Towards efficient and generic entanglement detection by machine learning

Jue Xu* and Qi Zhao†
QICI Quantum Information and Computation Initiative, Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
(Dated: November 5, 2022)

Detection of entanglement is an indispensable step to practical quantum computation and communication. Compared with the conventional entanglement witness method based on fidelity, we propose a flexible, machine learning assisted entanglement detection protocol that is robust to different types of noises and also experimental-friendly. In this protocol, an entanglement classifier for a generic entangled state is obtained by training a classical machine learning model with a synthetic dataset. The dataset contains classical features of two types of states and their labels (either entangled or separable). The classical features of a state, that is expectation values of a set of k-local Pauli observables, are estimated sample-efficiently by the classical shadow method. In numerical simulation, our classifier can detect the entanglement of 4-qubit GHZ states with coherent noise and W states mixed with large white noise, with high accuracy.

I. INTRODUCTION

Entanglement [1] is the key ingredient of quantum teleportation [2], quantum cryptography [3], quantum computation [4], and quantum metrology [5]. However, decoherence and imperfections are inevitable in real-world devices, which means the interaction between a quantum system and classical environment would significantly affect entanglement quality and diminish quantum advantage in applications. For practical purpose, it is essential to detect entanglement in certain quantum physical system. This problem has been widely studied [6], but still far from being perfectly solved. From theoretical perspective, even given the full density matrix of a general state, it is computationally intractable to determine whether the state is entangled by classical [7] or quantum computation [8]. Not alone, the sample complexity to fully recover the density matrix of a state from experiments grows exponentially with dimension [9, 10]. So, a more realistic scenario is to determine whether a prepared state from experiments is entangled or not, with the prior knowledge of the state. This task for many entangled states of practical interest can be efficiently solved by measuring few observables called entanglement witness [11–13]. Though attempts such as [14, 15] have been made to enhance robustness to noise, there is no generic, efficient solution to detect entanglement beyond witness [16]. Moreover, it is generally challenging to reduce the measurement efforts (sample complexity) of entanglement witnesses for non-stabilizer states [17].

The goal of this paper is to find an efficient and generic way to detect entanglement of a target state. Machine learning (ML) is a powerful tool for such purpose. As we know, many ML techniques including both classical and quantum machine learning models have been proposed for classification tasks in physics, such as classification of phases and prediction of ground states [18–20]. Entanglement detection as a typical classification problem has been studies by ML techniques, such as determining separability by Neural Network (NN) [21, 22] and deriving generic entanglement witnesses by Support Vector Machine (SVM) [23, 24]. Nevertheless, these prior machine learning assisted methods only explore white noise robustness without considering other types of noises happened in experiments. And the sample efficiency of experimental implementation for these ML-derived classifiers have not been discussed.

In this work, a ML classifier is obtained by training SVM with a synthetic dataset on a classical computer. The dataset consists of two types of states, one is a set of certain target entangled states subject to randomly sampled noise and the other is a set of randomly sampled separable states. To be easily implemented in experiments, each state is characterized by its expectation values of Pauli observables, called classical features. The classical features $\mathbf{x}$ and the label $y \in \{-1, 1\}$ of a state consist a datapoint in a dataset $\{(\mathbf{x}, y)\}$. Within the framework of SVM, classification capability can be boosted by nonlinear kernel method and unimportant features can be eliminated programmatically. Furthermore, we restrict the Pauli observables to k-local such that classical features are estimated by the classical shadow method [25] with affordable sample complexity. In numeric simulation of 4-qubit GHZ state and W state, the kernel SVM classifier exhibits better robustness to white noise than conventional fidelity witnesses and also robust to coherent noise which is more realistic in experiments but not widely studied.

This paper is organized as follows: in Section II, we briefly present necessary definitions about multipartite entanglement, related entanglement detection problems, and mainstream methods for these problems; Section III demonstrates our end-to-end protocol including two parts: learning an entanglement witness for a generic state from synthetic data and efficient estimation of classical features of states from experiments; at last, numerical simulation results are discussed in Section IV.

* juexu@cs.umd.edu
† zhaoqi@cs.hku.hk

II. PRELIMINARIES

A. Multipartite entanglement

Large scale entanglement involving multiple particles maybe the main resource for quantum advantages in quantum computation and communication. Roughly, we say a quantum state ρ of n subsystems is *entangled* if it is not fully separable, i.e., the state cannot be written as the tensor product of all subsystems as $\rho = \rho_1 \otimes \cdots \otimes \rho_n$. Clearly, the simple statement 'the state is entangled' would allow that only two of the particles are entangled while the rest is in a product state, which is very weak entanglement. So, the more interesting entanglement property is bipartite separability:

Definition 1 (bi-separable)**.** A pure state $|\psi\rangle$ is bipartite separable (bi-separable) if and only if it can be written as a tensor product form $|\psi\rangle_{\text{bi}}^{\mathcal{P}} = |\phi_A\rangle \otimes |\phi_B\rangle$ with some bi-partition $\mathcal{P} = \{A, B \equiv \bar{A}\}$. A mixed state ρ is bi-separable if and only if it can be written as a convex combination of pure bi-separable states, i.e., $\rho_{\text{bi}} = \sum_i p_i \, |\psi_i\rangle\langle\psi_i|_{\text{bi}}^{\mathcal{P}_i}$ ($\mathcal{P}_i$ can be different partitions) with a probability distribution $\{p_i\}$. The set of all bi-separable states is denoted as $\mathcal{S}_{\text{bi}}$.

Definition 2 (GME)**.** On the contrary, if a state $\rho \notin \mathcal{S}_{\text{bi}}$, it possesses genuine multipartite entanglement (GME).

GME implies that all subsystems are indeed entangled with each other, so it is the strongest form of entanglement. Whereas, there is another restricted way for generalizing bi-separability to mixed states: if it is a mixing of pure bi-separable states with the same partition $\mathcal{P}_2$, and we denote the state set as $\mathcal{S}_{\text{bi}}^{\mathcal{P}_2}$. It is practically interesting to study entanglement under certain partition, because it naturally indicates the quantum information processing capabilities among a real geometric configuration. We have a formal definition for entanglement concerning partitions:

Definition 3 (full entanglement)**.** A state ρ possesses full entanglement if it is outside of the separable state set $\mathcal{S}_{\text{bi}}^{\mathcal{P}_2}$ for any partition, that is, $\forall \mathcal{P}_2 = \{A, \bar{A}\}, \rho \notin \mathcal{S}_{\text{bi}}^{\mathcal{P}_2}$.

For a state with full entanglement, it is possible to prepare it by mixing bi-separable states with different bipartitions, so full entanglement is weaker than GME but still useful in practice.

B. Entanglement detection

After introducing the definitions about entanglement, the next basic question is how to determine entanglement of a state efficiently. Despite clear definitions, it is a highly non-trivial question for a general state. For a general review on this subject, we refer readers to [6]. One of the most widely studied problems in this area is bi-separability.

Problem 1 (separability)**.** Given a density matrix [26] ρ, to determine if it is bi-separable (in $\mathcal{S}_{\text{bi}}$).

