Universal Adversarial Examples and Perturbations for Quantum Classifiers

Weiyuan Gong[1] and Dong-Ling Deng[1,2,*]

[1]*Center for Quantum Information, IIIS, Tsinghua University, Beijing 100084, People's Republic of China*
[2]*Shanghai Qi Zhi Institute, 41th Floor, AI Tower, No. 701 Yunjin Road, Xuhui District, Shanghai 200232, China*

Quantum machine learning explores the interplay between machine learning and quantum physics, which may lead to unprecedented perspectives for both fields. In fact, recent works have shown strong evidences that quantum computers could outperform classical computers in solving certain notable machine learning tasks. Yet, quantum learning systems may also suffer from the vulnerability problem: adding a tiny carefully-crafted perturbation to the legitimate input data would cause the systems to make incorrect predictions at a notably high confidence level. In this paper, we study the universality of adversarial examples and perturbations for quantum classifiers. Through concrete examples involving classifications of real-life images and quantum phases of matter, we show that there exist universal adversarial examples that can fool a set of different quantum classifiers. We prove that for a set of k classifiers with each receiving input data of n qubits, an $O(\frac{\ln k}{2^n})$ increase of the perturbation strength is enough to ensure a moderate universal adversarial risk. In addition, for a given quantum classifier we show that there exist universal adversarial perturbations, which can be added to different legitimate samples and make them to be adversarial examples for the classifier. Our results reveal the universality perspective of adversarial attacks for quantum machine learning systems, which would be crucial for practical applications of both near-term and future quantum technologies in solving machine learning problems.

Keywords: Quantum machine learning, quantum classifiers, adversarial examples, measure concentration, quantum no free lunch theorem

INTRODUCTION

Machine learning, or more broadly artificial intelligence, has achieved dramatic success over the past decade [1, 2] and a number of problems that were notoriously challenging, such as playing the game of Go [3, 4] or predicting protein structures [5], have been cracked recently. In parallel, the field of quantum computing [6] has also made remarkable progress in recent years, with the experimental demonstration of quantum supremacy marked as the latest milestone [7, 8]. The marriage of these two fast-growing fields gives birth to a new research frontier—quantum machine learning [9–11]. On the one hand, machine learning tools and techniques can be exploited to solve difficult problems in quantum science, such as quantum many-body problems [12], state tomography [13], topological quantum compiling [14], structural and electronic transitions in disordered materials [15], non-locality detection [16], and classification of different phases of matter and phase transitions [17–24]. On the other hand, new quantum algorithms running on quantum devices also possess the unparalleled potentials to enhance, speed up, or innovate machine learning [9–11]. Notable examples along this direction include the Harrow-Hassidim-Lloyd algorithm [25], quantum principal component analysis [26], quantum generative models [27–29], and quantum support vector machines [30], etc. Without a doubt, the interaction between machine learning and quantum physics will benefit both fields [11].

In classical machine learning, it has been shown that classifiers based on deep neural networks are rather vulnerable in adversarial scenarios [31, 32]: adding a tiny amount of carefully-crafted noises, which are even imperceptible to human eyes and ineffective to traditional methods, into the original legitimate data may cause the classifiers to make incorrect predictions at a notably high confidence level. A cel-

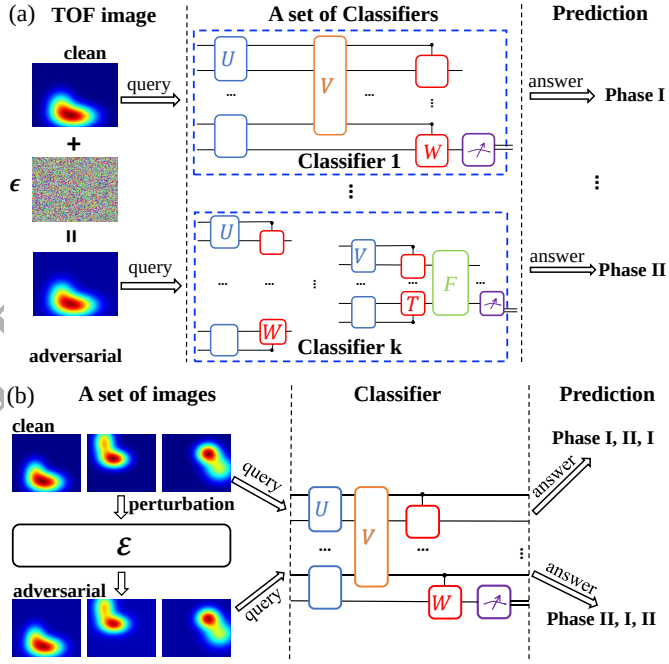

FIG. 1. A schematic illustration of universal adversarial examples and perturbations. (a) Universal adversarial examples: a set of quantum classifiers can be trained to assign phase labels to different time-of-flight images, which can be obtained directly in cold atom experiments. Adding a small amount of carefully crafted noise to a certain image could make it become a universal adversarial example, namely the new crafted image could deceive all the classifiers in the set. (b)Universal adversarial perturbations: adding the same carefully-constructed noise to a set of images could make them all become adversarial examples for a given quantum classifier.

