# Towards efficient and generic entanglement detection

Jue Xu* and Qi Zhao†
(Dated: October 4, 2022)

Detection of entanglement is an indispensable step for practical quantum computation and communication. In this work, we propose an end-to-end, machine learning assisted entanglement detection protocol. In this protocol, an entanglment witness for a generic entangled state is trained by classical SVM with synthetic data which consist of classical features of quantum state and their labels. In real experiments, classical features of a state, that is expectation values of a set of Pauli measurements, are estimated by sample-efficient methods such as classical shadow.

## I. INTRODUCTION

Entanglement [1] is the key ingredient of quantum computation [2], quantum communication, and quantum cryptography [3]. However, decoherence is inevitable in real-world, which means the interaction between a quantum system and classical environment would greatly affect entanglement quality and diminish quantum advantage. So, for practical purpose, it is essential to benchmark (characterize) entanglement (structures) of certain target states in actual experiments. Machine techniques including quantum machine learning [4] models have been widely applied to classification tasks in physics, such as classification of phases and prediction of ground state [5] [6].

The goal of this paper is to find an efficient and generic way to achieve it. Assume we would like to distinguish an entangled state incluing its 'vicinity' from all separable states, our method derive such a classifier by fitting a dataset sampled from target states with their labels (entangled or not). Specifically, our pipeline starts from evaluation of expectations of a subset of all possible $n$-qubit pauli operators. The set of expectation values that serves as classical features of a quantum state, together with its label, consist of a data point. A classical machine learning classifier is obtained by training with this dataset. With the trained classifier at hand, it is expected that brand new samples from real experiments can be classified with high accuracy, where classical features of quantum states are esitmated by classical shadow method [7] with affordable samples complexity.

This paper is organized as follows: in Section II, we briefly present necessary definitions about entanglement structures and mainstream entanglement detection methods; Section III demonstrates our end-to-end protocol including two parts: learning an entanglement witness and estimating classical features; at last, numerical simulation results are discussed in Section IV.

* juexu@cs.umd.edu
† zhaoqi@cs.hku.hk

## II. PRELIMINARIES

**Notation 1.** If no ambiguity, we omit the tensor products between subsystems and the hats on operators for readability, e.g., $|\psi_A\rangle |\psi_B\rangle \equiv |\psi_A\rangle \otimes |\psi_B\rangle$ and $X^{(1)}Z^{(3)} \equiv \hat{X} \otimes \mathbb{1} \otimes \hat{Z}$.

### A. Entanglement structures

Large scale entanglement involving multiple particles maybe the main resource of quantum advantages in quantum computation and communication. Roughly, we say a quantum state is entangled if it is not fully separable.

**Definition 1** (fully separable)**.** An $n$-particle (qubit) pure state $|\psi_f\rangle$ is fully separable if it can be written as the tensor product of subsystems $\{A_1, \ldots, A_n\}$, i.e., $|\psi_f\rangle = \bigotimes_{i=1}^{n} |\phi_{A_i}\rangle$. Then, a mixed state $\rho_f$ is fully separable if it can be written as a convex combination of fully separable pure states.

The simple statement 'The state is entangled' would still allow that only two of the qubits are entangled while the rest is in a product state. Firstly, we consider the simplest entanglement structure: bipartite separable case. Consider a bipartite system $AB$ with the Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$, where $\mathcal{H}_A$ has dimension $d_A$ and $\mathcal{H}_B$ has dimensional $d_B$, respectively.

**Definition 2** (bipartite separable)**.** A pure state $|\psi\rangle$ is bipartite (bi-)separable if it can be written as a tensor product form $|\psi_{\text{bi}}\rangle = |\phi_A\rangle \otimes |\phi_B\rangle$, where $\mathcal{P}_2 = \left\{A, B \equiv \bar{A}\right\}$ is a bipartition of the qubits in the system. A mixed state $\rho$ is separable if and only if it can be written as a convex combination of pure bi-separable states, i.e., $\rho_{\text{bi}} = \sum_i p_i |\psi_i\rangle\langle\psi_i|_{\text{bi}}$ with probability distribution $\{p_i\}$. The set of all bi-separable states is denoted as $S_{\text{bi}}$.

On the other hand, if a state is not a convex combination of any (partition) biseparable states, it means that all qubits in the system are indeed entangled with each other. This is the strongest form of entanglement, called genuine multipartite entanglement, formally defined as

**Definition 3** (GME)**.** If a state is not in $S_{\text{bi}}$, it possesses genuine multipartite entanglement (GME).

There is another restricted way for the extension to mixed states. In contrast to $S_{\text{bi}}$, a state is $\mathcal{P}_2$-separable, if it is a mixing of pure separable states with a same partition $\mathcal{P}_2$, and we denote the state set as $S_{\text{bi}}^{\mathcal{P}_2}$. It is practically interesting to study entanglement structure under partitions, because bi-separability between two subsystems naturally indicates loss of quantum information processing capabilities among certain geometric configuration in quantum commnuication or computation.

**Definition 4** (full entanglement)**.** A state $\rho$ possesses full entanglement if it is outside of the separable state set $S_{\text{bi}}^{\mathcal{P}_2}$ for any partition, that is, $\forall \mathcal{P}_2 = \left\{A, \bar{A}\right\}, \rho \notin S_{\text{bi}}^{\mathcal{P}_2}$.

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.

### B. Entanglement detection

Entanglemtn detection for a general state is a highly non-trivial problem. For a general review on this subject, we refer readers to [8]. The most widely studied problem maybe bi-separability.

**Problem 1** (separability)**.** Given a state $\rho$ in its density matrix repsentation, to determine if it is bipartite separable.

It is not hard to prove that if a state is bipartite separable, then it must have positive partial transpose (PPT), that is, the partially transposed (PT) density matrix $\rho_{AB}^{\mathsf{T}_A}$ is PSD [9]. By contrapositive, we have a criterion for entanglement, that is

**Theorem 1** (PPT criterion)**.** *If the smallest eigenvalue of partial transpose $\rho_{AB}^{\mathsf{T}_A}$ is negative, then the state is entangled (cannot be bi-separable) with respect to the partition $\mathcal{P} = \{A, B\}$.*

We should notice that PPT is a necessary and sufficient condition for separability for low-dimensional systems (a universal classifier when $d_A d_B \leq 6$) [10]. A natural question is whether it is possible to solve separability approximately. By relaxing the defintion (promise a gap), a reformulation of this problem in the theoretic CS language is

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

Classically, the hardness of determining the bipartite separability.

**Theorem 2** ([11])**.** *Weak membership problem for separability is NP-Hard for $\epsilon = 1/\operatorname{poly}(D)$ with respect to Euclidean norm and trace norm. [12] [13] while there exists a quasipolynomial-time algorithm with respect to $\|\cdot\|_{\text{LOCC}}$ (and $\|\cdot\|_2$?) [14].*

Even we know the complete information about a general state, it is hard to determine its separability. However, it does not rule out the possibility to solve it efficiently with stronger promise (approximation) or quantum algorithms, even machine learning powered by data.

#### 1. Direct detection and hardness

Unfortunately, no general solution for the separability problem is known. Another widely used and powerful one (criteria) is the $k$-symmetric extension hierarchy based on SDP [15], but computationally intractable with growing $k$. In order to apply the PPT criterion, the full density matrix must be available. However, **??** requires an exponential number of measurements.