It is not hard to prove that if a state is bi-separable regarding $\mathcal{P} = \{A, B\}$, then it must have positive partial transpose (PPT), i.e., the partially transposed (PT) [27] density matrix $\rho_{AB}^{\intercal_A}$ is positive, semidefinite [28] [29, 30]. By contrapositive, we have a sufficient condition for (bipartite) entanglement, that is if the smallest eigenvalue of partial transpose $\rho_{AB}^{\intercal_A}$ is negative (NPT), then the state is entangled (cannot be bi-separable with $\mathcal{P} = \{A, B\}$). We should mention that PPT criterion is a necessary and sufficient condition for separability only when the system dimension is low ($d_A d_B \leq 6$ where d_A and d_B are the dimensions of two bipartite subsystems respectively) [30]. Therefore, no general solution for the separability problem is known. Then, a natural question is whether it is possible to solve separability approximately. By relaxing the defintion (promise a gap between two types of states), a reformulation of separability in the theoretic computer science language is

Problem 2 (Weak membership problem for separability)**.** Given a density matrix ρ with the promise that either (i) $\rho \in \mathcal{S}_{\text{bi}}$ or (ii) $\|\rho - \rho_{\text{bi}}\| \geq \epsilon$ with certain norm, decide which is the case.

Unfortunately, even we are given the complete information about a state and promised a gap (error tolerance ϵ), it is still hard to determine separability approximately by classical computation. Weak membership problem for separability is NP-Hard for $\epsilon = 1/\operatorname{poly}(d_A, d_B)$ with respect to Euclidean norm and trace norm [31] [7] [32], while there exists a quasipolynomial-time algorithm with respect to certain norm [33]. A notable numeric method is the powerful criteria called k-symmetric extension hierarchy based on SDP [34] [35] [36], which also becomes computationally intractable with growing k. The quantum complexity (hardness) of a series of related separability testing problems were studied in the framework of quantum interactive proofs [8]. Nevertheless, these hardness results do not rule out the possibility to solve it efficiently with a stronger promise (approximation) or by machine learning (heuristic) techniques powered by data.

1. *Entanglement witness based on fidelity*

A (realistic) variant of separability is how to determine bi-separable given copies of an unknown state (from experiments) rather than its full density matrix. In this case, the sample complexity should be considered besides computational complexity. Since the input to this problem is quantum data (states), directly estimating spectrum or entanglement monotone functions of the reduced density matrix $\rho_A := \operatorname{Tr}_B(\rho_{AB})$ [37] [38] [39], e.g., purity, negativity, and entanglement entropy, by quantum measurement and circuits [40] [41] is a good option (without fully recovering density matrices). However, this line of

work does not provide capability beyond theoretical complexity bounds (though usually efficient for one-side test). In this paper, the problem we study is another variant:

Problem 3 (entanglement detection with prior knowledge)**.** Given copies of an unknown state ρ (from experiments) that is promised either (i) $\rho \in \mathcal{S}_{\text{bi}}$ or (ii) in 'proximity' of a target $|\psi_{\text{tar}}\rangle$, determine which is the case.

The typical scenario for this problem is one aims to prepare a pure entangled state $|\psi_{\text{tar}}\rangle$ in experiments and would like to detect (verify) it as true multipartite entangled. While the preparation is not perfect, it is reasonable to assume that the prepared mixed state ρ_{pre} is in the proximity of the target state, that is, $|\psi_{\text{tar}}\rangle$ undergoes noise channels restricted to white noise, bit/phase-flip error, or random local unitary.

This problem is supposed to be solved more efficiently, because we have a much stronger promise than the separability problem. The usual method for it is constructing an observable W called entanglement witness such that

$$\text{Tr}(W\rho_{\text{bi}}) \geq 0 \text{ and } \text{Tr}(W|\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}|) < 0 \tag{1}$$

which means that the witness W has a positive expectation value on all separable states. Hence, a negative expectation value implys the presence of entanglement (GME). It can be proved, for every entangled state, a witness can always be constructed, but no entanglement witness works for all entangled states [42]. So, entanglement witness only provides one-side test for separability. For instance, the Bell (CHSH) inequalities originally proposed to rule out local hidden variable models, can be regarded as an entanglement witness for many 2-qubit entangled states [43]. A Bell inequality can be considered as a linear combination of Pauli observables $W_{\text{Bell}} := \mathbf{w}_{\text{Bell}} \cdot \mathbf{O}_{\text{Bell}}$ such that only entangled states ρ have $|\text{Tr}(\rho W_{\text{Bell}})|$ greater than a threshold [44].

While various methods for constructing an entanglement witness exist, the most common one is based on the fidelity between a prepared state ρ_{pre} to the target (pure entangled) state $|\psi_{\text{tar}}\rangle$

$$W_\psi = \alpha\mathbb{1} - |\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}| \tag{2}$$

where $\alpha = \max_{\rho_{\text{bi}}} \text{Tr}(\rho_{\text{bi}}|\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}|)$ is the maximal fidelity between separable states and the target entangled state such that for every separable state $\text{Tr}(\rho_{\text{bi}}W) \geq 0$. This kind of fidelity witness classifies states as either (1) the fidelity $\text{Tr}(\rho_{\text{pre}}|\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}|) < \alpha$; or (2) the fidelity $\text{Tr}(\rho_{\text{pre}}|\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}|) \geq \alpha$ implies $\rho \notin \mathcal{S}_{\text{bi}}$ [45]. For instance, assume the target state is $|\text{GHZ}\rangle := \frac{1}{\sqrt{2}}(|0\rangle^{\otimes n} + |1\rangle^{\otimes n})$, the maximal overlap between GHZ and bi-separable states is 1/2, such that the witness Eq. (2) with $\alpha = 1/2$ certifies tripartite entanglement [46]. We call Eq. (2) as projector-based fidelity witness [11]. In order to effectively measure a witness in an experiment, it is preferrable to decompose the projector term into a sum of locally measurable observables such as [47]. Meanwhile, for graph states (stabilizer states, i.e., a large class of entanglement states), a witness can be constructed by very few local measurement settings (LMS) [12] [13] [48] and implemented in experiments [49] [50], but non-local measurements are usually required for non-stabilizer cases (e.g., W state) [17] [23].

III. END-TO-END ENTANGLEMENT DETECTION PROTOCOL

A. Motivation: Beyond fidelity witness

In most studies of fidelity witness, the robustness measure of a fidelity witness is its tolerence to white noise:

$$\rho = (1 - p_{\text{noise}})|\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}| + p_{\text{noise}}\frac{\mathbb{1}}{2^n} \tag{3}$$

where the limit of white noise (i.e., maximal p_{noise} s.t. $\text{Tr}(\rho W_\psi) < 0$) indicates the robustness of the witness. In general, there are entangled states mixed with large white noise that cannot be detected by conventional methods. For example, the maximally-entangled Bell state can maximally violate the CHSH inequality, but Bell states mixed with white noise don't violate the CHSH inequality when $1 - 1/\sqrt{2} < p_{\text{noise}} < 2/3$ despite they are still entangled in this regime. For 3-qubit GHZ states mixed with white noise, we can analytically compute the white noise threshold for NPT (implies bipartite entanglement): when $p_{\text{noise}} < 0.8$, the states cannot be bi-separable with respect to any partition (that is full entanglement). However, the conventional fidelity witness only detects GME when $p_{\text{noise}} < 4/7$ for GHZ states [6]. So, it would be practically interesting to have a witness for this white noise regime $p_{\text{noise}} \in [4/7, 0.8)$ [51] that beyond the capability of conventional fidelity witnesses.

Other than white noise, more realistic noise happened in (photonic) experiments is coherent noise, e.g., local rotations. Take n-qubit GHZ state as an example, unconscious phase accumulation and rotation on the first control qubit can be modeled as [15]

$$|\text{GHZ}(\phi,\theta)\rangle = \cos\theta\,|0\rangle^{\otimes n} + e^{\text{i}\phi}\sin\theta\,|1\rangle^{\otimes n}. \tag{4}$$

In certain noise regime (see Fig. 3 in [15]), $|\text{GHZ}(\phi,\theta)\rangle$ cannot be detected by conventional fidelity witness because coherent noises diminish the fidelity but not change entanglement property.