ebrated example that clearly showcases the vulnerability of deep learning was observed by Szegedy *et al.* [33], where an image of a panda will be misclassified as a gibbon after adding an imperceptible amount of noises. The crafted input samples that would deceive the classifiers are called adversarial examples. Now, it is widely believed that the existence of adversarial examples is ubiquitous in classical machine learning—almost all learning models suffer from adversarial attacks, regardless of the input data types and the details of the neural networks [31, 32]. More recently, the vulnerability of quantum classifiers has also been studied, sparking a new research frontier of quantum adversarial machine learning [34–39]. In particular, Ref. [34] explored different adversarial scenarios in the context of quantum machine learning and have demonstrated that, with a wide range of concrete examples, quantum classifiers are likewise highly vulnerable to crafted adversarial examples. This emergent research direction is growing rapidly, attracting more and more attentions across communities. Yet, it is still in its infancy and many important issues remain unexplored.

In this paper, we consider such an issue concerning the universality of adversarial examples and perturbations for quantum classifiers. We ask two questions: (i) whether there exist universal adversarial examples that could fool a set of different quantum classifiers? (ii) whether there exist universal adversarial perturbations, which when added to different legitimate input samples could make them become adversarial examples for a given quantum classifier? Based on extensive numerical simulations and analytical analysis, we give affirmative answers to both questions. For (i), we prove that, by exploring the concentration of measure phenomenon [40], an $O(\frac{\ln k}{2^n})$ increase of the perturbation strength is enough to ensure a moderate universal adversarial risk for a set of k quantum classifiers with each receiving input data of n qubits; For (ii), we prove that, based on the quantum no free lunch theorem [41, 42], the universal adversarial risk is bounded from both below and above and approaches unit exponentially fast as the number of qubits for the quantum classifier increases. We carry out extensive numerical simulations on concrete examples involving classifications of real-life images and quantum phases of matter to demonstrate how to obtain universal adversarial examples and perturbations in practice.

UNIVERSAL ADVERSARIAL EXAMPLES

To begin with, we first introduce some concepts and notations. Consider a classification task in the setting of supervised learning, where we assign a label $s \in S$ to an input data sample $\rho \in \mathcal{H}$, with S being a countable label set and $\mathcal{H}$ the set of all possible samples. The training set is denoted as $\mathcal{S}_N = \{(\rho_1, s_1), ..., (\rho_N, s_N)\}$, where $\rho_i \in \mathcal{H}$, $s_i \in S$, and N is the size of the training set. Essentially, the task of classification is to learn a function (called a hypothesis function) $h : \mathcal{H} \to S$, which for a given input $\rho \in \mathcal{H}$ outputs a label s. We denote the *ground truth* function as $t : \mathcal{H} \to S$, which gives the true classification for any $\rho \in \mathcal{H}$. For the purpose in this paper, we suppose that after the training process the hypothesis function match the ground truth function on the training set, namely $h(\rho) = t(\rho), \forall \rho \in \mathcal{S}_N$. We consider a set of k quantum classifiers $\mathcal{C}_1, ..., \mathcal{C}_k$ with corresponding hypothesis functions h_i $(i = 1, ..., k)$ and introduce the following definitions to formalize our results.

Definition 1. We suppose the input sample ρ is chosen from $\mathcal{H}$ according to a probability measure μ and $\mu(\mathcal{H}) = 1$. For h_i, we define $\mathcal{E}_i = \{\rho \in \mathcal{H} | h_i(\rho) \neq t(\rho)\}$ as the misclassified set, and the *risk* for $\mathcal{C}_i$ is denoted as $\mu(\mathcal{E}_i)$.

