Quantum noise protects quantum classifiers against adversaries

Yuxuan Du,[1] Min-Hsiu Hsieh,[2] Tongliang Liu,[1] Dacheng Tao,[1] and Nana Liu[3,4,5,*]

[1]*UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, University of Sydney, Australia*
[2]*Hon Hai Quantum Computing Research Center, No. 32, Jihu Road, Neihu District,*
Taipei 114, Taiwan
[3]*Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China*
[4]*Ministry of Education, Key Laboratory in Scientific and Engineering Computing, Shanghai Jiao Tong University, Shanghai 200240, China*
[5]*University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai 200240, China*

(Received 18 August 2020; revised 9 March 2021; accepted 8 April 2021; published 27 May 2021)

Noise in quantum information processing is often viewed as a disruptive and difficult-to-avoid feature, especially in near-term quantum technologies. However, noise has often played beneficial roles, from enhancing weak signals in stochastic resonance to protecting the privacy of data in differential privacy. It is then natural to ask: Can we harness the power of quantum noise that is beneficial to quantum computing? An important current direction for quantum computing is its application to machine learning, such as classification problems. One outstanding problem in machine learning for classification is its sensitivity to adversarial examples. These are small, undetectable perturbations from the original data where the perturbed data is completely misclassified in otherwise extremely accurate classifiers. They can also be considered as worst-case perturbations by unknown noise sources. We show that by taking advantage of depolarization noise in quantum circuits for classification, a robustness bound against adversaries can be derived where the robustness improves with increasing noise. This robustness property is intimately connected with an important security concept called differential privacy, which can be extended to quantum differential privacy. For the protection of quantum data, this quantum protocol can be used against the most general adversaries. Furthermore, we show how the robustness in the classical case can be sensitive to the details of the classification model, but in the quantum case the details of the classification model are absent, thus also providing a potential quantum advantage for classical data. This opens the opportunity to explore other ways in which quantum noise can be used in our favor, as well as identifying other ways quantum algorithms can be helpful in a way which is distinct from quantum speedups.

DOI: 10.1103/PhysRevResearch.3.023153

I. INTRODUCTION

Noise in quantum information processing has long been viewed as a feature to avoid and remove, notably in quantum computation. However, in the noisy intermediate-scale quantum (NISQ) era of near-term quantum computing [1], the presence of noise is inevitable. The focus is both on reducing the effects of quantum noise, for example, using error mitigation [2,3], and for finding protocols whose integrity can nevertheless withstand this noise. However, a parallel approach can be taken to instead study noise under a positive lens. In classical information processing, noise is actively leveraged in many applications, including strengthening security and privacy using differential privacy [4], enhancing weak signals using stochastic resonance [5], improving signal resolution after truncating data with dithering [6], and speeding convergence rates in neural networks [7]. Can we look at quantum noise in this same positive light and use it to our advantage?

One important proposed application of these quantum devices is performing machine-learning tasks like classification [8,9] and classification algorithms can be less vulnerable against noise. One reason behind this is that classification only has few possible outputs and machine learning can still provide accurate classification in the classical world despite the messiness of real-life data like images and sound recordings. Indeed, a recent work [10] showed how quantum binary classifiers can be made robust against common sources of quantum noise by choosing a right encoding of classical data into quantum states.

However, despite being tolerant to small amounts of noise with known sources, classification algorithms are generally not protected against unknown worst-case noise sources, such as adversarial attacks. In fact, classification algorithms in machine learning are often very sensitive to adversarial attacks and this presents a key obstacle for the future development of classical machine learning [11]. These adversaries perturb the original data point by only a small undetectable amount, yet the new data point, known as an adversarial example, is completely misclassified in otherwise extremely accurate

[*] Nana.Liu@quantumlah.org

classifiers. This observation presents an impetus for the vibrant field called adversarial machine learning [12,13] and this has recently been extended to the quantum domain in adversarial quantum learning [14–16]. While many important methods focus on finding more robust versions of existing algorithms [17], including on quantum devices [14,16], this approach is generally vulnerable to counterattacks and doesn't provide theoretical guarantees against all possible adversaries [18].

We take a different approach that does not require inventing new algorithms to improve robustness yet can provide a robustness guarantee against any unknown perturbation, such as from an adversary. We begin from our intuition that noise is a kind of scrambling mechanism. It can scramble the effects of disturbances made to one's original data, for instance, by adversaries, thus diminishing the effects adversarial attacks can have. Therefore, we can ask whether noise, instead of hindering the computation, can, in fact, assist in the presence of adversarial attacks?

More specifically, noise in the classical realm has been associated with improving the privacy of algorithms, providing a property called differential privacy [4]. Differential privacy is the property of an algorithm whose output cannot distinguish small changes in the initial data set, like the presence or absence of one party's data point, hence in this way preserving privacy of that party. This is, in fact, the very property we want in making our algorithm robust against adversarial examples, which are small changes to the initial data set that induce misclassification. For instance, adding classical noise to induce differential privacy has been used to protect classical classifiers against adversaries in Ref. [19].

We demonstrate that by including depolarization in one's quantum circuit for classification, we can achieve quantum differential privacy and, in turn, be able to provide robustness bounds in the presence of adversaries which were not possible before. This is the most natural mechanism to exploit noise to protect quantum data, which appear in condensed-matter systems, quantum communication networks, quantum simulation, quantum metrology, and quantum control. In addition, we show how the robustness bound in the classical case can be sensitive to the details of the classification model but in the quantum case this bound is dependent only on the number of possible class categories and no other feature of the classification model. We later provide an example of how this property can be useful in a security application.

We begin by defining classification, adversarial examples, and differential privacy. Then we demonstrate how adding depolarization noise in quantum classifiers can induce quantum differential privacy which can, in turn, provide protection against adversarial examples.

II. BACKGROUND

We briefly review the classification problem in both the classical and quantum domains before introducing the concept of adversarial examples. We then define classical and quantum differential privacy, which we later employ as a key tool to achieve robustness of our classifier against adversarial examples.

A. Classification task

A classification task is a mapping from a set of classical or quantum input states to a label chosen from a finite set. If the size of this finite set is $K \geqslant 2$, we have a K-multiclass classification problem [20]. $K = 2$ is the special case of binary classification, e.g., given images of only ants or cicadas, to decide which picture belongs to which insect.

Definition 1 (K-multiclass classification). The algorithm $\mathcal{A} : \Sigma \to \mathcal{C}$ is called a K-multiclass classification algorithm if it maps the set of input states Σ onto the set $\mathcal{C} = \{0, ..., K-1\}$. Let the state $\sigma \in \Sigma$ and $C \in \mathcal{C}$. If $\mathcal{A}(\sigma) = C$, then C is the predicted class label assigned to σ.

In machine learning, the algorithm $\mathcal{A}$ does not need to be predefined and can instead be learned through a training data set $\mathcal{D}$. This data set $\mathcal{D} = \{\sigma_i, \mathbf{Y}(\sigma_i)\}_{i=1}^{M}$ consists of M pairs of input states σ_i and their corresponding class labels represented by the K-dimensional vector $\mathbf{Y}(\sigma_i)$. Its$k$th entry $\mathbf{Y}_k(\sigma_i) = 1$ if the class label of σ_i is k and every other entry of $\mathbf{Y}_k(\sigma_i)$ is zero otherwise. To learn $\mathcal{A}$, we first define a parameterized function $f(\boldsymbol{\theta}, \sigma_i) \in \mathbb{R}^K$, where $\boldsymbol{\theta}$ are free parameters that can be tuned. The learning happens as $\boldsymbol{\theta}$ is optimized to minimize the empirical risk,

$$\min_{\boldsymbol{\theta}} \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(f(\boldsymbol{\theta}, \sigma_i), \mathbf{Y}(\sigma_i)), \tag{1}$$

where $\mathcal{L}$ refers to a predefined loss function. The goal in learning is to minimize this empirical risk Eq. (1) for one's given training data set $\mathcal{D}$, where the optimized parameters are denoted $\boldsymbol{\theta}^*$. Given test state σ, we can define $\mathbf{y}(\sigma) = f(\boldsymbol{\theta}^*, \sigma)/\|f(\boldsymbol{\theta}^*, \sigma)\|_1$ as the score vector among K labels, where $\|\cdot\|_1$ denotes the l_1 norm and $\mathbf{y}(\sigma) \in \mathbb{R}^K$ is the normalized vector of $f(\boldsymbol{\theta}^*, \sigma)$. Then the kth entry of the vector function $f(\boldsymbol{\theta}^*, \sigma) = \mathbf{y}_k(\sigma) \in [0, 1]$ can be interpreted as the probability that σ is assigned the label k. Then the learned classification algorithm $\mathcal{A}$ outputs the class label C for an input state σ using the condition

$$C \equiv \arg\max_k \mathbf{y}_k(\sigma) \equiv \mathcal{A}(\sigma), \tag{2}$$

where the final class label C is decided by identifying the class label with the highest corresponding probability.

For the quantum K-multiclass classification task with quantum test state σ, we can employ a quantum circuit, see Fig. 1(a), to compute $\mathbf{y}(\sigma)$ instead of using a classical circuit. We can identify $\mathbf{y}_k(\sigma)$ to be the probability of the final measurement outcome of the quantum circuit being k,

$$\mathbf{y}_k(\sigma) = \mathrm{Tr}(\Pi_k \mathcal{E}(\sigma \otimes |a\rangle\langle a|)), \tag{3}$$

where Π_k is a positive-operator valued measure (POVM), $\mathcal{E}$ is a quantum operation that contains information about the trained parameters $\boldsymbol{\theta}^*$ [21], and $|a\rangle\langle a|$ is an ancilla. However, precise values of the probabilities $\mathbf{y}_k(\sigma)$ can only be obtained in the infinite sampling regime. This means that if only N measurements are allowed at the output of the circuit, we can only obtain an estimated value $\mathbf{y}_k^{(N)}(\sigma)$ of the output probabilities.

B. Adversarial examples

Adversarial examples are attacks on input examples to classification problems that lead to misclassification. In

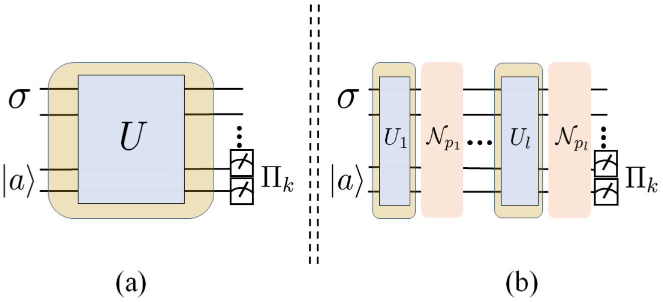

FIG. 1. (a) A generic quantum circuit to estimate $\mathbf{y}_k(\sigma)$, which is the probability that test state σ is assigned a class label k in a K-multiclass classification problem. $|a\rangle$ is an ancilla state where $\sigma \otimes |a\rangle\langle a|$ is D-dimensional and Π_k is D_{meas}-dimensional, where $D_{\text{meas}} \geqslant K$. With finite N measurements at the output, one obtains an estimate $\mathbf{y}_k^{(N)}(\sigma)$ for $\mathbf{y}_k(\sigma)$. (b) Adding depolarization noise channels $\mathcal{N}_{p_i}$ along the circuit, where $i = 1, ..., l$, the output in the $N \to \infty$ sampling limit becomes $\tilde{\mathbf{y}}_k(\sigma)$. With finite N measurements at the output, one obtains the estimate $\tilde{\mathbf{y}}_k^{(N)}(\sigma)$. See text for details.

particular, these include worst-case attacks where the adversary can craft small imperceptible perturbations $\sigma \to \rho$ about a given correctly classified input σ that result in misclassification [22]. This means that while the true labels σ and ρ are identical, if ρ is an adversarial example, $\mathcal{A}$ will class them differently. We can define adversarial examples more formally as follows [23].

Definition 2 (Adversarial example). Suppose we are given a well-trained classification function $\mathcal{A}(\cdot)$ as defined in Eq. (2), an input example (σ, C), a distance metric $h(\cdot, \cdot)$, and a small enough threshold value L. Then ρ is said to be an adversarial example if the following is true:

$$(\mathcal{A}(\sigma) = C) \wedge (\mathcal{A}(\rho) \neq C) \wedge (h(\sigma, \rho) \leqslant L). \tag{4}$$

If σ, ρ are classical states, suitable distance metrics are the l_p norms, so $h(\sigma, \rho) = ||\sigma - \rho||_p$. If σ, ρ are quantum states, we will use the trace distance $h(\sigma, \rho) = \tau(\sigma, \rho) = \text{Tr}(|\rho - \sigma|)/2$.

In the rest of this paper, we will use Greek letters to refer to quantum states and bold Roman letters to refer to classical states unless otherwise specified.

C. Differential privacy

Differential privacy is an important concept in computer science that quantifies the sensitivity of the outputs of algorithms to changes in their input data. The less sensitive it is, the better the algorithm can preserve the privacy of the input data. Here we can formulate the definition of classical differential privacy as follows [4].

Definition 3 (Classical differential privacy). Suppose $\mathcal{M}$ is a classical algorithm that takes as input entries $\mathbf{x} \in X$ of some classical database X and outputs values belonging to the set $\mathcal{S}$. Then $\mathcal{M}$ is said to satisfy *classical* (ϵ, δ)*-differential privacy* if, for all $\mathbf{x} \in X$, $\mathbf{x}' \in X'$, which are separated by a small distance, e.g., Hamming distance $h(\mathbf{x}, \mathbf{x}') \leqslant 1$ and all measurable sets $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$,

$$\Pr(\mathcal{M}(\mathbf{x}) \in \mathcal{S}) \leqslant e^{\epsilon} \Pr(\mathcal{M}(\mathbf{x}') \in \mathcal{S}) + \delta, \tag{5}$$

where $Pr(\cdot)$ denotes the probability of $(\cdot)$ and $\epsilon, \delta > 0$. We call (ϵ, δ) the *privacy budget* for the algorithm.

Informally, this definition says that for two input data points separated by a small distance, a small privacy budget means that the output of the algorithm differs very little, hence the input information is partially kept private. The selection of this distance $h(\cdot, \cdot)$ varies depending on the task, e.g., Hamming distance or l_p distance [4]. A natural distance $h(\cdot, \cdot)$ for quantum data is the trace distance, which we can employ in a definition for quantum differential privacy [24] which we will use throughout this paper. An alternative definition for quantum differential privacy [25] does not require quantum data σ and τ to be close in trace distance but rather that ρ is obtainable by applying a quantum operation on only a single register of σ. See also Ref. [26] for a related definition applied to probably approximate correct learning. However, for our purposes of working directly with quantum states σ and ρ, the use of trace distance is the most appropriate.

Suppose $\mathcal{M}(\sigma, \Pi_{\mathcal{S}})$ is a quantum algorithm that takes input state σ, applies a quantum operation $\mathcal{E}$ before applying the POVM $\{\Pi_k\}$, where the set of final measurement results $k \in \mathcal{S}$. These set of outcomes are then observed with probability $\Pr(\mathcal{M}(\sigma, \Pi_{\mathcal{S}}) \in \mathcal{S}) = \sum_{k \in \mathcal{S}} \text{Tr}(\Pi_k \mathcal{E}(\sigma))$. By analogy with Definition 1, we can write a definition of quantum differential privacy following Zhou and Ying [24].

