Enhancing quantum adversarial robustness by randomized encodings

Gong, W.; Yuan, D.; Li, W.; Deng, D.L.

Enhancing quantum adversarial robustness by randomized encodings

Weiyuan Gong,[1,2] Dong Yuan,[1,*] Weikang Li,[1,3] and Dong-Ling Deng[1,4,5,†]
[1]*Center for Quantum Information, IIIS, Tsinghua University, Beijing 100084, China*
[2]*School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*
[3]*Instituut-Lorentz, Universiteit Leiden, P.O. Box 9506, 2300 RA Leiden, The Netherlands*
[4]*Shanghai Qi Zhi Institute, No. 701 Yunjin Road, Xuhui District, Shanghai 200232, China*
[5]*Hefei National Laboratory, Hefei 230088, China*

(Received 3 January 2023; accepted 15 March 2024; published 4 April 2024)

The interplay between quantum physics and machine learning gives rise to the emergent frontier of quantum machine learning, where advanced quantum learning models may outperform their classical counterparts in solving certain challenging problems. However, quantum learning systems are vulnerable to adversarial attacks: adding tiny carefully crafted perturbations on legitimate input samples can cause misclassifications. To address this problem, we propose an effective approach to protect quantum learning systems from adversarial attacks by randomly encoding the legitimate data samples through unitary or quantum error correction encoders that are unknown to the attackers. In particular, we rigorously prove that both global and local random unitary encoders lead to exponentially vanishing gradients (i.e., barren plateaus) for any variational quantum circuits that aim to add adversarial perturbations, independent of the input data and the inner structures of adversarial circuits and quantum classifiers. In addition, we prove a rigorous bound on the vulnerability of quantum classifiers under local unitary adversarial attacks. We show that random black-box quantum error correction encoders can protect quantum classifiers against local adversarial noises and their robustness increases as we concatenate error correction codes. To quantify the robustness enhancement, we adapt quantum differential privacy as a measure of the prediction stability for quantum classifiers. Our results establish versatile defense strategies for quantum classifiers against adversarial perturbations, which provide valuable guidance to enhance the reliability and security for both near-term and future quantum learning technologies.

DOI: 10.1103/PhysRevResearch.6.023020

I. INTRODUCTION

The flourish of machine learning has led to unprecedented opportunities and achieved dramatic success in both research and commercial fields [1,2]. Some notoriously challenging problems, ranging from predicting protein structures [3] and weather forecasting [4] to playing the game of Go [5,6], have been cracked recently. Meanwhile, the field of quantum computation has also made tremendous progress in recent years [7,8], giving rise to unparalleled opportunities to speed up, enhance, or innovate machine learning [9–12]. Within this vein, ideas and concepts from the physics domain have been utilized as core ingredients for quantum machine-learning algorithms [13–20]. Notable examples in this direction include the Harrow-Hassidim-Lloyd algorithm [13], quantum principal component analysis [14], quantum generative models [12,15,16], quantum support vector machines [17], and variational quantum algorithms based on parametrized quantum circuits [18–21], etc. Yet, an important issue regarding quantum learning systems concerns their reliability and security in adversarial scenarios, especially for noisy intermediate-scale quantum (NISQ) devices [22]. Here, we introduce a series of defense strategies by randomly encoding legitimate data samples and analytically show their adversarial robustness in a rigorous fashion (see Fig. 1 for illustration).

Adversarial machine learning is an emerging frontier that studies the vulnerability of machine learning systems and develops defense strategies against adversarial attacks [23,24]. In the classical scenario, the prediction of a deep neural network can be susceptible to tiny carefully-crafted noises, which are even imperceptible to human eyes, added to the legitimate input data [25–29]. These adversarial noises can be generated by either a malicious adversary or the worst-case experimental noise from an unknown source. Recent works have demonstrated that quantum learning systems are vulnerable under adversarial settings similar to their classical counterparts [30–32], sparking a new interdisciplinary research frontier of quantum adversarial machine learning [30–35]. From the theoretical aspect, even an exponentially small perturbation can cause a moderate adversarial risk for a given quantum classifier [32]. Furthermore, it has been shown that there exist universal adversarial attacks for multiple quantum classifiers or input data samples [31]. More recently, quantum

*yuand21@mails.tsinghua.edu.cn
†dldeng@tsinghua.edu.cn

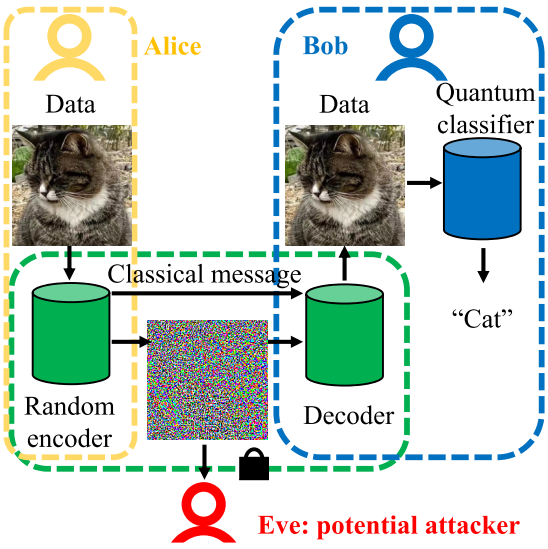

FIG. 1. An illustration for exploiting randomized encoding to defend against adversarial attacks. In the quantum learning task, Alice prepares an input data sample and sends it to Bob for classification. To protect the legitimate data against the potential adversary Eve, Alice, and Bob share a codebook and Alice randomly chooses an encoder in the codebook to transform the original data into encoded data, from which Eve can barely obtain any useful information. Then Alice sends Bob the encoded quantum data and classical information about the encoder. Bob receives the messages, translates the encoded quantum data into the original figure, and performs the classification.

adversarial learning has been experimentally demonstrated with both large-scale real-life datasets and quantum datasets on superconducting quantum devices [33]. To improve the robustness of quantum machine learning algorithms and defend against adversarial attacks, a straightforward approach is to employ a quantum-adaptive adversarial training [30]. However, adversarial training in general requires generations of a large number of adversarial samples and may only perform well for the same attacking method that generates those samples.

In classical adversarial learning, randomness is suggested to be the possible resource for developing defense strategies against adversarial perturbations [36–42]. However, these results are mostly empirical and there has been no unified and rigorous framework for employing randomness in this context. In quantum computation, quantum error correction (QEC) codes are widely used to detect and correct experimental errors. However, the errors that can be corrected are assumed to be local while adversarial perturbations are either carefully engineered or worst-case noises. In addition, the vanishing gradients (i.e., barren plateaus) for quantum circuits with random parameters give potential protection from most commonly used gradient-based adversarial algorithms [30,43]. A potential approach to achieve provable adversarial robustness for quantum classifiers is to combine randomness with QEC and barren plateaus phenomenon studied in quantum computation.

In this paper, we propose an approach employing a randomized encoding procedure to protect the quantum learning systems from potential adversarial perturbations. Under practical adversarial learning scenarios, adversarial perturbations can originate from either carefully crafted perturbations created by the attackers that have full access to the gradient information [30] or the worst-case experimental noises from unknown resources [44]. We show the effectiveness of our scheme by using two concrete types of random encoders to mask the gradient information from the adversary and improve the robustness of quantum learning algorithms. The first type uses random unitary encoders and is more practical for NISQ devices, whereas the second type exploits QEC encoders that are necessary for the future fault-tolerant quantum computation.

For the first type, we rigorously prove that a random global unitary encoder that satisfies the 2-design property [45] leads to exponentially small gradients for adversarial variational circuits, and thus creates barren plateaus [46–60] that may hinder gradient-based algorithms in generating adversarial perturbations. We further prove that even random encoders that can be decomposed into tensor products of unitary 2-design blocks of smaller sizes can generate barren plateaus for the adversaries as well. To benchmark the performance, we carry out numerical simulations concerning the classification of topological phases of the cluster-Ising model [61,62] with different loss functions and system sizes. For the second type of encoders, we consider local adversarial perturbations generated by worst-case experimental noises. We prove a lower bound for the adversarial risk in this setting based on the concentration of measure phenomenon in the high-dimensional space. We analytically show that a random black-box QEC [63,64] encoding procedure can improve the robustness of quantum learning systems for local unitary attacks. In particular, we show that it is sufficient to concatenate only $O(\log\log(n))$ levels of QEC encoders to bound the adversarial risk below a constant value. We adapt quantum differential privacy (QDP) [65–67] to measure the robustness of quantum classifiers against adversarial perturbations. We prove an information-theoretic upper bound for the adversarial risk of quantum learning algorithms satisfying differential privacy.

The randomized encoding approach introduced in this paper is distinct from previous literature that either exploit deterministic encoders for binary classification [68] or adds white noises [44]. Compared to the deterministic encoder scheme which uses amplitude and phase encoding, our approach uses variational unitary circuits that are more experimentally compatible for NISQ devices. Whereas adding white noise may diminish the performance of the quantum classifiers, our approach will not influence the accuracy of classification algorithms. Furthermore, in contrast to classical algorithms that employ randomness against adversarial attacks, our approaches provide rigorous theoretical bounds rather than empirical performance benchmarks. Our results not only establish a profound connection among QEC, QDP, barren plateau phenomenon, and quantum adversarial robustness, but also provide practical defense strategies that may prove valuable in future applications of quantum learning technologies.

The paper is organized as follows. In Sec. II, we introduce the basic concepts and general framework for quantum adversarial learning. In Sec. III, we present two theorems demonstrating that both global and local randomized unitary encoders on input data samples can lead to vanishing

gradients, which may hamper gradient-based algorithms from creating adversarial perturbations. We provide numerical evidence concerning classifications on the phases of the cluster-Ising model to benchmark the effectiveness of our approach. In Sec. IV, we give two theorems, one proving the vulnerability of quantum classifiers against local unitary adversarial perturbations, the other demonstrating that black-box QEC encoders can effectively defend the local unitary adversarial noises on the input data samples. Finally, in Sec. V, we discuss several open problems and conclude the paper.

II. BASIC CONCEPTS AND GENERAL FRAMEWORK

Machine-learning technologies have recently achieved remarkable breakthroughs in various real-world applications [1,2] including natural language processing [69], automated driving [43], and medical diagnostics [70]. Meanwhile, serious concerns have also been raised about the integrity and security of such technologies in various adversarial scenarios [23,25,26]. For instance, the medical recognition software from a medical diagnostics or a sign-recognition system from a self-driving car may cause catastrophic medical or traffic accidents if they are not robust against some occasional modifications (which may even be imperceptible to human eyes) in identifying medical scans or traffic images [71]. To address these vital problems and concerns, the field of adversarial machine learning has been developed to construct and defend the potential adversarial manipulations against machine-learning systems under different scenarios [72]. The field has attracted considerable attention and there are rapid developments for both attack and defense strategies in different adversarial settings. For simplicity and concreteness, we will only focus our discussion on the setting of supervised learning, although generalizations to unsupervised or reinforcement learning settings are possible and worth systematic future investigations.

On the one hand, there have been a number of algorithms proposed to transfer the adversarial attack problem into an optimization one and solve the corresponding problem or its variants through optimization strategies [25,27–29,72–77]. We divide the adversarial attacks into black-box and white-box attacks according to the amount of information known by the adversary about the target classifier. In the white-box setting, the attacker has full information about the inner structure and algorithm of the classifier. Whereas, in the black-box setting, the attacker possesses only partial or even no information about the classifier. A crucial piece of information under adversarial settings is the gradient information about the classifier. The gradients can be calculated based on the inner structure, algorithm, and the loss function of the classifier. In the white-box setting, various algorithms such as the fast gradient sign method (FGSM) [74], basic iterative method (BIM) [29], projected gradient descent (PGD) [74], and momentum iterative method (MIM) [78] have been developed based on the gradient information. In the black-box setting, algorithms that exploit the transferability property of neural-network classifiers have been developed, including the transfer attack [28], substitute model attack [73,75], and zeroth-order optimization attack [77] methods. On the other hand, a number of defense strategies against adversarial attacks have been developed as well. Some notable examples include adversarial training [79], defense generative adversarial network [80,81], and knowledge distillation [82,83]. These algorithms have achieved satisfactory robustness performance against particular types of adversarial attacks. In general, we *cannot* expect a defense strategy that can promote the robustness of all machine-learning algorithms against any adversarial attacks as long as the adversary knows the information about the classifier. An alternative protocol to protect the classifier is to hide the information from the attackers. Some algorithms along this direction include adding random noise or transformations which smooth the gradients and the landscape of the loss function [36–41]. As a trade-off, these approaches, in general, would increase the difficulty in training the classifier.

Quantum classifiers are analogs of classical classifiers, which aim to solve classification problems with quantum devices [84]. In this paper, we propose a defense strategy for quantum classifiers against adversarial attacks through randomized encoders. We start with a brief introduction to the basic concepts, notations, and ideas of quantum classifiers and quantum adversarial learning. In general, a quantum classification task in the supervised learning setting aims to assign a label $s \in S$ to a pure state input quantum data sample $|\psi\rangle \in \mathcal{H}$, with S being a countable label set and $\mathcal{H}$ being a subspace of the entire Hilbert space. The supervised learning procedure aims to learn a function (called a hypothesis function) $h : \mathcal{H} \to S$ that outputs a label $s \in S$ for each input state $|\psi\rangle \in \mathcal{H}$. To achieve this goal, we parametrize the hypothesis function with $\boldsymbol{\theta} \in \Xi$, where Ξ is the parameter space. We train the classifier with a set of training data $\mathcal{T}_N = \{(|\psi\rangle^{(1)}, s^{(1)}), ..., (|\psi\rangle^{(N)}, s^{(N)})\}$, where $|\psi\rangle^{(i)}$ and $s^{(i)}$ $(i = 1, ..., N)$ are the input states and the corresponding labels. This procedure is usually achieved by minimizing a chosen loss function $\min_{\boldsymbol{\theta}\in\Xi} L_N(\boldsymbol{\theta})$ over the parameter space Ξ, with $L_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} L(h(|\psi\rangle^{(i)}; \boldsymbol{\theta}), s^{(i)})$ denoting the loss function averaged over the training set. A number of different quantum classifiers with different structures, loss functions, and optimization methods have been proposed [17,85–98]. Each approach bears its pros and cons, and the choice of the classifiers depends on the specific problem. A straightforward approach to construct a quantum classifier, known as a variational quantum classifier [85,86,88], is to exploit variational quantum circuits [18–20] to optimize the loss function analogously to quantum support vector machines [96]. There exist a number of different variants on the structures of the variational quantum circuits, including hierarchical quantum classifiers [93] and quantum convolutional neural networks [91], etc.