Definition 2. Consider a metric over $\mathcal{H}$ with the distance measure denoted as $D(\cdot)$. Then the ϵ-expansion of a subset $\mathcal{H}' \subseteq \mathcal{H}$ is defined as: $\mathcal{H}'_\epsilon = \{\rho | D_{\min}(\rho, \mathcal{H}') \leq \epsilon\}$, where $D_{\min}(\rho, \mathcal{H}')$ denotes the minimum distance between ρ and any $\rho' \in \mathcal{H}'$. In the context of adversarial learning, a perturbation within distance ϵ added to the legitimate input sample $\rho \in \mathcal{E}_{i,\epsilon} = \{\rho' | D_{\min}(\rho', \mathcal{E}_i)\}$ can shift it to some misclassified one for the quantum classifier $\mathcal{C}_i$. Hence, we define the adversarial risk for $\mathcal{C}_i$ as $\mu(\mathcal{E}_{i,\epsilon})$. Similarly, the universal adversarial risk for a set of k quantum classifiers is defined as $R = \mu(\mathcal{E}_\epsilon)$, where $\mathcal{E}_\epsilon = \cap_{i=1}^k \mathcal{E}_{i,\epsilon}$ denotes the set of universal adversarial samples.

For technique simplicity and convenience, we focus on $\mathcal{H} = SU(d)$ (the special unitary group) with the Hilbert-Schmidt distance $D_{\text{HS}}(\rho, \rho')$ and Haar probability measure [43]. We mention that the input data ρ can be either classical or quantum in general. We treat both cases on the same footing since we can always encode the classical data into quantum states. We also note that any input state could be prepared by acting a unitary transformation on a certain initial state (e.g., the $|00\cdots0\rangle$ state) and hence the classification of quantum states is in some sense equivalent to the classification of unitary transformations. Now, we are ready to present one of our main results.

Theorem 1. Consider a set of k quantum classifiers $\mathcal{C}_i$, $i = 1, ..., k$ and let $\mu(\mathcal{E})_{\min}$ be the minimum risk among $\mu(\mathcal{E}_i)$. Suppose $\rho \in SU(d)$ and a perturbation $\rho \to \rho'$ occurs with $D_{\text{HS}}(\rho, \rho') \leq \epsilon$, then we can ensure that the universal adversarial risk is bounded below by R_0 if

$$\epsilon^2 \geq \frac{4}{d} \ln \left[\frac{2k}{\mu(\mathcal{E})_{\min}(1 - R_0)} \right]. \tag{1}$$

Proof. We give the main idea and intuition here. The full proof is a bit technically involved and thus left to the Supplementary Materials. The first step is to prove that for a single quantum classifier $\mathcal{C}_i$, we can ensure that its adversarial risk is bounded below by $R_{0,i}$ if $\epsilon^2 \geq \frac{4}{d} \ln \left[\frac{2}{\mu(\mathcal{E}_i)(1-R_{0,i})} \right]$. This can be done by exploring the concentration of the measure phenomenon for $SU(d)$ equipped with the Haar measure and Hilbert-Schmidt metric [35]. Next, we use the De Morgan's laws in set theory to deduce that $\mu(\mathcal{E}_\epsilon) \geq 1 - k + \sum_{i=1}^k \mu(\mathcal{E}_i)$. In the last step, we choose $R_{0,i} = \frac{k-1+R_0}{k}$ and replace $\mu(\mathcal{E}_i)$ by $\mu(\mathcal{E})_{\min}$ to increase ϵ a little bit for each $\mathcal{C}_i$. This leads to Eq. (1) and complete the proof.

