Single-shot quantum machine learning

Erik Recio-Armengol [1,2] Jens Eisert [3,4,5] and Johannes Jakob Meyer [3]

[1] *ICFO-Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels (Barcelona), Spain*

[2] *Eurecat, Centre Tecnològic de Catalunya, Multimedia Technologies, 08005 Barcelona, Spain*

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

[4] *Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany*

[5] *Helmholtz-Zentrum Berlin für Materialien und Energie, 14109 Berlin, Germany*

(Received 26 July 2024; revised 26 November 2024; accepted 26 March 2025; published 15 April 2025)

Quantum machine learning aims to improve learning methods through the use of quantum computers. If it is to ever realize its potential, many obstacles need to be overcome. A particularly pressing one arises at the prediction stage because the outputs of quantum learning models are inherently random. This creates an often considerable overhead, as many executions of a quantum learning model have to be aggregated to obtain an actual prediction. In this work, we analyze when quantum learning models can evade this issue and produce predictions in a near-deterministic way—paving the way to single-shot quantum machine learning. We give a rigorous definition of single shotness in quantum classifiers and show that the degree to which a quantum learning model is near deterministic is constrained by the distinguishability of the embedded quantum states used in the model. Opening the black box of the embedding, we show that if the embedding is realized by quantum circuits, a certain depth is necessary for single shotness to be even possible. We conclude by showing that quantum learning models cannot be single shot in a generic way and trainable at the same time.

DOI: 10.1103/PhysRevA.111.042420

I. INTRODUCTION

Machine learning is a burgeoning field where rapid advances regularly overturn assumptions on what can and cannot be learned by classical computers. This ongoing success story spurs interest in *quantum machine learning*, its intersection with quantum computing, another field that has recently seen tremendous technical progress. Investigating if quantum computers can be used to construct learning models that somehow outperform their classical counterparts has become one of the principal avenues of research in quantum machine learning.

While the intrinsic quantum nature of such models gives them at least some theoretical potential to push beyond the boundary of what is classically possible [1–4], it also comes with inherent downsides. Previous research in that direction mostly focused on the training stage, where issues such as barren plateaus [5] complicate the optimization of quantum learning models. Issues do, however, also appear at the inference stage, i.e., when a prediction is to be produced by a quantum learning model. A quantum learning model that deserves that name needs to perform some sort of manipulation of a quantum system. To extract a classical label, however, a measurement needs to be performed. It is the nature of quantum mechanics that the outcomes of such measurements are inherently probabilistic. For the model to solve a learning problem, for example, to correctly classify an image, the output needs to be deterministic. In practice, a quantum learning model is run a large number of times to circumvent the probabilistic outcomes of measurements and to extract a prediction, e.g., in the form of an expectation value of a suitably chosen observable. This issue, an instance of what has been called the *measurement problem,*[1] even persists if the model is run on a fault-tolerant quantum computer.

The nondeterministic nature of quantum mechanics means that quantum learning models usually come with overheads that are not present in classical learning models. This limits their appeal in a stark way, especially since this not only increases the *time* it takes to produce a prediction, but also the *cost*. Considering a full quantum machine learning pipeline, being able to reduce the measurement overhead for a given quantum model would lower this cost at all stages. It could also be useful considering nonrepeatable setups where a quantum state is produced and needs to be classified with one shot, for instance, when detecting objects from the scattered state of light.

In this work, we study if this particular limitation can be circumvented, specifically under which conditions it is possible to extract a prediction from a single run of a quantum learning model (see Fig. 1). We establish a rigorous definition of such *single-shot models* and use tools from quantum hypothesis testing to understand their ultimate limitations. If a quantum classifier is constructed using a quantum circuit, we establish that *single shotness* can only be possible if the circuit

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

[1] This is not to be confused with the measurement problem encountered in quantum chemistry applications, where the overhead comes from having to measure many noncommuting observables.

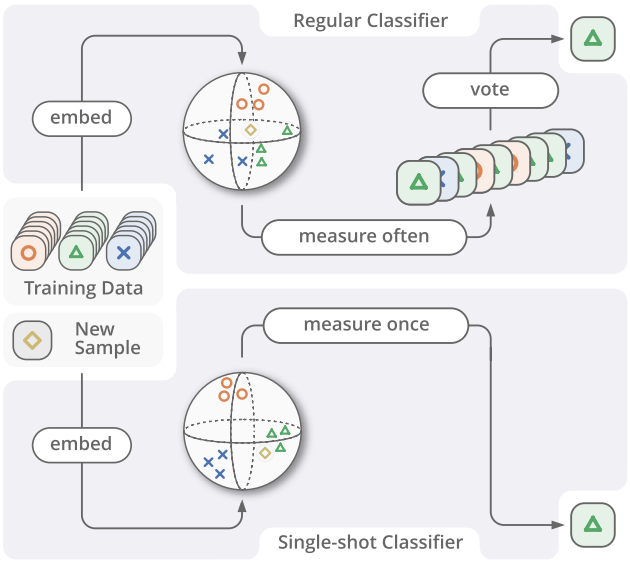


FIG. 1. A depiction of the intuitive difference between a regular quantum classifier (top panel) and a single-shot one (bottom panel). In both cases, data are embedded into a quantum system through a quantum feature map. In a regular classifier, the procedure that extracts the label has to be repeated often and then aggregated into a prediction by a majority vote. In a single-shot classifier, on the other hand, a single pass through the quantum model is sufficient to extract the label near deterministically.

has a certain depth. We treat both the noise-free and the noisy cases. We complement our result by establishing that learning models that exhibit the single-shot property generically are overly expressive and therefore hard to train.

To comprehend the full potential of quantum learning models, efforts have been put into studying their potential and limitations. The fact that quantum learning models, in particular quantum classifiers, can be analyzed through the lens of multihypothesis testing has been put to use before in Refs. [6–12]. Reference [11] remarks that there is a trade-off between expressivity and generalization performance. Reducing the measurement overhead of a quantum learning model has already been attempted by trying to reduce the variance of observables through training [13]. The continuity properties of quantum circuits have also been investigated in Ref. [14].

This work is structured as follows: We start the work with expositions on quantum classifiers (Sec. II) and quantum multihypothesis testing (Sec. III). We go on to introduce the concept of a single-shot quantum learning model in Sec. IV. In the subsequent Sec. V, we use the described concepts and tools to come up with bounds that establish that shallow models cannot have the single-shot property. Finally, in Sec. VI, we show that a quantum learning model that achieves a generic single-shot property is hard to learn. We conclude the paper in Sec. VII.

II. QUANTUM CLASSIFIERS

A classification task can be formalized by considering data $\boldsymbol{x}$ from a data domain $\mathcal{X}$ with labels y from a label domain $\mathcal{Y}$ which we assume to be discrete. Note that the data domain should not be seen as the ambient space, which is usually $\mathbb{R}^d$, but the subset of it which corresponds to actually valid data points. In the framework of statistical learning theory, we assume that there is a joint distribution of data and labels $p(\boldsymbol{x}, y)$ from which we obtain samples to train our classifier. We will also use the marginal distribution of the data alone, $p(\boldsymbol{x}) = \sum_{y\in\mathcal{Y}} p(\boldsymbol{x}, y)$.

A quantum classifier consists of a *quantum feature map* $\boldsymbol{x} \mapsto \rho(\boldsymbol{x})$ that embeds the data into the Hilbert space of a quantum system, and a procedure to extract predictions for the labels from the embedded states (compare Fig. 1). The feature map is usually implemented through a parametrized quantum circuit.

The procedure that assigns labels to embedded data points consists of measurements and arbitrary postprocessing. A particular example that is encountered often in the literature is to measure the expectation value of an observable, e.g., a Pauli-Z operator, and then decide for a label based on the expectation value. On a formal level, we can absorb both the measurements and the postprocessing into a positive operator-valued measurement (POVM) whose elements correspond to the different labels, $\{\Pi_y\}_{y\in\mathcal{Y}}$. Note that both the embedding and the postprocessing can be trainable, in the sense that they depend on further parameters $\boldsymbol{\theta}$, for example, the angles of a parametrized quantum circuit so that $\rho(\boldsymbol{x}) = \rho_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathcal{U}(\boldsymbol{\theta}, \boldsymbol{x})[\rho_0]$.

A quantum classifier is a particular instance of what we refer to as a *probabilistic classifier*, as it—contrary to, for example, a classical neural network—outputs labels nondeterministically. Abstractly speaking, the output of the classifier f is not a label, but a random variable $\hat{f}(\boldsymbol{x})$ with values in $\mathcal{Y}$ and distribution $\mathbb{P}[\hat{f}(\boldsymbol{x}) = y] = \mathrm{Tr}[\Pi_y \rho(\boldsymbol{x})]$. The probabilistic nature is problematic because we need to assign a unique label to every data point. In quantum machine learning, this problem is circumvented by running the classification procedure multiple times and taking a majority vote to predict the most likely label,

$$f(\boldsymbol{x}) := \operatorname*{argmax}_{y'} \mathrm{Tr}[\rho(\boldsymbol{x})\Pi_{y'}]. \tag{1}$$

This is also implicitly happening when the expectation value of an observable is used to determine a label, as multiple runs of a quantum experiment are usually needed to faithfully approximate the expectation value.

III. QUANTUM MULTIHYPOTHESIS TESTING

Quantum classifiers solve a task that is similar in spirit to *quantum multihypothesis testing*. This refers to the task of distinguishing quantum states from a set of known states, $\{\rho_j\}_{j=1}^r$ [15]. It is an important tool used in quantum information theory because many tasks that seem unrelated to hypothesis testing *a priori* can be reduced to it. Bounds on the best possible achievable precision in the multihypothesis testing task then yield corresponding bounds for the initial task. As we will shortly see, quantum classifiers are no exception to this.

In multihypothesis testing, our job is to design a measurement that upon input of the state ρ_j, produces the output j with high probability. We can model the most general quantum measurement by a POVM $\{\Pi_j\}_{j=1}^M$ whose operators

satisfy $\Pi_j \geqslant 0$ and $\sum_{j=1}^{M} \Pi_j = \mathbb{I}$. In the Bayesian version of quantum multihypothesis testing, we are given not only the quantum states $\{\rho_j\}_{j=1}^r$, but also the probabilities with which we are given the state, $\{p_j\}_{j=1}^r$. Our job is to find a measurement that minimizes the Bayesian multihypothesis testing error,

$$P_e^*\left(\{p_j\rho_j\}_{j=1}^r\right) = \min_{\substack{\{\Pi_j\}\\ \text{POVM}}} \left\{1 - \sum_{j=1}^{r} p_j \operatorname{Tr}[\Pi_j \rho_j]\right\}. \tag{2}$$

If we do not have knowledge about the prior probabilities of the different states, we can still ask for a worst-case guarantee. This yields the minimax hypothesis testing error which can be guaranteed for any prior distribution of the states,

$$\overline{P}_e^*\left(\{\rho_j\}_{j=1}^r\right) = \min_{\substack{\{\Pi_j\}\\ \text{POVM}}} \max_{1\leqslant j\leqslant r} \{1 - \operatorname{Tr}[\Pi_j \rho_j]\}. \tag{3}$$

The optimal measurement in both the Bayesian and the minimax setting can be computed with a semidefinite program (SDP) [16].