Definition 4 (Quantum differential privacy). The quantum algorithm $\mathcal{M}$ satisfies (ϵ, δ)*-quantum differential privacy* if for all input quantum states σ and ρ with $\tau(\sigma, \rho) < \tau_D$, where τ_D is any upper bound of $\tau(\sigma, \rho)$, and for all measurable sets $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ (equivalently, for every $\Pi_S \subseteq \{\Pi_k\}_{k=0}^{K-1}$):

$$\Pr(\mathcal{M}(\rho, \Pi_{\mathcal{S}}) \in \mathcal{S}) \leqslant e^{\epsilon} \Pr(\mathcal{M}(\sigma, \Pi_{\mathcal{S}}) \in \mathcal{S}) + \delta. \tag{6}$$

For the rest of the paper, we focus on the case $\delta = 0$, which is referred to as ϵ*-quantum differential privacy*. To illustrate a simple example, suppose we have a binary classification problem where we choose the POVM $\{\Pi_0, \Pi_1 = \mathbf{1} - \Pi_0\}$. The probability σ is assigned class labels $k = 0, 1$ by a quantum binary classifier is $\tilde{\mathbf{y}}_0(\sigma) \equiv \text{Tr}(\Pi_0(\mathcal{E}(\sigma)))$ and $\tilde{\mathbf{y}}_1(\sigma) = 1 - \tilde{\mathbf{y}}_0(\sigma)$, respectively. Then if $\mathcal{M}$ satisfies ϵ-quantum differential privacy, Definition 4 requires that we must satisfy

$$e^{-\epsilon} \leqslant \frac{\tilde{\mathbf{y}}_k(\rho)}{\tilde{\mathbf{y}}_k(\sigma)} \leqslant e^{\epsilon}. \tag{7}$$

III. IMPROVING ROBUSTNESS OF QUANTUM CLASSIFIERS AGAINST ADVERSARIES BY ADDING NOISE

In this section, we show how the presence of depolarization noise in quantum circuits for classification improves robustness against adversarial examples. We begin with our definition of adversarial robustness.

Definition 5 (Adversarial robustness). Let the test state σ have the class label $\mathcal{A}(\sigma)$ under a classification algorithm $\mathcal{A}$. Then $\mathcal{A}$ is said to possess *adversarial robustness of size* τ_D if for all σ that is perturbed $\sigma \to \rho$ by an unknown source where $\tau(\sigma, \rho) \leqslant \tau_D$, the class label of ρ does not change, i.e., $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$.

We must emphasize here the difference between robustness bounds against a known noise source versus an unknown adversary. Protection against an unknown adversary is a robustness guarantee against a worst-case scenario, whereas commonly appearing known noise sources are usually far from the worst-case scenario.

Our goal is to demonstrate how a naturally occurring known noise source can be used to protect a quantum classifier against worst-case adversarial perturbations. This can be done in three main steps. We first show the robustness of quantum classifiers to this known noise source, then demonstrate how this gives rise to quantum differential privacy for the classifier. Finally, we prove how quantum differential privacy can be used to derive a theoretical bound against general adversaries.

One such naturally occurring quantum noise source is the depolarization noise channel $\mathcal{N}_p$, which acts on a D-dimensional state Σ_D like

$$\mathcal{N}_p(\Sigma_D) = p\frac{\mathbb{I}_D}{D} + (1-p)\Sigma_D, \tag{8}$$

where $\mathbb{I}_D$ is the $D \times D$ identity matrix and $p \in [0, 1]$. Before the final measurement, we can represent our quantum classifier as a unitary U gate acting on an input state $\sigma \otimes |a\rangle\langle a|$, as represented in Fig. 1(a). We can then add $\mathcal{N}_{p_i}$ after each unitary U_i where $U = U_1 \ldots U_l$ and $i = 1, \ldots, l$. Here l is the total number of depolarization channels with noise parameters $p_i > 0$. This noisy circuit is depicted in Fig. 1(b). The output of this noisy K-multiclass classification circuit given test state σ can be written as

$$\tilde{\mathbf{y}}_k(\sigma) \equiv \mathrm{Tr}(\Pi_k \mathcal{N}_{p_l}(U_l(\ldots\mathcal{N}_{p_1}(U_1(\sigma \otimes |a\rangle\langle a|)U_1^\dagger)\ldots)U_l^\dagger)), \tag{9}$$

where it can be shown [27] that for $p \equiv 1 - \prod_{i=1}^{l}(1-p_i)$:

$$\tilde{\mathbf{y}}_k(\sigma) = \frac{p}{K} + (1-p)\mathbf{y}_k(\sigma). \tag{10}$$

This leads to the interesting observation that the noisy test score $\tilde{\mathbf{y}}_k(\sigma)$ is independent of where depolarization channels are placed in the circuit. Furthermore, the effect of all depolarization channels with parameters p_i can be replaced by a single depolarization channel with parameter $p \equiv 1 - \prod_{i=1}^{l}(1-p_i)$. For the rest of this paper, we will for simplicity replace the effect of all noise parameters p_i with p unless stated otherwise. We emphasize that we don't need to include depolarization at every layer of our circuit. We can also turn off depolarization at every layer except to a single layer j by setting $p = p_j$ and $p_{i \neq j} = 0$. A special case is having depolarization noise added only to the input state so $j = 1$.

Before achieving our goal, we first need Eq. (10) to prove the following lemma showing that the K-multiclass classification algorithm performed by the noisy circuit is robust against depolarization noise for any $0 \leqslant p_i < 1$. This is a generalization of a recent result from LaRose and Coyle [28] to the case of K-multiclass classification.

Lemma 1. Let $\mathbf{y}_k(\sigma)$ denote the output for the noiseless circuit in Fig. 1(a), i.e., $p_i = 0$ for all i. Then if the class label C is assigned to σ by the noiseless circuit, i.e., $C = \arg\max_k \mathbf{y}_k(\sigma)$, then the same label is also assigned by the noisy circuit, which has $p_i > 0$ for at least one i. This means $\arg\max_k \tilde{\mathbf{y}}_k(\sigma) = C$ for any σ and $0 \leqslant p_i < 1$. Furthermore, if $\arg\max_k \tilde{\mathbf{y}}_k(\sigma) = C$, then $C = \arg\max_k \mathbf{y}_k(\sigma)$.

Proof of Lemma 1. For details, please see Appendix A. ■

The above result demonstrates robustness of quantum classifiers against depolarization noise if one has access to the exact probabilities $\tilde{\mathbf{y}}_k(\sigma)$. However, this is only possible in the limit of infinite sampling. If one is only able to sample the circuit N times, one instead obtains only the estimated values $\tilde{\mathbf{y}}_k^{(N)}(\sigma)$. Then to guarantee robustness against depolarization noise to high probability, we find the following required sampling complexity N increases only with increasing depolarization noise parameter p, but is not dependent on the dimensionality of σ.

Lemma 2. Let the predicted classification label of σ using the noiseless K-multiclass classification circuit be C. This means we can define $\xi \equiv \mathbf{y}_C(\sigma) - \max_{k\neq C} \mathbf{y}_k(\sigma)$, where $\xi > 0$. In the corresponding circuit with depolarization noise parameters $p_1, \ldots, p_l$, one samples the circuit N times for each k to obtain the estimates $\tilde{\mathbf{y}}_k^{(N)}(\sigma)$. Then σ is also labeled C with probability at least β if the sample complexity $N \sim 1/[8\xi^2(1-p)^2)\ln(2/(1-\beta))]$, where $p \equiv 1 - \prod_{i=1}^{l}(1-p_i)$.

Proof of Lemma 2. A basic sketch of the proof is the following. It can be shown that $\eta \equiv \tilde{\mathbf{y}}_C(\sigma) - \max_{k\neq C} \tilde{\mathbf{y}}_k(\sigma) = p\xi$. Thus one requires sufficient N to resolve the difference $\tilde{\mathbf{y}}_C^{(N)}(\sigma) - \tilde{\mathbf{y}}_k^{(N)}(\sigma)$ to within 2η. We then employ Hoeffding's inequality [29] to bound the sample complexity. Please see Appendix B for details. ■

Given Lemmas 1 and 2, we want to show that the accuracy of the noisy quantum classifier also does not suffer. Suppose all our training and test states are sampled from some unknown distribution $\mathcal{D}$. The accuracy of a classifier can be defined as the probability $P(C = T)$ that the predicted label C of a state σ sampled from $\mathcal{D}$ is equivalent to the true label of the state, which we call T. Let C and $\tilde{C}$ denote the predicted label by the noiseless and noisy classifier, respectively. Then the following theorem holds:

Theorem 1. The accuracy of a noisy classifier $P(\tilde{C} = T)$ and the accuracy of a noiseless classifier $P(C = T)$ are connected by the relation $P(\tilde{C} = T) = P(\tilde{C} = C)(2P(C = T) + [1 - P(C = T)]$.

Proof of Theorem 1. See Appendix C for details. ■

In the infinite sampling limit, Lemma 1 gives $P(\tilde{C} = C) = 1$. This means that the accuracy of the noisy classifier is not degraded at all since $P(\tilde{C} = T) = P(C = T)$. In the finite sampling limit, Lemma 2 gives $P(\tilde{C} = C) \geqslant 1 - 2\exp[-8N\xi^2(1-p)^2]$, so $P(\tilde{C} = T)$ approaches $P(C = T)$ quickly as N grows.

Although we have shown that depolarization noise does not affect the outcome of the quantum classifier significantly, we note that results from literature on fault-tolerant computation does put limits on the size and depth of a quantum circuit with added depolarization noise and other noise sources, and these limits come from an extra requirement that gates need to be operated below an error threshold for fault tolerance. We can leave to more detailed future investigation modifications that can be made if fault-tolerance conditions on each gate are considered.

We also note that our Theorem 1 is a good theoretical starting point to investigate the question of trade-offs between robustness and accuracy, which is a key observation in a free lunch theorem [30]. Our result will allow one to see conditions under which trade-offs are possible and when they are absent. We leave a full investigation to future work.

Now we show how adding depolarization noise also gives rise to quantum differential privacy for our algorithm. This is an application of a result from Zhou and Ying [24] for our quantum classifier.

Lemma 3. Let the algorithm $\mathcal{M}$ correspond to the K-multiclass classification circuit defined in Fig. 1(b) with depolarization noise channels $\mathcal{N}_{p_i}$, where $i = 1, \ldots, l$ and $p \equiv 1 - \prod_{i=1}^{l}(1 - p_i)$, and measurement operators $\{\Pi_k\}_{k=1}^{K}$. Then for two quantum test states σ and ρ obeying $\tau(\sigma, \rho) \leqslant \tau_D$ with $0 \leqslant \tau_D \leqslant 1$, $\mathcal{M}$ satisfies ϵ-quantum differential privacy where

$$\epsilon = \ln\left(1 + D_{\text{meas}}\frac{(1-p)\tau_D}{p}\right) \tag{11}$$

and $D_{\text{meas}} \geqslant K$ is the dimension of the operators $\{\Pi_k\}_{k=1}^{K}$.

Proof of Lemma 3. This is equivalent to Theorem 3 from Ref. [24] applied to our quantum classifier, but we extend to the case where we can apply multiple depolarization channels $\mathcal{N}_{p_i}$. For details, please see Appendix D. ■

Lemma 3 states that the privacy budget ϵ in the presence of depolarization noise decreases with increasing $p \equiv 1 - \prod_{i=1}^{l}(1 - p_i)$, hence higher depolarization noise parameters give greater differential privacy. Furthermore, this privacy is independent of where one inserts depolarization noise because the product $\prod_{i=1}^{l}(1 - p_i)$ is invariant under permutation of its factors. It is also independent of any details of the classifier except D_{meas}, which serves as an upper bound to the number of class labels in our classifier. We will return to these points later.

Using the results of Lemmas 1 and 3, the following theorem demonstrates that by increasing the strength of depolarization noise in our circuit, this also increases our K-multiclass classifier's robustness against adversarial examples.

Theorem 2 (infinite sampling case). We begin with our K-multiclass classification circuit with depolarization noise parameters p_i where $i = 1, \ldots, l$ and $p \equiv 1 - \prod_{i=1}^{l}(1 - p_i)$. Let infinite sampling of the output be allowed, so we can find $\tilde{\mathbf{y}}_k(\rho)$ for $k = 0, \ldots, K-1$ for any test state ρ given. Suppose $\tilde{\mathbf{y}}_C(\sigma) > e^{2\epsilon}\max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma)$ holds, where $\epsilon = \ln(1 + D_{\text{meas}}(1-p)\tau_D/p)$, which implies that σ is assigned the class label C, i.e., $C = \arg\max_k \tilde{\mathbf{y}}_k(\sigma) = \arg\max_k \mathbf{y}_k(\sigma)$. Then ρ is also labeled as C, i.e., $C = \arg\max_k \tilde{\mathbf{y}}_k(\rho) = \arg\max_k \mathbf{y}_k(\rho)$ for any ρ where $\tau(\sigma, \rho) \leqslant \tau_D$.

Proof of Theorem 2. Please refer to Appendix E for the proof. ■

This means that if a test state σ undergoes an arbitrary adversarial perturbation $\sigma \to \rho$, the classification of ρ will remain identical to that of σ for a larger range of $\tau(\sigma, \rho)$ if p increases. Furthermore, if τ_D remains constant, then the extra condition required of the input state $\tilde{\mathbf{y}}_C(\sigma) > e^{2\epsilon}\max_{j\neq C}\tilde{\mathbf{y}}_j(\sigma)$ also becomes easier to satisfy as p increases. A similar result holds for the finite sampling case.

Theorem 3 (Finite sampling case). Suppose one samples the output of the circuit N times for the estimation of each $\tilde{\mathbf{y}}_k(\sigma)$. Let $\tilde{\mathbf{y}}_C^{(N)}(\sigma) - \zeta > e^{2\epsilon}\max_{k\neq C}(\tilde{\mathbf{y}}_k^{(N)}(\sigma) + \zeta)$ where $\epsilon = \ln[1 + D_{\text{meas}}(1-p)\tau_D/p]$, which implies σ has the class label C. Then the class label of ρ is also C, i.e., $C = \arg\max_k \mathbf{y}_k(\rho) = \arg\max_k \tilde{\mathbf{y}}_k(\rho)$ to probability at least $1 - 2\exp(-2N\zeta^2)$ for any ρ where $\tau(\sigma, \rho) \leqslant \tau_D$. This also implies $\tilde{\mathbf{y}}_C^{(N)}(\rho) + \zeta > \max_{k\neq C}\tilde{\mathbf{y}}_k^{(N)}(\rho) - \zeta$ to probability at least $1 - 2\exp(-2N\zeta^2)$.

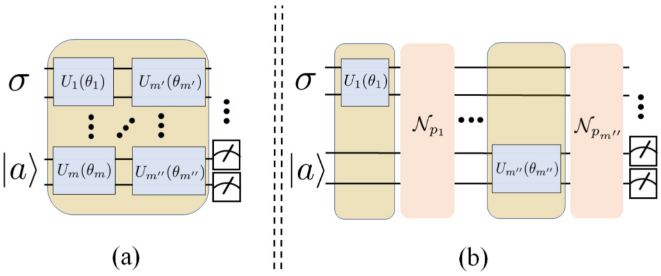

FIG. 2. *Noiseless and noisy QNN circuits.* (a) The basic scheme of QNN (noiseless). The trainable unitary $U(\boldsymbol{\theta})$ (yellow region) is composed of the product of parameterized single-qubit gates and fixed two-qubit gates $U_i(\theta_i)$ where $i = 1, \ldots, m''$ and in the diagram above $1 \leqslant m \leqslant n \leqslant m' \leqslant m'' \leqslant nl$ where l is the depth of the circuit and $n = \log_2 D$. The test state is σ and the ancilla state is $|a\rangle$. (b) Our protocol for QNN (noisy) where the depolarization channels $\mathcal{N}_{p_i}$ (pink region) are added to the noiseless QNN circuit.