The above theorem implies that for a set of k quantum

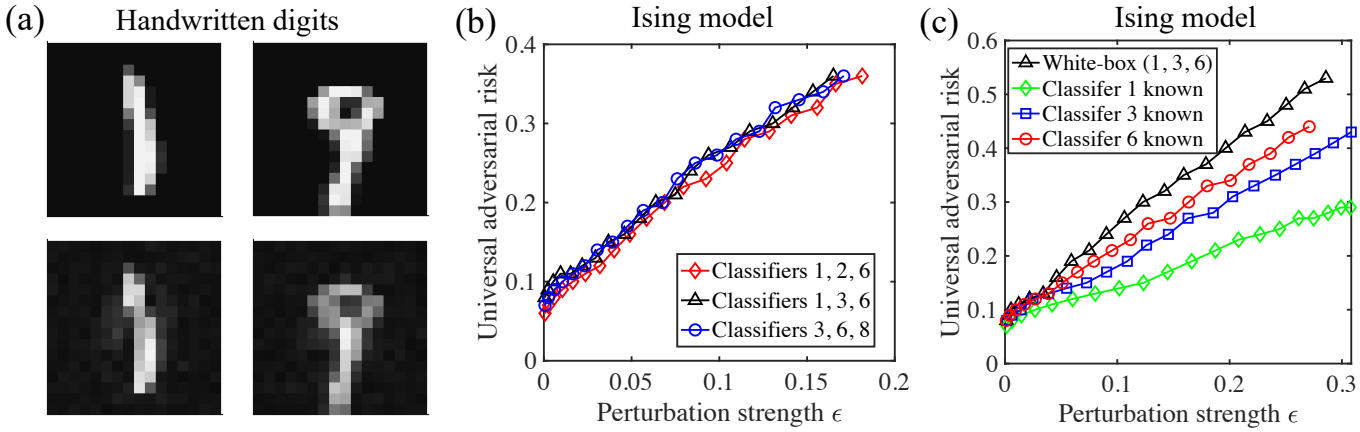

FIG. 2. Numerical results on universal adversarial examples. In this figure, the adversarial examples are obtained through the qBIM algorithm with step size $\alpha = 0.02$. (a) Illustrating samples: the clean (first row) and the corresponding universal adversarial handwritten digit images (second row) that can deceive all eight quantum classifiers. (b) The universal adversarial risk as a function of the perturbation strength ϵ for different subsets of the classifiers in classifying the ground states of the 1D transverse field Ising model. Here, we consider the white-box attack scenario and the universal adversarial risk is defined as the ratio of test samples that deceive all three classifiers in each subset. (c) Results for attacking a subset of classifiers consisting of the first, third, and sixth classifiers in classifying the ground states of the Ising model, under a white-box black-box hybrid setting. Here, we assume that only one of the classifiers is known to the attacker, and for comparison the black curve with triangles plots the result for the white-box attack case. For more details, see the Supplementary Material.

classifiers with each receiving input data of n qubits (thus $d = 2^n$), an $O(\frac{\ln k}{2^n})$ increase of the perturbation strength would guarantee a moderate universal adversarial risk lower bounded by R_0. As n increases, the lower bound of ϵ approaches zero exponentially. In other words, an exponentially small adversarial perturbation could result in universal adversarial examples that can deceive all k classifiers with constant probability. This is a fundamental feature of quantum classifiers in high dimensional Hilbert space due to the concentration of the measure phenomenon, independent of their specific structures and the input datasets.

Although the above theorem indicates the existence of universal adversarial examples in theory, it is still unclear how to obtain these universal examples in practice. To deal with this issue, in the following we provide concrete examples involving classifications of hand-writing digit images and quantum phases with extensive numerical simulations. We mention that, in the classical adversarial machine learning literature, universal adversarial examples have also been shown to exist in real applications. For instance, in Ref. [44] it is shown that an attacker can fool (such as dodging or impersonation) a number of the state-of-the-art face-recognition systems by simply wearing a pair of carefully-crafted eyeglasses. For our purpose, we consider a set of eight quantum classifiers with different structures, labeled by numbers from 1 to 8. The classifiers 1 and 2 are two quantum convolutional neural networks (QCNNs) [45] and the classifiers $3-8$ are other typical multi-layer variational quantum circuits with depths from five through ten. The detailed descriptions of these quantum classifiers are given in the Supplementary Materials.

The first example we consider is the classification of handwritten-digit images in the MNIST dataset [46], which is a prototypical testbed for benchmarking various machine learning scenarios. This dataset consists of gray-scale images of handwritten digits from 0 through 9, with each of them contains 28×28 pixels. We reduce the size of the images to 16×16, so that we can simulate the learning and attacking process of the quantum classifiers with moderate classical computational resources. We use amplitude encoding to map the input images into quantum states and the cross-entropy as the loss function for training and adversarial attacking. After training, we use the quantum-adapted basic iterative method (qBIM) [47] to obtain the adversarial examples. The details of the training and adversarial attacking process are provided in the Supplementary Materials. We mention that our quantum classifiers can achieve comparable training and validation accuracy as for classical classifiers. In Fig. 2 (a), we display two universal adversarial examples for digits 1 and 9, which can deceive *all* eight quantum classifiers at a high-confidence level. Notably, these universal adversarial examples only differ from the original legitimate ones slightly and they can be easily identified by human eyes. In fact, the fidelity between the adversarial and legitimate samples is about 96%, which is fairly high given that the Hilbert dimension involved is not very large ($d = 256$ for this case).