To formally characterize the cases beyond fidelity witness, Weilenmann et. al [16] [52] coined the term *unfaithful states* which systematically analyze 2-qudit entangled state mixed with white noise that cannot be detected by fidelity witness. They found that for $d \geq 3$ that almost all states in the Hilbert space are unfaithful. Subsequently, Güthe et. al [53] [54] gave a formal definition: a 2-qudit state ρ_{AB} is faithful if and only if there are local unitary transformations U_A and U_B such that $\langle\phi^+|U_A \otimes U_B\rho_{AB}U_A^\dagger \otimes U_B^\dagger|\phi^+\rangle > \frac{1}{d}$. Consequently, they

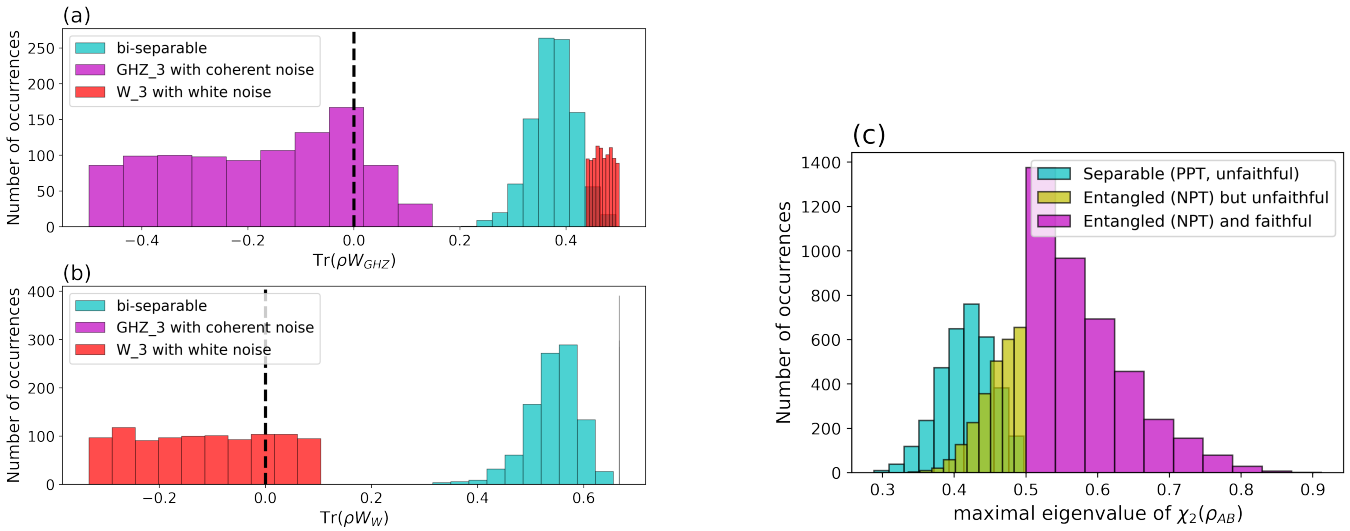

FIG. 1. Examples of the entanglement states cannot be detected by conventional fidelity witnesses. (a) GHZ states with coherent noise sampled with $\theta = \pi/3$ and $\phi \in [0.5\pi, 0.6\pi]$ cannot be detected by the GHZ projector fidelity witness W_{GHZ} (c.f. Eq. (2)). Entangled states should be on the left of the dashed vertical line, i.e., have negative expectation value of the witnesses $\text{Tr}(\rho W)$. (b) Similarly, W states with large white noise $p_{\text{noise}} \in [8/21, 0.5]$ cannot be detected by W_{w}. And we can see W states with white noise has $\text{Tr}(\rho_{\text{W}} W_{\text{GHZ}}) > 0$, vice versa. (c) Unfaithfulness of 2-qubit states: 10^4 randomly sampled 2-qubit states are categorized according to the minimal eigenvalue of partial transpose $\rho_{AB}^{\intercal_A}$ and the maximal eigenvalue of $\chi_2(\rho_{AB})$.

found a necessary and sufficient condition for 2-qubit unfaithfulness: a 2-qubit state ρ_{AB} is faithful if and only if the maximal eigenvalue of

$$\mathcal{X}_2(\rho_{AB}) = \rho_{AB} - \frac{1}{2}(\rho_A \otimes I + I \otimes \rho_B) + \frac{1}{2} I \otimes I \quad (5)$$

is larger than 1/2. We can see in (c) of Fig. 1, even for 2-qubit states, nonnegligible portion of randomly sampled states are unfaithful but still entangled (NPT).

Although there are variants of witness, such as nonlinear witness [14] and post-processing [55], designed to remedy the shortcomings of conventional fidelity witness respectively, it would be meaningful in practice to find a generic method to construct witnesses (classifiers) for entanglement detection with prior knowledge. Machine learning techniques suit the needs well because supervised learning can be regarded as a powerful nonlinear post-processing tool.

B. Training a generic witness via kernel SVM

One basic task in classical machine learning (ML) is binary classification, such as cat/dog images classification. In this case, the input to a ML algorithm is a (training) dataset $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{m}$ consists of m data points, where each data point is a pair of feature vector $\mathbf{x} \in \mathbb{R}^d$ of d features and its label $y \in \{-1, 1\}$. For example, the feature $\mathbf{x}$ of an image is a flatten vector of all pixel values and the label $y = -1$ for CAT images (1 for DOG). It is clear that separability or entanglement detection with prior knowledge problem are exactly such binary classification problems where each quantum state has a binary label, such as either 'ENTANGLED' or 'SEPARABLE'. The features $\mathbf{x}$ of a quantum state ρ can be the entries of its density matrix, or more realistically, the expectation values of selected observables.

With the surge of research on machine learning, ML algorithms have been proposed for classification tasks related to entanglement. Lu et. al [21] trained a (universal) separability classifier by classical neural network where features of $\mathbf{x}$ are the entries of density matrices. For the similar purpose, Ma and Yung [22] generalized Bell inequalities to a Bell-like ansatz $W_{\text{ml}} := \mathbf{w}_{\text{ml}} \cdot \mathbf{O}_{\text{Bell}}$ where the optimal weights $\mathbf{w}_{\text{ml}}$ are obtained via optimizing a neural network. And they found the tomographic ansatz

$$\text{Tr}(\rho W_{\text{ml}}) \equiv \langle W_{\text{ml}} \rangle := \mathbf{w}_{\text{ml}} \cdot \langle \mathbf{O}_\sigma \rangle \,, \ \forall \sigma \in \{I, X, Y, Z\}^n \quad (6)$$

where the feature $\mathbf{x}_{\rho,\boldsymbol{\sigma}} := \langle \mathbf{O}_\sigma \rangle$ is the vector of expectations of all 4^n Pauli observables [56], not only has better performance than the Bell-like ansatz, also required [57] for training a universal separability classifier. It is worth noting that training such a universal classifier for high-dimensional systems is a difficult optimization problem if the gap between two state sets is small (weak promise).

In our paper, we focus on the entanglement detection with prior knowledge problem with training data. In other words, we derive the entanglement witness (classifier) for certain target states with desired entannglement structure by fitting a synthetic dataset.

