White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/179888/

Version: Published Version

Article:
Banchi, Leonardo, Pereira, Jason and Pirandola, Stefano (2021) Generalization in Quantum Machine Learning: a Quantum Information Standpoint. PRX Quantum. 040321. ISSN: 2691-3399

https://doi.org/10.1103/PRXQuantum.2.040321

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Generalization in Quantum Machine Learning: A Quantum Information Standpoint

Leonardo Banchi,[1,2,*] Jason Pereira,[3,1] and Stefano Pirandola[3]

[1] *Department of Physics and Astronomy, University of Florence, Via Giovanni Sansone 1, Sesto Fiorentino (FI) I-50019 , Italy*

[2] *INFN Sezione di Firenze, Via Giovanni Sansone 1, Sesto Fiorentino (FI) I-50019, Italy*

[3] *Department of Computer Science, University of York, York YO10 5GH, United Kingdom*

(Received 4 March 2021; accepted 30 September 2021; published 1 November 2021)

Quantum classification and hypothesis testing (state and channel discrimination) are two tightly related subjects, the main difference being that the former is data driven: how to assign to quantum states $\rho(x)$ the corresponding class c (or hypothesis) is learnt from examples during training, where x can be either tunable experimental parameters or classical data "embedded" into quantum states. Does the model generalize? This is the main question in any data-driven strategy, namely the ability to predict the correct class even of previously unseen states. Here we establish a link between quantum classification and quantum information theory, by showing that the accuracy and generalization capability of quantum classifiers depend on the (Rényi) mutual information $I(C{:}Q)$ and $I_2(X{:}Q)$ between the quantum state space Q and the classical parameter space X or class space C. Based on the above characterization, we then show how different properties of Q affect classification accuracy and generalization, such as the dimension of the Hilbert space, the amount of noise, and the amount of neglected information from X via, e.g., pooling layers. Moreover, we introduce a quantum version of the information bottleneck principle that allows us to explore the various trade-offs between accuracy and generalization. Finally, in order to check our theoretical predictions, we study the classification of the quantum phases of an Ising spin chain, and we propose the variational quantum information bottleneck method to optimize quantum embeddings of classical data to favor generalization.

DOI: 10.1103/PRXQuantum.2.040321

I. INTRODUCTION

Quantum information and machine learning are two very active areas of research that have become increasingly interconnected [1–4]. In this panorama, many works have considered and designed learning models that are built by using quantum states and algorithms [5–19]. Some of these proposals have focused on learning classical data by exploiting the capability of quantum machines to easily perform computations that are in principle unfeasible using classical computers [5,6]. Other works have instead focused on "quantum data," i.e., information embedded in quantum states or quantum channels.

For the latter case, a fundamental model with particular relevance is that of quantum channel discrimination. This is known to have nontrivial implications for quantum sensing [20], in tasks such as the detection of targets [21–23] or the readout of memories [24,25]. Recently, it has been applied to study the model of channel position finding [26], associated with absorption spectroscopy [27], and more sophisticated problems of barcode decoding and pattern recognition [12]. In the latter, the use of quantum light sources was shown to drastically reduce the error affecting the supervised classification of images, even when the output measurements are not optimized.

The main difference between (supervised) quantum machine learning (QML) and quantum hypothesis testing (QHT), such as state and channel discrimination [28], is the role of prior information. In QHT all possible sets of states and their prior probabilities are known. This is not the case in any machine learning approach, where instead prior information is in the form of *samples*, that is, a collection of correctly classified states. These samples are not enough to cover all possible cases and the most important question in data-driven strategies is to check for generalization: *after having trained the model using a few known examples, can the model accurately classify unseen data?*

*leonardo.banchi@unifi.it

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.

In previous QML literature, the generalization capabilities were numerically verified by computing the classification error over a testing set. Some theoretical bounds were studied in Refs. [10,11,29,30], but only for particular classifiers or regression models, while intuitive geometrical characterizations were given in Refs. [7,8], yet without a formal proof.

Here we study generalization in QML classification tasks using tools from quantum information theory. In Sec. II we first establish a fruitful link between QML and QHT. This allows us to bound the QML classification error by exploiting some QHT results, and to study the role of the unknown prior in QHT. In Sec. III we introduce the main technical result of this paper: quantities linked to either the training or testing errors can be bounded by the quantum mutual information between some suitable quantum states and classical variables. Based on the study of these quantities, in Sec. IV we show different implications of our theoretical bounds: we introduce a quantum version of the bias-variance trade-off, which defines fundamental limitations on the testing error for finite amounts of data; we then show how to use results developed in the quantum communication and cryptography literature to study how to optimally embed classical information onto quantum states; finally, we show how different properties of the quantum states affect the classification accuracy and generalization, such as the dimension of the Hilbert space, the amount of noise, and the amount of neglected information via, e.g., pooling layers. Our results are based on the study of the linear loss function, yet we show how similar conclusions can be obtained in a loss-independent framework, by defining a quantum version of the information bottleneck principle [31]. In Sec. V we consider different applications of our theoretical results. We first study the quantum phase recognition problem of an exactly solvable quantum Ising chain and then define the variational quantum information bottleneck method to train quantum embeddings of classical data for good generalization. Conclusions are drawn in Sec. VI. The mathematical derivations of our results, as well as the extension to multiary classification, are presented in the appendices.

II. QUANTUM HYPOTHESIS TESTING VERSUS SUPERVISED CLASSIFICATION

We study the classification of either quantum states, as in Fig. 1(a), or classical data, as in Fig. 1(b), using the framework of QHT. Let us first consider the simpler case where a quantum device can only be in N_C possible states $\{\rho_c\}_{c=1,\dots,N_C}$ for some integer N_C. The possible values of c are called hypotheses in QHT, or *classes* in this paper. An experimentalist (Alice) performs a measurement on the device with a positive operator-valued measurement (POVM) $\{\Pi_c\}$, whose outcome is the predicted value of c. Via Naimark's dilation theorem, such a POVM can be effectively implemented as shown in Fig. 1(c), namely by using an ancillary system whose Hilbert space dimension is equal to N_C, by first applying a unitary circuit that couples the state and the ancilla, and then performing a projective measurement $|c\rangle\langle c|$ on the ancillary system. In the most general setting, the states ρ_c are not orthogonal and Alice cannot discriminate between them with a single measurement. When the device can be reinitialized in the same state, Alice can use N copies $\rho_c^{\otimes N}$ and the probability of wrong discrimination can decrease exponentially with N [32]. We remark that the common approach of using N measurement "shots" is just a particular case, possibly nonoptimal, of the above general framework, with independent measurements on each copy.

Unlike QHT, in QML classification tasks there are different states that belong to the same class. The number of classes N_C is finite, but the available states $\rho(x)$ are possibly infinite. In this paper the *inputs* x model tunable classical parameters. For instance, consider a device that, depending on parameters x, outputs either entangled ($c=1$) or separable ($c=0$) states [33], or a many-body system that may be in different phases c depending on some external magnetic fields x [9]. We may also be interested in classifying classical data x (e.g., images) using a quantum algorithm, to look for *algorithmic* quantum advantage [11,34] (faster classification), or classifying quantum channels using quantum probes to look for quantum advantage in accuracy [12] (fewer measurements). When dealing with classical inputs x, the quantum embedding circuit can be written as in Fig. 1(b) with x-dependent and x-independent gates $U_i(x)$ and V_j, which may be optimized during training. Finally, a mathematically related, yet physically different problem consists in classifying *physical* objects using quantum sensors [12,21–27], as in the example shown in Fig. 1(d). There, $\rho(x)=\mathcal{E}_x[\rho_{\text{in}}]$ describes the state received by a quantum detector, where ρ_{in} is the input probe state of light, possibly entangled, and $\mathcal{E}_x$ describes how the photons are scattered depending on the objects x living in the physical world.

In all the examples described above, we are interested in learning the unknown functional relation $c=f(x)$ between a classical input x and output class c, yet through measurements on a quantum device. The motivations can be quite diverse and range from quantum device characterization depending on external parameters to the use of quantum algorithms to classify classical data. Following common practices in theoretical machine learning, we assume that all possible pairs of data (c,x) follow some unknown probability distribution $P(c,x)$, so data pairs are independent samples from P [see Fig. 1(b)]. Formally, our ignorance can be modeled using mixed states

$$\rho_c=\sum_x P(x|c)\rho(x), \tag{1}$$

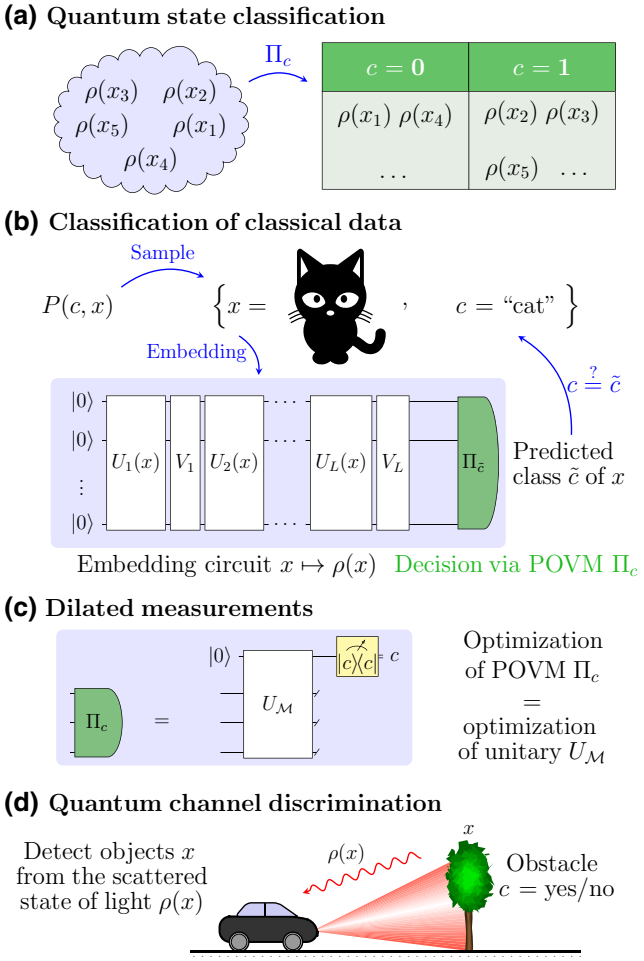

FIG. 1. (a) Example binary classification of quantum states that depend on some external parameters x. (b) Classification of classical data using a quantum embedding circuit with L layers. The classical data are sampled from an *unknown* distribution $P(c,x)$, where x describes, e.g., images of animals and c specifies the kind of animal, e.g., a cat. The classical input x is embedded into a quantum state $\rho(x)$ via layers of x-dependent and x-independent gates. A POVM $\{\Pi_c\}$ is performed at the end of the circuit. The predicted class of x corresponds to the measurement outcome c. (c) Any POVM can be expressed as a unitary circuit followed by a projective measurement $|c\rangle\langle c|$ on a suitably large ancillary system. (d) In quantum channel discrimination, the images x live in the physical world; a quantum probe senses the outside world and $\rho(x)$ is the scattered state of light collected by the detector, which depends on the outside objects. The detector then classifies the image with a POVM, as in (c).

where $P(x|c)$ is the unknown conditional probability. For N copies, such states read $\rho_c^{(N)} = \sum_x P(x|c)\rho(x)^{\otimes N}$. With a slight abuse of notation, to simplify the mathematical expressions, we may hide the dependence on N inside $\rho(x)$, namely as $\rho(x) = \tilde{\rho}(x)^{\otimes N}$ for some $\tilde{\rho}(x)$. The main difference between QHT and the classification problem studied in this paper is that we take measurements on the instances $\rho(x)$ rather than on the discrete states ρ_c. Measurements are still described via a POVM $\{\Pi_c\}$, possibly acting on N copies, constructed such that its outcome c is the predicted class x. Such a quantum classifier is probabilistic: given an input x, the predicted class c is found with probability $p_Q(c|x) = \mathrm{Tr}[\Pi_c\rho(x)]$. Other classifiers can be built using different techniques of quantum decision theory [19,28], for instance by repeating the measurement many times and taking the most likely class, or by defining an observable $\mathcal{M} = \sum_c m_c\Pi_c$ for certain real numbers m_c, and then assigning a certain class depending on the expectation value $\mathrm{Tr}[\mathcal{M}\rho(x)]$. For instance, for binary classification problems with $c = \{0,1\}$, we may set $m_0 = -1$, $m_1 = 1$ and then assign the class depending on the sign of $\mathrm{Tr}[\mathcal{M}\rho(x)]$ [7]. We remark that exact expectation values on real hardware can only be obtained in the limit of infinitely many shots, namely for $N \to \infty$ copies.

Since in QML the probability distribution is unknown, a (sub)optimal classifier must be built from a finite amount of training data. Is the trained classifier able to predict the correct class of previously unseen data? To answer this question, in the next sections we use tools from quantum information theory to formally study the two main errors, namely the approximation and generalization errors, which rigorously formalize the empirical testing error (see Fig. 2). We call $\rho(x)$ parametric quantum states (PQSs) that depend on some tunable classical parameters x and refer to the mapping $x \mapsto \rho(x)$ as *quantum embedding*. We study how different properties of the embedding affect accuracy or generalization, as schematically shown in Fig. 3, and then introduce fundamental limitations on the errors that we may expect for a given data distribution and finite training samples.

A. Training and testing with linear loss

We first formalize the various sources of error that may prevent generalization. Readers already familiar with this topic may skip this section and refer to Fig. 2 for the notation.