Recent research has shown that quantum classifiers also suffer from the vulnerability problem under adversarial attacks [30–32,35], with an experimental demonstration marked as recent progress [33]. Unlike the training procedure, finding an adversarial example for quantum classifiers can be regarded as a different optimization program on the input data space. Specifically, our goal is to discover the unitary perturbation U_δ within a restricted region Δ close to identity, which after being added to the legitimate input states, will maximize the loss function:

$$\max_{\delta\in\Delta} L(h(U_\delta|\psi\rangle^{(i)}; \boldsymbol{\theta}), s^{(i)}). \tag{1}$$

In the white-box setting, the inner structures of quantum classifiers and the loss functions are known by the attackers. Hence, the attackers can solve the optimization problem in Eq. (1) by exploiting the gradient information of the loss functions. There have been several algorithms to attack quantum classifiers, such as quantum-adaptive BIM, FGSM, MIM algorithms [30], etc.

The defense strategy under these quantum adversarial settings remains largely unexplored, with most attention concentrated on proving the robustness of a given classifier [68,99,100]. Some notable algorithms to boost the robustness of a quantum classifier, such as adversarial training [30] and adding random noises [44], still suffer from white-box adversarial attacks or the loss of classification accuracy.

Here we propose an effective approach to protect the quantum machine learning systems using black-box randomized encoders in adversarial settings. Our essential idea is illustrated in Fig. 1. We interpret the classification task as a three-party protocol, in which Alice prepares a legitimate quantum input data sample, Bob receives the data sample and performs the classification, and Eve is the potential adversary performing adversarial manipulations. To protect the quantum classification model from adversarial attacks, we assume that Alice and Bob share a codebook $C = \{p_i, E_i\}$ consisting of different encoders E_i with the corresponding decoders, and the probability distribution $\{p_i\}$ of choosing E_i. The agreement on the codebook can be realized by quantum key distribution [101] or quantum teleportation [63].

We assume that Alice is trusted and honest in the sense that she sends correctly encoded quantum data samples and correct classical messages about the random encoders to Bob. All the adversarial attacks by Eve are added during the transmission of encoded quantum messages from Alice to Bob. We remark that in our protocols while the quantum classifier is a white box for the potential attacker Eve, the encoder and decoder should be black boxes that are unknown by Eve. If Eve has the concrete information of the encoder, she can simply decode the quantum messages and add corresponding adversarial perturbations to the data samples. We mention that it is generally hard and requires exponentially many samples to learn the description of the unitary encoder [102,103] for Eve that only gets access to the encoded data. Alice and Bob can further switch among many different codebooks to prevent Eve from sampling the data using the same encoder for a large number of times. Besides, our defending protocols are not restricted to this three-party communication scenario. For example, the adversarial perturbations can come from the worst-case experimental noises occurring on some experimentally prepared quantum states that are waiting to be classified.

We rigorously show that by randomly choosing an encoder from the codebook, the encoded quantum data can be robust against adversarial perturbations. Roughly speaking, the random transformation induced by the encoder masks the information that can be obtained by the adversary, thus mitigating the adversarial risk. Specifically, we consider two types of codebooks shared by Alice and Bob in the following two sections concerning the variational quantum machine learning on NISQ devices and the fault-tolerant quantum machine learning in the future. We provide analytical bounds for the robustness of protected quantum machine-learning systems under adversarial settings.

III. DEFENDING ADVERSARIAL ATTACKS WITH BARREN PLATEAUS

We first consider the case of adding a random unitary transformation as an encoder. As shown in Eq. (1), the goal of the attackers in the quantum classification task is to implement the optimal unitary perturbation U_δ that maximizes the loss function L. To achieve this goal, most attacking algorithms assume the white-box access to the gradients of the loss functions. For instance, the quantum-adaptive BIM, FGSM, and MIM algorithms all hinge on calculating the gradients of the loss function L. To protect the quantum classifiers against the gradient-based attacking algorithms, an effective approach is to mask the gradient information of the loss functions from the attackers through some specially designed encoding schemes. Intuitively, random unitaries satisfying the unitary 2-design property could yield exponentially small expectation values when we compute their first and second moments. In particular, it has been rigorously proved that randomly parametrized quantum variational circuits satisfying the unitary 2-design could result in exponentially vanishing gradients [46], which is known as the barren plateau phenomenon. Therefore, it is natural to consider utilizing random unitary encoders to mask the gradient information from the attackers. Under the random unitary encoding scheme, the expected gradient values obtained by potential attackers will be reduced to be exponentially small.

Formally, we note that any adversary attack can be effectively implemented as adding a L-layer parametrized variational quantum circuit (PVQC) $U(\boldsymbol{\theta})$, as shown in Fig. 2(a). More concretely, we can write the adversarial PVQC as

$$U(\boldsymbol{\theta}) = U(\theta_1, ..., \theta_L) = \prod_{l=1}^{L} U_l(\theta_l) W_l, \tag{2}$$

where $U_l(\theta_l) = \exp(-i\theta_l A_l)$ is the parametrized variational component in each layer, A_l is a Hermitian operator, and W_l is a unitary operator that represents the fixed component in each layer. We assume the classifier $V(\Theta)$ is well-trained with parameters Θ. It can be a general unitary operator such as a P-layer PVQC shown in the figure. To perform a prediction, we simply measure some particular qubits at the output after the classifier and assign labels according to the measurement outcomes. Given an input pure state $|\psi\rangle_{\text{in}}$, the loss function can be regarded as an expectation value over a Hermitian operator H. For the legitimate input and the adversarial input, the loss functions can be written as $L(\Theta) = \langle\psi|_{\text{in}} V^\dagger(\Theta) H V(\Theta) |\psi\rangle_{\text{in}}$ and $L(\Theta;\boldsymbol{\theta}) = \langle\psi|_{\text{in}} U^\dagger(\boldsymbol{\theta}) V^\dagger(\Theta) H V(\Theta) U(\boldsymbol{\theta}) |\psi\rangle_{\text{in}}$, respectively.

To protect the quantum classifier $V(\Theta)$ from the adversarial PVQC $U(\boldsymbol{\theta})$, we exploit a random encoder E and the corresponding decoder $E^\dagger$ to encrypt the legitimate data sample $|\psi\rangle_{\text{in}}$. We note that the codebook $C = \{p_i, E_i\}$ contains a particular set of encoders with probability distribution $\{p_i\}$. We assume that C is unitary 2-design [45], namely, that the first and the second moments are equivalent to the

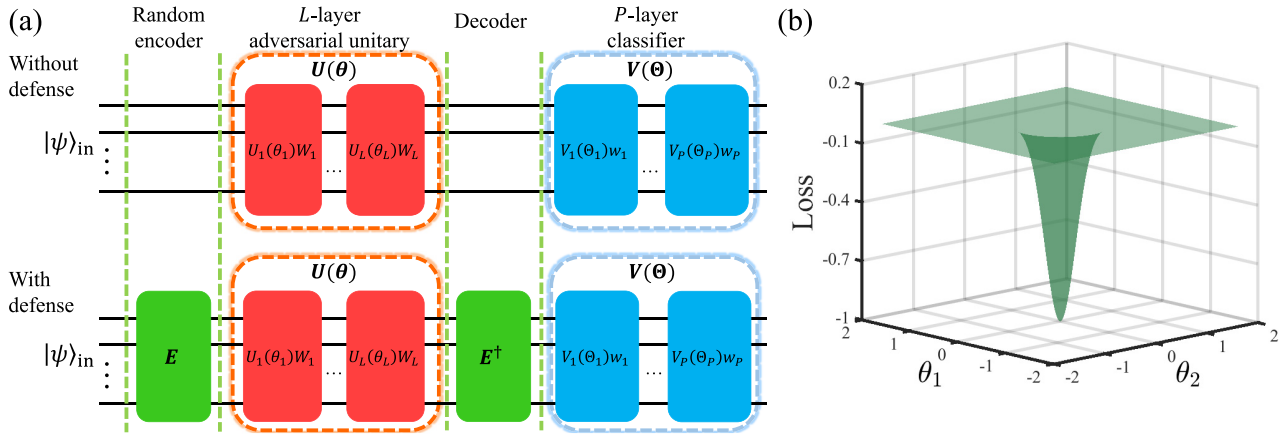

FIG. 2. (a) An illustration of exploiting a random unitary encoder to defend against adversarial attack from a parametrized variational quantum circuit. In the scenario without adversarial attacks, an input state $|\psi\rangle_{\text{in}}$ is input to the parametrized variational classifier $V(\Theta)$ directly, while in the adversarial scenario, a parametrized adversarial variational circuit $U(\boldsymbol{\theta})$ is used to add an evasion attack [72], as sketched in the upper panel. In the lower panel, a random unitary encoder E and the corresponding decoder $E^\dagger$ are added before and after $U(\boldsymbol{\theta})$ to protect the data sample against potential adversary Eve. (b) By using a random global unitary encoder, the landscape for any adversarial circuit exhibits a barren plateau (i.e., vanishing gradients) regardless of the inner structure of the circuit. The variables θ_1 and θ_2 are variational parameters of $U(\boldsymbol{\theta})$.

corresponding moments with respect to the Haar measure $d\mu_H(E)$,

$$\sum_i p_i E_i^{\otimes t} M E_i^{\dagger\otimes t} = \int d\mu_H(E) E^{\otimes t} M E^{\dagger\otimes t}, \quad t = 1, 2, \quad (3)$$

where M is an arbitrary operator. As shown in Refs. [104–107], quantum circuits can implement unitary 2-design efficiently—a circuit with only $O(n^2)$ $[O(n)]$ gates is sufficient for attaining exact (approximate) unitary 2-design. The type of gates in the quantum circuits can be further restricted to single-qubit rotations and nearest-neighbor entangling gates. Therefore, such a random encoder $E_i \in C$ can be efficiently realized by a PVQC with $O(n^2)$ gates. In this case, the loss function given a fixed encoder E_i can be represented by

$$L(\Theta, E_i; \boldsymbol{\theta}) = \langle\psi|_{\text{in}} E_i^\dagger U^\dagger E_i V^\dagger H V E_i^\dagger U E_i |\psi\rangle_{\text{in}}, \quad (4)$$

where $U \equiv U(\boldsymbol{\theta})$ and $V \equiv V(\Theta)$ are parametrized with $\boldsymbol{\theta}$ and Θ, respectively. In the adversarial setting, we assume that the adversarial PVQC is initialized with $\boldsymbol{\theta}_0$ such that $U(\boldsymbol{\theta}_0) = I$, i.e., the adversary starts from a legitimate quantum sample and explores the gradient direction to maximize the value of the loss function. We denote $\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta})$ to be the gradient of $L(\Theta, E_i; \boldsymbol{\theta})$ with respect to each parameter θ_l, $l = 1, ..., L$ in the adversarial PVQC. Now, we are ready to present our first theorem regarding the expectation and variance on each $\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta})$.

Theorem 1. Suppose we exploit a randomly chosen global unitary encoder E_i from a unitary 2-design codebook $C = \{p_i, E_i\}$. The expectation and variance of the derivatives of the loss function defined in Eq. (4) with respect to any component $\theta_l \in \boldsymbol{\theta}$ satisfy the following (in)equalities:

$$\mathbb{E}_{E_i \in C}\left[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)\right] = 0, \quad (5)$$

$$\text{Var}_{E_i \in C}\left[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)\right] \leqslant \frac{2\text{Tr}\left(A_l^2\right)}{d^2 - 1} \text{Tr}\left(\rho H_V^2\right), \quad (6)$$

where $\boldsymbol{\theta}_0$ are the initial parameters for the adversarial PVQC with $U(\boldsymbol{\theta}_0) = I$, $\rho = |\psi\rangle_{\text{in}}\langle\psi|_{\text{in}}$ is the density matrix of the input state, A_l is the Hermitian operator of the parametrized variational component in the lth layer, $H_V = V^\dagger H V$, and $d = 2^n$ is the dimension of the Hilbert space.

Sketch of proof. We give a brief sketch of the essential idea here. The full proof is technically involved and thus left to Appendix A. As we assume that the random encoder $C = \{p_i, E_i\}$ satisfies the unitary 2-design properties and $U(\boldsymbol{\theta}_0) = I$, we can obtain the expectation and variance of the gradients by calculating the first and second moments using the Haar integral. To derive the analytical results, the Haar integrals are calculated by Schur-Weyl duality [108]. We first prove that the $\mathbb{E}_{E_i \in C}[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)]$ is an integral over the first moment and thus vanishes. We then calculate the variance of the gradient using $\text{Var}_{E_i \in C}[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)] = \mathbb{E}_{E_i \in C}[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)^2] - \mathbb{E}_{E_i \in C}[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)]^2 = \mathbb{E}_{E_i \in C}[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)^2]$. The variance can thus be obtained by a second moment integral, which results in an exponentially small value and yields Eq. (6). ■

This theorem guarantees that by choosing a random unitary encoder from the codebook C, we can bound the variance of gradients for any parameter in any potential adversarial PVQC circuits with an exponentially small value. By using the Chebyshev's inequality, this theorem indicates that the probability of finding a gradient along any direction of amplitude larger than a fixed constant is exponentially small. Specifically, with probability at least $1 - \delta$, the absolute value of the gradient $\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)$ is upper bounded by the exponentially small value $\sqrt{[2\text{Tr}(A_l^2)\text{Tr}(\rho H_V^2)]/[\delta(d^2 - 1)]}$. It has been proved in Ref. [32] that the vulnerability of a quantum classifier also grows exponentially with system size n and perturbations of only $O(\sqrt{1/d})$, $d = 2^n$ can render a considerable adversarial risk. However, this result does not crack the security guarantee in our protocol because we prove that the gradient vanishes at a more rapid speed $O(1/d)$. The

exponentially small gradients for the adversarial PVQC lead to the barren plateau, which requires exponentially large precision and iteration steps for the adversary that exploits any gradient-based algorithm to construct an adversarial example. Therefore, this algorithm protects the quantum machine learning systems by masking the gradient information from the attackers. We emphasize that our protection encoder can be efficiently realized using a circuit containing only $O(n^2)$ gates to satisfy the unitary 2-design requirement, which is roughly the same scaling as most quantum classifiers commonly used in practice.