The usefulness of hypothesis testing stems from the fact that we have good lower bounds on the multihypothesis testing error. In the case of binary hypothesis testing, $r = 2$, we even have a closed-form solution. The Helstrom-Holevo theorem [17] relates the minimal error of binary hypothesis testing to the trace distance of the quantum states,

$$P_e^*[p\rho, (1-p)\sigma] = \tfrac{1}{2} - \tfrac{1}{2}\|p\rho - (1-p)\sigma\|_1. \tag{4}$$

We immediately see that the above expression vanishes for orthogonal states and, assuming without loss of generality that $p \geqslant 1/2$, takes its maximal value $1 - p$ for $\rho = \sigma$, where $1 - p$ amounts to the probability of success for random guessing.

For more than two states, we do not have closed-form solutions. We can, however, still retain the interpretability in terms of pairwise distances given by the Holevo-Helstrom theorem through the following lemma, which is inspired by a similar construction in Ref. [18].

Lemma 1. Multihypothesis testing to binary reduction. For a given quantum multihypothesis testing problem between states $\{p_i\rho_i\}_{i=1}^r$, we have that

$$\begin{aligned} &P_e^*\left(\{p_i\rho_i\}_{i=1}^r\right) \\ &\quad\geqslant \sum_{i=1}^{r} p_i \max_{1\leqslant j\neq i\leqslant r} P_e^*\left(\frac{p_i}{p_i+p_j}\rho_i, \frac{p_j}{p_i+p_j}\rho_j\right) &(5)\\ &\quad\geqslant \min_{1\leqslant i\leqslant r} \max_{1\leqslant j\neq i\leqslant r} P_e^*\left(\frac{p_i}{p_i+p_j}\rho_i, \frac{p_j}{p_i+p_j}\rho_j\right). &(6)\end{aligned}$$

The proof is presented in Appendix A. The corresponding result in the minimax setting is straightforward, as the worst-case error can only improve when considering a hypothesis test among fewer alternatives, and hence

$$\overline{P}_e^*\left(\{\rho_j\}_{j=1}^r\right) \geqslant \max_{i\neq j} \overline{P}_e^*(\rho_i, \rho_j). \tag{7}$$

Combining the above results with the Holevo-Helstrom theorem, we obtain a lower bound on the error for the multihypothesis testing problem in terms of the pairwise distances of the quantum states. We note that these lower bounds are not necessarily the tightest possible. But they allow us to later connect the limitations on single-shot quantum learning models to interpretable quantities based on the trace distance of quantum states.

Reduction to multihypothesis testing for quantum classifiers

Quantum classifiers can be naturally analyzed through the lens of quantum multihypothesis testing. In this section, we reproduce the derivation of this connection which has already been successfully applied to analyze various properties of quantum classifiers [6,8–12]. To do so, we first have a look at the probability that a given quantum classifier outputs the correct label when it is used once. This is simply the expected accuracy of the classifier. In the following calculation, we will split the parameter space according to labels, i.e., we define $\mathcal{X}_{y'} \coloneqq \{\boldsymbol{x} \,|\, y(\boldsymbol{x}) = y'\}$, where $y(\boldsymbol{x})$ is the true label of the data point $\boldsymbol{x}$:

$$\begin{aligned} \mathbb{P}[\text{success}] &= \int_{\mathcal{X}} \mathrm{d}p(\boldsymbol{x}, y)\, \mathbb{P}[\hat{f}(\boldsymbol{x}) = y] &(8)\\ &= \sum_{y'\in\mathcal{Y}} \int_{\mathcal{X}_{y'}} \mathrm{d}p(\boldsymbol{x})\, \operatorname{Tr}[\Pi_{y'}\rho(\boldsymbol{x})] &(9)\\ &= \sum_{y'\in\mathcal{Y}} \operatorname{Tr}\left[\Pi_{y'} \int_{\mathcal{X}_{y'}} \mathrm{d}p(\boldsymbol{x})\, \rho(\boldsymbol{x})\right] &(10)\\ &= \sum_{y'\in\mathcal{Y}} \operatorname{Tr}[\Pi_{y'} p_{y'} \rho_{y'}]. &(11)\end{aligned}$$

In the last step, we defined the average state for every class,

$$p_{y'} \coloneqq \int_{\mathcal{X}_{y'}} \mathrm{d}p(\boldsymbol{x}), \tag{12}$$

$$\rho_{y'} \coloneqq \frac{1}{p_{y'}} \int_{\mathcal{X}_{y'}} \mathrm{d}p(\boldsymbol{x})\, \rho(\boldsymbol{x}). \tag{13}$$

If we compare the above with Eq. (2), we see that a quantum classifier is equivalent to a quantum multihypothesis testing problem for the average states for each class. Therefore, we can conclude that the error of a quantum classifier is at least the optimal achievable error,

$$\mathbb{P}[\text{error}] \geqslant P_e^*(\{p_y\rho_y\}_{y\in\mathcal{Y}}). \tag{14}$$

The fact that the embedded states for different classes have to be well distinguishable to achieve reasonable accuracy—also measured in terms of other loss functions—has already been observed in prior works [6,8,11].

IV. SINGLE-SHOT QUANTUM MACHINE LEARNING

The fact that we have to run a quantum classifier multiple times, explained in Sec. II, leads to the so-called measurement problem, meaning that there are additional overheads compared to classical approaches which stem from the probabilistic nature of quantum mechanics. The purpose of this work is to formalize when we can circumvent this issue by enforcing that the classifier is so sure about the assigned label that running it only once is sufficient, which is when we call the classifier "single shot."

A. Single-shot classifiers

With the notation we introduced, we can define when a classifier is single shot: exactly when the probability that a random label $\hat{f}(\boldsymbol{x})$ is identical to the majority label $f(\boldsymbol{x})$ assigned by the classifier is very high. Formally, we define the following:

Definition 1. Single-shot probabilistic classifier (Bayesian). For data points $\boldsymbol{x}$ distributed according to a distribution $p(\boldsymbol{x})$ supported on a data domain $\mathcal{X}$, we say a probabilistic classifier f is δ-single shot if

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\{\mathbb{P}[\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x})]\} \geqslant 1-\delta, \tag{15}$$

where the expectation value is taken over the random distribution of data points.

Note that $\mathbb{P}[\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x})] = \max_{y\in\mathcal{Y}} \mathrm{Tr}[\Pi_y \rho(\boldsymbol{x})]$, which means we can rewrite this as

$$\int \mathrm{d}p(\boldsymbol{x}) \max_{y\in\mathcal{Y}} \mathrm{Tr}[\Pi_y \rho(\boldsymbol{x})] \geqslant 1-\delta. \tag{16}$$

It is worthwhile to spend some time to clarify multiple important points: First, being *single-shot* is a property of the classifier that is independent of the labels being actually correct. The single-shot property only captures how *deterministic* the outcomes of a classifier are. Consider, for example, a classifier that always outputs the same label; it will be perfectly single shot, i.e., have $\delta = 0$, but will have bad classification accuracy. Second, the single-shot property is necessarily dependent on the learning problem through the data distribution $p(\boldsymbol{x})$: a classifier that has this property for one problem will usually not have it for a different learning problem. Third, how strict the above single-shot definition actually is crucially depends on the allowed failure probability δ, as *every* classifier will have this property if the failure probability δ can be chosen arbitrarily close to 1. We are, naturally, interested in the case where δ is rather small. Fourth, we wish to emphasize that being single shot is a property that we wish to have for a *trained* classifier. It is not a property that will be present upon random initialization and has to be enforced through an appropriate training procedure in one way or another.

We refer to Sec. IV A as the *Bayesian* single-shot definition because we average over the distribution of the inputs $\boldsymbol{x} \sim p(\boldsymbol{x})$. We can obtain a more stringent definition that removes this dependence by minimizing over all possible distributions $p(\boldsymbol{x})$ supported on the data domain $\mathcal{X}$. Here, again, we emphasize that the data domain should be understood not as the whole ambient space of the data (usually $\mathbb{R}^d$), but as the subset corresponding to valid inputs. In this case, it is intuitively clear that the "worst" distribution is the one that always gives our classifier the data points for which $\mathbb{P}[\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x})]$ is minimal. This leads us to the following *agnostic* definition of being single shot:

Definition 2. Single-shot probabilistic classifier (agnostic). For data points $\boldsymbol{x}$ from a data domain $\mathcal{X}$, we say a probabilistic classifier f is agnostically $\bar{\delta}$-single shot if

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x})] \geqslant 1-\bar{\delta}, \tag{17}$$

where the probability is taken over the inherent randomness of the classifier.

Analogously to Eq. (16), we can rewrite this as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{y} \mathrm{Tr}[\Pi_y \rho(\boldsymbol{x})] \geqslant 1-\bar{\delta}. \tag{18}$$

The agnostic definition immediately implies the Bayesian one with the same $\delta = \bar{\delta}$, but it is much more stringent. This is especially clear if we assume the classifier to be continuous in the data, in which case the agnostic single-shot classification is not possible with a meaningfully small value of $\bar{\delta}$ if there are two sets of points with distinct labels assigned by the classifier that touch in parameter space, as continuity would imply that at the point where the label flips, both labels are equally likely and thus single-shot classification is not possible.

It is, however, quite intuitive that even a continuously parametrized classifier can be agnostically single shot on subsets of the data domain, and if the points we see are likely to come from that domain, we still have a pretty good Bayesian single-shot performance. Exactly this thought is formalized by the following lemma, which is just a simple application of the union bound.

Lemma 2. Restriction lemma. Let $\hat{f}(\boldsymbol{x})$ be a probabilistic classifier that is agnostically $\bar{\delta}$-single shot relative to a subset of the data domain $\bar{\mathcal{X}} \subseteq \mathcal{X}$. Then, for any data distribution, the classifier is Bayesian δ-single shot with

$$\delta \leqslant \bar{\delta} + \mathbb{P}[\boldsymbol{x} \notin \bar{\mathcal{X}} \,|\, \boldsymbol{x} \sim p(\boldsymbol{x})]. \tag{19}$$

B. Geometric picture

It is instructive to visualize the concept of single-shot quantum classification. To do so, we can define some specific sets of inputs. First, we define the set of all inputs that the classifier assigns the same majority label,

$$\tilde{\mathcal{X}}_y \coloneqq \{\boldsymbol{x} \in \mathcal{X} \,|\, f(\boldsymbol{x}) = y\}. \tag{20}$$

Next, for a given failure probability δ, we define the subset of the above set on which the classifier is agnostically δ-single shot according to Sec. IV A,

$$\tilde{\mathcal{X}}_y^{\delta} \coloneqq \{\boldsymbol{x} \in \mathcal{X} \,|\, f(\boldsymbol{x}) = y,$$
$$\mathbb{P}[\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x}) \,|\, \boldsymbol{x}] \geqslant 1-\delta\}. \tag{21}$$

It is quite natural that $\delta \leqslant \delta'$ implies that $\tilde{\mathcal{X}}^{\delta} \subseteq \tilde{\mathcal{X}}^{\delta'}$. Plotting the sets $\tilde{\mathcal{X}}_y$ amounts to plotting the decision boundaries of the classifier, whereas plotting the sets $\tilde{\mathcal{X}}_y^{\delta}$ shows the regions in which the classifier is additionally "sure" about its decisions.

