \qquad (171)$$

if n is large enough; in fact we can take $n = \frac{Tk}{\epsilon^2} \times O(\max(\log(Tkm/\delta), (\log m + C_1)^2))$, with $T = O\left(\frac{\log d}{\epsilon^3} \left(\log\log d + \log\frac{1}{\epsilon} \right) \right)$ and $k = O(\log(T/\delta))$.

Proof. The algorithm for ERE runs as follows. Prepare the classical guess ρ_0^* as in Lemma 4. Divide the product states into T batches of $2k$ product states. For $t = 1, \dots, T$, run the algorithm of Lemma 5 on the corresponding batch. If a concept c is selected, use it as an update for the algorithm in Lemma 4 and continue to the next t, otherwise terminate and update the collection of μ_c as obtained from ρ_t^*.

By Lemma 3, if $n \ge 6Tkl$ we can obtain $2Tk$ samples without replacement of length l from $[n]$. By identifying $x_{c,i} = \mathrm{Tr}[\Pi_i^{(c)}\rho_i]$, we get $2Tk$ product states

$$\varrho_s = \rho_{s,1} \otimes \dots \otimes \rho_{s,l}, \qquad s = 1, \dots, 2Tk, \qquad (172)$$

and a collection of lists of projectors $\{\Pi_{s,j}^{(c)}\}_{c=1,\dots,m,s=1,\dots,2Tk,j=1,\dots,l}$ such that

$$\left| \frac{1}{l} \sum_{j=1}^{l} \mathbb{E}_{\rho_{s,j}}[\Pi_{s,j}^{(c)}] - \mu_c \right| \le \epsilon/4, \, \forall c \in [m], s \in [2Tk]. \qquad (173)$$

with probability $p_{\mathrm{err}}^{(1)} \le 2Tkme^{-2l\epsilon^2/64}$.

If Eq. (173) is true, then condition Eq. (161) is satisfied and we can apply the algorithm of Lemma 5 to obtain, if $\max_c |\hat{\mu}_c - \mu_c| \ge 2\epsilon$ an estimate $\hat{\mu}_c$, $|\hat{\mu}_c - \mu_c| \ge \epsilon/2$ with probability of error $p_{\mathrm{err}}^{(2)} \le (0.97^k + 2(k+1)e^{-l\epsilon^2/6})$. Otherwise, if $\max_c |\hat{\mu}_c - \mu_c| \le 2\epsilon$ and we do not get any $|\hat{\mu}_c - \mu_c| \ge \epsilon/2$ we are satisfied. Anyway, after $T = O\left(\log d/\epsilon^3 \left(\log\log d + \log(1/\epsilon) \right) \right)$ updates we also have $|\hat{\mu}_c - \mu_c| \le \epsilon/2$ by Lemma 4. The probability of error is then less than $Tp_{\mathrm{err}}^{(2)}$.

We can make $p_{\mathrm{err}}^{(1)} + Tp_{\mathrm{err}}^{(2)} \le \delta$ with T and k as in the theorem statement and

$$l = O(\max(\log(Tkm/\delta)/\epsilon^2, (\log m + C_2)^2/\epsilon^2))$$

and thus n as stated. ∎

C. Empirical risk minimization for state-valued functions

A key subroutine we will need to introduce is based on *hypothesis selection*, which is a way of choosing a classical or quantum probability distribution that best fits some observed data. We will use a generalized version of the algorithm of Ref. [4] for hypothesis selection (which applied to IID states) to find the empirical risk minimizer, with the loss given by the trace distance, in a set of candidate state-valued processes.

Theorem 12 (Quantum empirical risk minimization for state-valued functions (Theorem 3, refined)). Let $\mathcal{C} = \{\sigma_i :$

$[n] \to D(\mathcal{H}^{(d)})\}_{i=1}^m$ be a class of state-valued functions and

$$\varrho = \rho_1 \otimes \cdots \otimes \rho_n. \quad (174)$$

There exists an algorithm, which given ϱ outputs i such that

$$\frac{1}{n} \sum_{s=1}^n d_{\mathrm{tr}}(\sigma_i(s), \rho_s) \leq 3\eta + 4\epsilon, \quad (175)$$

where

$$\eta := \min_{i \in [m]} \frac{1}{n} \sum_{s=1}^n d_{\mathrm{tr}}(\sigma_i(s), \rho_s), \quad (176)$$

with probability of error less than δ if n is large enough, i.e., we can take in fact

$$n = \frac{Tk}{\epsilon^2} \times O(\max(\log(Tkm/\delta), (\log m + C_1)^2)),$$

with $T = O\left(\log d/\epsilon^3 \left(\log \log d + \log \frac{1}{\epsilon}\right)\right)$ and $k = O(\log(T/\delta))$.

Proof. Let us define the following classical-quantum states for $k = 1, \ldots, m$:

$$\sigma_k := \frac{1}{n} \sum_{s=1}^n |s\rangle\langle s| \otimes \sigma_k(s), \quad (177)$$

and

$$\rho := \frac{1}{n} \sum_{s=1}^n |s\rangle\langle s| \otimes \rho_s, \quad (178)$$

Then $\eta = \min_i d_{\mathrm{tr}}(\sigma_i, \rho)$ and the algorithm has to output a hypothesis k such that $d_{\mathrm{tr}}(\sigma_k, \rho) \leq 3\eta + \epsilon$.

The algorithm works as follows: for each s, k and $i < j$ define the Helstrom projectors

$$A_{ij}(s) := (\sigma_i(s) - \sigma_j(s))_+, \quad (179)$$

where $(\cdot)_+$ is the projector on the positive part of the argument, and their block sum $A_{ij} := \sum_{s \in [n]} |s\rangle\langle s| \otimes A_{ij}(s)$. By construction, these projectors satisfy

$$d_{\mathrm{tr}}(\sigma_i, \sigma_j) = \mathrm{Tr}(\sigma_i A_{ij}) - \mathrm{Tr}(\sigma_j A_{ij}). \quad (180)$$

The algorithm then uses Algorithm 2 on ϱ to perform ERE of the projector-valued functions $\{A_{ij}(s) : i < j\}$,

outputting estimates μ_{ij} such that, with probability at least δ, it holds

$$\left| \frac{1}{n} \sum_{s=1}^n \mathrm{Tr}[\rho_s A_{ij}(s)] - \mu_{ij} \right| \leq 2\epsilon. \quad (181)$$

We also denote

$$v_{kij} := \frac{1}{n} \sum_{s=1}^n \mathrm{Tr}[\sigma_k(s) A_{ij}(s)]. \quad (182)$$

Finally, the algorithm employs a classical subroutine to minimize the quantity

$$\Delta_k := \max_{i<j} |v_{kij} - \mu_{ij}|, \quad (183)$$

finding $k^* := \mathrm{argmin}_k \Delta_k$. We can show that this is a good enough candidate, i.e., $d_{\mathrm{tr}}(\sigma_{k^*}, \rho)$ is sufficiently small. Indeed, if we define the optimal hypothesis attaining Eq. (176) as $i^* := \mathrm{argmin}_i d_{\mathrm{tr}}(\sigma_i, \rho)$, by the triangle inequality it holds

$$d_{\mathrm{tr}}(\sigma_{k^*}, \rho) \leq d_{\mathrm{tr}}(\sigma_{k^*}, \sigma_{i^*}) + d_{\mathrm{tr}}(\sigma_{i^*}, \rho) \quad (184)$$

$$= |\mathrm{Tr}(\sigma_{k^*} A_{k^* i^*}) - \mathrm{Tr}(\sigma_{i^*} A_{k^* i^*})| + \eta \quad (185)$$

$$\leq |\mathrm{Tr}(\sigma_{k^*} A_{k^* i^*}) - \mu_{k^* i^*}|$$
$$+ |\mu_{k^* i^*} - \mathrm{Tr}(\sigma_{i^*} A_{k^* i^*})| + \eta. \quad (186)$$

The first term in Eq. (186) can be bounded in the following way:

$$|\mathrm{Tr}(\sigma_{k^*} A_{k^* i^*}) - \mu_{k^* i^*}| = |v_{k^* k^* i^*} - \mu_{k^* i^*}| = \Delta_{k^*} \quad (187)$$

$$\leq \Delta_{i^*} = \max_{i<j} |\mathrm{Tr}(\sigma_{i^*} A_{ij}) - \mu_{ij}| \quad (188)$$

$$\leq \max_{i<j} |\mathrm{Tr}(\sigma_{i^*} A_{ij}) - \mathrm{Tr}(\rho A_{ij})|$$

$$+ \max_{i<j} |\mathrm{Tr}(\rho A_{ij}) - \mu_{ij}| \quad (189)$$

$$\leq \eta + 2\epsilon. \quad (190)$$

Note that the last two inequalities above are implied by Eq. (181) and therefore the overall result holds with probability at least $1 - \delta$. Proceeding similarly for the second term in Eq. (186) we obtain

$$|\mathrm{Tr}(\sigma_{i^*} A_{k^* i^*}) - \mu_{k^* i^*}| \leq \eta + 2\epsilon \quad (191)$$

as well.

Therefore, by taking $\epsilon' = \epsilon/3$ we can conclude that $d_{\mathrm{tr}}(\sigma_{k^*}, \rho) \leq 3\eta + 4\epsilon$ with probability at least $1 - \delta$. ∎

V. STATISTICAL LEARNING FOR CLASSICAL-QUANTUM PROCESSES

The previous section presented theorems for ERM on finite concept classes. In this section we extend our tools to learning even infinite concept classes. The main result of this section is a statistical learning theorem (Theorem 4) for classical-quantum processes, which gives conditions on their learnability in terms of the covering number of the concept class $\mathcal{C}$ from which they are drawn. Our proof is constructive and provides an explicit algorithm to learn the given concept class. Note however that the algorithm relies on constructing empirical covering nets, which can be demanding; hence we expect that faster algorithms can be found to improve the performance. We also remark that the following results are a consequence of the guarantees on our ERM algorithms for both projector-valued and state-valued concept classes, proved in Sec. IV, together with standard classical statistical learning theory guarantees on uniform convergence of the empirical risks in terms of the growth functions Γ_1 (see Secs. II D and II E). It is important to note that our sufficient condition for learnability is instead expressed in terms of the growth functions $\gamma_{1,q} \geq \Gamma_1$, which are sufficient to guarantee both ERM and uniform convergence.

In Sec. V A we formulate a lemma that identifies, even for continuous-valued concept classes, a finite cardinality concept class on which we apply the algorithm of the previous section. In Sec. V B, we prove our main Theorem 4 for concept classes that map to projector-valued functions. Through shadow tomography, this can be extended to estimating all the empirical risks, and we do this in Sec. V C. In Sec. V D, we prove Theorem 4 for concept classes that map to state-valued functions. In Appendix C we also show that the algorithm of Ref. [7] to learn concept classes that output pure states, in the realizable case, works also when the growth function is slowly growing.

A. Uniform convergence

In the rest of this section we will perform ERE and/or ERM with respect to an ϵ-net of the concept class, which depends on the classical data. The following lemma ensures that ERE and ERM also gives a good solution for the estimation and minimization of the true risks.

Lemma 6. Given l_0 copies of

$$\rho = \sum_{x \in \mathcal{X}} \mathcal{D}(x)|x\rangle\langle x| \otimes \rho(x),$$

by looking at the classical register, we can find a finite subset $\mathcal{C}_\epsilon$ of the concept class $\mathcal{C}$ with loss function (L_s or L_p) such that for every $c \in \mathcal{C}$, there is a (known) $c^*(c) \in \mathcal{C}_\epsilon$

such that

$$|R(c) - R(c^*(c))| \leq \epsilon, \qquad (192)$$

with cardinality at most $\gamma_{1,q}(l_0, \epsilon/2, \mathcal{C})$, and at the same time

$$\forall c \in \mathcal{C} : |R(c) - \hat{R}(c)| < \epsilon/4 \qquad (193)$$

with probability of error

$$p_{\mathrm{err}} \leq 4\gamma_{1,q}(2l_0, \epsilon/64, \mathcal{C})e^{-\frac{l_0\epsilon^2}{512}}. \qquad (194)$$

Proof. This is a simple consequence of Theorem 6. In fact, given a sample S consisting of l_0 samples of the random variable $x \in \mathcal{X}$ associated to the classical register, we have that

$$\Pr_{S \sim \mathcal{D}^{l_0}} [\exists c \in \mathcal{C} : |R(c) - \hat{R}(c)| \geq \epsilon/4] \qquad (195)$$

$$\leq 4\gamma_{1,q}(2l_0, \epsilon/64, \mathcal{C})e^{-\frac{l_0\epsilon^2}{512}}. \qquad (196)$$

By taking an $\epsilon/2$-net $\mathcal{C}_{\epsilon/2}(\vec{x})$ according to the appropriate loss function L on the data $\vec{x}$, we have that for every $c \in \mathcal{C}$ there is $c^*(c) \in \mathcal{C}_{\epsilon/2}(\vec{x})$ such that, via Eqs. (68), (72)

$$|\hat{R}(c) - \hat{R}(c^*(c))| \leq \epsilon/2. \qquad (197)$$

If both conditions are satisfied, we have

$$|R(c) - R(c^*(c))| \leq |R(c) - \hat{R}(c)|$$
$$+ |R(c^*(c)) - \hat{R}(c^*(c))| + |\hat{R}(c) - \hat{R}(c^*(c))| \qquad (198)$$

$$\leq |R(c) - \hat{R}(c)| + |R(c^*) - \hat{R}(c^*)| + \epsilon/2 \qquad (199)$$

$$\leq \sup_{c \in C} 2|R(c) - \hat{R}(c)| + \epsilon/2 \qquad (200)$$

$$\leq \epsilon. \qquad (201)$$

The bound on the cardinality comes from the definition of $\gamma_{1,q}$. $\blacksquare$

Via this lemma, whenever we want to solve the risk estimation for a classical quantum source, we can use part of the classical data to extract a good ϵ-net on the space of *true risks* (meaning exactly that for every concept c there is a known concept c_c^* with true risk ϵ-close), and use the quantum data to obtain the best hypothesis on this finite cardinality concept class, using the algorithms shown in the previous sections. *Remark:* As the cardinality of the ϵ-net on the data can increase with the size of the sample, one could also construct the ϵ-net only on a part of the classical data, such that uniform convergence is ensured at the desired level. If $\lim_{n \to \infty} \log \gamma_{1,q}(2n, \epsilon/64, \mathcal{C})/n = 0$, there will be a finite l_0 such that $p_{\mathrm{err}} \leq \delta$. The cardinality of the ϵ-net will be then $\gamma_{1,q}(l_0, \epsilon/2, \mathcal{C})$. This observation is used in Appendix D.