Proof of Theorem 3. We employ Hoeffding's inequality [29] to show $\tilde{\mathbf{y}}_k^{(N)}(\sigma) - \zeta \leqslant \tilde{\mathbf{y}}_k(\sigma) \leqslant \tilde{\mathbf{y}}_k^{(N)}(\sigma) + \zeta$ is true to probability at least $1 - 2\exp(-2N\zeta^2)$. This relates the finitely sampled estimates $\tilde{\mathbf{y}}_k^{(N)}(\sigma)$ to $\tilde{\mathbf{y}}_k(\sigma)$ from infinite sampling. Then we can apply the results of Theorem 2 for infinite sampling to prove our results. Please see Appendix F for details of the proof. ■

As special examples, we now explore the robustness property of two discriminative learning models for binary classification: quantum neural network (QNN) and quantum kernel classifiers.

A. Quantum neural network

The QNN, proposed by Ref. [31], is a building block for various quantum learning models [31–36]. The basic scheme of QNN is illustrated in Fig. 2(a), which is a special case of the circuit in Fig. 1(a). The D-dimensional quantum input state is $\sigma \otimes |a\rangle\langle a|$, where σ refers to either the training or test states and $|a\rangle$ is an ancilla. The trainable unitary $U(\boldsymbol{\theta}) \in \mathbb{C}^{D\times D}$ is then applied, which consists of trainable single-qubit gates and fixed two-qubit gates. Our protocol for QNN, as shown in Fig. 2(b), employs the depolarization channels $\mathcal{N}_{p_i}$ that can appear within the QNN circuit before final measurements with POVM $\{\Pi_k\}$.

The typical application of QNN is for binary classification, broadly used in Refs. [31,33,35,36], where one makes single-qubit measurements using $\{\Pi_0, \Pi_1 = \mathbf{1} - \Pi_0\}$ and $D_{\text{meas}} = 2$. We can apply Theorem 2 directly to our scenario and we have the following corollary.

Corollary 1. Let the given input σ be given the classification label 0 and define $\tilde{\mathbf{y}}_0(\sigma)/\tilde{\mathbf{y}}_1(\sigma) \equiv B$. In binary classification, QNN, with depolarization channels $\mathcal{N}_{p_i}$ and $p \equiv 1 - \prod_{i=1}^{l}(1 - p_i)$, is robust against any perturbations $\sigma \to \rho$ with $\tau(\sigma, \rho) < \tau_D$ and $\epsilon = \ln(1 + 2(1-p)\tau_D/p)$, if

$$B > \exp(2\epsilon). \tag{12}$$

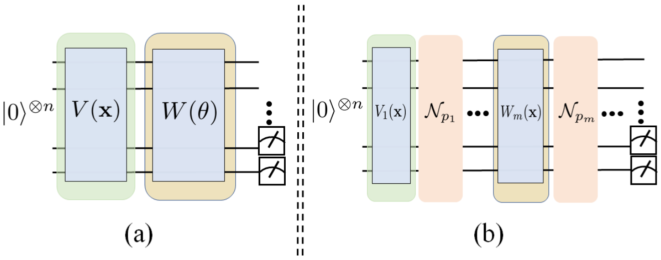

FIG. 3. *Noiseless and noisy quantum kernel classifiers.* (a) A basic scheme of the quantum kernel classifier. The unitary $V(\mathbf{x})$ (green region) takes $|0\rangle^{\otimes n} \rightarrow V(\mathbf{x})|0\rangle^{\otimes n}$, where $n = \log_2 D$. The trainable unitary $W(\boldsymbol{\theta})$ (yellow region) is composed of trainable single-qubit gates and fixed two qubits gates, which has the same architecture as in QNNs. For example, at the end we can measure in the basis $|0\rangle^{\otimes n}$ and this circuit can be used to compute the kernel $K(\boldsymbol{\theta}, \mathbf{x}) \equiv \langle 0|^{\otimes n} W(\boldsymbol{\theta}) V(\mathbf{x})|0\rangle^{\otimes n}$. (b) For our protocol, we can include depolarization noise channels $\mathcal{N}_{p_i}$ (pink region) anywhere along the quantum kernel classifier.

Since $D_{\text{meas}} = 2$ for binary classification, we note that the privacy budget ϵ is now *independent* of the dimension of the problem. Therefore, even as the feature dimension of the input σ grows, it does not affect the robustness of the classifier against adversarial examples so long as some depolarization noise with $0 < p < 1$ has been added to the circuit. This independence is an interesting contrast to the result in Ref. [15] which states that robustness should decrease as dimensionality of σ grows. This contradiction is resolved by observing that, unlike in Ref. [15], which places no constraints on distribution from which the input states σ are selected, here we have Eq. (12), which imposes a constraint.

In the finite sampling limit, we can employ Theorem 3 to apply to our binary classifier and we have the following corollary.

Corollary 2. Let the input σ be given the classification label 0 and define $(\tilde{\mathbf{y}}_0^{(N)}(\sigma) - \zeta)/(\tilde{\mathbf{y}}_1^{(N)}(\sigma) + \zeta) \equiv B$, where the probabilities are estimated using N samples of the quantum circuit. Then, if

$$B > \exp(2\epsilon), \tag{13}$$

the binary classification performed by the QNN circuit, with depolarization channels $\mathcal{N}_{p_i}$ and $p \equiv 1 - \prod_{i=1}^{l}(1 - p_i)$, is robust to adversarial attacks $\sigma \rightarrow \rho$ with the probability at least $1 - 2\exp(-2N\zeta^2)$, where $\tau(\rho, \sigma) \leqslant \tau_D$ and $\epsilon = \ln(1 + 2(1 - p)\tau_D/p)$.

B. Quantum kernel classifier

The main idea of kernel methods is to map complex input data $\mathbf{x}$ to a higher-dimensional feature space that can then be efficiently separated [20]. The generic form of a quantum kernel classifier [34,37,38] is shown in Fig. 3. The output of the kernel classifier can be written as $K(\boldsymbol{\theta}, \mathbf{x}) \equiv \langle 0|^{\otimes n} W(\boldsymbol{\theta}) V(\mathbf{x})|0\rangle^{\otimes n}$, where $K(\boldsymbol{\theta}, \mathbf{x})$ is identified with a classical kernel with test state $\mathbf{x}$ and weight vector captured by the trained $\boldsymbol{\theta}$ values. Here $W(\boldsymbol{\theta})$ contains the trainable parameters with the aim of minimizing the predefined loss function where the optimal occurs at $\boldsymbol{\theta}^*$ and $V(\mathbf{x})|0\rangle^{\otimes n}$ refers to the kernel state that maps the input data into the higher-dimensional feature space. Thus, the probability of obtaining the measurement values all 0 after applying $\Pi_0 \equiv (|0\rangle\langle 0|)^{\otimes n}$ in the noiseless circuit is given by $\mathbf{y}_0(\mathbf{x}) = \langle 0|^{\otimes n} W(\boldsymbol{\theta}^*) V(\mathbf{x})|0\rangle^{\otimes n}$.

For a binary classification problem, the class label of $\mathbf{x}$ is 0 if $\mathbf{y}_0(\mathbf{x}) > \mathbf{y}_1(\mathbf{x}) \equiv 1 - \mathbf{y}_0(\mathbf{x})$. In this case, $D_{\text{meas}} = D$, thus the privacy budget becomes $\epsilon = \ln(1 + D(1 - p)\tau_D/p)$. which grows with increasing dimensionality D of the input state. Corollaries 1 and 2 then hold for the quantum kernel classifier with this modified ϵ.

IV. NUMERICAL SIMULATIONS

We now conduct numerical simulations to illustrate our protocol for a binary QNN classifier. In particular, by leveraging the depolarization channel, we show how a trained QNN binary classifier has the ability to achieve certified robustness under bounded-norm adversarial attacks at testing time. We perform numerical simulations first on a low-dimensional data set [39] and then a high-dimensional data set [34]. In this section, we introduce our training data set and the preprocessing step. We then explain the attack method that is used to evaluate the performance of our protocol. Lastly, we analyze the performance of our proposed protocol.

A. Preprocessing and training procedure

1. Low-dimensional data set

We first choose to conduct our numerical simulations on the Iris data set [39], which has been broadly used in classical machine learning. The Iris data set $\mathcal{D}_I = \{\sigma_i, c_i^*\}_{i=1}^{150} \in \mathbb{R}^{150\times 4} \times \mathbb{R}^{150}$ consists of three different types of Iris flowers (setosa, versicolor, and virginica), where examples (belonging to setosa) with label $c_i^* = 0$ are linearly separable with respect to examples (belonging to versicolor) with label $c_i^* = 1$.

Next, we remove all examples belonging to virginica and denote the data set that only contains label $c_i^* = 0$ and $c_i^* = 1$ as $\mathcal{D}$, i.e., the cardinality of $\mathcal{D}$ is 100. Then we set the fourth entry of all examples as 0. Afterward, we apply l_2 normalization to each example, i.e., $\|\sigma_i\|_2 = 1$ for any $\sigma_i \in \mathcal{D}$. Then we need to efficiently encode this classical data into quantum states [40]. We can then carry out the amplitude encoding method [41] to encode the normalized σ_i into a quantum state.

Given the preprocessed data set $\mathcal{D}$, we randomly split it into a training data set $\mathcal{D}_{\text{Tr}}$ and a test data set $\mathcal{D}_{\text{Te}}$ with $n \equiv |\mathcal{D}_{\text{Tr}}| = 60$, $|\mathcal{D}_{\text{Te}}| = 40$, and $\mathcal{D} = \mathcal{D}_{\text{Tr}} \cup \mathcal{D}_{\text{Te}}$. In the training procedure, we randomly sample an example (σ_i, c_i^*) from $\mathcal{D}_{\text{Tr}}$ and forward σ_i to a binary QNN classifier. For details on the circuit, see Appendix I. We employ the squared loss function to train this QNN, i.e.,

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(c_i^* - \bar{c}_i)^2, \tag{14}$$

where $\bar{c}_i = \max_k \mathbf{y}_k(\sigma_i) \in [0, 1]$ is the score vector of QNN as formulated in Sec. II A and $\mathbf{y}_k(\sigma)$ denotes the ideal output of the QNN. We use the zeroth-order gradient method [37] to optimize trainable parameters $\boldsymbol{\theta}$ of the QNN to minimize the loss function $\mathcal{L}$. We set the number of training epochs to 50. The learning rate is set to 0.01 and the total number of trainable parameters is 24. Figure 4 illustrates the training

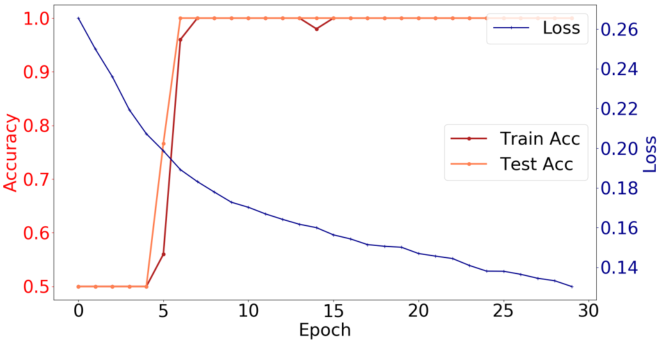

FIG. 4. *Learning performance for low-dimensional data set.* The blue, red, and orange lines, respectively, show the variation of loss, the training accuracy, and the test accuracy with respect to the number of epochs. The loss continuously decreases for longer epochs, while the training accuracy and test accuracy increases and converges sharply around epoch 5.

loss, training accuracy, and test accuracy. Both the training and test accuracy converges to 100% after 15 epochs (see Appendix I for more implementation details).

2. *High-dimensional data set*

Next, we also explore how quantum noise contributes to defending adversarial attacks for a higher-dimensional data set when the dimension is 20. The construction rule of the data set $\mathcal{D}$ is based on the proposal Ref. [34]. In particular, the data set $\mathcal{D}=\{\boldsymbol{x}_i, y_i\}_{i=0}^{N-1}$ contains in total $N=400$ examples. In each element of the set $\{\boldsymbol{x}_i, y_i\}$, $\boldsymbol{x}_i\in\mathbb{R}^{20}$ represents the data feature and y_i denotes the corresponding label. Let U_E be a specific embedding quantum circuit U_E that maps the classical data $\boldsymbol{x}_i$ to a 10-qubit quantum state, i.e., for all $i\in\{0,\ldots,N-1\}$,

$$U_E|0\rangle^{\otimes 10}=|g(\boldsymbol{x}_i)\rangle\in\mathbb{C}^{2^{10}}. \tag{15}$$

The label of $\boldsymbol{x}_i$ is labeled as $y_i=1$ if

$$\langle g(\boldsymbol{x}_i)|V^\dagger\Pi V|g(\boldsymbol{x}_i)\rangle\geqslant 0.5+\Delta, \tag{16}$$

where $V\in SU(2^{10})$ is a unitary operator (highlighted by the dark blue box in Fig. 11 in Appendix J), $\Pi=\mathbb{I}\otimes\sigma_Z$ is the measurement operator, and the gap Δ is set as 0.15. The label of $\boldsymbol{x}_i$ is assigned as $y_i=-1$ if

$$\langle g(\boldsymbol{x}_i)|V^\dagger\Pi V|g(\boldsymbol{x}_i)\rangle\geqslant 0.5-\Delta. \tag{17}$$

At the data-preprocessing stage, data set $\mathcal{D}$ is divided into the training data sets $\mathcal{D}_{\rm Tr}$ with size $N_{\rm Tr}=200$ and the test data set $\mathcal{D}_{\rm Te}$ with $N_{\rm Te}=200$. Once the splitting is completed, we employ the binary QNN classifier introduced in the main text to learn the data set $\mathcal{D}_{\rm Tr}$. In the training procedure, we continuously update trainable parameters $\boldsymbol{\theta}$ in $U(\boldsymbol{\theta})$ to minimize the squared loss $\mathcal{L}$ in Eq. (14). The implementation of the binary QNN is exhibited in the left panel of Fig. 11 in Appendix J.

The performance of the binary QNN is shown in Fig. 5. The training loss converges to 0.2 after three epochs. Meanwhile, after five epochs, both the training and test accuracies are above 98%. The simulation results indicate that under the ideal setting, the employed quantum classifier can well learn a hyperplane to correctly separate data with different labels.

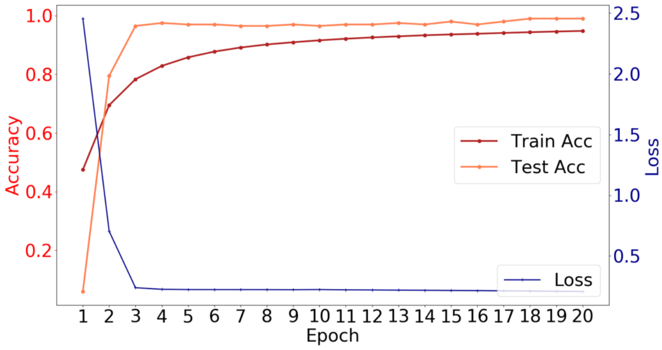

FIG. 5. *The learning performance for high-dimensional data set.* The blue, red, and orange lines respectively show the variation of loss, the training accuracy, and the test accuracy with respect to the number of epochs. The loss decreases sharply at around epoch 3 for longer epochs, the training accuracy increases continuously and test accuracy converges sharply around epoch 3.