In supervised learning the available data are split between a training and a testing set. Both these sets are composed of pairs (c_k, x_k), namely inputs x_k and their true class c_k, but are used differently. We consider a training set $\mathcal{T} = \{(c_k, x_k)\}_{k=1,\ldots,T}$ with T pairs, and similarly a testing set $\mathcal{T}'$ with T' pairs. In the training part a model is optimized in order to minimize a suitable distance between the true class c and the predicted class for all possible pairs $(c,x) \in \mathcal{T}$ in the training set. For given PQSs $\rho(x)$ and POVM $\{\Pi_c\}$, the quantum model predicts a class $\tilde{c}$ with probability $\mathrm{Tr}[\Pi_{\tilde{c}}\rho(x)]$, as in Fig. 1. If c is the true class of x, the linear *loss* is defined as the probability of misclassification, namely the probability that the predicted class $\tilde{c}$ is

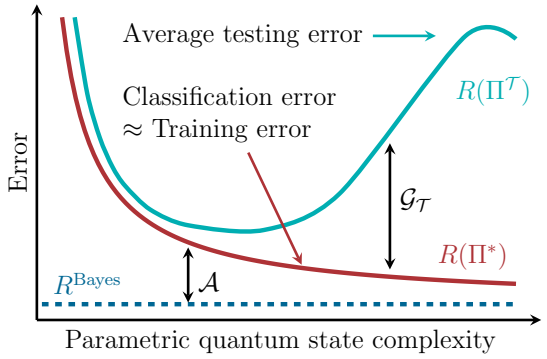

FIG. 2. Summary of the error sources for given parametric quantum states $\rho(x)$ and a finite number of training samples. The classification error $R(\Pi^*)$ is the average loss with the unknown optimal measurement Π^*. The average testing error $R(\Pi^{\mathcal{T}})$ replaces Π^* with the POVM $\Pi^{\mathcal{T}}$ estimated from the training set $\mathcal{T}$ via empirical risk minimization. The testing error $R^{\mathcal{T}'}(\Pi^{\mathcal{T}})$ is a finite sample approximation of $R(\Pi^{\mathcal{T}})$ over the testing set $\mathcal{T}'$. The difference between the average testing error and the Bayes risk R^{Bayes} is split into the approximation error $\mathcal{A}$ and the generalization error $\mathcal{G}^{\mathcal{T}}$. The training error typically behaves as $R(\Pi^*)$.

different from the true class c,

$$\ell(c,x) = \sum_{\tilde{c}\neq c} \text{Tr}[\Pi_{\tilde{c}}\rho(x)] = 1 - \text{Tr}[\Pi_c\rho(x)], \quad (2)$$

where the second equality follows from $\sum_c \Pi_c = \mathbb{1}$. The linear loss allows us to link QML to QHT [19,28].

Training is done via empirical risk minimization, where the empirical risk is the average loss over all possible pairs

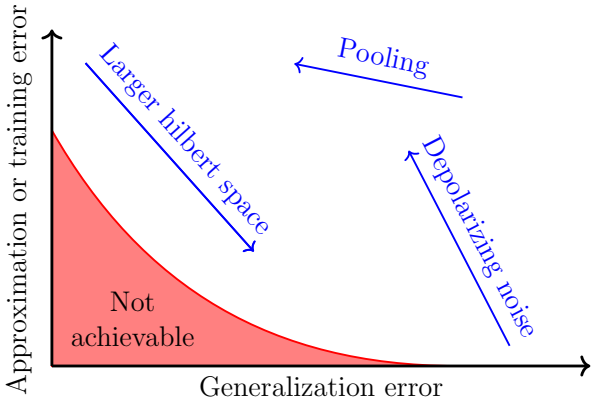

FIG. 3. Summary of some of the main conclusions of this paper. The approximation and generalization errors are respectively mathematically related to the training and testing errors over some datasets, and they cannot be simultaneously minimized (bias-variance trade-off). We use quantum information quantities to bound these errors and show how they are affected by the dimensionality of the quantum Hilbert space, noise, and "information pooling."

(c_k, x_k) in the training set

$$R^{\mathcal{T}}(\Pi) = \frac{1}{T}\sum_{(c_k,x_k)\in\mathcal{T}} \ell(c_k,x_k), \quad (3)$$

and the minimization is over the parameters of the model, namely the POVM and, in some applications, also the embedding. In general, such minimization does not have an analytic solution, except for a few notable cases. For binary classification problems, where $c = \{0,1\}$ can only take two distinct values, the optimal $\mathcal{T}$-dependent POVM, $\Pi^{\mathcal{T}} = \text{argmin}_\Pi[R^{\mathcal{T}}(\Pi)]$, is the Helstrom measurement [35], which is extensively used in QHT. The Helstrom measurement operator $\Pi_0^{\mathcal{T}}$ ($\Pi_1^{\mathcal{T}}$) is the projection onto the eigenspace of positive (negative) eigenvalues of $(T_0/T)\rho_0^{\mathcal{T}} - (T_1/T)\rho_1^{\mathcal{T}}$, where T_c is the number of inputs in the training set with class c and $\rho_c^{\mathcal{T}}$ is a mixture of all the states $\rho(x)$ with inputs in $\mathcal{T}$ and fixed class c. Although not necessary, to simplify the equations, we always assume that the training set contains an equal number of inputs per class, so $T_c/T = 1/2$. Using the optimal Helstrom measurement, the minimum empirical risk can be written analytically in terms of the trace distance between the two average states $\rho_0^{\mathcal{T}}$ and $\rho_1^{\mathcal{T}}$:

$$R^{\mathcal{T}}(\Pi^{\mathcal{T}}) = \frac{1}{2}\left(1 - \frac{1}{2}\|\rho_0^{\mathcal{T}} - \rho_1^{\mathcal{T}}\|_1\right). \quad (4)$$

The above quantity is what defines the *training error* for a given PQS $\rho(x)$, namely the average loss over the training set. From the above equation, zero training error is possible only when $\|\rho_0^{\mathcal{T}} - \rho_1^{\mathcal{T}}\|_1 = 2$, which happens when $\rho_0^{\mathcal{T}}$ and $\rho_1^{\mathcal{T}}$ have orthogonal support. We show in Appendix A that similar conclusions also hold when the number of classes is greater than two.

Does the model generalize? Empirically we need to check how the model performs with inputs not present in the training set. This is normally done by studying the testing error $R^{\mathcal{T}'}(\Pi^{\mathcal{T}})$, which is similar to Eq. (3), but where the samples are taken from the testing set $\mathcal{T}'$ and the POVM $\Pi^{\mathcal{T}}$ is the one minimizing the empirical risk. In order to define the generalization error more formally, we need first to define the true average classification error

$$R(\Pi) = \mathop{\mathbb{E}}_{(c,x)\sim P} \ell(c,x) = 1 - \sum_c P(c)\text{Tr}[\Pi_c\rho_c], \quad (5)$$

where in the last expression we use the chain rule $P(c,x) = P(x|c)P(c) = P(c|x)P(x)$, $P(x) = \sum_c P(c,x)$, $P(c) = \sum_x P(c,x)$ and the definition of Eq. (1). The training error (3) is an empirical approximation of the classification error (5) where the formal average over all possible pairs (c,x) is substituted with a finite average over the training set. The optimal classification POVM, $\Pi^* = \text{argmin}_\Pi R(\Pi)$, is in general different from the $\Pi^{\mathcal{T}}$ that we get from empirical

risk minimization. Overfitting happens when this difference is significant, namely when Π^* and $\Pi^{\mathcal{T}}$ disagree on the class of an input not present in the training set. The generalization error, also called the *estimation error*, is defined as $R(\Pi^{\mathcal{T}}) - R(\Pi^*)$, namely as a difference between two classification errors, where in one case we use the true classifier and in the other we use the classifier built from the training set $\mathcal{T}$. Note that the testing error $R^{\mathcal{T}'}(\Pi^{\mathcal{T}})$ is an empirical approximation of $R(\Pi^{\mathcal{T}})$.

In order to have a low testing error, we need to have both a low generalization error and a low classification error $R(\Pi^*)$. The lowest possible classification error is obtained with the (normally unknown) Bayes classifier, with corresponding Bayes risk R^{Bayes}—see the next section for a formal definition. The difference between the average testing error $R(\Pi^{\mathcal{T}})$ and R^{Bayes} can be written as

$$R(\Pi^{\mathcal{T}}) - R^{\text{Bayes}} = \mathcal{G}^{\mathcal{T}} + \mathcal{A}, \tag{6}$$

with

$$G^{\mathcal{T}} = R(\Pi^{\mathcal{T}}) - R(\Pi^*), \tag{7}$$

$$\mathcal{A} = R(\Pi^*) - R^{\text{Bayes}}. \tag{8}$$

In Eq. (6) the difference between the average testing error and the Bayes risk has been split into two positive terms (see also Fig. 2): $\mathcal{G}^{\mathcal{T}}$ is the previously defined generalization error while $\mathcal{A}$ is called the *approximation error*. A standard result of statistical learning theory, dubbed the bias-variance trade-off [36], shows that it is impossible to minimize both $\mathcal{A}$ and $\mathcal{G}$. Simple classifiers may escape from overfitting but have a bias in the resulting predictions, while too complex classifiers lead to overfitting and a higher variance in the predictions. We remark that these complexity analyses cannot explain the success of deep learning, where models with millions of parameters generalize well in spite of their complexity. There are some explanations of why deep learning works in particular models [37,38], but this is still a subject of intensive research. Moreover, quantum models that can be trained in near-term quantum hardware are quite far from the regime where deep learning operates, so in this paper we focus on models of "moderate complexity."

III. QUANTUM INFORMATION BOUNDS FOR SUPERVISED LEARNING

In this section we study bounds on the approximation and generalization errors using tools from quantum information theory, the main theoretical results of this paper. In Sec. IV we study how different properties of the PQSs $\rho(x)$ affect these errors, while in Sec. V we study more practical applications.

A. Generalization error

Employing tools from statistical learning theory [36] and quantum information, in Appendix A we prove one of our main results, which loosely states that the generalization error typically goes to zero when the number of training pairs T is much larger than a quantity linked to the mutual information between the PQSs and the classical parameters.

Theorem 1. *For a given embedding $x \mapsto \rho(x)$ and any $\delta > 0$, with probability at least $1 - \delta$, the generalization error is bounded as*

$$G^{\mathcal{T}} \leq 2\sqrt{\frac{\mathcal{B}}{T}} + \sqrt{\frac{2\log(1/\delta)}{T}}, \tag{9}$$

where T is the size of the training set,

$$\mathcal{B} = \left(Tr\sqrt{\sum_x P(x)\rho(x)^2}\right)^2 = 2^{I_2(X:Q)} \tag{10}$$

depends on the embedding, and $P(x)$ is the (unknown) prior probability for an input x.

We refer to $\mathcal{B}$ as the *generalization bound*, which constrains how large the generalization error can be for a fixed number T of training pairs. The inequality (9) applies to binary classification problems, but its general form, derived in Appendix A, is equivalent to inequality (9) up to a constant that depends on the number of (equiprobable) classes—see Theorem 2 in Appendix A. The inequality (9), with the explicit form of $\mathcal{B}$ in Eq. (10), represents one of the central results of this paper, as it links the generalization error to properties of the embedding that are measured by information theoretic quantities. Indeed, the quantity found in the second equality of Eq. (10) is the 2-Rényi mutual information between subsystems X and Q of the classical-quantum state

$$\rho_{\text{CXQ}} = \sum_{cx} P(c,x)\, |cx\rangle\langle cx| \otimes \rho(x). \tag{11}$$

For general α and subsystems A and B, the α-Rényi mutual information [39] is defined as

$$I_\alpha(A{:}B) = \frac{\alpha}{\alpha - 1} \log_2 \text{Tr}\sqrt[\alpha]{\text{Tr}_A\left(\rho_A^{(1-\alpha)/2}\rho_{AB}^{\alpha}\rho_A^{(1-\alpha)/2}\right)}. \tag{12}$$

For $\alpha \to 1$, one recovers the quantum mutual information $I_1(A{:}B) \equiv I(A{:}B) = H(A) + H(B) - H(AB)$, where $H(A) = S(\rho_A)$ and $S(\rho) = -\text{Tr}[\rho \log_2 \rho]$ is the von Neumann entropy. Although $I_\alpha(A{:}B)$ for $\alpha \neq 1$ does not satisfy all of the properties of $I_1(A{:}B)$, it does satisfy the data processing inequality [39], namely $I_\alpha(A{:}B) \geq I_\alpha(A'{:}B')$ under

local quantum channels $\mathcal{E}^{A/B\to A'/B'}$, a central ingredient in quantum information theory.

In Eq. (11) we have introduced three Hilbert spaces: the quantum space Q where the PQSs $\rho(x)$ live, the class space C spanned by $\{|c\rangle\}_{c=1,\dots,N_C}$, and the input space X spanned by $\{|x\rangle\}$ for all possible values of x, namely where each input x is mapped onto a different orthogonal state $|x\rangle$. For instance, if the inputs x are made of classical images with n pixels, each with a 16 bit color, then $|x\rangle$ lives in a space of $4n$ qubits. For continuous inputs, e.g., when $\rho(x)$ is an equilibrium state of a many-body system and x some external parameters, one must consider a suitably regularized infinite-dimensional Hilbert space. Here, for simplicity, we assume that X is discrete and can be represented using N_X classical bits, so ρ_{CXQ} lives in an $N_C 2^{N_X+N_Q}$-dimensional Hilbert space.

Inequalities as in Eq. (9) are common in statistical learning theory and show that, with high probability, a model generalizes well whenever $T\to\infty$. The importance of Eq. (10) is in quantifying when the size T of the training set is "large." According to our analysis, a training set is large whenever $T\gg 2^{I_2(X:Q)}$, namely when $\log_2(T)$ is much larger than the number of bits required to describe the information shared between the input distribution and the quantum embedding, as measured by $I_2(X{:}Q)$.