Problem 4 (learning an entanglement witness)**.**

- **Input**: a dataset $\{(\rho^{(i)}, y^{(i)})\}$ consist of randomly

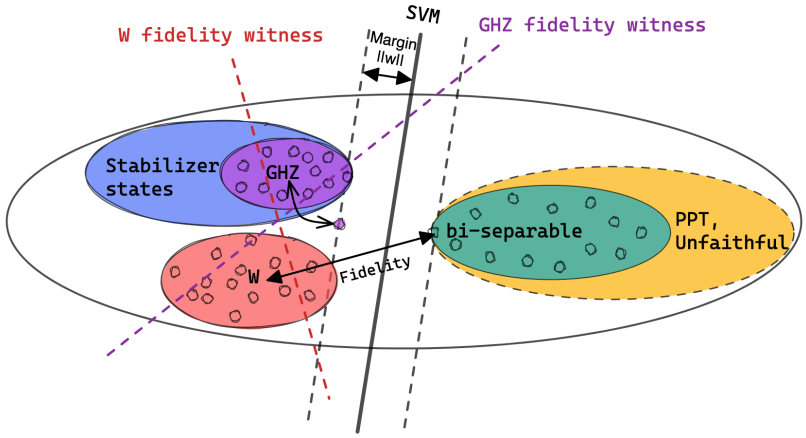

FIG. 2. Schematic diagram for different entanglement detection methods: the colored ellipses without black boundary indicate the vicinity (white noise) of certain entangled state such as GHZ, W states. Conventional fidelity witnesses for different states are depicted by colored dash lines (hyperplanes in feature space). Entangled state with large white noise or coherent noise (local rotation depicted by a curve) cannot be detected by conventional fideilty witnesses. SVM without kernel is a hyperplane separating two sets of colored dots (synthetic dataset). The data points on the boundaries (dashed black lines) are called support vectors. The distance between the SVM hyperplane and boundary is the margin to be minimized via optimization. The PPT criterion is a nonlinear but one-side classifier without prior knowledge.

sampled entangled states ρ around $|\psi_{\text{tar}}\rangle$ with label $y=-1$ and randomly sampled separable states with label 1.

- **Output**: a learned classifier $f(\mathbf{x}_{\rho,\tilde{\boldsymbol{\sigma}}})$ with high training accuracy where $\tilde{\boldsymbol{\sigma}}$ is a subset of all Pauli observables and $\mathbf{x}_{\rho,\tilde{\boldsymbol{\sigma}}}$ is a vector of corresponding expectation values.

The learning an entanglement witness problem has also been studied by classical ML [23] [24], but by a technique different from Neural Network (NN), called Support Vector Machine (SVM) [58]. A classification task performed by SVM can be formulated as a convex optimization problem: find a hyperplane parametrized by $(\mathbf{w}, b)$ in a feature space (a linear function f) that maximizes the margin between two decision boundaries subject to the constraint that two types of datapoints are separated (on the two sides of the hyperplane, see Fig. 2)

$$\max_{\mathbf{w}} \|\mathbf{w}\|_2^2 \text{ s.t. } \forall i,\ y^{(i)}\cdot(\mathbf{w}\cdot\mathbf{x}^{(i)}+b)\geq 1. \tag{7}$$

where $\mathbf{w}$ is the (not necessarily normalized) normal vector to the hyperplane and b is a bias term similar to α in Eq. (2). Therefore, the predicted label is given by the sign of the inner product (projection) of the hyperplane and the feature vector $\mathbf{x}$, i.e., $y=f(\mathbf{x})=\text{sign}(\mathbf{w}\cdot\mathbf{x}+b)$ (c.f. Eq. (1) and Eq. (6)). Geometrically, both SVM witness and conventional fidelity witness are hyperplanes in feature spaces, but the SVM witness is more flexible because the classifier $(\mathbf{w}, b)$ can be numerically derived through optimization for any generic target state. And it only requires local Pauli observables (measurements) O_σ that is feasible in most experiments, even when the target state is a non-stabilizer state.

The SVM formalism allows for the programmatic elimination of features [59], i.e., reducing the cost of experimental measurements (samples). We start with the feature vector of all k-local Pauli observables, then we randomly eliminate one feature such that the training accuracy remains high enough with the new feature vector $\tilde{\mathbf{x}}$. By repeating this procedure, we obtain a classifier $f(\mathbf{x}_{\rho,\tilde{\boldsymbol{\sigma}}})$, where $|\tilde{\boldsymbol{\sigma}}|=M$ is the minimal number of Pauli observables required for classification. The algorithm summarized in Algorithm. III.1.

Algorithm III.1: train a witness via kernel SVM

input : dataset $\left\{\left(\rho^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}$, minimal number of features: M, and tolerance ϵ
output: a classifier $f(\mathbf{x}_{\rho,\tilde{\boldsymbol{\sigma}}})$

$\mathbf{x}^{(i)} := \text{Tr}\left(\rho^{(i)} O_{\boldsymbol{\sigma}}\right), \forall i$ // evaluate all k-local Pauli observables and shuffle
while accuracy $< \epsilon$ or $len(\mathbf{x}) > M$ **do**
  for j in range$(len(\mathbf{x}))$ **do**
    /* eliminate j-th feature */
    $\forall i$, let $\tilde{\mathbf{x}}^{(i)}$ be $\mathbf{x}^{(i)}$ without the j-th feature
    /* Train SVM with the new feature vectors */
    accuracy, classifier $=$ SVM$(\left\{\left(\tilde{\mathbf{x}}^{(i)}, y^{(i)}\right)\right\}_{i}^{m})$
    if accuracy $\geq \epsilon$ **then**
      $\mathbf{x}^{(i)} := \tilde{\mathbf{x}}^{(i)}$ and then **break**
    else if accuracy $< \epsilon$ and $j = len(\mathbf{x})$ **then**
      /* If cannot find a classifier with less features, then output the last classifier with high accuracy */
      return a classifier $f(\mathbf{x}_{\rho,\tilde{\boldsymbol{\sigma}}})$
return a classifier $f(\mathbf{x}_{\rho,\tilde{\boldsymbol{\sigma}}})$ with $|\tilde{\boldsymbol{\sigma}}| = M$

A key drawback of conventional witnesses is its lin-

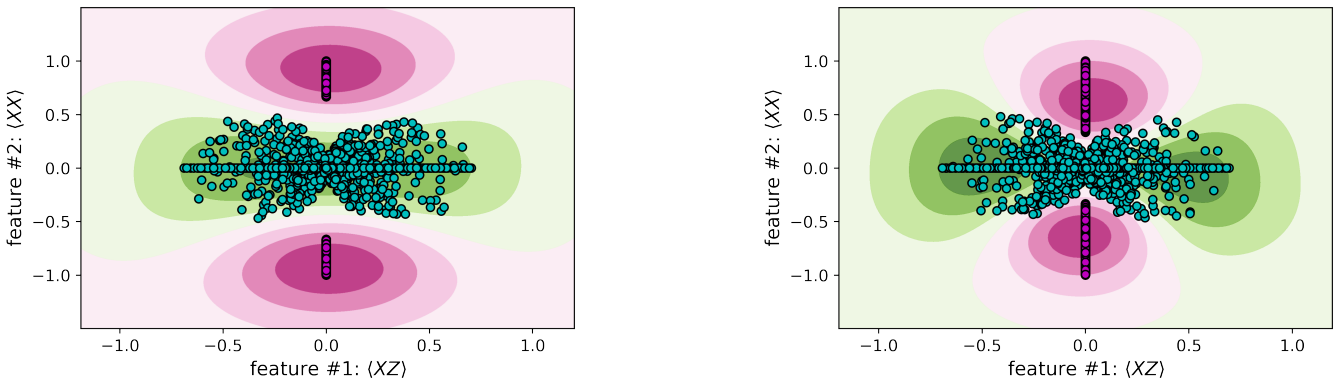

FIG. 3. The two-dimensional embedding (a low-dimensinoal feature space $\langle XZ\rangle$ VS $\langle XX\rangle$) of 2-qubit states: green dots represent randomly sampled separable states, while pink ones represent entangled Bell states mixed with white noise in the range (left figure) $p_{\text{noise}} \in [0, 1/3]$ and (right figure) $p_{\text{noise}} \in [0, 2/3]$. The colored shade indicates the nonlinear decision boundary of the RBF kernel SVM classifier. When the white noise is larger, the gap between two sets of data points is smaller such that training a classifier becomes harder.