B. Evaluation metrics and adversarial attack methods

To evaluate the performance of our protocol, we adopt an adversarial attack method that is widely employed in classical machine learning. It is known as the iterative-fast gradient sign method (I-FGSM) with l_2-bounded norm [42–44] that aims to attack the test data set $\mathcal{D}_{\rm Te}$ to make incorrect predictions when using a trained classifier. If we denote the original input by $\mathbf{x}$ and the adversarial example at the tth updating step when using the I-FGSM by $\mathbf{x}'(t)$, then

$$\begin{aligned}\mathbf{x}'_{(0)}&=\mathbf{x},\\ \mathbf{x}'_{(t+1)}&=\mathbf{x}'(t)+\alpha\cdot\text{sign}(\nabla_{\mathbf{x}}\mathcal{L}),\end{aligned} \tag{18}$$

where $\alpha=L/T$ is the learning rate with $\|\mathbf{x}-\mathbf{x}'\|_2\leqslant L$ and $\mathcal{L}$ is the loss function formulated in Eq. (14).

C. Adversarial attack at test time

1. *Low-dimensional data set*

Here we employ our trained classifier and the adversarial attack method formulated above to quantify the performance of our protocol. Recall that Corollaries 1 and 2 are the special cases of Theorems 2 and 3 when applied to binary QNN classifiers and work in the regime of using infinite and finite sampling of the output probabilities, respectively. Here we explore how our protocol protects the binary QNN classifier against adversarial attacks under these two settings.

The infinite sampling case. At testing time, we randomly sample an example $(\rho=|\mathbf{x}\rangle\langle\mathbf{x}|,\tilde{y})$ from $\mathcal{D}_{Te}$ to investigate its robustness τ_D with respect to different levels of depolarization noise p. Without loss of generality, the original test example has label $\tilde{y}=0$. We set three different values of p and τ_D: $\{p^{(1)}=0.5,\tau_D^{(1)}=0.02\}$; $\{p^{(2)}=0.1,\tau_D^{(1)}=0.02\}$ and $\{p^{(1)}=0.5,\tau_D^{(2)}=0.2\}$. From Eq. (11), their corresponding privacy budgets are $\epsilon_1=1.04$, $\epsilon_2=1.36$ and $\epsilon_3=1.4$. Given our input ρ, the outputs of our trained classifier with added depolarization noise are $\Pr(\tilde{\mathbf{y}}^{(1)}(\rho)=0)=54.46\%$, $\Pr(\tilde{\mathbf{y}}^{(2)}(\rho)=0)=58.04\%$, and $\Pr(\tilde{\mathbf{y}}^{(3)}(\rho)=0)=54.46\%$, where the corresponding constants B defined in Corollary 1 is $B^{(1)}=1.20$, $B^{(2)}=1.38$, and $B^{(3)}=1.20$, respectively.

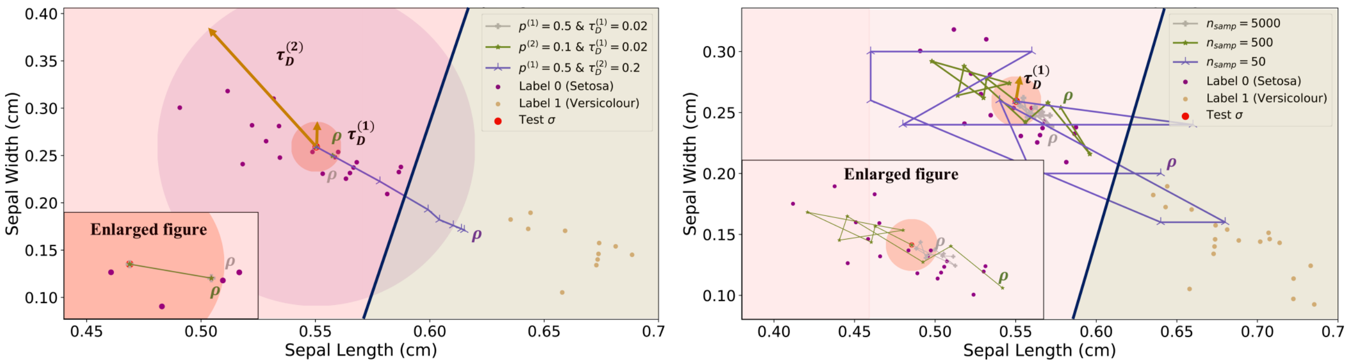

FIG. 6. *The robustness of our protocol to adversarial examples (low-dimensional data set).* Left: The left panel illustrates a bounded-norm attack on the Iris data set in the infinite sampling case. The inner and outer circle regions indicate the robustness values $\tau_D^{(1)} = 0.02$ and $\tau_D^{(2)} = 0.2$, respectively. The thick purple line is a trained hyperplane of our QNN classifier. The dotted arrows indicate how an adversary iteratively attacks the input ρ under three different settings of $\{p^{(1)} = 0.5, \tau_D^{(1)} = 0.02\}$, $\{p^{(2)} = 0.1, \tau_D^{(1)} = 0.02\}$, and $\{p^{(1)} = 0.5, \tau_D^{(2)} = 0.2\}$, where the aim of the adversary is to induce the classifier to output the wrong prediction. The inner plot enlarges the part of the central figure near the test example. Right: The right panel illustrates the bounded-norm attack in the finite precision case. The circle region indicates the robustness value $\tau_D^{(1)} = 0.02$ and $p^{(1)} = 0.5$. The dotted arrows indicate the path of an adversary that iteratively attacks the input under $n_{\text{samp}} = 50$, 500, and 5000, where the adversary aims to induce the classifier to output the wrong prediction. The inner plot enlarges the part of the central figure near the test example.

Following the condition for robustness in Eq. (12), we have confidence that the classifier is robust to adversarial attacks if $B > e^{2\epsilon}$. A simple comparison indicates that robustness is guaranteed when $\{p = 0.5, \tau_D = 0.02\}$, since $B^{(1)} > e^{2\epsilon_1} = 1.08$ while $B^{(2)} < e^{2\epsilon_2} = 1.85$ and $B^{(3)} < e^{2\epsilon_3} = 1.96$.

To validate the correctness of our theoretical results, we employ I-FGSM to attack our trained classifier, where we identify the l_2-norm bound with its corresponding τ_D value. The left panel of Fig. 6 demonstrates the simulation results and Table I shows the final test score of the attacked input. The classifier with the first setting $\{p^{(1)} = 0.5, \tau_D^{(1)} = 0.02\}$ is robust to the bounded-norm adversarial attacks, where the predicted label of $\tilde{\mathbf{x}}$ is still 0. For the third setting when $\{p^{(3)} = 0.5, \tau_D^{(3)} = 0.2\}$, the adversary can easily perturb the input and lead the classifier to give the wrong prediction. In particular, the adversary can easily perturb the input to cross the classification boundary, as highlighted by the purple line. For the second setting with $\{p^{(2)} = 0.1, \tau_D^{(2)} = 0.05\}$, the classifier correctly predicts the label, while our protocol cannot provide any promises, since Theorem 2 and Corollary 1 provide only sufficient conditions for robustness. The above three simulation results are then in accordance with our theoretical results.

The finite sampling case. The only difference in the finite sampling case is the acquisition of the output of our trained classifier. The same test example ($\rho = |\mathbf{x}\rangle\langle\mathbf{x}|, \tilde{y}$) is employed. The hyperparameters are set as $\{p^{(1)} = 0.5, \tau_D^{(1)} = 0.02\}$ and from Eq. (11) the privacy budget $\epsilon = 1.04$ is fixed. We set three different sampling number values n_{samp} to explore how n_{samp} affects the robustness guarantees, where $n_{\text{samp}}^{(1)} = 50$, $n_{\text{samp}}^{(2)} = 500$, and $n_{\text{samp}}^{(3)} = 5000$. The corresponding three approximated test scores are $\Pr(\tilde{\mathbf{y}}^{(1)} = 0) = 0.515$, $\Pr(\tilde{\mathbf{y}}^{(2)} = 0) = 0.529$, and $\Pr(\tilde{\mathbf{y}}^{(3)} = 0) = 0.552$. The corresponding parameters B are $B^{(1)} = 1.06$, $B^{(2)} = 1.124$, and $B^{(3)} = 1.23$ with respect to $n_{\text{samp}}^{(1)}$, $n_{\text{samp}}^{(2)}$, and $n_{\text{samp}}^{(3)}$. Following the results of Theorem 3 and Corollary 2, with probability at least $1 - 2\exp(-2n_{\text{samp}}\zeta^2)$, the trained classifier with added depolarization noise is robust to adversarial attacks if $B > e^{2\epsilon}$. By setting $\zeta = 0.95$, a simple inspection shows that $n_{\text{samp}} = 5000$ guarantees robustness. Analogous to the infinite sampling case, we employ a bounded-norm adversary to confirm the correctness of our theory result, where the simulation results are shown in the right panel of Fig. 6.

Given the test data set $\mathcal{D}_{\text{Te}}$, we randomly select three test examples and explore how the maximum robustness τ_D changes with varied p according to Eq. (11), which we can rewrite as

$$\tau_D = \frac{(e^{\epsilon} - 1)p}{D_{\text{meas}}(1 - p)}. \tag{19}$$

Figure 7 illustrates how τ_D scales with different p for three different test examples with $D_{\text{meas}} = 2$. Note that the constants ϵ are different for the three test examples and the test examples satisfy the condition in Eq. (12). In the same figure, we also

TABLE I. We list the test scores of the selected test examples (in the low-dimensional data set) $\{\rho, \tilde{y}\}$ after bounded-norm adversarial attacks in both the infinite and finite sampling cases. The parameters $p = 0$, $\tau_D = 0$ refer to the test score under in the absence of any depolarization noise in the circuit. The other parameter settings are $\{p^{(1)} = 0.5, \tau_D^{(1)} = 0.02\}$, $\{p^{(2)} = 0.1, \tau_D^{(1)} = 0.02\}$, and $\{p^{(1)} = 0.5, \tau_D^{(2)} = 0.2\}$.

Infinite Precision Case ($n_{\text{samp}} = \infty$)	$p = 0, \tau_D = 0$	$p^{(1)}, \tau_D^{(1)}$	$p^{(2)}, \tau_D^{(1)}$	$p^{(1)}, \tau_D^{(2)}$
$\tilde{\mathbf{y}}_0(\rho)$	58.92% (label 0)	52.96% (label 0)	57.11% (label 0)	49.42% (label 1)
Finite Precision Case ($p^{(1)}, \tau_D^{(1)}$)	–	$n_{\text{samp}} = 50$	$n_{\text{samp}} = 500$	$n_{\text{samp}} = 5000$
$\tilde{\mathbf{y}}_0(\rho)$	–	44.32% (label 1)	55.80% (label 0)	53.88% (label 0)

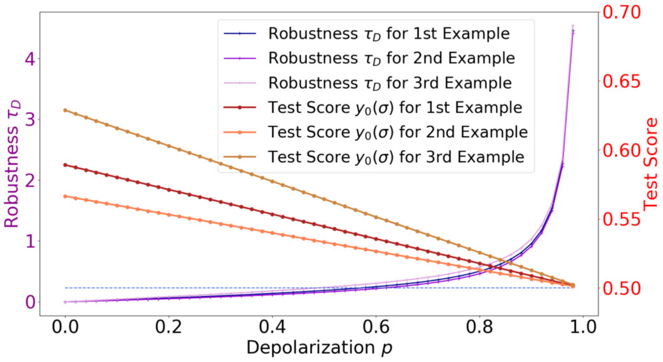

FIG. 7. *Low-dimensional data set.* For three different test examples, we see how the simulated results for robustness τ_D and test scores $\tilde{\mathbf{y}}_0(\sigma)$ varies with respect to p. All three test examples have label 0, where the test score is above the blue dotted line. The closed-form expressions for the variation of τ_D and $\tilde{\mathbf{y}}_0(\sigma)$ are shown in Eqs. (19) and (20).

plot how the test score $\tilde{\mathbf{y}}_0(\sigma)$ varies with p, coming from Eq. (10) for the case of binary classification $K = 2$:

$$\tilde{\mathbf{y}}_k(\sigma) = p/2 + (1-p)\mathbf{y}_k(\sigma). \tag{20}$$

For more details on the implementation of the classifier and performance analysis of our protocol, please see Appendix I.

2. *High-dimensional data set*

Similarly to above, given our test data set $\mathcal{D}_{\mathrm{Te}}$, we randomly select three test examples and explore how the maximum robustness τ_D changes with varied p according to Eq. (19). Figure 8 illustrates how τ_D scales with different p for three different test examples with $D_{\mathrm{meas}} = 2$. For more details of the implementation of our classifier, please see Appendix J.

V. ADVANTAGES OF PROTOCOL

Adversarial settings naturally occur when data needs to be delegated to different parties, for instance, in a client-server setting and in multiparty computing. When this data is in the form of quantum states before processing using a quantum classifier, our protocol currently provides the only existing method to protect the general quantum classifier against arbitrary adversarial examples and also includes a theoretically provable bound. Furthermore, it can take advantage of certain existing quantum noise in a quantum classifier, like depolarization noise, to provide protection against adversarial examples thus obviating the need for error correction or error mitigation if no other noise sources are present. Moreover, even if the test score is diminished in the presence of depolarization noise, its original value in the absence of any quantum noise can be retrieved by simply increasing the number of times one samples from the classifier. This sample complexity increases with the amount of exisiting depolarization noise and is independent of the dimension of the state itself.

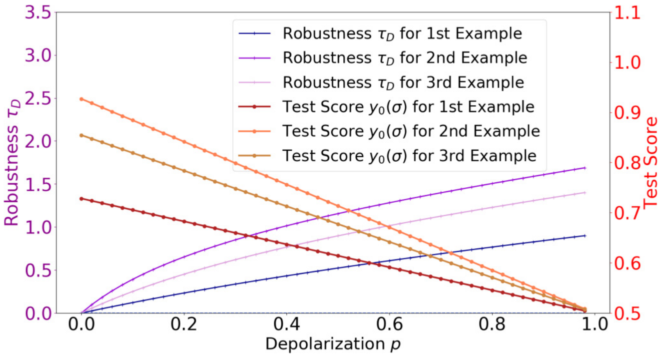

FIG. 8. *High-dimensional data set.* For three different test examples, we see how the simulated results for robustness τ_D and test scores $\tilde{\mathbf{y}}_0(\sigma)$ varies with respect to p. All three test examples have the same label, where the test score is above the blue dotted line.

Utilizing quantum noise like depolarization noise also has certain advantages over classical methods for classical data in improving robustness against adversarial examples. We discuss this below.

A. Comparison to the best known classical protocol

While in the quantum case the theoretical bound on robustness is independent of the details of the classification model and is simple to compute, this is not true in the best known classical protocol. Before elaborating on this quantum advantage, we briefly review the classical results.

