Robust in practice: Adversarial attacks on quantum machine learning

Haoran Liao,[1,2,*] Ian Convy,[3,2] William J. Huggins,[3,2] and K. Birgitta Whaley[3,2]

[1]*Department of Physics, University of California, Berkeley, California 94720, USA*
[2]*Berkeley Quantum Information and Computation Center, University of California, Berkeley, California 94720, USA*
[3]*Department of Chemistry, University of California, Berkeley, California 94720, USA*

(Received 22 October 2020; revised 26 February 2021; accepted 1 March 2021; published 28 April 2021)

State-of-the-art classical neural networks are observed to be vulnerable to small crafted adversarial perturbations. A more severe vulnerability has been noted for quantum machine learning (QML) models classifying Haar-random pure states. This stems from the concentration of measure phenomenon, a property of the metric space when sampled probabilistically, and is independent of the classification protocol. To provide insights into the adversarial robustness of a quantum classifier on real-world classification tasks, we focus on the adversarial robustness in classifying a subset of encoded states that are smoothly generated from a Gaussian latent space. We show that the vulnerability of this task is considerably weaker than that of classifying Haar-random pure states. In particular, we find only mildly polynomially decreasing robustness in the number of qubits, in contrast to the exponentially decreasing robustness when classifying Haar-random pure states and suggesting that QML models can be useful for real-world classification tasks.

DOI: 10.1103/PhysRevA.103.042427

I. INTRODUCTION

Quantum machine learning (QML) protocols, by exploiting quantum mechanics principles, such as superposition, tunneling, and entanglement [1], have given hope of outperforming their classical counterparts, even with noisy intermediate-scale quantum (NISQ) [2] hardware in the near-term [3]. For classification tasks where statistical patterns can be revealed in complex feature spaces, the high-dimensional Hilbert space of sizable quantum systems offers a naturally advantageous starting ground for QML models. However, many state-of-the-art classical machine learning models, such as deep neural networks with complicated internal feature mappings, have been shown vulnerable to small crafted perturbations to the input, namely to, adversarial examples [4,5]. These are intentional worst-case perturbations to the original samples with an imperceptible difference, that are nevertheless misclassified by the classifier. This not only raises questions as to why well-performing classifiers suffer from such instabilities but also poses security threats to machine learning applications that emphasize reliability, such as in spam filtering [6]. To understand this unreliable behavior, the transferability of these attacks across different architecture and the robustness against perturbations has led to extensive investigations in the classical machine learning community in recent years [7–9]. Notably, some geometric and probabilistic arguments, based on curvatures of decision boundaries [10] and the concentration of measure [11–15], have been employed to quantify the risk of adversarial attacks in various settings. It has been shown that any classifier will have an adversarial robustness that is increasingly reduced by the dimension of the space on which it classifies, given the concentration of measure phenomenon in certain metric probability spaces [11]. This has raised attention in the QML community where the models take advantage of the high dimensionality of quantum systems [16–19].

The concentration of measure is a phenomenon that describes the fact that, in certain metric probability spaces, points tend to gather around the boundaries of subsets having at least one half of the probability measure. As a result, there is generically a high probability of obtaining values close to the average for any reasonably smooth function that is evaluated on the distribution [20–24]. This means that when samples are selected from such a concentrated space, the confidences predicted by the classifier tends to accumulate around the critical value separating the correct and incorrect classes. As such, small targeted perturbations can then easily move the samples across the decision boundary. In particular, it has been recognized that this phenomenon can lead to extreme vulnerabilities of any quantum classifier on high-dimensional Haar-random pure states [16]. Nevertheless, there is no indication of whether such vulnerability exists when classifying on a subset of encoded pure states in a realistic task, such as using a quantum classifier on classical images encoded in pure states.

In this paper, we approach the task of classifying quantum states from a geometric perspective. The quantum classifier divides the Hilbert space into subsets, each of which belongs to a certain class. We use this perspective here to study aspects of the problem that are relevant to practical applications of QML. In a practical classification task, such as in recognizing natural images, the samples to be classified can be generated from a Gaussian latent space by one of a number of commonly-used generative models [25–30]. The success of these models for real-world data generation ensures that

[*] haoran.liao@berkeley.edu

the focus on QML models classifying a subset of encoded pure states, where these states are sampled from a distribution that is *smoothly* mapped from a Gaussian latent space [15], will yield insight into the vulnerability of QML models in a real-world classification task. This contrasts with the previous analysis of the vulnerabilities when classifying Haar-random pure states [16].

We demonstrate that the *adversarial robustness* over this generated distribution decreases as $O(1/\sqrt{n})$ in the number of qubits n, with the scaling measured in the trace norm. This decline in the robustness is mild, indicating that a quantum classifier can be robust to attacks on high dimensional quantum states. In contrast, when considering *prediction-change* adversarial settings where the inputs are pure states drawn Haar-randomly, we show that the robustness decreases as $O(1/2^n)$ in the number of qubits n, implying extreme vulnerabilities to attacks in high-dimensional quantum systems. This second case parallels the result of Ref. [16], which considered *error-region* adversarial settings and found the robustness also decreases as $O(1/2^n)$ here. However, we argue that the extreme vulnerability in this setting is not of concern in practice, since the states to be classified are always sampled from a distribution over some subsets of states, rather than from the Haar-random distribution over the entire set of pure states.

The rest of the paper is structured as follows. In Sec. II, we introduce the setups and preliminaries in both classical and quantum adversarial attacks. In Sec. III, we describe the prediction-change adversarial setting, which is often more relevant to real-world classification tasks than the previously employed error-region adversarial setting. We then derive the prediction-change adversarial robustness of any quantum classifier on Haar-randomly distributed pure states and explain its practical limitations. In Sec. IV, we derive the main results on the adversarial robustness of any quantum classifier classifying a smoothly generated distribution over a subset of encoded pure states of interest, and propose a feasible modification to any quantum classifier to lower bound unconstrained adversarial robustness. In Sec. V, a summary and discussion of the derived robustness over the two types of distribution are presented.

II. BACKGROUND

A. Classical adversarial attacks

Classical adversarial attacks were introduced to analyze the instability of deep neural networks caused by a small change to the input sample. Classically, the confidence is often quantified as the probability corresponding to the label class in the output normalized discrete distribution, e.g., the largest softmax value in the output vector in a multiclass logistic-regression convolutional neural network. As numerically shown in various works, such an attack results in a significant drop in the confidence in the correct class [4,8,31,32], and may also bring a significant increase in the confidence in an incorrect class [5]. So far, some arguments have been proposed to explain the vulnerabilities of various classifiers to adversarial attacks and their transferability [5,14,33–35], yet no conclusive consensus has been established [36].

The most common type of adversarial attack is the evasion attack where the adversary does not interfere with the training phase of a classifier and perturbs only the testing samples [7]. The adversary can devise white-box attacks if it possesses total knowledge about the classifier architecture. Otherwise, it can devise black-box attacks relying on the transferability [7,8]. We shall focus here on white-box evasion attacks.

We introduce some notations and definitions used in this paper. Let $(\mathcal{X}, d, \mu)$ denote the sample set $\mathcal{X}$ with a metric d and a probability measure μ. The notation $x \leftarrow \mu$ denotes that a sample x is drawn with a probability measure μ. $\mathcal{L}$ denotes the countable label set. For a subset $\mathcal{S} \subseteq \mathcal{X}$, we let $d(x, \mathcal{S}) = \inf\{d(x, y)|y \in \mathcal{S}\}$ and let $B_\epsilon(x) = \{x'|d(x, x') \leqslant \epsilon\}$ be the ϵ-neighborhood of x, where d is the metric on $\mathcal{X}$. We also let $\mathcal{S}_\epsilon = \{x|d(x, \mathcal{S}) \leqslant \epsilon\}$ be the ϵ-expansion of $\mathcal{S}$. h is a hypothesis or a trained classifier that maps each $x \in \mathcal{X}$ to a predicted label $l \in \mathcal{L}$. c is the ground-truth function that maps each $x \in \mathcal{X}$ to a correct label $l \in \mathcal{L}$. h^l denotes the set of samples classified as label l, namely, $h^l = \{x \in \mathcal{X}|h(x) = l\}$. The error region $\mathcal{M}$ is the set of samples on which the hypothesis disagrees with the ground-truth, namely, $\mathcal{M} = \{x|h(x) \neq c(x)\}$. We define the risk as $R(h, c) = \Pr_{x\leftarrow\mu}[h(x) \neq c(x)] = \mu(\mathcal{M})$.

The two relevant types of evasion attacks studied here are based on the error region and the prediction change. In an error-region attack, the ground-truth function c is accessible and an attack occurs when a perturbation in the sample causes h to disagree with c. In contrast, a prediction-change attack emphasizes the instability of h. Here, an attack occurs when a perturbation results in a different prediction by h, and c is irrelevant. The precise definitions of these two types of attacks are as follows.

Definition 1. The error-region adversarial risk under ϵ-perturbation is the probability of drawing a sample such that its ϵ-neighborhood intersects with the error region,

$$R_\epsilon^{\mathrm{ER}}(h, c, \mu) = \Pr_{x\leftarrow\mu}[\exists x' \in B_\epsilon(x)|h(x') \neq c(x')].$$

Definition 2. The prediction-change adversarial risk under ϵ-perturbation is the probability of drawing a sample such that its ϵ-neighborhood contains a sample with a different label,

$$R_\epsilon^{\mathrm{PC}}(h, \mu) = \Pr_{x\leftarrow\mu}[\exists x' \in B_\epsilon(x)|h(x) \neq h(x')],$$

or equivalently,

$$R_\epsilon^{\mathrm{PC}}(h, \mu) = \Pr_{x\leftarrow\mu}\Big[\min_{x'\in\mathcal{X}}\{d(x', x)|h(x') \neq h(x)\} \leqslant \epsilon\Big].$$

In either type of attack, we refer to the nearest misclassified examples as the adversarial examples. We say that h is more robust if the induced risk of either type is lower for a certain ϵ-perturbation. We shall refer to the minimal ϵ-perturbation to x resulting in an adversarial example as the adversarial perturbation or the robustness of x with h. In contrast, we shall quantify the adversarial robustness of h as the size of ϵ necessary for the adversarial risk of h to be upper bounded by some constant. The main result of this paper is an upper bound on the adversarial robustness of any quantum classifier when the input states are smoothly generated from a Gaussian latent space.

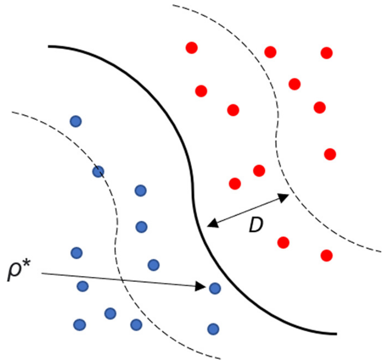

FIG. 1. The solid curve depicts the decision boundary of a quantum classifier. The states in blue are classified in a different class from the states in red. The metric is the trace distance. The trace distance between any pair of states generates an upper bound on the difference between their quantum classification confidences. Thus, ρ^*, the state closest to the decision boundary, is the ideal target of a prediction-change adversarial attack if the adversary aims to achieve misclassifications with minimal perturbations. On the other hand, if the adversary aims to maximize confidence change to any state with associated perturbations of size up to D, then all states between the dashed lines can be perturbed to be misclassified, while all other states can be perturbed to get closer to the boundary, resulting in overall decreased confidence in predicting the correct class. The concentration of measure phenomena implies that for a sufficiently large class, samples tend to lie near the decision boundary.

B. Quantum adversarial attacks

For this work, a quantum classifier is a quantum channel $\mathcal{E}$ that assigns labels l with some set of positive-operator-valued measures (POVMs) $\{\Pi_l\}$. The quantum classifier takes in an ensemble of identically prepared copies of a state and assigns the state a label l. The confidence of a prediction is quantified as the expectation value of the POVM for the prediction l, namely, $\mathrm{tr}(\mathcal{E}(\rho)\Pi_l)$ for an input density matrix ρ. We do not consider the number of copies of a state that is required to implement any specific quantum classification protocol. To measure the perturbation size, the natural choice of metric on quantum states—the trace distance—can be shown to generate an upper bound on the difference between their quantum classification confidence (see Appendix A), which implies that no small variation can induce a large swing in the predictive confidence. This property of the trace distance is a consequence of its interpretation as the achievable upper bound on the total variation distance [37] between probability distributions arising from measurements performed on those quantum states [38]. Furthermore, we show in Appendix A that the Hilbert-Schmidt norm, the Bures distance, and the Hellinger distance between two quantum states all generate a similar upper bound. As a result, in quantum adversarial attacks, the adversary either perturbs the states near the decision boundary minimally to seek misclassification, or aims to maximize confidence change to any state with associated perturbations that are upper bounded by some considerable size in these norms, as illustrated in Fig. 1. Our work analyzes primarily the risks due to the former objective. In Appendix B, we also propose a method for the latter objective exploiting the reversibility of parametrized quantum circuits (see, e.g., Refs. [39,40]). We note that the latter adversarial setting is justified, since to assess the security of a classifier under attack, it is reasonable—given a feasible space of modifications to the input data—to assume that the adversary aims to maximize the classifier's confidence in wrong predictions, rather than merely perturbing minimally in size [8].