earity because many real-world datasets are not linearly-separable in a low-dimensional feature space. Despite the nonlinear witness [14] proposed, its experimental implementation is more challenging than linear ones. The good news is, within the framework of SVM, nonlinearity can be easily achieved by the so-called kernel method [60]. The main idea is mapping the features $\mathbf{x}$ to a higher dimensional space via a feature map $\phi(\mathbf{x})$ such that they can be linearly separated in the high-dimensional feature space. The kernel function $k(\mathbf{x}, \mathbf{x}') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ measures the similarity between two input data points in a high-dimensional feature space because a kernel can be written as an inner product $\langle \phi(\mathbf{x}), \phi(\mathbf{x}')\rangle$. The commonly used kernel is the radial basis function (RBF) kernel which a Gaussian function $k_{\text{rbf}}(\mathbf{x}, \mathbf{x}') := \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|_2^2\right)$ with l_2 Euclidean norm and a parameter γ. Since the RBF kernel SVM is convex, the optimal classifier function will be found if it exists for the input dataset. It can be easily observed in Fig. 3 that two kinds of data points are clearly classified by a nonlinear (RBF kernel) SVM classifier, though it is not linearly separable in this 2-dimensional space.

Witnesses	# observables	weights	comment
Conventional fidelity	few LMS	fixed	one-side
SVM (kernel)	$\ll 4^n - 1$	trained	flexible
Tomographic (NN)	$4^n - 1$	trained	universal

TABLE I. Comparison of conventional fidelity witness, tomographic classifier, and SVM witness.

We compare the characters of different kinds of witnesses in Table I. The conventional fidelity witness only need few local measurement settings for stabilizer states, but it is a one-side test. The tomographic witness trained by NN only need the promise that there is a gap between entangled and separable states (almost universal), but it requires complete information of a state ($4^n - 1$ features). Between these two cases, the SVM witness has stronger classification capability than conventional fidelity witnesses and do not need as many classical features as the tomgraphic witness. However, these prior ML witnesses only consider the robustness to white noise and cannot be directly applied to experiments. In numerical simulation, we can efficiently evaluate classical features by direct calculation, but in actual experiments, entries of a density matrix are not explicitly known. Instead, we need to estimate observables (classical features) by repeat measurements, which we are going to discuss in next section.

C. Sample-efficient expectation estimation methods

The brute force approach to fully characterize a state in an experiment is quantum state tomography [61] [62]. With a recovered densitry matrix, we can directly calculate classcial features or separability measures, but full tomography is experimentally demanding. Even adaptive or collective measurements (and post-processing) allowed [63], rigorous analysis [9] [10] showed that $\Omega(D^2/\epsilon^2)$ measurements (copies) are required for recovering a $D \times D$ density matrix with error tolerence ϵ measured by trace distance. Now that full tomography is intractable for large systems, a workaround is to extract (partial) information about a state without fully recovering it:

Problem 5 (shadow tomography)**.** Given m copies (samples) of an unknown D-dimensional state and M known 2-outcome measurements $\{E_1, \ldots, E_M\}$, to estimate $\forall i, \text{Tr}(\rho E_i)$ within additive error ϵ with success probability at least $1 - \delta$.

Since shadow tomography can be implemented with

$\tilde{\mathcal{O}}(\log^4 M \cdot \log D \cdot \log 1/\delta \cdot \epsilon^{-4})$ copies [64] [65], we can estimate M classical features (Pauli observables) for $f(\mathbf{x}_{\rho,\tilde{\boldsymbol{\sigma}}})$ in a samples-efficient manner. However, Aaronson's shadow tomography procedure is very demanding in terms of quantum hardware (in the collective preparation and measurement on $\rho^{\otimes m}$). To be more feasible for current experiments, Huang et. al [25] introduced classical shadow (CS) scheme which we apply in our protocol.

Algorithm III.2: estimate Pauli observables by randomized classical feature

input : R copies of ρ and selected observables $O_{\tilde{\boldsymbol{\sigma}}}$
output: estimation of $\mathbf{x}_{\rho,\tilde{\boldsymbol{\sigma}}} := \mathrm{Tr}(\rho O_{\tilde{\boldsymbol{\sigma}}})$

Sample R Pauli measurements $P \in \{X, Y, Z\}^{\otimes n}$
for $i = 1, 2, \ldots, R$ **do**
  // apply single-copy measurement P to a copy ρ
  $\rho \mapsto U_P \rho U_P^\dagger \mapsto |\mathbf{b}\rangle$ with $b_j \in \{0, 1\}, \forall j \in [n]$
  // inverse channel $\mathcal{M}^{-1}(\rho') = (3\rho' - I)$
  $\rho_{\mathrm{cs}}^{(i)} = \bigotimes_j \left(3U_j^\dagger |b_j\rangle\langle b_j| U_j - I\right)$
$\mathrm{CS}(\rho, R) = \left\{\rho_{\mathrm{cs}}^{(1)}, \ldots, \rho_{\mathrm{cs}}^{(R)}\right\}$ // classical shadow
// estimate features for SVM from classical shadow
return $\mathbf{x}_{\rho,\tilde{\boldsymbol{\sigma}}} = \textsc{Expectation}(\mathrm{CS}(\rho, R) O_{\tilde{\boldsymbol{\sigma}}})$

The classical shadow of a state ρ (a set of snapshots ρ_{cs}) is a succinct classical description of a state ρ, which can be used to estimate the expectations of a sef of observables with a reasonably small number of copies of the state. To construct the randomized classical shadow, we firstly need to uniformly sample R Pauli measurements $P \in \{X, Y, Z\}^{\otimes n}$ (assume the state ρ of n qubits). Then, we apply single-copy measurement P to a copy of ρ, i.e., each measurement measures all qubits in Pauli X, Y, or Z-basis according to P. Specifically, we apply the transformation $\rho \mapsto U\rho U^\dagger$ where $U^\dagger P U = \Sigma$ is the eigendecomposition of P and then measure this rotated stated in computational basis (collapse to $|\mathbf{b}\rangle \in \{|0\rangle, |1\rangle\}^{\otimes n}$). The expectation values of a set of observables can be approximated A snapshot ρ_{cs} can be constructed by taking the inverse of the quantum channel $\mathcal{M}^{-1}(U^\dagger |\mathbf{b}\rangle\langle\mathbf{b}| U)$. By repeating this procedure R times, we obtain R snapshots of ρ which can be used to estimate the expectation values of a Pauli observable by an empirical average over R classical shadows, i.e., $o = \mathrm{Tr}(O\rho_{\mathrm{cs}})$ obeys $\mathbb{E}[o] = \mathrm{Tr}(O\rho)$. The algorithm is summarized in Algorithm. III.2.

The size of the classical shadow scales $\mathcal{O}(\log(M)3^k/\epsilon^2)$ to approximate M k-local Pauli observables with error tolerance ϵ [25], so this scheme has advantage the small k and large M (many quite local observables). There are several variants of classical shadow method [66–68]. The derandomized version [67] is the refinement of the original randomized protocol which provides better performance for k-local observables. The core idea of derandomized version is to select more Pauli measurements compatiable with k-local Pauli observables to be estimated. The main result of this work is a procedure for identifying "good" Pauli measurements that allow for accurately predicting many (fixed) Pauli expectation values. This procedure is designed to interpolate between two extremes: (i) completely randomized measurements (good for predicting many local observables) and (ii) completely deterministic measurements that directly measure observables sequentially (good for predicting few global observables). From the perspective of entanglement witness, the classical shadow method finds a effective local measurement settings for a sets of k-local Pauli observables. The entanglement detection by estimating p_3-PPT with classical shadow [69] and comparison of classical shadow methods [70] have been done experimentally.