Similar to the PPT condition, the $p_3$-PPT condition applies to mixed states and is completely independent of the state in question. From this data set, the PT-moments $p_n$ can be estimated without having to reconstruct the density matrix $\rho_{AB}$, and with a significantly smaller number of experimental runs $M$ than required for full quantum state tomography.

Rather than qualitatively determining (bi)separability, there are measures to quantify entanglement. For $\operatorname{Tr}(\rho_A^m)$, an important application of multivariate trace estimation [16] is to entanglement spectroscopy [17] [18] [19] - deducing the full set of eigenvalues of $\rho_A$. The smallest eigenvalue of diagnoses whether $\psi_{AB}$ is separable or entangled [20]. The well-known identity (related to the replica trick originating in spin glass theory)

$$\operatorname{Tr}(U^\pi(\rho_1 \otimes \cdots \otimes \rho_m)) = \operatorname{Tr}(\rho_1 \cdots \rho_m) \tag{1}$$

where the RHS is the multivariate trace and $U^\pi$ is a unitary representation of the cyclic shift permutation. Direct entanglement detections, can be employed as subroutines in quantum computation and realized by quantum circuits because it is naturally quantum data. unknown how to generalize to multipartite. estimate the spectrum by quantum constant depth circuits, but this method only works for bipartite. solve the quantum problem by quantum, without full tomography.

#### 2. Entanglement witness based on fidelity

Another approach for detecting multipartite entanglement is using entanglement witnesses. does not require any a priori knowledge about the quantum state. This is a key distinction from entanglement witnesses, which can be more powerful, but which usually require detailed prior information about the state. In a typical experiment one aims to prepare a pure state, $|\psi_{\text{tar}}\rangle$, and would like to detect it as true multipartite entangled. While the preparation is never perfect, it is still expected that the prepared mixed state is in the proximity of $|\psi_{\text{tar}}\rangle$. Different from separability problem, we assume prior knowledge of the target state $|\psi\rangle$, that is, $|\psi_{\text{tar}}\rangle$ undergoes some

noise channels: white noise, bit/phase-flip error, or random local unitary

**Problem 3** (entanglement detection)**.** Given an unknown state $\rho$ (from experiments) is promised in 'proximity' of a target $|\psi_{\text{tar}}\rangle$, determine whether $\rho$ possesses 'useful' entanglement (such as GME, full entanglement, $S_b^{\mathcal{P}}$, intactness, depth ...) or not.

Bell inequalities as the oldest tool to detect entanglement [21]. Originally, Bell inequalities were designed to rule out local hidden variable (LHV) models. It can also detect many entangled 2-qubit states. CHSH inequality $\hat{\mathbf{p}} = (\mathbb{1}, ab, ab', a'b, a'b')$

$$a = Z, a' = X, b = (X - Z)/\sqrt{2}, b = (X + Z)/\sqrt{2}, \quad (2)$$

features as the input. CHSH operator $W_{\text{CHSH}} := \mathbf{w}_{\text{CHSH}} \cdot \hat{\mathbf{p}}$ with $\mathbf{w}_{\text{CHSH}} = (\pm 2, 1, -1, 1, 1)$. However, even for two-qubit systems there exist entangled states which do not violate any Bell inequality. 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.

**Definition 5** (entanglement witness)**.** Given a speific entangled state $|\psi_{\text{tar}}\rangle$, its entanglement witness $W$ is an obseverable such that

$$\text{Tr}(W\rho_{\text{bi}}) \geq 0 \text{ and } \text{Tr}(W|\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}|) < 0 \quad (3)$$

In general, an entanglement witness is an observable which has a positive expectation value on all separable states, hence a negative mean value implys the presence of entanglement. There is no entanglement witness that detects all entangled states [22]. see Fig. 3 for relations.

While various methods for constructing an entanglement witness exist, one of the most common is based on the fidelity of a state to a pure entangled state. The usual way to construct entanglement witnesses using the knowledge of this state is

$$W_\psi = \alpha\mathbb{1} - |\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}| \quad (4)$$

where $\alpha$ is the smallest constant such that for every product state $\text{Tr}(\rho W) \geq 0$. This kind of witness is projector-based witness [23]. However, it is generally difficult to evaluate the quantity $\text{Tr}(\rho_{\text{pre}}|\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}|)$ by the direct projection, because the target state is an entangled state. $W = \alpha\mathbb{1} - |\text{GHZ}\rangle\langle\text{GHZ}|$ for any state $\rho_s$ with only bipartite entanglement, $\text{Tr}(O\rho_s) \leq 0.5$ ($\geq 0.5$ certifies tripartite entanglement), while for any state $\rho_s$ with at most $W$-type entanglement, $\text{Tr}(O\rho_s) \leq 0.75$. $\text{Tr}(O\rho) > 0.75$ certifies that $\rho$ has GHZ-type entanglement [24].

In order to measure the witness in an experiment, it is supposed be decomposed into a sum of locally measurable operators. For graph states (stabilizer states), witness can be constructed by very few local measurement settings (tradeoff tolerance) [25] [26] [27].

The common robustness measure of fidelity wintess is the tolerence of white noise.

$$\rho' = (1 - p_{\text{noise}})|\psi_{\text{tar}}\rangle\langle\psi_{\text{tar}}| + 2^{-n} \cdot p_{\text{noise}}\mathbb{1} \quad (5)$$

where the limit of (maximal) $p_{\text{noise}}$ indicates the robustness of the witness. There is a tradeoff between white noise tolerance (robustness) and efficiency (number of measurements). For non-stabilizer case, more careful analysis is required [28] [29].

### 3. *Beyond fidelity witness*

For GHZ and W state with white noise, we can analytically compute the white noise threshold for the PPT criterion (NPT for bipartite entanglement). When $p_{\text{noise}} < 0.8$, GHZ states have full entanglement. However, the conventional fidelity witness detects GME when $p_{\text{noise}} < 1/2$. It would be practically interesting to have a witness for full entanglement.

Other than white noise, more practical noise happened in experiments is coherent noise, e.g., local unitary/rotation. Take GHZ as an example, unconscious phase accumulation and rotation on the first control qubit can be modeled as [30]

$$|\text{GHZ}(\phi, \theta)\rangle = \cos\theta\,|0\rangle^{\otimes n} + e^{i\phi}\sin\theta\,|1\rangle^{\otimes n}. \quad (6)$$

In some noise parameter regime, $|\text{GHZ}(\phi, \theta)\rangle$ cannot be detected by conventional fidelity witness (need post measurement processing).

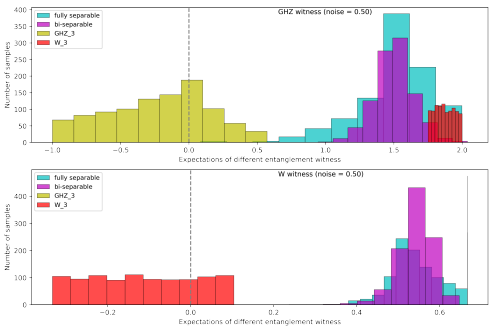


FIG. 1: Entanglement cannot be detected by fidelity witness (large white noise, coherent noise)

To formally characterize the states beyond fidelity witness, Weilenmann et. al [31] coined the term *unfaithful states* which systematically analyze entangled (2-qudit) state mixed with white noise cannot be detected by fidelity witness. faithful states are useful for quantum teleportation. [32] [33] [34]

**Definition 6** (unfaithful state)**.** a state $\rho_{AB}$ is faithful if and only if there are local unitary transformations $U_A$

and $U_B$ such that

$$\langle\phi^+|U_A\otimes U_B\rho_{AB}U_A^\dagger\otimes U_B^\dagger|\phi^+\rangle > \frac{1}{d} \tag{7}$$

For 2-qubit systems, unfaithfulness is identified by

$$X_2(\rho_{AB}) = \rho_{AB} - \frac{1}{2}(\rho_A\otimes\mathbb{1}+\mathbb{1}\otimes\rho_B)+\frac{1}{2}\mathbb{1}\otimes\mathbb{1} \tag{8}$$

if and only if the smallest eigenvalue of $X_2(\rho_{AB})$ is greater than $1/2$.