B. Learnability of quantum processes that map to projectors (Proof of Theorem 4 for loss function L_p)

We now present one of our main technical results about learning unknown quantum processes without input control. This was Theorem 4, which we reproduce below for convenience:

Theorem 13 ((Theorem 4, repeated) Learning quantum processes via ERM). Suppose the concept class $\mathcal{C}$ consists of classical-quantum processes mapping to projectors or states and let $\epsilon > 0$ be the accuracy parameter. Furthermore, let $S = (x_i, \rho(x_i))_{i=1}^n$ be the training set, with $x_i \xleftarrow{\mathcal{D}} X$ and $\rho(\cdot)$ an unknown classical-quantum channel.

Then, the appropriate ERM algorithm of Theorems 1, 3, run on an ϵ-net of the concept class $\mathcal{C}$ (according to the appropriate pseudometric determined by $x_1, \ldots, x_n$), provide an *agnostic* learning algorithm $\mathcal{A} : \mathcal{X}^n \times \mathcal{L}(\mathcal{H}^{(d)})^{\otimes n} \to \mathcal{C}$. This algorithm outputs a hypothesis $\mathcal{A}(S)$ satisfying, for some fixed $\eta, \xi \geq 1$, and n large enough,

$$\Pr_S[R_\rho(\mathcal{A}(S)) - \eta \inf_{h \in \mathcal{C}} R_\rho(h) < \xi\epsilon] > 1 - \delta, \quad (202)$$

if

$$\lim_{n \to \infty} \frac{\log^2 \gamma_{1,q}(n, \epsilon, \mathcal{C})}{n} = 0, \quad \forall \epsilon > 0. \quad (203)$$

In particular, this applies to risks defined via the loss functions L_p (in this case $\eta = 1$, $q = \infty$) and L_s (in this case $\eta = 3$, $q = 1$) for projector-valued and state-valued concept classes $\mathcal{C}$, respectively.

We prove this theorem in two parts. In this section, we prove that Eq. (203) gives a sufficient condition for learning concept classes of projector-valued functions (i.e., using loss function L_p). In Sec. V D, we will prove that the same condition suffices for learning a concept class whose functions map to mixed states (i.e., using loss function L_s). Note that, in the case where the functions in the concept class map to pure states, this latter subcase would follow almost immediately from the former.

We will actually prove a more refined, quantitative statement:

Theorem 14 (Learning projector-valued functions via ERM). Suppose the concept class $\mathcal{C}$ consists of quantum processes mapping to projectors and let $\epsilon > 0$ be the accuracy parameter, and suppose that

$$\lim_{n \to \infty} \frac{\log^2 \gamma_{1,\infty}(n, \epsilon, \mathcal{C})}{n} = 0, \quad \forall \epsilon > 0. \quad (204)$$

Given as input a training set $S = (x_i, \rho(x_i))_{i=1}^n$ with $x_i \xleftarrow{\mathcal{D}} X$ and $\rho(\cdot)$ an unknown classical-quantum channel, there

is an *agnostic* learning algorithm $\mathcal{A} : \mathcal{X}^n \times \mathcal{L}(\mathcal{H}^{(d)})^{\otimes n} \to [0,1]^{\mathcal{C}}$ (i.e., the algorithm of Theorem 10 using an ϵ-net of $\mathcal{C}$ as concept class), such that it outputs c^* together with an estimate $\hat{\mu}_{c^*}$ of $\hat{R}_\rho(h)$ and

$$\Pr_S[|\hat{\mu}_{c^*} - \inf_{c \in \mathcal{C}} R_\rho(c)| \geq 7\epsilon \cup |\hat{\mu}_{c^*} - \hat{R}_\rho(c)| \geq 6\epsilon] =: p_{\text{err}}. \quad (205)$$

With $n = 6Tkl$, for large enough $T = O(\log \frac{1}{\epsilon})$, $k = O(\log(1/\delta \log(1/\epsilon)))$, there exist constants C_1, C_2, C_3 such that, as long as l satisfies

$$(\log \gamma_{1,\infty}(6Tkl, \epsilon/2, \mathcal{C}) + C_2)^2 \leq C_1 l\epsilon^2/9, \quad (206)$$

we have

$$p_{\text{err}} \leq \frac{\delta}{2} + C_3 Tk\gamma_{1,\infty}(6Tkl, \epsilon/2, \mathcal{C})e^{-\frac{l\epsilon^2}{72}}$$
$$+ 4 \log \gamma_{1,\infty}(12Tkl, \epsilon/64, \mathcal{C})e^{-\frac{6Tkl\epsilon^2}{512}}. \quad (207)$$

Once this is established, Part 1 of Theorem 4 follows by redefining ϵ and taking l large enough.

We emphasize that this ThresholdSearch subroutine differs from the original ones [3,4] because it does not require identical copies of the state to output a concept above threshold. Exploiting this algorithm and the convergence of the empirical risk to the true risk as stated in Theorem 6, we can finally prove Theorem 4 for the case of projectors.

Proof. We remind readers that the empirical risk for projector-valued functions can be written as

$$\hat{R}_\rho(c) = \frac{1}{n} \sum_{i=1}^n (1 - \text{Tr}[\rho(x_i)\Pi^{(c)}(x_i)]). \quad (208)$$

The algorithm looks at the classical register $\vec{x}$ and obtains an ϵ-net of the empirical risk $\mathcal{C}|_{\vec{x}}$, which is also a 2ϵ-net on the true risks, with probability of error bounded as in Lemma 6

$$p_{\text{err, unif}}(6Tkl) \leq 4\gamma_1(12Tkl, \epsilon/64, \mathcal{C})e^{-\frac{6Tkl\epsilon^2}{512}}. \quad (209)$$

Then it runs the algorithm of Theorem 10 on the product state obtained from the full dataset $\vec{x}$, with the concepts in $\mathcal{C}|_{\vec{x}}$. In this way we obtain c^* such that $|\hat{\mu}_{c^*} - \max_{c \in \mathcal{C}|_{\vec{x}_0}} \hat{R}(c)| \leq 6\epsilon$ and $\hat{\mu}_{c^*} - \hat{R}(c^*) \leq 6\epsilon$, implying $\hat{R}(c^*) < \max_{c \in \mathcal{C}|_{\vec{x}_0}} \hat{R}(c) + 6\epsilon$ with probability

bounded as in the proof of Theorem 10

$$p_{\text{err,ERM}} \le T(0.97^k + (k+1)e^{-l\epsilon^2/72}) + 2Tkme^{-2l\epsilon^2/64}. \tag{210}$$

If this is true we have that, for the selected concept c^*

$$R(c^*) \le \hat{R}(c^*) + \epsilon/4 \tag{211}$$

$$\le \inf_{c \in \mathcal{C}|_{\bar{x}}} \hat{R}(c) + 6\epsilon + \epsilon/4 \tag{212}$$

$$\le \inf_{c \in \mathcal{C}} \hat{R}(c) + 6\epsilon + 3\epsilon/4 \qquad (\epsilon/2 \text{ net})$$

$$\le \inf_{c \in \mathcal{C}} R(c) + 7\epsilon. \tag{213}$$

The probability of error is then upper bounded as

$$p_{\text{err,unif}} + p_{\text{err,ERM}}. \tag{214}$$

With the choices made for the parameters, we obtain the thesis. ∎

C. Shadow tomography for classical-quantum states

We can also estimate the empirical risks of all the concepts in a class, a task that is strictly related to shadow tomography. By uniform convergence, these will be close to the true risks. In fact, we can show an *improved* algorithm for shadow tomography of classical-quantum states. Using this algorithm, we cannot only find the minimum empirical risk in the concept class, but also simultaneously estimate empirical risk for all concepts.

Theorem 15 ((Theorem 5, refined) Improved shadow tomography of classical-quantum states). Consider a collection of projector-valued functions $\mathcal{C} = \{f_c : x \to \Pi^{(c)}(x)\}$ with domain $\mathcal{X}$ and image in the projectors of a Hilbert space of dimension d. Given access to n copies of a classical-quantum state $\rho = \sum_{x \in \mathcal{X}} \mathcal{D}(x)|x\rangle\langle x| \otimes \rho(x)$, the algorithm of Theorem 2 used with an ϵ-net of $\mathcal{C}$ outputs values $\{\mu_c\}_{c \in \mathcal{C}}$ such that, except with probability δ, for all $f_c \in \mathcal{C}$ it holds

$$\left| \sum_x \mathcal{D}(x)\text{Tr}[\rho(x)\Pi^{(c)}(x)] - \mu_c \right| \le 3\epsilon \tag{215}$$

if either one of these conditions holds:

(a) $\mathcal{C}$ is finite. Then the minimal number of copies satisfies $n = \tilde{O}(\frac{\log^2 m \log d \log(1/\delta)}{\epsilon^5})$.

(b) $\mathcal{C}$ is infinite but it holds

$$\lim_{n \to \infty} \frac{\log^2 \gamma_{1,\infty}(n, \epsilon, \mathcal{C})}{n} = 0, \qquad \forall \epsilon > 0 \tag{216}$$

and the number of copies is large enough.

In particular, with $n = 6Tk$, for large enough $T = O\left(\log d/\epsilon^3 \left(\log \log d + \log(1/\epsilon)\right)\right)$ and $k = O(\log(T/\delta))$, there exist constants C_1, C_2, C_3 such that, as long as l satisfies

$$(\log \gamma_{1,\infty}(6Tkl, \epsilon/2, \mathcal{C}) + C_2)^2 \le C_1 l\epsilon^2/9, \tag{217}$$

have that the probability of error is bounded as

$$p_{\text{err}} \le \frac{\delta}{2} + C_3 Tk\gamma_{1,\infty}(6Tkl, \epsilon/2, \mathcal{C})e^{-\frac{l\epsilon^2}{32}}$$
$$+ 4\log \gamma_{1,\infty}(12Tkl, \epsilon/64, \mathcal{C})e^{-\frac{6Tkl\epsilon^2}{512}}. \tag{218}$$

Proof. The only difference with Theorem 20 is to replace the ERE subroutine with the shadow tomography subroutine to estimate the empirical risks. The probability of error of the ERE, $p_{\text{err,ERE}}$, becomes bounded as in the proof of Theorem 2, with $m = \gamma_{1,\infty}(6Tkl, \epsilon/2, \mathcal{C})$:

$$p_{\text{err,ERE}} \le T((0.97)^k + 2(k+1)e^{-l\epsilon^2/6}) + 2Tk\gamma_{1,\infty}(6Tkl, \epsilon/2, \mathcal{C})e^{-2l\epsilon^2/64}.$$

On the other hand, the probability of error of the uniform convergence is the same as in the previous theorem. By summing the two probabilities of error we get the bound in the thesis. Notice that with Theorem 2 we get estimates of the empirical risks at precision 2ϵ. By Theorem 6 this translates into a precision 3ϵ on the full class. ∎

D. Learnability of quantum processes that map to states (Theorem 4 with loss function L_s)

In this section, we will prove that quantum processes that output states are also efficiently learnable (having shown the analogous statement for processes that output projectors in Sec. V B). This also constitutes the second half of the statement of Theorem 4. In this setting the empirical risk takes the form

$$\hat{R}_\rho(h) = \frac{1}{n} \sum_{i=1}^n d_{\text{tr}}(\sigma_h(x_i), \rho(x_i)). \tag{219}$$