IV. NUMERICAL SIMULATION AND DISCUSSION

We generate quantum state samples, construct quantum circuits, and manipulate quantum objects numerically by QuTiP Python library [71, 72]. We generate multi-partite entangled states (synthetic data) including: Bell states, 3-qubit GHZ with coherent noise Eq. (4) and W states with white noise Eq. (3), see Fig. 4 for the samples used in Fig. 1. The the noise parameters are randomly (uniformly) sampled from a range. In contrast to entangled states, we generate random separable states for different number of qubits by tensoring random (sampled by Haar measure, calling rand_dm(N,dims) function from QuTiP) density matrices of subsystems. For example, there are three different partitions $\rho_A \otimes \rho_{BC}$, $\rho_{AB} \otimes \rho_C$, and $\rho_B \otimes \rho_{AC}$ for 3-qubit separable states. It is not necessary to prepare the mixed separable states as convex combination of separable states with different partitions because SVM can correctly classifiy a mixture if it can classify each case.

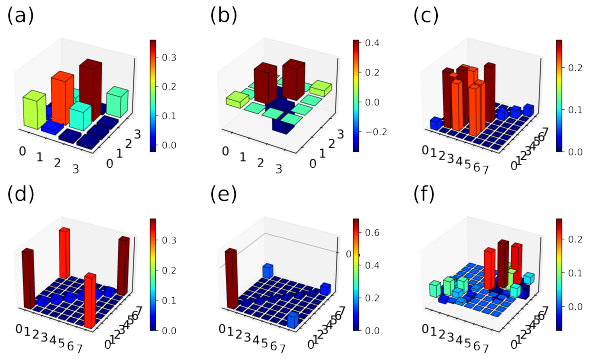

FIG. 4. The real part of sampled density matrices: (a) a random 2-qubit state; (b) Bell state (singlet) with white noise; (c) 3-qubit W state with white noise; (d) 3-qubit GHZ state with white noise; (e) 3-qubit GHZ state with coherent noise; (f) a random 3-qubit bi-separable state. The 4-qubit states used in the numeric simulation Fig. 5 are the natural extensions of (c), (d), (e), (f).

For the machine learning part, we make use of scikit-learning Python package [73] to train SVM with RBF kernel. For training a 4-qubit SVM classifier with accu-

racy 0.999, we generate 10^4 states for each kind of states. bi-separable states including $\rho_1 \otimes \rho_{234}$ and $\rho_{12} \otimes \rho_{34}$..... Fig. 1 and Fig. 5 show that conventional fidelity witnesses cannot correctly classify when GHZ states with coherent noises $\theta = \pi/3, \phi > \pi/2$ (even mixed with white noise $p_{\text{noise}} \in [0, 0.1]$) and W states mixed with white noise $p_{\text{noise}} > 8/21$, while the SVM classifier can classify them with hight accuracy.

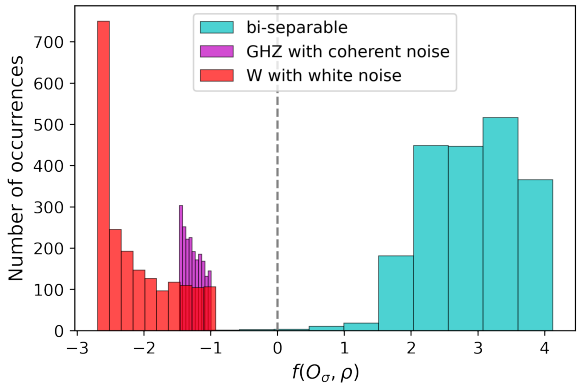

FIG. 5. The states beyond detection by fidelity witnesses (GHZ state with coherence noise $\theta \in [0, \pi/3], \phi \in [0, 0.6\pi]$, and W state with large white noise $p_{\text{noise}} \in [0, 0.5]$) can be classified by the kernel SVM classifier with high accuracy.

One set of features found by the kernel SVM are $\mathbf{x} = \text{Perm}(\langle XIIX\rangle, \langle YIIZ\rangle, \langle IIZZ\rangle, \langle ZXII\rangle)$. permutations $\text{Perm} := [(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)]$ and by symmetry of qubits of GHZ and W states.... The error of estimating these features VS the size of shadow is shown in Fig. 6. [74] (tradeoff between robustness and meaurement efficiency) The derandomized version performs better than randomized shadow and independent measurement. The comparable size classical shadow has been implemented in photonic experiments [70].

In conclusion, our protocol is flexible and sample-efficient in detecting entanglement around certain entangled state by training a kernel SVM classifier with a synthetic dataset. We test our protocol for 4-qubit GHZ states with coherent noise and W states with large white noise. We also show that the features for training such machine learned classifier can be estimated by a sample-efficient scheme....

There are also several potential directions for future research: (1) It is of practical interest to find rigorous proof or perform numerical simulation for the dataset size and number of features (required for high training accuracy) scaling with the system size (more than 4 qubits); (2) It is meaningful to test more kernels, such as graph kernel [75], shadow kernel [20], and neural tangent kernel [76], for better performance of the kernel SVM. And quantum kernel methods [77–79] also provide advantages over classical counterparts. (3) The task of estimating expectation values can also be achieved efficiently by both classical [80–82] and quantum machine learning [20, 83].

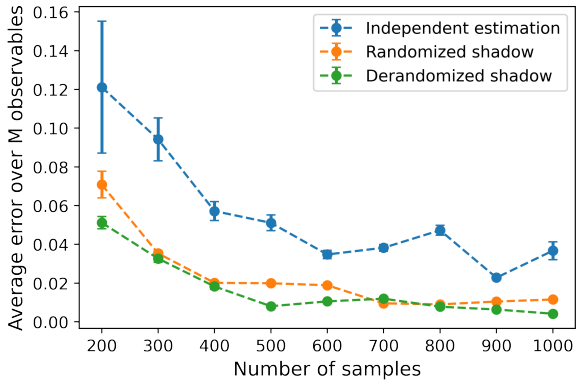

FIG. 6. Average error of estimating expectation values of 30 2-local 4-qubit Pauli observables $\sum_i (o_i - \text{Tr}(O_i\rho))^2/M$ by different estimation methods VS number of samples (the error bar indicates the standard deviation for each case). The blue line represents estimating each Pauli observables independently. The orange line is the estimation by the randomized classical shadow, while the green one is the derandomized version of classical shadow.

Huang et. al rigorously showed that, for achieving accurate prediction on all $4^n - 1$ Pauli observables the exponential quantum advantage over classical ML is possible [84]. Training a more powerful (universal) classifier with all Pauli observables as features might be interesting.

ACKNOWLEDGMENTS

[1] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, Rev. Mod. Phys. **81**, 865 (2009), arXiv:quant-ph/0702225.

[2] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, Phys. Rev. Lett. **70**, 1895 (1993).

[3] A. K. Ekert, Phys. Rev. Lett. **67**, 661 (1991).

[4] H. J. Briegel, D. E. Browne, W. Dür, R. Raussendorf, and M. V. den Nest, Nature Phys **5**, 19 (2009), arXiv:0910.1116.