Following the results of Ref. [19], classical ϵ-differential privacy of a classification algorithm is obtained by adding noise sampled from the Laplacian distribution $\mathcal{N}(z,\kappa)$ to the trained classifier. This is commonly known as the Laplace mechanism. For numerical functions [45], the only other common method to attain differential privacy is the Gaussian mechanism, which adds noise sampled from the Gaussian distribution. However, this leads to classical (ϵ,δ)-differential privacy where $\delta \neq 0$, so it cannot be directly compared to our quantum scenario where $\delta = 0$. The Laplacian distribution used in the Laplace mechanism can be written as

$$\mathcal{N}(z,\kappa) = \frac{\sqrt{2}}{2\kappa}\exp\left(\frac{-|z|}{\sqrt{2}\kappa}\right), \text{with } \kappa = \frac{\Delta f L}{\epsilon}, \tag{21}$$

where κ refers to the variance of the Laplacian distribution and L is the upper-bounded l_2 norm between original input $\mathbf{x}$ and attacked input $\mathbf{x}'$ such that classical ϵ-differential privacy is preserved. The sensitivity Δf of the function $f(\cdot)$ applied at a layer of the neural network classifier just before the Laplacian noise is injected is defined as

$$\Delta f = \max_{\mathbf{x},\mathbf{x}'} ||f(\mathbf{x}) - f(\mathbf{x}')||_2/||\mathbf{x}-\mathbf{x}'||_2. \tag{22}$$

The classical protocol runs in the following way. In the testing phase, the adversarial example $\mathbf{x}'$, where $\|\mathbf{x}-\mathbf{x}'\|_2 \leqslant L$ and $\mathbf{x}$ is the original test example, is inserted into the trained classifier $\mathbf{y}(\cdot)$. The predicted label for $\mathbf{x}'$ is obtained by invoking $\mathbf{y}(\mathbf{x}')$ a total of N times. For every run of $\mathbf{y}(\mathbf{x}')$, the noise $z_{i,j}$ with $i = 1, \ldots, N$ is independently sampled from $\mathcal{N}_L(z,\kappa)$ and applied to the input to some layer j of the neural network realizing the classifier. Let N_k denote the number of times that the predicted label is k, so the probability of the predicted label being k is given by N_k/N. Then, similarly to Theorem 3, we can write the following condition for robustness of the K-class classifier under the Laplace mechanism.

Lemma 4 (modified from Ref. [19]). Let $\mathbf{x}$ be the input to the K-multiclass classifier, which is endowed with classical ϵ-differential privacy under the Laplace mechanism, with $\epsilon = \Delta f L/\kappa$, as formulated in Eq. (21). Let C be the label of $\mathbf{x}$. Then with probability at least $1 - \zeta$, the classifier is robust to any adversarial example $\mathbf{x}'$ with $\|\mathbf{x} - \mathbf{x}'\|_2 \leqslant L$ if

$$L = \frac{\epsilon\kappa}{\Delta f} < \frac{\kappa}{2\Delta f} \ln \left(\frac{\frac{N_C}{N} - \sqrt{\frac{1}{2N} \ln\left(\frac{2}{1-\zeta}\right)}}{\max_{k \neq C} \frac{N_k}{N} + \sqrt{\frac{1}{2N} \ln\left(\frac{2}{1-\zeta}\right)}} \right). \quad (23)$$

This means that this best available classical theoretical bound to L depends on Δf, which, in general, is dependent on both the details of the classification model used and the layer of the neural network in which the Laplacian noise is injected. However, in the quantum scenario with depolarization noise, we see that the robustness bound is independent of both U, the circuit realizing the quantum classifier, as well as the location or locations of noise injection. This means that the adversarial robustness bound is universal for all quantum classifiers.

We can see this from the fact that the final state of the quantum circuit after applying depolarization noise in layers 1 to l depends only on the product $\prod_{i=1}^{l}(1 - p_i)$, which is independent of U and invariant under any re-ordering of the layers. This simplicity in the quantum case results from two facts: that the noisy part of depolarization noise lies in injecting a maximally mixed channel with a certain probability and that unitary U operations realizing any quantum classifier are unital (i.e., the identity operator $\mathbf{1}$ remains invariant under U). On the other hand, there is no known classical equivalent of this property that also gives rise to differential privacy.

The dependence of Δf on the details of the classifier in the most general cases also leads to a difficulty in the computation of Δf and is often intractable except in the simplest cases [19]. This means that, unlike in the quantum case, the corresponding classical bound on robustness L cannot be derived in closed form from Eq. (23) in the most general case.

However, in special simple cases like in kernel methods, we can provide quantitative examples of this quantum advantage. As a simple illustration, we can look at the binary classifier for the kernel perceptron, which can be written as

$$\mathbf{y}(\mathbf{x}) = \begin{pmatrix} \mathbf{y}_0(\mathbf{x}) \\ 1 - \mathbf{y}_0(\mathbf{x}) \end{pmatrix}, \quad \mathbf{y}_0(\mathbf{x}) = \sum_{i=1}^{M} w_i^* y_i^* K(\mathbf{x}_i^*, \mathbf{x}), \quad (24)$$

where $\{(\mathbf{x}_i^*, y_i^*)\}_{i=1}^{M}$ are the M training examples and $\{w_i^*\}_{i=1}^{M}$ are trained parameters of the classifier. We can consider the polynomial kernel

$$K(\mathbf{x}_i^*, \mathbf{x}) = (\mathbf{x}_i^* \cdot \mathbf{x})^n, \quad (25)$$

where n is the kernel degree and $n = 1$ is the special case of the linear kernel. We now have the following theorem.

Theorem 4. We have a binary classifier $\mathbf{y}(\mathbf{x}) = (\mathbf{y}_0(\mathbf{x}), 1 - \mathbf{y}_0(\mathbf{x}))^T$, where $\mathbf{y}_0(\mathbf{x}) = \sum_{i=1}^{M} w_i^* y_i K(\mathbf{x}_i^*, \mathbf{x})$ with the polynomial kernel $K(\mathbf{x}_i^*, \mathbf{x}) = (\mathbf{x}_i^* \cdot \mathbf{x})^n$. Let $\mathbf{x}$ denote all correctly labeled test examples. We now implement the Laplace mechanism in this classifier where the sensitivity is $\Delta f \equiv ||\mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')||_2/||\mathbf{x} - \mathbf{x}'||_2$ and the privacy budget is $\epsilon = \Delta f L/\kappa$. Let us choose $\tilde{\mathbf{y}}_0(\mathbf{x}) > \exp(2\epsilon)\tilde{\mathbf{y}}_1(\mathbf{x})$ and define $B \equiv \mathbf{y}_0(\mathbf{x})/\mathbf{y}_1(\mathbf{x})$. We can define the function $g(\cdot)$ for our noisy classifier where $g(B) = \tilde{\mathbf{y}}_0(\mathbf{x})/\tilde{\mathbf{y}}_1(\mathbf{x})$. Then the classifier is robust under any adversarial example $\mathbf{x}'$ where $||\mathbf{x} - \mathbf{x}'||_2 \leqslant L$ and

$$L \leqslant \frac{1}{M} \frac{\kappa}{2\sqrt{2}n \max\{|w_i^* y_i|\}_{i=1}^{M}} \ln g(B). \quad (26)$$

Proof of Theorem 4. We compute an upper bound for Δf in terms of classification model parameters in $\mathbf{y}(\mathbf{x})$ and use $L = \epsilon\kappa/\Delta f < \kappa \ln g(B)/(2\Delta f)$. Please see Appendix G for details. ■

From this, we see that we can guarantee only a smaller robustness bound for a more nonlinear kernel (i.e., higher n). We can also use a quantum classifier below to realize the same polynomial kernel and find a robustness bound that is now independent of degree of nonlinearity of the kernel.

Theorem 5. We have a kernel perceptron binary classifier $\mathbf{y}(\sigma) = (\mathbf{y}_0(\sigma), \mathbf{y}_1(\sigma) = 1 - \mathbf{y}_0(\sigma))^T$ that is realized by a quantum circuit in the absence of noise and takes the form in Fig. 2 with $D_{\text{meas}} = 2$. Without losing generality, we can assume the class label of σ is 0. Now we add depolarization noise channels $\mathcal{N}_{p_i}$ to the classifier where $i = 1, ..., l$ to create a noisy classifier $\tilde{\mathbf{y}}(\sigma)$. Let us choose $\tilde{\mathbf{y}}_0(\sigma) > \exp(2\epsilon)\tilde{\mathbf{y}}_1(\sigma)$ and define $B \equiv \mathbf{y}_0(\sigma)/\mathbf{y}_1(\sigma)$. Then the noisy classifier is robust under any adversarial perturbation $\sigma \to \rho$ such that $\tau(\sigma, \rho) \leqslant \tau_D$, where

$$\tau_D < \frac{B - 1}{4(B + 1) + 8(1 - p)/p} \quad (27)$$

for $p = 1 - p \in (0, 1/2)$ and

$$\tau_D^2 < \frac{B - 1}{4(B + 1)(1 - p)/p + 8(1 - p)^2/p^2} \quad (28)$$

for $p = 1 - p \in [1/2, 1)$.

Proof of Theorem 5. We use the expression for ϵ-quantum differential privacy with depolarization noise that relates τ_D with ϵ and relate ϵ to the fraction B. Please see Appendix H for details. ■

The trace distance τ_D can be turned into a corresponding l_2 norm distance L if an encoding of the classical data $\mathbf{x}$ into a quantum state $\sigma_{\mathbf{x}}$ is chosen. For instance, we can choose the most widely used amplitude encoding $\mathbf{x} \to \sum_{i=1}^{D} x_i|i\rangle$ where x_i is the ith element of $\mathbf{x}$ and we assume for simplicity the normalization $||\mathbf{x}||_2 = 1$. Then the trace distance $\tau(\sigma_{\mathbf{x}}, \sigma_{\mathbf{x}'}) = \sqrt{1 - \text{Tr}(\sigma_{\mathbf{x}}\sigma_{\mathbf{x}'})} = \sqrt{1 - (\mathbf{x} \cdot \mathbf{x}')^2}$ and $l_2 \equiv ||\mathbf{x} - \mathbf{x}'||_2 = \sqrt{2 - 2(\mathbf{x} \cdot \mathbf{x}')}$. Therefore we can write $\tau(\sigma_{\mathbf{x}}, \sigma_{\mathbf{x}'}) = l_2\sqrt{1 - l_2^2/2} \geqslant l_2$. This means Theorem 5 still holds if we replace τ_D with L and can compare results directly with Theorem 4 with the same chosen constant B. Then we see how the robustness bound in the classical case is dependent on details of the kernel function like the nonlinearity n whereas the robustness bound can be completely independent of the kernel function.

While in Theorems 4 and 5 we have provided only sufficient though not necessary conditions for robustness, this was only for the purpose of illustrating a clearer interpretation of robustness in terms of a model parameter like the degree of nonlinearly n. Necessary conditions can also be found since we already have the exact expressions for L and τ_D. We know that the former is dependent on the details of the classification model through Δf in the most general case whereas the latter is dependent only on p, D_{meas}, and ϵ, which can be chosen to be constants independent of the details of the kernel or any other classifier. This latter property of τ_D we have already learned is not consistent with any known classical mechanism for differential privacy.

Another advantage of the quantum mechanism is that depolarization noise can occur naturally in quantum systems, especially for NISQ devices, whereas the Laplace mechanism needs to be artificially injected into the classifier.

We note that while there are other classical methods of adding noise, like drop-out, that may be more analogous to depolarization noise, it is not clear if it can really be a differential private mechanism and it is not currently used in any state-of-the-art protocols.

This feature in the quantum protocol where the bound is not directly dependent on the classification model can be advantageous in the following setting, which is independent of other quantum advantages like quantum speedups. Suppose a quantum company sells different quantum classifiers to different clients but they sell them as black boxes and wish to keep any details of their circuit hidden. The client might indirectly infer some details by performing circuit tomography, but this is notoriously resource intensive. It is also not always possible for classifiers since the device contains a final projective measurement whose dimension K is less than the dimension of the input, D so there is irretrievable information loss.

Now the client receives quantum states from other sources and wishes to protect the classifier against adversaries. However, each client has a different threshold τ_D they are willing to work with. Classically, every client must implement a separate protocol for his/her device to guarantee robustness against adversaries. However, model independence of adversarial robustness is an advantage because now any client can simply add depolarization noise anywhere in their circuit by the desired amount without opening their black box or spending large resources in inferring information about their black box. We call this a possible security advantage because we can secure the classifier against adversaries of certain sizes without having information about the classifier revealed (which the company might also not want) or spending extra client resources in finding the appropriate protocol.

While we cannot rigorously prove this quantum advantage in comparison to more general classical cases and can only currently show for our particular example, we hope this will give a motivation to the community to examine these alternative quantum advantages in future works. We also note that, irrespective of the comparison to the classical case, our protocol is still relevant for when the incoming data are quantum states and cannot be replaced by a classical protocol.

It is also intriguing to investigate other types of quantum noise that can similarly protect quantum classifiers against adversaries. To show this is not limited to depolarization noise, we can show that the Pauli channel acting on σ like $\mathcal{N}_{\text{Pauli}}(\sigma) = p_{id}\sigma + \sum_{i=x,y,z}(\mathbf{1}^{\otimes D-1} \otimes \sigma_i)\sigma(\mathbf{1}^{\otimes D-1} \otimes \sigma_i)$ also have the dual properties of leaving the classification robust as well as having ϵ-quantum differential privacy when $0 < p_x + p_y < 1/2$, where the privacy budget is $\epsilon = \ln((1-2(p_x+p_y))\tau_D/(p_x+p_y)+1)$ (see Appendix K for details). It remains exciting work for future investigation to see if other natural sources of quantum noise can be harnessed for adversarial protection without compromising on the accuracy of the classification.

VI. DISCUSSION

We demonstrated how depolarization noise placed anywhere in a quantum circuit used for classification can be exploited to protect the classification algorithm against arbitrary worst-case attacks like adversarial examples. A theoretical bound for robustness can be proved without any assumptions on the type of adversary or the classification model and applies to both quantum and classical data. This bound relies on a new relationship we introduced between quantum differential privacy and adversarial robustness in the quantum setting. In particular, depolarization noise allows the theoretical robustness bound to be dependent only on the number of classes in the classification model and no other feature of the classifier. However, all known classical noise that can give rise to differential privacy results in robustness bounds that would generally depend on more details of the classification model, for instance, the degree of nonlinearity of the classification boundary.

This result raises many intriguing possibilities for exploring other naturally occurring quantum noise sources that could offer similar advantages against adversarial attacks, which become pertinent concerns as quantum data are shared in a future quantum internet. We see that the fruitful merging of concepts in security and quantum machine learning potentially leads to quantum advantages that are independent of quantum speedups. This also highlights how noise in the NISQ era for quantum computation can be used as a positive feature and can be employed in parallel with other methods to demonstrate quantum advantage.

ACKNOWLEDGMENTS

N.L. is grateful to Barry Sanders (University of Calgary), Ryan LaRose (Michigan State University), Dong-Ling Deng (Tsinghua University), Jeongho Bang (Korea Institute for Advanced Study), and Srinivasan Arunachalam (IBM Research) for very interesting and fruitful discussions. N.L. also thanks Dong-Ling Deng for giving useful feedback on this paper during her stay in Tsinghua University. N.L. acknowledges funding from the Shanghai Pujiang Talent Grant No. 20PJ1408400 and the NSFC International Young Scientists Project No. 12050410230. N.L. is also supported by the Innovation Program of the Shanghai Municipal Education Commission No. 2021-01-07-00-02-E00087.

APPENDIX A: PROOF OF LEMMA 1

Here we prove that if the noiseless quantum classifier assigns σ to the class C, i.e., $C = \arg\max_k \mathbf{y}_k(\sigma)$, then the noisy circuit with depolarization noise also assigns σ to the class C, i.e. $C = \arg\max_k \tilde{\mathbf{y}}_k(\sigma)$. This is equivalent to the condition that if $\mathbf{y}_C(\sigma) > \max_{k\neq C} \mathbf{y}_k(\sigma)$, then $\tilde{\mathbf{y}}_C(\sigma) > \max_{k\neq C} \tilde{\mathbf{y}}_k(\sigma)$.