We can now prove Theorem 4 for state-valued functions, with loss function L_s. In fact, we prove a refined statement:

Theorem 16 (Refinement of Theorem 4 for state-valued functions). Suppose the concept class $\mathcal{C}$ consists of quantum processes mapping to states and let $\epsilon > 0$ be the accuracy parameter. Suppose that

$$\lim_{n \to \infty} \frac{\log^2 \gamma_{1,1}(n, \epsilon, \mathcal{C})}{n} = 0, \qquad \forall \epsilon > 0. \tag{220}$$

Given as input a training set $S = (x_i, \rho(x_i))_{i=1}^n$ with $x_i \xleftarrow{\mathcal{D}} X$ and an unknown $\rho(x)$, there is a learning algorithm

$\mathcal{A} : \mathcal{X}^n \times \mathcal{L}(\mathcal{H}^{(d)})^{\otimes n} \to \mathcal{C}$ obtained from the algorithm of Theorem 12 applied to an ϵ-net of $\mathcal{C}$ (according to the appropriate pesudometric defined by $x_1, \ldots, x_n$), such that

$$\Pr_{S}[R_\rho(\mathcal{A}(S)) - 3 \inf_{c \in \mathcal{C}} R_\rho(c) < 6\epsilon] =: 1 - p_{\text{err}}, \quad (221)$$

and setting $n = 6Tkl$, for large enough $T = O\left(\log d / \epsilon^3 (\log \log d + \log(1/\epsilon))\right)$, $k = O(\log(T/\delta))$, there exist $C_1, C_2, C_3, C_4 > 0$ constants such that as long as

$$(\log \gamma_{1,1}(6Tkl, \epsilon/2, \mathcal{C}) + C_2)^2 \leq C_1 l \epsilon^2, \quad (222)$$

we have

$$p_{\text{err}} \leq \frac{\delta}{2} + C_3 \gamma_{1,1}(12Tkl, \epsilon/16, \mathcal{C}) e^{-6Tkl\epsilon^2/128}$$

$$+ C_4 Tk (\gamma_{1,1}(6Tkl, \epsilon/2, \mathcal{C}))^2 e^{-\frac{l\epsilon^2}{32}}. \quad (223)$$

Proof of Theorem 16. The proof is a straightforward generalization of the proof of Theorem 12 using the algorithm of Theorem 15. In this case, $\mathcal{C}$ is possibly of infinite cardinality. The algorithm works as follows. Following the notation of Theorem 15, first we measure the classical register of the copies of ρ, obtaining a vector $\vec{x}$ and an ϵ-net $\mathcal{C}_{\vec{x}}$, with cardinality less than the covering number $\gamma_{1,1}(6Tkl, \epsilon/2, \mathcal{C})$. By Theorem 6, this $\epsilon/2$-net gives an ϵ-net for the true risks, with probability of error bounded as

$$p_{\text{err,unif}} \leq 4\gamma_{1,1}(12Tkl, \epsilon/16, \mathcal{C}) e^{-6Tkl\epsilon^2/128}. \quad (224)$$

Therefore, we have that if we can find c^* such that $\hat{R}(c^*) \leq 3 \inf_{c \in \mathcal{C}_{\vec{x}}} \hat{R}(c) + 4\epsilon$, then

$$R(c^*) \leq \hat{R}(c) + \epsilon/4 \quad (225)$$

$$\leq 3 \inf_{c \in \mathcal{C}_{\vec{x}}} \hat{R}(c) + 4\epsilon + \epsilon/4 \quad (226)$$

$$\leq 3 \inf_{c \in \mathcal{C}} \hat{R}(c) + 4\epsilon + 7\epsilon/4 \quad (\epsilon/2 \text{ net})$$

$$\leq 3 \inf_{c \in \mathcal{C}} R(c) + 6\epsilon. \quad (227)$$

Therefore, we also have that $R(c^*) \leq 3 \inf_{c \in \mathcal{C}} R(h) + 6\epsilon$ with high probability. We can then apply Theorem 12 to the quantum data with the concept class $\mathcal{C}_{\vec{x}}$. The probability of error of the algorithm of Theorem 12 is in fact the probability of error of the ERE algorithm for the projector-valued concept class constructed from $\mathcal{C}_{\vec{x}}$ as in Theorem 12, therefore, we have again

$$p_{\text{err,ERM}} \leq T((0.97)^k + 2(k+1)e^{-l\epsilon^2/6})$$

$$+ 2Tk\gamma_{1,1}^2(6Tkl, \epsilon/2, \mathcal{C}) e^{-2l\epsilon^2/64}. \quad (228)$$

Putting all together we obtain the full error bound. ∎

VI. CLASSES WITH BOUNDED COVERING NUMBERS

In this section we present some models of physically motivated concept classes for which upper bounds on their covering numbers can be found. We find these bounds on covering numbers using continuity and bounds on covering numbers of real functions and matrices. This opens up the question of whether it is possible to define a combinatorial dimension (such as the VC dimension or the fat-shattering dimension) corresponding to the covering numbers we introduced.

Whenever needed in the following, we employ a real-valued function class $\mathcal{F}$, and we assume that for this function class we have a bound on the covering number in terms of some combinatorial dimension, for example, the fat-shattering dimension (see Ref. [13] for definitions and proofs of these bounds). It holds that, for a class of functions $\mathcal{F} : \mathbb{R} \to [0, B]$ with fat-shattering dimension $D = \text{fat}_{\mathcal{F}}(\epsilon/4)$, the covering number is bounded as

$$\Gamma_p(n, \epsilon, \mathcal{F}) < 2 \left(\frac{4nB^2}{\epsilon^2}\right)^{D \log_2(4eBn/(D\epsilon))}. \quad (229)$$

Furthermore, we can have finite covering numbers for sets of $k \times k$ matrices, seen as functions with trivial input to $\mathbb{C}^k \times \mathbb{C}^k$. These techniques were used for the purpose of studying generalization bounds in Ref. [31], and uniform convergence for learning quantum channels on multiqubit systems of finite size, in the controlled input setting, in Ref. [8]. The following lemma is crucial.

Lemma 7 (Size of an ϵ net over unitary and bounded hermitian matrices).

$$N_{in}(\epsilon, B_R(x), ||\cdot||) \leq \left(1 + \frac{2R}{\epsilon}\right)^K \leq \left(\frac{3R}{\epsilon}\right)^K, \quad (230)$$

where $B_R(x)$ is a norm ball of radius $R \geq \varepsilon$ about some point $x \in \mathbb{R}^K$. We can use this to calculate the size of ϵ-nets over matrices with respect to the operator norm, by using the fact that for a set of unitaries S_U in dimension k, $S_U \subset B_1(0)$ with $K = (2k)^2$, and for a set of hermitian matrices S_H in dimension k with norm bounded by b, which we denote by $\mathcal{M}_b^{(k)}$, $\mathcal{M}_b^{(k)} \subset B_b(0)$ with $K = k^2$.

$$|N_{in}(\epsilon, S_U, ||\cdot||)| \leq |N_{in}(\epsilon/2, B_1(0), ||\cdot||)|$$

$$\leq \left(\frac{6}{\epsilon}\right)^{4k^2}, \quad (231)$$

$$|N_{in}\left(\epsilon, \mathcal{M}_b^{(k)}, ||\cdot||\right)| \leq |N_{in}(\epsilon/2, B_b(0), ||\cdot||)|$$

$$\leq \left(\frac{6b}{\epsilon}\right)^{k^2}. \quad (232)$$

The bound on the covering number of the Euclidean ball is given in Ref. [15], and the argument generalizes to any norm in a finite-dimensional space. The first inequalities are a consequence of triangular inequality. For specific matrix classes one can have improved bounds.

In the following, we first exhibit concept classes based on circuits, which are valuable since they are associated to quantum states and measurement that can actually be produced efficiently on a quantum computer: not only the data that we receive could be states of this form, but also the resulting output hypothesis would be efficiently produced for any future need. In a further subsection we also present concept classes that are more motivated by physical scenarios, ideally exhibiting toy models for the setting described in the introduction, inspired by sensing and hamiltonian learning.

A. Quantum circuits

We consider concept classes based on circuits. By choosing circuits that do not depend on the data, we will first give several examples of concept classes with covering numbers that are not just slowly growing, but in fact bounded by a quantity independent of n, the length of the sample. Then, including a finite number of gates that depend on the data through real functions with finite fat-shattering dimension, we will obtain concept classes that are slowly growing with n.

1. Quantum circuits that give rise to state-valued functions

First, we will look at an example of a concept class that maps to quantum states. We remind readers that this corresponds to choosing as the loss function the trace distance L_s. Consider the concept class $\mathcal{C}_m \subseteq \{c : \mathcal{X} \to \mathcal{L}(\mathcal{H}_m)\}$ consisting of m-qubit quantum circuits acting on arbitrary input states. That is,

$$\mathcal{C}_m = \{c_U(x) := U\rho(x)U^\dagger\}_{U\in S_m}, \tag{233}$$

where $\rho : \mathcal{X} \to D(\mathcal{H}_m)$ is a fixed process preparing a mixed state that depends on a classical random variable $\mathcal{X}$, and U is an arbitrary m-qubit unitary chosen from a set S_m. For instance, we could be interested in studying processes corresponding to quantum circuits given by a particular architecture (specified by S_m) acting on pure computational-basis states. This class is obtained in our formalism by setting $\mathcal{X} = \{0, 1\}^m$ and noting that the process that prepares the pure computational basis state $|x\rangle\langle x|$ corresponds to $\rho(x) = |x\rangle\langle x|$. Specific relevant examples of S_m will be discussed later.

Then an upper bound on $\gamma_{1,1}(n, \epsilon, \mathcal{C}_m)$ is given by the size of an $\epsilon/8$-net over the m-qubit unitaries in S_m, where the distance metric is taken to be the spectral norm $\|\cdot\|$ [see Eq. (44)].

In the following, we remind readers that n is the number of data points observed, while m is a parameter of the concept class—the number of qubits in the circuits in the concept class.

Proposition 2. $\gamma_{1,1}(n, \epsilon, \mathcal{C}_m) \leq |N_{in}(\epsilon, S_m, \|\cdot\|)|$

Proof. To see this, observe that for an arbitrary state ρ and unitaries U, V

$$\frac{1}{2}\|U\rho U^\dagger - V\rho V^\dagger\|_1 \leq \|U - V\|. \tag{234}$$

Indeed, with the spectral decomposition $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, we have

$$\frac{1}{2}\|U\rho U^\dagger - V\rho V^\dagger\|_1$$

$$= \frac{1}{4}\|(U + V)\rho(U^\dagger - V^\dagger) + (U - V)\rho(U^\dagger + V^\dagger)\|_1 \tag{235}$$

$$\leq \frac{1}{4}\sum_i p_i \left(\|(U + V)|\psi_i\rangle\langle\psi_i|(U^\dagger - V^\dagger)\|_1 \right.$$

$$\left. + \|(U - V)|\psi_i\rangle\langle\psi_i|(U^\dagger + V^\dagger)\|_1\right) \tag{236}$$

$$= \frac{1}{4}\sum_i p_i \left(\|(U + V)|\psi_i\rangle\|_2\|(U - V)|\psi_i\rangle\|_2 \right.$$

$$\left. + \|(U - V)|\psi_i\rangle\|_2\|(U + V)|\psi_i\rangle\|_2\right) \tag{237}$$

$$\leq \frac{1}{2}\max_{|\psi_i\rangle}\|(U + V)|\psi_i\rangle\|_2 \max_{|\psi_j\rangle}\|(U - V)|\psi_j\rangle\|_2 \tag{238}$$

$$\leq \frac{1}{2}\|U + V\|\|U - V\| \tag{239}$$

$$\leq \|U - V\|, \tag{240}$$

where in the third line we have used Fact 1.

Therefore, for any two functions $c_U, c_V \in \mathcal{C}_m$, we have that

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{2}\left\|c_U(x_i) - c_V(x_i)\right\|_1 \tag{241}$$

$$= \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\left\|U\rho(x_i)U^\dagger - V\rho(x_i)V^\dagger\right\|_1 \tag{242}$$

$$\leq \|U - V\|, \tag{243}$$

where the second inequality is inequality (240).