We also stress that the adversarial PVQC is restricted to a small neighborhood of the identity operator, thus itself does not satisfy unitary 2-design. The barren plateaus faced by the adversary are induced by our random encoding scheme. This is in sharp contrast to the barren plateaus for variational quantum circuits studied in the previous literature [46,47], where the variational quantum circuits themselves are required to be unitary 2-design.

Theorem 1 can be further extended to other codebooks. For example, we consider another model, where the encoder E_i can be written as a tensor product of m-qubit blocks ($m < n$) with each block satisfying unitary 2-design. We show that by using these encoders, one can create barren plateaus for the adversary PVQC with a less stringent requirement. Without loss of generality, we assume that $n = m\xi$ and $E_i = \bigotimes_{j=1}^{\xi} E_i^j$ such that ensemble $\{p_i^j, E_i^j\}$ forms a unitary 2-design for all j. We can similarly decompose the operator A_l in each layer of the adversarial PVQC as

$$A_l = \sum_k c_k \bigotimes_{j=1}^{\xi} A_{l,k}^j. \tag{7}$$

We assume that $\sum_{k,k'} c_k c_{k'}$ is bounded, $\mathrm{Tr}(A_{l,k}^{j2}) \leqslant 2^m, \forall l, i$, and $A_{l,k}^j$ is traceless. We remark that this assumption is reasonable, in the sense that it is satisfied by most commonly used quantum variational circuits (e.g., those in [47,49]). We have the following theorem:

Theorem 2. Assume we exploit a randomly chosen encoder E_i, which can be written as the tensor product of ξ m-qubit blocks independently chosen from unitary 2-design codebook $C = \{p_i^j, E_i^j\}$ ($j = 1, ..., \xi$). We assume the operators in adversarial PVQC can be decomposed as Eq. (7) and $A_{l,k}^j$ is traceless with $\mathrm{Tr}(A_{l,k}^{j2}) \leqslant 2^m, \forall l, i, k$. The expectation and variance of the derivatives of the loss function defined in Eq. (4) with respect to any component $\theta_l \in \boldsymbol{\theta}$ satisfies the following (in)equalities:

$$\mathbb{E}_{E_i^j \in C}\left[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)\right] = 0, \tag{8}$$

$$\mathrm{Var}_{E_i^j \in C}\left[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)\right] \leqslant O\left(\left(\frac{2^m + 1}{2^{2m} - 1}\right)^{\xi}\right). \tag{9}$$

Sketch of proof. We sketch the main idea for the proof here and leave the technical details in Appendix A. As we assume that each block E_i^j in the encoder satisfies unitary 2-design independently, we can obtain the expectation and variance of the loss function by calculating the Haar integral separately on each E_i^j. According to the decomposition in Eq. (7), we regard A_l as a summation of terms that are tensor products of operators on each block and calculate these terms separately. In the first step, we derive the zero expectation on the gradients in Eq. (8) by calculating the first moment similar to Theorem 1. Next, we compute the variance of the gradients by calculating the second moment Haar integral for each E_i^j. The result for the integral contains 2^{ξ} terms. We can derive the upper bound for the variance in Eq. (9) based on the assumption that $A_{l,k}^j$ is traceless with $\mathrm{Tr}(A_{l,k}^{j2}) \leqslant 2^m, \forall l, i, k$. ■

Theorem 2 indicates that, under reasonable assumptions on the adversarial PVQC, even if the encoder only satisfies unitary 2-design on each of the subspace $SU(2^m)$ for any $m \geqslant 2$, the variance of the gradients for adversarial PVQC still decreases exponentially as the system size increases. By exploiting this scheme, we can reduce the gate count required in Theorem 1 from $O(n^2)$ to $O(\xi m^2) = O(n)$. Compared with Theorem 1, the codebook requires fewer experimental resources at the price of a larger upper bound on the variance for the gradients. We mention that this encoding scheme carries over to adversarial PVQCs with other inner structures, although we can only analytically derive the variance bound under some constraints for the adversary due to technical difficulties.

We stress that our approach does not rely on any specific properties of the quantum classifiers $V(\Theta)$. It does not require that $V(\Theta)$ is unitary 2-design and applies to arbitrary quantum classifiers. Therefore, we can avoid the barren plateau landscape when training the quantum classifier by using shallow circuits or some quantum circuits with specific structures that are not unitary 2-design, such as quantum convolutional neural networks [49,91]. Even though we only rigorously prove the case for the loss function that can be regarded as an expectation value over the Hermitian operator H, our method can also effectively protect quantum classifiers equipped with other loss functions. This claim is supported by the numerical results using Kullback-Leibler (KL) divergence [109] in the subsequent paragraphs.

To verify that the scaling results in the above theorem are valid for quantum machine-learning models with modest system sizes and different loss functions, we carry out numerical simulations on classifying quantum phases of matter, which is widely concerned for quantum classification tasks [110]. In particular, we consider the ground states of the cluster-Ising model [61,62],

$$H(\lambda) = -\sum_{i=2}^{n-1} \sigma_{i-1}^x \sigma_i^z \sigma_{i+1}^x + \lambda \sum_{i=1}^{n} \sigma_i^y \sigma_{i+1}^y, \tag{10}$$

where $\sigma_i^{\alpha}, \alpha = x, y, z$ denotes the Pauli matrices on the ith qubit and λ is the interaction strength. Here, we take the open boundary condition. This model features a phase transition at $\lambda = 1$, between the cluster phase for $0 < \lambda < 1$ and the antiferromagnetic phase for $\lambda > 1$. We sample the Hamiltonian with a different parameter λ from 0 to 2 and compute the corresponding ground states. We then construct the data set using these ground states with the corresponding labels. We carry out the classification task using variational quantum classifiers of varying systems sizes from four to fourteen qubits and circuit depth being ten. We consider two types of

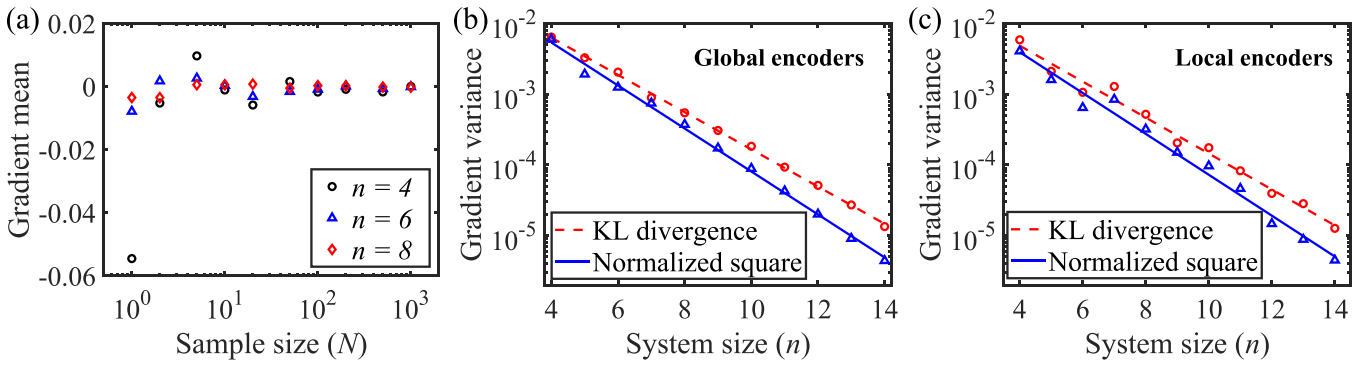

FIG. 3. Numerical results for the gradients of adversarial variational circuits. (a) The mean values of $\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)$ as functions of sample size N for different system sizes n. The loss function is taken as the KL divergence. The mean values of gradients are averaged over all parameters in the adversarial parametrized variational quantum circuits. (b) The average variances of $\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)$ for the KL divergence and normalized square loss as functions of the system size n. The encoders used here are global random parametrized variational quantum circuits. (c) Similar to (b), whereas the encoders used here are tensor products of two-qubit random parametrized variational blocks.

loss functions for the classifiers: (i) the normalized square loss $1 - |\langle\phi|\psi\rangle_{\text{out}}|^2$, where $|\psi\rangle_{\text{out}}$ is the output state at the end of the circuit in Fig. 2 and $|\phi\rangle$ is the state encoded by the target labels; (ii) the KL divergence between $|\psi\rangle_{\text{out}}$ and $|\phi\rangle$. We construct the encoder via a PVQC of four layers and sample the gradients from an adversarial PVQC of four layers. The results are obtained by averaging over variational encoders and adversarial PVQC with random parameters and input data samples. Further details for numeric results are provided in Appendix B. As shown in Fig. 3(a), the expectation values of the gradient along any directions in the adversarial PVQC converges to zero rapidly as we increase the number of samples, which is consistent with Eq. (5). From Fig. 3(b), we can observe that the variance of the gradients decays exponentially as the system size increases from four to fourteen qubits. The outcome from this numerical simulation fits the results for the global encoder setting given by Eq. (6). In Fig. 3(c), we perform numerical experiments for the local encoder settings at $m = 2$ in Eq. (9). We construct the encoder by using a PVQC that can be written as a tensor product of two-qubit blocks (each satisfying unitary 2-design) by randomly changes the parameters within the block. The two-qubit blocks are set to be a two-layer variational quantum circuit with the inner structure described in Appendix B. We observe that the variance of gradient approaches zero rapidly as the system size increases. The numerical result shows the exponential decay of gradients predicted in Eq. (9).

IV. DEFENDING LOCAL ADVERSARIAL NOISES BY BLACK-BOX QUANTUM ERROR CORRECTION

As we mentioned in the previous section, the adversarial perturbation can be regarded as experimental noises in the worst case. Under realistic experimental settings, most operations and noises are local [7,111]. Therefore, in this section we consider the case in which both the adversarial perturbation and state preparation can be written as tensor products of single-qubit rotations [30]. This setting is widely employed in qubit-encoding quantum computation and machine learning [112]. We consider the quantum classifier model C mentioned in Sec. II. We first analytically evaluate the vulnerability of quantum classifiers against such local adversarial perturbations. We suppose the quantum classifier $h : \mathcal{H} \to S$ maps the locally encoded data from $\bigotimes_{i=1}^{n} SU(2)$ to a label set $S = \{s_1, .., s_K\}$ that contains K labels. We assume that the input data sample g_ρ is chosen from $\mathcal{H}$ according to a probability measure $\mu(\cdot)$. We denote $\mu(h^{-1}(s_k))$ to be the fraction of data that will be assigned the label s_k by the classifier. We now introduce the following measure of adversarial risk:

Definition 1. Consider a hypothesis function $h : \mathcal{H} \to S$. Suppose the input data $|\psi\rangle$ is chosen from $\mathcal{H}$ according to the measure $\mu(\cdot)$ and an adversarial attack $|\psi\rangle \to |\psi\rangle', \forall |\psi\rangle \in \mathcal{H}$ occurs under the constraint $d(|\psi\rangle, |\psi\rangle') \leqslant \epsilon$. We denote $M = \{|\psi\rangle \in \mathcal{H} \mid h(|\psi\rangle) \neq h(|\psi\rangle')\}$ to be the set containing all the states that can be made as adversarial data samples. The adversarial risk is defined as $\mu(M)$.

We consider the set of input states that can be encoded by a local unitary operator on a certain initial state (e.g., the $|0\rangle^{\otimes n}$ state) and thus the classification of the quantum data is equivalent to the classification of special unitary groups $\bigotimes_{i=1}^{n} SU(2)$. For technical simplicity, we assume that the input data sample g_ρ is uniformly chosen from $\mathcal{H}$ according to the Haar measure for each qubit $\mu_H^{\otimes n}(\cdot)$ and denote $\mu_H^{\otimes n}(h^{-1}(s_k))$ to be the fraction of data that will be assigned the label s_k by the classifier. For two states $\rho = g_\rho |0\rangle^{\otimes n}$ and $\sigma = g_\sigma |0\rangle^{\otimes n}$, where $g_\rho = \bigotimes_{i=1}^{n} g_\rho^i$ and $g_\sigma = \bigotimes_{i=1}^{n} g_\sigma^i$ are chosen from $\bigotimes_{i=1}^{n} SU(2)$, we exploit the normalized Hamming distance to measure the difference between g_ρ and g_σ:

$$d_{\text{NH}}(g_\rho, g_\sigma) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[g_\rho^i \neq g_\sigma^i\right]. \tag{11}$$

This normalized Hamming distance measures the fraction of unequal g_ρ^i and g_σ^i for all the qubits. We can then deduce the following theorem concerning the effectiveness of local unitary adversarial attack:

Theorem 3. Consider a quantum classifier that maps an input sample from $\bigotimes_{i=1}^{n} SU(2)$ to a K-label set $S = \{s_1, ..., s_K\}$. Suppose we choose an operator g_ρ from $\bigotimes_{i=1}^{n} SU(2)$ according to the Haar measure on each qubit $\mu_H^{\otimes n}(\cdot)$. Without loss of generality, we assume $\mu_H^{\otimes n}(h^{-1}(s_1)) \geqslant \mu_H^{\otimes n}(h^{-1}(s_2)) \geqslant ... \geqslant \mu_H^{\otimes n}(h^{-1}(s_K))$. There always exists a perturbation $g_\rho \to g_{\rho'}$

with $d_{\mathrm{NH}}(g_\rho, g_{\rho'}) \leqslant \tau$, such that the adversarial risk is greater than $R \in (0, 1)$ if

$$\tau^2 \geqslant \frac{1}{n} \min_{k=2,3,\ldots,K} \ln\left[\frac{4k}{\mu_H^{\otimes n}(h^{-1}(s_k))(1-R)}\right]. \quad (12)$$

Sketch of proof. We provide the intuition here and the technical details for the full proof are provided in Appendix. C. Notice that the input data space $\bigotimes_{i=1}^n SU(2)$, when equipped with the Haar measure $\mu_H^{\otimes n}(\cdot)$ and the normalized Hamming distance $d_{\mathrm{NH}}(\cdot)$, forms a $(2, n/2^n)$-Levy family [113,114]. By exploiting the concentration of measure phenomenon on the Levy family [115,116], we show that the measure of all data within distance τ from a subset $\mathcal{H}' \subseteq \bigotimes_{i=1}^n SU(2)$ can be bounded below by R', if $\tau^2 \geqslant \frac{1}{n}\ln[\frac{4}{\mu_H^{\otimes n}(\mathcal{H}')(1-R')}]$. We then use De Morgan's law to prove that in the case of a K-label classification, by choosing $k = 2, 3, \ldots, K$ that minimizes $\ln[\frac{4k}{\mu_H^{\otimes n}(h^{-1}(s_k))(1-R)}]$, we can bound the adversarial risk below by R. This ends the proof of the theorem. ■

The above theorem indicates that, for any quantum classifiers receiving input data of n qubits, an adversarial attack that only changes a fraction $O(\frac{1}{\sqrt{n}})$ of qubits will result in a moderate adversarial risk bounded below by R. As the system size n increases, the vulnerability of a quantum classifier becomes more severe even for local unitary adversarial attacks. It has been shown in Ref. [32] that in the setting of global encoding quantum data from SU(d) ($d = 2^n$) and global adversarial perturbation, an perturbation of $O(\frac{1}{\sqrt{d}})$ strength under the Hilbert-Schmidt distance measure can guarantee a moderate adversarial risk. Compared with this global case, the adversarial risk under a local unitary attack is not as severe since additional constraints has been assumed for possible attacks. However, for large quantum machine learning systems, Eq. (12) still shows that the prediction is unstable even under a tiny noise. We remark that Eq. (12) still holds for other distance measures, such as the normalized Hilbert-Schmidt distance $d_{\mathrm{NHS}}(g_\rho, g_\sigma) = \frac{1}{n}\sum_{i=1}^n d_{\mathrm{HS}}(g_\rho^i, g_\sigma^i)$ that calculates the average Hilbert-Schmidt distance between each qubit. This follows from the fact that the normalized Hilbert-Schmidt distance is always bounded above by the normalized Hamming distance.