Using Eq. (8),

$$\mathcal{N}_p(\sigma) = \frac{p}{D}\mathbb{I}_D + (1-p)\sigma, \tag{A1}$$

we can rewrite Eq. (9) as

$$\begin{aligned}\tilde{\mathbf{y}}_k(\sigma) &\equiv \mathrm{Tr}(\Pi_k\mathcal{N}_{p_l}(U_L(\ldots\mathcal{N}_{p_1}(U_1(\sigma\otimes|a\rangle\langle a|)U_1^\dagger)\ldots)))\\ &= \mathrm{Tr}\left(\Pi_k p\frac{\mathbb{I}}{D}\right) + (1-p)\mathbf{y}_k(\sigma)\\ &= \frac{p}{K} + (1-p)\mathbf{y}_k(\sigma),\end{aligned} \tag{A2}$$

where $k = 1, 2, \ldots, K$ and $p \equiv 1 - \prod_{i=1}^{l}(1-p_i)$. The second line can be readily derived by induction. Then if $\mathbf{y}_C(\sigma) > \max_{k\neq C}\mathbf{y}_k(\sigma)$, Eq. (A2) implies

$$\begin{aligned}\tilde{\mathbf{y}}_C(\sigma) &= \frac{p}{K} + (1-p)\mathbf{y}_C(\sigma)\\ &> \frac{p}{K} + (1-p)\max_{k\neq C}\mathbf{y}_k(\sigma)\\ &= \max_{k\neq C}\left(\frac{p}{K} + (1-p)\mathbf{y}_k(\sigma)\right) = \max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma).\end{aligned} \tag{A3}$$

Conversely, if $\tilde{\mathbf{y}}_C(\sigma) > \max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma)$, then from Eq. (A2) it is clear that $\mathbf{y}_C(\sigma) > \max_{k\neq C}\mathbf{y}_k(\sigma)$ is true also. ■

APPENDIX B: PROOF OF LEMMA 2

From Lemma 1, we know that if σ is labeled as C in the noiseless circuit, then in the infinite sampling limit this label is maintained in the corresponding circuit with depolarization noise, so $\tilde{\mathbf{y}}_C(\sigma) > \max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma)$. However, in the finite sampling limit with sample complexity N, we only have access to the estimate $\tilde{\mathbf{y}}_k^{(N)}(\sigma)$. So, we want to find the smallest N so $\tilde{\mathbf{y}}_C^{(N)}(\sigma) > \max_{k\neq C}\tilde{\mathbf{y}}_k^{(N)}(\sigma)$ with probability at least β.

From results in Lemma 1, we see that since $\tilde{\mathbf{y}}_k(\sigma) = 1 - p/K + p\mathbf{y}_k(\sigma)$ and $\xi \equiv \mathbf{y}_C(\sigma) - \max_{k\neq C}\mathbf{y}_k(\sigma)$, then $\eta \equiv \tilde{\mathbf{y}}_C - \max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma) = p\xi$. Thus we need large enough sampling to resolve the difference $\tilde{\mathbf{y}}_C^{(N)}(\sigma) - \max_{k\neq C}\tilde{\mathbf{y}}_k^{(N)}(\sigma)$ to at least $2\eta = 2p\xi$. It is then sufficient to find N that estimates $\tilde{\mathbf{y}}_k^{(N)}(\sigma)$ to precision 2η. To find N, we can employ Hoeffding's inequality in the following.

Lemma A. (Hoeffding's inequality [29]) Let $Z_1, \ldots, Z_N$ be independent bounded random variables with $Z_i \in [a, b]$ for all $i \in [N]$, where $-\infty < a \leqslant b < \infty$. Then the probability

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N}Z_i - \mathbb{E}(Z_i)\right| \leqslant \zeta\right) \geqslant 1 - 2\exp\left(-\frac{2N\zeta^2}{(b-a)^2}\right). \tag{B1}$$

In our case, we can use $b - a = 1$, $\zeta = 2\eta$, $(1/N)\sum_{i=1}^{N}Z_i = \tilde{\mathbf{y}}_k^{(N)}(\sigma)$ and $\mathbb{E}(Z_i) = \tilde{\mathbf{y}}_k(\sigma)$. Thus if we require the probability $\Pr(|\tilde{\mathbf{y}}_k^{(N)}(\sigma) - \mathbf{y}_k(\sigma)| < 2\eta) \geqslant \beta$, it is sufficient to require $1 - 2\exp(-8N(1-p)^2\xi^2) \sim \beta$ or, equivalently,

$$N \sim \frac{1}{8(1-p)^2\xi^2}\ln\left(\frac{2}{1-\beta}\right). \tag{B2}$$

■

APPENDIX C: PROOF OF THEOREM 1

Let C_σ be the class label output of our noiseless model for input state σ. However, this model does not necessarily coincide with the ground truth for σ, which we present by T_σ. Thus the model is correct for σ if $T_\sigma = C_\sigma$ and is incorrect for σ if $C_\sigma \neq T_\sigma$. For both our training and our test states, let us sample from the states σ from some (usually unknown) distribution $\mathcal{D}$. Then, by *accuracy in the absence of noise,* we mean the probability that $T_\sigma = C_\sigma$ when $\sigma \sim \mathcal{D}$, which can be denoted

$$A \equiv P_{\sigma\sim\mathcal{D}}(C_\sigma = T_\sigma) \equiv P(C = T), \tag{C1}$$

where we have dropped the subscripts for convenience. Now let us include noise in our model so $C_\sigma \to \tilde{C}_\sigma$. Then the *accuracy in the presence of noise* can denoted by the probability

$$\tilde{A} \equiv P_{\sigma\sim\mathcal{D}}(\tilde{C}_\sigma = T_\sigma) \equiv P(\tilde{C} = T). \tag{C2}$$

There is *another* type of accuracy which we can call *robustness accuracy*, which refers to the probability that the model itself gives rise to the same prediction after adding noise, irrespective of the relationship to the ground truth. This robustness accuracy we can denote by

$$A^* = P_{\sigma\sim\mathcal{D}}(\tilde{C}_\sigma = C_\sigma) \equiv P(\tilde{C} = C). \tag{C3}$$

Usually it is A and $\tilde{A}$ that is of interest for generalization performance, so by relating $\tilde{A}$ with A, we can determine by what amount accuracy degrades in the presence of noise. However, we will see that this relationship will also depend on A^* and it is through A^* that we can include information about the type of noise that is added.

Let there be K classes, so each of $C, \tilde{C}, T$ can take values $0, 1, \ldots, K-1$. So, now we can rewrite $\tilde{A}$ as

$$\begin{aligned}\tilde{A} &\equiv P(\tilde{C} = T) = \sum_{j=0}^{K-1}P(\tilde{C} = T = j)\\ &= \sum_j P(\tilde{C} = j|T = j)P(T = j) = \frac{1}{K}\sum_j P(\tilde{C} = j|T = j)\\ &= \frac{1}{K}\sum_j P(\tilde{C} = j|C = j)P(C = T)\\ &\quad + \frac{1}{K}\sum_j\sum_{k\neq j}P(\tilde{C} = j|C = k \neq j)P(C \neq T)\\ &= \frac{1}{K}\left(A\sum_j P_{jj} + (1-A)\sum_j\sum_{k\neq j}P_{jk}\right),\end{aligned} \tag{C4}$$

where $P_{ij} \equiv P(\tilde{C} = i|C = j)$. We also used the generic assumption in the first line that the samples we have are unbiased in the sense that there are as many states of one class as the other determined by the ground truth, so $P(T = j) = 1/K$ for any j.

From normalization of probability, we can write $P_{kk} = 1 - \sum_{j\neq k}P_{jk}$, which allows us to rewrite $\sum_j\sum_{k\neq j}P_{jk} = \sum_k(\sum_{j\neq k}P_{jk}) = \sum_k(1 - P_{kk}) = K - \sum_j P_{jj}$. Now inserting this into Eq. (C4), we find

$$\tilde{A} = \frac{1}{K}\left(A\sum_j P_{jj} + (1-A)\right)\left(K - \sum_j P_{jj}\right)$$
$$= \frac{1}{K}(1-2A)\sum_j P_{jj} + (1-A). \tag{C5}$$

Now using the generic assumption in the second line that the samples we have are unbiased also with respect to the mode, so $P(C=j) = 1/K$, we see that

$$\sum_j P_{jj} = \sum_j \frac{P(\tilde{C} = C = j)}{P(C=j)} = KP(\tilde{C} = C) = KA^*. \tag{C6}$$

Therefore, we can write the general relation for *any* noise type as

$$\tilde{A} = A^*(2A-1) + (1-A). \tag{C7}$$

So, now we see that the source of the decrease in accuracy $\tilde{A}$ due to noise itself is due *only* to $A^* = P(\tilde{C} = C)$. The rest of the accuracy dependence is on the accuracy in the noiseless case A which is independent of noise.

It is through the behavior of A^* that makes depolarization noise quite special, since in the *infinite sampling* limit $N \to \infty$, then $\tilde{C}_\sigma = C_\sigma$ for all states σ, which gives $A^* = 1$. This was the result already proved in Lemma 1 with details in Appendix A. Therefore, if we allow $N \to \infty$ sampling of the quantum circuit, inserting $A^* = 1$ into Eq. (C7) gives

$$\tilde{A} = (2A-1) + (1-A) = A, \tag{C8}$$

so the accuracy in the absence of noise is the same as the accuracy in the presence of depolarization noise! Note that this is special to depolarization noise having the property $\tilde{C} = C$ and is not true for other noises. However, there are also other classes of noises that can help the model remain robust. Some other robustness properties have been investigated in Ref. [24] for other types of noises and their impact on accuracy can be found through Eq. (C7).

However, in the *finite sampling* limit, $A^* \neq 1$ for depolarization noise. We already proved in Lemma 2 that $A^* > 1 - 2\exp(-2N\zeta^2)$ where N is the number of samples we take from the quantum circuit and ζ is the precision to which we determine the quantity $\tilde{\mathbf{y}}_0(\sigma)$. Therefore, we see that in the finite sampling limit, the accuracy in the presence of noise degrades as

$$\tilde{A} > [1 - 2\exp(-2N\zeta^2)](2A-1) + (1-A), \tag{C9}$$

where $\tilde{A} \to A$ exponentially quickly as N grows, so accuracy in the presence of depolarization noise is not in fact compromised very much. Thus, all effects of depolarization noise on accuracy can be remedied by taking more measurements (efficiently).

APPENDIX D: PROOF OF LEMMA 3

This proof follows Zhou and Ying [24], applied to the case where the dimension of the final projector is $D_{\rm meas}$ and we can apply multiple depolarization channels $\mathcal{N}_{p_i}$ for $i = 1, ..., l$ where $p \equiv 1 - \prod_{i=1}^{l}(1-p_i)$. To show ϵ-differential privacy, we must show that when $\tau(\sigma,\rho) \leqslant \tau_D$, the following relation must hold, i.e.,

$$e^{-\epsilon} \leqslant \frac{\tilde{\mathbf{y}}_k(\rho)}{\tilde{\mathbf{y}}_k(\sigma)} \leqslant e^{\epsilon}, \tag{D1}$$

where from Eq. (9):

$$\tilde{\mathbf{y}}_k(\rho) = {\rm Tr}(\Pi_k(\mathcal{N}_{p_l}(U_l(...\mathcal{N}_{p_1}(U_1(\rho)U_1^\dagger)...U_l^\dagger)))). \tag{D2}$$

By employing the definition of depolarization noise with noise parameter p acting on an arbitrary quantum state σ, from Eq. (8),

$$\mathcal{N}_p(\sigma) = \frac{p}{D}\mathbb{I}_D + (1-p)\sigma, \tag{D3}$$

we can derive

$$\tilde{\mathbf{y}}_k(\rho) = \frac{p(D-D_{\rm meas})}{D}{\rm Tr}(\Pi_k)$$
$$+ (1-)p{\rm Tr}(\mathbb{I}_{D-D_{\rm meas}} \otimes \Pi_k U(\rho)U^\dagger), \tag{D4}$$

and similarly for $\tilde{\mathbf{y}}_k(\sigma)$. From this, we can write

$$\frac{\tilde{\mathbf{y}}_k(\rho)}{\tilde{\mathbf{y}}_k(\sigma)} - 1 = (1-p){\rm Tr}(U(\rho-\sigma)U^\dagger)\mathbb{I}_{D-D_{\rm meas}} \otimes \Pi_k)$$
$$\Big/\left(\frac{p(D-D_{\rm meas})}{D}{\rm Tr}(\Pi_k) + F\right)$$
$$\leqslant \frac{(1-p)\tau_D{\rm Tr}(\mathbb{I}_{D-D_{\rm meas}} \otimes \Pi_k)}{\frac{p(D-D_{\rm meas})}{D}{\rm Tr}(\Pi_k)}$$
$$= \frac{1-p}{p}D_{\rm meas}\tau_D, \tag{D5}$$

where $F \equiv (1-p){\rm Tr}(\mathbb{I}_{D-D_{\rm meas}} \otimes \Pi_k U(\sigma)U^\dagger) > 0$. In the first inequality, we used the relation ${\rm Tr}(U(\rho-\sigma)U^\dagger\Lambda_k) \leqslant \tau_D{\rm Tr}(\Lambda_k)$ and the inequality $\tau(U(\sigma)U^\dagger, U(\rho)U^\dagger) \leqslant \tau(\sigma,\rho) \leqslant \tau_D$ [24,46].