Now consider the set $\mathcal{C}_m^{(\epsilon)} \subseteq \mathcal{C}_m$ induced by the ϵ-net of S_m (denoted as $S_m^{(\epsilon)}$), and

$$\mathcal{C}_m^{(\epsilon)} := \left\{c_V : c_V(\vec{x}) = V\rho(x_i)V^\dagger\right\}_{V\in S_m^{(\epsilon)}}. \tag{244}$$

Equation (243) implies that, if for all $U \in S_m$, $\exists V \in S_m^{(\epsilon)}$ such that $\|U - V\| < \epsilon$, then for all $c_U \in \mathcal{C}_m$, $\exists c_V \in \mathcal{C}_m^{(\epsilon)}$

such that $\|c_U - c_V\|_{1,\vec{x}} < \epsilon$. This is a sufficient condition for $\mathcal{C}_m^{(\epsilon)}$ to be an ϵ-net of $\mathcal{C}_m$.   ■

Proposition 2 makes an important reduction: to upper bound $\gamma_{1,1}(n, \epsilon, \mathcal{C}_m)$, where $\mathcal{C}_m$ is a concept class induced by a set of m-qubit quantum circuits S_m, it suffices to upper bound the size of an ϵ-net in the spectral norm over the set S_m.

We now compute this covering number for circuit classes S_m corresponding to particular architectures of relevance.

(a) Consider the class of one-dimensional local random quantum circuits on m qubits of depth ℓ, which was defined by Brandao, Harrow, and Horodecki [16] as applying a Haar-random two-qubit nearest-neighbor gate on a uniformly random pair of neighboring qubits at each time step $t = 1, \dots \ell$. (Note that the distribution from which the circuits are drawn is actually immaterial to us; the concept class merely consists of the support of that distribution.) We shall call this class $LQC(m, \ell)$, and we will call the induced loss function class $\mathcal{G}_{LQC(m,\ell)}$.

Corollary 1 (Bound on covering net size for 1D random local quantum circuits).

$$N_{in}(\epsilon, LQC(m, \ell), \|\cdot\|) \leq \left[m \times \left(\frac{6\ell}{\epsilon} \right)^{32} \right]^\ell, \quad (245)$$

and hence by Proposition 2,

$$\Gamma_1(n, \epsilon, \mathcal{G}_{LQC(m,\ell)}) \leq \gamma_{1,1}(n, \epsilon, \mathcal{C}_{LQC(m,\ell)})$$
$$\leq \left[m \times \left(\frac{6\ell}{\epsilon} \right)^{32} \right]^\ell. \quad (246)$$

Proof. The argument follows along the lines of that made in [8,31], and so we only sketch it briefly. Let $\mathcal{N}_{\tilde{\epsilon}}$ be an $\tilde{\epsilon} = \epsilon/\ell$-net over a single two-qubit unitary. Let us now consider the net

$$\tilde{\mathcal{N}}_\epsilon = \left\{ U_{1,p(1)} U_{2,p(2)} \dots U_{d,p(\ell)} \middle| U_i \in \mathcal{N}_{\tilde{\epsilon}}, p(i) \in [\ell] \right\}, \quad (247)$$

where $U_{i,p(i)}$ denotes the unitary U_i acting on qubits $p(i), p(i) + 1$. As a special case of Lemma 7, an ϵ-net over the set of two-qubit unitaries [a set that lies within $B_{\|\mathbb{1}_{(\mathbb{C}^2)^{\otimes 2}}\|}(0)$, a set of matrices with $4 \times 4 \times 2$ parameters], has size $(6/\epsilon)^{32}$. Combining this with the fact that there are $m - 1$ choices of qubit pairs on which the unitary can act, $\tilde{\mathcal{N}}_\epsilon$ has size $\left[(6/\tilde{\epsilon})^{32} \cdot (m - 1) \right]^\ell$. An arbitrary circuit in

$LQC(m, \ell)$ has the form

$$V = V_1, \dots V_\ell, \quad (248)$$

where each V_i is a two-qubit nearest-neighbor gate. We now show that $\tilde{\mathcal{N}}_\epsilon$ is an ϵ-net over the set $LQC(m, \ell)$. For V_i acting on qubits $p(i), p(i) + 1$, there exists some $U_i \in \mathcal{N}_{\tilde{\epsilon}}$ such that $\|U_i - V_i\| < \tilde{\epsilon}$, and we can enforce that it acts on qubits $p(i), p(i) + 1$. This means that $U_1 \dots U_\ell \in \tilde{\mathcal{N}}_\epsilon$, and moreover

$$\|V_1 \dots V_\ell - U_1 \dots U_\ell\|$$
$$\leq \|V_1 \dots V_{\ell-1} V_\ell - V_1 \dots V_{\ell-1} U_\ell\|$$
$$\quad + \|V_1 \dots V_{\ell-1} U_\ell - U_1 \dots U_{\ell-1} U_\ell\| \quad (249)$$
$$\leq \|V_\ell - U_\ell\| + \|V_1 \dots V_{\ell-1} - U_1 \dots U_{\ell-1}\| \quad (250)$$
$$\leq \epsilon/\ell + \|V_1 \dots V_{\ell-1} - U_1 \dots U_{\ell-1}\| \quad (251)$$
$$\leq \epsilon, \quad (252)$$

where the first inequality is the triangle inequality, the second inequality follows from unitary invariance of the spectral norm, the third inequality follows from the assumption that $\mathcal{N}_{\tilde{\epsilon}}$ is an ϵ/ℓ-net and the last inequality follows from repeating the above three steps on the remaining term in Eq. (251).   ■

(a) Exactly analogously, we could also consider the set $S_b(m, \ell)$ of all "brickwork" local quantum circuits, where every pair of neighboring qubits has a nearest-neighbor two-qubit gate acting on them at every time step. We then obtain

$$\Gamma_1(n, \epsilon, \mathcal{G}_{S_b(m,\ell)}) \leq \gamma_{1,1}(n, \epsilon, \mathcal{C}_{S_b(m,\ell)}) \leq \left(\frac{6m\ell}{\epsilon} \right)^{32m\ell}. \quad (253)$$

(b) Alternatively, we could consider choosing our unitaries from the set of all possible unitaries on m qubits. Let us call this set S_m. Immediately from Lemma 7 we have

$$\gamma_{1,1}(n, \epsilon, \mathcal{C}_{S_m}) \leq \left(\frac{6}{\epsilon} \right)^{2^{2m+2}}. \quad (254)$$

Observe that in all of these cases, the covering number bound (which determines learnability of the corresponding class) is independent of the number of sampled points, n.

2. Quantum circuits that give rise to projector-valued functions

In the previous subsection, we computed the covering number of state-valued concept classes. In this subsection, we will give an example of a concept class consisting

of *projector*-valued functions, and compute its covering number.

Side note: An immediate example of a projector-valued concept class is, in fact, the state-valued concept classes obtained by circuits applied to computational basis states, as a special case of the settings considered in the last subsection. This is because the concepts output pure states which are rank-1 density matrices, i.e., rank-1 projectors. Even for higher ranks, the calculation of the covering number for this concept class proceeds similarly. Some slight tweaks are needed to account for the use of a different loss function (namely L_p) and notion of distance between concepts (namely the spectral norm instead of the trace norm).

We present a class of projector-valued concepts that have an explicit x dependence but each of them is based on a single circuit. We fix a projector-valued function $\Pi(x)$ and we define $\mathcal{C}_\Pi \subseteq \{c : \mathcal{X} \to \mathcal{L}(\mathcal{H}_m)\}$, where S_m is a set of circuits:

$$\mathcal{C}_\Pi = \{\Pi_U := U\Pi(x)U^\dagger\}_{U \in S_m}. \qquad (255)$$

To calculate the covering numbers, it suffices to find the size of an $\epsilon/8$-net over the isometries $S_m = \{V_h\}_h$ according to the diamond distance. To see that this suffices, we will need two standard consequences of the definition of operator norms, as follows:

Proposition 3. For any Π_1 and Π_2 which are projectors onto $\mathcal{L}(\mathcal{H}_m)$,

$$\max_{\rho \in D(\mathcal{H}_m)} |\mathrm{Tr}[\rho(\Pi_1 - \Pi_2)]| = \|\Pi_1 - \Pi_2\|. \qquad (256)$$

Proof. $\Pi_1 - \Pi_2$ is Hermitian, hence it is diagonalizable and it has real eigenvalues, so $\|\Pi_1 - \Pi_2\|$ equals the largest absolute values among those of the eigenvalues of $\Pi_1 - \Pi_2$, say $|\lambda|$, with corresponding eigenvector $|\lambda\rangle$. Then, $\|\Pi_1 - \Pi_2\| = |\mathrm{Tr}[|\lambda\rangle\langle\lambda|(\Pi_1 - \Pi_2)]|$. Conversely, by the duality between trace norm and operator norm, $\max_{\rho \in D(\mathcal{H}_m)} |\mathrm{Tr}[\rho(\Pi_1 - \Pi_2)]| \leq \|\Pi_1 - \Pi_2\|$. ∎

Proposition 4. For any two unitaries $U, V : \mathcal{L}(\mathcal{H}_m) \to \mathcal{L}(\mathcal{H}_m)$,

$$\|UXU^\dagger - VXV^\dagger\| \qquad (257)$$
$$\leq \|UX(U^\dagger - V^\dagger)\| + \|(U - V)XV^\dagger\| \qquad (258)$$
$$\leq 2\|X\|\|U - V\|. \qquad (259)$$

Let us fix any internal ϵ-net $S_m^{(\epsilon)}$ of cardinality $N_{in}(\epsilon, S_m, \|\cdot\|)$ over the set of unitaries S_m. We can then

prove that the set induced by $S_m^{(\epsilon)}$

$$\mathcal{C}_\Pi^{(\epsilon)} := \Big\{\Pi_V : \mathcal{X}^n \to L(\bigotimes_{i=1}^n \mathcal{H}_m),$$
$$\Pi_V(\vec{x}) = V\Pi(x)V^\dagger\Big\}_{V \in S_m^{(\epsilon)}} \qquad (260)$$

lifts to a 2ϵ-net over $\mathcal{C}_\Pi$, as follows: we will show that for any $\Pi_V \in \mathcal{C}_\Pi$, we can find $\Pi_U \in \mathcal{C}_\Pi^{(\epsilon)}$ at most distance 2ϵ away. Let $V \in S_m$ be the unitary associated with Π_V. Then, choose U to be the nearest point in $S_m^{(\epsilon)}$ to V. Then we claim that Π_U suffices to be such a point:

$$\frac{1}{n} \sum_{i=1}^n \|\Pi_V(x_i) - \Pi_U(x_i)\|$$
$$\leq \frac{1}{n} \sum_{i=1}^n 2\|\Pi(x_i)\| \cdot \|V - U\| \leq 2\epsilon \qquad (261)$$

where the second to last inequality follows from Proposition 4 and the last one from the definition of ϵ-net.

Finally, exactly analogously to Proposition 2 and Proposition 1, we obtain from Eq. (261):

Proposition 5 (Bounding Γ_1 and $\gamma_{1,\infty}$ for projector-valued function classes). For the class of projector-valued functions $\mathcal{C}_\Pi$

$$\gamma_{1,\infty}(n, \epsilon, \mathcal{C}_\Pi) \leq N_{in}(\epsilon/2, S, \|\cdot\|). \qquad (262)$$

B. Data-dependent circuits

The previous subsections have studied concept classes based on circuits that are independent of the data. As a result, the upper bounds on the covering numbers do not depend on the size of the dataset. Without deep changes in the structure of the proofs of the bounds, we can also treat cases where this dependence is in fact explicit. Such data-dependent circuits are considered in variational quantum machine learning models with data reuploading (see, e.g., Ref. [39]). Consider circuits described by a circuit model S of k two-qubit-gates, as in the previous sections, and modify it by inserting in specific places in the sequence a number k' of gates of the form $e^{iH_{j'}g_{j'}(x)}$, $j' = 1, \ldots, k'$, $\|H_i\| \leq b$, $H_{j'}$ fixed and $g_{j'}(x)$ belonging to a concept class $\mathcal{F}$ with fat-shattering dimension D, where the functions in the class map to $[0, B]$ as explained in the beginning of this section. Each circuit in the concept class will be made of a sequence of k data-independent qubit gates and k' data-dependent gates, in specified positions. Since for

any unitarily invariant matrix norm we have [40]

$$||e^X - e^Y|| \leq ||X - Y||e^{||X-Y||}e^{||X||}, \quad (263)$$

we have that, choosing $B = 1/b$, whenever $|g_i(x) - g_i'(x)| \leq \epsilon/(4bet), \epsilon < 1$

$$||e^{iH_ig_i(x)} - e^{iH_ig_i'(x)}||$$

$$\leq ||H_ig_i(x) - H_ig_i'(x)||e^{||H_ig_i(x)-H_ig_i'(x)||}e^{||H_ig_i(x)||}$$

$$\leq b||g_i(x) - g_i'(x)||e^{b||g_i(x)-g_i'(x)||}e^1$$

$$\leq b\frac{\epsilon}{4bet}e^{b\frac{\epsilon}{4bet}}e^1 \leq \frac{\epsilon}{2t}. \quad (264)$$

Now, take two circuits U and V such that the non-data-dependent parts are such that $\sum_{j=1}^{k} ||U_j - V_j|| \leq \epsilon/2$, and also $||g_{j'}(x) - g_{j'}'(x)||_{1,\vec{x}} \leq \epsilon/(4bek')$, for every j, j'. By triangular inequalities, for two-state valued concepts c_U, c_V associated to circuits U, V, repeating the reasoning behind Eq. (249) and following equations, we get

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left\|c_U(x_i) - c_V(x_i)\right\|_1 \leq \frac{1}{2n}\sum_{i=1}^{n}||U(x_i) - V(x_i)||$$

$$\leq \sum_{j=1}^{k}||U_j - V_j||$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\sum_{j'=1}^{k'}||e^{ig_{j'}(x_i)H_{j'}} - e^{ig_{j'}'(x_i)H_{j'}}||$$

$$\leq \epsilon/2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j'=1}^{k'}\frac{\epsilon}{2k'}$$

$$\leq \epsilon/2 + \epsilon/2 = \epsilon. \quad (265)$$

The same inequality holds if we take projector-valued classes based on the circuits. This means that the covering number for a data-dependent class of circuits $\mathcal{C}'$ constructed from a qubit model S is bounded as

$$\gamma_{1,\infty}(n, \epsilon, \mathcal{C}') \leq N_{in}(\epsilon/2, S, || \cdot ||_S)$$

$$\times 2^{k'}\left(\frac{64ne^2(k')^2}{\epsilon^2}\right)^{k'D\log_2(16e^2k'n/(D\epsilon))}. \quad (266)$$

where $||U - V||_S := \sum_{i=1}^{k}||U_i - V_i||$ with $\{U_i\}$ and $\{V_i\}$ being the gates of U and V, respectively, and $N_{in}(\epsilon/2, S, || \cdot ||_S)$ can be bounded as in the previous section.