QEC codes [63,64] are widely used in quantum computation to protect the computation from local noises and are believed to be a crucial building block in the future implementation of quantum computers. Inspired by this, it would be natural to think whether QEC codes can effectively protect quantum machine-learning systems from local unitary adversarial attacks. A typical QEC procedure contains an encoder E, an error correction circuit $\mathcal{C}$, and a decoder D. The encoder E encodes a logical quantum state $|\psi\rangle_L$ from the logical Hilbert space $\mathcal{H}_L$ into a physical quantum state $|\psi\rangle_p$ from physical Hilbert space $\mathcal{H}_p$. When errors occur, we perform an error correction on physical qubits to correct particular types of errors and use the decoder to recover the original logical state ρ_L. A popular choice for QEC is the $[n_0, k, t]$ code [63], which encodes each k logical qubits into n_0 physical qubits and is able to correct $\lfloor \frac{t}{2} \rfloor$ erroneous physical qubits for each logical qubit. Without loss of generality, we consider the case when $k = 1$ and $t = 3$. The corresponding QEC codes can correct one local error for each logical qubit. Since the QEC codes were originally designed to correct realistic experimental errors, it is natural to ask whether the QEC codes can help to defend the adversarial perturbations that act locally on the physical qubits. When the attackers have no prior knowledge of the encoding scheme, we can intuitively expect the QEC codes to correct a significant proportion of such adversarial errors.

We consider the model in Fig. 4(a). The adversarial settings are similar to Sec. III except that we assume both logical state $|\psi\rangle_L$ and the adversarial attack are local. To protect the quantum machine-learning systems from such adversarial attacks, we consider applying a QEC encoder after the state preparation stage and the corresponding decoder before the classification stage. The QEC encoder is a black-box oracle for the adversary and thus can be regarded as a *random* encoder that encodes each logical qubit into physical qubits and is able to correct particular types of local errors on these physical qubits. We remark that the assumption of a random QEC encoder can be experimentally practical. A straightforward approach is to permutate the physical qubits randomly such that the adversary does not know the corresponding encoding structures between logical qubits and physical qubits. The fact that short random circuits are good QEC codes [117] indicates that random QEC codes can also be realized using circuits only containing $O(n \log n)$ gates.

To quantify the enhancement of adversarial robustness, we adapt the idea of QDP and utilize it as a measure for adversarial robustness [67,118,119]. Differential privacy [66] is the property of an algorithm whose outputs cannot be distinguished when inputting neighboring data sets. We can thus measure the sensitivity and vulnerability of the algorithm when changing the input using the differential privacy. For two input data samples that are separated by a small distance, differential privacy bounds the distance of the outputs after the algorithm. A formal definition of QDP is given below:

Definition 2. (Quantum differential privacy [67]) Consider the quantum algorithm $\mathcal{Q}$ and a measurement $\mathcal{M}$ on its output. The algorithm $\mathcal{Q}$ is said to be $(\epsilon(\tau), \gamma)$-QDP if for all input quantum states ρ, σ that satisfy $d(\rho, \sigma) \leqslant \tau$, the following inequality holds for any possible subset Y of all possible outcomes of the measurement:

$$\Pr[\mathcal{M}(\mathcal{Q}(\rho)) \in Y] \leqslant e^{\epsilon(\tau)} \Pr[\mathcal{M}(\mathcal{Q}(\sigma)) \in Y] + \gamma, \quad (13)$$

where $\epsilon(\tau)$ is a function of distance τ.

For technical simplicity, we focus on the case $\gamma = 0$, referred to as ϵ-QDP. As shown in Fig. 4(b), if we consider two neighboring x and x' input quantum data, the ϵ-QDP property bounds the difference between the probability distributions $\{p_1, p_2, p_3, p_4\}$ and $\{p_1', p_2', p_3', p_4'\}$ after the classification by bounding each $p_i/p_i' \in (e^{-\epsilon}, e^{\epsilon})$. We focus on the locally encoded states in this section and exploit the normalized Hamming distance as a distance measure. It is shown in Refs. [44,67,118] that adding any amount of white noise can make the algorithm satisfy QDP under trace distance and normalized Hamming distance. Therefore, it is reasonable to assume that under experimental settings the quantum machine-learning system satisfies QDP, as we can always keep a tiny amount of quantum white noise whose influence is negligible. We remark that as distance τ increases, the corresponding $\epsilon(\tau)$ will always increase according to the

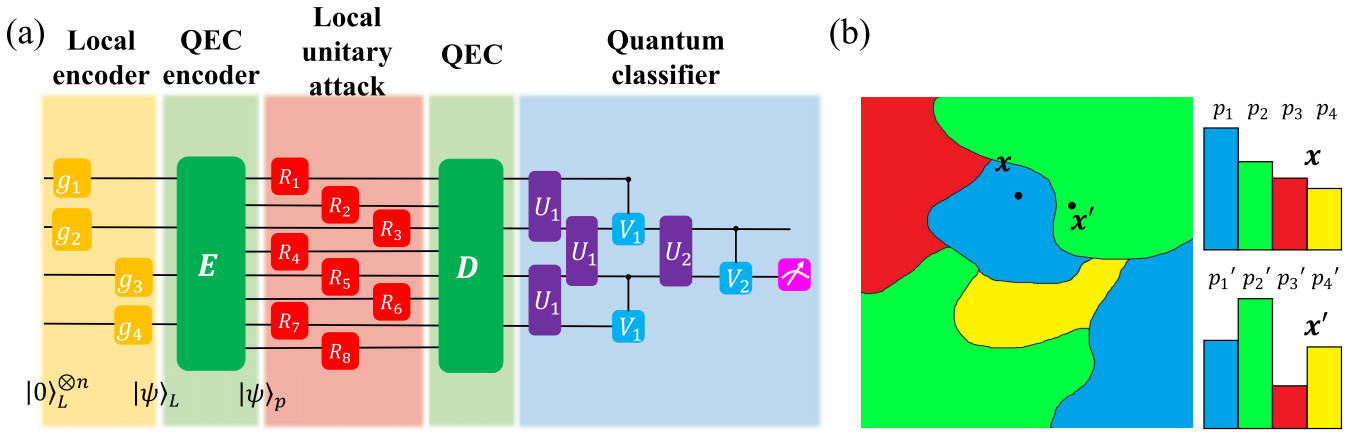

FIG. 4. (a) An illustration of exploiting black-box quantum error correction (QEC) encoders to defend against local unitary adversarial attacks. An initial state $|0\rangle^{\otimes n}$ is sequentially encoded by a local encoder and a QEC encoder into the logical state $|\psi\rangle_L$ and the physical state $|\psi\rangle_p$. The physical state is exposed to local adversarial attacks from potential adversaries. It then enters a QEC decoder and is classified by a quantum classifier. (b) A sketch of the connection between quantum differential privacy (QDP) and adversarial robustness. The quantum classifier maps input data x into a probability distribution $\{p_1, ..., p_4\}$ and predicts its label according to the maximum likelihood. The different labels are distinguished by different colors. Given a perturbed data x' with $d(x, x') \leqslant \tau$, ϵ-QDP limits the shift on probability distribution by bounding $|\ln(p_i'/p_i)| \leqslant \epsilon$ $(i = 1, \ldots, 4)$.

definition given by Eq. (13). For a quantum classification algorithm $\mathcal{Q} : \mathcal{H} \to \mathbb{R}^{|S|}$ that maps a quantum data sample to a probability distribution on the label set S, the QDP property in Definition 2 has a direct connection with the quantum adversarial risk in Definition 1. The adversarial risk for $\mathcal{Q}$ can be explicitly written as

$$\mathbb{E}_{|\psi\rangle\in\mathcal{H}}\left[\sup_{\delta:\|\delta\|\leqslant\tau}\Pr_{s_1\sim\mathcal{Q}(|\psi\rangle),s_2\sim\mathcal{Q}(|\psi\rangle+\delta)}(s_1\neq s_2)\right], \tag{14}$$

where δ is the adversarial perturbation and the expectation value is averaged over all choices of $|\psi\rangle$ according to the measure $\mu(\cdot)$. By Jensen's inequality, the probability for two random variables chosen from probability distribution $P = (p_1, ..., p_{|S|})$ and $Q = (q_1, ..., q_{|S|})$ having the same value can be bounded by $\sum_i p_i q_i \leqslant \exp(\sum_i p_i \log(q_i)) = \exp(-d_{KaL}(P, Q) - H(P))$. Here, $H(\cdot)$ is the Shannon entropy and $d_{\mathrm{KL}}(\cdot)$ is the KL divergence between two distributions. Noticing that the algorithm $\mathcal{Q}$ is $\epsilon(\tau)$-QDP, $P = \mathcal{Q}(|\psi\rangle)$, $Q = \mathcal{Q}(|\psi\rangle + \delta)$, and $\|\delta\| \leqslant \tau$, $|\ln(p_i/q_i)|$ is bounded by $\epsilon(\tau)$. As a result, the KL divergence between P and Q is bounded above by $2\epsilon(\tau)^2$ [120]. Therefore, we can derive the following information-theoretic upper bound for the adversarial risk of $\mathcal{Q}$ that is $\epsilon(\tau)$-QDP.

Proposition 1. Assume we have a quantum classifier $\mathcal{Q} : \mathcal{H} \to \mathbb{R}^{|S|}$ that satisfies $\epsilon(\tau)$-QDP. When performing an adversarial attack $|\psi\rangle \to |\psi\rangle'$ with $d(|\psi\rangle, |\psi\rangle') \leqslant \tau$, the adversarial risk $R(\tau)$ is bounded above by

$$R(\tau) \leqslant 1 - e^{-2\epsilon(\tau)^2}\mathbb{E}_{|\psi\rangle\in\mathcal{H}}[e^{-H(\mathcal{Q}(|\psi\rangle))}], \tag{15}$$

where the expectation value is averaged over all $|\psi\rangle$ chosen uniformly from $\mathcal{H}$ according to the measure $\mu(\cdot)$.

It is worthwhile to mention that $\mathbb{E}_{|\psi\rangle\in\mathcal{H}}[e^{-H(\mathcal{Q}(|\psi\rangle))}]$ is a constant that only depends on the property of the classifier $\mathcal{Q}$ itself. When the classifier is well-trained and can provide the correct label with a large confidence, $\mathbb{E}_{|\psi\rangle\in\mathcal{H}}[e^{-H(\mathcal{Q}(|\psi\rangle))}]$ is close to 1. In particular, if $\mathcal{Q}$ predicts the labels with a unity confidence, $\mathbb{E}_{|\psi\rangle\in\mathcal{H}}[e^{-H(\mathcal{Q}(|\psi\rangle))}] = 1$. As we decrease $\epsilon(\tau)$, the adversarial risk $R(\tau)$ will decrease polynomially. This indicates that one can improve the adversarial robustness by simply amplifying the QDP property (i.e., decreasing the ϵ parameter). This leads us to the next theorem concerning the amplification of QDP using the black-box QEC encoding:

Theorem 4. Suppose we use a quantum classifier that satisfies $\epsilon(\tau)$-QDP under the normalized Hamming distance. We apply a QEC encoder which encodes each logical qubit into n_0 physical qubits and is able to correct an arbitrary error on one of these qubits. Assume that the inner structure of the QEC code is unknown by the adversary. We randomly choose arbitrary ρ and σ on the physical qubits with $d_{\mathrm{NH}}(\rho, \sigma) \leqslant \tau$, then for any subset Y of all possible outcomes of the measurement $\mathcal{M}$, $\Pr[\mathcal{M}(\mathcal{Q}(\rho)) \in Y] \leqslant e^{\epsilon_{\mathrm{QEC}}}\Pr[\mathcal{M}(\mathcal{Q}(\sigma)) \in Y]$, where

$$\epsilon_{\mathrm{QEC}} = \epsilon\left(\frac{n_0(n_0-1)\tau^2}{\delta}\right), \tag{16}$$

with probability at least $1 - \delta$.

Proof. The normalized Hamming distance measures the fraction of qubits that are different from the legitimate data. Assume an adversarial attack $\rho \to \rho'$ occurs on the physical qubits such that $d_{\mathrm{NH}}(\rho, \rho') \leqslant \tau$. Under a black-box QEC procedure, a logical qubit becomes erroneous if it contains more than two erroneous physical qubits. Therefore, the expected fraction of logical qubits affected is $O(n_0(n_0-1)\tau^2)$ [63]. As a consequence, to achieve the bound in Theorem 3 on the logical qubits, the adversary should alter at least $O(\frac{1}{n^{1/4}})$ fraction of physical qubits in expectation. The QEC encoder mitigates the adversarial risk in expectation.