B. Approximation error

Fixing the embedding $x\mapsto\rho(x)$ is like fixing the model class in classical machine learning, e.g., a neural network with a given architecture and a certain number of nodes. The difference between the minimum classification error with a given architecture and the theoretical minimum over all possible architectures, namely the Bayes risk, is the approximation error (8). For a known $P(c,x)$ and a given x, the Bayes classifier picks the class that maximizes $P(c|x)$. The corresponding Bayes risk for binary classification problems with $P(c)=1/2$ is then

$$R^{\mathrm{Bayes}}=1-\frac{1}{2}\sum_x \max\{P(x|0),P(x|1)\}=\frac{1-\Delta}{2},$$

where

$$\Delta=\frac{1}{2}\sum_x\big|P(x|0)-P(x|1)\big|. \tag{13}$$

Using the definition of the approximation error (8) and the classification error $R(\Pi^*)$, which is analogous to Eq. (4) but with the states (1), we find that the approximation error for quantum binary classification problems can be written as

$$\mathcal{A}=R(\Pi^*)-R^{\mathrm{Bayes}}=\Delta-\frac{\|\rho_0-\rho_1\|_1}{2}. \tag{14}$$

It is simple to show that $0\le\mathcal{A}\le\Delta$. Indeed, the upper bound is trivial, and can be achieved when $\rho_0=\rho_1$. As for the lower bound, by defining $\rho_c^{XQ}=\sum_x P(x|c)|x\rangle\langle x|\otimes\rho(x)$, we first note by explicit calculation that $\|\rho_0^{XQ}-\rho_1^{XQ}\|_1=2\Delta$. Then by the contractivity of the trace distance over quantum channels we find that $\|\rho_0^{Q}-\rho_1^{Q}\|_1\le\|\rho_0^{XQ}-\rho_1^{XQ}\|_1$, where $\rho_c^Q=\mathrm{Tr}_X[\rho_c^{QX}]\equiv\rho_c$, thus showing that $\mathcal{A}\ge 0$. The approximation error $\mathcal{A}$ can be interpreted as a generalization of the probability of error in QHT, where the difference is due to the measurements over the instances $\rho(x)$ rather than over the discrete hypothesis states (1). The two errors coincide up to a multiplicative factor when $X\equiv C$.

In the previous section, we have showed how to bound the generalization error using the mutual information between subsystems X and Q in Eq. (11). We can use entropies to bound the average classification error $R(\Pi^*)$, and hence $\mathcal{A}$. Indeed, using the quantum Chernoff bound [40] and explicit calculations we get $R(\Pi^*)\le\frac{1}{2}\mathrm{Tr}[\sqrt{\rho_0}\sqrt{\rho_1}]=2^{-I_{1/2}(C:Q)}-\frac{1}{2}$, which shows that a low classification error is possible when $I_{1/2}(C{:}Q)$ is large. A more general result, valid for any number of classes, can be found using conditional entropies [41],

$$R(\Pi^*)=1-2^{-H_{\min}(C|Q)}\le 1-\frac{2^{I(C:Q)}}{N_C}, \tag{15}$$

which is valid for states of the form given in Eq. (11), where $H_{\min}(C|Q)\le H(C|Q)$ is the min-entropy of C conditioned on Q, which is smaller than von Neumann's conditional entropy [42]. The second inequality comes from $H(C|Q)=H(C)-I(C{:}Q)$ and $H(C)=\log_2 N_C$ for a classification problem with N_C classes. Since Q and C are classically correlated, $I_1(C{:}Q)\le H(C)$ and a small risk is possible when the mutual information between Q and C is large.

To conclude, small $\mathcal{G}$ is possible for small $I_2(X{:}Q)$ while small $\mathcal{A}$ is possible for large $I(C{:}Q)$. In the following section, we build upon these theoretical bounds to study how different properties of the PQSs $\rho(x)$ affect the approximation and generalization errors.

IV. BIAS-VARIANCE TRADE-OFF FOR QUANTUM MACHINE LEARNING

The bias-variance trade-off is a central result in machine learning, stating that it is impossible to minimize both the approximation and the generalization errors. Models with lots of parameters and structure are expected to have low approximation error, potentially at the cost of poor generalization (overfitting). On the other hand, a low-dimensional model with few parameters would be easier to learn, but it might not reliably classify the data.

We study the bias-variance trade-off using quantum information. Remember that the difference between the

average testing error and the Bayes risk can be written as a sum of two positive terms [see Eq. (6)]: the generalization error $\mathcal{G}$ and the approximation error $\mathcal{A}$. The latter is the difference between the average classification error R and the Bayes risk, while the former can be bounded by the generalization bound $\mathcal{B}$ from Eq. (10), and can be made arbitrarily small by considering arbitrarily large training sets with $T \to \infty$. For finite and fixed T, we show that it is impossible to minimize both $\mathcal{A}$ and $\mathcal{G}$, and that different properties of PQSs $\rho(x)$ affect the approximation and generalization errors, as schematically shown in Fig. 3.

Thanks to our framework, many characterizations of PQSs $\rho(x)$ will be formally derived in the next section using tools from quantum information. We use, in particular, the contractivity of the trace distance under quantum channels $\mathcal{E}^{Q\to Q'}$ [43], mapping from the space Q to the space Q', i.e., $\|\mathcal{E}^{Q\to Q'}(\rho-\sigma)\|_1 \leq \|\rho-\sigma\|_1$, the data processing inequality [39] $I_2(X{:}Q) \geq I_2(X{:}Q')$, and finally the bounds satisfied by $\mathcal{B}$, studied in Appendices A and B,

$$1 \leq \mathcal{B} \leq 2^{\min\{H_2(X),N_Q\}}, \tag{16}$$

where $D = 2^{N_Q}$ is the dimension of the embedding Hilbert space, i.e., N_Q is the number of qubits in the PQS, and $H_\alpha[X] = [\alpha/(\alpha-1)]\log_2[\sum_x P(x)^{1/\alpha}]$ is the Rényi entropy of the classical input distribution.

A. Properties of quantum embeddings

Here we focus on the mapping $x \mapsto \rho(x)$ and discuss some properties and desirable features.

It is impossible to minimize both $\mathcal{B}$ and R: the optimal embedding is the one that discards all the *irrelevant* information from the input space X that is not necessary to predict the class C. Indeed, according to Eqs. (10) and (15), $I_2(X{:}Q)$ must be small while $I(C{:}Q)$ must be large. We now show that it is impossible to minimize both $\mathcal{B}$ and R by studying the two extreme cases where the information about X is either fully discarded or fully maintained.

Constant embeddings provide zero generalization error, but the largest approximation error: indeed, the generalization error (7) is defined as the distance between the risk obtained by minimizing the empirical loss over the training data and the true average loss. For constant embeddings, this difference is zero: if we restrict ourselves to classifiers that always produce a constant answer then it is trivial to learn this classifier from data, but the average classification error will be as high as 50%. Mathematically, from the definition (10) and the bounds (16), it is clear that the minimum $\mathcal{B}$ is achieved with the trivial constant embedding, $\rho(x) = \rho$ for all x, but such a trivial embedding provides both the largest $R = 1/2$ and the largest training error $R^{\mathcal{T}} = 1/2$ from Eq. (4). Moreover, using mutual information, the constant embedding is the only embedding for which the space Q is uncorrelated from both C and X in Eq. (11) and so $I(X{:}Q) = I(C{:}Q) = 0$.

Basis encoding guarantees zero approximation error, but the largest generalization error. Basis encoding [2, 44] is defined as $\rho(x) = |x\rangle\langle x|$, namely different inputs are mapped onto orthogonal vectors on a suitably large Hilbert space. No information is lost, or hidden, in the quantum embedding, so using Eqs. (14) and (16), we get the lowest possible approximation error $\mathcal{A} = 0$, meaning that the average loss can achieve the Bayes risk. However, for the same reason, we also get the largest $\mathcal{B} = 2^{H_2(X)}$, since $X = Q$. Therefore, basis encoding allows us to reach the theoretical minimum classification error, but it requires a large embedding space and many ($T \gg \mathcal{B}$) training pairs to avoid overfitting.

High-dimensional embeddings may have lower approximation error: in Appendix A (Theorem 3) we show that if we define an embedding by taking N copies of a simpler one, i.e., if we consider $x \mapsto \rho(x)^{\otimes N}$, then $\mathcal{A} \to 0$ for $N \to \infty$ as long as $F[\rho(x),\rho(y)] \neq 0$ for $x \neq y$, where $F(\rho,\sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1$ is the fidelity between two quantum states. Intuitively, this happens because asymptotically it is possible to correctly discriminate all the states $\rho(x)$ via QHT [12,32], effectively achieving a basis encoding for $N \to \infty$. If $\rho(x)$ is an N_Q qubit state then optimized embeddings using $N \times N_Q$ qubits can only have a lower approximation error than $\rho(x)^{\otimes N}$, since the latter is a particular case. However, high-dimensional embeddings may suffer from poor generalization, as $\mathcal{B}$ may be larger. A numerical check of this prediction is shown in Fig. 4 for a binary discrimination problem with Gaussian priors. We see that, as the number of qubits increases, the classification error risk quickly decreases but the generalization bound increases. This is consistent with our numerical prediction. According to our Theorem 3, for many copies, $\mathcal{A}$ decreases exponentially with N. The asymptotic behavior of $I(X{:}Q)$ is still an open question, but we note that, for local measurements, $I(X{:}Q)$ can be bounded by the mutual information between an input random variable and N output observations, which may display two different regimes $\mathcal{O}(\log N)$ or $\mathcal{O}(N)$ [45]. Therefore we conjecture that $\mathcal{B}$ may slowly increase with N (e.g., polynomially) for particular datasets and embeddings, as we observe numerically for small N.

Low-entropy datasets and low-dimensional embeddings can in principle generalize well: this is a trivial consequence of Eq. (16), when the entropy of the dataset is measured by $H_2[X]$. The statement about the dimension can be made a bit more precise by focusing not just on the dimension of the Hilbert space, but on how much the information is distributed within the Hilbert space. For instance, let us assume a pure state embedding $\rho(x) = U(x)|0\rangle\langle 0|U(x)^\dagger$ with a unitary embedding circuit $U(x)$. If the embedding is such that the input information is "fully scrambled" in a d-dimensional subspace, with $d \ll 2^{N_Q}$, then we may write $\sum_x P(x)\rho(x)^2 \approx \mathbb{1}_d/d$. Substituting this approximation in

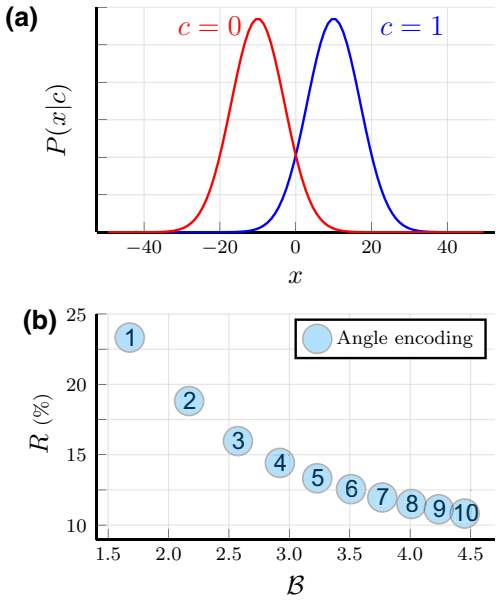

FIG. 4. (a) Example data distributions for two different classes $c = \{0, 1\}$: Gaussian distributions with different means ± 10 and the same standard deviation 7. The corresponding Bayes risk is $R^{\text{Bayes}} \simeq 7.6\%$. (b) Average classification error R (5) versus generalization bound $\mathcal{B}$ (10) for the angle encoding $\rho(x) = |\psi(x)\rangle\langle\psi(x)|^{\otimes N_Q}$, where $|\psi(x)\rangle = \cos(x/2)|0\rangle + \sin(x/2)|1\rangle$ and the x values have been normalized such that $x \in [0, 2\pi]$. The number inside each circle represents the value of $N_Q = 1, \ldots, 10$.

Eq. (10) we get

$$\mathcal{B} \approx \mathcal{O}(d). \tag{17}$$

Therefore, an embedding is capable of generalization not just when built using few qubits, but rather when it "scrambles" information in a small subspace of the full N_Q-qubit Hilbert space.

Geometric characterization: there is an intuitive geometrical characterization of "good" embeddings (see, e.g., Refs. [7,8]). A good embedding is possible when the fidelity between two embedded states is small if the inputs are from different classes and high if the inputs are from the same class, as schematically shown in Fig. 5. This intuitive picture can be explicitly proved using our results. Indeed, using the Fuchs–van de Graaf inequality and the strong concavity of the fidelity, we get

$$\mathbb{E}_{\mathcal{T}} R^{\mathcal{T}}(\Pi^{\mathcal{T}}) \leq \tfrac{1}{2}\mathbb{E}^{\mathcal{T}}[F(\rho_0^{\mathcal{T}}, \rho_1^{\mathcal{T}})] \leq \tfrac{1}{2}F(\rho_0, \rho_1), \tag{18}$$

so, on average, low training errors and low classification errors are possible when the fidelity between the two averaged states ρ_0 and ρ_1 is small. Moreover, in Appendix B we show that $\mathcal{B} \leq [\sum_c \sqrt{P(c)\mathcal{B}_c}]^2$ and, for pure state

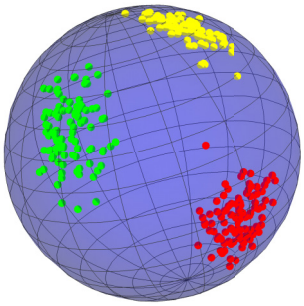

FIG. 5. Geometric visualization of a good embedding, where each point in the sphere represents a state. Points belonging to different classes are plotted with different colors. A good embedding should cluster points with the same class and separate points belonging to different classes.

embeddings,

$$\mathcal{B}_c \leq 1 + \sqrt{(r_c^2 - r_c)(1 - \text{Tr}[\rho_c^2])}, \tag{19}$$

where r_c is the rank of ρ_c. The general case is discussed in Appendix B. The above inequality shows that low generalization error is possible when the average embedding states ρ_c have low rank and/or high purity. Since ρ_c is an ensemble of embeddings for inputs from the same class, the above requirement is satisfied when $\rho(x)$ effectively maps all the inputs from the same class to the same state. More precisely, for pure state embeddings $\rho(x) = U(x)|0\rangle\langle 0|U(x)^\dagger$, the purity can be written as

$$\text{Tr}[\rho_c^2] = \sum_{x,y} P(x|c)P(y|c)F[\rho(x), \rho(y)]^2, \tag{20}$$

where $F[\rho(x), \rho(y)] = |\langle 0|U(y)^\dagger U(x)|0\rangle|$ is the fidelity. Therefore, good generalization is possible whenever $F[\rho(x), \rho(y)]$ is large for all possible pairs (x, y) of inputs with the same class, namely when $\rho(x)$ and $\rho(y)$ are always geometrically close in the embedding Hilbert space. Combining this with Eq. (18) we see that a desirable feature to get a good embedding is that the fidelity between two embedded states is small if the inputs are from different classes and high if the inputs are from the same class, as in Fig. 5.