The above discussion concerns the vulnerability of quantum classifiers in classifying classical data (images). Yet, unlike classical classifiers that can only take classical data as input, quantum classifiers can also directly classify quantum data (states) produced by quantum devices. This is a notable distinction between quantum and classical classifiers. To demonstrate the existence of universal adversarial examples for quantum classifiers in classifying quantum data, we consider classifying the ground state of the one-dimensional (1D)

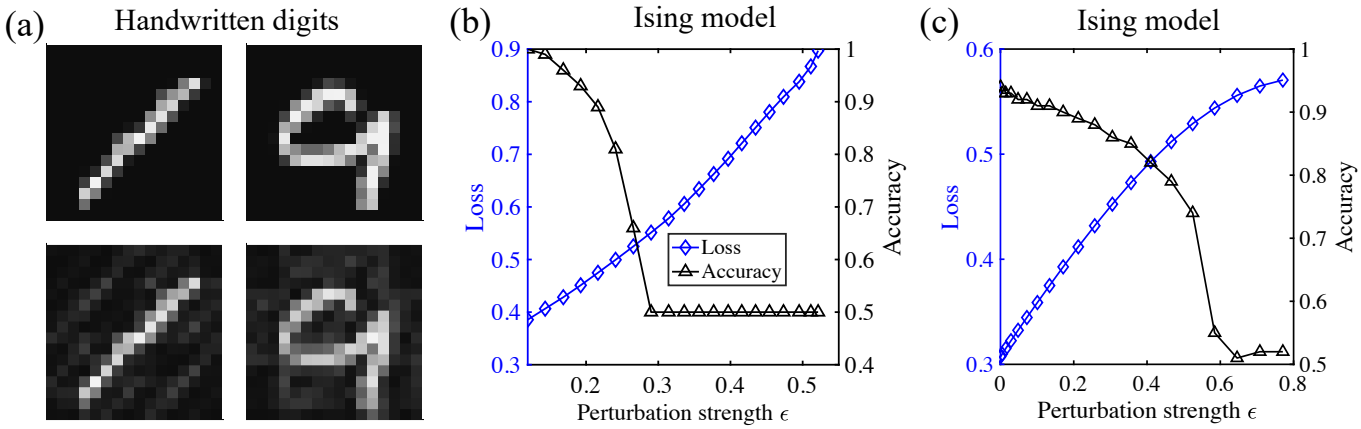

FIG. 3. Numerical results on universal adversarial perturbations. Similar to Fig. 2, in this figure the adversarial perturbations are also obtained by the qBIM algorithm with step size $\alpha = 0.02$. (a) Illustrating samples: the clean (first row) and corresponding adversarial examples (second row) that can fool the second quantum classifier, which is a quantum convolutional neural network. These two adversarial images (second row) are obtained by adding the same perturbation to the original legitimate ones shown in the first row. (b) The loss and accuracy as functions of the perturbation strength ϵ for the second classifier in classifying the ground states of H_{Ising}. (c) A similar result for the eighth classifier. Throughout this figure, the white-box attack is considered. For more details, see the Supplementary Material.

transverse field Ising model:

$$H_{\text{Ising}} = -\sum_{i=1}^{L-1} \sigma_i^z \sigma_{i+1}^z - J_x \sum_{i=1}^{L} \sigma_i^x, \tag{2}$$

where J_x denotes the strength of the transverse field and σ_i^x and σ_i^z are the Pauli matrices for the i-th spin. This Hamiltonian maps to free fermions via Jordan-Wigner transformation and is exactly solvable. Its ground state features a quantum phase transition at $J_x = 1$, between paramagnetic phase with $J_x > 1$ and ferromagnetic phase with $0 < J_x < 1$. We consider classifying these two different phases by the eight quantum classifiers mentioned above, with the ground state as input data. We sample the Hamiltonian with varying J_x from 0 to 2 and compute their corresponding ground states. These quantum states with their corresponding labels form the dataset required.

In Fig. 2(b), we consider three subsets of quantum classifiers in classifying the ground states of H_{Ising}, under the white-box attack setting (namely the attacker has full information about the learned model and the learning algorithm). We find that universal adversarial examples indeed exist for classifying quantum states, regardless of the internal structures of the classifiers. As the perturbation strength ϵ increases, the universal adversarial risk increases roughly linearly with ϵ. With a perturbation strength $\epsilon = 0.18$, we find that 37% of the test samples could become universal adversarial examples for each subset of the classifiers. In Fig. 2(c), we consider a white-box black-box hybrid scenario, where the attacker knows only the full information about one classifier in the subset and does not have any information about the rest ones. The motivation of this consideration is to study the transferability of universal adversarial examples. From Fig. 2(c), we find that even with limited partial information, the adversary is still able to create universal adversarial examples, indicating a notable transferability property of these examples. The universal adversarial risk also increases linearly with ϵ, but it is noticeably smaller than that for the white-box case. This is consistent with the intuition that the more information the attacker has the easier to create adversarial examples.