To satisfy Eq. (D2), we upper bound this final term by $e^\epsilon - 1$ and find the privacy budget

$$\frac{1-p}{p}D_{\rm meas}\tau_D \leqslant e^\epsilon - 1 \Rightarrow \epsilon = \ln\left(1 + D_{\rm meas}\tau_D\frac{1-p}{p}\right). \tag{D6}$$

APPENDIX E: PROOF OF THEOREM 2

Here we prove that if $\tilde{\mathbf{y}}_k(\sigma) > e^{2\epsilon}\max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma)$ where $\epsilon = \ln[1 + D_{\rm meas}(1-p)\tau_D/p]$, then $\tilde{\mathbf{y}}_C(\rho) > \max_{k\neq C}\tilde{\mathbf{y}}_k(\rho)$ for all ρ where $\tau(\sigma,\rho) \leqslant \tau_D$. First we employ Lemma 3, which states that given depolarization noise with parameter p, the algorithm implemented by the noisy circuit has ϵ-quantum differential privacy. Then, from Eq. (7) following Definition 4, we see that in our case it states

$$e^{-\epsilon} \leqslant \frac{\tilde{\mathbf{y}}_k(\rho)}{\tilde{\mathbf{y}}_k(\sigma)} \leqslant e^{\epsilon}, \tag{E1}$$

which holds true for when $\epsilon = \ln[1 + D_{\rm meas}(1-p)\tau_D/p]$ and all ρ where $\tau(\sigma,\rho) \leqslant \tau_D$. Then if we insert $\tilde{\mathbf{y}}_k(\sigma) > e^{2\epsilon}\max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma)$ into the above, we can write

$$\tilde{\mathbf{y}}_k(\rho) \geqslant e^{-\epsilon}\tilde{\mathbf{y}}_k(\sigma) > e^{\epsilon}\max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma). \tag{E2}$$

Then, from the left-hand side inequality in Eq. (E1), we find

$$\tilde{\mathbf{y}}_k(\rho) \geqslant \max_{k\neq C} \tilde{\mathbf{y}}_k(\rho). \tag{E3}$$

From Lemma 1, we see that this is also equivalent to the claim $\mathbf{y}_k(\rho) \geqslant \max_{k\neq C} \mathbf{y}_k(\rho)$. ■

APPENDIX F: PROOF OF THEOREM 3

From Hoeffding's inequality [see Eq. (B1) in Lemma A of Appendix B], it is clear that

$$\tilde{\mathbf{y}}_k^{(N)}(\sigma) - \zeta \leqslant \tilde{\mathbf{y}}_k(\sigma) \leqslant \tilde{\mathbf{y}}_k^{(N)}(\sigma) + \zeta \tag{F1}$$

to probability greater than $1 - 2\exp(-2N\zeta^2)$. In the statement of Theorem 3, we assume $\tilde{\mathbf{y}}_C^{(N)}(\sigma) - \zeta > e^{2\epsilon}\max_{k\neq C}(\tilde{\mathbf{y}}_k^{(N)}(\sigma) + \zeta)$. Inserting the above, this implies

$$\begin{aligned}\tilde{\mathbf{y}}_C(\sigma) &\geqslant \tilde{\mathbf{y}}_C^{(N)}(\sigma) - \zeta \\ &> e^{2\epsilon}\max_{k\neq C}\left(\tilde{\mathbf{y}}_k^{(N)}(\sigma) + \zeta\right) \geqslant e^{2\epsilon}\max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma)\end{aligned} \tag{F2}$$

is true to probability at least $1 - 2\exp(-2N\zeta^2)$. From Theorem 2, we know that the above inequality $\tilde{\mathbf{y}}_C(\sigma) > e^{2\epsilon}\max_{k\neq C}\tilde{\mathbf{y}}_k(\sigma)$ leads to the condition $C = \arg\max_k \tilde{\mathbf{y}}_k(\rho) = \arg\max_k \mathbf{y}_k(\sigma)$ for all ρ where $\tau(\sigma, \rho) \leqslant \tau_D$ and $\epsilon = \ln(1 + D_{\text{meas}}(1-p)\tau_D/p)$.

Then using Eq. (F1) again in the condition $C = \arg\max_k \tilde{\mathbf{y}}_k(\rho)$, equivalent to $\tilde{\mathbf{y}}_C(\rho) > \max_{k\neq C}\tilde{\mathbf{y}}_k(\rho)$, we find

$$\tilde{\mathbf{y}}_C^{(N)}(\sigma) + \zeta > \max_{k\neq C}\tilde{\mathbf{y}}_k^{(N)}(\sigma) + \zeta \tag{F3}$$

to probability at least $1 - 2\exp(-2N\zeta^2)$. ■

APPENDIX G: PROOF OF THEOREM 4

We first observe that for integers $n > 0$ and numbers u and v we have $u^n - v^n = (u - v)\sum_{j=0}^{n} u^j v^{n-1-j}$. If we assume $|u|, |v| \leqslant G$, this then implies $|u^n - v^n| \leqslant |u - v|nG^{n-1}$. Let $q(u_i) = a_i u_i^n$ so

$$\begin{aligned}\left|\sum_{i=1}^{M} q(u_i) - q(v_i)\right| &\leqslant \sum_{i=1}^{M}|q(u_i) - q(v_i)| \\ &\leqslant \sum_{i=1}^{M}|a_i|\left|u_i^n - v_i^n\right| \leqslant \sum_{i=1}^{M}|a_i||u_i - v_i|nG_i^{n-1},\end{aligned} \tag{G1}$$

where all $|u_i|, |v_i| \leqslant G_i$. In our case, we can define $a_i = w_i^* y_i$, $u_i = \mathbf{x}_i^* \cdot \mathbf{x}$, $v_i = \mathbf{x}_i^* \cdot \mathbf{x}'$. Suppose we fix a normalization $||\mathbf{x}_i^*||_2 = 1 = ||\mathbf{x}||_2 = ||\mathbf{x}'||_2$. From the Cauchy-Schwarz inequality, $|u_i| = |\mathbf{x}_i^* \cdot \mathbf{x}| \leqslant ||\mathbf{x}_i^*||_2||\mathbf{x}||_2 = 1$ and similarly $|v_i| \leqslant 1$, so it is sufficient for us to choose $G_i = 1$. We now want to compute the sensitivity for the kernel perceptron model where the sensitivity is defined in Eq. (22),

$$\Delta f = \max_{\mathbf{x},\mathbf{x}'}||f(\mathbf{x}) - f(\mathbf{x}')||_2/||\mathbf{x} - \mathbf{x}'||_2, \tag{G2}$$

where in our case of the polynomial kernel $f(\mathbf{x}) = \mathbf{y}(\mathbf{x}) = (\mathbf{y}_0(\mathbf{x}), 1 - \mathbf{y}_0(\mathbf{x}))^T$ and $\mathbf{y}_0(\mathbf{x}) = \sum_{i=1}^{M} w_i^* y_i^* (\mathbf{x}_i^* \cdot \mathbf{x})^n$. Then it is straightforward to show

$$\Delta f = \max_{\mathbf{x},\mathbf{x}'}\sqrt{2}\frac{\left|\sum_{i=1}^{M} w_i^* y_i^*(K(\mathbf{x}_i^*, \mathbf{x}) - K(\mathbf{x}_i^*, \mathbf{x}'))\right|}{||\mathbf{x} - \mathbf{x}'||_2}. \tag{G3}$$

Using Eq. (G1) for the polynomial kernel, we obtain

$$\begin{aligned}\Delta f &\leqslant \sum_{i=1}^{M}\frac{|a_i||u_i - v_i|}{||\mathbf{x} - \mathbf{x}'||_2}n = \sum_{i=1}^{M}\frac{|a_i||\mathbf{x}_i^* \cdot (\mathbf{x} - \mathbf{x}_i')|}{||\mathbf{x} - \mathbf{x}'||_2}n \\ &\leqslant \sum_{i=1}^{M}\frac{|a_i|||\mathbf{x}_i^*||_2||\mathbf{x} - \mathbf{x}_i'||_2}{||\mathbf{x} - \mathbf{x}'||_2}n \\ &= \sum_{i=1}^{M}|w_i^* y_i|n \leqslant M\max\{|w_i^* y_i|\}_{i=1}^{M}n,\end{aligned} \tag{G4}$$

where we used the normalization $||\mathbf{x}_i^*||_2 = 1$ in the last line. In the special case of the linear kernel (or $n = 1$), we have $\Delta f \leqslant M\max\{|w_y y_i|\}_{i=1}^{M}$.

From Eq. (21) in the text,

$$\kappa = \frac{\Delta f L}{\epsilon}. \tag{G5}$$

This means that the classifier is robust against all adversarial examples $\mathbf{x}'$, where $||\mathbf{x}' - \mathbf{x}||_2 \leqslant L = \kappa\epsilon/\Delta f$. In our theorem, we require the condition $g(B) \equiv \tilde{\mathbf{y}}_0(\mathbf{x})/\tilde{\mathbf{y}}_1(\mathbf{x}) > \exp(2\epsilon)$ where $B \equiv \mathbf{y}_0(\mathbf{x})/\mathbf{y}_1(\mathbf{x})$, which gives $\epsilon < (1/2)\ln g(B)$. Together with $1/\Delta f \geqslant 1/(M\max\{|w_i^* y_i|\}_{i=1}^{M}n)$ from Eq. (G4), this implies $||\mathbf{x}' - \mathbf{x}||_2 < \kappa\ln g(B)/(2\sqrt{2}M\max\{|w_i^* y_i|\}_{i=1}^{M}n)$ is a sufficient condition for robustness.

APPENDIX H: PROOF OF THEOREM 5

Following the results of Lemma 1, when the depolarization noise layer is inserted into the trained model just before the final measurement, the classifier $y(\sigma)$ has the ϵ-differential privacy property where

$$\begin{aligned}e^{-\epsilon}\tilde{\mathbf{y}}_0(\sigma) < \tilde{\mathbf{y}}_0(\rho) \leqslant e^{\epsilon}\tilde{\mathbf{y}}_0(\sigma), \\ e^{-\epsilon}\tilde{\mathbf{y}}_1(\sigma) \leqslant \tilde{\mathbf{y}}_1(\rho) \leqslant e^{\epsilon}\tilde{\mathbf{y}}_1(\sigma).\end{aligned} \tag{H1}$$

Now, if the initial class label of σ is 0, to correctly predict the attacked input ρ requires

$$\tilde{\mathbf{y}}_0(\rho) > \tilde{\mathbf{y}}_1(\rho) = 1 - \tilde{\mathbf{y}}_0(\rho). \tag{H2}$$

In combination with Eqs. (H1), this robustness condition is equivalent to $(1 + e^{2\epsilon})\tilde{\mathbf{y}}(\sigma) > e^{2\epsilon}$ or

$$\tilde{\mathbf{y}}_0(\sigma)/\tilde{\mathbf{y}}_1(\sigma) > e^{2\epsilon}. \tag{H3}$$

By including depolarization channels with corresponding depolarization parameters p_1, ..., p_l, we can write

$$\tilde{\mathbf{y}}_0(\sigma) = p/2 + (1 - p)\mathbf{y}_0(\sigma), \tag{H4}$$

where $p = 1 - p$. Then inserting Eq. (H4) into Eq. (H3), we find

$$\begin{aligned}&\frac{p}{2} + (1 - p)\mathbf{y}_0(\sigma) > e^{2\epsilon}\left(\frac{p}{2} + (1 - p)\mathbf{y}_1(\sigma)\right) \\ &\quad\Leftrightarrow \frac{\frac{p}{2} + (1 - p)\mathbf{y}_0(\sigma)}{\frac{p}{2} + (1 - p)\mathbf{y}_1(\sigma)} > e^{2\epsilon} \\ &\quad\Leftrightarrow 1 + \frac{(1 - p)[\mathbf{y}_0(\sigma) - \mathbf{y}_1(\sigma)]}{\frac{p}{2} + (1 - p)\mathbf{y}_1(\sigma)} > \left(1 + 2\frac{1 - p}{p}\tau_D\right)^2\end{aligned}$$

$$\Leftrightarrow \frac{(1-p)[\mathbf{y}_0(\sigma)-\mathbf{y}_1(\sigma)]}{\frac{p}{2}+(1-p)\mathbf{y}_1(\sigma)} > 4\frac{1-p}{p}\tau_D + 4\frac{(1-p)^2}{p^2}\tau_D^2, \tag{H5}$$

where we used $\epsilon = \ln(1+2\frac{1-p}{p}\tau_D)$ in the second line, which is a result from Lemma 1. We can distinguish the following cases:

(1) If $(1-p)\tau_D/p < 1$, we replace the right side of Eq. (H5) by its upper bound, i.e.,

$$\begin{aligned}\frac{(1-p)[\mathbf{y}_0(\sigma)-\mathbf{y}_1(\sigma)]}{\frac{p}{2}+(1-p)\mathbf{y}_1(\sigma)} &> 8\frac{1-p}{p}\tau_D\\ \Leftrightarrow \frac{[\mathbf{y}_0(\sigma)-\mathbf{y}_1(\sigma)]}{4+8\frac{(1-p)}{p}\mathbf{y}_1(\sigma)} &> \tau_D.\end{aligned} \tag{H6}$$

(2) If $(1-p)\tau_D/p > 1$, or equivalently $p \in (0, 1/2)$, we replace the right side of Eq. (H5) by its upper bound, i.e.,

$$\begin{aligned}\frac{(1-p)[\mathbf{y}_0(\sigma)-\mathbf{y}_1(\sigma)]}{\frac{p}{2}+(1-p)\mathbf{y}_1(\sigma)} &> 8\frac{(1-p)^2}{p^2}\tau_D^2\\ \Leftrightarrow \frac{[\mathbf{y}_0(\sigma)-\mathbf{y}_1(\sigma)]}{4\frac{(1-p)}{p}+8\frac{(1-p)^2}{p^2}\mathbf{y}_1(\sigma)} &> \tau_D^2.\end{aligned} \tag{H7}$$

The definition $B \equiv \mathbf{y}_0(\sigma)/\mathbf{y}_1(\sigma)$ implies

$$\mathbf{y}_0(\sigma) = B\mathbf{y}_1(\sigma) \Leftrightarrow \mathbf{y}_0(\sigma) = B/(1+B). \tag{H8}$$

Inserting this into Eqs. (H6) and (H7), we have

$$\frac{B-1}{4(B+1)+8\frac{(1-p)}{p}} > \tau_D \tag{H9}$$

for the first case $p \in (0, 1/2)$, and

$$\frac{B-1}{4\frac{(1-p)}{p}(B+1)+8\frac{(1-p)^2}{p^2}} > \tau_D^2 \tag{H10}$$

for the second case $p \in [1/2, 1)$.

APPENDIX I: NUMERICAL SIMULATION DETAILS: IRIS DATA SET

In this Appendix, we explain how our quantum classifier is implemented and then use a generic metric to evaluate the performance of our defense protocol.

1. Implementation of quantum classifier

Our quantum classifier is composed of four main ingredients, i.e., the unitary U_{prep} for state preparation, the parameterized quantum circuits $U(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are to be optimized, the final projective measurements $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$ in the σ_z basis and the depolarization channel $\mathcal{N}_p$ that is conditionally applied at testing time. Note that it doesn't matter where the depolarization channel is placed in the circuit since results only depend on the product $\prod_{i=1}^{l}(1-p_i)$, where $p \equiv 1-\prod_{i=1}^{l}(1-p_i)$. Our circuit is shown in Fig. 9, composed of two qubits, where each entry of the classical input vector is separately encoded into the amplitude of the quantum state in the computational basis. The state preparation unitary U_{prep}, i.e., the computation of parameters $\{\mathbf{x}_i\}_{i=1}^{5}$, follows from Ref. [47]. This $U(\boldsymbol{\theta})$ is composed of five layers, where each layer consist of trainable single-qubit gates and two-qubits gates as shown in the upper right panel of Fig. 9, highlighted by the dashed box. The layers are then sequentially applied to form $U(\boldsymbol{\theta})$ [21]. The mathematical representation of $U(\boldsymbol{\theta}_{i,1}) = R_Z(\boldsymbol{\theta}_{i,1})R_Y(\boldsymbol{\theta}_{i+1,1})R_Z(\boldsymbol{\theta}_{i+2,1})$ and the total number of trainable parameters is 25.

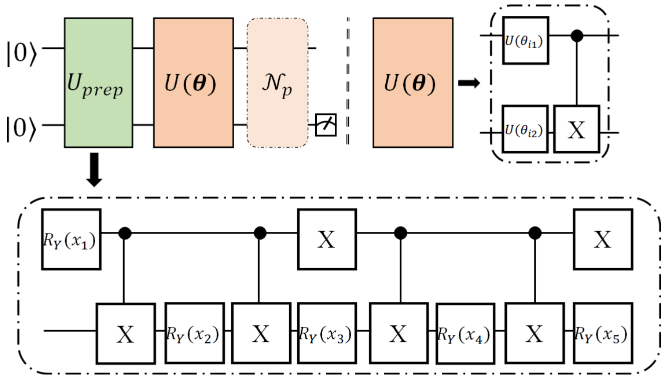

FIG. 9. *Our binary QNN classifier.* The upper left panel shows the main structure of the quantum classifier and the upper right panel illustrates one layer in the implementation of the trainable unitary and we employ five layers in total. The lower panel is the quantum circuit that encodes the classical input data into a quantum state.