C. Learning physical processes

We now present some concept classes inspired by physical scenarios. We will always assume that the quantum register has dimension d.

We use facts about matrix continuity to obtain a bound on the covering number for our state-valued and projector-valued concept classes. In particular, we need the following.

(a) For any unitarily invariant matrix norm we have [40]

$$||e^X - e^Y|| \leq ||X - Y||e^{||X-Y||}e^{||X||}. \quad (267)$$

(b) [Lemma 16 in Ref. [41]]. For any Hermitian matrices H, H',

$$||\frac{e^{-H}}{\text{Tr}[e^{-H}]} - \frac{e^{-H'}}{\text{Tr}[e^{-H'}]}||_1 \leq 2(e^{||H-H'||} - 1). \quad (268)$$

(c) Weyl's perturbation theorem [42]. Let A and B be Hermitian matrices with ordered eigenvalues, respectively, $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$ and $\lambda_1(B) \geq \lambda_2(B) \geq \cdots \geq \lambda_n(B)$. Then

$$\max_{i=1,\ldots,n}|\lambda_i(A) - \lambda_i(B)| \leq ||A - B||. \quad (269)$$

(d) Given two projectors E, F, $||E - F|| = ||E^\perp F||$ ([42], Exercise VII.1.11).

(e) ([42], Theorem VII.3.2). Let A and B be Hermitian operators, and let S_1, S_2 any two intervals such that their distance $\sup_{x \in S_1, y \in S_2}|x - y| = \delta > 0$. Let E be the projector on eigenspaces of A with eigenvalues contained in S_1, and F be the projector on eigenspaces of B with eigenvalues contained in S_2. Then, for any unitarily invariant norm $|||\cdot|||$,

$$|||EF|| \leq \frac{\pi}{2\delta}|||A - B|||. \quad (270)$$

(f) (Corollary of previous points) Given two Hermitian matrices A and B such that $||A - B|| \leq \epsilon$ and such that A does not have eigenvalues in the interval $(\theta - 2\epsilon, \theta + 2\epsilon)$, [and thus B does not have eigenvalues in the interval $(\theta - \epsilon, \theta + \epsilon)$ by Weyl perturbation's theorem] and E, F the projectors on the eigenspaces associated to the eigenvalues less than $\theta - \epsilon$ satisfy

$$||E - F|| \leq \frac{\pi}{4\epsilon}||A - B||. \quad (271)$$

This follows from the previous point since E coincides with the projector on the eigenspaces of A with eigenvalues contained in $(-\infty, \theta - \epsilon]$ and $F^\perp$ coincides with the projector on the eigenspaces of B with eigenvalues contained in $[\theta + \epsilon, +\infty)$, and the two sets $(-\infty, \theta - \epsilon]$ and $[\theta + \epsilon, +\infty)$ have distance 2ϵ.

We are now equipped to prove the following propositions. Let $\mathcal{F}$ be a function class with fat-shattering dimension D, containing functions $g \in F$ such that $|g(x)| \le B$. We remind that $\mathcal{M}_b^{(d)}$ indicates Hermitian matrices of operator norm bounded by b.

Proposition 6 (Gibbs states of a perturbed Hamiltonian). Consider a concept class given by a set of Gibbs states obtained by perturbing a Hamiltonian with a field-dependent term, but the specific dependence on the field is not known, namely

$$\mathcal{C} = \{f \,|\, f(x) = \frac{e^{-H_0 - g(x)V}}{\text{Tr}[e^{-H_0 - g(x)V}]}, \; g \in \mathcal{F}, V \in \mathcal{M}_b^{(d)}\}. \tag{272}$$

Its covering number is bounded as

$$\gamma_{1,1}(n, 2e^{\epsilon(B+b)}\epsilon(b+B), \mathcal{C})$$
$$\le 2\left(\frac{4nB^2}{\epsilon^2}\right)^{D\log_2(4eBn/(D\epsilon))} \times \left(\frac{6b}{\epsilon}\right)^{d^2}. \tag{273}$$

Proof. Let $\mathcal{F}_\epsilon$ be an ϵ-net of $\mathcal{F}$ and $\mathcal{M}_{b,\epsilon}^{(d)}$ an ϵ-net of $\mathcal{M}_b^{(d)}$. Then, for every $g \in \mathcal{F}$ there is $g \in \mathcal{F}_\epsilon$ such that $||g - g'||_{1,\vec{x}} \le \epsilon$ and for every $V \in \mathcal{M}_b^{(d)}$ there is $V' \in \mathcal{M}_{b,\epsilon}^{(d)}$ such that $||V - V'|| \le \epsilon$. Let us consider the corresponding functions $f(x) = \frac{e^{-H_0 - g(x)V}}{\text{Tr}[e^{-H_0 - g(x)V}]} \in \mathcal{C}$, $f'(x) = \frac{e^{-H_0 - g'(x)V'}}{\text{Tr}[e^{-H_0 - g'(x)V'}]} \in \mathcal{C}$. Then, using Eq. (268) we obtain

$$||\frac{e^{-H_0 - g(x)V}}{\text{Tr}[e^{-H_0 - g(x)V}]} - \frac{e^{-H_0 - g'(x)V'}}{\text{Tr}[e^{-H_0 - g'(x)V'}]}||_1$$
$$\le 2(e^{||g(x)V - g'(x)V'||} - 1)$$
$$\le 2(e^{(||V(g(x) - g'(x))|| + ||(V - V')g'(x)||)} - 1)$$
$$\le 2e^{\epsilon(B+b)}(||V(g(x) - g'(x))|| + ||(V - V')g'(x)||). \tag{274}$$

Therefore,

$$||f - f'||_{1,1,\vec{x}} \le 2e^{\epsilon(B+b)}\epsilon(b+B), \tag{275}$$

which implies that $\mathcal{C}_\epsilon = \{f \,|\, f(x) = \frac{e^{-H_0 - g(x)V}}{\text{Tr}[e^{-H_0 - g(x)V}]}, \; g \in \mathcal{F}_\epsilon, V \in \mathcal{M}_{b,\epsilon}^{(d)}\}$, is a $2e^{\epsilon(B+b)}\epsilon(b+B)$-net for $\mathcal{C}$. ∎

Proposition 7 (Phase shifts with position-dependent depth). Consider a spatially local channel that acts as a power of an unknown unitary channel at position x according to some classical variable at position x (for example, a thickness), measured as $g(x)$, which we probe

with some position-dependent state $\rho(x)$, possibly entangled with a different reference system for each x. We can model this class as either a state-valued class, but also as a projector-valued function class if the states $\rho(x)$ are pure:

$$\mathcal{C} = \{f \,|\, f(x) = U^{g(x)}\rho(x)(U^\dagger)^{g(x)} \; g \in \mathcal{F}, \tag{276}$$
$$U = e^{iH} \in \text{U}(d), H \in \mathcal{M}_b^{(d)}\}. \tag{277}$$

The covering number of this class is bounded as

$$\gamma_{1,1}(n, 2e^{\epsilon(B+b)+Bb}\epsilon(b+B), \mathcal{C}) \tag{278}$$
$$\le 2\left(\frac{4nB^2}{\epsilon^2}\right)^{D\log_2(4eBn/(D\epsilon))} \times \left(\frac{6b}{\epsilon}\right)^{d^2}, \tag{279}$$

and the same bound holds for $\gamma_{1,\infty}(n, 2e^{\epsilon(B+b)+Bb}\epsilon(b+B), \mathcal{C})$ when $\rho(x)$ are pure.

Proof. Let $\mathcal{F}_\epsilon$ be an ϵ-net of $\mathcal{F}$ and $\mathcal{M}_{b,\epsilon}^{(d)}$ an ϵ-net of $\mathcal{M}_b^{(d)}$. Then, for every $g \in \mathcal{F}$ there is $g \in \mathcal{F}_\epsilon$ such that $||g - g'||_{1,\vec{x}} \le \epsilon$ and for every $H \in \mathcal{M}_b^{(d)}$ there is $H' \in \mathcal{M}_{b,\epsilon}^{(d)}$ such that $||H - H'|| \le \epsilon$. Let us consider the corresponding functions $f(x) = U^{g(x)}\rho(x)(U^\dagger)^{g(x)} \in \mathcal{C}$, $f'(x) = U'^{g'(x)}\rho(x)(U'^\dagger)^{g'(x)} \in \mathcal{C}$.

Using Eq. (267) therefore, since $||H|| \le b$ and $|g(x)| \le B$

$$||e^{iHg(x)} - e^{iH'g'(x)}||$$
$$\le ||Hg(x) - H'g'(x)||e^{||Hg(x) - H'g'(x)||}e^{||Hg(x)||}$$
$$\le (||H(g(x) - g'(x))|| + ||(H - H')g'(x)||)e^{3(B+b)}. \tag{280}$$

Moreover, applying Eq. (234) (or Proposition 4 for the projector-valued case) to the pseudometric for our class, we obtain that for both $q = 1, \infty$

$$||f - f'||_{1,q,\vec{x}} = \frac{1}{n}\sum_{i=1}^n ||e^{iHg(x_i)}\rho(x_i)e^{-iHg(x_i)}$$
$$- e^{iH'g'(x_i)}\rho(x_i)e^{-iH'g'(x_i)}||_q \tag{281}$$

$$\le 2e^{\epsilon(B+b)+Bb}\epsilon(b+B), \tag{282}$$

which implies that $\mathcal{C}_\epsilon = \{f \,|\, f(x) = U^{g(x)}\rho(x)(U^\dagger)^{g(x)}, \; g \in \mathcal{F}_\epsilon, V \in \mathcal{M}_{b,\epsilon}^{(d)}\}$, is an $2e^{\epsilon(B+b)+Bb}\epsilon(b+B)$-net for $\mathcal{C}$. ∎

Proposition 8 (Projectors onto low-energy subspace of a position-dependent Hamiltonian). Consider a class that projects on the eigenspaces of low energy of a set of unitary equivalent Hamiltonians, defining Π_E as the projector

on eigenstates of H_0 with energy less than E. We can model this as

$$\mathcal{C} = \{f \,|\, f(x) = U^{g(x)} \Pi_E (U^\dagger)^{g(x)} \, g \in \mathcal{F},$$
$$U = e^{iH} \in \mathrm{U}(d), H \in \mathcal{M}_b^{(d)}\}. \tag{283}$$

The covering number of this class is bounded as

$$\gamma_{1,\infty}(n, 2e^{3(B+b)}\epsilon(b+B), \mathcal{C})$$
$$\leq 2 \left(\frac{4mB^2}{\epsilon^2}\right)^{D \log_2(4eBn/(D\epsilon))} \times \left(\frac{6b}{\epsilon}\right)^{d^2}. \tag{284}$$

Proof. The proof is identical to the one of Proposition 7 in the projector-valued case, exchanging $\rho(x)$ with $\Pi_E(x)$. ∎

Proposition 9 (Projectors on low-energy subspace of perturbed Hamiltonians). Similarly, we can consider a class that projects on the eigenspaces of low energy of a set of perturbed Hamiltonians, defining Π_E as the projector on eigenstates of $H_0 + g(x)V$ with energy less than E, where H_0 does not have eigenvalues in the interval $[E - 2\delta, E + 2\delta]$, denoted $P_E(H_0 + g(x)V)$.

$$\mathcal{C} = \{f \,|\, f(x) = \Pi_E(H_0 + g(x)V), \; g \in \mathcal{F}, V \in \mathcal{M}_{\delta/B}^{(d)}\}, \tag{285}$$

where we choose parameters in such a way that the perturbation is small, $||g(x)V|| < \delta$. The covering number of this class is bounded as

$$\gamma_{1,\infty}\left(n, \frac{\pi\epsilon}{4\delta}\left(B + \delta\frac{1}{B}\right), \mathcal{C}\right)$$
$$\leq 2 \left(\frac{4mB^2}{\epsilon^2}\right)^{D \log_2(4eBn/(D\epsilon))} \times \left(\frac{6b}{\epsilon}\right)^{d^2}. \tag{286}$$

Proof. Let $\mathcal{F}_\epsilon$ be an ϵ-net of $\mathcal{F}$ and $\mathcal{M}_{b,\epsilon}^{(d)}$ an ϵ-net of $\mathcal{M}_b^{(d)}$. Then, for every $g \in \mathcal{F}$ there is $g \in \mathcal{F}_\epsilon$ such that $||g - g'||_{1,\vec{x}} \leq \epsilon$ and for every $H \in \mathcal{M}_b^{(d)}$ there is $H' \in \mathcal{M}_{b,\epsilon}^{(d)}$ such that $||H - H'|| \leq \epsilon$. Let us consider the corresponding functions $f(x) = \Pi_E(H_0 + g(x)V) \in \mathcal{C}, f'(x) = \Pi_E(H_0 + g'(x)V') \in \mathcal{C}$.