There are two natural setups of adversarial attacks in QML that can be specified. The first is when the input data to the classifier is already quantized and any data transmitted through the quantum communication network comes from an untrusted party. In this case, the adversary, who may be the sender or an interceptor, can perform an attack either by perturbing each of the transmitted density matrices, or by intercepting a fraction of the copies of the state and substituting them entirely (see Appendix A). In a broader context, our analysis can be extended to include the instability of classifying quantum states subject to decoherence. We focus on this first setup in the current paper. The second setup is when the input to the quantum classifier is classical. The quantum classifier encodes the classical data before classifying. Since the adversary is perturbing the classical input data, it is effectively attacking classically. If one views such a quantum classifier as a black-boxed hypothesis function that maps each input to a class, any classifier-agnostic classical analysis of adversarial robustness can then be directly applied. For example, Ref. [10] analyzes the robustness of any classifier against random or semi-random perturbations, provided the curvature of the decision boundary is sufficiently small, while Ref. [15] analyzes the adversarial robustness of any classifier when classical input vectors are smoothly mapped from a Gaussian latent representation.

C. Quantum data encoding

We now explain the feature maps used throughout the paper. Considering a normalized positive vector $\vec{u}$ of length n, without loss of generality, we intuitively refer to it as a gray-scale image with n pixels throughout the paper. We focus on a particular set of encoding schemes where the normalized gray-scale value of each pixel, i.e., $u_i \in [0, 1]$, $i = 1, \dots, n$, is featurized into a qubit-encoding state $|\phi_i\rangle$. The product state $|\phi\rangle$ to be classified is a tensor product state of these qubit-encoded pixels in the 2^n-dimensional Hilbert space [41–44], namely,

$$|\phi\rangle = \bigotimes_{i=1}^{n} |\phi_i\rangle = \bigotimes_{i=1}^{n} \left[\cos\left(\frac{\pi}{2}u_i\right)|0\rangle + \sin\left(\frac{\pi}{2}u_i\right)|1\rangle\right]. \quad (1)$$

The qubit-encoding states, Eq. (1), do not require a quantum random access memory (QRAM) [45] and are efficient in time to prepare. Other schemes including amplitude encoding (see, e.g., Ref. [46]) are not considered here. We note that some of our results are general and independent of the encoding scheme. We further generalize Eq. (1) to qudits. In this case each pixel is mapped to a Hilbert space of higher dimension $d \geqslant 2$, with the coefficient of the jth component of the ith qudit state given by

$$|\phi_i\rangle_j = \sqrt{\binom{d-1}{j-1}} \cos^{d-j}\left(\frac{\pi}{2}u_i\right) \sin^{j-1}\left(\frac{\pi}{2}u_i\right). \quad (2)$$

These qudit states are special cases of what are known as spin-coherent states [41], and the qubit states in Eq. (1) correspond to $d = 2$.

D. Concentration of measure phenomenon

To describe this phenomenon, let $\Sigma \subseteq \mathcal{X}$ be a Borel set [47]. The concentration function, defined as

$$\alpha(\epsilon) = 1 - \inf_{\Sigma\subseteq\mathcal{X}} \left\{\mu(\Sigma_\epsilon)\middle|\mu(\Sigma) \geqslant \tfrac{1}{2}\right\}, \tag{3}$$

has a smaller value when more points are aggregated in the ϵ-expansion of a sufficiently large set Σ, for a fixed ϵ. Informally, a space $\mathcal{X}$ exhibits a concentration of measure if $\alpha(\epsilon)$ decays very fast as ϵ grows, and we shall refer to it as a concentrated space. This is true for a simple example—the standard Gaussian distribution $[\mathbb{R}, \ell^2, \mathcal{N}(0, 1)]$. Looking at the Borel set $\Sigma = (-\infty, 0)$ whose probability measure is $1/2$, the cumulative density outside its ϵ-expansion, namely, $\mathbb{R} \setminus \Sigma_\epsilon = (\epsilon, +\infty)$, decreases at least as fast as $\exp(-\epsilon^2/2)$ by the tail bound [48]. One can invoke isoperimetric inequality [49] to show that this clustering occurs around any Borel set with measure at least $1/2$ and applies to any canonical m-dimensional Gaussian measure in the Euclidean space (see Appendix G). More formally, a family of N-dimensional spaces with corresponding concentration functions $\alpha_N(\cdot)$ is called a (k_1, k_2)-normal Lévy family if $\alpha_N(\epsilon) \leqslant k_1 \exp(-k_2^2\epsilon^2 N)$, where k_1 and k_2 are particular constants. Consequently, the measure is more concentrated for a higher dimension. Two notable normal Lévy families are $\mathbb{SU}(N)$ and $\mathbb{SO}(N)$, both of which are equipped with the Hilbert-Schmidt norm L^2 and the Haar probability measure ν [50,51]. An implication of this phenomenon is that when points x are drawn from a highly concentrated space, for any function f varying not rapidly, we have $f(x) \approx \langle f\rangle$ with high probability. Lévy's Lemma [20,21] constitutes a specific example of this.

E. Related work

The work in Ref. [11] considered any normal Lévy family and derived the robustness for error-region adversarial attacks. The results show that for a nice classification problem [52], if $\mu(\mathcal{M}) = \Omega(1)$, the size of perturbations must be $O(1/\sqrt{N})$ to have the error-region adversarial risk upper bounded by some constant, where N is the dimension of the concentrated space. References [12,13] studied some specific concentrated spaces and revealed the same scaling.

Reference [16] transforms the classification of pure states $|\phi\rangle$ into that of unitaries U in $|\phi\rangle = U|\vec{0}\rangle$ for some fixed initial state $|\vec{0}\rangle$. These quantum classifiers then classify samples drawn from $\mathbb{SU}(N)$ equipped with the Haar probability measure ν and the Hilbert-Schmidt norm, which is a $(\sqrt{2}, 1/4)$-normal Lévy family. Therefore, if $\mu(\mathcal{M}) > 0$, the necessary condition on the perturbation size for the error-region adversarial risk to be bounded above by $1 - \gamma$ for some $\gamma \in [0, 1]$ is $O(1/\sqrt{N})$. Precisely, to have $R_\epsilon^{\mathrm{ER}}(h, c, \nu) \leqslant 1 - \gamma$, the ϵ-perturbation to any unitary must be upper bounded as [53]

$$\epsilon \leqslant \sqrt{\frac{4}{N}}\left[\sqrt{\ln\left(\frac{\sqrt{2}}{\mu(\mathcal{M})}\right)} + \sqrt{\ln\left(\frac{\sqrt{2}}{\gamma}\right)}\right]. \tag{4}$$

III. PROBLEMS WITH PRACTICAL CLASSIFICATIONS

The result in Eq. (4) claims that when classifying unitaries in $\mathbb{SU}(N)$ with the Haar measure, given that an adversary can devise white-box attacks and $\mu(\mathcal{M})$ is not exponentially suppressed by N, the robustness of any quantum classifier decreases polynomially in the dimension of the input N. This is daunting since the input has a dimension $N = d^n$ which is exponential in the number of qudits.

To apply any result related to Eq. (4), a ground-truth function c on $\mathbb{SU}(N)$ is needed to obtain the risk $\mu(\mathcal{M})$. However, c may not be easily defined in a real-world machine learning task. For instance, it is challenging to define what constitutes a mistake for visual object recognition. After adding a perturbation to an image, it likely no longer corresponds to a photograph of a real physical scene [54]. Furthermore, it is difficult to define the labels for images undergoing gradual semantic change. All of these factors complicate the evaluation of $\mu(\mathcal{M})$. It thus motivates us to focus on prediction-change adversarial risks (see, e.g., Refs. [10,13,54]) to avoid requiring access to the ground-truth. The following theorem and corollary then apply.

Theorem 1. Let $\mathbb{SU}(N)$ be equipped with the Haar measure ν and the Hilbert-Schmidt norm L^2. For any hypothesis $h : \mathbb{SU}(N) \to \mathcal{L}$ that is not a constant function, let $\eta \in [0, 1/2]$ determine the measure of the dominant class such that $\nu(h^l) \leqslant 1 - \eta, \forall l \in \mathcal{L}$. Suppose $U \in h^l$, $V \notin h^l$ and a perturbation $U \to V$ occurs, where $\|U - V\|_2 \leqslant \epsilon$. If the prediction-change adversarial risk $R_\epsilon^{\mathrm{PC}}(h, \nu) \leqslant 1 - \gamma$, then ϵ must satisfy

$$\epsilon \leqslant \sqrt{\frac{4}{N}}\left[\sqrt{\ln\left(\frac{2\sqrt{2}}{\eta}\right)} + \sqrt{\ln\left(\frac{2\sqrt{2}}{\gamma}\right)}\right]. \tag{5}$$

It is evident from Eq. (5) that the upper bound on the size of the perturbation ϵ is suppressed as the dimension N of the space increases. It is also suppressed when the measure of the dominant class $(1 - \eta)$ decreases and when the tolerance on the adversarial risk $(1 - \gamma)$ decreases.

Corollary 1. With $\rho = U|\vec{0}\rangle\langle\vec{0}|U^\dagger$ and $\sigma = V|\vec{0}\rangle\langle\vec{0}|V^\dagger$, Eq. (5) translates to a necessary upper bound in the trace norm between the pure-state density matrices

$$\|\rho - \sigma\|_1 \leqslant \frac{4}{N}\lambda_1,$$

where $N = d^n$, $\lambda_1 = [\ln(2\sqrt{2}/\eta)]^{1/2} + [\ln(2\sqrt{2}/\gamma)]^{1/2}$ with η and γ defined in Theorem 1, and the upper bound scales as $\Omega(d^{-n})$. With the qudit encoding in Eq. (2), a naive translation of this necessary upper bound to that in the ℓ^1 norm of the encoding vectors u and v gives,

$$\|u - v\|_1 \leqslant \frac{2n}{\pi}\cos^{-1}\left[\left(1 - \frac{2}{N}\lambda_1\right)^{\frac{1}{(d-1)n}}\right],$$

where the upper bound scales as $\Omega(d^{-n/2}\sqrt{n})$.

The proofs can be found in Appendices D and E. The interpretation of Theorem 1 and Corollary 1 is clear: Given that no class occupies Haar-measure 1, any quantum classifier on quantum states is more vulnerable to prediction-change adversarial attacks on higher-dimensional pure states drawn Haar-randomly, with the robustness decaying exponentially in the number of qudits.

In what follows, we apply this theorem to a practical task by presenting two perspectives on the application, to illustrate the limitations of the theorem. Suppose that the objective of the practical task is to classify a subset of quantum states, for example, the pure product states in Sec. II C that encode images displaying a digit 0 or 1. On one hand, if we label unitaries not related to an actual image, together with unitaries associated with noisy images not displaying a digit 0 or 1, in a third-class, this class will have measure 1, since the set of all unitaries that evolve the initial $|\vec{0}\rangle$ to some final pure product state $|\phi\rangle$ has Haar measure 0 in $\mathbb{SU}(N)$ [55]. For example when $n=1$, this can be seen by recognizing that the encoded states $\{|\phi\rangle\}$ correspond to only a fraction of the great circle passing through $|0\rangle$ and $|1\rangle$ on the Bloch Sphere. This labeling renders Theorem 1 useless for any h trained in this way because $\eta = 0$. On the other hand, if we train a binary h to classify half of $\mathbb{SU}(N)$, including unitaries corresponding to 0-digit images, to $l=0$, and the other half, including unitaries corresponding to 1-digit images, to $l=1$, then $\eta = 1/2$. Using Eq. (5) then gives $O(1/\sqrt{d^n})$ robustness against prediction-change adversarial attacks, again suggesting extreme vulnerabilities in high dimensions.

However, the interpretation of this result is not of practical interest, for the following reasons. We emphasize that in applying Theorem 1 or Eq. (4), the notion of adversarial risks by Definition 2 represents the probability of perturbing a Haar-randomly selected unitary by some ϵ to its adversarial example. It does not represent, for instance, the probability of perturbing a particular unitary associated with a real image to its adversarial example, nor does it represent the risk of attacking a unitary drawn from any other distribution over some subset. Therefore, if the task is to train and generalize a quantum classifier on a subset of quantum states with some distribution, this theorem cannot claim vulnerabilities that are exponential in the number of qudits. It is also noted that, as far as how Eq. (4) and Theorem 1 are formulated, the perturbed states cannot be mixed states since the latter are mapped from $|\vec{0}\rangle\langle\vec{0}|$ by a completely positive and tracing preserving (CPTP) maps rather than by unitaries. In Sec. IV, we shall see that this is an example of an *in-distribution* attack, which applies to scenarios where both the original and perturbed states are pure.