UNIVERSAL ADVERSARIAL PERTURBATIONS

In the above discussion, we demonstrate, with both theoretical analysis and numerical simulations, that there exist universal adversarial examples that could deceive a set of distinct quantum classifiers. We now turn to the second question and show that there exist universal adversarial perturbations that can be added to different legitimate samples and make them adversarial to a given quantum classifier $\mathcal{C}$. Without loss of generality, we may consider a unitary perturbation $\hat{\epsilon} : \mathcal{H} \to \mathcal{H}$ as means of adversarial attack for all input samples. We denote the misclassified set as $\mathcal{E} = \{\rho \in \mathcal{H} | h(\rho) \neq t(\rho)\}$ and consequently the unitary adversarial set as $\mathcal{E}_{\hat{\epsilon}} = \{\hat{\epsilon}^{-1}(\rho) | \rho \in \mathcal{E}\}$.

Theorem 2. For an adversarial perturbation with unitary operator $\hat{\epsilon}$ and n samples $\rho_1, ..., \rho_n$ chosen from $\mathcal{H}$ according to the Haar measure, the performance of the quantum classifier $\mathcal{C}$ with $\hat{\epsilon}(\rho_1), ..., \hat{\epsilon}(\rho_n)$ as input samples is bounded by:

$$|R_E - \mu(\mathcal{E})| \leq \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\delta}\right)} \tag{3}$$

with probability at least $1 - \delta$ $(0 < \delta < 1)$. Here R_E is the empirical error rate defined as the ratio of the misclassified samples and $\mu(\mathcal{E})$ is the risk for $\mathcal{C}$. In addition, the expectation of the risk over all ground truth t and training set $\mathcal{S}_N$ is bounded below by:

$$\mathbb{E}_t[\mathbb{E}_{\mathcal{S}_N}[\mu(\mathcal{E})]] \geq 1 - \frac{d'}{d(d+1)}(N^2 + d + 1), \tag{4}$$

where $d = \dim(\mathcal{H})$ is the dimension of the input data and $d' = |S|$ is the number of output labels.

Proof. We only sketch the major steps here and leave the details of the full proof to the Supplementary Materials. Noting that unitary transformations are invertible, the unitary perturbation operator $\hat{\epsilon}$ will transfer samples in $\mathcal{E}_{\hat{\epsilon}}$ into the misclassified set $\mathcal{E}$, and we can therefore deduce that $\mu(\mathcal{E}) = \mu(\mathcal{E}_{\hat{\epsilon}})$. Then from the definition of $\mu(\mathcal{E})$, the Ineq. (3) follows straightforwardly by applying the Hoeffding's inequality [48]. The derivation of the Ineq. (4) relies on the recent works about reformulation of the no free lunch theorem in the context of quantum machine learning [41, 42] (see the Supplementary Materials for details).

This theorem indicates that in the limit $d \to \infty$, the expectation of the risk for a general quantum classifier goes to unit, independent of its structure and the training algorithm. For a fixed d, the lower bound of such an expectation decreases as the number of the output labels or the size of the training set increase. Adding an identical adversarial unitary perturbation to all possible data samples will not increase the risk on average. However, it is still possible for such a perturbation to increase the ratio of misclassified samples for a given finite set of n original samples. In the following, we carry out numerical simulations and show how to obtain the universal adversarial perturbations in classifying images of handwritten digits and the ground states of the 1D transverse field Ising model. To implement the unitary perturbation $\hat{\epsilon}$, we add an additional variational layer before the original quantum classifiers. After training, we fix the variational parameters of the given classifier $\mathcal{C}$ and optimize the parameters of the perturbation layer through the qBIM algorithm to maximize the loss function for a given set of n original samples.