They found that for $d\geq 3$ that almost all states in the Hilbert space are unfaithful. Although there are nonlinear witnesses which also can detect entanglement in unfaithful states, they usually require more measurements [] [35]. Moreover, they can only be applied to bipartite systems, which means they cannot be generalized to detect genuine entanglement in multipartite states. It is natural to ask how nonlinear entanglement witness [36] and the kernel method (nonlinear boundary) in machine learning can be applied.

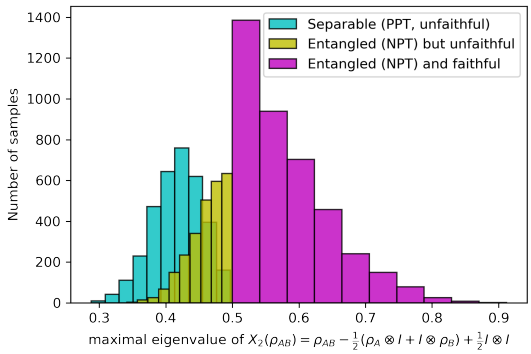


FIG. 2: Faithfulness of random 2-qubit states: ...

## III. END-TO-END ENTANGLEMENT DETECTION PROTOCOL

In the past several years, classical machine learning algorithms have been proposed for classification tasks related to entanglement. The idea is to feed the classifier by a large amount of sampled trial states as well as their corresponding class labels. The quantum extension of this problem (classficiation/pattern recognition) is to replace the data points $\mathbf{x}_i$ with density matrices of quantum states $\rho_i$. Specifically, a quantum state classifier outputs a "label" associated with the state, for example, entangled or "unentangled".

Lu et. al [37] trained a universal bipartite separable classifier by classical neural network where input dataset are (randomly sampled) density matrices with labels?. For the similar purpose, Ma and Yung [38] trained Bell (inequlaity) like ansatz and tomographic witness by neural network. an ansatz for entanglement witness Different Bell inequalities can be regarded as entanglement witness for different types of entanglement in a multi-party entangled state.

$$W_{\text{ansatz}} := \sum_{\mathbf{p}\in\{I,X,Y,Z\}^n} w_{\mathbf{p}}\bigotimes_i^n \mathbf{p}_i \tag{9}$$

c.f. **??** a linear Bell-like predictor by generalizing the CHSH operator $W_{\text{ml}} := \mathbf{P}\cdot\mathbf{w}_{\text{ml}}$ in Eq. (2) where the coefficients (or weights) $\mathbf{w}$ are determined by machine learning. It is not surprised that tomographic ansatz has better performance (also required [39]) for universal bi-separability. However, training a universal classifier (or GME) for multi-qubit, high-dimensional system is hard if the gap between two sets is small.

In our paper, we focus on the task entanglement detection with training data. In other word, we derive an entanglement witness for desired states with entannglement structure by fitting a dataset.

**Problem 4** (Learn an entanglement witness)**.** Assume we would like to derive a witness for certain entanglement structure of a state $\rho$

- **Input**: synthetic data consist of density matrices $\rho$ with corresponding labels $y$
- **Output**: the minimal classifier $\mathbf{w}_{\text{ml}}$ with high training accuracy

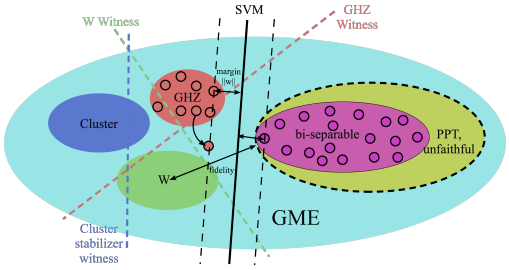


FIG. 3: entanglement witness, PPT criterion, SVM (kernel)?. convex hull... The witness is depicted by the line in state space $\text{Tr}(\rho W)=0$

### A. Training a generic witness via SVM

On the other hand, entanglement detection is also studied by classical machine learning algorithm [29], but a different technique called Support Vector Machine (SVM) [40]. SVM has the very similar structure to entanglement witness. feature: $\mathbf{x}_{\vec{k}}$ expectation of Pauli strings. This method the ability to obtain witnesses that require only local measurements even when the target state is a non-stabilizer state $W$ state (normally need nonlocal measurements).

An SVM allows for the construction of a hyperplane $\langle W\rangle = \sum_k w_k\mathbf{x}_k$ that clearly delineates between separable states and the target entangled state this hyperplane

(linear SVM and witness) is a weighted sum of observables (features) whose coefficients are optimized during the training of the SVM. We focus on kernel methods rather than neural networks, as they not only provide provable guarantees, but are also very flexible in the functions they can learn. For example, recent advancements in theoretical machine learning show that training neural networks with large hidden layers is equivalent to the neural tangent kernel [41].

the training of an SVM is convex; if a solution exists for the given target state and ansatz, the optimal SVM will be found. this SVM formalism allows for the programmatic removal of features, i.e., reducing the number of experimental measurements, in exchange for a lower tolerance to white noise, in a manner similar to [??]. SVM, (universal), 4 qubit [42]. In actual experiments, we don't know entries of a density matrix and only focus on white noise. Instead, we need measurements or classical shadow as features of machine learning algorithms. However, these SVM algorithms cannot be directly applied to experiments because they didn't address the problem of estimating classical features efficiently which we will discuss in Section III B 1.

| | # observables | weights | input state |
|---|---|---|---|
| fidelity witness | few local | fixed | partial |
| Bell (CHSH) inequality | constant | fixed | unknown |
| tomographic classifier | $4^n-1$ | trained | unknown |
| SVM kernel witness | $\ll 4^n-1$ | trained | partial |

TABLE I: Comparison of CHSH inequality, conventional fidelity witness, and machine learning witness ansatz.

**Algorithm III.1:** train witness via SVM

**input :** states with labels $\{(\rho_i, y_i)\}_i^m$
**output:** a classifier $\mathbf{w}_{\text{ml}}$

Evaluate all Pauli observables $\mathbf{x} := \text{Tr}(P_{\mathbf{x}}\rho_i)\ \forall \rho_i$
**for** $j = 1, 2, \ldots, 4^n$ **do**
  **while** *accuracy not high enough* **do**
    randomly select $j$ features $\tilde{\mathbf{x}}$ from $\mathbf{x}$
    accuracy, classifier = SVM$(\{(\tilde{\mathbf{x}}_i, y_i)\}_i^m)$
  **return** classifier $\mathbf{w}_{\text{ml}}$
Test the classifier on test dataset

### B. Sample-efficient expectation estimation methods

To apply classical machine learning classifier (SVM) in actual experiments, we need classical features of quantum states. Nevertheless, we cannot directly evaluate expectation values as in numerical simulation because we do not know density matrices of incoming states. The brute force approach is to fully characterize a state by performing quantum state tomography [43] and then calculating classcial features or separability measures from the recovered density matrix.