Noisy operations: we focus on what happens when $\rho(x) = (1 - \epsilon)U(x)|0\rangle\langle 0|U(x)^\dagger + \epsilon \mathbb{1}/2^{N_Q}$, namely when the embedding discussed in the previous example is degraded by depolarising noise with strength ϵ. Again, assuming that the average fully scrambles information in a d-dimensional subspace, then

$$\sum_x P(x)\rho(x)^2 \approx \left[(1 - \epsilon)^2 + \frac{2(1 - \epsilon)\epsilon}{2^{N_Q}}\right]\frac{\mathbb{1}_d}{d} + \frac{\epsilon^2}{2^{2N_Q}}\mathbb{1},$$

from which

$$\mathcal{B} \approx [\sqrt{d}(1-\epsilon+\epsilon/2^{N_Q}) + (1-d/2^{N_Q})\epsilon]^2 \simeq d(1-\epsilon)^2.$$

From the above equation, we see that the generalization error does not increase with noise. It actually decreases when $1 \ll d \ll 2^{N_Q}$, as for large ϵ the embedding approaches the constant embedding, which has the lowest generalization error but the highest classification and approximation errors.

Kernels close to identity are not good for generalization: quantum embeddings can be used to define quantum kernels [5–7,11]. These "kernels" are nothing but the fidelity between two states. Sometimes working with kernels rather than quantum states can be beneficial. In Appendix C we show that, for pure state embeddings $\rho(x) = |\psi(x)\rangle\langle\psi(x)|$, we can express the generalization bound quantity as

$$\mathcal{B} = (\mathrm{Tr}\sqrt{K})^2, \tag{21}$$

where K is the normalized kernel operator, whose matrix elements are

$$K_{x,y} = p(x)\langle\psi(x)|\psi(y)\rangle. \tag{22}$$

Accordingly, the generalization bound can be computed from the eigenvalues η_k of the normalized kernel matrix as $\mathcal{B} = (\sum_k \sqrt{\eta_k})^2$. Note that the study of normalized kernel eigenvalues was also recently proposed as a possible explanation of the generalization capabilities of deep learning [38]. From the above expression, one readily finds that worse generalization performances, according to our bounds (9) and (16), are obtained when $\langle\psi(x)|\psi(y)\rangle \simeq \delta_{x,y}$, namely when different states have almost orthogonal support. Indeed, when $\langle\psi(x)|\psi(y)\rangle = \delta_{x,y}$, Eq. (21) results in the upper bound of Eq. (16).

Pooling may help: we have seen that a large embedding Hilbert space may favor the classification accuracy, yet hinder generalization. According to Eq. (16), the generalization bound may approach its largest value when N_Q, i.e., the number of qubits in the embedding, is as large as $H_2(X)$. What about the minimum number of qubits? For binary classification problems, a good embedding can be obtained even with $N_Q = 1$. Indeed, the simplest embedding that achieves the minimum Bayes risk is $\rho^{\mathrm{Bayes}}(x) = |\tilde{c}\rangle\langle\tilde{c}|$, where $\tilde{c} = \mathrm{argmax}_c P(x|c)$ is, for a given x, the class with largest conditional probability. Clearly, it is impossible to construct this embedding, as the probabilities $P(x|c)$ are unknown, but the above example shows that a good embedding is possible with a single qubit. Although the state before the measurement must be as low dimensional as possible, we may start from a large dimensional embedding and then iteratively *throw away information*, either via measuring some qubits and then applying a different unitary on the remaining ones depending on the measurement result or, equivalently, by applying a conditional gate and then discarding some qubits via a partial trace.

Since the generalization error depends only on the dimension of the final Hilbert space, one can use pooling to iteratively reduce the number of qubits, using different layers, eventually leaving a single qubit for measurements. Promising forms of pooling have been proposed as a basis for quantum convolutional neural networks (QCNNs) [9,46], where the pooling layers are constructed using a reverse multiscale-entanglement-renormalization-ansatz circuit, whose depth depends logarithmically on the total number of qubits. QCNNs have some desirable features, such as the ability to distinguish states corresponding to complex phases of matter [9], and the lack of barren plateaus in their parameter landscape [47], which aids training. Our analysis shows that QCNNs, or other embeddings built by iteratively pooling information, also have good generalization capabilities.

B. Quantum information bottleneck

In the previous section we showed that it is impossible to minimize both the approximation and the generalization errors, when these are defined starting from the linear loss (2). We now show how the generalization-approximation trade-off can also be understood from information theoretic principles that are independent of the choice of loss function. In classical settings, a method designed for this purpose is the information bottleneck (IB) principle [31,48], whose aim is to find the "best" compressed representation Z of the input X that nonetheless has all the relevant information required to predict the class C. The amount of compression can be quantified using the classical mutual information $I(X{:}Z)$, while $I(C{:}Z)$ quantifies the residual information between C and Z. In order to have accurate classification, $I(C{:}Z)$ must be large, while to have good compression, $I(X{:}Z)$ must be small. The information bottleneck principle finds a compromise between accuracy and compression by minimizing the Lagrangian $\mathcal{L}_{\mathrm{IB}} = I(X{:}Z) - \beta I(C{:}Z)$ for a certain value of β. The parameter β allows us to *explore* different regimes and to favor either accuracy or compression. When $\beta = 0$, the minimization of $\mathcal{L}_{\mathrm{IB}}$ achieves the best compression, without caring about correct classification; while for $\beta \to \infty$, the minimization of $\mathcal{L}_{\mathrm{IB}}$ achieves optimal classification without compression. Quantum generalizations of the information bottleneck principle were considered for quantum communication problems in Refs. [49,50]. Here we apply the IB principle to the different problem of finding the optimal embedding.

We focus on the state (11) where X and C are respectively the classical spaces of inputs and classes, and Q is the quantum embedding Hilbert space. We then define the

quantum IB Lagrangian as

$$\mathcal{L}_{\rm IB} = I(X{:}Q) - \beta I(C{:}Q), \tag{23}$$

where $I(A{:}B) = I_{\alpha\to 1}(A{:}B)$ in Eq. (12) is the quantum mutual information. Both $I(X{:}Q)$ and $I(C{:}Q)$ can be expressed using Holevo's accessible information [51]. In Eqs. (9), (10), and (15), we have shown that good generalization is possible whenever $I_2(X{:}Q)$ is small, while low classification error is possible when $I(C{:}Q)$ is large. These conclusions were found for the linear loss (5). We may assume that $I_\alpha(X{:}Q)$ defines a family of generalization bounds for different loss functions, so the minimization of the generalization error is consistent with the minimization of $I(X{:}Q)$, according to some metric, while the maximization of $I(C{:}Q)$ is consistent with the accurate prediction of C from Q. For a particular value of β, the optimal embedding is then obtained as $\min_{\rho(x)} \mathcal{L}_{\rm IB}$. From the definition, we find the explicit form of the IB Lagrangian as

$$\mathcal{L}_{\rm IB} = (1-\beta) S\Big[\sum_x P(x)\rho(x)\Big] - \sum_x P(x) S[\rho(x)] + \beta \sum_c P(c) S[\rho_c] + \sum_x \tilde{\lambda}_x \mathrm{Tr}[\rho(x)] + \eta, \tag{24}$$

where $S[\rho]$ is the von Neumann entropy of ρ, the $\tilde{\lambda}_x$ are Lagrange multipliers to force correct normalization, and η contains all the terms that are independent of the embedding. The optimal embedding corresponds to a minimum of $\mathcal{L}_{\rm IB}$, which satisfies $\partial \mathcal{L}_{\rm IB}/\partial \rho(z) = 0$. By explicit computation we find that the above condition defines a recursive equation for the optimal embedding

$$\tilde{\lambda}_z \rho(z) = e^{(1-\beta)\log\bar{\rho} + \beta \sum_c P(c|z) \log \rho_c} \rho(z), \tag{25}$$

where $\bar{\rho} = \sum_c P(c)\rho_c$ and λ_z is directly related to $\tilde{\lambda}_z$ and is needed to enforce normalization. Alternatively, by restricting to pure state embeddings $\rho(x) = |\psi(x)\rangle\langle\psi(x)|$, we get

$$\tilde{\lambda}_z |\psi(z)\rangle = e^{(1-\beta)\log\bar{\rho} + \beta \sum_c P(c|z) \log \rho_c} |\psi(z)\rangle. \tag{26}$$

From Eqs. (25) and (26) we see that, for $\beta = 0$, we get a constant embedding, while for large β, the optimal embedding for a given x is iteratively obtained from one of the eigenvectors of $\sum_c P(c|x) \log \rho_c$ with the largest eigenvalue, or a mixture of them.

A numerical solution of the IB equations is shown in Fig. 6, where "pure" and "mixed" refer to either Eq. (26) or (25), which are solved for two Gaussian distributions and a single-qubit embedding, using a fixed number of iterations (1000). The Bayes risk is plotted as a reference, giving the smallest possible value of R that can be obtained with any embedding. For a fixed $1 \le \beta \le 3$, we first compute the optimal embedding via either Eq. (26) or (25), and then compute the classification error R from Eq. (5) and the generalization bound $\mathcal{B}$ from Eq. (10). Recall that, for a given $P(c,x)$, the classification error is equivalent to the approximation error $\mathcal{A}$, up to the constant Bayes risk (8). Figure 6(a) shows the approximation-generalization trade-off for the different regimes that we have explored by varying β. For low values of β, we get an almost constant embedding with a large classification error (up to 50%) and low $\mathcal{B}$, while for large values of $\beta \ge 2$, we find that R approaches the theoretical lower bound (Bayes risk), but at the expense of a larger generalization error, as $\mathcal{B}$ gets close to the theoretical upper bound (16). We point out though that, for this particular example, with a single-qubit embedding and two Gaussian priors, the generalization error is always low due to the bound (16).

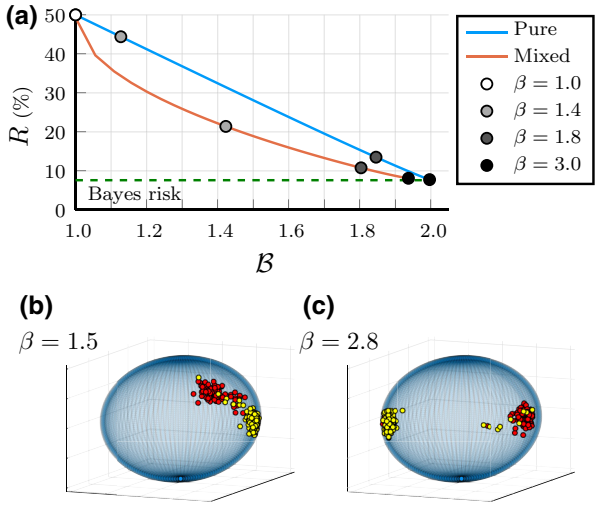

FIG. 6. (a) Average risk (approximation error) R (5) versus generalization error $\mathcal{B}$ (10) for the optimal embeddings obtained by solving the IB equations using either pure (26) or mixed (25) single-qubit states for different values of $\beta \in [1,3]$. Some values of R and $\mathcal{B}$ for particular β are also shown with markers. (b),(c) Bloch sphere visualizations of the quantum embeddings $\rho(x)$ for different values of x sampled from either $P(x|0)$ (red) or $P(x|1)$ (yellow), obtained by solving the IB Eq. (26) for single-qubit encodings, using the two shown values of β. In all three subfigures, the data distributions are those of Fig. 4(a).

The properties of the optimal embedding are shown in Figs. 6(b) and 6(c). In particular, in panel (b) we observe that data belonging to different classes are clustered, but not well separated from each other. On the other hand, for larger values of β, points belonging to different classes are typically very far apart in the Bloch sphere, though there are still some points in the wrong cluster. This prediction is consistent with the analysis of the fidelity between two different embeddings discussed in the previous section and sketched in Fig. 5: a good embedding is such that

$F[\rho(x), \rho(y)]$ is large whenever x and y belong to the same class and small otherwise.

V. APPLICATIONS

In this section we study two different applications of our theoretical results. The first one deals with "quantum data," where the parametric quantum states $\rho(x)$ are fixed by the problem. The second one focuses on the classification of classical data, where the quantum embedding $x \mapsto \rho(x)$ can be optimized. In this latter case, we propose the variational quantum information bottleneck (VQIB) method for optimizing embeddings in order to favor generalization.