Naturally, we also care about the true labels of the data points and, as such, we will use the following notation to denote all data points with a given label:

$$\mathcal{X}_y \coloneqq \{\boldsymbol{x} \in \mathcal{X} \,|\, y(\boldsymbol{x}) = y\}. \tag{22}$$

Figure 2 shows an exemplary plot of some sets for a trivalent classification problem for points in $\mathbb{R}^2$ that clarifies the nature of the sets $\tilde{\mathcal{X}}_y^{\delta}$, $\tilde{\mathcal{X}}_y$, and $\mathcal{X}_y$.

C. Relation to sample complexity

We can relate our notion of a probabilistic classifier being single shot back to the notion of sample complexity that we usually encounter when dealing with the measurement problem in quantum machine learning models (see, also, Ref. [12]). This can be achieved by considering that the process of using m repetitions of the same "weak" classifier

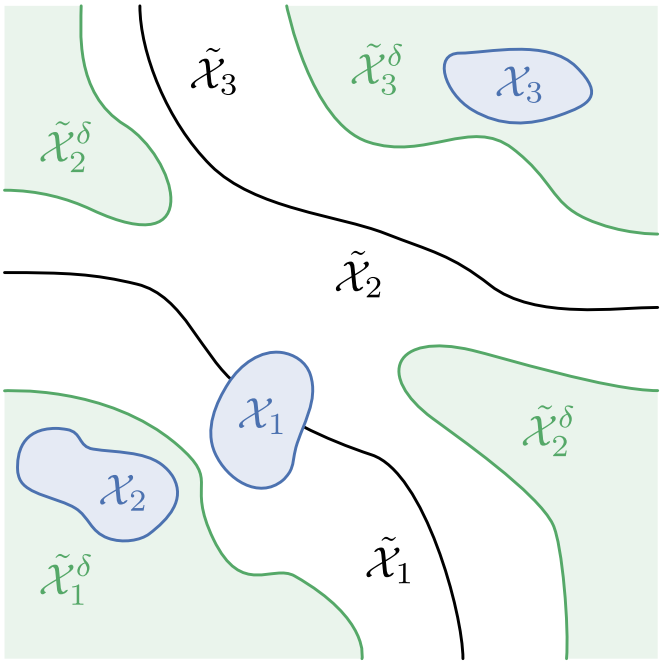


FIG. 2. To gain intuition about the notion of single-shot classification, we look at a problem where points in $\mathcal{X} = \mathbb{R}^2$ are classified into three classes. The sets $\tilde{\mathcal{X}}_i$ correspond to the different labels assigned by the classifier (boundaries in black). A subset of each of these sets, $\tilde{\mathcal{X}}_i^\delta$, corresponds to the inputs for which these labels are assigned in a δ-single-shot way (green). The supports of the data points with true labels are given by $\mathcal{X}_i$ (blue). We observe three different situations: Data from class 1 are not always correctly classified and labels are not assigned in a δ-single-shot way. Data from class 2 are incorrectly classified, but labels are assigned in a single-shot way. Data from class 3 are correctly classified in a single-shot way.

followed by a majority vote can be viewed as a single model that should then have the single-shot property, which amounts to a low failure probability. Mathematically speaking, we have the following lemma about boosting the success probability of a weak classifier, i.e., one that has the single-shot property only with a high failure probability.

Lemma 3. Let f be a probabilistic classifier on a data domain $\mathcal{X}$. Then, m independent repetitions of f together with a majority vote form a new classifier that is agnostically $\bar{\delta}$-single shot, with

$$\bar{\delta} \leqslant |\mathcal{Y}| \exp\big(-\tfrac{1}{2}m\Delta^2\big), \tag{23}$$

where Δ is the smallest gap between the two largest label probabilities $\mathrm{Tr}[\Pi_y \rho(\boldsymbol{x})]$ for all $\boldsymbol{x}$.

Proof. We define the random variables $\{\hat{r}_y(\boldsymbol{x})\}_{y\in\mathcal{Y}}$ as the ratio of the m classifiers that output the label y, i.e.,

$$\hat{r}_y(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \chi[\hat{f}_i(\boldsymbol{x}) = y], \tag{24}$$

where χ is the indicator function, understanding that $\mathbb{E}[\hat{r}_y(\boldsymbol{x})] = \mathrm{Tr}[\Pi_y\rho(\boldsymbol{x})]$. From Hoeffding's inequality, we have the two one-sided estimates,

$$\mathbb{P}\{\hat{r}_y(\boldsymbol{x}) - \mathrm{Tr}[\Pi_y\rho(\boldsymbol{x})] \geqslant \epsilon\} \leqslant \exp(-2m\epsilon^2), \tag{25}$$

$$\mathbb{P}\{\mathrm{Tr}[\Pi_y\rho(\boldsymbol{x})] - \hat{r}_y(\boldsymbol{x}) \geqslant \epsilon\} \leqslant \exp(-2m\epsilon^2). \tag{26}$$

We could have used a tighter estimate deriving from the Chernoff-Hoeffding theorem on the sum of i.i.d. Bernoulli random variables, but the improvements are not necessary for our argument. We can guarantee a successful outcome of the majority vote if we choose ϵ to be half the gap between the largest and the second-largest value of $\mathrm{Tr}[\Pi_y\rho(\boldsymbol{x})]$. Formally, we can define the gap as

$$\Delta(\boldsymbol{x}) := \min_{\mathcal{Y}\ni y'\neq y} \max_{y\in\mathcal{Y}} \mathrm{Tr}[(\Pi_y - \Pi_{y'})\rho(\boldsymbol{x})], \tag{27}$$

and hence choose $\epsilon = \epsilon(\boldsymbol{x}) = \Delta(\boldsymbol{x})/2$. We have to use Hoeffding's inequality both to assure that $\hat{r}_y(\boldsymbol{x})$ associated to the largest value is not too low, as well that all of the other options are not too large. Combining them through the union bound leads us to the bound

$$\bar{\delta}(\boldsymbol{x}) \leqslant |\mathcal{Y}| \exp\big[-\tfrac{1}{2}m\Delta(\boldsymbol{x})^2\big]. \tag{28}$$

Taking the worst case over $\boldsymbol{x}$ and defining

$$\Delta = \min_{\boldsymbol{x}\in\mathcal{X}} \Delta(\boldsymbol{x}) \tag{29}$$

yields the statement of the lemma. ■

We wish to note that the dependence on the gap instead of the failure probability of the classifier is necessary. We can connect them and obtain a lower bound on the gap $\Delta \geqslant 2\epsilon$ if the classifier used in the majority voting procedure has failure probability $\bar{\delta} = \frac{1}{2} + \epsilon$.

A Bayesian version of the argument of Sec. IV C proceeds similarly, but the resulting right-hand side is

$$\delta \leqslant |\mathcal{Y}| \int \mathrm{d}p(\boldsymbol{x}) \exp\left[-\frac{1}{2}m\Delta(\boldsymbol{x})^2\right], \tag{30}$$

which is not as easy to interpret and to rearrange to obtain a lower bound on m for desired $\delta(\boldsymbol{x})$, which is why we think it better to use the agnostic version of Sec. IV C together with the restriction argument of Sec. IV A to arrive at a Bayesian version of the result. There, the restriction of the parameter space $\tilde{\mathcal{X}}$ should be chosen such that a lower bound on the gap Δ can be guaranteed for all $\boldsymbol{x} \in \tilde{\mathcal{X}}$.

V. ULTIMATE LIMITS OF SINGLE-SHOT QUANTUM MACHINE LEARNING

Having defined the single-shot property, it is natural to ask if it can actually be achieved for a given embedding $\rho(\boldsymbol{x})$ and, vice versa, what resources are necessary to realize an embedding that allows for single-shot classification. This section is devoted to exploring the ultimate limits of single-shot quantum classifiers by reducing the single-shot classification task to a multihypothesis testing task. After establishing said reduction, we open the black box of the embedding and establish lower bounds on the gate complexity that a quantum circuit needs to establish the single-shot property.

A. Reduction to hypothesis testing

In Sec. III, we have seen that the error probability of quantum classifiers can be lower bounded through a reduction to multihypothesis testing. We can employ a similar argument to bound the best value of the single-shot error probability δ. The subtle difference lies in the fact that we are no longer

concerned with classifying the data well with respect to the true labels $y(\boldsymbol{x})$ as they do not figure in the definition of the single-shot property. Instead, we want the classifier to work with respect to the labels *assigned by the classifier itself*, i.e., $f(\boldsymbol{x})$ as in Eq. (1). After this change of perspective, the argument can, however, proceed analogously.

To quantify the limits of the Bayesian single-shot property, we split the parameter space according to the majority labels assigned by the classifier and recall the definition $\tilde{\mathcal{X}}_{y'} \coloneqq \{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) = y'\}$ from Eq. (20). We obtain the following theorem:

Theorem 1. Bayesian error probability lower bound. Let f be a probabilistic classifier on a data domain $\mathcal{X}$ taking discrete values in $\mathcal{Y}$. For data points $\boldsymbol{x}$ distributed according to a distribution $p(\boldsymbol{x})$, we define the average states for the classes assigned by f as

$$\tilde{p}(y) \coloneqq \int_{\tilde{\mathcal{X}}_y} \mathrm{d}p(\boldsymbol{x}), \quad \tilde{\rho}_y \coloneqq \frac{1}{\tilde{p}(y)} \int_{\tilde{\mathcal{X}}_y} \mathrm{d}p(\boldsymbol{x})\, \rho(\boldsymbol{x}). \tag{31}$$

If the classifier f has the Bayesian single-shot property, the error probability δ has to fulfill

$$\delta \geqslant P_e^*(\{\tilde{p}_y \tilde{\rho}_y\}_{y\in\mathcal{Y}}). \tag{32}$$

Proof. We start from Sec. IV A and obtain

$$\begin{aligned} 1-\delta &\leqslant \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p(\boldsymbol{x})} \left\{\mathbb{P}[\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x})]\right\} && (33)\\ &= \int_{\mathcal{X}} \mathrm{d}p(\boldsymbol{x})\, \mathbb{P}[\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x})] && (34)\\ &= \int_{\mathcal{X}} \mathrm{d}p(\boldsymbol{x}) \max_{y'\in\mathcal{Y}} \mathrm{Tr}[\Pi_{y'} \rho(\boldsymbol{x})] && (35)\\ &= \sum_{y'\in\mathcal{Y}} \int_{\tilde{\mathcal{X}}_{y'}} \mathrm{d}p(\boldsymbol{x})\ \mathrm{Tr}[\Pi_{y'} \rho(\boldsymbol{x})] && (36)\\ &= \sum_{y'\in\mathcal{Y}} \mathrm{Tr}\left[\Pi_{y'} \int_{\tilde{\mathcal{X}}_{y'}} \mathrm{d}p(\boldsymbol{x})\, \rho(\boldsymbol{x})\right] && (37)\\ &= \sum_{y'\in\mathcal{Y}} \mathrm{Tr}[\Pi_{y'} \tilde{p}_{y'} \tilde{\rho}_{y'}] && (38)\\ &\leqslant P_s^*(\{\tilde{p}_y \tilde{\rho}_y\}_{y\in\mathcal{Y}}), && (39) \end{aligned}$$

with $P_s^*(\{\tilde{p}_y \tilde{\rho}_y\}_{y\in\mathcal{Y}})$ being the optimal Bayesian multihypothesis testing success probability. Rearranging and using $P_s^* = 1 - P_e^*$ yields the statement of the theorem. ■

In the above theorem, the data space $\mathcal{X}$ and the marginal distribution of the data $p(\boldsymbol{x})$ are the same as in Sec. III, only the average state $\tilde{\rho}_{y'}$ is now taken among the states which are assigned the same label by the classifier and is not related to the actual correct labels of the dataset. The lower bound on δ of the above theorem is achieved by the optimal measurement for the Bayesian multihypothesis testing problem between the states $\tilde{\rho}_y$ with prior probabilities $\tilde{p}_y$.