[5] V. Giovannetti, S. Lloyd, and L. Maccone, Science **306**, 1330 (2004), arXiv:quant-ph/0412078.

[6] O. Gühne and G. Toth, Physics Reports **474**, 1 (2009), arXiv:0811.2803 [cond-mat, physics:physics, physics:quant-ph].

[7] L. Gurvits, Classical deterministic complexity of Edmonds' problem and Quantum Entanglement (2003), arXiv:quant-ph/0303055.

[8] G. Gutoski, P. Hayden, K. Milner, and M. M. Wilde, Theory of Comput. **11**, 59 (2015), arXiv:1308.5788

[quant-ph].
[9] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu, IEEE Trans. Inform. Theory , 1 (2017).
[10] R. O'Donnell and J. Wright, in *Proc. Forty-Eighth Annu. ACM Symp. Theory Comput.* (ACM, Cambridge MA USA, 2016) pp. 899–912.
[11] M. Bourennane, M. Eibl, C. Kurtsiefer, S. Gaertner, H. Weinfurter, O. Guehne, P. Hyllus, D. Bruss, M. Lewenstein, and A. Sanpera, Phys. Rev. Lett. **92**, 087902 (2004), arXiv:quant-ph/0309043.
[12] G. Toth and O. Guehne, Phys. Rev. Lett. **94**, 060501 (2005), arXiv:quant-ph/0405165.
[13] G. Tóth and O. Gühne, Phys. Rev. A **72**, 022340 (2005).
[14] O. Gühne and N. Lütkenhaus, Phys. Rev. Lett. **96**, 170502 (2006).
[15] Y. Zhou, Phys. Rev. A **101**, 012301 (2020), arXiv:1907.11495 [quant-ph].
[16] M. Weilenmann, B. Dive, D. Trillo, E. A. Aguilar, and M. Navascués, Phys. Rev. Lett. **124**, 200502 (2020), arXiv:1912.10056 [quant-ph].
[17] Y. Zhang, Y. Tang, Y. Zhou, and X. Ma, Phys. Rev. A **103**, 052426 (2021), arXiv:2012.07606 [quant-ph].
[18] J. Carrasquilla and R. G. Melko, Nature Phys **13**, 431 (2017), arXiv:1605.01735.
[19] I. Cong, S. Choi, and M. D. Lukin, Nat. Phys. **15**, 1273 (2019), arXiv:1810.03787 [cond-mat, physics:quant-ph].
[20] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill, Science **377**, eabk3333 (2022), arXiv:2106.12627.
[21] S. Lu, S. Huang, K. Li, J. Li, J. Chen, D. Lu, Z. Ji, Y. Shen, D. Zhou, and B. Zeng, Phys. Rev. A **98**, 012315 (2018), arXiv:1705.01523 [quant-ph].
[22] Y.-C. Ma and M.-H. Yung, npj Quantum Inf **4**, 34 (2018), arXiv:1705.00813 [quant-ph].
[23] E. Y. Zhu, L. T. H. Wu, O. Levi, and L. Qian, Machine Learning-Derived Entanglement Witnesses (2021), arXiv:2107.02301 [quant-ph].
[24] S. V. Vintskevich, N. Bao, A. Nomerotski, P. Stankus, and D. A. Grigoriev, Classification of four-qubit entangled states via Machine Learning (2022), arXiv:2205.11512 [quant-ph].
[25] H.-Y. Huang, R. Kueng, and J. Preskill, Nat. Phys. **16**, 1050 (2020), arXiv:2002.08953 [quant-ph].
[26] A quantum (mixed) state ρ can be represented by a density matrix which is a Hermitian, positive semidefinite operator (matrix) of trace one. If the rank of ρ is 1, then the state is a pure state.
[27] The partial transpose (PT) operation acting on subsystem A is defined as $|k_A, k_B\rangle\langle l_A, l_B|^{\intercal_A} := |l_A, k_B\rangle\langle k_A, l_B|$ where $\{|k_A, k_B\rangle\}$ is a product basis of the joint system $\mathcal{H}_{AB}$.
[28] A matrix (operator) is positive, semidefinite (PSD) if all its eigenvalues are non-negative.
[29] A. Peres, Phys. Rev. Lett. **77**, 1413 (1996), arXiv:quant-ph/9604005.
[30] M. Horodecki, P. Horodecki, and R. Horodecki, Physics Letters A **223**, 1 (1996), arXiv:quant-ph/9605038.
[31] The Euclidean norm of a matrix A is defined as $\|A\|_2 := \sqrt{\mathrm{Tr}(A^\dagger A)}$. The trace norm of A is defined as $\|A\|_{\mathrm{Tr}} \equiv \|A\|_1 := \mathrm{Tr}(|A|) \equiv \mathrm{Tr}\left(\sqrt{A^\dagger A}\right)$. Correspondingly, trace distance between two density matrices is $d_{\mathrm{tr}}(\rho, \rho') := \frac{1}{2}\|\rho - \rho'\|_1$.
[32] S. Gharibian, Strong NP-Hardness of the Quantum Separability Problem (2009), arXiv:0810.4507 [quant-ph].
[33] F. G. Brandão, M. Christandl, and J. Yard, in *Proc. 43rd Annu. ACM Symp. Theory Comput. - STOC 11* (ACM Press, San Jose, California, USA, 2011) p. 343, arXiv:1011.2751 [quant-ph].
[34] A. C. Doherty, P. A. Parrilo, and F. M. Spedalieri, Phys. Rev. A **69**, 022308 (2004), arXiv:quant-ph/0308032.
[35] L. M. Ioannou, Quantum Inf. Comput. **7**, 335 (2007), arXiv:quant-ph/0603199.
[36] M. Navascues, M. Owari, and M. B. Plenio, Phys. Rev. A **80**, 052306 (2009), arXiv:0906.2731 [quant-ph].
[37] A. K. Ekert, C. M. Alves, D. K. L. Oi, M. Horodecki, P. Horodecki, and L. C. Kwek, Phys. Rev. Lett. **88**, 217901 (2002), arXiv:quant-ph/0203016.
[38] P. Horodecki and A. Ekert, Phys. Rev. Lett. **89**, 127902 (2002), arXiv:quant-ph/0111064.
[39] S. Johri, D. S. Steiger, and M. Troyer, Phys. Rev. B **96**, 195136 (2017), arXiv:1707.07658.
[40] Y. Wang, Y. Li, Z.-q. Yin, and B. Zeng, npj Quantum Inf **4**, 46 (2018), arXiv:1801.03782.
[41] Y. Quek, M. M. Wilde, and E. Kaur, Multivariate trace estimation in constant quantum depth (2022), arXiv:2206.15405 [hep-th, physics:quant-ph].
[42] T. Heinosaari and M. Ziman, *The Mathematical Language of Quantum Theory: From Uncertainty to Entanglement*, 1st ed. (Cambridge University Press, 2011).
[43] B. M. Terhal, Physics Letters A **271**, 319 (2000), arXiv:quant-ph/9911057.
[44] The Bell (CHSH) inequality (witness): $\mathbf{O}_{\mathrm{CHSH}} = (\mathbb{1}, ab, ab', a'b, a'b')$ with $a = Z, a' = X, b = (X - Z)/\sqrt{2}, b = (X + Z)/\sqrt{2}$ and $\mathbf{w}_{\mathrm{CHSH}} = (\pm 2, 1, -1, 1, 1)$.
[45] In other words, the trace distance $\|\rho_{\mathrm{pre}} - |\psi_{\mathrm{tar}}\rangle\langle\psi_{\mathrm{tar}}|\|_1 \leq \sqrt{1-\alpha}$ (c.f Weak membership problem for separability) because the fidelity and trace distance are related by the inequalities $1 - F \leq d_{\mathrm{tr}}(\rho, \rho') \leq \sqrt{1 - F^2}$.
[46] A. Acin, D. Bruss, M. Lewenstein, and A. Sanpera, Phys. Rev. Lett. **87**, 040401 (2001), arXiv:quant-ph/0103025.
[47] $W_{\mathrm{GHZ}_3} = \frac{1}{8}(3 * III - XXX - \mathrm{Perm}(IZZ) + \mathrm{Perm}(XYY))$ where $ZZI \equiv Z \otimes Z \otimes I$ and $\mathrm{Perm}(IZZ) \equiv ZZI + ZIZ + IZZ$ for readability.
[48] Y. Zhou, Q. Zhao, X. Yuan, and X. Ma, npj Quantum Inf **5**, 83 (2019).
[49] H. Lu, Q. Zhao, Z.-D. Li, X.-F. Yin, X. Yuan, J.-C. Hung, L.-K. Chen, L. Li, N.-L. Liu, C.-Z. Peng, Y.-C. Liang, X. Ma, Y.-A. Chen, and J.-W. Pan, Phys. Rev. X **8**, 021072 (2018).
[50] Y. Zhou, B. Xiao, M.-D. Li, Q. Zhao, Z.-S. Yuan, X. Ma, and J.-W. Pan, npj Quantum Inf **8**, 1 (2022).
[51] The corresponding white noise regime for W state is $p_{\mathrm{noise}} \in [8/21, 0.791)$.
[52] X.-M. Hu, W.-B. Xing, Y. Guo, M. Weilenmann, E. A. Aguilar, X. Gao, B.-H. Liu, Y.-F. Huang, C.-F. Li, G.-C. Guo, Z. Wang, and M. Navascués, Phys. Rev. Lett. **127**, 220501 (2021).
[53] O. Gühne, Y. Mao, and X.-D. Yu, Phys. Rev. Lett. **126**, 140503 (2021), arXiv:2008.05961 [quant-ph].
[54] G. Riccardi, D. E. Jones, X.-D. Yu, O. Gühne, and B. T. Kirby, Exploring the relationship between the faithfulness and entanglement of two qubits (2021), arXiv:2102.10121 [quant-ph].
[55] Y. Zhan and H.-K. Lo, Detecting Entanglement in Unfaithful States (2021), arXiv:2010.06054 [quant-ph].
[56] Denote $O_\sigma \in \{I, X, Y, Z\}^{\otimes n}$ for a Pauli observable. De-