By the Markov's inequality, one can choose ρ and ρ' from physical quantum states such that $d_{\mathrm{NH}}(\rho, \rho') \leqslant \tau$ and

$$d_{\mathrm{NH}}(D \circ \mathcal{C}(\rho), D \circ \mathcal{C}(\rho')) \leqslant \frac{n_0(n_0-1)\tau^2}{\delta}, \tag{17}$$

with probability at least $1-\delta$. By the definition of QDP, it is guaranteed with probability at least $1-\delta$ that $\Pr[\mathcal{M}(\mathcal{Q}(\rho)) \in Y] \leqslant e^{\epsilon_{\rm QEC}} \Pr[\mathcal{M}(\mathcal{Q}(\rho')) \in Y]$, where $\epsilon_{\rm QEC} = \epsilon(\frac{n_0(n_0-1)\tau^2}{\delta})$. ■

The above theorem indicates that for a quantum classifier satisfying QDP, a black-box QEC encoder can effectively amplify the QDP property with high probability. In particular, the QEC encoder can promote the $\epsilon(\tau)$-QDP quantum classifier to a new quantum classifier that satisfies $\epsilon(\frac{n_0(n_0-1)\tau^2}{\delta})$-QDP property for at least $1-\delta$ fraction of all possible input data samples. Under the normalized Hamming distance, a QEC encoder can always promote the robustness of the quantum classifier against perturbations added on the input quantum states with large probability, as long as $\tau n_0(n_0-1)/\delta \leqslant 1$. As n_0 is a constant for a fixed QEC encoder, this is a constant threshold for τ. We remark that a quantum algorithm satisfying QDP can be obtained through adding white noise [44]. However, white noise will erase the information required for classification and should be suppressed for variational quantum classifiers on NISQ devices. In contrast, our approach only assumes the existence of tiny noise and can guarantee ϵ-QDP for arbitrary small $\epsilon > 0$ by concatenating QEC encoders. Compared with the previous work [44] that simply relies on the white noise to produce QDP property, our approach prevents the risk of losing too much information due to noises. Theorem 4 opens a door for studying the promotion of adversarial robustness from the perspective of QEC. Intuitively, a QEC encoder can mitigate the adversarial risk from the local unitary adversarial attack by reducing the bound in Eq. (12) to $O(\frac{1}{n^{1/4}})$ with a large probability. This is because the QEC can reduce the error rate from p to p^2 in expectation [63]. We mention that it is necessary to keep the inner structure of the QEC encoder confidential to the potential adversary. If the QEC encoder is known by the attacker, the QEC circuit together with the quantum classifier can be regarded as an enlarged quantum classifier with system size nn_0. According to Theorem 3, such a quantum classifier is more vulnerable under adversarial attacks.

In fault-tolerant quantum computation, the errors are assumed to occur locally on each qubit for each quantum operation independently with probability below the threshold $p_{\rm thres}$. To mitigate the influence of these errors, multiple levels of QEC codes are concatenated to bound the error below an expected value ζ. It has been proved that $O(\log\log(1/\zeta))$ levels of QEC codes are enough [63]. As we mentioned in the previous sections, adversarial perturbations are not random experimental noises. Instead, these perturbations are either carefully engineered noises from hostile adversary or worst-case experimental noises. However, Theorem 4 indicates that we can always decrease the ϵ parameter in the ϵ-QDP property of the quantum classifier by concatenating additional levels of QEC codes. We have shown in Theorem 3 that adversarial perturbations with strength $O(\frac{1}{\sqrt{n}})$ can lead to a moderate adversarial risk under local unitary adversarial attacks. To reduce the potential adversarial risk, we should bound the distance $\tau' = \Omega(\frac{1}{\sqrt{n}})$ after concatenating the QEC codes. We can thus deduce the following corollary.

Corollary 1. We consider the classifier $\mathcal{Q}$ discussed in Theorem 4. After concatenating $L_{\rm QEC}$ levels of QEC codes, one can guarantee with high probability that $\mathcal{Q}$ satisfies $\epsilon(\frac{1}{\sqrt{n}})$-QDP for randomly chosen ρ and σ with $d_{\rm NH}(\rho,\sigma) \leqslant \tau$, as long as

$$L_{\rm QEC} \geqslant O(\log\log n). \tag{18}$$

The proof of this corollary follows from using Theorem 4 repeatedly and fixing the δ in each level as $\frac{\delta}{L_{\rm QEC}}$. This corollary indicates that only $O(\log\log n)$ layers of repeated QEC encoders can guarantee $\epsilon(O(\frac{1}{\sqrt{n}}))$-QDP for the quantum classifier. The number of levels of QEC required has a double logarithmic scaling over the system size n. We show through this theorem that fault-tolerant quantum computers with black-box QEC codes are robust against local adversarial attacks with a large probability. This result shows the effectiveness of QEC under the condition of even the worst-case experimental noises.

V. CONCLUSIONS AND OUTLOOK

In this paper, we proposed an effective approach to protect quantum learning systems in adversarial scenarios using randomized encoders that are unknown by the attackers. We rigorously proved that random unitary encoders forming a unitary 2-design set can create barren plateaus for any adversarial PVQC, which prevent the creations of adversarial perturbations. To benchmark the performance of our approach, we carried out numerical simulations on classifying topological phases of ground states for the clustered-Ising Hamiltonian. We remark that this approach is feasible on NISQ devices as the classifiers, adversarial circuits, and encoders can be implemented by variational quantum circuits. In addition, we proved that black-box QEC encoders unknown to the adversary can mitigate the adversarial risk by promoting the differential privacy against local unitary noises. Our results develop versatile defense strategies to enhance the reliability and security of quantum learning systems, which may have far-reaching consequences in applications of quantum artificial intelligence based on both near-term and future quantum technologies.

Many questions remain and warrant further investigations. For instance, our discussions in this paper mainly focus on quantum supervised learning scenarios. Yet, unsupervised and reinforcement learning approaches may also suffer from the vulnerability problem [72]. Thus, it will be interesting and important to develop similar defense strategies in the context of quantum unsupervised or reinforcement learning, where obtaining analytical performance guarantees in a rigorous fashion might be more challenging. In addition, how to extend our results to the scenario of quantum delegated learning with multiple clients [121] is well worth future studies. Finally, it is also of crucial importance to carry out an experiment to demonstrate our defense strategies against adversarial perturbations. This would be a key step toward secure and reliable quantum artificial intelligence technologies.

ACKNOWLEDGMENTS

We thank L.-M. Duan, S. Choi, L. Yu, Z. Liu, Z. Lu, W. Jiang, and S. Jiang for helpful discussions. This work is supported by the National Natural Science Foundation of China

(Grants No. T2225008 and No. 12075128), the Innovation Program for Quantum Science and Technology (Grant No. 2021ZD0302203), the Tsinghua University Dushi Program, and the Shanghai Qi Zhi Institute.

APPENDIX A: BARREN PLATEAU INDUCED BY RANDOM UNITARY ENCODINGS

In this Appendix, we provide the detailed proofs for Theorem 1 and Theorem 2 in the main text. We give further analytical and numerical results concerning creating barren plateaus for adversarial PVQC. To begin, we provide the formal definition of unitary t design. Consider a polynomial $P_{t,t}(U)$ with the homogeneous degree at most t in the entries of a unitary matrix U, and degree t in the complex conjugates of these entries. We can evaluate the average of $P_{t,t}(U)$ under the Haar measure with $\sum_{i=1}^{M} p_i P_{t,t}(U_i)$, where $\{U_1, ..., U_M\}$ equipped with probability $\{p_1, ..., p_M\}$ is called to form the unitary t design. The formal definition [45,46,105] is given as follows.

Definition 3. Let $P_{t,t}(U)$ be a polynomial of unitary U and its complex conjugate $U^\dagger$, with up to a given degree t. An ensemble $\{p_i, U_i\}$, $i = 1, ..., M$ is called a unitary t design if

$$\sum_{i=1}^{M} p_i P_{t,t}(U_i) = \int P_{t,t}(U) d\mu_H(U) \tag{A1}$$

holds for any possible $P_{t,t}(U)$, where $d\mu_H(U)$ is the Haar measure.

The above expression can be either exact or approximate, which corresponds to the exact unitary t design or the approximate unitary t design, respectively. The unitary t design indicates that the tth moments are (approximately) the same as the corresponding moments with respect to the Haar measure. The first and second moments over the Haar measure are given by Weingarten functions [108]:

$$\int U^\dagger O U d\mu_H(U) = \frac{\mathrm{Tr}(O)}{d} I, \tag{A2}$$

$$\int U^\dagger A U X U^\dagger B U d\mu_H(U) = \frac{d\mathrm{Tr}(AB) - \mathrm{Tr}(A)\mathrm{Tr}(B)}{d(d^2-1)} \mathrm{Tr}(X) I + \frac{d\mathrm{Tr}(A)\mathrm{Tr}(B) - \mathrm{Tr}(AB)}{d(d^2-1)} X, \tag{A3}$$

where $d = 2^n$ is the dimension of the unitary U. Below, we prove Theorem 1 in the main text.

Consider the adversarial PVQC equipped with parameters $\boldsymbol{\theta}$ and the random unitary encoder satisfying the unitary 2-design as shown in Fig. 2(a). When we fix an encoder E_i and initialize the adversarial PVQC with $\boldsymbol{\theta}_0$ such that $U(\boldsymbol{\theta}_0) = I$, the gradient of the loss function L can be written as

$$\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0) = i \langle \psi|_{\mathrm{in}} E_i^\dagger U_-^\dagger [A_l, U_+^\dagger E_i V^\dagger H V E_i^\dagger U_+] U_- E_i |\psi\rangle_{\mathrm{in}}, \tag{A4}$$

where $U_- = \prod_{i=1}^{l-1} \exp(-iA_i\theta_i) W_i$ and $U_+ = \prod_{i=l}^{L} \exp(-iA_i\theta_i) W_i$. At $\boldsymbol{\theta}_0$, $U_+ U_- = I$. Since we have assumed the codebook C equipped with the probability distribution $\{p_i, E_i\}$ forms unitary 2-design, the average gradients at $\boldsymbol{\theta}_0$ are calculated as zero:

$$\begin{aligned}
\mathbb{E}_{E_i \in C}\left[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)\right] &= i \int d\mu_H(E) \langle \psi|_{\mathrm{in}} E^\dagger U_-^\dagger [A_l, U_+^\dagger E V^\dagger H V E^\dagger U_+] U_- E |\psi\rangle_{\mathrm{in}} \\
&= i\mathrm{Tr}\left\{ \int d\mu_H(E) |\psi\rangle_{\mathrm{in}} \langle \psi|_{\mathrm{in}} E^\dagger U_-^\dagger [A_l, U_+^\dagger E V^\dagger H V E^\dagger U_+] U_- E \right\} \\
&= \frac{i}{2^n} \mathrm{Tr}[\rho(\mathrm{Tr}(U_-^\dagger A_l U_+^\dagger) V^\dagger H V - \mathrm{Tr}(U_+ A_l U_-) V^\dagger H V)] = 0,
\end{aligned} \tag{A5}$$

where $\rho = |\psi\rangle_{\mathrm{in}} \langle \psi|_{\mathrm{in}}$. Next, we prove the exponential decay variances of the gradients. Since $\mathrm{Var}_{E_i \in C}[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)] = \mathbb{E}_{E_i \in C}[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)^2] - \mathbb{E}_{E_i \in C}[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)]^2 = \mathbb{E}_{E_i \in C}[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)^2]$, the variances can be calculated by the second moment integral:

$$\begin{aligned}
\mathbb{E}_{E_i \in C}\left[\partial_{\theta_l} L(\Theta, E_i; \boldsymbol{\theta}_0)^2\right] &= -\mathrm{Tr}\left\{ \int d\mu_H(E) \rho E^\dagger U_-^\dagger [A_l, U_+^\dagger E V^\dagger H V E^\dagger U_+] U_- E \rho E^\dagger U_-^\dagger [A_l, U_+^\dagger E V^\dagger H V E^\dagger U_+] U_- E \right\} \\
&= \frac{2}{d(d^2-1)} \left[d\mathrm{Tr}\left(A_l^2\right) - \mathrm{Tr}^2(A_l) \right] \left[\mathrm{Tr}\left(\rho H_V^2\right) - \mathrm{Tr}^2(\rho H_V) \right] \leqslant \frac{2}{d^2-1} \mathrm{Tr}\left(A_l^2\right) \mathrm{Tr}\left(\rho H_V^2\right),
\end{aligned} \tag{A6}$$

where $H_V = V^\dagger H V$. This completes the proof of Theorem 1 in the main text.