In the agnostic setting, the single-shot property has to hold irrespective of the underlying distribution of the data. As such, we can bound it by constructing the "worst" possible distribution over the data, i.e., the one that minimizes the Bayesian success probability. It is clear that this distribution consists of putting all the probability on the single data point $\boldsymbol{x}$ for which the classifier struggles the most.[2] This strategy yields the following bound:

Theorem 2. Agnostic error probability lower bound. Let f be a probabilistic classifier on a data domain $\mathcal{X}$ taking discrete values in $\mathcal{Y}$. If the classifier f has the agnostic single-shot property, the error probability $\bar{\delta}$ has to fulfill

$$\bar{\delta} \geqslant \max_{\{\boldsymbol{x}_y \in \tilde{\mathcal{X}}_y\}_{y\in\mathcal{Y}}} \overline{P}_e^*[\{\rho(\boldsymbol{x}_y)\}_{y\in\mathcal{Y}}]. \tag{40}$$

Proof. With the reasoning described above, we see that for the adversarially chosen distribution, we have

$$\begin{aligned} 1-\bar{\delta} &\leqslant \min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x})] && (41)\\ &= \min_{\boldsymbol{x}\in\mathcal{X}} \max_{y'\in\mathcal{Y}} \mathrm{Tr}[\Pi_{y'} \rho(\boldsymbol{x})]. && (42) \end{aligned}$$

We can write

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{y'\in\mathcal{Y}} = \min_{y\in\mathcal{Y}} \min_{\boldsymbol{x}_y \in \tilde{\mathcal{X}}_y} \tag{43}$$

because we know, by definition of $\tilde{\mathcal{X}}_y$, which projector achieves the maximum over y'. This gives

$$1-\bar{\delta} \leqslant \min_{y\in\mathcal{Y}} \min_{\boldsymbol{x}_y \in \tilde{\mathcal{X}}_y} \mathrm{Tr}[\Pi_y \rho(\boldsymbol{x}_y)]. \tag{44}$$

To exchange the two minima, we optimize over sets of data points $\{\boldsymbol{x}_y \in \tilde{\mathcal{X}}_y\}_{y\in\mathcal{Y}}$, i.e., one for each possible label. This gives

$$\begin{aligned} 1-\bar{\delta} &\leqslant \min_{\{\boldsymbol{x}_y \in \tilde{\mathcal{X}}_y\}_{y\in\mathcal{Y}}} \min_{y\in\mathcal{Y}} \mathrm{Tr}[\Pi_y \rho(\boldsymbol{x}_y)] && (45)\\ &\leqslant \min_{\{\boldsymbol{x}_y \in \tilde{\mathcal{X}}_y\}_{y\in\mathcal{Y}}} \overline{P}_s^*[\{\rho(\boldsymbol{x}_y)\}_{y\in\mathcal{Y}}], && (46) \end{aligned}$$

where we recognized the definition of the minimax success probability for multihypothesis testing from Eq. (3) in the last step. The theorem statement follows by using $\overline{P}_s^* = 1 - \overline{P}_e^*$ and rearranging. ■

We note that we could, in theory, further improve the lower bound on the agnostic single-shot error probability in the above bound by including multiple states per label, which would result in a reduction to a composite multihypothesis testing problem. These problems are less studied and fewer analytical tools are available for them. For numerical investigations, this is, however, an adequate way to achieve tighter bounds in conjunction with semidefinite programming.

We now use the tools presented in Sec. III to connect the ultimate limits of single-shot quantum classification to more intuitive quantities. We apply the reduction to the binary case of Sec. III combined with the Holevo-Helstrom theorem of Eq. (4) to the Bayesian error probability lower bound of Sec. V A. This gives the following lower bound on the error probability in terms of the trace distance of the average states for the different classes:

$$\begin{aligned} \delta &\geqslant \sum_{y\in\mathcal{Y}} \tilde{p}_y \max_{y\neq y'\in\mathcal{Y}} \frac{1}{2} e(y, y') && (47)\\ &\geqslant \min_{y\in\mathcal{Y}} \max_{y\neq y'\in\mathcal{Y}} \frac{1}{2} e(y, y'), && (48) \end{aligned}$$

[2]It is important to emphasize that in this scenario, the adversarial distribution over the data is chosen *after* a POVM performing the classification was fixed; if it were the other way around, a different strategy of the adversary would be optimal.

where we defined

$$e(y, y') \coloneqq 1 - \left\| \frac{\tilde{p}_y}{\tilde{p}_y + \tilde{p}_{y'}} \tilde{\rho}_y - \frac{\tilde{p}_{y'}}{\tilde{p}_y + \tilde{p}_{y'}} \tilde{\rho}_{y'} \right\|_1. \tag{49}$$

This expression shows that the closer the average states of the different classes assigned by the classifier are, the worse the achievable error probability becomes and the less "single shot" our classifier must be. Therefore, in order to have a small error probability, the average states of different assigned labels have to be far apart.

In the agnostic case, we can directly reduce to a binary discrimination problem as trying to distinguish less states in a minimax fashion will certainly be easier, i.e., we have that

$$\overline{P}_e^*(\{\rho_j\}_{j=1}^r) \geqslant \max_{1 \leqslant i \neq j \leqslant r} \overline{P}_e^*(\rho_i, \rho_j). \tag{50}$$

While there is no closed-form expression for the minimax binary hypothesis testing error probability, it is reasonable to assume that it is not too far from the problem of distinguishing both states with prior probability $1/2$ for each state. Therefore,

$$\overline{P}_e^*(\rho_i, \rho_j) \geqslant P_e^*\left(\frac{1}{2}\rho_i, \frac{1}{2}\rho_j\right). \tag{51}$$

Combining this reasoning again with the Holevo-Helstrom bound and Sec. V A, we obtain the following lower bound for the agnostic single-shot error probability:

$$\begin{aligned} \bar{\delta} &\geqslant \max_{\substack{\boldsymbol{x},\boldsymbol{x}' \\ f(\boldsymbol{x}) \neq f(\boldsymbol{x}')}} \overline{P}_e^*[\rho(\boldsymbol{x}), \rho(\boldsymbol{x}')] && (52) \\ &= \max_{\substack{\boldsymbol{x},\boldsymbol{x}' \\ f(\boldsymbol{x}) \neq f(\boldsymbol{x}')}} P_e^*\left[\frac{1}{2}\rho(\boldsymbol{x}), \frac{1}{2}\rho(\boldsymbol{x}')\right] && (53) \\ &\geqslant \max_{\substack{\boldsymbol{x},\boldsymbol{x}' \\ f(\boldsymbol{x}) \neq f(\boldsymbol{x}')}} \frac{1}{2}\left[1 - \frac{1}{2}\|\rho(\boldsymbol{x}) - \rho(\boldsymbol{x}')\|_1\right] && (54) \\ &= \frac{1}{2} - \frac{1}{4} \min_{\substack{\boldsymbol{x},\boldsymbol{x}' \\ f(\boldsymbol{x}) \neq f(\boldsymbol{x}')}} \|\rho(\boldsymbol{x}) - \rho(\boldsymbol{x}')\|_1. && (55) \end{aligned}$$

From this, we can conclude that agnostic single shotness imposes a strong requirement on the distances of the embedded states for different classes predicted by the classifier, as we can reverse the inequality above to obtain

$$\tfrac{1}{2}\|\rho(\boldsymbol{x}) - \rho(\boldsymbol{x}')\|_1 \geqslant 1 - 2\bar{\delta} \tag{56}$$

for all $\boldsymbol{x}$ and $\boldsymbol{x}'$ belonging to different classes, as predicted by the classifier.

It is quite obvious from this bound that any embedding $\rho(\boldsymbol{x})$ that continuously varies over $\boldsymbol{x}$ cannot be agnostically single shot on the whole value range of $\boldsymbol{x}$. This emphasizes again that the agnostic definition of single shotness necessitates a restriction of the domain of $\boldsymbol{x}$, $\mathcal{X}$, as we also used in the lifting from agnostic to Bayesian in Sec. IV A.

Having established a connection of the single-shot error probability in both the agnostic and the Bayesian setting to the trace distance of the embedded states, we open up the black box of the embedding $\rho(\boldsymbol{x})$ and consider that it is realized by a variational quantum circuit. In that case, imposing a certain single-shot error probability necessitates a certain depth and/or width of the circuit, as we show below.

B. Noiseless quantum circuits

We will first consider the case where the data are embedded by the means of a noiseless quantum circuit. We can express any such circuit as a unitary composed from parametrized gates, $R_G(\alpha) = e^{-iG\alpha}$, where G is the *generator* of the gate, and fixed unitaries. In our case, we assume that the circuit is partitioned into L layers, where each layer corresponds to a subcircuit that encodes every parameter x_i once through a parametrized gate and can consist of arbitrary variationally parametrized gates and fixed unitaries otherwise. In Appendix B, we prove that parametrized quantum circuits have a continuity bound in terms of the arguments of the parametrized gates, which, as a corollary, yields the following result:

Lemma 4. Continuity of parametrized quantum states. Let $\rho_{\boldsymbol{\theta}}(\boldsymbol{x})$ be obtained from a parametrized quantum circuit of L layers, where, in each layer, the parameters $\{x_i\}_{i=1}^d$ are encoded through separate parametrized gates. Then, we have that

$$\|\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) - \rho_{\boldsymbol{\theta}}(\boldsymbol{x}')\|_1 \leqslant L\Delta\|\boldsymbol{x} - \boldsymbol{x}'\|_1, \tag{57}$$

where $\Delta \coloneqq \max_{G \in \mathcal{G}}[\lambda_{\max}(G) - \lambda_{\min}(G)]$ is the largest spectral spread of the generators used to encode the parameters $\{x_i\}_{i=1}^d$.