A. Quantum phase recognition

In quantum phase recognition [9] the task is to recognize the phases of matter of a quantum many-body system, by taking measurements on the quantum device itself, without having access to a classical description of its state. Here we focus on a paradigmatic exactly solvable model of quantum statistical mechanics, namely the one-dimensional transverse-field Ising model [52]

$$H = -\sum_{i=1}^{L} (\sigma_i^x \sigma_{i+1}^x + h\sigma_i^z), \tag{27}$$

where the $\sigma_j^{x,y,z}$ are the Pauli matrices acting on site j and we consider periodic boundary conditions, $\sigma_{L+1}^{\alpha} \equiv \sigma_1^{\alpha}$. For this model, the classical input is the magnetic field $h \equiv x$. In the thermodynamic limit $L \to \infty$, the model displays a quantum phase transition at the critical value $h = 1$, separating an ordered phase for $|h| < 1$ with twofold degenerate ground states from a disordered phase for $|h| > 1$ with unique ground state. The model can be exactly solved via fermionization [52]. To simplify our analysis for finite L, here we ignore the subtleties of the different fermion parity sectors by considering a small symmetry-breaking term that forces the ground state to have even parity. In that case, for even L, the ground state can be expressed as [53]

$$|\Phi_{\mathrm{GS}}(h)\rangle = \bigotimes_{k=1}^{L/2} [\cos(\theta_{k,h}/2)|00\rangle_k + \sin(\theta_{k,h}/2)|11\rangle_k], \tag{28}$$

where $|00\rangle_k$ and $|11\rangle_k$ are respectively the vacuum and occupied states by two fermion pairs with opposite momentum $k, -k$, and

$$\theta_{k,h} = \arccos\left(\frac{c_k - h}{1 + h^2 - 2hc_k}\right), \qquad c_k = \cos\frac{2\pi k}{L}. \tag{29}$$

From the above expression, it is trivial to compute the overlap $f(h, h') = \langle \Phi_{\mathrm{GS}}(h')|\Phi_{\mathrm{GS}}(h)\rangle = \prod_{k=1}^{L/2} \cos[(\theta_{k,h} - \theta_{k,h'})/2]$. In the thermodynamic limit the fidelity induced distance $1 - f(h, h+\epsilon)$ for small ϵ diverges at the critical point [53]. Therefore, we may expect that the fidelity between two states from the different phases become very small. This is indeed shown in Fig. 7(a). Geometrically, this means that the states belonging to different phases are clustered in distant areas of the Hilbert space, as in Fig. 5. However, $f(h, h')$ decreases exponentially in L for $h \neq h'$, so for large L, the matrix $f(h, h')$ is almost diagonal, thus signaling bad generalization performances according to our Eq. (21).

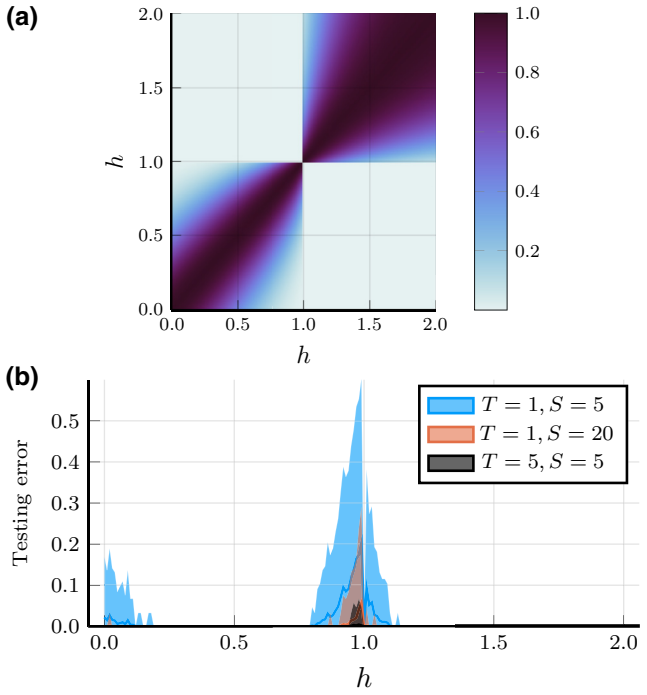

FIG. 7. (a) Fidelity between two ground states of the quantum Ising model with different values of the magnetic field h for $L = 100$. The model displays a quantum phase transition at the critical value $h = 1$, separating ordered ($|h| < 1$) and disordered ($|h| > 1$) phases. (b) Testing error in quantum phase recognition as a function of the magnetic field h. We use the fidelity classifier with a training set of T random elements per phase. Each fidelity is estimated via a SWAP test with S shots. For each h, the fidelity is calculated 1000 times. Solid lines represent the mean fidelity, while shaded areas are the confidence intervals within a standard deviation.

A scaling analysis of $\mathcal{B}$ as a function of L is beyond the scope of this work. In what follows we test our theoretical predictions for a fixed chain length $L = 100$. In this case, we consider a uniform distribution $P(h)$ over $[0, 2]$ and compute $\mathcal{B}$ from Eq. (21)—where x there is the magnetic field h. More specifically, we have discretized the interval such that Eq. (21) can be computed from the numerical eigenvalues, and we have observed that the result converges to $\mathcal{B} \simeq 5.9$ for 100 discretization points.

We then train a fidelity classifier [19] to recognize the phases of the quantum Ising model (27). In general, the fidelity classifier associates to an unknown state $|\psi\rangle$ the class of the state from the training set with highest fidelity with $|\psi\rangle$. Such fidelity can be estimated via the SWAP test using S shots, namely S copies of $|\psi\rangle$. Since the SWAP test measurement operator is idempotent, the result of the SWAP test is a Bernoulli random variable with mean F, the fidelity, and variance $F(1-F)/S$. The fidelity measurement provides a nonoptimal classification POVM, so this classifier is expected to perform slightly worse then the optimal strategies discussed theoretically in the previous sections.

For numerical simulations, we consider a training set with T random elements with $h > 1$ and T random elements with $h < 1$, and verify the quantum phase recognition problem by generating new testing states $|\Phi_{\rm GS}(h)\rangle$ for h uniformly distributed in $[0,2]$. In Fig. 7(b) we numerically observe that even with $T = 1$ the testing error is almost zero, except near the critical point. By increasing the number of shots, the fidelity is estimated more precisely, and given that states belonging to different phases have very low fidelity, as shown in Fig. 7(a), the testing error decreases. When $T \approx \mathcal{B}$, the training error is normally very low, except near the critical point. For $T = 10 \gg \mathcal{B}$, we always find zero training error, irrespective of the number of shots. Therefore, this analysis confirms the predictions of our Theorem 1.

B. Variational quantum information bottleneck

We now focus on using a quantum algorithm to classify classical data. In this case, the states $\rho(x)$ are not fixed by the problem, as in the previous section, and can be optimized together with the measurement POVM. The embedding $x \mapsto \rho(x)$ can be optimized by training a quantum circuit as in Fig. 1. More specifically, we consider one of the simplest yet most general classification circuits with a single-qubit classifier, dubbed "data reuploading" [8]: here we use a slightly modified version where the embedding is obtained as a composition of L layers of x-dependent single-qubit rotations around the y and z axes

$$|\psi_w(x)\rangle = \prod_{\ell=1}^{L} [R^z(w^{z\ell}\cdot x + w_0^{z\ell}) R^y(w^{y\ell}\cdot x + w_0^{y\ell})]|0\rangle, \quad (30)$$

where $R^\alpha(\theta) = e^{i\theta\sigma^\alpha}$, the σ^α are the Pauli matrices, and the weight tensor $w_k^{\alpha\ell}$ can be optimized during training.

Based on the quantum information bottleneck principle proposed in Sec. B we study the variational minimization of the QIB Lagrangian (24) with respect to the parametric states (30). For single-qubit states, the entropies in Eq. (24) can be expressed without loss of generality in terms of the purity as

$$S(\rho) = -(\lambda_- \log_2 \lambda_-) - (\lambda_+ \log_2 \lambda_+) =: s(\mathcal{P}), \quad (31)$$

where

$$\lambda_\pm(\rho) = \frac{1 \pm \sqrt{2\mathcal{P}(\rho) - 1}}{2} \quad (32)$$

are the eigenvalues of ρ, which depend only on the purity $\mathcal{P}(\rho) = \mathrm{Tr}[\rho^2]$. Since the state (30) is pure, $S[\rho(x)] = 0$ in Eq. (24). Moreover, in order to train the embedding, we approximate the averages over the distribution $P(c,x)$ with empirical averages over the elements of the training set $\mathcal{T}$, so from Eq. (24) we get

$$\mathcal{L}_{\rm IB}^{\mathcal{T}} = (1-\beta) s(\mathcal{P}_{\rm tot}) + \beta \sum_c \frac{T_c}{T} s(\mathcal{P}_c), \quad (33)$$

where constant terms have been neglected, and by explicit computation, the purities read

$$\mathcal{P}_{\rm tot} = \frac{T + 2\sum_{x<y}^{\mathcal{T}} F[\rho(x),\rho(y)]^2}{T^2}, \quad (34)$$

$$\mathcal{P}_c = \frac{T_c + 2\sum_{x<y}^{\mathcal{T}_c} F[\rho(x),\rho(y)]^2}{T_c^2}, \quad (35)$$

where $\sum^{\mathcal{T}}$ refers to the double sum over the elements $(c_x, x), (c_y, y)$ from the training set, while in $\sum^{\mathcal{T}_c}$ the sum is restricted over elements with class $c_x = c_y = c$. The ordering $x < y$ refers to the index of the inputs in the training set, and is used just to avoid double counting.

As an example for numerical simulations, we consider a binary classification problem with the 2-moons dataset shown in Fig. 8(a), where each point is described by two real coordinates $x \equiv (x_1, x_2)$. Moon points are organized in the two different patterns shown with different colors in Fig. 8(a), which represent the two classes. Data have been generated using a noise parameter 0.3, which makes the classification less deterministic. We generate a training set of 100 samples per class and optimize Eq. (33) using the Nelder-Mead algorithm with starting point $w_k^{\alpha\ell} = 0$ (constant embedding). In Figs. 8(b) and 8(c), we show the fidelity between two trained embeddings $F[\rho(x),\rho(y)]$, where training is performed using either $\beta = 30$ or $\beta = 1.5$. After training, we use the fidelity classifier [19] to study both the training and testing errors. Unlike the previous section, here we study an exact evaluation of the fidelity, which would require an infinite amount of measurement shots. The training error we get with the optimized embedding is always zero. This is consistent with our theoretical analysis (see Theorem 3 in Appendix

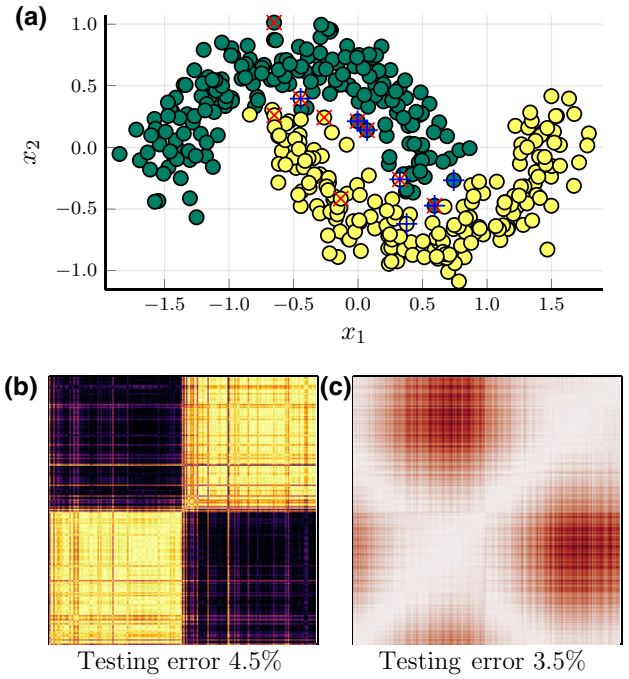

FIG. 8. (a) Union of the training and testing sets from a generated 2-moon dataset. Data (filled circles) with different colors belong to two different classes. Wrongly classified data in the testing set after training are marked with a red diagonal cross ($\beta = 30$) or with a blue cross ($\beta = 1.5$). (b) Fidelity F between two embeddings for $\beta = 30$, using the data from (a). Data points are ordered to first have all points from the first class and then all points from the second class. Dark black points represent $F \simeq 0$, while light yellow points represent $F \simeq 1$. (c) Fidelity F between two embeddings, as in (b) but for $\beta = 1.5$. White points represent $F = 1$ while dark red points have infidelity $1 - F \simeq 10^{-7}$.

A), as for $N \to \infty$ copies we may formally get zero approximation error.

As shown in Fig. 8(b), for large β, the trained embedding is able to separate most data points belonging to different classes into almost orthogonal quantum states. More precisely, the fidelity is almost zero for most inputs belonging to different classes, yet being mostly very high for states belonging to the same class, thus signaling good generalization. Indeed, by generating a testing set with 100 elements per class [also shown in Fig. 8(a)], we observe a testing error $\simeq 4.5\%$. With a much larger testing set of 10^4 points we get a testing error of $\simeq 2.6\%$.

Nonetheless, even better generalization can be obtained for $\beta = 1.5$, although the optimized embedding is almost constant, as shown in Fig. 8(c), with largest infidelity $\simeq 10^{-7}$. The testing errors over the testing sets of 100 or 10^4 elements per class described above are respectively 3.5% and 1.9%, both smaller than those obtained with larger β. The price to pay is that, due to the small infidelities, many more measurements are needed to estimate the fidelity with the due high precision for correct discrimination.

The wrongly classified samples in the smaller testing set are shown in Fig. 8(a) with a cross. We observe that, for the small $\beta = 1.5$, only the elements near the boundaries may be wrongly classified, while for the larger $\beta = 30$, in spite of neater class separation shown in Fig. 8(b), there are wrongly classified samples in the "bulk" of the moons. Something similar was also observed in the numerical simulation shown in Fig. 6(c).

Our analysis shows that the variational quantum information bottleneck method can be successfully used to train quantum embeddings with different generalization properties.

VI. CONCLUSIONS

We have introduced measures of complexity to *quantify* the generalization and approximation capability QML classification problems, either with general parametric quantum states $\rho(x)$ or quantum embeddings $x \mapsto \rho(x)$ of classical data x, when optimal measurements are performed on the system. One of the main results of this paper is the bound on the generalization error via the Rényi mutual information $I_2(X{:}Q)$ between the embedding space Q and the classical input space X. Thanks to our bound, overfitting does not occur when the number of training pairs T is much bigger than $2^{I_2(X:Q)}$. Moreover, we have shown how to bound the approximation error via the mutual information between the embedding space and the class space, and shown that the classification error can approach it lowest possible value (Bayes risk), in the limit of many measurement shots or large Hilbert spaces. Our bounds were obtained for the linear loss function, routinely employed in QHT, but different losses can be linked to the linear loss via bounds. We have also introduced an information bottleneck principle for quantum embeddings, which is independent of the choice of loss function and allows us to explore different trade-offs between accuracy and generalization.