We then prove Theorem 2 in the main text. We utilize the following Haar measure integral over tensor-product unitary matrices [108]:

$$\int_{U(d_1)} ... \int_{U(d_\xi)} d\mu_H(U_1) ... d\mu_H(U_\xi) (U_1 \otimes ... \otimes U_\xi)^\dagger X (U_1 \otimes ... \otimes U_\xi) = \mathrm{Tr}(X) \bigotimes_{j=1}^{\xi} \frac{I_j}{d_j}, \tag{A7}$$

where d_j is the dimension of the unitary matrix U_j and I_j is the identity of d_j dimensions. Therefore, the expectations of the gradients in Eq. (A4) can be calculated as

$$\mathbb{E}_{E_i^j\in C}\big[\partial_{\theta_l}L(\Theta,E_i;\boldsymbol{\theta}_0)\big]=\int_{U(2^m)}...\int_{U(2^m)}d\mu_H(E^1)...d\mu_H(E^\xi)\partial_{\theta_l}L\big(\Theta,E=\otimes_{j=1}^{\xi}E^j;\boldsymbol{\theta}_0\big)=0. \tag{A8}$$

Next, we calculate the variances for the gradients by $\mathrm{Var}_{E_i^j\in C}[\partial_{\theta_l}L(\Theta,E_i;\boldsymbol{\theta}_0)]=\mathbb{E}_{E_i^j\in C}[\partial_{\theta_l}L(\Theta,E_i;\boldsymbol{\theta}_0)^2]$. We utilize the following subspace Haar measure integral:

$$\begin{aligned}
&\int_{U(d_1)}(U_1\otimes I_{2,...,\xi})^\dagger(A_1\otimes I_{2,...,\xi})(U_1\otimes I_{2,...,\xi})X(U_1\otimes I_{2,...,\xi})^\dagger(B_1\otimes I_{2,...,\xi})(U_1\otimes I_{2,...,\xi})d\mu_H(U_1)\\
&=\sum_{\nu,\nu'=1}^{d_2d_3...d_\xi}\left[\int_{U(d_1)}U_1^\dagger A_1U_1(X^1_{\nu,\nu'})U_1^\dagger B_1U_1d\mu_H(U_1)\right]\otimes|\nu\rangle\langle\nu'|\\
&=\frac{1}{d_1\big(d_1^2-1\big)}[(d_1\mathrm{Tr}(A_1B_1)-\mathrm{Tr}(A_1)\mathrm{Tr}(B_1))\mathrm{Tr}_1(X)\otimes I_{d_1}+(d_1\mathrm{Tr}(A_1)\mathrm{Tr}(B_1)-\mathrm{Tr}(A_1B_1))X].
\end{aligned} \tag{A9}$$

In the second line, we decompose the operator X into $\sum_{\nu,\nu'}X^1_{\nu,\nu'}\otimes|\nu\rangle\langle\nu'|$ using the orthogonal basis $\{|\nu\rangle:\nu=1,2,...,d_2d_3...d_\xi\}$. By repeatedly using the above formula, we can evaluate the Haar measure integral of the form $\int_{U(d_1)}...\int_{U(d_\xi)}(U_1\otimes...\otimes U_\xi)^\dagger(A_1\otimes...\otimes A_\xi)(U_1\otimes...\otimes U_\xi)X(U_1\otimes...\otimes U_\xi)^\dagger(B_1\otimes...\otimes B_\xi)(U_1\otimes...\otimes U_\xi)d\mu_H(U_1)...d\mu_H(U_\xi)$. Now, we are ready to compute the variances for the gradients in the adversarial PVQC $U(\boldsymbol{\theta})$. We divide the variance into four terms:

$$\begin{aligned}
\mathrm{Var}_{E_i^j\in C}\big[\partial_{\theta_l}L(\Theta,E_i;\boldsymbol{\theta}_0)\big]=&-\mathrm{Tr}\left[\int_{U(2^m)}...\int_{U(2^m)}d\mu_H(E^1)...d\mu_H(E^\xi)\rho E^\dagger U_-^\dagger A_lU_+^\dagger EH_V\rho E^\dagger U_-^\dagger A_lU_+^\dagger EH_V\right]\\
&-\mathrm{Tr}\left[\int_{U(2^m)}...\int_{U(2^m)}d\mu_H(E^1)...d\mu_H(E^\xi)\rho H_VE^\dagger U_+A_lU_-E\rho H_VE^\dagger U_+A_lU_-E\right]\\
&+\mathrm{Tr}\left[\int_{U(2^m)}...\int_{U(2^m)}d\mu_H(E^1)...d\mu_H(E^\xi)\rho E^\dagger U_-^\dagger A_lU_+^\dagger EH_V\rho H_VE^\dagger U_+A_lU_-E\right]\\
&+\mathrm{Tr}\left[\int_{U(2^m)}...\int_{U(2^m)}d\mu_H(E^1)...d\mu_H(E^\xi)\rho H_VE^\dagger U_+A_lU_-E\rho E^\dagger U_-^\dagger A_lU_+^\dagger EH_V\right],
\end{aligned} \tag{A10}$$

where $E=E^1\otimes...\otimes E^\xi$, $H_V=V^\dagger HV$, and $\rho=|\psi\rangle_{\mathrm{in}}\langle\psi|_{\mathrm{in}}$. We calculate each term in Eq. (A10) using Eqs. (A7) and (A9):

$$\begin{aligned}
&\mathrm{Tr}\left[\int_{U(2^m)}...\int_{U(2^m)}d\mu(E^1)...d\mu(E^\xi)\rho E^\dagger U_-^\dagger A_lU_+^\dagger EH_V\rho E^\dagger U_-^\dagger A_lU_+^\dagger EH_V\right]\\
&=\frac{1}{(2^{2m}-1)^\xi}\sum_{k,k'}c_kc_{k'}\sum_{J\subseteq\{1,...,\xi\}}\prod_{j\in J}\left(\mathrm{Tr}\big(A^j_{l,k}A^j_{l,k'}\big)-\frac{1}{2^m}\mathrm{Tr}\big(A^j_{l,k}\big)\mathrm{Tr}\big(A^j_{l,k'}\big)\right)\\
&\quad\times\prod_{j\notin J}\left(\mathrm{Tr}\big(A^j_{l,k}\big)\mathrm{Tr}\big(A^j_{l,k'}\big)-\frac{1}{2^m}\mathrm{Tr}\big(A^j_{l,k}A^j_{l,k'}\big)\right)\mathrm{Tr}\left[\mathrm{Tr}_{j\in J}(H_V\rho)\otimes\bigotimes_{j\in J}I_{2^m}\cdot H_V\rho\right],
\end{aligned} \tag{A11}$$

$$\begin{aligned}
&\mathrm{Tr}\left[\int_{U(2^m)}...\int_{U(2^m)}d\mu(E^1)...d\mu(E^\xi)\rho H_VE^\dagger U_+A_lU_-E\rho H_VE^\dagger U_+A_lU_-E\right]\\
&=\frac{1}{(2^{2m}-1)^\xi}\sum_{k,k'}c_kc_{k'}\sum_{J\subseteq\{1,...,\xi\}}\prod_{j\in J}\left(\mathrm{Tr}\big(A^j_{l,k}A^j_{l,k'}\big)-\frac{1}{2^m}\mathrm{Tr}\big(A^j_{l,k}\big)\mathrm{Tr}\big(A^j_{l,k'}\big)\right)\\
&\quad\times\prod_{j\notin J}\left(\mathrm{Tr}\big(A^j_{l,k}\big)\mathrm{Tr}\big(A^j_{l,k'}\big)-\frac{1}{2^m}\mathrm{Tr}\big(A^j_{l,k}A^j_{l,k'}\big)\right)\mathrm{Tr}\left[\mathrm{Tr}_{j\in J}(\rho H_V)\otimes\bigotimes_{j\in J}I_{2^m}\cdot\rho H_V\right],
\end{aligned} \tag{A12}$$

$$\begin{aligned}
&\mathrm{Tr}\left[\int_{U(2^m)}...\int_{U(2^m)}d\mu(E^1)...d\mu(E^\xi)\rho E^\dagger U_-^\dagger A_lU_+^\dagger EH_V\rho H_VE^\dagger U_+A_lU_-E\right]\\
&=\frac{1}{(2^{2m}-1)^\xi}\sum_{k,k'}c_kc_{k'}\sum_{J\subseteq\{1,...,\xi\}}\prod_{j\in J}\left(\mathrm{Tr}\big(A^j_{l,k}A^j_{l,k'}\big)-\frac{1}{2^m}\mathrm{Tr}\big(A^j_{l,k}\big)\mathrm{Tr}\big(A^j_{l,k'}\big)\right)\\
&\quad\times\prod_{j\notin J}\left(\mathrm{Tr}\big(A^j_{l,k}\big)\mathrm{Tr}\big(A^j_{l,k'}\big)-\frac{1}{2^m}\mathrm{Tr}\big(A^j_{l,k}A^j_{l,k'}\big)\right)\mathrm{Tr}\left[\mathrm{Tr}_{j\in J}(H_V\rho H_V)\otimes\bigotimes_{j\in J}I_{2^m}\cdot\rho\right],
\end{aligned} \tag{A13}$$

$$\begin{aligned}
&\mathrm{Tr}\left[\int_{U(2^m)}\cdots\int_{U(2^m)} d\mu(E^1)\ldots d\mu(E^\xi)\rho H_V E^\dagger U_+ A_l U_- E\rho E^\dagger U_-^\dagger A_l U_+^\dagger E H_V\right]\\
&=\frac{1}{(2^{2m}-1)^\xi}\sum_{k,k'}c_k c_{k'}\sum_{J\subseteq\{1,\ldots,\xi\}}\prod_{j\in J}\left(\mathrm{Tr}\left(A_{l,k}^j A_{l,k'}^j\right)-\frac{1}{2^m}\mathrm{Tr}\left(A_{l,k}^j\right)\mathrm{Tr}\left(A_{l,k'}^j\right)\right)\\
&\quad\times\prod_{j\notin J}\left(\mathrm{Tr}\left(A_{l,k}^j\right)\mathrm{Tr}\left(A_{l,k'}^j\right)-\frac{1}{2^m}\mathrm{Tr}\left(A_{l,k}^j A_{l,k'}^j\right)\right)\mathrm{Tr}\left[\mathrm{Tr}_{j\in J}(\rho)\otimes\bigotimes_{j\in J} I_{2^m}\cdot H_V\rho H_V\right],
\end{aligned}\tag{A14}$$

where the $\sum_{J\subseteq\{1,\ldots,\xi\}}$ sums over all possible subsets of $\{1,\ldots,\xi\}$ and the operator A_l in each layer of the adversarial PVQC is decomposed as $A_l=\sum_k c_k\bigotimes_{j=1}^{\xi}A_{l,k}^j$ [Eq. (7) of the main text]. By summing up these four terms, we obtain the variance of the gradient as

$$\begin{aligned}
\mathrm{Var}_{E_i^j\in C}\left[\partial_{\theta_l}L(\Theta,E_i;\boldsymbol{\theta}_0)\right]&=\left(\frac{1}{2^{2m}-1}\right)^\xi\sum_{k,k'}c_k c_{k'}\sum_{J\subseteq\{1,\ldots,\xi\}}\prod_{j\in J}\left(\mathrm{Tr}\left(A_{l,k}^j A_{l,k'}^j\right)-\frac{1}{2^m}\mathrm{Tr}\left(A_{l,k}^j\right)\mathrm{Tr}\left(A_{l,k'}^j\right)\right)\\
&\quad\times\prod_{j\notin J}\left(\mathrm{Tr}\left(A_{l,k}^j\right)\mathrm{Tr}\left(A_{l,k'}^j\right)-\frac{1}{2^m}\mathrm{Tr}\left(A_{l,k}^j A_{l,k'}^j\right)\right)\mathrm{Tr}\left[\mathrm{Tr}_{j\in J}(\rho)\otimes\bigotimes_{j\in J} I_{2^m}\cdot H_V\rho H_V+\mathrm{Tr}_{j\in J}(H_V\rho H_V)\right.\\
&\quad\left.\otimes\bigotimes_{j\in J} I_{2^m}\cdot\rho-\mathrm{Tr}_{j\in J}(\rho H_V)\otimes\bigotimes_{j\in J} I_{2^m}\cdot\rho H_V-\mathrm{Tr}_{j\in J}(H_V\rho)\otimes\bigotimes_{j\in J} I_{2^m}\cdot H_V\rho\right].
\end{aligned}\tag{A15}$$

We bound the following term with a constant C_0, which does not increase with the total system dimension:

$$\begin{aligned}
C_0=\max_{J\subseteq\{1,\ldots,\xi\}}\mathrm{Tr}&\left[\mathrm{Tr}_{j\in J}(\rho)\otimes\bigotimes_{j\in J} I_{2^m}\cdot H_V\rho H_V+\mathrm{Tr}_{j\in J}(H_V\rho H_V)\otimes\bigotimes_{j\in J} I_{2^m}\cdot\rho-\mathrm{Tr}_{j\in J}(\rho H_V)\right.\\
&\left.\otimes\bigotimes_{j\in J} I_{2^m}\cdot\rho H_V-\mathrm{Tr}_{j\in J}(H_V\rho)\otimes\bigotimes_{j\in J} I_{2^m}\cdot H_V\rho\right].
\end{aligned}\tag{A16}$$

According to the assumption of Theorem 2, $A_{l,k}^j$ is traceless and $\mathrm{Tr}(A_{l,k}^{j2})\leqslant 2^m$, $\forall l,i,k$. We have

$$\mathrm{Tr}\left(A_{l,k}^j A_{l,k'}^j\right)\leqslant\sqrt{\mathrm{Tr}\left(A_{l,k}^{i2}\right)\mathrm{Tr}\left(A_{l,k'}^{i2}\right)}\leqslant 2^m,\tag{A17}$$

$$\left|\mathrm{Tr}\left(A_{l,k}^j\right)\mathrm{Tr}\left(A_{l,k'}^j\right)-\frac{1}{2^m}\mathrm{Tr}\left(A_{l,k}^j A_{l,k'}^j\right)\right|\leqslant 1.\tag{A18}$$

Hence, we bound the variance as

$$\mathrm{Var}_{E_i^j\in C}\left[\partial_{\theta_l}L(\Theta,E_i;\boldsymbol{\theta}_0)\right]\leqslant\left(\frac{1}{2^{2m}-1}\right)^\xi\sum_{k,k'}c_k c_{k'}\sum_{J\subseteq\{1,\ldots,\xi\}}\prod_{j\in J}2^m C_0\tag{A19}$$

$$=\sum_{k,k'}c_k c_{k'}\left(\frac{2^m+1}{2^{2m}-1}\right)^\xi C_0,\tag{A20}$$

which finishes the proof for Theorem 2.