Section V B shows us that a parametrized circuit embeds data points that are close into quantum states that are close as well. Intuitively, we can reason that the quantum classifier is going to have a hard time with two data points that are close but belong to different assigned labels. We formalize this in the following theorem:

Theorem 3. Lower bound on the single-shot error probability (Bayesian). Let $\hat{f}(\boldsymbol{x})$ be a quantum classifier with the task of classifying r groups of classical data for which we know its probability distribution. The single-shotness error probability is bounded by the worst average distance between two data classes,

$$\delta \geqslant \min_{1 \leqslant i \leqslant r} \max_{1 \leqslant j \neq i \leqslant r} \frac{\min(p_i, p_j)}{p_i + p_j}\left(1 - \frac{L\Delta d_{\text{avg}}^{ij}}{2}\right), \tag{58}$$

where

$$d_{\text{avg}}^{ij} \coloneqq (p_i + p_j) \int_{\mathcal{X}_i} \frac{dp(\boldsymbol{x}_i)}{p_i} \int_{\mathcal{X}_j} \frac{dp(\boldsymbol{x}_j)}{p_j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1 \tag{59}$$

is the expected distance of two data points sampled according to the data distribution, conditioned on them being from the classes i and j.

The proof is presented in Appendix C. The average distance d_{avg}^{ij} gives us a sense of how far apart the points of the two classes i and j are in the input data space.

In this theorem, we see that in order to have a low single-shot error probability, either we have that the different classes are located far apart in the input space or we have a sufficiently deep circuit. Therefore, fixing a dataset, this expression shows us that the depth of the circuit can be limiting to the classifier performance. To get some intuition, we assume that all r classes are equally likely, i.e., that $p = 1/r$, in which case we

can rearrange the expression in Eq. (58) to obtain

$$L \geqslant 2\frac{1-2\delta}{\Delta d_{\text{avg}}}, \tag{60}$$

where

$$d_{\text{avg}} = \max_{1\leqslant i\leqslant r} \min_{1\leqslant j\neq i\leqslant r} d^{ij}_{\text{avg}} \tag{61}$$

and $\delta \geqslant P_e^*$. As we can see, the length of the circuit is inversely proportional to the average distance between the classes. The closer they are, the longer the circuit will be needed in order to separate the states enough so they can be classified.

It is worth emphasizing at this point that even if the condition of the above theorem would permit a certain single-shot error probability, it is not necessarily the case (and usually not expected) that this single-shot error probability is actually achieved by the circuit.

C. Noisy circuits

Realistic circuits are far from perfect. In the NISQ era, they are plagued by the eponymous noise that limits their performance. To study the impact of noise on the single shotness of a quantum classifier, we study an error model where local depolarizing noise of survival probability p, $\mathcal{D}_p[\rho] = p\rho + (1-p)\omega$, with ω being the maximally mixed state, is applied after every step of computation of the classifier to every qubit. To keep track of this mathematically, we will use the notation ρ_t to denote the quantum state produced by the noisy embedding circuit after t steps of computation. In our model of the embedding circuit, we have so far been counting the number of layers, L, where each layer consists of a data embedding that we now assume to take one step of computation and fixed unitaries and variational gates that we assume to take ℓ steps of computation, yielding a total of $T = L(1+\ell)$ steps of computation in our parametrized circuit.

A quantum circuit affected by local depolarizing noise will eventually produce a quantum state that is very close to the maximally mixed state ω [19]. This result is established through a decay of the relative entropy between the circuit state after t steps of computation and yields the following estimate [19]:

$$D(\rho_t \,\|\, \omega) \leqslant p^{2t} D(\rho_0 \,\|\, \omega) \leqslant p^{2t} n, \tag{62}$$

where n is the number of qubits. Combining this with Pinsker's inequality then gives us a bound on the trace distance to the maximally mixed state,

$$\|\rho_t - \omega\|_1 \leqslant p^t \sqrt{2n}, \tag{63}$$

which decreases exponentially in t. By a simple application of the triangle inequality, we obtain an estimate of the distance between two noisy embedded states,

$$\|\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) - \rho_{\boldsymbol{\theta}}(\boldsymbol{x}')\|_1 \leqslant \|\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) - \omega\|_1 + \|\rho_{\boldsymbol{\theta}}(\boldsymbol{x}') - \omega\|_1 \tag{64}$$

$$\leqslant p^{L(1+\ell)}\sqrt{8n}. \tag{65}$$

This estimate, however, does not take into account the continuity of the circuits, so for close parameters, it can actually be overly optimistic. As noisy channels cannot increase the trace distance between states, we immediately see that the result of Sec. V B remains valid if noisy circuits are considered, albeit it will also be overly optimistic because it does not consider the noise in the system. A simple way of combining both estimates is to simply take the minimum of both, which gives

$$\|\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) - \rho_{\boldsymbol{\theta}}(\boldsymbol{x}')\|_1 \leqslant \min\{p^{L(1+\ell)}\sqrt{8n}, L\Delta\|\boldsymbol{x} - \boldsymbol{x}'\|_1\}. \tag{66}$$

We can improve on this estimate for shallow circuits. We perform a more involved analysis in Appendix D, which implies the following lemma:

Lemma 5. Consider a noisy embedding circuit with L layers such that

$$L \geqslant L_0 = \left\lceil 1 + \frac{1}{2(\ell+1)} \frac{\ln 2n}{\ln \frac{1}{p}} \right\rceil. \tag{67}$$

Then,

$$\|\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) - \rho_{\boldsymbol{\theta}}(\boldsymbol{x}')\|_1 \leqslant \Delta\|\boldsymbol{x} - \boldsymbol{x}'\|_1 \left(L_0 + \sqrt{2n} \frac{p^{L_0(\ell+1)} - p^{L(\ell+1)}}{1 - p^{\ell+1}} \right). \tag{68}$$

In the case $L \leqslant L_0$, the bound does not improve on the noiseless one.

Contrary to the noise-only estimate, the above bound does not decay to zero as $L \to \infty$, which means it can only improve on the estimate of Eq. (66) in a limited parameter region. It does, however, always improve on the naive continuity estimate given by Eq. (57). Figure 3 shows a comparison of the three different bounds.

Now, we can connect back to the setting of single-shot classification in a similar way as to Sec. V B. The derivation can proceed analogously; only the constant L is replaced with

$$L \to L_0 + \sqrt{2n} \frac{p^{L_0(\ell+1)} - p^{L(\ell+1)}}{1 - p^{\ell+1}}. \tag{69}$$

Combining all three bounds that we have derived for the noisy case gives us an upper bound on the achievable distance between the embedded average states per majority class predicted by the quantum classifier and, as such, allows us to

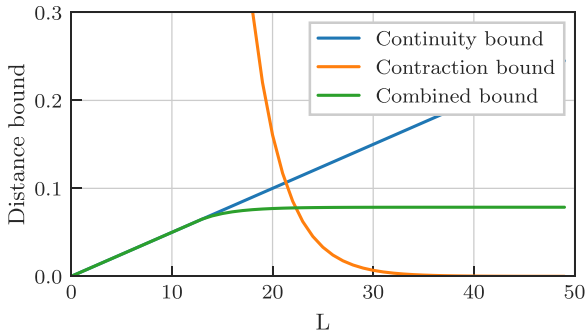


FIG. 3. Comparison of the three different bounds for noisy circuits. We can distinguish the linearly growing continuity bound of Eq. (57) (blue), the exponentially contracting bound of Eq. (65) (orange), and the geometrically growing combined bound of Eq. (68) (green). The constants used for the plot are $n = 1000$, $\|\boldsymbol{x} - \boldsymbol{x}'\|_1 = 0.005$, $\Delta = 1$, $\ell = 2$, and $p = 0.9$.

analyze the possibility of single-shot classification by noisy quantum classifiers.

VI. GENERICALLY ACCURATE SINGLE-SHOT MODELS ARE HARD TO LEARN

Having established a theoretical foundation of single-shot quantum machine learning, we have learned that single shotness is a property of a trained model that is not necessarily a given. It depends on the specific way a quantum learning model embeds classical data into quantum states, and only materializes if quantum states of the same predicted label cluster together—or, in other words, they are not too close to states of a different predicted label.

From this intuition, we see that it should, in theory, be possible to construct an embedding that enables single-shot classification for all possible labelings of a dataset. In machine learning language, this means that the model is expressive enough to *shatter* the data. This is achieved by embedding all sufficiently separated inputs into mutually orthogonal directions of Hilbert space. One expect that for example, we can construct such an embedding by using a large number of encoding gates interleaved with deep layers of random unitary gates. However, it is evident that such an embedding would result in poor generalization performance for any POVM inferred from the training data. Previously unseen data points would be embedded into the orthogonal complement of the embedded training data in Hilbert space. There, we cannot infer a good classification and the best strategy would be random guessing.

This "thought experiment" highlights an inherent problem of (quantum) learning models, namely, that there is a fine trade-off between expressivity, trainability, and generalization performance, which is the subject of the ongoing research in quantum machine learning [11,20–22]. In this section, we will show that this trade-off also influences the single-shot property of a learning model.

To this end, we consider models that can be considered as *generically accurate single-shot models*. These are models built from an embedding $\rho(\boldsymbol{x})$ such that for any possible labeling of the data, there exists a POVM $\{\Pi_y\}_{y\in\mathcal{Y}}$ that accurately classifies in a single-shot way. Note that contrary to the preceding sections, where we treated single shotness in a way that is completely detached from accuracy, here we additionally take it into account. For this notion to be sensible, we need to assume that all the data points fed into the model cannot be arbitrarily close to each other. Formally, we assume that all possible data points encountered by the model have a lower bound on their distance in some norm: $\|\boldsymbol{x}-\boldsymbol{x}'\| > \xi$. This does not restrict the generality of our argument, as very close data points would represent the same input. An example of this would be images: images that can be considered reasonably different would also have to have reasonably different pixel values. Intuitively, we expect that a *generically accurate single-shot model* should map all points that are separated by at least ξ to (approximately) orthogonal directions in Hilbert space.

To study this specific class of models, we rely on the results of Ref. [6]. There, the authors study the essentially similar problem of learning a POVM effect Π from examples of the form

$$\mathcal{D}_m = \{(\rho_i, \mathrm{Tr}[\rho_i\Pi])\}_{i=1}^m, \tag{70}$$

where the authors emphasize that the sample complexity does not change significantly when single measurement outcomes are given instead of expectation values. To connect this to our setting, we consider the setting of binary classification. A classical data point $\boldsymbol{x}$ can be seen as a classical description of $\rho(\boldsymbol{x})$, which means we already know the first half of the data tuples. The second half is then obtained by realizing that $\mathrm{Tr}[\rho(\boldsymbol{x})\Pi]$ can be seen as an instance of a label that should be either 0 or 1. As we assumed that there actually exists a POVM that realizes the desired classification, we can just assume that the given label—either 0 or 1—is the expected value of the targeted POVM effect Π that realizes the correct classification. The fact that the classifier is single shot then enforces that the expected value of the effect Π is actually identical to the target label. Hence, the training data of the classification task,

$$\mathcal{D}_m = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m, \tag{71}$$

can be seen as equivalent to the setup of Ref. [6] in the case of a generically accurate single-shot model.