Based on our theoretical results and bounds, we have studied different applications for both the classification of quantum and classical data. In particular, we have studied the classification of the quantum phases of an Ising spin chain and proposed the variational quantum information bottleneck to train quantum embeddings with good generalization properties.

Our analysis can be applied to models of moderate complexity, such as those that can be trained with near-term quantum hardware. It is currently an open question to understand whether quantum classifiers of very high complexity can mimic the generalization capabilities of classical deep learning.

ACKNOWLEDGMENTS

L.B. acknowledges support by the program "Rita Levi Montalcini" for young researchers. J.P and S.P. acknowledge funding from the EU Horizon 2020 Research and Innovation Action under Grant No. 862644 (FET-Open project: "Quantum Readout Techniques and Technologies," QUARTET). L.B. thanks M. Schuld and N. Killoran for helpful comments and discussions about the topics of this paper.

APPENDIX A: EXTENDED DERIVATION

1. Statistical learning theory

Here we show a brief overview of the tools from statistical learning theory [36] that we use throughout the manuscript. As in the previous chapters, we assume that there exists an abstract probability distribution that models the inputs and their corresponding classes $P(c,x)$. This distribution is obviously unknown, but by construction, the samples in the training set $\mathcal{T}$ are drawn independently from $P(c,x)$. Suppose now that we have built a classifier $h \in \mathcal{H}$, where $\mathcal{H}$ is the set of classifiers that we are considering. We may define the error due to misclassification via the loss function $\ell_h(c,x)$, which is zero if and only if c is the correct class of x. Training is done by minimizing the empirical risk, namely the average loss over the training set

$$R^{\mathcal{T}}(h) = \frac{1}{T}\sum_{k=1}^{T} \ell_h(c_k,x_k), \tag{A1}$$

while the "true" risk of a classifier h is given by

$$R(h) = \mathop{\mathbb{E}}_{(c,x)\sim P}[\ell_h(c,x)]. \tag{A2}$$

Supervised learning is practically done via empirical risk minimization, namely the optimal data-driven classifier is obtained from

$$h_{\mathcal{T}} = \mathop{\mathrm{argmin}}_{h\in\mathcal{H}} R^{\mathcal{T}}(h). \tag{A3}$$

The generalization error defines how $h^{\mathcal{T}}$ performs with unseen data, i.e., data not present in the training set. Formally the generalization error is then defined as $R(h^{\mathcal{T}}) - \inf_{h\in\mathcal{H}} R(h)$. Setting $h^* = \mathrm{argmin}_{h\in\mathcal{H}} R(h)$ as the true optimal classifier, we may bound the generalization error $\mathcal{G}$ as

$$\begin{aligned}
\mathcal{G} &= R(h^{\mathcal{T}}) - R(h^*) \\
&= R(h^{\mathcal{T}}) - R^{\mathcal{T}}(h^{\mathcal{T}}) + R^{\mathcal{T}}(h^{\mathcal{T}}) - R^{\mathcal{T}}(h^*) \\
&\quad + R^{\mathcal{T}}(h^*) - R(h^*) \\
&\le R(h^{\mathcal{T}}) - R^{\mathcal{T}}(h^{\mathcal{T}}) + R^{\mathcal{T}}(h^*) - R(h^*) \\
&\le 2\sup_{h\in\mathcal{H}} |R(h) - R^{\mathcal{T}}(h)|,
\end{aligned} \tag{A4}$$

where in the first inequality we used the fact that $h^{\mathcal{T}}$ is optimal for $R^{\mathcal{T}}$; therefore, $R^{\mathcal{T}}(h^{\mathcal{T}}) \le R^{\mathcal{T}}(h^*)$. The upper bound is known as the uniform deviation bound. It represents the maximum deviation between the true and empirical risks, Eqs. (A1)–(A2), maximized over the possible classifiers.

The goal of statistical learning theory is to study how much larger the risk $R(h^{\mathcal{T}})$ is than the Bayes risk, namely $R^{\mathrm{Bayes}} = \inf_h R(h)$, where the infimum is over all possible hypotheses, not restricted to $\mathcal{H}$. Then by summing and subtracting $R(h^*)$ we get

$$R(h^{\mathcal{T}}) - R^{\mathrm{Bayes}} = \mathcal{G} + \mathcal{A}, \tag{A5}$$

where $\mathcal{A} = R(h^*) - R^{\mathrm{Bayes}}$ is the approximation error, which depends on the hypothesis space $\mathcal{H}$. One of the central results of statistical learning theory is the following [36]: if ℓ has support in $[0,1]$ then, with probability at least $1-\delta$, we have

$$\mathcal{G} \le 4\mathcal{C}_T(\mathcal{H}) + \sqrt{\frac{2\log(1/\delta)}{T}}, \tag{A6}$$

where $\mathcal{C}_T(\mathcal{H})$ is the Rademacher complexity of $\mathcal{H}$, which is defined as

$$\mathcal{C}_T(\mathcal{H}) := \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h\in H} \frac{1}{T}\sum_{k=1}^{T} \sigma_k \ell_h(c_k,x_k)\right], \tag{A7}$$

where σ_k is a random variable that can take two possible values, ± 1, with the same probability $1/2$, and the notation $\mathcal{T}\sim P^T$ means that the T elements in the training set $\mathcal{T}$ are sampled independently from the distribution P. From Eq. (A6) we see that if the Rademacher complexity of $\mathcal{H}$ decreases with T then, for sufficiently large T, the model is able to generalize and correctly predict the class of a new input, not present in the training set $\mathcal{T}$.

2. Quantum Rademacher complexity

Let us calculate the Rademacher complexity of the quantum loss function introduced in Eq. (2), for which it is clear from the definition that $0 \le \ell(c_k,x_k) \le 1$, as requested. For a fixed embedding, defining $\mathcal{P}$ as the set of all possible POVMs, the Rademacher complexity of this quantum classifier (2) is

$$\begin{aligned}
\mathcal{C}_T(\mathcal{P}) &:= \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\{\Pi_c\}\in\mathcal{P}} \frac{1}{T}\sum_{k=1}^{T} \sigma_k \sum_{c\neq c_k} \mathrm{Tr}[\Pi_c \rho(x_k)]\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\{\Pi_c\}\in\mathcal{P}} \frac{1}{T}\sum_{k=1}^{T} \sigma_k \mathrm{Tr}[\Pi_{c_k} \rho(x_k)]\right],
\end{aligned} \tag{A8}$$

where in the second line we used the second equality in Eq. (2), the fact that the constant term [from substituting

in Eq. (2)] commutes with the sup and is averaged out by $\mathbb{E}_{\boldsymbol{\sigma}}$, and finally the fact that the minus sign can be removed by noting that $\boldsymbol{\sigma}$ and $-\boldsymbol{\sigma}$ have the same distribution. Let us define

$$Q^{\mathcal{T}}_{c,\boldsymbol{\sigma}} = \frac{1}{T}\sum_{k=1}^{T} \delta_{c,c_k}\sigma_k \rho(x_k); \tag{A9}$$

then by linearity we may rewrite Eq. (A8) as

$$\mathcal{C}_T(\mathcal{P}) := \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\{\Pi_c\}\in\mathcal{P}} \sum_c \mathrm{Tr}[\Pi_c Q^{\mathcal{T}}_{c,\boldsymbol{\sigma}}] \right]. \tag{A10}$$

In the following sections we show how to bound $\mathcal{C}_T(\mathcal{P})$ using quantities that can be easily computed for a given embedding $x \mapsto \rho(x)$. The main technical result that allows such simple expressions is the following

Lemma 1. *Let A_i be a set of operators and i a random variable with probability distribution p_i. Then*

$$\mathop{\mathbb{E}}_{i\sim p}(\|A_i\|_1) \le Tr\sqrt{\mathop{\mathbb{E}}_{i\sim p}(A_i A_i^\dagger)}, \tag{A11}$$

where $\mathop{\mathbb{E}}_{i\sim p} f(i) := \sum_i p_i f(i)$.

Proof. We define the positive operators $X_i := \sqrt{A_i A_i^\dagger}$. Thanks to the definition of the trace norm $\|A\|_1 = \mathrm{Tr}\sqrt{AA^\dagger}$ and the linearity of the trace, it is sufficient to prove that

$$\sum_i p_i X_i \le \sqrt{\sum_i p_i X_i^2}, \tag{A12}$$

where the operator inequality $Y \ge 0$ means that Y is a positive operator. The above inequality is proven as follows. Since the function $f(x) = x^2$ is operator convex [54], we may write

$$\left(\sum_i p_i X_i\right)^2 \le \sum_i p_i X_i^2. \tag{A13}$$

Moreover, since $g(x) = \sqrt{x}$ is operator monotone [54], we may take the square root of both sides of the above equation and get Eq. (A12). Note that a convex combination of positive matrices is also positive, so the left-hand side of Eq. (A12) is a positive operator, and thus is equal to the square root of its square. This completes the proof. ■

We are now ready to write the main result of this section, namely a bound that allows us to express the Rademacher complexity of the quantum classifier via the quantity that was introduced in Eq. (10) and Theorem 1. We focus on binary classification problems, where there are two possible classes, which we call 0 and 1, so $c \in \{0, 1\}$ and a POVM consists of two positive operators, Π_0 and $\Pi_1 = \mathbb{1} - \Pi_0$. Then, we extend the result to a general multiary classification problem with N_C classes.

Theorem 2. *For binary classification problems with fixed embedding $x \mapsto \rho(x)$ and POVM $\mathcal{P}_2 = \{\Pi, \mathbb{1} - \Pi\}$, we find that*

$$\mathcal{C}_T(\mathcal{P}_2) \le \frac{1}{2\sqrt{T}} Tr\sqrt{\sum_x P(x)\rho(x)^2} = \frac{1}{2}\sqrt{\frac{\mathcal{B}}{T}}, \tag{A14}$$

where $\mathcal{B}$ was defined in Eq. (10). For multiary classification problems with N classes, we get

$$\mathcal{C}_T(\mathcal{P}_{N_C}) \le \sqrt{\frac{N_C \mathcal{B}}{T}}, \tag{A15}$$

which is slightly larger than Eq. (A14) when $N_C = 2$.

Proof. We first focus on binary classification problems. Since constant terms are averaged out, we can write Eq. (A10), for $\Pi_0 = \Pi$ and $\Pi_1 = \mathbb{1} - \Pi$, as

$$\mathcal{C}_T(\mathcal{P}_2) = \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\Pi} \mathrm{Tr}[\Pi(Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}})] \right], \tag{A16}$$

where again the constant term is averaged out. The maximization over Π can be done by adapting the Helstrom theorem (see Theorem 13.2 of Ref. [55]):

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\Pi} \mathrm{Tr}[\Pi(Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}})] \right] \\
&\quad \le \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \sup_{\Pi} \{ |\mathrm{Tr}[\Pi_0(Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}})]| \\
&\qquad + |\mathrm{Tr}[\Pi_1(Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}})]| \} \\
&\quad \le \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \sup_{\Pi} \{ \mathrm{Tr}[\Pi_0 |Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}}|] \\
&\qquad + \mathrm{Tr}[\Pi_1 |Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}}|] \} \\
&\quad \le \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathrm{Tr}[|Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}}|]
\end{aligned}$$

with $|A| = \sqrt{AA^\dagger}$ and $\|A\|_1 = \mathrm{Tr}|A|$. In the first inequality, we are again able to average out the constant term, despite the fact it is within an absolute difference, because setting $\boldsymbol{\sigma} \to -\boldsymbol{\sigma}$ changes its sign whilst not affecting the other term in the absolute difference. The second inequality follows from the fact that $|\mathrm{Tr}[AB]| \le \mathrm{Tr}[A|B|]$ for any operator B and positive operator A. The third inequality

follows from the linearity of the trace and the fact that the elements of the POVM sum to the identity. Therefore,

$$\mathcal{C}_T(\mathcal{P}_2) \leq \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\frac{\|Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}}\|_1}{2} \right] = \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left\| \frac{1}{2T} \sum_{k=1}^{T} (\delta_{c_k,0} - \delta_{c_k,1}) \sigma_k \rho(x_k) \right\|_1 . \tag{A17}$$

An alternative proof of the above inequality without the 1/2 follows by using the definition [43] of the trace norm $\|A\|_1 = \max_{B:\|B\|_\infty \leq 1} \mathrm{Tr}[AB]$, and noting that $\|\Pi\|_\infty \leq 1$ for elements of the POVM. We now use Eq. (A11) to get

$$\mathcal{C}_T(\mathcal{P}_2) \leq \frac{1}{2} \mathrm{Tr} \sqrt{\mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} (Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}})^2}, \tag{A18}$$

and by explicit calculation

$$(Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}})^2 = \frac{1}{T^2} \sum_{k,j=1}^{T} (\delta_{c_k,0} - \delta_{c_k,1})(\delta_{c_j,0} - \delta_{c_j,1}) \times \sigma_k \sigma_j \rho(x_k) \rho(x_j). \tag{A19}$$

Since the σ_j are independent and have zero mean, we get

$$\mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} (Q^{\mathcal{T}}_{0,\boldsymbol{\sigma}} - Q^{\mathcal{T}}_{1,\boldsymbol{\sigma}})^2 = \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \sum_{k=1}^{T} \frac{(\delta_{c_k,0} - \delta_{c_k,1})^2}{T^2} \rho(x_k)^2 = \mathop{\mathbb{E}}_{(c,x)\sim P} \left[\frac{1}{T} \rho(x)^2 \right], \tag{A20}$$

where we have used the fact that $(\delta_{c_k,0} - \delta_{c_k,1})^2 = 1$ and that (c_k, x_k) are independent and identically distributed. Inserting the above equation into Eq. (A18) and using $P(x) = \sum_c P(c,x)$ we get Eq. (A14), which completes the first part of the theorem.