Using Eq. (271), we have

$$||\Pi_E(H_0 + g(x)V) - \Pi_E(H_0 + g'(x)V')||$$
$$\leq \frac{\pi}{4\delta}||Hg(x) - H'g'(x)||$$
$$\leq \frac{\pi}{4\delta}\left(\frac{\delta}{B}|g(x) - g'(x)| + B||H - H'||\right). \tag{287}$$

Therefore, we get

$$||f - f'||_{1,\infty,\vec{x}}$$
$$= \frac{1}{n}\sum_{i=1}^{n}||\Pi_E(H_0 + g(x_i)V) - \Pi_E(H_0 + g'(x_i)V')|| \tag{288}$$
$$\leq \frac{\pi\epsilon}{4\delta}\left(B + \delta\frac{1}{B}\right), \tag{289}$$

which implies that $\mathcal{C}_\epsilon = \{f \,|\, f(x) = \Pi_E(H_0 + g(x)V), \; g \in \mathcal{F}_\epsilon, V \in \mathcal{M}_{b,\epsilon}^{(d)}\}$, is a $\frac{\pi\epsilon}{4\delta}(B + \delta\frac{1}{B})$-net for $\mathcal{C}$. ∎

APPENDIX A: PROOF OF THRESHOLD REPORTING FOR PRODUCT STATE

Proof of Theorem 7. We will use the following facts already listed in Ref. [4]:

(a) For S a random variable and $f : \mathbb{R} \to \mathbb{R}$ 1-Lipschitz,

$$\mathrm{Var}[f(\mathsf{S})] \leq \mathrm{Var}[\mathsf{S}] \qquad \text{(A1)}$$

(Proposition 3.1 in Ref. [4]).

(b) Fix $0 \leq p \leq 1, q = 1 - p$. Then for $C = (e - 1)^2 \leq 3$, we have

$$q + pe^{2\lambda} \leq (1 + Cpq\lambda^2) \times (q + pe^\lambda)^2 \quad \forall \lambda \in [0, 1]. \qquad \text{(A2)}$$

(Lemma 3.2 in Ref. [4]).

(c) Let S be a discrete random variable, and let B be an event on the same probability space with $\Pr[B] < 1$. For each outcome s of S, define $f(s) = \Pr[B \mid \mathsf{S} = s]$. Then

$$d_{\chi^2}((\mathsf{S} \mid \overline{B}), \mathsf{S}) = \mathrm{Var}[f(\mathsf{S})] \big/ \Pr[\overline{B}]^2. \qquad \text{(A3)}$$

(Proposition 3.3 in Ref. [4]).

We take X to be an exponential random variable with average $\mathbb{E}[\mathsf{X}] = 1/\lambda$, and we assume $\lambda \leq 1/\sqrt{\sum_{i=1}^n p_i q_i}$ and $\lambda \leq 1$.

Note that

$$\begin{aligned} f(s) &= \Pr[\mathsf{X} > \theta n - s] = \min(1, g(s)), \\ g(s) &= \exp(-\lambda(\theta n - s)). \end{aligned} \qquad \text{(A4)}$$

Therefore, we have from Eq. (A3)

$$d_{\chi^2}((\mathsf{T} \mid \overline{B}), \mathsf{T}) = \mathrm{Var}[f(\mathsf{T})] \big/ \Pr[\overline{B}]^2 \leq \frac{16}{9} \mathrm{Var}[f(\mathsf{T})],$$

and we need to show that

$$\mathrm{Var}[f(\mathsf{T})] \lesssim \Pr[B]^2 \cdot \left(\sum_{i=1}^n p_i q_i\right) \lambda^2.$$

We can use the fact that $y \mapsto \min(1, y)$ is 1-Lipschitz, therefore from Eq. (A1) we have that $\mathrm{Var}[f(\mathsf{T})] \leq \mathrm{Var}[g(\mathsf{T})]$.

As in the original proof, $\mathrm{Var}[g(\mathsf{T})]$ can be computed using the moment-generating function, this time of the Poisson binomial distribution T, namely $\mathbb{E}[\exp(t\mathsf{T})] = \prod_{i=1}^n (q_i + p_i e^t)$:

$$\mathbb{E}[g(\mathsf{T})] = \mathbb{E}[\exp(-\lambda(\theta n - \mathsf{T}))]$$

$$= \exp(-\lambda \theta n) \times \prod_{i=1}^n (q_i + p_i e^\lambda),$$

$$\mathbb{E}[g(\mathsf{T})^2] = \mathbb{E}[\exp(-2\lambda(\theta n - \mathsf{T}))]$$

$$= \exp(-2\lambda \theta n) \times \prod_{i=1}^n (q_i + p_i e^{2\lambda}).$$

Thus

$$\begin{aligned} \mathrm{Var}[g(\mathsf{T})] &= \mathbb{E}[g(\mathsf{T})]^2 \times \left(\frac{\mathbb{E}[g(\mathsf{T})^2]}{\mathbb{E}[g(\mathsf{T})]^2} - 1\right) \\ &= \mathbb{E}[g(\mathsf{T})]^2 \times \left(\left(\prod_{i=1}^n \frac{q_i + p_i e^{2\lambda}}{(q_i + p_i e^\lambda)^2}\right) - 1\right) \\ &\leq \mathbb{E}[g(\mathsf{T})]^2 \times \left(\prod_{i=1}^n (1 + 3p_i q_i \lambda^2) - 1\right) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{(Eq. A2)} \\ &\leq \mathbb{E}[g(\mathsf{T})]^2 \times \left((1 + 3\frac{1}{n}\sum_{i=1}^n p_i q_i \lambda^2)^n - 1\right) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{(AM-GM)} \\ &\leq 3\mathbb{E}[g(\mathsf{T})]^2 \times \sum_{i=1}^n p_i q_i \lambda^2. \quad (\text{as } \lambda^2 \leq \tfrac{1}{\sum_{i=1}^n p_i q_i}) \end{aligned}$$

Furthermore

$$\mathbb{E}[g(\mathsf{T})] = \exp(-\lambda \theta n) \times \prod_{i=1}^n (q_i + p_i e^\lambda)$$

$$\leq \exp(-\lambda \theta n) \times \left(\overline{q} + \overline{p} e^\lambda\right)^n, \qquad \text{(A5)}$$

where $\overline{q} := 1/n \sum_{i=1}^n q_i$ and $\overline{p} := 1/n \sum_{i=1}^n p_i$, and we will show

$$\exp(-\lambda \theta n) \times \left(\overline{q} + \overline{p} e^\lambda\right)^n \leq D \Pr[B], \qquad \text{(A6)}$$

for some other constant $D > 0$

We consider two cases: $\overline{p} \geq 1/n$ and $\overline{p} < 1/n$.

a. Case 1: $\overline{p} \geq 1/n$. In this case we use that for a binomial random variable $\overline{\mathsf{T}}$ with mean $\overline{p} n$, we have $\Pr[\overline{\mathsf{T}} > \overline{p} n] \geq 1/4$ (see, e.g., Ref. [43]). Moreover, for the Poisson binomial distribution it holds that $\Pr[\mathsf{T} > k] \geq \Pr[\overline{\mathsf{T}} > k]$ if $0 \leq k \leq n\overline{p} - 1$ [44,45], therefore $\Pr[\mathsf{T} \geq \overline{p} n - 1] \geq \Pr[\overline{\mathsf{T}} > \overline{p} n - 1] \geq 1/4$.

Now observe that (i) $\theta \geq \overline{p} - 1/n$, since we are assuming $\Pr[B] = \Pr[\mathsf{T} + \mathsf{X} > \theta n] < 1/4$; and, (ii)

$$\Pr[B] \geq \Pr[\mathsf{T} \geq \overline{p} n - 1] \cdot \Pr[\mathsf{T} + \mathsf{X} \geq \theta n | \mathsf{T} \geq \overline{p} n - 1]$$

$$\geq \frac{1}{4} \Pr[\mathsf{X} \geq (\theta - \overline{p})n + 1]$$

$$\geq \frac{1}{4} \exp(-\lambda((\theta - \overline{p})n + 1)), \qquad \text{(A7)}$$

where the first inequality used independence of T and X, the second inequality the fact that $\Pr[T \geq \bar{p}n - 1] \geq 1/4$ and the third inequality used $(\theta - \bar{p})n + 1 \geq 0$ [by (i)]. Thus to establish our claim, it remains to show

$$\exp(-\lambda\theta n) \times (\bar{q} + \bar{p}e^{\lambda})^n \lesssim \exp(-\lambda((\theta - \bar{p})n + 1)),$$

which is equivalent to

$$\bar{q} + \bar{p}e^{\lambda} \lesssim \exp(\lambda\bar{p}).$$

But this last inequality indeed holds, as for $\lambda \leq 1$ we have $\bar{q} + \bar{p}e^{\lambda} < \bar{q} + \bar{p}(1 + 2\lambda) = 1 + 2\bar{p}\lambda$, and $1 + 2\bar{p}\lambda \leq 2\exp(\lambda\bar{p})$ always.

b. Case 2: $\bar{p} < 1/n$. Here $\bar{q} + \bar{p}e^{\lambda} < 1 + 2\bar{p}\lambda \leq 1 + 2/n$, and so $(\bar{q} + \bar{p}e^{\lambda})^n \lesssim 1$, which implies that Eq. (A6) follows from $\Pr[B] \geq \Pr[X > \theta n] = \exp(-\lambda\theta n)$. ∎

APPENDIX B: RESULTS ON SAMPLING WITHOUT REPLACEMENT

The following theorem due to Hoeffding holds [46] (see also Ref. [47]).

Theorem 17. Let $\mathcal{X} = x_1, \ldots, x_N$ a finite population of points, and $\{X_1, \ldots, X_n\}$ a random sample drawn without replacement from $\mathcal{X}$. Let

$$a = \min_{1 \leq i \leq N} x_i \qquad b = \max_{1 \leq i \leq N} x_i. \tag{B1}$$

Then it holds

$$\Pr\left(\sum_{i=1}^{n} X_i - n\mu \geq n\epsilon\right) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}. \tag{B2}$$

where $\mu := 1/N \sum_i x_i$ is the empirical mean of $\mathcal{X}$.

Now we apply the above theorem to M different finite populations of N points.

Theorem 18. Let $\mathcal{Y} = 1, \ldots, N$, and let indices $= \{X_1, \ldots, X_n\}$ a random sample drawn without replacement from $\mathcal{Y}$. Furthermore, consider M different finite populations of N numbers $\{\mathcal{X}_i\}_{i=1}^M = \{\{x_{i,j}\}_{j=1}^N\}_{i=1}^M$ with $0 \leq x_{i,j} \leq 1$, for every i, j, and for each $\mathcal{X}_i$, use the indices to choose a subset of the points $\{x_{i,j}\}_{j \in \text{indices}}$.

Then it holds

$$\Pr\left(\max_{i=1,\ldots,M} |\sum_{j=1}^{n} x_{i,X_j} - n\mu_i| \geq n\epsilon\right) \leq 2Me^{-2n\epsilon^2}, \tag{B3}$$

where $\mu_i := 1/N \sum_{j=1}^N x_{i,j}$ is the empirical mean of $\mathcal{X}_i$.

Proof. From Theorem 17, consider the $2M$ finite populations $\{\mathcal{X}_i, -\mathcal{X}_i\}_{i=1,\ldots,M}$ (the negative sign is to handle the absolute value). The indices determine the choice of a sample without replacement from each of the above finite populations. The proposition follows from the union bound. ∎

We can thus prove the following theorem.