IV. CLASSIFICATIONS ON GENERATOR OUTPUT DISTRIBUTIONS

A. Concentration in generated distributions

In practice, one is interested in the performance of a classifier on a distribution over some subset of meaningful samples, such as the subset of images displaying digits including the MNIST data set [56]. It is this distribution on which the adversarial risk should be computed to infer the extent of the vulnerability. To ensure that the probability measure on the classifier-input space covers meaningful samples, we resort to approximating the distribution over meaningful samples using the image of a smooth generator function on a concentrated latent space, trained on samples of interest [15]. Following convention, we refer to the latter as a real-data manifold. Such a generator can be a Normalizing Flow model [25–27] or the generator of a Generative Adversarial Network (GAN) [28–30], both with a Gaussian latent space, trained on the same data set that the classifier will be trained on. A generative model serving this purpose is also referred to as a spanner [57]. In this way, a major fraction of the samples in the generator output $\mathcal{S}$ can be related to samples of interest, despite the fact that the smoothness of the generator may introduce some samples off the real-data manifold, such as those undergoing gradual semantic change during interpolations. This generative setup can be generalized to multiple generators on the same latent space. However, each generator maps to a disjoint part of the real-data manifold, overcoming the problem of covering the off real-data manifold when the latent space is globally connected [58]. This generalization requires relaxing the demand that $\omega(0)=0$ in Eq. (6) below. As a result, no data off the real-data manifold is generated in $\mathcal{S}$.

The reason that we require the latent space to be concentrated is so that we can study the concentration of samples in the generator-output space resulting from the concentration of the latent space. This connection is made by the assumption that the generator is smooth, in the sense that it admits a modulus of continuity (i.e., it is uniformly continuous), namely, if there exists a monotone invertible function $\omega(\cdot)$ such that

$$\|g(z)-g(z')\| \leqslant \omega(\|z-z'\|_2), \quad \forall z, z' \in \mathcal{Z}, \tag{6}$$

where $\|\cdot\|$ is the metric equipped by the image of g. This is a weaker condition than the Lipschitz continuity which results when $\omega(\cdot)$ is a linear function. In this paper, we assume $\omega(\cdot)$ to yield a tight upper bound in Eq. (6), and we demand $\omega(\tau)$ to be small for small τ for a smooth generator. The idea is that any tendency to concentration of measure in the latent space is preserved by such a smooth mapping to its image, and the generated samples then follow a modified concentrated distribution. We can imagine that if some pairs of latent variables from different classes are within distance b across the class boundary in the generator domain, their generator images must be accordingly within distance at most $\omega(b)$ across the boundary. This can also display a clustering. Although the tendency to cluster is preserved, the extent to which the points in the generator image gather is mediated by the modulus of continuity. A tight upper bound with $\omega(\cdot)$ that yields distances larger than the typical distances in the output space means that generated samples can be further apart, and vice versa. As far as adversarial robustness is concerned, a larger $\omega(\cdot)$ is then favorable since it implies that larger perturbations are needed to definitively perturb a larger number of generated samples across decision boundaries.

In generating these to-be-classified samples, the fact that a large probability density resides near the decision boundary is not at odds with a trained classifier that predicts training samples with high confidence. The training samples comprise only a subset of the support of the generator-output distribution. High confidence training samples result from the classifier drawing the decision boundaries away from them. When such a decision boundary encloses a sufficiently large measure, it then inevitably encounters large probability densities—as dictated by the concentration of measure phenomenon on these distributions—that do not contribute to training. As a result, when generalizing to test samples that are similar to the training samples, some test samples may locate near the boundary and constitute vulnerable targets to adversarial attacks.

B. Robustness of QML models

We consider the quantum adversarial attack setup where the input to the classifier is already quantized and transmitted through a quantum communication network.

Let our latent space $\mathcal{Z}$ be, for example, $\mathbb{R}^m$ with the Euclidean metric ℓ^2 and the canonical m-dimensional Gaussian measure $\mathcal{N}_m \equiv \mathcal{N}(0, I_m)$. So this is a concentrated space. Let $z \leftarrow \mathcal{N}_m$ in $\mathcal{Z}$. Suppose that a smooth generator $g : \mathcal{Z} \to \mathcal{S} \subseteq \mathcal{X}$ is trained to generate a distribution ξ of concern, such as some distribution of natural images, on a subset $\mathcal{S}$ of $\mathcal{X}$. For a sample $g(z) \in \mathcal{S}$, we then have $\xi[g(z)] = \mathcal{N}_m(z)$.

Incorporated in the generator $g = g_2 \circ g_1$, g_1 maps the latent space to a subset of n-pixel natural images, g_2 then encodes the natural image into a density matrix defined in Eq. (2). That is, $g(z) = |\phi(z)\rangle\langle\phi(z)| = \rho(z) \in \mathcal{S} \subseteq \mathcal{X}$, where $\mathcal{S}$—the image of g—is a subset of all density matrices $\mathcal{X}$. The metric on density matrices is the trace norm L^1 unless otherwise specified. The probability measure ξ, which is a distribution mapped by g from the m-dimensional Gaussian measure $\mathcal{N}_m$ on $\mathcal{Z}$, is only supported on $\mathcal{S}$ over density matrices capturing the natural image distribution. Any quantum classifier h then classifies the $d^n \times d^n$ density matrices in $(\mathcal{X}, L^1, \xi)$. Let us denote the intermediate stage—the set of images with n pixels (normalized vectors with length n)—as $\mathcal{I}$, then the corresponding measure on $\mathcal{I}$ can be denoted as $\xi \circ g_2$. The metric on $\mathcal{I}$ is, for instance, the ℓ^1 norm. Diagrammatically, these mappings are

$$\underbrace{\mathcal{Z} \xrightarrow{g_1} \mathcal{I} \xrightarrow{g_2} \mathcal{S}}_{g} \subseteq \mathcal{X} \xrightarrow[h]{} \mathcal{L}.$$

It is noted that smoothness is a desirable property of generative models. It is hinted at by gradual transitions in the features in the generated samples, which imply that the generator has learned relevant factors of variation [59]. We are then justified in assuming that the real-data manifold on $\mathcal{I}$ can be covered by a smooth generator g_1 (see, e.g., Refs. [26–30]). In what follows, we show that the overall generator g, mapping from $\mathcal{Z}$ to the real-data manifold in the set of density matrices $\mathcal{X}$, is also smooth.

Proposition 1. Assuming that $g_1 : \mathcal{Z} \to \mathcal{I}$ is smooth with a modulus of continuity $\omega_1(\cdot)$ and the qudit encoding scheme, Eq. (2), is applied, then the generator $g = g_2 \circ g_1 : \mathcal{Z} \to \mathcal{S} \subseteq \mathcal{X}$ is also smooth and admits a modulus of continuity $\omega(\cdot)$ that is lower bounded as

$$\omega(\tau) \geqslant \sqrt{1 - \cos^{2n(d-1)}\left(\frac{\pi}{2n}\omega_1(\tau)\right)}, \quad \forall \tau \geqslant 0.$$

The proof can be found in Appendix F. In terms of the scaling with respect to n and d, when $\omega_1(\cdot)$ scales as $\Omega(1)$, for instance, when g_1 is Lipschitz continuous (e.g., the generator in Refs. [60,61]), Proposition 1 implies that the modulus of continuity of the overall generator g, i.e., $\omega(\cdot)$, scales as $\Omega(\sqrt{d/n})$. It is desirable to enforce Lipschitz continuity on some generators, for example when imposing spectral normalization [62] on the generator of a GAN to improve training [61].

A distinction can be made concerning whether the adversarial example σ must be also in the subset $\mathcal{S}$. If so, then the adversarial attack is called in-distribution, since the attacker only looks for an adversarial example within the data manifold $\mathcal{S}$. Otherwise, we call it an unconstrained adversarial attack since the perturbation is arbitrary in $\mathcal{X}$, i.e., it is not confined to the data manifold. We state the precise definitions, based on prediction-change adversarial risks in Definition 2, as follows.

Definition 3. An in-distribution adversarial attack, or a data-manifold attack, attempts to find the perturbation

$$\begin{aligned}\varepsilon_{\text{in}}(\rho) &= \min_{r \in \mathcal{Z}}\{\|g(z+r) - \rho\|_1 | h(g(z+r)) \neq h(\rho)\} \\ &= \min_{\sigma \in \mathcal{S}}\{\|\sigma - \rho\|_1 | h(\sigma) \neq h(\rho)\},\end{aligned}$$

which is within the data manifold $(\mathcal{S}, L^1, \xi)$. It induces an in-distribution adversarial risk,

$$R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) = \Pr_{\rho \leftarrow \xi}[\varepsilon_{\text{in}}(\rho) \leqslant \epsilon_{\text{in}}].$$

Definition 4. An unconstrained adversarial attack attempts to find

$$\varepsilon_{\text{unc}}(\rho) = \min_{\sigma \in \mathcal{X}}\{\|\sigma - \rho\|_1 | h(\sigma) \neq h(\rho)\},$$

which is in $(\mathcal{X}, L^1)$ not restricted to the data manifold $\mathcal{S}$. It induces an unconstrained adversarial risk,

$$R^{\text{PC}}_{\epsilon_{\text{unc}}}(h, \xi) \equiv R^{\text{PC}}_{\epsilon}(h, \xi) = \Pr_{\rho \leftarrow \xi}[\varepsilon_{\text{unc}}(\rho) \leqslant \epsilon].$$

It is noted that when the generator is surjective on $\mathcal{X}$, i.e., $\mathcal{S} = \mathcal{X}$, there is no distinction between the two types of attacks. The setups in Theorem 1 and Eq. (4) consider classifying on the subset of all pure-state density matrices in $\mathcal{X}$ on which a Haar-random distribution ν is supported. Since this requires both the original and perturbed states be pure, the adversarial risks are considered in-distribution, although we shall see in Sec. IV B that the same upper bound applies to the unconstrained robustness for a general quantum classifier.

In-distribution Adversarial Robustness

The following theorem, depending on the distribution to be classified as well as the specific classical-data generator g_1 in terms of $\omega_1(\cdot)$, then applies.

Theorem 2. Let $h : \mathcal{X} \to \mathcal{L}$ be any quantum classifier on the set of density matrices. Considering in-distribution adversarial attacks on the image of g, if $\xi(h^l) \leqslant 1/2, \forall l$, i.e., the classes are not too unbalanced, then for the prediction-change risk $R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) \leqslant 1 - \gamma$, the distance between two density matrices ϵ_{in} must satisfy

$$\epsilon_{\text{in}} \leqslant \omega\left[\sqrt{\ln\left(\frac{\pi}{2\gamma^2}\right)}\right], \tag{7}$$

where $\omega(\cdot)$ is the modulus of continuity in Proposition 1.

The proof can be found in Appendix G. This result is independent of the quantum data encoding scheme. It can be generalized to quantum classifiers with arbitrary decision boundaries, but in this case, the necessary upper bound on the in-distribution robustness will not have a closed-form (see Appendix G). This upper bound is saturated when Eq. (6) is tight and the quantum classifier induces linearly separable regions in the latent space, namely, when $h \circ g$ is a linear function on $\mathcal{Z}$, giving rise to the maximally robust quantum classifier. The nonsaturation of this upper bound when class regions are not linearly separable in $\mathcal{Z}$ can be seen in the example of the standard Gaussian in Sec. II D above. Suppose one looks at $\Sigma' = (-\infty, -2\delta) \cup (0, 2\delta)$ for some $\delta > 0$, which has the same probability measure 1/2 as $\Sigma = (-\infty, 0)$ but is not linearly separable in $\mathbb{R}$. The measure outside the δ-expansion of Σ', i.e., $\mathbb{R} \setminus \Sigma'_\delta = (3\delta, +\infty)$, is smaller than that outside of the δ-expansion of Σ, namely, $\mathbb{R} \setminus \Sigma_\delta = (\delta, +\infty)$, implying more concentration outside and near Σ' than Σ.

The nonsaturation of this upper bound for nonlinearly separable classification regions in $\mathcal{Z}$ also implies that it is prone to misclassification with an increasing number of equiprobable classes. The proof for cases with at least five equiprobable classes can be found in Appendix G. Informally, more equiprobable classes lead to more boundaries, enclosing classes with sufficiently large total measure, that border distinct classes. Then within a fixed distance beyond more of those boundaries, there are more samples subject to some prediction change.

We note that this upper bound is usually not saturated in practice, since a quantum classifier is usually linear, such as a parametrized quantum circuit and a unitary tensor network, while the generator g is usually nonlinear, given that g_1 is usually nonlinear and g_2, the quantum feature map, is nonlinear. Classically, some highly nonlinear state-of-the-art neural networks have robustness one or two orders of magnitude smaller in the ℓ^2 norm on some data sets than the corresponding upper bound derived with similar arguments [15]. It would be interesting to examine the amount of deviation from the upper bound for QML models in future works.

Theorem 2 implies that when the quantum states to be classified encode classical data generated with a modulus of continuity scaling as $\Omega(1)$, the in-distribution robustness of any quantum classifier decreases polynomially in the number of qudits n and increases polynomially in the qudit dimension d. To see this, we first note that according to Proposition 1, when $\omega_1(\cdot) = \Omega(1)$, which applies to generators such as those enforcing Lipschitz continuity, $\omega(\cdot)$ is lower bounded by a function that scales as $\Omega(\sqrt{d/n})$. This means that the upper bound on the perturbation size ϵ_{in} between any two in-distribution states, i.e., the right hand side of Eq. (7), is then asymptotically bounded from below by $\sqrt{d/n}$.