For the multiary classification problem with N_C equiprobable classes, using Hölder's inequality in Eq. (A10), and noting that $\|\Pi_c\|_\infty \leq 1$, we may write

$$\begin{aligned} \mathcal{C}_T(\mathcal{P}_{N_C}) &\leq \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \sum_c \|Q^{\mathcal{T}}_{c,\boldsymbol{\sigma}}\|_1 \\ &\leq N_C \mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{c} \|Q^{\mathcal{T}}_{c,\boldsymbol{\sigma}}\|_1 \\ &\leq N_C \mathrm{Tr} \sqrt{\mathop{\mathbb{E}}_{\mathcal{T}\sim P^T} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{c} (Q^{\mathcal{T}}_{c,\boldsymbol{\sigma}})^2} \\ &\leq N_C \mathrm{Tr} \sqrt{\frac{1}{N_C T} \sum_{c,x} P(c,x) \rho(x)^2} \\ &\leq \sqrt{\frac{N_C}{T}} \mathrm{Tr} \sqrt{\sum_x P(x) \rho(x)^2}, \end{aligned} \tag{A21}$$

where in the second line we have substituted the sum over c with an average where c is sampled from the uniform distribution with $p_c = 1/N_C$, in the third line we use Eq. (A11), in the fourth line we perform the averages as in Eq. (A20), and in the last line we simply employ the marginal distribution, as in Eq. (A14). ■

3. Bound on the approximation error

In this section we focus on the approximation error (14) and prove the following important result.

Theorem 3. *Given some x-dependent states $\rho(x)$, if we define an embedding using N copies $x \mapsto \rho(x)^{\otimes N}$ then $\mathcal{A} \to 0$ for $N \to \infty$ as long as $F[\rho(x), \rho(y)] \neq 0$ for $x \neq y$. Moreover,*

$$\lim_{N\to\infty} \frac{\log \mathcal{A}}{N} \leq \log F_{max} \tag{A22}$$

where $F_{max} = \max_{x\neq y} F[\rho(x), \rho(y)]$.

Proof. From the definition of the approximation error $\mathcal{A} = R(h^*) - R^{\mathrm{Bayes}}$, we write

$$R = \sum_x \sum_{c\neq c'} P(c,x) \mathrm{Tr}[\Pi^*_{c'} \rho(x)], \tag{A23}$$

$$R^{\mathrm{Bayes}} = \sum_x \sum_{c\neq c'} P(c,x) \ell^{\mathrm{Bayes}}(c',x), \tag{A24}$$

where $\ell^{\mathrm{Bayes}}(c,x) = \delta_{c,b(x)}$ and $b(x) = \mathrm{argmax}_c P(c|x)$. Note that the second summation is over all c and c' such that $c \neq c'$. Therefore,

$$\mathcal{A} = \sum_x \sum_{c\neq c'} P(x,c) \{\mathrm{Tr}[\Pi^*_{c'} \rho(x)] - \delta_{c',b(x)}\}. \tag{A25}$$

The approximation error is calculated using the optimal measurement for a given encoding $\rho(x)$; however, we can upper bound it by replacing this optimal measurement with a suboptimal measurement. We may find a suboptimal strategy as follows: we know that there always exists some POVM Π_x that obeys [12,32]

$$q(x) \mathrm{Tr}[\Pi_y \rho(x)] \leq \sqrt{q(x)q(y)} F[\rho(x), \rho(y)] \tag{A26}$$

for any probability distribution $q(x)$. From Π_x we can then construct the POVM Π_c as

$$\Pi_c = \sum_x \delta_{c,b(x)} \Pi_x, \tag{A27}$$

namely we first try to learn the value of x and then perform the standard Bayesian classification to get the class.

Now there are many possibilities to find bounds depending on the choice of $q(x)$ in Eq. (A26). Here we chose to use $q(x) = p(x|c)$, namely we train the measurements to recognize all inputs within a certain class, and then check whether the Bayes classifier predicts a different result. We may write

$$\begin{aligned}\mathcal{A} &\leq \sum_x \sum_{c\neq c'} P(c)P(x|c)\left[\sum_y \delta_{c',b(y)}\mathrm{Tr}[\Pi_y\rho(x)] - \delta_{c',b(x)}\right] \\ &= \sum_x \sum_{c\neq c'} P(c)\left[\sum_y \delta_{c',b(y)}\mathcal{F}^c_{xy} - P(x|c)\delta_{c',b(x)}\right] \\ &= \sum_{x\neq y}\sum_{c\neq b(y)} P(c)\mathcal{F}^c_{xy}, \end{aligned} \tag{A28}$$

where, in the last line, the summation is over all x and y such that $x \neq y$, and where

$$F^c_{xy} = \sqrt{P(x|c)P(y|c)}F[\rho(x),\rho(y)]. \tag{A29}$$

The upper bound (A28) is typically too large to be practical. However, it can be used to show an important result. If we define an embedding as $x \mapsto \rho(x)^{\otimes N}$ then $F^c_{xy} = \sqrt{P(x|c)P(y|c)}F[\rho(x),\rho(y)]^N \to 0$ for $N \to \infty$ as long as $F[\rho(x),\rho(y)] \neq 0$. Moreover, since $F[\rho(x),\rho(y)] \leq F_{\max}$, we get Eq. (A22). ■

Thanks to the above theorem, we see that taking copies of a simple embedding guarantees that $\mathcal{A} \to 0$ for $N \to \infty$, as we observe in the numerical simulations shown in Fig. 4.

APPENDIX B: FURTHER INEQUALITIES

In this section we discuss other inequalities and connections with other entropic quantities. We first recall the following inequality, which will be extensively used in this section:

$$\mathrm{Tr}\sqrt{\sum_i X_i} \leq \sum_i \mathrm{Tr}\sqrt{X_i}. \tag{B1}$$

This inequality is valid for any set of positive operators X_i (see Ref. [56] for a proof).

We first discuss the risk (2) and empirical risk (3) for multiary classification problems with N_C classes. In this case there is no known analytic form of the optimal POVM, but suboptimal choices can be constructed using pretty good measurements [12,57]: calling T_c the number of samples in the training set with class c, we may write $R^{\mathcal{T}}(\Pi) = 1 - \sum_c (T_c/T)\mathrm{Tr}[\Pi_c\rho_c^{\mathcal{T}}]$ and the error for an optimal measurement can be bounded as [12,57]

$$R^{\mathcal{T}}(\Pi^{\mathcal{T}}) \leq \sum_{c\neq c'} \frac{\sqrt{T_cT_{c'}}}{T}F(\rho_c^{\mathcal{T}},\rho_{c'}^{\mathcal{T}}), \tag{B2}$$

where $F(\rho,\sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1$ is the quantum fidelity. Using the strong concavity of the fidelity,

$$\mathbb{E}^{\mathcal{T}}R^{\mathcal{T}}(\Pi^{\mathcal{T}}) \leq \sum_{c\neq c'} \frac{\sqrt{T_cT_{c'}}}{T}F(\rho_c,\rho_{c'}), \tag{B3}$$

which is a multiclass generalization of Eq. (18). Therefore, even for classification problems with $N_C > 2$, low risk is possible when the fidelity between the average states with inputs belonging to different classes is low.

As for the generalization error, we see that the complexity $\mathcal{B}$ defined in Eq. (10) does not depend on the classes. Nonetheless, using Eq. (B1) we also get the inequalities

$$\mathcal{B} \leq \left(\sum_c \sqrt{P(c)\mathcal{B}_c}\right)^2, \tag{B4}$$

$$\mathcal{B}_c = \left(\mathrm{Tr}\sqrt{\sum_x P(x|c)\rho(x)^2}\right)^2. \tag{B5}$$

Setting $\sigma_c = \sum_x P(x|c)\rho(x)^2 = \sum_{i=1}^{r_c}\lambda_i|\lambda_i\rangle\langle\lambda_i|$, where r_c is the rank of σ_c, we find that

$$\mathcal{B}_c = \left(\sum_i \sqrt{\lambda_i}\right)^2 = \sum_i \lambda_i + \sum_{i\neq j}\sqrt{\lambda_i\lambda_j}. \tag{B6}$$

Thanks to Jensen's inequality, for every set of positive x_i of size n, we can write $\sum_{i=1}^n \sqrt{x_i} \leq \sqrt{n\sum_{i=1}^n x_i}$. In Eq. (B6), the number of terms in the second sum is $r_c^2 - r_c$, so

$$\begin{aligned}\mathcal{B}_c &\leq \sum_i \lambda_i + \sqrt{(r_c^2 - r_c)\sum_{i\neq j}\lambda_i\lambda_j} \\ &= \sum_i \lambda_i + \sqrt{(r_c^2 - r_c)\left(\sum_{i,j}\lambda_i\lambda_j - \sum_i \lambda_i^2\right)} \\ &= \mathrm{Tr}[\sigma_c] + \sqrt{(r_c^2 - r_c)(\mathrm{Tr}[\sigma_c]^2 - \mathrm{Tr}[\sigma_c^2])}. \end{aligned} \tag{B7}$$

For pure state embeddings, $\sigma_c = \rho_c$ and we get Eq. (19).

Another interesting bound can be found by applying the Cauchy-Schwarz inequality $\mathrm{Tr}[\sqrt{X}]^2 \leq \mathrm{Tr}[X]\mathrm{Tr}[\mathbb{1}]$ to Eq. (10). We find that

$$\mathcal{B} \leq D\sum_x P(x)\mathrm{Tr}[\rho(x)^2] \leq D, \tag{B8}$$

where D is the dimension of the embedding Hilbert space.

We now study the embedding $\rho(x)$ using tools from quantum information. We define an extended tripartite mixed ρ_{CXQ} state as in Eq. (11), where CX is the data Hilbert space, C is spanned by the classes $|c\rangle$ and X by the labels $|x\rangle$, while Q is the Hilbert space of the quantum embedding $\rho(x)$. We now introduce the Rényi conditional mutual information of ρ_{CXQ} following Proposition 8 of Ref. [39],

$$I_\alpha(X{:}Q|C) = \frac{\alpha}{\alpha-1}\log_2 \mathrm{Tr}\Big\{\Big[\rho_C^{(\alpha-1)/2} \times \mathrm{Tr}_X\Big(\rho_{CX}^{(1-\alpha)/2}\rho_{\mathrm{CXQ}}^{\alpha}\rho_{CX}^{(1-\alpha)/2}\Big)\rho_C^{(\alpha-1)/2}\Big]^{1/\alpha}\Big\}, \tag{B9}$$

where $\rho_{CX} = \mathrm{Tr}_Q[\rho_{\mathrm{CXQ}}] = \sum_{xc} P(c,x)|cx\rangle\langle cx|$ and $\rho_C = \mathrm{Tr}_X[\rho_{XC}] = \sum_c P(c)|c\rangle\langle c|$. When multiplying operators that span different Hilbert spaces, it is implicit that the operators take a tensor product with the identity on the spaces that they do not span [e.g., $\rho_{CX}\rho_{\mathrm{CXQ}} = (\rho_{CX}\otimes \mathbb{1}_Q)\rho_{\mathrm{CXQ}}$]. By explicit computation,

$$\begin{aligned}\mathrm{Tr}_X\Big(\rho_{CX}^{(1-\alpha)/2}\rho_{\mathrm{CXQ}}^{\alpha}\rho_{CX}^{(1-\alpha)/2}\Big) &= \sum_{cx} P(c,x)^{1-\alpha}P(c,x)^{\alpha}|c\rangle\langle c|\otimes\rho(x)^{\alpha}\\ &= \sum_{cx} P(c,x)|c\rangle\langle c|\otimes\rho(x)^{\alpha},\end{aligned} \tag{B10}$$

$$\begin{aligned}\rho_C^{(\alpha-1)/2}\mathrm{Tr}_X\Big(\rho_{CX}^{(1-\alpha)/2}\rho_{\mathrm{CXQ}}^{\alpha}\rho_{CX}^{(1-\alpha)/2}\Big)\rho_C^{(\alpha-1)/2} &= \sum_{cx} P(c,x)P(c)^{\alpha-1}|c\rangle\langle c|\otimes\rho(x)^{\alpha}\\ &= \sum_{cx} P(x|c)P(c)^{\alpha}|c\rangle\langle c|\otimes\rho(x)^{\alpha},\end{aligned} \tag{B11}$$

and

$$I_\alpha(X{:}Q|C) = \frac{\alpha}{\alpha-1}\log_2 \times\Big\{\sum_c P(c)\mathrm{Tr}\Big(\Big[\sum_x P(x|c)\rho(x)^{\alpha}\Big]^{1/\alpha}\Big)\Big\}. \tag{B12}$$

We note a similarity between the above expression and the quantities that are found in the generalization bound (B5). Indeed, for a uniform distribution over N_C classes, $P(c) = 1/N_C$, i.e., when all classes are equally likely, we find from Eq. (B5) that

$$\mathcal{B} \le 2^{I_2(X:Q|C)}N_C, \tag{B13}$$

and thus show a direct link between the generalization bound and the Rényi conditional mutual information of ρ_{CXQ}. Therefore, good generalization is possible whenever $I_2(X{:}Q|C)$ is small or, for large training sets, $T \gg 2^{I_2(X:Q|C)}N_C$.