Theorem 19. Let $\mathcal{Y} = \{1, \ldots, N\}$, $N \geq 3Kn$ and let the K sets of indices $\{\{X_{nk+i}\}_{i=1}^n\}_{k=0,\ldots,K-1}$ be random samples drawn without replacement from $\mathcal{Y}$. Furthermore, consider m different finite populations of N numbers $\{\{x_{i,j}\}_{j=1}^N\}_{i=1}^m$, with $0 \leq x_{i,j} \leq 1$. From each population $\{x_{i,j}\}_{j=1}^N$ obtain K subsets of size n, as $\{x_{i,X_{nk+1}}, \ldots, x_{i,X_{nk+n}}\}_{k=0,\ldots,K-1}$. Then it holds

$$\Pr\left(\max_{\substack{i=1,\ldots,m \\ k=0,\ldots,K-1}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \geq n\epsilon\right) \leq 2Kme^{-2n\epsilon^2/4}, \tag{B4}$$

where $\mu_i = 1/N \sum_{j=1}^N x_{i,j}$.

Proof. We compute an upper bound recursively using

$$\Pr\left(\max_{\substack{i=1,\ldots,m \\ k=0,\ldots,t}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \geq n\epsilon\right) \leq \Pr\left(\max_{\substack{i=1,\ldots,m \\ k=0,\ldots,t-1}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \geq n\epsilon\right)$$

$$+ \Pr\left(\max_{i=1,\ldots,m} |\sum_{j=1}^{n} x_{i,X_{nt+j}} - n\mu_i| \geq n\epsilon \Big| \max_{\substack{i=1,\ldots,m \\ k=0,\ldots,t-1}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \leq n\epsilon\right). \tag{B5}$$

Defining $\mu_{i,t}$ such that

$$N\mu_i = (N - tn)\mu_{i,t} + \sum_{k=0}^{t-1}\sum_{j=1}^{n} x_{i,X_{nk+j}}, \qquad \text{(B6)}$$

given the condition $\max\limits_{\substack{i=1,\ldots,m \\ k=0,\ldots,t-1}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \leq n\epsilon,$

we have

$$N\mu_i \leq (N - tn)\mu_{it} + tn\mu_i + tn\epsilon \qquad \text{(B7)}$$

$$N\mu_i \geq (N - tn)\mu_{it} + tn\mu_i - tn\epsilon, \qquad \text{(B8)}$$

and therefore, for $t \leq K - 1$

$$\mu_{it} \geq \frac{N - tn}{N - tn}\mu_i - \frac{tn}{N - tn}\epsilon \geq \mu_i - \frac{K-1}{2K+1}\epsilon \geq \mu_i - \epsilon/2$$
$$\mu_{it} \leq \frac{N - tn}{N - tn}\mu_i + \frac{tn}{N - tn}\epsilon \leq \mu_i + \frac{K-1}{2K+1}\epsilon \leq \mu_i + \epsilon/2, \qquad \text{(B9)}$$

meaning that

$$|n\mu_{it} - n\mu_i| \leq n\epsilon/2, \qquad \text{(B10)}$$

and

$$\Pr\left(\max_{i=1,..,m} |\sum_{j=1}^{n} x_{i,X_{nt+j}} - n\mu_i| \geq n\epsilon \,\Big|\, \max_{\substack{i=1,\ldots,m \\ k=0,\ldots,t-1}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \leq n\epsilon\right)$$

$$= \Pr\left(\max_{i=1,..,m} |\sum_{j=1}^{n} x_{i,X_{nt+j}} - n\mu_{it} + n\mu_{it} - n\mu_i| \geq n\epsilon \,\Big|\, \max_{\substack{i=1,\ldots,m \\ k=0,\ldots,t-1}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \leq n\epsilon\right)$$

$$\leq \Pr\left(\max_{i=1,..,m} |\sum_{j=1}^{n} x_{i,X_{nt+j}} - n\mu_{it}| + |n\mu_{it} - n\mu_i| \geq n\epsilon \,\Big|\, \max_{\substack{i=1,\ldots,m \\ k=0,\ldots,t-1}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \leq n\epsilon\right)$$

$$\leq \Pr\left(\max_{i=1,..,m} |\sum_{j=1}^{n} x_{i,X_{nt+j}} - n\mu_{it}| \geq n\epsilon/2 \,\Big|\, \max_{\substack{i=1,\ldots,m \\ k=0,\ldots,t-1}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \leq n\epsilon\right). \qquad \text{(B11)}$$

Using Theorem 18 we have

$$\Pr\left(\max_{i=1,..,m} |\sum_{j=1}^{n} x_{i,X_{nt+j}} - n\mu_{it}| \geq n\epsilon/2 \,\Big|\, \max_{\substack{i=1,\ldots,m \\ k=0,\ldots,t-1}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \leq n\epsilon\right) \leq 2me^{-\frac{2n\epsilon^2}{4}}, \qquad \text{(B12)}$$

and therefore

$$\Pr\left(\max_{\substack{i=1,\ldots,m \\ k=0,\ldots,t}} |\sum_{j=1}^{n} x_{i,X_{nk+j}} - n\mu_i| \geq n\epsilon\right) \leq 2Kme^{-\frac{2n\epsilon^2}{4}}. \qquad \text{(B13)}$$

∎

APPENDIX C: REALIZABLE LEARNING WITH PURE STATES

We first consider learning in the setting where there is a guarantee that the unknown concept $f : \mathcal{X} \to D(\mathcal{H})$ comes from the concept class $\mathcal{C}$. This setting was first considered in Ref. [7]. Additionally, Ref. [7] gives two algorithms for learning in the realizable setting, one for pure states and one for mixed states. In this section we

show that the pure-state algorithm can be straightforwardly improved if one looks at growth functions.

We will use the shorthands

$$\Delta(c, c') = \sum_x \mathcal{D}(x) d_{\text{tr}}(c(x), c'(x))$$

$$\tilde{\Delta}(c, c', \vec{x}) = \frac{1}{|\vec{x}|} \sum_{i=1}^{|\vec{x}|} d_{\text{tr}}(c(x_i), c'(x_i)). \tag{C1}$$

First let us make some definitions.

Definition 9 (Measurement output distribution). For any state $\sigma \in S$, and for a d-outcome POVM $\mathcal{M} := \{M_z\}_{z=1}^d$ where $M_z \in \mathbb{C}^d$, define the measurement output distribution $D_{\mathcal{M},\sigma} : [d] \to [0, 1]$ as the distribution over outcomes upon measuring the state σ with $\mathcal{M}$:

$$D_{\mathcal{M},\sigma}(z) = \text{Tr}[\sigma M_z]. \tag{C2}$$

Definition 10 (ϵ-far set). For any $c \in \mathcal{C}$, define the ϵ-far set: $\mathcal{C}_{\geq\epsilon}(c) := \{h \in \mathcal{C} : \Delta(h, c) \geq \epsilon\}$.

The error occurs if maximum likelihood chooses a concept h, which is 2ϵ-far from the true one (i.e., $\Delta(h, c^*)$), which could happen for two reasons:

(1) There is a concept h such that $\Delta(h, c^*) > 2\epsilon$ but $\tilde{\Delta}(h, c^*, \vec{x}) < \epsilon$, i.e., the sampled points do not give a good approximation of h's distance from the true concept, which affects the construction of a good covering.

(2) There is a concept h such that $\Delta(h, c^*) > 2\epsilon$ but maximum likelihood on samples from $D_{\otimes_i \mathcal{M}^{(i)}, \otimes_i c^*(x_i)}$ outputs h, i.e., the sampled points

---

ALGORITHM 4.   Learning channels mapping to pure states

Input: $(x_1, |c^*(x_1)\rangle), (x_2, |c^*(x_2)\rangle), \ldots, (x_n, |c^*(x_n)\rangle)$

1: Restrict the concept class $\mathcal{C}$ into $\mathcal{C}|_{\vec{x}}$.
2: Do a random d-outcome orthonormal basis measurement $\mathcal{M}^{(i)}$ on each output state $|c^*(x_i)\rangle$ (if necessary augmenting the space with auxiliary variables) so that the measurement on the overall state is $\bigotimes_{i=1}^n \mathcal{M}^{(i)} := \{\bigotimes_{i=1}^n M_{z_i}^{(i)} \mid \vec{z} \in [d]^n\}$. Let the outcome of the measurement $\mathcal{M}^{(i)}$ be $z_i \in [d]$.
3: Output the concept $h \in \mathcal{C}|_{\vec{x}}$ that is most likely to have produced the measurement outcomes recorded in Line 2:

$$h = \arg\max_{c \in \mathcal{C}|_{\vec{x}}} \prod_{i=1}^T D_{\mathcal{M}^{(i)}, c(x_i)}(z_i) \quad \text{(Maximum Likelihood)}$$

$$\tag{C3}$$

---

do not give a good approximation of the measurement output distribution of the true concept. The probability of this happening is smaller than

$$1 - TV(D_{\otimes_i \mathcal{M}^{(i)}, \otimes_i c^*(x_i)}, D_{\otimes_i \mathcal{M}^{(i)}, \otimes_i h(x_i)}), \tag{C4}$$

as proven in Ref. [7].

To bound the probability of error from the first source, we note that for fixed $h \in \mathcal{C}|_{\vec{x}}$,

$$\Pr_{\vec{x} \sim \mathcal{D}} \left[\tilde{\Delta}(h, c^*, \vec{x}) < \epsilon \text{ and } \Delta(h, c^*) \geq 2\epsilon \right]$$

$$\leq \Pr_{\vec{x} \sim \mathcal{D}} \left[\left| \tilde{\Delta}(h, c^*, \vec{x}) - \Delta(h, c^*) \right| \geq \epsilon \right] \leq 2 \exp(-2\epsilon^2 n),$$

$$\tag{C5}$$

where the second inequality follows from Höffding's inequality. Hence, with probability $1 - \exp(-2\epsilon^2 T)$ over the distribution $\mathcal{D}$, the samples $x_1, \ldots x_T$ are such that

$$\Delta(h, c^*) \geq 2\epsilon \Rightarrow \tilde{\Delta}(h, c^*, \vec{x}) \geq \epsilon. \tag{C6}$$

Let us now condition on this event and bound the probability of error from the second source. We use the fact [48] that for any two d-dimensional quantum states $|\psi\rangle, |\psi'\rangle$ with $r := \text{rank}(|\psi\rangle - |\psi'\rangle)$ and a random orthogonal measurement basis M, with probability $1 - \exp(-kd/r)$, which can be assumed to be sufficiently close to 1 (by padding auxiliary quantum registers) over the choice of the basis,

$$TV(D_{\mathcal{M}, |\psi\rangle}, D_{\mathcal{M}, |\psi'\rangle}) \geq \frac{k}{\sqrt{2}} d_{\text{tr}}(|\psi\rangle, |\psi'\rangle), \tag{C7}$$

for k a universal constant. Condition on this event too. Hence for any $h \in \mathcal{C}_{>2\epsilon}(c^*)$,

$$\frac{1}{n} \sum_{i=1}^n TV(D_{M, h(x_i)}, D_{M, c^*(x_i)})$$

$$> \frac{k}{\sqrt{2}} \frac{1}{n} \sum_{i=1}^n d_{\text{tr}}(|h(x_i)\rangle, |c^*(x_i)\rangle) \geq \frac{k}{\sqrt{2}} \epsilon. \tag{C8}$$

The first inequality follows by linearity of Eq. (C7). The second inequality follows from the conditioning event Eq. (C6). Finally, the probability of maximum likelihood on samples from $D_{\otimes_i \mathcal{M}^{(i)}, \otimes_i c^*(x_i)}$ erroneously assigning them to $D_{\otimes_i \mathcal{M}^{(i)}, \otimes_i h(x_i)}$ is at most

$$1 - TV(D_{\otimes_i \mathcal{M}^{(i)}, \otimes_i c^*(x_i)}, D_{\otimes_i \mathcal{M}^{(i)}, \otimes_i h(x_i)})$$

$$\leq 2^{-\Omega(n \cdot (\frac{1}{n} \sum_{i=1}^n TV(D_{\mathcal{M}, h(x_i)}, D_{\mathcal{M}, c^*(x_i)}))^2)}$$

$$\leq 2^{-\Omega(n \cdot k^2 \epsilon^2 / 2)}. \tag{C9}$$

The first inequality follows from Lemma 6 of Ref. [7]; the second inequality follows from Eq. (C8).

Having accounted for both sources of error for a fixed hypothesis h in the "bad" set $\mathcal{C}|_{\vec{x}} \cap \mathcal{C}_{>2\epsilon}(c^*)$, we now take a union bound over all bad hypotheses. Thus the total probability of error is at most

$$
\begin{aligned}
\left|\mathcal{C}|_{\vec{x}} \cap \mathcal{C}_{>2\epsilon}(c^*)\right| & \left(2^{-2\epsilon^2 T} + 2^{-\Omega(T \cdot k^2 \epsilon^2/2)}\right) \\
& \leq G(T) 2^{-\Omega(\epsilon^2 T)}. \quad \text{(C10)}
\end{aligned}
$$

Thus, there exists a finite number of samples that makes the right-hand-side of Eq. (C10) smaller than δ if

$$
\lim_{T \to \infty} \frac{1}{\epsilon^2} \log\left(\frac{G(T)}{\delta}\right) / T = 0. \quad \text{(C11)}
$$

Note that if $G(T) \leq \mathcal{C}$ for all $T < \infty$ one recovers the sample complexity of Ref. [7].