As such, the vulnerability increases slightly with a larger number of qudits n and by contrast, decreases slightly with qudits of higher dimension $d \geqslant 2$. When the encoded classical data manifold comes from generators for which Lipschitz continuity is not enforced, it requires numerical approximations of the modulus of continuity $\omega_1(\cdot)$ to determine its scaling in the output space, before obtaining the robustness scaling. Compared to Theorem 1 where samples are Haar-random pure states, the states to be classified here, which characterise the adversarial risk, are similar to those considered in practical tasks. Specifically, they are a subset of encoded states with a distribution smoothly generated from a Gaussian latent space. Theorem 2 demonstrates that, contrary to previous claims [16], there is no guarantee that quantum classifiers are exponentially more vulnerable to in-distribution attacks in higher-dimensional Hilbert space. We shall now show that the theorem applies to unconstrained attacks as well.

Unconstrained Adversarial Robustness

Unconstrained adversarial attacks are arbitrary perturbations in $\mathcal{X}$ to a sample ρ. In a broader context in which the instability of the quantum classifier is concerned, this may derive from density matrices subject to decoherence in a classification task. It is clear that $\varepsilon_{\text{unc}}(\rho) \leqslant \varepsilon_{\text{in}}(\rho), \forall \rho \in \mathcal{X}$ and thus by changing the in-distribution perturbations in Theorem 2 to unconstrained ones the same bound in Eq. (7) applies.

We argue that there does not exist a tighter upper bound that holds for general quantum classifiers for unconstrained robustness. Consider a particular family of quantum classifiers that project any state onto the data manifold, namely, to map any state to its closest in-distribution state, before classifying. These classifiers can be shown to satisfy $1/2\varepsilon_{\text{in}}(\rho) \leqslant \varepsilon_{\text{unc}}(\rho) \leqslant \varepsilon_{\text{in}}(\rho), \forall \rho \in \mathcal{X}$ [63]. Even in the worst case where $\varepsilon_{\text{unc}}(\rho) = 1/2\varepsilon_{\text{in}}(\rho), \forall \rho \in \mathcal{X}$, their unconstrained robustness is as large as half of the in-distribution one. We stress that, although robust, such a quantum classifier is inefficient in our setting since there is no apparent tractable way to obtain the closest pure product state to an arbitrary state.

Inspired by this strategy, we propose that one can construct a family of efficient quantum classifiers $\tilde{h}$ on n-qubit density matrices $\mathcal{X}$ with unconstrained robustness $\varepsilon_{\text{unc}}(\rho)$ lower bounded for any $\rho \in \mathcal{X}$. To be specific, we construct $\tilde{h}$ from any h with the following procedure.

Let the original sample $\rho \in \mathcal{S}$ be a pure product-state density matrix with n qudits as in Eq. (1). A perturbation $\epsilon_{\text{unc}} \equiv \epsilon$ leads to a sample $\sigma \in \mathcal{X}$. First, we perform single qubit tomography on every qubit of σ to reconstruct a product-state density matrix from these single qubits. We denote this mapping as $P : \mathcal{X} \to \mathcal{X},\ \sigma \mapsto \bigotimes_{i=1}^{n} \text{tr}_{\{j\neq i\}}(\sigma)$, where $\text{tr}_{\{j\neq i\}}(\sigma)$ means tracing out all but the ith qubit of σ. Subsequently, we numerically fit the pixel values $\{u_i\}$ to $P(\sigma)$ to find its closest density matrix $\tilde{\sigma}$ within our data manifold $\mathcal{S}$. We use a symbol $\tilde{\sigma}$ to represent the density matrix attained from this procedure. $\tilde{\sigma}$ is then replacing σ when fed to the quantum classifier h. We now have the following theorem.

Theorem 3. For every n-qubit $\rho \in \mathcal{S} \subseteq \mathcal{X}$, let $\tilde{\rho}$ be the density matrix within the data manifold attained from the above procedure. For any quantum classifier h, let $\tilde{h} : \mathcal{X} \to \mathcal{L}$ be such that $\tilde{h}(\rho) = h(\tilde{\rho})$. Then

$$2 - 2\left(1 - \frac{\varepsilon_{\text{in}}(\rho)^2}{16}\right)^{\frac{1}{n_e}} \leqslant \varepsilon_{\text{unc}}(\rho) \leqslant \varepsilon_{\text{in}}(\rho),$$

TABLE I. Summary of the adversarial robustness, namely, the size of perturbations necessary for the adversarial risk to be upper bounded by some constant, of any quantum classifier obtained within the prediction-change adversarial attack setting. In this setting, the prediction-change adversarial risk over the Haar-random distribution ν, $R_\epsilon^{\mathrm{PC}}(h, \nu)$, and over a smoothly generated distribution ξ, $R_\epsilon^{\mathrm{PC}}(h, \xi)$, are both upper bounded by $(1-\gamma)$ (column 0). d denotes the qudit dimension in Eq. (2) and n denotes the number of encoded qudits or the length of the encoding vectors (number of pixels in the image classification example). Parameters λ_1 and λ_2 are defined as $\lambda_1 = [\ln(2\sqrt{2}/\eta)]^{1/2} + [\ln(2\sqrt{2}/\gamma)]^{1/2}$ and $\lambda_2 = \sqrt{\ln[\pi/(2\gamma^2)]}$. Row 1 summarizes the adversarial robustness when a pure state ρ sampled from the Haar-random distribution ν is perturbed to a state σ. The robustness is shown both in the trace norm (column 1), as well as in its translation to the robustness measured in the ℓ^1 norm of the set of encoding vectors (from Corollary 1 of Theorem 1) (column 2). Both upper bounds decrease exponentially in n. Row 2 summarizes the adversarial robustness when a pure state ρ sampled from a smoothly generated distribution ξ from a Gaussian latent space is perturbed to a state σ (column 1), and the robustness when the intermediate generated vector $\vec{u}$ is perturbed to $\vec{v}$ (column 2) (from Proposition 1 and Theorem 2). Note that when the robustness in adversarially perturbing a vector scales as $\Omega(1)$, e.g., when the intermediate vectors are generated Lipschitz continuously, the robustness in perturbing an encoded pure state scales as $\Omega(\sqrt{d/n})$.

$\leqslant 1-\gamma$	$\\|\rho - \sigma\\|_1 \leqslant$	$\\|\vec{u} - \vec{v}\\|_1 \leqslant$
$R_\epsilon^{\mathrm{PC}}(h, \nu)$	$4d^{-n}\lambda_1 = \Omega(d^{-n})$	$\frac{2n}{\pi}\cos^{-1}[(1-2d^{-n}\lambda_1)^{\frac{1}{(d-1)n}}] = \Omega(d^{-\frac{n}{2}}\sqrt{n})$
$R_\epsilon^{\mathrm{PC}}(h, \xi)$	$\omega(\lambda_2) \geqslant \sqrt{1-\cos^{2n(d-1)}\left[\frac{\pi}{2n}\omega_1(\lambda_2)\right]} = \Omega\left(\sqrt{\frac{d}{n}}\right)$	$\omega_1(\lambda_2) = \Omega(1)$

where $n_e = n$ for even n and $n_e = n + 1$ for odd n.

The proof can be found in Appendix H. We note that the procedure can be applied to any product state encoding scheme. This procedure yields an explicit lower bound to the unconstrained adversarial perturbation when it is possible to estimate the in-distribution adversarial perturbation by, for example, sampling in the latent space [64] or gradient descent search in the latent space [57] before mapping to the density matrices. This $\tilde{h}$ constructed from h amounts to a feasible tomographic preprocessing of input states. It guarantees that the unconstrained robustness of each sample ρ is bounded from below and may be used as a defense strategy against unconstrained adversarial attacks in practice. However, we note that when n is large, this lower bound can be several orders of magnitude smaller than the upper bound.

V. DISCUSSION

A summary of the upper bounds on the prediction-change adversarial robustness over pure states sampled from the Haar-random distribution ν and a smoothly generated distribution ξ, is presented in Table I.

In this work, we first showed the prediction-change adversarial robustness over Haar-randomly distributed pure states, and compared this with the previously demonstrated error-region robustness of Ref. [16] over the same distribution. Both types of adversarial robustness show similar extreme vulnerabilities exponential in the number of qudits. However, in this work we have argued that these vulnerabilities for Haar-random pure states are not of practical interest. This is because, in practice, the adversarial risk of a quantum classifier should be computed on a distribution over some subset of meaningful states, such as a subset of qubit encoding states featurizing some images, to infer the extent of the vulnerability. In general, practical quantum classification tasks classify a subset of encoded states with some commonly used qubit encoding scheme. For such tasks, we have shown that we can use the concentration of measure phenomenon to derive the robustness of any quantum classifiers in situations where the distribution of states to be classified can be smoothly generated from a Gaussian latent space, as quantified in Eq. (6). In this situation, we have shown that one finds only a mildly polynomially decreasing robustness in the number of such encoded qubits, specifically with scaling as $O(\sqrt{1/n})$ in the trace norm.

As noted for Theorem 2, it is the upper bound on the perturbation size necessary for the adversarial risk to be bounded from above that scales as $\Omega(\sqrt{1/n})$. This upper bound is usually not tight and the actual adversarial robustness could therefore be smaller. We have also proposed a feasible modification of any quantum classifier with product-state inputs—namely, by performing single qubit tomography before numerically fitting the closest encoded qubit state—to obtain a lower bound on the unconstrained robustness and to defend against unconstrained adversarial attacks.

Most importantly, our analysis provides QML protocols some relief from adversarial attacks in real-world tasks. For example, when classifying on some qubit states encoding MNIST images, the robustness decreases only as $O(\sqrt{1/n})$, in contrast to the extreme vulnerability of quantum classifiers in classifying Haar-random pure states (Theorem 1 and Ref. [16]). In future, it will be interesting to experimentally compare the adversarial robustness of particular QML models for real-world data on a distribution of states smoothly mapped from a Gaussian latent space with the bounds that we have derived here.

We note that the polynomially decreasing robustness in n is derived from the qudit encoding scheme. The concentration of measure due to the Gaussian isoperimetric inequality for the latent space only contributes to the argument of Eq. (7). It will be interesting to investigate whether a different encoding scheme can give better scaling in the robustness, and also to determine whether quantum data that derives naturally from a distribution other than the Haar-random distribution is robust to attacks. In Appendix B, we propose a method to perform white-box adversarial attacks on classically intractable input states with QML models. It will be interesting to further explore white-box attacks, assuming that the adversary is capable of devising these. In practice, with current NISQ-era hardware, it will also be useful to examine how robust QML models are against adversarial attacks under noise and decoherence.

ACKNOWLEDGMENTS

H.L. was supported by the National Aeronautics and Space Administration under Grant/Contract/Agreement No. 80NSSC19K1123 issued through the Aeronautics Research Mission Directorate. I.C. was supported by the US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Quantum Algorithm Teams Program, under Contract No. DE-AC02-05CH11231. W.J.H. was supported by Siemens #FutureMakers Fellowship 045695.

APPENDIX A: CONFIDENCE DIFFERENCE AND DISTANCE BETWEEN STATES

We show that the predictive confidence difference in any QML protocol is upper bounded by the distance between the input density matrices up to some constant factor, where this distance is measured in the trace norm L^1, the Hilbert-Schmidt norm L^2, the Bures distance, or the Hellinger distance.

Considering density matrices ρ and σ, the trace norm between them is defined to be $\|\rho-\sigma\|_1 = \mathrm{tr}(|\rho-\sigma|)$. Consider a set of POVMs $\{\Pi_l\}$ and a quantum channel $\mathcal{E}$ such that $\mathcal{E}(\rho)=\sum_i M_i\rho M_i^\dagger$ and $\sum_i M_i^\dagger M_i = I$. We have

$$\begin{aligned}\mathrm{tr}(\mathcal{E}(\rho)\Pi_l)-\mathrm{tr}(\mathcal{E}(\sigma)\Pi_l) &= \mathrm{tr}\left[\sum_i M_i(\rho-\sigma)M_i^\dagger\Pi_l\right]\\ &= \mathrm{tr}\left[(\rho-\sigma)\sum_i M_i^\dagger\Pi_l M_i\right]\\ &\equiv \mathrm{tr}[(\rho-\sigma)\mathcal{E}^*(\Pi_l)].\end{aligned}$$

We note that $\mathcal{E}^*$ is the dual map of $\mathcal{E}$ and $\{\mathcal{E}^*(\Pi_l)\}$ is still a set of POVMs, since $\mathcal{E}^*(\Pi_l)$ is hermitian, nonnegative because $\mathrm{tr}[\rho\mathcal{E}^*(\Pi_l)] = \mathrm{tr}[\mathcal{E}(\rho)\Pi_l] \geqslant 0$, and complete because $\sum_{i,l} M_i^\dagger\Pi_l M_i = \sum_i M_i^\dagger M_i = I$.