note $\mathbf{x}_{\rho,\boldsymbol{\sigma}} := (\mathrm{Tr}(\rho O_{\sigma_1}), \ldots, \mathrm{Tr}(\rho O_{\sigma_M}))$ for expectations of M Pauli observables $\boldsymbol{\sigma} \subseteq \{I, X, Y, Z\}^n$ with respect to the state ρ.

[57] D. Lu, T. Xin, N. Yu, Z. Ji, J. Chen, G. Long, J. Baugh, X. Peng, B. Zeng, and R. Laflamme, Phys. Rev. Lett. **116**, 230501 (2016), arXiv:1511.00581 [quant-ph].

[58] C. Cortes and V. Vapnik, Mach Learn **20**, 273 (1995).

[59] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Machine Learning **46**, 389 (2002).

[60] T. Hofmann, B. Schölkopf, and A. J. Smola, Ann. Statist. **36**, 10.1214/009053607000000677 (2008).

[61] J. Altepeter, E. Jeffrey, and P. Kwiat, in *Advances In Atomic, Molecular, and Optical Physics*, Vol. 52 (Elsevier, 2005) pp. 105–159.

[62] Quantum state tomography refers to the task of recovering the density matrix of an unknown D-dimensional state ρ within error tolerence ϵ, given the ability to prepare and measure copies of ρ.

[63] Adaptive measurements are the intermediate between independent measurements and collective (entangled) measurements, in which the copies of ρ are measured individually, but the choice of measurement basis can change in response to earlier measurements.

[64] $\tilde{\mathcal{O}}$ hides a polylog factor. A full tomography require estimate D^2 measuremments (observables) with additive error $\epsilon \ll 1/D$ for all E_i, so the sample complexity of shadow tomography is compatiable with lower bounds of full quantum state tomography.

[65] S. Aaronson, in *Proc. 50th Annu. ACM SIGACT Symp. Theory Comput.*, STOC 2018 (Association for Computing Machinery, New York, NY, USA, 2018) pp. 325–338, arXiv:1711.01053.

[66] C. Hadfield, S. Bravyi, R. Raymond, and A. Mezzacapo, Commun. Math. Phys. **391**, 951 (2022), arXiv:2006.15788 [quant-ph].

[67] H.-Y. Huang, R. Kueng, and J. Preskill, Phys. Rev. Lett. **127**, 030503 (2021), arXiv:2103.07510 [quant-ph].

[68] S. Chen, W. Yu, P. Zeng, and S. T. Flammia, PRX Quantum **2**, 030348 (2021), arXiv:2011.09636 [quant-ph].

[69] A. Elben, R. Kueng, H.-Y. Huang, R. van Bijnen, C. Kokail, M. Dalmonte, P. Calabrese, B. Kraus, J. Preskill, P. Zoller, and B. Vermersch, Phys. Rev. Lett. **125**, 200501 (2020), arXiv:2007.06305 [cond-mat, physics:quant-ph].

[70] T. Zhang, J. Sun, X.-X. Fang, X.-M. Zhang, X. Yuan, and H. Lu, Experimental quantum state measurement with classical shadows (2021), arXiv:2106.10190 [physics, physics:quant-ph].

[71] J. R. Johansson, P. D. Nation, and F. Nori, Computer Physics Communications **184**, 1234 (2013), arXiv:1110.0573.

[72] B. Li, S. Ahmed, S. Saraogi, N. Lambert, F. Nori, A. Pitchford, and N. Shammah, Quantum **6**, 630 (2022), arXiv:2105.09902 [quant-ph].

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[74] The open-source code for classical shadow with the code from `https://github.com/hsinyuan-huang/predicting-quantum-properties`.

[75] S. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt, J. Mach. Learn. Res. **11**, 1201 (2010), arXiv:0807.0093.

[76] A. Jacot, F. Gabriel, and C. Hongler, Neural Tangent Kernel: Convergence and Generalization in Neural Networks (2020), arXiv:1806.07572 [cs, math, stat].

[77] M. Schuld and N. Killoran, Phys. Rev. Lett. **122**, 040504 (2019), arXiv:1803.07128 [quant-ph].

[78] M. Schuld, Supervised quantum machine learning models are kernel methods (2021), arXiv:2101.11020 [quant-ph, stat].

[79] Y. Liu, S. Arunachalam, and K. Temme, Nat. Phys. **17**, 1013 (2021), arXiv:2010.02174 [quant-ph].

[80] X. Gao and L.-M. Duan, Nat Commun **8**, 662 (2017), arXiv:1701.05039 [cond-mat, physics:quant-ph].

[81] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Nature Phys **14**, 447 (2018), arXiv:1703.05334.

[82] Y. Zhu, Y.-D. Wu, G. Bai, D.-S. Wang, Y. Wang, and G. Chiribella, Flexible learning of quantum states with generative query neural networks (2022), arXiv:2202.06804 [quant-ph].

[83] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Nat Commun **12**, 2631 (2021), arXiv:2011.01938 [quant-ph].

[84] H.-Y. Huang, R. Kueng, and J. Preskill, Phys. Rev. Lett. **126**, 190505 (2021), arXiv:2101.02464 [quant-ph].