Now, given this equivalence, Ref. [6] posits that at least

$$m \geqslant O\left(\frac{D}{\epsilon^2}\right) \tag{72}$$

samples are necessary to learn a POVM effect on a D-dimensional Hilbert space to precision ϵ in operator norm. The notion and precision of the approximation are not crucial to the argument we present here, which rests on the dimension dependence. An ϵ-approximation to the POVM effects in operator norm, however, immediately guarantees that the precision of the implemented binary classifier is close to the target within an additive error of 2ϵ via the matrix Hölder inequality.

We are left to analyze what the dimension of the Hilbert space is that we need to consider. To this end, we go back to our assumption that inputs that are ξ apart in some norm should be mapped to different corners of the Hilbert space. This means we have that D is approximately the number of balls in the norm of our choosing that we need to cover the input space $\mathcal{X}$. If the space has some sort of characteristic length L and a dimension d, we expect to need

$$D \sim O\left(\left[\frac{L}{\xi}\right]^d\right) \tag{73}$$

dimensions, which is inefficient in the size of the input as we expected from our thought experiment. This is also in line with the conclusions of Ref. [11], which show that overly expressive models have bad generalization performance.

VII. CONCLUSION

If quantum machine learning is ever to become useful, many obstacles have to be overcome. One of them is the *measurement problem* that directly arises from the probabilistic nature of quantum mechanics. Predictions from quantum learning models are inherently random and, as such, many executions of a quantum learning model have to be aggregated to obtain an actual prediction.

In this work, we proposed the concept of *single-shot* quantum machine learning. We established a rigorous definition of when a quantum classifier can circumvent the measurement problem by providing labels in a near-deterministic way.

Exploiting a close connection between quantum classifiers and quantum multihypothesis testing allowed us to establish that *how* single shot a quantum classifier can be is fundamentally constrained by the distinguishability of the embedded quantum states. Considering that embeddings are usually realized by quantum circuits, we additionally showed how a certain depth of these circuits is necessary to realize a single-shot classifier.

Finally, we additionally established that a learning model cannot be single shot in a generic way, i.e., for all possible labelings of the input. In that case, it would become too expressive and there would be no meaningful way for such a model to generalize well.

This work can only be considered the start of our way to overcome the measurement problem. It invites future research in many directions. First of all, we only considered classification tasks in this work. A generalization to regression would surely also be possible and similar fundamental bounds could be expressed in terms of quantum metrology reductions instead of multihypothesis testing [23].

Beyond extending the definitions and fundamental treatment of single-shot quantum machine learning, it is an immediately pressing question to understand, how can we construct models that are single shot? How can we enforce this property during training to obtain accuracy and single shotness? A possible inspiration in this direction is to consider ways of training that are immediately related to the distinguishability of the embedded quantum states [8]. Can we also find explicit constructions for learning models or embedding circuits that enforce a certain degree of single shotness?

ACKNOWLEDGMENTS

The authors wish to thank Matthias Caro for his support in obtaining the continuity bounds used in this work. This work has been supported by the DFG (CRC 183, FOR 2724), the BMBF (Hybrid), the BMWK (PlanQK, EniQmA), the Munich Quantum Valley (K-8), QuantERA (HQCC), and the Einstein Foundation Berlin (Einstein Research Unit on Quantum Devices). This work has also been funded by the DFG under Germany's Excellence Strategy—The Berlin Mathematics Research Center MATH+ (EXC-2046/1, Project ID No. 390685689). It is further supported by the Government of Spain (Severo Ochoa CEX2019-000910-S, FUNQIP and European Union NextGenerationEU PRTR-C17.I1), Fundació Cellex, Fundació Mir-Puig, Generalitat de Catalunya (CERCA program), and European Union (PASQuanS2.1, Grant No. 101113690). E.R. is a fellow of Eurecat's "Vicente López" Ph.D. grant program.

APPENDIX A: MULTIHYPOTHESIS TESTING TO BINARY REDUCTION

This Appendix is devoted to the proof of the reduction from multihypothesis testing to binary hypothesis testing presented in the main text.

Proof of Sec. III. When facing a multihypothesis testing problem, additional side information can only improve our performance. Imagine now that a state ρ_i is selected with probability p_i. A third party then selects the label j of the other state in the set that is most difficult to distinguish from ρ_i given the prior probabilities, i.e.,

$$j = \operatorname*{argmax}_{1\leqslant j\leqslant r} P_e^*(p_i\rho_i, p_j\rho_j). \tag{A1}$$

The third party then gives us the information that the state we need to detect is either ρ_i or ρ_j. The best strategy we can then perform is the optimal strategy for the binary discrimination of ρ_i and ρ_j, leading to the bound

$$\begin{aligned} &P_e^*\left(\{p_i\rho_i\}_{i=1}^r\right) \\ &\quad\geqslant \sum_{i=1}^r p_i \max_{1\leqslant j\neq i\leqslant r} P_e^*\left(\frac{p_i}{p_i+p_j}\rho_i, \frac{p_j}{p_i+p_j}\rho_j\right) &\text{(A2)}\\ &\quad\geqslant \min_{1\leqslant i\leqslant r}\max_{1\leqslant j\neq i\leqslant r} P_e^*\left(\frac{p_i}{p_i+p_j}\rho_i, \frac{p_j}{p_i+p_j}\rho_j\right), &\text{(A3)}\end{aligned}$$

where we divided the prior probabilities in the error probability term by $p_i + p_j$ to account for the fact that we conditioned on the fact that it is either p_i or p_j, and the p_i in front is the probability of obtaining i in the first place. We lower bound the convex combination by the smallest of the summed-up terms to arrive at the lemma. ■

APPENDIX B: CONTINUITY BOUNDS FOR VARIATIONAL QUANTUM CIRCUITS

This Appendix is devoted to establishing the desired continuity estimates for quantum embeddings. Let us first rigorously define what we understand as a parametrized quantum circuit.

Definition 3. Parametrized quantum circuit. A parametrized quantum circuit is a unitary transformation obtained by successively applying gates on a quantum computer, where the parametrized gates are time evolutions under self-adjoint generating Hamiltonians G_i,

$$U(\boldsymbol{\theta}) = \prod_{i=1}^P e^{-i\theta_i G_i} V_i, \tag{B1}$$

where the nonparametrized gates V_i can be arbitrary. We denote the set of all generators as $\mathcal{G}$.

We can then establish the following fundamental continuity estimate for parametrized quantum circuits in diamond norm.

Theorem 4. Diamond norm continuity. Let $U(\boldsymbol{\theta})$ denote a parametrized quantum circuit with set of generators $\mathcal{G}$ as in Appendix B and $\mathcal{U}(\boldsymbol{\theta})[\rho] = U(\boldsymbol{\theta})\rho U^\dagger(\boldsymbol{\theta})$ its associated channel. Then, the channel obeys the following continuity bound:

$$\|\mathcal{U}(\boldsymbol{\theta})^{\otimes m} - \mathcal{U}(\boldsymbol{\theta}')^{\otimes m}\|_\diamond \leqslant m\Delta\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1, \tag{B2}$$

where Δ is the largest spectral spread of all generating Hamiltonians,

$$\Delta = \max\{\lambda_{\max}(G) - \lambda_{\min}(G) \mid G \in \mathcal{G}\}. \tag{B3}$$

There exist circuits that saturate the above bound.

Proof. We first establish continuity of a single parametrized gate in operator norm. Observe that

$$\|R_G(\theta) - R_G(\theta')\|_\infty = \|e^{-i\theta G} - e^{-i\theta' G}\|_\infty \tag{B4}$$

$$= \|e^{-i(\theta-\theta')G} - \mathbb{I}\|_\infty \tag{B5}$$

$$= \sup_{\lambda\in\mathrm{spec}(H)} |\exp[-i\lambda(\theta-\theta')] - 1| \tag{B6}$$

$$\leqslant \sup_{\lambda\in\mathrm{spec}(H)} |\lambda(\theta-\theta')| \tag{B7}$$

$$\leqslant \|G\|_\infty |\theta-\theta'|, \tag{B8}$$

where the first equality is the definition of the rotation gate, the second equality follows from unitary invariance of the operator norm, the third equality from the fact that the identity commutes with the unitary, and the first inequality from the upper bound $|\exp(-ix) - 1| \leqslant |x|$ for $x \in \mathbb{R}$.

We can use the freedom of adding a global phase, which corresponds to adding a multiple of the identity to the generator, to optimize this bound. The minimal operator norm is

$$\min_{r\in\mathbb{R}} \|G + r\mathbb{I}\|_\infty = \frac{\lambda_{\max}(G) - \lambda_{\min}(G)}{2} =: \frac{\Delta(G)}{2}, \tag{B9}$$

where we defined the spectral spread $\Delta(H)$.

This reasoning easily expands to the full circuit $U(\boldsymbol{\theta})$, which can be seen by noting that a straightforward application of the triangle inequality and the unitary invariance gives us a way to "chop away" gates one after another:

$$\|UR_G(\theta) - U'R_G(\theta')\|_\infty$$

$$= \|UR_G(\theta-\theta') - U'\|_\infty \tag{B10}$$

$$= \|UR_G(\theta-\theta') - U' + U - U\|_\infty \tag{B11}$$

$$\leqslant \|UR_G(\theta-\theta') - U\|_\infty + \|U - U'\|_\infty \tag{B12}$$

$$= \|R_G(\theta-\theta') - \mathbb{I}\|_\infty + \|U - U'\|_\infty \tag{B13}$$

$$\leqslant \frac{\Delta(G)}{2}|\theta-\theta'| + \|U - U'\|_\infty. \tag{B14}$$

Applying this relation recursively yields the bound

$$\|U(\boldsymbol{\theta}) - U(\boldsymbol{\theta}')\|_\infty \leqslant \sum_{i=1}^{P} \frac{\Delta(G_i)}{2}|\theta_i - \theta_i'|$$

$$\leqslant \left(\max_i \frac{\Delta(G_i)}{2}\right)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1. \tag{B15}$$

Lemma 5 of Ref. [21] now allows us to use the operator norm estimate to bound the diamond norm,

$$\|\mathcal{U}(\boldsymbol{\theta}) - \mathcal{U}(\boldsymbol{\theta}')\|_\diamond \leqslant 2\|U(\boldsymbol{\theta}) - U(\boldsymbol{\theta}')\|_\infty$$

$$\leqslant [\max_i \Delta(G_i)]\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1. \tag{B16}$$

To conclude the bound claimed in the theorem statement, we make use of some fundamental properties of the diamond norm (see, e.g., Proposition 3.48 of Ref. [24]) which we can use to treat the tensor products of channels,

$$\|\mathcal{N}\otimes\mathcal{M} - \mathcal{N}'\otimes\mathcal{M}'\|_\diamond$$

$$\leqslant \|\mathcal{N}\otimes\mathbb{I} - \mathcal{N}'\otimes\mathbb{I}\|_\diamond + \|\mathbb{I}\otimes\mathcal{M} - \mathbb{I}\otimes\mathcal{M}'\|_\diamond \tag{B17}$$

$$= \|\mathcal{N} - \mathcal{N}'\|_\diamond + \|\mathcal{M} - \mathcal{M}'\|_\diamond. \tag{B18}$$

The above implies that

$$\|\mathcal{N}^{\otimes m} - \mathcal{M}^{\otimes m}\|_\diamond \leqslant m\|\mathcal{N} - \mathcal{M}\|_\diamond, \tag{B19}$$

which indeed implies the desired statement.