For each particular measurement, we can expand in its eigenbasis $\mathcal{E}^*(\Pi_l)=\sum_k b_k|\phi_k\rangle\langle\phi_k| \equiv \sum_k b_k P_k$. Let $\{|\psi_i\rangle\}$ and $\{\lambda_i\}$ be the eigenbasis and eigenvalues of $(\rho-\sigma)$, so $\|\rho-\sigma\|_1=\sum_i|\lambda_i|\in[0,2]$. We then expand $\mathcal{E}^*(\Pi_l)=\sum_{i,j,k} b_k a_{ik}|\psi_i\rangle a_{jk}^*\langle\psi_j|$ such that $\sum_i |a_{ik}|^2=1, \forall k$ and $\sum_k b_k = \mathrm{tr}[\mathcal{E}^*(\Pi_l)] \geqslant 0$ due to the nonnegativity. We have

$$\begin{aligned}\mathrm{tr}[(\rho-\sigma)\mathcal{E}^*(\Pi_l)] &= \mathrm{tr}\left[(\rho-\sigma)\sum_{i,j,k} b_k a_{ik}|\psi_i\rangle a_{jk}^*\langle\psi_j|\right]\\ &= \sum_k b_k \mathrm{tr}\left[\sum_{i,j} a_{ik}a_{jk}^*\langle\psi_j|(\rho-\sigma)|\psi_i\rangle\right]\\ &= \sum_{i,k} b_k|a_{ik}|^2\lambda_i \leqslant \sum_k b_k\|\rho-\sigma\|_1\\ &= \mathrm{tr}[\mathcal{E}^*(\Pi_l)]\|\rho-\sigma\|_1. \qquad \text{(A1)}\end{aligned}$$

Therefore,

$$|\mathrm{tr}[\mathcal{E}(\rho)\Pi_l]-\mathrm{tr}[\mathcal{E}(\sigma)\Pi_l]| \leqslant \mathrm{tr}[\mathcal{E}^*(\Pi_l)]\|\rho-\sigma\|_1.$$

When $\mathrm{tr}[\mathcal{E}^*(\Pi_l)]$ is not too large the above inequality suggests that the confidence change will be small when the trace norm between the two density matrices is small. However, $\mathrm{tr}[\mathcal{E}^*(\Pi_l)]$ may be very large in high dimensions and in that case, the upper bound becomes very weak.

We resort instead to the physical interpretation of trace distance being a generalization of the classical total variation distance. The trace distance between two quantum states is an achievable upper bound on the total variation distance between probability distributions arising from measurements performed on those states [38]:

$$\frac{1}{2}\|\rho-\sigma\|_1 = \frac{1}{2}\max_{\{\Pi_l\}}\sum_l|\mathrm{tr}[(\rho-\sigma)\Pi_l]|,$$

where the maximization is over all POVMs $\{\Pi_l\}$ and the factor of 2 is to restrict the maximal trace distance to be 1. Using the contractive property of the trace norm under any CPTP map, we conclude that the trace norm constitutes an upper bound to the sum of confidence changes of all POVMs:

$$\sum_l|\mathrm{tr}[\mathcal{E}(\rho-\sigma)\Pi_l]| \leqslant \|\mathcal{E}(\rho)-\mathcal{E}(\sigma)\|_1 \leqslant \|\rho-\sigma\|_1. \qquad \text{(A2)}$$

Considering the Hilbert-Schmidt norm defined as $\|\rho-\sigma\|_2^2 = \mathrm{tr}[(\rho-\sigma)^2]$, if we regard $\|\rho-\sigma\|_2$ as the inner product of the two vectors $(1,1,\cdots,1)$ and $(|\lambda_0|,|\lambda_1|,\cdots,|\lambda_{N-1}|)$, then from the Cauchy-Schwarz inequality we find $\|\rho-\sigma\|_1 \leqslant \sqrt{N}\|\rho-\sigma\|_2$. But this bound is very weak in high dimensional Hilbert space. A better upper bound is given in Ref. [65] that $\|\rho-\sigma\|_1 \leqslant 2\sqrt{R}\|\rho-\sigma\|_2$, where $R=\mathrm{rank}(\rho)\mathrm{rank}(\sigma)/[\mathrm{rank}(\rho)+\mathrm{rank}(\sigma)]$. This implies that, even when one state is full rank, if the other state is low rank, then the Hilbert-Schmidt norm is of the same order of magnitude as the trace norm. This is the case when we consider any perturbation to an encoded pure state density matrix ρ whose rank is 1. Combined with Eq. (A2), we arrive at a similar upper bound,

$$\sum_l|\mathrm{tr}[\mathcal{E}(\rho)\Pi_l]-\mathrm{tr}[\mathcal{E}(\sigma)\Pi_l]| \leqslant 2\sqrt{R}\|\rho-\sigma\|_2.$$

Considering the Bures distance defined as $\|\rho-\sigma\|_{\mathrm{B}}^2 = 2[1-\sqrt{F(\rho,\sigma)}]$, it is an extension to mixed states of the Fubini-Study distance for pure states [66]. We have

$$\begin{aligned}\|\rho-\sigma\|_1 &\leqslant 2\sqrt{1-\left(1-\tfrac{1}{2}\|\rho-\sigma\|_{\mathrm{B}}^2\right)^2}\\ &= 2\sqrt{\|\rho-\sigma\|_{\mathrm{B}}^2-\tfrac{1}{4}\|\rho-\sigma\|_{\mathrm{B}}^4} \leqslant 2\|\rho-\sigma\|_{\mathrm{B}},\end{aligned}$$

where the first inequality is proven in Refs. [66,67] and saturated for pure states. Therefore, together with Eq. (A2), we conclude that

$$\sum_l|\mathrm{tr}[\mathcal{E}(\rho)\Pi_l]-\mathrm{tr}[\mathcal{E}(\sigma)\Pi_l]| \leqslant 2\|\rho-\sigma\|_{\mathrm{B}}. \qquad \text{(A3)}$$

Finally, considering the Hellinger distance defined as $\|\rho-\sigma\|_{\mathrm{H}}^2 = 2-2\mathrm{tr}(\sqrt{\rho}\sqrt{\sigma})$, it is shown that $\|\rho-\sigma\|_{\mathrm{B}} \leqslant \|\rho-\sigma\|_{\mathrm{H}}$ [66] and thus, the same upper bound applies by changing $\|\rho-\sigma\|_{\mathrm{B}}$ to $\|\rho-\sigma\|_{\mathrm{H}}$ in Eq. (A3).

In QML, if ρ and σ are close in these norms and are separated by any class boundary, say between class $l=s$ and class $l=t$, then $\mathrm{tr}[\mathcal{E}(\rho)\Pi_s] > \mathrm{tr}[\mathcal{E}(\sigma)\Pi_s]$, while $\mathrm{tr}[\mathcal{E}(\rho)\Pi_t] <$

$\mathrm{tr}[\mathcal{E}(\sigma)\Pi_t]$. This suggests that no small perturbation to density matrices in these norms can significantly change the measurement outcome and thus, alter the prediction, unless the original sample is near the boundary. In other words, viewing $\mathrm{tr}[\mathcal{E}(\rho)\Pi_s]$ as the confidence of predicting $l = s$, it implies that no small perturbations can result in a high-confidence sample in one class perturbed to a low-confidence sample in the same class, or a high-confidence sample in a different class.

APPENDIX B: ADVERSARIAL ATTACKS EXPLOITING QUANTUM CLASSIFIER REVERSIBILITY

We propose a method to perform adversarial attacks in our first setup in Sec. II B on quantized data. This method can be carried out on a quantum hardware when the computation is classically intractable. We assume a noiseless QML model for this analysis, so the quantum channel is unitary. Considering, for example, the unitary tree tensor network (TTN) in Ref. [39] among other types of parametrized unitary quantum circuits, the adversary can run it reversely starting from a density matrix with any designated wrong class label $l = t$ such that $\mathrm{tr}(\sigma'\Pi_t) = 1$ while $\mathrm{tr}(\sigma'\Pi_{l\neq t}) = 0$. Any qubit that is traced out in the forward direction is initialized to an arbitrary state and passes through the network in the reverse direction. The output of the reversal circuit is a set of density matrices $\{U^\dagger\sigma'U\} \equiv \{\sigma\}$ such that $\mathrm{tr}(U\sigma U^\dagger\Pi_t) = 1$ whereas $\mathrm{tr}(U\sigma U^\dagger\Pi_{l\neq t}) = 0$. Thus, this set of density matrices will be classified in the wrong class by the POVM Π_t with high-confidence. Suppose that the original samples are $\{\rho\}$ in the class $s \neq t$ and $\mathrm{tr}(U\rho U^\dagger\Pi_s) = 1/2 + \delta$ with some $\delta \in (0, 1/2]$. The adversary then replaces an ϵ-portion of the transmitted quantum states $\{\rho\}$ with the $\{\sigma\}$ to attack the receiver.

To achieve a prediction change, the adversary demands $\mathrm{tr}(U[(1-\epsilon)\rho + \epsilon\sigma]U^\dagger\Pi_s) < 1/2$. This requires

$$\epsilon > 1 - \frac{1}{1 + 2\delta}, \tag{B1}$$

which means that the portion of $\{\rho\}$ being substituted with $\{\sigma\}$ increases with higher-confidence of $\{\rho\}$. We note that this effectively creates a perturbation of size

$$\begin{aligned}\|\rho - [(1-\epsilon)\rho + \epsilon\sigma]\|_1 &\geqslant \epsilon \sum_l |\mathrm{tr}[U(\rho - \sigma)U^\dagger\Pi_l]| \\ &= \epsilon\left\{\sum_{l\neq t}\mathrm{tr}(U\rho U^\dagger\Pi_l) + [1 - \mathrm{tr}(U\rho U^\dagger\Pi_t)]\right\} \\ &= \epsilon[2 - 2\mathrm{tr}(U\rho U^\dagger\Pi_t)] \geqslant \epsilon(1 + 2\delta),\end{aligned}$$

where the first inequality follows from Eq. (A2). As a result, a misclassification by the attack demands a perturbation of size $\|\rho - [(1-\epsilon)\rho + \epsilon\sigma]\|_1 \geqslant 2\delta$, where we substituted in Eq. (B1).

APPENDIX C: PROOF OF EQ. (4)

We present a condensed proof based on the proof to Theorem 3.7 in Ref. [11]. Let $\epsilon_1 > \sqrt{1/(Nk_2)\ln[k_1/\mu(\mathcal{M})]}$ and $\epsilon_2 > \sqrt{1/(Nk_2)\ln(k_1/\gamma)}$. Then the concentration function satisfies $\alpha(\epsilon_1) < \mu(\mathcal{M})$ and $\alpha(\epsilon_2) < \gamma$. As such, by directly applying Part 2 of Theorem 3.2 in Ref. [11], we conclude $R_\epsilon^{\mathrm{ER}}(h, c, \nu) > 1 - \gamma$ for $\epsilon = \epsilon_1 + \epsilon_2$. It can be shown that $\mathbb{SU}(N)$ is a $(\sqrt{2}, 1/4)$-normal Lévy family and so $k_1 = \sqrt{2}$ and $k_2 = 1/4$ [16]. The contrapositive statement on $R_\epsilon^{\mathrm{ER}}(h, c, \nu) \leqslant 1 - \gamma$ then gives the necessary condition Eq. (4).

APPENDIX D: PROOF OF THEOREM 1

Proof. We let $\epsilon_1 > \sqrt{1/(Nk_2)\ln(2k_1/\eta)}$ and $\epsilon_2 > \sqrt{1/(Nk_2)\ln(2k_1/\gamma)}$, then the concentration function satisfies $\alpha(\epsilon_1) < \eta/2$ and $\alpha(\epsilon_2) < \gamma/2$. Therefore, by applying Part 1 of Theorem A.2 in Ref. [11], we conclude that for $\epsilon = \epsilon_1 + \epsilon_2$, $R_\epsilon^{\mathrm{PC}}(h, \nu) > 1 - \gamma$. For completeness, we present our explained version of the proof below.