APPENDIX B: MORE NUMERICAL RESULTS

In this Appendix, we provide the details for our numerical simulations. The structure of PVQCs we used in numerical simulations is shown in Fig. 5(a). In this P-layer PVQC classifier, we first prepare the input state as an $(m+n)$-qubit state $|\psi\rangle_{\mathrm{in}}\otimes|0\rangle^{\otimes m}$, where $|\psi\rangle_{\mathrm{in}}$ is the quantum state that encodes the data to be classified and $|0\rangle^{\otimes m}$ are the ancillary qubits for measurement outputs. Then we apply P layers of unitary quantum operations with each layer containing two rotation units and one entangling unit. Each rotation unit performs an Euler rotation in the single-qubit Bloch sphere and each entangling unit entangles different qubits using CNOT gates between each pair of neighboring qubits. We can adjust the rotation angles and these angles are collectively regarded as variational parameters Θ. The final output state can be written as

$$|\psi(\Theta)\rangle=\left(\prod_{i=1}^{P}U_i\right)|\psi\rangle_{\mathrm{in}}\otimes|0\rangle^{\otimes m},\tag{B1}$$

where $U_i=[\prod_{j=1}^{n+m}Z(\theta_{i,d}^j)X(\theta_{i,c}^j)]U_{\mathrm{ent}}[\prod_{j=1}^{n+m}Z(\theta_{i,b}^j)X(\theta_{a,1}^j)]$ denotes the quantum operation for the ith layer and U_{ent} denotes the entangling unit. For adversarial attacks and encoders, we set $m=0$. For the quantum classifier, we employ $P=10$ for different system sizes. As the ground states

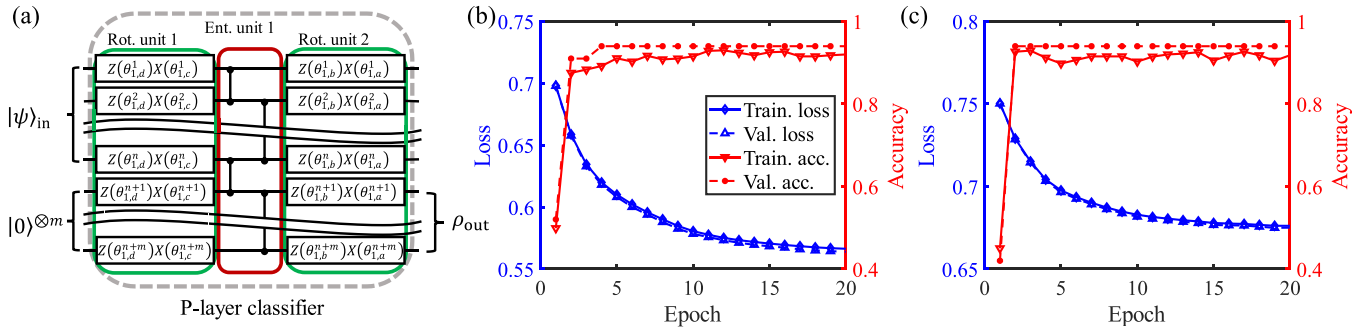

FIG. 5. (a) The illustration for P-layer variational quantum circuits to construct encoders and classifiers in the numerical simulations. Each layer contains two single-qubit rotation units (the green boxes) and one entangling unit (the red box). In each rotation unit, we perform and Euler angle rotation $Z(\theta^k_{i,u})X(\theta^k_{i,v})$. $(u, v) = (d, c)$ or (b, a) distinguishes the two rotation units. $i = 1, 2, ..., P$ denotes the index of the layer. $k = 1, 2, ..., m + n$ denotes the qubit index. (b), (c) The loss and accuracy averaged over the training set and the validation set as the function of epoch, for the 12-qubit quantum classifier with (b) the KL divergence and (c) the normalized square loss as the loss function. Each epoch consists of ten iterations.

of the cluster Ising model has two phases, we set $m = 1$. After the classifier, we measure the output ancillary qubit ρ_{out} and compute $\Pr(y = m) = \text{Tr}(\rho_{\text{out}}|m\rangle\langle m|)$, $m = 0, 1$. We assign a label $y = 0$ if $\Pr(y = 0) \geqslant \Pr(y = 1)$ and $y = 1$ otherwise.

For the adversarial PVQC $U(\boldsymbol{\theta})$, we require that the initial parameters satisfy $U(\boldsymbol{\theta}_0) = I$. Therefore, we employ an alternative version of variational quantum circuits. To guarantee that $U(\boldsymbol{\theta}_0) = I$, we add a complex conjugate entangling unit at the end of each layer, and each U_i in Eq. (B1) becomes $U_i = [\prod_{j=1}^{n+m} Z(\theta^j_{i,d})X(\theta^j_{i,c})]U_{\text{ent}}[\prod_{j=1}^{n+m} Z(\theta^j_{i,b})X(\theta^j_{a,1})]U^\dagger_{\text{ent}}$. We then initialize the adversarial PQVC with all rotation angles being zero.

In the numerical simulations, we exploit the quantum-adapted KL divergence $L_{KL}(h(|\psi\rangle_{\text{in}}; \Theta), \boldsymbol{p})$ and the normalized square loss $L_{NS}(h(|\psi\rangle_{\text{in}}; \Theta), |y\rangle)$:

$$L_{KL}(h(|\psi\rangle_{\text{in}}; \Theta), \mathbf{p}) = -\sum_{i=1}^{2} p_k \log q_k, \tag{B2}$$

$$L_{NS}(h(|\psi\rangle_{\text{in}}; \Theta), |y\rangle) = 1 - |\langle y|\rho_{\text{out}}|y\rangle|^2, \tag{B3}$$

where $\mathbf{q} = (q_1, q_2)$ denotes the diagonal elements of the output state ρ_{out} and $\mathbf{p} = (1, 0), (0, 1)$ for $y = 0, 1$. During the training procedure of the quantum classifier, we exploit the gradient-based Adam optimization algorithm [122,123] to minimize the empirical loss function $L_N(\Theta) = \frac{1}{N}\sum_{i=1}^{N} L(h(|\psi\rangle; \Theta), y)$ over N training samples. To calculate the gradients of the loss function, we employ the definition $\partial L_N(\theta)/\partial\theta = \lim_{\epsilon\to 0} \frac{1}{2\epsilon}[L_N(\theta + \epsilon) - L_N(\theta - \epsilon)]$ and estimate the value by choosing a small $\epsilon = 10^{-10}$. In Figs. 5(b) and 5(c), we display the averaged loss and accuracy in the training procedure for the 12-qubit classifier. The overfitting risk is low [124] as the loss and accuracy are close for validation data samples and training data samples. The numerical simulations in this paper were implemented based on the Yao.jl extension [125], the Flux.jl [126] and the Zygote.jl [127] packages using the Julia programming language [128].

APPENDIX C: ANALYTICAL DERIVATIONS FOR THE ADVERSARIAL RISK OF LOCAL UNITARY ATTACKS

In this Appendix, we give a detailed proof for Theorem 3 in the main text. We introduce the concepts of the concentration function and the Levy family, and some basic results regarding the adversarial machine learning.

Definition 4. For a subset $\mathcal{H}' \subseteq \mathcal{H}$, a *$\tau$-extension* for $\mathcal{H}'$ under the distance metric D is defined as $\mathcal{H}'_\tau = \{x \in \mathcal{H} | D(x, \mathcal{H}') \leqslant \tau\}$. The *concentration function* for a probability measure μ is defined as $\alpha(\tau) = 1 - \inf\{\mu(\mathcal{H}'_\tau) | \mu(\mathcal{H}) \geqslant \frac{1}{2}\}$. A d-dimensional space equipped with the distance metric D and probability measure μ is called an (l_1, l_2)-Levy family if

$$\alpha(\tau) = l_1 e^{-l_2\tau^2 d}. \tag{C1}$$

We next introduce the following lemma showing that $\bigotimes_{i=1}^{n}$ SU(2) equipped with the Haar measure on each SU(2) and the normalized Hamming distance is a Levy family.

Lemma 1. (Example 2.6 in Ref. [113]) Given a probability measure space $(\mathcal{X}, \mu)$, we consider the tensor product space $\mathcal{X}^{\otimes n}$ equipped with $\mu^{\otimes n}$ and the normalized Hamming distance d_{NH}. The ensemble $(\mathcal{X}^{\otimes n}, \mu^{\otimes n}, d_{\text{NH}})$ forms a $(2, \frac{n}{\dim(\mathcal{X})^n})$-Levy family with the concentration function satisfying:

$$\alpha(\tau) = 2e^{-\tau^2 n}. \tag{C2}$$

We refer to Refs. [113,114] for the proof of the above lemma. The above lemma implies that the tensor space $\bigotimes_{i=1}^{n} SU(2)$ with the Haar measure on each qubit $\mu_H^{\otimes n}$ and the normalized Hamming distance forms a $(2, \frac{n}{2^n})$-Levy family. We recap the following lemma regarding the lower bound of the measure of a τ-extension for a subspace $\mathcal{H}'$.

Lemma 2. (Theorem 3.6 in Ref. [115]) For a subspace $\mathcal{H}'$ chosen from a (l_1, l_2)-Levy family $(\mathcal{H}, D, \mu)$, $\dim(\mathcal{H}) = d$, the measure of a τ-extension of $\mathcal{H}'$ is greater than R if τ satisfies

$$\tau^2 \geqslant \frac{1}{l_2 d} \ln\left[\frac{l_1^2}{\mu(\mathcal{H}')(1 - R)}\right]. \tag{C3}$$

Proof. We briefly recap the proof given in Ref. [32,115]. We decompose τ into two parts $\tau = \tau_1 + \tau_2$. We choose τ_1 with $\mu(\mathcal{H}') > l_1 e^{-l_2 \tau_1^2 d}$. There are two cases concerning whether $\mu(\mathcal{H}') \leqslant \frac{1}{2}$:

(1) For the case when $\mu(\mathcal{H}') \leqslant \frac{1}{2}$, assuming $\mu(\mathcal{H}'_{\tau_1}) \leqslant \frac{1}{2}$, we have $\mu(\mathcal{H}\backslash\mathcal{H}'_{\tau_1}) \geqslant \frac{1}{2}$. For simplicity, we denote $\overline{\mathcal{H}}'_{\tau_1} = \mathcal{H}\backslash\mathcal{H}'_{\tau_1}$. By the definition of Levy family, we deduce that $\alpha(\tau_1) \geqslant 1 - \mu(\overline{\mathcal{H}}'_{\tau_1}) = \mu(\mathcal{H}') > \alpha(\tau_1)$, which leads to a contradiction.

(2) For the case when $\mu(\mathcal{H}') > \frac{1}{2}$, it is straightforward to see $\mu(\mathcal{H}'_{\tau_1}) > \frac{1}{2}$.

Therefore, by choosing such τ_1 we can guarantee that $\mu(\mathcal{H}'_{\tau_1}) > \frac{1}{2}$. Next, we consider the τ_2-extension of $\mathcal{H}'_{\tau_1}$ with $\tau_2^2 \geqslant \frac{1}{l_2 d} \ln[\frac{l_1}{(1-R)}]$. Applying the definition of the Levy family we can prove the lemma as $\mu(\mathcal{H}'_{\tau_1+\tau_2}) > 1 - \alpha(\tau_2) \geqslant R$ and $\tau^2 \geqslant \tau_1^2 + \tau_2^2 = \frac{1}{l_2 d} \ln[\frac{l_1^2}{\mu(\mathcal{H}')(1-R)}]$. ■

Now we start to prove Theorem 3 in the main text. We notice that $(\bigotimes_{i=1}^n SU(2), d_{\mathrm{NH}}, \mu_H^{\otimes n})$ forms a $(2, \frac{n}{2^n})$-Levy family and $\dim((\bigotimes_{i=1}^n SU(2)) = 2^n$. Hence, for any subspace $\mathcal{H}' \subseteq \bigotimes_{i=1}^n SU(2)$, any τ-extension of $\mathcal{H}'$ has measure at least R^* if

$$\tau^2 \geqslant \frac{1}{n} \ln\left[\frac{4}{\mu_H^{\otimes n}(\mathcal{H}')(1-R^*)}\right]. \tag{C4}$$

Given $k = 2, 3, ..., K$, any data sample in the intersection of the τ-extensions $\{h^{-1}(s_i)_\tau\}_{i=1}^k$ can be transformed into a data sample in any $h^{-1}(s_i)$, when a perturbation $\rho \to \rho'$ of amplitude τ occurs. The adversarial attack can thus change the labels for all the data samples in this intersection set. By the De Morgan's law, the measure of this intersection set satisfies

$$\begin{aligned}\mu_H^{\otimes n}\left(\cap_{i=1}^k h^{-1}(s_i)_\tau\right) &\geqslant 1 - \sum_{i=1}^k \mu_H^{\otimes n}(\mathcal{H}\backslash h^{-1}(s_i)_\tau) \\ &= \sum_{i=1}^k \mu_H^{\otimes n}(h^{-1}(s_i)_\tau) - (k-1). \end{aligned}\tag{C5}$$

After setting $R^* = \frac{k-1+R}{k}$ and $\mathcal{H}' = h^{-1}(s_k)$ in Eq. (C4), we deduce that adversarial risk is bounded below by $\mu_H^{\otimes n}(\cap_{i=1}^k h^{-1}(s_i)_\tau) \geqslant R$. By choosing the minimal value of all $k = 2, 3, ..., K$, we finish the proof for Theorem 3 in the main text.

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[2] M. Jordan and T. Mitchell, Machine learning: Trends, perspectives, and prospects, Science **349**, 255 (2015).

[3] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, Improved protein structure prediction using potentials from deep learning, Nature (London) **577**, 706 (2020).

[4] S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, R. Prudden, A. Mandhane, A. Clark, A. Brock, K. Simonyan, R. Hadsell, N. Robinson, E. Clancy, A. Arribas, and S. Mohamed, Skilful precipitation nowcasting using deep generative models of radar, Nature (London) **597**, 672 (2021).

[5] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, Mastering the game of go with deep neural networks and tree search, Nature (London) **529**, 484 (2016).

[6] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, Mastering the game of go without human knowledge, Nature (London) **550**, 354 (2017).

[7] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, Quantum supremacy using a programmable superconducting processor, Nature (London) **574**, 505 (2019).

[8] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu *et al.*, Quantum computational advantage using photons, Science **370**, 1460 (2020).

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

[10] S. D. Sarma, D.-L. Deng, and L.-M. Duan, Machine learning meets quantum physics, Phys. Today **72**, 48 (2019).

[11] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, Quantum Boltzmann machine, Phys. Rev. X **8**, 021050 (2018).

[12] X. Gao, Z.-Y. Zhang, and L.-M. Duan, A quantum machine learning algorithm based on generative models, Sci. Adv. **4**, eaat9004 (2018).

[13] A. W. Harrow, A. Hassidim, and S. Lloyd, Quantum algorithm for linear systems of equations, Phys. Rev. Lett. **103**, 150502 (2009).

[14] S. Lloyd, M. Mohseni, and P. Rebentrost, Quantum principal component analysis, Nat. Phys. **10**, 631 (2014).

[15] S. Lloyd and C. Weedbrook, Quantum generative adversarial learning, Phys. Rev. Lett. **121**, 040502 (2018).

[16] L. Hu, S.-H. Wu, W. Cai, Y. Ma, X. Mu, Y. Xu, H. Wang, Y. Song, D.-L. Deng, C.-L. Zou, and L. Sun, Quantum generative adversarial learning in a superconducting quantum circuit, Sci. Adv. **5**, eaav2761 (2019).

[17] M. Schuld and N. Killoran, Quantum machine learning in feature Hilbert spaces, Phys. Rev. Lett. **122**, 040504 (2019).

[18] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv:1411.4028.

[19] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. Obrien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[20] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New J. Phys. **18**, 023023 (2016).