2. Evaluation

An evaluation metric broadly used in classical adversarial learning is the conventional accuracy, which measures the prediction accuracy of the test data set under adversarial attacks with respect to different bounded norms [19,48]. The mathematical expression for the conventional accuracy Acc_c is

$$\text{Acc}_c = \frac{\sum_{i=1}^{|D_{\text{Te}}|}\mathbf{1}_{\bar{c}_i=c_i^*}}{|D_{\text{Te}}|}, \tag{I1}$$

where $|D_{\text{Te}}|$ is the size of the test data set, $\bar{c}_i$ and c_i^* are the predicted and real labels of the ith test example. Here $\mathbf{1}_{\bar{c}_i=c_i^*}$ is the indicator function, which takes the value 1 when $\bar{c}_i = c_i^*$ and is 0 otherwise. Using the depolarization noise $p = 0.5, 0.8$, and $\tau_D = 0.015$, we explore the trade-off between adversarial robustness and the conventional accuracy for our classifier. Let $L \in (0, 0.7]$ and $n_{\text{samp}} = 300$. The number of iterations used to generate adversarial attacks is set to 50 without early stopping. Figure 10 illustrates the simulation results under $p = 0, 0.5, 0.8$. We can see how our protocol increases the robustness against l_2 norm attacks with increasing p. For instance, the conventional accuracy of our baseline $(p = 0)$ drops to zero when $L = 0.4$, while the conventional accuracy remains nonzero for both $p = 0.5$ and $p = 0.8$. In addition, a larger depolarization noise p promises a better robustness against large L. Specifically, when $L = 0.1$, the conventional accuracy when $p = 0.8$ is slightly less than when $p = 0.5$. However, with increased L, the conventional accuracy when $p = 0.8$ outperforms the case when $p = 0.5$. Also when $L = 0.5$, both baseline and $p = 0.5$ cases have the zero conventional accuracy, while the setting $p = 0.8$ gives nonzero conventional accuracy.

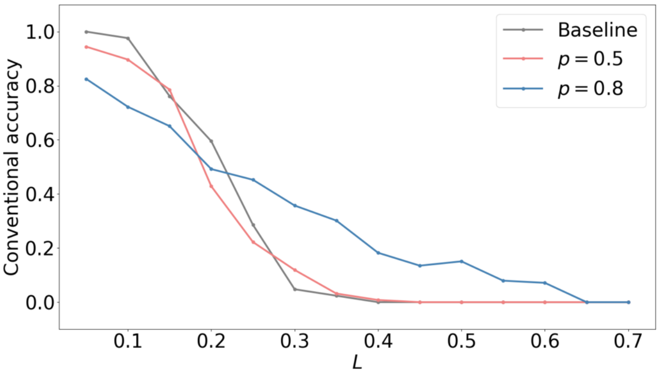

FIG. 10. *Conventional accuracy for different depolarization noise p.* We denote L as the maximum l_2 bounded-norm used in the adversarial attack. The conventional accuracy corresponding to $p = 0.5, 0.8$ is with respect to L is in red and blue, respectively. The label baseline refers to the conventional accuracy with when $p = 0$.

APPENDIX J: NUMERICAL SIMULATION DETAILS: LBM DATA SET

Here we give details of our numerical simulation for the data set based on Ref. [34].

The implementation of the binary QNN is exhibited in the left panel of Fig. 11. Specifically, the encoding unitary U_E is defined as

$$U_E = U_{\text{etg}}\big[\otimes_{i=11}^{20} R_Y(x_i)\big] U_{\text{etg}}\big[\otimes_{i=1}^{10} R_Y(x_i)\big], \tag{J1}$$

where U_{etg} (highlighted by the dark blue line) refers to the entanglement layer such that CNOT gates are applied to the adjacent qubits in sequence. The right panel presents the trainable unitary $U(\boldsymbol{\theta}) = \prod_{l=1}^{2} U_l(\boldsymbol{\theta})$. Mathematically, the lth layer satisfies $U_l(\boldsymbol{\theta}) = U_{\text{etg}}[\otimes_{i=1}^{10} U(\boldsymbol{\theta}_{li})]$ with $\boldsymbol{\theta}_{li} = [\alpha, \beta, \gamma]$ and $U(\boldsymbol{\theta}_{li}) = R_Z(\gamma) R_Y(\beta) R_Z(\alpha)$. The hyperparameter setting is as follows. The layer number of variational quantum circuits is set as $L = 2$. The total number of trainable parameters is 60. The number of epochs used in classical optimization is 20. The gradient descent optimizer is employed to optimize these parameters. The random seed is set as 1.

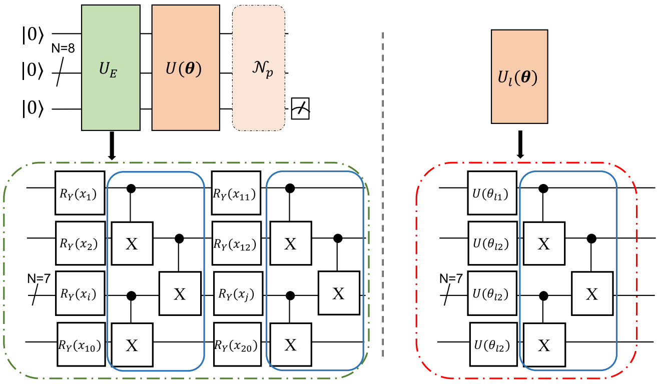

FIG. 11. *The implementation of QNN used to learn the synthetic data set $\mathcal{D}$.* The left panel shows the main architecture of QNN. The right panel exhibits the basic component to construct the trainable unitary $U_l(\boldsymbol{\theta})$.

APPENDIX K: THE ROBUSTNESS AND PRIVACY BUDGET OF THE PAULI NOISE CHANNEL

An important class of examples is the single-qubit Pauli channel (which includes the bit-flip and dephasing channels as special cases) acting on input state σ as

$$\mathcal{N}_{\text{Pauli}}(\sigma) = p_{id}\sigma + \sum_{i=x,y,z} p_i(\mathbf{1}^{\otimes D-1} \otimes \sigma_i)\sigma(\mathbf{1}^{\otimes D-1} \otimes \sigma_i), \tag{K1}$$

where $p_{id} + p_x + p_y + p_z = 1$. We first show the conditions under which a quantum classifier is robust against this noise. For simplicity, we use binary classification, which can be straightforwardly extended to multiclass classification. Let the output of the classifier of the noiseless circuit be $\mathbf{y}_0(\sigma) = \text{Tr}(\Pi_0\sigma)$, where $\Pi_0 = \mathbf{1}^{\otimes D-1} \otimes |0\rangle\langle 0|$ and $\mathbf{y}_1(\sigma) = \text{Tr}(\Pi_1\sigma)$, where $\Pi_1 = \mathbf{1}^{\otimes D-1} \otimes |1\rangle\langle 1|$. If the Pauli channel above is added only to the final state (which we now denote as σ for simplicity of notation), then the output of the noisy classifier becomes

$$\tilde{\mathbf{y}}_0(\sigma) = \text{Tr}[\Pi_0\mathcal{N}_{\text{Pauli}}(\sigma)] = p_{id}\text{Tr}(\Pi_0\sigma) + \sum_{i=x,y,z} \text{Tr}(\tilde{\Pi}_{0,i}\sigma), \tag{K2}$$

where $\tilde{\Pi}_{0,i} \equiv (\mathbf{1}^{\otimes D-1} \otimes \sigma_i)\Pi_0(\mathbf{1}^{\otimes D-1} \otimes \sigma_i)$. Then, generalizing the single-qubit result in Theorem 1 of Ref. [10] to our multiqubit case, we still have $\tilde{\Pi}_{0,x} = \Pi_1 = \tilde{\Pi}_{0,y}$ and $\tilde{\Pi}_{0,z} = \Pi_0$. Inserting these into Eq. (K2), we arrive at the same result as Theorem 1 in Ref. [10]:

$$\tilde{\mathbf{y}}_0(\sigma) = [1 - 2(p_x + p_y)]\text{Tr}(\Pi_0\sigma) + (p_x + p_y). \tag{K3}$$

Then it is straightforward to show that the conditions (i) $\mathbf{y}_0(\sigma) > 1/2$ implies $\tilde{\mathbf{y}}_0(\sigma) > 1/2$ and (ii) if $\mathbf{y}_0(\sigma) < 1/2$ implies $\tilde{\mathbf{y}}_0(\sigma) < 1/2$ are both true when $p_x + p_y \leqslant 1/2$.

Note that this robustness condition holds always for dephasing noise where $p_{id} = 1 - p_z$ and $p_x = p_y = 0$. Bit-flip noise is the case where $p_{id} = 1 - p_x$ and $p_y = 0 = p_z$. Therefore, robustness also holds for bit-flip noise when $p_x \leqslant 1/2$.

Now we show how the Pauli channel also gives rise to quantum differential privacy. The proof follows very similarly to the depolarization channel case in Appendix D. So, now for $k = 0, 1$, we have

$$\begin{aligned}\frac{\tilde{\mathbf{y}}_k(\rho)}{\tilde{\mathbf{y}}_k(\sigma)} - 1 &= \frac{[1 - 2(p_x + p_y)]\text{Tr}[\Pi_k(\rho - \sigma)]}{[1 - 2(p_x + p_y)]\text{Tr}(\Pi_k\sigma) + p_x + p_y}\\ &\leqslant \frac{[1 - 2(p_x + p_y)]\tau_D}{p_x + p_y} \leqslant e^{\epsilon} - 1,\end{aligned} \tag{K4}$$

where we used $0 \leqslant p_x + p_y \leqslant 1/2$ and the same inequalities as in Appendix D. Here τ_D is also the upper bound to the trace distance between the input states of the circuit. Therefore, the Pauli channel also has quantum differential privacy when $0 \leqslant p_x + p_y \leqslant 1/2$, where the privacy budget is

$$\epsilon = \ln\left((1 - 2(p_x + p_y))\tau_D/(p_x + p_y) + 1\right). \tag{K5}$$

Thus, if the noise is added at the end of the circuit, all cases of Pauli noise where $0 < p_x + p_y < 1/2$ satisfies both robustness against noise and gives rise to quantum differential privacy. From the above expression, we can also see that the adversarial robustness bound is independent of any details of the classifier.

We can also look at wider classes of examples by noting that the above condition of robustness against noise is not strictly necessary (although it would be desired). Robustness against noise as defined in our paper means that the accuracy of the model *does not diminish at all* as noise is added. If we used a noise model that our classification model is not completely robust against, we would need to compromise on the accuracy. This would mean that a trade-off with quantum differential privacy would need to be considered. How much the accuracy would be compromised in the case of incomplete robustness is captured in Theorem 1. Thus, if the accuracy obtained after noise is added is still above one's threshold, it is still possible to use our technique to utilize noise to obtain quantum differential privacy. Therefore, this gives the possibility of using more general noise models that give protection against adversaries.

Given that even the question of which noise models gives rise to robustness or incomplete robustness for quantum classification has not been fully investigated in the literature (which would be an independent useful work in its own right). This means that generalization of the results in this paper requires more dedicated analysis. Therefore, a more systematic and complete characterization of what types of noises would be beneficial for robustness against adversaries would be the subject of future work.

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

[2] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, Phys. Rev. X **8**, 031027 (2018).

[3] K. Temme, S. Bravyi, and J. M. Gambetta, Error Mitigation for Short-Depth Quantum Circuits, Phys. Rev. Lett. **119**, 180509 (2017).

[4] C. Dwork, Differential privacy, *Encyclopedia of Cryptography and Security* (Springer, Boston, MA, 2011), pp. 338–340.

[5] L. Gammaitoni, P. Hänggi, P. Jung, and F. Marchesoni, Stochastic resonance, Rev. Mod. Phys. **70**, 223 (1998).

[6] L. Roberts, Picture coding using pseudo-random noise, IRE Trans. Inf. Theory **8**, 145 (1962).

[7] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, How to escape saddle points efficiently, in *Proceedings of the International Conference on Machine Learning, 6–11 August 2017* (International Convention Centre, Sydney, Australia, 2017), Vol. 70, pp. 1724–1732.

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

[9] E. Grant, M. Benedetti, S. Cao, A. Hallam, J. Lockhart, V. Stojevic, A. G. Green, and S. Severini, Hierarchical quantum classifiers, npj Quantum Inform. **4**, 1 (2018).

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

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in *International Conference Learning Representations April 14th to April 16th 2014* (Banff, Canada, 2014).

[12] 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, New York, 2011), pp. 43–58.

[13] A. Kurakin, I. J. Goodfellow, and S. Bengio, Adversarial machine learning at scale, in *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017* (Toulon, France, 2017).

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

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

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

[17] I. Goodfellow, P. McDaniel, and N. Papernot, Making machine learning robust against adversarial inputs, Commun. ACM **61** (2018).

[18] X. Yuan, P. He, Q. Zhu, and X. Li, Adversarial examples: Attacks and defenses for deep learning, IEEE Trans. Neural Net. Learn. Syst. **30**, 2805 (2019).

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

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).

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

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, arXiv:1412.6572.

[23] M. Sharif, L. Bauer, and M. K. Reiter, On the suitability of lp-norms for creating and preventing adversarial examples, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, Salt Lake City, UT, 2018), pp. 1605–1613.

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

[25] 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* (ACM, New York, 2019).

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

[27] This is an extension from Theorem 2 in Ref. [10] to beyond $K = 2$ and follows by an inductive application of Eq. (9).

[28] This appears in Theorem 2 in Ref. [10]. Also see Ref. [10] for a list of common types of noise that binary quantum classifers are naturally robust against as well as interesting encoding

strategies to induce robustness when the classifiers are not naturally robust.

[29] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning* (MIT Press, Cambridge, MA, 2018).

[30] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, Robustness may be at odds with accuracy, arXiv:1805.12152.

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

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

[33] W. Huggins, P. Patil, B. Mitchell, K. B. Whaley, and M. Stoudenmire, Towards quantum machine learning with tensor networks, Quantum Sci. Technol. **4**, 024001 (2019).

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

[35] M. Benedetti, D. Garcia-Pintos, O. Perdomo, V. Leyton-Ortega, Y. Nam, and A. Perdomo-Ortiz, A generative modeling approach for benchmarking and training shallow quantum circuits, npj Quantum Inform. **5**, 1 (2019).

[36] P.-L. Dallaire-Demers and N. Killoran, Quantum generative adversarial networks, Phys. Rev. A **98**, 012324 (2018).

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

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

[39] R. A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics **7**, 179 (1936).

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

[41] M. Mottonen, J. J. Vartiainen, V. Bergholm, and M. M. Salomaa, Transformation of quantum states using uniformly controlled rotations, Quant. Inf. Comp. **5**, 467 (2005).

[42] Y. Liu, X. Chen, C. Liu, and D. Song, Delving into transferable adversarial examples and black-box attacks, in *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017* (ICLR, Toulon, France, 2017).

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

[44] 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 2018* (ICLR, Vancouver, Canada, 2018).

[45] For nonnumerical functions, the exponential mechanism is employed.

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

[47] M. Plesch and Č. Brukner, Quantum-state preparation with universal gate decompositions, Phys. Rev. A **83**, 032302 (2011).

[48] E. Wong and Z. Kolter, Provable defenses against adversarial examples via the convex outer adversarial polytope, in *International Conference on Machine Learning* (Stockholm, Sweden, 2018), pp. 5283–5292.