Let $\epsilon = \epsilon_1 + \epsilon_2$. By assumption that $\nu(h^l) \leqslant 1 - \eta, \forall l \in \mathcal{L}$, it can be easily verified by contradiction that $\exists l_1 \in \mathcal{L}$ s.t. $\nu(h^{l_1}) \in (\eta/2, 1/2]$. Let $h^{l_1,C} = \mathcal{X} \setminus h^{l_1}$. On one hand, we know that $\nu(h^{l_1}) > \eta/2 > \alpha(\epsilon_1)$ where the last inequality is given by our assumption. We prove by contradiction that $\nu(h_{\epsilon_1}^{l_1}) > 1/2$. Suppose not, then we have for $\mathcal{S} = \mathcal{X} \setminus h_{\epsilon_1}^{l_1}$, $\nu(\mathcal{S}) = 1 - \nu(h_{\epsilon_1}^{l_1}) \geqslant 1/2$. Then by the definition of the concentration function in Eq. (3), $\nu(\mathcal{S}_{\epsilon_1}) \geqslant 1 - \alpha(\epsilon_1)$. Combining with what we obtained that $\nu(h^{l_1}) > \alpha(\epsilon_1)$, we have $\nu(\mathcal{S}_{\epsilon_1}) + \nu(h^{l_1}) > 1$. Thus, $\exists x \in \nu(\mathcal{S}_{\epsilon_1}) \cup \nu(h^{l_1})$. This implies $\exists y \in \mathcal{S}|d(y, x) \leqslant \epsilon_1$. But this same y must also be in $h_{\epsilon_1}^{l_1}$ since the same x is also in h^{l_1}. However, this raises a contradiction since $\mathcal{S}$ and $h_{\epsilon_1}^{l_1}$ are disjoint by definition, i.e., $\nexists y | y \in \mathcal{S}, y \in h_{\epsilon_1}^{l_1}$. Now, $\nu(h_{\epsilon_1}^{l_1}) > 1/2$ means, by the definition of the concentration function in Eq. (3), as well as the assumption that $\gamma/2 > \alpha(\epsilon_2)$, we have $\nu(h_\epsilon^{l_1}) \geqslant 1 - \alpha(\epsilon_2) > 1 - \gamma/2$.

On the other hand, knowing that $\nu(h^{l_1,C}) \geqslant 1/2$, we have that $\nu(h_{\epsilon_2}^{l_1,C}) > 1 - \gamma/2$ followed by simply replacing the $h_{\epsilon_1}^{l_1}$ in the previous sentence with $h^{l_1,C}$ since they both have measure at least 1/2. We then also have $\nu(h_\epsilon^{l_1,C}) > 1 - \gamma/2$. Hence, using the inequality $\mu(\cap_{i=1}^n A_i) \geqslant \sum_{i=1}^n \mu(A_i) - (n - 1)$, one can conclude that $\nu(h_\epsilon^{l_1} \cap h_\epsilon^{l_1,C}) > 1 - \gamma$ and so, by the prediction-change risk's definition, $R_\epsilon^{\mathrm{PC}}(h, \nu) \geqslant \nu(h_\epsilon^{l_1} \cap h_\epsilon^{l_1,C}) > 1 - \gamma$.

It can be shown that $\mathbb{SU}(N)$ is a $(\sqrt{2}, 1/4)$-normal Lévy family and so $k_1 = \sqrt{2}$ and $k_2 = 1/4$ [16]. The contrapositive statement on $R_\epsilon^{\mathrm{PC}}(h, \nu) \leqslant 1 - \gamma$ then gives the necessary condition Eq. (5).

APPENDIX E: PROOF OF COROLLARY 1

Proof. We have from Theorem 1 that the necessary condition for $R_\epsilon^{\mathrm{PC}}(h, \nu) \leqslant 1 - \gamma$ on $\mathbb{SU}(N)$ is $\|U - V\|_2 \leqslant \sqrt{4/N}\lambda_1$ where $\lambda_1 = [\ln(2\sqrt{2}/\eta)]^{1/2} + [\ln(2\sqrt{2}/\gamma)]^{1/2}$. Let $\sigma = V|\vec{0}\rangle\langle\vec{0}|V^\dagger$. From the Proof of Theorem 3 in Ref. [16], we have $\|U - V\|_2^2 \geqslant 2N(1 - |\langle\phi|\psi\rangle|)$. The Fuchs-van de Graaf inequality for pure states is

$$2 - 2\sqrt{F(\rho, \sigma)} \leqslant \|\rho - \sigma\|_1 = 2\sqrt{1 - F(\rho, \sigma)}, \tag{E1}$$

where the fidelity $F(\rho, \sigma) = |\langle\phi|\psi\rangle|^2$. Based on Eq. (E1), we obtain

$$2N(1 - |\langle\phi|\psi\rangle|) \geqslant \frac{2NT(\rho, \sigma)^2}{(1 + |\langle\phi|\psi\rangle|)} \geqslant NT(\rho, \sigma)^2,$$

where T is the trace distance. As such, we need

$$\sqrt{\frac{4}{N}\lambda_1} \geqslant \|U-V\|_2 \geqslant \sqrt{N}\,T(\rho,\sigma) = \frac{\sqrt{N}}{2}\|\rho-\sigma\|_1,$$

which gives $\|\rho-\sigma\|_1 \leqslant 4/N\lambda_1 = 4d^{-n}\lambda_1$.

We translate this upper bound on the distance between two density matrices to that between their encoding vectors $g_1(z)$ and $g_1(z')$. Altogether with the necessary condition and Eq. (E1), we have

$$4d^{-n}\lambda_1 \geqslant \|\rho-\sigma\|_1 \geqslant 2 - 2\sqrt{F(\rho,\sigma)}. \tag{E2}$$

For density matrices $\rho,\sigma\in\mathcal{X}$ respective to two images, we have $\rho = |\phi\rangle\langle\phi| = \bigotimes_i|\phi_i\rangle\bigotimes_i\langle\phi_i| = \bigotimes_i|\phi_i\rangle\langle\phi_i| = \bigotimes_i\rho_i$ and $\sigma = \bigotimes_i|\psi_i\rangle\langle\psi_i| = \bigotimes_i\sigma_i$, which are mapped from images $g_1(z) = \vec{s}$ and $g_1(z') = \vec{t}$, respectively. All i-indices run from 1 to n. And $|\phi_i\rangle$ and $|\psi_i\rangle$ are featurized from pixels of value s_i and t_i, respectively. It can be shown by induction that

$$F(\rho,\sigma) = \prod_i \cos^{2(d-1)}\left(|s_i-t_i|\frac{\pi}{2}\right). \tag{E3}$$

For $d=2$, we have that $F(\rho,\sigma) = \mathrm{tr}(\bigotimes_i\rho_i\bigotimes_i\sigma_i) = \prod_i\mathrm{tr}(\rho_i\sigma_i) = \prod_i|\langle\phi_i|\psi_i\rangle|^2 = \prod_i\cos^2(|s_i-t_i|\pi/2)$. It then suffices to show $\langle\phi_i|\psi_i\rangle = \cos^{d-1}(|s_i-t_i|\pi/2)$ for the qudit encoding $d>2$. We drop all $\pi/2$ factors and the subscripts i in s_i and t_i hereafter. Suppose for $d=k$, we have $\langle\phi_i|\psi_i\rangle$ equal to

$$\sum_{j=1}^{k}\binom{k-1}{j-1}\cos^{k-j}(s)\cos^{k-j}(t)\sin^{j-1}(s)\sin^{j-1}(t)$$
$$= \cos^{k-1}(s-t). \tag{E4}$$

Then for $d=k+1$, we have $\langle\phi_i|\psi_i\rangle$ equal to

$$\sum_{j=1}^{k+1}\binom{k}{j-1}\cos^{k+1-j}(s)\cos^{k+1-j}(t)\sin^{j-1}(s)\sin^{j-1}(t)$$
$$= \cos(s)\cos(t)\left[\sum_{j=1}^{k}\beta\binom{k}{j-1}\cos^{k-j}(s)\cos^{k-j}(t)\right.$$
$$\left.\times\,\sin^{j-1}(s)\sin^{j-1}(t)\right] + \sin(s)\sin(t)\left[\sum_{j=2}^{k+1}(1-\beta)\right.$$
$$\left.\times\binom{k}{j-1}\cos^{k+1-j}(s)\cos^{k+1-j}(t)\sin^{j-2}(s)\sin^{j-2}(t)\right], \tag{E5}$$

where $\beta = (k+1-j)/k$.

Identifying the two expressions in the square brackets as both equal to Eq. (E4), we obtain the desired outcome $\langle\phi_i|\psi_i\rangle = \cos^k(s-t)$, and the induction completes.

Combining Eqs. (E2) and (E3), we have

$$4d^{-n}\lambda_1 \geqslant 2 - 2\prod_i\cos^{d-1}\left(|s_i-t_i|\frac{\pi}{2}\right)$$
$$\geqslant 2 - 2\cos^{(d-1)n}\left(\frac{\sum_i|s_i-t_i|}{n}\frac{\pi}{2}\right), \tag{E6}$$

where the last inequality follows from the inequality $\cos^n(\sum_i x_i/n) \geqslant \prod_i\cos(x_i)$. It can be readily shown for $n\geqslant 2$ using the following trick. Consider any pair x_i and x_j and let x_m be their arithmetic average so $x_i = x_m + d$ and $x_j = x_m - d$ for some $d\neq 0$. Then $\cos(x_i)\cos(x_j) = \cos(x_m + d)\cos(x_m - d) = \cos^2(x_m) - \sin^2(d) \leqslant \cos^2(x_m)$. Therefore, one can maximize the overall cosine product, while maintaining the sum of the arguments, by replacing any pair $\cos(x_i)$ and $\cos(x_j)$ with $\cos(x_m)$ and $\cos(x_m)$, and successively replacing every pair till every factor converges to $\cos(\sum_i x_i/n)$ with the same argument.

Solving for $\sum_i|s_i-t_i| = \|g_1(z)-g_1(z')\|_1$ in Eq. (E6) yields the upper bound on the perturbation size in $(\mathcal{I},\ell^1)$.

APPENDIX F: PROOF OF PROPOSITION 1

Proof. We decompose g into $g_2\circ g_1$ where $g_1:(\mathcal{Z},\ell^2)\to(\mathcal{I},\ell^1)$ is desired to be smooth in practice. It can be generalized to ℓ^p norm on $\mathcal{I}$ and similar proof follows since the ℓ^p norm of any given vector does not grow with p. We have $\|g_1(z)-g_1(z')\|_1 \leqslant \omega_1(\|z-z'\|_2), \forall z,z'\in\mathcal{Z}$.

We show that it is also smooth for the qudit encoding $g_2:(\mathcal{I},\ell^1)\to(\mathcal{X},L^1)$ as in Eq. (2). Applying the qudit feature map and similar to that in Appendix E, it can be shown that

$$\|\rho-\sigma\|_1 = 2\sqrt{1-\prod_i\cos^{2(d-1)}\left(|s_i-t_i|\frac{\pi}{2}\right)}. \tag{F1}$$

Since $\omega(\cdot)$ is used in an upper bound in Theorem 2, we need to obtain the scaling of a lower bound to $\omega(\cdot)$. The $\omega(\cdot)$ that forms a tight upper bound in Eq. (6) must have $\omega(\|z-z'\|_2)$ upper bounding Eq. (F1) for arbitrary $z,z'\in\mathcal{Z}$. Hence, it is equivalent to find the scaling of a lower bound to Eq. (F1). That is, we have $\forall z,z'\in\mathcal{Z}$,

$$\omega(\|z-z'\|_2) \geqslant 2\sqrt{1-\prod_i\cos^{2(d-1)}\left(|s_i-t_i|\frac{\pi}{2}\right)}$$
$$\geqslant 2\sqrt{1-\cos^{2(d-1)n}\left(\frac{\sum_i|s_i-t_i|}{n}\frac{\pi}{2}\right)}$$
$$= 2\sqrt{1-\cos^{2(d-1)n}\left(\frac{\pi}{2n}\|g_1(z)-g_1(z')\|_1\right)},$$

where the second inequality follows from the inequality $\cos^n(\sum_i x_i/n) \geqslant \prod_i\cos(x_i)$ proven for Eq. (E6). Since the above inequality holds for any z,z' such that $\|z-z'\|_2 = \tau$ for any τ, and since we assume $\omega(\cdot)$ forms a tight upper bound in Eq. (6), g is smooth with

$$\omega(\tau) \geqslant \sqrt{1-\cos^{2n(d-1)}\left(\frac{\pi}{2n}\omega_1(\tau)\right)}, \quad \forall\tau>0.$$

In terms of the scaling with respect to n and d, if $\omega_1(\cdot) = \Omega(1)$, such as when g_1 is Lipschitz continuous, we have $\omega(\cdot) = \Omega(\sqrt{d/n})$.

APPENDIX G: PROOF OF THEOREM 2

Proof. If letting $\epsilon_{\mathrm{in}} \geqslant \omega(\sqrt{\ln[\pi/(2\gamma^2)]})$, then $\gamma \geqslant \sqrt{\pi/2}\exp[-\omega^{-1}(\epsilon_{\mathrm{in}})^2/2]$. By the definition of the generator and the latent space, we have $\mathcal{N}_m[g^{-1}(\rho)] = \xi(\rho),\ \forall\rho\in\mathcal{S}\subseteq$

$\mathcal{X}$. Let us define $h^i_{\rightarrow} = \{\rho \in h^i | d(\rho, \cup_{j\neq i} h^j) \leqslant \epsilon_{\text{in}}\}$ which is the set of density matrices that are at positive distance at most ϵ_{in} from $\cup_{j\neq i} h^j$, then following Definition 3,

$$R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) = \Pr_{\rho \leftarrow \xi} \{\min_{\sigma \in \mathcal{S}}[\|\sigma - \rho\|_1 | h(\sigma) \neq h(\rho)] \leqslant \epsilon_{\text{in}}\}$$
$$= \xi(\cup_i h^i_{\rightarrow}) = \mathcal{N}_m[g^{-1}(\cup_i h^i_{\rightarrow})], \tag{G1}$$

since $h^i_{\rightarrow}$ are disjoint for different class i. Hence, it can be shown that $R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) \geqslant 1 - \gamma$ when $\xi(h^i) \leqslant 1/2, \forall i$ from Theorem 1 in Ref. [15]. The contrapositive yields the necessary condition Eq. (7). For completeness, we present our condensed version of the proof below.