APPENDIX D: LEARNING VIA RESTRICTED RESTRICTED EMPIRICAL RISK ESTIMATION

The following variant of risk minimization via ERM on a projector-valued concept class, which is not necessarily an ϵ-net on the full input data shows that risk minimization can be done even if a less restrictive condition on $\gamma_{1,\infty}(n, \epsilon, \mathcal{C})$ holds. Risk estimation for projector-valued concept classes and risk minimization for state-valued concept classes can also be realized with a similar modification.

Theorem 20 (Learning quantum processes via ERM, alternative statement). Suppose the concept class $\mathcal{C}$ consists of quantum processes mapping to projectors and let $\epsilon > 0$ be the accuracy parameter, and suppose that

$$
\lim_{n \to \infty} \frac{\log \gamma_{1,\infty}(n, \epsilon, \mathcal{C})}{n} = 0, \qquad \forall \epsilon > 0. \quad \text{(D1)}
$$

Given as input a training set $S = (x_i, \rho(x_i))_{i=1}^n$ with $x_i \xleftarrow{\mathcal{D}} X$ and $\rho(\cdot)$ an unknown classical-quantum channel, there is an *agnostic* learning algorithm $\mathcal{A} : \mathcal{X}^n \times \mathcal{L}(\mathcal{H}^{(d)})^{\otimes n} \to \mathcal{C}$ such that

$$
\Pr_S[R_\rho(\mathcal{A}(S)) - \inf_{h \in \mathcal{C}} R_\rho(h) < 9\epsilon] =: 1 - p_{\text{err}}. \quad \text{(D2)}
$$

Writing $n = Tkl$, for large enough $T = O(\log(1/\epsilon))$, $k = O(\log(1/\delta \log(1/\epsilon)))$, and $m_0 \leq Tkl$, such that

$$
4\gamma_{1,q}(2m_0, \epsilon/64, \mathcal{C}) e^{-\frac{m_0 \epsilon^2}{512}} \leq \delta/4, \quad \text{(D3)}
$$

there exist constants C_1, C_2, C_3 such that, if l satisfies

$$
(\log \gamma_{1,\infty}(m_0, \epsilon/2, \mathcal{C}) + C_2)^2 \leq C_1 l \epsilon^2/9, \quad \text{(D4)}
$$

we have

$$
\begin{aligned}
p_{\text{err}} \leq {} & \frac{\delta}{2} + 4\gamma_{1,\infty}(6Tkl, \epsilon/2, \mathcal{C}) e^{-\frac{Tkl\epsilon^2}{32}} \\
& + C_3 \log \frac{1}{\epsilon} \log\left(\frac{1}{\delta} \log \frac{1}{\epsilon}\right) \gamma_{1,\infty}(m_0, \epsilon/2, \mathcal{C}) e^{-\frac{l\epsilon^2}{72}}. \\
& \hspace{9cm} \text{(D5)}
\end{aligned}
$$

Once this is established, it can be verified that p_{err} can be made arbitrarily small if

$$
\lim_{n \to \infty} \frac{\log \gamma_{1,\infty}(n, \epsilon, \mathcal{C})}{n} = 0, \qquad \forall \epsilon > 0.
$$

Proof. The algorithm takes the first m_0 classical variables $\vec{x}_0$ and obtains an ϵ-net on the true risks $\mathcal{C}|_{\vec{x}_0}$, with probability of error bounded as in Lemma 6

$$
p_{\text{err, net}}(m_0) \leq 4\gamma_1(2m_0, \epsilon/64, \mathcal{C}) e^{-\frac{m_0 \epsilon^2}{512}}. \quad \text{(D6)}
$$

Then it runs algorithm 2 on the product state obtained from the full dataset $\vec{x}$, with the concepts in $\mathcal{C}|_{\vec{x}_0}$. In this way we obtain c^* such that $\hat{R}(c) < \max_{c \in \mathcal{C}|_{\vec{x}_0}} \hat{R}(c) + 6\epsilon$ with probability bounded as in the proof of Theorem 10

$$
\begin{aligned}
p_{\text{err,ERM}} \leq {} & T(0.97^k + (k+1)e^{-l\epsilon^2/72}) \\
& + 2Tk\gamma_1(m_0, \epsilon/2, \mathcal{C}) e^{-2l\epsilon^2/64}. \quad \text{(D7)}
\end{aligned}
$$

Finally, we require that uniform convergence occurs also for the full dataset (not just the first m_0 data). This happens with probability

$$
\begin{aligned}
p_{\text{err,unif}} = {} & \Pr_{S \sim \mathcal{D}^n}[\exists c \in \mathcal{C} : |R(c) - \hat{R}(c)| \geq \epsilon] \\
& \leq 4\gamma_{1,\infty}(12Tkl, \epsilon/8, \mathcal{C}) e^{-\frac{6Tkl\epsilon^2}{32}}. \quad \text{(D8)}
\end{aligned}
$$

If this is true we have that, for the selected concept c

$$
\begin{aligned}
R(c) & \leq \hat{R}(c) + \epsilon & \text{(D9)} \\
& \leq \sup_{c \in \mathcal{C}|_{\vec{x}_0}} \hat{R}(c) + 7\epsilon & \text{(D10)} \\
& \leq \sup_{c \in \mathcal{C}|_{\vec{x}_0}} R(c) + 8\epsilon & \text{(D11)} \\
& \leq \sup_{c \in \mathcal{C}} R(c) + 9\epsilon. & \text{(D12)}
\end{aligned}
$$

The probability of error is then upper bounded as

$$
p_{\text{err,net}} + p_{\text{err,ERM}} + p_{\text{err,unif}}. \quad \text{(D13)}
$$

With the choices made for the parameters, we obtain the thesis. ∎

[1] R. P. Feynman, Quantum mechanical computers, Found. Phys. **16**, 507 (1986).

[2] S. Aaronson, The learnability of quantum states, Proc. R. Soc. A Math. Phys. Eng. Sci. **463**, 3089 (2006).

[3] S. Aaronson, Shadow tomography of quantum states, SIAM J. Comput. **49**, STOC18-368 (2020).

[4] C. Bădescu and R. O'Donnell, in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC 2021)* (Association for Computing Machinery, New York, NY, USA, 2021), pp. 1398–1411.

[5] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill, Provably efficient machine learning for quantum many-body problems, Science **377**, eabk3333 (2022).

[6] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).

[7] K.-M. Chung and H.-H. Lin, in *16th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2021). Leibniz International Proceedings in Informatics (LIPIcs)* (Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021), Vol. 197, pp. 3:1–3:22.

[8] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, and J. Preskill *et al.*, Quantum advantage in learning from experiments, Science **376**, 1182 (2022).

[9] M. C. Caro, Binary classification with classical instances and quantum labels, Quantum Mach. Intell. **3**, 1 (2021).

[10] C.-Y. Lai and H.-C. Cheng, Learning quantum circuits of some T gates, IEEE Trans. Inf. Theory **68**, 3951 (2022).

[11] V. N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Networks **10**, 988 (1999).

[12] M. J. Kearns and R. E. Schapire, Efficient distribution-free learning of probabilistic concepts, J. Comput. Syst. Sci. **48**, 464 (1994).

[13] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations* (Cambridge University Press, Cambridge, 1999).

[14] M. Wolf, Mathematical foundations of supervised learning (2021), https://mediatum.ub.tum.de/doc/1723378/1723378.pdf.

[15] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, 2018).

[16] F. G. S. L. Brandão, A. W. Harrow, and M. Horodecki, Local random quantum circuits are approximate polynomial-designs, Commun. Math. Phys. **346**, 397 (2016).

[17] S. Aaronson and G. N. Rothblum, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (STOC 2019) (Association for Computing Machinery, New York, NY, USA, 2019), pp. 322–333.

[18] H.-Y. Huang, R. Kueng, and J. Preskill, Information-theoretic bounds on quantum advantage in machine learning, Phys. Rev. Lett. **126**, 190505 (2021).

[19] L. Banchi, J. Pereira, and S. Pirandola, Generalization in quantum machine learning: A quantum information standpoint, PRX Quantum **2**, 040321 (2021).

[20] H.-Y. Huang, S. Chen, and J. Preskill, Learning to predict arbitrary quantum processes, PRX Quantum **4**, 040337 (2023).

[21] M. C. Caro, Learning quantum processes and hamiltonians via the Pauli transfer matrix, arXiv:2212.04471.

[22] D. Aharonov, J. Cotler, and X.-L. Qi, Quantum algorithmic measurement, Nat. Commun. **13**, 887 (2022).

[23] K. Chen, Q. Wang, P. Long, and M. Ying, Unitarity estimation for quantum channels, IEEE Trans. Inf. Theory **69**, 5116 (2023).

[24] S. Chen, J. Cotler, H. Huang, and J. Li, in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* (Denver, CO, USA, 2022), pp. 574–585.

[25] K. Bu, D. E. Koh, L. Li, Q. Luo, and Y. Zhang, Statistical complexity of quantum circuits, Phys. Rev. A **105**, 062431 (2022).

[26] H. Cai, Q. Ye, and D.-L. Deng, Sample complexity of learning parametric quantum circuits, Quantum Sci. Technol. **7**, 025014 (2022).

[27] C. M. Popescu, Learning bounds for quantum circuits in the agnostic setting, Quantum Inf. Process. **20**, 286 (2021).

[28] M. C. Caro, E. Gil-Fuster, J. J. Meyer, J. Eisert, and R. Sweke, Encoding-dependent generalization bounds for parametrized quantum circuits, Quantum **5**, 582 (2021).

[29] Y. Du, Z. Tu, X. Yuan, and D. Tao, Efficient measure for the expressivity of variational quantum algorithms, Phys. Rev. Lett. **128**, 080506 (2022).

[30] C. Gyurik, D. Vreumingen, van, and V. Dunjko, Structural risk minimization for quantum linear classifiers, Quantum **7**, 893 (2023).

[31] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, Generalization in quantum machine learning from few training data, Nat. Commun. **13**, 4919 (2022).

[32] M. Rosati, A learning theory for quantum photonic processors and beyond, arXiv:2209.03075.

[33] Y. Quek, S. Arunachalam, and J. A. Smolin, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Curran Associates, Inc., 2021), Vol. 34, pp. 20503–20515. https://proceedings.neurips.cc/paper_files/paper/2021/file/abdbeb4d8dbe30df8430a8394b7218ef-Paper.pdf.

[34] This selection is influenced by Ref. [4], as our algorithms are heavily based on their work. We refer the reader to Ref. [4] for a more extended presentation. Note also that we refer to Lemmas 2.5 and 2.6 appearing in the v2 of the arXiv post.

[35] C. A. Fuchs and C. M. Caves, Mathematical techniques for quantum communication theory, Open Syst. Inf. Dyn. **3**, 345 (1995).

[36] P. Sen, in *2012 IEEE International Symposium on Information Theory Proceedings* (IEEE, Cambridge, MA, USA, 2012), pp. 736–740.

[37] M. M. Wilde, Sequential decoding of a general classical-quantum channel, Proc. R. Soc. A: Math., Phys. Eng. Sci. **469**, 20130259 (2013).

[38] The difference between a pseudometric and a metric is that the former is allowed to be zero for pairs of distinct points.

[39] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Data re-uploading for a universal quantum classifier, Quantum **4**, 226 (2020).

[40] C. Van Loan, The sensitivity of the matrix exponential, SIAM J. Numer Anal. **14**, 971 (1977).

[41] F. G. S. L. Brandao and K. M. Svore, in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* (Berkeley, CA, USA, 2017), pp. 415–426.

[42] R. Bhatia, *Matrix Analysis* (Springer New York, New York, NY, 1997).

[43] B. Doerr, An elementary analysis of the probability that a binomial random variable exceeds its expectation, Stat. Probab. Lett. **139**, 67 (2018).

[44] W. Hoeffding, On the Distribution of the Number of Successes in Independent Trials, The Annals of Mathematical Statistics **27**, 713 (1956).

[45] W. Tang and F. Tang, The Poisson Binomial Distribution–Old & New, arXiv:1908.10024.

[46] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. **58**, 13 (1963).

[47] R. Bardenet and O.-A. Maillard, Concentration inequalities for sampling without replacement, Bernoulli **21**, 1361 (2015).

[48] P. Sen, in *Proceedings of the 21st Annual IEEE Conference on Computational Complexity (CCC'06)* (IEEE Computer Society, USA, 2006), pp. 274–287.