The tightness of the bound is established by considering the saturating example of parallel Z rotations on n qubits,

$$U(\boldsymbol{\theta}) = \bigotimes_{i=1}^{n} R_Z(\theta_i). \tag{B20}$$

We will use the following fact for self-adjoint and unitary generators:

$$R_Z(\theta) = \mathbb{I}\cos\theta - iZ\sin\theta. \tag{B21}$$

We consider an infinitesimal perturbation around $\boldsymbol{\theta} = 0$, for which we have that

$$\bigotimes_{j=1}^{n} R_Z(\mathrm{d}\theta_j) = \bigotimes_{j=1}^{n}(\mathbb{I}\cos\mathrm{d}\theta_j - iZ\sin\mathrm{d}\theta_j) \tag{B22}$$

$$= \bigotimes_{j=1}^{n}(\mathbb{I} - iZ\mathrm{d}\theta_j) \tag{B23}$$

$$= \bigotimes_{j=1}^{n}\mathbb{I} - \sum_{j=1}^{n} iZ_j\mathrm{d}\theta_j + O\big(\|\mathrm{d}\boldsymbol{\theta}\|_1^2\big), \tag{B24}$$

where we defined $Z_j = \mathbb{I}^{\otimes j-1}\otimes Z\otimes\mathbb{I}^{\otimes n-j}$. With this at hand, we see that

$$\|U(0) - U(\mathrm{d}\boldsymbol{\theta})\|_\infty = \left\|\mathbb{I} - \bigotimes_{j=1}^{n} R_Z(\mathrm{d}\theta_j)\right\|_\infty$$

$$= \left\|\sum_{j=1}^{n} Z_j\mathrm{d}\theta_j\right\|_\infty + O\big(\|\mathrm{d}\boldsymbol{\theta}\|_1^2\big). \tag{B25}$$

As $\sum_{i=1}^{n} Z_i\mathrm{d}\theta_i$ has terms acting on independent tensor factors, we have a simple structure for the eigenstates which we can understand as labeled by bit strings $\mathbf{b}$, such that

$$\sum_{j=1}^{n} Z_j\mathrm{d}\theta_j\,|\mathbf{b}\rangle = \sum_{j=1}^{n}(-1)^{b_j}\mathrm{d}\theta_j\,|\mathbf{b}\rangle\,. \tag{B26}$$

We see that the eigenvalues are all possible signed sums of the $\mathrm{d}\theta_j$, which means that

$$\left\|\sum_{j=1}^{n} Z_j\mathrm{d}\theta_j\right\|_\infty = \max_{\mathbf{b}}\sum_{j=1}^{n}(-1)^{b_j}\mathrm{d}\theta_j = \|\mathrm{d}\boldsymbol{\theta}\|_1. \tag{B27}$$

Therefore,

$$\|\mathcal{U}(0) - \mathcal{U}(\mathrm{d}\boldsymbol{\theta})\|_\diamond = 2\|U(0) - U(\mathrm{d}\boldsymbol{\theta})\|_\infty$$

$$= 2\|\mathrm{d}\boldsymbol{\theta}\|_1 + O(\|\mathrm{d}\boldsymbol{\theta}\|^2), \tag{B28}$$

which—noting that $\Delta(Z) = 2$—saturates the bound as claimed. We now showed the saturation specifically for $m = 1$, but a similar construction works for general m by just repeating the above construction in parallel. ■

We can now use this result to prove Sec. V B of the main text.

Proof of Sec. V B. In our setup, we encounter a parametrized quantum state with two kinds of parameters: the data $\boldsymbol{x}$ and

the variational parameters $\boldsymbol{\theta}$. As the variational parameters are equal for the two circuits that we compare, we can disregard them and absorb them into the fixed unitaries of the circuit. The encoding of the data is repeated L times in our example. This we can take care of by viewing each of these instances as different parameters $\{\tilde{x}_j\}_{j=1}^{Ld}$ and noting that $\|\tilde{\boldsymbol{x}}\|_1 = L\|\boldsymbol{x}\|_1$. Hence, the diamond norm continuity of Appendix B implies that

$$\|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{x}')\|_\diamond \leqslant L\Delta\|\boldsymbol{x} - \boldsymbol{x}'\|_1, \tag{B29}$$

where the maximal spectral spread Δ is now calculated only over the gates encoding the parameters. We can conclude the bounds on the output states through a chain of straightforward inequalities,

$$\|\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) - \rho_{\boldsymbol{\theta}}(\boldsymbol{x}')\|_1 = \|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{x})[\rho_0] - \mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{x}')[\rho_0]\|_1 \tag{B30}$$

$$\leqslant \|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{x}')\|_\diamond \|\rho_0\|_1 \tag{B31}$$

$$\leqslant L\Delta\|\boldsymbol{x} - \boldsymbol{x}'\|_1. \tag{B32}$$

■

APPENDIX C: PROOF OF SEC. V B—CONTINUITY BOUND FOR THE NOISELESS CASE

Because we reduced the task of single-shot classification to multihypothesis testing in Sec. V A, it suffices to establish a lower bound on the error probability of the multihypothesis test between the average states of the majority labels.

Putting together Sec. III and Eq. (4), we find

$$P_e^* \geqslant \min_{1\leqslant i\leqslant r} \max_{1\leqslant j\neq i\leqslant r} P_e^*\left(\frac{p_i}{p_i+p_j}\rho_i, \frac{p_j}{p_i+p_j}\rho_j\right) \tag{C1}$$

$$= \min_{1\leqslant i\leqslant r} \max_{1\leqslant j\neq i\leqslant r} \frac{1}{2}\left(1 - \left\|\frac{p_i}{p_i+p_j}\rho_i - \frac{p_j}{p_i+p_j}\rho_j\right\|_1\right). \tag{C2}$$

Analyzing the elements inside the trace distance and renaming

$$p_+ = \frac{p_i}{p_i+p_j}, \quad \rho_+ = \rho_i \quad \mathcal{X}_+ = \mathcal{X}_i \tag{C3}$$

$$p_- = \frac{p_j}{p_i+p_j}, \quad \rho_- = \rho_j \quad \mathcal{X}_- = \mathcal{X}_j, \tag{C4}$$

we see that

$$\|p_+\rho_+ - p_-\rho_-\|_1 = \left\|\int_{\mathcal{X}_+} \mathrm{d}p(\boldsymbol{x}_+)\,\rho(\boldsymbol{x}_+) - \int_{\mathcal{X}_-} \mathrm{d}p(\boldsymbol{x}_-)\,\rho(\boldsymbol{x}_-)\right\|_1 \tag{C5}$$

$$= \left\|\int_{\mathcal{X}_+} \mathrm{d}p(\boldsymbol{x}_+)\,\rho(\boldsymbol{x}_+)\int_{\mathcal{X}_-}\frac{\mathrm{d}p(\boldsymbol{x}_-)}{p_-} - \int_{\mathcal{X}_-} \mathrm{d}p(\boldsymbol{x}_-)\rho(\boldsymbol{x}_-)\int_{\mathcal{X}_+}\frac{\mathrm{d}p(\boldsymbol{x}_+)}{p_+}\right\|_1 \tag{C6}$$

$$= \left\|\frac{1}{p_+p_-}\int_{\mathcal{X}_+} \mathrm{d}p(\boldsymbol{x}_+)\int_{\mathcal{X}_-} \mathrm{d}p(\boldsymbol{x}_-)\,(p_+\rho(\boldsymbol{x}_+) - p_-\rho(\boldsymbol{x}_-))\right\|_1 \tag{C7}$$

$$\leqslant \frac{1}{p_+p_-}\int_{\mathcal{X}_+} dp(\boldsymbol{x}_+)\int_{\mathcal{X}_-} dp(\boldsymbol{x}_-)\|p_+\rho(\boldsymbol{x}_+) - p_-\rho(\boldsymbol{x}_-)\|_1, \tag{C8}$$

where we have applied the triangle inequality. Looking further into the one-norm,

$$\|p_+\rho(\boldsymbol{x}_+) - p_-\rho(\boldsymbol{x}_-)\|_1 = \|p_+\rho(\boldsymbol{x}_+) - p_-\rho(\boldsymbol{x}_-) - p_-\rho(\boldsymbol{x}_+) + p_-\rho(\boldsymbol{x}_+)\|_1 \tag{C9}$$

$$= \|(p_+ - p_-)\rho(\boldsymbol{x}_+) + p_-[\rho(\boldsymbol{x}_+) - \rho(\boldsymbol{x}_-)]\|_1 \tag{C10}$$

$$\leqslant |p_+ - p_-| + p_-\|\rho(\boldsymbol{x}_+) - \rho(\boldsymbol{x}_-)\|_1, \tag{C11}$$

where we have again applied the triangle inequality. Notice that we could have used p_+ instead of p_- in the added terms; therefore,

$$\|p_+\rho(\boldsymbol{x}_+) - p_-\rho(\boldsymbol{x}_-)\|_1 \leqslant |p_+ - p_-| + \min(p_+, p_-)\|\rho(\boldsymbol{x}_+) - \rho(\boldsymbol{x}_-)\|_1. \tag{C12}$$

Now everything is ready to apply the bound from Sec. V B:

$$\|\rho(\boldsymbol{x}_+) - \rho(\boldsymbol{x}_-)\|_1 \leqslant L\Delta\|\boldsymbol{x}_+ - \boldsymbol{x}_-\|_1. \tag{C13}$$

Putting everything together yields

$$P_e^* \geqslant \min_{1\leqslant i\leqslant r}\max_{1\leqslant j\neq i\leqslant r}\frac{1}{2}\left(1-\frac{|p_i-p_j|}{p_i+p_j}-L\Delta\frac{\min(p_i,p_j)}{p_i+p_j}\int_{\mathcal{X}_+}\frac{dp(\boldsymbol{x}_+)}{p_+}\int_{\mathcal{X}_-}\frac{dp(\boldsymbol{x}_-)}{p_-}\|\boldsymbol{x}_+-\boldsymbol{x}_-\|_1\right) \tag{C14}$$

$$= \min_{1\leqslant i\leqslant r}\max_{1\leqslant j\neq i\leqslant r}\frac{1}{2}\left(1-\frac{|p_i-p_j|}{p_i+p_j}-L\Delta\frac{\min(p_i,p_j)}{p_i+p_j}d_{\text{avg}}^{ij}\right) \tag{C15}$$

$$= \min_{1\leqslant i\leqslant r}\max_{1\leqslant j\neq i\leqslant r}\frac{\min(p_i,p_j)}{p_i+p_j}\left(1-\frac{L\Delta d_{\text{avg}}^{ij}}{2}\right). \tag{C16}$$

In the above derivation, we used the following identity:

$$1-\frac{|p_i-p_j|}{p_i+p_j} = \begin{cases}\frac{p_i+p_j-(p_i-p_j)}{p_i+p_j} & \text{if } p_i>p_j\\ \frac{p_i+p_j-(p_j-p_i)}{p_i+p_j} & \text{if } p_i<p_j\end{cases} \tag{C17}$$

$$= \begin{cases}\frac{2p_j}{p_i+p_j} & \text{if } p_i>p_j\\ \frac{2p_i}{p_i+p_j} & \text{if } p_i<p_j\end{cases} \tag{C18}$$

$$= \frac{2\min(p_i,p_j)}{p_i+p_j}. \tag{C19}$$

APPENDIX D: PROOF OF THE NOISY CONTINUITY BOUND

We will derive a result that will have Sec. V C as a corollary.