We write the cumulative distribution function of the standard Gaussian distribution $\mathcal{N}(0, 1)$ as $\Phi(x) = 1/\sqrt{2\pi} \int_{-\infty}^{x} \exp(-u^2/2) du$.

Theorem 4 (Gaussian isoperimetric inequality [20,49]). Let $\mathcal{N}_m$ be the canonical Gaussian measure on $\mathbb{R}^m$. Let $\Sigma \subseteq \mathbb{R}^m$ be any Borel set and let $\Sigma_\epsilon = \{z \in \mathbb{R}^m | \exists z' \in \Sigma \text{ s.t. } \|z - z'\|_2 \leqslant \epsilon\}$. If $\mathcal{N}_m(\Sigma) = \Phi(a)$ then $\mathcal{N}_m(\Sigma_\epsilon) \geqslant \Phi(a + \epsilon)$.

Lemma 1 ([15]). Let $p \in [1/2, 1]$, we have for all $\eta > 0$,

$$\Phi(\Phi^{-1}(p) + \eta) \geqslant 1 - (1 - p)\sqrt{\frac{\pi}{2}} e^{-\frac{\eta^2}{2}}. \tag{G2}$$

If $p = 1 - 1/K$ for $K \geqslant 5$ and $\eta \geqslant 1$, then we have

$$\Phi\left(\Phi^{-1}(1 - \frac{1}{K}) + \eta\right) \geqslant 1 - \frac{1}{K}\sqrt{\frac{\pi}{2}} e^{-\frac{\eta^2}{2}} e^{-\eta\sqrt{\log\left(\frac{K^2}{4\pi \log(K)}\right)}}. \tag{G3}$$

We first introduce the following sets in the latent space $(\mathbb{R}^m, \ell^2, \mathcal{N}_m)$: $H^i = g^{-1}(h^i)$ and $H^i_{\rightarrow} = \{z \in H^i | d(z, \cup_{j\neq i} H^j) \leqslant \omega^{-1}(\epsilon_{\text{in}})\}$. We note that $H^i_{\rightarrow} \bigcup \cup_{j\neq i} H^j$ is the set of points that are at distance at most $\omega^{-1}(\epsilon_{\text{in}})$ from $\cup_{j\neq i} H^j$. Then by Theorem 4 applied with $\Sigma = \cup_{j\neq i} H^j$ and $a = a_{\neq i} \equiv \Phi^{-1}[\mathcal{N}_m(\cup_{j\neq i} H^j)]$, we have $\mathcal{N}_m(H^i_{\rightarrow}) + \mathcal{N}_m(\cup_{j\neq i} H^j) \geqslant \Phi[a_{\neq i} + \omega^{-1}(\epsilon_{\text{in}})]$. Rearranging, $\mathcal{N}_m(H^i_{\rightarrow}) \geqslant \Phi[a_{\neq i} + \omega^{-1}(\epsilon_{\text{in}})] - \Phi(a_{\neq i})$. As $H^i_{\rightarrow}$ are disjoint for different class i, we have

$$\mathcal{N}_m(\cup_i H^i_{\rightarrow}) \geqslant \sum_{i=1}^{K} \{\Phi[a_{\neq i} + \omega^{-1}(\epsilon_{\text{in}})] - \Phi(a_{\neq i})\}.$$

By the definition of $\omega(\cdot)$, we have $g(H^i_{\rightarrow}) \subseteq h^i_{\rightarrow}$. It leads to $\mathcal{N}_m(g^{-1}(h^i_{\rightarrow})) \geqslant \mathcal{N}_m(H^i_{\rightarrow})$ and $\mathcal{N}_m[\cup_i g^{-1}(h^i_{\rightarrow})] \geqslant \mathcal{N}_m(\cup_i H^i_{\rightarrow})$. Therefore, we obtain the result for arbitrary decision boundary,

$$\mathcal{N}_m(\cup_i g^{-1}(h^i_{\rightarrow})) \geqslant \sum_{i=1}^{K} \{\Phi[a_{\neq i} + \omega^{-1}(\epsilon_{\text{in}})] - \Phi(a_{\neq i})\}.$$

Equivalently by Eq. (G1),

$$R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) \geqslant \sum_{i=1}^{K} \{\Phi[a_{\neq i} + \omega^{-1}(\epsilon_{\text{in}})] - \Phi(a_{\neq i})\}.$$

Suppose $\xi(h^i) = \mathcal{N}_m(H^i) \leqslant 1/2$ and $\mathcal{N}_m(\cup_{j\neq i} H^j) \geqslant 1/2, \forall i$. Using Eq. (G2) in Lemma 1 in the second inequality below,

$$R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) \geqslant \sum_{i=1}^{K} \big(\Phi\{\Phi^{-1}[\mathcal{N}_m(\cup_{j\neq i} H^j)] + \omega^{-1}(\epsilon_{\text{in}})\}$$
$$-\mathcal{N}_m(\cup_{j\neq i} H^j)\big)$$
$$\geqslant \sum_{i=1}^{K} \left\{1 - [1 - \mathcal{N}_m(\cup_{j\neq i} H^j)]\sqrt{\frac{\pi}{2}} e^{\frac{-\omega^{-1}(\epsilon_{\text{in}})^2}{2}}\right.$$
$$\left. -\mathcal{N}_m(\cup_{j\neq i} H^i)\right\}$$
$$= \left(1 - \sqrt{\frac{\pi}{2}} e^{\frac{-\omega^{-1}(\epsilon_{\text{in}})^2}{2}}\right) \sum_{i=1}^{K} [1 - \mathcal{N}_m(\cup_{j\neq i} H^i)]$$
$$= 1 - \sqrt{\frac{\pi}{2}} e^{\frac{-\omega^{-1}(\epsilon_{\text{in}})^2}{2}} > 1 - \gamma,$$

provided that $\gamma > \sqrt{\pi/2} \exp[-\omega^{-1}(\epsilon_{\text{in}})^2/2]$. The contrapositive yields the results in our Theorem 2 that $\epsilon_{\text{in}} \leqslant \omega(\sqrt{\ln[\pi/(2\gamma^2)]})$ is necessary for $R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) \leqslant 1 - \gamma$.

When there are at least five equiprobable classes [15], substituting Eq. (G3) in Lemma 1 into the above inequality yields

$$R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) \geqslant 1 - \sqrt{\frac{\pi}{2}} e^{\frac{-\omega^{-1}(\epsilon_{\text{in}})^2}{2}} e^{-\epsilon_{\text{in}}\sqrt{\log\left(\frac{K^2}{4\pi \log(K)}\right)}}.$$

Hence, the in-distribution robustness of h decreases with the number of equiprobable classes.

Alternatively, a numerically looser upper bound on ϵ_{in} can be derived from the fact that $(\mathbb{R}^m, \ell^2, \mathcal{N}_m)$ resembles a normal Lévy family but the concentration function decays independently of N. By Theorem 4, any Borel set Σ there such that $\mathcal{N}_m(\Sigma) = \Phi(a)$ satisfies $\mathcal{N}_m(\Sigma_\epsilon) \geqslant \Phi(a + \epsilon)$. In particular, for all Borel sets A with measure at least 1/2, we have $a \geqslant 0$ and thus, $1 - \mathcal{N}_m(A_\epsilon) \leqslant 1 - \Phi(\epsilon) \leqslant \exp(-\epsilon^2/2)$ where the last inequality follows from the Gaussian tail bound. By definition of the concentration function in Eq. (3), $\alpha(\epsilon) = \sup_A\{1 - \mathcal{N}_m(A_\epsilon)\} \leqslant \exp(-\epsilon^2/2)$.

By substituting the statement and the proof of Theorem 1 with $k_1 = 1$ and $k_2 = 1/\sqrt{2}$ and $N = 1$, we have the following. Let $\eta \in [0, 1/2]$ be such that $\mathcal{N}_m(H^l) = \xi(h^l) \leqslant 1 - \eta$, $\forall l \in \mathcal{L}$. If $\epsilon_{\text{in}} \geqslant \omega(\sqrt{\ln(4/\gamma^2)} + \sqrt{\ln(4/\eta^2)})$, then by acting $\omega^{-1}(\cdot)$, which is a strictly increasing function, on both sides, we obtain $\omega^{-1}(\epsilon_{\text{in}}) \geqslant \sqrt{\ln(4/\gamma^2)} + \sqrt{\ln(4/\eta^2)}$. This implies that $R^{\text{PC}}_{\omega^{-1}(\epsilon_{\text{in}})}(h, \mathcal{N}_m) \geqslant 1 - \gamma$. Since $R^{\text{PC}}_{\omega^{-1}(\epsilon_{\text{in}})}(h, \mathcal{N}_m) \leqslant R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi)$ (this is equivalent to $g(H^i_{\rightarrow}) \subseteq h^i_{\rightarrow}$), it therefore implies $R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) \geqslant 1 - \gamma$. The contrapositive yields, for $R^{\text{PC}}_{\epsilon_{\text{in}}}(h, \xi) \leqslant 1 - \gamma$, it is necessary to have $\epsilon_{\text{in}} \leqslant \omega(\sqrt{\ln(4/\gamma^2)} + \sqrt{\ln(4/\eta^2)})$. When $\eta = 1/2$, it can be verified that this necessary upper bound is looser than that in Theorem 2 for the same γ.

APPENDIX H: PROOF OF THEOREM 3

Proof. We have the mapping to obtain a product state density matrix $P : \mathcal{X} \to \mathcal{X}$, $\sigma \mapsto \bigotimes_{i=1}^{n} \text{tr}_{\{j\neq i\}}\sigma$ where n is the number of qubits and $\text{tr}_{\{j\neq i\}}(\sigma)$ means tracing out all but the ith qubit of σ. This is not a CPTP map on the set of $d^n \times d^n$

density matrices $\mathcal{X}$ since it is nonlinear. Nonetheless, it can be viewed as a CPTP map Λ on $\mathcal{X}^{\otimes n}$ as $\Lambda : \mathcal{X}^{\otimes n} \to \mathcal{X}$, $\sigma^{\otimes n} \mapsto \mathrm{tr}_{\{j\neq i\}}([\sigma^{\otimes n}]_i)$ where $[\sigma^{\otimes n}]_i$, $i \in \{1,\dots,n\}$, denotes the ith copy of σ in $\sigma^{\otimes n}$, since, Λ involves only partial tracing on every σ copy in $\sigma^{\otimes n}$. In particular, for a product state $\rho^{\otimes a}$ with the integer $a \geqslant 1$, $\Lambda(\rho^{\otimes a}) = \rho$.

Consider $\rho \in \mathcal{S} \subseteq \mathcal{X}$ an n-qubit density matrix, namely, $\rho = g(z)$ for some $z \in \mathcal{Z}$. Let $\sigma \in \mathcal{X}$. We have

$$\|\rho - P(\sigma)\|_1 = \|\Lambda(\rho^{\otimes n}) - \Lambda(\sigma^{\otimes n})\|_1 \leqslant \|\rho^{\otimes n} - \sigma^{\otimes n}\|_1$$
$$\leqslant 2\sqrt{1 - F(\rho^{\otimes n}, \sigma^{\otimes n})} = 2\sqrt{1 - F(\rho,\sigma)^n},$$

where the first inequality follows from the contractive property of the trace norm under any CPTP map and the last equality follows from the multiplicativity of fidelity with respect to tensor products. By Eq. (E1), we have $F(\rho,\sigma) \geqslant (1 - \|\rho - \sigma\|_1/2)^2$. Substituting in, we obtain

$$\|\rho - P(\sigma)\|_1 \leqslant 2\sqrt[2n]{1 - \left(1 - \frac{\|\rho - \sigma\|_1}{2}\right)}.$$

Let $\tilde{\sigma} \in \mathcal{S}$ be the closest in-distribution sample to $P(\sigma)$, which can be found by fitting parameters $\{u_i\}$ in Eq. (1). Therefore, $\|P(\sigma) - \tilde{\sigma}\|_1 \leqslant \|P(\sigma) - \rho\|$. We then obtain

$$\|\rho - \tilde{\sigma}\|_1 \leqslant \|\rho - P(\sigma)\|_1 + \|P(\sigma) - \tilde{\sigma}\|_1$$
$$\leqslant 4\sqrt[2n]{1 - \left(1 - \frac{\|\rho - \sigma\|_1}{2}\right)}. \tag{H1}$$