The major results are shown in Fig. 3. In Fig. 3(a), we display two adversarial examples for digits 1 and 9, which are obtained by adding the same unitary perturbation to the original images and can fool the classifier 2 (one of the QCNN classifiers mentioned above). We mention that the fidelity between the original and crafted images is relatively small (about 78%) compared with the examples given in Fig. 2(a), but the crafted images remain easily identifiable by human eyes. In Fig. 3(b), we consider adding the same unitary perturbation to all the test samples of the ground states of H_{Ising} in a white-box attack setting for classifier 2. From this figure, it is clear that the accuracy drops rapidly at first as we increase the perturbation strength, and then maintains at a fixed finite value (about 0.5). This is consistent with the Ineq.(3) that R_E has an upper bound around $\mu(\mathcal{E})$. We mention that the loss keeps increasing as the perturbation strength increases, even in the region where the accuracy becomes flattened. This counterintuitive behavior is due to the fact that the loss function (cross-entropy) is continuous, whereas the accuracy is defined by the ratio of correctly classified samples whose labels are assigned according to the largest output probability. Fig. 3(c) shows similar results as in Fig. 3(b), but for a different quantum classifier (i.e., the classifier 10 mentioned above).

We remark that in our numerical simulations the Hilbert dimension involved is not very large due to limited classical computational resources. Consequently, a larger perturbation is needed to create the adversarial examples. As in Fig. 3(a), the perturbation is perceptible to human eyes. However, this is by no means a pitfall in principle and can be circumvented by simulating larger quantum classifiers. As noisy intermediate-scale quantum devices now become available in laboratories [7], this may also be resolved by running the protocol in real quantum devices. In addition, although we only focus on two-category classifications for simplicity in this paper, the extension to multi-category classifications and other adversarial scenarios is straightforward.

DISCUSSION AND CONCLUSION

This work only reveals the tip of the iceberg in the fledgling field of quantum adversarial machine learning. Many important questions remain unexplored and demand further investigations. First, this work shows that the existence of universal adversarial examples is a fundamental feature of quantum learning in high-dimensional space in general. However, for a given learning task, the legitimate samples may only occupy a tiny subspace of the whole Hilbert space. This brings about the possibility of defending against adversarial attacks. In practice, how to develop appropriate countermeasures feasible in experiments to strengthen the reliability of quantum classifiers still remains unclear. In addition, unsupervised and reinforcement learning approaches may also suffer from the vulnerability problem [49]. Yet, in practice it is often more challenging to obtain adversarial examples in these scenarios. The study of quantum adversarial learning in the unsupervised or reinforcement setting is still lacking. In particular, how to obtain adversarial examples and perturbations and study their universality properties for quantum unsupervised or reinforcement learning remains entirely unexplored and is well worth future investigations. Finally, it would be interesting and important to carry out an experiment to demonstrate the existence of universal adversarial examples and perturbations. This would be a crucial step toward practical applications of quantum technologies in artificial intelligence in the future, especially for these applications in safety and security-critical environments, such as self-driving cars, malware detection, biometric authentication, and medical diagnostics [50].

In summary, we have studied the universality of adversarial examples and perturbations for quantum classifiers. We proved two relevant theorems: one states that an $O(\frac{\ln k}{2^n})$ increase of the perturbation strength is already sufficient to ensure a moderate universal adversarial risk for a set of k quantum classifiers, and the other asserts that, for a general quantum classifier, the empirical error rate is bounded from both below and above and approaches to unit exponentially fast as the size of the classifier increases. We carried out extensive numerical simulations on concrete examples to demonstrate the existence of universal adversarial examples and perturbations for quantum classifiers in reality. Our results uncover a new aspect about the vulnerability of quantum machine learn-

ing systems, which would provide valuable guidance for practical applications of quantum classifiers based on both near-term and future quantum technologies.

SUPPLEMENTARY DATA

Supplementary data are available at NSR online.

ACKNOWLEDGEMENTS

We thank Sirui Lu, Weikang Li, Xun Gao, Si Jiang, Wenjie Jiang and Nana Liu for helpful discussions.

FUNDING

This work is supported by the start-up fund from Tsinghua University (Grant. No. 53330300320), the National Natural Science Foundation of China (Grant. No. 12075128), and the Shanghai Qi Zhi Institute.

AUTHOR CONTRIBUTIONS

D.-L. D. proposed and supervised the project. W.-Y. G. carried out the numerical simulations and all authors discussed the results and participated in writing the manuscript.

Conflict of interest statement. None declared.

* dldeng@tsinghua.edu.cn

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

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

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

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

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

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

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

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

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

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

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

[12] G. Carleo and M. Troyer, "Solving the quantum many-body problem with artificial neural networks," Science **355**, 602 (2017).

[13] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, "Neural-network quantum state tomography," Nat. Phys. , 1 (2018).

[14] Y.-H. Zhang, P.-L. Zheng, Y. Zhang, and D.-L. Deng, "Topological Quantum Compiling with Reinforcement Learning," Phys. Rev. Lett. **125**, 170501 (2020).