Theorem 5. Let

$$\rho_t = \mathcal{D}_p^{\otimes n}\mathcal{U}_t\mathcal{D}_p^{\otimes n}\mathcal{U}_{t-1}\ldots\mathcal{U}_1\mathcal{D}_p^{\otimes n}[\rho_0], \tag{D1}$$

$$\sigma_t = \mathcal{D}_p^{\otimes n}\mathcal{V}_t\mathcal{D}_p^{\otimes n}\mathcal{V}_{t-1}\ldots\mathcal{V}_1\mathcal{D}_p^{\otimes n}[\rho_0] \tag{D2}$$

be the outputs of two quantum circuits of depth t such that $\|\mathcal{U}_i-\mathcal{V}_i\|_\diamond\leqslant\Gamma_i\leqslant\Gamma_{\max}$. Then,

$$\|\rho_t-\sigma_t\|_1\leqslant\Gamma_{\max}\times\begin{cases}t & \text{if } t\leqslant t_0\\ t_0+\sqrt{2n}\frac{p^{t_0}-p^t}{1-p} & \text{if } t>t_0,\end{cases} \tag{D3}$$

where

$$t_0=\left\lceil 1+\frac{1}{2}\frac{\ln 2n}{\ln\frac{1}{p}}\right\rceil. \tag{D4}$$

Proof. We will use the same strategy that establishes the continuity bound of Appendix B; however, in the course of the derivation, we will make use of the fact that the maximally mixed state is a fixed point of any unitary, and hence $(\mathcal{V}^\dagger\mathcal{U}-\mathbb{I})[\omega]=0$, to improve our estimate:

$$\|\rho_t-\sigma_t\|_1$$

$$= \left\|\mathcal{D}_p^{\otimes n}(\mathcal{U}_t[\rho_{t-1}]-\mathcal{V}_t[\sigma_{t-1}])\right\|_1 \tag{D5}$$

$$\leqslant \|\mathcal{U}_t[\rho_{t-1}]-\mathcal{V}_t[\sigma_{t-1}]\|_1 \tag{D6}$$

$$= \|\mathcal{V}_t^\dagger\mathcal{U}_t[\rho_{t-1}]-\sigma_{t-1}\|_1 \tag{D7}$$

$$= \|(\mathcal{V}_t^\dagger\mathcal{U}_t-\mathbb{I})[\rho_{t-1}]+\rho_{t-1}-\sigma_{t-1}\|_1 \tag{D8}$$

$$= \|(\mathcal{V}_t^\dagger\mathcal{U}_t-\mathbb{I})[\rho_{t-1}-\omega]+\rho_{t-1}-\sigma_{t-1}\|_1 \tag{D9}$$

$$\leqslant \|\mathcal{V}_t^\dagger\mathcal{U}_t-\mathbb{I}\|_\diamond\min\{\|\rho_{t-1}\|_1,\|\rho_{t-1}-\omega\|_1\} + \|\rho_{t-1}-\sigma_{t-1}\|_1 \tag{D10}$$

$$\leqslant \Gamma_t\min\{1,p^{t-1}\sqrt{2n}\}+\|\rho_{t-1}-\sigma_{t-1}\|_1 \tag{D11}$$

$$\leqslant \sum_{i=1}^{t}\Gamma_i\min\{1,p^{i-1}\sqrt{2n}\} \tag{D12}$$

$$\leqslant \Gamma_{\max}\sum_{i=1}^{t}\min\{1,p^{i-1}\sqrt{2n}\}. \tag{D13}$$

To complete our proof, note that we have a geometric series for the terms that are powers in p, which gives

$$\sum_{i=1}^{t}p^{i-1}=\frac{1-p^t}{1-p}. \tag{D14}$$

As the terms p^{i-1} are monotonically decreasing in i, the noisy estimate will be better as soon as $i\geqslant t_0$, where t_0 solves

$$1\geqslant p^{t_0-1}\sqrt{2n}\ \Leftrightarrow\ t_0=\left\lceil 1+\frac{1}{2}\frac{\ln 2n}{\ln\frac{1}{p}}\right\rceil. \tag{D15}$$

The final bound is thus an improved version of the continuity version that saturates asymptotically,

$$\|\rho_t-\sigma_t\|_1\leqslant\Gamma_{\max}\times\begin{cases}t & \text{if } t\leqslant t_0\\ t_0+\sqrt{2n}\left(\frac{1-p^t}{1-p}-\frac{1-p^{t_0}}{1-p}\right) & \text{if } t>t_0\end{cases} \tag{D16}$$

$$= \Gamma_{\max}\times\begin{cases}t & \text{if } t\leqslant t_0\\ t_0+\sqrt{2n}\frac{p^{t_0}-p^t}{1-p} & \text{if } t>t_0.\end{cases} \tag{D17}$$

■

We can now specialize the above theorem to the setting of the main text.

Proof of Sec. V C. In the setting of Sec. V C, the total circuit depth is $t=L(\ell+1)$, but only L of the computational steps where data is encoded have a nonzero upper bound on the diamond norm distance given by $\Gamma_{\max}\leqslant\Delta\|\boldsymbol{x}-\boldsymbol{x}'\|_1$. We therefore obtain an estimate of the form

$$\|\rho_{\boldsymbol{\theta}}(\boldsymbol{x})-\rho_{\boldsymbol{\theta}}(\boldsymbol{x}')\|_1\leqslant\Gamma_{\max}\sum_{i=1}^{L}\min\{1,p^{\ell(i-1)}\sqrt{2n}\}, \tag{D18}$$

where i now counts the number of layers. The geometric series then takes the form

$$\sum_{i=1}^{L} p^{(i-1)(\ell+1)} = \sum_{i=1}^{L} (p^{\ell+1})^{i-1} = \frac{1-p^{L(\ell+1)}}{1-p^{\ell+1}}, \tag{D19}$$

and we need to solve the number of layers, L_0, for

$$1 \geqslant p^{(L_0-1)(\ell+1)}\sqrt{2n} \;\Leftrightarrow\; L_0 = \left\lceil 1 + \frac{1}{2(\ell+1)} \frac{\ln 2n}{\ln \frac{1}{p}} \right\rceil. \tag{D20}$$

Combining everything, we obtain the stated bound. ■

[1] R. Sweke, J.-P. Seifert, D. Hangleiter, and J. Eisert, On the quantum versus classical learnability of discrete distributions, Quantum **5**, 417 (2021).

[2] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, Nat. Phys. **17**, 1013 (2021).

[3] N. Pirnay, R. Sweke, J. Eisert, and J.-P. Seifert, Superpolynomial quantum-classical separation for density modeling, Phys. Rev. A **107**, 042416 (2023).

[4] C. Gyurik and V. Dunjko, Exponential separations between classical and quantum learners, arXiv:2306.16028.

[5] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[6] H.-C. Cheng, M.-H. Hsieh, and P. C. Yeh, The learnability of unknown quantum measurements, Quantum Inf. Comput. **16** 615 (2016).

[7] S. Arunachalam and R. de Wolf, Optimal quantum sample complexity of learning algorithms, J. Mach. Learn. Res. **19**, 1 (2018).

[8] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, Quantum embeddings for machine learning, arXiv:2001.03622.

[9] M. Weber, N. Liu, B. Li, C. Zhang, and Z. Zhao, Optimal provable robustness of quantum classification via quantum hypothesis testing, npj Quantum Inf. **7**, 76 (2021).

[10] L. Banchi, J. Pereira, and S. Pirandola, Generalization in quantum machine learning: A quantum information standpoint, PRX Quantum **2**, 040321 (2021).

[11] Y. Du, Y. Yang, D. Tao, and M. H. Hsieh, Problem-dependent power of quantum neural networks on multiclass classification, Phys. Rev. Lett. **131**, 140601 (2023).

[12] L. Banchi, J. L. Pereira, S. T. Jose, and O. Simeone, Statistical complexity of quantum learning, Adv. Quantum Technol. 2300311 (2024).

[13] D. A. Kreplin and M. Roth, Reduction of finite sampling noise in quantum neural networks, Quantum **8**, 1385 (2024).

[14] J. Berberich, D. Fink, D. Pranjić, C. Tutschku, and C. Holm, Training robust and generalizable quantum models, Phys. Rev. Res. **6**, 043326 (2024).

[15] S. Khatri and M. M. Wilde, Principles of quantum communication theory: A modern approach, arXiv:2011.04672.

[16] H. Yuen, R. Kennedy, and M. Lax, Optimum testing of multiple hypotheses in quantum detection theory, IEEE Trans. Inf. Theory **21**, 125 (1975).

[17] C. W. Helstrom, Quantum detection and estimation theory, J. Stat. Phys. **1**, 231 (1969).

[18] K. M. R. Audenaert and M. Mosonyi, Upper bounds on the error probabilities and asymptotic error exponents in quantum multiple state discrimination, J. Math. Phys. **55**, 102201 (2014).

[19] D. S. França and R. García-Patrón, Limitations of optimization algorithms on noisy quantum devices, Nat. Phys. **17**, 1221 (2021).

[20] Y. Du, M. H. Hsieh, T. Liu, S. You, and D. Tao, Learnability of quantum neural networks, PRX Quantum **2**, 040337 (2021).

[21] 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).

[22] E. Gil-Fuster, J. Eisert, and C. Bravo-Prieto, Understanding quantum machine learning also requires rethinking generalization, Nat. Commun. **15**, 2277 (2024).

[23] J. J. Meyer, S. Khatri, D. S. França, J. Eisert, and P. Faist, Quantum metrology in the finite-sample regime, arXiv:2307.06370.

[24] J. Watrous, *The Theory of Quantum Information*, 1st ed. (Cambridge University Press, Cambridge, 2018).