We can interpret the space Q in Eq. (11) as compression of the input into a quantum state. Assuming that the conclusions from Ref. [48] (which were originally formulated for the classical Shannon entropy) can be trivially extended to Rényi entropies, optimal compression happens when

$$I_\alpha(X{:}Q|C) = I_\alpha(X{:}C|Q) = I_\alpha(C{:}Q|X) = 0. \tag{B14}$$

A zero conditional mutual information means that the three systems form a Markov chain: conditioning over one of the three systems makes the other two mutually independent. Quantum mechanically $I(X{:}Q|C) = 0$ if $\rho_{XQC} = \rho_{XC}^{1/2}\rho_C^{-1/2}\rho_{QC}\rho_C^{-1/2}\rho_{XC}^{1/2}$ [39], which is not generally satisfied by the state (11). According to Ref. [48], we should both minimize $I(X{:}Q|C)$ and maximize $I(C{:}Q)$. The Rényi generalization of $I(C{:}Q)$ can be obtained using a similar expression to Eq. (B9), by applying the expression for conditional mutual information to a state that is independent of the conditioning system [i.e., by calculating $I(C{:}Q|X)$ for the state $\rho_{CQ}\otimes\rho_X$]. We get the expression

$$\begin{aligned}I_\alpha(C{:}Q) &= \frac{\alpha}{\alpha-1}\log_2 \mathrm{Tr}\Big\{\Big[\mathrm{Tr}_C\Big(\rho_C^{(1-\alpha)/2}\rho_{CQ}^{\alpha}\rho_C^{(1-\alpha)/2}\Big)\Big]^{1/\alpha}\Big\}\\ &= \frac{\alpha}{\alpha-1}\log_2 \mathrm{Tr}\Big(\Big[\sum_c P(c)\rho_c^{\alpha}\Big]^{1/\alpha}\Big),\end{aligned} \tag{B15}$$

where ρ_c was introduced in Eq. (1). Simpler expressions that are directly connected with the fidelity may be obtained for $\alpha = 1/2$. For pure state embeddings, $\rho(x)^{\alpha} = \rho(x)$, and we get

$$\begin{aligned}I_{1/2}(X{:}Q|C) &= -\log_2\sum_c P(c)\mathrm{Tr}\Big(\Big[\sum_x P(x|c)\sqrt{\rho(x)}\Big]^2\Big)\\ &= -\log_2\sum_c P(c)\mathcal{F}(c),\end{aligned} \tag{B16}$$

where

$$\mathcal{F}(c) = \sum_{x,y} P(x|c)P(y|c)F[\rho(x),\rho(y)]^2 \tag{B17}$$

is the average squared fidelity between embeddings of inputs from the same class c. Therefore, $\mathcal{F}(c)$ should be maximized to minimize $I_{1/2}(X{:}Q|C)$. In other words, low conditional mutual information is possible when $F[\rho(x),\rho(y)] \simeq 1$ for x and y belonging to the same class.

Moreover,

$$-I_{1/2}(C{:}Q) = \log_2 \mathrm{Tr}\left[\sum_c P(c)\sqrt{\rho_c}\right]^2 = \log_2\left[\sum_c P(c)^2 + \sum_{c\neq c'} P(c)P(c')\mathcal{F}(c,c')\right], \tag{B18}$$

where

$$\mathcal{F}(c,c') = \mathrm{Tr}[\sqrt{\rho_c}\sqrt{\rho_{c'}}] \leq F(\rho_c,\rho_{c'}) \tag{B19}$$

is related to the fidelity between the average states ρ_c and $\rho_{c'}$ for different classes. Since the logarithm is monotonic, the minimization of $-I_{1/2}(C{:}Q)$ is possible by minimizing $F(\rho_c,\rho_{c'})$. Therefore, using Rényi entropies with $\alpha = 1/2$, we recover the same conclusions as in the previous section: *a good embedding is one for which the fidelity between two embedded states is small if the inputs are from different classes and high if the inputs are from the same class.*

APPENDIX C: CONNECTION TO KERNELS

We prove the following theorem that allows us to express functions on an ensemble of pure states

Lemma 2. *If the function* $f : [0,1] \mapsto \mathbb{R}$ *admits a series representation then*

$$Trf\left(\sum_x p(x)|\psi(x)\rangle\langle\psi(x)|\right) = Tr[f(\tilde{K})], \tag{C1}$$

where the operator $\tilde{K}$ *has entries*

$$\tilde{K}_{x,y} = \sqrt{p(x)p(y)}\langle\psi(x)|\psi(y)\rangle. \tag{C2}$$

The proof trivially follows from the matrix power

$$\mathrm{Tr}\left(\sum_x p(x)|\psi(x)\rangle\langle\psi(x)|\right)^n = \sum_{\{x_j\}}\prod_{j=1}^n p(x_j)\langle\psi(x_n)|\psi(x_1)\rangle\cdots\langle\psi(x_{n-1})|\psi(x_n)\rangle = \sum_{\{x_j\}}\tilde{K}_{x_n,x_1}\tilde{K}_{x_1,x_2}\cdots\tilde{K}_{x_{n-1},x_n} = \mathrm{Tr}\tilde{K}^n. \tag{C3}$$

When $|z-1|<1$, the square root function admits the expansion $\sqrt{z} = \sum_{k=0}^{\infty}[(-1)^k/k!](-1/2)_k(z-1)^k$, where $(a)_k = a(a+1)\cdots(a+k-1)$ is the Pochammer symbol. Since $(z-1)^k$ can be expressed as a sum of z^n for $n \leq k$, we can apply the above lemma to show that, for ensemble of pure states,

$$\mathcal{B} = (\mathrm{Tr}\sqrt{\tilde{K}})^2 = (\mathrm{Tr}\sqrt{K})^2, \tag{C4}$$

which is Eq. (21). Note that the matrices $\tilde{K}$ and K are related by a similarity transformation and, accordingly, have the same eigenvalues.

For almost diagonal kernel matrices, we may write $\tilde{K} = K_d + K_o$, where $(K_d)_{xy} = p(x)\delta_{xy}$ is the diagonal part and K_o the off-diagonal part, as in Eq. (C2) with $x \neq y$. If the off-diagonal elements are much smaller $[\mathcal{O}(\epsilon)]$ than the diagonal ones, then we may expand $\sqrt{\tilde{K}} = \sqrt{K_d} + \epsilon K_1 + \epsilon^2 K_2 + \mathcal{O}(\epsilon^3)$. Taking the square on either side we find that

$$(K_1)_{xy} = \frac{(K_o)_{xy}}{\sqrt{p(x)}+\sqrt{p(y)}}, \tag{C5}$$

$$(K_2)_{xy} = -\frac{(K_1^2)_{xy}}{\sqrt{p(x)}+\sqrt{p(y)}}, \tag{C6}$$

and

$$\mathrm{Tr}\sqrt{K} = 2^{H_2(X)/2} - \frac{1}{4}\sum_{x\neq y}\frac{\sqrt{p(x)p(y)}}{\sqrt{p(x)}+\sqrt{p(y)}}F_{xy}^2 + \mathcal{O}(\epsilon^4), \tag{C7}$$

where $F_{xy} = |\langle\psi(x)|\psi(y)\rangle|$ and $2^{H_2(X)} = \mathrm{Tr}[\sqrt{K_d}]$.

[1] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature **549**, 195 (2017).

[2] M. Schuld, Supervised Learning with Quantum Computers (Springer, 2018).

[3] V. Dunjko and H. J. Briegel, Machine learning & artif. intell. in the quantum domain: A review of recent progress, Rep. Prog. Phys. **81**, 074001 (2018).

[4] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[5] M. Schuld and N. Killoran, Quantum Machine Learning in Feature Hilbert Spaces, Phys. Rev. Lett. **122**, 040504 (2019).

[6] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature **567**, 209 (2019).

[7] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, Quantum embeddings for machine learning, arXiv:2001.03622 (2020).

[8] 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).

[9] I. Cong, S. Choi, and M. D. Lukin, Quantum convolutional neural networks, Nat. Phys. **15**, 1273 (2019).

[10] C. Ciliberto, A. Rocchetto, A. Rudi, and L. Wossnig, Fast quantum learning with statistical guarantees, arXiv:2001.10477 (2020).

[11] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning, arXiv:2011.01938 (2020).

[12] L. Banchi, Q. Zhuang, and S. Pirandola, Quantum-Enhanced Barcode Decoding and Pattern Recognition, Phys. Rev. Appl. **14**, 064026 (2020).

[13] L. Banchi, N. Pancotti, and S. Bose, Quantum gate learning in qubit networks: Toffoli gate without time-dependent control, Npj Quantum Inf. **2**, 1 (2016).

[14] J. Gray, L. Banchi, A. Bayat, and S. Bose, Machine-Learning-Assisted Many-Body Entanglement Measurement, Phys. Rev. Lett. **121**, 150503 (2018).

[15] H.-Y. Huang, R. Kueng, and J. Preskill, Information-theoretic bounds on quantum advantage in machine learning, arXiv:2101.02464 (2021).

[16] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[17] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, Parameterized quantum circuits as machine learning models, Quantum Sci. Technol. **4**, 043001 (2019).

[18] J.-G. Liu and L. Wang, Differentiable learning of quantum circuit born machines, Phys. Rev. A **98**, 062324 (2018).

[19] S. Gambs, Quantum classification, arXiv:0809.0444 (2008).

[20] S. Pirandola, B. R. Bardhan, T. Gehring, C. Weedbrook, and S. Lloyd, Advances in photonic quantum sensing, Nat. Photon. **12**, 724 (2018).

[21] S. Lloyd, Enhanced sensitivity of photodetection via quantum illumination, Science **321**, 1463 (2008).

[22] S.-H. Tan, B. I. Erkmen, V. Giovannetti, S. Guha, S. Lloyd, L. Maccone, S. Pirandola, and J. H. Shapiro, Quantum Illumination with Gaussian States, Phys. Rev. Lett. **101**, 253601 (2008).

[23] S. Barzanjeh, S. Guha, C. Weedbrook, D. Vitali, J. H. Shapiro, and S. Pirandola, Microwave Quantum Illumination, Phys. Rev. Lett. **114**, 080503 (2015).

[24] S. Pirandola, Quantum Reading of a Classical Digital Memory, Phys. Rev. Lett. **106**, 090504 (2011).

[25] G. Ortolano, E. Losero, S. Pirandola, M. Genovese, and I. Ruo-Berchera, Experimental quantum reading with photon counting, Sci. Adv. **7**, eabc7796 (2021).

[26] Q. Zhuang and S. Pirandola, Entanglement-enhanced testing of multiple quantum hypotheses, Commun. Phys **3**, 103 (2020).

[27] H. Shi, Z. Zhang, S. Pirandola, and Q. Zhuang, Entanglement-Assisted Absorption Spectroscopy, Phys. Rev. Lett. **125**, 180502 (2020).

[28] J. Bae and L.-C. Kwek, Quantum state discrimination and its applications, J. Phys. A: Math. Theor. **48**, 083001 (2015).

[29] H.-C. Cheng, M.-H. Hsieh, and P.-C. Yeh, The learnability of unknown quantum measurements, arXiv:1501.00559 (2015).

[30] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, The power of quantum neural networks, arXiv:2011.00027 (2020).

[31] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, arXiv:physics/0004057 (2000).

[32] A. Montanaro, Pretty simple bounds on quantum state discrimination, arXiv:1908.08312 (2019).

[33] S. Lu, S. Huang, K. Li, J. Li, J. Chen, D. Lu, Z. Ji, Y. Shen, D. Zhou, and B. Zeng, Separability-entanglement classifier via machine learning, Phys. Rev. A **98**, 012315 (2018).

[34] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, arXiv:2010.02174 (2020).

[35] C. W. Helstrom and C. W. Helstrom, *Quantum Detection and Estimation Theory* (Academic press, New York, 1976), Vol. 3.

[36] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms (Cambridge university press, 2014).

[37] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, Proc. Natl. Acad. Sci. **116**, 15849 (2019).

[38] A. Canatar, B. Bordelon, and C. Pehlevan, Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks, Nat. Commun. **12**, 1 (2021).

[39] M. Berta, K. P. Seshadreesan, and M. M. Wilde, Rényi generalizations of the conditional quantum mutual information, J. Math. Phys. **56**, 022205 (2015).

[40] K. M. Audenaert, J. Calsamiglia, R. Munoz-Tapia, E. Bagan, L. Masanes, A. Acin, and F. Verstraete, Discriminating States: The Quantum Chernoff Bound, Phys. Rev. Lett. **98**, 160501 (2007).

[41] R. Konig, R. Renner, and C. Schaffner, The operational meaning of min-and max-entropy, IEEE Trans. Inf. Theory **55**, 4337 (2009).

[42] M. Tomamichel, R. Colbeck, and R. Renner, A fully quantum asymptotic equipartition property, IEEE Trans. Inf. Theory **55**, 5840 (2009).

[43] J. Watrous, *The Theory of Quantum Information* (Cambridge University Press, 2018).

[44] E. Farhi and H. Neven, Classification with quantum neural networks on near term processors, arXiv:1802.06002 (2018).

[45] D. Haussler and M. Opper *et al.*, Mutual information, metric entropy and cumulative relative entropy risk, Ann. Stat. **25**, 2451 (1997).

[46] S. Oh, J. Choi, and J. Kim, in *2020 International Conference on Information and Communication Technology Convergence (ICTC)* (IEEE, 2020), p. 236.

[47] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of barren plateaus in quantum convolutional neural networks, arXiv:2011.02966 (2020).

[48] I. Fischer, The conditional entropy bottleneck, arXiv:2002.05379 (2020).

[49] S. Salek, D. Cadamuro, P. Kammerlander, and K. Wiesner, Quantum rate-distortion coding of relevant information, IEEE Trans. Inf. Theory **65**, 2603 (2018).

[50] N. Datta, C. Hirche, and A. Winter, in *2019 IEEE International Symposium on Information Theory (ISIT)* (IEEE, 2019), p. 1157.

[51] A. S. Holevo, Bounds for the quantity of information transmitted by a quantum communication channel, Problemy Peredachi Informatsii **9**, 3 (1973).

[52] G. B. Mbeng, A. Russomanno, and G. E. Santoro, The quantum ising chain for beginners, arXiv:2009.09208 (2020).

[53] P. Zanardi, M. Cozzini, and P. Giorda, Ground state fidelity and quantum phase transitions in free fermi systems, J. Stat. Mech.: Theory Exp. **2007**, L02002 (2007).

[54] E. Carlen, Trace inequalities and quantum entropy: An introductory course, Entropy Quantum **529**, 73 (2010).

[55] I. Bengtsson and K. Życzkowski, *Geometry of Quantum States: an Introduction to Quantum Entanglement* (Cambridge university press, 2017).

[56] K. M. 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).

[57] H. Barnum and E. Knill, Reversing quantum dynamics with near-optimal quantum and classical fidelity, J. Math. Phys. **43**, 2097 (2002).