[21] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio *et al.*, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

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

[23] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, Adversarial machine learning, in *Proceedings of the 4th ACM workshop on Security and artificial intelligence* (ACM, Chicago, 2011), pp. 43–58.

[24] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, Adversarial attacks and defences: A survey, arXiv:1810.00069.

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

[26] D. J. Miller, Z. Xiang, and G. Kesidis, Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks, Proc. IEEE **108**, 402 (2020).

[27] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in *Second International Conference on Learning Representations* (ICLR, Banff, 2014).

[28] I. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, in *Proceedings of the 3rd International Conference on Learning Representations* ((ICLR, San Diego, 2015).

[29] A. Kurakin, I. Goodfellow, and S. Bengio, Adversarial examples in the physical world, arXiv:1607.02533.

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

[31] W. Gong and D.-L. Deng, Universal adversarial examples and perturbations for quantum classifiers, Natl. Sci. Rev. **9**, nwab130 (2021).

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

[33] W. Ren, W. Li, S. Xu, K. Wang, W. Jiang, F. Jin, X. Zhu, J. Chen, Z. Song, P. Zhang *et al.*, Experimental quantum adversarial learning with programmable superconducting qubits, Nat. Comput. Sci. **2**, 711 (2022).

[34] J. Guan, W. Fang, and M. Ying, Robustness verification of quantum classifiers, in *International Conference on Computer Aided Verification* (Springer, Virtual Event, 2021), pp. 151–174.

[35] H. Liao, I. Convy, W. J. Huggins, and K. B. Whaley, Robust in practice: Adversarial attacks on quantum machine learning, Phys. Rev. A **103**, 042427 (2021).

[36] B. Li, C. Chen, W. Wang, and L. Carin, Certified adversarial robustness with additive noise, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (NeurIPS, Vancouver, Canada, 2019), pp. 9464–9474.

[37] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, Mitigating adversarial effects through randomization, in *Proceedings of the 6th International Conference on Learning Representations* ((ICLR, Vancouver, Canada, 2018).

[38] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, Countering adversarial images using input transformations, in *Proceedings of the 6th International Conference on Learning Representations* (ICLR, Vancouver, Canada, 2018).

[39] J. Cohen, E. Rosenfeld, and Z. Kolter, Certified adversarial robustness via randomized smoothing, in *Proceedings of the 36th International Conference on Machine Learning* (PMLR, Long Beach, 2019), pp. 1310–1320.

[40] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, Certified robustness to adversarial examples with differential privacy, in *2019 IEEE Symposium on Security and Privacy (SP)* (IEEE, San Francisco, 2019), pp. 656–672.

[41] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh, Towards robust neural networks via random self-ensemble, in *Proceedings of the European Conference on Computer Vision* (ECCV, Munich, Germany, 2018), pp. 369–385.

[42] R. Pinot, L. Meunier, A. Araujo, H. Kashima, F. Yger, C. Gouy-Pailler, and J. Atif, Theoretical evidence for adversarial robustness through randomization, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (NeurIPS, Vancouver, Canada, 2019), pp. 11860–11870.

[43] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, A survey of deep learning techniques for autonomous driving, J. Field Robot. **37**, 362 (2020).

[44] Y. Du, M.-H. Hsieh, T. Liu, D. Tao, and N. Liu, Quantum noise protects quantum classifiers against adversaries, Phys. Rev. Res. **3**, 023153 (2021).

[45] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, Symmetric informationally complete quantum measurements, J. Math. Phys **45**, 2171 (2004).

[46] 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).

[47] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nat. Commun. **12**, 1791 (2021).

[48] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, Effect of barren plateaus on gradient-free optimization, Quantum **5**, 558 (2021).

[49] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of barren plateaus in quantum convolutional neural networks, Phys. Rev. X **11**, 041011 (2021).

[50] A. Arrasmith, Z. Holmes, M. Cerezo, and P. J. Coles, Equivalence of quantum barren plateaus to cost concentration and narrow gorges, Quantum Sci. Technol. **7**, 045015 (2022).

[51] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, Nat. Commun. **12**, 6961 (2021).

[52] M. Cerezo and P. J. Coles, Higher order derivatives of quantum neural networks with barren plateaus, Quantum Sci. Technol. **6**, 035006 (2021).

[53] K. Sharma, M. Cerezo, L. Cincio, and P. J. Coles, Trainability of dissipative perceptron-based quantum neural networks, Phys. Rev. Lett. **128**, 180505 (2022).

[54] Z. Holmes, A. Arrasmith, B. Yan, P. J. Coles, A. Albrecht, and A. T. Sornborger, Barren plateaus preclude learning scramblers, Phys. Rev. Lett. **126**, 190501 (2021).

[55] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, Entanglement-induced barren plateaus, PRX Quantum **2**, 040316 (2021).

[56] T. L. Patti, K. Najafi, X. Gao, and S. F. Yelin, Entanglement devised barren plateau mitigation, Phys. Rev. Res. **3**, 033090 (2021).

[57] A. Uvarov and J. D. Biamonte, On barren plateaus and cost function locality in variational quantum algorithms, J. Phys. A: Math. Theor. **54**, 245301 (2021).

[58] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, An initialization strategy for addressing barren plateaus in parametrized quantum circuits, Quantum **3**, 214 (2019).

[59] C. Zhao and X.-S. Gao, Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus, Quantum **5**, 466 (2021).

[60] Z. Liu, L.-W. Yu, L.-M. Duan, and D.-L. Deng, Presence and absence of barren plateaus in tensor-network based machine learning, Phys. Rev. Lett. **129**, 270501 (2022).

[61] W. Son, L. Amico, R. Fazio, A. Hamma, S. Pascazio, and V. Vedral, Quantum phase transition between cluster and antiferromagnetic states, Europhys. Lett. **95**, 50001 (2011).

[62] P. Smacchia, L. Amico, P. Facchi, R. Fazio, G. Florio, S. Pascazio, and V. Vedral, Statistical mechanics of the cluster Ising model, Phys. Rev. A **84**, 022304 (2011).

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

[64] D. Gottesman, Stabilizer codes and quantum error correction, Ph.D. thesis, California Institute of Technology, 1997.

[65] C. Hirche, C. Rouzé, and D. S. França, Quantum differential privacy: An information theory perspective, IEEE Trans. Inf. Theory **69**, 5771 (2023).

[66] C. Dwork and J. Lei, Differential privacy and robust statistics, in *Proceedings of the forty-first annual ACM Symposium on Theory of Computing* (ACM, Bethesda, 2009), pp. 371–380.

[67] L. Zhou and M. Ying, Differential privacy in quantum computation, in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)* (IEEE, Santa Barbara, 2017), pp. 249–262.

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

[69] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, IEEE Signal Process. Mag. **29**, 82 (2012).

[70] I. Kononenko, Machine learning for medical diagnosis: History, state of the art and perspective, Artif. Intell. Med. **23**, 89 (2001).

[71] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, Adversarial attacks on medical machine learning, Science **363**, 1287 (2019).

[72] Y. Vorobeychik and M. Kantarcioglu, *Adversarial Machine Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning (Springer, Cham, 2018).

[73] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, Practical black-box attacks against machine learning, in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS 17 (ACM, Abu Dhabi,United Arab Emirates, 2017), pp. 506–519.

[74] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, Towards deep learning models resistant to adversarial attacks, in *Proceedings of the 6th International Conference on Learning Representations* (ICLR, Vancouver, Canada, 2018).

[75] N. Papernot, P. McDaniel, and I. Goodfellow, Transferability in machine learning: From phenomena to black-box attacks using adversarial samples, arXiv:1605.07277.

[76] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, The limitations of deep learning in adversarial settings, in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (IEEE, Saarbrucken, Germany, 2016), pp. 372–387.

[77] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec'17 (ACM, New York, 2017), pp. 15–26.

[78] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, Boosting adversarial attacks with momentum, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR, Salt Lake City, 2018), pp. 9185–9193.

[79] A. Kurakin, I. J. Goodfellow, and S. Bengio, Adversarial machine learning at scale, in *Proceedings of the Fifth International Conference on Learning Representations* (ICLR, Toulon, France, 2017).

[80] I. 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 Advances in Neural Information Processing Systems* (NeurIPS, Montreal, Canada, 2014), pp. 2672–2680.

[81] P. Samangouei, M. Kabkab, and R. Chellappa, Defense-gan: Protecting classifiers against adversarial attacks using generative models, in *Proceedings of the 6th International Conference on Learning Representations* (ICLR, Vancouver, Canada, 2018).

[82] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in *2016 IEEE Symposium on Security and Privacy (SP)* (IEEE, San Jose, 2016), pp. 582–597.

[83] G. Hinton, O. Vinyals, and J. Dean, Distilling the knowledge in a neural network, arXiv:1503.02531.

[84] W. Li and D.-L. Deng, Recent advances for quantum classifiers, Sci. China Phys. Mech. Astron. **65**, 220301 (2022).

[85] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Circuit-centric quantum classifiers, Phys. Rev. A **101**, 032308 (2020).

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

[87] M. Schuld, M. Fingerhuth, and F. Petruccione, Implementing a distance-based classifier with a quantum interference circuit, Europhys. Lett. **119**, 60002 (2017).

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

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

[90] D. Zhu, N. M. Linke, M. Benedetti, K. A. Landsman, N. H. Nguyen, C. H. Alderete, A. Perdomo-Ortiz, N. Korda, A. Garfoot, C. Brecque, L. Egan, O. Perdomo, and C. Monroe, Training of quantum circuits on a hybrid quantum computer, Sci. Adv. **5**, eaaw9918 (2019).

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

[92] K. H. Wan, O. Dahlsten, H. Kristjánsson, R. Gardner, and M. Kim, Quantum generalisation of feedforward neural networks, npj Quant. Inf. **3**, 36 (2017).

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

[94] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, A grover-search based quantum learning scheme for classification, New J. Phys. **23**, 023020 (2021).

[95] A. V. Uvarov, A. S. Kardashin, and J. D. Biamonte, Machine learning phase transitions with a quantum processor, Phys. Rev. A **102**, 012415 (2020).

[96] P. Rebentrost, M. Mohseni, and S. Lloyd, Quantum support vector machine for big data classification, Phys. Rev. Lett. **113**, 130503 (2014).

[97] C. Blank, D. K. Park, J.-K. K. Rhee, and F. Petruccione, Quantum classifier with tailored quantum kernel, npj Quant. Inf. **6**, 41 (2020).

[98] F. Tacchino, C. Macchiavello, D. Gerace, and D. Bajoni, An artificial neuron implemented on an actual quantum processor, npj Quant. Inf. **5**, 26 (2019).

[99] N. Wiebe and R. S. S. Kumar, Hardening quantum machine learning against adversaries, New J. Phys. **20**, 123019 (2018).

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

[101] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus, and M. Peev, The security of practical quantum key distribution, Rev. Mod. Phys. **81**, 1301 (2009).

[102] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning, Nat. Commun. **12**, 2631 (2021).

[103] H. Zhao, L. Lewis, I. Kannan, Y. Quek, H.-Y. Huang, and M. C. Caro, Learning quantum states and unitaries of bounded gate complexity, arXiv:2310.19882.

[104] A. W. Harrow and R. A. Low, Random quantum circuits are approximate 2-designs, Commun. Math. Phys. **291**, 257 (2009).

[105] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).

[106] F. G. Brandao, A. W. Harrow, and M. Horodecki, Local random quantum circuits are approximate polynomial-designs, Commun. Math. Phys. **346**, 397 (2016).

[107] A. W. Harrow and S. Mehraban, Approximate unitary t-designs by short random quantum circuits using nearest-neighbor and long-range gates, Commun. Math. Phys. **401**, 1531 (2023).

[108] L. Zhang, Matrix integrals over unitary groups: An application of Schur-Weyl duality, arXiv:1408.3782.

[109] S. Kullback and R. A. Leibler, On information and sufficiency, Ann. Math. Stat. **22**, 79 (1951).

[110] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill, Provably efficient machine learning for quantum many-body problems, Science **377**, eabk3333 (2022).

[111] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, M. Gong, C. Guo, C. Guo, S. Guo, L. Han, L. Hong, H.-L. Huang, Y.-H. Huo, L. Li, N. Li *et al.*, Strong quantum computational advantage using a superconducting quantum processor, Phys. Rev. Lett. **127**, 180501 (2021).

[112] V. Giovannetti, S. Lloyd, and L. Maccone, Quantum random access memory, Phys. Rev. Lett. **100**, 160501 (2008).

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

[114] M. Talagrand, Concentration of measure and isoperimetric inequalities in product spaces, Publ. Math. Inst. Hautes Etudes Sci. **81**, 73 (1995).

[115] 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 30th AAAI International Conference on Algorithmic Learning Theory* (AAAI, Honolulu, 2019), Vol. 98, pp. 1–29.

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

[117] W. Brown and O. Fawzi, Short random circuits define good quantum error correcting codes, in *2013 IEEE International Symposium on Information Theory* (IEEE, Istanbul, Turkey, 2013), pp. 346–350.

[118] S. Aaronson and G. N. Rothblum, Gentle measurement of quantum states and differential privacy, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (STOC, Phoenix, 2019), pp. 322–333.

[119] S. Arunachalam, A. B. Grilo, and H. Yuen, Quantum statistical query learning, arXiv:2002.08240.

[120] C. Dwork, G. N. Rothblum, and S. Vadhan, Boosting and differential privacy, in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science* (IEEE, Las Vegas, 2010), pp. 51–60.

[121] W. Li, S. Lu, and D.-L. Deng, Quantum federated learning through blind quantum computing, Sci. China Phys. Mech. Astron. **64**, 100312 (2021).

[122] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[123] J. R. Sashank, K. Satyen, and K. Sanjiv, On the convergence of adam and beyond, in *Proceedings of the 6th International Conference on Learning Representations* (ICLR, Vancouver, Canada, 2018).

[124] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. **15**, 1929 (2014).

[125] X.-Z. Luo, J.-G. Liu, P. Zhang, and L. Wang, Yao.jl: Extensible, efficient framework for quantum algorithm design, Quantum **4**, 341 (2020).

[126] M. Innes, Flux: Elegant machine learning with Julia, J. Open Source Softw. **3**, 602 (2018).

[127] M. Innes, Don't unroll adjoint: differentiating SSA-form programs, arXiv:1810.07951.

[128] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, Julia: A fresh approach to numerical computing, SIAM Rev. **59**, 65 (2017).