However, full quantum state tomography is experimentally and computationally demanding. The canonical representation (decomposition) of a $n$-qubit density matrix is $\rho = 2^{-n}\sum_{\mathbf{p}\in\{I,X,Y,Z\}^n} t_{\mathbf{p}} \bigotimes_i^n \mathbf{p}$ [44]. Since there are $4^n$ terms in the decomposition, navie quantum state tomography based on independent measurements need at least exponential copies. Rigorous analysis [45] [46] proved that $\Omega(Dr/\epsilon^2)/\log(D/r\epsilon)$ copies/measurements is required for $D$-dimensional ($D = 2^n$ for $n$-qubit systems) rank-$r$ density matrix with error $\epsilon$ (trace distance) if adaptive measurements allowed [47].

Now that full tomography is expensive, a natural question is whether it is possible to extract a bunch of information about a state without fully recover the state. The answer is yes. In quantum mechanics, interesting properties are often linear functions of the underlying density matrix $\rho$, such as classical features $\langle O\rangle = \text{Tr}(O\rho)$ for entanglement witness [48]. This enables the possibility to shadow tomography, the idea proposed by Aaronson [49].

**Problem 5** (shadow tomography)**.** $\mathbb{P}[E_i \text{ accept } \rho] = ?\,\text{Tr}(E_i\rho)$

- **Input**: copies of an unknown $D$-dimensional state $\rho$ and $M$ known 2-outcome measurements $\{E_1, \ldots, E_M\}$
- **Output**: estimate $\mathbb{P}[E_i \text{ accept } \rho]$ to within additive error $\epsilon$, $\forall i \in [M]$, with $\geq 2/3$ success probability.

It is proved that shadow tomography can be implemented in samples-efficient (copies) manner.

**Theorem 3** ([49])**.** *It is possible to do shadow tomography using* $\tilde{\mathcal{O}}(\log^4 M \cdot \log D \cdot \epsilon^{-4})$ *copies. [50] sample complexity lower bound* $\Omega(\log(M)\cdot\epsilon^{-2})$*,*

Since Aaronson's shadow tomography procedure is very demanding in terms of quantum hardware, Huang et. al [7] introduce classical shadow (CS) method that is more friendly to experiments. In our pipeline, we focus on classical shadow method.

#### 1. *Estimate expectations via classical shadow*

A classical shadow is a succinct classical description of a quantum state, which can be extracted by performing reasonably simple single-copy measurements on a reasonably small number of copies of the state. The classical shadow attempts to approximate this expectation value by an empirical average over $T$ independent samples, much like Monte Carlo sampling approximates an integral. Directly measuring $M$ different entanglement witnesses requires a number of quantum measurements that scales (at least) linearly in $M$. In contrast, classical shadows get by with $\log(M)$-many measurements only. classical shadows are based on random Clifford measurements and do not depend on the structure of the concrete witness in question. In contrast, direct estimation crucially depends on the concrete witness in question and may be considerably more difficult to implement.

**Definition 7** (classical shadow)**.** The classical shadow of a state $\rho$ is a ... such that we can predict the linear function with it

$$o_i = \mathrm{Tr}(O_i \rho_{\mathrm{cs}}) \text{ obeys } \mathbb{E}[o] = \mathrm{Tr}(O_i \rho) \tag{10}$$

**Algorithm III.2:** estimate features by CS

**input :** samples of $\rho$ and $\mathbf{P}_{\mathrm{ml}}$
**output:** estimation of $\mathrm{Tr}(\rho \mathbf{P}_{\mathrm{ml}})$

**for** $i = 1, 2, \ldots, N$ **do**
  $\rho \mapsto U \rho U^\dagger$ // apply a random unitary
  $\mapsto |b\rangle ...$ // measurement
  $\rho_{\mathrm{cs}} = \mathcal{M}^{-1}\left(U^\dagger |b\rangle\langle b| U\right)$ // measurement outcome $|b\rangle \in \{0,1\}^n$, $\mathcal{M}$ quantum channel
$S(\rho, N) = \{\rho_{\mathrm{cs}_1}, \ldots, \rho_{\mathrm{cs}_N}\}$ // call this array the classical shadow of $\rho$
**return** $\mathrm{Tr}(\mathbf{P}_{\mathrm{ml}} \rho)$ // estimate features for SVM

Given a quantum state $\rho$, a classical shadow is created by repeatedly performing a simple procedure: Apply a unitary transformation $\rho \mapsto U \rho U^\dagger$, and then measure all the qubits in the computational basis ...$|b\rangle$. its classical shadow (snapshots) $\rho_{\mathrm{cs}}$ is

$$\rho_{\mathrm{cs}} := \mathcal{M}^{-1}\left(U^\dagger \left|\hat{b}\right\rangle\!\left\langle\hat{b}\right| U\right) \tag{11}$$

where $|b\rangle$ is ... $U \rho U^\dagger$. The number of times this procedure is repeated is called the size of the classical shadow. The transformation $U$ is randomly selected from an ensemble of unitaries, and different ensembles lead to different versions of the procedure that have characteristic strengths and weaknesses. Classical shadows with size of order $\log(M)$ suffice to predict $M$ target functions $\{O_1, \ldots, O_M\}$.

**Theorem 4** ([7])**.** *Any procedure based on a fixed set of single-copy local measurements that can predict, with additive error $\epsilon$, $M$ arbitrary $k$-local linear function $\mathrm{Tr}(O_i \rho)$, requires at least (lower bound) $\Omega(\log(M) 3^k/\epsilon^2)$ copies of the state $\rho$. [51] $\Omega(\log(M) \max_i \mathrm{Tr}(O_i^2)/\epsilon^2)$*

The classical shadow size required to accurately approximate all reduced $r$-body density matrices scales exponentially in subsystem size $r$, but is independent of the total number of qubits $n$. Derandomized variant of classical shadow [52] is the refinement of the original randomized protocol, but not necessarily guarantees better performance for global observables (involving all subsystems).

### 2. *Estimate expectation by machine learning*

The task of estimating expectation value can also be achieved by machin learning with training data. The quantum ML algorithm accesses the quantum channel $\mathcal{E}_\rho$ multiple times to obtain multiple copies of the underlying quantum state $\rho$. Each access to $\mathcal{E}_\rho$ allows us to obtain one copy of $\rho$. Then, the quantum ML algorithm performs a sequence of measurements on the copies of $\rho$ to accurately predict $\mathrm{Tr}(P_{\mathbf{x}} \rho), \forall \mathbf{x} \in \{I, X, Y, Z\}^n$.

Huang et. al rigorously show that, for any quantum process $\mathcal{E}$, observables $O$, and distribution $\mathcal{D}$, and for any quantum ML model, one can always design a classical ML model achieving a similar average prediction error such that $N_C$ (number of experiments?) is larger than $N_Q$ by at worst a small polynomial factor. In contrast, for achieving accurate prediction on all inputs, exponential quantum advantage is possible.