Recall that for the quantum classifier $\tilde{h}$, $\tilde{h}(\sigma) = h(\tilde{\sigma})$. Taking minimum over all σ such that $\tilde{h}(\sigma) \neq \tilde{h}(\rho)$ [i.e., $h(\tilde{\sigma}) \neq h(\rho)$],

$$\varepsilon_{\mathrm{in}}(\rho) \leqslant \min\{\|\rho - \tilde{\sigma}\|_1\}$$
$$\leqslant 4\sqrt[2n]{1 - \left(1 - \frac{\min\{\|\rho - \sigma\|_1\}}{2}\right)}, \tag{H2}$$

we obtain

$$\varepsilon_{\mathrm{in}}(\rho) \leqslant 4\sqrt[2n]{1 - \left(1 - \frac{\varepsilon_{\mathrm{unc}}(\rho)}{2}\right)}. \tag{H3}$$

Notice that to obtain an inequality between $\varepsilon_{\mathrm{in}}(\rho)$ and $\varepsilon_{\mathrm{unc}}(\rho)$ like in Eq. (H3), it is sufficient to have Eq. (H2) hold after taking the minimum, and it is not necessary to have Eq. (H1) hold for a generic σ. Since n-qubit density matrices which are separable with respect to some equal bipartition of the system, denoted as $\{\rho_b\}$, form a dense subset [68], we can effectively realize the same minimum in Eq. (H2) over $\sigma \in \{\rho_b\}$ such that $\tilde{h}(\sigma) \neq \tilde{h}(\rho)$ instead. For equal bipartite states, the number of copies to make a CPTP map Λ' acting on them to obtain $P(\sigma)$ reduces to $n/2$ if n is even and reduces to $(n+1)/2$ if n is odd. For instance, given a 4-qubit σ whose qubit 1 is only entangled with 2 and qubit 3 is only entangled with 4, $\Lambda'(\sigma^{\otimes 2}) = \mathrm{tr}_{\{1,3\}}(\sigma) \otimes \mathrm{tr}_{\{2,4\}}(\sigma) = P(\sigma) = \Lambda(\rho^{\otimes 4})$. Therefore, we can replace the exponent $1/(2n)$ in Eq. (H3) with $1/n$ for even n and $1/(n+1)$ for odd n.

We recall $\varepsilon_{\mathrm{unc}}(\rho) \leqslant \varepsilon_{\mathrm{in}}(\rho)$, $\forall \rho \in \mathcal{X}$ and rearrange,

$$2 - 2\left(1 - \frac{\varepsilon_{in}(\rho)^2}{16}\right)^{\frac{1}{n_e}} \leqslant \varepsilon_{\mathrm{unc}}(\rho) \leqslant \varepsilon_{\mathrm{in}}(\rho),$$

where $n_e = n$ for even n and $n_e = n + 1$ for odd n.

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

[2] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[3] Y. Xia, W. Li, Q. Zhuang, and Z. Zhang, Quantum-Enhanced Data Classification with a Variational Entangled Sensor Network, arXiv:2006.11962 [Phys. Rev. X (to be published)].

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in *Proceedings of the ICLR* (ICLR, 2014).

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, in *Proceedings of the ICLR* (ICLR, 2015).

[6] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, Adversarial classification, in *Proceedings of the ACM* (ACM, 2004) p. 99.

[7] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, A survey on adversarial attacks and defences, CAAI Transactions on Intelligence Technology **6**, 25 (2021).

[8] B. Biggio and F. Roli, Wild Patterns: Ten years after the rise of adversarial machine learning, Pattern Recogn. **84**, 317 (2018).

[9] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, Adversarial machine learning, in *Proceedings of the ACM* (ACM, 2011) pp. 43–57.

[10] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, Robustness of classifiers: From adversarial to random noise, in *Proceedings of the NIPS* (ACM, 2016), pp. 1632–1640.

[11] S. Mahloujifar, D. I. Diochnos, and M. Mahmoody, The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure, in *Proceedings of the AAAI*, Vol. 33 (PKP Publishing, 2019), pp. 4536–4543.

[12] J. Gilmer, L. Metz, F. Faghri, S. S Schoenholz, M. Raghu, M. Wattenberg, I. Goodfellow, and G. Brain, The relationship between high-dimensional geometry and adversarial examples, ICLR Workshop, arXiv:1801.02774 (2018).

[13] D. I. Diochnos, S. Mahloujifar, and M. Mahmoody, Adversarial risk and robustness: General definitions and implications for the uniform distribution, in *Proceedings of the NIPS* (ACM, 2018), pp. 10380–10389.

[14] A. Fawzi, O. Fawzi, and P. Frossard, Analysis of classifiers' robustness to adversarial perturbations, Mach. Learn. **107**, 481 (2018).

[15] A. Fawzi, H. Fawzi, and O. Fawzi, Adversarial vulnerability for any classifier, in *Proceedings of the NIPS* (ACM, 2018) pp. 1186–1195.

[16] N. Liu and P. Wittek, Vulnerability of quantum classification to adversarial perturbations, Phys. Rev. A **101**, 062331 (2020).

[17] S. Lu, L.-M. Duan, and D.-L. Deng, Quantum adversarial machine learning, Phys. Rev. Res. **2**, 033212 (2020).

[18] Y. Du, M.-H. Hsieh, T. Liu, D. Tao, and N. Liu, Quantum noise protects quantum classifiers against adversaries, arXiv:2003.09416 (2020).

[19] J. Guan, W. Fang, and M. Ying, Robustness verification of quantum machine learning, arXiv:2008.07230 (2020).

[20] M. Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs, Vol. 89 (American Mathematical Society, Providence, RI, 2001).

[21] Vitali D. Milman, Gideon Schechtman, *Asymptotic Theory of Finite Dimensional Normed Spaces: Isoperimetric Inequalities in Riemannian Manifolds*, Lecture Notes in Mathematics (Springer, Berlin, 2002).

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

[23] S. Popescu, A. J. Short, and A. Winter, Entanglement and the foundations of statistical mechanics, Nat. Phys. **2**, 754 (2006).

[24] M. P. Müller, D. Gross, and J. Eisert, Concentration of measure for quantum states with a fixed expectation value, Commun. Math. Phys. **303**, 785 (2011).

[25] D. J. Rezende and S. Mohamed, Variational inference with normalizing flows, in *Proceedings of the PMLR*, Vol. 37 (PMLR, 2015) pp. 1530–1538.

[26] D. P. Kingma and M. Welling, Auto-encoding variational Bayes, in *Proceedings of the ICLR* (ICLR, 2014).

[27] L. Dinh, J. Sohl-Dickstein, and S. Bengio, Density estimation using real NVP, in *Proceedings of the ICLR* (ICLR, 2017).

[28] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Proceedings of the NIPS* (ACM, 2014), pp. 2672–2680.

[29] P. Bojanowski, A. Joulin, D. Lopez-Paz, and A. Szlam, Optimizing the latent space of generative networks, in *Proceedings of the PMLR*, Vol. 80 (PMLR, 2018), pp. 600–609.

[30] M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein Generative adversarial networks, in *Proceedings of the PMLR*, Vol. 70 (PMLR, 2017) pp. 214–223.

[31] A. Kurakin, I. J. Goodfellow, and S. Bengio, Adversarial machine learning at scale, in *Proceedings of the ICLR* (ICLR, 2017).

[32] A. Kurakin, I. J. Goodfellow, and S. Bengio, Adversarial examples in the physical world, in *Proceedings of the ICLR* (ICLR, 2019).

[33] S. Bubeck, E. Price, and I. Razenshteyn, Adversarial examples from computational constraints, in *Proceedings of the PMLR*, Vol. 97 (PMLR, 2019) pp. 831–840.

[34] Z. Charles, H. Rosenberg, and D. Papailiopoulos, A geometric perspective on the transferability of adversarial directions, in *Proceedings of the PMLR*, Vol. 89 (PMLR, 2018) pp. 1960–1968.

[35] L. Engstrom, J. Gilmer, G. Goh, D. Hendrycks, A. Ilyas, A. Madry, R. Nakano, P. Nakkiran, S. Santurkar, B. Tran, D. Tsipras, and E. Wallace, Adversarial examples are not bugs, they are features, in *Proceedings of the NIPS* (ACM, 2019) pp. 125–136.

[36] J. P. Göpfert, A. Artelt, H. Wersing, and B. Hammer, Adversarial attacks hidden in plain sight, Adv. Intell. Data Anal. **18**, 235 (2020).

[37] Informally, total variation distance is the largest possible difference between the probabilities that the two distributions can assign to the same event.

[38] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, UK, 2010).

[39] W. Huggins, P. Patil, B. Mitchell, K. Birgitta Whaley, and E. Miles Stoudenmire, Towards quantum machine learning with tensor networks, Quant. Sci. Technol. **4**, 24001 (2019).

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

[41] E. M. Stoudenmire and D. J. Schwab, Supervised learning with quantum-inspired tensor networks, in *Proceedings of the NIPS* (ACM, 2016), pp. 4799–4807.

[42] E. Grant, M. Benedetti, S. Cao, A. Hallam, J. Lockhart, V. Stojevic, A. G. Green, and S. Severini, Hierarchical quantum classifiers, Quant. Info. **4**, 65 (2018).

[43] S. Cao, L. Wossnig, B. Vlastakis, P. Leek, and E. Grant, Cost-function embedding and dataset encoding for machine learning with parameterized quantum circuits, Phys. Rev. A **101**, 052309 (2020).

[44] J. Martyn, G. Vidal, C. Roberts, and S. Leichenauer, Entanglement and tensor networks for supervised image classification, arXiv:2007.06082 (2020).

[45] V. Giovannetti, S. Lloyd, and L. Maccone, Quantum Random Access Memory, Phys. Rev. Lett. **100**, 160501 (2008).

[46] R. LaRose and B. Coyle, Robust data encodings for quantum classifiers, Phys. Rev. A **102**, 032420 (2020).

[47] Borel sets are sets that can be constructed from open or closed sets through countable union, countable intersection, and relative complement .

[48] R. Vershynin, Four lectures on probabilistic methods for data science, arXiv:1612.06661 (2017).

[49] C. Borell, The Brunn-Minkowski inequality in Gauss space, Invent. Math. **30**, 207 (1975).

[50] M. Gromov and V. D. Milman, A topological application of the isoperimetric inequality, Am. J. Math. **105**, 843 (1983).

[51] T. Giordano and V. Pestov, Some extremely amenable groups related to operator algebras and ergodic theory, J. Inst. Math. Jussieu **6**, 279 (2007).

[52] The precise definition of a nice classification problem can be found in Definition 2.3 in Ref. [11].

[53] A concise proof of Eq. (4) can be found in Appendix C.

[54] G. F. Elsayed, N. Papernot, S. Shankar, A. Kurakin, B. Cheung, I. Goodfellow, and J. Sohl-Dickstein, Adversarial examples that fool both computer vision and time-limited Humans, in *Proceedings of the NIPS* (2018) pp. 3910–3920.

[55] R. Lockhart, Low-rank separable states are a set of measure zero within the set of low-rank states, Phys. Rev. A **65**, 064304 (2002).

[56] Y. LeCun, C. Cortes, and C. Burges, MNIST Handwritten Digit Database, http://yann.lecun.com/exdb/mnist/.

[57] A. Jalal, A. Ilyas, C. Daskalakis, and A. G. Dimakis, The robust manifold defense: Adversarial training using generative models, arXiv:1712.09196 (v5 2019).

[58] M. Khayatkhoei, M. K. Singh, and A. Elgammal, Disconnected manifold learning for generative adversarial networks, in *Proceedings of the NIPS* (ACM, 2018) pp. 7343–7353.

[59] A. Radford, L. Metz, and S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in *Proceedings of the ICLR* (ICLR, 2016).

[60] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J.-H. Jacobsen, Invertible residual networks, in *Proceedings of the PMLR*, Vol. 97 (PMLR, 2019), pp. 573–582.

[61] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, Self-attention generative adversarial networks, in *Proceedings of the PMLR*, Vol. 97 (PMLR, 2019) pp. 7354–7363.

[62] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, Spectral normalization for generative adversarial networks, in *Proceedings of the ICLR* (ICLR, 2018).

[63] It is proven in Theorem 2 in Ref. [15].

[64] Z. Zhao, D. Dua, and S. Singh, Generating natural adversarial examples, in *Proceedings of the ICLR* (ICLR, 2018).

[65] P. J. Coles, M. Cerezo, and L. Cincio, Strong bound between trace distance and Hilbert-Schmidt distance for low-rank states, Phys. Rev. A **100**, 022103 (2019).

[66] D. Spehner, F. Illuminati, M. Orszag, and W. Roga, Geometric measures of quantum correlations with Bures and Hellinger distances, in *Lectures on General Quantum Correlations and their Applications*, edited by F. F. Fanchini, D. de Oliveira Soares Pinto, and G. Adesso, Quantum Science and Technology (Springer International Publishing, Cham, Switzerland, 2017), pp. 105–157.

[67] A. S. Holevo, On quasiequivalence of locally normal states, Theor. Math. Phys. **13**, 1071 (1972).

[68] R. Orus and R. Tarrach, Weakly-entangled states are dense and robust, Phys. Rev. A **70**, 050101(R) (2004).