[15] V. L. Deringer, N. Bernstein, G. Csányi, C. B. Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, and S. R. Elliott, "Origins of structural and electronic transitions in disordered silicon," Nature **589**, 59 (2021).

[16] D.-L. Deng, "Machine learning detection of bell nonlocality in quantum many-body systems," Phys. Rev. Lett. **120**, 240402 (2018).

[17] Y. Zhang and E.-A. Kim, "Quantum Loop Topography for Machine Learning," Phys. Rev. Lett. **118**, 216401 (2017).

[18] J. Carrasquilla and R. G. Melko, "Machine learning phases of matter," Nat. Phys. **13**, 431 (2017).

[19] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, "Learning phase transitions by confusion," Nat. Phys. **13**, 435 (2017).

[20] L. Wang, "Discovering phase transitions with unsupervised learning," Phys. Rev. B **94**, 195105 (2016).

[21] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, "Machine learning phases of strongly correlated fermions," Phys. Rev. X **7**, 031038 (2017).

[22] Y. Zhang, A. Mesaros, K. Fujita, S. Edkins, M. Hamidian, K. Ch'ng, H. Eisaki, S. Uchida, J. S. Davis, E. Khatami, *et al.*, "Machine learning in electronic-quantum-matter imaging experiments," Nature **570**, 484 (2019).

[23] W. Lian, S.-T. Wang, S. Lu, Y. Huang, F. Wang, X. Yuan, W. Zhang, X. Ouyang, X. Wang, X. Huang, L. He, X. Chang, D.-L. Deng, and L. Duan, "Machine learning topological phases with a solid-state quantum simulator," Phys. Rev. Lett. **122**, 210503 (2019).

[24] M. S. Scheurer and R.-J. Slager, "Unsupervised machine learning and band topology," Phys. Rev. Lett. **124**, 226401 (2020).

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

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

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

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

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

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

[31] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A sur-

vey," arXiv:1810.00069 (2018).
[32] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognition **84**, 317 (2018).
[33] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Second International Conference on Learning Representations (ICLR, Banff, Canada, 2014) (2014).
[34] S. Lu, L.-M. Duan, and D.-L. Deng, "Quantum adversarial machine learning," Phys. Rev. Res. **2**, 033212 (2020).
[35] N. Liu and P. Wittek, "Vulnerability of quantum classification to adversarial perturbations," Phys. Rev. A **101**, 062331 (2019).
[36] Y. Du, M.-H. Hsieh, T. Liu, D. Tao, and N. Liu, "Quantum noise protects quantum classifiers against adversaries," arXiv:2003.09416 (2020).
[37] P. Casares and M. Martin-Delgado, "A quantum active learning algorithm for sampling against adversarial attacks," New J. Phys. **22**, 073026 (2020).
[38] J. Guan, W. Fang, and M. Ying, "Robustness verification of quantum machine learning," arXiv:2008.07230 (2020).
[39] H. Liao, I. Convy, W. J. Huggins, and K. B. Whaley, "Adversarial robustness of quantum machine learning models," arXiv:2010.08544 (2020).
[40] M. Ledoux, *The concentration of measure phenomenon*, 89 (American Mathematical Soc., 2001).
[41] K. Poland, K. Beer, and T. J. Osborne, "No free lunch for quantum machine learning," arXiv:2003.14103 (2020).
[42] K. Sharma, M. Cerezo, Z. Holmes, L. Cincio, A. Sornborger, and P. J. Coles, "Reformulation of the no-free-lunch theorem for entangled data sets," arXiv:2007.04900 (2020).
[43] J. G. Ratcliffe, S. Axler, and K. Ribet, *Foundations of hyperbolic manifolds*, Vol. 149 (Springer, 2006).
[44] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 acm sigsac conference on computer and communications security* (2016) pp. 1528–1540.
[45] I. Cong, S. Choi, and M. D. Lukin, "Quantum convolutional neural networks," Nat. Phys. **15**, 1273 (2019).
[46] *The mnist database of handwritten digits* (1998).
[47] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv:1607.02533 (2016).
[48] W. Hoeffding, "Probability inequalities for sums of bounded random variables," J. Am. Stat. Assoc. **58**, 13 (1963).
[49] Y. Vorobeychik and M. Kantarcioglu, "Adversarial machine learning," Synthesis Lectures on Artificial Intelligence and Machine Learning **12**, 1 (2018).
[50] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," Science **363**, 1287 (2019).