**Theorem 5** ([53])**.** *To predict expectations of all Pauli observables of an $n$-qubit system $\rho$, classical ML models require $2^{\Omega(n)}$ copies of $\rho$ , there is a quantum ML model using only $\mathcal{O}(n)$ copies. $\mathcal{O}(\log(M/\delta)\epsilon^{-4})$ copies of the unknown quantum state $\rho$. ($M = 4^n$ implies linear copy for full tomography???)*

[7] [6] [53] [54] ... the required amount of training data scales badly with $\epsilon$. This unfortunate scaling is not a shortcoming of the considered ML algorithm, but a necessary feature. [55] [56] generative neural network [57]

| | circuit/sample complexity |
|---|---|
| shadow tomography | Theorem 3 (exponential circuit?) |
| classical shadow | Theorem 4 (experiment friendly) |
| classical/quantum ML | Theorem 5 (quantum advantage) |

TABLE II: complexity (measures) of different expectation estimation methods

## IV. NUMERICAL SIMULATION

### A. Data preparation and state generation

We generate quantum state samples, construct quantum circuits, and manipulate quantum objects numerically by QuTiP library [58] [59]. multi-partite entangled states generation: Bell, GHZ, W state, graph (cluster) state separable states generation: 2-qubit: bipartite $\rho_A \otimes \rho_B$ where $\rho_A$ is a random density matrix sampled (Haar measure); random multi-qubit pure states: $\rho_A \otimes \rho_{BC}$, $\rho_C \otimes \rho_{AB}$, and $\rho_B \otimes \rho_{AC}$. noice channels: white noise Eq. (5), coherent noise Eq. (6). dataset size: $10^4$ for 3-qubit case.

For the machine learning part, we make use of scikit-learning package [60] to train SVM with the rbf kernel.

### B. Classification accuracy and comparison

We consider a set of different regularization parameters,...

The goal of RFE is to eliminate non-essential features by recursively considering smaller and smaller subsets of

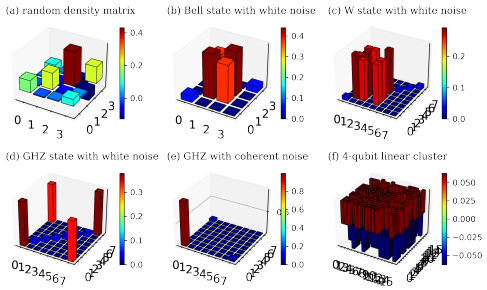


FIG. 4: Data praparation: (a) random 2-qubit density matrix; (b) Bell state with white noise; (c) 3-qubit W state with white noise; (d) 3-qubit GHZ state with white noise; (e) GHZ state with coherent noise; (f) 4-qubit linear cluster.

the original features using a greedy algorithm. Initially, RFE takes the SVM we trained and ranks the coefficients by their magnitudes, with the lowest one pruned away; then the model is trained again with the remaining features. Fig. 5. Fig. 6.

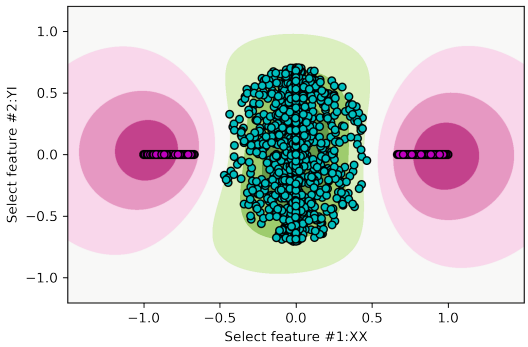


FIG. 5: two-dimensional embedding (non-linear kernel): feature space

performance of different methods:

## V. EXPERIMENTS [TODO]

Related experiments: photonic implementation with a few qubits (generation, verification) [61]; fully entangled graph state (ring of 16 qubits) IBM by measuring negativity [62]; optical lattice [63] (homogeneous, restricted measurement, different noise channels; detect GME, full entanglement); experiments of classical shadow [64]; detect entanglement by estimating $p_3$-PPT by classical shadow [19].

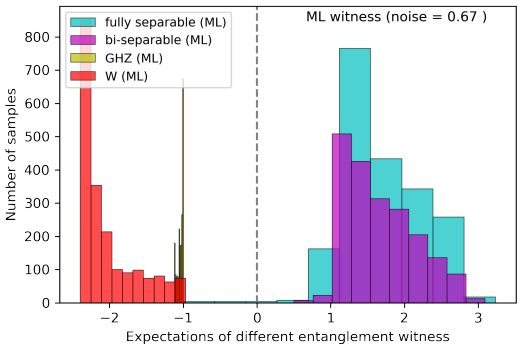


FIG. 6: ML witness for the state cannot be detect by fidelity witness (large white noise), non-stabilizer (W) state

## VI. CONCLUSION AND DISCUSSION

Possible future research directions: (1) rigorous proof for dataset size and number of features (required for high training accuracy) scaling with the system size; (2) better kernel options such as graph kernel, quantum kernel, shadow kernel and neural tangent kernel; (3) quantum machine learning for estimating all classical features (tomography) efficiently; (4) if we have all classical features, is it possible to train a universal classifier or with weaker promise; (5) can we estimate concurrence by quantum circuit

## ACKNOWLEDGMENTS

[1] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, Rev. Mod. Phys. **81**, 865 (2009), arXiv:quant-ph/0702225.
[2] H. J. Briegel, D. E. Browne, W. Dür, R. Raussendorf, and M. V. den Nest, Nature Phys **5**, 19 (2009), arXiv:0910.1116.
[3] F. Xu, X. Ma, Q. Zhang, H.-K. Lo, and J.-W. Pan, Rev. Mod. Phys. **92**, 025002 (2020), arXiv:1903.09051.
[4] I. Cong, S. Choi, and M. D. Lukin, Nat. Phys. **15**, 1273 (2019), arXiv:1810.03787 [cond-mat, physics:quant-ph].

[5] J. Carrasquilla and R. G. Melko, Nature Phys **13**, 431 (2017), arXiv:1605.01735.

[6] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill, Science **377**, eabk3333 (2022), arXiv:2106.12627.

[7] H.-Y. Huang, R. Kueng, and J. Preskill, Nat. Phys. **16**, 1050 (2020), arXiv:2002.08953 [quant-ph].

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

[9] A. Peres, Phys. Rev. Lett. **77**, 1413 (1996), arXiv:quant-ph/9604005.

[10] M. Horodecki, P. Horodecki, and R. Horodecki, Physics Letters A **223**, 1 (1996), arXiv:quant-ph/9605038.

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

[12] L. M. Ioannou, Quantum Inf. Comput. **7**, 335 (2007), arXiv:quant-ph/0603199.

[13] A. C. Doherty, P. A. Parrilo, and F. M. Spedalieri, Phys. Rev. A **69**, 022308 (2004), arXiv:quant-ph/0308032.

[14] F. G. S. L. Brandao, M. Christandl, and J. Yard, A quasipolynomial-time algorithm for the quantum separability problem (2011), arXiv:1011.2751 [quant-ph].

[15] M. Navascues, M. Owari, and M. B. Plenio, Phys. Rev. A **80**, 052306 (2009), arXiv:0906.2731 [quant-ph].

[16] Y. Quek, M. M. Wilde, and E. Kaur, Multivariate trace estimation in constant quantum depth (2022), arXiv:2206.15405 [hep-th, physics:quant-ph].

[17] 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.

[18] S. Johri, D. S. Steiger, and M. Troyer, Phys. Rev. B **96**, 195136 (2017), arXiv:1707.07658.

[19] 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].

[20] P. Horodecki and A. Ekert, Phys. Rev. Lett. **89**, 127902 (2002), arXiv:quant-ph/0111064.

[21] B. M. Terhal, Physics Letters A **271**, 319 (2000), arXiv:quant-ph/9911057.

[22] T. Heinosaari and M. Ziman, *The Mathematical Language of Quantum Theory: From Uncertainty to Entanglement*, 1st ed. (Cambridge University Press, 2011).

[23] 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.

[24] A. Acin, D. Bruss, M. Lewenstein, and A. Sanpera, Phys. Rev. Lett. **87**, 040401 (2001), arXiv:quant-ph/0103025.

[25] G. Toth and O. Guehne, Phys. Rev. Lett. **94**, 060501 (2005), arXiv:quant-ph/0405165.

[26] G. Tóth and O. Gühne, Phys. Rev. A **72**, 022340 (2005).

[27] Y. Zhou, Q. Zhao, X. Yuan, and X. Ma, npj Quantum Inf **5**, 83 (2019).

[28] Y. Zhang, Y. Tang, Y. Zhou, and X. Ma, Phys. Rev. A **103**, 052426 (2021), arXiv:2012.07606 [quant-ph].

[29] E. Y. Zhu, L. T. H. Wu, O. Levi, and L. Qian, Machine Learning-Derived Entanglement Witnesses (2021), arXiv:2107.02301 [quant-ph].

[30] Y. Zhou, Phys. Rev. A **101**, 012301 (2020), arXiv:1907.11495 [quant-ph].

[31] M. Weilenmann, B. Dive, D. Trillo, E. A. Aguilar, and M. Navascués, Phys. Rev. Lett. **124**, 200502 (2020), arXiv:1912.10056 [quant-ph].

[32] O. Gühne, Y. Mao, and X.-D. Yu, Phys. Rev. Lett. **126**, 140503 (2021), arXiv:2008.05961 [quant-ph].

[33] 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].

[34] 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).

[35] Y. Zhan and H.-K. Lo, Detecting Entanglement in Unfaithful States (2021), arXiv:2010.06054 [quant-ph].

[36] O. Gühne and N. Lütkenhaus, Phys. Rev. Lett. **96**, 170502 (2006).

[37] 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].

[38] Y.-C. Ma and M.-H. Yung, npj Quantum Inf **4**, 34 (2018), arXiv:1705.00813 [quant-ph].

[39] 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].

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

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

[42] 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].

[43] Informally, quantum state tomography refers to the task of estimating complete description (density matrix) of an unknown $D$-dimensional state $\rho$ within error tolerence $\epsilon$, given the ability to prepare and measure copies of $\rho$.

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

[45] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu, IEEE Trans. Inform. Theory , 1 (2017).

[46] R. O'Donnell and J. Wright, in *Proc. Forty-Eighth Annu. ACM Symp. Theory Comput.* (ACM, Cambridge MA USA, 2016) pp. 899–912.

[47] Intermediate between independent measurements and unrestricted (also called "collective" or "entangled") measurements are adaptive measurements in which the copies of $\rho$ are measured individually, but the choice of measurement basis can change in response to earlier measurements.

[48] Nonlinear functions: entropy; multivariate functions: $\mathrm{Tr}(\rho_1 \cdots \rho_m)$, quantum kernel $\mathrm{Tr}(\rho\rho')$, quadratic $\mathrm{Tr}(O\rho_i \otimes \rho_j)$, fidelity $F(\rho, \rho')$.

[49] 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.

[50] Additive error $\epsilon \ll 1/D$.

[51] Known fundamental lower bounds state that classical shadows of exponential size (at least) $T = \Omega(2^n/\epsilon^2)$ are required to $\epsilon$-approximate $\rho$ in trace distance.

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

[53] H.-Y. Huang, R. Kueng, and J. Preskill, Phys. Rev. Lett. **126**, 190505 (2021), arXiv:2101.02464 [quant-ph].
[54] 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].
[55] X. Gao and L.-M. Duan, Nat Commun **8**, 662 (2017), arXiv:1701.05039 [cond-mat, physics:quant-ph].
[56] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Nature Phys **14**, 447 (2018), arXiv:1703.05334.
[57] 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].
[58] J. R. Johansson, P. D. Nation, and F. Nori, Computer Physics Communications **184**, 1234 (2013), arXiv:1110.0573.
[59] B. Li, S. Ahmed, S. Saraogi, N. Lambert, F. Nori, A. Pitchford, and N. Shammah, Quantum **6**, 630 (2022), arXiv:2105.09902 [quant-ph].
[60] 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).
[61] 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).
[62] Y. Wang, Y. Li, Z.-q. Yin, and B. Zeng, npj Quantum Inf **4**, 46 (2018), arXiv:1801.03782.
[63] Y. Zhou, B. Xiao, M.-D. Li, Q. Zhao, Z.-S. Yuan, X. Ma, and J.-W. Pan, npj Quantum Inf **8**, 1 (2022).
[64] 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].
[65] M. Hein, W. Dür, J. Eisert, R. Raussendorf, M. V. den Nest, and H.-J. Briegel, Entanglement in Graph States and its Applications (2006), arXiv:quant-ph/0602096.
[66] L. Bai, L. Rossi, A. Torsello, and E. R. Hancock, Pattern Recognition **48**, 344 (2015).

## Appendix A: Definitions

**Definition 8** (density matrix)**.** A quantum (mixed) state $\rho$ can be represented by a density matrix which is a Hermitian, PSD operator (matrix) of trace one. If the rank of $\rho$ is 1, then the state is a pure state $\rho \equiv |\psi\rangle\langle\psi|$.

**Definition 9** (POVM)**.** A positive-operator valued measurement (POVM) $M$ consists of a set of positive operators that sum to the identity operator $\mathbb{1}$. When a measurement $M = \{E_1, \ldots, E_k\}$ is applied to a quantum state $\rho$, the outcome is $i \in [k]$ with probability $p_i = \mathrm{tr}(\rho E_i)$. observables ... $\mathbb{E}[x] \equiv \langle O_x \rangle := \mathrm{tr}(O_x \rho)$

**Definition 10** (PSD)**.** A matrix (operator) is positive, semidefinite (PSD) if all its eigenvalues are non-negative.

**Definition 11** (reduced density matrix)**.** reduced density matrix $\rho_A = \mathrm{Tr}_B(\rho_{AB})$

**Definition 12** (partial transpose)**.** [10] The partial transpose (PT) operation - acting on subsystem $A$ - is defined as

$$|k_A, k_B\rangle\langle l_A, l_B|^{\intercal_A} := |l_A, k_B\rangle\langle k_A, l_B| \tag{A1}$$

where $\{|k_A, k_B\rangle\}$ is a product basis of the joint system $\mathcal{H}_{AB}$.

**Definition 13** (Schmidt measure)**.** Consider the following bipartite pure state, written in Schmidt form:

$$|\psi\rangle = \sum_i^r \sqrt{\lambda_i}\, |\phi_i^A\rangle \otimes |\phi_i^B\rangle \tag{A2}$$

where $\{|\phi_i^A\rangle\}$ is a basis for $\mathcal{H}_A$ and $\{|\phi_i^A\rangle\}$ for $\mathcal{H}_B$. The strictly positive values $\sqrt{\lambda_i}$ in the Schmidt decomposition are its *Schmidt coefficients*. The number of Schmidt coefficients, counted with multiplicity, is called its *Schmidt rank*, or Schmidt number. Schmidt measure is minimum of $\log_2 r$ where $r$ is number of terms in an expansion of the state in product basis. (Schmidt rank ?? $\mathrm{SR}^A(\psi) = \mathrm{rank}(\rho_\psi^A)$)

**Definition 14** (entropy)**.** In quantum mechanics (information), the von Neumann *entropy* of a density matrix is $H_N(\rho) := -\mathrm{Tr}(\rho \log \rho) = -\sum_i \lambda_i \log(\lambda_i)$; In classical information (statistical) theory, the Shannon entropy of a probability distribution $P$ is $H_S(P) := -\sum_i P(x_i) \log P(x_i)$.

**Definition 15** (entanglement entropy)**.** The bipartite *von Neumann entanglement entropy* $S$ is defined as the von Neumann entropy of either of its reduced density matrix $\rho_A$. For a pure state $\rho_{AB} = |\Psi\rangle\langle\Psi|_{AB}$, it is given by

$$E(\Psi_{AB}) = S(\rho_A) = -\mathrm{Tr}(\rho_A \log \rho_A) = -\mathrm{Tr}(\rho_B \log \rho_B) = S(\rho_B) \tag{A3}$$

where $\rho_A$ and $\rho_B$ are the reduced density matrix for each partition. With Schmidt decomposition (Eq. (A2)), the entropy of entanglement is simply $-\sum_i p_i^2 \log(p_i)$. the $n$th Renyi entropy, $S_n = \frac{1}{n-1} \log(R_n)$ where $R_n = \mathrm{Tr}(\rho_A^n)$

**Example 1.** The Schmidt measure for any multi-partite GHZ states is 1, because there are just two terms. Schmidt measure for 1D, 2D, 3D-cluster state is $\lfloor \frac{N}{2} \rfloor$. Schmidt measure of tree is the size of its minimal vertex cover[??]. other entanglement measures...

**Definition 16** (fidelity)**.** Given a pair of states (target $\rho$ and prepared $\rho'$), Uhlmann fidelity $F(\rho,\rho') := \mathrm{Tr}\left(\sqrt{\sqrt{\rho}\rho'\sqrt{\rho}}\right) \equiv \left\|\sqrt{\rho}\sqrt{\rho'}\right\|_1$, where $\sqrt{\rho}$ dentoes the positive semidefinite square root of the operator $\rho$. (infidelity $1-F(\rho,\rho')$) For any mixed state $\rho$ and pure state $|\psi\rangle$, $F(\rho,|\psi\rangle\langle\psi|) = \sqrt{\langle\psi|\rho|\psi\rangle} \equiv \sqrt{\mathrm{Tr}(\rho\,|\psi\rangle\langle\psi|)}$ which can be obtained by the Swap-test[?]. linear fidelity or overlap $F(\rho,\rho') := \mathrm{tr}(\rho\rho')$.

**Notation 2** (norm)**.** Schatten p-norm $\|x\|_p := (\sum_i |x_i|^p)^{1/p}$. Euclidean norm $l_2$ norm; Spectral (operator) norm $\|\mathbf{x}\|_\infty$; Trace norm $\|A\|_{\mathrm{Tr}} \equiv \|A\|_1 := \mathrm{Tr}(|A|) \equiv \mathrm{Tr}\left(\sqrt{A^\dagger A}\right)$, $|A| := \sqrt{A^\dagger A}$, $p=1$; Frobenius norm $\|A\|_F := \sqrt{\mathrm{Tr}(A^\dagger A)}$, $p=2$; Hilbert-Schmidt norm $\|A\|_{HS} := \sqrt{\sum_{i,j} A_{ij}^2} =? \sum_{i\in I} \|Ae_i\|_H^2$; Hilbert-Schmidt inner product $\langle A,B\rangle_{\mathrm{HS}} := \mathrm{Tr}\left(A^\dagger B\right)$, Frobenius inner product $\langle A,B\rangle_{\mathrm{F}} := \mathrm{Tr}\left(A^\dagger B\right)$? (in finite-dimensionala Euclidean space, the HS norm is identical to the Frobenius norm) Although the Hilbert-Schmidt distance is arguably not too meaningful, operationally, one can use Cauchy-Schwarz to relate it to the very natural trace distance. shadow norm ...

**Definition 17** (distance)**.** For mixed states, trace distance $d_{\mathrm{tr}}(\rho,\rho') := \frac{1}{2}\|\rho-\rho'\|_1$. For pure states, $d_{\mathrm{tr}}(|\psi\rangle,|\psi'\rangle) := \frac{1}{2}\||\psi\rangle\langle\psi| - |\psi'\rangle\langle\psi'|\|_1 = \sqrt{1-|\langle\psi|\psi'\rangle|^2}$. fidelity and trace distance are related by the inequalities

$$1 - F \le D_{\mathrm{tr}}(\rho,\rho') \le \sqrt{1-F^2} \tag{A4}$$

variation distance of two distribution $d_{var}(p,p') := \frac{1}{2}\sum_i |p_i - p_i'| = \frac{1}{2}\|p-p'\|_1$. $l_2$ distance ... Hellinger distance ... HS distance $D_{\mathrm{HS}}(\rho,\rho') := \|\rho-\rho'\|_{\mathrm{HS}} = \sqrt{\mathrm{Tr}((\rho-\rho')^2)}$

**Definition 18** (stabilizer)**.** An observable $S_k$ is a stabilizing operator of an $n$-qubit state $|\psi\rangle$ if the state $|\psi\rangle$ is an eigenstate of $S_k$ with eigenvalue 1, A stabilizer set $S = \{S_1,\ldots,S_n\}$ consisting of $n$ mutually commuting and independent stabilizer operators is called the set of stabilizer "generators".

Many highly entangled $n$-qubit states can be uniquely defined by $n$ stabilizing operators which are locally measurable, i.e., they are products of Pauli matrices. A stabilizer $S_i$ is an $n$-fold tensor product of $n$ operators chosen from the one qubit Pauli operators $\{\mathbb{1}, X, Y, Z\}$.

**Example 2** (GHZ)**.** For GHZ state: $|\mathrm{GHZ}\rangle := \frac{1}{\sqrt{2}}(|0\rangle^{\otimes n} + |1\rangle^{\otimes n})$, the projector based witness

$$W_{\mathrm{GHZ}_3} = \frac{1}{2}\mathbb{1} - |\mathrm{GHZ}\rangle\langle\mathrm{GHZ}| \tag{A5}$$

requires four measurement settings. For three-qubit GHZ state [25], the local measurement witness

$$W_{\mathrm{GHZ}_3} := \frac{3}{2}\mathbb{1} - X^{(1)}X^{(2)}X^{(3)} - \frac{1}{2}\left(Z^{(1)}Z^{(2)} + Z^{(2)}Z^{(3)} + Z^{(1)}Z^{(3)}\right) \tag{A6}$$

This witness requires the measurement of the $\left\{\hat{\sigma}_x^{(1)},\hat{\sigma}_x^{(2)},\hat{\sigma}_x^{(3)}\right\}$ and $\left\{\hat{\sigma}_z^{(1)},\hat{\sigma}_z^{(2)},\hat{\sigma}_z^{(3)}\right\}$ settings. For $n$-qubit case, detect genuine $n$-qubit entanglement close to $\mathrm{GHZ}_n$

$$W_{\mathrm{GHZ}_n} = (n-1)\mathbb{1} - \sum_{k=1}^{n} S_k(\mathrm{GHZ}_n) \tag{A7}$$

where $\hat{S}_k$ is the stabilizer ... [26]

**Definition 19** (cluster state)**.** 1D four qubits

$$\left|\psi_4^{1D}\right\rangle = \frac{1}{2}(|{+}0{+}\rangle + |{+}0{-}\rangle + |{-}1{+}\rangle - |{-}1{-}\rangle) \tag{A8}$$

The entanglement in a graph state is related to the topology of its underlying graph [65].

**Remark 1.** LU, LC equivalence, local operations and classical communication (LOCC),

**Definition 20** (graph state)**.** Given a simple graph (undirected, unweighted, no loop and multiple edge) $G=(V,E)$, a graph state is constructed as from the initial state $|+\rangle^{\otimes n}$ corresponding to $n$ vertices. Then, apply controlled-Z gate to every edge, that is $|G\rangle := \prod_{(i,j)\in E} \mathsf{cZ}_{(i,j)} |+\rangle^{\otimes n}$ with $|+\rangle := (|0\rangle + |1\rangle)/\sqrt{2}$.

| | $\|\mathrm{GHZ}_3\rangle$ | $\|W_3\rangle$ | $\|CL_3\rangle$ | $\|\psi_2\rangle$ | $\|\mathcal{D}_{2,4}\rangle$ | $\|\mathrm{GHZ}_n\rangle$ | $\|W_n\rangle$ | $\|G_n\rangle$ |
|---|---|---|---|---|---|---|---|---|
| maximal overlap $\alpha$ | 1/2 | 2/3 | 1/2 | 3/4 | 2/3 | 1/2 | $(n-1)/n$ | 1/2 |
| maximal $p_{\text{noise}}$ | 4/7 | 8/21 | 8/15 | 4/15 | 16/45 | $1/2\cdot(1-1/2^n)^{-1}$ | $1/n\cdot(1-1/2^n)^{-1}$ | $1/2\cdot(1-1/2^n)^{-1}$ |
| # local measurements | 4 | 5 | 9 | 15 | 21 | $n+1$ | $2n-1$ | depend on graphs |

TABLE III: Results on local decompositions of different entanglement witnesses for different states. [8]

## Appendix B: Machine learning background

Notations: The (classical) training data (for supervised learning) is a set of $m$ data points $\left\{(\mathbf{x}^{(i)}, y^{(i)})\right\}_{i=1}^{m}$ where each data point is a pair $(\mathbf{x}, y)$. Normally, the input (e.g., an image) $\mathbf{x} := (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ is a vector where $d$ is the number of *features* and its *label* $y \in \Sigma$ is a scalar with some discrete set $\Sigma$ of alphabet/categories. For simplicity and the purpose of this paper, we assume $\Sigma = \{-1, 1\}$ (binary classification).

### 1. Support vector machine

SVM is a typical supervised learning algorithm for classification. Taking the example of classifying cat/dog images, supervised learning means we are given a dataset in which every image is labeled either a cat or a dog such that we can find a function classifying new images with high accuracy. More precisely, the training dataset is a set of pairs of features X and their labels y. In the image classification case, features are obtained by transforming all pixels of an image into a vector. In SVM, we want to find a linear function, that is a hyperplane which separates cat data from dog data. So, the prediction label is given by the sign of the inner product (projection) of the hyperplane and the feature vector. We can observe that the problem setting of image classification by SVM is quite analogous to entanglement detection, where input data are quantum states now and the labels are either entangled or separable.

**Definition 21** (SVM)**.** Given a set of (binary) labeled data, support vector machine (SVM) is designed to find a hyperplane (a linear function) such that maximize the margin between two partitions...

$$\max_{\mathbf{w}} \|\mathbf{w}\|^2 \text{ s.t. } \forall i, y^{(i)} \cdot (\mathbf{w} \cdot \mathbf{x} + b) \geq 1. \tag{B1}$$

Lagrange multipliers $\alpha$

$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i^m \alpha^{(i)} \left(\mathbf{w} \cdot \mathbf{x}^{(i)} + b\right) + \sum_i^m \alpha^{(i)} \tag{B2}$$

#### a. *kernel method*

However, note that SVM is only a linear classifier. while most real-world data, such as cat/dog images and entangled/separable quantum states are not linearly separable. For example, with this two dimension dataset, we are unable to find a hyperplane to separate red points from the purple points very well. Fortunately, there is a very useful tool called kernel method or kernel trick to remedy this drawback. The main idea is mapping the features to a higher dimensional space such that they can be linearly separated in the high dimensional feature space. Just like this example, two dimensional data are mapped to the three dimensional space. Now, we can easily find the separating plane. With SVM and kernel methods, we expect to find a generic and flexible way for entanglement detection. kernel

**Definition 22** (kernel)**.** In general, the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ measures the similarity between two input data points by an inner product

$$k(\mathbf{x}, \mathbf{x}') := \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \tag{B3}$$

If the input $\mathbf{x} \in \mathbb{R}^d$ (conventional machine learning task, e.g., image classification), the feature map $\phi(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^n$ $(d < n)$ from a low dimensional space to a higher dimensional space. The corresponding kernel (Gram) matrix $\mathbf{K}$ is PSD.

**Example 3** (kernels)**.** Some common kernels: the polynomial kernel $k_{\text{poly}}(\mathbf{x}, \mathbf{x}') := (1 + \mathbf{x} \cdot \mathbf{x}')^q$ with feature map $\phi(\mathbf{x})$ ... The Gaussian kernel $k_{\text{gaus}}(\mathbf{x}, \mathbf{x}') := \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|_2^2\right)$ with an infinite dimensional feature map $\phi(\mathbf{x})$. An important feature of kernel method is that kernels can be computed efficiently without evaluating feature map (might be infinite dimension) explicitly.

**Definition 23** (graph kernel)**.** given a pair of graphs $(G, G')$, *graph kernel* is $k(G, G') =$. quantum graph kernel $k(G, G') = |\langle G|G'\rangle|^2$ ?? [66]

### b. *quantum kernel*

**Definition 24** (quantum kernel)**.** quantum kernel with quantum feature map $\phi(\mathbf{x}) : \mathcal{X} \to |\phi(\mathbf{x})\rangle\langle\phi(\mathbf{x})|$

$$k_Q(\rho, \rho') := |\langle\phi(\mathbf{x})|\phi(\mathbf{x}')\rangle|^2 = \left|\langle 0|U^\dagger_{\phi(\mathbf{x})}U_{\phi(\mathbf{x}')}|0\rangle\right|^2 =? \operatorname{Tr}(\rho\rho') \equiv \langle\rho, \rho'\rangle_{\text{HS}} \tag{B4}$$

where $U_{\phi(\mathbf{x})}$ is a quantum circuit or physics process that encoding an input $\mathbf{x}$. In quantum physics, quantum kernel is also known as transition amplitude (quantum propagator);

### c. *neural network and kernel*

**Definition 25** (neural tagent kernel)**.** neural tangent kernel [41]: proved to be equivalent to deep neural network [55] in the limit ...

$$k_{\text{NT}}\left(S_T(\rho_l), \tilde{S}_T(\rho_{l'})\right) = \left\langle \phi^{(\text{NT})}(S_T(\rho_l)), \phi^{(\text{NT})}(\tilde{S}_T(\rho_l))\right\rangle \tag{B5}$$